PROPOSAL AND IMPLEMENTATION OF CHURN PREDICTION SYSTEM
FOR TELECOMMUNICATIONS COMPANY

Field of study: Systems Engineering and Informatics

Ing. Jan Manděk

Ostrava, July, 2018
PROPOSAL AND IMPLEMENTATION OF CHURN PREDICTION SYSTEM
FOR TELECOMMUNICATIONS COMPANY

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

Ostrava, July, 2018
Author of doctoral dissertation: Ing. Jan Mandák
Department of Systems Engineering
Faculty of Economics
VŠB – Technical University of Ostrava

Department of Systems Engineering
Faculty of Economics
VŠB – Technical University of Ostrava

The defense of the doctoral dissertation will be held before a committee for the defense of doctoral dissertation in the field of study Systems Engineering and Informatics at the Faculty of Economics, VŠB – TU Ostrava, Sokolská tř. 33.

The doctoral dissertation can be consulted at the Department of doctoral studies, Havlíčkovo nábřeží 38a, room E 410.
# Table of contents

1 Introduction .......................................................................................................................... 2

2 Objective and structure of doctoral dissertation ............................................................... 3

3 Content of doctoral dissertation ....................................................................................... 4

4 Procedure of doctoral dissertation ................................................................................... 6

5 Methods applied in doctoral dissertation ......................................................................... 7

6 Summary of results and conclusion .................................................................................. 13

7 List of references ............................................................................................................. 16

8 List of author’s publications and research ....................................................................... 22

9 Summary ............................................................................................................................ 24

10 Shrnutí ............................................................................................................................... 26
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, telecommunications 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.

Motivation of the thesis

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.
2 Objective and structure of doctoral dissertation

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.

Structure of the thesis

The dissertation thesis consists of three main chapters. The first one (chapter 2) is focused on the definition of theoretical background for management of relationships with customers in telecommunications. The quantitative methods used in the thesis are described in the next chapter 3. Methods of cluster analysis such as Gower distance, partitioning around medoids or Silhouette coefficient and linear (logistic regression) and non-linear (decision trees, random forests) classification models are explained in detail. The last chapter 4 is the most important one and deals with the creation of churn prediction system with the use of methods and approaches described in the chapter 3. The predictive ability of the proposed system is verified on two different data sets both of approximately 50,000 customers from the selected European telecommunications company.
3 Content of doctoral dissertation

1 Introduction

1.1 Purpose and motivation

1.2 Main and secondary objectives of work

1.3 Current state of examination

2 Customer relationship management in telecommunications

2.1 Customer satisfaction and its measurement

2.2 Customer loyalty and its measurement

2.3 Relationship marketing

2.4 Preventing churn in telecommunications

3 Description of applied methods and approaches

3.1 Data mining methodology – CRISP DM

3.2 Cluster analysis

3.2.1 Gower coefficient of similarity

3.2.2 Hierarchical clustering

3.2.3 Partitioning around medoids

3.2.4 Determining optimal number of clusters

3.3 Classification models

3.3.1 Logistic regression

3.3.2 Decision trees

3.3.3 Random forests

3.4 Evaluating predictive performance

3.4.1 Estimating future performance
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
4 Procedure of doctoral dissertation

The doctoral dissertation thesis can be divided into three logical parts:

- theoretical background of customer relationship management, mainly measurement of satisfaction and loyalty of customers,
- description of applied classification methods and methods of cluster analysis,
- proposal and implementation of churn prediction system.

At the beginning it is necessary to emphasize the importance of customer satisfaction and loyalty and its measurement, because they have a direct impact on the company economic effectiveness. Loyal customers are a company asset because they are characteristic by repeated purchases and positive references. The “relationship marketing” - a subfield of marketing focused on relationships with customers – is also important within the prediction of leaving customers.

The second part deals with a description of quantitative methods suitable for distinguishing customers to there who are about to leave and those who are about to stay. The logistic regression from the linear classification models and decision trees and random forests from non-linear classification models are described. Because the data set is large and consists of approx. 50,000 customers, it could be useful to divide them into homogeneous clusters. Therefore, also methods of cluster analysis such as Gower distance, $k$-medoids algorithm or Silhouette coefficient are explained. The quality of the estimated models is assessed by various performance measures, for instance hit rate, sensitivity or AUC, which also need to be introduced.

The quantitative methods are further applied to the data set of 50,000 customers and used as a core of the proposed churn prediction system. Two approaches are compared. Within the first approach, classification models are estimated using the entire training data set. The second approach is based on division of training data set into clusters and successive estimation of classification models for each cluster. Both approaches are compared from the predictive ability point of view. Because the predicted probabilities of churn are not sufficient for employees in retention department, it is necessary to put them in the context. The customer characteristics, importances of variables and predictions are for this purpose stored in the database, which is further connected to the visualization tool Qlik Sense. The final output of the system is the churn prediction dashboard with various visualizations. With these information employees are able to see customers at risk of churning and can create and target personalized retention campaigns to the right group of customers.
5 Methods applied in doctoral dissertation

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 those who are not. The creation of the system follows the traditional CRISP DM methodology, which consists of business understanding, data understanding, data preparation, modelling, evaluation and deployment.

The dependent variable is categorical with 2 classes (churn/non-churn), therefore classification methods are used in the 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. A 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 the thesis, the customers are first clustered and the classification models are estimated for individual clusters with the goal of improvement of predictive performance. Because the variables in the data set are of mixed types, Gower coefficient is used to compute the similarity matrix between individual customers. The partitioning around medoids algorithm (PAM) is then used to divide the data into homogeneous groups.

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 goal 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 the areas, where objects have tendency to group (Meloun and Militký, 2006).

Because in our data sets the variables are of a mixed type (numerical, categorical, binary), we will further describe only selected methods suitable for this situation – Gower coefficient of similarity, PAM algorithm and Silhouette coefficient.

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 it is not possible to do a comparison because of missing information or because of non-existent character in both \(i\) and \(j\) in the case of dichotomous variables. 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} (1)

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.

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 $k$.

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

**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 cluster with the representative object to which it is the most similar. The partitioning method algorithm is based on minimization of the sum of the dissimilarities between each object and its corresponding medoid (Han and Kamber, 2001):

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|,$$  \hspace{1cm} (2)

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 define 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, $o_{random}$;
5. compute the total cost, $S$, of swapping representative object, $o_j$, with $o_{random}$;
6. if $S < 0$ then swap $o_j$ with $o_{random}$ 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 an analyst needs to specify the number of clusters $k$ before run of the algorithm.

**Determining optimal number of clusters**

Oftentimes we want to divide a dataset into smaller homogeneous groups and from the prior knowledge it is not apparent how many groups should be there. There are several options how to select a number of clusters, e.g. the elbow method based on within clusters sum of squares, Calinski 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.

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 indicates that the object is well matched to its own cluster and poorly matched to 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)

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.

Classification models

As it has been previously mentioned, 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 are easy to implement 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 it. The random forests model need also a much longer time to train them.

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 a class of models called generalized linear models (Zumel and Mount, 2014). 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 is estimated by an equation that is linear in the regression parameters:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$  \hspace{1cm} (4)

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), but linear regression doesn’t 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.

We can express logistic regression by logistic function (James et al., 2013).

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}}$$

(5)

which can be further transformed into form:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}.$$  

(6)

The expression $$\frac{\pi(x)}{1-\pi(x)}$$ is called the odds and can take on any value between 0 and ∞. Values close to 0 indicate very low and values close to ∞ indicate very high probability. To make the right-hand side linear, we can modify equation (6) to the form:

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

(7)

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

The unknown regression coefficients $$\beta = (\beta_1, \beta_2, ... \beta_p)$$ in equation (7) 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.

The usual practice after estimation of the model coefficients is to assess the significance of explanatory variables (Hosmer and Lemeshow, 2000). It can be done e.g. by the likelihood ratio test, which compares likelihood of model with reduced number of variables and full model. Another test is the Wald test, which calculates a $$z$$ statistic (8), which is for $$i$$-th variable computed as:

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

(8)

This $$z$$ value is then squared yielding a Wald statistic with a chi-square distribution.

An important step after the model building is assessing the fit of the model. It can be done by computation of performance measures such as accuracy, hit rate or sensitivity, but there are also some statistics specific for logistic regression, e.g. the Hosmer-Lemeshow test or several pseudo-$$R^2$$ measures.

**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 be 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 as nodes (Williams, 2011).

A great benefit of decision trees is that its structure is in a human-readable format (Lantz,
This provides an 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.

**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. The reasons for stopping of growing the tree can be e.g. 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 a 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 \( x \), then the proportion of the second class is \( (1-x) \).

**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 \( (S_2) \):

\[
GiniGain(F) = Gini(S_1) - Gini(S_2)
\]

Because the data are after split divided into more than one partition, the 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:

\[
Gini(S) = \sum_{i=1}^{n} w_i Gini(p_i)
\]

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

**Random forests**

One of the common weaknesses of simpler models is a training variance. The 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 a 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 of 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. Because 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 a 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 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:
   - randomly select a subset of mtry variables from the p total features (typically, the number of candidate predictors is \( m \approx \sqrt{p} \)),
   - 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 a 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.

In comparison to logistic regression or to single decision tree, the random forest model is difficult to interpret. One possibility to make the model clearer is to estimate the “importance” of a variable v. The values of variables are randomly permuted in the out-of-bag samples, and the corresponding decrease in each tree’s accuracy 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 prediction of 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 to help them with variable reduction (for creating smaller, faster trees) and also for business representatives who can see what is driving the dependent variable.
6 Summary of results and 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 as Gower distance matrix, k-medoids algorithm or Silhouette coefficient and classification models logistic regression, decision trees and random forests together with performance metrics like hit rate, sensitivity or AUC can be showed to students in real applications.

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.
7 List of references


Articles in scientific journals and contributions from conferences


Norms


R packages


8 List of author’s publications and research


9 Summary

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.

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 as Gower distance matrix, k-medoids algorithm or Silhouette coefficient and classification models logistic regression, decision trees and random forests together with performance metrics like hit rate, sensitivity or AUC can be showed to students in real applications.

**Key words:**
Logistic regression, decision trees, random forests, k-medoids, R, MySQL, Qlik Sense, telecommunications, customer churn
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 musí 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 předvídání toho kdo, kdy a proč se chystá opustit společnost.

Hlavním cílem doktorské disertační práce je navrhnout a implementovat systém pro předvídání odchodu zákazníků, který pomůže vybrané telekomunikační společnosti tento počet snížit a lépe porozumět zákaznické bázi. Dílčí cíle jsou: shrnutí 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 data miningu 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 metriky 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 předvídání odchodu zákazníků. Vstupní data ve formátu CSV jsou nahrána do statistického nástroje R, kde jsou zákazníci rozděleni do shluků a také odhadnuty modely logistické regrese, rozhodovacího stramu a náhodných lesů pro celý tréninkový datový soubor a jednotlivé shluky. Zákaznické charakteristiky, predikované pravděpodobnosti odchodu a důležitost proměnných jsou uloženy do relační databáze MySQL a použity pro tvorbu dashboardu pro vizualizačním nástroji Qlik Sense. Dashboard je poskytlý uživatelům jako uživatelsky přívětivý nástroj pro pochopení chování zákazníků.

Prínos doktorské disertační práce ke studijnímu oboru „Systémové inženýrství a informatika“ může být rozdělen do 3 úrovní – metodické, empirické a pedagogické. Metodický přínos práce spočívá v návrhu a verifikaci systému pro evaluaci a predikci odchodu zákazníků v oblasti telekomunikací. Navržený komplexní systém je založen na 4 hlavních prvcích:

- aplikace systémového přístupu,
- klasifikace zákazníků telekomunikační firmy ve vztahu k odchodu,
- využití lineárních i nelineárních klasifikačních modelů a shlukové analýzy,
- vizualizace výsledků.

Z empirického hlediska je přínos práce založen na verifikaci navržené metodiky nejen pro celý soubor zákazníků, ale také pro individuální segmenty zákazníků. Dalším benefitem vyplývajícím ze shlukové analýzy je definice profilů shluků a doporučení, jak jednat se specifickými zákaznickými segmenty. Detekce a porozumění důvodů odchodů zákazníků a celkový přehled o chování zákazníků umožněný díky vizualizaci je také klíčový. Z podnikového hlediska je nejdůležitějším benefitem systému finanční přínos získaný zachráněním zákazníků, kteří se chystali firmu opustit.

10 Shrnutí
Opomenut by neměl být ani pedagogický přínos práce. Kvantitativní metody z oblasti shlukové analýzy (Gowerova matice vzdáleností, metoda k-medoidů, Silhouette koeficient) a klasifikační modely (logistická regrese, rozhodovací stromy, náhodné lesy) společně s výkonnostními metrikami pro měření kvality modelů (sensitivity, precision, plocha pod ROC křivkou - AUC) mohou být ukazovány studentům jako příklad uplatnění kvantitativních modelů v podnikové praxi.

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