UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Developing an ensemble approach for predicting customer churn in telecommunication industry

Author:
Nazar TODOSHCHUK

Supervisor:
Farnoush RESHADI

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Nazar TODOSHCHUK, declare that this thesis titled, "Developing an ensemble approach for predicting customer churn in telecommunication industry" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Alone we can do so little; together we can do so much."

Helen Keller

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Developing an ensemble approach for predicting customer churn in telecommunication industry

by Nazar TODOSHCHUK

Abstract

Customer satisfaction and retention are key goals and, at the same time challenges, for most of the modern companies which try to keep up with the times. To identify and retain the customers who are most likely to 'break ties' with the company, the latter spend much financial and technological resources. Those include advanced machine learning algorithms for customer churn prediction. This thesis explores a number of different common ML algorithms, including logistic regression, support vector machines, decision tree, random forest and XGBoost, which predict customer churn in wireless telecommunication industry. To mitigate the risks of non-accurate predictions, an ensemble algorithm is developed based on the weighted voting approach. In this thesis the performance of ensemble algorithm will be compared to those of all above mentioned to rank them by prediction accuracy and choose the best-performing one.

Acknowledgements

First of all, I would like to express my greatest gratitude to my project advisor Farnoush Reshadi for the valuable advice and feedback throughout the entire time of writing this thesis. I also want to thank my loved ones who supported and inspired me during my university study. While writing a thesis in conditions of war, it is impossible not to mention those who protect the people of Ukraine from invaders. My sincere thanks to Ukrainian army who at the risk of their own lives keep the peace in our country.

Contents

List of Figures

List of Tables

List of Abbreviations

CRM	**C**ustomer **R**elationship **M**anagment
SVM	**S**upport **V**ector **M**achines
CP	**C**ost **P**arameter
ML	**M**achine **L**earning

Dedicated to the brave nation of Ukrainians…

Chapter 1

Introduction

1.1 Motivation

The concept of customer relationship management has gained much popularity over the last decades due to its ability to significantly improve the process of interaction between the company and the customer. Since this process directly impacts the business interests of the companies, including financial ones, most of them are now trying to shift their business models from product-oriented to customer-oriented ones. From the company's perspective, the main benefit the customer relationship management can potentially bring up is the optimization of all marketing-related processes, starting from the customer acquisition and ending with the customer retention.

Intermediate processes, like customer segmentation and customer churn prediction, are no less important. The customer segmentation process aims at dividing the whole customer base into specific groups based on the number of features, including demographic, psychographic, geographic, and behavioral ones. The customer churn prediction process aims at detecting customers who are likely to quit the business with the company.

The customer churn prediction process mainly applies to those companies that provide subscription-based services. Their key goal is to retain their customers as long as possible, since retaining the existing customers is more profitable and less costly for them than acquiring the new ones. (Kim, Park, and Jeong, 2004) Those companies are devoting a lot of time, money, and efforts to building a proper churn management strategy which will help them to maximize their customers' loyalty, detect those customers who are most likely to churn in the nearest future (it is where customer churn prediction is applied), and take proactive actions to prolong their lifetime.

1.2 Problem

Since it is vitally necessary for a business to retain its customers as long as possible, it is crucial to have an extremely accurate churn prediction algorithm that will further assist in making business decisions for each individual customer. The problem with using a single classification algorithm is that it can not always be reliable. For example, in the case when customer behaviour significantly differs from the common one, the result can be unpredictable. It is also associated with the risks related to poor generalisation performance of classification algorithms having a good performance on the training data (Polikar, 2006).

When making an important decision, people usually tend to have several different opinions from different experts to come up with the final solution. Similar

considerations may be applied to the customer churn prediction problem. And it is where an ensemble approach becomes useful. The general idea behind this approach comes down to combining the results of multiple data mining models into one, which usually outperforms single models in terms of accuracy.

1.3 Objectives

The main goals of this thesis can be described as:

1. Build multiple classification algorithms for churn prediction, compare them using appropriate evaluation criteria, and define the most accurate one. Later, based on the best-performing algorithm, select the most important variables contributing to the customer churn.

2. Build an ensemble algorithm and compare its results to the best classification algorithm.

3. Cluster the customers who have eventually churned. Based on the clusterization results, describe the customer segments.

The main results of the thesis include the business recommendations for improving customer churn management strategy.

1.4 Domain background

The wireless telecommunications sector is highly competitive and even though there are three major players in the US, Verizon, AT & T, and T Mobile US, the market counts for a total of about 26000 companies. The combined annual revenue for them in 2021 was more than $300 billion (First research, 2022). The offerings of plans with unlimited text, voice, and data usage look attractive to customers, which allows to generate enormous revenues and develop an industry further. Companies entice to join them by proposing more appealing plans, discounts for newcomers, and advanced developments in the industry. In such a competitive market, a major task for small-sized companies consists in retaining the existing customers rather than acquiring new ones.

Chapter 2

Related literature

2.1 Segmentation

With the customer-centric approach taking the leadership, most of today's companies accumulate more and more customer data with new features introduced to describe the customer. Hence, traditional techniques, like multiple regression analysis, can not deliver the desired accurate and insightful result. Researchers describe the reason for this as follows: simple statistical methods use only one characteristic of the customer, but this can only help to discover potentially highly-valuable customers, and no more than that. (Chen et al., 2006)

Instead, researchers propose to use data mining techniques, like association rule, decision tree, and neural network, to build a complex segmentation model. The data mining process, also called sample learning, constructs the mapping relationship between the attribute space and conception spaces. This process is dynamic, so it automatically segments the new customers and uses this information for training. As a result of the trained segmentation model, one gets a segmentation rule that easily handles the newcomers. (Chen et al., 2006)

Another research with a different approach to customer segmentation segments the customers of one telecommunication company based on the lifetime value (LTV) metric. (Hwang, Jung, and Suh, 2004) It introduces a new approach to calculating the LTV: while other researchers use two dimensions for segmenting the customers, current value and potential value (Verhoef and Donkers, 2001), Hwang et al. add customer loyalty as the third one, which significantly changes the calculations. A number of different data mining techniques, including logistic regression, decision tree, and neural network, were applied.

However, one of the most popular clustering techniques today remains k-means. It is a centroid-based algorithm in which each data point is placed in exactly one of the k non-overlapping clusters, which are selected before the algorithm runs (Ezenkwu, Ozuomba, and Kalu, 2015) According to those authors, there is a stable algorithm to perform k-means clustering, which, if performed properly, could generate significant results. In particular, in the work mentioned above, this algorithm helped to achieve a 95% clustering accuracy.

2.2 Customer churn rate

Customer churn in telecommunication industry is defined as a process of a customer switching from one service provider to another. Since the major source of profit are customers, so customer churn plays a significant role in the survival and development of telecommunication industry. (Dahiya and Bhatia, 2015)

As mentioned before, Hwang et.al (Hwang, Jung, and Suh, 2004) decided to add a third dimension to the LTV calculation - customer loyalty. They used an interesting approach to define it. In particular, they defined this metric as the opposite of the customer churn rate. To calculate the latter they applied a number of variables, which were chosen based on their statistical significance, to a number of different models, including logistic regression, decision tree, and neural network. Even though all three models did not significantly differ in terms of misclassification rate, the logistic regression was chosen for its performance advantages.

In another research on predicting customer churn (Owczarczuk, 2010) four data mining models were used: linear regression, Fisher linear discriminant analysis, logistic regression, and decision tree. The purpose of such a choice is the possibility to easily interpret the models, as opposed to the random forest or support vector machines. The author also stated that such models are easier to debug when dealing with a huge amount of customer features. As a result of the research, logistic regression outperformed linear regression and Fisher discriminant analysis but was of the same performance as the decision tree in the short term. However, Owzcarczuk stated that the decision tree gets outdated with the introduction of new features and new customers. It needs frequent updates, consequences of which are additional time and expenditures. Thus, by taking into account this fact, the author preferred logistic regression among other models.

Artificial neural networks and their extensions are more and more frequently used to predict customer churn rates.(Tsai and Lu, 2009) Tsai and Lu took advantage of mixing neural networks in order to achieve better prediction accuracy in their work. To be more precise, they used two hybrid models, SOM+ANN and ANN+ANN, and compared them to the baseline ANN model. In the first case, it was a combination of clustering (SOM) and classification technique, while in the second - a combination of two classification techniques. In hybrid models, the first model, which is either clustering or classification, serves as an outlier detector, which filters out unrepresentative training data. Then this result is used to train the second model. As a result of the research, ANN + ANN model outperformed the baseline ANN model in terms of prediction accuracy and Type I and Type II errors, and thus showed the hybrid model to be superior to a single neural network.

2.3 Ensemble approach

There are many ways of developing an ensemble algorithm for classification tasks. For example, Kumari et.al used a soft voting classifier for diabetes mellitus prediction. (Kumari, Kumar, and Mittal, 2021) In particular, they used a logistic regression, naive Bayes, and random forest algorithms, as the base three classifiers, and then combined them into the ensemble with the soft voting classifier. The latter implied taking the average of probabilities of a customer belonging to a certain group obtained from those three algorithms. In case the resulting probability was bigger than some specific threshold, e.g. 0.5, the ensemble assigned the customer to the respective group.

Contrary to the soft voting classifier, the hard voting one deals with the binary outcomes of the models and decides on the principle of majority. In classification problems, a hard voting classifier works like a median of results obtained from the base classifiers.

A more advanced approach to developing ensemble or hybrid classifiers is presented in the research conducted by Keramati et al. (Keramati et al., 2014) The chosen base models were decision tree, k-nearest neighbor, support vector machines, and artificial neural network. For the hybrid classifier, they used a score-based technique. This technique assigned each of the interior models a score, which was calculated from the validity and reliability of the models. The latter, in their turn, were obtained from the average and variance of $F1$ *scores* of the models. To get the final decision of the classifier the score of each base model was multiplied by the respective decisions of the models, 0 or 1. Then the summation of that was multiplied by variable ϕ. The obtained result was compared to half of the overall scores. In case the first was bigger, the decision was 1 and 0, otherwise. It is important to note that by changing the variable ϕ one decides on how strict the algorithm should be, which is quite similar to soft and hard classification approaches on the conceptual level.

Chapter 3

Data

3.1 General overview

The data, used in this thesis, was kindly provided by the WPI Business School and was synthesized for educational purposes. It describes the customers of the company in the US wireless telecommunication industry. The dataset contains 2094 observations, where each observation is described by 20 features. To collect the data the CRM system was used, which allowed for gathering the following information:

- **Income:** the customer's annual income in U.S. Dollars (customers have to report their income every year, so the numbers all represent recent customers' income).
- **Gender:** 0 = male, 1 = female.
- **Account Length:** how many months the customer has been with the company. For customers that have already churned, this number shows how many months they were with the company before terminating their contract.
- **Plan Type:** the type of plan the customer has with the company:
  - **Plan 1:** Unlimited text, talk, and internet with 40 GB of Tethering.
  - **Plan 2:** Unlimited text, talk, and internet with 15 GB of Tethering.
  - **Plan 3:** Unlimited text, talk, and internet without Tethering.
  - **Plan 4:** Unlimited text and talk, 1GB internet without Tethering.
  - **Plan 5:** Unlimited text and talk, 3GB internet without Tethering.
  - **Plan 6:** Unlimited text and talk, 5GB internet without Tethering.
  - **Plan 7:** Unlimited text and talk, 7GB internet without Tethering.
  - **Plan 8:** Unlimited text and talk, 10GB internet without Tethering.
  - **Plan 9:** Unlimited text and talk, 15GB internet without Tethering.
  - **Plan 10:** Pay per use.
- **Voice Mail:** Is the customer using the voice mail plan.
- **Voice Mail Messages:** How many voice mail messages has the customer received since joining the company.
- **Day Minutes:** How many minutes has the customer talked during the day on the phone since joining the company.
- **Day calls:** Number of calls the customer talked on the phone during the day since joining the company.

- **Evening Minutes:** How many minutes has the customer talked during evening hours on the phone since joining the company.
- **Evening calls:** Number of calls the customer talked on the phone during evening hours since joining the company.
- **Night Minutes:** How many minutes has the customer talked during night hours on the phone since joining the company.
- **Night calls:** Number of calls the customer talked on the phone during night hours since joining the company.
- **International Plan:** Whether the customer has signed up to use the international plan.
- **International Minutes:** How many minutes has the customer talked to an international number since joining the company.
- **International Calls:** Number of calls the customer has made to an international number since joining the company.
- **Customer Service Calls:** Number of calls the customer has made to the customer service since joining the company.
- **Has Phone:** Whether the customer has bought a phone from the company or not.
- **Phone monthly payment:** How much money in U.S. dollars is the customer paying monthly to purchase a phone from the company.
- **Phone Payment left:** How many more months does the customer have to pay their phone monthly payment to own the phone.
- **Churn:** whether the customer is still working with the company or has terminated their contract with the company.

3.2 Data exploration

To discover the data and each feature in more detail, the data exploration was done using data visualizations.

The first point of interest was the distribution of churners and non-churners in the data set.

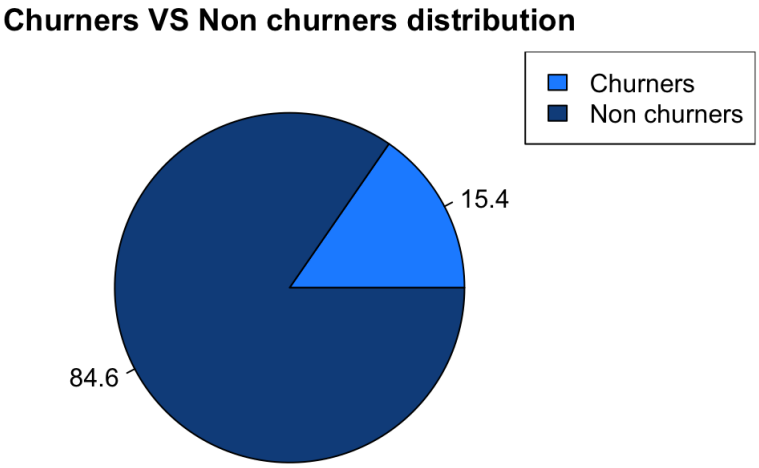

FIGURE 3.1: Distribution of churners and non-churners

The Figure 3.1 revealed that people, who eventually left the company, make up only 15.4% of the data. That is why, it can be stated that the data is imbalanced in terms of churners.

Another point of interest was the average duration of calls per month during the day, evening, and night for the customers with various plan types. Since the customers with the plan of type 10 pay per use, there was a hypothesis that they might talk less than those with an unlimited number of minutes.

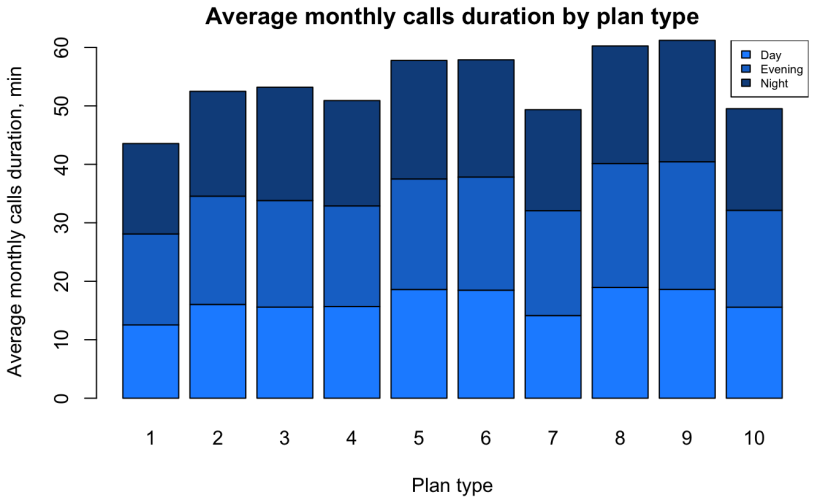

FIGURE 3.2: Average monthly calls duration by plan type

Indeed, the above Figure 3.2 showed that the average duration of calls for customers with the plan of type 10 is less than the ones for customers with all other plan types, except plan 1.

The last point of interest was the number of times customers addressed customer service or support in another way.

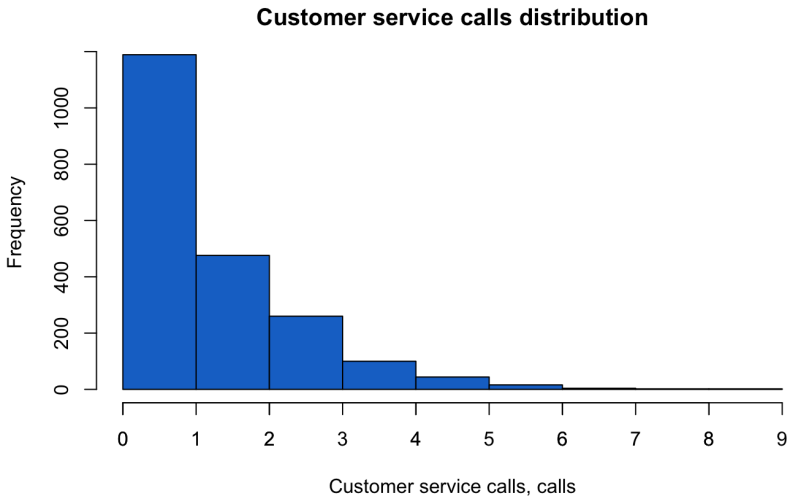

FIGURE 3.3: Customer service calls distribution

The Figure 3.3 showed that most of the customers have never called a customer support service at all.

3.3 Data preprocessing

First of all, it was decided to filter out the rows that contained missing values. There was also an empty column in the data, which got removed as well.

The next step in data preprocessing was detecting the outliers. For that reason the box plots of some features were plotted on Figure 3.4.

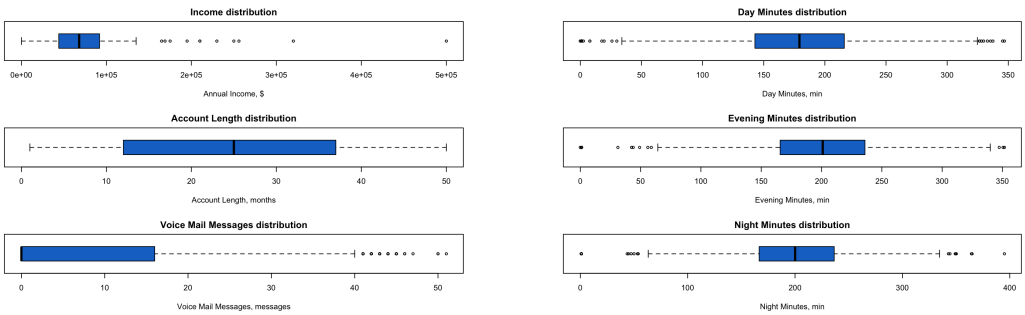

FIGURE 3.4: Boxplots for outlier detection

The above Figure 3.4 showed that there were many outliers among different columns. Thus, it was impossible to filter them out using the visual method. It is also important to note that for some features like *Voice Mail Messages* or *International Minutes* the mean was 0 since those services were rarely used by the customers. Hence, those customers, who used those services, were treated as outliers. The example is demonstrated in Figure 3.5.

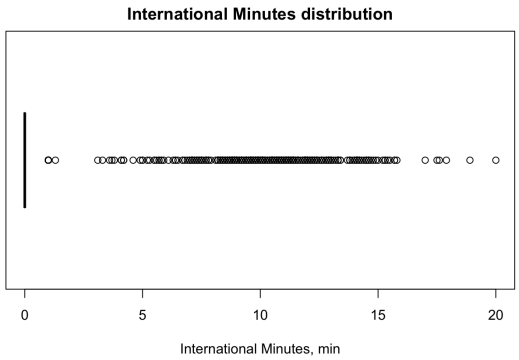

FIGURE 3.5: Boxplot for International Minutes

For further research those customers were not considered as outliers since it contradicted the logical assumptions.

For outlier detection and removal, the z-score method was used. It is based on the z-score metric, which shows how far from the population mean the data point is. More precisely, how many standard deviations the data point is below or above the mean. For a single data point, the z-score is calculated as:

$$Z = \frac{x - \mu}{\sigma} \tag{3.1}$$

where x is a single data value, μ is a population mean (the mean across the column) and σ is a standard deviation of the population. The z-score method states that the observation is considered to be an outlier if its z-score is less than -3 or higher than 3. z-score values were calculated for each cell in the data frame. Based on the procedure described above, the outliers were filtered out. As a result, 235 outliers were detected and removed, which resulted in a decrease in the total number of observations from 2093 to 1858. What is more important, it also influenced the overall number of churners in the data, cutting the number of them from 322 to 245. Since then, accuracy can no longer be an appropriate evaluation criterion for future models as the percentage of churned users became smaller.

3.4 Choosing optimal features

In order to avoid multicollinearity among features, which is one of the assumptions of logistic regression which will be used in the research, a correlation plot was built. This plot, displayed on Figure 3.6, indicates the correlation among all features in the dataset. Examining this plot one can see that there is a strong correlation between the following features:

1. *Voice Mail* and *Voice Mail Messages*
2. *International Plan*, *International Minutes*, and *International Calls*
3. *Has Phone* and *Phone Payment Left*

Voice Mail, *International Plan*, and *Has phone* were all binary variables that intuitively correlated with features that measure the number of voice mail messages sent, international minutes talked, and money that needed to be paid for the phone, respectively. For those reasons, those binary features, which are less significant,

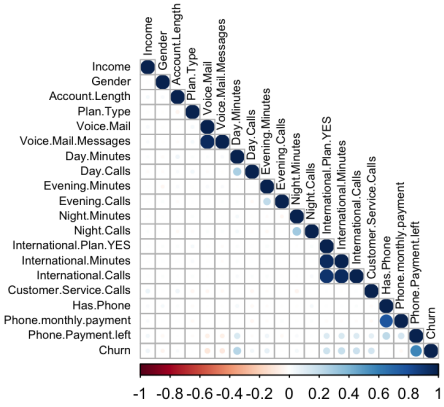

FIGURE 3.6: Correlation plot for all variables

were omitted for further research. There was also a correlation found between *Day Calls* and *Day Minutes*, *Evening Calls* and *Evening Minutes*, *Night Calls* and *Night Minutes*, which was again quite intuitive, since the more calls the users have - the more the overall duration of calls was expected. To avoid the correlation three new variables were introduced by dividing the minutes by the number of calls:

1. *dcall_avg*: the average call duration in the daytime,
2. *ecall_avg*: the average call duration in the daytime,
3. *ecall_avg*: the average call duration in the evening time.

Unfortunately, all of them showed statistical insignificance with respect to the churn. That is why it was decided to go with the *Day Minutes*, *Evening Minutes*, and *Night Minutes* variables as those that can explain more about the customer's behavior.

Chapter 4

Methodology

In order to train the models and later evaluate their performance, the data was split among the training and testing sub-samples by using a stratified random splitting method. This approach ensures that the distribution of the target variables, in this case, customers whose churn attribute is 1, is equal in both training and testing datasets. This approach also excludes the possibility of prediction errors related to the heterogeneous data split.

4.1 Logistic regression

Since churn rate, which is tried to be predicted, is a binary dependent variable, a simple linear probability model is not applicable. However, one can use an alternative, a binary response model, in particular. There are two types of binary response models, probit and logit, but only the last one will be explored and used in this thesis. The logistic binary response model mainly consists of two parts: in the first one it estimates the probability p of a binary outcome, and in the second links this probability p to a linear equation using logit function, also called log odds ratio, which actually transforms probabilities to real numbers from $-\infty$ to $+\infty$:

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{4.1}$$

4.2 Support vector machines

Support vector machines (SVM) is a supervised machine learning algorithm that is widely used for classification problems. The main goal of SVM is to find a hyperplane in n-dimensional space, where n is the number of features, which will separate the data points between two classes. In two-dimensional space a hyperplane takes the form of a line, while in three-dimensional - a form of a plane. The algorithm chooses a hyperplane which maximises the boundary between two classes, which must be equal from both sides. The intuition behind it is when new samples of data are provided to the algorithm, the expectation of correct classification increases.

When the problem of classification is non-linear, SVM uses a kernel function for transformation of a low-dimensional input space into a higher-dimensional one. In this thesis, a polynomial kernel of degree two was used, which computes the two-dimensional relationships between each pair of observations and then uses those relationships to create a hyperplane, which is called a support vector classifier.

There are a number of parameters, which can be adjusted to improve the performance of the algorithm. One of them is a *cost* parameter, which is used to learn and adjust how strictly the model should perform in terms of avoiding misclassifications. When the *cost* parameter is set to be low, the model tries to maximise the

boundaries and, as a result, is more tolerant to misclassifications. On the other hand, when the *cost* parameter is set to be high, the model treats the outliers in a more aggressive way, which might lead to the smaller margins and overfitting, as a result. This parameter, alongside *gamma* and $coef0$, which are used in the polynomial kernel function, are tuned using cross validation.

4.3 Decision tree

A decision tree classifier is another popular supervised machine learning algorithm that is based on the principle of continuous data splitting with respect to some parameter. The first step in building a decision tree is choosing a root node with the feature that provides the most information gain for the algorithm. This gain can be obtained from subtracting the average entropy value of the children nodes from the entropy value of the parent node, with entropy being the measure of disorder, or impurity, in the dataset. After selecting a root node, at each step, the algorithm recursively creates binary branches and calculates the impurity of each split.In case the purity reaches 100%, the algorithm stops and makes the node to be a leaf.

In order to improve the performance of the algorithm on the test data, the value of the *complexity parameter* was chosen manually. This parameter is the minimum threshold value of improvement in the relative error that should be reached in order for the split to take place. In case after splitting the error is not reduced by the value of the complexity parameter, the split is rejected. From the practical point of view, this parameter is used to avoid overfitting on the training data and improve the prediction accuracy of the test data. In this thesis, based on the Figure 4.1 below, the value of the *complexity parameter* was set to be 0.01. As a result, the cross-validation error of the decision tree was minimal.

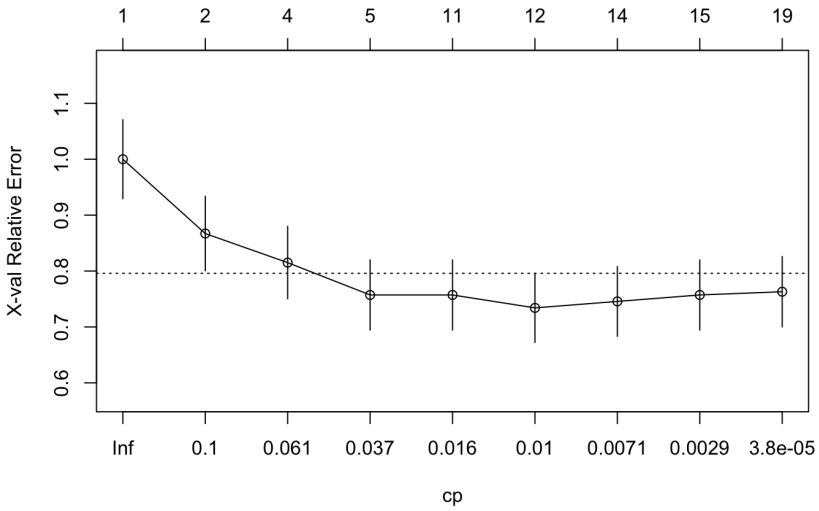

FIGURE 4.1: Optimal *cp* parameter for decision tree

4.4 Random forest

After introducing the decision tree algorithm, it is time to introduce the random forest one, which incorporates the usage of a number of uncorrelated decision trees.

The random forest classification algorithm mainly uses the power of the joint decision, rather than the individual one, which helps to minimize the risks of making a wrong final decision.

Contrary to a single decision tree, a random forest does not strongly depend on the data it was trained on, since it uses a bootstrap aggregation. The latter allows the decision trees inside the random forest to randomly sample the observations from the dataset with replacement, which results in different trees.

The reason for the forest to be called random is the way the features for training are selected. In particular, each tree inside the random forest has the ability to choose those features only from the specified randomly generated pool. This allows the decision trees inside the random forest to be as uncorrelated and as diversified as possible.

In order to improve the performance of random forest on test data, function tuneRF from the randomForest package was used. It helped to choose the best *mtry*, which is the number of randomly sampled variables at each split, for the algorithm. From Figure 4.2 below, which plotted the values of *mtry* against the out of bag error, one can assume that the value 3 of *mtry* is the best choice as such that minimizes the OOB error.

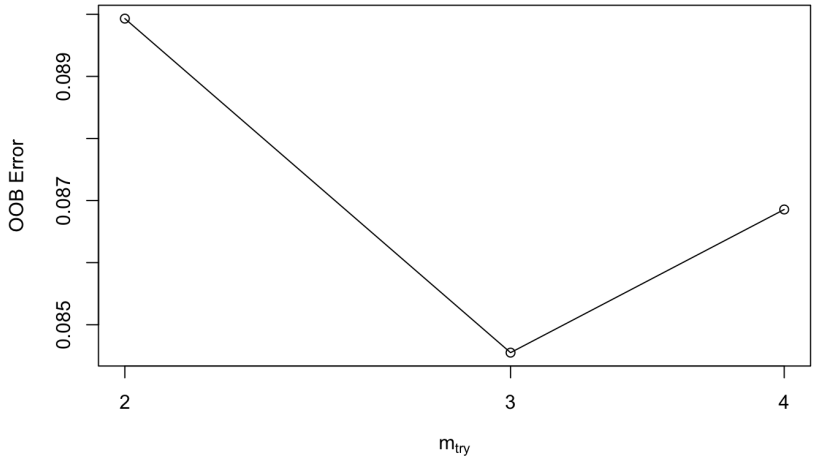

FIGURE 4.2: Optimal *mtry* parameter for random forest

4.5 XGBoost

The term XGBoost stands for extreme gradient boosting. The reason for the algorithm to be called extreme is the more effective and accurate way of gradient boosting implementation. The gradient boosting is called so because of the gradient descent algorithm application, which iteratively improves the logarithmic loss of the function by updating weights when adding a new decision tree to the model. This model is based on the ensemble learning approach, which combines the predictions of weaker models to come up with an accurately predicted target variable. The main difference between XGBoost and gradient boosting is that the first one builds the trees in parallel, while the second - sequentially. When comparing XGBoost to other tree based models, it shows to be superior in terms of prediction accuracy.

To maximise the performance of the XGBoost algorithm on the test data, the cross-validation technique was applied to choose the optimal number of *nrounds*, the maximal number of boosting iterations.

4.6 Models evaluation criteria

Since, as was mentioned before, the data is imbalanced with respect to the target variable in both training and test data, one can not use accuracy as an appropriate evaluation criterion for the models. Instead, the *F*1 *score*, as a universal metric, can be used to evaluate the performance of all models to compare them. In order to later analyze the variables that contributed to the churn prediction the most, the SHAP method was used with the term SHAP standing for Shapley Additive Explanations. This method is mainly used for interpreting the machine learning models to understand what features contributed to the model prediction the most.

As explained in the work by Lundberg et al. (Lundberg, Erion, and Lee, 2018) for tree-based models like a random forest or XGBoost, the standard feature attribution metrics are all inconsistent and are not individualized for each separate prediction. Contrary to this, the SHAP approach satisfies the properties of local accuracy, missingness, and consistency. In general, SHAP calculates the marginal contribution of each feature to the prediction of the target variable. It does so by decomposing the prediction into an additive contribution of each feature. The latter is calculated using the Shapley value, the term that came from game theory which is used to define the contribution of each player in a coalition game. In the case of this thesis, SHAP helped to understand what variables contributed the most to the prediction of customer churn.

4.7 Ensemble approach

The methodology used in the ensemble approach in this thesis was self-created after learning more about other possible realizations. It is based on voting using the *F*1 *score* metric of the implemented models. To fully understand this metric, it is necessary to get an in-depth understanding of the way the *F*1 *score* is obtained and the reason it is applicable in this thesis.

First of all, one should define the basic terms, True Positives, True Negatives, False Positives, and False Negatives:

1. **True Positives (TP):** the correctly predicted positive values, the value of the churn attribute in the dataset is 1 and the predicted value is 1.

2. **True Negatives (TN):** the correctly predicted negative values, the value of the churn attribute in the dataset is 0 and the predicted value is 0.

3. **False Positives (FP):** the incorrectly predicted negative values, the value of the churn attribute in the dataset is 0 and the predicted value is 1.

4. **False Negatives (FN):** the incorrectly predicted positive values, the value of the churn attribute in the dataset is 1 and the predicted value is 0.

The performance evaluation criteria for models include *Accuracy*, *Precision*, *Recall*, and *F*1 *score*. *Accuracy* is a ratio of correctly predicted data instances to the total

number of data instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Precision is a ratio of correctly predicted positive data instances to the total predicted positive data instances:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

Recall is a ratio of correctly predicted positive data instances to the total number of positive data instances in the dataset:

$$Recal = \frac{TP}{TP + FN} \tag{4.4}$$

And last but not the least - *F*1 *score*, which is the harmonical mean of the *Precision* and *Recall*. It is considered to be more useful than those two metrics, while being a compromise between them, especially when the data is imbalanced with respect to the target class:

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.5}$$

Returning back to the way the ensemble approach was created, each implemented model was given a specific score, which was calculated as a ratio between the *F*1 *score* of a particular model and the sum of the *F*1 *scores* of all models:

$$Model\ score = \frac{F1\ score(model)}{Total\ sum\ of\ F1\ scores\ for\ all\ models} \tag{4.6}$$

Each of the implemented models predicted whether each customer in the dataset will churn or not. The decision to churn is denoted as C and N otherwise. Then for the ensemble to make a final decision on a specific observation, the following metrics were obtained:

1. For those models, which predicted churning: $SUM(Model\ score \times C)$.
2. For those models, which predicted non-churning: $SUM(Model\ score \times N)$.

Then the following voting method was applied: in case the first sum is larger than the second one, the ensemble predicts churning, otherwise - non-churning.

4.8 K-means

K-means is one of the most popular clustering techniques used nowadays. The main idea behind it comes down to grouping similar, in terms of distance, items into k clusters. In other words, each item aims to be as close as possible to the items from the same cluster and at the same time - as far as possible from the items from the other clusters. As it was mentioned before, there exists a specific stable algorithm that helps to achieve significant accuracy results.

However, before running it, one should decide on the number of clusters k, which often reveals to be one of the most challenging tasks in the whole process. Since one can still choose this number randomly, a better decision is to use some predefined methods, like Elbow or Silhouette ones. The first one bases on the within-cluster

sum of the square ($WCSS$) metric:

$$WCSS = \sum_{C_k}^{C_n} (\sum_{d_i in C_i}^{d_m} distance(d_i, C_k)^2) \tag{4.7}$$

where C is the cluster centroids and d is the data point in each cluster. This metric tends to be the highest when the number of clusters k is equal to 1 and decreases as the number of clusters k decreases. The general rule for this method is to pick up that number of clusters k after which the values of the $WCSS$ metric start decreasing non-abruptly.

The second method bases on the silhouette coefficient. For a single observation the silhouette coefficient is calculated as:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4.8}$$

where $S(i)$ is the silhouette coefficient of the data point i, $a(i)$ is thee average distance between i and all other data points in the cluster it belongs. $b(i)$ is the average distance from i to all clusters to which i does not belong.

For each cluster, the average silhouette coefficient of its observations is calculated. After that, the average silhouette coefficient among all clusters is obtained. This metric varies from -1 to 1, with 1 being the best-case scenario when the item is as close as possible to the items from its cluster and as far as possible from the items from the other clusters, and -1 being the worst-case scenario. The general rule for this method is to pick up the number of clusters k with the highest value of the average silhouette coefficient.

After deciding on the optimal number of clusters k, one can proceed with the following algorithm:

1. Randomly initialize k centers of clusters or, in other words, centroids.
2. Calculate the distance from all items to the initialized centroids and assign those items to the closest cluster.
3. Initialize k new centroids by taking the mean of all items in the cluster.

This algorithm continues iterating until all of the items converge and changes in the positions of centroids become zero.

In this thesis, the customers who have eventually churned were clustered in order to understand the reason why they did so and try to avoid losing customers of similar characteristics.

Performing k-means clustering on the high-dimensional data might be a problem. Since every feature creates its own dimension, k-means might perform poorly by not focusing on the critical variables. To address this type of problem, the principal component analysis might be applied to reduce the number of dimensions.

Chapter 5

Results

5.1 Logistic regression

For the logistic regression the following formula was used:

$$\textbf{\textit{Churn}} \sim Income + Gender + Account.Length + Plan.Type + Voice.Mail.Messages + Day.Minutes + Evening.Minutes + Night.Minutes + International.Minutes + Phone.Monthly.Payment + Customer.Service.Calls$$

After running the regression on the training data, the following estimates of the coefficients alongside the values of the standard errors, *z-values*, and *p-values* were obtained:

	Estimate	Std. Error	z value	Pr($>\|z\|$)	
(Intercept)	-9.497e+00	8.894e-01	-10.678	< 2e-16	***
Income	5.371e-06	3.475e-06	1.546	0.12218	
Gender	-9.708e-02	1.845e-01	-0.526	0.59880	
Account.Length	-2.052e-03	6.452e-03	-0.318	0.75043	
Plan.Type	-7.364e-03	3.196e-02	-0.230	0.81775	
Voice.Mail.Messages	-2.268e-02	7.954e-03	-2.851	0.00436	**
Day.Minutes	1.952e-02	1.942e-03	10.053	< 2e-16	***
Evening.Minutes	8.412e-03	1.987e-03	4.232	2.31e-05	***
Night.Minutes	4.695e-03	1.891e-03	2.482	0.01305	*
International.Minutes	1.434e-01	1.853e-02	7.736	1.02e-14	***
Phone.Monthly.Payment	-8.578e-03	4.287e-03	-2.001	0.04540	*
Customer.Service.Calls	4.353e-01	7.082e-02	6.147	7.90e-10	***

TABLE 5.1: Logistic regression estimated coefficients

From the Table 5.1, one can state that seven of eleven variables were revealed to be statistically significant with respect to churn. In particular, variables *Voice Mail Messages*, *Day Minutes*, *Evening Minutes*, *Night Minutes*, *International Minutes*, *Phone Monthly Payment* and *Customer Service Calls* had a *p-value* less or equal to 0.05. To evaluate the accuracy of this model the regression was run on the testing data. The following confusion matrix can summarise the result of this regression:

Predicted / Actual	0	1
0	475	56
1	10	16

From this matrix one can also obtain the values of the specific evaluation criteria, which will be later used for all of the models.

Model	Logistic regression
Precision	0.61538
Recall	0.22222
F1	0.32653

TABLE 5.2: Logistic regression accuracy metrics

As can be seen from Table 5.2, the value of the *F*1 *score* metric was revealed to be quite low, which led to the conclusion that the logistic regression performed poorly in this case. However, it can still be treated as a benchmark for the other models.

5.2 SVM

For the support vector machines model with a polynomial kernel of degree two the same, as in the previous model, formula was used. After applying the cross-validation technique, the parameters *cost*, *gamma*, and $coef0$ were chosen to be equal to 0.1, 1, and 0.5, respectively. To evaluate the performance of the SVM model, it was run on the testing data, which resulted in the following confusion matrix:

Predicted / Actual	0	1
0	478	37
1	7	35

As in the case with the logistic regression, from this matrix one can obtain the values of the following evaluation criteria:

Model	SVM
Precision	0.83333
Recall	0.48611
F1	0.61404

TABLE 5.3: Support vector machines accuracy metrics

As can be seen from Table 5.3, the SVM model performed better than the logistic regression in terms of all of the criteria. In particular, the value of the *F*1 *score* for the SVM model was revealed twice as high as the value of this metric for the logistic regression.

5.3 Decision tree

For the decision tree model the same formula, as in the previous model, was used. After running the model on the train data, only the six variables were actually used in the decision tree construction: *Day Minutes*, *Customer Service Calls*, *Voice Mail Messages*, *International Minutes*, *Evening Minutes*, *Phone Monthly Payment* with *Day Minutes* being the root node. The visualization of the decision tree is provided is Figure 5.1.

Again, to evaluate the performance of the decision tree model, it was run on the testing data, which resulted in the following confusion matrix:

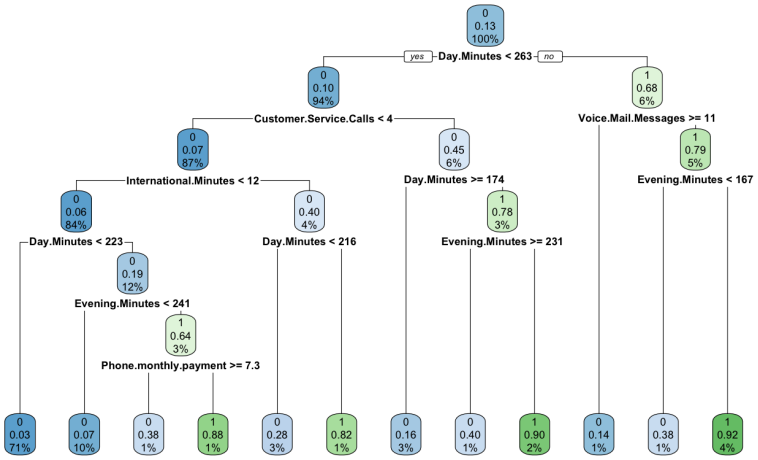

FIGURE 5.1: Decision tree for churn prediction

Predicted / Actual	0	1
0	473	30
1	12	42

The values of the evaluation criteria were as follows:

Model	Decision Tree
Precision	0.77778
Recall	0.58333
F1	0.66667

TABLE 5.4: Decision tree accuracy metrics

This model slightly outperformed the SVM in terms of *Recall* and *F1 score* but not in terms of precision. It can be explained by the high number of False Positive predictions made.

5.4 Random Forest

For the random forest model the same formula, as in the previous models, was used. The value of *mtry* was chosen to be 3 and there were 500 trees built in the forest. After running this model on the training data the most important features that affected the construction of the trees were obtained. The top five of them by the Gini index were *Day Minutes*, *Evening Minutes*, *Customer Service Calls*, *International Minutes* and *Night Minutes*. The confusion matrix for the random forest model run on the test data was the following:

Predicted / Actual	0	1
0	479	32
1	6	40

The values of the evaluation criteria were as follows:

Model	Random Forest
Precision	0.86957
Recall	0.55556
F1	0.67797

TABLE 5.5: Random Forest accuracy metrics

The random forest performance was better than the performance of a single decision tree in terms of *Precision* and *F*1 *score*. It performed worse in terms of *Recall*, which can be explained by the higher number of False Negatives predicted.

5.5 XGBoost

For the XGBoost model, the formula remained the same as in the previous models. As cross-validation proposed, the maximum number of boosting iterations was set to be 41. The confusion matrix for the XGBoost model run on the test data looked the following way:

Predicted / Actual	0	1
0	474	25
1	11	47

While the performance evaluation metrics looked as follows:

Model	XGBoost
Precision	0.81034
Recall	0.65278
F1	0.72308

TABLE 5.6: XGBoost accuracy metrics

As it can be seen from the Table 5.6, the XGBoost model outperformed all five models in terms of the main evaluation criteria - *F*1 *score*.

5.6 Variables importance

Thus, it was decided to run more analysis on this model in order to find out what variables contributed the most to the prediction of churn. To do that the SHAP value of each observation was calculated and plotted with respect to the features that are present in the data. The plot can be seen in Figure 5.2

In this plot, for every feature, one dot belongs to a certain customer in each row. The position of the x represents the feature impact on the model's prediction for a certain customer, while the color represents the value of that feature for that customer. XGBoost uses a logarithmic loss function, so the x-axis is measured in log-odds. That means that a higher SHAP value corresponds to a higher probability of customer to churn.

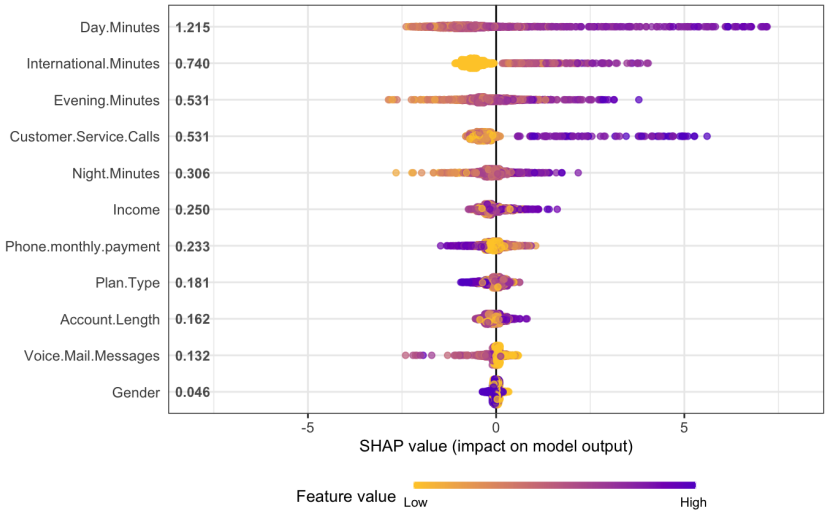

FIGURE 5.2: Variables importance in XGBoost model

Analyzing the plot, it can be said that there are four features that are the most important in predicting customer churn: *Day Minutes*, *International Minutes*, *Evening Minutes*, and *Customer Service Calls*. Each of them was analyzed separately.

Firstly, the *Day Minutes* SHAP values were plotted against the *Day Minutes* actual values.

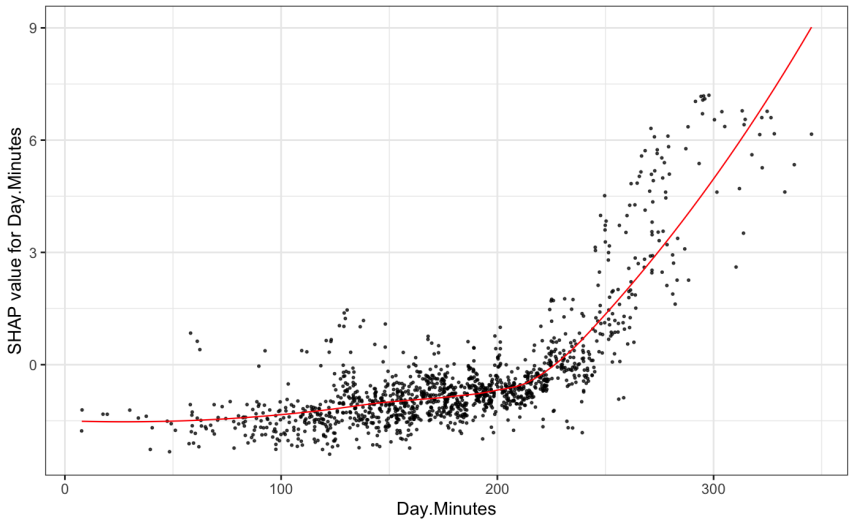

FIGURE 5.3: SHAP values for *Day Minutes* variable

Here the x-axis represents the values of *Day Minutes*, while the y-axis represents how much the feature changed the log-odds of the customer being a churner. Each dot belongs to a certain customer in the training dataset. From the Figure 5.3, one can assume that customers that have more than 250 minutes talked in the daytime are much more likely to churn than those who have less. It was also interesting to understand whether other variables influenced the importance of the *Day Minutes* variable.

For that reason, the color scale representing the amount of *Evening Minutes* talked was added to the plot.

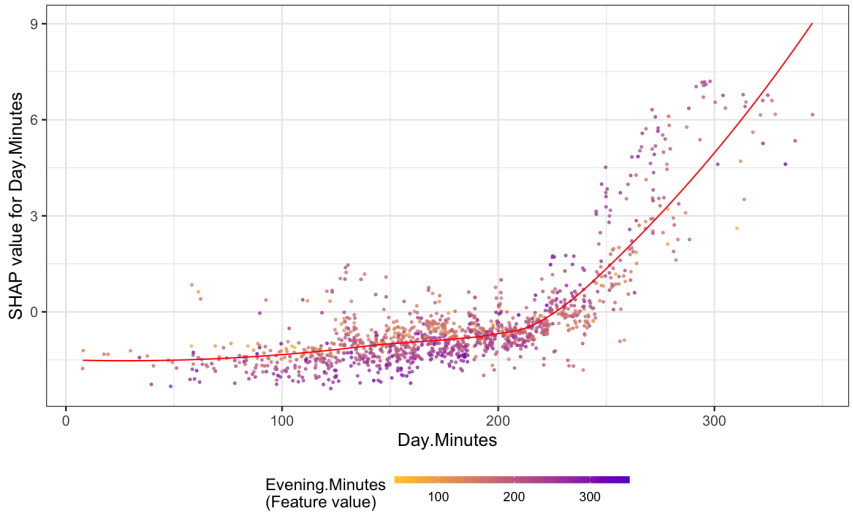

FIGURE 5.4: Impact of *Evening Minutes* feature value on SHAP values for *Day Minutes* variable

Indeed, Figure 5.4 indicates that the probability of churning for people who talked for more than 250 minutes in the daytime becomes higher if a person has also talked for more than 200 minutes in the evening time. For people that talked for less than 200 minutes in the evening time, the probability of churning decreases That is an interesting finding that could be used for business implications.

A similar plot was created for the *Customer Service Calls* variable.

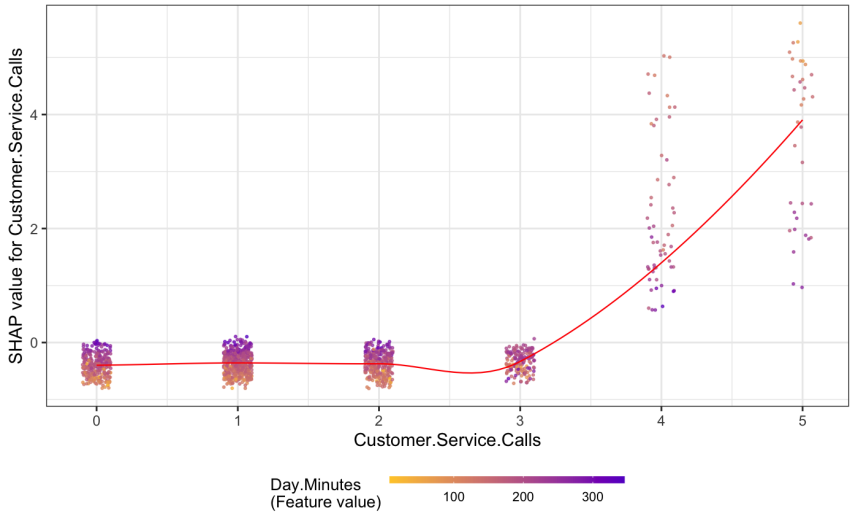

FIGURE 5.5: SHAP values for *Customer Service Calls* variable

Here, the x-axis represents the number of calls a customer had to the customer service, the y-axis represents the log-odds of the customer to churn, and the dot color indicates the number of minutes the customer talked during the daytime. From the Figure 5.5, the following information can be obtained: the probability of churn for the customers that called customer service more than three times is much higher than for those who did less. What is more important: the fewer minutes the customer has talked in the daytime, the higher the possibility of that customer quitting the business with the company.

For *International Minutes* and *Evening Minutes* variables analogous plots were created. In this plot, the x-axis represents the total number of minutes the customer

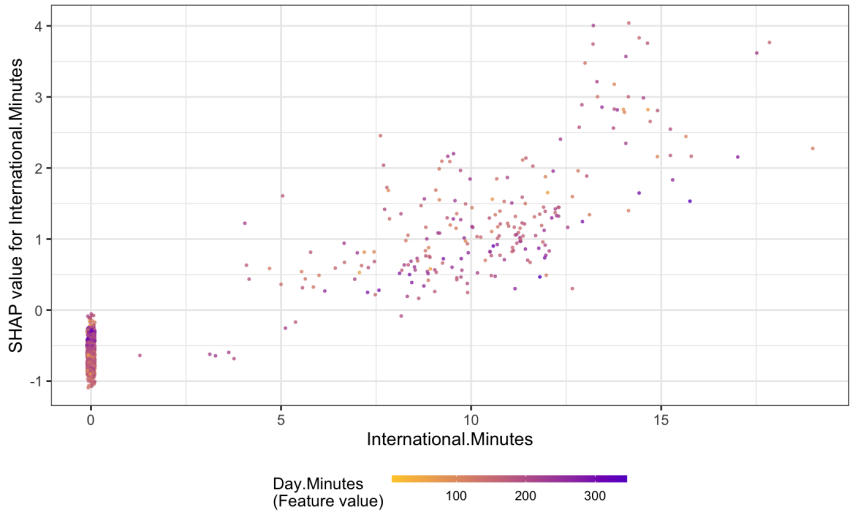

FIGURE 5.6: SHAP values for *International Minutes* variable

has talked to an international number, with the y-axis and the color of the dots representing the same as in the previous example. One can definitely say that the customers who do not talk internationally have a lower probability of churning than those who do. The general trend is that the more minutes the customers talk internationally the higher the probability of their churn. The number of minutes talked during the daytime is not connected with the number of international minutes.

The plot for *Evening Minutes* was the following:

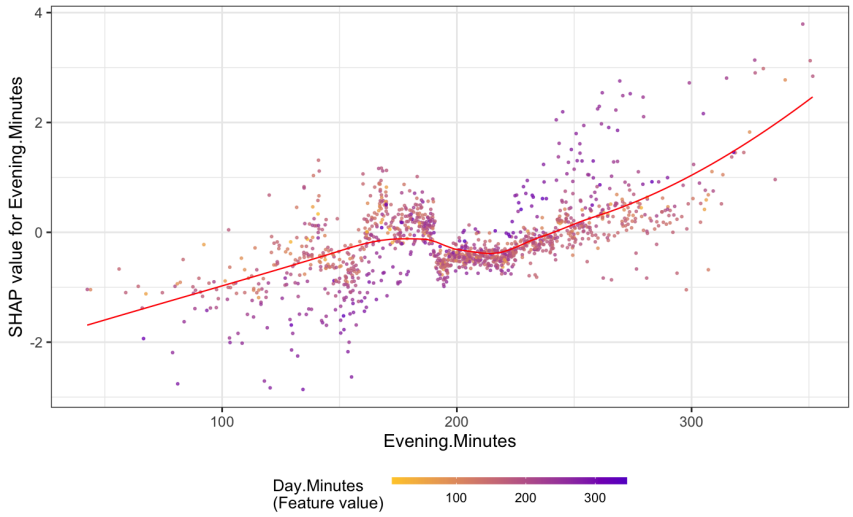

FIGURE 5.7: SHAP values for *Evening Minutes* variable

Analyzing it does not provide a clear understanding of how the probability of churning is connected to the number of minutes talked during the evening time and thus no useful insights can be obtained.

5.7 Ensemble approach

To evaluate the performance of the created ensemble, the models predicted the target variable churn for all observations in the dataset. The prediction of the ensemble was also made using the voting method described in the methodology section. The metric used to evaluate the performance of an ensemble was as in all the other models - the *F*1 *score*. To calculate the latter, the confusion matrix for the ensemble was built and looked the following way:

Predicted / Actual	0	1
0	1610	80
1	3	165

One can see that the ensemble is more accurate regarding Type I errors than the Typee II errors. Even though there were 80 False Negatives predictions made, the evaluation metrics are commendable. In particular, the *Precision* value, in this case, shows how many customers who churned were identified correctly out of all customers who churned.

Model	Ensemble
Precision	0.98214
Recall	0.67347
F1	0.79903

TABLE 5.7: Ensemble accuracy metrics

The table of *F*1 *scores* for all implemented models can be seen below:

Model	Logit	SVM	Decision Tree	Random Forest	XGBoost	Ensemble
F1 Score	0.3265	0.6140	0.6666	0.6949	0.7230	0.7990

TABLE 5.8: Comparison of *F*1 *Scores* for all models

As in the research of Keramati et al. (Keramati et al., 2014), mentioned earlier, where the hybrid ensemble approach was developed to predict customer churn in the telecommunication industry, the ensemble created in this thesis outperformed all of the implemented models in terms of the *F*1 *score*, the universal metric. It is important to note, that the researchers managed to achieve a *Precision* score similar to the above mentioned, while the *F*1 *score* and *Recall* metrics were close to 1. The possible reason for better performance in their case might hide behind the different data and as a result, explanatory variables used. That resulted in models with higher *F*1 *scores* and as a result, the better performing ensemble algorithm that used the models' results to come up with the final decision.

5.8 Clustering

Since the data used to cluster the churners was high-dimensional, an attempt to reduce the number of dimensions using the PCA algorithm was done. The importance of principal components alongside respective variance that was explained in the data are listed below:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.3470	1.1417	1.0442	1.0319	0.9891	0.91617	0.90317	0.81794	0.65156
Proportion of variance	0.2016	0.1448	0.1211	0.1183	0.1087	0.09326	0.09063	0.07434	0.004717
Cumulative Proportion	0.2016	0.3464	0.4676	0.5859	0.6946	0.78786	0.87849	0.95283	1.00000

TABLE 5.9: Descriptive statistics for principal components

The four principal components with eigenvectors higher than 1 together were able to explain nearly 58% of the variance in the dataset. Performing clustering on the latter principle components resulted in clusters with a low Silhouette coefficient. Thus, another technique, namely the feature selection was used to reduce the number of dimensions. This technique finds such a combination of variables for the different number of centroids that maximizes the Silhouette width for the possible clusters. In this case, only two combinations of variables produced reasonable results. Those combinations were:

1. *Day Minutes* and *International Minutes*.
2. *Day Minutes*, *International Minutes*, and *Customer Service Calls*.

Even though the average Silhouette width for the first combination was higher, it was decided to continue with the second one for the sake of better interpretability of obtained clusters and, as a result, better business recommendations.

The next step was to define the optimal number of clusters k. The Elbow method indicated that the best number of clusters was 3 or 4.

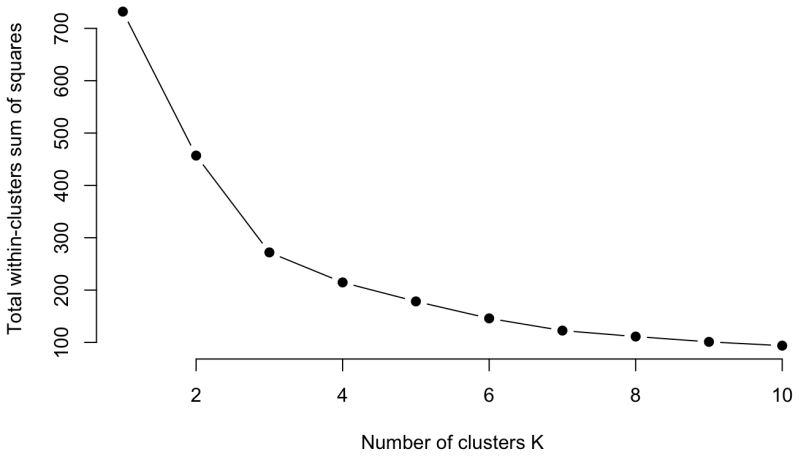

FIGURE 5.8: Elbow method to define optimal number of k

To identify this number more precisely, the Silhouette method was used. It defined k to be equal to 4. This result can be explored on the Figure 5.9. The average Silhouette width for that number of clusters k was 0.46. This result can be explored on the Figure 5.10. In addition, the total variance explained by the clustering in the dataset of upper mentioned three variables was 70.7%.

After that the four clusters were visualized on the Figure 5.11.

And the means of the scaled variables *Day Minutes*, *International Minutes*, and *Customer Service Calls* for the four clusters were provided.

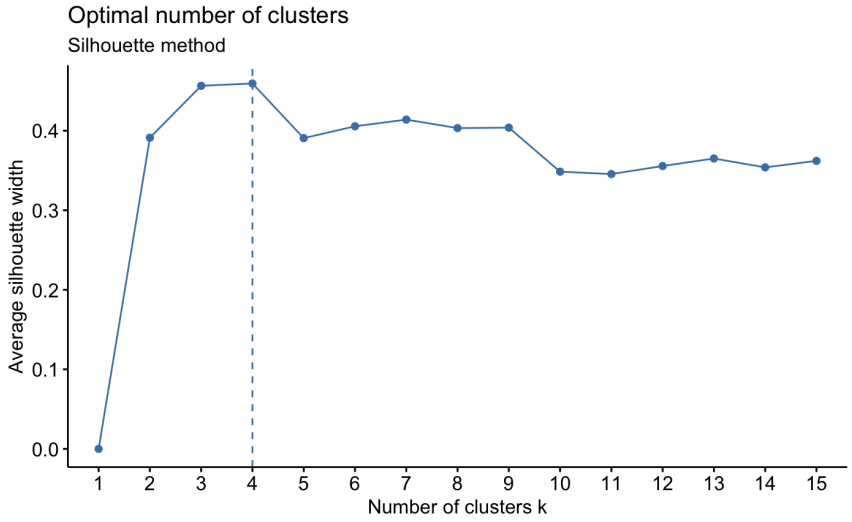

FIGURE 5.9: Silhouette method to define optimal number of k

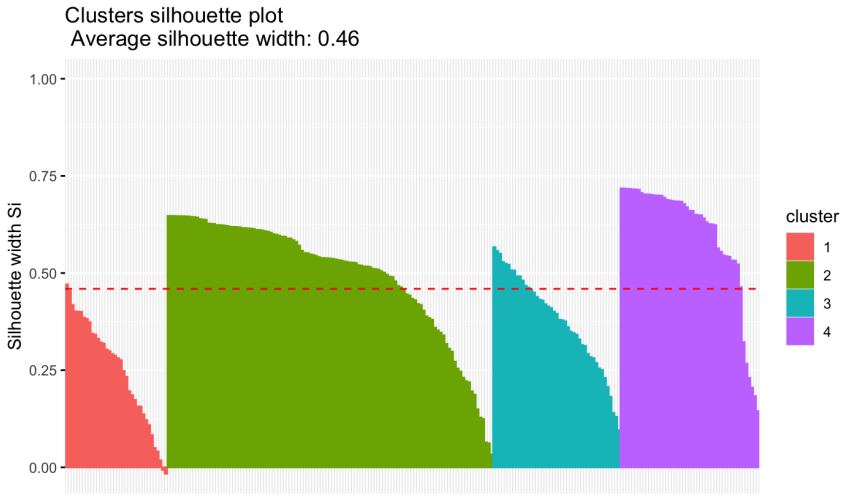

FIGURE 5.10: Clusters Silhouette Plot

Interpretation of those clusters became a customer segmentation.

Cluster 1- *International talkers*

This type of customers talked the least during the daytime and the most to international numbers. One can assume that they have friends or family abroad and mainly talk with them.

Cluster 2 - *Day talkers*

Those customers have talked the most in the daytime among other users. Also, they almost never spoke to the international numbers.

Cluster 3 - *Heavy users*

Those customers use the services of the company very actively. They never called customer support. Thus this type of customers can be considered a prototype of the ideal one for the company.

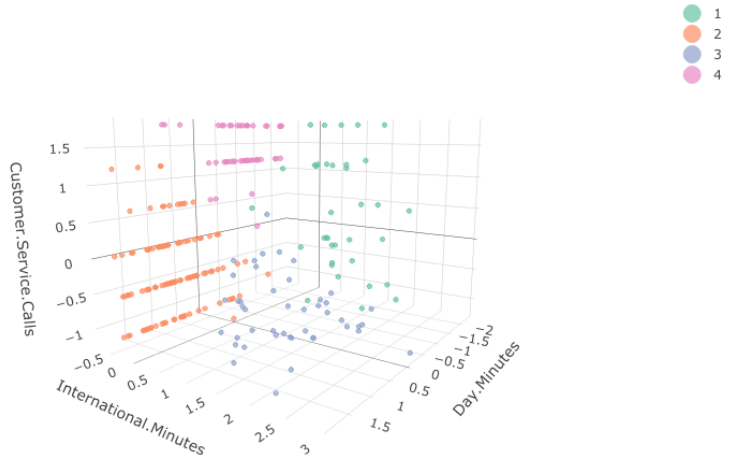

FIGURE 5.11: Visualization of 4 churners clusters in 3-dimensional space

Cluster / Variable	Day Minutes	International Minutes	Customer Service Calls
1	-1.0930388	1.3095389	0.4343530
2	0.6160607	-0.6653610	-0.4337388
3	0.4143949	1.3927979	-0.7551460
4	-1.0233746	-0.6796489	1.3923433

TABLE 5.10: Means for variables in 4 clusters

Cluster 4 - *Non — active complainers*
Those customers talked for a very small amount of time during the daytime and also do not talk to international numbers at all. In addition to being non-active talkers in the daytime, this type of customers complained the most. That is why, they can be considered the worst ones for the company. This cluster represents the process of involuntary churn (Krull, 2020), which means that the customers in this segment churn due to the reasons that are out of the control of the business. It can not take any actions to prevent their churn and it is not cost-efficient to spend the company's resources to retain these customers.

Chapter 6

Conclusions

In this thesis, five data mining classification models were built: logistic regression, support vector machines, decision tree, random forest, and XGBoost. The comparison of their prediction evaluation metrics showed the XGBoost algorithm to be the most accurate one.

Then, in order to understand what variables contributed to the prediction of customer churn in the XGBoost algorithm, the SHAP method was used. It showed, that from among all variables the four stood out significantly: *Day Minutes*, *International Minutes*, *Customer Service Calls*, and *Evening Minutes*.

Next, the ensemble algorithm that used $F1$ *scores* of implemented models and the voting method described in Chapter 4 was developed. As expected, it showed significant improvement in prediction accuracy compared to the implemented models.

Finally, based on the three variables that contributed the most value to the churn prediction, k-means clustering on customers that were churners was done. The latter resulted in four clusters, which were named and described in Chapter 5.

6.1 Business recommendations

6.1.1 Improve the customer support service

Since the customers that contacted customer support more than three times are more likely to churn, it can be assumed that they did not receive the desired service level, in particular appropriate answers to their questions. This could potentially be solved by introducing the FAQ page on the company's website along with a 24/7 support chat-bot service. Other improvement points for decreasing the churn rate could be the higher qualification level of support workers and detailed text instructions for complex cases.

6.1.2 Encourage active customers with benefits

The above results showed that with the increasing number of minutes talked in the daytime and internationally, the probability of customers quitting the company increases accordingly. To mitigate the risks of the latter, as well as to improve the relationships with those customers, the company should introduce a kind of loyalty program for them. In particular, to decide on the specific benefit, e.g. discounts, bonus points, etc., the company should conduct a series of A/B tests.

6.1.3 Collect better data

The analysis, conducted in this thesis, showed that the data the company collects is not enough for revealing causal relationships between customer behaviour and churn. Segmentation of churned customers, based on this data, also showed up to be quite a challenging task that still remained an open question. To solve this problem, the company should take full advantage of CRM usage and start collecting better data from their customers. For example, they can collect more demographic information, better and more detailed variables that capture consumers' usage behavior and transaction history. On another note, the company should try to diversify its range of plan types since they are very similar and only differ in terms of internet usage. By doing so, it could potentially increase its customers' satisfaction rate because they will choose the plan type in accordance with their specific needs. The latter could also potentially be a basis for insightful customer segmentation.

6.1.4 Conduct surveys

The analysis I ran in this thesis offered some insights as to why customers churn. However, my findings were limited due to the low number of records in the data, collinearity of the variables recorded in the dataset, and unclarity of the recorded variables. Thus, I recommend the company to collect more insights from their customers by conducting surveys both on existing customers and those who have already left. The company can survey previous company customers to better understand why they churned. They can also survey their existing customers to understand how satisfied they are with the company's services and whether they need the company to improve any aspect of their business. This could potentially help to understand the main customers' preferences and dissatisfactions which later contribute to their decision of switching to the competitor. Specific attention during survey conduct should be paid to customers that have already talked for nearly 200 minutes during the day and evening times.

6.1.5 Customer segments applications

To exploit the knowledge obtained after performing churned customers segmentation the company should take into consideration the following business recommendations. For the *Heavy customers* that never contacted the customer support, the company must take an initiative to frequently contact them via email, phone, or other appropriate channels to understand their satisfaction level and manage their expectations. For *Daily* and *International talkers*, the above mentioned surveys should be conducted on a regular basis. That would help to reveal the troubles such kinds of customers are facing and what could be done to prolong the cooperation with the company. For $Non-active\ complainers$ no actions should be taken and the company should let such customers go. Since it is assumed that they do not bring much value to the company, retaining those customers could be more costly than attempting to satisfy them.

The ensemble algorithm, that was developed in this thesis showed to outperform all of the implemented data mining models in terms of prediction accuracy. Thus, it is highly recommended to use it in the future for customer churn prediction and take proactive actions to avoid losing customers.

6.2 Future work

The ensemble algorithm, developed in this thesis, showed to outperform all of the implemented data mining models in terms of prediction accuracy. However, the comparison of its results with related works in this area indicates that there is still room for improvements. Among them for the future work it is recommended to:

1. Experiment with different combinations of ML classification techniques to include in the ensemble for predicting customer churn in the same industry.
2. Investigate the performance of developed approach in a different business industry. Make corrections taking into consideration the domain specificity.
3. Experiment with the voting method weights by including other accuracy and reliability metrics.

Bibliography

Chen, Yun et al. (2006). "Customer segmentation in customer relationship management based on data mining". In: *International Conference on Programming Languages for Manufacturing*. Springer, pp. 288–293.

Dahiya, Kiran and Surbhi Bhatia (2015). "Customer churn analysis in telecom industry". In: *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1–6. DOI: 10.1109/ICRITO.2015.7359318.

Ezenkwu, Chinedu Pascal, Simeon Ozuomba, and Constance Kalu (2015). "Application of K-Means algorithm for efficient customer segmentation: a strategy for targeted customer services". In.

First research, website (2022). *Wireless Telecommunications Services Industry Profile, website*. URL: https://www.firstresearch.com/industry-research/Wireless-Telecommunications-Services.html (visited on 05/31/2022).

Hwang, Hyunseok, Taesoo Jung, and Euiho Suh (2004). "An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry". In: *Expert systems with applications* 26.2, pp. 181–188.

Keramati, Abbas et al. (2014). "Improved churn prediction in telecommunication industry using data mining techniques". In: *Applied Soft Computing* 24, pp. 994–1012.

Kim, Moon-Koo, Myeong-Cheol Park, and Dong-Heon Jeong (2004). "The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services". In: *Telecommunications policy* 28.2, pp. 145–159.

Krull, Alex (2020). *Subscriber Retention and Understanding Involuntary vs. Voluntary Churn*. URL: https://recurly.com/blog/subscriber-retention-and-understanding-involuntary-vs.-voluntary-churn/ (visited on 05/31/2021).

Kumari, Saloni, Deepika Kumar, and Mamta Mittal (2021). "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier". In: *International Journal of Cognitive Computing in Engineering* 2, pp. 40–46.

Lundberg, Scott M, Gabriel G Erion, and Su-In Lee (2018). "Consistent individualized feature attribution for tree ensembles". In: *arXiv preprint arXiv:1802.03888*.

Owczarczuk, Marcin (2010). "Churn models for prepaid customers in the cellular telecommunication industry using large data marts". In: *Expert Systems with Applications* 37.6, pp. 4710–4712.

Polikar, Robi (2006). "Ensemble based systems in decision making". In: *IEEE Circuits and systems magazine* 6.3, pp. 21–45.

Tsai, Chih-Fong and Yu-Hsin Lu (2009). "Customer churn prediction by hybrid neural networks". In: *Expert Systems with Applications* 36.10, pp. 12547–12553.

Verhoef, Peter C and Bas Donkers (2001). "Predicting customer potential value an application in the insurance industry". In: *Decision support systems* 32.2, pp. 189–199.