DOCTORAL DISSERTATION

Proposal and Implementation of Churn Prediction System for Telecommunications Company

Study program: Systems Engineering and Informatics
Field of study: Systems Engineering and Informatics

Author: Ing. Jan Mandák

Ostrava, 2018
DECLARATION

I hereby declare that I wrote the entire doctoral dissertation myself, including attachments. I have included all information sources in the references and have cited them appropriately in the doctoral dissertation.

Ostrava, 12th July, 2018
Acknowledgement

Mainly, I would like to thank my advisor, Jana Hančlová, for her support, valuable advice, feedback and positivism, which helped me a lot during the whole Ph.D. studies. Special thanks belong also to my colleagues from the department of Systems Engineering and to my family and friends for their support and patience throughout my doctoral studies. My thanks go also to my girlfriend who always encouraged me and discussed with me tricky linguistic issues.
Abstract

The telecommunications industry is a large and important part of the sector of information and communication technologies. Because of its highly competitive nature, it is very common for customers to switch to another service provider or to non-renew a commitment. This behavior of customers is called customer churn. It is an expensive business problem since acquiring new customers costs five to six times more than retaining the existing ones. With the still decreasing costs of data storage, telecommunication companies have an access to various customer related data sources, which can be used to create predictive models helpful to identify who, when and why is about to leave the company. The main objective of the dissertation thesis is to propose and implement churn prediction system, which helps selected telecommunications company to reduce the number of churning customers and better understand the customer base. The partial goals are to summarize current theoretical, methodological and empirical results and to process raw data, divide customers into clusters, estimate and compare selected classification models, determine the key factors driving the churn, create customer knowledge database and visualize the data in selected visualization tool. Firstly, the methodological part of the thesis is focused on the data mining methodology CRISP-DM. Then methods of cluster analysis utilized in the thesis such as Gower distance and $k$-medoids algorithm and classification models – logistic regression, decision trees and random forests are described. Performance measures for comparison of predictive ability of classification algorithms are also introduced. The last part deals with an estimation of future performance of predictive models - approaches such as training and testing data set, cross-validation or bootstrap sampling. The application part of the thesis is devoted to the proposal of churn prediction system. Input data in CSV files are loaded into statistical tool R. Customers are then divided into clusters and logistic regression, decision tree and random forest models are estimated for the entire training data set as well as for each cluster. Customer characteristics, predicted probabilities of churn and variable importances are stored to MySQL relational database and these data are used to create a dashboard in the visualization tool Qlik Sense. This dashboard is provided to business users as a user-friendly tool for understanding the customer behavior.

Keywords: Logistic regression, decision trees, random forests, k-medoids, R, MySQL, Qlik Sense, telecommunications, customer churn

JEL classification: C38, C52, C53, L96
Abstrakt

Telekomunikační sektor je důležitou částí sektoru informačních a komunikačních technologií. Díky vysoké konkurenci je v tomto sektoru běžné, že zákazníci přecházejí k jinému poskytovateli služeb nebo neobnovují své smlouvy. Toto chování zákazníků se označuje jako “churn“. Telekomunikační společnosti musejí na udržení zákazníků vynakládat nemalé prostředky, které jsou však 5 až 6 krát menší než náklady na získání nových zákazníků. Se stále cenově dostupnějšími a většími datovými úložišti mají telekomunikační společnosti k dispozici obrovské množství informací o zákaznících, které mohou být využity pro tvorbu prediktivních modelů užitečných pro předvídat toho kdo, kdy a proč se chystá opustit společnost. Hlavním cílem doktorské disertační práce je navrhnout a implementovat systém pro předvídat odchází zákazníků, který pomůže vybrané telekomunikační společnosti tento počet snížit a lépe porozumět zákaznické bázi. Dílčí cíle jsou shromáždění současné teoretické, metodologické a empirické výsledky a zpracovat vstupní data, rozdělit zákazníky do shluků, odhadnout a porovnat vybrané klasifikační modely, určit klíčové faktory odchodů zákazníků, vytvořit zákaznickou databázi znalostí a nakonec data vizualizovat ve vybraném nástroji. Metodologická část práce je zaměřena nejprve na popis metodologie vzniku CRISP-DM. Dále jsou popsány metody shlukové analýzy jako Gowerova vzdálenost či metoda k-medoidů, aplikované klasifikační modely – logistická regrese, rozhodovací stromy a náhodné lesy a také metody pro porovnání prediktivních schopností klasifikačních algoritmů. Poslední část se zabývá odhadem budoucí výkonnosti modelu – přístupy jako rozdělení datového souboru na trénovací a testovací, křížovou validaci nebo bootstrapingem. Aplikační část práce je věnována návrhu systému pro předvídat odchodu zákazníků. Vstupní data ve formátu CSV jsou nahrána do statistického nástroje R, kde jsou zákazníci rozděleni do shluků a odhadnuty modely logistické regrese, rozhodovacího stroje a náhodných lesů pro celý tréninkový datový soubor a jednotlivé shluky. Zákaznické charakteristiky, predikované pravděpodobnosti odchodu a důležitosti proměnných jsou uloženy do relační databáze MySQL a použity pro tvorbu dashboardu ve vizualizačním nástroji Qlik Sense. Dashboard je poskytnut byznys uživatelům jako uživatelsky přívětivý nástroj pro pochopení chování zákazníků.

Klíčová slova: Logistická regrese, rozhodovací stromy, náhodné lesy, metoda k-medoidů, R, MySQL, Qlik Sense, telekomunikace, odcházející zákazníci

JEL klasifikace: C38, C52, C53, L96
Contents

1 Introduction ................................................................................................................................. 9
   1.1 Purpose and motivation ........................................................................................................ 9
   1.2 Main and secondary objectives of work ........................................................................... 10
   1.3 Current state of examination ............................................................................................ 11

2 Customer relationship management in telecommunications ................................................. 16
   2.1 Customer satisfaction and its measurement ................................................................... 16
   2.2 Customer loyalty and its measurement .......................................................................... 20
   2.3 Relationship marketing ..................................................................................................... 22
   2.4 Preventing churn in telecommunications ......................................................................... 24

3 Description of applied methods and approaches ................................................................. 27
   3.1 Data mining methodology – CRISP DM ......................................................................... 27
   3.2 Cluster analysis .................................................................................................................. 31
      3.2.1 Gower coefficient of similarity .............................................................................. 34
      3.2.2 Hierarchical clustering ......................................................................................... 35
      3.2.3 Partitioning around medoids ................................................................................. 37
      3.2.4 Determining optimal number of clusters ............................................................... 38
   3.3 Classification models .......................................................................................................... 39
      3.3.1 Logistic regression ................................................................................................. 39
      3.3.2 Decision trees ......................................................................................................... 43
      3.3.3 Random forests ...................................................................................................... 47
   3.4 Evaluating predictive performance .................................................................................. 49
      3.4.1 Estimating future performance .............................................................................. 52
4 Application of quantitative methods for churn prediction.................................55

4.1 Description of the data..................................................................................56
4.2 Exploratory data analysis...............................................................................60
4.3 Predictive models trained on all data set ........................................................67
  4.3.1 Logistic regression model ........................................................................67
  4.3.2 Decision tree model ..............................................................................70
  4.3.3 Random forest ......................................................................................72
  4.3.4 Comparison of variable importance .......................................................74
  4.3.5 Predictive performance computed from testing data set .......................75
4.4 Cluster analysis ............................................................................................79
4.5 Predictive models trained for individual clusters .........................................83
  4.5.1 Predictive models for cluster 1 ..............................................................84
  4.5.2 Predictive models for cluster 2 ..............................................................87
  4.5.3 Predictive models for cluster 3 ..............................................................90
  4.5.4 Predictive models for cluster 4 ..............................................................93
4.6 Comparison of results ..................................................................................97
  4.6.1 Comparison of predictive performance among clusters ....................97
  4.6.2 Comparison of “all data set” approach vs. “clustering” approach ..........104
4.7 Proposal of churn prediction system ............................................................107

5 Conclusion .......................................................................................................115

References ..........................................................................................................119

List of abbreviations and symbols .......................................................................126

The list of attachments .........................................................................................130

Appendix .............................................................................................................131

  A.1 User guide for dashboard in Qlik Sense ....................................................131
1 Introduction

The telecommunications industry is a large and important sector within the sector of information and communication technologies. It consists of telecommunications companies and internet service providers and it plays a crucial role in the development of information society. The biggest revenue in the sector is still generated by telephone calls, but thanks to the rapidly developing network technology and high speed internet access, telecommunication sector is nowadays more about texts (messaging, emailing), images and video streaming. The telecommunications industry is highly competitive with many alternative providers, so for customers it is easy to change their provider. In each country there are typically at least three mobile network operators and there can also be many mobile virtual network operators. Another specific of this industry is heavy regulation by national or supranational authorities. Because of intense competition and significant regulatory pressures, the price of mobile services tends to decrease over time. However, thanks to the more mobile users in emerging markets and more data usage, global mobile revenue remains on a positive trend.

The industry is significantly changing and it is expected that the change will continue. The biggest opportunity lies in a new revenue areas such as mobile data and companies do not hesitate to invest huge amounts of money into them. It is expected that the high level of competition among mobile network operators will continue. For companies it will be necessary to invent new and new value-added services to satisfy customer needs and to gain a competitive advantage against competitors. Also the pressure from regulators is expected to continue, e.g. General Data Protection Regulation (GDPR) in Europe has been implemented in May 2018. It is also clear that a usage of customer and other types of data will bring a competitive advantage to companies. Telecommunication operators store a huge amount of data. In order to retain or improve its position in the market, it will be necessary to unlock a hidden value in the data using Big Data and Advanced Analytics technologies. One important use case where big data can help mobile operators is the reduction of customer churn.

1.1 Purpose and motivation

The doctoral dissertation thesis focuses on the solving of the problem of customers leaving telecommunication companies. Decision support system will be designed to deal with this problem.
The proposed system will be created for a European company, but thanks to the usage of open-source tools for analysis, data storage and visualization, it can be used in any company. The only prerequisite is the data set containing customer’s characteristics and information whether they left the company or stayed. The main purpose of the thesis is to create a churn prediction system which will help selected company to reduce the amount of customers leaving their business – churners.

The initial idea to focus the doctoral dissertation thesis on the theme of churn modelling came up from requirements of customers of the author’s employer. Number of scientific papers focused on this theme confirmed the author’s thought that it is a topical business problem which touches businesses from various sectors. For instance telecommunication provider’s fight against churning customers, manufacturers who provide services based on contracts need to predict contract renewal and departments of human resources face employee turnover. Customer relationship managers responsible for the number of customers leaving the company need to find various ways how to reduce the number of churners.

1.2 Main and secondary objectives of work

The main objective of the dissertation thesis is to create and implement a system for modelling and prediction of churning customers for the selected European telecommunication company. The proposed system should serve as a quantitative support for decision makers responsible for customer relationship management (CRM). The CRM managers and their subordinates will have an access to the tool, which will help them to identify the customers with high probability of churning in the period of the next 45 days, determine the main reasons of churn, make the proactive actions to prevent the customers from churning, reduce the churn rate, retain the customer base and last but not least explore groups of customers with similar characteristics.

On a regular basis of 45 days the system must be able to:

- prepare dependent and independent variables from the source data,
- divide customer into reasonable number of segments,
- characterize each customer segment,
- estimate several classification models using selected approaches,
- determine the key factors influencing the probability of customers to churn,
- compare predictive accuracy of linear and non-linear classification models,
• select the best predictive model according to performance measures computed on the test data set unseen in the model training phase,
• store the customer characteristics, importance of variables and predictions to database,
• visualize the data in the selected visualization tool.

Another partial goal of the thesis is to create a summary of current theoretical, methodological and empirical results, which serves as a basis for selection of variables and quantitative methods or for comparison of results with other similar studies.

The proposed system is based on open source analytical tools. For data preparation and models training the R statistical programming language is used. The computed data and predictions are stored in MySQL database and the data are further visualized in Qlik Sense visualization tool. The proposed system could serve as a template for other businesses dealing with the problem of customer churn, not only from the telecommunications sector, but also from other sectors, such as banking, insurance or retail.

The creation and implementation of this churn prediction and modelling system brings several benefits. Firstly, there are benefits for which it is not so easy to quantify their financial benefit. Here we can assign e.g. deeper insight into customer base using customer segmentation and visualizations in Qlik Sense. Good understanding of customer characteristics and behavior is a key prerequisite for creation of services fitting the customer needs. Secondly, the savings connected with the saved customers can be easily calculated. Monthly savings of churn prediction model deployment are calculated as a difference between the average invoice of customers, which were correctly predicted to churn (TP – true positives) and the sum of average invoice of leaving customers, who weren’t predicted to churn (FN – false negatives) and the retention costs spent to save the customers predicted to churn:

\[
savings = TP \cdot \text{average invoice} - \left[ FN \cdot \text{average invoice} + \frac{FN \cdot \text{average invoice}}{(FP + TP) \cdot \text{retention costs}} \right]
\]  
(1.1)

1.3 Current state of examination

Customer churn prediction is a key issue in many fields such as telecommunications, internet service providers, e-commerce, retail, marketing, banking or financial Services. Responsible
persons, in each of these fields, study the behavior of customers and try to predict, whether a certain customer is likely to quit the service of his/her existing provider and join a new service provider.

Data mining, the process of discovering patterns in large datasets, is applied to help businesses to reduce the amount of churners. Predictive models based on analyzing the past behavior of customers with the goal to forecast how the customer will behave in the future represent a tool to tackle with this business problem. There are many models applicable on distinguishing churners and non-churners, e.g. association rules, classification models, clustering, regression or just various types of visualizations.

Joseph (1996) states that neural networks have a big potential to be applied to industrial and also business domains. Omar et al. (2014) used neural networks to predict churn in randomly selected 5,000 customers from a telecommunication company in Jordan. They found that monthly fees, total minutes of usage and 3G services are the most important drivers of churn. Also Sharma and Panigrahi (2011) applied neural networks in SPSS Clementine to the data set of 2427 customers from UCI Machine Learning repository (Lichman, 2013). They achieved accuracy 0.92 and found out that medium sized neural networks perform the best for the customer churn prediction. Another application of neural networks can be found in Jadhaw and Pawar (2011). Call related information (number of calls, duration of calls etc.) for 895 randomly selected users from an Indian telecommunication company was used to train a model. Churning customers, on the average, used their phone for fewer for the first few days and then their behavior changed.

Another frequent classification method for churn prediction is logistic regression. Olle and Cai (2014) gathered dataset of 2,000 subscribers from an Asian telecommunications operator. Location, age, tenure or tariff were some of the explanatory variables. Logistic regression model estimated in WEKA achieved precision 0.72 and recall 0.75. Gürsoy (2010) analyzed churn in a major telecommunication firm in Turkey. Logistic regression model was estimated in SPSS Clementine. The following explanatory variables were confirmed as statistically significant using Wald test: discount package, customer age or average length of call. Ahn et al. (2006) investigated determinants of customer churn in the Korean mobile telecommunications service market using logistic regressions. Their results indicate that call quality-related factors influence customer churn. Also heavy users are more likely to leave their service provider. Sebastian and Wagh (2017) gathered dataset with over 2,000 customers described by 22 variables. They achieved accuracy 0.8
by the use of backward stepwise logistic regression. They also made the results more understandable by visualization in Power BI. Modelling customer attrition in telecommunications is important also in African countries, Oghojafor (2012) state that e.g. in Nigeria the annual churn rate came up to 41%. They created a well-structured and compliant questionnaires and obtained 6,000 subscribers of telecommunications service providers. They distinguish churners from non-churners by questioning respondents whether they would you like to change their current service provider. Stepwise logistic regression model revealed that call expenses, providers’ advertisement medium, type of service plan, number of mobile connections and providers’ service facilities are the most important factors driving churn.

Decision trees are widely used in churn analysis because of their ability to model non-linear relationships in the data. Khalida et al. (2010) used decision tree analysis, specifically ID3 algorithm for churn prediction. The process of decision tree creation was described and applied to the sample of data from Malaysian telecommunications Company. Although they used only 3 independent variables (length of service, area and total minutes) they showed that decision tree provided a reasonable predictive value. Kirui et al. (2013) obtained a huge dataset of 112 attributes and 106,405 instances from a European telecommunications company. They prepared four types of features: customer profiles, traffic details, contract-related features and calls pattern features. 112 attributes were decreased to the 60 most important according to information gain and the accuracy of decision tree grew up. They also reported that traffic details, call pattern features and change in call pattern played more important role in churn prediction than customer profile data. Mandăk (2017) used decision tree along with logistic regression to predict customer churn in a European telecommunications company. His results showed that both models are able to catch churners, as all performance metrics were not lower than 0.9. The most important predictors were days till the end of contract, customer lifetime and age of customer.

Besides traditional classification approaches to churn prediction, there are also some less frequently used methods. Hun et al. (2006) conducted a churn analysis in the Taiwan telecommunications industry. Customer demographics (age, tenure, gender), customer internal data (plan type, monthly fee) and call details (call duration, call type) were used as an input data to segment customers into clusters using K-means clustering. They found that users who do not make calls to other users in the same network have a high probability to churn. They also discovered that users whose contract is going to expire in near future have a higher probability to churn. Khan et
al. (2010) also used *K-means clustering* for creation of customer groups for telecommunications Company in Iran. They gathered a dataset containing demographic, billing and usage data. The clustering revealed that billing and usage features had a higher explanatory effect on churn in comparison with demographic data. Another rarely used method for churn prediction is *social network analysis*. Richter et al. (2010) applied a social network based approach to predict the churn of subscribers. They created clusters of customers based on social groups and gave each member of the group a churn score based on his/her membership to the cluster. Only call related data were used in their study and they found out that social leaders can have a significant impact on the churn in their groups. They also found out that a social leader has three times greater probability of churn in comparison with other group members. Lu (2002) used survival analysis to predict which customers are likely to churn and when they are likely to churn. The author gathered data set of 41,374 customers from a telecommunications company of one state in the USA with variables such as age, gender, income, plan type, billing agency, number of weekly calls or customer contract records. Author claims that he is able to catch 90 % of the churners.

It is clear that the customer churn in telecommunications is an important and costly business problem. Many researchers tried to use various quantitative methods for prediction of churners. The information about studies focused on churn prediction in telecommunications are summarized in Tab. 1.1. It is clear that among the most frequent methods used for this business problem belong logistic regression, neural networks and decision trees. This finding is in accordance with Kamalraj and Malathi (2013).

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Method</th>
<th># of customers</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omar et al. (2014)</td>
<td>Neural networks</td>
<td>5 000</td>
<td>Jordan</td>
</tr>
<tr>
<td>Sharma and Panigrahi (2011)</td>
<td>Neural networks</td>
<td>2 427</td>
<td></td>
</tr>
<tr>
<td>Jadhaw and Pawar (2011)</td>
<td>Neural networks</td>
<td>895</td>
<td>India</td>
</tr>
<tr>
<td>Ahn et al. (2006)</td>
<td>Logistic regression</td>
<td>10 000</td>
<td>South Korea</td>
</tr>
<tr>
<td>Gürsoy (2010)</td>
<td>Logistic regression</td>
<td>1 000</td>
<td>Turkey</td>
</tr>
<tr>
<td>Oghojafor (2012)</td>
<td>Logistic regression</td>
<td>6 000</td>
<td>Nigeria</td>
</tr>
<tr>
<td>Olle and Cai (2014)</td>
<td>Logistic regression</td>
<td>2 000</td>
<td>Asia</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>Sample Size</td>
<td>Country</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Sebastian and Wagh (2017)</td>
<td>Logistic regression</td>
<td>2,000</td>
<td>India</td>
</tr>
<tr>
<td>Khalida et al. (2010)</td>
<td>Decision tree</td>
<td></td>
<td>Malaysia</td>
</tr>
<tr>
<td>Kirui et al. (2013)</td>
<td>Decision tree</td>
<td>106,405</td>
<td>Europe</td>
</tr>
<tr>
<td>Hun et al. (2006)</td>
<td>K-means clustering</td>
<td>160,000</td>
<td>Taiwan</td>
</tr>
<tr>
<td>Khan et al. (2010)</td>
<td>K-means clustering</td>
<td>2,685</td>
<td>Iran</td>
</tr>
<tr>
<td>Richter et al. (2010)</td>
<td>Social network analysis</td>
<td>16,000,000</td>
<td></td>
</tr>
<tr>
<td>Lu (2002)</td>
<td>Survival analysis</td>
<td>41,374</td>
<td>USA</td>
</tr>
</tbody>
</table>

**Table 1.1** Churn prediction studies

The above mentioned studies have some disadvantages:

- They usually use only 1 data set (there are not 2 independent training and testing samples),
- Except the study of Lu (2002), time dimension is missing in all of them (when will the customer leave),
- There are no reasons of churn (what are the main reasons of churn),
- They did not contain any visualization of the discovered findings.

This doctoral dissertation thesis answers all of the following questions: which customers are about to leave the company, when they will leave and what are the key factors influencing their behavior. Churn prediction dashboard, making the predictive results accessible for CRM or retention department of company, is also part of the thesis.
2 Customer relationship management in telecommunications

The telecommunications industry is rapidly changing with new technologies and value-added services occurring almost every month. As the technologies and services change very quickly, it is necessary to track and analyze all interactions of companies with existing and also potential customers. Monitoring and analyzing of customer satisfaction is the first step to understand customer’s needs and requirements and is really important for the company performance.

Customer satisfaction is described in detail at the beginning of this theoretical part of the thesis. Especially, the emphasis is put on the process of measuring and monitoring of customer satisfaction. Because it is not sufficient only to focus on customer satisfaction, the importance of customer loyalty is also discussed. Loyal customers are a dream for companies since their behavior is characterized by repeated purchases and positive references. The process supporting long-term relationships with customers (the so called relationship marketing) is also explained. Among some of its features belong the focus on stable customers, regular contact with customers, creation of value perceived by customers, long-term planning, high emphasis on quality of services provided to customers, effort to meet all customer expectations or quality of products and services. Customer churn in telecommunications is the last part of this chapter. Customer churn and its measurement are defined, the types of customer churn (voluntary, involuntary) and its dimensions (who, when and why will churn?) are at first introduced. The typical customer data which are available are also listed. In the end, the economic value of customer retention is also emphasized.

2.1 Customer satisfaction and its measurement

Any quality management system should have working feedback for monitoring of customer satisfaction with products and services. “Customer satisfaction” is in the norm ČSN EN ISO 9000:2016 (ČSN, 2016) defined as an opinion of a customer to the rate of fulfillment of his/her needs and expectations. The European Foundation for Quality Management defines customer satisfaction as a summary of customer’s feelings derived from the differences between his/her expectations and the perceived market reality (Nenadál, 2008).

Customer requirements are declared based on his/her immediate needs, past experiences, and information from the surroundings, including advertising (Nenadál, 2008). Usually the reality is worse than the customer requirements. The resulting gap is defined in the Fig. 2.1 as $X$. 
The size of this gap is an argument of function, which enables us to quantify the rate of customer satisfaction, denoted by $RCS$:

$$RCS = f(X)$$ \hspace{1cm} (2.1)

High rate of customer satisfaction is an assumption for their future loyalty, which is expressed by repeated purchases and positive references. On the contrary, low rate of customer satisfaction leads to warranty claims, complaints and to the loss of a customer. Both situations significantly influence economic performance of producers or service providers.
If a company wants to master the processes of measuring and monitoring of customer satisfaction, it is necessary to realize the following steps:

1. define who is a customer;
2. define indicators of customer satisfaction;
3. propose and create questionnaires for monitoring of customer satisfaction;
4. determine the sample size;
5. select suitable method for data gathering;
6. create procedures for data assessment, including quantification of the satisfaction rate;
7. use the results as an inputs for the processes of continuous improvement.

Let’s describe in more detail what the above mentioned steps are about. For companies it is absolutely necessary to realize **who their customer is**. Generally, there are two types of customers – internal customers (employees) and external customers (subcontractors, purchasers, end customers). Even though the following steps are related mainly to external customers, methods and procedures of monitoring of satisfaction are common for both internal and external customers.

The second, step after the clarification of who the customers are and how important they are, is the **definition of indicators of customer satisfaction**. One approach is called “listening to the voice of customers”. The requirements are defined on the basis of active participation of current and potential customers. The methods like interview, discussions in the focus groups or method of critical events are often used. The term “critical event” represents a particular customer statement related to very positive or very negative experience with the product or service. These critical events are further structured into indicators of customer satisfaction.

Proposal and creation of **questionnaires for monitoring of customer satisfaction** is another important step. Questionnaires are the most frequently used tool for gathering feedback. It is essential to pay attention to the preparation of them – the quality of the proposed questionnaire strongly influences the objectivity of gathered information. The number of questions should not be higher than 15. There are three basic formats of questionnaires – check-list, Likert format and numeric format.
The simplest one is check-list, which is based on reaction of customers in the form “Yes-No” or “Agree-Disagree” (see Tab. 2.1). This format enables only rough estimate of customer satisfaction.

<table>
<thead>
<tr>
<th>Were you satisfied with the service?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
</table>

**Table 2.1** Example of the check-list format

The Likert format should be applied wherever it is possible (Nenadál, 2004). Respondents have a multiscale possibility of evaluation. One extreme state expresses absolutely positive perception and another extreme state expresses absolutely negative perception (see Tab. 2.2)

<table>
<thead>
<tr>
<th>I am satisfied with the behavior of the service.</th>
<th>strongly agree</th>
<th>agree</th>
<th>neutral</th>
<th>disagree</th>
<th>strongly disagree</th>
</tr>
</thead>
</table>

**Table 2.2** Example of the Likert format

The total size of the customer base is usually too big to measure satisfaction of everyone. It is therefore desirable to create a representative sample. If it is possible, statistical approaches for sampling should be used, because only these ones guarantee the objectivity of the customer satisfaction measurement (Řezanková, 2017).

The next question is selection of appropriate method of data gathering, which should be selected based on the size of the customer sample, financial and human resources provided by the top management, technical capabilities or requirements on the questionnaire response rates. There are the following possibilities of data gathering:

- using telephone,
- using regular mail or email,
- by personal interviews with customers,
- by the “pro-forma” method, when customers can react immediately after provided services.

The evaluation of questionnaires is possible using support of standard statistical tools. There is an evident trend among companies to quantify customer satisfaction by various indices. Usually it is expected that the Likert scale is used (Nenadál, 2004). An advantage of usage of indices is the possibility to evaluate them in time and therefore to spot the trend in the development of customer satisfaction.
The company’s executives should utilize the **results of these measurements as inputs for continuous improvement processes**. Any negative trend has to be perceived as a warning before potential economic problems and also as an input for acceptance of improvement projects. It has been proved that the increasing satisfaction of customers is the most important factor for their loyalty and therefore for the long-term economic success of the company (see Fig. 2.2).

### 2.2 Customer loyalty and its measurement

Nenadál (2004) states that it is not sufficient to focus only on measuring *customer satisfaction*, but also on *customer loyalty*. Customer loyalty is defined by EFQM as “*a way of future customer behavior, demonstrated mainly by repeated purchases and positive references*”.

The first option for monitoring and measuring customer loyalty is through **measurement of future customer intentions**. This approach is based on following type of questions:

- What is the probability that you will repeat your purchase?
- Would you recommend our products/services to your partners and friends?
- Do you think that the volume of our sales will be increasing or decreasing?

The second approach to measure loyalty is **measurement of the so-called effective loyalty**. Fig. 2.2 reflects some connections between customer loyalty and its effect on economic performance of organizations.
It can be difficult to measure customer loyalty by economic indicators such as profit or change in cash flow, because they can be influenced by other market factors (Nenadál, 2008). For this purpose it is suitable to use indirect indicators of loyalty, such as:

- **Indicator of customer retention**
  \[
  CR = \frac{C_E}{C_B}
  \]
  \[ (2.2) \]
  where \( C_B \) denotes number of customers at the beginning of the year and \( C_E \) denotes number of customers at the end of the year. It is important to note that newly acquired customers must not be counted into the number of customers at the end of the year.

- **Indicator of long-term relationships**
  \[
  LTR = \frac{T}{C}
  \]
  \[ (2.3) \]
  where \( T \) is total time of relationships with customers counted as a sum of all concluded contracts and \( C \) is total number of customers with concluded contracts.

---

**Figure 2.2** Customer loyalty model (Nenadál, 2004, p. 138)
The last approach frequently used to measure customer loyalty is measurement according to acquired and lost customers. In some organizations it is necessary to monitor the counts of acquired and lost customers. It must be noted that rather than monitoring of the number of lost customers, it is more important to analyze their behavior and reasons for leaving the company.

2.3 Relationship marketing

It is important to create processes supporting long-term loyalty of customers (Christopher, 2000). The aim of relationship marketing is to build and develop long-term business relationships advantageous for both businesses and customers. The specifics of the relationship marketing are the following:

- orientation to stable customers,
- regular contact with customers,
- emphasis on creation of value perceived by customers,
- long-term planning,
- high emphasis on quality of service to customers,
- effort to meet all customer expectations,
- quality products and services, which are a matter of all employees.

The philosophy of relationship marketing is based on “quality” of market share, not on its absolute size. Its aim is to minimize loss of customers and to build a long-term partnership with customers.

An average lifetime of customer (in years) can be computed by this equation:

\[
\text{Average\_customer\_lifetime} = \frac{1}{1 - \text{customer\_stability}}.
\]  

(2.4)

There is a positive relationship between customer stability and average customer lifetime. When the customer stability is annually for example 90 % (a company loses 10 % of its customers every year), then the average customer lifetime is 10 years. Fig. 2.3 captures the relationship between customer stability and his/her average lifetime (Christopher, 2000, p. 88).
Figure 2.3 Relationship between customer lifetime and customer stability

If stability of customers is the main factor influencing long-term profitability and if the quality of the relationships with customers is the key driver of customer stability, what are the factors which have an impact on the quality of relationships with customers? We could find a lot of factors but probably the most important one is satisfaction with the level of provided services (see Fig. 2.4).

```
<table>
<thead>
<tr>
<th>Long-term profitability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability of customers</td>
</tr>
<tr>
<td>Quality of the relationship</td>
</tr>
<tr>
<td>Level of services provided to customers</td>
</tr>
</tbody>
</table>
```

Figure 2.4 Drivers of long-term profitability (Christopher, 2000, p. 66)
Organizations, which aim to be successful, must know what the customer requirements for a services are. Once the organizations have the information, then they can adjust its strategies to fulfill or exceed customer expectations.

2.4 Preventing churn in telecommunications

The term “customer churn” is used to indicate the customers who are about to leave for a new competitor, or end their subscriptions. Customer churn or customer attrition has become an important issue for organizations, particularly in subscription based businesses, where customers have contracts for a specific time range, typically for one year. Customer churn in telecommunications industry is really a hot topic, because the market is saturated and therefore it is difficult to attract new customers and moreover for customers it is relatively easy to switch to another company (Canale and Lunardon, 2014). It is generally accepted that acquiring a new customer is five to six times more costly than retaining the existing one. For telecommunication operators it is preferable to invest into existing customers and renew their trust rather than to attract new ones for who a higher churn rate is typical.

Churn management is the process of identifying customers who are valuable for company and are likely to churn, and executing proactive steps to retain them. Lazarov and Capota (2007) define the measurement of the number of customers moving out during a specific period of time as a Churn rate (see Fig. 2.5).

![Figure 2.5 Calculation of churn rate](image)

It is necessary to distinguish between two types of churn (Lazarov and Capota, 2007):

- **Voluntary churn** – the customer decides to cancel his contract and to switch to another provider. For companies it is necessary to know the reason of churn before applying the right retention strategy. The reasons for churn may include dissatisfaction with the quality of the service, too high costs, not competitive price plans, no rewards for customer loyalty, bad support, long time of problem solutions, privacy concerns etc.
• **Involuntary churn** – the company discontinues the contract itself, e.g. because of fraud or non-payment of invoices.

For a company involuntary churn can be perceived as healthy, because a company loses only non-profitable or problem causing customers. The type of churn we will be interested in is the voluntary churn.

Basically, in the churn management process, responsible business representatives should take care of these three dimensions:

• **WHO**, it means identifying those customers, who are about to churn,
• **WHEN** will the customer churn (in a week, month or year?),
• **WHY**, what is the reason of the customer churn.

Predictive analytics models applied in this thesis are able to answer all of these questions.

Dahia and Talwar (2015) emphasize that machine learning models work well if there is enough time spent to prepare meaningful features. Thus, having the right features is usually the most important thing for predictive models to be applicable. With the still decreasing costs of data storage, telecommunications companies have access to various data sources which can be beneficial for analysis of customer churn. Traditional transactional data stored in databases can be combined with unstructured data such as complaints or feedback scraped from social networks or call recordings. Therefore, it is necessary to invest time into feature engineering, because well prepared features can also help us with identification of reasons of churn.

Typically, there are four types of customer data available in telecommunications:

• **Customer demographics** – contain the information like name, address, age, gender, income, level of education etc.
• **Customer behavior** – these data capture which services and how often a customer use. Among typical information belongs service plan, contract information, credit score or payment history.
• **Customer perceptions** – can be measured with customer surveys and include data such as overall satisfaction, quality of service, satisfaction with problem handling, reputation of the company etc.
• **Call details** – include information such as average call length or number of calls to other networks.

These data are used to create predictive models with the use of algorithms such as logistic regression, decision trees, random forests or neural networks. The frequency of methods and models used for churn prediction were in detail described in the section Current state of examination.

It is also important to mention the financial part of the churn management process. Proper churn management can undoubtedly save a huge amount of money to company. Van Den Poel and Lariviere (2004) summarize the economic value of customer retention as follows:

• satisfied customers can bring new ones,
• long-term customers tend to buy more,
• long-term customers are less sensitive to competitors marketing offers,
• company can focus on satisfying existing customer’s needs,
• lost customers can share negative experience which can result in the negative image of the company.

It is without any doubt that every company in any business area, including telecommunications, must pay a huge attention to the relationships with their existing and potential customers.
3 Description of applied methods and approaches

The main goal of the proposed churn prediction system is to classify the existing customers into those who are likely to leave the company and into those who are not. The creation of the system follows traditional CRISP DM methodology which consists of business understanding, data understanding, data preparation, modelling, evaluation and deployment. The CRISP DM methodology is described at the beginning of this part of the thesis.

The dependent variable is categorical with 2 classes (churn/non-churn), therefore classification methods are used in this dissertation thesis. The first used method is logistic regression which represents the class of linear classification models. The second method used is decision tree representing non-linear classification models. Big advantage of these models is that the resulting models are quite easy to interpret, so we do not obtain only a probability of churn, but also the main factors and quantification of their strength and direction. A potential disadvantage of these models is their usually lower predictive performance in comparison with more sophisticated models. For this purpose the third model used in the thesis is random forest model, an ensemble of decision trees created using bagging procedure.

Because the data set consists of approx. 50,000 customers, it might be difficult to classify such a huge amount of customers. In the second part of the application part of this thesis, the customers are at first clustered and the classification models are estimated for individual clusters with the goal of improvement of predictive performance. Because of the fact that there are mixed types of variables in the data set, Gower coefficient is used to compute the similarity matrix between individual customers. Then algorithms such as hierarchical clustering and partitioning around medoids are used to divide the data into homogeneous groups.

3.1 Data mining methodology – CRISP DM

A methodology of data mining process is given by the Cross Industry Standard Process for Data Mining (Shearer, 2000), illustrated in Figure 3.1 (Fawcett and Provost, 2013).
The first step of the data mining process is **business understanding**. This step consists of understanding the project objectives, gathering business requirements, translating them into data mining problem and creating a plan to achieve the objectives.

**Data understanding** is the second step of data mining process. This phase starts with data collection and is followed by getting familiar with the data – the first insights or potential data quality problems are discovered.

**Data preparation** phase is a logical follower of data understanding. In this phase the raw data extracted from various places and systems are transformed into the form necessary for modelling tools. Data transformation contains for example data merging, creating new features, data cleaning – deleting/replacing outliers or feature selection.

In the **modeling** phase, prepared data are fed into various modeling algorithms. The parameters of these algorithms are usually optimized using cross validation. As some methods have specific requirements for the form of input data, it is often necessary to go back to the data preparation phase.
The created models have to be thoroughly evaluated in the evaluation phase of the project. It is important to assess all steps leading to creation of the model and also the resulting model has to achieve the business goals defined in the business understanding phase. Because data mining process is iterative, it is possible and actually it often happens that we have to go back to the business understanding phase to modify some steps before model deployment.

The last phase, deployment, is really important, because until the customer cannot use the model results, it has no business value. The clearest cases of deployment involve implementing a predictive model in some information system or business process. It is also common that the model predictions are visualized in reporting tools such as Qlik Sense, Tableau or Power BI.

Below we can see phases of data mining process accompanied by generic tasks and outputs (SPSS, 2000, p. 12).

- Business understanding
  - Determine Business Objectives
    - Background
    - Business Objectives
    - Business Success Criteria
  - Assess Situation
    - Inventory of Resources
    - Requirements, Assumptions and Constraints
    - Risks and Contingencies
    - Terminology
    - Costs and Benefits
  - Determine Data Mining Goals
    - Data Mining Goals
    - Data Mining Success Criteria
  - Produce Project Plan
    - Project Plan
    - Initial Assessment of Tools and Techniques
- Data Understanding
  - Collect Initial Data
- Initial Data Collection Report
  - Describe Data
    - Data Description Report
  - Explore Data
    - Data Exploration Report
  - Verify Data Quality
    - Data Quality Report

- Data Preparation
  - Select Data
    - Rationale for Inclusion/Exclusion
  - Clean Data
    - Data Cleaning Report
  - Construct Data
    - Derived Attributes
    - Generated Records
  - Integrate Data
    - Merged Data
  - Format Data
    - Reformatted Data

- Modeling
  - Select Modeling Techniques
    - Modeling Technique
    - Modeling Assumptions
  - Generate Test Design
    - Test Design
  - Build Model
    - Parameter Settings
    - Models
    - Models Descriptions
  - Assess Model
    - Model Assessment
- Revised Parameter Settings

**Evaluation**
- Evaluate Results
  - Assessment of Data Mining Results
  - Business Success Criteria
  - Approved Models
- Review Process
  - Review of Process
- Determine Next Steps
  - List of Possible Actions
  - Decision

**Deployment**
- Plan Deployment
  - Deployment Plan
- Plan Monitoring and Maintenance
  - Monitoring and Maintenance Plan
- Produce Final Report
  - Final Report
  - Final Presentation
- Review Project
  - Experience
  - Documentation

### 3.2 Cluster analysis

Cluster analysis belongs to the group called *unsupervised learning*, where we can assign methods used to discover unknown relationships in data. According to another categorization, it is a part of *multivariate statistics*. The objective of any clustering algorithm is to divide the observations of a data set into distinct groups, so the observations within each group are similar to each other, while observations in different groups are different from each other (James, 2013). Cluster analysis is useful especially in those areas, where objects tend to group (Meloun and Militký, 2006).
Because clustering is popular in many fields such as biology, medicine, business and marketing, computer science or social science, there are many clustering methods from which it is possible to select (Manďák, 2016b). The two best known and the most frequently used are *K-means clustering*, where we seek to partition the observations into predefined number of clusters, and *hierarchical clustering*, where we do not know in advance how many clusters we want, but we obtain a tree-like visual representation of the observations, called a *dendrogram*. “An advantage of *dendrogram* is that it allows us to view, at once, the clustering obtained for each possible number of clusters, from 1 to n” (James, 2013, pp. 386).

Before performing a cluster analysis, it is necessary to check some of its assumptions. “Traditional assumptions such as normality, homoscedasticity or linearity applied in other multivariate statistical methods are not so important in cluster analysis” (Meloun, Militký and Hill, 2005, pp. 316). There are three important assumptions of cluster analysis – the absence of multicollinearity, absence of outliers and comparability of units of variables.

By multicollinearity we mean a high correlation between two or more variables (Wooldridge, 2013). Hančlová (2012) proposes that variable causing multicollinearity should be removed or we should extend the data set. There are a few ways how to test multicollinearity – the best known one, which is applicable to numeric variables, is *Pearson’s correlation coefficient* (3.1)

\[
    r = \frac{\sum_{i=1}^{n}(x_i)(\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(\bar{y}-\bar{y})^2}}
\]  
(3.1)

The closer the coefficient to 0, the weaker is the relationship between two variables and the closer the coefficient in absolute value to 1, the stronger is the relationship between two variables.

Another possibility to test multicollinearity is *Variance Inflation Factor (VIF)*, see formula 3.2:

\[
    VIF_i = \frac{1}{1-R_i^2}
\]  
(3.2)

The coefficient of determination \(R_i^2\) is from the regression \(X_i = f(others\ X_j), x_i \neq x_j\). The value of *VIF* higher than 10 signalizes a strong multicollinearity (Wooldridge, 2013). In our case when the variables are of mixed types (numerical, categorical, binary), *generalized VIF* will be appropriate to measure multicollinearity among variables.
There are a lot of methods used to detect extreme values in data sets, such as distance based methods, density based methods or techniques based on clustering. One of the simplest and most frequently used methods is box plot diagram, where the first quartile, the second quartile and the inter-quartile range are computed. We should be careful in removing outliers since it can collapse the clustering solution (Meloun, Militký and Hill, 2005).

Usually the input data for clustering contains variables with different measurement units. Therefore, it is necessary to unify them into dimensionless number to make variables comparable. The standardization of variables is done using transformation into Z-score:

\[ u_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \]  

where \( u_{ij} \) is standardized variable for customer \( i \) and variable \( j \), \( x_{ij} \) is original value, \( \bar{x}_j \) is arithmetic mean of variable \( j \) and \( s_j \) is standard deviation of variable \( j \).

After preparation and transformation of the data set into the required form, when it fulfills necessary requirements, we can proceed to clustering. Generally, there are two main approaches to clustering – hierarchical clustering and partitional clustering. Hierarchical clustering can be further divided into agglomerative and divisive clustering (Meloun, Militký and Hill, 2005).

Among the most frequent methods of hierarchical clustering belong these methods:

- single linkage,
- complete linkage,
- average linkage,
- centroid.

Among the best-known methods of partitional clustering belong methods such as:

- k-means,
- k-medoids,
- fuzzy k-means.

Because in our data set the variables are of a mixed types (numerical, categorical, binary), we will further describe only selected methods suitable for this situation – Gower coefficient of similarity, partitioning around medoids algorithm and hierarchical clustering method.
3.2.1 Gower coefficient of similarity

The first step necessary for cluster analysis is the computation of similarity or dissimilarity matrix. There are three possibilities for types of data in Gower coefficient – dichotomous, qualitative and quantitative. By the term dichotomous Gower (1971) means characters that are either present or absent and whose absence in both of a pair of individuals is not taken as a match; when both levels of a two-level qualitative variable should be taken as a match, the levels are termed *alternatives*. Qualitative variables can take many levels, but in comparison with quantitative characters they do not form an ordered set.

Let’s describe how to compute similarity matrix for mixed data types. Two individuals $i$ and $j$ are compared on a variable $k$ and a score $s_{ijk}$ equal to zero is assigned when $i$ and $j$ are considered different, or a positive fraction or unity when they have some degree of similarity (Gower, 1971). Sometimes, in the case of dichotomous variables, it is not possible to make a comparison because of missing information or because of non-existent character in both $i$ and $j$. When it is possible to make a comparison, a quantity $\delta_{ijk}$ is equal to 1, and 0 otherwise. The similarity between $i$ and $j$ is calculated as the average score taken over all possible comparisons:

$$S_{ij} = \frac{\sum_{k=1}^{v} s_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}}.$$  \hspace{1cm} (3.4)

According to Gower (1971), there are three possible situations:

1) When $\delta_{ijk} = 0$ for all variables, $S_{ij}$ is undefined.

2) When all comparisons are possible, $\sum_{k=1}^{v} \delta_{ijk}$ is equal to the total number of variables $v$.

3) Otherwise $\delta_{ijk}$ is equal to the number of variables over which the comparison is made.

The scores $S_{ijk}$ are assigned in the following manner (Gower, 1971):

a) The presence or absence of the character for *dichotomous variables* is denoted by + and -, respectively. There are four different combinations of its value, see Table 3.1.

b) For *qualitative variables* $s_{ijk} = 1$ if the values of the variable $k$ are the same for the two individuals $i$ and $j$, and $s_{ijk} = 0$ if they differ.

c) For *quantitative variables* $s_{ijk} = 1 - |x_i - x_j|/R_k$, where $R_k$ is the range of variable $k$. 

34
The total similarity $S_{ij}$ between two individuals $i$ and $j$ ranges between 0 and 1; when the two individuals differ in no variables, then $S_{ij} = 1$ whereas $S_{ij} = 0$ when they differ maximally in all variables (see Tab. 3.1).

<table>
<thead>
<tr>
<th>Individual $i$</th>
<th>Values of character $k$</th>
<th>$j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>$s_{ijk}$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\delta_{ijk}$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 3.1** Scores and validity of dichotomous variables comparisons (Gower, 1971)

### 3.2.2 Hierarchical clustering

Hierarchical clustering methods are based on successive merging of objects and smaller clusters into another bigger clusters (Meloun and Militký, 2006). An advantage of these methods, in comparison with for example *k*-means clustering, is that there is no need to specify an exact number of clusters and that it results in an attractive tree-based representation of the observations, called a dendrogram (James et al., 2013). The first step of hierarchical clustering is to compute a base distance matrix between objects. The process in the case of agglomerative clustering is as follows:

1) Two objects with the smallest distance are merged into the first cluster.
2) A new distance matrix is calculated where this cluster is taken into account as a whole.
3) This process is repeated until all objects create one big cluster or until we get the predefined number of clusters.

The divisive clustering is reverse – at the beginning all objects are a part of one big cluster, by successive division we get a system of clusters until we end up in the stage of individual objects.

This algorithm is quite simple, but one issue has not been addressed. We have a concept of the dissimilarity between the pairs of observations. The question is how to define the dissimilarity between two clusters if one or both of the clusters contain/s multiple observations. The concept of dissimilarity between a pair of observations needs to be extended to a pair of groups of observations (James et al., 2013, pp. 394). This extension is achieved by developing the notion of linkage which
defines the dissimilarity between two groups of observations. There are four most common types of linkage – complete, average, single, and centroid, see Tab. 3.2 (James et al., 2013).

<table>
<thead>
<tr>
<th>Linkage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.</td>
</tr>
<tr>
<td>Single</td>
<td>Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.</td>
</tr>
<tr>
<td>Average</td>
<td>Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.</td>
</tr>
<tr>
<td>Centroid</td>
<td>Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B.</td>
</tr>
</tbody>
</table>

Table 3.2 Linkage types of clustering (James et al., 2013)

The average and complete linkages are generally preferred by statisticians to single linkage, as they tend to yield more balanced dendrograms (see Fig. 3.2).
3.2.3 Partitioning around medoids

Partitioning around medoids (PAM) was one of the first k-medoids algorithms (Han and Kamber, 2001). A big advantage of this method, in comparison with k-means algorithm, is that it is less sensitive to outliers. The reference point of the cluster, called medoid, is actual object, not the mean value of the objects like in k-means. Each remaining object is in the cluster with the representative object to which it is the most similar. The partitioning around medoids algorithm is based on minimization of the sum of dissimilarities between each object and its corresponding medoid (Han and Kamber, 2001):

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|,$$

(3.5)
where $E$ is the sum of the absolute error for all objects in the data set; $p$ is the point representing a given object in cluster $C_j$ and $o_j$ is the medoid of $C_j$. There are two things necessary to be defined for the PAM algorithm, the number of clusters $k$ and a data set containing $n$ objects. The output is a set of $k$ clusters. The process of the PAM algorithm is as follows (Han and Kamber, 2001):

1. arbitrarily choose $k$ objects in D as the initial representative objects (medoids);
2. repeat
3. assign each remaining object to the cluster with the nearest representative object;
4. randomly select a nonrepresentative object, $o_{\text{random}}$;
5. compute the total cost, $S$, of swapping representative object, $o_j$, with $o_{\text{random}}$;
6. if $S < 0$ then swap $o_j$ with $o_{\text{random}}$ to form the new set of $k$ representative objects;
7. until no change;

As it was mentioned at the beginning, the $k$-medoids method is more robust to the presence of noise and outliers because medoid is less influenced by them than mean. However, its processing time is more costly. In both methods, the analyst need to specify the number of clusters $k$ before run of the algorithm.

**3.2.4 Determining optimal number of clusters**

Oftentimes we want to divide a dataset into smaller homogeneous groups and from prior knowledge it is not apparent how many groups should be there. Unlike in hierarchical clustering, for a certain class of clustering algorithms such as $k$-means or $k$-medoids there is parameter $k$ that specifies into how many clusters we want to partition the data set. The correct number of clusters is often ambiguous and increasing number of clusters $k$ without penalty will always reduce the heterogeneity in particular clusters. There are several options how to select a number of clusters, e.g. the elbow method based on within clusters sum of squares, Calinsky criterion or Silhouette criterion (Zumel and Mount, 2014). It is also possible to utilize visualization methods like Stochastic Neighbor Embedding ($t$-SNE) algorithm. The Silhouette criterion is further described in detail because the implementation of PAM algorithm in R contains Silhouette criterion and because of recommendations of my colleagues from Finland.

Silhouette criterion is a method proposed by Rousseeuw (1987) to assess the validity of clustering solution. The silhouette value measures how similar an object is to its own cluster compared to other clusters. The range of Silhouette is from $-1$ to $+1$. Values close to 1 indicate that
the object is well matched to its own cluster and poorly matched to the neighboring clusters. The clustering is appropriate when majority of objects have a high value of Silhouette. The average Silhouette value across all objects is usually calculated and used to select the number of clusters. The Silhouette can be computed with any distance metric, Euclidean, Manhattan or Gower.

The value of Silhouette index is computed according to the following formula:

\[ s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \]  \hspace{1cm} (3.6)

where \( a(i) \) is the average distance between object \( i \) and all other data within the same cluster, \( b(i) \) is the lowest average distance of object \( i \) to all points in any other cluster, of which \( i \) is not a member.

### 3.3 Classification models

As mentioned at the beginning of the third part of the thesis, three classification methods are used for predicting customer churn – logistic regression, decision trees and random forests. It is important to compare their advantages and disadvantages. A big advantage of logistic regression and decision tree models is that the resulting models are easy to interpret and they are not time consuming when it comes to model estimation time. In logistic regression we can even determine the direction of the impact of an individual variable (positive, negative) and we can also compute the effect of one unit increase of independent variable to probability of churn. Because decision trees are essentially a set of IF-THEN rules, they can be easily implemented into information systems. The disadvantage of both logistic regression and decision tree is their generally lower predictive performance in comparison with more complex models. The third used model – random forest model is known by its generally high (and in the most cases the highest) predictive performance. The disadvantage of this model is that it is a black box, it consists of many decision trees and it is difficult to understand them. The random forests model also need a much longer time to train them.

#### 3.3.1 Logistic regression

Logistic regression is a classification method used in many fields, such as biomedical research, business and finance, criminology, ecology, engineering, health policy or linguistics. Logistic regression is a member of the class of models called generalized linear models (Zumel and Mount,
The aim of generalized linear models for a binary dependent variable and linear regression models for a continuous variable is to estimate a regression equation that relates the expected value of the dependent variable $y$ to one or more predictor variables, denoted by $x$ (Heeringa et al., 2010). In linear regression the expected value of $y$ is the conditional mean of $y$ given a vector of covariates $x$, and it is estimated by an equation that is linear in the regression parameters:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$  (3.7)

When $y$ is a binary variable with possible values 0 and 1, $E(y|x) = \pi(x)$ is the conditional probability that $y = 1$ given the covariate vector $x$.

A naïve approach is to model $\pi(x)$ as a linear function of $x$ (Heeringa et al., 2010). Fig. 3.3 shows a simple example of comparison of linear and logistic regression for binary outcome. In this case churn is predicted based on year of birth. As it is clear from the Fig. 3.3, linear regression does not capture the relationship between $y$ and $\pi(x)$ and moreover it may produce predictions that are outside the permissible range 0-1. The alternative would be a nonlinear function that yields a regression model that is linear in the coefficients and it is possible to transform the resulting predicted values to the range 0-1. These functions are called in the terminology of generalized linear models link functions (Heeringa et al., 2010). The two most common link functions used to model binary survey variables are the logit and the probit. Because logit link function will be used to model the churn in this thesis, only the logistic regression is further described. As mentioned above, the predicted values of a logistic regression model are between 0 and 1 (see Fig. 3.3).

![Figure 3.3 Comparison of linear and logistic regression](image-url)
We can express logistic regression by logistic function (James et al., 2013):

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}},
\]

which can be further transformed into form:

\[
\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_1 x_1 + \cdots + \beta_p x_p}.
\]

The expression \(\frac{\pi(x)}{1 - \pi(x)}\) is called the odds and can take on any value between 0 and \(\infty\). Values close to 0 indicate very low and values close to \(\infty\) indicate a very high probability. To make the right-hand side linear, we can modify equation (3.9) to the form:

\[
\log \left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.
\]

The left-hand side is called the log-odds or logit and can take values from the interval \((-\infty; \infty)\).

The unknown regression coefficients \(\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)\) in equation (3.10) have to be estimated based on training data. The method used for estimation is called maximum likelihood (James et al., 2013). The idea behind the method is that we are trying to find the estimates for coefficients \(\beta\) so that the predicted probability \(\pi(x)\) of our variable of interest corresponds as closely as possible to the real values. For simplicity, let’s assume a model only with constant and one explanatory variable. A mathematical equation representing the idea can be expressed by likelihood function (James et al., 2013):

\[
l(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}.
\]

The estimates of \(\boldsymbol{\beta}\) coefficients are chosen to maximize the likelihood function (3.11). Mathematically it is easier to work with the log of equation (3.11). The log likelihood function is defined as:

\[
L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i)\ln[1 - \pi(x_i)]\}.
\]

To find the value of \(\boldsymbol{\beta}\) that maximizes \(L(\boldsymbol{\beta})\) we differentiate \(L(\boldsymbol{\beta})\) with respect to \(\beta_0\) and \(\beta_1\) and set the resulting expressions equal to zero (Hosmer and Lemeshow, 2000, p. 9). The form of the so-called likelihood equations is as follows:

\[
\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0,
\]
The value of $\mathbf{\beta}$ given by the solution of equations is called the maximum likelihood estimate.

The usual practice after estimation of the model coefficients is to assess the significance of explanatory variables (Hosmer and Lemeshow, 2000). One approach to testing the significance of variables is based on the comparison of predictive accuracy of the model with and without a given variable. It is clear that if the predictions of the model contain a given variable, they are more accurate than the predictions of the model without that variable, this variable is significant. The difference between the observed and predicted values is meant by a predictive accuracy.

A likelihood ratio test is frequently used to evaluate the statistical significance of one or more logistic parameters (Heeringa et al., 2010). Under the null hypotheses $H_0: \beta_j = 0$ (single parameter) or $H_0: \mathbf{\beta}_q = \mathbf{0}$ (with $q$ parameters), the test statistic $G$ follows a chi-square distribution with either 1 (for a single parameter) or $q$ degrees of freedom:

$$G = -2 \ln \left[ \frac{L(\mathbf{\hat{\beta}}_{MLE})_{reduced}}{L(\mathbf{\hat{\beta}}_{MLE})_{full}} \right],$$

(3.15)

where $L(\mathbf{\hat{\beta}}_{MLE})$ is the likelihood under the model evaluated at the maximum likelihood estimates of $\mathbf{\beta}$. The reduced model is the model excluding the $q$ regression parameters to be tested, while the full model is the model including $q$ regression parameters. Both models should be fitted using exactly the same set of observations for this test to be valid (Heeringa et al., 2010, p. 241).

Another test, which can be used to test the statistical significance of the coefficients $\mathbf{\beta}$ in the model is called Wald test. Wald test calculates a $Z$ statistic (3.16), which is for $i$-th variable computed as:

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

(3.16)

This $z$ value is then squared yielding a Wald statistic with a chi-square distribution. However, several authors have identified problems with the use of the Wald statistic. Menard (1995) warns that standard error is inflated for large coefficients what leads to lowering the Wald statistic (chi-square) value. Agresti (1996) states that the likelihood-ratio test is more reliable for small sample sizes than the Wald test.
An important step after the model building is assessing the fit of the model. It can be done by approaches explained in the Chapter 3.4, but there are also some statistics specific for logistic regression, e.g. the Hosmer-Lemeshow test. The Hosmer-Lemeshow test is a goodness of fit test, which tells how well the data fits the model. Specifically, it calculates if the observed event rates match the expected event rates in a population subgroups. Data are first regrouped by ordering the predicted probabilities and forming \( g \) number of groups. The Hosmer-Lemeshow test statistic (3.17) is calculated with the following formula:

\[
G_{HL}^2 = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j(1-E_j/n_j)}
\]

where \( n_j \) = number of observations in the \( j^{th} \) group, \( O_j \) = number of observed cases in the \( j^{th} \) group and \( E_j \) = number of expected cases in the \( j^{th} \) group. This statistics follows a chi-squared distribution with \((g - 1)\) degrees of freedom.

There have also been proposed several pseudo-\( R^2 \) measures, but these measures tend to be incorrectly confused with \( R^2 \) values in linear regression. Heeringa et al. (2010) and Hosmer and Lemeshow (2000) agreed that they may be used by an analyst to compare the fits of alternative models, but should not be cited as a measure of fit in scientific papers.

### 3.3.2 Decision trees

*Decision trees* utilize a tree-logic to make predictions and can be expressed as a set of IF-THEN rules (Ledolter, 2013). The tree is generally presented upside down, with the root at the top and the leaves at the bottom. Starting from the root, the tree splits from the single trunk into two or more branches. Each branch itself might further split into two or more branches. This continues until we reach a leaf, which is a node that is not further split. We refer to the split of a branch as a node of the tree. The root and leaves are also referred to as nodes (Williams, 2011). Simple example of decision tree created with the subset of observations and explanatory variables is in the Fig. 3.4. The first number in the square above the decision rule says, whether there are mostly churners (1) or non-churners (0) in the given region. For instance number 0 in the root node means, that in the entire dataset majority of customers are non-churners. The second two numbers represent proportion of churners and non-churners, e.g. in the root node we can see that 91 % of the customers stayed with the company and 9 % left the company. The last number shows percentage of customers in the given region from the whole customer base.
A great benefit of decision trees is that its structure is in a human-readable format (Lantz, 2013; Manďák, 2016a). This provides insight into how and why the model works or does not work well for a particular task. Decision trees are therefore appropriate for applications, where the classification mechanism need to be transparent, e.g. for legal reasons. The suitable use cases for applying decision trees include:

- credit scoring models, where the criteria causes acceptance/rejection of an applicant need to be clearly documented,
- marketing models such as customer satisfaction or customer churn, which will be shared with decision makers.

Recursive partitioning is a method to grow a classification tree (Grus, 2015). This approach is known also as divide and conquer, because it splits the whole data set into subsets, which are then split again into smaller subsets, and so on and so forth, until some stopping criterion is reached. Reasons for stopping of growing the tree can be situations, when

- nearly all of the observations at the node have the same class,
- there are no remaining features to divide the data,
- the tree has grown to a predefined size limit.
The first question in building the decision tree is which variable to split upon. We look for variable which splits the data the way that the resulting subsets contain observations primarily of a single class. The degree to which a subset of observations contains a single class is known as purity (Lantz, 2013). There are various measures of purity such as Entropy, Cross-Entropy or Gini index. Because the CART algorithm based on Gini index is used in the thesis, we further describe only Gini index.

Mathematically, Gini index is defined as follows:

\[
Gini(S) = \sum_{i=1}^{c} p_i (1 - p_i),
\]

where \( S \) denotes a given subset of data, \( c \) denotes number of class levels and \( p_i \) refers to the proportion of values falling into class level \( i \). In the case of two classes, if we know that the proportion of observations of one class is \( x \), than the proportion of the second class is \( (1-x) \). We can then plot possible values of Gini index, see Fig. 3.5.

![Gini index of node purity](image)

**Figure 3.5** Gini index of node purity (Lantz, 2013, p. 126)

The peak in the figure (\( Gini = 0.5 \)) illustrates a situation, when the proportion of classes is exactly 50 % - 50 %. As one class dominates the other, the Gini index decreases to zero.
*Gini gain* is a measure for determining the best variable to split the data set upon (Lantz, 2013). It calculates the change in *Gini index* that would result from a split on each possible feature. The *Gini gain* for a variable $F$ is computed as the difference between the *Gini index* in the subset before the split ($S_1$) and the subset resulting from the split ($S_2$):

$$
\text{GiniGain}(F) = \text{Gini}(S_1) - \text{Gini}(S_2)
$$

(3.19)

Because the data are after split divided into more than one partition, the $\text{Gini}(S_2)$ needs to consider the total *Gini impurity* across all of the partitions. It is done by weighting the *Gini index* of a given subset. The weight is simply the proportion of observations falling into the subset:

$$
\text{Gini}(S) = \sum_{i=1}^{n} w_i \text{Gini}(P_i)
$$

(3.20)

There is a simple rule for *Gini gain* for a given feature is: the higher – the better.

Decision trees have the following advantages over the more classical approaches like linear or logistic regression (Zaki and Meira, 2014):

- trees are easy to explain to people, even easier than linear regression,
- trees can be easily displayed graphically,
- decision trees can handle qualitative predictors without the need to create dummy variables,
- there are no distributional assumptions or preprocessing necessity,
- most implementations handle missing data,
- the method is robust to redundant and nonlinear data,
- the algorithm excludes unimportant features.

On the other hand, decision trees have also some weaknesses:

- generally lower level of predictive accuracy compared to advanced classification approaches like random forests,
- tendency to overfit, so there is a need to prune the tree,
- high training variance – samples drawn from the same population can produce trees with different structures and different prediction accuracy,
- they are often biased towards splits on features having a large number of levels,
- can have a trouble with modeling some relationships due to axis-parallel splits,
- large trees can be difficult to interpret.
3.3.3 Random forests

One of the common weaknesses of simpler models is the training variance. Training variance means that small changes in the training set result in models that make substantially different predictions (Zumel and Mount, 2014). Decision trees can exhibit this effect. Reducing training variance and sensitivity to overfitting can be done by technique called bagging or models like random forest. In this chapter, general terms like bagging or out-of-bag error estimation are explained and then random forest algorithm is described.

Bootstrap aggregation, or bagging, is a general-purpose procedure for reduction the variance of machine learning method (James et al., 2013). Model created using bagging is also less likely to overfit the data. A natural way to reduce the variance and to increase the prediction accuracy is to take many training sets from the population, build a separate model using each training set and average the resulting predictions. As we generally do not have access to multiple training sets, we can bootstrap. Bootstrap means taking repeated random samples with replacement from the training data set. For classification and qualitative outcome, the simplest possible solution is to take a majority vote. For a given test observation, we record the class predicted by each of the model and the overall prediction is the most commonly occurring class among these predictions.

Out-of-bag error estimation is a technique used by random forest model to create multiple training and test sets. Each bagged tree uses two thirds of the observations for model training. The remaining one-third of the observations are called out-of-bag (OOB) observations. We then predict the value for the \( i \)-th observation using each of the trees where this observation was OOB. To obtain a final prediction, as stated above, we take a majority vote.

Random forests enhance the bagging procedure by randomizing the set of variables that each tree is allowed to use. This should lead to de-correlating of the trees. The process of building a random forest model is as follows (Zumel and Mount, 2014):

1) Draw a bootstrapped sample from the training data.
2) For each sample, create a decision tree, and at each node of the tree:
   1) randomly select a subset of \( mtry \) variables from the \( p \) total features (typically, the number of candidate predictors is \( m \approx \sqrt{p} \)),
   2) pick the best variable and the best split from that set of \( mtry \) variables,
   3) continue until the tree is fully grown.
The reason for this limitation of predictors is clear. Suppose that there is one very strong predictor in the data set and number of weaker predictors. If the algorithm could choose from all of the predictors, then all of the decision trees would be similar and the predictions would be correlated. Graphical representation of random forest algorithm can be seen in the Fig. 3.6.

![Random Forest Diagram](image)

**Figure 3.6** Random forest - an ensemble of decision trees (James et al., 2013, p. 320)

In comparison with logistic regression or with single decision tree, the random forest model is difficult to interpret. One possibility to make the model clearer is to estimate the “importance” of the variable $v$. The values of variables are randomly permuted in the out-of-bag samples, and the corresponding decrease in the accuracy of each tree is estimated. If the average decrease over all of the trees is large, then the variable is considered important – it makes a big difference in predicting the outcome. If the average decrease is small, then the variable does not make much difference to the outcome (Hastie et. al, 2008). The knowledge of the most important variables is beneficial for data scientists as it helps them with variable reduction (for creating smaller, faster trees) and also for business representatives who can see what is driving the dependent variable.
The following Tab. 3.3 lists the strengths and weaknesses of random forest models (Lantz, 2013).

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>• An all-purpose model that performs well on most problems.</td>
<td>• Unlike a decision tree, the model is not easily interpretable.</td>
</tr>
<tr>
<td>• Can handle noisy or missing data; categorical or continuous features.</td>
<td>• May require some work to tune the model.</td>
</tr>
<tr>
<td>• Selects only the most important features.</td>
<td></td>
</tr>
<tr>
<td>• Can be used on the data with an extremely large number of features and examples.</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3 Strengths and weaknesses of random forests (Lantz, 2013)

3.4 Evaluating predictive performance

For the classification tasks, where the class which we want to predict is much less frequent, the most known performance metric – accuracy – is not sufficient. Two other metrics, hit rate (precision) and sensitivity (recall), are much more important. There are also another metrics, e.g. F-score or AUC (Area under ROC curve), which can be used to determine the classification power of various models.

Accuracy, hit rate or sensitivity can be computed from the information available in confusion matrix (Lantz, 2013). It is a table that categorizes predictions according to whether they match the actual value in the data (see Table 3.4). One dimension indicates the possible categories of predicted values, while the other dimension indicates the same for actual values. There are four categories in 2 x 2 confusion matrix:

• True Positive (TP): Correctly classified as the class of interest;
• True Negative (TN): Correctly classified as not the class of interest;
• False Positive (FP): Incorrectly classified as the class of interest;
• False Negative (FN): Incorrectly classified as not the class of interest.
Table 3.4 Confusion matrix (Lantz, 2013, p. 298)

As a logical following step, performance metrics, which can be computed from confusion matrices, are described (Lantz, 2013).

- **Accuracy** – the proportion of predictions, which were right. It is the sum of true positives and true negatives divided by the total number of observations:
  \[
  \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{3.21}
  \]

- **Error rate** – the proportion of incorrectly classified examples. It can be easily computed by one minus accuracy:
  \[
  \text{Error rate} = \frac{FP+FN}{TP+TN+FP+FN} = 1 - \text{accuracy}. \tag{3.22}
  \]

- **Hit rate** – the proportion of predictions of the class of interest which were right. In case of churn prediction, it is a proportion of customers where model predicted that they should churn, and they really churned. Formally it is ratio of the true positives and sum of the true positives and false positives:
  \[
  \text{Hit rate (precision)} = \frac{TP}{TP+FP}. \tag{3.23}
  \]

- **Sensitivity** - the proportion of correctly classified churners. Formally it is calculated as the ratio of the true positives and sum of the true positives and false negatives:
  \[
  \text{Sensitivity (recall)} = \frac{TP}{TP+FN}. \tag{3.24}
  \]

- **Specificity** - the proportion of correctly classified non-churners. Formally it is calculated as the ratio of the true negatives and sum of the true negatives and false positives:
  \[
  \text{Specificity} = \frac{TN}{TN+FP}. \tag{3.25}
  \]
- **F-score** - a measure of model performance that combines precision and recall using harmonic mean into a single number. It is known also as F1 score or F-measure.

\[
F-score = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}
\]  

(3.26)

Visualization of performance of machine learning models is often better than thinking about only a single pair of statistics such as precision and recall (Lantz, 2013). The pROC package in R provides easily used functions for creating visualizations of the performance statistics of classification models (Robin et al., 2011). For creation of visualization two vectors are necessary – the first one must contain the class values predicted and the second one must contain the estimated probability of the positive class. The **ROC curve** (Receiver Operating Characteristic) is commonly used to examine the tradeoff between the detection of true positives, while avoiding the false positives (Lantz, 2013). **ROC curves** are useful for comparing different classifiers, since they take into account all possible thresholds (James, 2013). A typical **ROC curve** is showed in the Fig. 3.7. On the vertical axis there is True Positive Rate (sensitivity) and on the horizontal axis there is False Positive Rate (1 – specificity). The **ROC curve** is plotted based on the values of sensitivity and 1 – specificity for different cut-off points. Three hypothetical classifiers are compared in the Fig. 3.7. A classifier with no predictive value detects true positives and false positives with exactly the same rate. The perfect classifier has a curve that passes through the point at 100 percent true positive rate and 0 percent false positive rate. This classifier detects all of the true positives before it incorrectly classifies any negative result. The test classifier, which falls somewhere between perfect and useless classifier, is typical for most real-world classifiers (Lantz, 2013).
The overall performance of a classifier summarized over all possible thresholds is given by the area under the (ROC) curve (AUC). AUC ranges from 0.5 (classifier with no predictive value) to 1 (a perfect classifier). A convention for interpreting AUC scores uses a system similar to academic letter grades (Lantz, 2013):

- 0.5 – 0.6 = F (no discrimination)
- 0.6 – 0.7 = D (poor)
- 0.7 – 0.8 = C (acceptable/fair)
- 0.8 – 0.9 = B (excellent/good)
- 0.9 – 1.0 = A (outstanding)

### 3.4.1 Estimating future performance

In order to estimate the future performance of predictive models, we have to compute performance statistics described above not based on the training set, but on the different data set, which was not used in the process of building a model and which is called test set. A classifier which assigns every training instance into the correct class would be the most probably unable to make predictions on data which it has never seen before. These approaches should test the ability of created models to generalize. For example the caret package in R offers a number of various option for this purpose.
Three methods which are further used in the thesis are described. The first one is \textit{training/testing} method, where the data are divided into training and testing data sets. The second one is the \textit{cross-validation}, because this approach is used to tune the hyper parameters of decision tree and random forest model. The third method is \textit{bootstrap sampling}, a technique used in the process of building a random forest model.

The procedure of partitioning dataset into training and testing dataset is called \textit{the holdout method} (Lantz, 2013). As it is visible in the Fig. 3.8 below, the \textit{training dataset} is randomly sampled to create the model, which is applied to \textit{testing dataset} for generating predictions. Typically two-thirds of observations are in the training dataset and the remaining one-third is in the testing dataset, but this proportion is not strictly defined and can be chosen by an analyst, e.g. based on amount of data available.

\begin{figure}
\centering
\includegraphics[width=0.7\textwidth]{fig3_8}
\caption{Division into training and testing data set}
\end{figure}

\textit{K-fold cross-validation} randomly divides the data into \(k\) completely separate random partitions called folds (Zummel and Mount, 2014). The number of folds \(k\) can be set to any number, the most common is to set this value to 5 or 10. For each of the \(k\) folds, a machine learning model is built on 90 percent of the data (blue cells in the Tab. 3.5) and the remaining fold is used for model evaluation (see example of 5-fold cross validation in the Tab. 3.5). After the process of training and scoring is done \(k\)-times, we can compute an average performance across folds or plot the performance statistics using histogram to see the variance.
Table 3.5 5-fold cross validation

For logistic regression there are no parameters to tune, but for decision tree there is the complexity parameter \( cp \) and for random forest there is the parameter \( mtry \). 10-fold cross validation is used to tune these parameters, because we do not know in advance their optimal value (Kuhn, 2016). The process of parameter tuning has the following steps:

Define sets of model parameter values to evaluate
for each parameter set do
  for each resampling iteration do
    Hold-out specific samples
    Fit the model on the remainder
    Predict the hold-out samples
  end
  Calculate the average performance across hold-out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set

The last technique, described here, for estimating future performance, is bootstrap sampling. Bootstrap sampling is based on creation of several randomly selected training and testing datasets, which are followingly used to compute the performance of the model (James et al., 2013). This technique is different from cross-validation, because bootstrap allows examples to be selected multiple times through the process of sampling with replacement. From the original dataset of \( n \) examples, the bootstrap procedure will create one or more new training datasets that also contain \( n \) examples, some of which are repeated. The corresponding testing datasets are then constructed from the set of examples not selected for the respective training datasets.
4 Application of quantitative methods for churn prediction

This section of the thesis is devoted to application of models and approaches described in Section 3 on real data set of approx. 50,000 customers from the selected European telecommunication company. Two data sets were available, a training data set, which purpose is to train classification models, and a testing data set to evaluate and compare the predictive performance of used linear and non-linear models. The variables, which could help to reveal leaving customers, were thoroughly selected on the basis of cooperation with customer service managers and IT database experts. Customer managers wanted to see mainly two categories of data – demographic data, such as age, customer lifetime or type of account, and service usage data, such as consumption of mobile data, voice or monthly invoice paid. They had another interesting suggestions for input variables such as number of calls to other networks, but this information was not easily extracted from the data warehouse.

The input variables divided into demographic group and service usage group are described and descriptive statistics are showed at the beginning of this part. Then the variables are analyzed graphically using histograms, overlaid density plots or bar plots to uncover interesting patterns. Two approaches to predictive modelling are compared. The first one is based on the estimation of logistic regression, decision tree and random forest models using the entire training data set. The estimated models are further described and explained. The testing data set is then used to calculate performance statistics to check the behavior of these models on unseen data. The second approach utilizes cluster analysis before the creation of predictive models. The training data set is at first divided into clusters using Gower distance and partitioning around medoids algorithm. A logistic regression, decision tree and random forest models are then estimated for each cluster. To make results of both approaches comparable, it is necessary to assign customers in testing data set into clusters. It is done based on minimal distance from cluster medoids. At the end of this Section the predictive performance of the two approaches is compared.

The final step is the proposal of the churn prediction system. At first, it is necessary to create tables with customer data, predictions and estimated models in relational database MySQL. It is done using connection of statistical programming environment R and database MySQL (Mandák, 2015). The tables in the database are further connected into visualization tool Qlik Sense which opens up the analytical results to the business users.
4.1 Description of the data

Data for the thesis were obtained from the selected European telecommunications operator. Raw input data for this task are stored in various production systems. The process of preparing data into the form suitable for modelling consists of the three traditional steps (ETL) – extract data from source systems, transform it and load it into database. IT experts from the company further provide the data in CSV format and make it accessible for data scientists. Three R packages are used to process the data in R – `plyr` (Wickham, 2011), `dplyr` (Wickham and Francois, 2016) and `reshape2` (Wickham, 2007). There are two types of input variables, according to its original systems – demographic data (see Tab. 4.1) and data related to the usage of various services (see Tab. 4.2).

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthYear</td>
<td>Customer’s year of birth</td>
</tr>
<tr>
<td>delinquent</td>
<td>Whether a customer has problem with paying bills</td>
</tr>
<tr>
<td>duration</td>
<td>How long a customer is the company’s customer</td>
</tr>
<tr>
<td>accType</td>
<td>Type of account</td>
</tr>
<tr>
<td>contractDuration</td>
<td>Date when contract ends</td>
</tr>
<tr>
<td>city</td>
<td>Indication, whether a customer is from the 5 biggest cities</td>
</tr>
</tbody>
</table>

Table 4.1 Demographic variables

There are following variables in the service usage group of input variables, see Tab. 4.2. It is important to note that all averages are monthly averages.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>avgInvoice</td>
<td>Average amount of invoice</td>
</tr>
<tr>
<td>avgExtra3</td>
<td>Average overpayment for last 3 months</td>
</tr>
<tr>
<td>avgExtra6</td>
<td>Average overpayment for last 6 months</td>
</tr>
<tr>
<td>avgExtra12</td>
<td>Average overpayment for last 12 months</td>
</tr>
<tr>
<td>avgData</td>
<td>Average data consumed (GB’s)</td>
</tr>
<tr>
<td>avgVoice</td>
<td>Average voice consumed (min’s)</td>
</tr>
<tr>
<td>avgSMS</td>
<td>Average number of SMS consumed</td>
</tr>
<tr>
<td>avgMMS</td>
<td>Average number of MMS consumed</td>
</tr>
<tr>
<td>VAS</td>
<td>Whether the customer has value-added services</td>
</tr>
<tr>
<td>portout</td>
<td>Whether the customer left the company or not</td>
</tr>
</tbody>
</table>

Table 4.2 Service usage variables
A few rows were randomly selected from the training data set (see Tab. 4.3 and 4.4) to get a picture of how the data look like.

<table>
<thead>
<tr>
<th>birthYear</th>
<th>contractDuration</th>
<th>delinquent</th>
<th>accType</th>
<th>avgInvoice</th>
<th>avgExtra12</th>
<th>avgExtra6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1957</td>
<td>-362</td>
<td>FALSE</td>
<td>A</td>
<td>281.42</td>
<td>140.48</td>
<td>177.45</td>
</tr>
<tr>
<td>1985</td>
<td>-593</td>
<td>FALSE</td>
<td>A</td>
<td>333.80</td>
<td>25.90</td>
<td>41.58</td>
</tr>
<tr>
<td>1984</td>
<td>123</td>
<td>FALSE</td>
<td>A</td>
<td>370.90</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1955</td>
<td>18</td>
<td>FALSE</td>
<td>A</td>
<td>485.22</td>
<td>56.89</td>
<td>84.44</td>
</tr>
<tr>
<td>1970</td>
<td>96</td>
<td>FALSE</td>
<td>A</td>
<td>1388.21</td>
<td>89.65</td>
<td>137.19</td>
</tr>
<tr>
<td>1933</td>
<td>255</td>
<td>FALSE</td>
<td>A</td>
<td>126.75</td>
<td>5.19</td>
<td>6.69</td>
</tr>
</tbody>
</table>

Table 4.3 Example of the training data set – part1

<table>
<thead>
<tr>
<th>avgExtra3</th>
<th>avgData</th>
<th>avgVoice</th>
<th>avgSms</th>
<th>avgMms</th>
<th>vas</th>
<th>city</th>
<th>duration</th>
<th>portout</th>
</tr>
</thead>
<tbody>
<tr>
<td>138.40</td>
<td>26.99</td>
<td>72.97</td>
<td>12</td>
<td>0</td>
<td>FALSE</td>
<td>no</td>
<td>2855</td>
<td>No</td>
</tr>
<tr>
<td>77.85</td>
<td>9343.69</td>
<td>0.00</td>
<td>25</td>
<td>0</td>
<td>TRUE</td>
<td>no</td>
<td>867</td>
<td>No</td>
</tr>
<tr>
<td>0.00</td>
<td>4788.45</td>
<td>0.00</td>
<td>6</td>
<td>1</td>
<td>FALSE</td>
<td>no</td>
<td>1041</td>
<td>No</td>
</tr>
<tr>
<td>110.25</td>
<td>183.90</td>
<td>140.84</td>
<td>5</td>
<td>0</td>
<td>FALSE</td>
<td>no</td>
<td>2729</td>
<td>No</td>
</tr>
<tr>
<td>85.83</td>
<td>55023.45</td>
<td>0.00</td>
<td>36</td>
<td>3</td>
<td>TRUE</td>
<td>no</td>
<td>2501</td>
<td>No</td>
</tr>
<tr>
<td>9.67</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>FALSE</td>
<td>no</td>
<td>5289</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4.4 Example of the training data set – part2

You can see descriptive statistics (minimum, the first quartile, median, mean, the third quartile and maximum) of numeric variables in the Tab. 4.5. There are clearly visible some outliers, e.g. maximal values for avgSms and avgMms. Outliers were removed and replaced by medians.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthYear</td>
<td>-6.890</td>
<td>-2.583</td>
<td>-1274</td>
<td>-1695</td>
<td>-366</td>
<td>4866</td>
</tr>
<tr>
<td>contractDuration</td>
<td>0.16</td>
<td>331.69</td>
<td>431.37</td>
<td>567.95</td>
<td>674.85</td>
<td>746.64</td>
</tr>
<tr>
<td>avgInvoice</td>
<td>0.00</td>
<td>9.35</td>
<td>46.31</td>
<td>92.19</td>
<td>25.16</td>
<td>6 133.47</td>
</tr>
<tr>
<td>avgExtra12</td>
<td>0.00</td>
<td>7.00</td>
<td>48.44</td>
<td>100.91</td>
<td>136.43</td>
<td>6 533.47</td>
</tr>
<tr>
<td>avgExtra6</td>
<td>0.00</td>
<td>2.00</td>
<td>42.18</td>
<td>101.82</td>
<td>136.16</td>
<td>9 666.17</td>
</tr>
<tr>
<td>avgData</td>
<td>0.00</td>
<td>366.70</td>
<td>2 505.60</td>
<td>10 746.80</td>
<td>8 968</td>
<td>60 602.80</td>
</tr>
<tr>
<td>avgVoice</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>49.15</td>
<td>64.00</td>
<td>2 723.72</td>
</tr>
<tr>
<td>avgSms</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.85</td>
<td>0.00</td>
<td>594.55</td>
</tr>
<tr>
<td>avgMms</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
<td>0.00</td>
<td>680</td>
</tr>
<tr>
<td>duration</td>
<td>15</td>
<td>1335</td>
<td>2196</td>
<td>2639</td>
<td>3833</td>
<td>7070</td>
</tr>
</tbody>
</table>

Table 4.5 Descriptive statistics of numeric input variables
The counts of individual levels for categorical variables are visible in the Tab. 4.6. The majority of customers have the account of type A and have no problems with paying bills. Roughly three fifths of customers use value-added services and only one fifth of customers is from the 5 biggest cities.

<table>
<thead>
<tr>
<th>accType</th>
<th>delinquent</th>
<th>vas</th>
<th>city</th>
<th>portout</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>true</td>
<td>49 948</td>
<td>31 623</td>
<td>yes</td>
</tr>
<tr>
<td>B</td>
<td>false</td>
<td>112</td>
<td>18 437</td>
<td>no</td>
</tr>
<tr>
<td>D</td>
<td>false</td>
<td>116</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4.6 Descriptive statistics of categorical input variables

Missing data were present in both numeric and categorical variables. An R package *Amelia* was used to impute these values (Honaker et al., 2011). This package uses multiple imputation and sophisticated EMB algorithm which combines expectation minimization algorithm and bootstrapping.

For employees in CRM department it can be interesting to see, how the typical customer which left company looks like. This information is obtained when only customers with portout variable equal to yes are selected and mean and median for numeric variables and mode for categorical and binary variables are computed. We can see these statistics in the Tab. 4.7.

<table>
<thead>
<tr>
<th>Numeric variable</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthYear</td>
<td>1973</td>
<td>1973</td>
</tr>
<tr>
<td>contractDuration</td>
<td>84</td>
<td>71</td>
</tr>
<tr>
<td>avgInvoice</td>
<td>423</td>
<td>546</td>
</tr>
<tr>
<td>avgExtra12</td>
<td>67</td>
<td>79</td>
</tr>
<tr>
<td>avgExtra6</td>
<td>65</td>
<td>86</td>
</tr>
<tr>
<td>avgExtra3</td>
<td>69</td>
<td>90</td>
</tr>
<tr>
<td>avgData</td>
<td>3 482</td>
<td>4 568</td>
</tr>
<tr>
<td>avgVoice</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>avgSms</td>
<td>3.9</td>
<td>4.3</td>
</tr>
<tr>
<td>avgMms</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>duration</td>
<td>825</td>
<td>1379</td>
</tr>
</tbody>
</table>

Table 4.7 Profile of typical churner – numeric variables
The most frequent values – modes – were calculated for categorical variables. From the Table 4.8 it is clear that a typical churner have the account of type A, does not have problems with paying bills, does not use any value-added services and does not come from the five biggest cities in the country.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>accType</td>
<td>A</td>
</tr>
<tr>
<td>delinquent</td>
<td>false</td>
</tr>
<tr>
<td>vas</td>
<td>false</td>
</tr>
<tr>
<td>city</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 4.8 Profile of typical churner – categorical variables

In creation of predictive models it is common to divide data sets into training data set and testing data set. Data sets in this thesis were created with respect to the time to simulate the behaviour of predictive models in a real world application. At first, explanatory variables were computed in the middle of August 2017. Because on 16.8.2017 we know the date of future portout, we can easily compute the customers who left company in 45 days and those who did not. The period of 45 days was selected based on consultations with customer – they need time to prepare retention activities. Then the same process was applied on the creation of the testing data set (see Fig. 4.1). In each data set there are approximately 50,000 customers. Finally, training data set is used to create predictive models and testing data set is used to test the predictive performance.

Figure 4.1 Training and testing data sets
In the following part of the thesis the variables in the training data set are graphically analyzed in order to explore some interesting patterns and to get a view of how the data look like.

### 4.2 Exploratory data analysis

Exploratory data analysis is typically applied before modeling and can help to inform about the development of more complex predictive models. Histograms are typically plotted for numeric variables to see their distribution, and bar plots are commonly used for categorical variables to see the counts of categories. Overlapping density plots can be used for comparison of differences of numeric variables grouped by categorical (binary) dependent variables.

Two density plots for variables where the differences between churners and non-churners are the biggest are showed. Density plots in the Fig. 4.2 and 4.3 were created using `ggplot2` package in R (Wickham, 2009).

![Figure 4.2 Overlaid density plot for duration](image)
Figure 4.3 Density plots for birth year

Blue density curve represents customers who left company, pink density curve represents customers who stayed. It is evident that churners are younger and are with company for a shorter time.

The distribution of values of independent variables is visible in histograms. From the Fig. 4.4 we can see that the most frequent birth year of customers is between 1960 and 1970. It is also evident that the usage variables (data, voice, invoice, sms, mms and over payment) follow Pearson Chi-Square distribution. The negative values of contract duration mean that the contract for a given customer ended in past. The positive values of contract duration have customers whose contract was still active when the dataset was being prepared. Another interesting variable in the data set is duration – how long a customer is with the company. The biggest group of customers (approx. 10,000 out of 50,000) is with company around 2,000 days ≈ 5.5 years. Then the frequency of customers slightly decreases for increasing duration.
Figure 4.4 Histograms of explanatory variables

The categorical variables are graphically represented by bar plots. The numbers inside the bars represent absolute frequencies of a given category in the whole data set. Majority of customers have no problem with paying their bills (delinquent = FALSE) and have the account of type A. Almost 37% of customers utilize value-added services and 20% are from the 5 biggest cities in the country (see Fig. 4.5).
It is important to note that the proportion of churners in the training data set is very low (835 out of 50,060, it means 1.668 %), see Fig. 4.6.
An important step in the preparation of the explanatory variables for modelling is to check the multicollinearity among explanatory variables. In the case of the presence of various data types (numeric, categorical, and binary) it is suitable to use generalized variance inflation factor (Fox and Monette, 1992) for the multicollinearity check. It is recommended to think about exclusion of variables with GVIF higher than 10. In the Tab. 4.9 we can see that variables avgExtra12, avgExtra6 and avgExtra3 exceed this limit. After the elimination of avgExtra6 and avgExtra3 all GVIF coefficients are lower than 2 and there is no multicollinearity among them.

<table>
<thead>
<tr>
<th>Variable</th>
<th>GVIF</th>
<th>GVIF^(1/(2*Df))</th>
<th>GVIF^(1/(2*Df))</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthYear</td>
<td>1.068372</td>
<td>1.033621</td>
<td>1.033538</td>
</tr>
<tr>
<td>contractDuration</td>
<td>1.024397</td>
<td>1.012125</td>
<td>1.012096</td>
</tr>
<tr>
<td>delinquent</td>
<td>1.030736</td>
<td>1.015252</td>
<td>1.014706</td>
</tr>
<tr>
<td>accType</td>
<td>1.003811</td>
<td>1.000951</td>
<td>1.000959</td>
</tr>
<tr>
<td>avgInvoice</td>
<td>1.253302</td>
<td>1.119510</td>
<td>1.109283</td>
</tr>
<tr>
<td>avgExtra12</td>
<td>10.514094</td>
<td>3.242544</td>
<td>1.044081</td>
</tr>
<tr>
<td>avgExtra6</td>
<td>38.937620</td>
<td>6.240002</td>
<td></td>
</tr>
<tr>
<td>avgExtra3</td>
<td>28.024005</td>
<td>5.293770</td>
<td></td>
</tr>
<tr>
<td>avgData</td>
<td>1.185410</td>
<td>1.088765</td>
<td>1.087832</td>
</tr>
<tr>
<td>avgVoice</td>
<td>1.087612</td>
<td>1.042886</td>
<td>1.036974</td>
</tr>
<tr>
<td>avgSms</td>
<td>1.552652</td>
<td>1.246055</td>
<td>1.244545</td>
</tr>
<tr>
<td>avgMms</td>
<td>1.540898</td>
<td>1.241329</td>
<td>1.241656</td>
</tr>
<tr>
<td>vas</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>city</td>
<td>1.025887</td>
<td>1.012861</td>
<td>1.012610</td>
</tr>
<tr>
<td>duration</td>
<td>1.062902</td>
<td>1.030971</td>
<td>1.030722</td>
</tr>
</tbody>
</table>

Table 4.9 Generalized Variance Inflation Factor

A nice visualization of correlation coefficients among variables is possible using corplot package in R (Wei, 2017). Variables related to payments over regular fee (avgExtra12, avgExtra6 and avgExtra4) are highly correlated, so it will be sufficient to use only one of them in predictive models. In the first Fig. 4.7, the size of correlation coefficients is represented by ellipses, the flatter the ellipse, the stronger the correlation. The Fig. 4.8 expresses the values of correlation coefficients also in absolute values.
Figure 4.7 Correlation matrix of numeric variables (method = ellipse)
4.3 Predictive models trained on all data set

This part of the thesis is focused on detailed explanation of the three classification models trained on the entire training data set – logistic regression, decision tree and random forest. The Caret package (Kuhn, 2016) was used to create each of these models. For decision tree and random forest the Caret package was also utilized for tuning model parameters – complexity parameter \( cp \) for decision tree and \( mtry \) – number of variables available for splitting at each tree node for random forest.

### 4.3.1 Logistic regression model

The first trained model, logistic regression, is a representative of linear methods for classification. The \textit{Train} function from Caret package (Kuhn, 2016) with method \textit{glm} and family \textit{binomial} was used to train a logistic regression model. The estimated model has the following form:

\[
\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p
\]

where \( p \) is the probability of the event occurring, and \( x_1, x_2, \ldots, x_p \) are the predictor variables.

The correlation matrix of numeric variables is shown in Figure 4.8.

![Correlation matrix of numeric variables](image)

**Figure 4.8** Correlation matrix of numeric variables (method = number)
\[
\log \left( \frac{\pi(\text{portout})}{1-\pi(\text{portout})} \right) = \ -16.38 + 0.0083 \cdot \text{birthYear} - 0.0037 \cdot \text{contractDuration} + 0.9964 \cdot \text{delinquenttrue} - 16.02 \cdot \text{accTypeD} - 0.9608 \cdot \text{accTypeA} - 0.0001 \cdot \text{avgInvoice} - 0.0001 \cdot \text{avgExtra12} - 0.0000 \cdot \text{avgData} - 0.0012 \cdot \text{avgVoice} + 0.0039 \cdot \text{avgSms} - 0.0176 \cdot \text{avgMms} - 19.02 \cdot \text{vastrue} - 0.0426 \cdot \text{cityyes} - 0.0004 \cdot \text{duration}
\]

All estimation results of logistic regression model are listed in the Table 4.10.

|                | Estimate  | Std. Error | z value | \( Pr(Z > |z|) \) |
|----------------|-----------|------------|---------|------------------|
| Intercept      | -1.64e+01 | 5.31e+00   | -3.08   | 0.002 **         |
| birthYear      | 8.34e-03  | 2.68e-03   | 3.11    | 0.002 **         |
| contractDuration | -3.70e-03 | 1.25e-04   | 29.57   | 0.000 ***        |
| delinquent = true | 9.96e-01  | 2.61e-01   | 3.82    | 0.000 ***        |
| accType = D    | -1.60e+01 | 1.63e+03   | -0.01   | 0.992            |
| accType = A    | -9.61e-01 | 4.88e-01   | -1.97   | 0.049 *          |
| avgInvoice     | -6.13e-05 | 1.33e-04   | -0.46   | 0.644            |
| avgExtra12     | -1.08e-04 | 1.93e-04   | -0.56   | 0.577            |
| avgData        | -2.23e-07 | 1.48e-06   | -0.15   | 0.880            |
| avgVoice       | -1.23e-03 | 4.20e-04   | -2.94   | 0.003 **         |
| avgSms         | 3.94e-03  | 2.25e-03   | 1.75    | 0.080            |
| avgMMS         | -1.76e-02 | 5.00e-02   | -0.35   | 0.725            |
| vas = True     | -1.90e+01 | 1.85e+02   | -0.10   | 0.918            |
| city = yes     | -4.26e-02 | 1.03e-01   | -0.41   | 0.679            |
| duration       | -3.70e-04 | 3.25e-05   | -11.39  | 0.000 ***        |

**Table 4.10** Estimation results for logistic regression model

We can see also values of *Wald statistics* (\( z – value \)) in the Table 4.10 with estimated results. According to the \( p \)-values corresponding to the \( z \)-statistic, we can determine which variables are statistically significant at the 5% level of significance:
• birthYear: The $p$-value is 0.002. There is a strong statistical evidence that year of birth is related to the probability of churn.
• contractDuration: The $p$-value is 0.002. There is a strong statistical evidence that contract duration is related to the probability of churn.
• delinquent: The $p$-value is lower than 0.0005. There is a strong statistical evidence that delinquency of customer is related to the probability of churn.
• accType: The $p$-value is 0.005. There is a strong statistical evidence that account type A is related to the probability of churn.
• avgVoice: The $p$-value is 0.003. There is a strong statistical evidence that average voice consumed is related to the probability of churn.
• duration: The $p$-value is lower than 0.0005. There is a strong statistical evidence that duration of customer in company is related to the probability of churn.

It is without any doubt that knowledge of variables which have a statistically significant impact on churn is important. However, it is also necessary to quantify their impact and to define whether their influence is positive or negative. Let’s demonstrate the process of calculation of the effect of individual explanatory variables by the introduction of odds ratio. For example for dichotomous independent variable delinquency, the relationship between the odds ratio and the regression coefficient is $OR = e^\beta_t$ (Hosmer and Lemeshow, 2000). When we know that the estimated regression coefficient of the variable delinquency is equal to 0.9964, the odds ratio can be computed as

$$e^{0.9964} = 2.71.$$  \hspace{1cm} (4.2)

The value of odds ratio 2.71 means that when a customer is delinquent, then he/she is almost three times likely to churn than a non-delinquent customer, ceteris paribus.

The direction of the impact of explanatory variables can be assessed by the plus or minus sign before the estimated regression coefficients. Table 4.11 shows division into variables with negative and positive impact on the probability of churn. The positive impact for numeric variables means that higher values of these variables increase the probability of churn and the negative impact means that higher values decrease the probability of churn. For example if the customer is younger (birthYear is higher), then the probability of churn increases. If the customer is longer time with the company (duration is higher), then the probability of churn decreases.
<table>
<thead>
<tr>
<th>positive impact</th>
<th>β coefficient</th>
<th>negative impact</th>
<th>β coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthYear</td>
<td>8.34e-03</td>
<td>accType = D</td>
<td>-1.60e+01</td>
</tr>
<tr>
<td>contractDuration</td>
<td>3.69e-03</td>
<td>accType = A</td>
<td>-9.61e-01</td>
</tr>
<tr>
<td>delinquent = TRUE</td>
<td>9.96e-01</td>
<td>avgInvoice</td>
<td>-6.13e-05</td>
</tr>
<tr>
<td>avgSms</td>
<td>3.94e-03</td>
<td>avgExtra12</td>
<td>-1.08e-04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>avgData</td>
<td>-2.23e-07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>avgVoice</td>
<td>-1.23e-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>avgMMS</td>
<td>-1.76e-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vas = TRUE</td>
<td>-1.90e+01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>city = YES</td>
<td>-4.26e-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>duration</td>
<td>-3.70e-04</td>
</tr>
</tbody>
</table>

Table 4.11 Direction of effect of predictors on probability to churn

### 4.3.2 Decision tree model

The second trained model, decision tree, is a representative of non-linear methods for classification. The `train` function from *Caret* package (Kuhn, 2016) with the method `rpart` was used to estimate the decision tree model. The shortcut `rpart` stands for *recursive partitioning*, approach which is used for decision tree building and which was described in the methodological part. Decision tree for classification of customers to churners and non-churners is showed in the Fig. 4.9. Only three variables were used to create this decision tree – duration, contractDuration and vas. These three variables have the biggest explanatory power. As it is clear from the Fig. 4.9, one variable can be used multiple times in the process of building the decision tree. In our case contractDuration variable is used three times.
As mentioned at the beginning of this chapter, the complexity parameter $cp$ of the decision tree model was tuned using 10-fold cross validation. This option is specified in the `trainControl` function from `caret` package (Kuhn, 2016), see Fig. 4.10.

```
library(caret)
fitcontrol <- trainControl(# 10-fold cv
                          method = "cv",
                          number = 10,
                          classProbs=TRUE)
```

The following Fig. 4.11 shows various values of complexity parameter on the $x$ axis and corresponding values of accuracy on the $y$ axis. The highest accuracy was achieved with the value of complexity parameter equals 0.0909, so the final model was trained using this value.
4.3.3 Random forest

Fig. 4.12 shows information about the random forest model – number of observations to train the model, number of predictors, number of classes of dependent variable, the different values of \textit{mtry} parameter and their corresponding \textit{accuracies}.
Random Forest
50060 samples
13 predictor
2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 45055, 45055, 45055, 45054, 45053, 45053, 45054, ...
Resampling results across tuning parameters:

<table>
<thead>
<tr>
<th>mtry</th>
<th>Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.9918897</td>
<td>0.6874440</td>
</tr>
<tr>
<td>3</td>
<td>0.9952857</td>
<td>0.8531815</td>
</tr>
<tr>
<td>4</td>
<td>0.9953855</td>
<td>0.8608083</td>
</tr>
<tr>
<td>5</td>
<td>0.9953256</td>
<td>0.8596406</td>
</tr>
<tr>
<td>6</td>
<td>0.9954255</td>
<td>0.8628616</td>
</tr>
<tr>
<td>7</td>
<td>0.9954654</td>
<td>0.8642383</td>
</tr>
<tr>
<td>8</td>
<td>0.9954055</td>
<td>0.8621547</td>
</tr>
<tr>
<td>9</td>
<td>0.9954055</td>
<td>0.8619813</td>
</tr>
<tr>
<td>10</td>
<td>0.9954255</td>
<td>0.8626347</td>
</tr>
<tr>
<td>11</td>
<td>0.9953256</td>
<td>0.8593137</td>
</tr>
<tr>
<td>12</td>
<td>0.9953256</td>
<td>0.8592190</td>
</tr>
<tr>
<td>13</td>
<td>0.9952856</td>
<td>0.8579093</td>
</tr>
<tr>
<td>14</td>
<td>0.9952856</td>
<td>0.8575670</td>
</tr>
<tr>
<td>15</td>
<td>0.9952257</td>
<td>0.8563595</td>
</tr>
</tbody>
</table>

Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 7.

Figure 4.12 Output of the random forest model

The following Fig. 4.13 shows various values of mtry parameter on the x axis and corresponding values of accuracy on the y axis. The highest accuracy was achieved with value of complexity parameter equals 0.0909, so the final model was trained using this value.
4.3.4 Comparison of variable importance

A big advantage of caret package in $R$ is a possibility to compute importance of explanatory variables and therefore see which factors drive the customer churn. For logistic regression absolute value of the $t$-statistic and for decision tree and random forest mean decrease in accuracy are computed (Kuhn, 2016). In short, the more the accuracy of the model decreases, because of the exclusion of a single variable, the more important that variable seems to be. Measures of importance are scaled to have a maximum value of 100 (see Fig. 4.14). The results are quite similar. All algorithms are mainly based on customer lifetime and duration of customer’s contract. One big difference is also visible, importance of value-added services for decision tree and random forest is very high, whereas for logistic regression this variable seems to be unimportant. On the other hand, delinquency of customer is a useful variable for logistic regression, while it is not the case of decision tree and random forest.
4.3.5 Predictive performance computed from testing data set

Three predictive models trained using logistic regression, decision tree and random forest algorithms are used to predict churn of customers in the testing data set. This step is really important, because in production use the models will predict on data unseen in the phase of model building. The ratio of churned customers in testing data set is similar to the ratio in training data set – 966 out of 50,446, it means 1.915 %. The whole workflow of predictive performance assessment is illustrated in the Fig. 4.15. At first, training data set is used to create predictive models. Then the customers from the testing data set serve as an input to these models. Predictions of churn are then computed and finally compared with the real values. Performance measures such as accuracy, hit rate, sensitivity, specificity, F-score and Area Under and ROC curve (AUC) are calculated using comparisons of predicted and real values.
For the classification problems, where one class is imbalanced (in our case churners) the most used performance metric – *accuracy* – is not sufficient. Two other metrics, *hit rate (precision)* and *sensitivity (recall)*, are much more important. All these metrics can be computed from information available in *confusion matrix* (Lantz, 2013). The threshold for logistic regression is usually set up to 0.5. Hosmer and Lemeshow (2000) recommend to choose the threshold in the intersection of sensitivity and specificity. In this case, the threshold was set to value 0.051 (see Fig 4.16.). It means that a customer with probability of churn higher than or equal to 0.051 is predicted to churn and a customer with probability lower than 0.051 is predicted to stay.

![Figure 4.15 Assessment of predictive performance using test set](image)

*Figure 4.15* Assessment of predictive performance using test set

![Figure 4.16 Determining optimal threshold value – intersection of sensitivity and specificity](image)

*Figure 4.16* Determining optimal threshold value – intersection of sensitivity and specificity
Table 4.12 summarizes the confusion matrices for our three compared models.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>stay</td>
<td>leave</td>
<td>stay</td>
</tr>
<tr>
<td></td>
<td>stay</td>
<td>leave</td>
<td>stay</td>
</tr>
<tr>
<td>leave</td>
<td>leave</td>
<td>stay</td>
<td>leave</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>stay</th>
<th>leave</th>
<th>stay</th>
<th>leave</th>
<th>stay</th>
<th>leave</th>
</tr>
</thead>
<tbody>
<tr>
<td>stay</td>
<td>46582</td>
<td>63</td>
<td>48149</td>
<td>61</td>
<td>48179</td>
<td>111</td>
</tr>
<tr>
<td>leave</td>
<td>1756</td>
<td>1680</td>
<td>189</td>
<td>1682</td>
<td>159</td>
<td>1632</td>
</tr>
</tbody>
</table>

Table 4.12 Confusion matrices

Confusion matrices in Table 4.5 were used to compute the performance statistics — Accuracy, Hit rate, Sensitivity, Specificity and F-score (see Tab. 4.13). Due to the imbalanced nature of the data set it is not unexpected that accuracy is high for all models. As mentioned above, Hit rate and Sensitivity are performance measures of our interest. The highest value of Hit rate is achieved by random forest model (0.9112). Hit rate for decision tree is almost the same as for random forest – 0.8990. If we use decision tree or random forest, around 90 % of our predictions for customers to churn are correct. Sensitivity measures the ability of the model to catch customers who, in reality, left the company. Results are quite impressive, decision tree is able to catch 96.5 % and random forest 93.63 % of such customers. Also F-score, which combines Hit rate and Sensitivity into one measure, is almost the same for decision tree and random forest – 0.9308 and 0.9236, respectively. The F-score is slightly higher for decision tree, it could mean that less sophisticated model is more suitable for this business case.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9637</td>
<td>0.5111</td>
<td>0.9639</td>
<td>0.9637</td>
<td>0.6488</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9950</td>
<td>0.8990</td>
<td>0.9650</td>
<td>0.9961</td>
<td>0.9308</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9946</td>
<td>0.9112</td>
<td>0.9363</td>
<td>0.9967</td>
<td>0.9236</td>
</tr>
</tbody>
</table>

Table 4.13 Performance metrics

Another approach to measure the performance of classifiers is to use ROC curves or corresponding performance measure Area under Curve (AUC). The pROC package in R was used for creation of ROC curves and for computation of AUC measure. Two vectors are necessary – the
predicted class values and the estimated probabilities of the positive class. Comparison of ROC curves for our three compared classification models can be seen in the Fig. 4.17.

![Comparison of ROC curves](image)

**Figure 4.17** Comparison of ROC curves

From the above Fig. 4.17 it is evident that classification ability of all three compared models is excellent. Also the appropriate AUC statistics are very close to 1, which confirms a great classification ability (see Tab. 4.14).

<table>
<thead>
<tr>
<th>Model type</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9883</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9928</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9962</td>
</tr>
</tbody>
</table>

**Table 4.14** Comparison of AUC values

We have seen, from the testing of predictive performance based on test set unseen by the model, that all classification algorithms are able to reliably detect customers who are about to leave the company. Decision tree and random forest achieved high values of all performance metrics. Logistic regression was a little bit weaker in hit rate and F-score metrics.
4.4 Cluster analysis

The goal of the cluster analysis is to partition the training data set into smaller and homogeneous sets of data. The aim of the clustering is revealing of interesting characteristics of clusters and improving predictive performance of models created for prediction of customer churn.

Because in our case of mixed data types the clustering algorithms based on Euclidean distance (e.g. \textit{k-means} or Ward’s method) are not applicable, there are two recommended algorithms which are able to cope with \textit{Gower’s distance} – \textit{hierarchical clustering} and \textit{partitioning around medoids} (Rocco, 2016). Hierarchical clustering algorithms with various linkage types (single, average, complete) were at first tried, but did not produce satisfying results. Therefore, the \textit{PAM} algorithm was used to divide training data set into clusters.

It is known that in \textit{PAM} algorithm we don’t know in advance what the optimal number of clusters is. One possibility is to run this algorithm for various number of clusters, let’s say from 2 to 10 and then pick the number of clusters proposed by \textit{average Silhouette width}. In \textit{R} there is a package \textit{fpc} (Hennig, 2015) containing \textit{PAM} algorithm and one of its useful features is computation of \textit{Silhouette index} for each observation. The next step is calculation and comparison of \textit{average Silhouette width} for all observations and the optimal number of clusters is obtained for the highest value of the \textit{average Silhouette width}. The values of \textit{average Silhouette width} for 2 to 10 clusters are shown in the Fig. 4.18. The highest average silhouette width (0.544) is achieved for 4 clusters.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig4.18.png}
\caption{Average Silhouette width for 2 to 10 clusters}
\end{figure}
Optimal number of clusters can be obtained also with the usage of visualization methods. The t-SNE method (Stochastic Neighbor Embedding) is a technique useful for visualizing high-dimensional data. The t-SNE algorithm models the probability distribution of neighbors (the set of nearest points) around each point. This probability distribution is modelled in high-dimensional space as a Gaussian distribution while it is modelled as a $t$-distribution in the 2-dimensional space. The goal of the algorithm is to find a mapping onto the 2-dimensional space that minimizes the differences between these two distributions over all points. The main parameter controlling the fitting is called *perplexity* and is equivalent to the number of neighbors considered when matching the original and fitted distributions for each point (Maaten and Hinten, 2008).

There are two assumptions for this algorithm – the data must be numeric (e.g. categorical variables should be converted by binary encoding) and should be normalized so the data are in the same scale. See the visualization of training data set using t-SNE method in the Fig. 4.19.

Figure 4.19 Visualization of clusters using t-SNE method
Probably the first thing we are interested in after clustering process is the detection of a cluster size – in our case it is represented by the number of customers in each cluster. The biggest cluster is the cluster 1 containing 38,546 customers, then cluster 2 with 8,579 customer, the third biggest is cluster 4 with 2,104 customers and the smallest cluster is cluster 3 with 831 customers (see Fig. 4.20).

![Number of customers in particular clusters](image)

**Figure 4.20** Number of customers in particular clusters

An important step after clustering is the creation of cluster profiles – the definition of characteristic features of each cluster. For this purpose, customers representing clusters, in the terminology of cluster analysis the so-called *medoids*, are compared and described. The interesting differences between clusters are highlighted by orange color in Tables 4.15 and 4.16.

In the first cluster there is a majority of customers – 38,546 out of 50,060 customers in the training data set. The typical representative of this cluster was born in 1970 and he/she, compared to the other clusters, is the most loyal customer of the company. From the financial point of view, this cluster is not so interesting, because the average monthly invoice is the lowest. The consumption of data and voice is also low and value-added services are not used.

The second cluster, from the point of view of number of customers, is quite numerous – 8,579 customers fall into this cluster. A huge attention should be given to the customers of this cluster,
because their average monthly invoice is 3 to 5 times higher than in other three clusters. This cluster is also characteristic by usage of value-added services.

There are two areas which make the third cluster different from the other ones – the highest overpayments and high consumption of data, voice and sms. The high amount of overpayments may signalize that customers in this cluster have inappropriate contracts which should be revised. This cluster consists of 831 customers.

The last cluster is made of young customers – the year of birth of the typical representative is 1993. It is well-known that there is a trend amongst young people using telecommunications services. They do not use traditional services such as voice, sms and mms and instead they use mainly mobile internet and various additional services.

<table>
<thead>
<tr>
<th>cluster</th>
<th>birthYear</th>
<th>delinquent</th>
<th>accType</th>
<th>avgInvoice</th>
<th>avgExtra12</th>
<th>avgData</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1970</td>
<td>FALSE</td>
<td>A</td>
<td>311.46</td>
<td>7.88</td>
<td>1276.07</td>
</tr>
<tr>
<td>2</td>
<td>1970</td>
<td>FALSE</td>
<td>A</td>
<td>1697.45</td>
<td>44.03</td>
<td>18451.65</td>
</tr>
<tr>
<td>3</td>
<td>1969</td>
<td>FALSE</td>
<td>A</td>
<td>455.42</td>
<td>537.39</td>
<td>142470.19</td>
</tr>
<tr>
<td>4</td>
<td>1993</td>
<td>FALSE</td>
<td>A</td>
<td>526.38</td>
<td>33.26</td>
<td>57457.95</td>
</tr>
</tbody>
</table>

Table 4.15 Representatives (medoids) of clusters

<table>
<thead>
<tr>
<th>cluster</th>
<th>avgVoice</th>
<th>avgSms</th>
<th>avgMms</th>
<th>vas</th>
<th>city</th>
<th>duration</th>
<th># of customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.96</td>
<td>0.00</td>
<td>0.00</td>
<td>FALSE</td>
<td>yes</td>
<td>1688</td>
<td>38546</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>TRUE</td>
<td>yes</td>
<td>991</td>
<td>8579</td>
</tr>
<tr>
<td>3</td>
<td>544.31</td>
<td>105.57</td>
<td>4.57</td>
<td>FALSE</td>
<td>yes</td>
<td>1095</td>
<td>831</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>TRUE</td>
<td>yes</td>
<td>1086</td>
<td>2104</td>
</tr>
</tbody>
</table>

Table 4.16 Representatives (medoids) of clusters

The training data set was successfully divided into four meaningful clusters. An input distance matrix containing Gower’s distances was clustered using PAM algorithm. The partition into four clusters was defined based on the value of Silhouette index. The visualization using t-SNE method confirmed that four clusters are reasonable. Based on medoids, it was revealed that there occur substantial differences among clusters.
4.5 Predictive models trained for individual clusters

In the previous chapter the process of division of the training data set into four clusters was described in detail. In the following phase we use the created clusters and estimate predictive models specific for each clusters. For each cluster there are three models – logistic regression, decision tree and random forest. The total number of predictive models is therefore 12. The whole process is described in Fig. 4.21 and consists of these steps:

1) Divide training data set into $k$ clusters.
2) Estimate 3 predictive models for each cluster.
3) Compute the distance between data in the test set and the cluster medoids.
4) Assign cluster to the observation from the testing data set based on the minimal distance.
5) For the test set use predictive models corresponding to their assigned cluster.
6) Assess predictive performance of all models on the test data set.

![Diagram](image)

Figure 4.21 Combination of clustering and classification
Before estimation of predictive models, let’s compare the churn rates among individual clusters (see Fig. 4.22). The values of churn rates are comparable among clusters ranging from 1.28 % in cluster 4 to 1.9 % in cluster 2. For completeness, the absolute number of customers in a given cluster is also showed.

![Figure 4.22 Number of customers and churn rate for individual clusters](image)

### 4.5.1 Predictive models for cluster 1

Before description of the estimated models it should be noted that the first cluster is the biggest of all clusters and consists of 38,546 customers. The first estimated model is logistic regression model, see estimation results in Tab. 4.17. It is interesting that some variables, which are statistically significant for the model trained on all training data set are statistically insignificant for model trained for cluster 1, and vice versa. The differences in statistical significance of individual variables, in comparison with model trained on all training data set, are following:

- variable birthYear in cluster 1 is statistically insignificant,
- in cluster 1 accType B is significant instead of accType A,
- avgInvoice in cluster 1 is statistically significant,
- avgData in cluster 1 is statistically significant.

Based on these findings we can clearly see that the clustering makes sense as in clusters are likely slightly different drivers of churn.
|                           | Estimate | Std. Error | z value | $Pr(Z > |z|)$ |
|---------------------------|----------|------------|---------|-------------|
| Intercept                 | -9.194e+00 | 6.100e+00 | -1.507  | 0.131797    |
| birthYear                 | 4.225e-03  | 3.096e-03 | 1.365   | 0.172304    |
| contractDuration          | 3.663e-03  | 1.386e-04 | 26.425  | 0.000000 ***|
| delinquent = true         | 1.301e+00  | 3.353e-01 | 3.882   | 0.000104 ***|
| accType = B               | 1.208e+00  | 5.300e-01 | 2.279   | 0.022669 *  |
| accType = D               | -1.533e+01 | 1.708e+03 | -0.009  | 0.992839    |
| avgInvoice                | -7.991e-04 | 2.340e-04 | -3.416  | 0.000636 ***|
| avgExtra12                | -1.357e-04 | 2.316e-04 | -0.586  | 0.557960    |
| avgData                   | 7.760e-05  | 1.979e-05 | 3.922   | 0.000088 ***|
| avgVoice                  | -1.024e-03 | 4.626e-04 | -2.213  | 0.026896 *  |
| avgSms                    | 5.344e-03  | 2.332e-03 | 2.292   | 0.021909 *  |
| avgMMS                    | -1.975e-02 | 5.547e-02 | -0.356  | 0.721722    |
| vas = True                | -1.890e+01 | 2.330e+02 | -0.081  | 0.935365    |
| city = yes                | -7.906e-02 | 1.186e-01 | -0.667  | 0.504875    |
| duration                  | -3.533e-04 | 3.602e-05 | -9.808  | 0.000000 ***|

**Table 4.17** Logistic regression model for cluster 1

As a random forest model is an ensemble of many decision trees and the model itself, unlike in logistic regression, does not have any coefficients, we can observe only these information about random forest model - number of observations to create a model, number of predictors, number of classes of dependent variables and information about cross-validation (see Fig. 4.23).

38546 samples  
13 predictor  
2 classes: 'No', 'Yes'

No pre-processing  
Resampling: Cross-Validated (10 fold)  
Summary of sample sizes: 34691, 34692, 34692, 34691, 34692, 34691, ...

**Figure 4.23** Random forest output from R for cluster 1
There is a hyperparameter *mtry* in a random forest model – number of variables randomly sampled as candidates at each split – which need to be optimized. This is done using cross-validation. In the random forest model for cluster 1 *mtry* parameter was set to 9 (see Fig. 4.24)

![Figure 4.24 Tuning of *mtry* parameter for random forest model in cluster 1](image)

The last estimated model for cluster 1 is decision tree model. Its structure and used variables are the same as in the decision tree model trained on all training data set (see Fig. 4.25).

![Figure 4.25 Decision tree for cluster 1](image)
There is a comparison of variable importance for logistic regression, decision tree and random forest estimated for cluster 1 in the Fig. 4.26. It is clearly visible that contractDuration is the most important predictor for all models followed by duration and value-added services.

![Variable importance](image)

**Figure 4.26** Variable importance for models in cluster 1

### 4.5.2 Predictive models for cluster 2

Similarly as in previous chapter, at first let’s repeat that there are 8,579 customers in the cluster 2 and that this cluster is the second biggest. We can see estimated results of logistic regression model for cluster 2 in the Table 4.18. The first thing which is visible at the first sight, and which makes this model much different from the model for cluster 1, is that there are only three statistically significant variables. These variables are contractDuration, avgVoice and duration.

|                       | Estimate | Std. Error | z value | $Pr(Z > |z|)$ |
|-----------------------|----------|------------|---------|--------------|
| Intercept             | -2.258e+01 | 1.594e+01  | -1.416  | 0.156759     |
| birthYear             | 1.101e-02  | 8.032e-03  | 1.371   | 0.170335     |
| contractDuration      | 3.933e-03  | 3.618e-04  | 10.869  | < 2e-16 ***  |
| delinquent = true     | 8.084e-01  | 5.087e-01  | 1.589   | 0.112077     |
| accType = B           | -4.310e-01 | 1.213e+00  | -0.355  | 0.722340     |
Table 4.18 Logistic regression model for cluster 2

Another model estimated for cluster 2 is random forest model (see Fig. 4.27). Again, hyperparameter \textit{mtry} was optimized using cross-validation to the value 6 (see Fig. 4.28).

\begin{verbatim}
8579 samples
13 predictor
2 classes: 'No', 'Yes'

NO pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 7721, 7721, 7721, 7721, 7721, 7721, ...
\end{verbatim}

\textbf{Figure 4.27} Random forest output from \textit{R} for cluster 2

\textbf{Figure 4.28} Tuning of \textit{mtry} parameter for random forest model in cluster 2
The decision tree model for cluster 2 consists again of 3 variables – duration, contractDuration and vas. The structure is similar to the decision tree model for cluster 1. The only difference is the order of the variable vas in the tree (see Fig. 4.29).

![Decision tree for cluster 2](image)

**Figure 4.29** Decision tree for cluster 2

We can see a comparison of variable importance for logistic regression, decision tree and random forest models estimated for cluster 1 in the Fig. 4.30. Again contractDuration is the most important predictor for all models. The second most important variable is variable duration expressing customer lifetime. It is worth mentioning that in case of logistic regression variable avgVoice also plays an important role in predicting customer churn.
4.5.3 Predictive models for cluster 3

The third cluster consists of only 831 customers. Nevertheless, it is a very important cluster, because there is a large number of potential churners. Therefore, this segment should be seriously taken into account. As we know from the cluster analysis, this cluster is characteristic by high overpayments and this unpleasant fact could cause churn of these customers.

The following Table 4.19 shows estimated regression coefficients as well as their standard errors, z values and p values. The only statistically significant variable is contractDuration. This is probably caused by the low number of observations in the cluster (831). If we were less strict in the assessment of statistical significance, also the variable avgExtra12, representing overpayments, could be accepted as statistically significant (p-value 0.128).
|                     | Estimate | Std. Error | z value | \( Pr(Z > |z|) \) |
|---------------------|----------|------------|---------|-----------------|
| Intercept           | 7.178e+01| 6.154e+01  | 1.166   | 0.243           |
| birthYear           | -3.629e-02| 3.118e-02  | -1.164  | 0.244           |
| contractDuration    | 4.182e-03| 1.048e-03  | 3.989   | 0.000 ***       |
| delinquent = true   | 8.218e-01| 1.540e+00  | 0.534   | 0.594           |
| accType = B         | -1.923e+01| 7.339e+04  | 0.000   | 1.000           |
| avgInvoice          | -3.216e-04| 6.807e-04  | -0.472  | 0.637           |
| avgExtra12          | 4.112e-03| 2.703e-03  | 1.521   | 0.128           |
| avgData             | -5.166e-06| 3.854e-06  | -1.340  | 0.180           |
| avgVoice            | -4.638e-04| 1.951e-03  | -0.238  | 0.812           |
| avgSms              | -4.297e+01| 2.468e+03  | -0.017  | 0.986           |
| avgMMS              | 1.114e+01 | 2.137e+04  | 0.001   | 1.000           |
| vas = True          | -2.791e+01| 5.418e+04  | -0.001  | 1.000           |
| city = yes          | 9.952e-03| 9.610e-01  | 0.010   | 0.992           |
| duration            | -3.129e-04| 3.086e-04  | -1.014  | 0.311           |

**Table 4.19** Logistic regression model for cluster 3

Figures 4.31 and 4.32 show information about random forest model for cluster 3. In this case 9 variables are randomly sampled as candidates at each split of individual decision tree.

```
831 samples
13 predictor
2 classes: 'No', 'Yes'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 748, 748, 748, 748, 749, 747, ...
```

**Figure 4.31** Random forest output from R for cluster 3
The decision tree model for the cluster 3 is due to the low number of observations constructed only from one variable – duration. If a customer from cluster 3 is customer of the company for more than approx. one year (351 days) than the customer is predicted to leave the company (see Fig. 4.33)
Importance of individual predictors in cluster 3 models is quite different than in the previous two clusters – variable contractDuration is no longer the most important one for all models. The variable duration is the key predictor for random forest and decision tree models. There are also some variables which do not play any role in the models, such as avgSms, avgMms, accType or city (see Fig. 4.34).

![Variable importance for models of cluster 3](image)

**Figure 4.34** Variable importance for models of cluster 3

### 4.5.4 Predictive models for cluster 4

We know from the results of cluster analysis that in the last (the fourth) cluster there are young customers who use mainly mobile internet and value-added services. There are 2,104 customers in this segment. There are estimated results of logistic regression model can be found in the Tab. 4.20.

Again, probably due to the low number of observations in this cluster, there are only two variables which are statistically significant at 5 % level of significance – contractDuration and duration.
|                     | Estimate  | Std. Error | z value | Pr(Z > |z|) |
|---------------------|-----------|------------|---------|--------|
| Intercept           | 1.077e+01 | 4.039e+01  | 0.267   | 0.7896 |
| birthYear           | -6.050e-03| 2.035e-02  | -0.297  | 0.7663 |
| contractDuration    | 3.448e-03 | 6.727e-04  | 5.126   | 0.0000 |
| delinquent = true   | -1.188e-01| 1.187e+00  | -0.100  | 0.9203 |
| accType = B         | -1.424e+01| 9.501e+03  | -0.001  | 0.9988 |
| avgInvoice          | 9.000e-04 | 5.598e-04  | 1.608   | 0.1079 |
| avgExtra12          | -3.654e-04| 1.831e-03  | -0.200  | 0.8418 |
| avgData             | -1.129e-05| 1.568e-05  | -0.720  | 0.4713 |
| avgVoice            | -3.672e-03| 3.809e-03  | -0.964  | 0.3350 |
| avgSms              | 2.638e-02 | 2.309e-02  | 1.142   | 0.2533 |
| avgMMS              | -1.375e-01| 3.530e-01  | -0.390  | 0.6968 |
| vas = True          | -2.072e+01| 1.058e+03  | -0.020  | 0.9844 |
| city = yes          | -2.997e-01| 5.559e-01  | -0.539  | 0.5898 |
| duration            | -4.838e-04| 2.412e-04  | -2.006  | 0.0449 * |

Table 4.20 Logistic regression model for cluster 4

Figures 4.35 and 4.36 show information about random forest model for cluster 3. In this case 13 variables are randomly sampled as candidates at each split of individual decision tree.

2104 samples
13 predictor
2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 1894, 1893, 1893, 1893, 1895, 1894, ...

Figure 4.35 Random forest output from R for cluster 4
Figure 4.36 Tuning of mtry parameter for random forest model in cluster 2

The decision tree model for cluster 4 is constructed from 4 variables – duration, contractDuration, vas and avgExtra12 (see Fig. 4.37).

Figure 4.37 Decision tree for cluster 4
The last figure in the part devoted to the description of models for individual clusters is comparison of variable importance. We can see that the importances of variables are similar to those for models in cluster 1, 2 or models for the whole training data set – the most important variable is contractDuration followed by duration, value-added services and avgInvoice (see Fig. 4.38).

![Graph showing variable importance for models of cluster 4](image)

**Figure 4.38** Variable importance for models of cluster 4
4.6 Comparison of results

The next step after the estimation of predictive models for each cluster is an assessment of their predictive performance. For this purpose confusion matrix is created from values predicted on the test data set and real values of the test set. Confusion matrices are further used as a base for calculation of performance metrics – accuracy, hit rate, sensitivity, specificity and F-score. An ROC curves are also plotted and AUC statistics computed. The thresholds for logistic regression models are defined based on intersection of sensitivity and specificity, as proposed by Hosmer and Lemeshow (2000).

In the first part of this chapter three estimated predictive models – logistic regression, decision tree and random forest are compared in each cluster. Then in the second part the initial approach when predictive models were trained on the entire training data set is compared with the “clustering” approach in which 3 models, one for each cluster, were trained. The prediction results for individual clusters are aggregated in order to make comparison with the initial approach possible. By the aggregation we mean sum of confusion matrices for all clusters and predictive models and calculation of performance statistics using these aggregated data.

4.6.1 Comparison of predictive performance among clusters

There are confusion matrices for cluster 1 in the Table 4.21. From the first look they seem similar with one weakness for logistic regression model – this model has high number of false positives (137).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>stay</td>
<td>stay</td>
<td>stay</td>
</tr>
<tr>
<td>leave</td>
<td>8042</td>
<td>1</td>
<td>8164</td>
</tr>
<tr>
<td>leave</td>
<td>137</td>
<td>101</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8167</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>99</td>
</tr>
</tbody>
</table>

Table 4.21 Confusion matrices for cluster 1

The high number of false positives for logistic regression has an impact on hit rate, which is equal to 0.4244 (see Tab. 4.22). Also F-score is for logistic regression model lower than for decision tree
or random forest – 0.5941. The predictive performance of decision tree and random forest is strong, as suggested by all performance metrics.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9833</td>
<td>0.4244</td>
<td>0.9910</td>
<td>0.9833</td>
<td>0.5941</td>
<td>0.9889</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9981</td>
<td>0.8707</td>
<td>0.9902</td>
<td>0.9982</td>
<td>0.9266</td>
<td>0.9941</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9983</td>
<td>0.9000</td>
<td>0.9706</td>
<td>0.9987</td>
<td>0.9340</td>
<td>0.9945</td>
</tr>
</tbody>
</table>

**Table 4.22** Performance statistics for cluster 1

As we know from the previous chapters, to make logistic regression comparable with decision tree and random forest we need to determine probability threshold. In the case of cluster 1 it is in intersection of sensitivity and specificity and equals to 0.075 (see Fig. 4.39).

![Figure 4.39 Determining optimal cutoff value for logistic regression in cluster 1](image)

The Fig. 4.40 shows an ROC curves of random forest, logistic regression and decision tree models trained on data from cluster 1. All curves are close to the ideal state what implies that all models have a high predictive ability.
Figure 4.40 ROC curves for models in cluster 1

The situation in confusion matrices for models in cluster 2 is similar to the previous one in cluster 1 – logistic regression model contains really high number of false positives – 1,526 (see Tab. 4.23). The lowest value of false positives is achieved by random forest model (134) but at the expense of a higher number of false negatives (119 vs. 74 in logistic regression and 52 in decision tree).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>27888</td>
<td>29244</td>
<td>29280</td>
</tr>
<tr>
<td>leave</td>
<td>1526</td>
<td>1384</td>
<td>134</td>
</tr>
</tbody>
</table>

Table 4.23 Confusion matrices for cluster 2

Again a high number of false positives has an impact on both hit rate and F-score of logistic regression. Performance statistics for decision tree and random forest are as well as in cluster 1 very high.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9482</td>
<td>0.4756</td>
<td>0.9492</td>
<td>0.9481</td>
<td>0.6337</td>
<td>0.9811</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9928</td>
<td>0.8921</td>
<td>0.9643</td>
<td>0.9942</td>
<td>0.9268</td>
<td>0.9929</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9918</td>
<td>0.9090</td>
<td>0.9184</td>
<td>0.9954</td>
<td>0.9137</td>
<td>0.9961</td>
</tr>
</tbody>
</table>

Table 4.24 Performance statistics for cluster 2
As it is clear from the Fig. 4.41, probability threshold for logistic regression model was set to 0.057 at the intersection of sensitivity and specificity.

![Figure 4.41 Determining optimal cutoff value for logistic regression in cluster 2](image)

All ROC curves are close to the ideal state and this fact implies that predictive ability of all three models is good. Only ROC curve for logistic regression is slightly worse than the curves for decision tree and random forest (see Fig. 4.42).

![Figure 4.42 ROC curves for models in cluster 2](image)
The third cluster is, from the number of customers point of view, the smallest one – only 75 customers from the test data set were assigned to this cluster. In confusion matrices visible in Table 4.25 we can see that there are no false positives for decision tree and random forest.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>36</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>leave</td>
<td>24</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 4.25** Confusion matrices for cluster 3

Performance statistics for cluster 3 are showed in Tab. 4.26. According to all metrics it seems that random forest model is the best one.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.6133</td>
<td>0.2941</td>
<td>0.6667</td>
<td>0.6000</td>
<td>0.4081</td>
<td>0.6317</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.8800</td>
<td>1.0000</td>
<td>0.4000</td>
<td>1.0000</td>
<td>0.5714</td>
<td>0.7000</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9467</td>
<td>1.0000</td>
<td>0.7333</td>
<td>1.0000</td>
<td>0.8462</td>
<td>0.9967</td>
</tr>
</tbody>
</table>

**Table 4.26** Performance statistics for cluster 3

The probability cutoff of logistic regression model was set to the value 0.01.

![Figure 4.43 Determining optimal cutoff value for logistic regression in cluster 3](image)
The best performance of random forest model is confirmed also in Fig. 4.44. The AUC value for random forest model is equal to 0.9967, which is higher than in the case of logistic regression (AUC = 0.6317) and also in the case of decision tree (AUC = 0.7).

Figure 4.44 ROC curves for models in cluster 3

As it is clear from the Tab. 4.27 which shows confusion matrices of models for cluster 4, logistic regression again suffers from the high number of false positives. On the other hand, the number of false negatives is really low and it means that logistic regression model was able to predict 166 out of 168 customer who, in reality, left the company. Decision tree and also random forest models have less false positives, but more false negatives.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>stay</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>leave</td>
<td>10418</td>
<td>2</td>
<td>10683</td>
</tr>
<tr>
<td>leave</td>
<td>267</td>
<td>166</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4.27 Confusion matrices for cluster 4
The values of the hit rate and F-score for logistic regression are lower than for other two models (see Tab. 4.28), but the sensitivity is the highest for logistic regression. The disadvantage of decision tree model trained on data from cluster 4 is that it is able to catch only 1/2 of customers who really left company. Also AUC value for decision tree is the lowest of three compared models.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9752</td>
<td>0.3834</td>
<td>0.9881</td>
<td>0.9750</td>
<td>0.5524</td>
<td>0.9924</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9923</td>
<td>0.9775</td>
<td>0.5179</td>
<td>0.9998</td>
<td>0.6770</td>
<td>0.7529</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9964</td>
<td>0.9329</td>
<td>0.8274</td>
<td>0.9991</td>
<td>0.8770</td>
<td>0.9966</td>
</tr>
</tbody>
</table>

Table 4.28 Performance statistics for cluster 4

We can see that the values of sensitivity and specificity are changing based on various probability cutoffs in the Fig 4.45. The value of probability threshold for the 4th cluster was set to 0.02.

![Figure 4.45 Determining optimal cutoff value for logistic regression in cluster 3](image)

The final comparison of ROC curves shows that random forest and logistic regression models, in the case of cluster 4, are more appropriate than decision tree model. Let’s repeat that the main imperfection of decision tree model is that it is not able to capture 50% of churners.
4.6.2 Comparison of “all data set” approach vs. “clustering” approach

After a detailed comparison of predictive models for particular clusters, it is important to compare the predictive ability of both approaches. Just for repetition, the first approach is based on a training of one logistic regression model, one decision tree and one random forest model using the entire training data set. These three models are applied to the whole test data set. The second approach utilizes results of cluster analysis. The training set is divided into 4 clusters and three models are trained for each cluster. Distances of customers in the test set from cluster medoids are computed and a given customer is assigned to the cluster where the distance to the cluster medoid is the lowest. The test set is thus enriched by cluster membership for each customer, so we can use an appropriate models trained for a specific cluster. Because the test set consists of exactly the same customers for both approaches, we can finally compare them and see, whether the clustering of customers before the estimation of predictive models leads to an increase in predictive performance.

Similarly as in the previous comparisons, confusion matrices serving as a basis for computation of performance statistics are constructed (see Tab. 4.29 and 4.30). The aggregated confusion matrices, for approach based on clustering, were created simply by summing up true
negatives, false negatives, false positives and true positives across all clusters. There are no big differences between the tables at the first sight, what could imply that there will not be any significant differences between the approaches. For logistic regression we can observe that the number of wrong predictions is slightly higher for the clustering approach – both false negatives and false positives are higher. If we compare confusion matrices for decision trees and random forest, the biggest difference is in higher number of false negatives. Based on the above mentioned findings we can expect that in the clustering approach, predictive models should be slightly worse than in all data set approach. Just for completeness, for logistic regression model trained on all data set the probability threshold was set to 0.051, probability thresholds for logistic regression models in individual clusters were described in the corresponding chapters 4.5.1 to 4.5.4.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>stay</td>
<td>leave</td>
<td>stay</td>
</tr>
<tr>
<td>leave</td>
<td>1756</td>
<td>1680</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>46582</td>
<td>48149</td>
<td>48179</td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>61</td>
<td>111</td>
</tr>
</tbody>
</table>

*Table 4.29* Confusion matrices for models trained on all training data set

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Logistic regression</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>stay</td>
<td>stay</td>
<td>leave</td>
<td>stay</td>
</tr>
<tr>
<td>leave</td>
<td>1954</td>
<td>1661</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>46384</td>
<td>48151</td>
<td>48182</td>
</tr>
<tr>
<td></td>
<td>82</td>
<td>143</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>1680</td>
<td>1682</td>
<td>1632</td>
</tr>
</tbody>
</table>

*Table 4.30* Aggregation of confusion matrices for individual clusters
Performance statistics computed from the above confusion matrices are available in Tables 4.31 and 4.32. The pattern is the same for both approaches – all models have a strong predictive ability with values of all performance statistics not lower than 0.9. There is one exception for both approaches – logistic regression. Its hit rate and F-score are significantly lower than for decision tree and random forest. To explain the value of hit rate 0.51, it means that 51% of the predictions of churn were correct, whereas 49% were wrong. It has a negative financial effect for the company due to wasted money on retention campaigns, which were focused on customers, who do not plan to leave the company.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9637</td>
<td>0.5111</td>
<td>0.9639</td>
<td>0.9637</td>
<td>0.6488</td>
<td>0.9883</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9950</td>
<td>0.8990</td>
<td>0.9650</td>
<td>0.9961</td>
<td>0.9308</td>
<td>0.9928</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9946</td>
<td>0.9112</td>
<td>0.9363</td>
<td>0.9967</td>
<td>0.9236</td>
<td>0.9962</td>
</tr>
</tbody>
</table>

Table 4.31 Performance statistics for “all data set” approach

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.9593</td>
<td>0.4595</td>
<td>0.9530</td>
<td>0.9596</td>
<td>0.6200</td>
<td>0.9819</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.9934</td>
<td>0.8953</td>
<td>0.9180</td>
<td>0.9961</td>
<td>0.9065</td>
<td>0.9854</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.9938</td>
<td>0.9111</td>
<td>0.9111</td>
<td>0.9968</td>
<td>0.9111</td>
<td>0.9973</td>
</tr>
</tbody>
</table>

Table 4.32 Performance statistics for “clustering” approach

The fact that both approaches are comparable is illustrated in the Tab. 4.33. The difference between both approaches is computed for each metric and each model. It is clear that the deviations are very low – an average difference in metrics for logistic regression is 0.0200, 0.0153 for decision tree and 0.0077 for random forest. The values of performance statistics are slightly higher for approach based on training models using all training data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Hit rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-score</th>
<th>AUC</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.0044</td>
<td>0.0516</td>
<td>0.0109</td>
<td>0.0041</td>
<td>0.0288</td>
<td>0.0064</td>
<td>0.0200</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.0016</td>
<td>0.0037</td>
<td>0.047</td>
<td>0</td>
<td>0.0243</td>
<td>0.0074</td>
<td>0.0153</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.0008</td>
<td>1E-04</td>
<td>0.0252</td>
<td>-1E-04</td>
<td>0.0125</td>
<td>-0.0011</td>
<td>0.0077</td>
</tr>
</tbody>
</table>

Table 4.33 Differences in performance statistic between both approaches
4.7 Proposal of churn prediction system

The main goal of this thesis is the *churn prediction system* which should serve as a quantitative support for decision makers responsible for customer relationship management. The system consists of the following subsystems:

1) computation of explanatory variables,
2) model training,
3) scoring existing customers using created models,
4) storing predictions into *MySQL* database,
5) connection of database to *Qlik Sense* and visualization.

The proposed *churn prediction system* (see Fig 4.47) is built on open source software, namely *R*, programming language for statistical computing and graphics, *R Studio*, an integrated development environment for *R*, *MySQL*, an relational database management system, and *Qlik Sense*, a platform for data analysis and visualization. Based on this fact, the system could serve as a template for other companies dealing with churning customers to create their own system with minimal costs.

![Churn prediction system diagram](image)

**Figure 4.47** Churn prediction system

The data set for model training is provided in a CSV format. The data are loaded into *R* and further processed into the form needed to train the predictive models – one row for each customer.
and one column for each variable. The second subsystem is focused on training of classification models – logistic regression, decision tree and random forest. The trained models are used to score the probability of churn for customers in the testing data set – customers unseen in the process of the model training. The customer data along with predictions and variable importances are stored in three tables in MySQL database. The data model is available in the Fig. 4.48.

![E-R diagram for customer data, predictions and model variables](image)

**Figure 4.48** E-R diagram for customer data, predictions and model variables

Connection of the R and MySQL database is provided using RMySQL package (Ooms et al., 2017). Function `dbWriteTable` can be used to store R data type data frame as a relational table in MySQL. The first table stored to MySQL database is the Customers table (see Fig. 4.49) with characteristics listed in the “Description of the data” part of the thesis. This table will serve as a data source for Customer characteristics sheet in Qlik Sense dashboard, where the business users will clearly see the overall picture of the customer features. Business users have also a possibility to filter by cluster and therefore can see differences among them.
The aim of the second table – *Predictions* – is to store the values predicted by three utilized models. The column *customer_id* stores the unique identification of customer, the column *model* stores a utilized model and the column *predicted_value* stores predicted probability of customer to leave the company (see Fig. 4.50). The last column *timestamp_pred* stores the date of the prediction. This information is important for tracking the development of tendency of customers to churn.

The probability of churn of some customers could increase, so the business users can further explore the data in *Qlik Sense* dashboard to see which factors have caused this increase.

The last table *Model_variables* has 4 columns (see Fig. 4.51). The columns *model* and *variable* are self-explanatory, the column *importance* contains relative importance of a given variable for a given model, and the column *coefficient* is filled only for logistic regression model since there are no coefficients for decision tree and random forest. This table is used in the *Variable*
importance sheet in the Qlik Sense dashboard and is vital for business users, because they can see the factors driving churn and for logistic regression also the direction of the impact.

**Figure 4.51** Table Model_variables in MySQL database

Once we have the tables and relational data model in database prepared, we can proceed to data visualization in Qlik Sense. Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it is processed. In our case, Qlik Sense was selected as a tool for data visualization. Qlik Sense is a platform for data analysis, which enables to analyze data and make data discoveries on your own. You can share knowledge and analyze data in groups and across organizations, ask and answer your own questions and follow your own paths to insight. Qlik Sense also enables you and your colleagues to reach decisions collaboratively.

The first sheet – Customer – contains basic summary information about the customers and their probabilities to churn (see Fig. 4.52). There are 3 KPI’s – the number of customers with churn probability higher than 0.9, amount of money, which will company lose each month if customers with probability to churn > 0.9 leave the company, and the number of customers. We can see also histogram of churn probabilities divided into 10 bins ranging from 0 - 0.1 to 0.9 - 1. The table in the bottom right corner lists customers with their probability to churn and their average monthly invoice. The red cells in this column denotes that the average invoice for a given customer is higher than the 3rd quartile – retention teams should focus their activities and efforts mainly on these profitable customers. On the top there are 6 filtering options – accType, birthYear, city, vas, duration and cluster to explore the customer data more deeply.
Figure 4.52 Sheet Customers in Qlik Sense
Fig. 4.53 displays Variable importance sheet in Qlik Sense. The aim of this sheet is to provide decision makers an answer to the question what are the main reasons why customers leave/stay the company. We can see here comparison of variable importance of individual variables across classification models, negative and positive drivers according to the estimated beta coefficients from logistic regression, the most important variables summed across classification models and also individual values of variable importance for all 3 models.
Visualization of customer characteristics is visible in the Fig. 4.54. The data source for this sheet is table Customers in MySQL database. On the top of the dashboard there are again filtering options to drill down to the desired level of granularity. Under the filter there are various KPI’s (key performance indicators) – the average values of numeric variables. Under the averages, there are also another 5 KPI’s – number of delinquent customers and number of customer with account type A, B, C, D. At the bottom of the sheet there are 2 histograms, one for year of birth and one for duration. On the right side we can see pie charts showing percentage of customer utilizing value-added services and percentage of customers living in the top 5 biggest cities of the country.
The last sheet in the dashboard contains “raw” data in the same format like they are stored in the database or like business users are used to see them in Excel spreadsheets (see Fig. 4.55). This sheet could be beneficial for employees of the company, who are in a direct contact with customers and need to have detailed information about them in hands. Above the table in this sheet there are again filtering options for accType, birthYear, city, vas and duration for drill downs.

![Figure 4.55 Sheet Customer detail in Qlik Sense](image)

<table>
<thead>
<tr>
<th>custo...</th>
<th>birth...</th>
<th>value added services</th>
<th>city</th>
<th>account type</th>
<th>delinqu...</th>
<th>average invoice</th>
<th>average overpayment 12</th>
<th>average overpayment 6</th>
<th>average overpayment 3</th>
<th>average data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1951</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>537</td>
<td>41</td>
<td>39</td>
<td>31</td>
<td>16.81</td>
</tr>
<tr>
<td>2</td>
<td>1969</td>
<td>True</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>1162</td>
<td>29</td>
<td>47</td>
<td>25</td>
<td>19.012</td>
</tr>
<tr>
<td>3</td>
<td>1979</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>374</td>
<td>99</td>
<td>102</td>
<td>89</td>
<td>47.14</td>
</tr>
<tr>
<td>4</td>
<td>1965</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>1277</td>
<td>22</td>
<td>36</td>
<td>19</td>
<td>9.469</td>
</tr>
<tr>
<td>5</td>
<td>1902</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>3106</td>
<td>201</td>
<td>205</td>
<td>174</td>
<td>73.8</td>
</tr>
<tr>
<td>6</td>
<td>1967</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>291</td>
<td>348</td>
<td>268</td>
<td>382</td>
<td>3302</td>
</tr>
<tr>
<td>7</td>
<td>1967</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>187</td>
<td>175</td>
<td>183</td>
<td>191</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1979</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>127</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1953</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>127</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1932</td>
<td>False</td>
<td>no</td>
<td>A</td>
<td>false</td>
<td>196</td>
<td>157</td>
<td>155</td>
<td>126</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1936</td>
<td>True</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>547</td>
<td>19</td>
<td>15</td>
<td>8</td>
<td>34.88</td>
</tr>
<tr>
<td>12</td>
<td>1964</td>
<td>True</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>1833</td>
<td>206</td>
<td>248</td>
<td>288</td>
<td>176.11</td>
</tr>
<tr>
<td>13</td>
<td>1993</td>
<td>True</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>592</td>
<td>65</td>
<td>56</td>
<td>63</td>
<td>68.08</td>
</tr>
<tr>
<td>14</td>
<td>1958</td>
<td>True</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>393</td>
<td>44</td>
<td>64</td>
<td>65</td>
<td>18.69</td>
</tr>
<tr>
<td>15</td>
<td>1954</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>567</td>
<td>44</td>
<td>84</td>
<td>64</td>
<td>39.76</td>
</tr>
<tr>
<td>16</td>
<td>1959</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>392</td>
<td>118</td>
<td>115</td>
<td>89</td>
<td>76</td>
</tr>
<tr>
<td>17</td>
<td>1966</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>434</td>
<td>43</td>
<td>86</td>
<td>98</td>
<td>38.63</td>
</tr>
<tr>
<td>18</td>
<td>1964</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>727</td>
<td>61</td>
<td>82</td>
<td>155</td>
<td>32.59</td>
</tr>
<tr>
<td>19</td>
<td>1999</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>136</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1968</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>574</td>
<td>326</td>
<td>263</td>
<td>334</td>
<td>8.597</td>
</tr>
<tr>
<td>21</td>
<td>1931</td>
<td>False</td>
<td>yes</td>
<td>A</td>
<td>false</td>
<td>378</td>
<td>368</td>
<td>422</td>
<td>462</td>
<td>0</td>
</tr>
</tbody>
</table>
5 Conclusion

Satisfied customers are a key prerequisite for long-term success of businesses, especially in saturated markets such as telecommunications. An ideal scenario is the situation when satisfied customers become loyal customers who repeatedly renew their contracts and recommend the company to other people in their surroundings. A branch of marketing called relationship marketing deals with long-term relationships with customers. Long-term relationships with customers undoubtedly bring huge financial and non-financial benefits to the company. Companies in telecommunications, where it is easy to switch to another service provider, have to continuously monitor customer needs, requirements and wishes and modify their services in order to be successful in the market.

The main objective of the thesis was to propose and implement churn prediction system for the selected European telecommunications Company. The purpose of the system is to help to reduce number of leaving customers and enable deep understanding of customer base. Technically the proposed system consists of three technologies – statistical programming environment R, relational database management system MySQL and visualization tool Qlik Sense. An advantage of the system is that all components are open source so companies in any other business area can get inspired and can create their own system for prediction of leaving customers with relatively low costs.

The first step of the creation of the system was to prepare data. It means create meaningful features from the data set, both independent variables and dependent variable expressing that a given customer left the company 45 days after the day of computation predictors. An exploratory data analysis, performed to explore an interesting relationships and patterns in the data, revealed that younger customers and customers of the company for a shorter time tend to churn more. Independent variables were also tested for the presence of multicollinearity using Spearman’s correlation coefficient (for numeric variables) and Generalized VIF (for all variables). Variables representing overpayments over regular tariff computed for the last 3, 6 and 12 months were strongly correlated and only monthly average computed from last the 12 months remained in the data set.
When the data were prepared for modelling, two approaches were compared. The first one was based on estimation of logistic regression, decision tree and random forest models using the entire training data set of 50,000 customers. The second one utilized cluster analysis before creation of predictive models. The estimated models were further described and explained and the testing data set of 50,000 customers was used to check the quality of models on unseen data.

Logistic regression, decision tree and random forest models were at first estimated using the entire training data set. *Wald tests* for individual regression coefficients confirmed that birth year, days till the end of contract, delinquency and average number of sent sms have statistically significant positive effect on churn of customer; statistically significant negative impact on churn was proved for account types A and D, average monthly invoice, mobile data, voice, mms, overpayment, value-added services, duration and residence of the customer in the 5 biggest cities in the country. Decision tree model was constructed from variables duration, contract duration and value-added services. As previously mentioned, it is really important to know the reasons of churn. It is possible using importance of variables in predictive models. The most important driver of churn is the number of days till the end of contract – the closer the end, the higher is the probability of churn. The second most important driver is customer duration – customer who spent shorter time with company tend to churn more. Another important aspect decreasing probability to churn is value-added services – customers who use them are likely more satisfied and churn less. The predictive performance tested using independent test set showed that all three estimated models have a great predictive ability, values of all performance metrics except logistic regression weren’t lower than 0.9. The logistic regression had lower values of hit rate (0.5111) and F-score (0.6488) due to the high portion of false positives.

The first step of the second approach used in the thesis was to partition training data set into clusters. Because of the presence of mixed data types (numeric, categorical, binary), *Gower distance* along with *Partitioning around Medoids (PAM)* algorithm were used. Four clusters were selected as optimal. The criteria were the highest value of *Silhouette index* and the visualization *t-SNE* method. The first biggest cluster (38,546 customers) was characterized by loyal customers with the lowest average monthly invoice. The second cluster consisted of 8,579 customers should be thoroughly monitored because of the highest monthly invoice. The third cluster was the smallest one with 831 customers and was typical of the highest overpayments and usage of mobile data.
The last cluster represented by 2,104 customers contained the youngest customers who used mobile data and value-added services.

Three predictive models for each cluster were estimated after division of the training data set. Logistic regression model for cluster 1 revealed that, for this cluster, variables accType = B, avgInvoice and avgData are statistically significant unlike in the logistic regression model estimated on the whole training data set. Logistic regression models for clusters 2-4 were characteristic by considerably lower number of statistically significant variables, mainly contractDuration and duration played a role. It seems that it was caused by the lower number of observations in these clusters. Decision tree models were similar to the decision tree in the first approach except cluster 3, where only duration was used to create the tree. From the point of view of predictive performance, similar pattern as in the first approach was found – great predictive ability of all models except the lower hit rate and F-score for logistic regression models. Customers in the testing data set were assigned to the corresponding clusters based on the minimal distance from the cluster medoids.

From the comparison of the first approach, which was based on training models on the entire training data set and the second approach, which was based on training models for individual clusters, it came out that there is only minimal difference between them. The values of performance statistics varied in hundredths. But we cannot state that this will be the case of all other churn prediction models. It is therefore necessary to test various models and approaches because each data set is unique.

An integral and from the usability point of view the most important part of the proposed system is its deployment – accessibility of the system to the employees in marketing and retention departments. Deployment of the system was done using storage of the customer characteristics, predicted probabilities of churn and model characteristics into MySQL database and subsequent connection of the database to the visualization tool Qlik Sense. Business users can therefore see the relationships and interesting patterns in the data.

The contribution of the doctoral dissertation to the field of study “Systems Engineering and Informatics” can be divided into three levels – methodical, empirical and pedagogical. Methodical contribution of the thesis resides in the proposal and verification of the system for evaluation and
prediction of customer churn in the area of telecommunications. The proposed complex system is based on 4 main features:

- application of system approach,
- classification of customers from telecommunications company in relation to churn,
- usage of linear, but also non-linear classification methods and cluster analysis,
- visualization of results.

From the empirical point of view, the contribution of this thesis lies in the verification of the methodology not only for the entire sample of customers, but also for individual customer segments. Another benefit which comes out from cluster analysis is definition of cluster profiles and recommendation of how to deal with specific customer segments. Detection and understanding of the reasons of churn and overview about behavior of customers made accessible thanks to the visualization are also crucial. The most important benefit of the system, from the business point of view, is the financial benefit reached by the rescue of at risk customers.

There is also a pedagogical contribution of the thesis which should not be forgotten. The quantitative methods from the area of cluster analysis (Gower distance matrix, k-medoids algorithm, Silhouette coefficient) and classification models (logistic regression, decision trees, random forests) together with performance metrics (hit rate, sensitivity, AUC) can be showed to students as an examples of application of quantitative methods in real business practice.

Although it was proved that the predictive performance of the proposed churn prediction system is very good, there are some ways how to improve the system. Sensitivity analysis applied to estimated regression coefficients could be applied to see the robustness of the model. Additional classification models could be used to predict churn, e.g. neural networks which are often cited in literature. Because churn prediction is unbalanced in nature, some of the techniques to balance the data could be employed – for instance SMOTE, up-sampling or down-sampling. From the technical point of view, the run of an R script could be automated using taskscheduleR package in R, which would help an analyst to minimize his/her work if new data are placed in specific folder.
References

Specialized books


**Articles in scientific journals and contributions from conferences**


**Norms**


**R packages**


### List of abbreviations and symbols

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Area under Curve</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and regression trees</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-separated values</td>
</tr>
<tr>
<td>CRISP-DM</td>
<td>Cross Industry Standard Process for Data Mining</td>
</tr>
<tr>
<td>CR</td>
<td>Customer Retention</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>EFQM</td>
<td>European Foundation for Quality Management</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract, Transform, Load</td>
</tr>
<tr>
<td>FN</td>
<td>False Negatives</td>
</tr>
<tr>
<td>FP</td>
<td>False Positives</td>
</tr>
<tr>
<td>GDPR</td>
<td>General Data Protection Regulation</td>
</tr>
<tr>
<td>GVIF</td>
<td>Generalized Variance Inflation Factor</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>LTR</td>
<td>Long-term Relationship</td>
</tr>
<tr>
<td>OOB</td>
<td>Out of Bag Observation</td>
</tr>
<tr>
<td>PAM</td>
<td>Partitioning Around Medoids</td>
</tr>
<tr>
<td>RCS</td>
<td>Rate of customer satisfaction</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>rpart</td>
<td>Recursive partitioning</td>
</tr>
<tr>
<td>SNE</td>
<td>Stochastic Neighbor Embedding</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>VIF</td>
<td>Variance Inflation Factor</td>
</tr>
</tbody>
</table>
\( a(i) \)  Average distance between object \( i \) and all other data within the same cluster

\( b(i) \)  The lowest average distance of object \( i \) to all points in any other cluster

\( \hat{\beta}_i \)  Estimate of regression parameter

\( C \)  Total number of customers with concluded contracts

\( C_B \)  Number of customers at the beginning of the year

\( C_E \)  Number of customers at the end of the year

\( C_j \)  Cluster \( j \)

\( cp \)  Complexity parameter for decision tree

\( E \)  Sum of the absolute error for all objects in the data set

\( E_j \)  Number of expected cases in the \( j^{th} \) group

\( E(y|x) = \pi(x) \)  Conditional probability that \( y = 1 \) given the covariate vector \( x \)

\( \delta_{ijk} \)  Indicator whether is possible comparison of objects \( i \) and \( j \) on variable \( k \)

\( G \)  Test statistic of likelihood ratio test

\( G^2_{HL} \)  Test statistic of Hosmer-Lemeshow test

\( Gini(S) \)  Gini coefficient computed for subset of data \( S \)

\( GiniGain(F) \)  Gini gain computed for variable \( F \)

\( \log \left( \frac{\pi(x)}{1-\pi(x)} \right) \)  The log-odds (logit)

\( l(\beta) \)  Likelihood function

\( \ln[l(\beta)] \)  Log-likelihood function

\( L(\hat{\beta}_{MLE}) \)  Likelihood under the model evaluated at the maximum likelihood estimates

\( mtry \)  Number of predictors sampled for splitting at each node

\( n_j \)  Number of observations in the \( j^{th} \) group

\( O \)  Number of observed cases in the \( j^{th} \) group
$o_j$ Medoid of cluster $j$

$p$ Point representing a given object in cluster $C_j$

$\frac{\pi(x)}{1-\pi(x)}$ Odds

$\hat{SE}(\beta_i)$ Standard error of the estimate

$s(i)$ Silhouette coefficient for object $i$

$S_{ijk}$ Similarity of observations $i$ and $j$ on variable $k$

$S_{ij}$ Total similarity between objects $i$ and $j$

$s_j$ Standard deviation

$T$ Total time of relationships with customers

$u_{ij}$ Standardized variable for customer $i$ and variable $j$

$\bar{x}_j$ Arithmetic mean

$x_{ij}$ Original value

$Z$ Test statistic of Wald test
I declare that

– I was (a) aware of the fact that my dissertation is fully covered by Act No. 121/2000 Coll. – Copyright Act, particularly Sec. 35 – use of the work within civil and religious ceremonies, in school performances and school use of the work and Sec. 60 – school work;

– I acknowledge that VŠB – Technical University of Ostrava (the VŠB-TUO) has the right to use the dissertation in a non-profit manner for its internal needs (Sec. 35 para. 3);

– I agree that the dissertation will be electronically archived in the Central Library of VŠB-TUO and one copy shall be filed with the dead of the dissertation work. I agree that bibliographic data of the dissertation will be published in the information system of VŠB-TUO;

– It was agreed that I shall conclude a license agreement with VŠB-TUO, in case of interest on its part, containing permission to use the work in accordance with Sec. 12 para. 4 of the Copyright Act;

– It was agreed that I can use the work, dissertation work, or license its use only with the approval of VŠB-TUO, which is authorized in this case to ask me to make a reasonable contribution to the costs that were incurred by VŠB-TUO towards the creation of the work (up to the actual amount).

In Ostrava on 12.7.2018

Jan Mandžák
The list of attachments

A.1  User guide for dashboard in Qlik Sense
Appendix

A.1 User guide for dashboard in Qlik Sense

This user guide provides a manual how to access the sheets within the dashboard, how to navigate between them, how to make selections and also shortly describes the logic of Qlik Sense.

What is Qlik Sense?

Qlik Sense is a data analysis platform enabling users to analyse data and make data discoveries. Users can share knowledge and analyse data in groups and across organizations. It does not require predefined and static reports. When a user make selection, Qlik Sense instantly updates every visualization and view in the app with a newly calculated set of data and visualizations specific to selections. The Qlik Sense associative data model functions like a one huge relational table – every single selection has effect on all other data.

Sheets within the dashboard

The environment in the figure below is called the App overview. Each dashboard consists of one or more sheets. The nature of sheets is similar to the sheets in Excel – each one is composed of charts and tables focused on a specific area. There are 4 sheets in churn prediction dashboard – customers, customer characteristics, variable importance and customer detail.

![Figure A.1 App overview](image-url)
A sheet is opened by clicking on the left mouse button on the icon of any sheet. A user can switch between sheets either by choosing the desired sheet by clicking on the sheet name (in this example grey button “Variable importance”) or by clicking on the left and right arrows. A user can also create a new sheet in the case that necessary rights are assigned to him.

![Figure A.2 Switching between sheets](image)

**Making selections in sheets**

Qlik Sense is known as a data discovery tool thanks to the various possibilities of selections. There are many ways how to filter the data, the most used option is probably the filter pane. A user can filter data in all sheets using filter pane based on the following variables: account type, birth year, city, value-added services, customer lifetime or cluster.

![Figure A.3 Filter pane](image)

Another option is to make desired selection directly in the chart. Here we selected customers with value-added services in the pie chart.
A user can make selections not only in charts but also in tables, see Fig. A.5. Here we want to see e.g. only information about importance of variable birthYear.

![Figure A.4 Selection in the chart](image)

The selected values appears in the top left corner of the sheet (see Fig. A.6). Keep in mind that you can select only values from dimensions, not from measures. So for example you can select customers with value-added services from big cities, but cannot select customers e.g. with birth year higher than 1968. There are another possibilities how to make this selection possible in Qlik Sense which are beyond the scope of this tutorial.

![Figure A.5 Selection in the table](image)
Exporting data to excel

Sometimes a user need to export the data to Excel. Exporting data from table or other chart objects from Qlik Sense to Excel is easy. In the figure below there is an example of an export of data from table in Qlik Sense. All you need to do is to click on the right mouse button inside the chart object. Three options appears, see Fig. A.7.

After clicking on the option Export data, another options appears, see Fig. A.8.
There are three options – export as an image, export to PDF or export data. The last one downloads the data into an xlsx file. A user receives the following confirmation of successful download (see Fig. A.9). The last step is to click on the link Click here to download your data file.

Figure A.9 Confirmation of successful download