Digital transformation and business predictions

Lebefromm, Uwe

Doctoral thesis / Disertacija

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Economics and Business / Sveučilište u Rijeci, Ekonomski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:192:897644

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

SVEUČILIŠTE U RIJECI
EKONOMSKI FAKULTET

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Economics and Business - FECRI Repository

UNIRI DIGITALNA KNJIŽNICA

Image not found or type unknown

UNIVERSITY OF RIJEKA
FACULTY OF ECONOMICS AND BUSINESS

Uwe Lebefromm

DIGITAL TRANSFORMATION AND BUSINESS PREDICTIONS

DOCTORAL THESIS

Rijeka, 2021

UNIVERSITY OF RIJEKA
FACULTY OF ECONOMICS AND BUSINESS

Uwe Lebefromm

DIGITAL TRANSFORMATION AND BUSINESS PREDICTIONS

DOCTORAL THESIS

Mentor: Prof. Neda Vitezić, PhD

Rijeka, 2021

SVEUČILIŠTE U RIJECI
EKONOMSKI FAKULTET

Uwe Lebefromm

DIGITALNA TRANSFORMACIJA I POSLOVNO PREDVIĐANJE

DOKTORSKI RAD

Rijeka, 2021.

Supervisor: Prof. Neda Vitezić, PhD

The doctoral thesis has been defended on Monday, October 18th, 2021, at the University of Rijeka, Faculty of Economics and Business, before the Committee of the following members:

1. Full Professor Marija Kaštelan Mrak, PhD, President of the Board, Faculty of Economics and Business, University of Rijeka, Croatia,
2. Full Professor Mojca Marc, PhD, University of Ljubljana, School of Economics and Business, member,
3. Full Professor Mira Dimitrić, PhD, member.

Thesis Acknowledgments

During the preparation of the thesis, I received excellent support from my mentor Prof. Neda Vitezić, PhD. With her expertise, it was possible to identify and implement the scientific requirements for this thesis. The expertise of my mentor was invaluable in the formatting of the thesis topic and methodology in particular. The organizational process of this thesis was planned and implemented with her support. The excellent contacts of my mentor to the marina industry in Croatia made it possible to give this thesis a solid empirical basis. I owe the success of this thesis to the inexhaustible scientific spirit of my mentor.

I would like to acknowledge the Dean of the Faculty of Business and Economics of the University of Rijeka, Prof. Alan Host, PhD. With the conclusion of a contract for scientific cooperation with the marina company Adriatic Croatia International Club, the data that are so important for an empirical thesis were made available.

I would also like to acknowledge my language editor, Ms. Denisse Mandekić, who checked and corrected the English grammar of my thesis.

I would also like to acknowledge Ms. Dolores Kraljić Načinović, Expert Associate for Controlling and Ervin Pliskovac, Head of Information Technology and Security in the company Adriatic Croatia International Club (ACI). They were very supportive and provided the available empirical data. They also organized interviews during the processing of the case studies at ACI.

I would also like to thank my colleagues and my manager, Christoph Rendel, Head of SAP Product Learning Center of Excellence Finance and & Industry at SAP Germany. In countless situations in my daily work, my special situation due to the preparation of a scientific paper was considered and they made the effort to compensate for it.

Last, but not least, I would like to thank my wife Elzbieta. Over a long period of time, she understood that I was very busy, had a lot of patience, and always encouraged me.

SUMMARY

The digital transformation is increasingly determining everyday professional life. The increasing digitization of operational processes is taking place across all industries. The implementation of data digitization and company networking offers the possibility of optimizing operational processes using this technology. Technical implementation can increase the operational success of a company if it results in better decisions of the controlling and management. Such data-driven decision-making processes are the core of this thesis. First of all, this process includes the transformation of company data into decision-relevant information. This doctoral dissertation demonstrates that this is possible by using machine learning to develop prediction models to provide reliable predictions about future customer behavior. Predictions that allow a validated division of customers into customer groups enable target group-oriented customer relationship management. For this purpose, the explanatory variables were identified in the prediction models which have a significant influence on the target variable set for the respective model. In the next step, it could be proven that the consideration of the forecast models considerably improves the decisions of the controlling. This has set the basis for the development of stochastic decision models based on data science in this thesis. The marina industry in Croatia offered a suitable scenario as there is intense national and international competition in this industry. Decisions on berth allocation, improved predictability of expected revenues and better knowledge of who the company's customers really are will significantly increase the company's competitiveness.

Keywords: Machine Learning, Predictive Analytics, Controlling Decision-Making Process, Marina Industry

Contents

1 INTRODUCTION 1

1.1 Problem and Subject of the Thesis 1

1.2 Scientific Hypothesis 5

1.3 Objective of the Thesis 7

1.4 Thesis Design 8

1.5 Scientific Thesis Review 10

1.6 Methods 14

1.7 Thesis Structure 16

2 THE IMPACT OF DIGITIZATION ON CONTROLLING 18

2.1 Challenge to Controller 19

2.1.1 Controller as Service Provider for the Management 20

2.1.2 The Organizational Position of the Controller 21

2.2 Digitalization Concepts 23

2.2.1 From a Business Model to a Digital Business Model 23

2.2.2 The Impact of Digitalization on the Controlling Process Model 28

2.3 Controllers' Information Technology in the Digital Age 31

2.3.1 Digital Transformation in Controllers' Information Technology 32

2.3.2 Real-Time Business Analytics 34

2.4 Predictive analytics 36

2.4.1 Learning from Historical Data 37

2.4.2 The Rectangle Method 42

2.4.3 Application of Predictive Analytics 44

3 THEORIES SUPPORTING PREDICTION AND DECISION-MAKING 46

3.1 Support Vector Machine for Classification and Regression 46

3.2 Decision Theory 56

3.2.1 Probability Theory Based on Bayes' Theorem 57
3.2.2 Information Status and Future Scenario 60
3.2.3 Scenario Probability According to the Bayes' Theorem 62
4 EMPIRICAL RESEARCH – PREDICTIVE ANALYTICS 65
4.1 Requirements for Empirical Research 65
4.2 Creation and Validation of the Model 70
4.3 The Learning Process in Predictive Analytics 74
4.4 Description of the Prediction Models 76
4.4.1 Model A – Client Type 'PRIVATE' with a Yearly Contract 77
4.4.2 Model B – Client Type 'FIRM' with a Yearly Contract 86
4.4.3 Model C – Contract Renewal 92
4.5 Prediction with Regression Analysis 98
4.5.1 Model D – Late Payment 99
4.5.2 Model E – Contract Amount 108
4.5.3 Model F – Customer Classification 115
5 THE INFLUENCE OF PREDICTIVE MODELS ON DECISION BEHAVIOR . 125
5.1 Expert Survey with the Presentation of Case Studies 126
5.2 Case Study Solutions 128
a) Solution Case Study A – Yearly Contract with a Private Client 128
b) Solution Case Study B – Yearly Contract with a Firm 129
c) Solution Case Study C – Contract Renewal Private Client 130
d) Solution Case Study D – Poor Payment Behavior 131
e) Solution Case Study E – Contract Amount 132
f) Solution Case Study F – Customer Group 134
5.3 Evaluation of the Hit Rate of the Participants 135

6 STOCHASTIC DECISION MODELS BASED ON DATA SCIENCE FOR THE MARINA INDUSTRY IN CROATIA 139

a) Sub-Decision Model A – Private Client with a Yearly Contract 140

b) Sub-Decision Model B – Business Customer with a Yearly Contract 143

c) Sub-Decision Model C – Private Client with Contract Renewal 145

d) Decision Model ABC – Marina Industry – Berth Allocation 148

e) Decision Model D – Marina Industry – Payment Behavior 150

f) Decision Model E – Vessel Length and Contract Amount 153

g) Decision Model F – Customer Classification 155

7 PROOF OF HYPOTHESIS AND CONTRIBUTION OF THE RESEARCH 159

7.1 Proof of Hypothesis 159

7.2 Contribution of the Research 163

8 CONCLUSION AND OUTLOOK 165

Scientific Apparatus 167

List of Figures 167

List of Tables 169

List of Abbreviations 170

Reference List 172

1 INTRODUCTION

This thesis was created to analyze the possibilities of the digital transformation for decision-making processes in the economy and the business of the marina industry in Croatia and whether it is suitable and needed for the optimization of these processes. In this thesis it could be proven that the possibilities of data science resulting from the digital transformation do not only significantly improve the level of information of controlling and management (Artamonow, M., 2017, page 35); (Gleich, R. et al., 2018); (Gleich, R. et al., 2014); (Langmann, Chr., 2019); (Nasca, D. et. al., 2018, page 73); (Roßmeisl, E. et. al., 2014), but also lead to a changed decision-making behavior (Busemeyer, J. R. et al., 1993, pages 432–459); (Dörsam, P., 2013, page 43). It also has an impact on controlling and decision making process in controlling (Gänßlein, S., 2017, page 21); (Jeschke, B. G., 2017, page 57); (Möller, K., 2017); (Steinke, K.-H. et al., 2017); (Tschandl, M. et al., 2017, pages 20–23.); (Tschand., M. et al., 2018, pages 27–48). The influence of the digital transformation leads to completely new innovative business models that represent a new competitive situation compared to the established business models (Bloomberg, J., 2018); (Gassmann, O. et al., 2019 page 28); (Handelsblatt, 2019); (Hoffmeister, C., 2015); (Progroscewska, P. J., 2016); (Spur, G. et al., 1997, page 577). This research does not aim to develop new business models, but to place existing business models in a new database. The business model of renting berths in the marina industry in Croatia is an essential economic factor. A company in this branch is in the national but also international competition. The optimization of the efficiency of decision-making processes which leads to a sustainable business with long-term customer relationships and secured income is not only in the interest of companies but also of the government.

1.1 Problem and Subject of the Thesis

Production networking on a global level and the worldwide availability of business data, which can be processed and made available at an unprecedented processing speed, have internationalized the competition among companies. Companies are facing the challenge of keeping up with the rapid technological development in all technical areas. Operational decisions must be made quickly. Tactical decisions should find the best solution from several alternatives. Strategic decisions are to be made to secure the long-term existence and growth of the company. At all decision-making levels, the decisions must be made in an understandable and substantiated way. This requirement for decision quality can only be met if all decision-

relevant information are available in a short time. This does not only affect statistical data, but also related forecast data. The development in information technology has made it possible to calculate prediction models based on mathematical-statistical methods in machine learning times over a period of a minutes. Such a system, SAP PREDICTIVE ANALYTICS®, is used in this thesis to gain reliable knowledge from the available data using the statistical methods of clustering, regression analysis and classification. The issue of who makes the decisions in the company is controversial. The KPMG[1] team examined the future of the Chief Financial Officer (CFO). Four extreme scenarios were designed (Weber, J., 2020, page 23) for the year 2040.

Scenario I: R.I.P. CFO (Rest in Peace). There will be no independent finance department and no CFO in the future. Artificial Intelligence provides the decision-relevant data. The decisions are made exclusively by the Chief Executive Officer (CEO).

Scenario II: Robo CFO scenario. The analysis refers to robotic process automation (RPA). "*The so-called software robot mimics the behavior of a human user. The software bots can, e.g., interact with large ERP systems such as SAP and Oracle or only smaller applications such as the Microsoft Office suite*" (Safar, M., 2020). The employee's job is to work with the self-programming algorithms of the RPA in the financial area with data. The decisions are made by RPA (Weber, J., 2020, page 24).

Scenario III: Chief Freelance Officer. The companies work with a pure project organization. External employees are commissioned if necessary. Chief Finance Officer becomes Chief Freelance Officer. The CFO coordinates the projects and hires freelancers. KPMG analysts speak of a fluid organization in which the freelancers are given decision-making powers (Weber, J., 2020, page 24).

Scenario IV: Chief Experience Officer (CXO). The Chief Financial Officer becomes the Chief Experience Officer. The CXO collects experience data, evaluates, and passes them on to other company departments, for example marketing. KPMG analysts say that the CFO (CXO) replaces the CEO. In an empirical thesis, the analysts found that this fourth scenario, at 31%, is the most likely scenario (Weber, J., 2020, page 25–26).

In this thesis there are similarities to scenario IV (CXO). However, the experience data from learning algorithms based on the statistical learning theory are used to generate prediction models. The prediction models are used as the basis for making controlling and management

[1] The name KPMG emerged from the merger of Peat Marwick International (PMI) and Klynveld Main Goerdeler (KMG).

decisions regarding the applicants for a berth i.e., which applicants should be awarded. Another decision to be made is which of the existing customers should be awarded contract renewal. The decision models developed in this thesis define the decision-making process for business decisions. Digitization is therefore not a matter of replacing decision-makers with computer programs. However, customer data evaluation results can be used to create customer-oriented offers and customer-oriented marketing campaigns. If the controlling staff can use the computer applications in such a way that reliable predictions are generated and forward-looking decisions are made, then the controller will develop into a digital leader. Therefore, the author of this thesis agrees with the statement made by the analysts at KPMG that a digital leader creates trust in the digital transformation within their company (Weber, J., 2020, page 27). Controlling is transforming into a competence center for future-oriented analyses and, due to the short processing time of the SAP Predictive Analytics computer application, can guarantee a quick data update in the learning phase. This is the basis for agile and reliable controlling.

The area of this thesis is the effects of digital transformation on corporate management and the associated business processes of corporate planning, forecasting and the resulting decision-making processes. The theory that relates to this effect is the decision theory about the formation of a probability judgment in correlation with information acquisition and information evaluation in the decision-making process. In the case of risk, information are used as the basis for an initial probability judgment, which is revised to a new probability judgment after the acquisition of further information. This decision theory has its origin in the Bayes theorem that describes that the revision of the probability judgment based on further information should be done rationally. The Bayes theorem used for a-posteriori decision is decribed in the following literature: (Bamberg, G. et al., 2012, page 136); (Bayes, Th., 1763); (Bayes, 2008, in: Timmerding, H. E. (Ed.), 1908); (Beck, H., 2014); (Daper, D., 2005); (Feindt, M. et al., 2015, page 59); (Held, L., 2008); (Laux, H. et al., 2014); (Laux, H. et al. 2018); (Moris, D., 2017); (Stiegler, St., M., 1983 and 2018); (Gillenkirsch, R., 2018).

The business case used in this thesis is the decision-making behavior of a company in the marina industry in Croatia. The possibilities of data science associated with the digital transformation could be used to develop meaningful prediction models as a decision-making basis. In this thesis, 38,000 data sets by the marina company were used, which made it possible to develop prediction models with optimal predictive power and prediction confidence. While the marina company's 38,000 records were used for finding explanatory variables, target variables were set by the author of this thesis to develop the predictive models. The author of this thesis

collected the target variables from the employees of the marina company through expert interviews. Quantitative thesis to create prediction models and qualitative thesis to gain new knowledge about the decision-making behavior in controlling have also been carried out by the author of this thesis. The theory which is the basis for quantitative thesis has its origins in the statistical learning theory of Vladimir N. Vapnik and A. Ya. Chevronenkis (Vapnik, V. N., 1988), (Vapnik, V. N., Chevronenkis, A. Ya., 1991, pages 284–305), (Vapnik, V. N., reprint 2018). The theory on which the quality thesis depends has its origin in the Bayes theorem about a-posteriori decisions (Bayes, T., 1763). The methodology used for the expert interviews is a method of empirical thesis. The interviews were carried out as guided interviews using the case thesis method. References to the interview method: (Stier, W., 2018, page 184); (Schnell, R. et al., 2018). Reference to the case thesis methodology: (Borschardt, A. et al., 2007).

The mathematical method for the implementation of the statistical learning theory in machine learning takes place via the support vector machine. The origin is the Bayes classifier. The support vector machine has become the thesis field of machine learning because of its excellent performance. When creating the prediction models in this thesis, the learning time duration was about five minutes. References to the Bayes classifier and support vector machine: (Fischer, J., 2007); (Dianzi, K. et al., 2011, pages 2–10); (Krzanowski, W. J., Hand, D. J., 2009); (Russel, St., Norvig, P., 2012); (Kuhn, M., Kjell, J., 2013); (Kellerher, J. D., 2015); (Hurwitz, J., Kirsch, D., 2018); (Runkler, A., 2015).

The mathematical-statistical methods, which are the basis of the prediction models in this thesis are clustering, regression and classification (Auer von, L., 2014, page 19), (Georgi, H.-O., 2015, page 353). The information technology application used to create prediction models based on the learning machine is SAP Predictive Analytics®. Knowledge on how to use the system was taken from: (Bakhshaliyeva, N. et al, 1017, pages 257–286); (Butsmann, J. et al., 2019, page 373–396); (Charbert, A. et al., 2017); (Horváth, P. et al., 2015, page 363); (Kießwetter, M. et al., 2007); (Mindsquare, 2019). Regression analysis of the continuous explanatory variables was used. Clustering and classification have been selected as methodology for discrete explanatory variables (Stier, W., 1999); (Hyndman, R. J.; Athanasopoulos, G., 2014); (Auer von, L., 2014); (Lawrence, K. D.; Klimberg, R. K., 2018).

New knowledge has been gained from the investigation of decision-making behavior using the results of prediction models. This knowledge has been used to develop stochastic decision models based on data science (DDM). The results of the doctoral thesis follow the topic of

“Digital Transformation and Business Forecasting“ accepted by the Committee. As a result, however, the thesis shows that business forecasting has evolved into business prediction. With the result of this doctoral thesis in the form of stochastic decision models based on data science, a method was developed to operationalize and automate operational and tactical decisions. Through the developed methods for converting manual decision-making processes into optionally automatable decision-making processes, this doctoral thesis provides a contribution to the innovation in controlling in the digital age. The business case used is a marina company in Croatia. The generic and methodical approach offers the possibility of applying this method to other industries. It requires other industry-specific and company-specific target variables and explanatory variables. The possible automation of the decision-making processes by using machine learning as a decision factor shortens the decison-making processes and promotes the agility and flexibility of the company. The results of machine learning with high value of predictive power and prediction confidence tell the controller how to decide. However, there are also risks associated with the digital transformation of business processes. Classic barriers to market entry, such as high capital expenditure and know-how, are softened or eliminated by the possibilities to act digitally. A business model such as: “Rent and rent out berths digitally“ is just one example of how competition can arise from digital invaders. The search for constantly new data sources must support the requirement to be able to react quickly to changed customer behavior with one's own range of services. One of the examples is the trend towards charter boats. The results of this doctoral thesis can be used as a basis for the introduction of new decision-making processes that no longer reflect traditional styles and for their organization as a cross-sectional function across all company departments (Hölscher, B., 2017, page 115).

1.2 Scientific Hypothesis

The **main hypothesis** of the doctoral thesis is the following:

Main Hypothesis H = By using predictive analytics and theory of probability it is possible to develop stochastic decision models based on data science for effective decision making.

Using a company-specific data set, this can be proven by meaningful results of machine learning models and their use in plausible and understandable decision models for the company and can be used as a template for the industry the company belongs to. This contributes to the decision theory about the formation of a probability judgment and the evaluation of information

in general and for the marina industry in particular. To prove the main hypothesis, several auxiliary hypotheses were developed. Five of these will be tested by quantitative methods: H1, H1.1, H1.2, H1.3 and H1.4. Hypothesis H2 will be proven by using qualitative methods.

Main Hypothesis H1 = Using a marina company data set, it is possible to build prediction models whose significance can be proven by high values (0.995 to 0.998) of prediction power and prediction confidence. It is hypothesized that such results can also be generated with customer data from the marina industry in Croatia.

Such hypothesis implies several **auxiliary hypotheses.**

Auxiliary Hypothesis H 1.1. = Adjusted prediction models provide reliable prediction for contracting models leading to long-time customer relationships with the company.

This is based on long-time contract models which are yearly contract and renewed contract. This hypothesis will be proven by establishing a list of theoretically based distinction features among business contracting models. Groups are distinguished by applying the mathematical-statistical method of clustering relating to the characteristics of the customers and applicants. This relates to the first three prediction models A, B and C.

Auxiliary Hypothesis H.1.2 = Adjusted prediction models provide reliable prediction about applicants who are likely to be late payers.

Late-payment customer groups are identified with days payable outstanding. Scoring has been designed to set the boarder for this customer group. Regarding the continuous variable “days payable outstanding”, the mathematical-statistical method of multiple linear regression is used. The accuracy of the model can be measured by the key performance indicators “predictive power” and “predictive confidence”.

Auxiliary Hypothesis H.1.3 = Adjusted prediction model provides reliable prediction about the correlation of sales with customers who have specific characteristics, which will lead to higher sales.

Regarding the usage of the continuous variable of sales, the mathematical-statistical method of linear multiple regression is used. Customer groups with higher sales can be identified using the decision-tree method.

Auxiliary Hypothesis H.1.4 = Adjusted prediction model provides reliable prediction about a group of customers and sales groups with a specific range of sales.

A specific customer characteristic can be found to create useful customer groups regarding the sales volume. Such customer groups could be selected for customer-group-specific marketing campaigns and therefore increase the efficiency of the campaign. With regard to this hypothesis, the mathematical-statistical method for clustering is used. The accuracy of the model can be measured by the key performance indicators specificity and sensitivity.

Main Hypothesis H2 = By gaining additional information, it is possible to have a more accurate prediction and effective decision-making process.

This can be proven by using the hypothesis-testing case thesis. Case studies relating to the hypotheses are presented to the previously surveyed decision-makers and after presentation of the prediction models results. The changed decision-making behavior can be proven by measuring the correct decisions compared to the known results of the prediction models.

1.3 Objective of the Thesis

The objective of this thesis was based on thesis questions that are answered by the thesis results. The digital transformation leads to a change in methodology and paradigm in the compilation of decision-making bases and the decision-making processes. The aim of this thesis was to find answers to two central questions:

Research question I

By using machine learning, is it possible to extract knowledge from data to generate decision-relevant information in the form of prediction models, which have a very high significance according to decision-making criteria of the examined marina company? Does the targeted use of machine learning lead to reliable and meaningful predictions with measurable accuracy?

Research question II

Can a stochastic decision model based on data science be reliable enough for the predictive purposes of controllers and decision-making process of the marina industry?

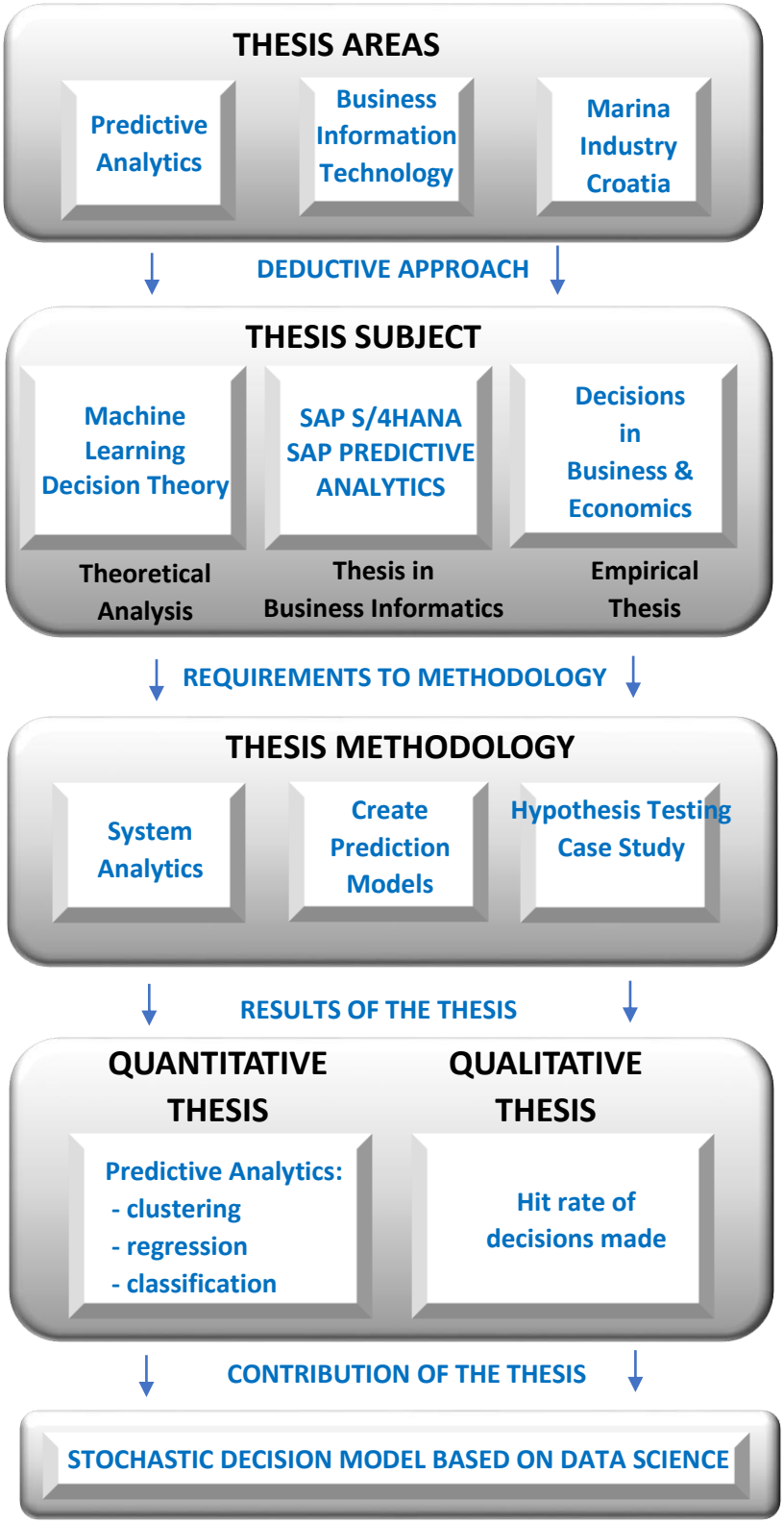
The **main goal** is to develop a method for implementing digital transformation in operational and tactical decision-making processes in controlling and management. The used business case is the marina industry in Croatia. The marina industry is characterized by complexity and agility (Gračan, D., 2016), (Kovačić, M., 2009), (Luković, T., 2013), (Peović, K., 2018). Mastering them requires the use of future-oriented technology and methodology in order to keep pace in

the age of digitization. This main goal is to be achieved by developing decision models that enable digitization of manual decision-making processes. The decision models essentially have the results of digital prediction models as the decision criterion. This reduces the duration of the decision-making process to a minimum and increases the agility of companies. Resilience is in vogue not only in times of crisis caused by the global outbreak of COVID-19 coronavirus but also to secure the competitiveness of a company in the age of digital transformation in the competition of digital business models (Schäffer, U., 2020); (Hoffmeister, C., 2015). The **first derived goal** is the creation of the prediction models. The scientific methods used to create the prediction models are mathematical and statistical methods of clustering, regression and classification. The statistical evaluation of the provided data is carried out using machine learning based on pattern recognition. The predictions are future customer behavior and the detection of customer groups as well as groups of applicants. For this purpose, the resilience of the prediction models must be proven using key figures for predictive power and prediction confidence. The **second derived objective** is the identification of realistic and representative target variables for the prediction models. These representative target variables are to be discovered based on the exemplary business case through expert interviews. The target variables should represent operational and tactical business relevance for the company. The target variables are the starting point for the development of the decision-making process and represent the respective decision criterion of the examined decision-making process. The **third derived goal** is to demonstrate that the knowledge of the results of the predictive models changes the decision-making behavior. The aim is to demonstrate significantly better decision-making. This proof is to be determined via the recorded hit rate for the alternatives correctly identified as TRUE.

1.4 Thesis Design

This thesis deals with the use of pattern recognition in business decisions. The theoretical framework is, therefore, data science and the associated statistically reliable prediction of future events. The future events in this thesis relate to the future actions of the customers of a marina company. Decisions by the company's controllers and managers are derived from the expected customer behavior to maximize the benefits resulting from the decisions.

Figure 1: Thesis Design



Source: Author

First of all, this thesis covers the following areas: The new possibilities of information technology to store data from external and internal accounting together and thus make reconciliation processes of the results in financial accounting and management accounting superfluous. Furthermore, the possibilities of information technology include the expansion of a pure reporting system for predictive analytics. The next step is the processing of the learning theory on which predictive analytics is based. This is implemented in predictive analytics with the algorithm of the support vector machine, which is explained in the thesis. The third pillar of the thesis is the decision theory on which business decisions are based. Thomas Bayes' decision theory is presented and applied in the context of this thesis. The proven higher correct-decision hit rate is the motivation for developing stochastic decision models. The application of the results from predictive forecasts in business decision-making processes is a new development in the marina industry and expands the current thesis.

1.5 Scientific Thesis Review

The thesis analyses the current state by exploring the usage of the Bayes' theorem to build models, conclusions, and predictions that can be found in many theses, papers, and publications. An example is given for predictive methods, decision support, and decision methods.

Aki Vehtari and Janne Ojanen investigated Bayesian predictive methods for model assessment, selection, and comparison. A central conclusion from the investigations is the following: If the goal is to estimate the predictive performance of a Bayesian model, Bayesian cross-validation should be used (Vehtari, A. et al., 2012, page 216). The Bavarian cross-validation divides the data and the associated explanatory variables into training data and validation data. Vehtari and Ojanen state that this method is known as an unbiased and asymptotically true estimate of generalization utility. The computer solution used in the context of this thesis to generate the prediction models works according to the same principle as the Bayesian cross-validation. When the model is generated, the system divides the data used into training data and validation data. The division is determined by the system. The performance indicators predictive power and prediction confidence are calculated based on validation.

G. Subbalakshmi, K. Ramesh and M. China Rao (Subbalakshmi, K. et al., 2011) developed in their thesis a decision model for decision support in the case of heart disease. The so-called Decision Support in Heart Disease Prediction System (DSIHDPS) is using the data mining modeling technique called Naïve Bayes. There are 15 explanatory variables in the model, like

age, sex, blood pressure, serum cholesterol, blood sugar, etc. The target variable is the likelihood of the patients getting heart disease. The decision support system is implemented as a web-based questionnaire application. Based on user answers, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database (Subbalakshmi, K. et al., 2011, page 171). According to the authors, the decision model can sustainably improve the quality of clinical decisions. The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naïve Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state (Subbalakshmi, K. et al., 2011, page 172)[2].

The arguments given by the authors of the thesis on why to use the Bayes' theorem are that Bayes' algorithm is used to create models with predictive capabilities and provide new ways of exploring and understanding data. According to the authors, Bayes should be used when the data volume is high, the attributes are independent and when a more efficient output is required as compared to other methods of output. The developed system is using the historical heart disease database. The system learns from the evidence by calculating the correlation between the target variable, which is the risk of heart disease, and the explanatory variables. In their conclusion, the authors mention that the model could be further enhanced and expanded. It could incorporate other data mining techniques, using categorical data, as well as continuous data.

The application of the Bayes' theorem takes place in this thesis in the analog form. On the one hand, Bayes is used in the generated models for clustering, regression, and classification. On the other hand, taking into account the results of the prediction model significantly improve the decisions of the surveyed controllers of the marina company. The current state of thesis for the usage of data science to create a decision model is presented in the following text:

Adele Diederich (Diederich, A., 1997) extended and generalized the so-called "decision-field theory" by Busermeyer and Townsend and developed the so-called "Multi Attributed Dynamic Decision Model (MADD)". Therefore, Diederich created information-processing models and explained the correlation of a less time-consuming decision-making process with its less accurate decision. Diederich defined the relationships between two decision alternatives and

[2] Subbalakshmi, K. et al., 2011, page 176.

the factors influencing the decisions in a two-dimensional transition matrix in the canonical form. The decision model described how the influencing factors are combined in the decision-making process to make the decision. The decision-maker starts the decision-making process with an initial decision. With the successive consideration of the influencing factors, the previous decision is reconsidered. If the decision-maker is under time pressure, his decision tends to be an initial decision. In the 1990s the world was still far from the possibilities of data science.

The presented thesis on digital transformation and business prediction uses the support vector machine algorithm to develop the prediction models. As in the MADD, the support vector machine follows the successive consideration of the available explanatory variables and checks the contribution of the variables to the target variables. However **–** and this is the difference to the age of digital transformation **–** the selection process takes place at machine speed and in n-dimensional transition matrix.

Murat Kucukvar, Mehdi Noori, Gokhan Egilmez, and Omer Tatari (Kucukvar, M. et al., 2014) developed the so-called "triple bottom line (TBL)" model to find the optimal allocation of different pavement types to fix the highways in the United States (U.S.). The explanatory variables of the decision scenario are stochastically determined data when using HMA (Hot Mixed Asphalt) versus WMA (Warm Mixed Asphalt). These explanatory variables relate to emissions into the environment, energy consumption, and water withdrawal. On the other hand, there are also socio-economic variables such as import, income, gross profit, taxes, employment, and gross operating surplus. The result of their thesis is a programmed model which, based on stochastic results, creates the compromise of deciding between the use of WMA and HMA for the US highways and therefore a compromise between a business decision and a sustainable decision. The stochastic data that have been used were databases from the BEA – Bureau of Economic Analysis 2002, GFN – Global Footprint Network, EIA – Energy Information Administration 2011, and BLS – Bureau of Labor Statistics 2002. The developed programmed model is an example of multi-criteria decision-making using stochastic data.

Swee S. Kuik, Toshiya Kaihara and Nobutada Fujii (Kuik, Sw. S. et al., 2015) developed a stochastic decision model of a remanufactured product with warranty. The decision scenario is the warranty policy. The decision model examines statistical data to calculate the cost-benefit ratio for the manufacturer of products requiring warranty services. The background of the decision scenario is the relationship between prices for guarantee contracts with customers on

the one hand and guarantee costs for the manufacturer of the products on the other. If the prices for guarantee contracts are too high, this weakens the purchasing power of the customers and the competitiveness of the manufacturers if the demand for the products decreases. If prices are too low, the company realizes losses through warranty services, which also weakens the competitiveness of these companies. The decision model examines the development of the guarantee costs for two types of guarantee contracts. Type I contract provides for a time-related guarantee. Type II, on the other hand, provides for an agreement based on the number of guarantee cases. The explanatory variables of the model are the development of the guarantee costs compared to the income from the guarantee contracts. The target variable is the resulting profit or loss. The calculations based on the stochastic model have shown that the manufacturers achieve a better result with the type I guarantee and therefore, they must decide in favor of the type I guarantee.

The quality of the forecasting models and the proof of significantly better decisions when considering the results of prediction were used in this thesis as a basis to develop decision models for the controlling of the marina company. The idea of a prediction-based decision model is not new.[3] Dr Livia Maglić, Assistant Professor at the Faculty of Maritime Studies at the University of Rijeka, published in April 2019 a decision model to find potential locations for the construction of marinas in the County of Primorje-Gorski-Kotar in the northwest of Croatia. The method used is Promethee and Gaia, which helps the decision-makers to better understand the available choices and the necessary compromises. The thesis by the author of this investigation did not reveal any evidence of an available data-science-based decision model for the marina industry. In general, this refers to the marina industry and it particularly refers to the marina industry in Croatia. This thesis, therefore, makes the following contribution: The demonstration of the of a-posteriori decisions theory made according to Bayes. With the use of machine learning, stochastic decision models were developed that support and improve the decisions of controlling and management. Routine decisions could be implemented by using

[3] Price, R. K., Vojinovic, Z.: Urban Hydro-Informatics: Data, Models and Decision Support for Integrated Urban Water Management. IWA publishing, 2011.

Malic, Livia, Varaždinac, Patricija, Škiljan, Ivona: Multi-Criterion Decision Model for Marina Location Selection in the County of Primorje-Gorski-Kotar. Published in Thesis Gate: https://www.thesisgate.net/publication/331624993

these decision models as information applications in an automated manner. These stochastic decision models represent a contribution to the digitalization of controlling.

1.6 Methods

Both quantitative and qualitative thesis methods are used in this thesis. Quantitative thesis means the evaluation of the data provided by the marina company and, based on this data set, the development of prediction models. The prediction models are based on the statistical methods of clustering, classification, and regression. Qualitative thesis method has been used in this thesis by defining case studies to examine business decisions about the case studies with and without knowledge of the results of the predictive models. The inductive procedure implies the following sequence:

The evaluation of the observed values will be initiated based on statistical data of a marina company used as training and test data to create prediction models. For this purpose, target variables are defined and the correlation of explanatory variables to the target variable is calculated. The so-called “predictive power” calculates the information content of the explanatory variables in the target variable. Variables with less significant influence are extracted. The prediction models calculated based on the training data are validated on the basis of test data, of which the results for the target variables are known. The validation leads to the calculated sensitivity and specificity. The sensitivity demonstrates the number of TRUE observations that have been detected as TRUE. The specificity demonstrates the number of FALSE observations that have been detected as FALSE. It is hypothesized that the patterns recognized in the training data also apply to new data. The so-called “predictive confidence” calculates the probability that the same good results of the evaluation of the prediction model with the test data can be expected with new data. The theory is that the behavior of the clients in the past will be the same in the future. The qualitative research in this thesis was carried out using an expert interview. The experts are the employees of the marina company with specific knowledge of the marina industry and the marina company in which they are employed in controlling and management. Case studies were developed for the interview created by experts. The description below introduces the decision-making process of the controllers according to the following scenario:

$$S_1: Customer\ is\ solvent\ -\ I_1: Prediction\ that\ the\ customer\ is\ solvent.$$

$$S_2: Customer\ is\ not\ solvent\ -\ I_2: Prediction\ that\ the\ customer\ is\ insolvent.$$

Based on the previous information status, the controlling department makes the following decisions:

The controller thinks, with 80% probability, that a specific customer is solvent:

$$p\ (S_1) = 0.8$$

The controller thinks, with 20% probability, that a specific customer is insolvent.

$$p\ (S_2) = 0.2$$

Prediction:

The probability that the prediction for S_1 is correct is about 100%

Predicted: I_1, actually: S_1. $p\ (I_1/S_1) = 100\%$

The probability that the prediction for S_1 is not correct is about 0%.

Predicted: I_2, actually: S_1: $p\ (I_2/S_1) = 0\%$

The probability that the prediction for S_2 is correct is about 70%

Predicted: I_2, actually: S_2. $p\ (I_2/S_2) = 70\%$

The probability that the prediction for S_2 is not correct is about 30%.

Predicted: I_1, actually: S_2: $p\ (I_1/S_2) = 30\%$

The calculation of a-posteriori decision probabilities results in:

decision solvent and actually solvent:

$$p(I_1/S_1) = \frac{1.0 * 0.8}{1.0 * 0.8 + 0.3 * 0.2} = 0{,}93$$

decision solvent and actually insolvent:

$$p(I_1/S_2) = \frac{0.3 * 0.2}{1.0 * 0.8 + 0.3 * 0.2} = 0{,}07$$

The probability of a correct a-posteriori decision is significantly higher than a-priori:

a-priori: 0.8 a-posteriori: 0.93.

Insolvent and actually solvent decision:

$$p(I_2/S_1) = \frac{0 * 0.8}{0 * 0.8 + 0.7 * 0.2} = 0$$

Insolvent and actually insolvent decision:

$$p(I_2/S_2) = \frac{0.7 * 0.2}{0 * 0.8 + 0.7 * 0.2} = 1.0$$

1.7 Thesis Structure

The technological framework in this thesis is defined by the computer application SAP Predictive Analytics as part of the innovative SAP S / 4 HANA solution released by the world market leader SAP in 2015. SAP S/4HANA was designed as a platform for digital transformation. All innovative computer application solutions that support the digitization of a company are summarized under the product term SAP LEONARDO. This includes solutions for the Internet of Things (IoT) and the associated possibility of connecting a wide variety of physical and virtual objects. This also includes machine learning with SAP Predictive Analytics and the associated possibility of providing a prediction service with the Clustering Service and Classification Service, with which robust forecast models can be generated and reliable forecasts about future developments can be calculated. The machine learning solution based on artificial intelligence, SAP Predictive Analytics, is used in this thesis. All the prediction models presented in this thesis were developed using SAP Predictive Analytics. The data used in this thesis do not come from the operational application of accounting and controlling, but from customer data of a marina company in Croatia. Therefore, the concepts of accounting and controlling in the SAP S / 4HANA system are not discussed in this thesis; it rather explains the application solution SAP Predictive Analytics.

The business framework in this thesis is defined by the controlling of a service company and the decisions to be made by controlling and management about customer contracts and customer relationship management. Therefore, this thesis examined how the role of controlling

changes in the age of digital transformation. It is explained that business models are transforming into digital business models. The role of information technology in the business processes of controlling was examined for this pupose. The focus is on the collection and analysis of decision-relevant data. Machine learning and the underlying mathematical-statistical theory of structured risk minimization form the theoretical framework for this thesis. The final result of this thesis is formed by stochastic decision models, which are based on the results of the prediction models calculated using the machine learning method. The central topics in the theoretical processing of the thesis task are explained in this thesis in the following order: from application to theory. The conceptual relationships of the SAP Predictive Analytics application solution are presented first. It is specifically explained how the key figures prediction confidence and prediction robustness are calculated, which determines the quality of the prediction models. The implementation of the statistical learning theory with the support vector machine algorithm is explained in the second step. The method of classification is discussed up to the positioning of the statistical learning theory. According to the contribution of this thesis, the development of stochastic decision models for the marina industry in Croatia, the last point of the theoretical analysis covers decision theory. The focus is on the theory developed by Thomas Bayes about a-posteriori decisions.

The empirical part of this thesis includes six prediction models developed based on clustering, classification, and regression analysis. The prediction model was developed to predict of which applicants for a berth it can be expected to have creditworthiness and long-term customer relationship as well as contracts with a high contract volume. The next step is to find out whether the decisions about applicants for berths, contract renewals with existing customers and assignment of customers to customer groups as the basis of customer group-specific marketing campaigns with knowledge of the results of the prediction models are significantly improved. After the emergence of significantly better decisions, the basis for the development of stochastic decision models was found. According to the six prediction models, six decision models were also developed in this thesis. The first three decision models were combined in a further decision model. This results in a two-stage decision model. The results of this thesis are then summarized, and an outlook is given. The decision models can be used as a conceptual basis for programming digital decision-making processes. This thesis thus provides a contribution to digitization in controlling.

2 THE IMPACT OF DIGITIZATION ON CONTROLLING

Digitization is changing the controlling process in all areas. The availability of real-time data in all areas of the company leads to the so-called "democratization of data". Although this allows the management to collect classification-appropriate data, it also requires a stringent data model to promote objective reasoning. Jason Bloomberg published a discussion on the terms of digitization, digitalization, and digital transformation:

"*Digitization essentially refers to taking analog information and encoding it into zeroes and ones so that computers can store, process, and transmit such information. Digitalization is the use of digital technologies to change a business model and provide new revenue and value-producing opportunities. In reality, digital transformation requires the organization to deal better with change overall, essentially making change a core competency as the enterprise becomes customer-driven end-to-end."* (Bloomberg, J., 2018).

With a view to the definitions thesised by Bloomberg, this thesis relates to the definition that digital transformations do change the decision-making processes in controlling and management.

Hölscher offers another interesting perspective with a view of digital Darwinism. The struggle for survival will also take place in the digital economy. Think of the new cyberattack challenge. The one who can be more agile – also in controlling – wins. This applies to technological changes as well as the adjustment of business models to the expectations of customers (Hölscher, B., 2017, pages 95–95 and Weber, J., page 1 ff).

The central driver of digitization is always the law named by Gordon Moore in 1965, according to which the performance of computers doubles every 18 months. Thackray outlines: *"The original statement of Moore's law is that the number of transistors it is more economical to produce with double every two years."* (Thackray, A., 2015, page 507). "*If you were to manufacture a smartphone with the technology of the 1970s today, the smartphone would be 12 square meters in size.* (Gassmann, O., 2019, page 7).

2.1 Challenge to Controller

Controlling must define its position in the field of tension between data collection, data analysis and data science and management decisions. Merging data, for example, in external and internal accounting into common tables of the application solution creates new possibilities of integration in reporting, but also requires a new and clear design of responsibility in the organization and responsibility for the data. The fact that each department only manages its own data makes no sense in the digital age. Historically, data compilation and its graphical processing was one of the operational fields of activity in cost accounting until controlling. Data warehouse development resulted in the creation of central data pools from which controlling reports were developed. Setting the application solutions in business warehousing and the definition of key figures and their calculation was one of the exclusive tasks in operational controlling. Now, using the so-called Self-Service-Business-Intelligence tools in computer applications, a single software provides self-service data visualization, real-time analytics, and advanced development tools. "*Mit BW[4] auf HANA und dem Einsatz von BW Workspaces können Sie den ersten Schritt in Richtung Self Service Business Intelligence (Self Servie BI) gehen"* (Klostermann, O., 2015, page 153). Ostermann et al. make clear that the reporting tools enable the departments in enterprises to design their queries for reporting. This may apply to operational reporting. Reporting with tactical and strategic key figures, however, requires specialist knowledge for identifying the correct database, developing data models and generating and interpreting forecast models. Here, the experts in data science and controlling work closely together to develop decision-making bases for management. Basically, this enables the management at all management levels to select data in real-time without the involvement of controlling and to summarize them in reports and graphics. SAP has developed the Digital Boardroom with the following goal: the basic idea is to use a consistent, up-to-date information base in management meetings and to get away from standard PowerPoint presentations, which are still a reality in many meetings today. However, there is a risk that the data in the boardroom will grow so extraordinary, that the overview would be lost. Additionally, every member of the board could collect data supporting their own goals. Therefore, strict data modeling is required.

The challenge for the controller is, therefore, acquiring knowledge in various disciplines. A controller must be a professional discussion partner with data scientist to be able to provide

[4] BW – Business Warehouse

specifications for the modeling of prediction models. The controller must be familiar with the mathematical and statistical principles to understand the model parameters. The controller must be able to implement the strategic specifications of the company management in cooperation with the data scientists in data science projects to be able to develop workable predictions. The controller must be able to convert the results of the predictions into expert reports as a decision-making basis for the management. Forecast updating using machine support is the basis for agile controlling.

2.1.1 Controller as Service Provider for the Management

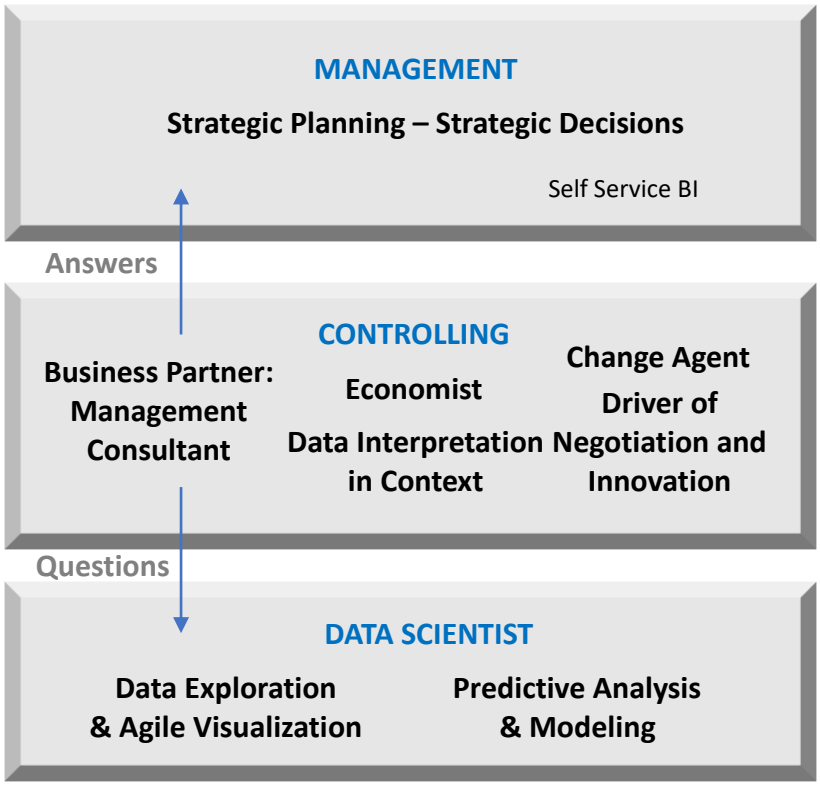
The role of the controller will focus on analyzing data in context. The controller knows the value drivers of the KPI and can refer to them in the interpretation of the numbers. While artificial intelligence (AI) is used in statistical analysis and forecasting, the interpretation and evaluation of numbers in context are considered soft factors, and this is the task for controllers. Management could use IT tools for self-controlling. This would increase the need for coordination, because it is important to know who uses which numbers and when. Experience shows that management does not like to use self-controlling. Therefore, controllers will still be in demand as service providers (Gänßlein, O., 2017, page 21). The ever-increasing opportunities in information technology and the associated data analysis and data prediction possibilities lead to the requirement for the controllers to expand their skills. Although data scientists are responsible for developing pattern recognition and forecasting models, the controllers must be able to interpret data. Weber mentions that the focus of the controllers' scope of tasks is transforming into a business partner of the management. This could relate to the preparation of expert reports. The controller's role as a management service provider is the guiding principle of the International Controller Group IGC (Weißenberger, 2011, page 33). One of the controller's tasks in the capacity of a business partner is support in the planning and implementation of digital transformation in the company. This affects almost all resources in company operations. Examples include moving the back office into automated computer-controlled processes, digital communication, the use of digital infrastructure, and a new understanding of customers, business relationships, and value chains (Gassmann et al., 2019, page 28). According to Gassmann et al., digital business models are using the agility and leverage of intangible resources, digital channels, set industry standards by disclosing intellectual property, and encase their products in service systems. The scope of controllers' tasks resulted from the access to and analysis of digital data, digital customer communication.

Digital transformation transforms business models into digital business models and will create new digital business models. The subject of a digital business model can be a virtual product; for example, a digital three-dimensional model of a product prototype. The simulation of product properties in theoretical scenarios is called virtual product development. Controllers cannot and do not want to replace data scientists, computer scientists, and market sales representatives. However, they can advise on how to implement an efficient and flexible control of value-added chains and how to optimize the associated decision-making processes. The controllers' role lies in supporting the development of management concepts in distributed virtual product development. This includes the cooperative and simultaneous approach in IT-supported project management. Management concepts for distributed virtual product development are explained in Spur, G.: The Virtual Product. (Spur, G., 1997, page 577).

2.1.2 The Organizational Position of the Controller

An important challenge the controller is facing is his role in data analysis in general and the prediction of future business developments. Weber notes (Weber et al., 2019) that these topics are currently still viewed-by marketing and supply chain management from the point of view of controllers. However, if controllers stay out of this, what is stopping the management to directly consult data scientists and bypass the controllers? The answer to this question lies in the expert know-how of controllers in terms of business relationships. Controllers provide the economic design for high-quality IT systems (Gänßlein, S., in Horváth, P., 2017, page 21). From the perspective of the author of this thesis, controllers determine the parameters in the development of predictive models.

Figure 2: The Role of Controlling



Source: Author

Controllers must be able to present complex facts in accounting and the accounting to the management in such a way that the integration of all influencing variables becomes clear (compare to Grünert, L., in Horváth, P., 2017, page 17). As a service provider for the management, controller should be informed about the corporate strategy and derive tactical plans and actions to implement the strategy. The collection of mostly unstructured data must be changed into structured information with a significant relationship to the corporate strategy. A controller makes a decisive contribution to ensuring the rationality of management decisions (Irrek, W., 2002, page 46). A controller assumes information needs of the management about future developments and communicates specific requirements for the development of forecast models to data scientist. The data scientist ensures the creation of reliable statements with the development of forecast models, which can be proven by key figures on prediction power (high relationship between input variables and target variable) and robustness (high accordance between the known results and the predicted results).

Conclusion

Controllers' position in business organizations is changing from data gatekeepers to democratized data advisors. Generation of an unimaginable flood of data leads to the fact that the pure key figure calculation for corporate management is no longer sufficient. The influences, not only of customers, suppliers, and competitors but also the dynamics of the markets, are too diverse. Market knowledge and corporate policy alone are no longer sufficient. A controller must also be a partner in a dialogue with proven experts in data science, as well as the management. The range of yesterday's controller's know-how is from cost management to the design of key performance indicators. The controller of today, on the other hand, needs the know-how and expertise ranging from business design from data models, business processes, to business models. Business models should be tested constantly, from a critical angle, and methodically.

2.2 Digitalization Concepts

The concepts for implementing digitalization in controlling relate to the merging of big data, artificial intelligence, and business intelligence. This is not just a technical task which is solved by the manufacturers of computer applications. The scope of tasks in controlling involves economic design. The tasks include the identification of the drivers of business results, the development of key figures and performance indicator systems. The central aspect is central data management, whereby controllers are responsible for the field of semantic data modeling.

2.2.1 From a Business Model to a Digital Business Model

Digitalization business models will play a decisive role in the future competitiveness of a company. A business model is the definition of a logical structure in order to achieve a competitive advantage and a positive economic result with a sellable product. A business model will be successful if it is innovative. This is the area in which controllers have the central role as innovators, in which they contribute to the digitization of existing business processes or the development of new business processes. The path to a digital business model relates to the flexibility of the value chains, the use of optimization potential, decentralized control, and the use of real-time information to support decisions (Gassmann, O., 2019, page 28). This thesis starts at the point of real-time information for decision support, primarily through the

possibilities of simplifying the data models of modern business applications in information technology, as described in the previous chapter regarding the system SAP S / 4HANA. More in focus, however, is the use of data science models to generate real-time information as decision support. In this thesis, forecast analyses are modeled and analyzed with the IT application SAP Predictive Analytics. Decision models for the marina industry in Croatia are developed based on the prediction results. The implementation of digital transformation in a company requires a multi-stage development process. Business processes change. Acting when necessary turns into forward-looking action. It is necessary to design an implementation plan. Such a plan represents a kind of roadmap, the stages of which must be reached successively. Therefore, roadmaps that describe and demonstrate the individual steps of the process have been developed for the planning and organization of the digitization process. This work refers to the Roadmap Industry 4.0 (Tschandl, M., 2017, page 22). In the first step, possible fields of action are examined. Process analyses can be carried out to detect whether there are any weak spots in the company processes and whether there is potential for optimization. A typical example of such a weak spot is the processing of analog data. Time-consuming processing, the lack of compatibility between different analog data formats and personal dependency of the processes are reasons for requiring improvement. In the second step, the digitalization competencies of the company are examined. Technical facilities, knowledge of employees and planning and organization of business processes define the status of the company. Step three includes the determination of the target state by developing technical and organizational solutions to optimize the business processes. Controllers can support the digitization process in a company through methodological competence (Tschandl, M., 2018, page 45). As part of this thesis, the methodological focus is on the development and application of predictive models. The decision-makers' experience is supported by robust predictions, which are calculated based on modern information technology. It is the task of controlling to identify suitable forecasting methods, to determine the appropriate target variables and to draw sustainable conclusions from them. The use of the digital transformation methods in the form of prediction does not only digitize the business processes; it also digitizes the entire business model and creates new digital business models. The development of a digital business model unfolds in several steps (Hoffmeiser, 2017). It starts with a fractal company. A fractal in a company is an independently operating organizational unit. Fractals define their own goals and are characterized by self-organization and their own dynamics. Fractals developed as a separate business model can be connected into a framework. Standards are required for the interfaces. There is an interesting approach that argues that digital business models consist of three control loops: input,

processing, and output. Digital business models are machines, usually in the combination of hardware and software, which digitally process an input – probably transferred from analog to digital – and use algorithms to perform the tasks associated with the input. The digital result can also be converted into an analog edition. The networking of digital business models and their IT-technical application takes place via the Internet. The networks can be closed or open. In closed networks, only actors from certain organizations have access to the network. In between, there is a cooperative network in which the actors have all the prerequisites. For example, MyTaxi requires proof of passengers' transportation permit (cooperative), and UBER does not require proof (market). In times of digitalization, customers expect more and more services around the centrally offered product. Configuration is important. To select a marina, digital information systems about current offers, additional offers for berths, weather conditions, etc. would only be conceivable in a cost-effective manner via a digital information system. This would be the customer side. On the side of employees in the planning of customer service and controlling, real-time information have a decisive role in decision support. Real-time information refer to customer characteristics which turn out to be decision-relevant, as shown in statistical analyses. Real-time information refer to machine learning, which is used here. Statistical methods are used to let computers learn from data. The algorithms for pattern recognition are found gradually, but with a high-speed assignment of characteristics to characteristic classes. (Hoffmeister, 2015, page 199). To convert the existing business models into digital business models, a project team must be formed that will devise the creation of implementation concepts and plan and organize the implementation process. In the first step, an assessment will be carried out involving all company departments that will examine the company's products and business processes for their digitization potential. In this initial workshop, all project employees are informed about the project and the existing business models are critically examined. Therefore, data about the organization, key figures and products in the business, as well as services, are needed for the initial workshop. A controller complements the assessment with methodical advice. The approved methods include St. Gallen Business Navigator (Gassmann et al., 2017), the Business Model Canvas (and the so-called "3C model" (Tschandl, M., 2017, page 34). As an example, this thesis explains the Business Model Navigator. According to this method, a business model does not have to be completely reinvented but can be created by recombining the elements of the existing business model. An existing business model can be transferred to other business fields or other industries. Another possibility is the application to new products. The combination of existing business models is also conceivable. The method is the development of the so-called "magic triangle with

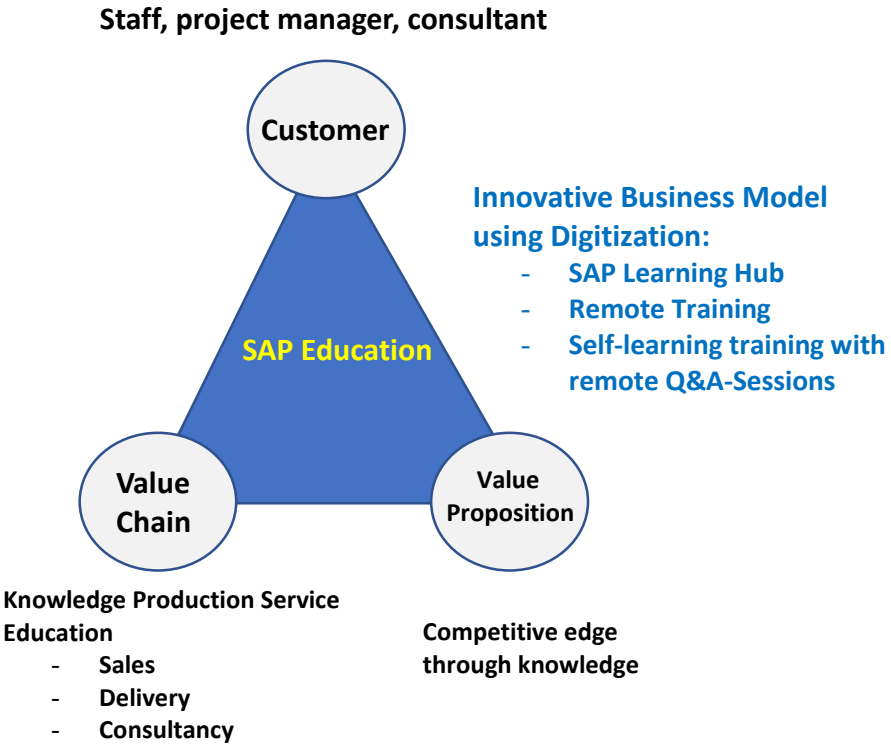
questions" (Progoscewska, P. J., 2016, pages 17–19): Who are the customers? What value-added do we promise them? Which distribution channels do we serve them? What influences the customers? Who is the human being behind the customer? What is the value proposition?

What values are created for the customers and how are they perceived? Which of our products meet which customer requirements? Which unique selling points can we claim for us? What is the value chain? Which instances are involved in the value chain? Which competencies and key activities are necessary? Which are the most important preliminary products and from which suppliers are they delivered? Company-specific, unmet customer wishes can be identified through the assessment. Examples: a customer wants a time reduction and cost reduction. Solution: A webinar does not require travel and associated travel expenses. Solution: Webinar. The customer would like to have the opportunity to access the training session with time flexibility. Note: The examples are derived from the environment of SAP[5] Education. New business ideas result from the interaction of the considered elements. This means that a purely technical innovation will lead to success if it adds value to the customer. A key element in the analysis of the influencing factors is the recognition of trends. Thus, the conversion of external and internal accounting is a clear trend which runs through the leading business application solutions. Against this background, the consideration of such a trend in the further development of own computer applications is of great importance. An existing solution must be considered in the concept of financial accounting and management accounting. This does not only affect the computer application, but also the organization within the company. Business models could therefore consist of offering a corresponding change management consultation in addition to the computer application. For brainstorming, the St. Gallen Model 55 provides card decks with sample business models. At first, the selection criteria are to be determined according to which 6–8 cards and pattern business models are selected from an entire deck of cards. An attempt is then made to transfer the business model pattern to one’s own company. Two basic principles of pattern adaptation are proposed: the similarity principle and the confrontation principle. The similarity principle searches for similar industries and business scenarios and verifies how many matches there are in the sample scenarios with their business scenarios. If there is a high number of matches, it is considered whether the business model template could be applied to your own business scenario. In the confrontation principle, it is the other way around. It searches for industries or business scenarios that differ significantly from their business scenarios. It is now being investigated how one's organization would deal with this pattern scenario. Since the

[5] SAP – Systems, Applications and Products in Data Processing

participants generally only know their own industry, a high degree of creativity is required here. From the perspective of the author of this thesis, it has similarities with business engineering. The method motivates free thinking in other dimensions. Free thinking in other dimensions has a specific expression: “Design Thinking”.

Figure 3: Business Model Navigator - Assessment



Source: Author

The implementation of design thinking involves the consolidation of individual ideas, the check for completeness and proposals for action in organizational and technical terms. The consolidation of individual ideas refers to uncovering contradictions. It is necessary to investigate whether contradictions can be resolved, partial aspects remain or whether necessary priorities have been set. A central consideration for deciding on a business model is digitization capability. Data management is considered in this thesis as an example. A very low digitizing ability could be detected in the following scenario: The master data are stored in different departments. Data are stored in different media. Data quality is not ensured by a prescribed update process. A high degree of digitization, on the other hand, results from using a Master Data Management (MDM) system. The MDM links all critical data (regardless of which

database they are in) to the so-called “master file” as the central reference file. This allows SAP users to determine the sources of master data, clean up the master data record by duplicates, merge distributed data, and much more (Mindsquare, IT-Consultancy, 2019).

Conclusion

Thinking in controlling has evolved from process-oriented thinking to thinking in business models. This also applies to the management. In this respect alone, both instances have become business partners. Digitization has driven the evolution of general business models into digital business models. A digital business model includes the integration of communication and information technology networking. The transformation of companies in the digital age requires a process that incorporates all business factors. While business management factors are implemented with technology, the "human being" factor must be convinced. Controllers assume the role of methodological competence in the transformation process. Without the methodical guidance of controllers, the process is critical and without success.

In this thesis, a digital business model could be the following: A company in the marina industry offers a cooperative digital business model that customers and potential customers can use to request services such as a berth. The inquiries are first processed digitally by pre-selecting prospective customers based on predefined models using pattern recognition and classifying them into applicant groups using the IT system. The results are used to make decisions on berth allocation.

2.2.2 The Impact of Digitalization on the Controlling Process Model

One of the impacts of digitalization on the controlling processes is automation and standardization of processes in controlling, faster data acquisition and data analysis as well as an improvement of planning and budgeting processes (Nasca, D., 2018, page 73), (Gleich, R., 2014, page 73). In this thesis, agile controlling is examined. However, agility in controlling does not replace a set of rules that must be used to make fair and objective arguments. Such a set of rules requires a data model that is coordinated down to the board level. The business concept of the company data model leads to a semantic data model in theory and practice. Data management, both in person and in information technology, is responsible for creating and updating a data model. The core of data management is the development of an efficient semantic data model. The data model represents the business design for transferring the exponentially increasing flood of data in the context of digitization into business-related information. These

include the harmonization of data in business and technical terms. Business harmonization refers to a uniform terminology (standardization), a thematic grouping, for example, according to customer groups and product groups and pattern recognition in unstructured data. This last aspect is one of the key points of this thesis. The following changes in data management can be observed through a digital transformation (Steinke, K.-H., 2017, page 58): new data collection and evaluation processes. The analysis of real-time data from social media through structuring and classification (author's note) triggers the process of business analytics. Web-based applications also enable data evaluation by the management, whereby controlling is responsible for keeping the applications at a high level to ensure the informational value. Further development of the analysis tools in the context of digitization also influences the objectives and control in controlling according to the motto: The goals increase with the possibilities. Agile controlling is the answer to changing environmental scenarios. This thesis is carried out by agile controlling with machine support for calculating future customer behavior. The strategic framework for agile controlling is set by orienting corporate goals to the digital business model, which requires the implementation of manual business processes in digital business models. This refers to a business model-oriented objective and the derived strategic initiatives. These initiatives determine the thesis and development projects as well as the necessary investments. Planning is oriented towards a rolling forecast calculation and simulations. Planning is standardized up to outsourcing of the operative planning processes in the shared service center and expert center. The data collection processes with bottom-up planning will be supplemented in the long term by top-down planning and will also be replaced to a large extent. The planned objectives are calculated by simulation models and drive the planning. The advantages of digital forecasting are obvious: speed, flexibility, mathematical proof, effectiveness, and efficiency. Machine learning, data mining and permanent validation of the forecasting results lead to a more realistic representation of the forecast models (Nasca, D., 2018, page 86). The risk – not only of digital data management but also the technical digitization of business processes – lies in the fact that one's business model must be tested to see if it fits the digitalization. Hölscher (2017, page 12) believes the crucial task is that of the middle management to achieve the probability of success in absorbing the effects of digitization in the business process to the business model all-in-all and transferring the changes to the top management. Here is an example of an unsuccessful transformation to the digital world. Kodak introduced the first digital camera as early as 1991 and was a pioneer in this technology. However, top management ignored the smartphone and digitization of its business model. It is one of the most spectacular misconceptions in economic history. The world's largest producer

of photo films in 1991 had no idea that, shortly thereafter, with the introduction of digital camera technology, Kodak's core business would become almost superfluous. When the brand manufacturer realized this, it was too late. He did not want to endanger his ancestral business, so he was hesitant to embrace digital technology. Instead of massively investing in the market itself, Kodak brought the chip-equipped cameras into stores together with Nikon, with moderate success. It was soon overtaken by Canon, Sony, Panasonic, and other competitors from the Far East that produced their models at a much lower price. (Handelsblatt, business pater, 2019). On July 6th, Kodak filed for bankruptcy. Agile controlling has become a trademark for controlling in the digital age. Agility as a solution is based on personal interaction. However, this should not lead to a daily commotion; controlling must be able to respond appropriately to the dynamics of the environment. In general, agility is understood as adaptability of organizations. If personal agility is defined as a central characteristic of agility, and a business relationship is still a long-term, economic-oriented interaction between economic agents, management and controllers become business partners. Thus, agility appears as an aspect of a business partnership between controllers and management. The question of implementing agility is usually reduced to planning coordination. However, controlling supports the management in their efforts to make rational decisions (Irrek, W., 2002, page 46) and (Weber, J., 2001, page 112). Regardless of the coordination framework of the plan, controllers are usually only there if a company is on a growth threshold and plans to replace personal instructions. If a controller only wants to implement the adaptation to digitization by increasing the dominant plan coordination, it reaches its limits when it comes to substantial adaptability and innovative capability. Substantial changes require the so-called “self-coordination by mutual agreement” of the coworkers without the inclusion of the instance, defined scope of action, and qualified, responsible, acknowledged, and team players. With digitalization, the complexity of business processes has increased in operational, tactical, and strategic terms. The increasing complexity cannot be explained by highly aggregated key figures. A systemic understanding is required instead.

Conclusion

Digital data management is characterized by simplicity, flexibility, integration as well as orientation towards the interests of the recipients and their participation. The importance of controlling will increase. The central role of classifying the predictions calculated from

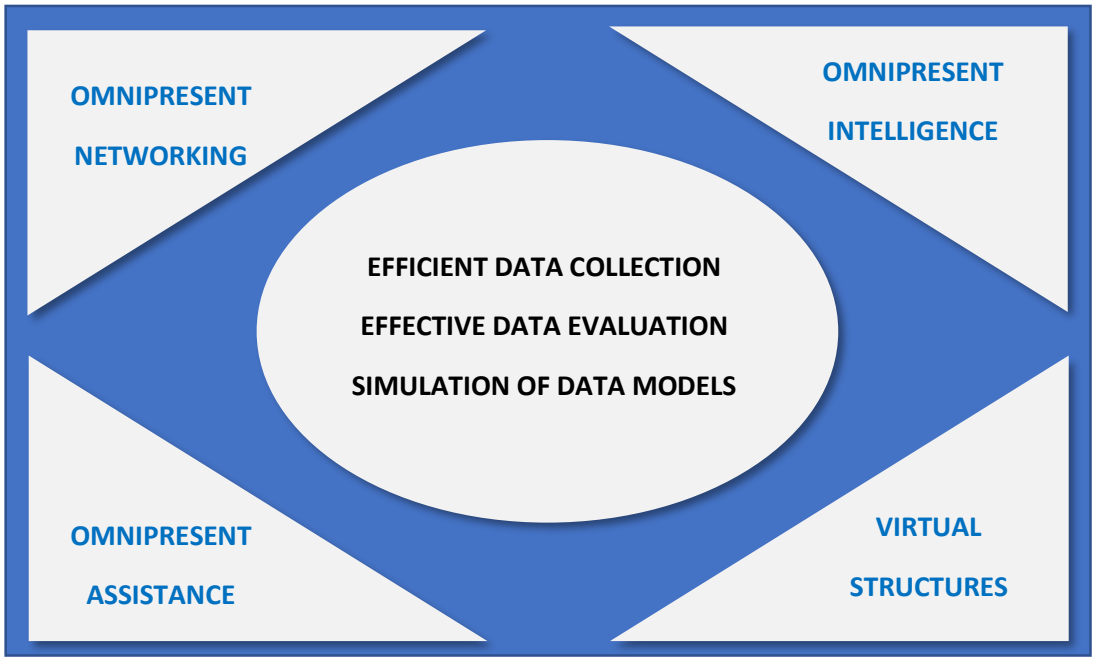
simulation models and their interpretation in the context of the company and the corporate environment requires the controller to be a business instance. Data become transparent only through the evaluations of controllers. As far as the risks of digitization are concerned, controllers are indispensable as coordinating authorities, innovation drivers, and consultants or business partners of the management.

2.3 Controllers' Information Technology in the Digital Age

Access to information is characterized by the ubiquity of networking, intelligence, and assistance (Rossmeisl, E., 2014, page 141). The technical possibilities of the fourth industrial revolution are in the fusion of industrial manufacturing processes with information technology. The advanced development of sensor technology allows every production step and every resource to be planned, monitored, and predicted in further development. This places flexibility as well as agility on a completely different level, the digital level. This development was accelerated with the transition from the third to the fourth industrial revolution. Now, however, there is global networking of all production factors. The technically possible rapid adaptation to changing requirements also requires adequate controlling. Scenarios must be developed and simulated to measure agility. These include, for example, the well-known stress tests, in which companies are – as a simulation – exposed to special situations to investigate the available regulatory mechanisms. Networking naturally leads to a technical dependency. Therefore, omnipresence is an important component for effective information access and efficient evaluations (Rossmeisl, E., 2014, page 143). Digitization has been leading to a fusion of business processes with IT on a global scale. Networking creates digital business models whose implementation and control become more and more complex. The property of digital business models is the digital optimization of business processes. According to Hoffmeister (2015, page 2), the elements of a digital business model are the following: business transaction, offering of a digital system, requesting a digital system, a digital service, a digital compensation, repeatability of the transaction. In this thesis, business transaction refers to the allocation of berths to applicants. An information system created on the basis of the decision models developed in this thesis could completely digitize this business process. The applicants and customers of the company could use an online service as a digital system for filing requests from their perspective and offering a digital system from the perspective of the company. Additional digital services could be provided to increase the functionality of the systems, like additional leisure activities. The effort that is initially invested in the development and

establishment of a decision-making process based on prediction models is compensated by higher effectiveness and efficiency. In any case, a digitized business process is repeatable.

Figure 4: Paradigms on the Internet



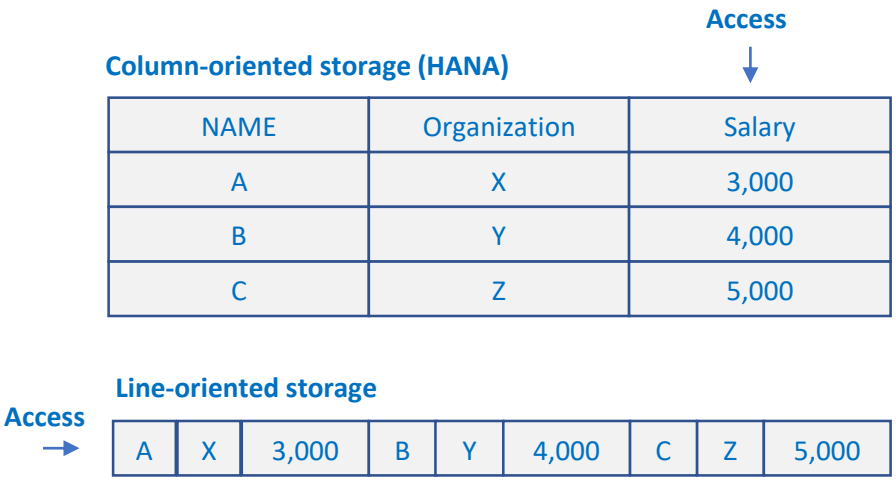
Source: Roßmeisl, E., 2014, page 143, adapted by the author of this thesis

2.3.1 Digital Transformation in Controllers' Information Technology

Digitization is the technical translation of any information into a digital format (Butsmann, J., 2019, page 20). Transforming analog data into digital data allows access to all data through the usage of technology. However, the resulting data are usually in an unstructured form. With the data mining method, information systems extract information from data and thus generate knowledge (compare Runkler, A., 2015, page 2). If digitized data are recorded and processed by all instances of a business process, it results in the creation of a consistent digital business process. There are several definitions in the literature on the concept of digitalization. Gartner defines: "Digitalization, according to Gartner, Inc., is the process of employing digital technologies and information to transform business operations" (Gartner Group, 2020, blog). Digital transformation describes a fundamental transformation of a company into a fully networked digital organization. All business factors, products, production factors, services, and processes are being redesigned or newly developed and adapted to the requirements of the digital economy (Hölscher, B., page 43). In conclusion, digitization refers to data, digitalization to processes, and digital transformation to organizational units, companies, and even industries.

Digital transformation leads to innovative technologies that displace and replace existing technologies. This includes the use of information technology such as cloud computing and application solutions with a digital core. (Butsmann, J., 2019, page 24) Butsmann summarizes the characteristics of a digital core of the application solution SAP S/4HANA as follows: Integrated system, a real-time operating system that provides data in real-time, a common database for all applications in all business areas. Data are the fuel of digital business processes, including controlling. Due to their architecture with a number of tables for each business module and performance based on hardware technology, conventional systems are unable to process these high volumes of data in real-time. Controlling in the digital age depends on the use of this technology, since the company needs to keep pace with corporate development. This relates to real-time data availability and a change in controlling from the analysis of past data to the prediction of future events. Replacing many interfaces of heterogeneous system landscapes with an integrated system by simplifying the data model has created the advantage of consistent and readily available data. The next step was further development of the application solution with the intelligent use of data for forecasts. All company processes are controlled and support to management's decision-making processes is provided through the usage of embedded analytics, (Butsmann, J., 2019, page 35). Rapid data availability has become possible because basic technology for very large main memory in the computers has made it possible to develop an in-memory database. There is also another database design, from a row-oriented to a column-oriented architecture. The following figure provides an example. In a row-oriented storage, the records are stored sequentially. A program that calculates an average salary must read the data sets completely. For a column-oriented storage, on the other hand, only the columns of salaries must be read. This shortens the execution time of the program to a fraction of the time. The availability of the column-oriented database and the technically possible in-memory computing have led to high-speed aggregation and parallelization of processes in the database (Butsmann, J., 2019, page 36). With the use of this technology, it has become possible to have analyses and transactions on the same platform.

Figure 5: Column-oriented database



Source: Author

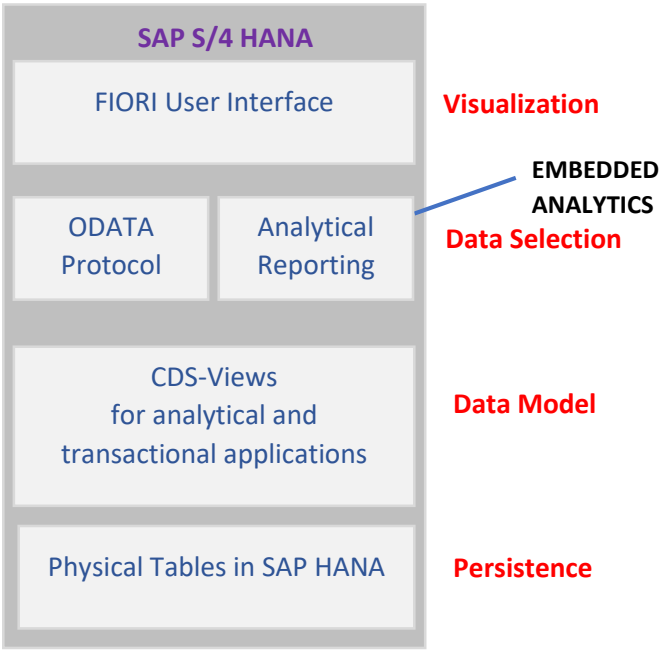
Without transferring data to a data warehouse platform, the analysis is available in real-time (nanoseconds). In classical systems with their database platform and their data warehouse platform, data had to be transferred for evaluation purposes. Different techniques were used to increase the performance. This includes, for example, data aggregation, in which data were redundantly pre-compressed. A business process consists of an analytical and a transactional part in most cases. If we post a sales order income, the expected revenue could be posted into an extension ledger for predictive accounting. Predictive accounting could increasingly simulate different sales scenarios like the analysis of which customer groups take note of the marketing information and which ignore this information. This leads to an update of the target groups that need to be addressed. Additionally, predictive accounting supports smart prediction.

2.3.2 Real-Time Business Analytics

In general, analytics is the collection, structuring and visualization of data in order to understand and make the right decisions. This involves linking the transactions to analytics in the following ways: linking to a technical platform, linking within the transaction, and linking to a common user interface. As a database solution, HANA offers data reading with high-speed aggregation without the need for data aggregation and creates virtual data models. Using virtual data models, it is possible to abstract and process the data of different origins, without the need for any reformatting nor storage. This means that even large amounts of data can be processed and evaluated during the execution of transactions. Linking a transaction to an analysis provides

decision support during the transaction and the analyses of the executed transaction. Machine learning is used for decision support so that in the future routine decisions can be made by the learning system itself. In these standard processes, the human factor only takes on the role of monitoring. The linking of transaction and analysis on a single screen was done at SAP by implementation of the user interface Fiori. Transactional data, key performance indicators and reports can be processed together on a web-based page (Butsmann, J., 2019, page 48). The monitoring of business processes requires operational reporting, which selects the data in real-time. SAP S / 4HANA uses the so-called “Core Data Services (CDS)” to select data in real-time. The CDS View forms the basis for all transactional and analytical applications. Once data have been modeled using a CDS view, they become available to the application. With the assistance of CDS views, data are prepared by the database, not by the transaction. CDS views use a standardized data protocol, the ODATA (open data). ODATA is a web standard that enables consumption in SAP's web-based user interface FIORI. The SAP S/4HANA concept enables the integration of analytical data in the transactions. Various SAP applications are available for analysis in embedded analytics. The relocation of operations to the database, for example, refers to data aggregation at run-time in the database as opposed to the classic approach of pre-aggregating the data in transactions. In this context, a code pushdown is spoken into the database. Instead of telling the database which data are needed for processing, the statement is passed to the database refering to how the data are to be processed, like the order of the already processed raw material. The access to business data is provided using the so-called “virtual data model”, i.e., a data model, which is formed at runtime of the programs. The virtual data model transforms the data into well-defined, easy-to-understand views. These data can then be consumed directly from HTML5 (Basis for SAP FIORI). The architecture of embedded analytics with SAP HANA was illustrated in a figure by Butsmann (Butsmann, J., 2019, page 56).

Figure 6: Architecture SAP S/4HANA



Source: Butsmann, J., 2019, edited by the author

Conclusion

The technical core of future-oriented controlling is a database which provides real-time data. This allows ongoing business processes to be evaluated and forecasts to be created. An example of these database requirements is provided by SAP HANA in the Business Suite SAP S/4HANA. Due to the in-memory technique, redundant data management is no longer necessary. Due to the timeliness of the data, controlling can react quickly so that agile controlling can be implemented technically.

2.4 Predictive analytics

"The management – controller relationship has changed in the digital age. In the past, it seems that controllers were DCOs – Data Collecting Officers – and now, controllers are MCPs – Management Consulting Partners. With the new role of controlling positioned as an interface between data scientist and the management, it is the responsibility of the controlling to avoid misinterpretation of data and find decision-relevant key data" (Vitezić, 2018, page 21). Empirical evaluations and forecasts are important for giving decision-makers an insight into

the business context. Only the knowledge of the relationships between INPUT and OUTCOME enables sound management decisions. While OUTPUT indicators measure the results of measures taken, OUTCOME indicators measure the benefits and thus the success of the project.

2.4.1 Learning from Historical Data

The solution to a problem starts with defining the problem itself: creating a so-called "problem specification". At the same time, this specification forms the basis for development of a predictive model. Real-world objects are analyzed, relationships between objects are revealed, operations are searched for, and, ultimately, the results are presented in a formal notation. An example of a statistical method is the least-squares method as the standard mathematical method for calculating a functional relationship of observed values reference. In this case, a mathematical function is identified for a data point cloud whose graph is as close as possible to data points. The data may represent physical measures or economic quantities. The least-squares method is then used to determine the curve parameters so that the sum of the square deviations of the curve from the observed points is minimized. The deviations are called residuals. In the role of an analyst, controllers assume the task of interpreting the results from statistical calculations. Controllers bring their experience to company data handling. Controllers know how to use the data and convert them into useful decision-relevant information. They are business partners of the management because the availability of big data has increased the accuracy of the data and their strategic importance. From the various forms of the mathematical-statistical analysis, the methods of prescriptive and exploratory analytics are distinguished. In prescriptive analytics, statistical data are evaluated in order to make statements about future developments. In the exploratory analysis, knowledge gained from the statistics is not used to make predictions for new data, but to understand the relationships in the previous data (Charbert, A., 2017, pages 28–29). However, it must be measured how well the created prediction model represents the correlations of the characteristics in the given data. The prediction model must also be tested and examined to discover how well the prediction results could be transferred to new data. The accuracy of the model indicates the influence of the explanatory variables on the target variable. The performance of the predictive models can be measured by two key performance indicators:

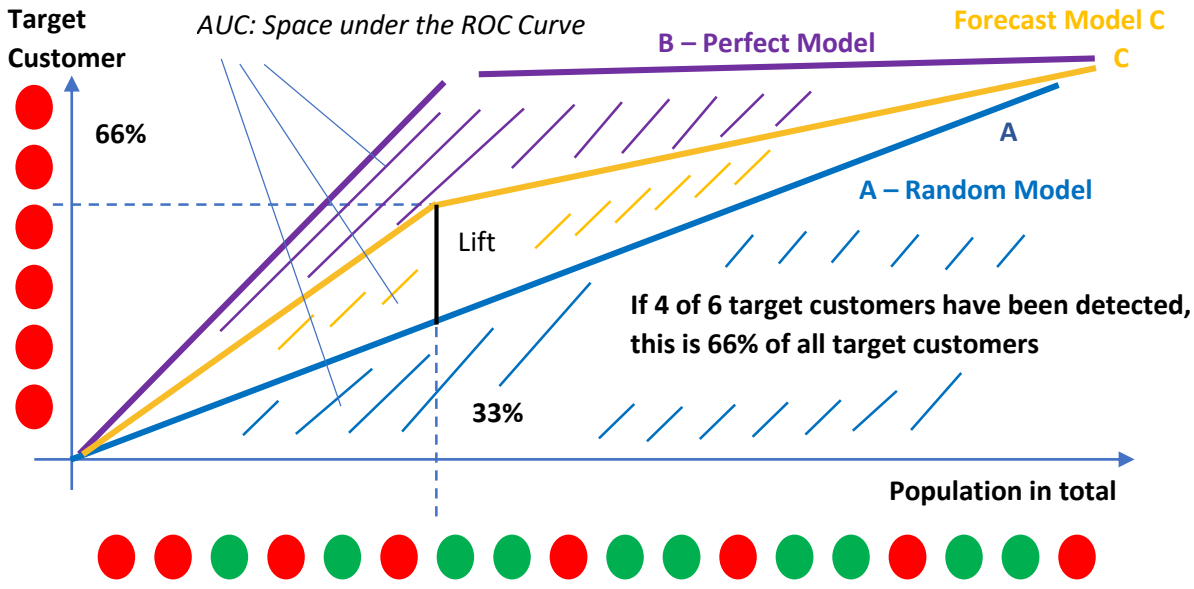
1. Predictive power (KI) or predictive capability, which indicates the accuracy of the model. The accuracy is measured by the correlation of the input variables (explanatory variables) to

the target variable. It defines the significance of the input variables for explaining the target variable. KI has a range of 0.0 to almost 1.0, usually 0.998. The higher the value of the AI, the more accurate the model. However, not all available explanatory variables of the database should be used. Only the essential variables that contribute to the target variable should be considered (Chabert, A., 2017, page 124). Only then can decisions be made based on these remaining explanatory variables. If all available explanatory variables of the used dataset would be part of the prediction model, then the value of KI would be 1.0. But, in this case, it would not be possible to concentrate on the important variables. The slogan for such a model would be “All variables somehow explain all target variables”. However, such a model is not useful.

2. Prediction confidence (KR) or reliability of the prognosis characterizes the statistical robustness of the model. KR has a range of 0.0 to 1.0. Only a value above 0.95 reveals that it is a robust model (Chabert, J., 2017, page 124). Robustness measures the prediction accuracy of the prediction model when applied to data whose results are known and thus the prediction can be compared with the actual development.

The abbreviations KI and KR come from the American company KXEN INC., which was founded in 1998 and taken over by SAP in 2013. The results of the forecasts are visualized by the model graphics. The graph of the estimation curve results from the model calculation. The graph of the random curve results from values that have been observed at random. This means that finding or not finding an observed value in terms of prognosis is coincidental. The probability of a "hit" is therefore 50% for the random curve. The definition of a prognosis model is intended to eliminate chance and increase the probability of finding a "hit", if possible, by more than 95%. A forecasting model contains the parameters which should find the target customers, i.e., those who are interested in the company's products. The parameters of the model include personal characteristics, for example, age, education, occupation, marital status, etc. The prognosis model is designed to calculate the influence of the parameters on the interest in the company's products. Once the key parameters have been identified, product marketing can be targeted. In the following example is the entire population with a summed-up value of 18, from target customers with the value of 6 and non-target customers with the value of 12.

Figure 7: Random Model and Forecast Model



Lift is the difference of the random model and the own forecast model.

Source: Author

With random addressing of people, only 33% of the target customers would be reached. A model with random selection, in which no characteristics of the persons are specifically addressed, is represented by random model A. If the analysis of statistical data with mathematical-statistical functions for pattern recognition (data mining) shows a preference of certain characteristics of persons on the basis of which the persons have made a decision on the company's products or comparable products, then the probability to find the target customers is higher than 50%. In the present example, with 33% selection from all persons, 66% of the target customers are already found. In a perfect model, with 33% selection, 100% of the target customers would already be found. From a business perspective, target customers can be found faster and marketing costs can be saved. Revenues can be increased by using forecast models. As previously mentioned, the calculation of the prognosis ability (accuracy of the prognosis) and prognosis reliability (its application on new data) takes place through the characteristics KI and KR. Statistically collected data are divided into two groups: a) Statistical data (STAT-I), where the result is known (customers), from which the model is trained (estimation) and b) statistical data (STAT-II) as part of the data used to validate the model (validation). The first step, the so-called “training” of the model is done. The mathematical-statistical method used

by the system of predictive analysis is the theory of empirical risk minimization developed by Vladimir Vapnik and Alexey Chervonenkis (Vapnik, V., Chervonenkis, A., 1991 (English translation). Vapnik is the main developer for the machine learning algorithm, the support vector machine. This algorithm is used for clustering and classification, the basis for the models in this thesis. There are many examples of scientific theses that used the statistical learning theory. One of the examples is the thesis of consumer purchasing behavior extraction using the support vector machine which is based on the statistical learning theory (Yi Zuo et al., 2014, pages 1464–1473). The consumer shopping behavior was analyzed regarding the way consumers were spending time in a certain area of the supermarket, and which customer groups prefered which supermarket areas. The contribution lies in the clustering of the consumers so that the target group for specific products could be identified. Another thesis using the support vector machine (SVM) has been published by Yang (2012, pages 1489–1496). To provide a better service to network users, the customers' network behavior has been classified into browsing news, downloading shared resources, and real-time communication. With the identification of such customer clusters, the network services could be better addressed to specific customer groups.

Predictive power (KI) and predictive robustness (KR) are calculated to prove the quality of the prediction models. A target variable y should be predicted based on input variables x. It is assumed that the probability for the target and input variables is fixed and known. It defines a set of mathematical-statistical functions as well as an error function. Example: In the case of linear regression analysis, the derivation of a parameter for the absolute values and the slope is the objective function and the quadratic deviation is the error. The goal is to find from the selected functions the one that minimizes the expected error. In the case of regression analysis, the function with the least square deviation.

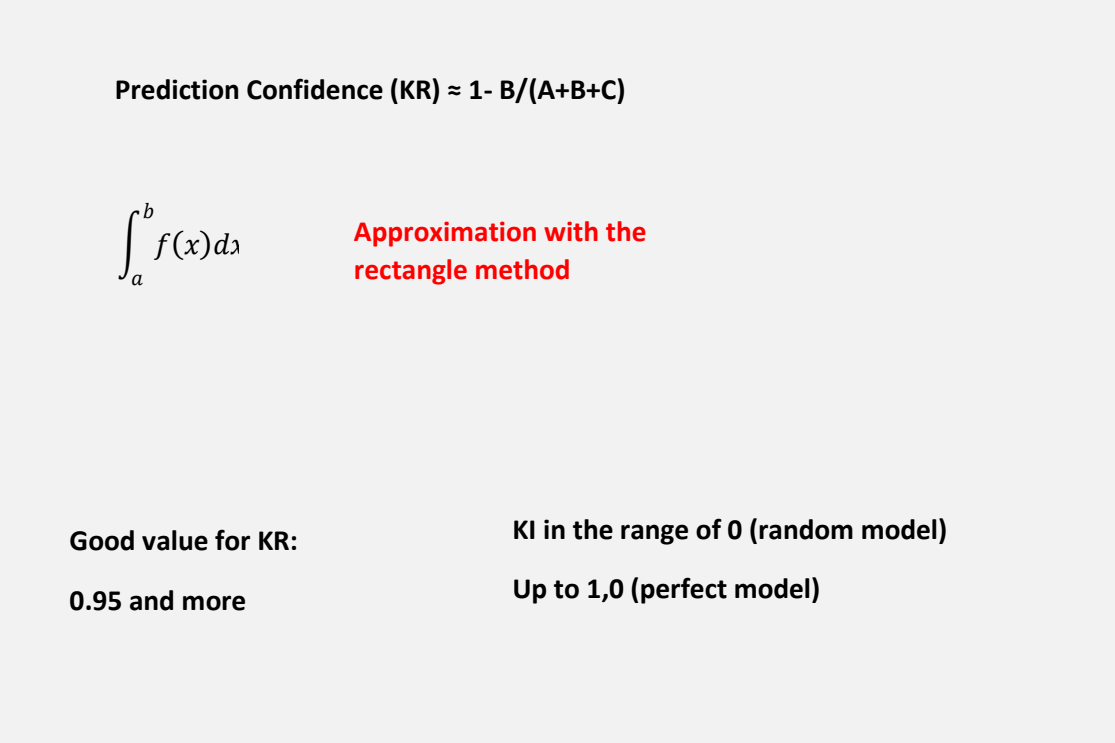
KI is calculated as follows:

The areas A + B + C shown in the previous figure represent the population.

The ratio of the area of the forecasting model C and the perfect model B to the population A + B + C is the measure of predictive accuracy of the model and thus the result of training the model (estimation). The ratio of the area of the forecasting model C to the population A + B + C is a measure for the verification of the forecasting accuracy of the model (validation). As part of the validation process, it is intended to examine, with regard to the known statistical results, whether the predictions calculated by the model on the known statistical results can be

confirmed. The robustness of the forecast model is its applicability to statistical data whose result is unknown, and it takes place via the calculation of KR. The KR results in the following:

Figure 8: Prediction Confidence and Prediction Power



Source: Author

As an example of the robustness, the following series of numbers are given:

x = {1, 1, 3, 8, 9, 9, 10}

The arithmetic mean is: $\bar{x} = \frac{41}{7} = 5{,}9$

The central value median: $x_{med} = 8$

Adding an extra value of 1000 (outliers) changes the arithmetic mean, but not the central value:

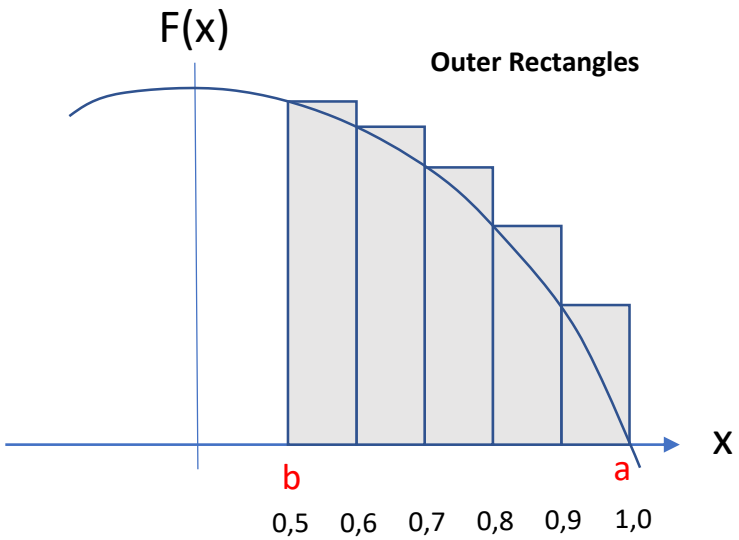
Central Value: $x_{med} = 8$ Arithmetic Mean: $\bar{x} = \frac{1032}{7} = 147$

In contrast to the arithmetic mean, the central value is robust. The identification of outliers is therefore one of the necessary standard functions of a forecast calculation.

2.4.2 The Rectangle Method

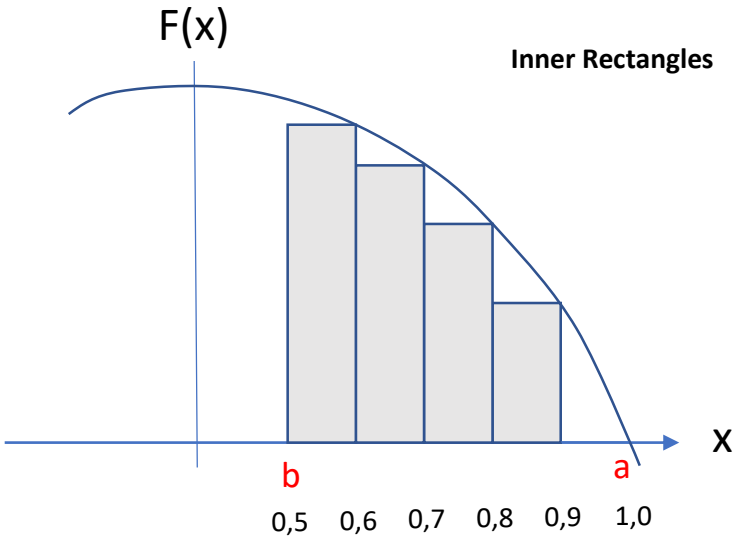
The area calculation is done by an approximation method, the so-called “rectangle method”. For this purpose, the area of the function is divided into a number of rectangles whose areas are summed up. It results in a sum of rectangles which are above the function, and a sum of rectangles lying below the function. The mean value is formed from both values. The following illustrations provide an example.

Figure 9: Rectangle Method – Outer Rectangles



Source: Author

Figure 10: Rectangle Method – Inner Rectangles



Source: Author

$$Space = \frac{\boldsymbol{space\ outer}\ \textbf{rectangles}\ \boldsymbol{RE} + \textbf{space inner rectangles}}{\mathbf{2}}$$

$$= \frac{\mathbf{0,245 + 0,17}}{\mathbf{2}}$$

$$= \mathbf{0,2075}$$

Sensitivity and Specificity

To verify the accuracy of the model, it is examined which of the correct target objects are identified by the model as such and which of the wrong target objects have been recognized by the model. The goal is a forecasting model that minimizes the wrong decisions and maximizes the correct decisions. Correct decisions are measured by the HIT rate , wrong decisions by the MISS rate.

Sensitivity of a test (HIT RATE)

The ratio of the correct decisions n_{ss} (suitable and selected) to the total number of suitable candidates $n_{ss} + n_{ns}$ (suitable and selected + suitable but not selected).

$$\text{Sensitivity} = \frac{n_{ss}}{n_{ss}+n_{ns}} = \frac{n_{ss}}{n_s} \quad \frac{\boldsymbol{Detected\ number\ of\ suitable\ applicants}}{\boldsymbol{Total\ number\ of\ applicants}}$$

Specificity of a test (MISS RATE)

Ratio of correct rejections n_{nn} (not suitable, not selected) compared to the total number of unsuitable candidates $n_{nn} + n_{sn}$ (non-applicant applicants, not selected and unsuitable candidates, but selected).

$$\text{Specificity} = \frac{n_{nn}}{n_{nn}+n_{sn}} = \frac{n_{nn}}{n_n} \quad \frac{\boldsymbol{Detected\ number\ of\ not\ suitable\ applicants}}{\boldsymbol{Total\ number\ of\ not\ suitable\ applicants}}$$

Sensitivity and specificity are calculated for each parameter of the forecasting model. It is important to separate the decision-relevant parameters from statistically insignificant influences, as otherwise, the result is an over-trained model. This is like memorizing sentences without knowing their meaning. In this case, the forecasting model must be critically reviewed (Feindt, M., 2015, page 59).

The thesis focuses on a-priori and a-posteriori probability according to the Bayes' theorem.*"Bayesian modeling has been applied successfully to many classifications or diagnostic testing problems in standard situations when the classifier score S is either binary (e.g., Dendukuri and Joseph, 2001, pages 158–167) or ordinal (Peng and Hall, 1996; Hellmich et al., 1998), but there is little material available to guide the Bayesian enthusiast when S is continuous."* (Krzanowski, W. J., 2009, page 125)

2.4.3 Application of Predictive Analytics

The computer application for data science and prediction SAP PREDICTIVE ANALYTICS is used in this thesis. This section summarizes the presentation of the different aspects of performing prediction analytics by the computer application. How to use the application for prediction is explained in the empirical part of this thesis.

The SAP Predictive Analytics application tracks several of SAP's goals (Bakhshaliyeva, N., 2017, page 78):

- Operationalization of predictive models,
- Making a common computer application for the user groups Data Scientist, Business Analyst, and Business User usable,
- Using big data and applying them in real-time using the SAP S / 4HANA database's in-memory technology.

SAP Predictive Analytics provides two basic evaluation modules: Automated Analytics and Expert Analytics. Automated Analytics is used for standard analyses based on automated data preparation and modeling processes with predefined algorithms (Bakhshalileva, N., 2017, page 82). The available mathematical-statistical methods include the creation of classification and regression models, clustering models, time series analyses and association rules. For the application, additional methods of artificial intelligence, e. g. the generation of artificial neural networks by given patterns and learning rules are part of the module Expert Analytics. In this module, own R-routines[6] can also support the generation of models. Automated analytics is used in this thesis. The focus in this thesis is not on optimization of mathematical-statistical models, but investigation of the influence of the information status changed by the prediction

[6] R-routines: Free programming language for statistical calculations with an integrated development workbench and graphical user interface.

on the controlling decisions. Models with sufficiently accurate confidence can be developed using Automated Analytics. The application of SAP Predictive Analytics is used by different user groups. Business users consume the results of the prediction without having to understand their calculation basis. The calculation logic plays a subordinate role, and the task concentrates on the preparation of detailed reports as well as *ad hoc* reports. The second user group are business analysts. Their mission is to develop models to solve business issues related to questions with the forecast of future events in the company and the business environment, using the methods available in automated analytics. In addition to their interpretation in the context of the company, the generated predictive models are used to develop recommendations for the management. This application group is the focus of this thesis. Whether the decisions based on the results of the prediction are made by the controlling or the management is an organizational issue. The focus of this thesis is the question of whether and to what extent the decision-making process is influenced by the prediction. The third group of users concerns data scientists. These employees have programming skills for predictive modeling. This allows them to expand existing models or program individual models. The explanation of the predictive analyses for the empirical thesis and their implementation is given in Chapter 5 in direct connection with the modeling.

Conclusion

The methodology for controlling has been increased. While in earlier times the summarization of the past fiscal year and the testing of the balance sheet statement and profit and loss statement have been the main tasks, the perspective of controlling is now in the future. Since the digital transformation generates new technology to transfer all data into digital data, computer applications that have been available for centuries have now assumed an important role. The availability of big data was a prerequisite for prediction, which was not available before. Predictive accounting supports smart prediction. Smart prediction will be the basis for future-oriented controlling.

3 THEORIES SUPPORTING PREDICTION AND DECISION-MAKING

The thesis is focused on the implementation of the results from pattern recognition in predictive models. The objective in pattern recognition is to detect the best classifiers which lead to the classification of the observed objects and their characteristics. For discrete target variables, a classification leads to classes for which reliable predictions can be calculated. Procedures for optimal decision-making are developed based on the prediction models for classification. Business decisions refer to the fact that they are made differently for different classes. The theory underlying pattern recognition is the support vector machine developed by Vladimir N. Vapnik (Vapnik, V., 2018). Outgoing from the support vector machine, the structured risk minimization has been developed by Vapnik and Chervonenkis (1991, pages 107–149). The purpose of the probability theory is to demonstrate that the risk of a wrong decision can be calculated and even minimized by identifying the relevant risk factors. Based on the Bayes' theorem, learning from experience and the connection to a theory of minimal risk of wrong decisions by Vladimir Naumovich Wapnik[7] and Alexey Yakovlevich Chernovenkis[8] has been further developed into the statistical learning theory (Vapnik, V., 2018). Vapnik's support vector machine theory has been programmed in the computer application SAP PREDICTIVE ANALYTICS®. In the second part of this chapter, the Bayes' theorem is presented, which is the basic theory for the minimization of a wrong decision.

3.1 Support Vector Machine for Classification and Regression

The formulation of an application problem is restricted by the following constraints: class conditional probabilities are mostly unknown. If the true probability of x ∈X is unknown or if the probability p (x | y) does not have the estimated form, the estimated probability $\tilde{p}$ $((x|y)$ could be arbitrarily bad. The desired properties of the classifier are assumed. Therefore, a training data set is needed in which-the distribution of the parameters is known.

Training Data Set T: P $(\Theta = \Theta_0)$ and P ($\Theta = \Theta'$) are known.

The known distribution of the training data set could be derived for the validation data set to prove a predictive model and perform for a data set in which the distribution is unknown. The

[7] * December 6, 1936, Soviet-American mathematician
[8] * September 7, 1938; † September 21, 2014, in Moscow, Russian computer scientist

classifiers are calculated via parameter estimation. It is assumed that the distribution p (x | y) is a certain distribution with a finite number of parameters Θ_y.

The distribution of the parameters p (x | y) from the training data set T is estimated: $\tilde{p}\ ((x|y)$ using a mathematical-statistical function.

The training data set is split into training data set (70% of the data) and validation data set (30% of the data). The mathematical-statistical methodology used is pattern recognition. Simply put, pattern recognition is the separation of marked patterns into two classes. The support vector machine (SVM) as linear programming is doing just that. Using the pattern recognition classification, a set of classes Q will be used. An example for a set of classes Q is the class q_1 with interested customers and q_2 with customers who are uninterested in a new product. The classification approximates the assignment of input variables x to discrete output variables y, which defines the class q. It is therefore important, for example, to recognize which characteristics determine an interested or uninterested customer. This determines the probability that an input variable x belongs to one or the other class q_i. The variable x is then assigned to the class with the highest probability. The assignment of the characteristics to a class is done by the classifier. This is an algorithm in the form of a mathematical-statistical function. The assignment process is a learning process based on training data. The so-called "supervised learning" sets the classes. In unsupervised learning, classes must also be learned by the algorithm. One of the forms of unsupervised learning is clustering. To concentrate only on the essential features, a feature reduction takes place in the learning process. This can be done by calculating correlation coefficients that define the degree of the linear relationship of variables. Support vector machines (SVM) will be discussed below as an approximating criterion function. The subjects' decision depends on several decision parameters. A decision criterion could be ~~a~~ specific personal property of the decision-maker. The decision criteria are understood as training objects, which belong in each case to a certain class of objects. Each training object is considered a vector, which is part of a vector space. The training objects (decision criteria) should be separated by a plane. A two-dimensional plane is determined by a support vector and two direction vectors. A hyperplane is correspondingly multidimensional. SVM calculates the hyperplane that splits the training objects into two classes (for example, YES or NO). For example, the value on the abscissa (A) and the value on the ordinate (O) added together could yield the vector X = A + B. The value for A (x) and the value for O (y) define a vector, which in turn are considered support vectors for the plane. To describe the hyperplane plane exactly mathematically, only those support vectors are considered which are closest to

the hyperplane. Generalization should be kept within limits. If too many support vectors are considered, one speaks of an over-trained model (overfitting), in which no conclusions can be drawn from the obtained influencing variables because there are so many. If there are N training data available with (x_1, y_1), (x_2, y_2), … (x_N, y_N), i = 1, …, N and $y_i \in (+1, -1)$ and $x_i \in \mathbb{R}$, $x_i \in \mathbb{R}$ corresponds to the input data, $y_i \in (+1, -1)$ corresponds the two classes in which the input data are assigned. $\mathbb{R}$ defines the set of all possible values of the random variable X (Fischer, J., 2007, page 4). The hyperplane is intended to split a set of objects into two classes. To do this, we searched for a function which correctly classifies the input variables:

$$f(x_{i)} = y_i$$

The input variables are considered vectors in a multidimensional vector space because there are multiple variables. In the case of linear separability of x, a hyperplane can be defined by the observed random variable x, which is defined by a normal vector w and a displacement b. A separating hyperplane H is defined as follows (Fischer, J., 2007, page 10):

$$H: \{x \in R\ (w, x\ ) + b = 0\}$$

x: Element of the total of complex objects.

w: Normal vector as a generalization of the objects.

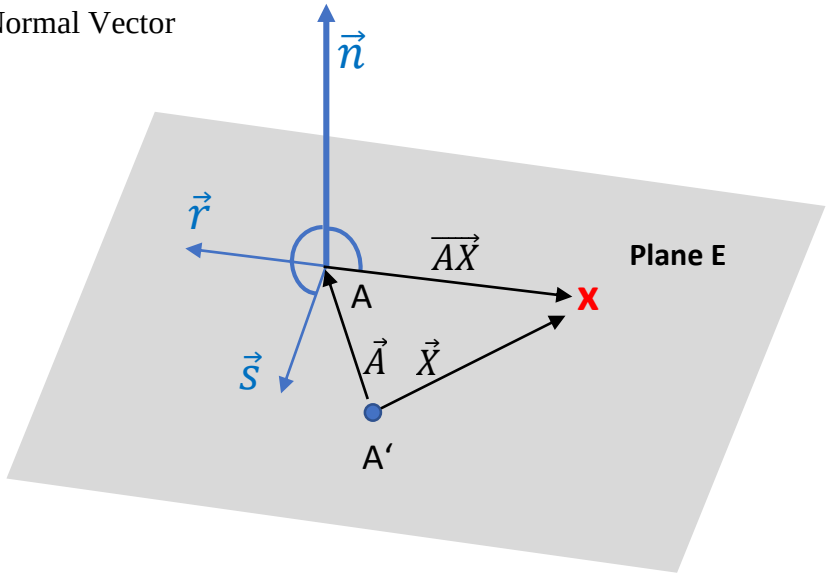
b: Shift

(W, x) is the scalar product with which two elements of a complex vector space are assigned a scalar (mathematical size). This allows geometric methods to be applied to abstract structures. Kowalczyk defined: *"A hyperplane is a set of points"* (Kowalczyk, A., 2017, page 25). The random sample survey results in the collection of random data are called random variables. The connection of the random variables is done by a vector that explains the commonalities of the random variables. This applies to the one-dimensional characteristic space R^1. The commonalities are defined in this application as a statistical context, abbreviated context. In the two-dimensional feature space R^2, the context is explained by a plane. In the multi-dimensional feature space R^n, the context is explained by a hyperplane. The hyperplane is a linear classifier that describes a finite set of learning patterns.

The objective is to define the hyperplane and to compute whether a point (observed random variable, like the characteristics of the overserved customers) lies exactly on the hyperplane. The plane equation can be defined in a normal form with a vector representation and with coordinate representation. The direction of a plane **E** is defined by two direction vectors $\vec{r}$ and

$\vec{s}$. The plane also could be defined by a normal vector $\vec{n}$ which is orthogonal to the plane. With a given plane **E**, the vector product of $\vec{r}$ and $\vec{s}$ defines the normal vector $\vec{n}$.

Figure 11: Normal Vector



A: Starting point of a normal vector

A’: Any starting point

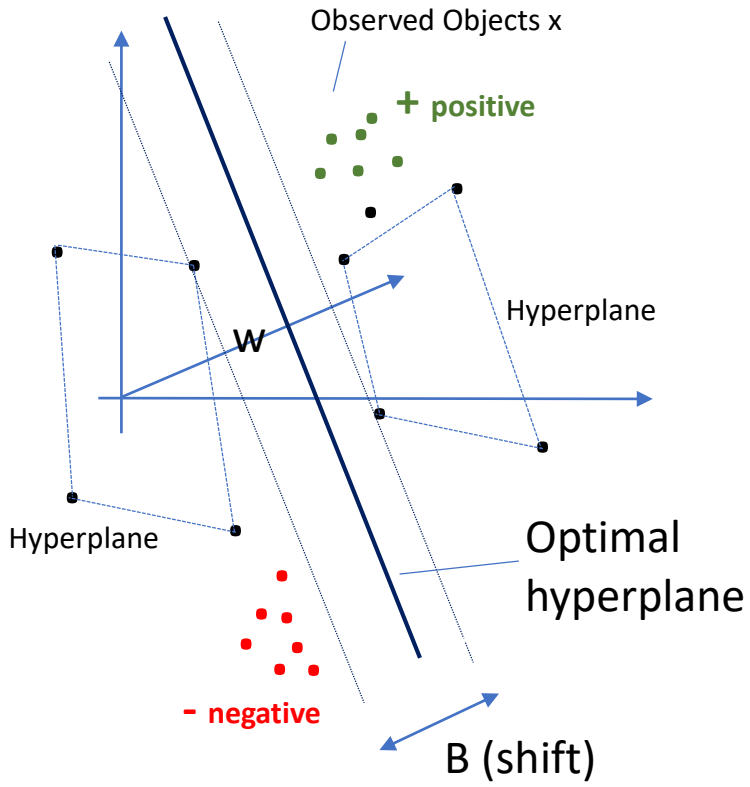
X: Observed random variable

Source: Heinert, M., 2010, page 4.

Any point X is part of the plane **E** when the connection vector $\overrightarrow{AX}$ starting from any starting point A’ on one side and the normal vector $\vec{n}$ on the other side are orthogonal **o**. On the other hand, if a random variable X is orthogonal to the normal vector $\vec{n}$, then the random variable X is part of the plane **E**. Whether two vectors are orthogonal to each other can be calculated with the scalar product.

These prediction models created in this thesis for classification are using the method of linear classification. The thesis is a subspace of the ambient space with one dimension less than the ambient space is in geometry named hyperplane.

Figure 12: Hyperplane

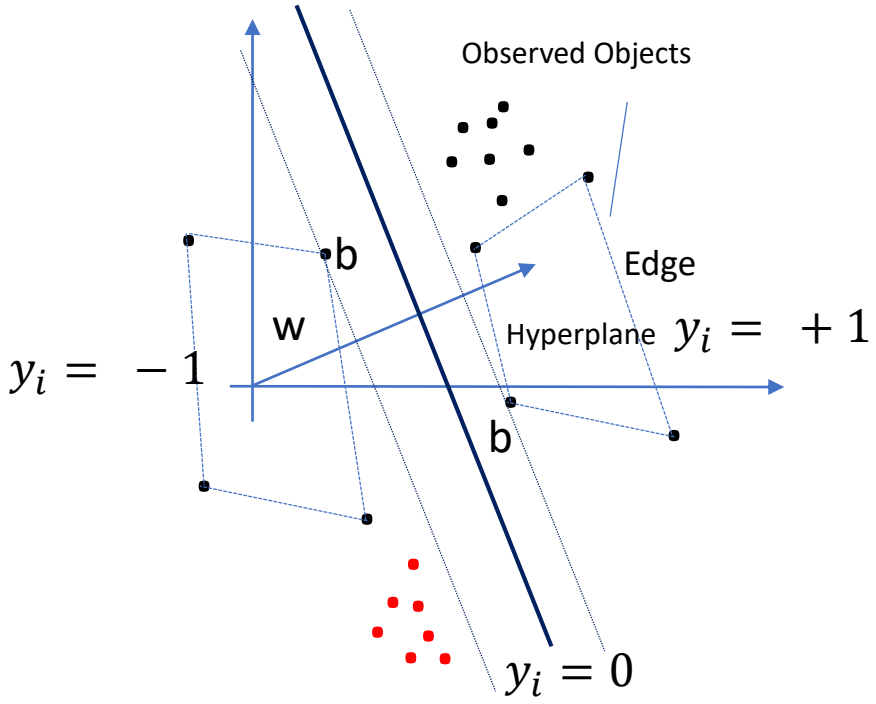


Source: Fischer, J., 2007, page 10. Modified by the author.

Since an exact description of the hyperplane is not possible, normalization takes place as a canonical form of the hyperplane. Therefore, the normal vector **w** is determined as a straight line through the origin, which defines a generalization. The minimization and classification in the hyperplane lead to the following correlations. Perpendicular to the normal vector **w,** the hyperplane which intersects the normal vector **w** at a distance **b** from the origin could be defined. Therefore, the normal vector **w** together with the distance **b** determines the hyperplane. Using the normal vector **w,** the condition is defined as follows:

$$Min_{i=1,\dots N} \mid (w, x_i) + b = 1$$

Figure 13: The Optimizing Problem in SVM



Source: Fischer, J., 2007, page 12.

A central problem of statistical learning theory is the ability to generalize. The empirical risk is that a small training error would lead to a big real error in pattern recognition.

The classification process is defined by the assignment function:

Classification Process ≡ Assignment Function $f\ (x,u): x \rightarrow y\ \in \{+1,-1\}$

x: Random variables from one of the two classes

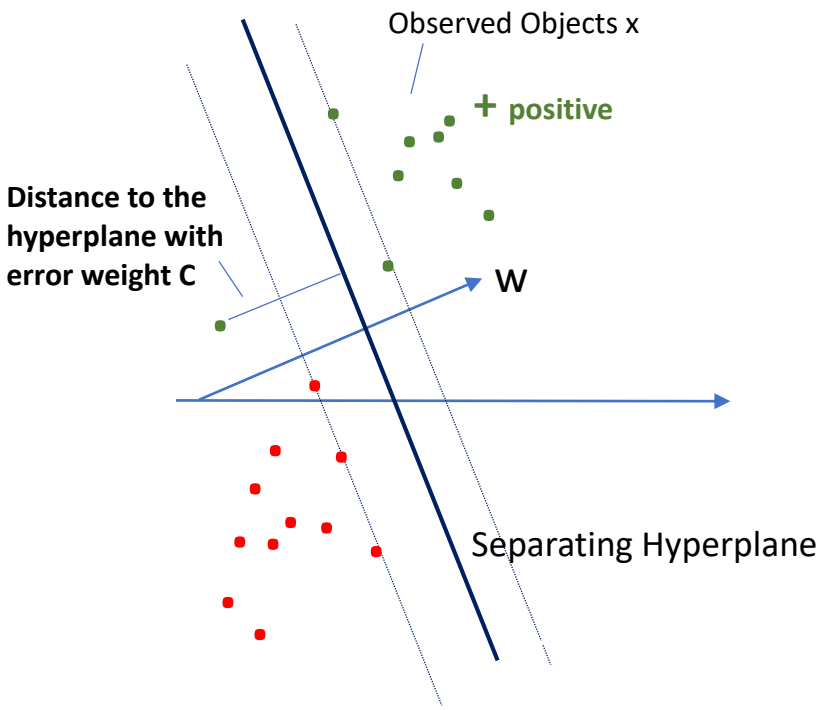
u: Parameter vector of the classifier

Linear classification is used in this thesis. The empirical risk (error rate) of a training data set with n observations of x_1, x_2, ..., x_n with a class membership to y_1, y_2, ... y_n is defined:

$$R_{emp}(n) = \frac{1}{2n}\sum_{i=1}^{n}|y_i - f(x_i,n)|\ \in\ [0,1]$$

A deterministic assignment function is needed that minimizes the expected risk of an assignment error. The central question is how well genuine risk can be estimated from empirical risk, structured risk minimization rather than empirical risk minimization (Fischer, J., 2007, page 8 ff).

Figure 14: Linear Separability with Accepted Errors



Source: Fischer, J., 2007, page 9. Modified by the author.

A linear separability will be used. Errors will be considered with a calculated penalty, using an error weight. The SVM seeks a hyperplane with maximum edge to determine the class of space as uniquely as possible. The optimization problem is the following: Maximize the edge to the optimal hyperplane so that the selected hyperplane represents the optimum. Because there are positive and negative distances to the optimal hyperplane, the optimizing problem is defined as follows:

Minimize the square distance of the hyperplane to the nearest point.

$$Min_{w\in\mathbb{R};b\in\mathbb{R}}: \frac{1}{2} \parallel w \parallel^2$$

with the condition of positive support vectors: $y_i \left((x_i, w) + b\right) \geq 1 \qquad \forall\, i = 1, \ldots N$

This is an optimization problem with several variables under constraints. The variables are the normal vector w and the distance b. This leads to the procedure of Lagrange multipliers (Joseph-Louis Lagrange[9]):

$$L(w, b, \alpha) = \frac{1}{2} \parallel w \parallel^2 - \sum_{i=1}^{N} \alpha_i (y_i \, (x_i, w) + b) - 1)$$

with the Lagrange multipliers $\alpha_i \geq 0$

The idea behind the process is the following: Every variable in the function is multiplied with a parameter, called the Lagrange multiplier L (compared to Kowalczyk, A., 2017, page 54). Every constraint is set equal to zero. For the variable w and b using the Lagrange multipliers, the following is defined:

$$\frac{\partial}{\partial b} = L(w, b, \partial) = 0 \qquad \frac{\partial}{\partial w} = L(w, b, \partial) = 0$$

The partial derivation for each constraint leads to a system of equations. The equation system can then be solved for example by the Gauss algorithm or in the insertion process. The solution of the system of equations yields the extreme points and thus the vector of the canonical hyperplane.

Prediction models are generated using learning procedures based on statistical learning theory. The aim is to divide observed objects into groups based on their characteristics. This is done through the so-called "classification". The learning algorithm for classifying data was developed by Vladimir N. Vapnik. The paradigm developed by V. Vapnik solves the following problem: In classic statistics, a large amount of statistical data and the associated a-priori information are necessary to calculate reliable statistical results and to draw reliable conclusions. The VC dimension of a set of indicator functions $\Phi_w, w \in \mathbb{W}$ is the largest number h of vectors that can be distributed to two different classes in each 2^h possible distribution using the set of functions (compare to Zhang, Ch., 2013, pages 1156–1160). The VC dimension is therefore the maximum number of patterns in an n-dimensional characteristic space $\mathbb{R}^n$, that can be correctly separated into two classes 1 (Vapnik, V., 2018, page 147). However, since fuzzy assignments from the set of functions are possible, the VC dimension defines the extent of unambiguous assignments (cardinality). The patterns are represented by hyperplanes as discussed above.

[9] Joseph Louis Lagrange, 1736–1813, Italian mathematician and astronomer

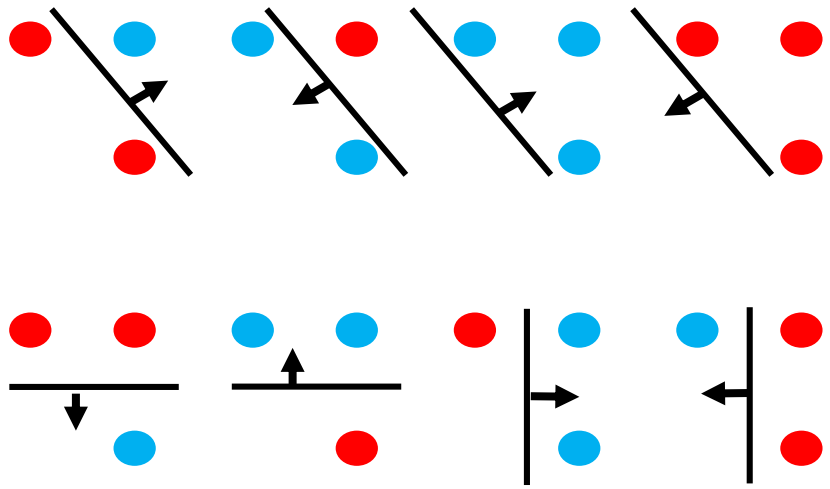
Example

A set of three indicator functions with $\Theta = \{\Theta_0.\Theta_1, \Theta_2\}$ could be chattered in the 2-dimensional characteristic space R^n. The 2^3 possibilities to define the patterns lead to three lines for a clear assignment into two classes.

In other words, three indicator functions in a two-dimensional characteristic space R^2 allow 8 patterns, which can be divided into 2 classes by 3 hyperplanes.

Therefore, the VC-dimension h = 3.

Figure 15: Shattering in the VC-Dimension



Source: Hlaváč, V., page 12.

As shown above, separating the observed features can be done in many ways. The way of separation results from the criterion according to which the separation is to take place. This criterion results from the target variable in the prediction model. A certain bandwidth, which results from the distance between the marks and the dividing lines, is permissible, but calculable. This results in the calculable risk with which future observed features lie in the calculated feature spaces. This leads to structured risk minimization according to Vapnik (Vapnik, 1991, pages 284–305). In the special case of regression, a series of random values can be described with a regression function of any complexity. The regression function defines the deterministic part of the prediction. However, the actual events, for example, decisions, are also influenced by random variables. This represents the stochastic part of the prediction. To be able

to calculate the risk of the stochastic influence, the following is considered: The logic consists of empirical logic and mathematical logic. To consider both, the description of the full risk is composed of the empirical risk and a confidence interval. The empirical risk is calculated using a mathematical-statistical function. The confidence interval is the range of values that were derived statistically from a random sample and that contains a calculated probability of an unknown but searched parameter of the characteristic space. The minimization of the guaranteed risk is determined by the empirical risk, defined using a mathematical-statistical function and the confidence interval. The confidence interval defines how accurate the measurements are, how robust the estimate is and how close the measurements are to the original estimate when the measurements are repeated. A multi-training data set T is given. The parameters "y" of the training data set "T" is classified by the classifier "x". The true, but unknown probability p (x, y) is substituted by $\tilde{p}\,(x, y)$. The real probability p (x, y) can be empirically predicted with a training data set divided into a training set and a validation set.

The Idea of Guaranteed Risk

The basic idea is to replace the true risk in forecasting through the empirical risk. Structural Risk Minimization (SRM) selects an infinite series of multi-dimensional statistical models. Then one minimizes the empirical risk in each model with an additional so-called "penalty term" for the size of the model to be able to control the capacity of the model. The preferred models are smaller in size (capacity) or complexity. A model with mathematical functions of higher-order polynomial becomes with increasing order even more complex and thus loses its importance. This prevents the usage of models which are so complex that they are no longer usable for evaluable statements. It prevents the so-called "over-fitting". This algorithm is performed by support vector machines explained in the previous section (Hlaváč, V., page 8).

Conclusion

The mathematical relationships in this section have been explained to provide an insight into how the support vector machine algorithm works. It was shown that the method of pattern recognition used in this thesis works with the vector geometry method. The observed features, for example, those of the customers, represent a multidimensional feature space. The SVM calculates multidimensional planes (hyperplanes) in such a way that the observed features can be separated so that they can be assigned to a hyperplane. In this way, the feature classes corresponding to distinguishable customer groups are found.

3.2 Decision Theory

A fundamental question in decision theory is the investigation of how decisions can be understood and how decisions can be predicted. Decision theory is a wide field of scientific research and goes in many directions. The investigation relates to cost-oriented decisions. The explanations about cost-oriented decisions that are often found in business administration are not pursued in this thesis. This thesis intends to contribute to an information-oriented decision theory. The subject of the thesis is the logical and empirical analysis of decision-making behavior (Bamberg, G., 2012, page 1). From a business point of view, a decision must then be defined with high importance, when there is a high impact on the performance and development of a company (Laux, H., 2014, page 3). The objective of this thesis is to develop predictive models for decision making. These decisions can be made safely and at/under risk. Secure decisions are made when complete information are available, which rarely occurs in practice. In this respect, decision models would only represent drastically simplified images of reality under security (Bamberg, 2012, page 41). The practice-relevant case are decisions under risk. Bamberg always discusses these risks as probabilities (Bamber, 2012, page 67). Decisions for which there is no objective evidence of the occurrence of risks would be made at high risk. This is where this thesis comes in. The occurrence of risks can be calculated. You can look at this positively or negatively. On the positive side, the risk does not occur, and on the negative, the risk occurs. By calculating the probability of the occurrence of risks, decisions can be made with a calculated risk, and thus better than with an incalculable risk. Such a risk is, for example, creditworthiness of the customer or termination of a contract. Decision theory has developed as interdisciplinary thesis and deals with the systematic investigation of the decision-making behavior of individuals and groups. The aim is to find out the system the decision-making process is based on. The decisions should be understandable and therefore objectively assessable. This leads to the decision-makers themselves making rational decisions with the knowledge of decision theory. Jürgen Weber believes that one of the central tasks of controlling is to support management in making rational decisions. (Weber et al. 1999, page 13). A rational decision is defined when the decision-maker considers all decision alternatives and weighs the commitment for a decision against the effects of the decision. The goal of descriptive decision theory is to empirically investigate decision-making behavior and to gain insights into decision-making behavior in defined decision-making situations. It is the goal of descriptive decision theory to be able to predict decision behavior in specific decision situations. In this thesis, decisions are made using a questionnaire in two decision-making situations: before knowing

the result of the predictive analysis of the company's customers and with knowledge of the analysis. This follows Bayes' theory of the posterior decision probability. According to this theory, the probability of correctly made decisions at a higher information level is significantly higher than without this knowledge. This is explained in chapters 4.2 and 4.3 of this thesis. The goal of the prescriptive decision theory is to support decision-makers in the respective specific decision-making situations. There is a certain abstraction of the specific decision situation (Laux, H., 2014, page 4). In this thesis, abstraction takes place via data compression. Data are compressed using mathematical-statistical methods of classification, clustering, and regression analysis. Decision-makers are supported by developing a decision-making model. In this decision model, the decision-making processes are optimized so that optimal decisions can be made regarding the target variables. The practical reference is made by deciding on the allocation of berths for a company in the marina branch. In this thesis, two groups of decisions are processed – external decisions of the company's customers and internal decisions of the company's controlling and management. Customer decisions relate to the renewal of the contract, the internal decisions to the allocation of berths to applicants. The controllers support the decision-makers in deciding on which of the potential customers requesting a berth they would assign a vacant berth to. The decision-making behavior of customers who have opted for an extension of the contract is explored in this thesis through the application of predictive analysis. Descriptive decision theories, therefore, reject the assumption of absolute rationality of human decisions. Human factors are also included, e. g. personal characteristics of the person and their living conditions. For this thesis, there is a data set with customer data used to develop a predictive model. The pattern recognition method is used to identify customer groups. This method is explained in the previous sections of this chapter. It is assumed that the results of the predictive analysis about the customers are included in the decisions.

3.2.1 Probability Theory Based on Bayes' Theorem

The Bayes' theorem has different aspects. The first interesting aspect is Bayes' concept of probability, which deviates from Bayes' theorem (Chapter 4.1.2). Bayes developed the Bayesian statistic and an estimator for the so-called "a-posteriori distribution" and finally a classifier which minimizes the likelihood of a wrong decision. The Bayesian probability concept interprets probability as a degree of personal conviction. This concept of probability differs from the so-called "frequentist probability concept, which interprets probability as

relative frequency. The frequentist probability concept calculates the probability from the relative occurrence of an observed feature of a high number of randomly conducted random experiments. The result is a-priori probability, which is acquired based on previous knowledge. This includes, for example, the frequency of a rolled number in a fair cube. Bayesian statistics work with a-posteriori probability. This describes the knowledge of an unknown environmental state Θ (scenario) after observing a random variable X which is statistically dependent on the scenario Θ. An unknown scenario **Θ** will be estimated with the observation of sampling **x** of the random variable **X**. The probability distribution of **Θ** is given before the observation of **x**, which is the "a-priori distribution". On the condition that **Θ** will take the value Θ_0, the probability function is written as:

$$f(\Theta_0)$$

There are two possible scenarios, TRUE and FALSE. This means whether

Θ will take the value Θ_0 or not. The "a-posteriori distribution" defines the distribution of **Θ** under the condition that the value **x** has been observed for the random variable **X**. Using Bayes' theorem, the "a-posteriori distribution" is calculated from a-priori-distribution based on the following aspects: The distribution of the random variable **x** of **X** in the sample is known, the condition has been set, that **Θ** will take the value Θ_0. For discrete a-priori distributions with, for example, binomial distribution, the following applies: The discrete "a-priori-probability" that **Θ** will take the value Θ_0 is defined as follows: $P(\Theta = \Theta_0)$**.** The "a-posteriori-probability" for Θ_0 describes the state of knowledge about an unknown environmental state θ a-posteriori, that is, after observing a random variable x of X that is statistically dependent on θ. The "a-posteriori-probability" for Θ_0 with **x** of **X** has been observed and is defined:

$$P(x)$$

The "a-posteriori probability" for scenario Θ_0 can be calculated according to Bayes.

(Held, L., 2008, page 264):

$$P(x) = \frac{f(\Theta_0) * P(\Theta = \Theta_0)}{\sum_{\Theta' \in \Theta} f(x = \Theta') * P(\Theta = \Theta')}$$

If the possible scenarios Θ' *and* Θ_0 are discretely distributed evenly, the following is applied:

$$P(\Theta = \Theta_0) = P(\Theta = \Theta') = 0{,}5$$

Namely, the probability function for the conditional distribution of the sample, multiplied by the probability that Θ will take the value Θ_0, is proportioned to the sum of the probabilities of all scenarios for Θ. The theorem about a-posteriori probability is explained with a numerical example. Customers are aware that either 40% or 60% of them are interested in the product. To find out whether it is more 40% or 60%, 11 customers (**n**) were interviewed. Four customers (**k**) were interested (license plate "red") and seven customers were not interested (license plate "black"), which is defined by scenario **Θ**. The random variable **X**: "Proportion of interested customers" is binomial distributed and could only have the value **Θ** = 0,4 or **Θ** = 0,6. This distribution is given in the following example: The probability of whether **Θ** will be 0,4 or 0,6 is discretely **distributed equally** and it therefore applies:

$$P\ (\Theta =\ 0{,}4) = \mathrm{P}\ (\ \Theta = 0{,}6) = \mathbf{0{,}5}$$

To find out whether the hit rate of four from the 11 interviewed customers represents 40% or 60% of the interested customers, the "a-posteriori-distribution" can be calculated based on the form for binomial distribution. Based on binomial distributed X = x, the probability for Θ = 0,4 is:

$$f\ (\Theta =\ \Theta_0) = \left(\frac{n}{k}\right) *\ \Theta_0^4 *\ (1 -\ \Theta_0)^7$$

$$f\ (\Theta = 0{,}4) = \left(\frac{11}{4}\right) * 0{,}4^4 * 0{,}6^7 = \left(\frac{11*10*9*8}{1*2*3*4}\right) *\ 0{,}0256 * 0.0279936 = 0{,}236$$

Based on binomial distributed X = x, the probability for Θ = 0,6 is:

$$f\ (\Theta = 0{,}6) = \left(\frac{11}{4}\right) * 0{,}6^4 * 0{,}4^7 = \left(\frac{11*10*9*8}{1*2*3*4}\right) *\ 0{,}1296 * 0.0016384 = 0{,}07001$$

Using the Bayes' theorem for discrete distribution, the "a-posteriori-probability" for Θ = 0,4 is:

$$P\ \ (x = 4) = \frac{0{,}236 * 0{,}5}{0{,}236 * 0{,}5 + 0{,}07 * 0{,}5} = 0{,}77$$

The "a-posteriori-probability" for Θ = 0,6 is:

$$P\ \ (x = 4) = \frac{0{,}07 * 0{,}5}{0{,}236 * 0{,}5 + 0{,}07 * 0{,}5} = 0{,}23$$

Results:

The likelihood that 40% of customers are interested is about 77%. Accordingly, the likelihood that 60% of customers are interested in the product is only 23%. This knowledge-gaining is also an objective of this thesis.

3.2.2 Information Status and Future Scenario

In common usage, statements about the occurrence of future scenarios "S_i" are often formulated as "it is more likely" or "it is quite sage", etc. In business, such statements are not useful because business decisions need a reliable basis (Laux, H., 2014, page 342). In this environment, it is necessary to specify estimates more precisely to make their effects predictable. To assess the assessment of a future scenario, decision theory distinguishes between the direct and the indirect method. The direct method is the direct questioning of the decision-maker. Even if a rough fuzziness can be reduced through well-developed interview methods and intelligent, self-controlling questionnaires, fuzziness cannot be eliminated. By contrast, the indirect method uses statistically determined probabilities to predict future decision-making behavior. The mathematical-statistical methods are used to improve the information status $\mathbf{I_i}$ of the decision-maker.

$\mathbf{I_i}$ identifies all possible information statuses with i = 1 ... n.

$\mathbf{S_j}$ identifies all imaginable scenarios with j = 1 ... n

The information is composed of individual indicators (i = 1, ... n) that are assumed to have an impact on the decision. Non-decision-related indicators should be eliminated for reasons of complexity reduction. These relationships are discussed in more detail in the following chapter about prediction with SAP Predictive Analytics. Given the information status, before the search and selection of further information, the likelihood of future scenarios is called **a-priori probability**. In the previous section, a-priori probability was justified with certain plausibility. With regard to decision theory, this definition is applied to the information status. If additional information are hypothesised and selected, the likelihood of the arrival of future scenarios is called **a-posteriori probability**. The a-posteriori probability characterizes the review and adaptation of a previous estimate of the likelihood that a scenario will occur. Therefore, there

is probability before and after the actualized information status[10]. Since the decision-maker getting an updated information status will determine the direction and extent of the previously defined likelihood of future scenarios, the predictive quality of the indicators is of crucial importance. Assuming, for example, that the name and contact data of a possible customer, who merely requests to rent a berth, are known. The likelihood that the potential customer will extend his contract is as high as the likelihood that the potential client will not extend the contract. The following applies:

$$I_1 = Information\ Status\ before\ Additional\ Information$$

$$I_2 = Information\ Status\ after\ Additional\ Information$$

$$S_1 = Customer\ will\ extend\ Contract$$

$$S_2 = Customer\ will\ not\ extend\ Contract$$

A: The decision-maker collects additional information without applying scientific methods of forecasting.

Probability of the scenario and information status:

$$p(S_1) = p(S_2) \qquad \text{and} \quad p(S_1) = p(S_2)$$

Conclusion: Such collected additional information have no impact on the probability of future scenarios.

B: Application of scientific forecasting methods

A decision-maker can create predictive analytics using appropriate computer applications such as SAP Predictive Analytics. It is possible to check the forecasting quality using key figures. This was explained in the previous section "3.3 SAP Predictive Analytics". If the predictive analysis calculates a higher probability than 0.5 for the potential customer to extend the contract on basis of the additional indicators of the potential customer, then the decision-maker should trust the forecast. The posterior probability is thus higher than the a-priori probability. In other words, the higher the forecast quality, the higher the probability of the arrival of the forecasted scenario (for example, the customer will extend his contract). Bayes' principle: maximizing the statistically expected value for the expected benefits when deciding. The expected value

[10] Laux et al. 2018, page 347, denote the subjective evaluation of the decision-maker on the arrival of a future scenario as a probability judgement. In this thesis, the expression "likelihood of future scenarios" is used.

expresses the probability of the expected benefits (Jesche, B., G., 2017, page 73). Of all alternative decisions, one decision is optimal, for which the sum of all weighted expectation values concerning the target variable is maximum (Dörsam, P., 2013, page 43). He states that the expected value results from the average of all expected values if the decision situation were repeated an infinite number of times. In this thesis, the target variables on which the expected value is based are customer groups with the highest share of sales, reliable payment behavior, and long-term contractual partnership.

3.2.3 Scenario Probability According to the Bayes' Theorem

The Bayes' theorem was already discussed in the context of probability distribution. One of the examples was the probability with which the customer's interest in a product can be expected. For two exclusionary scenarios, in the example of whether 40% or 60% of the customers are interested in the product, 11 customers were interviewed. Of the 11 interviewed customers, four were interested in the product. The interview resulted in a new information status with a-posteriori probability. To find out whether the four positive customer responses from 11 interviewed customers were expressing an interest in the product, the objective was to determine whether more than 40% of customers or more than 60% of customers were interested. The binomial distribution is used as a probability calculation (posteriori). The result of the calculation in Chapter 4.1, that over 40% of customers were more interested than 60% of customers, is not surprising. Ultimately, four positives of 11 replies account for 36.37%. However, by using the Bayes' theorem of disjoint discrete probabilities for a-posteriori probability, probability can be calculated accurately. The relationship between a-priori probability at information status I1 and a-posteriori probability at information status I2 can be described with the Bayes' theorem. Bayes describes the stochastic dependence between information status and scenario from a-priori probability to a-posteriori probability. According to the Bayes' theorem, the a-posteriori probability of alternative scenarios can be calculated from the a-priori probability of the scenarios by referring to the correctness of the information status. The opinion of the author of this thesis is that a-priori information status is one before prediction, and a-posteriori information status after prediction. For this purpose, the following is calculated: For each information status I_1, I_2, ... I_i, the respective probabilities of the possible alternative scenarios are calculated. At first, the possible information status is the status before and after prediction (I_1, $I_2 \in I$). However, the information status could be predicted over several periods and continuously evaluated. In this case, more than two information statuses are

possible. The scenarios may consist of several indicators. If the scenario consists of only one indicator, for example, if a contract is renewed or a contract is not renewed, the calculation is done for two scenarios, S_1 and S_2 in combination with each information status I_i. Regarding $\{I_1, I_2, \dots I_n\}$ the possible information status and $\{S_1, S_2, \dots S_n\}$ scenarios, the Bayes' theorem is defined (Laux, H., 2018, page 351), and changed by the author of this thesis:

$$p\,(S_s|I_i) = \frac{p\,(I_i|S_s) * p(S_s)}{\sum_{s=1}^{n} p\,(I_i|S_s) * p(S_s)} \qquad (i = 1, \dots n)$$

The a-posteriori status $p(S_s)$ is calculated from the a-priori status $p(\,S_s)$. The information status is the leading factor for the calculation of each combination with information status I and scenario S. In the case of two information statuses, two scenarios apply: Predict $\mathbf{I_1}$ or predict $\mathbf{I_2}$ concerning scenarios $\mathbf{S_1}$ and $\mathbf{S_2}$, the probability of the following combinations of I_i and S_i will be calculated (Laux, 2018, page 352).

Calculation of the probability of a-posteriori scenarios:

$$p\,(\,S_1|I_1) = \frac{p * 05}{p * 0.5 + (1-p)_* \, 0.5} = 0.5$$

$$p\,(S_2|I_1) = \frac{(1-p) * \; p * 05}{p * 0.5 + (1-p) \; * \; 0.5} = (1-p)$$

$$p\,(S_1|I_2) = \frac{(1-p) * 05}{(1-p) * \; 0.5 + p * 0.5} = (1-p)$$

$$p\,(S_2|I_2) = \frac{p * 05}{(1-p) * \; 0.5 + jp * 0.5} = p$$

The benefit for the controller is the following: Using the predictive analysis performance, the probabilities of a correct decision will increase compared to the decision made in the same decision scenario based on the previous information status. The prerequisite is that the results of the predictive analytics are reliable.

Conclusion

Risky decisions are a classic topic in decision theory. However, the decision theory is based on decision rules according to which decisions are made based on the willingness of decision-makers to take risks if a benefit is to be expected. This thesis, on the other hand, contributes to making risky decisions for which the risk can be calculated objectively. In this thesis, the basis for decision-making are prediction models. Decision models are generated using pattern recognition. The risk is minimized, because the used algorithm for pattern recognition is the support vector machine, which has its theoretical basis on the structured risk minimization theory.

4 EMPIRICAL RESEARCH – PREDICTIVE ANALYTICS

The practical part of this research is the implementation of the theoretical and conceptual foundations. The theoretical basis for generating meaningful prediction models is now based on the data provided by a marina company in Croatia. The theory of a-posteriori decisions made, which are of higher quality than decisions without an adequate information status, is implemented by presenting company's controllers in a case study with decisions they make with and without knowledge of the results of the prediction models.

4.1 Requirements for Empirical Research

A mandatory requirement for a successful predictive analysis is the availability of empirical data. Regarding the objective of this research, customer data were used as the basis for customer-related decisions. This was possible with the provision of customer master data from the Adriatic Croatia International Club in Croatia (ACI). ACI is the largest company in nautical tourism in the Mediterranean and, with 22 marinas, the leading marina operator.

"The company ACI d. d. started business as ADRIATIC CLUB YUGOSLAVIA Brijuni, a company for nautical tourism. The company's acronym at the time was ACY. It was established on July 1st, 1983, with the aim of implementing the program of the development of capacities and the accompanying offer of nautical tourism services on the eastern coast of the Adriatic Sea. In June 1994 the company underwent privatization and was registered as a joint-stock company with the new name ADRIATIC CROATIA INTERNATIONAL CLUB d.d. Opatija – or ACI d. d."[11]

With over 1000 islands, Croatia is a paradise for nautical tourists. However, there is also competition, not only in Croatia. Prices have been rising for many years. "Croatia has already overtaken Monaco in terms of daily prices"[12]. The Croatian state is also a price factor. While the sojourn fees increased drastically in 2018, the fees for larger ships have been reduced, while users of smaller ships will have to accept an increase again in 2019.[13] A leading objective in ACI's controlling is the optimal utilization of the berths and avoidance of disputes related to late or unsettled payments by clients. There is a strong demand for berths, so ACI can select the

[11] Compare to: https://www.aci-marinas.com/en/povijest/
[12] Compare to: https://marine-project.com/de/liegeplatze/marinas-kroatien/
[13] Compare to: „Yacht online ", download on January 5th, 2020. https://www.yacht.de/reise/news/kroatien-korrigiert-gebuehren-fuer-yachten-2019/a119124.html

applicants for vacant berths. The decisions are based on the experience of the ACI controllers who prepare the information and make suggestions to the managers, the director, and the board. On the other hand, nowadays, it is also possible to book a berth using an online service provided by the marina company. In relation to the online service, it is recommended to use an expert system which checks the characteristics of the applicants regarding corporate goals about long-time customer relationships, timely payments, and high sales volume. Therefore, customer data have been checked to find out which customer variables influence the conclusion of long-term contracts and guarantee good creditworthiness of the customers.

The development of nautical tourism in Croatia has been growing steadily for many years. Management and controlling are decisive factors in ensuring the continuation of positive development and that Croatian marinas can successfully defend themselves against the international competition (Luković, 2013, page 201, Luković, 2019, page 40). From the point of view of operational controlling, the goal is to minimize the business effort for the allocation of berths, negotiate contracts, and create payment reminders for customers who are in default. The tactical controlling perspective is aimed at the customer group with which the business can be increased and optimized. At the level of strategic controlling, it is a matter of securing the company in the long term by reducing payment defaults and achieving adequate company growth through continuous customer analysis with meaningful results. The IT application SAP Predictive Analytics is used as a decision support system. The methodology used is the case-study method. Gaining knowledge through case studies is one of the qualitative research methods. Experts are interviewed on selected topics – in this research, the allocation of berths and classification of customers in customer classes – in the form of decisions to be made. The results should support the research question – higher accuracy of a-posteriori decisions in comparison with a-priori decisions (Borchardt, 2007, page 34). To find a definition for the research method of case studies, Robert K. Yin clarifies the following in his book (Yin, R. K., 2018, page 14):

"The essence of case study, the central tendency among all types of case study, is that it tries to illuminate a decision or set of decisions: why they were taken, how they were implemented, and with what result."

The quote taken from Yin's book refers to W. Schramm (1971): "Notes on case studies of instructional media project". Working paper for the Academy of Educational Development, Washington, D.C.

"This definition thus cites cases of "decisions" as the major focus of case studies." (Yin, R. K., 2018, page 14). This research will follow this definition.

The definition of qualitative research with case studies according to Yin (2018) adopts a theory-based approach and does not exclude quantitative methods. According to Yin, the steps in qualitative research are Plan – Design – Collect – Analyze. When planning, a relevant situation for a case study has to be identified. In this research, the relevant situation of making decisions on whom to allocate a vacant berth was set as the basis. The case study design in this research is analyzing the decision behavior based on two different information statuses. Data collection means that this task is not only meant to sample a high volume of data, but also to analyze data quality relating to lack of plausibility and contradictions. The result of the case study in this research is the comparison of the quality of decisions with and without additional information. The interpretative paradigm is interpreted in this research in such a way that the subject of interpretation is the behavior of the customers of the marina company. This refers to the payment behavior of the customers and the conclusion of yearly contracts as well as the renewal of contracts. The existence of much more interesting variables than data points is handled through a predictive analytics system which detects the explanatory variables to the target variable. The observed characteristic values converge into clusters generated by the system. The case study is guided by the theory of significantly better a-posteriori decisions. In the first step, a data set was analyzed to find possible errors. The time range of the data set is from 2013 to 2019. With a volume of 26,652 data sets, it was possible to create prediction modules with very good values for the key performance indicators 'predictive power' and 'predictive robustness'. The data ensured reliable results using the statistical methods of clustering, regression analysis, and classification. As mentioned before, these customer data were used for customer-related decisions. The data set with customer data is important for the marina company to recognize its customers. These data can be used to gain important information about the characteristics of customers and customer behavior. Using pattern recognition as a data science method, customer patterns and customer segments can be identified. Findings in relation to customer behavior in this research refer to payment behavior, the intention to conclude long-term contracts or contract extensions, and the classification of customers into sales groups. The following table provides an overview of the data set made available by the marina company. The relevant variables are identified in the following table as explanatory variables, target variables or variables that are not relevant for this research.

Table 1: Analyzing the ACI Data Set

FIELD-ID	FIELD NAME	Useful as a variable in this research
Contract-ID	Identification number of the contract	Not important
Vessel ID	Identification number of the vessel	Not important
Vessel Name	Name of the vessel as given by the customer	Not important
Vessel Length	Length of the vessel in meters	Useful as an explanatory variable
Vessel Manuf. Date	The manufacturing date of the vessel	The manufacturing date is the basis for the age of the boat.
Vessel Age	Age of the boat	Useful as an explanatory variable
Contract Start Date	Start date of the contract	Useful as an explanatory variable
Contract End Date	End date of the contract	Not used
Client-ID	Client's number	Not important
Client Name	The family name of the client	Not important
Client Birthday	Date of birth of the client	Not important
Client Age	Client's age	Not important, because of too many different values which could not be categorized
Client Age Group	Classification of customers into age groups.	Useful as an explanatory variable
Client First Visit	Date of the first contact with the client	Not important
Client Address City	Client's city	Useful as an explanatory variable
Client Citizenship	Citizenship of the client	Useful as an explanatory variable
Foreign Guest	Detects whether the client is a	Not used

	foreigner (1) or resident (0).	
Client Type	Client Type F = Company (firm) or Client Type = P (private person)	Useful as an explanatory variable
Charter Vessel	Defines whether the client chartered the vessel (= TRUE) or has an own vessel (= FALSE)	Useful as an explanatory variable
Contract Type 1	- Yearly contract (YC) - Monthly contract MV	Useful as a target variable
Contract Type 2	- new - renewal	Useful as a target variable
Contract Type 3	Detailing the type of contract regarding duration and seasonal reference.	Not used
Delay	Number of days payable outstanding	Useful as a target variable
Contract Start Year	Year in which the contract was concluded	Not important
Contract Start Month	Month in which the contract was concluded	Not important
Debt Amount	The amount which the client owes to ACI	Useful as an explanatory variable
Debt amount in %	The ratio of debt amount to the contract amount	Not used

Debtor	Indicator: Debtor = 1 if debt amount > 0 Debtor = 0 if no debt amount has currently been detected.	Useful as an explanatory variable
Debt period	Number of days the client owes ACI.	Useful as an explanatory variable
Vessel Type	K = Catamaran MY = Motor Yacht S = Sailboat	Useful as an explanatory variable
Advance	TRUE, FALSE	Useful as an explanatory variable

Source: Author

Note: Negative Debt Amount

To gain a better understanding of the prediction model results, it is necessary to know why debt amounts in the customer data are lower than zero. If the invoice amount is lower than the contract amount, the debt amount is stored with a negative sign. Therefore, a negative debt amount means a **"residual claim"**.

4.2 Creation and Validation of the Model

This section introduces the algorithms used in the computer application SAP Predictive Analytics in Automated Analytics mode. From a theoretical point of view, it is based on structured risk minimization developed by Vladimir Vapnik and Alexey Chervonekis between 1960 and 1990 (Vapnik, V., 1991). The learning algorithm developed on this mathematical-statistical basis is the support vector machine (SVM), which was developed by Vapnik and his colleagues at the AT&T Bell laboratories between 1992 and 1997 (Cortes, C., Vapnik, V., 1995). The SVM algorithm is explained in the thesis. The target function is introduced with its parameters. Secondly, the thesis includes an introduction of the learning procedure used in the computer application SAP Predictive Analytics. Six models have been developed and explained below. The model design relating to the target variable and explanatory variables is presented below.

The objective function used for Model A, B, and C by machine learning is the k-Means algorithm to divide the data set into **k** partitions in such a way that the sum of the squared deviations from the cluster centroids μ_i is minimal. Mathematically, this corresponds to the optimization of the function (Ester, M., 2013):

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \quad \text{with constraint set in model A and model B.}$$

SSE – Sum of Squared Distance to Euclid Center, c_i = cluster i, x = observed variable value of the explanatory variables.

μ_i = cluster centroid, $\left\| x_j - \mu_i \right\|^2$ = Squared distance of the observed variable x_j to the cluster centroid μ_i

In an interview with the controllers, they explained which target variables are the focus of controlling in the company. It was also discussed which explanatory variables could be considered. However, on this basis, objective functions have the explanatory variables which the computer application has identified during the learning process in the process of creating the prediction models as those with the highest correlation to the respective objective variable.

Model A: Target variable: Yearly Contract. **Explanatory variables:** x_1 = Vessel Type; x_2 = Vessel Age; x_3 = Client Age; x_4 = Foreign Guest; x_5 = Prepayment; **Filter:** Client Type = P (Private)

Model B: Target variable: Yearly Contract. Explanatory variables: x_1 = Vessel Type; x_2 = Vessel Age; x_3 = Vessel Length; x_4 = Contract Type 2 (new contract or renewed contract); **Filter:** Client Type = F (Firm)

Model C: Target Variable: Contract Renewal. x_1 = Vessel Type; x_2 = Vessel Age; x_3 = Vessel Length; x_4 = Marina. **No Filter.**

Model D: The objective of Model D is to learn the characteristics of customers who make late payments. With zero-days maturity date, one is already in default of payment upon receipt of the invoice. Commercial standards usually allow a payment term of 14 days. Additionally, COVID-19 has changed the consumers' payment acceptance (Ford, 2020, page 2). To focus on the significant debit items within the scope of this research, a filter was set to number of days greater than or equal to 100 days in the model. Since the target variable is continuous, regression is used as a method. The multiple linear regression form used in this prediction for Model D is the following (Auer, L. v., 2016, page 157):

Target Function

$$y_{i,j} = \alpha_i x_i + \beta_i x_i + \gamma_i x_i + \delta_i x_i \text{ with i = 1, … n and constraint } y_i < 0$$

$y_i = number\ of\ days\ \ payable\ outstanding, invoice_i, client_j$

$\alpha = citizenship\ client_i$

β = *client type* $client_j$

$\gamma = vessel\ age\ client_j$

$\delta = contract\ type\ client_i$

$x_j = client_j$

The equation does not define a straight line but a multidimensional plane; the so-called "hyperplane". However, the polynomial degree was set to 1. The constraint about days payable outstanding greater than zero remains from possible values below zero if the customer paid in advance.

The objective of **Model E** is to identify clients who are expected to make bigger sales. The target variable in the given data set is CONTRACT AMOUNT. An insight into the file has shown that all values are greater than zero and that, therefore, there are no implausible values. A contract is not extended but renewed. If a contract is renewed, a new data record is created for the same customer. The calculation of sales per customer, therefore, requires the cumulation of contract amounts per customer ID. This is not done in this research. It is examined which customers hold the highest contract values.

The form of multiple linear regression used in this prediction for model E is the following (Auer, L. v., 2016, page 157):

Target Function

$$y_{i,j} = \alpha_j x_j + \beta_j x_i + \gamma_i x_i + \delta_i x_i \text{ with i = 1, … n and j = 1, … n}$$

$y_{i,j} = contract\ amount, contract_i\ with\ client_j$

α_j = *vessel type* $client_j$

β_j = *vessel length* $client_j$

γ_j = *vessel age* $client_j$

δ_j = *contract type 2* $client_j$

The investigation is performed for both client types. The contract amount has been set as the target variable in this model. A single client j may have several contracts.

The objective of **the Model F** is to classify the customers into different customer groups:

- Smaller order volume with a short duration of a few days,
- Medium order volume with an average duration of a few weeks,
- Larger order volume with a rather long-term duration of one to several years.

Classification of customers into these three customer groups can be the basis for the allocation of berths depending on their availability. In the development of the decision-making models, however, the aim is to differentiate between marketing campaigns. Customer group-specific marketing campaigns can reach the target groups better and thus be used more successfully.

A form of multiple linear regression used in this prediction for **model F** is the following (compare to: Auer, L. v., 2016, page 157):

Target Function

$$y_{i,j} = \alpha_j x_j + \beta_j x_i + \gamma_i x_i \text{ with i = 1, … n and j = 1, … n}$$

$y_{i,j} =$ *charter vessel*$_i$ *with* $client_j$

α_j = *vessel type* $client_j$

β_j = *vessel length* $client_j$

γ_j = *vessel manufacturing date*

The investigation is conducted for client types ‘PRIVATE’. The characteristic ‘CHARTER VESSEL’ has been set as the target variable in this model because, while creating and testing the prediction model, experience has shown that clients with a charter vessel prefer short-term contracts of a smaller volume and vice versa.

4.3 The Learning Process in Predictive Analytics

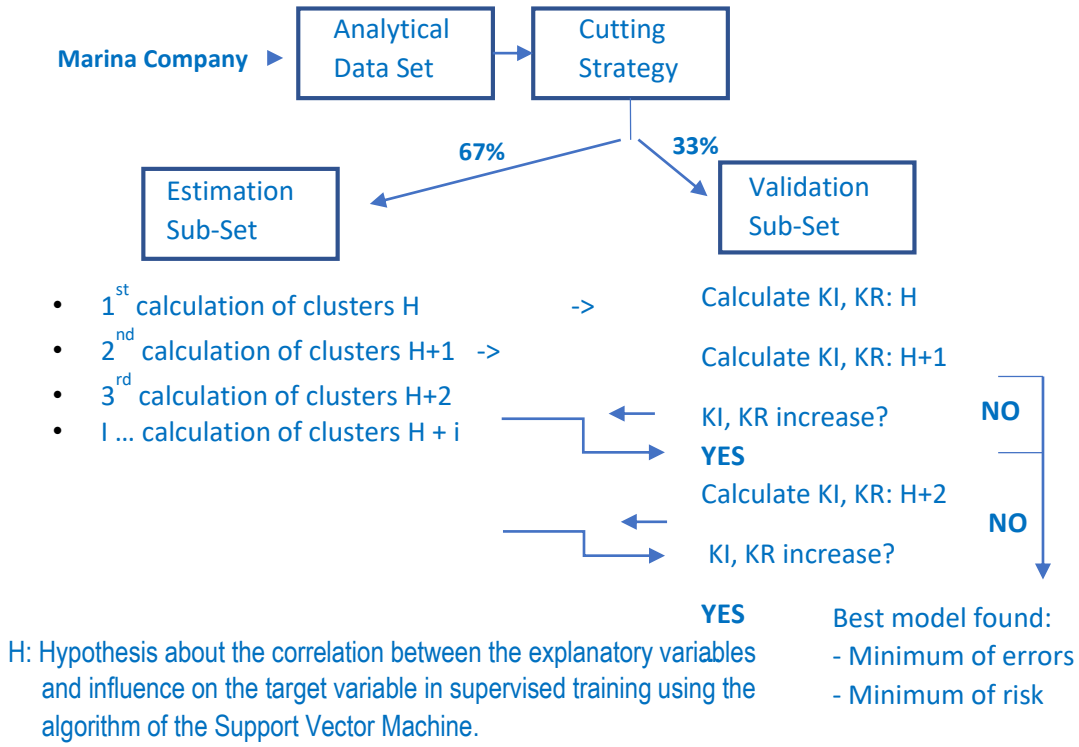
The solution SAP[14] Predictive Analytics is used in this research. SAP Predictive Analytics is a solution for statistical data analysis and evaluation and is developed based on the KXEN algorithm. The KXEN algorithm is based on the support vector machine (SVM). The so-called „supervised learning“ is used in this research because a target variable is given. Unsupervised learning without a defined target variable is not used in this research. Supervised learning means that the clusters are adjusted to provide a distribution that explains their influence on the target variable set in the models of this research. The explanatory variables, defined by the data set of the marina company in this research, are organized into groups with a similar influence on the target variable. The process of generating the models in this research is divided into a learning phase (ESTIMATION) and a validation phase (VALIDATION). The data set is divided randomly by the system into two sub-sets:

Estimation sub-set: In the first step, the variables which have not been excluded manually are tested for their significance to the target variable. If the significance is zero or very low, the algorithm will exclude the variables with the lowest significance as explanatory variables. A parameter can be used to set the degree of correlation at which an explanatory variable should be excluded. In the next iteration, a new model will be created using the remaining explanatory variables.

Validation sub-set: In the next step, the created models are evaluated using the evaluation data set. The best model with the highest values for predictive power KI and prediction confidence KR will be selected for the third step.

[14] SAP SE: SAP SE, based in Walldorf in Baden-Württemberg, is a listed software company. In terms of sales, SAP is the largest European (and non-American) as well as the world's third largest listed software company. July 20th, 2020.

Figure 16: The Learning Process in SAP Predictive Analytics

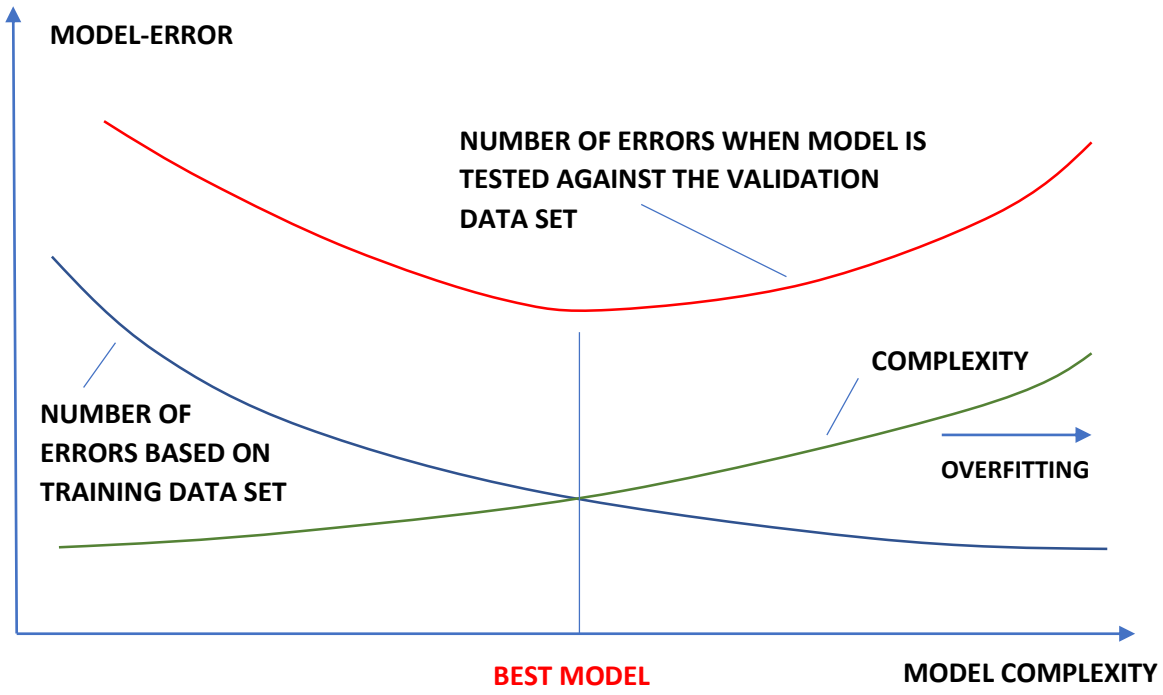


Source: Bakhsshaliyeva, N., page 116. Modified by the author.

The cutting strategy could be set manually. Charbert et al. (Charbert, 2018, page 163) propose using the default cutting strategy. Generally, 75% of data records are assigned to the estimation data sub-set and 25% to the validation data sub-set. A model with minimum complexity is created first. The errors related to the prediction of the target variables with the actual values of the target variables known in the training data-set are calculated. After the first generated model, the second model is generated by including additional variables with the complexity of h + 1. The second model is validated. If the error rate decreases, a third model is generated with the complexity h + 2. As soon as the error rate increases with another generated model, the previous model is determined by the algorithm as the best predictive model. In the third step, the performance indicators are calculated for the best model using the test data set. In this research, the principle of structured risk minimization (SRM) is applied with the SAP Predictive Analytics system (Chapter 4.1.4 in this research). The prediction model is sought based on the input variables and the target variables set in the model, which is, on the one hand, not too complex, and on the other hand, has a low number of errors (Bakhsshaliyeva, N., 2017, page 177). The parameter h stands for the complexity of the model. The complexity increases with

the number of variables. The best model after structured risk minimization is the one that has the lowest number of errors while avoiding excessive complexity.

Figure 17: Best Model According to the SRM Theory



Source: Bakhsshaliyeva, N., 2018, page 177.

Conclusion

The prediction models become efficient by applying the machine learning algorithm. The system divides the available data set into a training data set and a validation data set. Checking the generated model leads to a revision of the model by eliminating meaningless explanatory variables until no better result for predictive power (II) and predictive robustness (KR) has been found. Tested validation results lead to a reliable application of the prediction models for new data.

4.4 Description of the Prediction Models

Six prediction models were generated in this research. The models are described in this chapter. The first three models were developed using the clustering method since the target variable is binary. The fourth and fifth model are generated using the regression method since the target variable is an analog variable. The sixth model was generated with the classification that is

segmented in relation to an analog target variable. The underlying method for the first three prediction models is discussed in this section. The methods of the other prediction models are explained and described in the corresponding sections of this thesis. Clustering is the statistical method used for models A, B, and C. Clustering analysis is intended to identify similar objects and group them into clusters. The elements of a cluster have a mathematical closeness. A cluster is defined with records that are mathematically similar in comparison to the mathematical closeness of the elements to other clusters. The similarity is measured by the values of their attributes. (Charbert, A., 2017, page 271). The quality of the clusters is measured by how well a cluster explains a target variable. In supervised learning, clusters are defined. The learning process of the prediction model examines what information a variable provides to the target variable. The number of clusters is restricted. Otherwise, due to completely different characteristics, the same number of clusters as data records would be formed in each data record (Bakshaliyeva, N., 2017, page 196).

4.4.1 Model A – Client Type ‘PRIVATE’ with a Yearly Contract

The objective is to identify classes of customers with regard to the contract type. The classification parameters were set in the first step. The parameter (CONTRACT TYPE = 1) is used for supervised learning with target values “MV” (monthly contract) and “GV” (yearly contract). In model A the customer group ‘PRIVATE’ is identified as the customers intending to conclude a yearly contract. The customer group intending to conclude a yearly contract is preferred when the decision is to be made on applicants who should be allocated a free berth. The customer groups are defined by the learning process of the model, which compares combinations of parameter values given in the data set. The clustering algorithm finds centroids, which are the barycenter of the observed data. Therefore, a cluster is a group of customers with similar characteristics. The observations are grouped depending on their distances to the centroid of each cluster. A data file is selected first. To work with the file, the function ANALYZE must be performed to check the quality of the data set. SAP Predictive Analytics creates a description containing the description of the fields. The description file of the data set includes the description of the variables. This is an essential prerequisite for the computer application to work with the variables. The main important variable types are the following: Field STORAGE: This field defines variable type like NUMBER (computable), STRING (character strings), DATE (dates). Field VALUE: This field defines the value type. If a variable has the value CONTINUOUS , the variable is numeric and the mean, variances, etc. can be

calculated. If the variable has the value NOMINAL , it is a categorical variable and the only possible value for a string. Example: Client type P = Private Client. MISSING . It defines missing values. Example: Age and age group are only available for client type “P” (private), not for “F” (firms). According to SAP Predictive Analytics, the first row of the data set contains the variable names. The research in this clustering model is focused on the age group of the clients. ACI clients are classified into two groups: Private Clients (Private = P), who arrange the berth booking by themselves, and companies (Firms = F), whose employees arrange berth booking. With regard to the research question about a correlation of client age group with contract type, the filter ‘client type = P’ has been set. The target variable must be set next. Regarding the research question, the variable CONTRACT TYPE 1 is set as the target variable. This means that the cluster model should define the clusters with similar characteristics and their correlation to the target variable. The following question is answered: Which cluster variables have the highest influence on the target variable? Regarding the research question of which cluster variables have the highest influence on the target variable, variables that might not have any influence are excluded. It is recommended to exclude the Index-variable KxIndex in general. KxIndex is an internal counter for the variables. Variables that define any kind of identification numbers (ID) have been excluded because such variables cannot explain the correlation of a cluster to the target variable. These variables are too widely spread. With the setting of a target variable, the prediction is called supervised learning, because the result in which clients concluded a contract type 1 with yearly or monthly duration is known. In the next step, the minimum and maximum number of clusters are set as a restriction. This parameter of a maximum number of clusters has been set. The higher the number of clusters (segments), the lower the predictive confidence KR which defines the possibility to apply the prediction model to unsupervised data (data with unknown result). The lower the number of clusters, the lower the prediction power KI which defines how representative the model is compared to the ideal model. After several predictions run with a different number of ranges, it was found that the range of a minimum of 5 clusters up to a maximum of 8 clusters leads to the best result. The next parameter has been set to calculate SQL expressions. SQL stands for Standard Query Language. If this parameter is set, an additional cluster will be generated that contains the unassigned record. Example: Client type ‘FIRM’ does not have assigned values for age. It is possible to set advanced parameters. In the first run, the target key CONTRACT TYPE 1 is set to GV (yearly contract). The distance parameter defines the distance of two records in the cluster. The system-determined parameter is used in this approach. It is possible to overwrite the system-determined approach to the definition of centroid and their clusters, but this should

be carried out carefully because it might reduce the quality of the model. The clustering algorithm results in the creation of the centroid and its cluster. The clusters are defined by SQL expressions. Some observations are not defined by SQL expressions and are unassigned to a cluster. Some observations are defined by two SQL expressions and therefore overlap (see next figure). With the calculation of cross statistics, it is possible to visualize the profile of each explanatory variable for each cluster. The Encoding Strategy – Target Mean has been set. In this case, the target variable is nominal with value GV (yearly) or MV (monthly). In this case, the mean of the target, which is 50% yearly and 50% monthly contract, corresponds to the percentage of positive cases of the target variable for the input variable. 50% means random, and more than 50% defines a positive correlation.

The **statistical report** provides an overview of the remaining explanatory variables in the model for all sub-data sets. The negative value for the number of days payable outstanding (delay) means that the customer paid in advance. Additionally, negative values for invoice amount, debt amount, and the number of days payable outstanding are in the data set. As mentioned before, debt amount is calculated as the difference between the contract amount and the invoice amount. If an amount is smaller than the contract amount billed, the difference is saved as a residual claim with a negative sign.

Figure 18: Statistical Report, Model A

Variable	Data Set	Min	Max	Mean	Standard Deviation
Vessel length	Estimation	3	26	13.269	3.849
Vessel length	Validation	5	33	13.474	3.966
Vessel manufacture date	Estimation	1934	2019	2,003.24	12.878
Vessel manufacture date	Validation	1942	2019	2,004.01	12.111
Starost plovila	Estimation	0	80	12.506	12.911
Starost plovila	Validation	0	76	11.742	12.242
ClientId	Estimation	4354	7023416	2,724,900	2,132,720
ClientId	Validation	4370	6953298	2,788,770	2,113,510
Client age	Estimation	22	86	52.542	10.735
Client age	Validation	21	85	52.953	10.733
Contract amount	Estimation	184.5	179334	21,604	25,716.8
Contract amount	Validation	184.5	179334	22,534.7	25,775.6
Invoice amount	Estimation	-35469.6	167723	14,899	19,965.4
Invoice amount	Validation	-25591.2	149335	16,497.4	21,856.4
Ka?njenje	Estimation	-432	317	20.505	59.414
Ka?njenje	Validation	-112	346	21.047	64.21
Debt amount	Estimation	-175923	900.43	-6,504.23	15,942.1
Debt amount	Validation	-149292	1010.61	-5,970.56	16,087
Debt amount %	Estimation	-1.15	200	22.137	35.236
Debt amount %	Validation	-1.1	196.71	20.558	34.639

Source: Screenshot in SAP Predictive Analytics based on data created by the author.

Note: The figure above shows the results based on the original dataset with Croatian variable names. This applies to all figures about the results from the prediction models of the prediction system.

Legend: Starost plovila = Vessel Age. Kašnjenje = Delay.

Contract Amount – Invoice Amount = Debt Amount

If Invoice Amount < Contract Amount → Debt Amount is negative (residual claim)

If a customer pays in advance → negative invoice amount

Source: Screenshot in SAP Predictive Analytics based on data created by Lebefromm, U.

Figure 19: Store Payment in Advance

Contract amount	Payment date	Predujam	Invoice amount	Contract start ye	Debt amount
29476	01.05.2019 00:00	TRUE	-25055	2019	-54531
32999	03.07.2019 00:00	TRUE	-16499,49	2019	-49498,49
80826	26.04.2019 00:00	TRUE	-2180,16	2019	-83006,16
18587	28.05.2019 00:00	TRUE	-832,42	2019	-19419,42
39011	19.04.2019 00:00	TRUE	-311,52	2019	-39322,52

Advance – Payment in Advance = TRUE

Contract Amount + Invoice Amount (payment in advance) = Debt Amount

39476 + 25055 = 54531

Source: Screenshot of the data set created by the author.

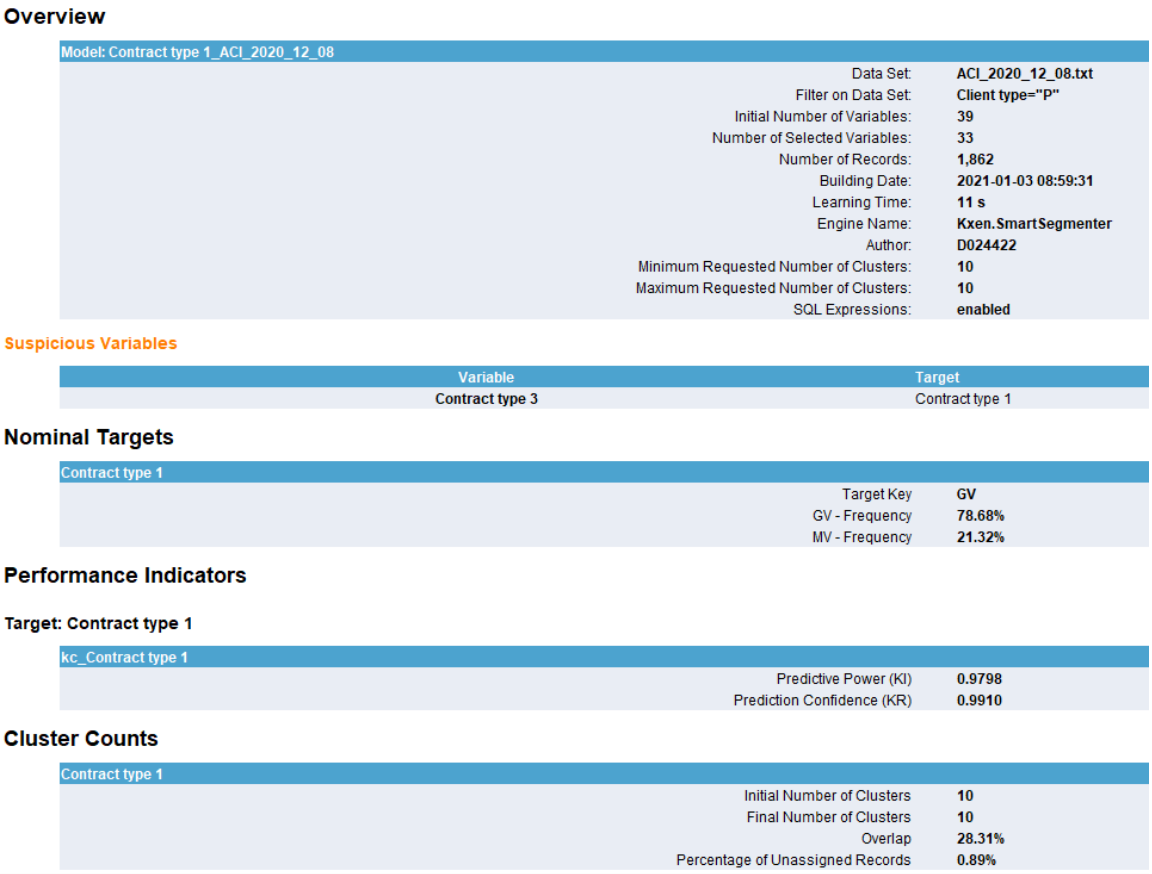
Legend: Predujam = Advance Payment

The **model overview** explains the quality of the generated model. The following information are provided:

Target	Key:	GV	(yearly	contract)
GV-Frequency:	78.68	%	(yearly	contract)
MV-Frequency:	21.32 % (monthly contract)			

The suspicious variable : Contract Type 3. Suspicious variable means that this variable is very correlated to the target variable. In the case of ACI data, contract type 3 specifies the correlation to season, region, one month or six months, etc. Every contract has its specification with contract type 3. Every data record has a characteristic value for contract type 3. A prediction of whether a contract-type 1 client has a contract type 3 using YES or NO answer is not considered in this research.

Figure 20: Model Overview: Model A



Overview

Model: Contract type 1_ACI_2020_12_08	
Data Set:	ACI_2020_12_08.txt
Filter on Data Set:	Client type="P"
Initial Number of Variables:	39
Number of Selected Variables:	33
Number of Records:	1,862
Building Date:	2021-01-03 08:59:31
Learning Time:	11 s
Engine Name:	Kxen.SmartSegmenter
Author:	D024422
Minimum Requested Number of Clusters:	10
Maximum Requested Number of Clusters:	10
SQL Expressions:	enabled

Suspicious Variables

Variable	Target
Contract type 3	Contract type 1

Nominal Targets

Contract type 1	
Target Key	GV
GV - Frequency	78.68%
MV - Frequency	21.32%

Performance Indicators

Target: Contract type 1

kc_Contract type 1	
Predictive Power (KI)	0.9798
Prediction Confidence (KR)	0.9910

Cluster Counts

Contract type 1	
Initial Number of Clusters	10
Final Number of Clusters	10
Overlap	28.31%
Percentage of Unassigned Records	0.89%

Source: Screenshot in SAP Predictive Analytics based on data created by the author.

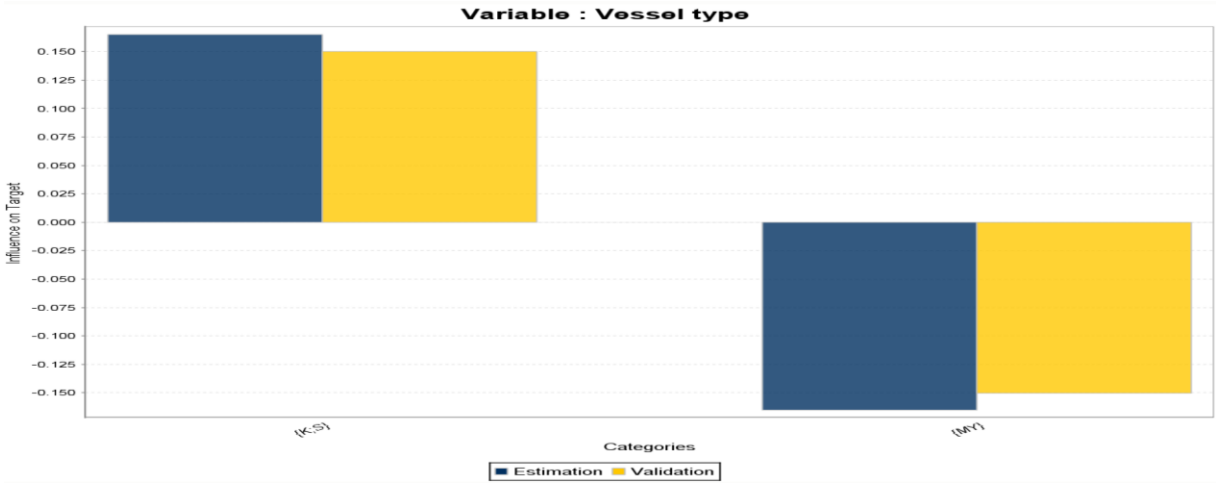
The calculated values for **predictive power** KI with the value of 0.9798 and **prediction confidence** KR with the value of 0.9910 fulfill the requirement for the significance of the model.

Category ‘Significance’

The five explanatory variables with the highest influence on the target variable “Yearly Contract” are introduced below. The influence could be positive or negative. An explanatory variable with a positive influence on the target variable is represented more frequently in relation to the target variable than the average of all variables and vice versa. Explanatory variables with the same influence are grouped.

Explanatory variable VESSEL TYPE. A positive influence on the conclusion of yearly contracts comes from clients using a sailing boat or catamaran.

Figure 21: Model A – Influence of the explanatory variable 'VESSEL TYPE'



Source: Screenshot in SAP Predictive Analytics based on data created by the author

The formulas used: Category 'Importance' = NP*BF/{NC} where NP is Normal Profit, BF is Bin Frequency (or category frequency), and NC is Normalization Constant. The calculation of the normalization constant (NC) differs by target data type.

Calculation of the normalization constant (NC) for binary targets (yearly contract):

(Target Frequency) * (1 - Target Frequency])

Normal profit can be calculated using the following formula for the binary targets.

The frequency f1 is the frequency of the least frequent target class TC1, and f2 = 1 -f1 is the frequency of the most frequent target class TC2. We can "associate" the least frequent target class TC1 with profit (TC1) equal to f2 and the most frequent target class with profit (TC2) equal to -f1. The normalized profit (TC1) and profit (TC2) have been chosen:

Profit (TC1) * probability (TC1) + profit (TC2) * probability (TC2) = 0

The normal profit of category C is calculated as follows:

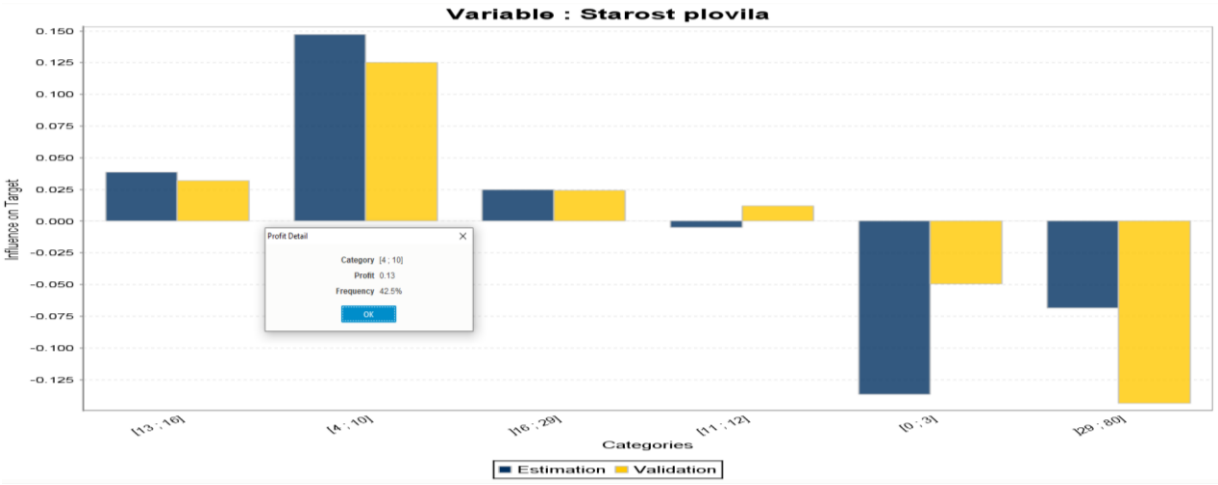
Normal Profit (C) = Profit (TC2) * P(TC2|C) + Profit (TC1) * P(TC1|C)

Where P(TC1|C) is the conditional probability of belonging to the least frequent target class given that the individual belongs to category C. This conditional probability is approximated by the frequency of TC1 within the individuals of category C.

The influence of the vessel types K and S is positive in all data sub-sets. Conclusion: Vessel types K and S should be preferred for the conclusion of a yearly contract.

Explanatory variable: VESSEL AGE. The influence of this explanatory variable is best with vessel age of about 4 to 10 years. 42.5 % of the boats holding a yearly contract are in this age range.

Figure 22: Model A – Influence of the Explanatory Variable: Vessel Age

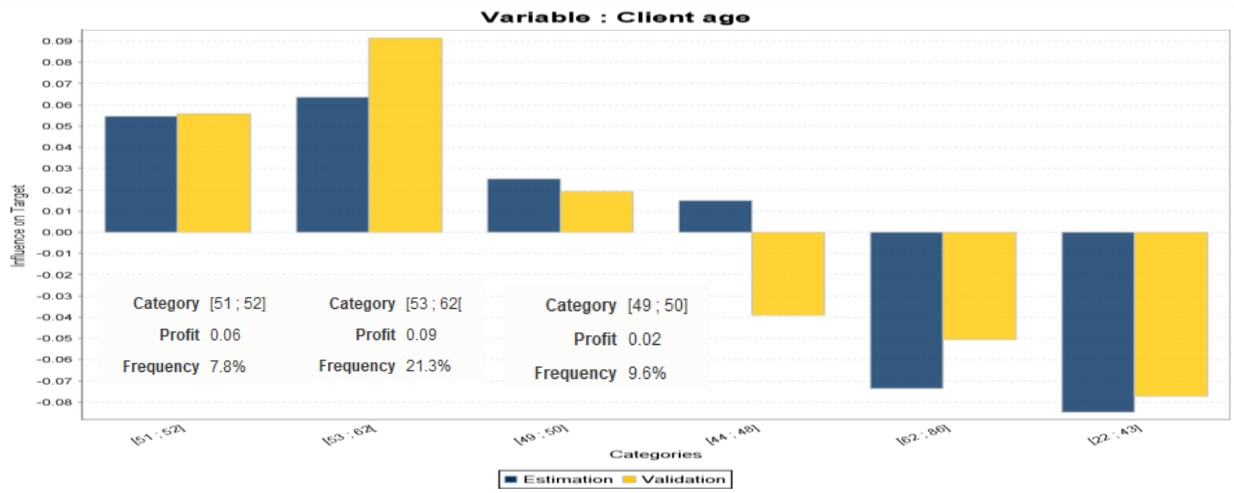


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Conclusion: Clients with younger boats should be preferred when detecting berth applicants with whom conclusion of a yearly contract is expected. No reliable statement can be made for applicants with older boats with the available data.

Explanatory Variable: CLIENT AGE

Figure 23: Model A – Influence of the Explanatory Variable 'Client Age'

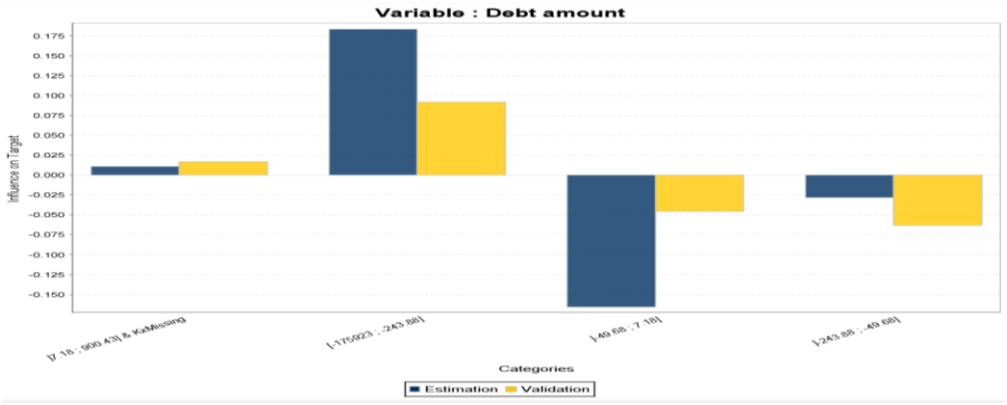


Regarding the clients' age, a clear statement can be made for both the estimation and the validation data sub-set. 38.7% of customers are in the age range from 49 to 62 years. This age group plays the largest role in the conclusion of yearly contracts. Conclusion: Marketing messages which should be addressed to the ones with whom a yearly contract is to be concluded should be designed for the age group 49–62 years.

Explanatory Variable: PREPAYMENT

As mentioned in the data description, prepayment is managed as a binary variable. The evaluation refers to the explanatory variable 'debt amount'. When evaluating the influence of these variables on the target variable 'yearly contract', it can be concluded that prepayments in the range of HRK 243.00 to HRK 175.00 have a tendency towards yearly contracts.

Figure 24: Model A – Influence of the Explanatory Variable 'Advance Payment'

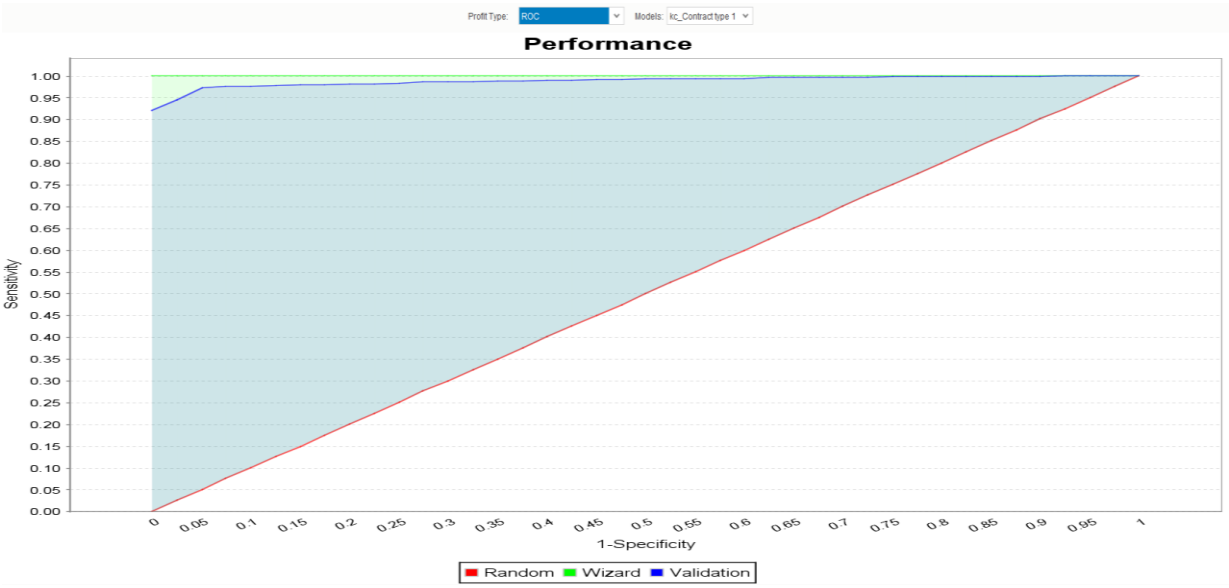


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Significance of the Model

The ROC (Receiver Operating Characteristics) graph is shown in the following figure. It portrays how well a model discriminates in terms of a tradeoff between sensitivity and specificity, or, effectively, between correct and mistaken detection given that the detection threshold varies.

Figure 25: ROC Curve, Model A



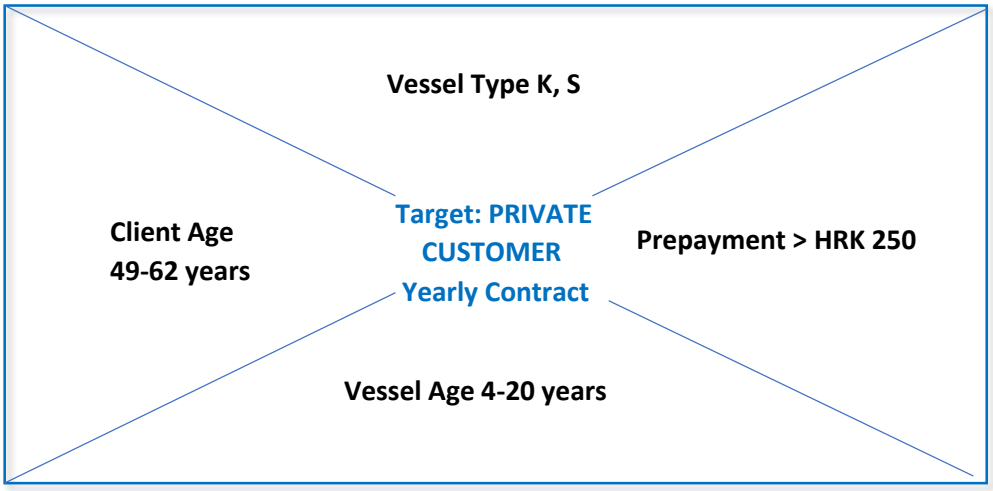
Source: Screenshot in SAP Predictive Analytics based on data created by the author

Sensitivity, which appears on the Y-axis, is the proportion of CORRECTLY identified signals (true positives out of all true positives in the validation dataset). [1 - Specificity], which appears on the X-axis, is the proportion of INCORRECT assignments to the signal class (false positives) incurred (out of all false positives in the validation dataset). Specificity, as opposed to [1 - specificity], is the proportion of CORRECT assignments to the class of NON-SIGNALS – true negatives. The prediction model is well above the random curve and very close to the wizard (perfect model), which is also reflected in the calculated values for predictive power KI and prediction confidence KR.

Summary of Model A

Four explanatory variables (vessel type, vessel age, client age, and prepayment) having a positive influence on the target variable YEARLY CONTRACT have been identified.

Figure 26: Explanatory variables with a positive influence on the target variable ‘Yearly Contract’ – Client Type: Private



Source: Author

Conclusion

Private customers are inclined to concluding a yearly contract if they are in the age group from 49 to 62 years, owning a sailing boat or catamaran that is between 4–20 years old, and if they made a prepayment in the amount higher than HRK 250.00.

4.4.2 Model B – Client Type ‘FIRM’ with a Yearly Contract

Regarding the client type ‘FIRM’, the variables ‘CLIENT AGE’ and ‘CLIENT AGE GROUP’ have been excluded from the explanatory variables when defining the target variable and explanatory variables. The same applies to the variable CLIENT ID. As in the previous model, the data set was selected and the step ‘Analyze and Save Description’ was executed. The target variable is set with contract type = 1 (YEARLY CONTRACT). In the next step, the minimum and the maximum number of clusters were set as a restriction. In this case, minimum of 5 clusters, maximum of 10 clusters. As described in the previous prediction, a compromise was found to create a model with good-quality values for prediction power KI and prediction confidence KR. The filter “GV” (yearly contract) was set in model B. The client type ‘FIRM’ has been set as a filter and is therefore constant. The figure below shows the statistical report for continuous variables. Client age refers to the age of the contact person of the company. The

variable ‘VESSEL AGE’ ranges from zero (built in 2019) to 123 years. Very old vessels are an exception, the number of data records with vessel age over 70 years is 40. The cutting strategy assigned two-thirds of the data set to the estimation data sub-set and one third to the validation data sub-set.

Figure 27: Statistical Report, Model B

Variable	Data Set	Min	Max	Mean	Standard Deviation
VesselID	Estimation	11173	7074726	2,726,370	2,308,650
VesselID	Validation	11171	7074357	2,753,140	2,318,890
Vessel length	Estimation	3	41	11.316	2.946
Vessel length	Validation	4	41	11.38	3.074
Vessel manufacture date	Estimation	1896	2019	1,997.89	10.877
Vessel manufacture date	Validation	1896	2019	1,997.68	11.438

Data Set			Number of Records		Total weight
Estimation			18,459		18459
Validation			6,325		6325
Client age	Estimation	18	89	58.659	11.392
Client age	Validation	19	89	58.472	11.39
Contract amount	Estimation	0	292380	11,672.2	16,120.5
Contract amount	Validation	0	399998	12,403.5	18,580.8
Invoice amount	Estimation	-33865.1	281967	8,209.13	12,914.8
Invoice amount	Validation	-29273.3	399998	8,852.14	15,279.3
Ka?njenje	Estimation	-528	419	10.561	61.406
Ka?njenje	Validation	-736	403	9.514	61.412
Debt amount	Estimation	-143832	28902.4	-3,305.4	9,493.82
Debt amount	Validation	-192456	5813.05	-3,332.61	10,136.4
Debt amount %	Estimation	-689.31	194.24	17.902	33.91
Debt amount %	Validation	-415.21	191.78	17.575	33.187

Source: Screenshot in SAP Predictive Analytics based on data created by the author.

Legend: Kašnjenje – Delay

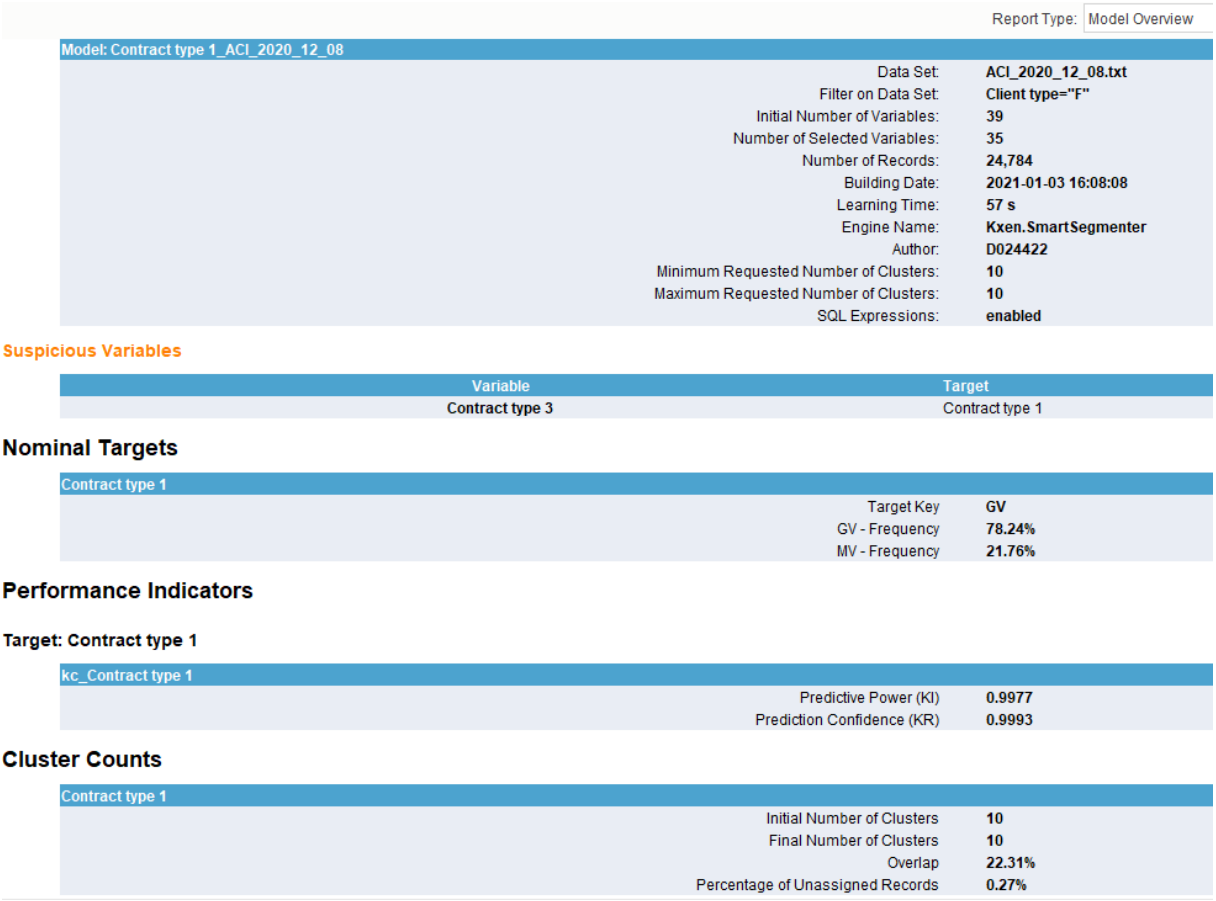
The **model overview** explains the quality of the generated model. The following information are provided:

Number of records selected: 24,784
Target Key: GV (yearly contract)
Predictive Power KI: 0.9977
Prediction Confidence: 0.9993

Ten clusters have been created with an overlap of about 22%. The frequency of yearly contracts with a value of 78.24% is higher than the value of 21.76% for monthly contracts. It must also be mentioned here that the negative values for the invoice amount and debt amount result from the prepayments made by the customers. The prepayment field only has the Boolean value TRUE or FALSE. The advance payments made by applicants and customers are posted directly in the data field “invoice amount” and “debt amount” and thus lead to a negative value. The variable ‘Contract Type 3’ is set as suspicious because it is highly correlated with the target variable. Contract Type 3 with a breakdown by months and time allocation within a year is

widely distributed. Five versions of this explanatory variable have a positive influence on the target variable, 7 versions have a negative influence. Because of this high correlation, this explanatory variable is not considered in the detected explanatory variables.

Figure 28: Overview, Model B



Report Type: Model Overview

Model: Contract type 1_ACI_2020_12_08

Data Set:	ACI_2020_12_08.txt
Filter on Data Set:	Client type="F"
Initial Number of Variables:	39
Number of Selected Variables:	35
Number of Records:	24,784
Building Date:	2021-01-03 16:08:08
Learning Time:	57 s
Engine Name:	Kxen.SmartSegmenter
Author:	D024422
Minimum Requested Number of Clusters:	10
Maximum Requested Number of Clusters:	10
SQL Expressions:	enabled

Suspicious Variables

Variable	Target
Contract type 3	Contract type 1

Nominal Targets

Contract type 1

Target Key	GV
GV - Frequency	78.24%
MV - Frequency	21.76%

Performance Indicators

Target: Contract type 1

kc_Contract type 1

Predictive Power (KI)	0.9977
Prediction Confidence (KR)	0.9993

Cluster Counts

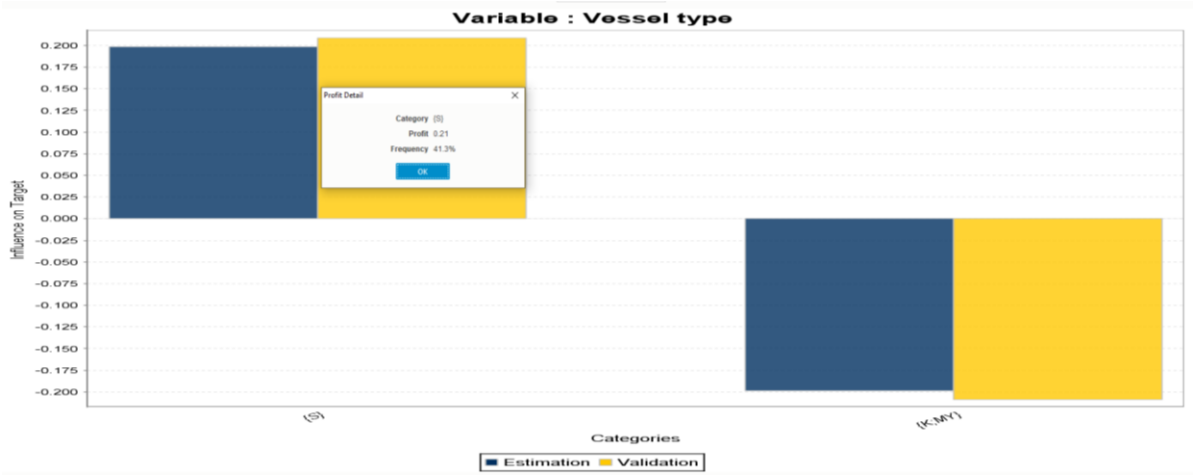
Contract type 1

Initial Number of Clusters	10
Final Number of Clusters	10
Overlap	22.31%
Percentage of Unassigned Records	0.27%

Source: Screenshot in SAP Predictive Analytics based on data created by the author.

The objective is to discover the characteristics of the companies that concluded a yearly contract. It is also interesting to analyze whether there is a difference between contracts concluded in different marinas of the company. The significance of the explanatory variables is presented below. At first, the significance of vessel characteristics like vessel type, vessel age, and vessel length is analyzed. Next, the influence of contract type 2 is analyzed to find out whether new contracts or renewed contracts have a significant influence on the target variable.

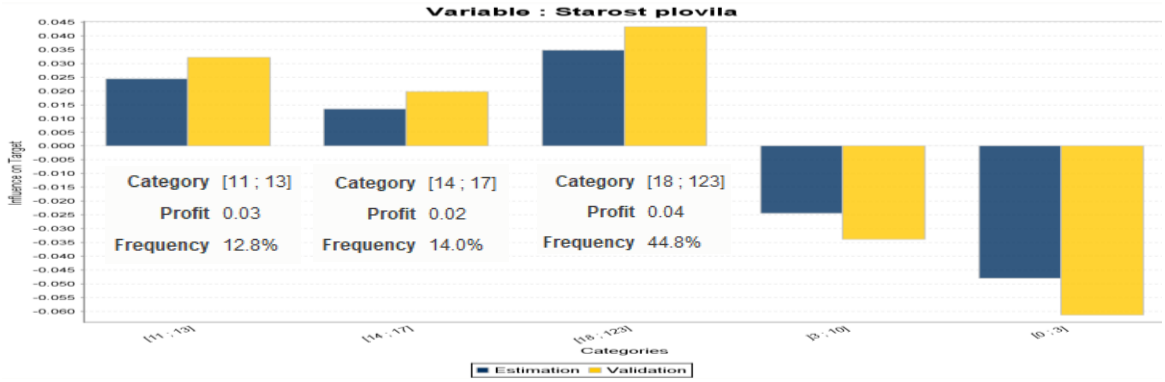
Figure 29: Model B – Influence of Vessel Type on Yearly Contract – Client Type: FIRM



Source: Screenshot in SAP Predictive Analytics based on data created by the author

Contrary to private customers, vessel type ‘catamaran’ has a negative influence on the target variable ‘YEARLY CONTRACT’. Companies with sailing boats tend to have more annual contracts than companies wanting to rent a berth for a motor yacht. The vessel type ‘catamaran’ plays a subordinate role in the number of boats.

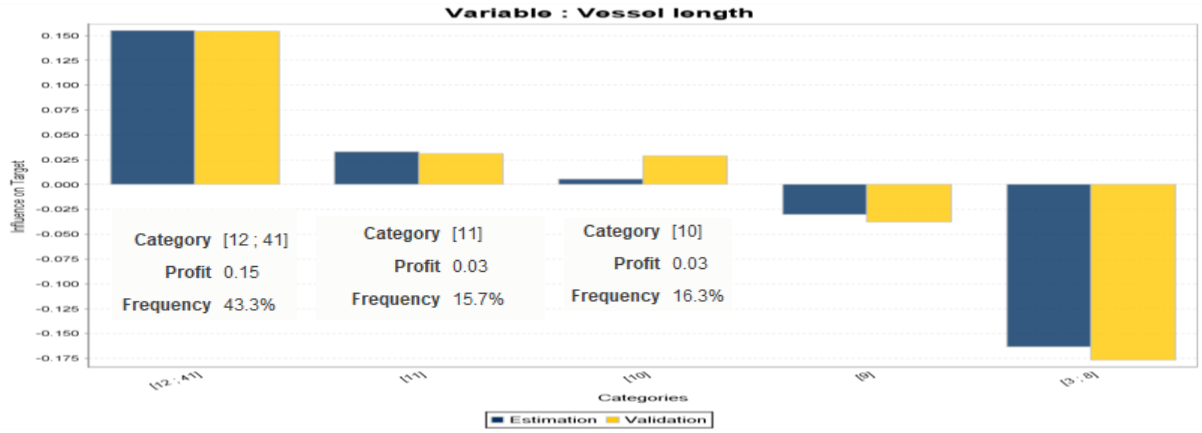
Figure 30: Model B – Influence of Vessel Age on Yearly Contract – Client Type: FIRM



Source: Screenshot in SAP Predictive Analytics based on data created by the author.

Vessels in the age range from 11 to 123 years represent 71.6% of contracts with a duration of one year. It is fair to say, the older the vessel, the more likely it is to have a yearly contract. It must be noted that very old vessels, those older than 100 years, make only 11 data records, together with the vessels older than 80 years, that make16 data records and, together with the vessels older than 70 years, make 40 data records out of more than 26,000 data records.

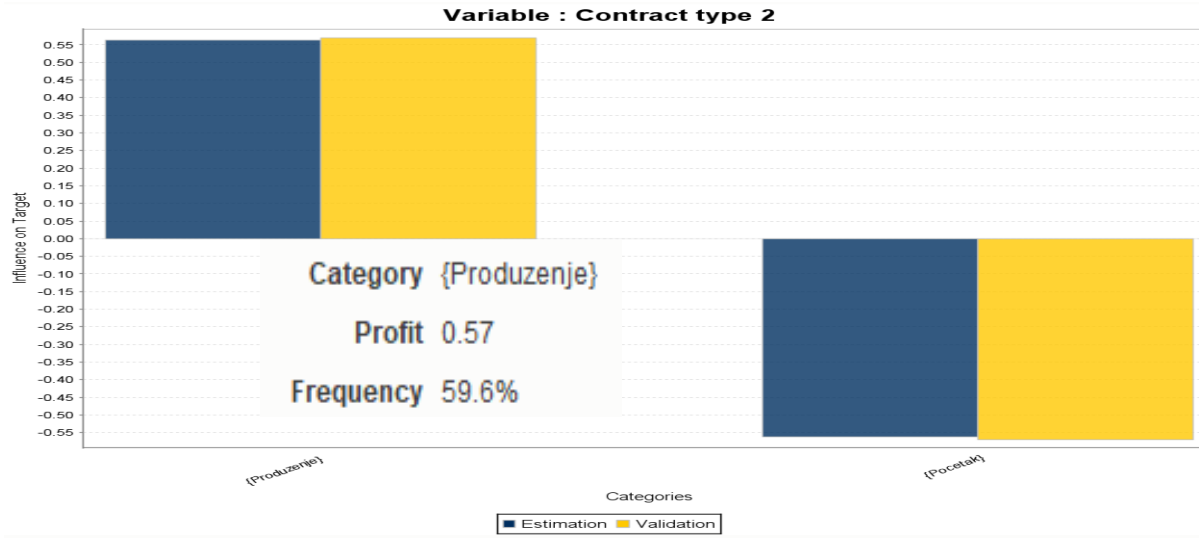
Figure 31: Model B – Influence of Vessel Length on Yearly Contract – Client Type: FIRM



Source: Screenshot in SAP Predictive Analytics based on data created by the author

It is fair to say that vessels longer than 9 meters are the ones that should be preferred when intending to conclude a yearly contract.

Figure 32: Model B – Influence of Contract Type NEW or RENEWAL to Yearly Contract – Client Type: FIRM



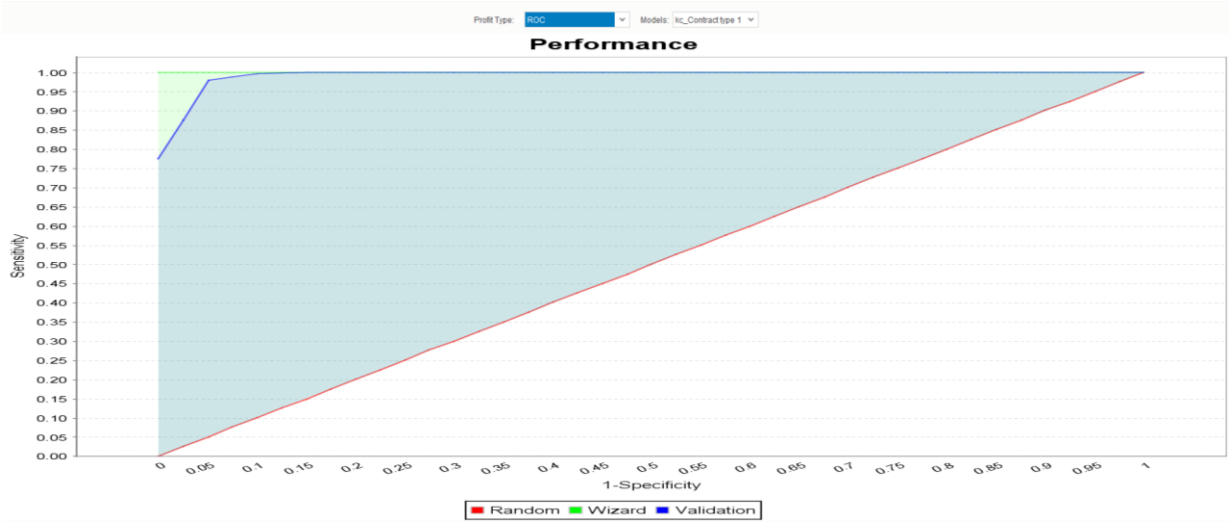
Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Produženje – Renewal

Vessels longer than 9 meters represent 71.6% of the contracts with a duration of one year. Regarding contract type 2 – new contract or contract renewal – 59.6% of yearly contracts are renewed.

Significance of the Model

Figure 33: ROC-Curve, Model B



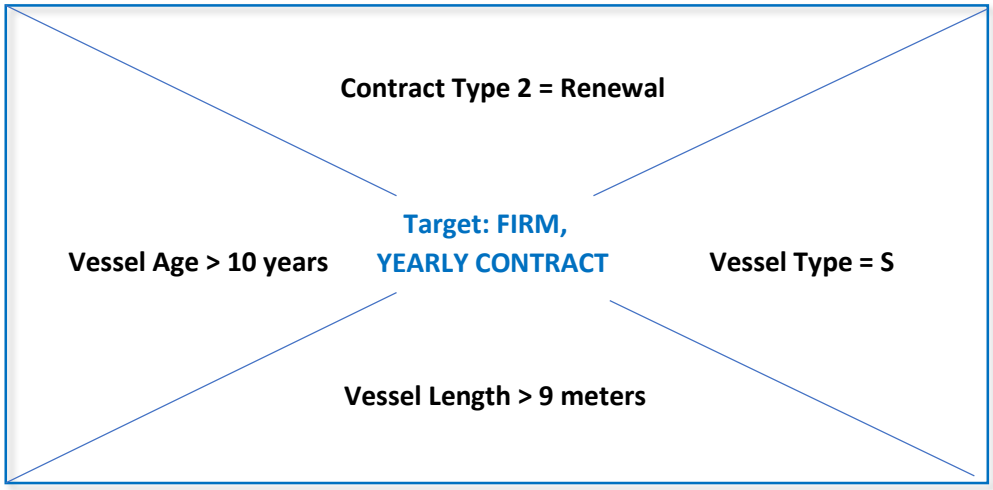
Source: Screenshot in SAP Predictive Analytics based on data created by the author

Sensitivity, which appears on the Y-axis, is the proportion of CORRECTLY identified signals. [1 – Specificity], which appears on the X-axis, is the proportion of INCORRECT assignments. The prediction model B is well above the random curve and very close to the wizard (perfect model), which is also reflected in the calculated values for KI and KR.

Summary of Model B

Four explanatory variables (vessel type, length, age, and contract type) that have a positive influence on the target variable have been identified. The most important explanatory variables for concluding yearly contracts with the client type ‘FIRM’ depend on the characteristics of the vessel. Neither the youngest nor the smallest vessels of sailing boat type would be preferred. The best option for a yearly contract is renewal of an existing contract.

Figure 34: Explanatory Variables with Positive Influence on Target Variable 'Yearly Contract' – Client Type: FIRM



Source: Author

Conclusion

When clients are companies using a sailing boat older than 10 years and longer than 9 meters with the intention to renew their contract, it is expected that these clients will conclude a yearly contract.

4.4.3 Model C – Contract Renewal

In times of global restrictions in tourism such as the COVID-19 pandemic, which has been spreading worldwide since 2019, it has become more difficult to acquire new customers. Against this background, and in line with maintaining long-term contractual relationships with customers, it is interesting to observe the characteristics of private applicants and customers who tend to renew existing contracts. With knowledge of customer characteristics, marketing campaigns can be designed for this target group. The prediction has been performed for the whole data set without setting any filters. The target variable CONTRACT TYPE 2 = RENEWAL was set in the selecting variables. Variables according to identification numbers like contract-ID, vessel-ID, contract-ID, and client-ID and monotonic variables, i.e., dates like contract start, contract end, payment date, invoice date, were excluded because such variables cannot explain any kind of influence on the target variable. The focus is on renewed contracts.

Therefore, the advanced settings for the target variable ‘contract type 2 = RENEWAL’ have been set in the specific parameters. 8,2587 records were selected. Predictive power KI with 0.9985 and prediction confidence KR with 0.9985 document a very good quality of the prediction model. Overlapping of 1.4% is acceptable. No filter has been set in model C, not even for client type. This stems from the results of the training and validation phase; the system grouped all client types into a single group and detected that this variable does not influence the target variable ‘contract type 2 – contract renewal’. The statistical report above shows negative values for the explanatory variables ‘DELAY’, ‘DEBT AMOUNT’, ‘DEBT AMOUNT PERCENTAGE’, and ‘INVOICE AMOUNT’.

Figure 35: Model C – Statistical Report and Data Size

Variable	Data Set	Min	Max	Mean	Standard Deviation
VesselID	Estimation	11171	7074159	2,754,350	2,297,810
VesselID	Validation	11173	7074726	2,738,710	2,328,210
Vessel length	Estimation	3	41	11.482	3.111
Vessel length	Validation	3	41	11.441	3.039
Vessel manufacture date	Estimation	1896	2019	1,998.26	11.209
Vessel manufacture date	Validation	1896	2019	1,998.15	11.319
Starost plovila	Estimation	0	122	17.681	11.292
Starost plovila	Validation	0	123	17.84	11.412
ClientId	Estimation	4348	7074140	2,656,310	2,260,380
ClientId	Validation	4348	7074330	2,641,810	2,306,850
Client age	Estimation	18	89	58.191	11.428
Client age	Validation	18	89	58.21	11.504
Contract amount	Estimation	0	399998	12,542.6	17,897.7
Contract amount	Validation	0	265566	12,607.5	17,283.1
Invoice amount	Estimation	-35469.6	399998	8,848.62	14,440.7
Invoice amount	Validation	-26021.8	192456	8,903.37	13,755.2
Ka?njenje	Estimation	-736	419	11.181	61.703
Ka?njenje	Validation	-432	403	10.525	60.572
Debt amount	Estimation	-192456	28902.4	-3,529.7	10,279.6
Debt amount	Validation	-143467	27345.2	-3,515.57	10,210.1
Debt amount %	Estimation	-667.32	200	18.24	33.649
Debt amount %	Validation	-689.31	191.5	17.654	34.381

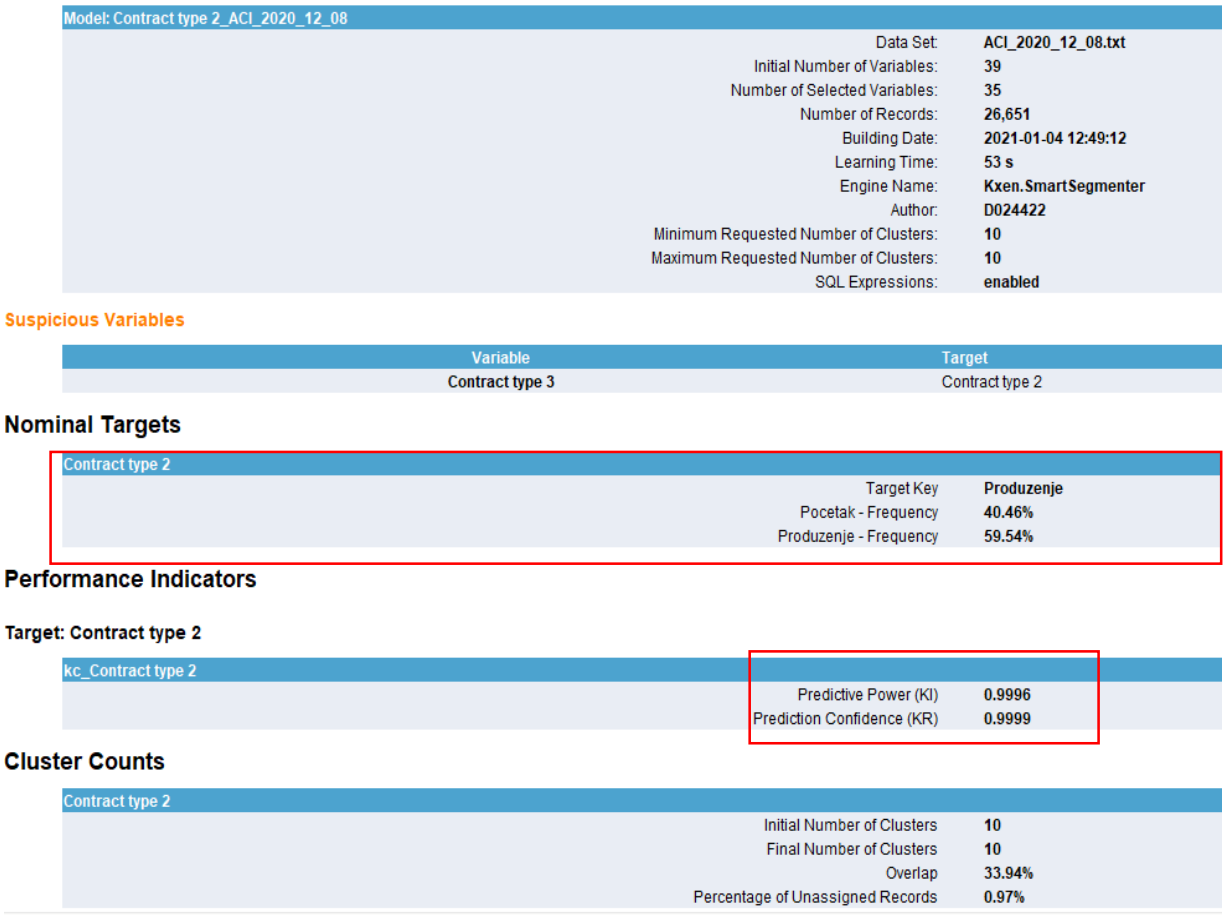
Data Set	Number of Records	Total weight
Estimation	19,864	19864
Validation	6,787	6787

Source: Screenshot in SAP Predictive Analytics based on data created by the author.

Legend: Starost plovila – Vessel age. Kašnjenje – Delay

All negative values are based on the data model that does not update advance payment from the clients in specific fields. For example, this would falsify a statement about average day sales outstanding, which is, however, not the subject of this research. All data records have been selected in this model. The model overview shows a high value of predictive power KI and prediction confidence KR. The ratio of renewed contracts to new contracts is about 3 to 2. Explanatory variables with a significant influence could be calculated with this tendency.

Figure 36: Model C – Overview



Model: Contract type 2_ACI_2020_12_08

Data Set:	ACI_2020_12_08.txt
Initial Number of Variables:	39
Number of Selected Variables:	35
Number of Records:	26,651
Building Date:	2021-01-04 12:49:12
Learning Time:	53 s
Engine Name:	Kxen.SmartSegmenter
Author:	D024422
Minimum Requested Number of Clusters:	10
Maximum Requested Number of Clusters:	10
SQL Expressions:	enabled

Suspicious Variables

Variable	Target
Contract type 3	Contract type 2

Nominal Targets

Contract type 2

Target Key	Produzenje
Pocetak - Frequency	40.46%
Produzenje - Frequency	59.54%

Performance Indicators

Target: Contract type 2

kc_Contract type 2

Predictive Power (KI)	0.9996
Prediction Confidence (KR)	0.9999

Cluster Counts

Contract type 2

Initial Number of Clusters	10
Final Number of Clusters	10
Overlap	33.94%
Percentage of Unassigned Records	0.97%

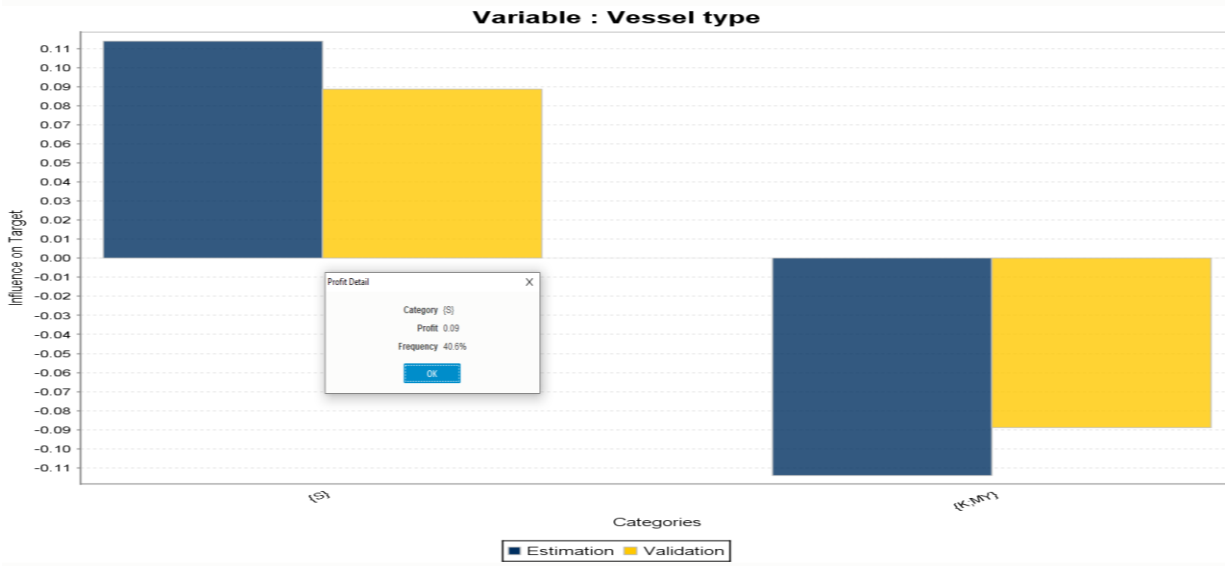
Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Početak – New Contract, Produženje – Renewal

Category Significance

If, in the current difficult times of a pandemic, the focus is on renewing existing contracts, it is important to identify the customers with the greatest chances of success. The investigation is focused on the explanatory variables for the vessel. This occurs because the system puts customer groups P and F together in a single group and therefore does not have any influence on the target variable from this feature. The following figure shows the message displayed. Nevertheless, citizenship analysis shows an interesting result. Compared with the previous evaluations, vessel type ‘sailing boat’ is ahead in terms of long-term customer relationships. One of the explanations could be that not every marina is suitable for sailing boats. If a customer finds a marina with good wind conditions, they are more likely to hold onto the berth.

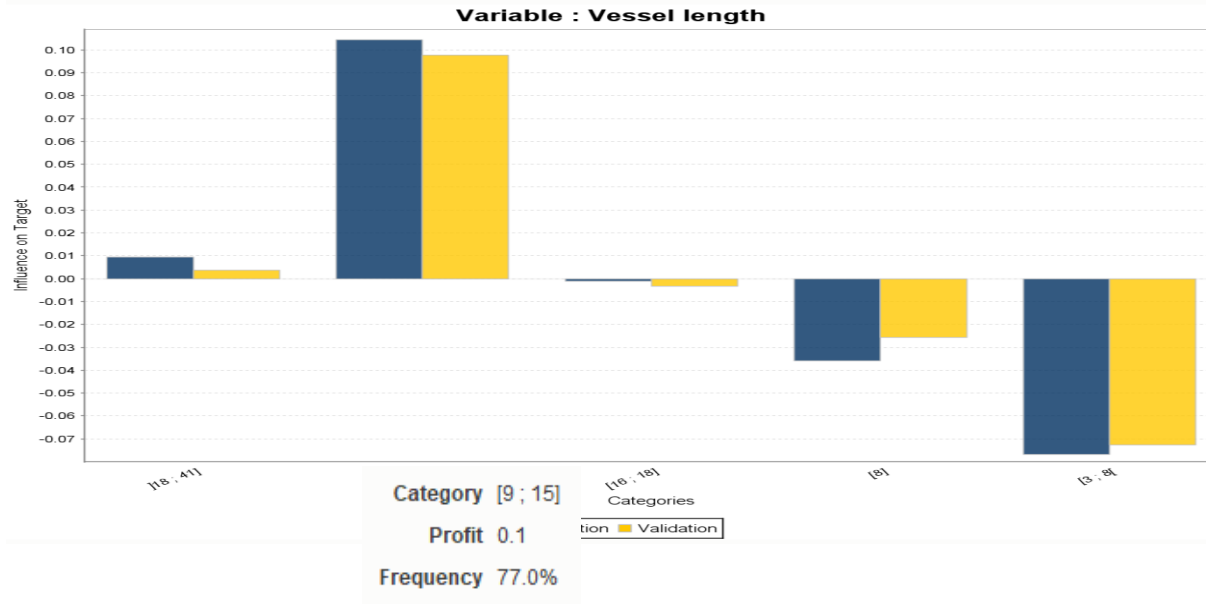
Figure 37: Model C – Influence of Vessel Type on Contract Renewal



Source: Screenshot in SAP Predictive Analytics based on data created by the author.

The ‘catamaran’ and ‘motorboat’ vessel types have a negative impact on the renewal of existing contracts. Catamarans play a subordinate role in terms of the proportion of vessel types. Motor yachts are less dependent on wind conditions than sailing boats. Regarding the size of the vessels, the obtained results are the same as in the previous investigations. Medium-sized vessels in the range of 9 to 15 meters are clear favorites when it comes to renewing existing contracts. Boats of this length represent 77% of the totals. This could also be an input for dimensioning of the berths. Owners of sailing boats, who prefers marinas suitable for sailing boats, could show that they are interested in free berths with a compatible berth length. Another aspect is the trend towards charter boats. An investigation into the average length of charter boats would be helpful here.

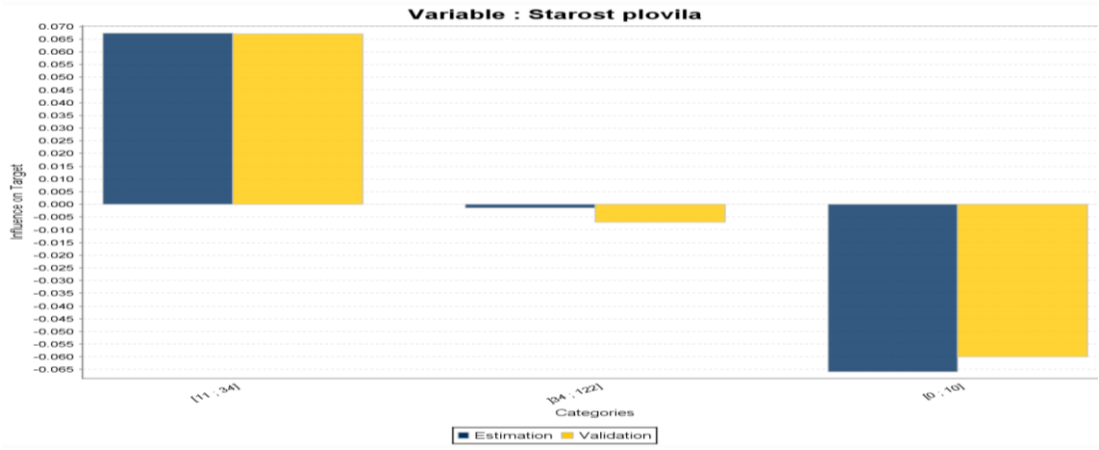
Figure 38: Influence of Vessel Length on Contract Renewal



Source: Screenshot in SAP Predictive Analytics based on data created by the author

In addition to medium-length boats, middle-aged boats also represent the strongest group of clients who tend to renew existing contracts. Owners of younger boats are evidently more inclined to exploring new areas and calling at other ports accordingly.

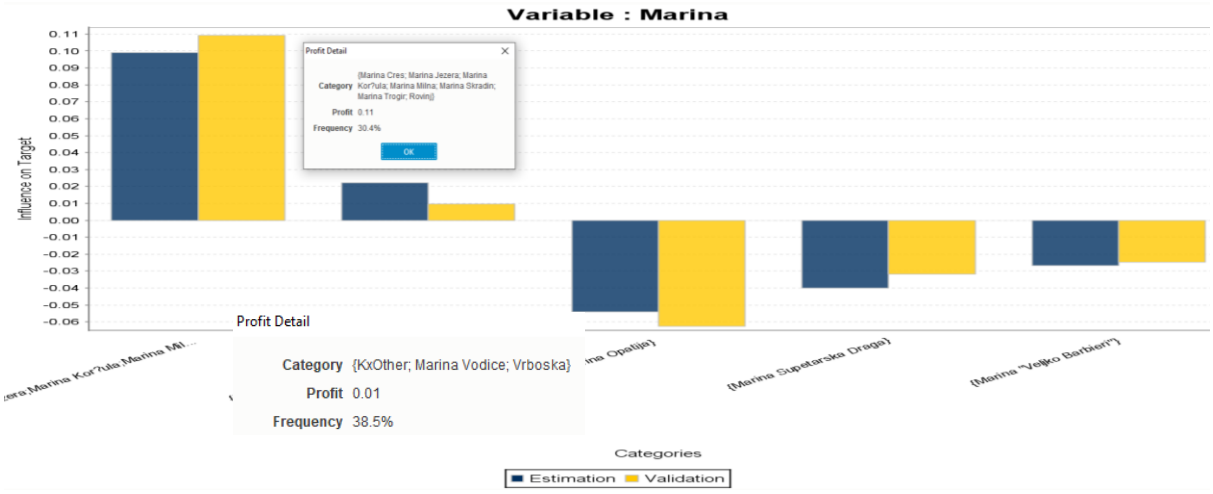
Figure 39: Influence of Vessel Age on Contract Renewal



Source: Screenshot in SAP Predictive Analytics based on data created by the author

It is interesting to examine the ports that have an advantage in contract renewals. The preferred ports for contract renewals are CRES, JEZERA, KORČULA, MILNA, SKRADIN, TROGIR, and ROVINJ. The ports of VODICE and VRBOSKA have a weaker but positive influence on contract renewals. A conclusion about positive wind conditions for sailboats can be drawn from this.

Figure 40: Model C – Influence of Marinas on Contract Renewal

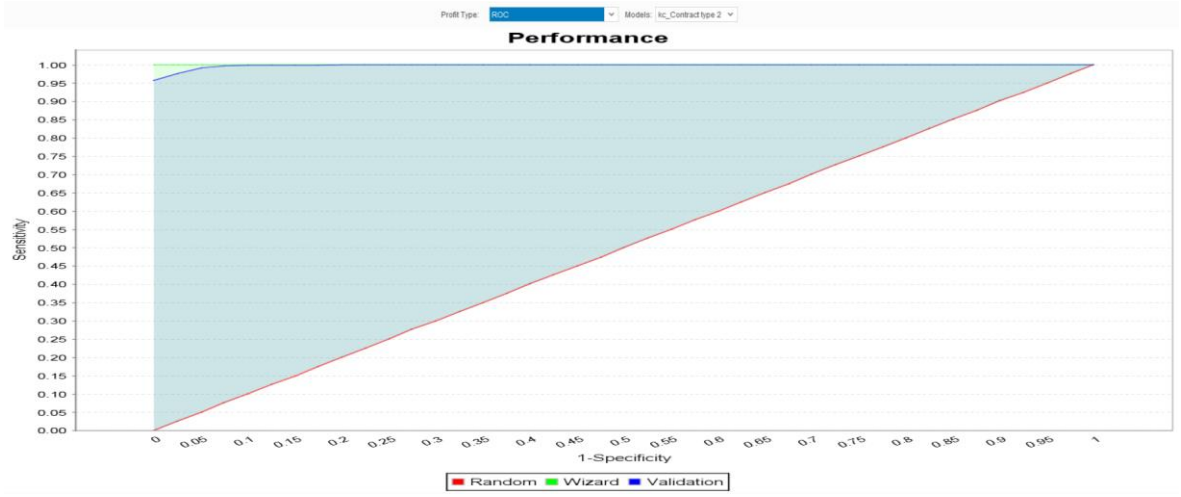


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Significance of the model

The values for KI and KR are very high. The ROC curve runs very closely to the wizard, i.e., the perfect model. However, the connections for renewed contracts also clearly point to sailing boats, suitable marinas, and the average length and average age of the boats.

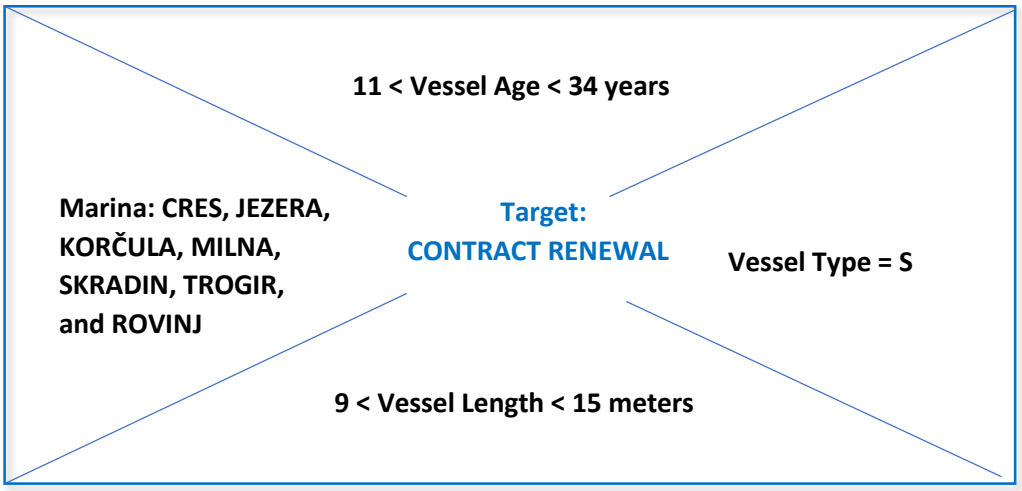
Figure 41: Model C – ROC Curve



Source: Screenshot in SAP Predictive Analytics based on data created by the author

Summary of Model C

Figure 42: Explanatory Variables with Positive Influence on Contract Renewal



Source: Author

A marketing campaign aiming for contract renewals should be addressed to clients who are using sailing boats between 9 and 15 meters in length and between 11 and 34 years old. The above-mentioned marinas should be advertised. As the controlling department of the marina company explained during the interview, the identified ports are very suitable for sailing boats due to wind conditions. However, these properties should not eliminate constant monitoring of the progress and constant updating of results with the availability of new data sets and correcting them if necessary. Nevertheless, tendencies are evident and decisions on the allocation of berths and the concept of marketing campaigns can be made according to the target group.

4.5 Prediction with Regression Analysis

Model D and Model E have a continuous target variable. Therefore, regression analysis is used instead of classification. Regression analysis is a method for estimating the value of output variables based on the value of input variables to explore how one variable affects another. Knowing the influence of input variables on output variables, the development of the output variables can be predicted based on the development of the input variables. SAP Predictive Analytics automatically generates a regression model when the target variable is continuous, like age or age group of the clients. If the target variable is binary , SAP Predictive Analytics automatically generates a classification model. A binary variable in the data set is ‘CHARTER VESSEL’ (TRUE, FALSE). Solving the problem begins with the definition of the problem

itself: creation of the so-called “specification” of the problem. At the same time, this specification forms the basis for the development of a prediction model. This means that real-world objects are analyzed, relationships between the objects uncovered, operations searched for and, finally, the results presented in a formal notation. An example of an estimation method is the ordinary least-squares method (OLS), the standard mathematical method for calculating a functional relationship of observed values. In this case, a mathematical function is sought for a data point cloud, the graph of which runs as close as possible to data points. Data can represent physical dimensions or economic quantities. The least-squares method is used to determine the curve parameters so that the sum of the square deviations of the curve from the observed points is minimized. The deviations are called residuals.

$$y_i = \alpha + \beta x_i + u_i \text{ with } i = 1,\ldots,N$$

The regression parameters are α and β. The random variable u_i is a disturbance variable. With multivariate linear regressions (linear regression with at least two variables) there can be a high correlation between the independent variables (multicollinearity). If variables significantly overlap, it is difficult to prove that the variables are scattered. The so-called “ridge regression” offers a solution. An additional parameter limits the influence of all regression coefficients and thus counteracts superimposed influences. This method is also called the shrinkage method . This ridge regression is also used in the SAP Predictive Analytics system (Bakhshaliyeva, N., 2017, page 44).

4.5.1 Model D – Late Payment

Late payment by customers is a factor that should not be underestimated. While on the one hand, many customers pay in advance, there is a noticeable delay in payment among many customers. The purpose of this analysis is to find out whether certain characteristics of customers indicate poor payment behavior. The goal of this model is to identify client characteristics which could explain why such clients are late payers. The target variable in this model is defined as the number of days between the payment due date and the date the data set was created by the marina company and given to the author of this research. Due to the fact that applicants for a berth and customers may be paying in advance, the value for DELAY is negative. Therefore, it is necessary to set the filter for the target variable: DELAY > 0. The statistical method of regression is used for the continuous variable ‘number of days payable outstanding’. Contrary to a pure investigation of correlations, the regression in this research

determines the significance of the given explanatory variables for the target variable. The same data set is used for the classification and regression analysis as in the previous clustering analysis. The task of data description was explained above. Data selection is limited to data records that show a positive payment delay. This is necessary because, due to prepayments, there are also records with negative payment delays. The variable which contains the number of days payable outstanding has the ID: DELAY, which is set as the target variable. All variables involving date, IDs, and names are excluded as explanatory variables. This should help concentrate on explanatory variables that can be taken into consideration by ACI's controlling department when allocating berths. The variable KxIndex was automatically set as excluded because this variable contains a data description.

Note: The selected variables as explanatory variables and the set of excluded variables could be saved and loaded for the next prediction. Variable weighting has not been set in order to prevent the model to become suspicious.

The modeling parameters are used to fine-tune the generation of the prediction model. The following parameters have been set:

Compute Decision Tree. The generation of a decision tree opens the possibility to put together the five variables with the highest explanatory contribution.

Enable Auto-Selection. This function excludes variables that have a negative impact on prediction power KI and prediction confidence KR.

The parameter AUTOSAVE leads to automatic saving of the model when generated. This prevents the loss of the model when SAP Predictive Analytics is closed without saving the model.

The option EXPORT KxSHELL SCRIPT is used to perform model generation as background processing. This is used in the processing of a very large numbers of data.

Another additional parameter that has been set is Polynomial Degree . The regression model generation leads to the generation of regression equations. In linear simple regression, the first equation explains the expected value and the second equation explains the change in the expected value. This parameter can be used to determine the order in which the regression functions can be generated. SAP notes that a higher-order polynomial does not necessarily lead

to a better forecast model[15]. Regarding SAP's arguments, this study uses the generation of 1st order polynomials.

Score Bin Count: Data binning is used to reduce different forms of variables. A typical example of this is replacing the variable 'age' with the variable 'age group'. Too few groups, but also too many groups diminish the forecast quality. Low prediction confidence variables are excluded by the system because such variables are difficult to transfer to new data sets. Correlations lower than the parameter are not detected as correlations and are ignored by the system. In this research, the default value of 50 is retained. A correlation above 50% is assumed to be a positive correlation.

Keep the first n correlations: If several correlating variables are detected in the learning process, it can be set how many variables are kept for the next iteration step.

Enable post-processing, within original target encoding: The regression result is used to calculate the standard mean values in the individual variable segments. This setting is the standard setting of SAP Predictive Analytics. It is used in this research. As part of the model generation, a step-by-step approach to the optimal solution is applied. New hypotheses are created during the learning phase with the addition and removal of variables. Such hypotheses are confirmed in the learning phase if the forecast quality improves. A hypothesis is rejected if forecast quality is reduced. Forecast quality is calculated with the key performance indicators KI and KR (see explanations above). With the successive consideration of variables, a new forecast model is created after each iteration step. The variable selection parameter can be used to determine how many parameters are to be retained for the next iteration step. As long as the quality parameters KI and KR increase due to the iterations, the process continues with the next iteration. If the quality parameters decrease by 5% in an iteration, the best forecast model has been found. The **statistical report** shows the mean and standard deviation for all continuous variables. The standard deviation is naturally high for all variables containing amounts and

[15]SAP Predictive Analytics Online help: "Using a higher degree of polynomial does not always guarantee better results than those obtained with a first-degree polynomial. In addition, the higher the degree of polynomial you select:

- the more time needed to generate the corresponding model
- the more time needed to apply the model to new datasets
- the harder it is to interpret the results of modeling.

identifications (ID). Variables containing names and variables with binary values (TRUE, FALSE) have been excluded by the system. 9,869 data records have been found by the system with a positive number of days payable outstanding. In relation to the total number of 26,652 data records, 37% of all selected customers are late payers – a problem not to be underestimated.

Figure 43: Model D – Statistical Report and Data Size

Variable	Data Set	Min	Max	Mean	Standard Deviation
VesselID	Estimation	11171	7038232	2,735,460	2,284,750
VesselID	Validation	11173	7047919	2,764,270	2,308,130
Vessel length	Estimation	3	33	11.813	3.187
Vessel length	Validation	3	33	11.903	3.257
Vessel manufacture date	Estimation	1896	2019	1,998.05	11.669
Vessel manufacture date	Validation	1934	2019	1,998.03	11.581
Starost plovila	Estimation	0	122	17.967	11.808
Starost plovila	Validation	0	83	18.078	11.696
ClientId	Estimation	4348	6991539	2,739,210	2,247,380
ClientId	Validation	4348	7047902	2,703,900	2,256,030
Client age	Estimation	18	89	56.82	11.623
Client age	Validation	18	88	57.034	11.569
Contract amount	Estimation	0	294303	14,662.6	19,993.4
Contract amount	Validation	125	265566	14,747.8	19,298.5
Invoice amount	Estimation	-35469.6	281967	7,761.39	14,063.4
Invoice amount	Validation	-33865.1	256107	7,889.81	13,717.5
Debt amount	Estimation	-192456	27345.2	-6,901.19	14,009.8
Debt amount	Validation	-149292	3971.55	-6,858.03	13,605.9
Debt amount %	Estimation	-689.31	200	35.527	42.164
Debt amount %	Validation	-69.48	190.68	34.694	40.264

Data Set	Number of Records	Total weight
Estimation	7,307	7307
Validation	2,562	2562

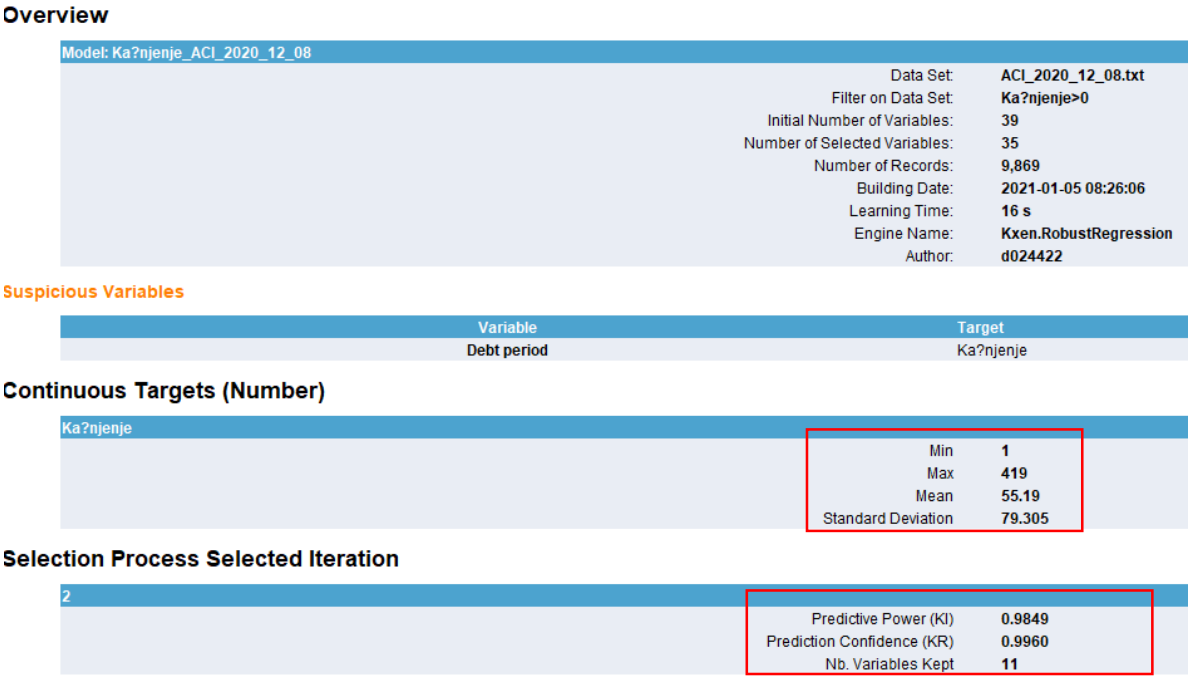
Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Starost plovila – Vessel age

The following figure displays the KI and KR values showing a high significance of the model. The debt period is evaluated as suspicious because of the high correlation with the target variable. The target variable is widely spread.

The value for standard deviation is high. As will be shown in the following evaluation of the graph of the model, this refers to the fact that there are clients with many days payable outstanding; up to 419 days.

Figure 44: Model D – Overview

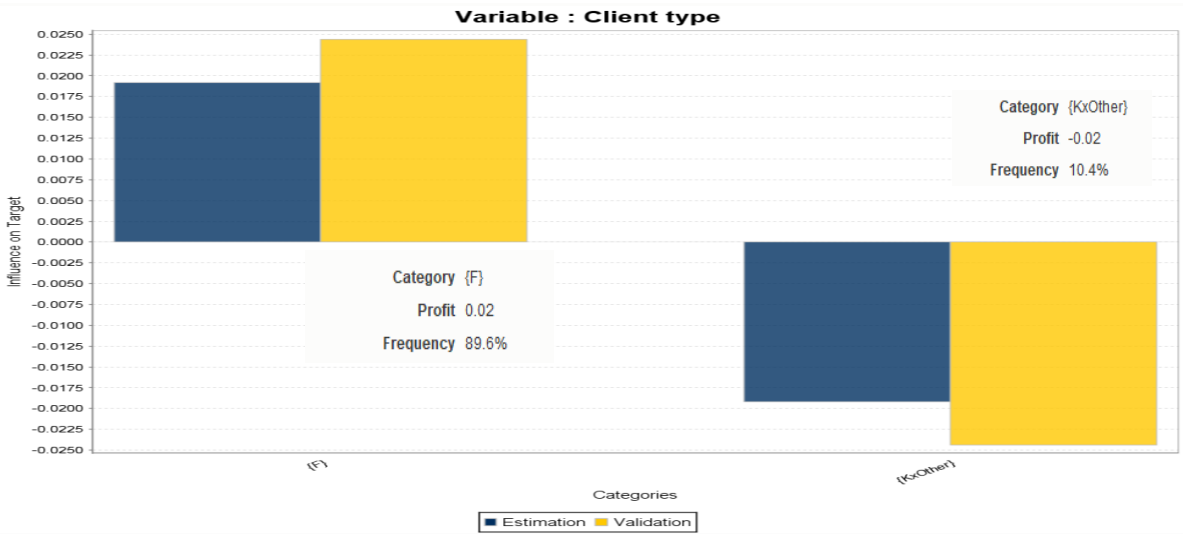


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Kašnjenje – Delay

The analysis of the **significance of explanatory variables** leads to client variables 'client type' and 'citizenship'. Vessel age has a significant influence, but not vessel type, because vessel types have been grouped. A fourth variable with a significant influence has been found with the variable 'contract type'.

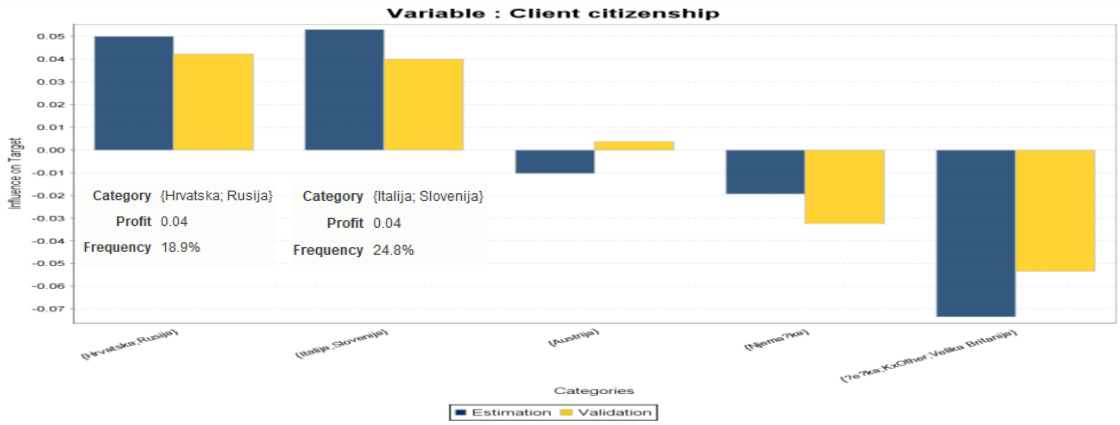
Figure 45: Model D – Influence of Variable ‘Client Type’ on Payment Behavior



Source: Screenshot in SAP Predictive Analytics based on data created by the author

With 89.6% of the selected data sets, the proportion of firms is significantly higher than the proportion of private clients. If a firm has acute financing needs, it is more likely that private clients will pay on time. Citizenship indicates an interesting result. 43.7% of late payers are from four countries. This could also be a selection criterion if the firm’s financial needs make it necessary.

Figure 46: Model D – Influence of Citizenship on Payment Behavior



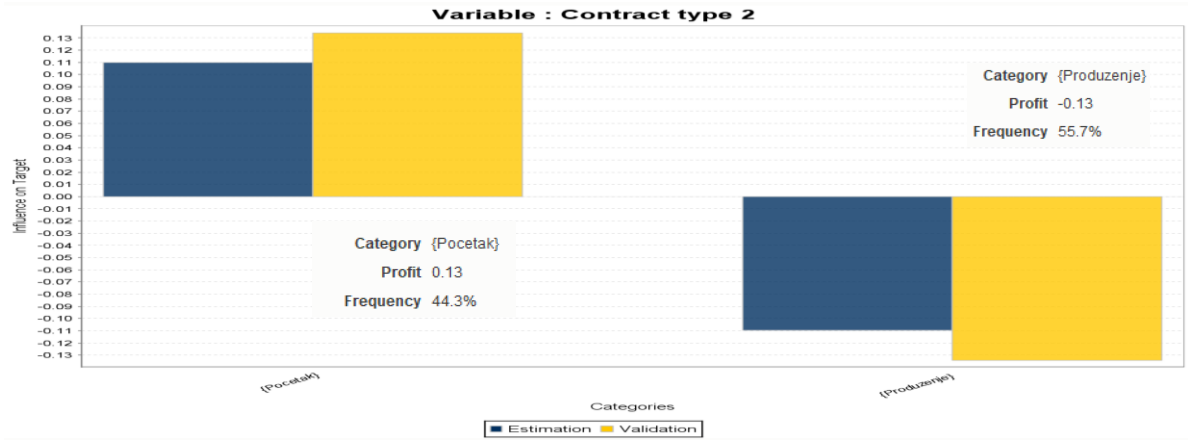
Source: Screenshot in SAP Predictive Analytics based on data created by the author.

Legend: Hrvatska – Croatia. Rusija – Russia. Italija – Italy. Slovenija – Slovenia

On the other hand, it should not be concluded that one should no longer conclude contracts with customers from these regions. It makes more sense to adapt the terms of payment, for example by agreeing to an advance payment. This could be combined with the rental price and

prepayments reducing the rental price. The prediction models provide the information, the resulting proposals concerning the terms of payment are in the scope of tasks of the controlling department, and the management are making the decision. 44.3% of the contracts with a positive influence on late payment are new contracts. On the other hand, the invoices for renewed berth contracts tend to be paid on time. However, it should be examined here whether different payment terms apply to renewed contracts in comparison with the new contracts. However, the trend is clear. Payment delay more often occurs in the case of a new contract (the Croatian expression is *novi ugovor*. The expression *početak* is used in the dataset, which stands for 'beginning'). In the case of contract renewal, the Croatian expression is *produženje*.

Figure 47: Model D – Influence of Contract Type on Payment Behavior

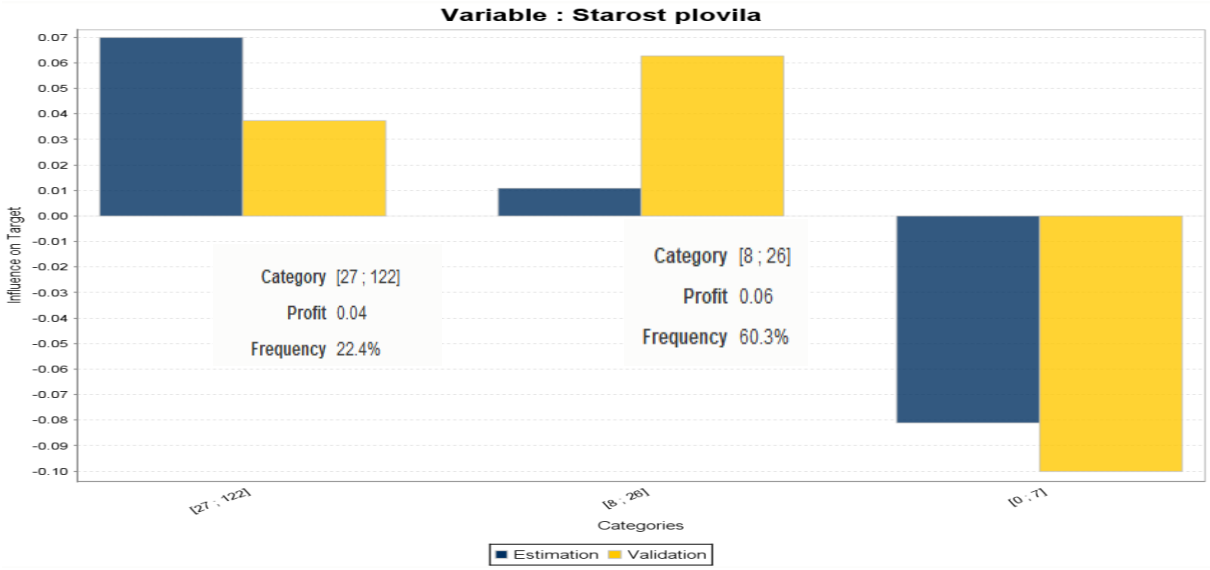


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Produženje – Renewal

The fourth explanatory variable is the age of the vessels. Owners of older boats display worse payment behavior than those owning younger vessels. 82.7% of late payers own a vessel that is 8 years old or older. It can be concluded that owners of younger boats are more solvent.

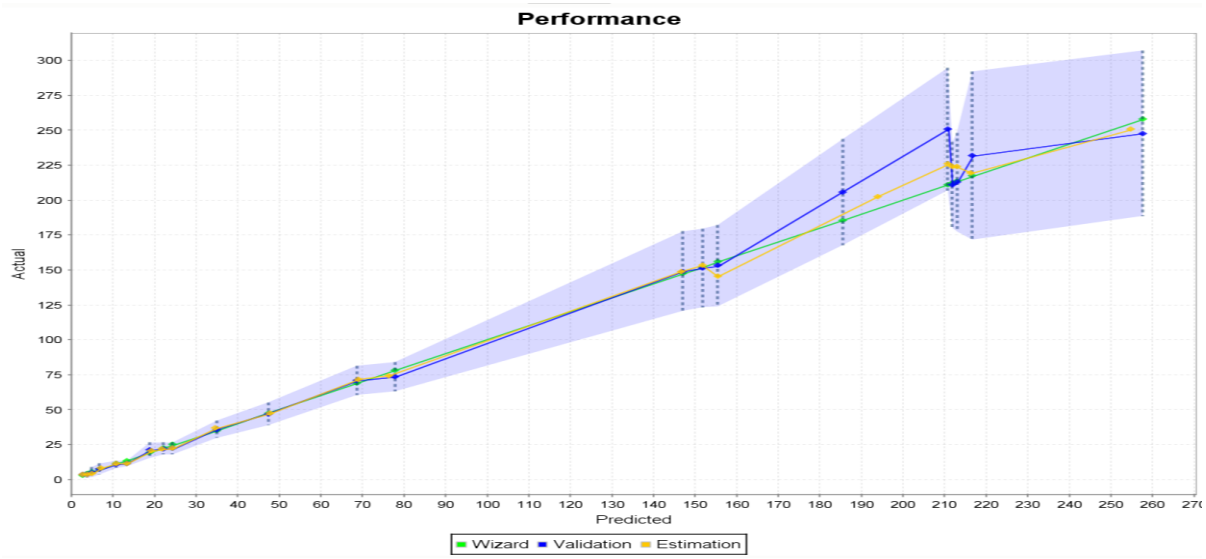
Figure 48: Model D – Influence of Variable ‘Vessel Age’ on Payment Behavior



Source: Screenshot in SAP Predictive Analytics based on data created by the author

It can be concluded that owners of younger boats are more solvent. In terms of the settings, a polynomial degree of 1 has been set. The documentation of SAP Predictive Analytics determines that a higher degree of polynomial does not always guarantee better results than those obtained with a polynomial degree of 1. The number of score bins is the standard setting of about 20. A higher or lower number of score bin counts would lead to poor model quality. Variables with low prediction confidence are excluded. The system uses an internal threshold to decide whether a variable has low prediction confidence. This threshold depends on dataset size and target distribution. With the correlation settings, the engine excluded the ones with the lowest correlation rate, thus keeping only the more significant ones. (Bakhshaliyea et al., 2016, page 144).

Figure 49: Model D – Performance



Source: Screenshot in SAP Predictive Analytics based on data created by the author

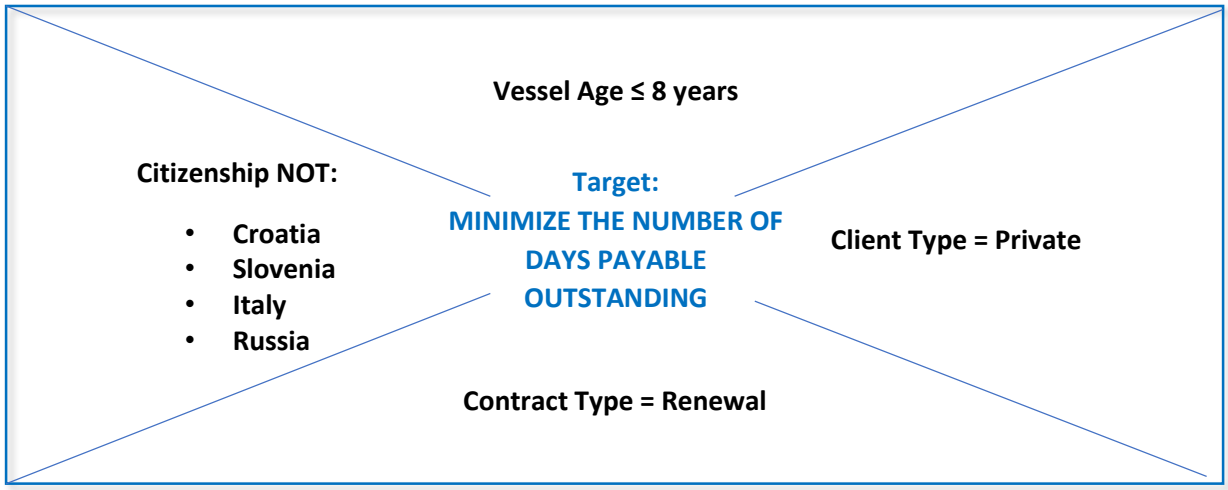
The figure displays the actual target values as a function of the predicted target values. The wizard curve (green line) means that all the predicted values are equal to the actual values. If the validation curve is going far from the wizard, the predicted value is suspicious (Bakhshaliyea et al., 2016, page 158). The figure above displays similar values for the estimation and validation of the wizard regarding 150 days payable outstanding. It is fair to say that the prediction model is completely suitable within 150 days payable outstanding. Additionally, the same applies to the range from 150 to 190 days; the deviation is still within an acceptable range. A greater number of days payable outstanding is difficult to predict.

Summary of Model D

A significant influence of explanatory variables on payment delay can be demonstrated up to the value of 150 measured in delay time. A payment delay is mainly to be expected from companies with boats older than 8 years and headquarters in Croatia, Slovenia, Italy or Russia concluding a new contract. However, it should not be interpreted from the results that the priority of concluding contracts with these target groups should be reduced. That would be a wrong conclusion. The findings are rather in favor of advance payment terms. The following figure reverses the results of the prediction models. The prediction model shows the characteristics of company customers who are late payers. The aim of the marina company, to conclude contracts with clients from whom it can expect timely payment, can be fulfilled by

clients having contrary characteristics to those detected in the prediction. They are summarized in the following figure.

Figure 50: Explanatory Variables with Positive Influence on Late Payment



Source: Author

4.5.2 Model E – Contract Amount

It is plausible that a longer boat also leads to the conclusion of berth contracts with a larger scope. Nevertheless, it is important to prove this with a regression analysis referring to ACI company data. It is therefore important to find out whether there is a demonstrable relationship between boat length and the scope of the contract. This knowledge is then used in a later classification analysis, with which the relationship between vessel type and contract volume can be demonstrated. The same data set is used for classification and regression analysis as in the previous clustering analysis. This investigation applies to the entire data set. It is assumed that the relationship between the attractiveness of the boat and the scope of the contract is the same for private and corporate customers. The selection of the variables relates to the contract amount as the target variable to examine the influence of the boat length and, together with the manufacturing date, the value of the boat with the contract amount. The same applies to vessel type. It is believed that more attractive sailboats and catamarans lead to more valuable contracts than is the case with customers with motor yachts.

The **statistical report** displays the minimal and maximum value for the estimator CONTRACT AMOUNT. In addition to the minimum and maximum, the mean and standard deviation are also given, showing plausible values. The mean value for the estimator “contract amount” shows that contract values in the six-digit range are rather an exception. The negative values

for the invoice amount, delay (number of days payable outstanding), and debt amount originate from the reversal of signs in the amounts when advance payments are made or when only partial amounts of the contract amount are invoiced.

Figure 51: Model E – Descriptive Statistical Report

Estimator rr_Contract amount

Data Set	Min	Max	Mean
Validation	68.45	265,107	12,606.8
Estimation	63.694	400,526	12,547.4

Variable	Data Set	Min	Max	Mean	Standard Deviation
VesselID	Estimation	11171	7074159	2,754,350	2,297,810
VesselID	Validation	11173	7074726	2,738,710	2,328,210
Vessel length	Estimation	3	41	11.482	3.111
Vessel length	Validation	3	41	11.441	3.039
Vessel manufacture date	Estimation	1896	2019	1,998.26	11.209
Vessel manufacture date	Validation	1896	2019	1,998.15	11.319
Starost plovila	Estimation	0	122	17.681	11.292
Starost plovila	Validation	0	123	17.84	11.412
ClientId	Estimation	4348	7074140	2,656,310	2,260,380
ClientId	Validation	4348	7074330	2,641,810	2,306,850
Client age	Estimation	18	89	58.191	11.428
Client age	Validation	18	89	58.21	11.504
Invoice amount	Estimation	-35469.6	399998	8,848.62	14,440.7
Invoice amount	Validation	-26021.8	192456	8,903.37	13,755.2
Ka?njenje	Estimation	-736	419	11.181	61.703
Ka?njenje	Validation	-432	403	10.525	60.572
Debt amount	Estimation	-192456	28902.4	-3,529.7	10,279.6
Debt amount	Validation	-143467	27345.2	-3,515.57	10,210.1
Debt amount %	Estimation	-667.32	200	18.24	33.649
Debt amount %	Validation	-689.31	191.5	17.654	34.381

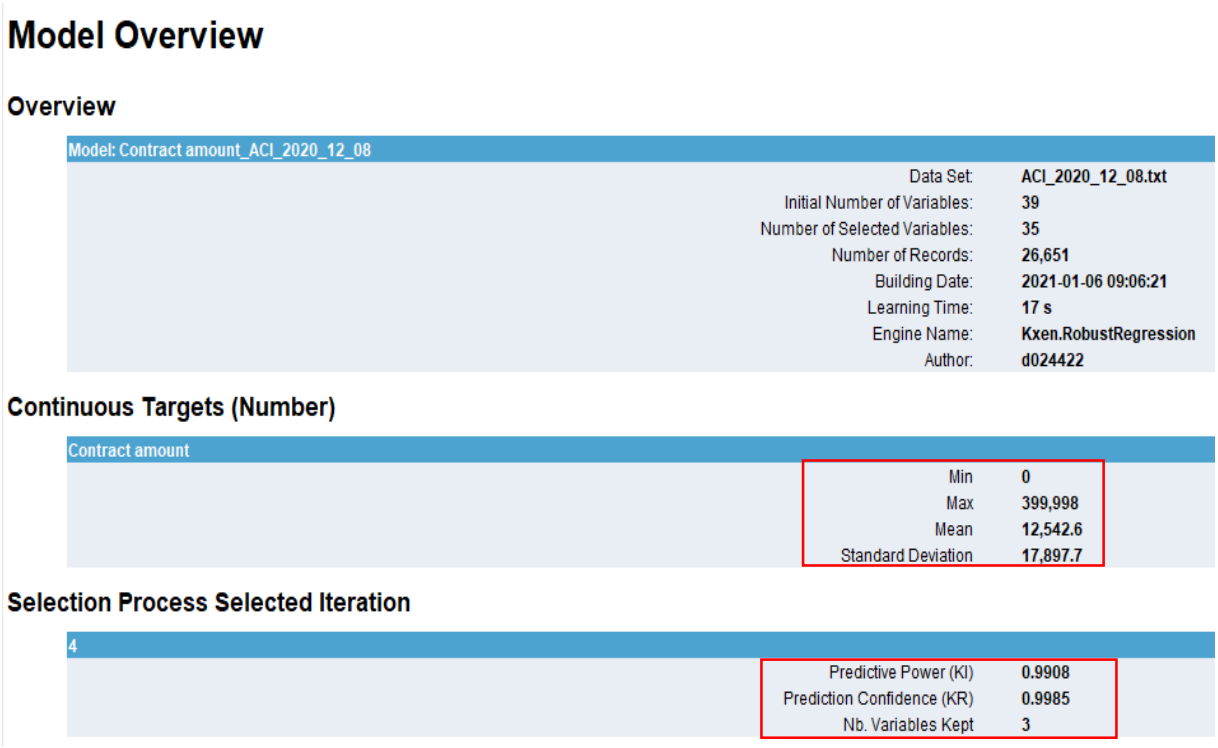
Data Set	Number of Records	Total weight
Estimation	19,864	19864
Validation	6,787	6787

Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Starost plovila – Vessel Age. Kašnjenje – Delay

Without setting a filter, the statistical report shows that the whole data set was selected for the prediction. The model overview displays acceptable values for KI and KR. The number of variables means that, during the iterations in the learning process, three variables contribute the most to the target variable ‘CONTRACT AMOUNT’.

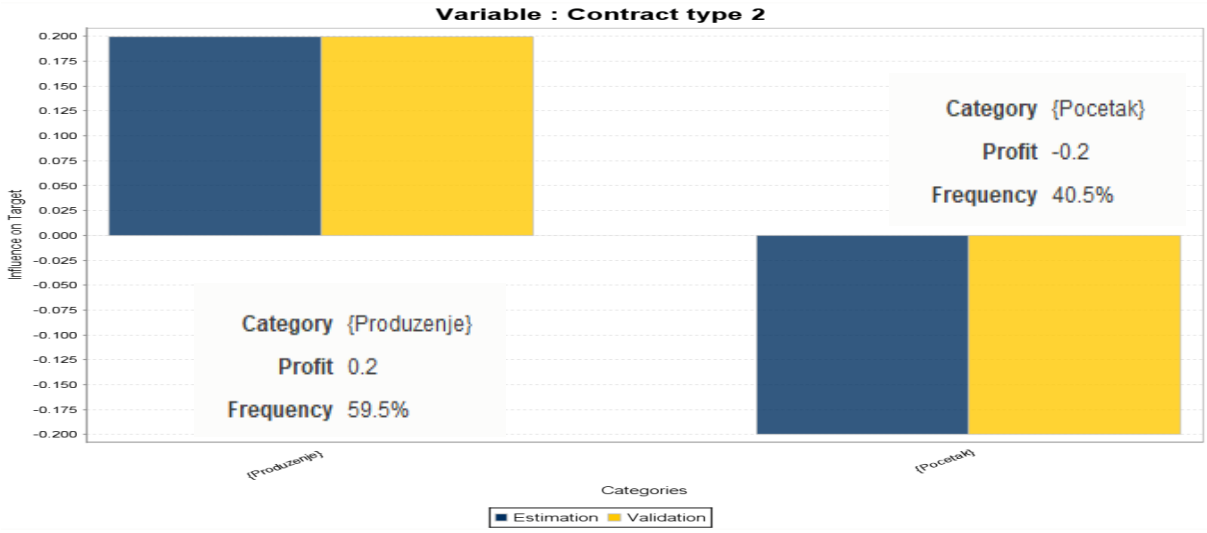
Figure 52: Model E – Overview



Source: Screenshot in SAP Predictive Analytics based on data created by the author

Although three variables contribute the most to the target variable, this does not mean that other variables have no significance. Invoice amount and debt amount naturally have a high correlation to the contract amount. However, only contract type 2 (renewal) is useful for an explanation. The figure below demonstrates that almost 60% of the sales volume has been generated with contract type renewal. Additionally, explanatory variables regarding the vessel have been detected as significant explanatory variables to make a decision on applicants from whom a high contract amount could be expected. New contracts tend to contain a smaller contract volume than renewed contracts. Assuming that new contracts are also new customers, positive customer experiences with the marina company play a decisive role when it comes to signing larger contracts.

Figure 53: Model E – Influence of Contract Type 2 on Contract Amount

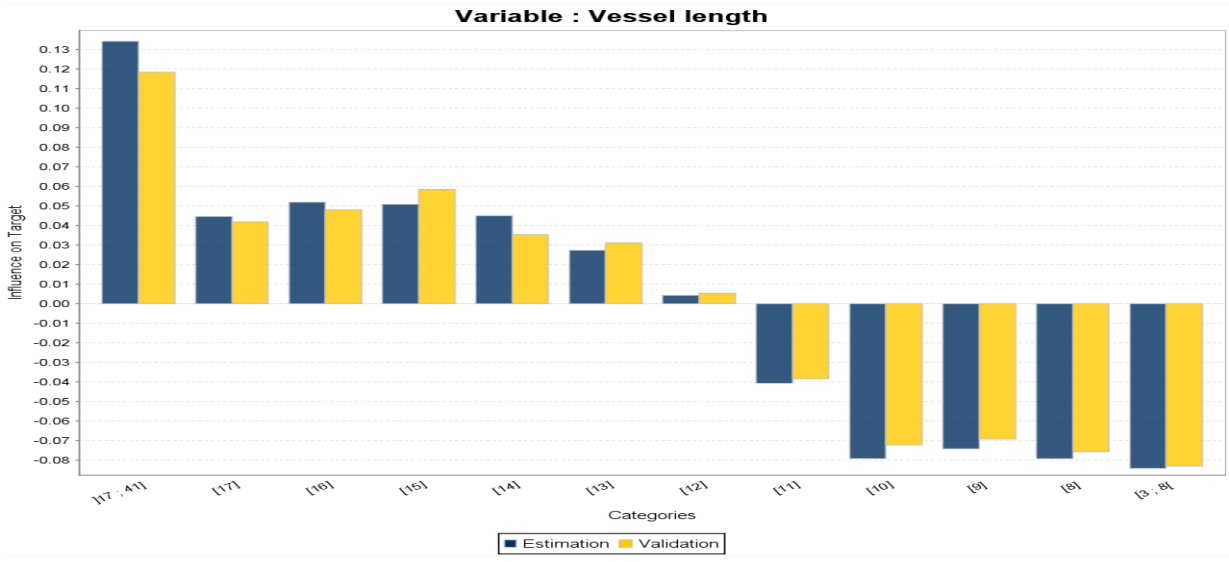


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Početak – New Contract. Produženje – Renewal

The analysis of vessel length shows that the owners of larger boats tend to conclude higher-quality contracts. This may seem plausible since a larger boat, naturally, also requires a larger berth. On the other hand, this could be compensated for by shorter idle times, which is not the case here. A positive correlation to the contract amount exists with vessel length from 12 to 41 meters.

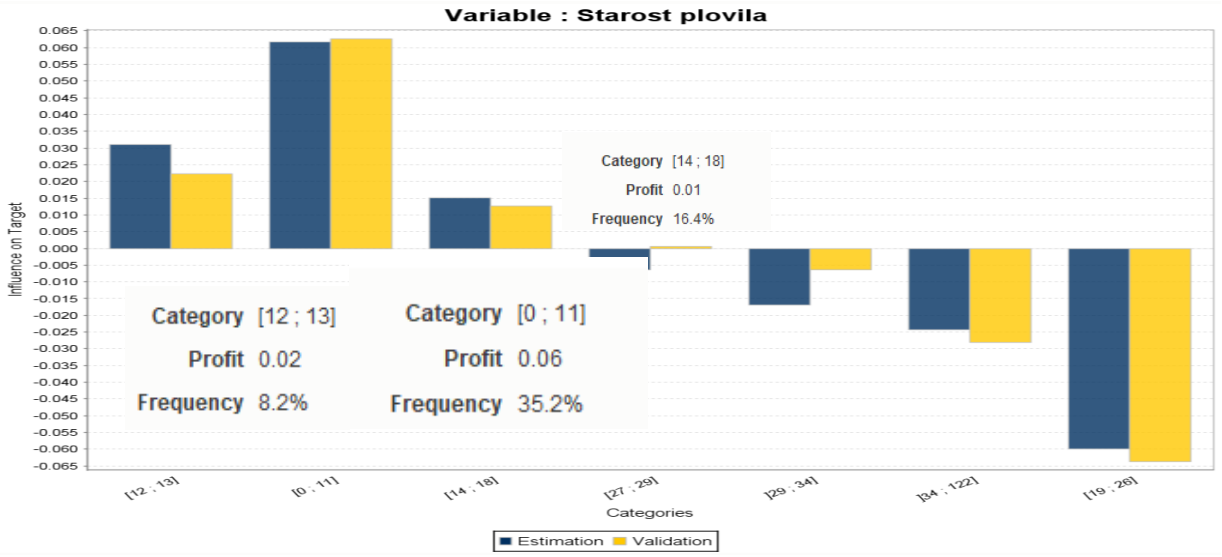
Figure 54: Model E – Influence of Vessel Length on Contract Amount



Source: Screenshot in SAP Predictive Analytics based on data created by the author

Analysis of vessel age shows that owners of younger boats are more likely to conclude contracts with higher volumes. Customer surveys will show whether this connection has to do with the creditworthiness of customers or with the fact that owners of older boats make fewer long-term plans. However, vessel age could be used to decide on the allocation of berths.

Figure 55: Model E – Influence of Vessel Age on Contract Amount

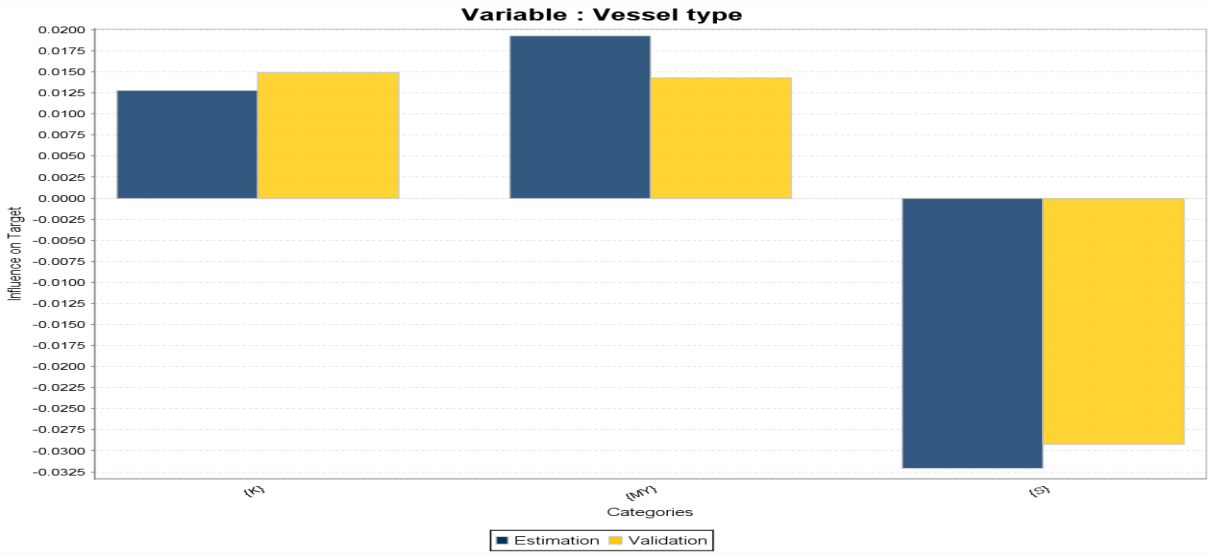


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Starost plovila: Vessel Age

While in previous evaluations it was more likely to identify sailing boats as the type of vessel whose owners renew contracts, the opposite is true for contract volume. The reason for this could be that among the customers with motor yachts, there are owners of large and luxurious motor yachts who tend to conclude high-value contracts.

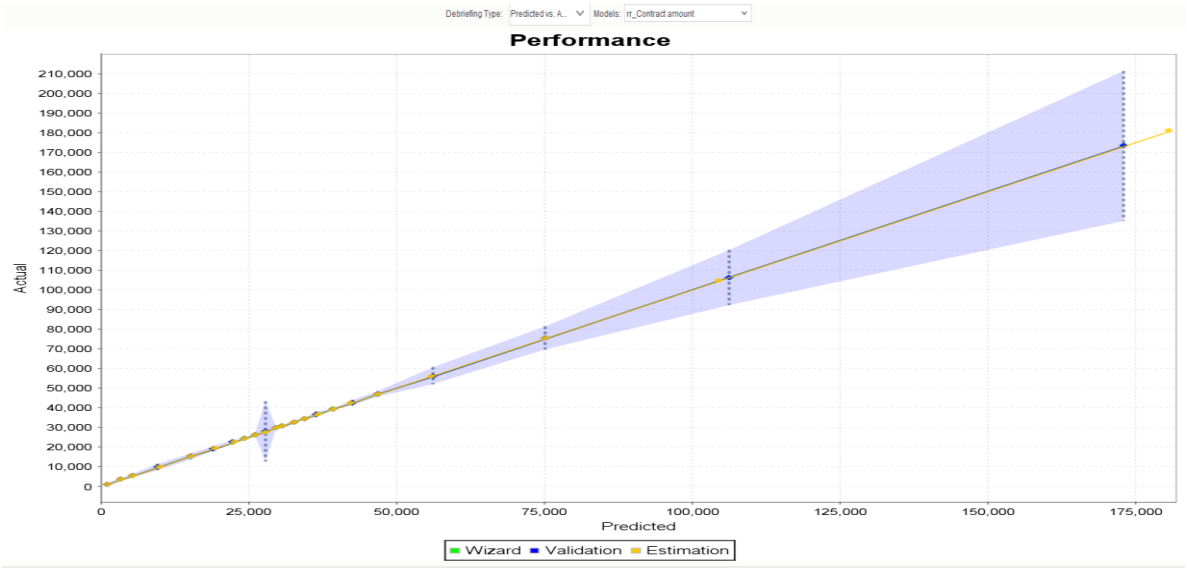
Figure 56: Model E: Influence of Vessel Type on Contract Amount



Source: Screenshot in SAP Predictive Analytics based on data created by the author

The evaluation of the performance of this regression shows that the forecast values largely correspond to the actual values up to a contract amount of HRK 50,000. For higher amounts, there is an increased distribution of the actual values. Nevertheless, the estimation and validation curves align very closely with the wizard. The graph is computed as follows: about 20 segments or bins of predicted values are built. Each of these segments represents roughly 5% of the whole population. For each of these segments, some basic statistics are computed on the actual value, such as the mean of the segment, the mean of the associated target, and the variance of the target within that segment. For each curve, a dot on the graph corresponds to the segment mean on the X-axis and the target mean on the Y-axis. (Bakhshaliyea et al., 2016, page 144).

Figure 57: Model E – Performance

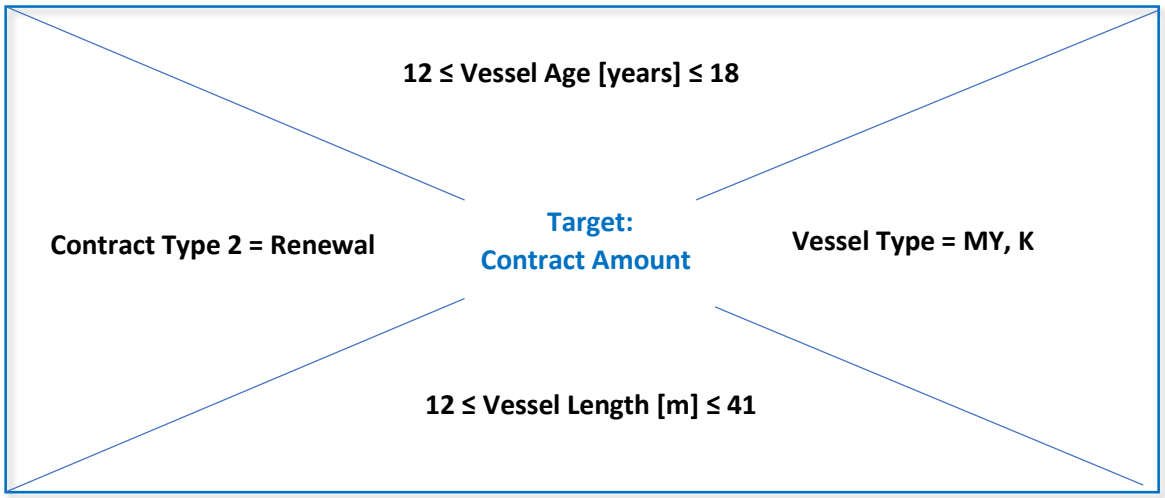


Source: Screenshot in SAP Predictive Analytics based on data created by the author

Summary of Model E

Due to certain characteristics, conclusion of higher-volume contracts can be predicted. Owners of motorboats and catamarans between 12 and 41 meters in length and 12 and 18 years old are particularly preferred when signing a contract renewal.

Figure 58: Model E – Summary of the Influence of Explanatory Variables



Source: Author

4.5.3 Model F – Customer Classification

The influence of characteristics of the clients' vessels is investigated within the previous predictions. Derived features are used in this model, which aims to classify the customers into customer groups. The basic ABC classification of customers: A – Customers are those customers from whom a high volume of orders and a long-term customer relationship can be expected. B – customers are expected to be in the middle range regarding these classification criteria. C – customers are to be assigned to the lower segment with smaller order volumes and rather short-term contracts. Nevertheless, the question arises, what exactly is an A, B, or C customer group for this marina company? To find out, a regression analysis was carried out, which shows the correlation between the properties of the customers' vessels and the order volume. It turns out that a plausible classification of customers into customer groups according to the ABC criterion can be carried out specifically for the marina company. A filter is set to client type "P" (private) to analyze whether the age of the client and, therefore, age group influences the preference for a charter vessel. Such a correlation between the age group of customers and the order volume could not be confirmed in this prediction model. It would therefore be possible to extend the investigation to include corporate customers. However, this was not done in this research. The findings from this investigation can, however, be used as a basis for further investigations for the customer type 'FIRM'. The target variable 'CHARTER VESSEL' has been set. The filter for the client type is set to "P" (PRIVATE CLIENT). Even if the vessel type 'CHARTER' does not represent the direct target variable, this characteristic was set as the target variable because a connection with the order volume could be demonstrated. In addition to the index KxIndex, the excluded variables are the variables for invoice ID and contract ID. In the first prediction, such variables were identified as monotonic variables. A monotonic variable increases and decreases in the same direction as the target variable, and at the same rate. For example, if you double the target variable, the monotonic variable will also exceed by a 100 %. The standard model parameters have been used in this prediction. The explanation given in the previous models also applies to this prediction. The **statistical report** displays the minimum and maximum values for the estimator 'CHARTER VESSEL'. In addition to the minimum and the maximum, the mean and standard deviation for the continuous variables showing plausible values is given. Negative values for the amounts like invoice amounts have been explained before in this research. If the client made a prepayment or has a credit note, or if they paid too early, it has been recorded as a negative value.

Figure 59: Model F – Statistical Descriptive Report

Estimator rr_Charter vessel

Data Set	Min	Max	Mean	Standard Deviation
Validation	-0.643	1.003	-0.001	0.422
Estimation	-0.64	0.997	-0.001	0.415

Variable	Data Set	Min	Max	Mean	Standard Deviation
Vessel length	Estimation	3	46	13.416	3.837
Vessel length	Validation	2	35	13.376	3.695
Client age	Estimation	0	86	52.593	10.879
Client age	Validation	1	85	52.699	10.877
Contract amount	Estimation	0	269999	39,042.5	29,808.3
Contract amount	Validation	0	261746	39,666.8	30,752.3
Contract amount due date	Estimation	2013-01-10 00:00:00	2020-01-01 00:00:00	2,016.61	1.766
Contract amount due date	Validation	2013-01-11 00:00:00	2019-08-22 00:00:00	2,016.7	1.793
Payment date	Estimation	2012-09-25 00:00:00	2019-07-07 00:00:00	2,016.32	1.937
Payment date	Validation	2012-09-25 00:00:00	2019-07-07 00:00:00	2,016.43	1.943
Invoice date	Estimation	2012-09-25 00:00:00	2019-07-07 00:00:00	2,016.32	1.937
Invoice date	Validation	2012-09-25 00:00:00	2019-07-07 00:00:00	2,016.44	1.941
Invoice amount	Estimation	-39504.1	269999	28,414	26,343.2
Invoice amount	Validation	-25770.6	261746	28,962	26,434.2
Kašnjenje	Estimation	-1058	586	34.935	67.441
Kašnjenje	Validation	-366	363	33.941	61.974
Debt amount	Estimation	-149477	16874.4	-10,569.2	17,809.4
Debt amount	Validation	-206438	240	-10,677.8	18,579.2
Debt amount %	Estimation	-150	185	23.456	31.613
Debt amount %	Validation	-11.05	300	23.318	30.946

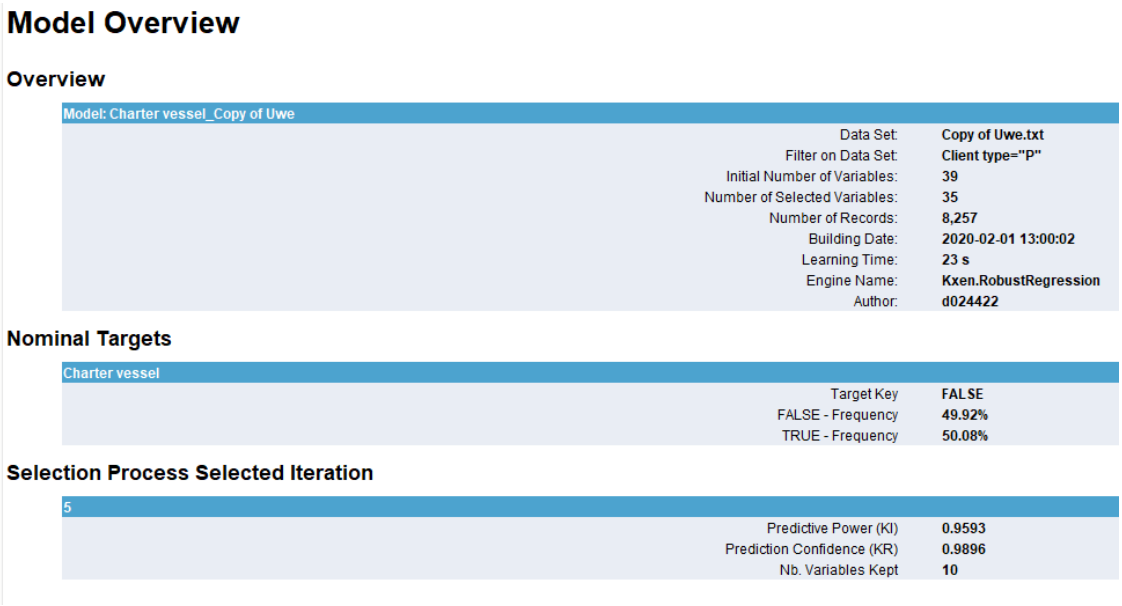
Data Set	Number of Records	Total weight
Estimation	6,108	6108
Validation	2,149	2149

Source: Screenshot in SAP Predictive Analytics based on data created by the author

Legend: Kašnjenje – Delay.

The model overview shows that 8,257 records have been selected. Predictive power KI with a value of about 0.9593 and predictive confidence KR with a value of 0.9896 meet the requirements. An interesting question, which will be discussed below, is the following: If the vessel type has any influence on the payment behavior or contract type 2 (RENEWAL – New Contract; New Contract – Renewal), clients who prefer an own vessel are the people who prefer renewing the contract.

Figure 60: Model Overview for the Classification ‘Charter Vessel’



Model Overview

Overview

Model: Charter vessel_Copy of Uwe	
Data Set:	Copy of Uwe.txt
Filter on Data Set:	Client type="P"
Initial Number of Variables:	39
Number of Selected Variables:	35
Number of Records:	8,257
Building Date:	2020-02-01 13:00:02
Learning Time:	23 s
Engine Name:	Kxen.RobustRegression
Author:	d024422

Nominal Targets

Charter vessel	
Target Key	FALSE
FALSE - Frequency	49.92%
TRUE - Frequency	50.08%

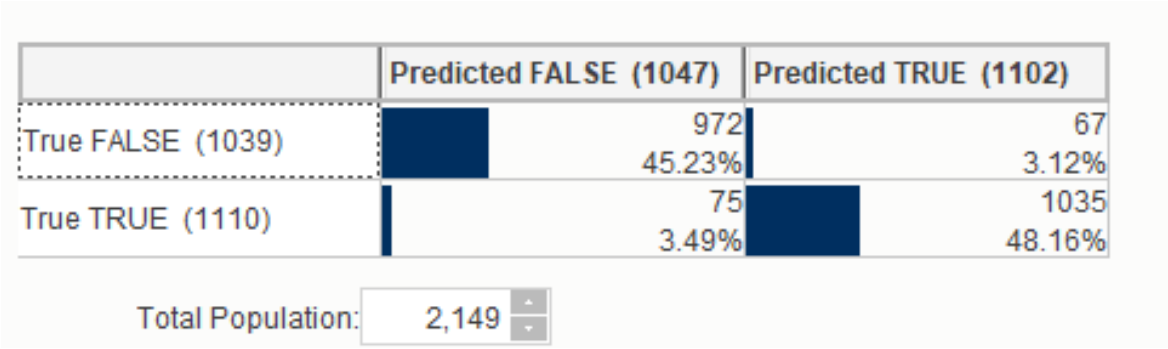
Selection Process Selected Iteration

5	
Predictive Power (KI)	0.9593
Prediction Confidence (KR)	0.9896
Nb. Variables Kept	10

Source: Screenshot in SAP Predictive Analytics based on data created by the author

Confusion Matrix

Figure 61: Confusion Matrix Classification ‘Vessel Type’



	Predicted FALSE (1047)	Predicted TRUE (1102)
True FALSE (1039)	972 45.23%	67 3.12%
True TRUE (1110)	75 3.49%	1035 48.16%

Total Population: 2,149

Source: Author

Legend for the previous figure is given in the following figure:

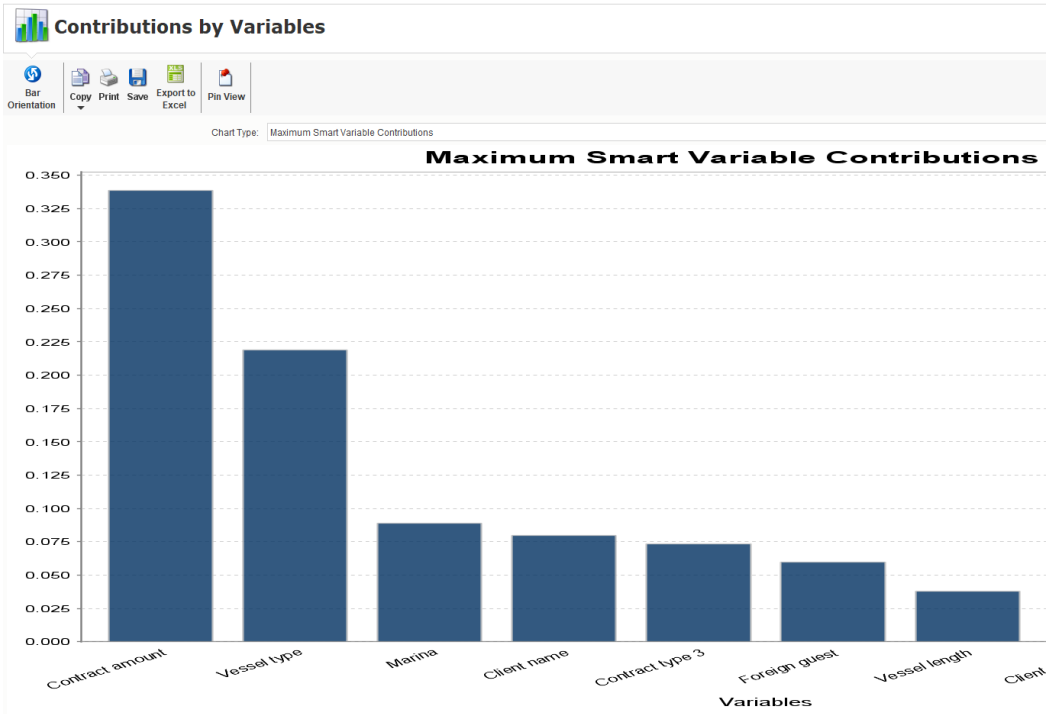
Table 2: Confusion Matrix

	Predicted (target category) Positive observations predicted	Predicted (Non-target category) Negative observations predicted
True (target category) Actual positive correlations	**Number of correctly predicted positive observations**	Number of actual positive observations with a negative prediction
True (non-target category) Actual negative observations	Number of actual negative observations with positive prediction	**Number of correctly predicted negative observations**

Source: Author

The confusion matrix of the prediction model shows that almost all the positive and negative observations were predicted correctly. With the analysis of the contribution of each variable to the target 'vessel type', the following picture shows that the contract amount has the highest influence on the target variable 'charter vessel' followed by 'vessel type'. This means that, if the decision-makers or decision-making in controlling prefer customers with their boats, attention should be devoted to vessel type.

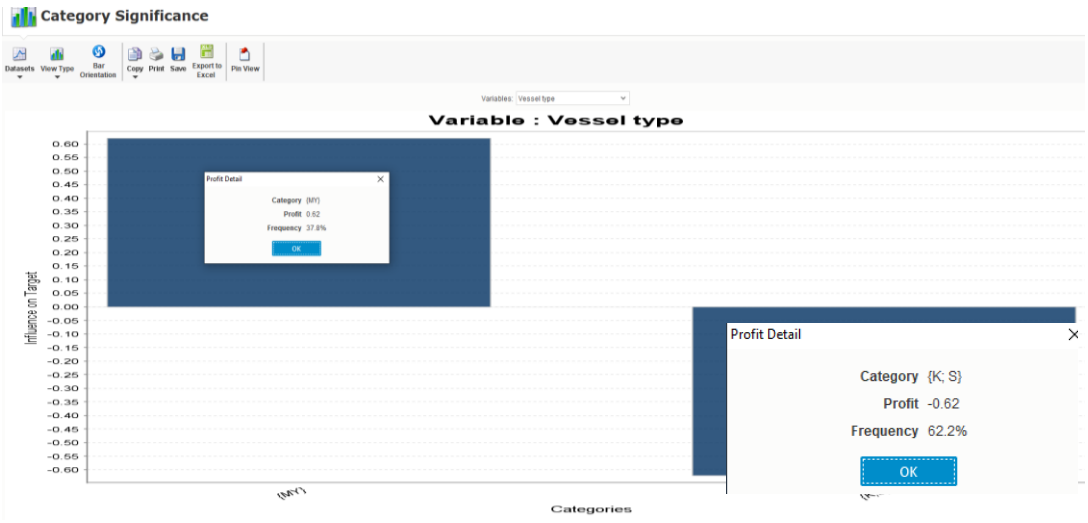
Figure 62: Variable Contributions to Classification ‘Vessel Type’



Source: Screenshot in SAP Predictive Analytics based on data created by the author

The following figure demonstrates a positive correlation between the vessel type ‘motorboat’ to the target variable ‘charter boat’ and vice versa. However, representative statistics on nautical tourism in Croatia indicate that the share of sailing boats in charter vessels has been increasing continuously for many years (YachtRent, 2014). However, it must also be considered that only private clients were considered in this research. In addition, this prediction model focuses on the classification, and not on the relationship between vessel type and charter vessel.

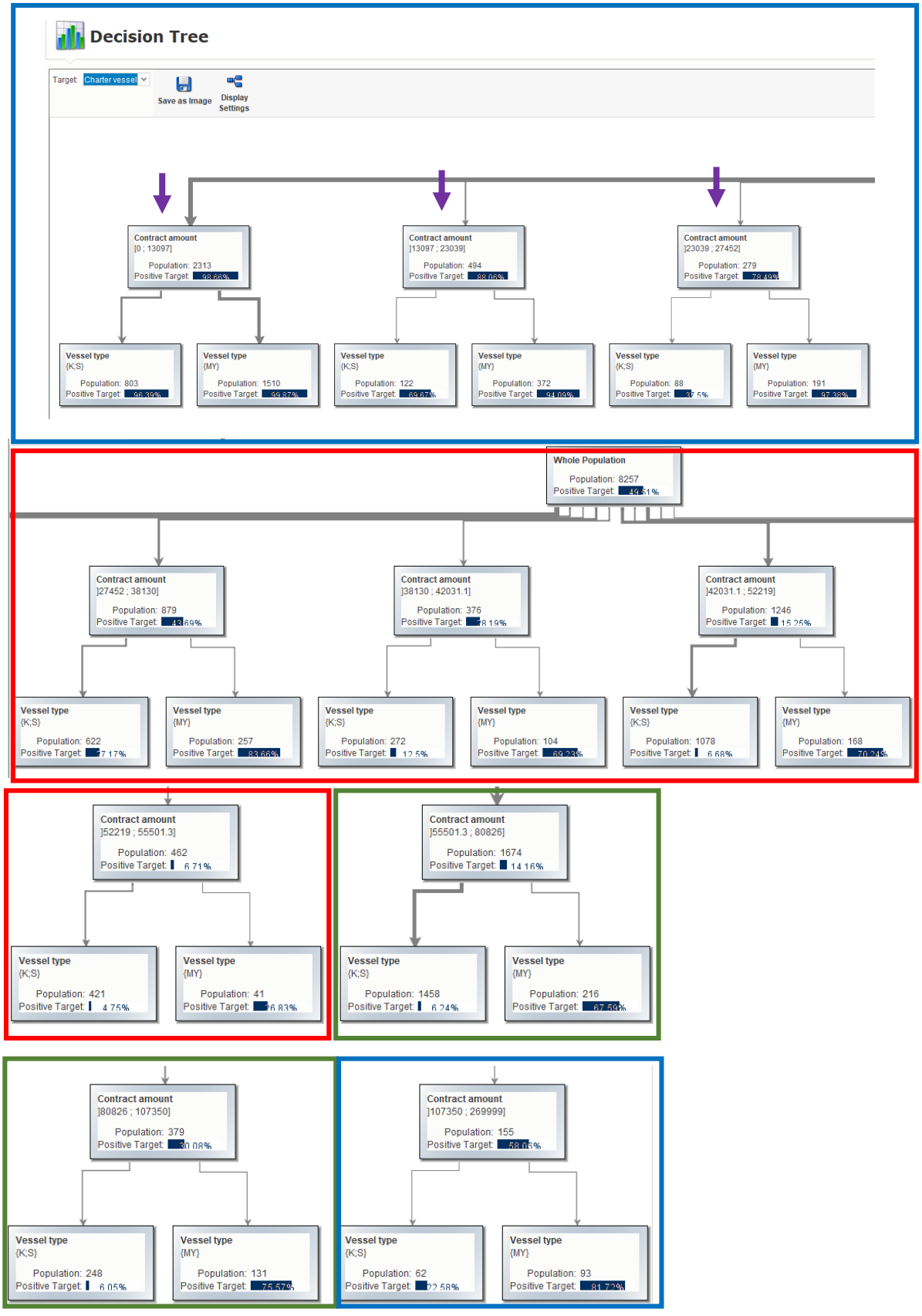
Figure 63: Category ‘Significance of Vessel Type for Contract Amount’



Source: Screenshot in SAP Predictive Analytics based on data created by the author.

The following figure shows the relationship between the variable ‘CONTRACT AMOUNT’ and the variable ‘VESSEL TYPE’. The first box shows the relationship between the vessel types and the sales ranges for the ranges HRK 0 to HRK 13,097.00, then HRK 13.097.00 to HRK 23,039.00 and finally, HRK 23,039.00 to HRK 27,452.00. The second box shows further relationship. The further course is the sales area from HRK 27,452.00 to HRK 38,130.00, etc. The decision tree shows results generated by the regression engine based on the five most contributing variables. The figure below shows the contribution of the explanatory variable ‘VESSEL TYPE’ to sales. On this basis, the author of this research grouped the sales areas into four areas and documented the role of vessel type.

Figure 64: Decision Tree Classification ‘Vessel Type’



Source: Screenshot in SAP Predictive Analytics based on data created by the author

The classification into customer groups was made using a decision tree generated by the system. The correlation between the contract amount and the preference to use a sailing boat can be analyzed with the decision tree of the prediction. The following correlation has been detected:

Table 3: Development of the Correlation Between Contract Amount and Vessel Type

Sales [HRK]	Population K (catamaran) & S (sailboat) [HRK]	Population MY (motor yacht) [HRK]	$\frac{K,S}{MY}$	Phase
0 – 13,097	803	1,510	0,53	I
13,097 – 23,039	122	372	0,32	I
23,039 – 27,452	88	191	0,46	I
27,452 – 38,130	622	257	2,42	II
38,130 – 42,031.1	272	104	2,61	II
42,031.1 – 52,219	1,078	168	6,42	II
52,219 – 55,501.3	421	41	10,26	II
55,501.3 – 80,826	1,458	216	6,75	III
80,826 – 107,350	248	131	1,89	III
107,350 – 269,999	62	93	0,67	IV

Source: Author

Up to the contract size of HRK 27,452, the proportion of customers with a motor yacht is larger than those with a catamaran or sailboat. In phase I, the proportion of K and S increases on average. In phase II, the proportion of K and S is higher than MY and increases. In phase III, the proportion of K and S is still higher than MY, but is decreasing. In phase IV, which represents the most expensive contracts, the proportion of MY is higher again. In conclusion, the more extensive the contract, the more likely the use of a catamaran or a sailboat. The owners of motor yachts, on the other hand, have very extensive contracts. The reason for this could be the fact that a customer with an extensive berth contract has an expensive luxury motor yacht. This could not be proven directly, because the data set does not contain the vessel price. However, the data set contains the vessel manufacturing date and the vessel length. Therefore, the correlation between vessel manufacturing date and vessel length in relation to the contract amount might be interesting to analyze. This has been proven in the corresponding previous prediction using regression analysis for the correlation of contract amount and vessel type.

Conclusion

The biggest business for the marina company lies in the sales area from HRK 27,452 to HRK 80,826. In this sales area, vessel type 'sailing boat' is dominant. If the controlling decides on berth allocation, the following is recommended: if the aim is conclusion of long-term contracts with higher sales, customers with sailboats should be preferred.

Figure 65: Customer Groups Based on Sales Volume



Source: Author

As a general conclusion, it can be stated that predictive analysis used in this research is a part of data science. The scientific methodology used here is pattern recognition based on methods of mathematics (structured risk minimization) and computer science (support vector machine). The result of the generated prediction models is methodically tested with the key performance indicators predictive power and predictive robustness. It is, therefore, not just a procedure, but a programmed algorithm that has its scientific basis in statistical learning theory.

5 THE INFLUENCE OF PREDICTIVE MODELS ON DECISION BEHAVIOR

The decision-making situations are often just as diverse as decision-making alternatives. The comparison of decision situations with decision alternatives results in a matrix (compare to BAMBERG12, page 110). The matrix was changed by the author of this research. The original decision matrix is in the German language and therefore the horizontal expressions are “z” for the German word “Zustände”. The author of this research changed this expression into “s”, which stands for “scenario”. Basically, a decision state is to be understood as the same as a decision scenario.

Figure 66: Decision Matrix

	s_1	…	s_n
a_1	$a_1 s_1$	⋯	$a_1 s_n$
…	⋮	⋱	⋮
a_n	$a_n s_1$	⋯	$a_n s_n$

Different scenarios s_1, … s_n are combined with different decision alternatives a_1, … a_n.

Source: Author

A direct comparison of all combinations of the occurring decision scenarios and decision alternatives is often not possible due to the heterogeneity of the data. To enable comparability, data must be structured. For example, an evaluation of the influence of customer age on their decision-making behavior is usually not meaningful because the exact age of the customers is too widely dispersed. Customers are therefore assigned to age groups. Grouping should be carried out in such a way that it can be assumed that an age group shares the same decision-making behavior. This assumption can be statistically proven by data science application of pattern recognition. It is concluded that a calculable decision base for choosing a decision alternative with the best expectation of future scenarios can only be developed with the use of predictive analysis. Otherwise, due to lack of comparability of unstructured data, decisions are subjective and difficult to understand. Decision processes can only be modeled and optimized using a statistically proven decision base.

5.1 Expert Survey with the Presentation of Case Studies

The data provided by Adriatic Croatia International Club (ACI) were used to develop prediction models using the software solution SAP Predictive Analytics. It has been used to develop verified predictions about customer behavior. The data refer to the contracts concluded between ACI and berth-renting customers. This includes data about the customers (age, age group, nationality), the boat (year of production, age, length of the boat) and the contracts (sales, type of contract). The study is based on ACI's interest in reaching good capacity utilization through long-term customers. Another aspect associated with good capacity utilization are revenues and profit. Significant factors influencing customer behavior and business figures were identified using the mathematical-statistical methods for clustering, regression calculation and classification. The objective of this research is to discover the influence of the prediction model results on controllers' decisions on assigning berths to applicants. In the first step, ACI controllers were surveyed without knowledge of the prediction results. The prediction results were presented in the second step. In the second phase, the same questions were asked again. According to the Bayes' theory, controllers should change their a-priori decisions into a-posteriori decisions. Both a-priori decisions and a-posteriori decisions are compared with statistically proven results of the prediction models and the hit rate is measured. The survey basis are previously developed prediction models, which are summarized in the following explanations.

Model A: Private clients who conclude a yearly contract. Berth allocation to private clients in order to conclude a yearly contract instead of a monthly contract. The influencing variables are client age and citizenship. The participants had multiple requests; there were 30 applicants for 10 berths. The goal was to allocate the berths to those more willing to sign a yearly contract (GV). The participants in the questionnaire had to decide on whom they would allocate the free berths.

Model B: Firms that conclude a yearly contract. Berth allocation to clients of client type **FIRM** in order to conclude a yearly contract instead of a monthly contract. The influencing variables are citizenship and vessel length. The participants in the questionnaire had multiple requests, 30 applicants for 10 berths. The goal was to allocate berths to those more willing to sign a yearly contract (GV). The participants had to decide on assigning 10 berths to clients of whom they expected the conclusion of a yearly contract.

Model C: Private clients who would like to renew their contract. Berth allocation to clients of client type **PRIVATE** in order to achieve contract renewal (RENEWAL). The influencing variables are age and vessel type. The participants in the questionnaire had multiple requests; 30 applicants for 10 berths. The goal is to allocate berths to those more willing to renew their contract. The participants had to decide on allocating 10 berths to clients of whom they expected contract renewal upon expiration of the first contract.

Model D: Poor payment behavior. The objective is to detect clients who are significantly late in their payments. The target variable is DELAY (days payable outstanding). To minimize financing costs and avoid financing gaps, customers who are at lower risk of late payments are preferred. The participants in the questionnaire had multiple requests; 30 applicants for 10 berths. The goal was to allocate berths for those with better payment behavior. Influencing variables are contract type 3, vessel type, and age of private clients. The participants had to decide on whom they would allocate the free berths.

Model E: Regression between vessel manufacturing date, vessel length, and contract amount. To increase sales and profits, it is useful to know which customers can generate higher sales. Customers with larger boats pay a higher price for a berth. However, sales can only be increased if these customers also book the berth for a comparatively longer period. The participants in the questionnaire had multiple requests; 30 applicants for 10 berths. The goal was to allocate berths to those with higher sales and profit. The influencing variables are vessel manufacturing date and vessel length. The participants had to decide on whom they would allocate the free berths.

Model F: Classification 'vessel type' with the contract amount. The decision scenario is the planning of a marketing campaign for three customer segments: customers in the lower sales segment, category A, customers in the middle sales segment, category B, and customers in the high sales segment, category C. The background of the decision is the requirement to define different marketing campaigns appropriate for customer groups. They are convinced that different marketing messages should be sent to these three customer groups: low-cost short-term contracts, berth contracts in the business class, and exclusive contracts. The following sales data according to vessel types were given to survey participants. The participants in the survey are four employees of the controlling department of the marina company. An external person also took part in the survey. The employees and the external person have many years of experience in the marina industry in general and the employees have many years of professional

experience. The participants were asked to assign the category A, B, and C to the customers in the list.

Table 4: Customer Sales

Client	**Customer characteristics**	**Key figures**	**Group into category A, B, and C**
1	Vessel type	Sales	???
2	Age group	Contract Renewal	???
3	Citizenship	Payment Behavior	???
…	…	…	…

Source: Author.

The same six models were given to the survey participants. The questionnaire was carried out to measure the difference of the hit rate with and without the knowledge of prediction results. A significant increase in the hit rate with knowledge of the prediction results has been demonstrated.

5.2 Case Study Solutions

The evaluation of the hit rates from the questionnaire depends on the comparison of the hit rates with and without the knowledge of prediction results.

a) Solution Case Study A – Yearly Contract with a Private Client

The following task was assigned to the survey participants: *The allocation of berths to private clients in order to conclude a yearly contract instead of a monthly contract. The influencing variables are client age and citizenship. You have multiple requests of 30 applicants for 10 berths. Your goal is to allocate berths to those who are more willing to sign a yearly contract (GV). Please assign the 10 berths to the clients from whom you expect a yearly contract.*

Table 5: Solution Case Study Model A

Client	Client Type	Client Age	Citizenship	Yearly Contract: GV Monthly Contract: MV **SOLUTION**
1	P	58	Slovenia	**GV**
2	P	23	Croatia	**MV**
3	P **F!**	23	Russia	**GV**
4	P	59	Luxembourg	**GV**
5	P **F!**	23	Austria	**MV**
6	P **F!**	30	Germany	**MV**
7	P	30 **in every age**	Slovenia	**GV**
8	P	36	Germany	**GV**
9	P	36	Croatia	**MV**
10	P	40	Croatia	**MV**
11	P	40 **in every age**	Germany	**GV**
12	P	40	Germany	**GV**
13	P	52	Croatia	**GV**
14	P	53	Slovenia	**MV**
15	P	53 **56!**	Belgium	**MV**
16	P	53 **51, 52**	Poland	**MV**
17	P	54	Croatia	**GV**
18	P	54	Germany	**GV**
19	P F!	55	Israel	**GV**
20	P	55	Croatia	**MV**
21	P	57	France	**GV**
22	P	57	Slovenia	**GV**
23	P	59	Croatia	**MV**
24	P	59	Poland	**GV**
25	P	60	Croatia	**GV**
26	P **F!**	60	New Zealand	**GV**
27	P	67	Germany	**MV**
28	P	67	Croatia	**MV**
29	P **F!**	71	Switzerland	**MV**
30	P	77	Croatia	**MV**

Source: Author

b) Solution Case Study B – Yearly Contract with a Firm

Allocation of berths to clients of client type FIRM**:** The objective is to conclude a YEARLY CONTRACT instead of a MONTHLY CONTRACT. The influencing variables are citizenship and vessel length. The task assigned to survey participants is identical to the one from the previous survey, but regarding client type FIRM.

Table 6: Solution Case Study Model B

Client	Client Type	Vessel Length (m)	Citizenship	Yearly Contract: GV Monthly Contract: MV **SOLUTION**
1	F	4	Italy	**MV**
2	F	4	Germany	**MV**
3	F	5	Croatia	**GV**
4	F	5	Germany	**MV**
5	F	10	Austria	**GV**
6	F	10	Croatia	**MV**
7	F	10	Slovenia	**GV**
8	F	11	Austria	**GV**
9	F	11	Croatia	**GV**
10	F	11	Germany	**GV**
11	F	11	Sweden	**GV**
12	F	11	Slovenia	**GV**
13	F	11	Italy	**MV**
14	F	12	Austria	**GV**
15	F	12	Germany	**GV**
16	F	12	Italy	**GV**
17	F	13	Austria	**GV**
18	F	13	Italy	**MV**
19	F	13	Croatia	**GV**
20	F	13	Hungary	**MV**
21	F	13	Czech Republic	**GV**
22	F	13	Slovenia	**MV**
23	F	14	Croatia	**MV**
24	F	15	Sweden	**GV**
25	F	15	Russia	**MV**
26	F	16	Austria	**GV**
27	F	16	Croatia	**GV**
28	F	29	Austria	**GV**
29	F	33	Croatia	**GV**
30	F	41	Germany	**GV**

Source: Author.

c) Solution Case Study C – Contract Renewal Private Client

Allocation of berths to clients of client type PRIVATE in order to achieve a renewal of the contract (RENEWAL). The influencing variables are age and vessel type. The task assigned to the survey participants was the following: *You have multiple requests; 30 applicants for 10 berths. Your goal is to allocate berths to those more willing to renew their contracts. Please assign the 10 berths to the clients of whom you expect a contract renewal when concluding a new contract. Please enter your decision in the table below.*

Table 7: Solution Case Study Model C

Client	Client Type	Age	Vessel Type	New Contract vs. Renewal
1	P	23	MY	NEW CONTRACT
2	P	28	MY	NEW CONTRACT
3	P	29	MY	NEW CONTRACT
4	P	30	MY	NEW CONTRACT*
5	P	31	MY	NEW CONTRACT
6	P	32	MY	**RENEWAL**
7	P	33	MY	NEW CONTRACT
8	P	34	MY	NEW CONTRACT**
9	P	56	S	**RENEWAL**
10	P	56	MY	**RENEWAL**
11	P	57	S	**RENEWAL**
12	P	58	MY	NEW CONTRACT
13	P	58	S	**RENEWAL**
14	P	59	MY	**RENEWAL**
15	P	59	S	**RENEWAL**
16	P	60	S	**RENEWAL**
17	P	62	MY	**RENEWAL**
18	P	62	S	**RENEWAL**
19	P	63	S	**RENEWAL**
20	P	64	S	**RENEWAL**
21	P	64	MY	NEW CONTRACT
22	P	65	S	**RENEWAL**
23	P	65	MY	**RENEWAL**
24	P	66	MY	**RENEWAL**
25	P	67	MY	NEW CONTRACT
26	P	67	S	**RENEWAL**
27	P	68	S	**RENEWAL**
28	P	68	MY	**RENEWAL**
29	P	69	S	**RENEWAL**
30	P	70	MY	**RENEWAL**

* If yearly contract (GV) RENEWAL
** 50% NEW CONTRACT, 50% RENEWAL

Source: Author

d) Solution Case Study D – Poor Payment Behavior

The objective is to detect the clients who are significantly late in payment. The target variable is DELAY (days payable outstanding). To minimize financing costs and avoid financing gaps, customers who are at lower risk of late payments are preferred. Not all clients use a charter vessel. The task assigned to the survey participants was the following: *You have multiple*

requests; 30 applicants for 10 berths. Your goal is to allocate berths to those of whom you expect better payment behavior. Influencing variables are contract type 3, vessel type, and age of private clients. Please enter your decision in the table below.

Table 8: Solution Case Study Model D

Client	Contract Type 3	Years of operation of the FIRM	DELAY*	Expected good payment behavior **SOLUTION**
1	M.6MJ	23	-3	**YES**
2	M.GVP	1st FIRM	Average	NO
3	M.ZIM	24	-1	**YES**
4	M.GV	27	-36	**YES**
5	M.GVP	2nd FIRM	-10	NO
6	M.MV	29	-9	**YES**
7	M.GV	3rd FIRM	224	NO
8	M.GVP	4th FIRM	0	**YES**
9	M.GVP	30	-1	**YES**
10	M.GV	31	253	NO
11	M.V2K.M	32	-2	**YES**
12	M.V2K.M	37	6	**YES**
13	M.GV	5th FIRM	-5	**YES**
14	M.GVP	43	187	NO
15	M.GV	44	-3	**YES**
16	M.GV	6th FIRM	-36	**YES**
17	M.GVP	44	-11	**YES**
18	M.GVP	7th FIRM	-1	**YES**
19	M.GVP	46	-29	**YES**
20	M.GVP	48	31	NO
21	M.GV	48	-2	**YES**
22	M.GV	8th FIRM	-68	NO
23	M.GV	50	54	NO
24	M.GV	51	14	NO
25	M.GVP	63	346	NO
26	M.GVP	9th FIRM	0	**YES**
27	M.GVP	63	229	NO
28	M.MV	64	44	NO
29	M.MV	65	-1	**YES**
30	M.GV	66	23	NO

- Minus means “prepayment”.

Source: Author

e) Solution Case Study E – Contract Amount

To increase sales and profits, it is interesting to know which customers can generate higher sales. The task assigned to the survey participants was the following: *You have multiple*

requests; 30 applicants for 10 berths. Your goal is to allocate berths to those with higher sales and profit. The influencing variables are the vessel manufacturing date and vessel length. Please enter whom you would allocate the free berths to. A large contract amount starts with HRK 44,851.00.

Table 9: Solution Case Study Model E

Client	Vessel Manufacturing Date (YEAR)	Vessel Length (m)	Contract Amount	Expected Large Contract Amount **SOLUTION**
1	2017	4	2,052	NO
2	2007	9	6,019	NO
3	2016	8	4,520	NO
4	1990	10	2,292	NO
5	1991	11	19,193	NO
6	2013	28	174,585	**YES**
7	2017	16	100,575	**YES**
8	2002	23	130,209	**YES**
9	2007	19	108,434	**YES**
10	1999	11	15,508	NO
11	2006	9	1,120	NO
12	1989	9	3,610	NO
13	2003	9	27,591	NO
14	1989	10	6,918	NO
15	1984	9	21,422	NO
16	1986	14	33,998	NO
17	2011	14	67,002	**YES**
18	2013	30	47,847	**YES**
19	2006	16	71,174	**YES**
20	2013	14	78,826	**YES**
21	1961	14	2,176	MP
22	2006	9	3,183	NO
23	2007	23	1,447	NO
24	2012	15	71,719	**YES**
25	2018	17	77,720	**YES**
26	2009	17	90,594	**YES**
27	1945	41	399,998	**YES**
28	2009	46	60,000	**YES**
29	2008	7	3,716	NO
30	1995	10	1,330	NO

Source: Author

f) Solution Case Study F – Customer Group

The objective of this case study is an ABC categorization of customers: A-Customers with a large annual turnover versus B-Customers with a medium-, and C-Customers with a rather small turnover. On this basis, customer relationship management can be better addressed, for example through extended services and offers for long-term contracts, which could be addressed more to A-customers.

The prediction using the classification method results in the following groups:

Group I (Segment A):	Lower sales segment with vessel type MY
Group II and III (Segment B):	Mid-range segment with vessel type S and K
Group III (Segment C):	High sales segment with vessel type MY

A different marketing message should be addressed to three different customer groups. To develop the marketing campaign according to target groups, marketing controlling must identify the target groups. Regarding this objective, the following survey was carried out:

The task assigned to the survey participants was the following: *You would like to start three different marketing campaigns with which you can target the following customer groups:*

Customers in the lower sales segment, category A

Customers in the middle sales segment, category B

Customers in the high sales segment, category C.

They are convinced that different marketing messages should be addressed to these three customer groups: low-cost short-term contracts, berth contracts in the business class, and exclusive contracts. The following sales data according to vessel types are available. Assign category I, II and III to the sample data.

Regarding the prediction results, the customer segment is allocated in the following table.

Table 10: Customer Segments

Client	Vessel Type	Sales	Customer Segment A, B, or C **SOLUTION**
1	MY	3183	**A**
2	S	3283	**A**
3	MY	3767	**A**
4	S	22968	**A**
5	K	23022	**A**
6	MY	25279	**A**
7	S	28256	**B**
8	MY	28359	**B**
9	S	28486	**B**
10	S	30427	**B**
11	S	31729	**B**
12	S	34050	**B**
13	MY	35008	**B**
14	S	39846	**B**
15	S	41374	**B**
16	MY	41386	**B**
17	K	41772	**B**
18	S	42031	**B**
19	S	48349	**B**
20	K	48558	**B**
21	K	48768	**B**
22	MY	96367	**B**
23	K	123184	**C**
24	K	125648	**C**
25	MY	130029	**C**
26	K	142458	**C**
27	MY	148085	**C**
28	MY	265565	**C**
29	MY	269998	**C**
30	MY	399998	**C**

Source: Author

5.3 Evaluation of the Hit Rate of the Participants

The HIT RATE is the key figure in this hypothesis-testing case study, which is used to measure the correctness of decision making of survey participants. The hypothesis of better a-posteriori decisions made by the participants in the survey is measured by the number of hits, which corresponds to the hits statistically proven by the prediction models. This means that a HIT is the correct selection of an applicant or a customer regarding the case study question.

Table 11: Hit Rate of Participants in the Case Study

Average Ratings of Participant I

Model	WITHOUT KNOWLEDGE PREDICTION		WITHIN KNOWLEDGE PREDICTION		% HIT WITHOUT PREDICTION	% HIT WITH PREDICTION
	HIT	FALSE	HIT	FALSE		
A	6	4	8	2	60	80
B	8	2	8	2	80	80
C	6	4	8	2	60	80
D	2	8	6	2	20	60
E	6	4	9	1	60	90
F	15	15	23	7	50	77

Average Ratings of Participant II

Model	WITHOUT KNOWLEDGE PREDICTION		WITHIN KNOWLEDGE PREDICTION		% HIT RATE WITHOUT PREDICTION	% HIT RATE WITH PREDICTION
	HIT	FALSE	HIT	FALSE		
A	7	3	6	5	70	60
B	7	3	10	0	70	100
C	9	1	10	0	90	100
D	4	6	6	4	40	60
E	7	3	7	3	70	70
F	21	9	29	1	70	97

Average Ratings of Participant III

Model	WITHOUT KNOWLEDGE PREDICTION		WITHIN KNOWLEDGE PREDICTION		% HIT RATE WITHOUT PREDICTION	% HIT RATE WITH PREDICTION
	HIT	FALSE	HIT	FALSE		
A	6	4	8	2	60	80
B	4	6	8	2	40	80
C	10	0	8	2	100	80
D	17	13	5	5	57	50
E	24	6	9	1	80	90
F	26	4	26	4	87	87

Average Ratings of Participant IV – External

Model	WITHOUT KNOWLEDGE PREDICTION		WITHIN KNOWLEDGE PREDICTION		% HIT RATE WITHOUT PREDICTION	% HIT RATE WITH PREDICTION
	HIT	FALSE	HIT	FALSE		
A	6	3	8	2	60	80
B	8	2	6	2	80	60
C	6	2	7	3	60	70
D	5	5	3	7	50	30
E	5	5	7	3	50	70
F	6	4	6	4	60	60

AVERAGE

Model	WITHOUT KNOWLEDGE PREDICTION		WITHIN KNOWLEDGE PREDICTION		% HIT RATE WITHOUT PREDICTION	% HIT RATE WITH PREDICTION
	HIT	FALSE	HIT	FALSE		
A	6,25	3,3	7,5	2,75	**62,5**	**75**
B	6,75	3,25	8	1,5	**67,5**	**80**
C	7,75	1,75	8,25	1,75	**77,5**	**82,5**
D	7	8	5	4,5	**39,25**	**50**
E	10,5	4,5	8	2	**65**	**80**
F	17	8	21	4	**66,75**	**80,25**

Source: Author

The evaluation shows a significantly higher hit rate of all survey participants across all case studies. It should be mentioned that the second round of the survey took place immediately after the presentation of the results of the prediction models. Nevertheless, it turned out that, after a brief study of the prediction results, the controllers of the marina company took the prediction results into account.

Numerical Results

For model A, the hit rate increased from 62,5% to 75%. Difference: 12,5%.

For model B, the hit rate increased from 67,5% to 80%. Difference: 12,5%.

For model C, the hit rate increased from 77,5% to 82,5%. Difference: 5%.

For model D, the hit rate increased from 39,25% to 50%. Difference: 10,75%.

For model E, the hit rate increased from 65% to 80%. Difference: 15%.

For model F, the hit rate increased from 66,75% to 80,25%. Difference: 13,5%.

Conclusion

The average increase in the hit rate is about 11,54%. The quality of the decisions will increase regarding berth allocation to clients with longer contract time, better payment behavior and higher sales as well as a better target-group oriented marketing campaign and a better overview of the client structure.

6 STOCHASTIC DECISION MODELS BASED ON DATA SCIENCE FOR THE MARINA INDUSTRY IN CROATIA

Random variables influence a decision with a stochastic decision model based on data science. Random variables relate to the properties of the applicants and customers. Since the probabilities for the occurrence of certain values of the random variables are known, decisions are made under risk, but not uncertainty. The probabilities are objective, given that they have been calculated using predictive models (BAMBERG12, page 67). In this research, the decision models are designed based on prediction results and the resulting prediction models.

Laux defines (LAUX18, page 54):

"Stochastic decision models capture multi-valued expectations about the characteristics of the decision-relevant data, whereby these characteristics are assigned probabilities of entry. Stochastic decision models, therefore, relate to risk situations".

This research follows this approach, but it goes a step further. The statistics this research is based on is pattern recognition in the data underlying the computation of the predictive models. The multivalued expectations result from the scatter of the observed characteristics. Customer properties are allocated as well as characteristics such as yacht length and age. The use of the prediction models leads to transferring of decisions under uncertainty to decisions under risk. The risk arises from the probability of occurrence of the predicted events. However, this risk is minimized by using pattern recognition based on the support vector machine and the associated structured risk minimization according to Vapnik and Chervonenkis (VAPNIK18). Very good values for prediction power and prediction confidence calculated in the prediction models reinforce this assumption.

The table below shows the relationship between the decision model and the forecast models used.

Table 12: Prediction Models Used for Decision Models

Decision Model	Used Prediction Model
Sub-Decision Model A – Private Client with a Yearly Contract	Prediction A – Client Type PRIVATE & Yearly Contract
Sub-Decision Model B – Business Customer with a Yearly Contract	Prediction B – Client Type FIRM & Yearly Contract
Sub-Decision Model C – Private Client with Contract Renewal	Prediction C – Client Type PRIVATE & Renewal Contract

Decision Model ABC – Marina Industry – Berth Allocation	Based on sub-decision models A, B, and C and therefore its prediction models.
Decision Model D – Marina Industry – Payment Behavior	Prediction D – Correlations in the Event of Late Payment
Decision Model E – Vessel Length and Contract Amount	Prediction E – Correlation of Contract Amount and Vessel Length
Decision Model F – Customer Classification	Prediction F – Correlation of Sales and Vessel Type

Source: Author

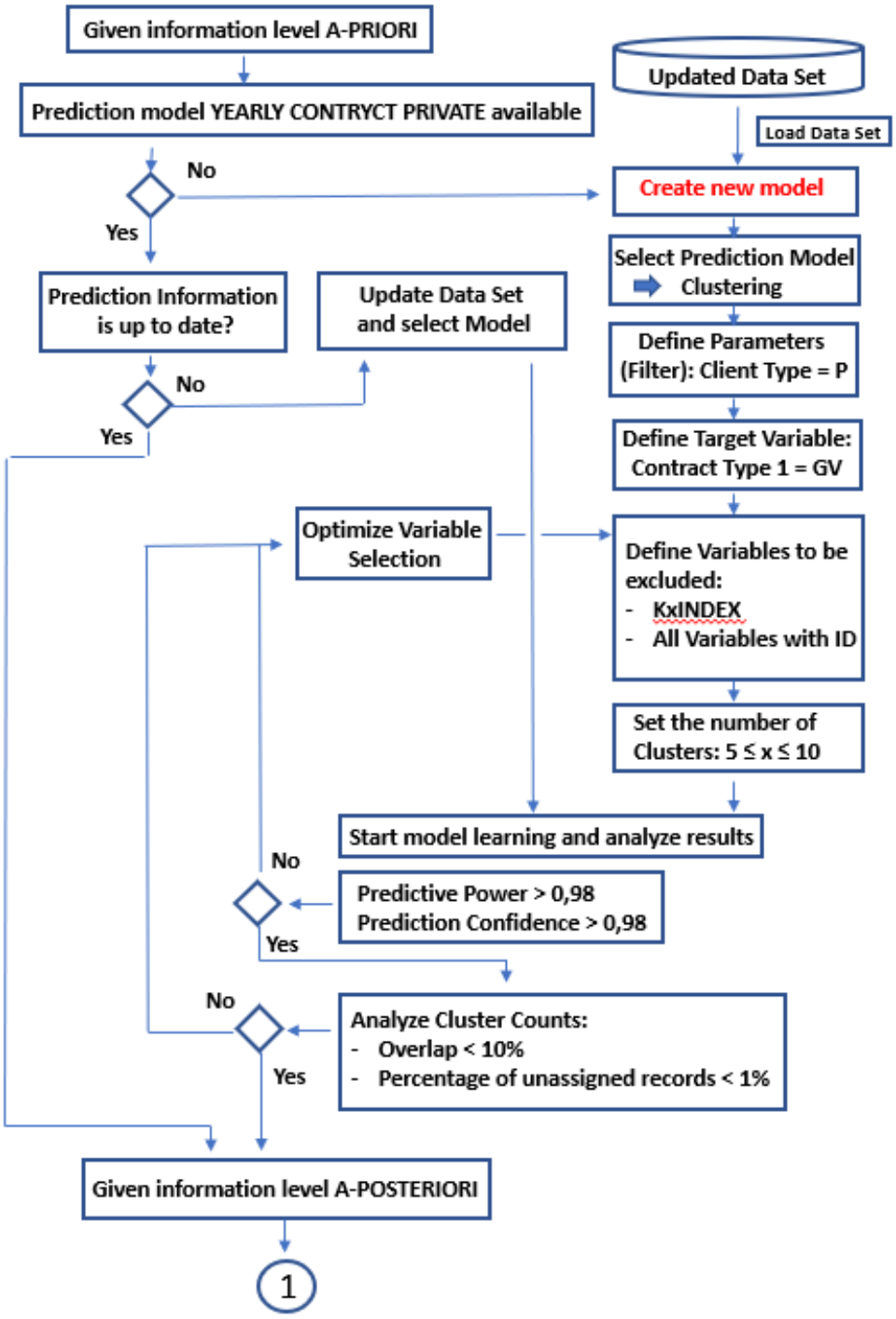
The decision scenario is allocation of a vacant berth. The decision objective is a long-time customer relationship with yearly contracts and contract renewal. Long-term customer relationships guarantee good occupancy of the berths and result in good sales.

a) Sub-Decision Model A – Private Client with a Yearly Contract

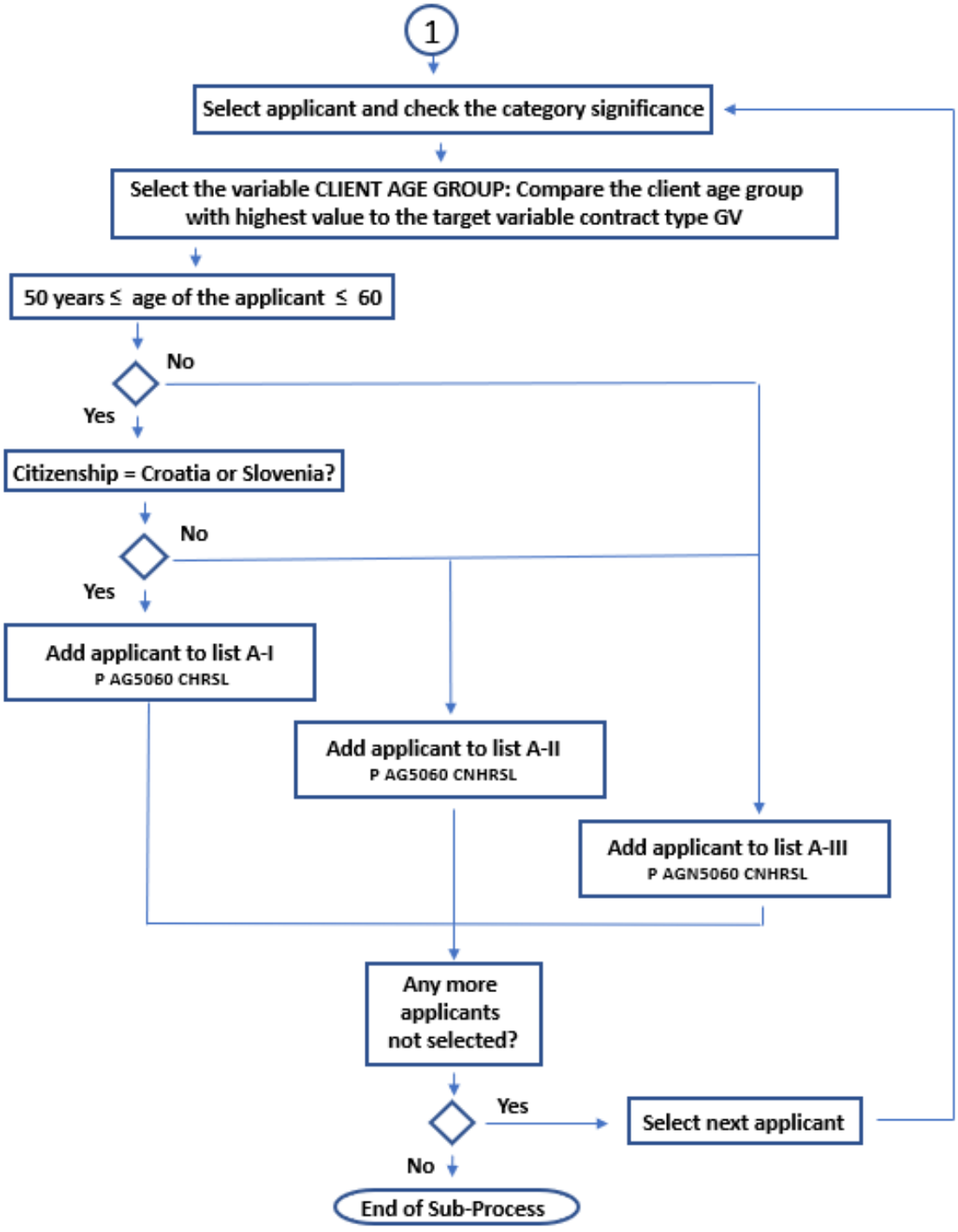
The objective of this decision-making process is to identify the berth applicants who are most likely to be inclined to conclude a yearly contract. For the evaluation of the applicant characteristics, a prediction model is generated based on the statistical method of clustering. The model calculates the relationship between the explanatory variables (applicant characteristics) and the target variable "yearly contract". All identity numbers such as applicant number, birthday, etc. are excluded from the selection because such variables are too widely spread and therefore no evaluable relationship can be calculated. If there is an updated forecast model, it could be used.

The explanatory variables, as calculated in the forecast models, are client type, age group, and citizenship. Only private applicants are selected from the existing data records. If an applicant optimally fulfills the other two criteria, this application is assigned to the Category A list. If an applicant belongs to the optimal age group but does not have optimal citizenship, he is assigned to group II. Applicants who only belong to the group of private clients are assigned to group III. The composition of groups I, II, and III defines a priority for this decision-making process. The following figure shows the process and the decision nodes for this decision process.

Figure 67: Decision Model: PRIVATE CLIENT – YEARLY CONTRACT



Source: Author



Explanation of Abbreviations

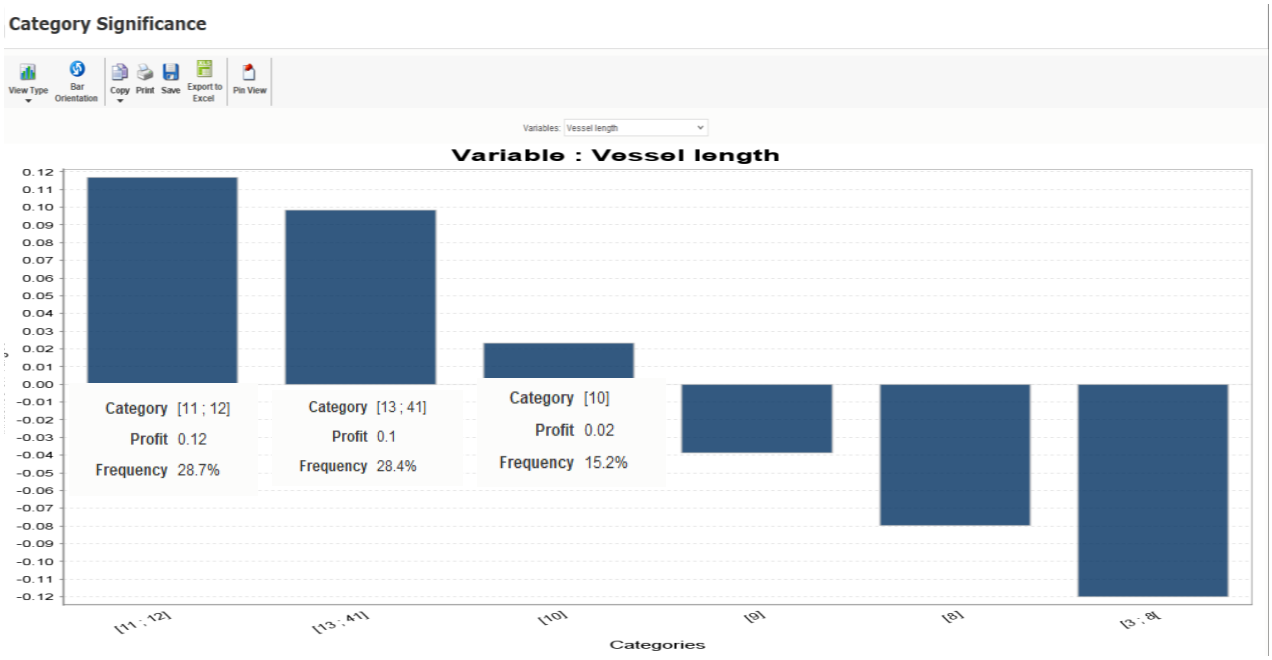
P	PRIVATE CLIENT
AG5060	AGE GROUP BETWEEN 50 – 60 YEARS
AGN5060	AGE GROUP NOT BETWEEN 50 – 60 YEARS
CHRSL	CITIZENSHIP CROATIA OR SLOVENIA
CNHRSL	CITIZENSHIP IS NOT CROATIA OR SLOVENIA

Source: Author

b) Sub-Decision Model B – Business Customer with a Yearly Contract

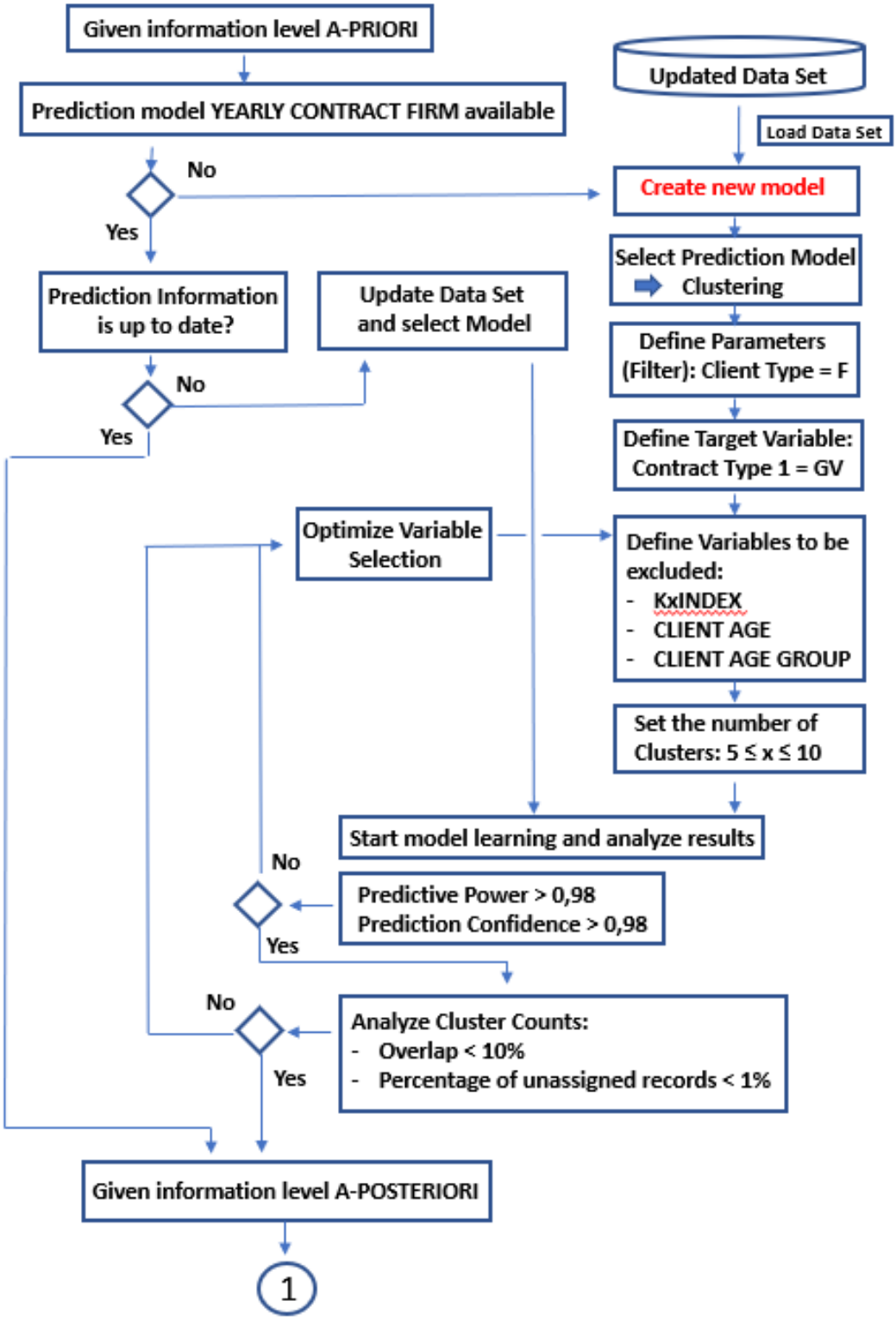
A forecast model with the target variable ‘yearly contract’ is required for this decision model. Only the client type “firm” is selected. The explanation variables ‘client age’ and ‘client age group’ are excluded, because such characteristics are not used for client type ‘FIRM’. The statistical method used is clustering. The number of clusters to be generated should be between 5 and 10 to increase the informative value of the clusters. Too many clusters restrict the options in the selection process for berth applicants. If the number of clusters is too small, selectivity is too low. In this evaluation scenario, there is a risk that clusters generated overlap too much. The overlap should not exceed 10%. The system SAP PREDICTIVE ANALYTICS is searching for the optimized number of clusters with minimized overlapping. If the overlap factor is a maximum of 10%, selectivity is acceptable. The prediction model detected that vessel length is the best explanatory variable for business clients intending to conclude a yearly contract. Applicants with vessel length between 11 and 41 meters should be preferred. The second choice is firms whose vessels are about 10 meters in length. Two lists are therefore created. Applicants whose vessel is between 11 and 41 meters in length have priority.

Figure 68: Explanatory variable ‘vessel length’ to target variable ‘YEARLY CONTRACT’

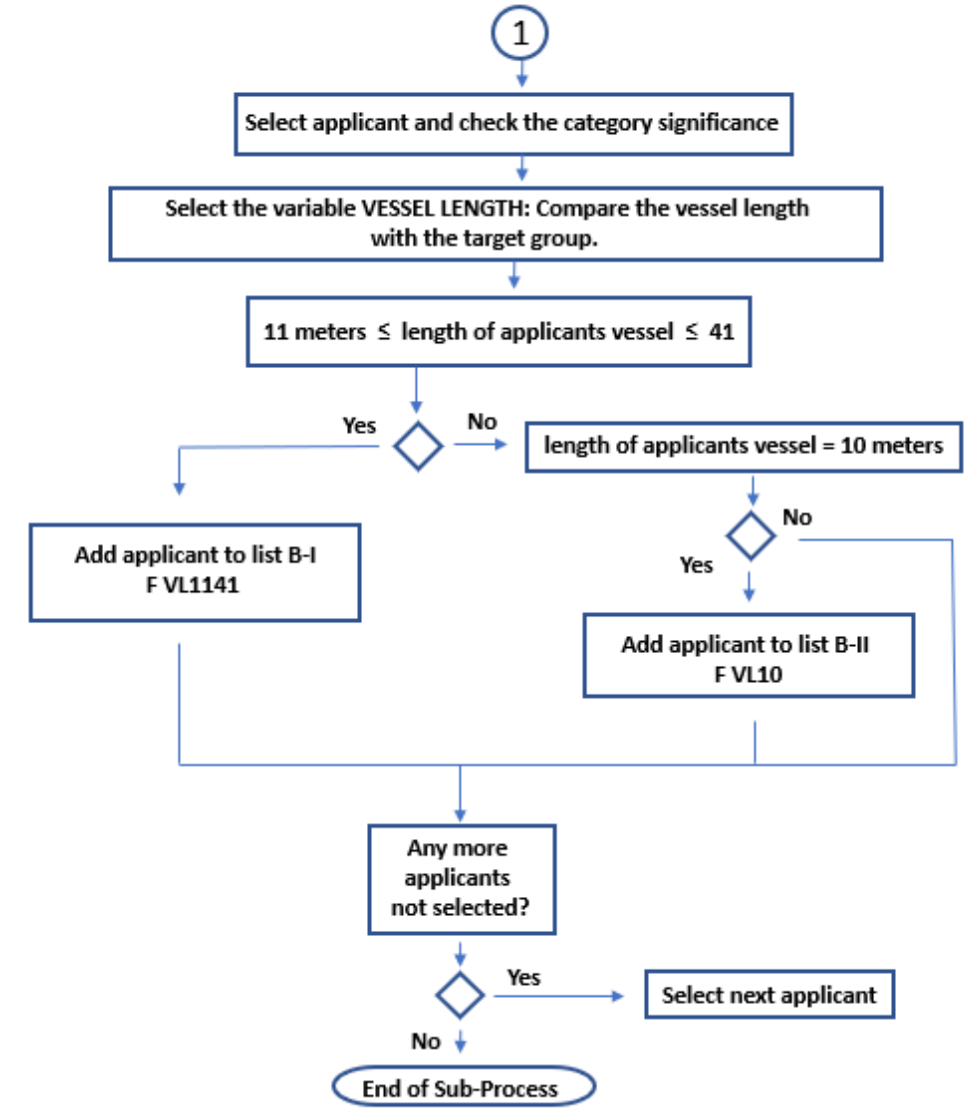


Source: Screenshot of SAP Predictive Analytics based on data created by Lebefromm, U.

Figure 69: Decision model: Client Type FIRM and YEARLY CONTRACT



Source: Author



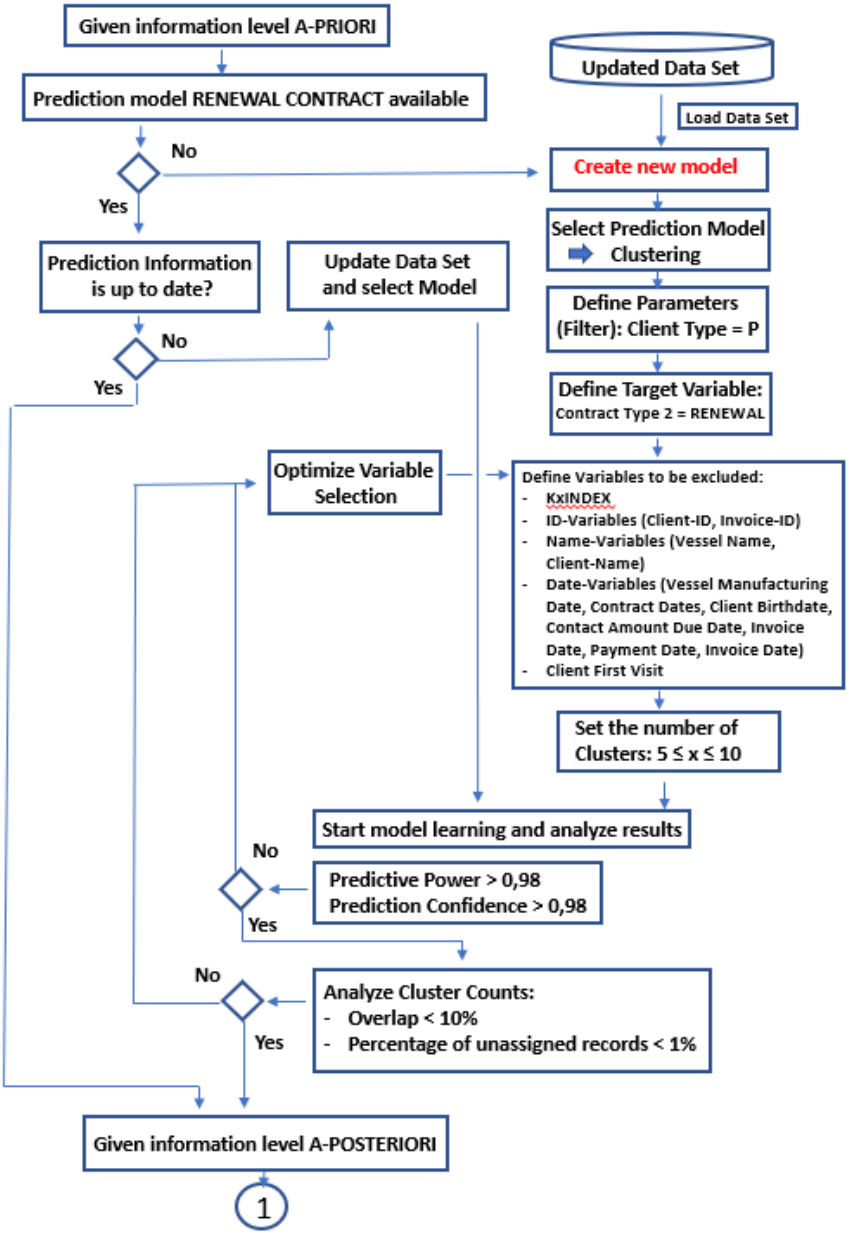
Source: Author

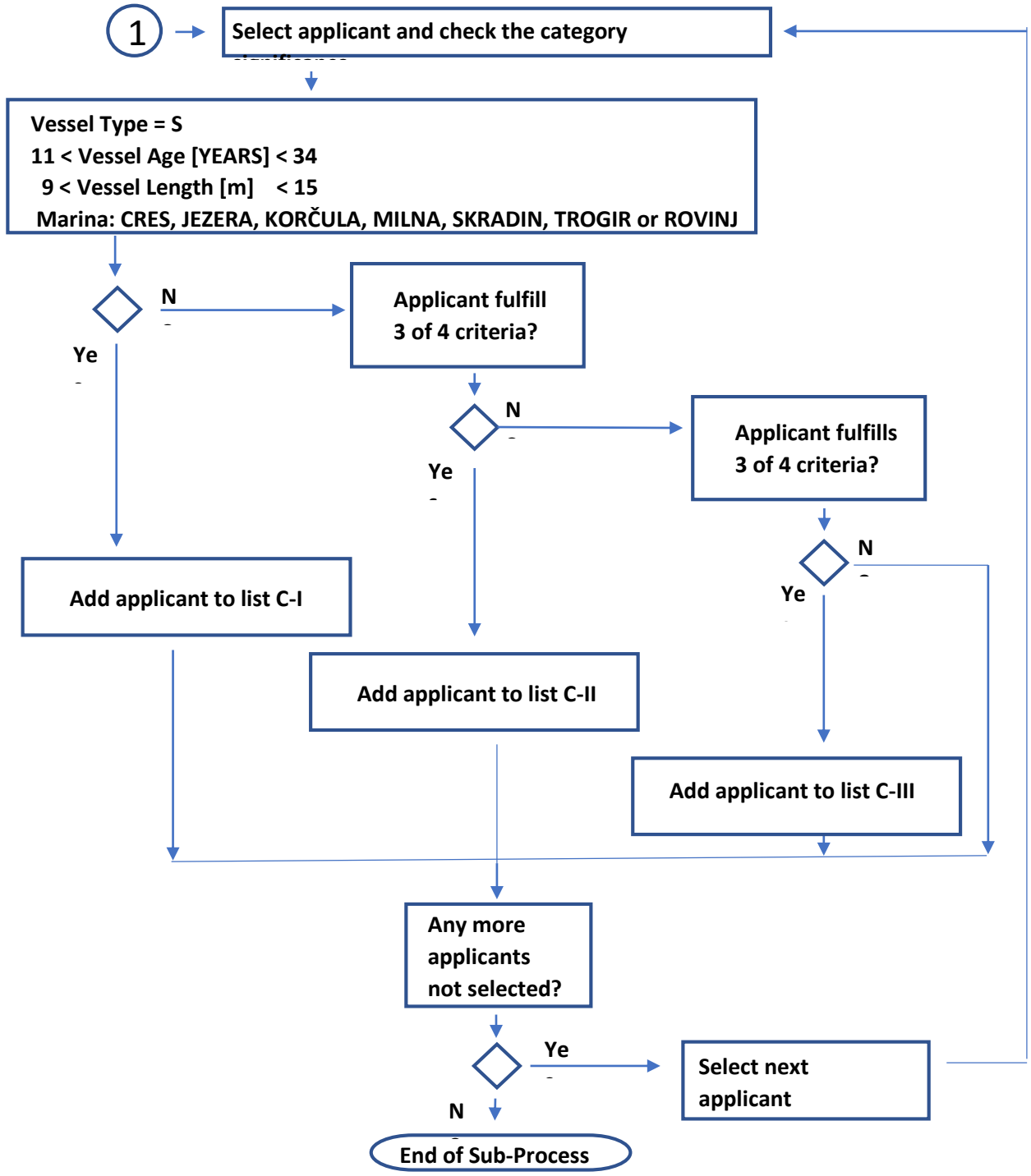
c) Sub-Decision Model C – Private Client with Contract Renewal

Customers who extend their contracts can become long-term customers. It is therefore interesting for the marina enterprise to recognize among the berth applicants those from whom an extension of the concluded contract can be expected by given probability. In this research, private clients have been selected and explanatory variables detected. A prediction revealed that

the age group in the range from 47 to 62 years has a positive correlation on the target variable contract renewal. Vessel type provides better information. Wind and wave conditions at berths may ensure frequent contract extensions. In any case, sailing boats are the clear favorite for contract renewals. For a new model, all ID variables were excluded.

Figure 70: Decision Model: CONTRACT RENEWAL





Legend:
F – CLIENT TYPE FIRM P – CLIENT TYPE PRIVATE S – SAILING BOAT
AG4756 – AGE GROUP BETWEEN 47 – 56 YEARS
AG5662 – AGE GROUP BETWEEN 56 – 62 YEARS
NOTE: The basic data has age groups – younger 50 years and younger 60 years. Therefore, an applicant of about 56 years of age can be seen as a member of age group AG5662.

Source: Author

With sufficiently good KI and KR and an acceptable overlap, the model can be used to decide which group should the applicants be assigned to. As the decision model shows, they can be assigned to the groups CI, CII and CIII. Group CI has the highest priority, Group CIII the lowest.

d) Decision Model ABC – Marina Industry – Berth Allocation

The decision models berth allocation, which refers to yearly contracts with private customers, yearly contracts with firms, and contract renewals with private customers are integrated into an overall decision model. This was done using a data matrix in which customers were assigned to three groups: PREMIUM, BUSINESS, and ECONOMY, based on their priorities. The priority level of each sub-model is assigned to a customer group. Ultimately, the customer group PREMIUM has the highest priority in all sub-models. Business group has the second-highest priority, and economy group the lowest priority.

Figure 71: Data Matrix for the Decision Model BERTH ALLOCATION

Decision Model Marina Industry - Decision Scenario: Allocation of Berth
Decision Objective: Long-Time Customer Relationship with Yearly Contract and Contract Renewal

Data Matrix

Decision - Subprocess	Applicant I Customer Class: Premium	Applicant II Customer Class : Business	Applicant III Customer Class: Economy
Decision Model Marina Industry Subprocess Model A Client-Type = P, Contract-Type = GV	Add applicant to list A-I P AG5060 CHRSL	Add applicant to list A-II P AG5060 CNHRSL	Add applicant to list A-III P AGN5060 CNHRSL
Decision Model Marina Industry Subprocess Model B Client-Type = F, Contract-Type = GV	Add applicant to list B-I F VL1141	Add applicant to list B-II F VL10	-
Decision Model Marina Industry Subprocess Model C Client-Type = P, Contract-Type = 2	Add applicant to list C-I P AG5662 S	Add applicant to list C-II P AG4756 S	Add applicant to list C-III P AGY47O56 S

Explanation of abbreviations

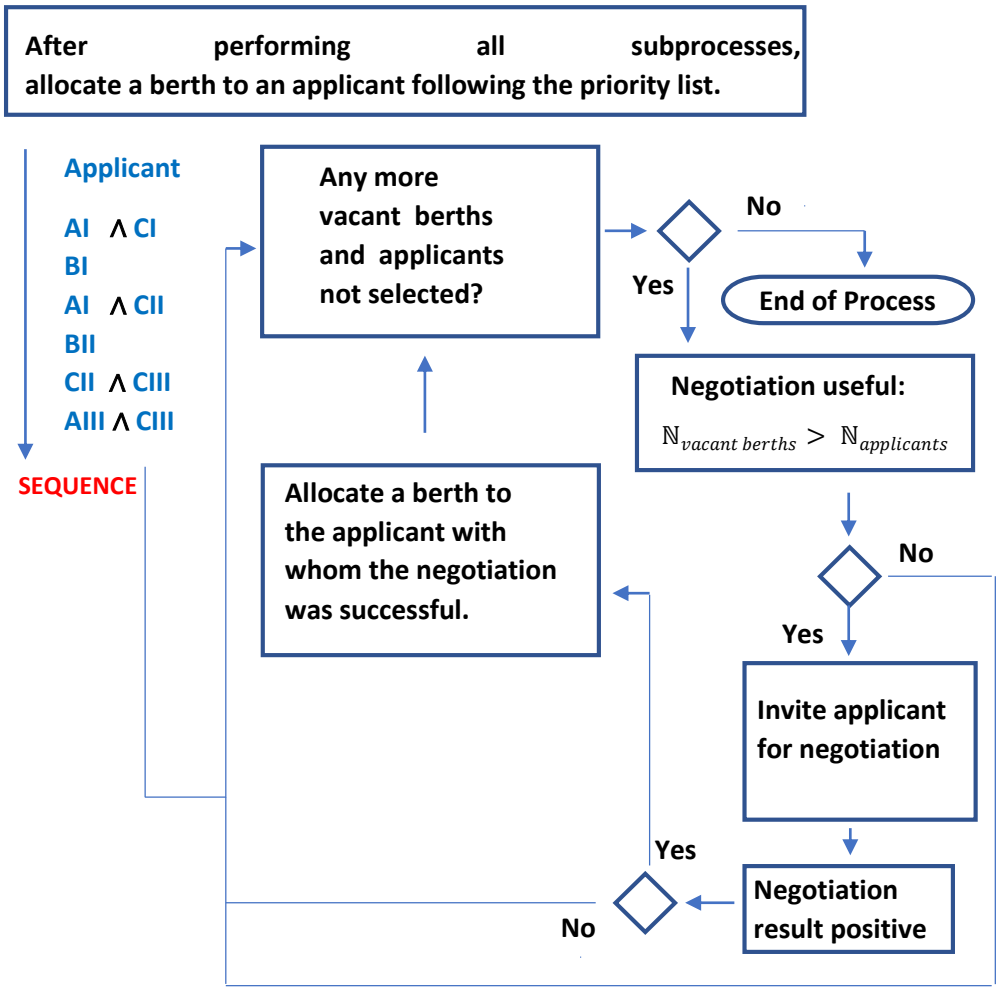
P	CLIENT TYPE: PRIVATE
AG5060	AGE GROUP BETWEEN 50 – 60 YEARS
AGN5060	AGE GROUP NOT BETWEEN 50 – 60 YEARS
CHRSL	CITIZENSHIP CROATIAN OR SLOVENIAN
CNHRSL	CITIZENSHIP NOT CROATIAN OR SLOVENIAN
F	CLIENT TYPE: FIRM
VL1141	VESSEL LENGTH BETWEEN 11 AND 41 METERS
V10	VESSEL LENGTH 10 METERS

NOTE: The basic data include age groups – younger than 50 years and younger than 60 years. Therefore, an applicant of about 56 years of age can be viewed as a member of age group AG5662.

Source: Author

The decision-making process follows a path that successively specifies deciding on an applicant using evaluation from the highest to the lowest priority.

Figure 72: Decision Path: BERTH ALLOCATION



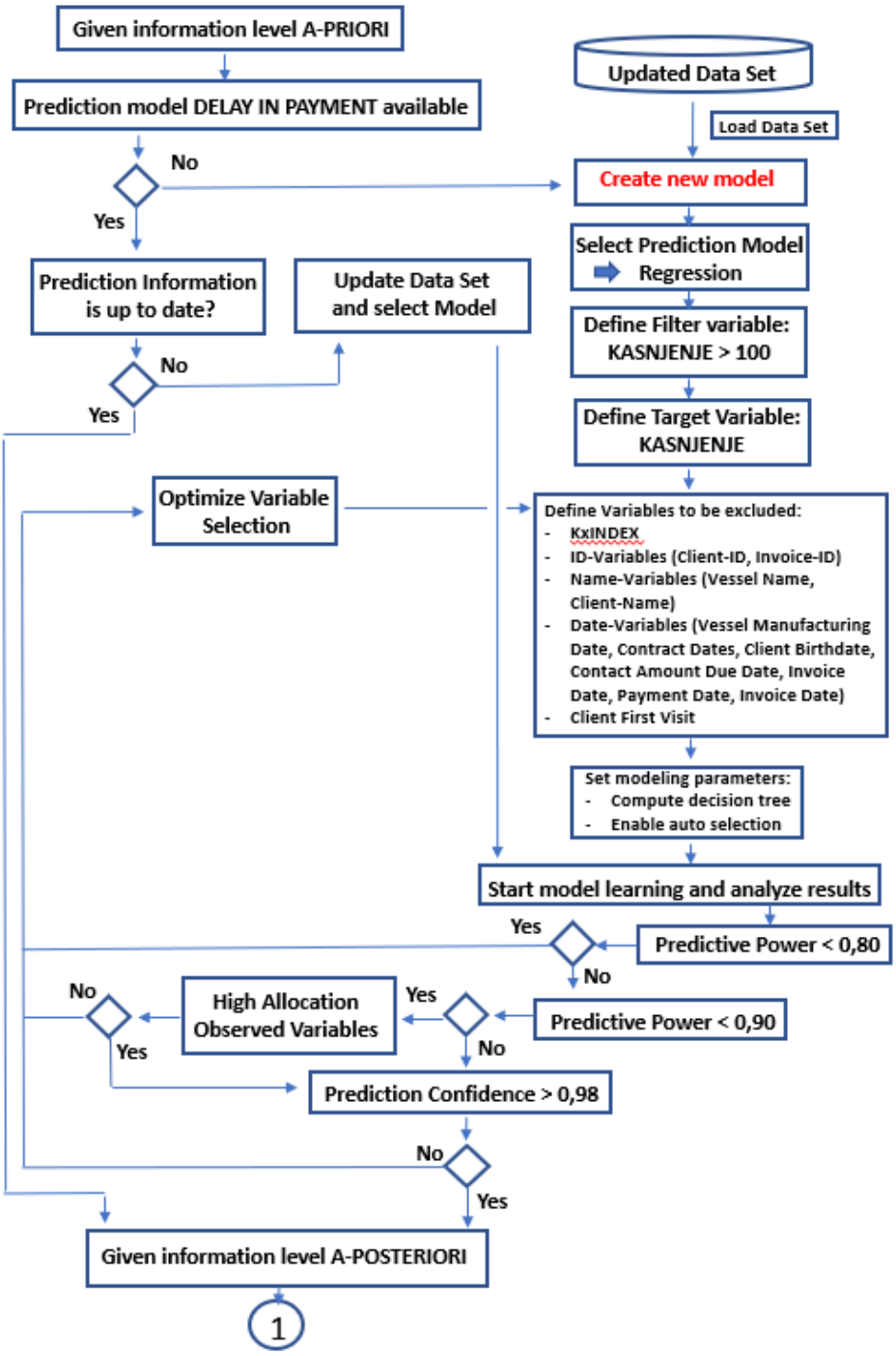
Source: Author

e) Decision Model D – Marina Industry – Payment Behavior

It is evident that the customers' payment behavior has a direct impact on the liquidity of the marina company.

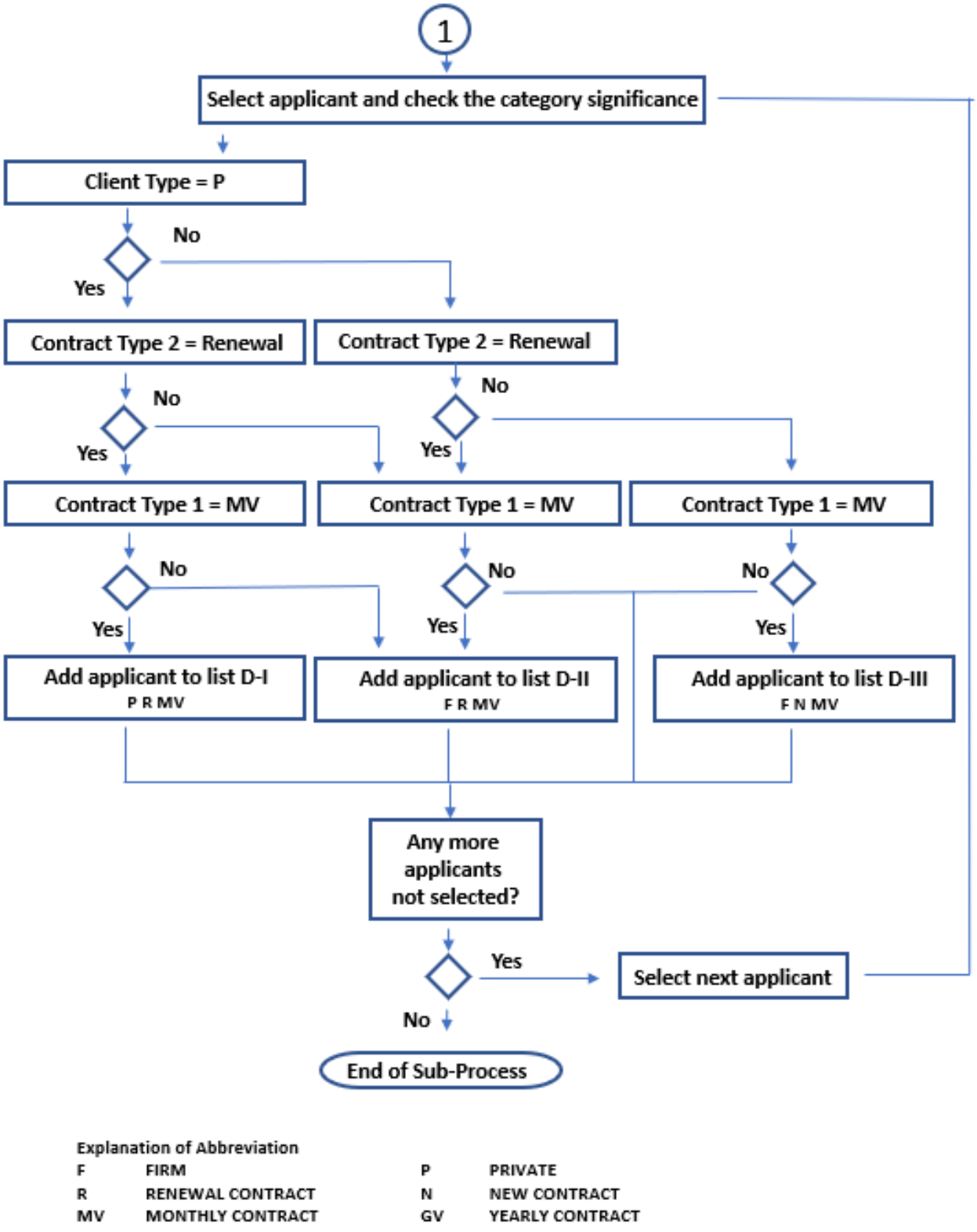
In a decision-making situation whose main goal is securing short-term liquidity, preference should be given to private applicants who wish to conclude a monthly contract or renew an existing contract. The prediction model developed for this decision situation uses regression analysis. The observed data are widespread, which leads to a lower predictive power KI with the value of 0.8091. According to the opinion of the developers of SAP Predictive Analytics found in the application documentation, the value is still acceptable if the trend is depicted. This is shown in the following figure.

Figure 73: Decision Model D – Payment Behavior



Legend: KAŠNJENJE – DELAY

Source: Author



Explanation of Abbreviation

F	FIRM	P	PRIVATE
R	RENEWAL CONTRACT	N	NEW CONTRACT
MV	MONTHLY CONTRACT	GV	YEARLY CONTRACT

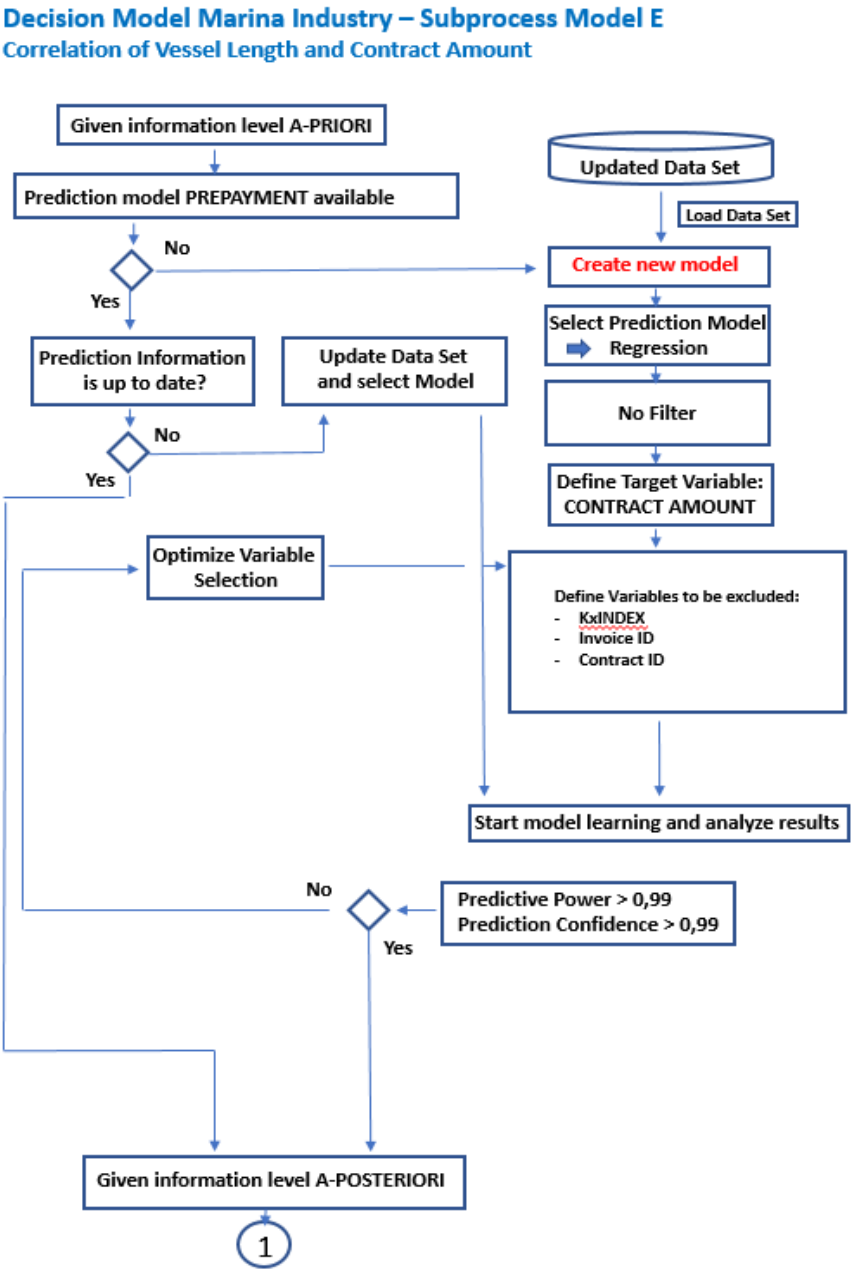
Source: Author

The clients are grouped into three groups. The first group, DI, has the highest priority because of the lowest probability for late payment, etc.

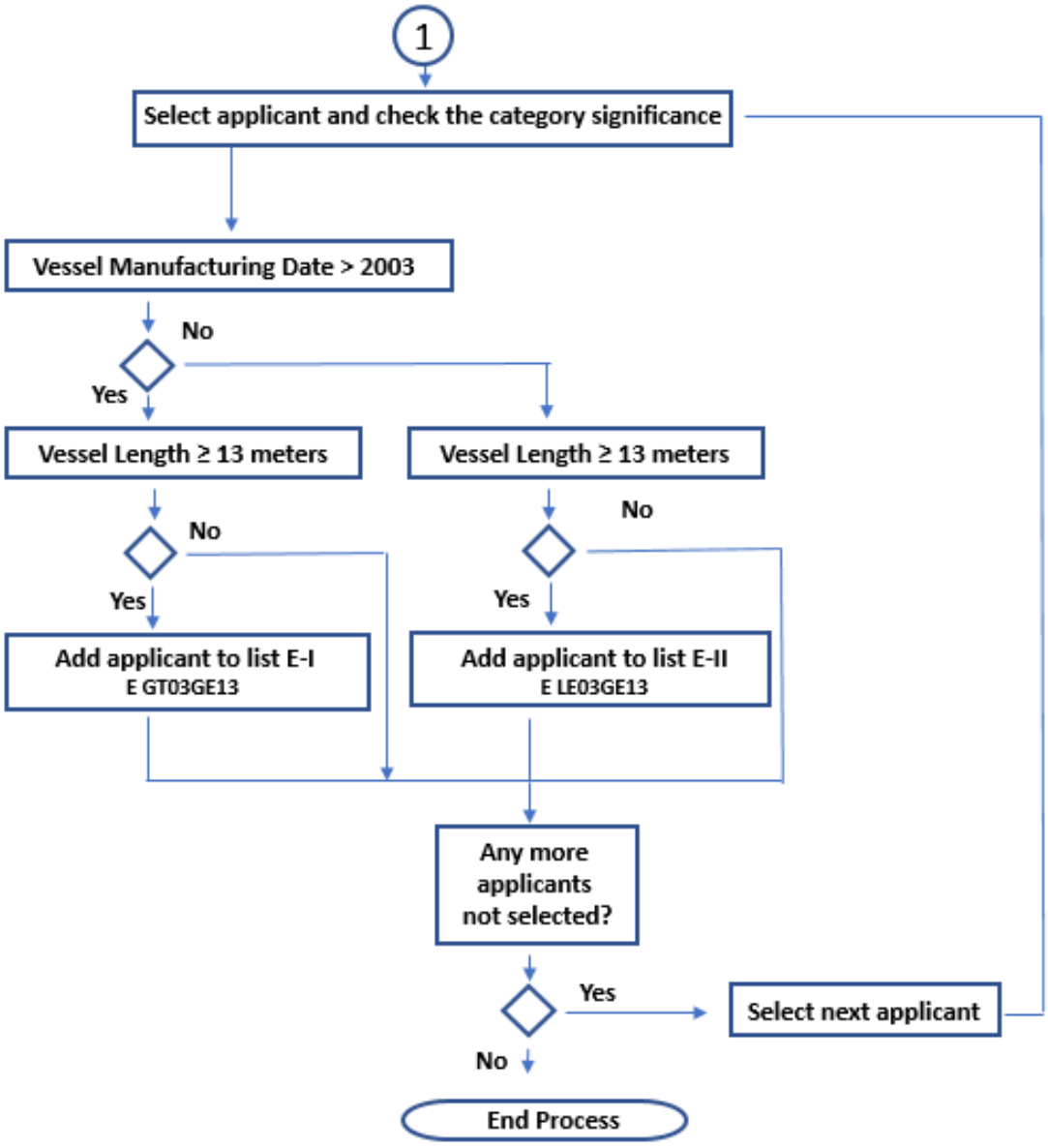
f) Decision Model E – Vessel Length and Contract Amount

It is no coincidence that berth applicants with a larger vessel also conclude higher-volume contracts. The reason might be that owners of larger vessels book shorter layover times. The same applies to owners of newer or older vessels.

Figure 74: Decision Model: Vessel Characteristics – CONTRACT AMOUNT



Source: Author



Explanation of Abbreviations

GT03	Greater Than 2003 (not old vessel)
GE13	Greater Equal 13 meters
LE03	Lower Equal 03 (older vessel)

Source: Author

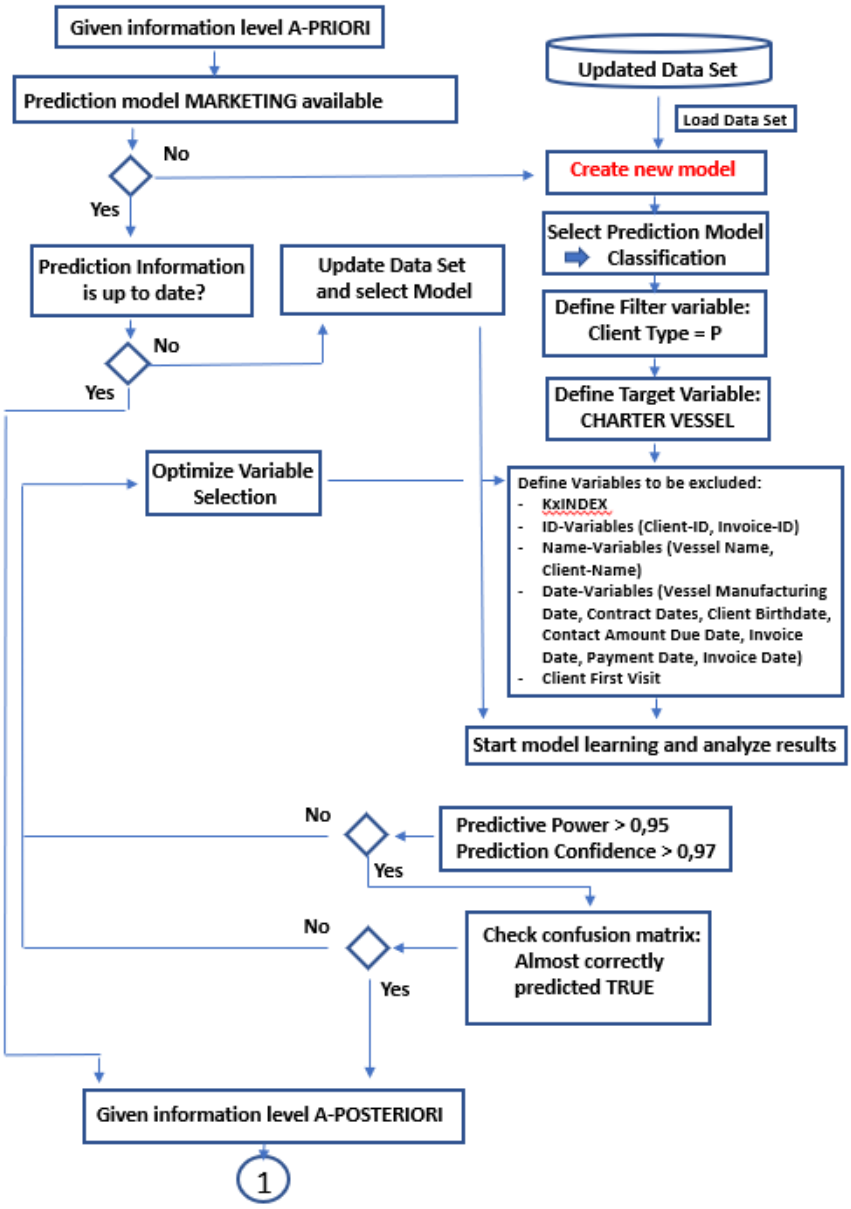
However, a regression analysis has shown that this is exactly the case: owners of newer and larger vessels conclude contracts with a higher order volume.

The decision model is based on the prediction results using the regression method. If the applicant has a vessel built after 2003, and if vessel length exceeds 13 meters, a higher sales volume could be expected. Priority II is assigned to applicants with a larger vessel, but of older manufacturing date.

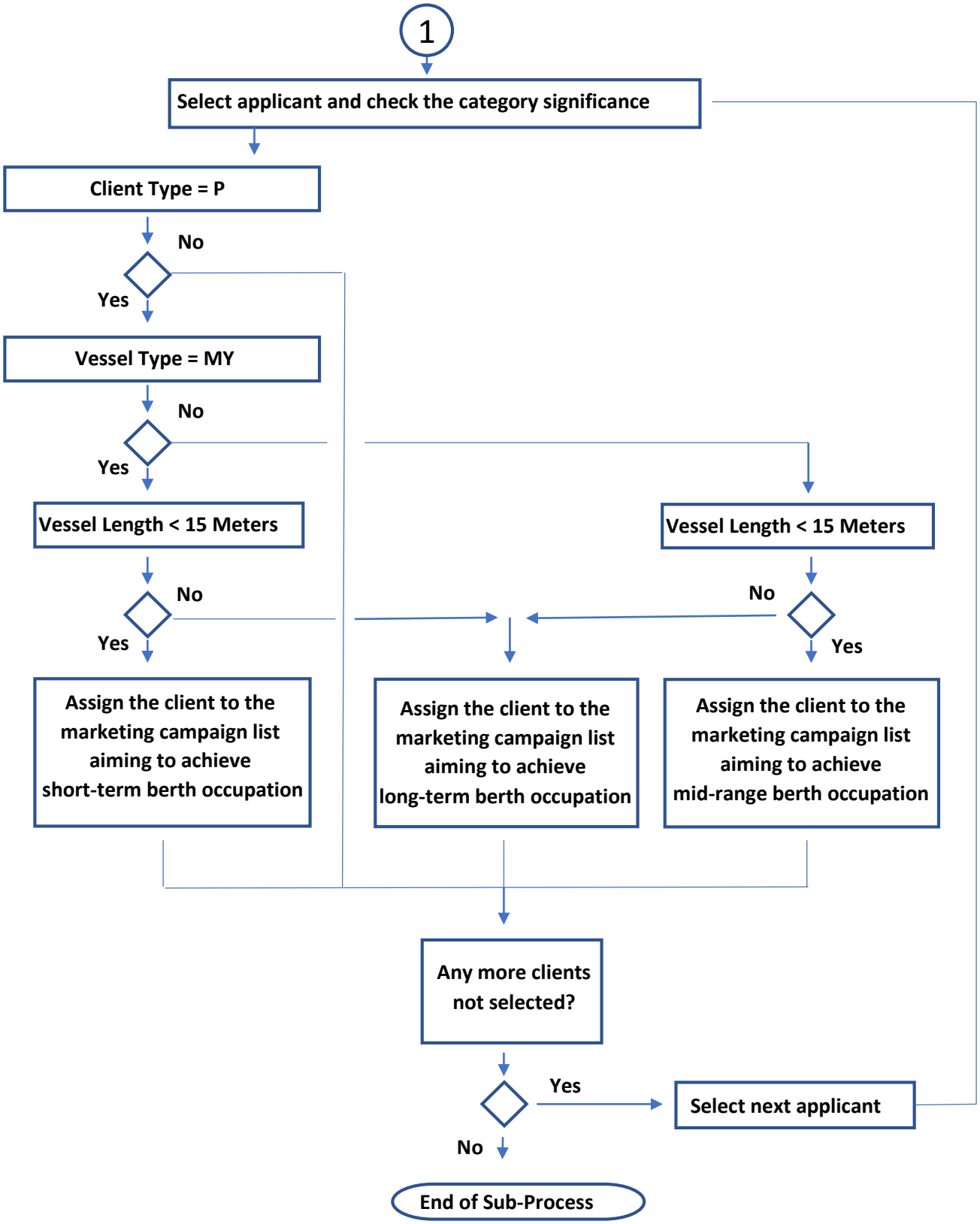
g) Decision Model F – Customer Classification

The concept and design of customer relationships and marketing are generally based on customer classification. An ABC categorization is common: A – customers with a long-term business relationship (more than 2 years) and high sales volume. B – customers with a medium-term business relationship (1 to 2 years) and medium sales volume. C – customers with a business relationship of up to 1 year. The forecast model based on the evaluation of the correlation between vessel type and sales volume has shown that customers with motorboats are assigned to the lower and upper segments. The largest share of sales in the middle segment is occupied by sailboats, only sporadically by catamarans. A very high turnover can be achieved with owners of luxury yachts.

Figure 75: Decision Model: CUSTOMER CLASSIFICATION



Source: Author



Source: Author

This results in the formation of the following customer groups for the decision model: owners of smaller motorboats (less than 15 meters) who are offered berths of a smaller size with shorter lay times. Owners of larger sailing boats and motor yachts (larger than 15 meters) who are offered larger berths with a long stay. Owners of smaller sailboats (less than 15 meters) who are offered medium-sized berths with a medium length of stay.

7 PROOF OF HYPOTHESIS AND CONTRIBUTION OF THE RESEARCH

7.1 Proof of Hypothesis

Main Hypothesis H

This thesis has demonstrated that reliable forecast models could be generated with the available data provided by the marina company. The revised data set fulfilled the requirement for high-quality data. The statistical report for each prediction model shows plausible values of the explanatory variables regarding the minimum value, maximum value, mean value, and standard deviation. Data anomalies were explained in the thesis; they result from the calculation of certain variables in the system of the marina company. It was made clear that prepayments by the customers of the marina company result in negative invoice amounts and a negative delay period. Furthermore, this also leads to negative claims and thus to liability of the marina company towards the customers to provide the services paid in advance. This peculiarity in the data was considered in the thesis in the way that the decision for or against an applicant would not be based on the specific extent of the correlation between explanatory variables and target variables, but only on whether a significantly positive or negative correlation was calculated. This was shown in the figures. The results of the prediction models were implemented as an integral part of the decision-making process. The grouping of the company's applicants and customers in different clusters based on the results of the prediction models leads to prioritization in the decision-making process. Applicants and customers whose characteristics have a higher correlation to the target variable are preferred. With the generation of the prediction models, decision-making processes could be developed in such a way that the objective can be clearly defined, the parameters for setting the prediction models are known and can thus be used as a decision criterion. All of this results in statistically reliable decision-making.

Hypothesis H 1

All prediction models were documented with a model overview. The respective figure for the model overview shows the calculation of the values for predictive power (KI) and predictive robustness (KR). High values could be demonstrated for all models via the correlation of the

most important explanatory variables to the target variables (KI). At the same time, there were also high values in the agreement of the predictive results and the known results in the test of the prediction models with validation data (KR). The hypothesis that such prediction results can also be generated with customer data from the marina industry in Croatia has been confirmed. The following table provides an overview of the calculated values of KI and KR for all six models.

Table 13: Overview of KI and KR for all prediction models

Model	KI	KR
A	0.9798	0.9910
B	0.9977	0.9993
C	0.9996	0.9999
D	0.9849	0.9960
E	0.9908	0.9985
F	0.9593	0.9896

Adequate values for KI and KR could be calculated in all prediction models. This means that the models can be viewed as meaningful and robust.

The significance of the prediction models was demonstrated with the so-called “ROC curve” (receiving operating characteristics). The figures presented in this thesis demonstrate a close relationship between the ROC curve of the prediction models and the ideal model (wizard) and with a large distance from the graph, which reflects random distribution.

Auxiliary Hypothesis H 1.1

A long-term customer relationship was defined in this research with the following target variables: contracts with a contract duration of at least one year and renewal of existing contracts. Since an explanatory variable such as age group does not play a role for corporate customers, a model for private customers and a model for corporate customers were explicitly generated with the target variable ‘yearly contract’. It could be proven that private customers in the age group 49–62, who made a prepayment of more than HRK 250.00, owning a vessel type catamaran or sailboat of vessel age between four and 20 years, intend to conclude a yearly contract. Private customers with these characteristics had the highest correlation with this type of contract. This was proven by prediction model A. Corporate customers wanting to rent a berth for sailboats, whose boats of more than 9 meters in length are more than 10 years old and

who renew an existing contract tend to have contracts with a term of one year. This was proven by the results of the prediction model B. In both customer groups, those whose intention is to renew an existing contract are the ones using a sailboat between 9 meters and 15 meters in length, of vessel age between 11 and 34 years and wanting to rent a berth at company marinas such as Cres, Jezera, Korčula, Milna, Skradin, Trogir, and Rovinj. This was proven by prediction model C.

Auxiliary Hypothesis H 1.2

With hypothesis H 1.2 it was asserted that late payers have similar characteristics. These similar properties could be identified with the prediction model D. A significant correlation with the following client characteristics could be predicted for the target variable ‘days payable outstanding’: Customer type ‘firm’, conclusion of new contracts, 8-year-old or older vessels, and client headquarters in Croatia, Slovenia, Italy, or Russia. In order to conclude contracts with clients from whom timely payments can be expected, the marina company should give preference to clients with opposite characteristics. However, implementation of the prediction model findings is not that easy. However, the prediction model results can support pricing; it should be adapted accordingly.

Auxiliary Hypothesis H 1.3

The hypothesis asserts that predictive analytics can make a valuable contribution to sales planning by precisely identifying customer groups that make a significant contribution to the sales volume. In prediction model E, a positive correlation with a high order volume and order type RENEWAL could be identified. Naturally, contracts with owners of larger boats have a higher contract volume. A positive correlation could also be identified with vessel age. Older boat owners are more likely to sign smaller contracts. Owners of one-to-eighteen-year-old boats represent more than a third of the sales volume. It was also concluded that the contracts of owners of motor yachts and catamarans have a positive correlation with higher order volumes. The characteristics of customers of whom a higher order volume can be expected could thus be used to better address such customers.

Auxiliary Hypothesis H 1.4

The hypothesis was set up that customer-related characteristics can be identified in the customer data set for private customers, which allows a meaningful assignment of customers into customer classes. It is useful in the sense that the marina company sales policy can be designed accordingly. With the prediction model F, a correlation between vessel type and turnover groups could be demonstrated. Four customer classes were formed within which a certain vessel type has priority. In the area of total turnover of HRK 27,452, larger turnover is achieved with the owners of motorboats. On the other hand, in the sales range from HRK 27,452 to 55,501.30, more sales are achieved with the owners of sailboats and catamarans with a rising trend in the share of sailboats and catamarans, in accordance with increasing sales. In the next sales class, from HRK 55,501.30 to HRK 107,350.00, the share of sailboats and catamarans is still higher than the share of motor yachts but has been continuously declining. The customer class with the highest turnover of more than HRK 107,350.00 prefers motor yachts.

Hypothesis H 2

According to Bayes' theorem about a-posteriori decisions, better decisions are made with consideration of prediction results. The hypothesis that better decisions are made with knowledge of the results of the prediction models (a-posteriori) was proven by measuring the hit rates of the analyzed candidates and customers according to certain characteristics with and without knowledge of the predictions. As shown in the previous corresponding chapter, the average hit rate increased by **11.54 percent**.

7.2 Contribution of the Research

The contribution of this research has a scientific and applicable aspect. Scientific contribution: this research makes the following contribution to methodological decision-making at the level of microeconomics:

A. This research confirms the usefulness of sophisticated computers tools (machine learning, predictive analytics) that are to be used for making accurate and reliable predictions in the case of a specific industry.

B. This research proves that, by using company-specific operative variables for business prediction, reliable predictions can be calculated about decision-relevant events in the future for controlling and management.

C. This research contributes to the theory of transaction costs in two ways. Firstly, the decision-making process follows the logic of deciding on the statistically best alternative. Therefore, the process can be automated, which makes it faster and more cost-efficient. Secondly, the use of statistically secured predictions helps increase the decision-makers' naturally limited perceptual capacity. The cost of obtaining information is reduced because, in this research, learning based on personal experience is replaced by machine learning.

D. This research explores the trend towards modernization of controlling and management in the age of digital transformation. The use of the data science method for pattern recognition leads to better identification of the company's target groups and thus to rational target group-related decisions.

E. Multi-level stochastic decision models were developed in this research, which makes the decision-making process more transparent and thus also comprehensible. The decision options developed in the decision models increase the flexibility and agility required by controlling in the age of digital transformation.

Applicable Contribution

F. Theoretically based and empirically proven developed stochastic prediction model supported by programmed stochastic methods provides a contribution that closes the gap in the non-existent stochastic decision-making models in the marina industry. The methodology created in this research is cross-industry.

G. The use of pattern recognition in the statistical data increases the predictability of future revenues and thus the planning of investments and finances.

8 CONCLUSION AND OUTLOOK

This dissertation deals with the effects of digital transformation on corporate management and the associated business processes of corporate planning, forecasting and the resulting decision-making processes. The main goal was to develop a model for implementing digital transformation in operational and tactical decision-making processes in controlling and management. Using quantitative and qualitative research methods it has been proven that the possibilities of data science resulting from digital transformation do not only significantly improve the level of information of controlling and management but also lead to a changed decision-making behavior. Quantitative research is based on the statistical learning and a-posteriori decision theory, and the mathematical method for the implementation of the statistical learning theory in machine learning takes place via the support vector machine. During initial analysis of the current information technology, it became clear that the machine learning methods and the prediction models developed on this basis result in a better supply of information for decision-makers. The mathematical-statistical methods which are the basis of the prediction models in this research are clustering, regression, and classification. Qualitative research is based on the case study method i.e., the business case used is the decision making behavior of the largest marina company in Croatia that offered its cooperation. With the conclusion of confidentiality agreements, the marina company made extensive customer data available with a sufficiently large database for the development of prediction models. To find out which exemplary business processes and related decisions should be used, interviews were conducted with the company's controlling staff and the department head responsible for information technology. It turned out that an essential part of the daily decision-making process is the allocation of berths. Although applicants and customers of the company can book berths online, this is a purely operational business process. However, it is possible to develop the transformation of such a booking system into an expert system based on machine learning. The object of the thesis was defined as follows: business processes in sales and customer relationship management using machine learning as decision support. The next step was to find out the characteristics of the applicants and customers of the marina company of which a long-term and successful business relationship can be expected. This was implemented in the thesis using the method of creating prediction models. Regarding the first research question, it could be proven that the prediction models extracted from all available explanatory variables are the ones that have a high significance for the decision criterion. The trustworthiness of the prediction models is proven by the key performance indicators prediction power and prediction

robustness. This also answered the second research question. The following question was whether the decision-makers would use the results of the prediction models and thus achieve better decision-making results. This hypothesis could be proven with a two-stage hypothesis-testing interview series. This provided a secure basis for the development of decision models in which the results of prediction models provided the decision criterion. An important contribution to the state-of-the-art in decision theory could thus be made. This thesis provides a decision model that uses the possibilities of machine learning and the associated method of pattern recognition.

The limits of this thesis are within the scope of the available environmental scenario. The decisions of the controlling department of the marina company do not necessarily have to represent the entire marina industry in Croatia or even other branches and countries. Another limit is in the qualitative examination of expert interviews. While in the case of extensive quantitative studies it is possible to draw conclusions about the population from the sample, this is not allowed with the comparatively small number of cases examined in this thesis. Nevertheless, in terms of an outlook, the results of this thesis contribute to the further development of decision theory under the influence of digital transformation. The decision models created are programmable and can therefore be implemented in an expert system. The methodology of this thesis can be used as a scientific approach to developing stochastic decision models based on data science for other branches and in other countries.

Scientific Apparatus

List of Figures

Figure 1: Thesis Design 9
Figure 2: The Role of Controlling 22
Figure 3: Business Model Navigator - Assessment 27
Figure 4: Paradigms on the Internet 32
Figure 5: Column-oriented database 34
Figure 6: Architecture SAP S/4HANA 36
Figure 7: Random Model and Forecast Model 39
Figure 8: Prediction Confidence and Prediction Power 41
Figure 9: Rectangle Method – Outer Rectangles 42
Figure 10: Rectangle Method – Inner Rectangles 42
Figure 11: Normal Vector 49
Figure 12: Hyperplane 50
Figure 13: The Optimizing Problem in SVM 51
Figure 14: Linear Separability with Accepted Errors 52
Figure 15: Shattering in the VC-Dimension 54
Figure 16: The Learning Process in SAP Predictive Analytics 75
Figure 17: Best Model According to the SRM Theory 76
Figure 18: Statistical Report, Model A 79
Figure 19: Store Payment in Advance 80
Figure 20: Model Overview: Model A 81
Figure 21: Model A – Influence of the explanatory variable 'VESSEL TYPE' 82
Figure 22: Model A – Influence of the Explanatory Variable: Vessel Age 83
Figure 23: Model A – Influence of the Explanatory Variable 'Client Age' 83
Figure 24: Model A – Influence of the Explanatory Variable 'Advance Payment' 84
Figure 25: ROC Curve, Model A 85
Figure 26: Explanatory variables with a positive influence on the target variable 'Yearly Contract' – Client Type: Private 86
Figure 27: Statistical Report, Model B 87
Figure 28: Overview, Model B 88
Figure 29: Model B – Influence of Vessel Type on Yearly Contract – Client Type: FIRM ... 89

Figure 30: Model B – Influence of Vessel Age on Yearly Contract – Client Type: FIRM 89
Figure 31: Model B – Influence of Vessel Length on Yearly Contract – Client Type: FIRM 90
Figure 32: Model B – Influence of Contract Type NEW or RENEWAL to Yearly Contract – Client Type: FIRM 90
Figure 33: ROC-Curve, Model B 91
Figure 34: Explanatory Variables with Positive Influence on Target Variable 'Yearly Contract' – Client Type: FIRM 92
Figure 35: Model C – Statistical Report and Data Size 93
Figure 36: Model C – Overview 94
Figure 37: Model C – Influence of Vessel Type on Contract Renewal 95
Figure 38: Influence of Vessel Length on Contract Renewal 96
Figure 39: Influence of Vessel Age on Contract Renewal 96
Figure 40: Model C – Influence of Marinas on Contract Renewal 97
Figure 41: Model C – ROC Curve 97
Figure 42: Explanatory Variables with Positive Influence on Contract Renewal 98
Figure 43: Model D – Statistical Report and Data Size 102
Figure 44: Model D – Overview 103
Figure 45: Model D – Influence of Variable 'Client Type' on Payment Behavior 104
Figure 46: Model D – Influence of Citizenship on Payment Behavior 104
Figure 47: Model D – Influence of Contract Type on Payment Behavior 105
Figure 48: Model D – Influence of Variable 'Vessel Age' on Payment Behavior 106
Figure 49: Model D – Performance 107
Figure 50: Explanatory Variables with Positive Influence on Late Payment 108
Figure 51: Model E – Descriptive Statistical Report 109
Figure 52: Model E – Overview 110
Figure 53: Model E – Influence of Contract Type 2 on Contract Amount 111
Figure 54: Model E – Influence of Vessel Length on Contract Amount 111
Figure 55: Model E – Influence of Vessel Age on Contract Amount 112
Figure 56: Model E: Influence of Vessel Type on Contract Amount 113
Figure 57: Model E – Performance 114
Figure 58: Model E – Summary of the Influence of Explanatory Variables 114
Figure 59: Model F – Statistical Descriptive Report 116
Figure 60: Model Overview for the Classification 'Charter Vessel' 117
Figure 61: Confusion Matrix Classification 'Vessel Type' 117

Figure 62: Variable Contributions to Classification ‘Vessel Type’ 119
Figure 63: Category ‘Significance of Vessel Type for Contract Amount’ 119
Figure 64: Decision Tree Classification ‘Vessel Type’ 121
Figure 65: Customer Groups Based on Sales Volume 123
Figure 66: Decision Matrix 125
Figure 67: Decision Model: PRIVATE CLIENT – YEARLY CONTRACT 141
Figure 68: Explanatory variable ‘vessel length’ to target variable ‘YEARLY CONTRACT’ 143
Figure 69: Decision model: Client Type FIRM and YEARLY CONTRACT 144
Figure 70: Decision Model: CONTRACT RENEWAL 146
Figure 71: Data Matrix for the Decision Model BERTH ALLOCATION 148
Figure 72: Decision Path: BERTH ALLOCATION 149
Figure 73: Decision Model D – Payment Behavior 151
Figure 74: Decision Model: Vessel Characteristics – CONTRACT AMOUNT 153
Figure 75: Decision Model: CUSTOMER CLASSIFICATION 156

List of Tables

Table 1: Analyzing the ACI Data Set 68
Table 2: Confusion Matrix 118
Table 3: Development of the Correlation Between Contract Amount and Vessel Type 122
Table 4: Customer Sales 128
Table 5: Solution Case Study Model A 129
Table 6: Solution Case Study Model B 130
Table 7: Solution Case Study Model C 131
Table 8: Solution Case Study Model D 132
Table 9: Solution Case Study Model E 133
Table 10: Customer Segments 135
Table 11: Hit Rate of Participants in the Case Study 136
Table 12: Prediction Models Used for Decision Models 139
Table 13: Overview of KI and KR for all prediction models 160

List of Abbreviations

ACDOCA	Accounting Documents Actual
ACI	Adriatic Croatia International Club
AH	Auxiliary Hypothesis
AI	Artificial Intelligence
BKPF	*Beleg-Kopf* (header of a line item)
BSEG	*Beleg-Segment* (posting line of a line-item)
CDS	Core Data Structure (program to select data for application)
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CM	Contribution Margin
CO	Controlling
COBK	Controlling *Beleg Kopf* (table in the SAP system with head of line items)
COEP	Controlling *Einzelposten* (table in the SAP system with CO line-item actuals)
COGM	Cost of Goods Manufactured
COGS	Cost of Goods Sold
COPA	Controlling Profitability Analysis
COSS	Controlling *Summensätze* (table in the SAP system with total values)
CRM	Customer Relationship Management
CXO	Chief Experience Officer
FAGLFLEXT	Financial Accounting General Ledger Flexible
FI	Financials
FIORI	SAP Trademark for computer application
GL	General Ledger
H	Hypothesis
HANA	High Analytical Appliance (in-memory database, developed from SAP)
HGB	*Handelsgesetzbuch*
HTML	Hyper Text Markup Language
ID	Identification Number

IFRS	International Financial Reporting Standard
K	Catamaran
KI	KXEN Information (key performance indicator for prediction)
KPI	Key Performance Indicator
KR	KXEN robustness (key performance indicator for prediction)
MDM	Master Data Management (SAP System to clarify system data)
ML	Material Ledger
MY	Motor Yacht
ODATA	Open Data (interface standard to transfer data)
OLAP	Online Analytical Processing
OLTP	Online Transactional Processing
PA	Profitability Analysis
PCS	Pieces
POC	Percentage of Completion
PSG	Profitability Segment
ROC	Receiver Operating Characteristic
RPA	Robotic Process Automation
S	Sailboat
S/4HANA	Simple for HANA (simplified application based on HANA database)
SAP	Systems, Applications and Products in Data Processing (SAP SE)
SCM	Supply Chain Management
SE	Societas Europaea (Latin) – European enterprise
S-Price	Standard-Price
SQL	Standard Query Language
SRM	Structural Risk Minimization
SVM	Support Vector Machine
US GAAP	United States General Accepted Accounting Principles
WIP	Work in Process

Reference List

1. Artamonow, M.: Auswirkungen von Industrie 4.0 auf das Controlling. Studylab Verlag, Norderstedt, 2017. Page 35.
2. Auer von, L.: Ökonometrie. Eine Einführung. 7. Aufl. 2016. Page 19. Page 157.
3. Awad, M., Khanna, R.: Efficient Learning Machines. Theories, Concepts and Applications for Engineers and Designers. Publisher: Apress Media LLC, New York, 2015.
4. Bakhshaliyeva, N., Chen, J. L., Dommer, U., Samlenski, E., Schmedt, H., Schulze, N., Wilczek, R.: SAP Predictive Analytics – Vorausschauende Analyse mit SAP, Rheinwerk, Bonn, 2017
5. Bamberg, G., Coenenberg, A., Krapp, M.: Betriebswirtschaftliche Entscheidgungslehre. 15th edition, 2012.
6. Bayes, Th., 1763: Bayes, Th.: *An Essay towards solving a Problem in the Doctrine of Chances. In: Philosophical Transactions*. Band 53, 1763, pages 370–418. An essay solving a problem in The doctrine of chances. By the late rev. Mr. Bayes F.R.S. Communicated by Price, R. in a letter to John Canton A.M.F.R.S.
7. Bayes, Th., 2008: Versuch zur Lösung eines Problems der Wahrscheinlichkeitsrechnung. Hrsg.: H. Timerding. Leipzig. 1908.
8. Beck, H.: Behavioral Economics. Eine Einführung. Springer-Gabler, Wiesbaden. 2014
9. Bieger, Th., zu Knyphausen-Aufseß, Krys, Chr.: Innovative Geschäftsmodelle. Konzeptionelle Grundlagen, Gestaltungsfelder und Unternehmerische Praxis. Springer Verlag, 2011.
10. Bloomberg, J.: Digitization, and Digital Transformation: Confuse them at your peril. Published 04/2018. https://www.forbes.com/sites/jasonbloomberg/2018/04/29/digitization-digitalization-and-digital-transformation-confuse-them-at-your-eril/#490513122f2c
11. Borchardt, A., Göthlich, St.: Erkenntnisgewinnung durch Fallstudien. In: Söhnke, A. et al.: Methodik der empirischen Forschung, Gabler Verlag, 2007. Page 33.
12. Busemeyer, J. R. & Townsend, J. T.: Decision field theory: A dynamic cognitive approach to decision-making in an uncertain environment. Psychological Review, 100, pages 432–459. 1993.

13. Butsmann, J., Crumbach, M., Franke, J., Köhler, B., Morgenthaler, J.: SAP S/4HANA Embedded Analytics – Architektur, Funktionen, Anwendungen. Rheinwerk, Bonn, 2019.
14. Charbert, A.; Forster, A.; Tessier, L.; Vezzosi, P.: SAP Predictive Analytics – The Comprehensive Guide. 2017. Boston, USA.
15. Cortes, C., Vapnik, V.: Support Vector, Networks. Kluwer Academic Publishers. Boston (USA). In: Machine Learning no. 20, 1995. Pages 273–297.
16. Dendukuri, N., Lawrence, J.: Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests. Published in Biometrics. Vol. 57, 2001, pages 158–167.
17. Dianzi, Keji, Daxue, Xuebao: An Overview of Theory and Algorithm of Support Vector Machines. In: Journal of the University of Electronic Science and Technology of China. Vol. 40/1, pages 2–10. The year 2011.
18. Diederich, A.: Dynamic Stochastic Models for Decision Making under Time Constraints. In: Journal of Mathematical Psychology 41, pages 260–274,1997
19. Dörsam, P.: Grundlagen der Entscheidungstheorie. 6. Edition. Heidenau. 2013. Page 43.
20. Draper, D.: Bayesian Modeling, Inference and Prediction. Department of Applied Mathematics and Statistics. University of California, Santa Cruz. 2005. Published by the University.
21. Ester, M., Sander, J.: Knowledge Discovery in Databases. Techniken und Anwendungen. Springer. Berlin. 2013
22. Feindt, M. and Kerzel, U.: Prognosen bewerten, Springer Gabler, 2015, page 59.
23. Fischer, J.: Support Vector Machines (SVM). Seminar „Statistische Lerntheorie und ihre Anwendungen“. Universität Ulm. June 12th, 2007.
24. Ford, D., Foliet, J.: Payment Acceptance will Never be the Same after the COVID-19 Pandemic. Published October 1st, 2020. Gartner Social Media Analytics.
25. Gänßlein, S.: Die Digitalisierung ist eine Chance für den Controller. In: Horvath, P.: Controlling, Zeitschrift für erfolgsorientierte Unternehmenssteuerung, Sonderausgabe 2017. Page 21.
26. Gassmann, O., Sutter, Ph.: Digitale Transformation gestalten: Geschäftsmodelle - Erfolgsfaktoren – Checklisten. 2. Auflage 2019. Hanser Verlag, München. Page 7 and page 28.
27. Georgii, H.-O.: STOCHASTIK. Einführung in die Wahrscheinlichkeitstheorie uns Statistik. 5. Auflage. Berlin und Boston. 2015. Page 240 and page 353.

28. Gillenkirsch, R., R.: Bayess Theorem. In: Gabler Wirtschaftslexikon. Revision von Bayes-Theorem *vom* 19. 02. 2018 –16:03
https://wirtschaftslexikon.gabler.de/definition/bayes-theorem-53898/version-276960
29. Gleich, R.; Tschandl, M. (Ed.): Digitalisierung und Controlling. Technologien, Instrumente, Praxisbeispiele. Haufe. Freiburg, München, Stuttgart. 2018.
30. Gleich, R., Grönke, K., Kirschmann, M., Leyk, J.: Controlling und Big Data. Anforderungen, Auswirkungen, Lösungen. Haufe. Freiburg, München. 2014.
31. Gračan, D.; Gregorić, M.; Martinić, T.: NAUTICAL TOURISM IN CROATIA: CURRENT SITUATION AND OUTLOOK; Conference Paper 2016. Page 1. In: Research Gate, https://www.researchgate.net/publication/325181470
32. Handelsblatt, Business Magazine, July 6th, 2019. https://www.handelsblatt.com/unternehmen/industrie/foto-unternehmen-kodak-selbst-laeutete-den-niedergang-ein/6083586-2.html?ticket=ST-4877389-c3o2Df925UnrpARScGc2-ap4
33. Hasendonckx, M.: Margin Analyis in S/4HANA. Presentation SAP SE. 2019
34. Heinert, M.: Support Vector Machines – Teil I: Ein theoretischer Überblick. Fachbeitrag. In: ZfV, 135 Jg., 2010.
35. Held, L.: Methoden der statistischen Inferenz – Likelihood und Bayes. Spektrum akademischer Verlag. Heidelberg, 2008.

Hlaváč, V.: Vapnik-Chervonankis learning theory. Czecz Technical University in Prague. Czesz Institute of Informatics, Robotics and Cybernetics. Original publication 1974. Teach-press EN. Download 2019. http://people.ciirc.cvut.cz/~hlavac/TeachPresEn/31PattRecog/

36. Hoffmeister, C.: Digital Business Modeling. Digitale Geschäftsmodelle entwickeln und strategisch verankern. Hanser. München. 2015.
37. Hölscher, B., Bert, F.: Digitales Dilemma – Unternehmen im Spannungsfeld zwischen Effizienz und Innovation. Arkadia, Hamburg, 2017. Pages 94–95.
38. Hölzlwimmer, A.: Integrierte Werteflüsse mit SAP S/4HANA. Rheinwerk Publishing. Bonn and Boston. 4. Auflage, 2021. Page 326.
39. Horváth, P., Gleich, R., Seiter M.: Controlling. 2015. 13. Auflage. Page 362.

40. Hurwitz, J., Kirsch, D.: Machine Learning for Dummies. Understanding Machine learning fundamentals. Make sense of machine learning algorithms. Build your data science team. Jon Wiley & Sons Corp. Hoboken (NE), 2018.
41. Hussy, W., Schreier, M. & Echterhoff, G.: Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor. 2. Auflage. Berlin, Heidelberg: Springer, 2013. Page 225
42. Hyndman Rob, J., Athanaspopoulos, G.: Forecasting – Principles and Practice. A comprehensive introduction to the latest Forecasting methods using R. Learn to improve Your forecast accuracy using dozens of real data Examples. Publisher OTEX. Melbourne 2014. Page 83.
43. Irrek, W.: Controlling als Rationalitätssicherung der Unternehmensführung – Denkanstöße zur jüngsten Entwicklung der Controlling-Diskussion. In: krp – Kostenrechnungspraxis, Business Magazine, 2002, page 46.
44. Jeschke, B. G.: Entscheidungsorientiertes Management. Einführung in eine konzeptionell fundierte pragmatische Entscheidungsfindung. Berlin, Boston. 2017. Page 57.
45. Kelleher, J. D., Mac Namee, B., D'Arcy, A.: Machine Learning for Predictive Data Analytics, Algorithms, Worked Examples and Case Studies. Cambridge, London. 2015. Page 323.
46. Kießwetter, M., Vahlkamp, D.: Data Mining in SAP Netweaver BI. Gallileo Press, Bonn, 2007.
47. Klostermann, O., Klein, R., O'Leary, J. W., Merz, M.: Praxishandbuch SAP BW. 2015. Rheinwerk Press. Bonn.
48. Kovačić, M: Criteria for selecting a location for a port of nautical tourism. Research Gate. 2009.
49. Kowalczyk, A.: Support Vector Machines. Succinctly. Published from Syncfusion Inc. Morrisville (USA), 2017.
50. Krzanowski, W. J.: Hand, D. J.: ROC Curves for Continuous Data. In: Monographs on Statistics and Applied Probability 111. CRC Press Taylor & Francis Group. London, New York. 2009
51. Kucukvar, M. Noori, M., Egilmez, G., Tatri, O.: Stochastic decision modeling for sustainable pavement designs. Springer, Berlin, Heidelberg, 2014.
52. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, Heidelberg, 2013.
53. Kuik, Sw. S., Kaihara, T., Fuji, N.: Stochastic Decision Model of the Remanufactured Product with Warranty. Proceedings of the International Multi-Conference of

Engineers and Computer Scientists. Vol II, IMECS, March 18–20, Hong Kong. 2015.

54. Kunze, Th., Schmalzing, K., Reinelt, D.: SAP S/4HANA Finance Customizing. Rheinwerk Publishing, Bonn. 2. Auflage, 2020.
55. Langmann, Chr.: Digitalisierung im Controlling. Springer Verlag, Heidelberg, 2019.
56. Laux, H., Gellenkirsch, R.M., Schlenk-Mathes, H.: Bildung eines Wahrscheinlichkeitsurteils und Bewertung von Informationen. 9. Auflage. Springer Verlag, Heidelberg, 2014.
57. Laux, H., Gillenkirch, R. M., Schenk-Mathes, H. Y.: Entscheidungstheorie. 10th Edition. Frankfurt am Main, 2018.
58. Lawrence, K. D., Klimberg, R. K.: Advances. Business and Management Forecasting. Vol. 12. Emerald Publishing. United Kingdom, North America, Japan, India, Malaysia, China. 2018. Page 63.
59. Leavy, P.: Research Design. Quantitative, Qualitative, mixed methods, arts-based and Community based participatory research-Approaches. Guilford Press, New York, London. 2017.
60. Lebefromm, U.: Profitability Analysis in SAP S/4HANA. Webinar online on January 21st, 2021.
61. Lebefromm, U., Mayer, U.: Die Strategie der Ergebnisrechnung. Webinar at SAP Financial Forum, March 23rd–24th, 2020 in Walldorf, Germany.
62. Luković, T.: Nautical Tourism. Dubronik, 2013.
63. Luković, T, Lapko, A., Vukovic, A.: Sources of Economic Development in Transition Economies. Lambert Academic Publishing, Mauritius. 2019, page 40.
64. Malic, Livia, Varaždinac, Patricija, Škiljan, Ivona: Multi-Criterion Decision Model for Marina Location Selection in the County of Primorje-Gorski-Kotar. Published in Research Gate: https://www.researchgate.net/publication/331624993
65. MindForest, Online Platform, 2020. https://www.mindforest.com/digital-pro-its-a-mindset/#_ftn1
66. Mindsquare, 2019: Homepage IT Consultancy. SAP Master Data Management. Download June 22nd, 2019. https://mindsquare.de/knowhow/sap-master-data-management/
67. Möller, K.: Controlling-Prozessmodell 2.0. Leitfaden für die Beschreibung und Gestaltung von Controlling-Prozessen. 2. Auflage. Haufe-Verlag. Freiburg – München – Stuttgart. 2017.

68. Morris, D.: BAYES THEOREM. A visual introduction for beginners. Blue Windmill Media. Printed by Amazon, Leipzig. 2017.
69. Nasca, D., Munck, J.-Chr., Gleich, R.: Controlling-Hauptprozesse: Einfluss der digitalen Transformation. Investigation April – June 2018. In: Gleich, R. et al., 2018. Digitalisierung und Controlling. Page 73.
70. Peović, K.: Die politische und wirtschaftliche Situation Kroatiens – die Peripherie Europas heute. Blog. Published in: "transform europe". Vienna. 2018.
71. Progroscewska, Paulina, Justina: Entwicklung eines Geschäftsmodells nach dem Prinzip des St. Galler Business Model Navigator für das HAWAI Projekt. Bachelorarbeit. Hochschule für angewandte Wissenschaften, Hamburg. 2016.
72. Roßmeisl, E.; Gleich, R.: Industrie 4.0: Neue Aufgaben für das Produktionsmanagement und - Controlling. In: Gleich, R. et al.: Controlling und Big Data. Haufe. München. 2014
73. Runkler, A.: Data Mining – Modelle und Algorithmen intelligenter Datenanalyse. 2nd Edition. Springer-Vieweg, Wiesbaden, 2015.
74. Russell, St., Norvig, P.: Künstliche Intelligenz. Ein moderner Ansatz. 3. Aufl. Pearson DE, München, 2012. Page 863.
75. Safar, M: Was ist eigentlich ein Software Roboter? White Paper. Weissenberg Business Consultin GmbH. Wolfsburg. Download June 14th, 2020. https://weissenberg-solutions.de/was-ist-ein-software-roboter/#:~:text=Der%20sogenannte%20Software%2D Roboter%20ahmt,den%20Anwendungsm%C3%B6glichkeiten %20keine%20Grenzen%20gesetzt.
76. SAP Training: S4F01 – Financial Accounting in SAP S/4HANA for SAP ERP FI Professionals. 2021. Col. 17, Figure 12, page 17.
77. Schäffer, U.: Levers of Organizational Resilience. In: Controlling & Management Review. Zeitschrift für Controlling & Management. 64, Jahrgang. Ausgabe 6-7. 2020. Pages 8–19.
78. Scherer,R., Willinger, M.: Parallele Rechnungslegung mit SAP. Galileo Press, Bonn, Boston. 2006.
79. Schmalzing, K., Löw, I.: Controlling in SAP S/4HANA. Das Praxishandbuch. Rheinwerk Publishing. 2019.
80. Schnell, R., Hill, P. B.: Methoden der empirischen Sozial-Forschung. 11. Auflage. 2018.

81. Schöb, O.: Ergebnisrechnung mit SAP. Prozesse und Customizing im Detail. Konzeption einer aussagekräftigen Ergebnisrechnung. Integration mit anderen SAP Komponenten. Galileo Press, Bonn und Boston. 2009. Page 370.
82. Schramm, W.: Notes on Case Studies of instructional media projects. Working paper for the Academy for Educational Development, Washington, D.C.
83. Sönke, A., Klapper, D., Konradt, U., Walter, A., Wolf, J. (Ed.), Methodik der empirischen Sozialforschung, Gabler Verlag, 2. Auflage, 2007.
84. Spur, G., Krause, F.-L.: Das virtuelle Produkt – Management der CAD-Technik. 1997. Hanser Verlag, München. Page 577.
85. Steinke, K.-H.; Schmidt, W.: Auf dem Weg zum Controlling 4.0 – Leitfaden für Controlling im Wandel. Haufe-Lexware, Freiburg, 2017
86. Stiegler, St. M.: Who discovered Bayes Theorem. The American Statistician. 1983. Vol. 37. No. 4
87. Stiegler, St. M.: Richard Price, the first Bayesian. Statistical Science, 2018. Vol. 33. No. 1. Pages 117–122. Institute of Mathematical Statistics.
88. Stier, W.: Empirische Forschungsmethoden. 2. Auflage. Springer Verlag, Berlin, Heidelberg 1999. Page 332.
89. Subbalakshmi, G., Ramesh, K., Chinna Rao, M.: Decision Support in Heart Disease Prediction System using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE). Vol. 2, No. 2, 2011.
90. Thackray, A., Brock, D.C., Jones, R.: Moore's Law: The Life of Gordon Moore, Silicon Valley's Quiet Revolutionary. 2015. Basic Books, New York.
91. Tschandl, M.; Peßl, E.; Baumann, S.: Roadmap Industrie 4.0 – Strukturierte Umsetzung von Smart Production und Services in Unternehmen. In: Wing Business, I, 2017. Page 20 – 23.
92. Tschandl, M., Kogleck, R.: Controller als Innovatoren: Von der Digitalisierungs-Roadmap zum neuen Geschäftsmodell. In: Gleich, R. et al.: Digitalisierung und Controlling – Technologien – Instrumente – Praxisbeispiele. Haufe, 2018. Pages 27–48.
93. Vapnik, V. N.: Inductive principals of statistics and learning theory. In: Yearbook of the Academy of Science of the USSR on Recognition, Classification and Forecasting (in Russian). Vol. 1, Nauka, Moscow. (English translation: V. N. Vapnik, Inductive principles of statistics and learning theory. In: Mathematical Perspectives on Neural Networks, Smolensky, Mozer and Rumelhart, eds., Lawrence, Erlbaum Associates. Springer New York, 1996.

94. Vapnik, V. N.: Statistical Learning Theory. A Volume in the Wiley Series on Adaptive and Learning Systems for Signal Processing. Communications, and Control. Simon Haykin, Series Editor. Wiley. 1998. Reprint 2018. London. New Delhi.
95. Vapnik, V. N., Chevronenkis, A. Ya.: The necessary and sufficient conditions for consistency of the method of empirical risk minimization. Yearbook of the Academy of Science of the USSR on Recognition, Classification and Forecasting. Vol. 2, Nauka, Moscow, pages 207–249 (in Russian). English translation: V. N. Vapnik and A. Ya. Chevronenkis (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization. Pattern Recognition. Image Anal.1(3), pages 284–305.
96. Varnholt, Norbert T., Hobberg, P., Gerhards, R., Wilms, St., Lebefromm, U.: Operatives Controlling und Kostenrechnung. Betriebswirtschaftliche Grundlagen und Anwendung mit SAP S/4HANA. 3. Auflage. Oldenbourg Verlag. 2020. Page 372.
97. Vehtari, A., Ojanen, O.: A survey of Bayesian predictive methods for model assessment, selection and comparison. In: Statistic Surveys. Vol. 6 2012. Pages 142–228.
98. Vitezić, N., Lebefromm, U.: Production Controlling in the Digital Age. University of Rijeka, Faculty of Economics and Business, 2018.
99. Von Auer, L.: Ökonometrie – Eine Einführung. 7. Auflage. Springer Verlag, Heidelberg. 2016.
100. Weber, J. et al., Digitalization: Eight Challenges for Controllers – How digitalization will change controlling and what controllers should do about it. In: WHU on Controlling, Academic insights for professionals in controlling & finance. https://www.whu-on-controlling.com/en/latest-thinking/digitalization/ Download October 7th, 2019.
101. Weber, J., Schäffer, U.: Einführung in das Controlling, Schäffer-Poeschel, Stuttgart, 2016.
102. Weber, J., Schäffer, U.: Rationalitätssicherung und Unternehmensführung, Schriften des Centers for Controlling & Management, Springer-Gabler, Wiesbaden, 2001.
103. Weber, J., Schäffer, U., Langenbach, W.: Gedanken zur Rationalitätskonzeption des Controlling. WHU Forschungspapier Nr. 70. 1999.
104. Weber, K.-H.: Schnelleinstieg ins Finanzwesen (FI) mit SAP S/4HANA. Espresso Tutorials, Gleichen, 2019. Page 86.
105. Weißenberger, B. E.: IFRS für Controller. Alles was Controller über IFRS wissen müssen. Publisher: Haufe, Freiburg, Berlin, München. 2. Auflage, 2011.

106. YachtRent: Overview of the Croatian Charter Fleet. June 19th, 2014. https//www.yacht-rent.com/overview-of-the-croatian-charter-fleet

107. Yang, Zh., Su, Xi: Customer Behavior Clustering using SVM. International Conference on Medical Physics and Biomedical Engineering. 2012. Published by Physics Procedia. Vol. 33. 1489–1496.

108. Yin, R. K.: Case Study Research and Applications. Design and Methods. 6th edition. SAGE publications inc. Los Angeles. 2018.

109. Zhang, Ch.: Structure of Indicator Function Classes with Finite Vapnik-Chervonenkis Dimensions. Publisher IEEE Computational Intelligence Society. Published: IEEE Transactions on Neuronal Networks and Learning Systems. Volume 24, issue 7th July 2013.

110. Zuo, Yi, Ali, A. B. M. Shawkat, Katsutoshi, Yada: Consumer purchasing behavior extraction using statistical learning theory. Published on the 18th International Conference on Knowledge-Based and Intelligent Information & Engineering. Gdynia, Poland. Publisher: Procedia Computer Science (35), 2014, pages 1464–1473.

STATEMENT

by which I, Uwe Lebefromm, index number: 1300, doctoral student at the Faculty of Economics and Business, University of Rijeka, as the author of a doctoral dissertation entitled: Digital Transformation and Business Predictions:

1. I declare that I prepared my doctoral dissertation independently under the mentorship of Prof. Neda Vitezić, PhD. In my work, I applied the methodology of scientific research and used the literature which is stated at the end of the paper. Other people's knowledge, attitudes, conclusions, theories, and laws that I am directly stating or paraphrasing in the dissertation I have cited and linked to the bibliographic units used in accordance with the provisions of the Ordinance on the preparation and equipping of doctoral theses of the University of Rijeka, Faculty of Economics and Business in Rijeka. The dissertation is written in the spirit of the English language.

2. I give permission for my work to be permanently stored in a publicly available digital, free of charge repository of the institution and the University and in the public Internet database of works of the National and University Library in Zagreb, in accordance with the obligation under the provision of Article 83, paragraph 11 of the Act on Scientific Activities and Higher Education (OG 123/03, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13, 101/14, 60/15).

I certify that the final version of the defended and completed doctoral dissertation will be submitted for dissertation storage. With this statement, as an author, I give my approval that my work, without compensation, be permanently made public and available free of charge to students and staff of the institution.

Uwe Lebefromm



_____