Vrije Universiteit

Master Business Analytics

Customer churn prediction for a telecom company

Author:
Sabine Goezinne

VU Supervisor: Evert Haasdijk
VU Second reader: Bram Gorissen

EY Supervisor: Johan Versaevel
EY Supervisor: Avinash Ramkhelawan

April 1, 2017

Vrije Universiteit
Faculteit der Exacte Wetenschappen
De Boelelaan 1081a
1081 HV Amsterdam

EY
ITRA - Data Analytics
Antonio Vivaldistraat 150
1083 HP Amsterdam

Preface

The final challenge before completing the master Business Analytics at the VU University is writing a thesis during an internship at a company of choice. This graduation internship took place at the department IT Risk & Assurance, data analytics of EY in Amsterdam from October 2016 till March 2017.

EY offered me the possibility to complete my thesis. Therefore, I would like to thank the whole team who gave advice and helped me with my research. I would also like to thank Bram Gorissen for being my second reader. Finally, I would like to give special thanks to Evert Haasdijk for his support and guidance throughout the internship. When things did not go as planned, he was always available for questions and gave helpful advice, for which I am very grateful.

Abstract

In this research we develop a model that uses data mining techniques with the goal to increase customer retention at a telecom company. We created prediction models using the data mining techniques logistic regression, classification tree, support vector machine and k-nearest neighbor using 80% of the customers for training and 20% for testing. This resulted in the classification tree generating the highest accuracy of 0.654 and the k-nearest neighbor the lowest accuracy of 0.508. We then applied 10 fold cross validation using 90% for training and 10% for testing and generated a confidence interval for each technique. The classification tree and the support vector machine performed very similar and the most optimal, with confidence intervals between 0.6074 and 0.6134 and between 0.6107 and 0.6164 respectively. Using 80% of the customers for training generated a different accuracy than when using 90% of the customers. When removing variables with missing values from the data set, the accuracy decreased for three of the four data mining techniques. Only the logistic regression method generated a higher accuracy when variables with missing values were removed. The customers received a probability to churn and were divided into one of five groups ranging from very likely to stay loyal to very likely to churn. Using statistical analysis, the groups were compared to test if there was any significant differences between the means of the groups. A number of variables (ranging from 17 till 160) had a significant different mean when any two groups were compared. This shows that a lot of variables have an influence on the predictions. The customers were also divided into segments with self-organizing maps and hierarchical clustering. Four groups were created. Although we could not determine if using these data mining techniques can actually increase customer retention, we did provide a model that can be applied at a telecom company to test this hypothesis.

Table of Contents

1 Introduction

In most industries, companies with the highest customer retention rates earn the highest profits (Reichheld et al, 2000). Focusing on increasing customer retention can thus be very beneficial to a company. Reichheld et al (2000) found that even a small increase of 5 percent in customer retention rates could result in a profit increase ranging from 25 to 100 percent. Customer retention can increase market share and repeat sales. There is less need to acquire new customers or replace old ones, which results in fewer costs. The company will perform better, which leads to an increase in the job satisfaction of the employees and will again increase customer retention due to the better service the employees will provide. Higher profits provide companies with the resources to invest in features that increase customer value. Reichheld and Teal (2001) also find that higher customer retention can lead to enormous competitive advantage, generate bonuses in growth and productivity and can lower the cost of capital.

Due to the benefits of customer retention, companies are more interested in retaining customers instead of acquiring new customers. The costs of acquiring new customers can range from five till ten times as high compared to the costs of retaining existing customers (Chu et al, 2007). Reichheld and Teal (2001) report that doing business with uninvested strangers can be less profitable than with people you understand and trust and is more efficient and predictable. Companies conclude that the best marketing strategy is to avoid that customers switch to another company and retain as many customers as possible. This applies to all companies in different sectors, like for example health insurers, energy suppliers and telecom companies.

Customers didn't always have the option to switch to another company. In the telecom industry, this option was introduced as of July 1997 when the Telecommunications Act was adjusted, allowing not only one operator, but allowing two new operators to also offer voice telephony. As of this moment, the telecommunications market in the Netherlands was fully liberalized (Inc IBP, 2016). The telecom companies were then competing with each other and the customers had the possibility to switch to another company. This decision to switch to another company is also known as churn. An independent Dutch company ACM that supervises the laws for telecommunication, transportation,

mail and energy, researched the churn rates in the Netherlands. They found that the annual churn rate ranged approximately from 5 till 11 percent of the consumers in the period from 2008 till 2010 (Ontwerpbesluit marktanalyse vaste telefonie 2012, 2011). They also predicted the market share of some of the largest Dutch telecom companies for single call services and stated that there will not be much difference from the previous years. One of the providers will probably lose some customers and another provider will gain a few. However, they found that for multiple call services, the largest provider dramatically lost customers (20-25%) in the period from 2012 till 2015. These customers switched to different smaller companies. Their prediction for the next four years is that this percentage will continue to drop for this provider and thus leading to a high customer churn rate in the following years (Consultatie marktanalyse vaste telefonie, 2016). The churn rates can differ for each telecom company. The annual churn rate for a telecom company can average between 10 and 67 percent (Taking a next best action approach to strengthening telecom customer relationships, 2015). As the cost of retaining an existing customer is lower than the cost of obtaining a new customer, the biggest challenge of telecom companies is to control their churn rates, thus optimizing their retention rates and try to predict the likelihood of this happening for each customer (Zhao et al, 2005). Every telecom company wants to optimize their profits and optimize the retention of customers that are of value to the company. Thus can it be beneficial to telecom companies to investigate how this retention rate can be optimized.

1.1 Research questions

The retention rate can be optimized using different techniques. In this thesis, we wanted to research if data mining is appropriate for this subject. The data mining techniques will be extensively discussed in the following chapters. This thesis thus discusses the following research question:

Can data mining techniques be used to optimize customer retention at a telecom company?

To find an answer to this question the behavior of the customers is first predicted. Data mining techniques will be researched to see which technique generates the most accurate prediction. The probability to churn can be higher

for some customers compared to others. To provide the telecom company with information about this probability, the customers are divided into groups. We extended the number of groups from two (churn and retention) to five ordinal groups. Hinkin (1995) shows that the five-point scale is one of the most used scale in other studies. Each group will have the same range of 0.2. Using this scale the retention team of the telecom company can target the customers in the group that can most likely be convinced to stay with the company and will result in optimizing the retention rate. Each customer is thus divided into one of the following five groups.

1. Are very likely to stay with current telecom company
2. Are likely to stay with current telecom company
3. Doubters
4. Are likely to switch telecom company
5. Are very likely to switch telecom company

The characteristics of the customers in each group are compared to see if there are any significant differences between customers who are likely to switch and customers who are not.

Next, all the customers are divided into multiple segments based on their characteristics. Knowing if there are any characteristics that relate to high risk switchers and knowing in what segment the customers belong can be useful for a telecom company. The company can use the groups to decide which customers to target and the segmentation to decide how to target these customers.

The following sub-questions will support the research question:

- Which data mining techniques predict customer behavior the most accurately?
- Are there significant differences between the characteristics of the customers in the high risk switching group and the characteristics of the customers in the low risk switching group?
- What types of customers can be found with segmentation?

1.2 Internship

The research was completed at the company EY. EY is an international organization that consists of accountants, lawyers and advisors. The department IT Risk & Assurance consists of a team of highly qualified advisors with a drive for analyzing large datasets and a continuous focus on adding value for the customer. They support and advise numerous multinationals about analyzing and optimizing financial business processes by using data analytics.

The dataset was retrieved from the Kaggle website. Kaggle provides numerous open datasets for everybody that is willing to discover and analyze open data. The datasets are diverse and range from health and government to games and dating trends. In this thesis the dataset of a telecommunication company was researched.

1.3 Structure of the thesis

This thesis consist of six chapters and three appendices. Following this introduction chapter, chapter 2 discusses the importance of customer data and analyzing it. Examples of data mining techniques are provided and we discuss which data mining techniques we will use and why these are chosen. Chapter 3 reports the data mining techniques. First, the techniques are presented that are used for predicting the risk of churn for each customer. Based on their risk the customers are assigned to one of the five groups. Next, the techniques are described that compare the groups to see if there are any differences in characteristics of the customers. Then, the technique is showed that divides the customers into segments. Chapter 4 describes the pre-processing of the data before data mining is applied. The data is cleaned by removing data with noise or missing values that cannot be used. Chapter 5 presents the results of our research. Finally, the conclusions and recommendations can be found in chapter 6.

2 Literature research

2.1 Importance of customer data

The increase of information on the internet has provided the opportunity for customers to easily find numerous products and companies that provide them. Because of all this information, customers that were loyal to a company can now easier be persuaded to switch to another company. Nowadays companies offer similar products and use comparable technologies, but they differentiate in their business processes (Davenport, 2006). Therefore companies should try and optimize their processes by analyzing customer data. Small companies have the opportunity to create personal relationships with their customers. Because of this, they can learn more about their clients and serve them better, which can lead to loyal customers (Linoff & Berry, 2011). Large companies do not have actual personal relationships with their clients. Instead, they rely on the data that is collected to learn more about their customers. More and more companies realize that the information they can collect from their customers is a key asset and therefore transform from a product-focused to a customer-centric organization. In every industry, companies want to understand each customer individually and use this to improve their marketing, sales and customer support operations (Linoff & Berry, 2011). There are different types of marketing approaches. The goal of mass marketing is to reach a lot of customers at once and expand the customer database. However, the costs of doing this can be high. Another approach is one-to-one marketing, where the focus lies on retaining the current customers, gaining a better perspective of the customer's needs and establishing a long-term relationship with these customers (Rygielski et al, 2002).

It is important to gain as much insight into their customers' characteristics as possible. Not only the characteristics of the customers that switch companies, but also the characteristics of the customers that are loyal. The company can then learn more about the reasons that lie behind their customers' choices. To know and understand a company's customers is also known as Customer Relationship Management (CRM). It entails selecting customers that are likely to be profitable and determining which are probably not worth targeting anymore. It involves determining which product to sell to which customer and determining what channel to use (Rygielski et al, 2002).

When a company has information about customer preferences and previous behavior, the company can make more accurate predictions about future customer behavior.

2.2 Predicting customer behavior

Analyzing customer data is fundamental for a company to optimize their customer retention and, as mentioned by Davenport (2006), if a company wants to be a leader in their field. He mentions that predictive modeling is used at these companies to identify the most profitable customers and the customers who are most likely to cancel their accounts.

There are different techniques that can be used. An example of how to optimize customer retention that Davenport (2006) mentions is that of the American UPS company. They researched usage patterns and complaints from customers to predict customer defections and used the multi attribute utility theory tool for predicting the decisions that customers will make and their behavior. This tool generates a simple hierarchy from a complex problem of quantitative and qualitative factors (Min, 1994). By doing this they dramatically increased their customer retention.

Another technique that can be used is data mining. There are a number of different data mining tasks that can be used for different situations. How these techniques can be implemented will be described in the following chapter. One of the things that data mining can be used for is to analyze the customer switching behavior. Wieringa and Verhoef (2007) researched the behavior of the customers of one of the three major energy suppliers in the Dutch energy market. They used the data mining techniques logistic regression and principal component analysis. They found that the reasons for customers to switch mainly depends on the relationship quality. This entails if the customers find the company trustworthy, if other people recommend the energy company and other quality aspects of the company, like quality of the products or the service. However, there are also other aspects that influence the switching decision, like the cost of switching, the risks and the attractiveness of switching which depends on the differences between energy suppliers. Data mining can also be used to identify valuable customers and predict their future behavior, as is

mentioned in Rygielski et al (2002). Doing this enables firms to make knowledge-driven decisions.

Bronner (2009) researched if the switching behavior can be predicted at a health insurance company using data mining techniques and in specific, logistic regression. Each customer was given a switching chance between 0 and 1. The customers with a switching chance of 0,73 and higher were approached to try and retain them. He found that the created model could predict the behavior of the customers well and he found a group of customers that had a switching chance that was three times higher than that of the average customer. Marketing campaigns were applied to this group which resulted in a 20-25% significantly higher customer retention compared to a control group. This is a result that we would also like to achieve with our research.

Another example of increasing customer retention is mentioned by Mozer et al (2000). They explored different data mining techniques to predict churn for a wireless telecom company. Using, logistic regression, classification trees and neural networks, they found subscribers who had a high churn probability and recommended the telecom company to contact them. Compared to a control group the churn rate dropped with 40%. Due to this intervention the company saved approximately $417 per churnable subscriber.

This example shows that in the telecom business costs can be saved by increasing customer retention. Due to these impressive results at other companies, we wanted to research if we could increase customer retention in our data set. To do this we first had to select data mining techniques to use for our research. We found studies that compared different data mining techniques and that also prove that using data mining is very effective when predicting customer behavior.

Risselada et al (2010) showed that logistic regression models and classification trees are the most commonly used models for customer churn prediction. They compared the results of these two techniques with the results of a combination of these techniques with a bagging procedure. The classification tree with a bagging procedure provided the best predictive performance on the data from a health insurance company. In this example the classification tree performed better than the logistic regression technique.

Moeyersoms and Martens (2015) also showed that the switching behavior can be predicted at an energy company using data mining techniques. The techniques that they used were logistic regression, classification trees and support vector machines. They created transformation techniques that can be applied to the data mining techniques to include high-cardinality attributes. High-cardinality attributes are categorical attributes like for example ZIP-code, family names and bank account numbers. One of the transformation techniques used large dimensions and the support vector machine was the only one that could handle the huge feature explosion. Adding the high-cardinality attributes significantly improved the predictive performance in all three data mining techniques. Leading to more accurately identifying the likely churners and improving the efficiency of customer churn campaigns. In this example the logistic regression technique generated the most accurate prediction, followed by the support vector machine and finally the classification tree.

Using data from a wireless telecom company to predict churn, Zhao et al (2005) found that the support vector machine technique performs well when comparing it to Naïve Bays, artificial neural networks (ANN) and Decision Trees. In their research they found an 87,15% accuracy rate when using Support Vector Machine, compared to an accuracy rate of 78,12%, 62% and 83,24% for the ANN, Decision Trees and Naïve Bays respectively.

Sree Hari Rao and Jonnalagedda (2012) also researched data from customers and experimented to extract behavioral patterns for customer retention. They used a variant of K-d tree with the k-nearest neighbor algorithm and applied this to data of a health insurance company. They found that their algorithm was executed considerably faster compared to software from Salford Systems that uses classification and regression trees.

There are many different data mining techniques to choose from when predicting customer churn. There is not one single technique that seems to perform best. In one example the classification tree performs better than the logistic regression and in another example it is the other way around. Thus it was chosen to research multiple data mining techniques in this thesis. As mentioned in the literature the techniques that were most researched were logistic regression and classification trees. The support vector machine will also be

applied, because the literature stated that it outperformed the previous mentioned techniques in some cases. Finally, the k-nearest neighbor is not mentioned much in literature, but there was one case where this technique performed better than the classification and regression tree. To investigate if this will also be the case in this research, the k-nearest neighbor technique will also be applied.

Thus, we chose to use logistic regression, classification trees, support vector machines and k-nearest neighbor for predicting customer behavior. The results of the different prediction techniques are compared based on the percentage of correct predictions. We want to find the technique with the highest accuracy for this dataset. What these prediction techniques exactly entails, will be explained in the next chapter. Due to the effectiveness of these techniques at different companies, it is interesting to test these methods on the data of a telecom company.

2.3 Customer characteristics and segmentation

The data mining techniques mentioned in the previous section generate a probability that the customers will churn. Based on this probability the customers are divided into groups. With this information the retention team of the telecom company can decide which customers to target. However, only knowing which customers are more likely to churn is not a solution for customer switching. The company still needs to apply a strategy to try and retain these customers. As mentioned before, information on customer preferences and characteristics is the key to improve customer retention. When it is found which characteristics are more common for customers who are likely to churn and which characteristics are more common for loyal customers, the company can apply directed offers that match their interests. The company's marketing strategy can be adjusted for every customer or for a group of customers with the same characteristics, leading to a more successful marketing strategy. The groups are therefore compared using statistical analysis to test if there are any significant differences between the characteristics of the customers in each group.

Another approach to learn more about the characteristics of the customers that churn, is to use customer segmentation. An example of customer segmentation is described by Wieringa and Verhoef (2007). We already mentioned that they predicted customer behavior, but they also segmented the customers of the energy company. They found four segments, with three segments having a low switching chance around 20% and one segment having a switching chance of 77%. The first segment consists of a large group of probably loyal customers that do not have a clear reason to switch. The second and third segments depend their decision on switching costs or the attractiveness of other suppliers respectively. The fourth and smallest segment of only 6% of all the customers, entails the customers that have the highest chance of switching. With this information the energy company could target their marketing offers carefully.

Customer segmentation can be applied using different data mining techniques. For example, Viveros et al (1996) and Verdu et al (2004) successfully applied self-organizing maps for segmentation using data collected in the health insurance industry and data of energy customers respectively. Their study provided a classification of customers with similar characteristics and they concluded that the identification was of high quality. Using k-means clustering, Hung et al (2006) modeled the customers into five segments. They found that combining customer segmentation with churn prediction helps telecom companies to design strategies that retain more valuable customers. This is also the goal of our research.

McCarty and Hastak (2007) used data from a mail order company and a non-profit organization that solicits contributions for segmentation. They found that some segmentation techniques tended to be superior to other techniques in some situations, but not in all. The results tended to differ based on different circumstances. This research showed that there is not one segmentation technique that performs better than other techniques. Dolnicar (2002) also found that every algorithm has its drawbacks and advantages and thus must be chosen based on the characteristics of the dataset that is used. She researched different techniques for segmentation including hierarchical clustering.

There are different segmentation techniques that can be used. As the literature states, the technique should be chosen based on the dataset. In the dataset that is used in this thesis, there are a lot of customers and variables. Therefore

we chose to use the self-organizing map to perform dimensionality reduction. Viveros et al (1996) and Verdu et al (2004) mention that this technique segment the customers with a high quality. The clustering techniques that were used in multiple studies were the k-means clustering and hierarchical clustering. Mingoti and Lima (2006) show that the k-means clustering method and the hierarchical clustering presented similar performance, so we chose to use both of the clustering methods in our research. K-means clustering was used to determine the number of clusters that would be suitable for our data set. Using the most optimal number of clusters, we clustered the self organizing map by applying hierarchical clustering.

The segmentation will be performed on all the customers to investigate what kind of customers there are and how to target them. Using this information the retention team will now not only know which customers to target, but can also make an informed decision on how to target the customers. As with the prediction technique, we are interested to find how these techniques will perform on the dataset of a telecom company.

3 Data mining techniques

Linoff and Berry (2011) define data mining as "a business process for exploring large amounts of data to discover meaningful patterns and rules". People use it to make sense of the data, create theories and predict what will happen in the future. Data mining is used to make more informed decisions. Problems can be solved and opportunities can be identified and taken advantage of. The amount of data has increased so fast over the past years that it is complex to search for patterns manually. Therefore the electronical data that is used for data mining is automated by computer (Hall et al, 2011).

Milovic and Milovic (2012) describes the following 5 main data mining tasks:

- Classification and regression - This entails creating a model that predicts target variables using a set of explained variables. The goal of classification is to find a function that classifies the data in one of multiple classes, where the target variables are usually discrete values. As with regression, the target variables are usually continuous.
- Association rule - This descriptive form of data mining includes finding associations in data sets using rules or implications.
- Cluster analysis - The goal of cluster analysis is to group similar elements in the same cluster. This is a process where the best grouping is realized by identifying variables based on their similarity.
- Text mining - The data that is used is mostly unstructured with no determined format or partially structured with some linkage with parts of the data. Text mining analyzes this data and creates a method for representation of this textual data.
- Link analysis - This form of data mining examines connections between elements based on the features of the element and the connections in which the element takes part in and provides a category.

In this research different clustering, classification and regression methods are examined. This will be further explained in the following chapters.

3.1 Predictive analytics

Four different techniques were chosen for predicting the behavior of the customers, namely logistic regression, classification trees, support vector machines and k-nearest neighbor. These techniques were chosen based on previous literature research and/or because they seem like they would fit to this particular dataset. These techniques will be compared to see how each technique performs. As mentioned before, the customers will be divided into one of five groups, from very likely to stay loyal to very likely to switch to another telecom company. Each method analyzes the customer data differently. How each method analyzes and divides the customers into groups will be explained in chapter 5.

3.1.1. Logistic regression

Logistic regression is a classification method that measures the relationship between variables by estimating a probability. The customers will thus receive a probability between 0 and 1, where the customers with the highest values are most likely to switch health insurer. The probability is estimated using the following formula:

$$p(churn(x_i)|w) = \frac{1}{1 + e^{-[w_0 + \sum_{i=1}^{N} w_i x_i]}}$$

Where w_i are the weights and x_i are the independent variables. Using the maximum likelihood estimation technique, the parameters for w are estimated (Saradhi & Palshikar, 2011). This is calculated with:

$$\ln L(w\,|Y_i) = -\sum_{i=1}^{n} \ln(1 + e^{(1-2Y_i)x_i w})$$

Where Y_i can take on the values 0 or 1. This formula searches for the value of w that generates the maximum value (King & Zeng, 2001).

3.1.2. Classification trees

This method uses a recursive divide-and-conquer process and was developed and refined by Quinlan (1986). This technique can be described using the following steps:

1. First select an attribute for the root node. This node receives a branch for each possible value.

2. Split the instances into subsets. One for each branch extending from the node.
3. Repeat this recursively for each branch, using only instances that reach the branch.
4. Stop when all instances at a node have the same class.

What remains to determine is on which attribute to split on (Hall et al, 2011). This will be further explained in chapter 5.

3.1.3. Support vector machines

Support Vector Machines is a binary classification method that separates the customers into two groups. The goal is to find an optimal hyper-plane that divides the two groups. This method uses a decision function that is specified by a subset of training samples, which are called the support vectors. These vectors are the instances that are closest to the maximum margin hyperplane. In order to find this hyper-plane, the norm of the vector w is minimized, where the vector w defines the separating hyper-plane. This is the same as maximizing the margin between the two groups (Zhao et al, 2005).

$$|| w || = \sqrt{w_1^2 + \cdots + w_n^2}$$

Which is also known as the Euclidean norm. This method has the constraints:

$$x_i \cdot w + b \geq +1 \; when \; d_i = +1$$
$$x_i \cdot w + b \leq -1 \; when \; d_i = -1$$

where d_i is the distance to the closest point. This can be rearranged to the following constraint optimization problem:

$$d_i * (x_i * w + b) \geq 1$$
$$\Phi(w) = \frac{1}{2} w^T w$$

Lagrangian multipliers are used to solve this problem (Haykin et al., 2009). In this case it will thus predict if a customer will churn or retains. However, as mentioned, the goal in this research is to divide the customers in five groups. Therefore the results of this method will be a bit different interpreted, which will be explained in chapter 5.

3.1.4. K-nearest neighbors

This technique is one of the simplest and oldest methods for pattern classification that still generates competitive results (Weinberger et al, 2005). In this method a small number k of nearest neighbors are located and used together to determine the class of the test instance through a simple majority vote (Hall et al, 2011). This technique has the following pseudocode:

For every Xi

Compute distance between Xi and x

End

Order the distances from lowest to highest

Select the k nearest instances to x

Assign to x the most frequent class

where x is the new instance to be classified. Devroye et al (2013) showed that k-nearest neighbor is asymptotically optimal for large k and n with k/n -> 0.

3.2 Risk Group Analysis

The prediction of the customer behavior have placed the customers in one of the five groups. Next will the characteristics of the customers be compared using statistical analysis, to see if there are significant differences between customers who are likely to churn and customers who are likely to retain.

3.2.1.Independent t-tests

In this test, the means are calculated from two groups using $m = \frac{\sum_i x_i}{n}$ and the variances are calculated using $s^2 = \frac{\sum_i (x_i - m)^2}{(n-1)}$ where n is the population size. The normalized distance between the two groups c and t are computed using:

$$t = \frac{(m_c - m_t)}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}}$$

This follows approximately a Student distribution with

$$f = \frac{\left[\left(s_c^2/n_c\right) + \left(s_t^2/n_t\right)\right]^2}{\frac{\left(s_c^2/n_c\right)^2}{n_c - 1} + \frac{\left(s_t^2/n_t\right)^2}{n_t - 1}}$$

degrees of freedom (Baldi, & Long, 2001).

3.2.2. Wilcoxon rank-sum test

This test assumes that the sample's distribution approaches a normal distribution. A U statistic will be calculated for each group using:

$$U_x = n_x n_y + \frac{n_x(n_x + 1)}{2} - R_x$$

$$U_y = n_x n_y + \frac{n_y(n_y + 1)}{2} - R_y$$

where n is the size of the population of a group and R is the sum of the ranks. U is thus the amount of observations, when observations in one group precede or follow observations in the other group when all the scores from one group are put in ascending order. The average of the two groups x and y is defined as $\mu = \frac{n_x n_y}{2}$ and the standard deviation as $\sigma = \sqrt{\frac{(n_x n_y)(N+1)}{12}}$ where $N = n_x + n_y$. The test statistic is $z = \frac{U-\mu}{\sigma}$ and in absolute values is the test statistic $|z| = \frac{|U_x + U_y|}{\sigma}$. The null hypothesis, the two groups come from the same population, is rejected if |z| is larger or equal to the value in the z-table (Nachar, 2008).

3.2.3. Welch's corrected unpaired t-test

The test statistic for this test is given by: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$ where $\bar{X}$ is the mean, s^2 is the variance and N is the size of the corresponding group. The degrees of freedom is approximated using: $f = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2 f_1} + \frac{s_2^4}{N_2^2 v_2}}$ where $f_i = N_i - 1$ and is the degrees of freedom of the corresponding group i (Algina et al, 1994). These statistics can be used with the t-distribution to test the null hypothesis that the two groups come from the same population.

3.3 Segmentation

The customers will be placed in different segments to see if there are different types of segments. Based on the type the retention team of the telecom company can decide how to target this segment of customers. The segmentation will be performed using self-organizing maps, k-means clustering and hierarchical clustering.

3.3.1. Self organizing maps

The goal of this technique is to create a one- or two dimensional discrete map by transforming an incoming signal pattern of an arbitrary dimension. This translation is performed adaptively in a topologically ordered fashion and consists of three processes (Haykin et al, 2009). First is the neuron selected with the highest activation value: Activation value $= w_j^T x$
where w_j is the weight vector of neuron j in the network. Next determines the winning neuron the spatial location of a topological neighborhood of excited neurons using the Gaussian function:

$$h_{j,i(x)} = e^{\frac{d_{j,i(x)}^2}{2\sigma^2}}$$

where i is the winning neuron, j are the excited neurons and d_{ij} is the lateral distance between neuron i and j. Finally, the synaptic weights are adjusted leading to the increase of the individual values of the excited neurons. This is done using the weight update function:

$$w_j(n+1) = w_j(n) + \eta(n)h_{j,i(x)}(n)(x(n) - w_j(n))$$

3.3.2. K-means clustering

To determine the optimal number of clusters, the k-means clustering technique is applied to the data set. The goal of k-means clustering is to find cluster centers that minimize the Euclidian distance from each point to the cluster center that is closest. The Euclidian distance is the total of the 2-norm distance, squared. The center is selected based on the mean of a cluster (Anderson et al, 2006). This algorithm has the following pseudocode (Hall et al, 2011):

Select k potential clusters
Repeat
    Assign each point to the closest cluster
    Compute the cluster centers again based on the mean

Until convergence

The k-means clustering techniques will be applied for different values of k. For each run, the total within sum of squares will be calculated. The within sum of squares is calculated using the following formula (Grimm, 1987):

$$D_k = \sum_{i=1}^{n_k} \sum_{j=1}^{m} (x_{k,i,j} - \bar{x}_{k,j})^2$$

Where n_k is the number of samples in cluster k, m the number of variables, $x_{k,i,j}$ the value of the jth variable of sample i in cluster k, and $\bar{x}_{k,j}$ the mean value of variable j in cluster k.

The total within sum of squares is for g clusters:

$$D = \sum_{k=1}^{g} D_k$$

Tibshirani et al (2001) mention that this value will decrease as the number of clusters increases, but there is a point where the decrease flattens. This is also known as the 'elbow' and indicates the optimal number of clusters.

3.3.3. Hierarchical clustering

The hierarchical clustering method can be applied in two different ways, top-down or bottom-up (Salvador & Chan, 2004). The hierarchical clustering method can be applied in a bottom-up manner by executing the following steps (Hall et al, 2011):

1. Make each instance into a trivial mini cluster.
2. Find the two closest clusters and merge them.
3. Repeat until all clusters have been merged into a single cluster.

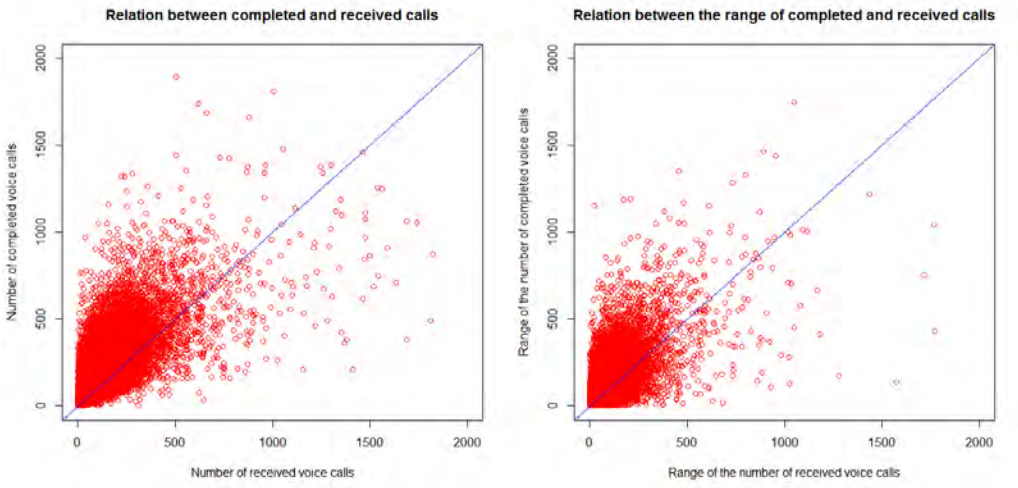
The top-down approach builds the hierarchy from its base. Then new clusters are created by uniting two clusters that were already created and that are closest to each other (Corpet, 1988). We will

4 Data description and preparation

4.1 Data Description

The open dataset was retrieved from the Kaggle website and contains data of 100,000 customers of a telecom company. Per customer there are 173 variables available, with 46 variables that contain information about the mean of the usage of the customers and there are 46 variables that contain information about the range of this same usage. The other 81 variables consists of general information. From the 173 variables there are 128 variables that are numeric and 45 that have character values.

The first step was to explore the dataset. We started with looking at the relation between the completed and received calls of the customers, as that is the first thing that comes to mind when thinking of a telecom company. It can be seen from graph 1 that the average customer calls more than that he receives phone calls. When looking at the mean, there are 84,689 of the 100,000 customers that call more than they receive calls, thus over 84%.



Graph 1a and b: Relation between completed and received calls, mean (1a) and range (1b)

The range of the customers is distributed a bit different. In this case over 72% (72,062 people) of the customers complete more calls than that they receive a call. The distribution of the mean and the range differs for some other variables as well. For example, the number of minutes of the completed calls

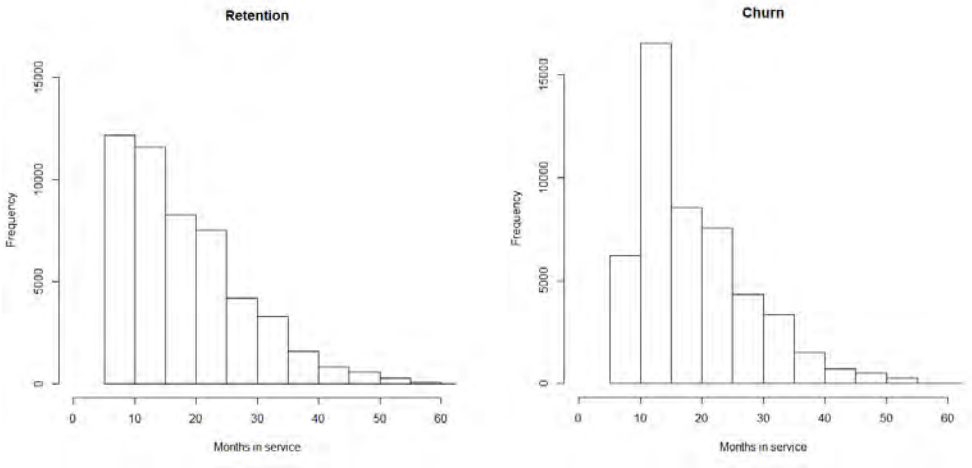
are over 80% of the time more than the number of minutes of the received phone calls. For the range, this is the case for 69% of the calls.

To see if there was a significant correlation between some of the variables we also applied two rank based correlation tests, the spearman rank and kendall's rank correlation test. Because of the large amount of variables, we chose random variables to see if there was any correlation. For example we wanted to see if there was a correlation between the months a customer was in service and if the customer churned. We found that the correlation is very low namely 0.044 (p-value $<$ 2.2e-16) with the spearman rank test and 0.036 (p-value $<$ 2.2e-16) with the Kendall rank correlation test.

When looking at the correlation between the number of days of the current equipment and the churn variable we found a correlation of 0.128 (p-value $<$ 2.2e-16) with the spearman rank test and 0.105 (p-value $<$ 2.2e-16) with the Kendall rank correlation test. When comparing the churn variable with other variables, we found almost no correlation between the variables.

4.2 Difference churn and retention

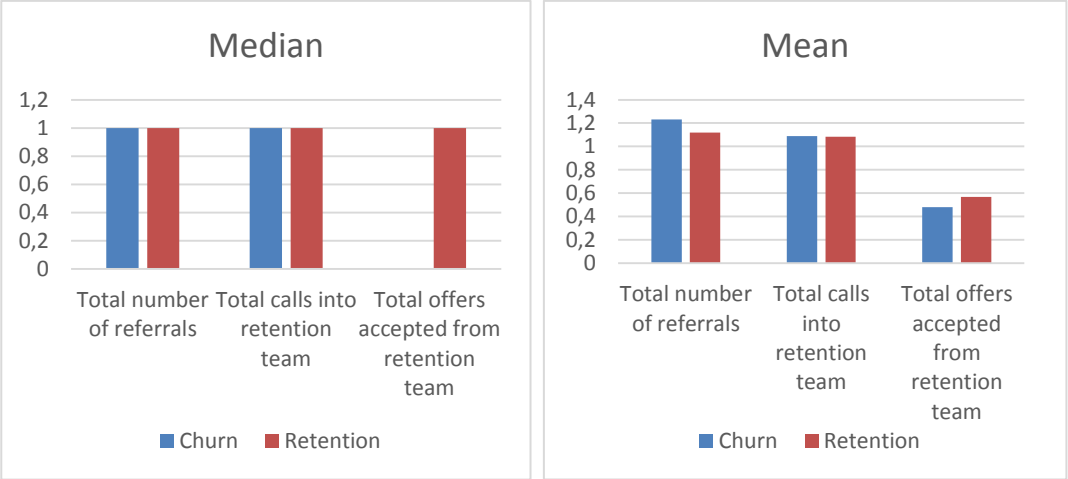
Of the 100,000 people there are 49,562 people that have churned and 50,438 people are (still) a current customer of this telecom company. It is not clear on what exact date a person joins and leaves the telecom company and we cannot see in which time period the data was collected. Instead there is a variable called months in service. The distribution of this variable is presented in the following graph.



Graph 2a and b: Frequency of the number of months in service

As can be seen from graph 2 there are a lot of people who churn when they are between ten and fifteen months in service. The retention graph shows the number of people that are currently in service. It shows that there are not many people that are longer than three years in service and that no one stays longer than five years with this telecom company.

As mentioned, the telecom company wants to retain as many profitable people as possible. The following graphs show what the telecom company has done so far to achieve this. The graphs show the median and the mean of the total number of actions per customer.



Graph 3a and b: Median and mean of actions by the retention team

There is not much difference between customers who churn and those who are loyal. We see in graph 3 that there are a bit more loyal customers who accepted offers from the retention team. This means that the offers help to retain people. However, we also see that there are customers that accepted an offer so they will stay with the company, but later decided to churn anyway. Unfortunately, we do not have information about the time between these two actions of the churned customer.

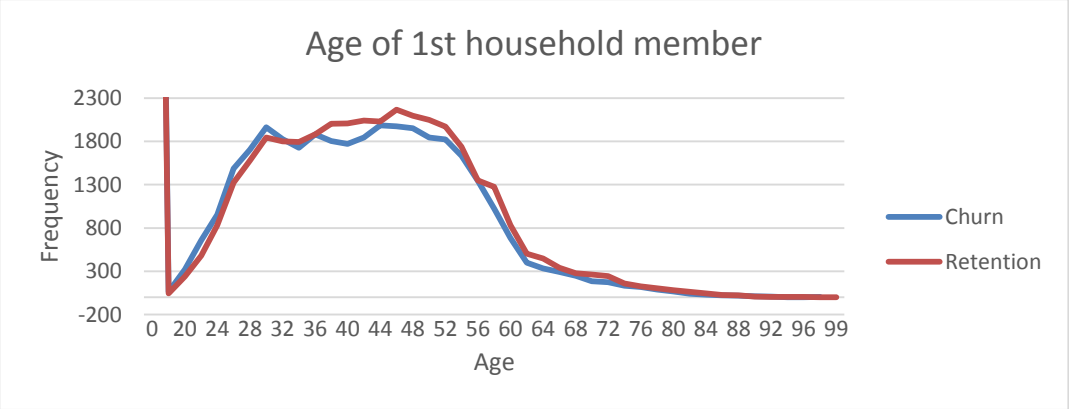
	Total Number of Referrals	Total Calls into Retention Team	Total Offers Accepted From Retention Team
Churn	2531	2931	1297
Retention	2688	1396	731

Table 1: Total values of actions from the retention team

When looking at the total amount of all customers, we see in table 1 that there are much more customers who were called by the retention team that still decided to churn. Of the 2931 customers that churned, there were 1297 offers accepted, but these customers later still decided to leave the telecom company. If we calculate the percentages, we find that 44.25% of the churned customers that received a call, accepted offers from the retention team compared to 52.36% of the loyal customers.

It is difficult to conclude if the calls from the retention team are useful or not, because we do not possess information on the time period of these actions. If the customers decided to stay a couple years after the retention call, the customers would probably be profitable and the calls would be effective. If the customers decided to leave right after the accepted offer, it would probably result in a loss of profit, due to the invested time of the retention team.

Looking further at the dataset we found that there was not much difference between the values of different variables of the customers that churned and the customers that did not. The following graph shows an example of this similarity between the frequencies in the dataset.



Graph 4: Frequency of the age of the first household member

This graph shows that the frequency of the age of the first household member is slightly higher for churners until the age of 32. After the age of 32, there are more people that retain in almost each age category. Besides the small difference, it can be seen that the frequencies are very similar.

Even though the difference is small, the same distinction can be found for the length of the residence and the handset price. In the beginning the frequency of the churned customers is higher and until a certain point the frequency of the loyal customers becomes higher and stays this way. Graph 5 shows that if the customer has a handset with a price up to $80 there is a slightly higher probability that the customer will churn and above $80 the customers is a bit more likely to stay loyal.

Graph 5: Frequency of the handset price

There are a couple of other variables that show a small difference when we compare churned customers with loyal customers. For example, when a customer has an account spending limit there is a slightly higher probability that that customer is loyal. 16.32% of the loyal customers have an account spending limit, compared to 11.51% of the churned customers, as can be seen in table 2.

	No	Yes		No	Yes	
Churn	43858	5703		88.49%	11.51%	100%
Retention	42205	8232		83.68%	16.32%	100%
Percentage difference	1.92%	18.15%		4.81%	-4.81%	

Table 2: Frequency of the account spending limit in numbers and in percentages

The same applies to the variable education level. We see that there is very little difference between churners and retained people, but when looking closer we see the same difference as with the before mentioned variables.

	1	2	3	4		1	2	3	4	
Churn	2628	2458	1263	200		40.13%	37.53%	19.29%	3.05%	100%
Retention	2720	2576	1447	230		39.01%	36.94%	20.75%	3.30%	100%
Percentage difference	1.72%	2.34%	6.79%	6.98%		1.12%	0.59%	-1.47%	-0.24%	

Table 3: Frequency of the education levels in numbers and percentages

In the description of the dataset it is not visible what each education level exactly entails. We assume that the value 1 is for lower educated customers and value 4 is for the highest educated customers. We see that the frequency of the loyal customers is higher than that of the churned customers. However there are also more loyal customers in the data set, so this is likely to happen. When we look at the percentages, we see that there are a bit more churned customers with a lower education level and a bit more loyal customers with a higher education level.

As with the before mentioned variables, the frequency of the churned customers is slightly higher in the first couple of values until a certain point and then the loyal customers have a slightly higher frequency. Even though the difference is small, it is noticeable.

4.3 Outliers and missing values

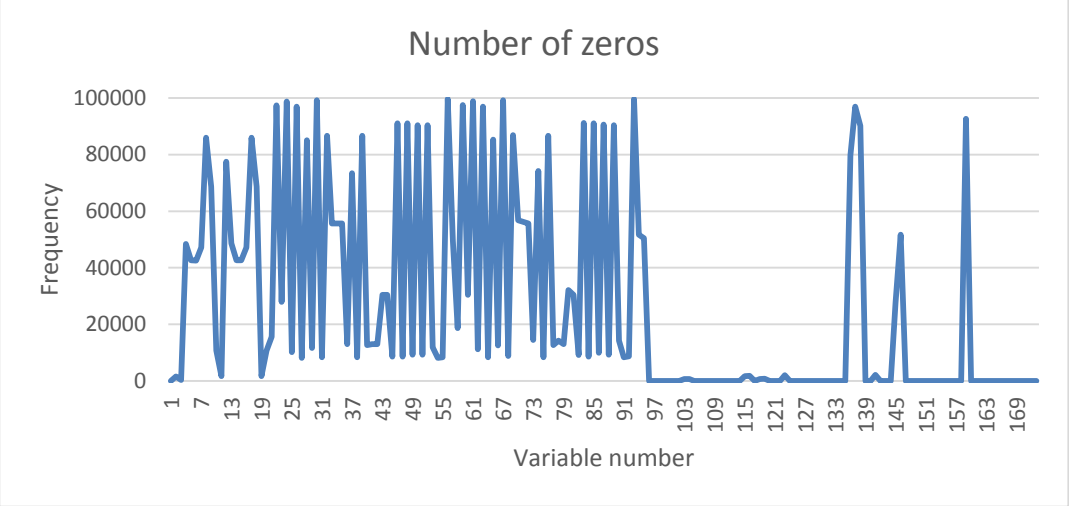
The data set contained a lot of missing values and outliers. We will first look at the outliers. The following table shows some examples for the numeric values of different variables.

	Churn						Retention				
	Min	Q1	Q2	Q3	Max		Min	Q1	Q2	Q3	Max
rev_Mean	-4	33	47	70	3843		-6	34	49	71	1223
mou_Mean	0	133	330	662	12207		0	169	381	743	7668
totmrc_Mean	27	30	43	57	410		-9	30	45	60	400
da_Mean	0	0	0	1	68		0	0	0	1	159
ovrmou_Mean	0	0	4	46	2756		0	0	2	39	4321
ovrrev_Mean	0	0	1	16	1102		0	0	1	13	896
vceovr_Mean	0	0	1	15	891		0	0	0	13	896

Table 4: Boxplot values for the first ten variables

It is visible that for both churned and loyal customers the maximum value is much higher than that of the third quantile of the entire data set. Because of the high outliers the median is always equal or lower than the mean of a variable, never higher. The outliers can also differ a lot between the churned customers and the loyal customers. The maximum value for the mean minutes of usage (mou_Mean) for the churned customers is 12207, while the maximum value for the loyal customers is only 7668 minutes of usage. The difference between the other quantiles is much smaller. In table 4 only the first seven variables are shown, but the distribution is very similar for the other numeric variables as well.

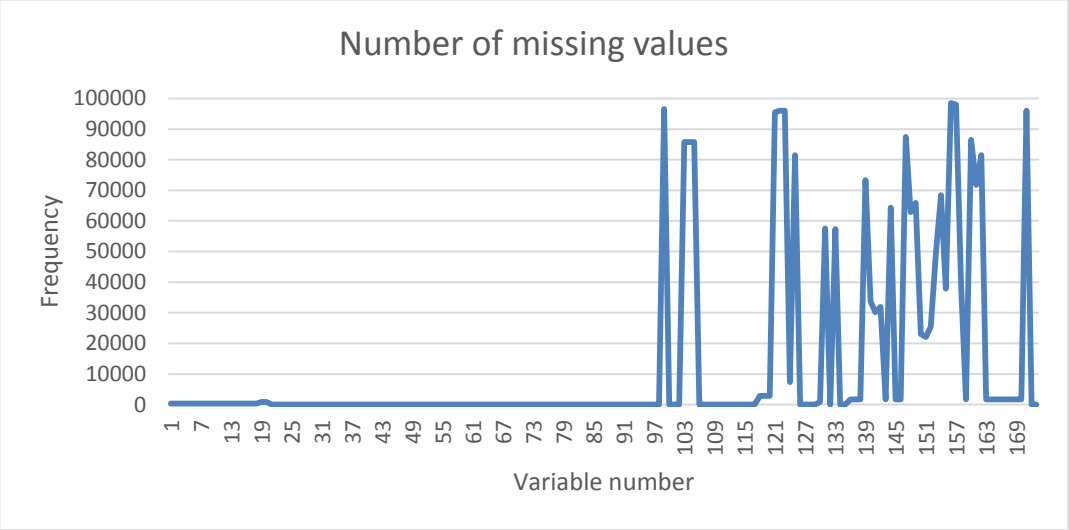
Besides outliers, there were also a lot of zeros in the dataset, which can be seen in graph 6.



Graph 6: Frequency of the number of zeros

The variables 1 until 99 are all numeric values and the variables 135 until 170 are all character values. See appendix C for the type of the remaining variables and for the description of the variable numbers. Some of the character variables were ranged for example from 0 till 4. This explains the high number of zeros for some of the character variables. However there is no clear explanation why there are so many zeros for the numeric variables. This is likely for the variable "overuse", because it is desirable for the customer that this value is zero. But this is remarkable for the number of received voice calls, namely 11676 people had zero received voice calls, but did have numerous of outgoing calls.

The number of missing values per variable is shown in graph 7. It can be seen that there are a lot of missing values for most of the character values.



Graph 7: Number of missing values per variable

There were some variables that had a relation in missing values and missed the exact amount of values as the other variables. For example, the numeric variables 1 till 18 had 357 missing variables for the same customers. These 357 customers had 74 numeric variables where the values were all zeros. The dataset contains a total of 8057 people where 74 numeric variables are all zeros.

When looking at the character variables in the dataset we saw that there were 15 character variables that had exactly 1737 missing values. For these 1737 customers 19 other character values were also missing.

A solution for all these missing values could be to remove the customers that did not have any missing variables. However, this was not an option, because there was not a single customer that had no missing values for all 173 variables. Therefore we decided to apply another technique to these missing values. When looking at the variables and the missing values we saw that we had to view every variable independently.

Removing variables with missing values

The predictive techniques that we used for this research can handle numeric variables easier than character variables. For every unique value in one single character variable, the logistic regression technique for example, creates a new

variable. The classification tree cannot handle character variables that have more than 32 different values, so we had to analyze these variables and decide what to do with them.

The variable 'credit class code' has 54 different character values. Creating a new variable for each of these values, with the logistic regression technique, greatly increases the execution time. When removing this variable from the dataset the accuracy of the model was not affected, but did lower the execution time. The variable 'communications service area' has 798 different values and also had a great effect on the execution time. There already exists a variable named area, which has fewer different values. The variable 'date of last phone swap' was removed, because there was already a variable "number of days of the current equipment", which is the same only expressed in number of months instead of a date. The variable Customer ID was removed, because this is only an index. In conclusion, we decided to remove the following variables.

ID	Description
101	Credit class code
126	Communications Service Area
133	Date of last phone swap
173	Unique customer ID

Even though some of the other variables have a lot of missing values, they seemed important for the predictions, for example, the variable 'total offers accepted from the retention team'. So we wanted to keep as much of the variables as possible in the first prediction run.

We were also interested to see if the accuracy of the prediction would improve if we didn't use the variables with a lot of missing values. We therefore decided to test the predictions in the following four steps:

1. Use almost all variables in the prediction (169 variables)
2. Use only variables that have less than 95000 missing values (162 variables)
3. Use only variables that have less than 50000 missing values (148 variables)
4. Use only variables that have less than 5000 missing values (138 variables)

Missing values of numeric variables

Next we had to decide what to do with the variables that we did want to keep in the dataset, but had missing values. A well know approach (Liu Peng, 2005) is to replace the numeric missing values with the mean or the median. As explained earlier in this chapter, there are a lot of outliers in the dataset. For this data set it seems more accurate to replace the missing values with the median. There are some numeric variables, however, that seemed to be registered differently than the other variables. These variables only seemed to be present when the value was greater than zero. We therefore decided to replace the following numeric missing values with zeros.

ID	Description
99	Number of courtesy credits
121	Total number of referrals
122	Total calls into retention team
123	Total offers accepted from retention team

We assumed that if a value was missing it would mean that the value would be zero. For example, the variable 'number of courtesy credits' would probably have a value if the customer would have received courtesy credits. We thus assumed that the customer does not have any credits and therefore replaced the missing values with zeros.

The missing values of the variable 'number of days since last retention call' were replaced with a high value, for example 99999. This is applied so this variable stays numeric, but indicates that there was no call from the retention team for a long time. We thus assume that when the value was missing no call has been made.

Missing values of character variables

A well know approach for character variables (Liu Peng, 2005) is to replace the missing values with the most frequent character. This seems appropriate for almost all character variables. There were some variables that had only one value or the values were missing. If we would replace the missing values with the most frequent character, all the values would be the same. In these cases, we assumed that a missing value does not mean that the opposite is true. For example, the variable 'Working woman in household' contains only values of 'Y' for yes. We assumed that not all the missing values therefore mean 'no'. It

could also be the case that the customer has not provided this information. The variable 'known number of vehicles' did not have a value zero. We assumed that there could also be customers that do not own a vehicle and so we also assumed that it could be possible that a missing values can mean that a customer has no car or that the customer has not provided this information yet. Therefore we created a new value named unknown. The missing values of the following variables were replaced with a new value 'Unknown'.

ID	Description
144	Mail Order Buyer
147	Working woman in household
148	Mail Responder
153	Known number of Vehicles
154	Dominant Vehicle Lifestyle
162	PC owner

There were a few character variables that have a high amount of only one value and only a few of a different value. If we would replace the missing values with the most frequent character, it would almost appear that this variable has only one value. For these three variables we also replaced the missing values with the variable 'unknown'.

ID	Description
151	Infobase match
156	Do not mail flag
157	Infobase no phone sol flag

4.4 Training and testing

The data set is divided so that 80% will be used for training and 20% for testing, which we shall call the 80/20 rule. Thus using 80,000 customers for training and 20,000 for testing the models.

The dataset will also be divided into ten subsets to use 10-fold cross validation so we can create a confidence interval for the accuracy of the predictions. Thus, in this situation 90% of the dataset (90,000 customers) will be used for training and 10% for testing (10,000 customers).

5 Results

In the following chapter the results of the used techniques in our research are presented and interpreted. First we will show the results of each predictive technique and then a comparison of these techniques. Next the results of the risk group analyses will be presented and we conclude with the results of the segmentation technique.

5.1 Predictive analytics

To measure the performance of the prediction models, a confusion matrix is used as shown in table 5. In this matrix a 1 means a customer will churn and a 0 means that a customer retains. We can assess the quality of the used models with this table.

	Predicted results	
Actual results	1	0
1	True positive	False negative
0	False positive	True negative

Table 5: Confusion matrix

If the prediction technique generates a probability instead of dividing the customers into binary classes, then we assume that the customers with a probability of higher than 0.5 are most likely to churn and shall be classified as such. Customers with a probability of 0.5 and lower shall be classified as a non-churner.

To create a 95% confidence interval for the accuracy of the prediction models, 10 fold cross validation is applied. A 95% confidence interval will also be presented for the runtime of creating the prediction model. Finally the accuracy of the prediction model will also be presented when removing some variables with missing values as explained in section 4.3.

5.1.1. Logistic regression

We start with the results of the logistic regression prediction method using the 80/20 rule as mentioned in section 4.4. The runtime of creating the model for the prediction was 2:09 minutes. Using this model, it generated a prediction which has an accuracy of 0.5078. This is only slightly higher than a random guess (50%) and is thus not an accurate prediction. The predictions of this model can be found in the following confusion matrix.

	Predicted	
Actual	Churn	Retention
Churn	6103	2219
Retention	7626	4052

Table 6: Confusion matrix, customers with a probability higher than 0.5 will churn

We see that the model predicts that a lot (7626) of customers will churn and 2219 customers to stay loyal, even though they did not.

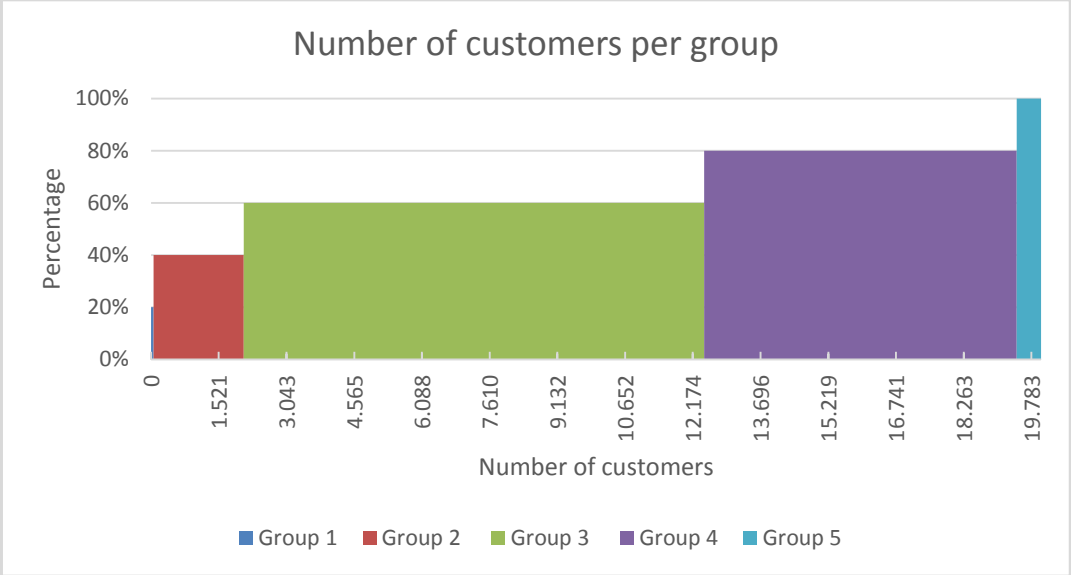
To try and increase the accuracy of this model we wanted to see what would happen if we assumed that customers with a probability of 0.7 or higher will churn instead of the before used probability of 0.5. This resulted in a prediction that generated an accuracy of 0.5942 and the resulting confusion matrix is very different, as can be seen in table 7.

	Predicted	
Actual	Churn	Retention
Churn	1431	6891
Retention	1226	10452

Table 7: Confusion matrix, customers with a probability higher than 0.7 will churn

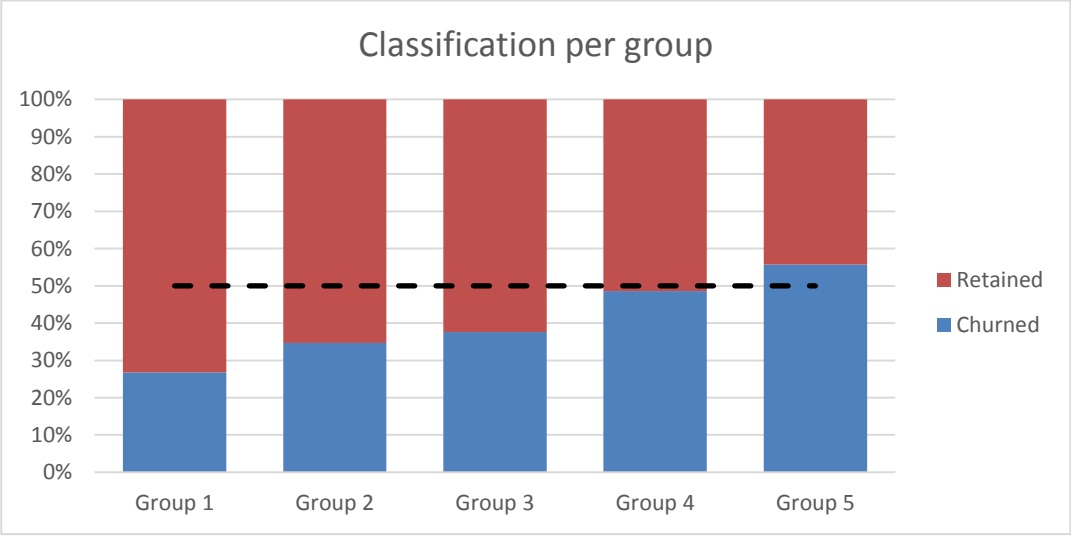
This model correctly predicts most of the customer that retain, but the prediction for the churned customers is less accurate. The actual numbers of churners is 8322 and there are 11678 loyal customers. Because there are more loyal customers in the test set than 8322 churned customers, it is likely that the accuracy of the model will increase when the probability used to classify the customers is increased. We decided to continue to use the assumption that customers with a probability higher than 0.5 will churn.

As mentioned in section 2.1, the retention team of the telecom company wants to know which customers to target to increase customer retention. To create a better overview of the likelihood to churn of each customer, we sorted the customers into five groups based on their probability to churn and can be seen in graph 8. The actual numbers can be found in appendix A.



Graph 8: Number of customers per groups

This graph shows that most of the 20,000 customers in the test set have a probability between 0.4 and 0.6. There is only a small amount of customers that are almost certain to churn (group 5) or retain (group 1).



Graph 9: Classification per group

Graph 9 shows the percentages of churned and retained customers per group. In group 1 are the customers that are very likely to retain. Of these customers 73% was classified as retained which is higher than a random percentage of 50%. In group 4 and 5 are the customers that are more likely to churn. However, the percentage of correctly classified customers is only slightly higher than random in group 5. In group 4 however, we see that a random guess performs better than our prediction. The logistic regression model correctly classified 49% of the customers, while this percentage should have been more. It is surprising that there are more customers classified as retained in group 3. We would assume this percentage to be around 50%. However, no conclusions can be made based on this group, because the distribution within group 3 is not visible. For example, if there were a number of customers with a probability just above 0.4, it would explain the difference.

Next we applied 10-fold cross validation to the data set using 90,000 customers for training, 10,000 for testing and using 169 variables. The mean of the 10 fold cross validation is 0.5931 and it has a 95% confidence interval between 0.5740 and 0.6120. The mean of the run time of this cross validation is 2:13 minutes and it has a 95% confidence interval between 1:58 and 2:28 minutes. Apparently using 90% of the dataset instead of 80% for training increases the accuracy of the model from 0.5078 to a mean of 0.5931.

To explore if we could increase the accuracy of the model further, we removed some of the variables with a lot of missing values. These variables could have a negative influence on the prediction model, because the missing values were replaced with other values. As explained in section 4.3, we will remove the variables that have more than 95,000 missing values, 50,000 and 5000 consecutively. The results are shown in table 8.

	Accuracy
Used all variables	0.5078
Used variables that have less than 95000 missing values	0.5695
Used variables that have less than 50000 missing values	0.5656
Used variables that have less than 5000 missing values	0.5966

Table 8: Accuracy when using different amount of variables

Table 8 shows that the accuracy increases when variables with missing values are removed. The accuracy is the highest when only the variables are used that have less than 5000 missing values. In this situation only 138 variables were used instead of 169 variables.

It is surprising to see that when using 90000 customers for training and 169 variables, the mean of the accuracy of the cross validation is 0.5931, but when using 80000 customers and 137 customers the accuracy is 0.5966. These results are very similar even though different approaches were applied.

5.1.2. Classification trees

Next we applied the classification tree prediction method to the data set, starting with the 80/20 rule. The runtime of creating the classification tree model is 2:04 minutes. The visual representation of this prediction model can be found in graph 10.

Graph 10: Classification tree with complexity parameter = 0.01

We see that this prediction model, is only based on the variable 'number of days of the current equipment'. If this is value is higher than 304.5 then the model predicts that a customer will churn and if the value is lower than 304.5 then a customer stays loyal. The model has an accuracy of 0.64 and the following confusion matrix.

	Predicted	
Actual	Churn	Retention
Churn	3511	4811
Retention	2390	9288

Table 9: Confusion matrix

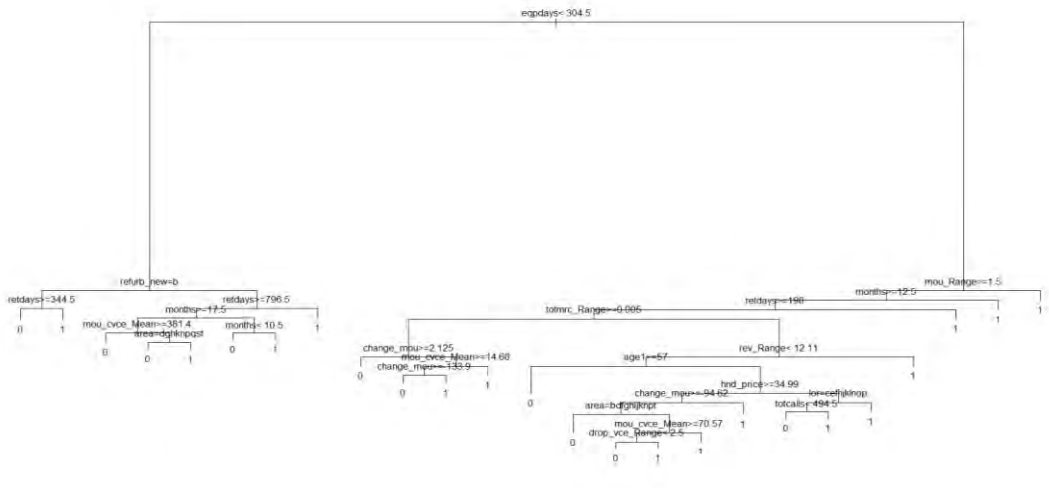
It is visible from table 9 that this prediction model classifies 9288 loyal customer correctly, but only 3511 churn customers. This model predicts that most of the customers (14,099) will retain.

The classification tree has a complexity parameter that can be adjusted to tune the model. This parameter has a small influence on the accuracy level. The following table presents the results when this parameter is adjusted.

Complexity parameter	Accuracy
0.01	0.6340
0.0014	0.6530
0.0013	0.6533
0.0012	**0.6538**
0.0011	**0.6538**
0.0010	0.6537
0.0009	0.6453
0.0001	0.6138

Table 10: Accuracy of predictions with different complexity parameters

Table 10 shows that the accuracy increases to 0.6538 if the parameter is adjusted to 0.0011 or 0.0012. After adjusting the tree with one of these complexity parameters, this resulted in the following classification tree and confusion matrix.



Graph 11: Classification tree with a complexity parameter of 0.0011

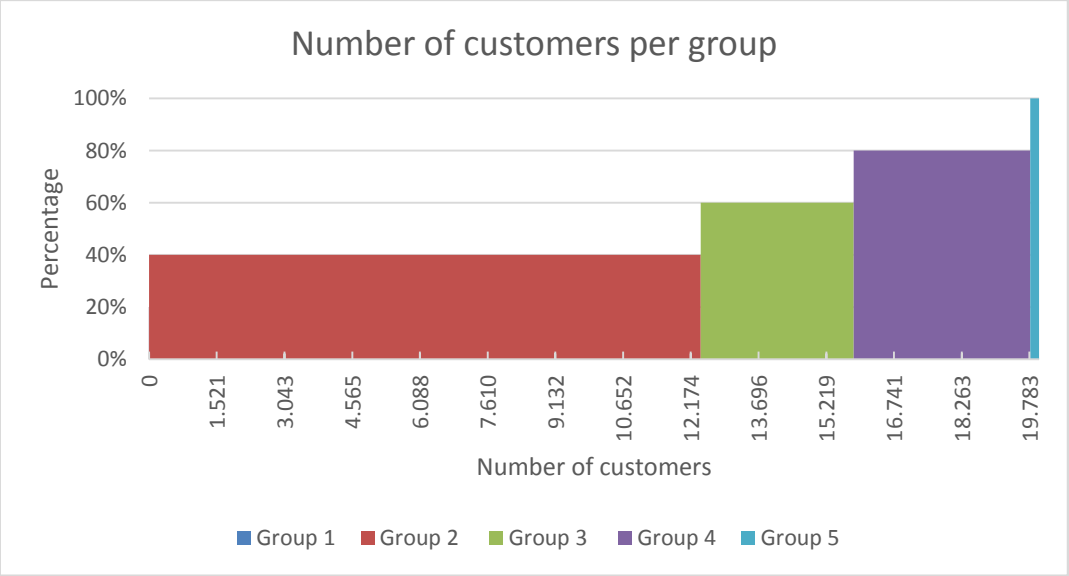
	Predicted	
Actual	Churn	Retention
Churn	3667	4655
Retention	2269	9409

Table 11: Confusion matrix

This model correctly classifies more customers than the previous model. The model used only 34 variables of the 169 variables. The variables that had the most influence on the model are the following:

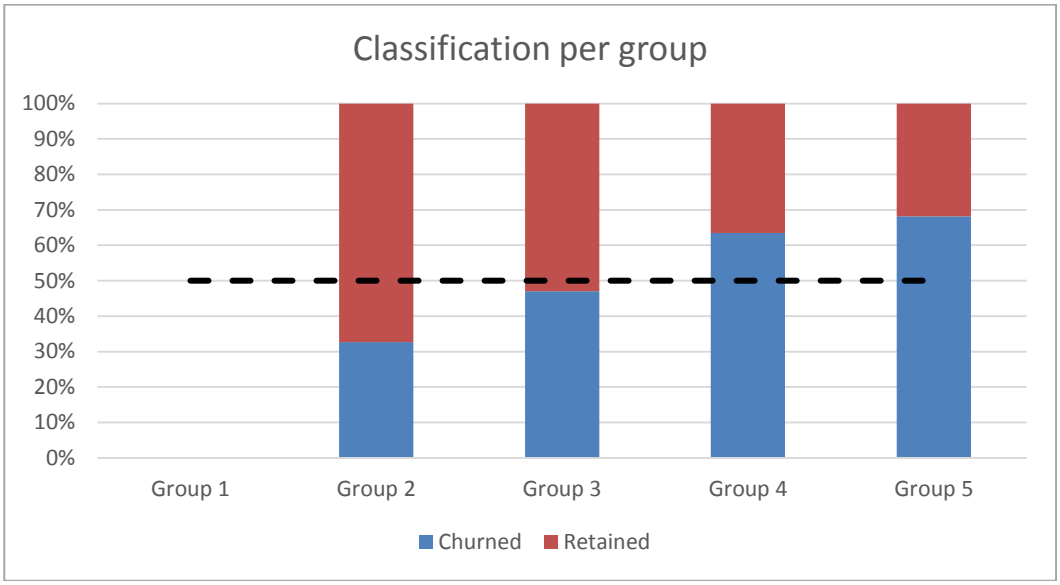
2. MINUTE_QTY
11. MINUTE_QTY
96. Months in Service
129. Handset Refurb or New
131. Previous Handset Price
132. # Handsets Issued
134. # Models Issued
171. Number of days since last retention call
172. Number of days of the current equipment

Each branch has a probability to occur and based on this, the probabilities are generated for each customer with which we divided them into one of the five groups. The results can be found in graph 12.



Graph 12: Number of customers per groups

We see that the classification tree does not generate a probability lower than 0.2, as there are no customers classified in group 1. Instead, most of the customers have a probability between 0.2 and 0.4, thus are likely to stay loyal.



Graph 13: Classification per group

It is visible from graph 13 that groups 2, 4 and 5 correctly classify more customers than a random guess. We also see that group 3 predicts in almost 50% of the customers that they will churn. Again, no conclusions can be made based on the results from group 3.

When applying 10 fold cross validation to the dataset we found a mean of 0.6104 which is lower than the initial 0.6538. The 95% confidence interval of the 10 fold cross validation is between 0.6074 and 0.6134. The mean runtime of creating the model for this cross validation is 3:46 minutes and has a 95% confidence interval between 3:41 and 3:51 minutes.
In this case, using 90% of the dataset for training instead of 80%, increases the runtime and slightly lowers the accuracy of the model.

The accuracy of the model when variables are removed with a lot of missing values are presented in table 12.

	Accuracy
Used all variables	0.6538
Used variables that have less than 95,000 missing values	0.6487
Used variables that have less than 50,000 missing values	0.6487
Used variables that have less than 5000 missing values	0.6489

Table 12: Accuracy when using different amount of variables

This table shows that the accuracy slightly decreases when fewer variables are used in the model. There is no difference in the accuracy when using variables that have less than 95,000 missing values compared to using variables that have less than 50,000 missing values, and only a difference of 0.02 when using variables that have less than 5000 missing values. Using all variables thus generates the most accurate prediction.

5.1.3. Support vector machines

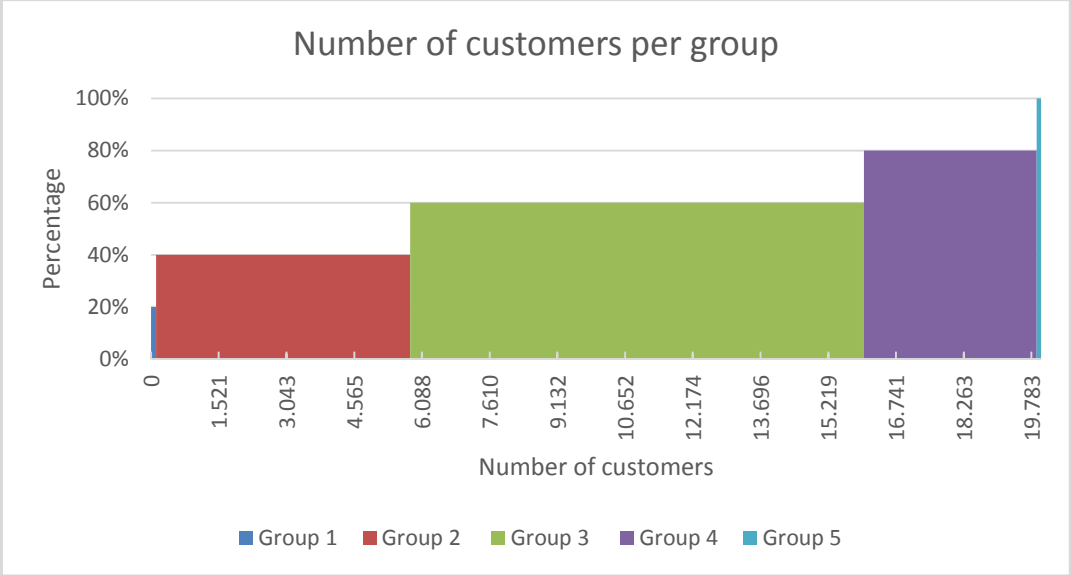
When we apply the support vector machine prediction method on the dataset, again using 80% for training and 20% for testing, we find an accuracy of 0.6156 and the following confusion matrix.

	Predicted	
Actual	Churn	Retention
Churn	4513	3809
Retention	3880	7798

Table 13: Confusion matrix

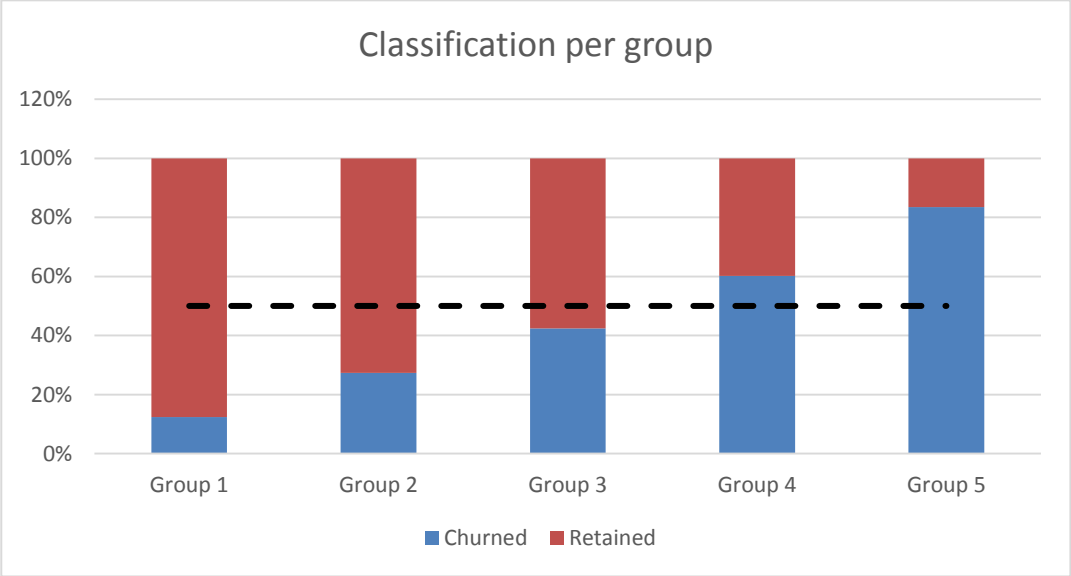
It can be seen from table 13 that there is not much difference between the incorrect predictions, namely 3809 and 3880. Most of the correct prediction classifications are for loyal customers. The execution time of creating the model is surprisingly long, namely 9:08 hours.

The separation of the customers into groups can be found in graph 14. We see that most of the customers (group 3) have a probability between 0.4 and 0.6, and only very few of the customers are most likely to churn (group 5) or retain (group 1).



Graph 14: Number of customers per groups

Graph 15 reports that especially groups 1 and 5 are correctly classified. Over 84% compared to the random guess of 50%. The percentage gain is the lowest in group 4.



Graph 15: Classification per group

A mean of 0.6135 was generated when applying 10 fold cross validation to the dataset with a 95% confidence interval between 0.6107 and 0.6164. The accuracy created with the 80/20 rule is inside the confidence interval, but higher than the mean of the 10 fold cross validation. The mean runtime to create the

model has lowered to 4:12 hours with a 95% confidence interval between 3:20 and 5:04 hours.

When we removed some variables with missing values, we see from table 14 that the accuracy decreases. This method generates a higher accuracy when all variables are used for the prediction.

	Accuracy
Used all variables	0.6156
Used variables that have less than 95000 missing values	0,6061
Used variables that have less than 50000 missing values	0,6066
Used variables that have less than 5000 missing values	0.6072

Table 14: Accuracy when using different amount of variables

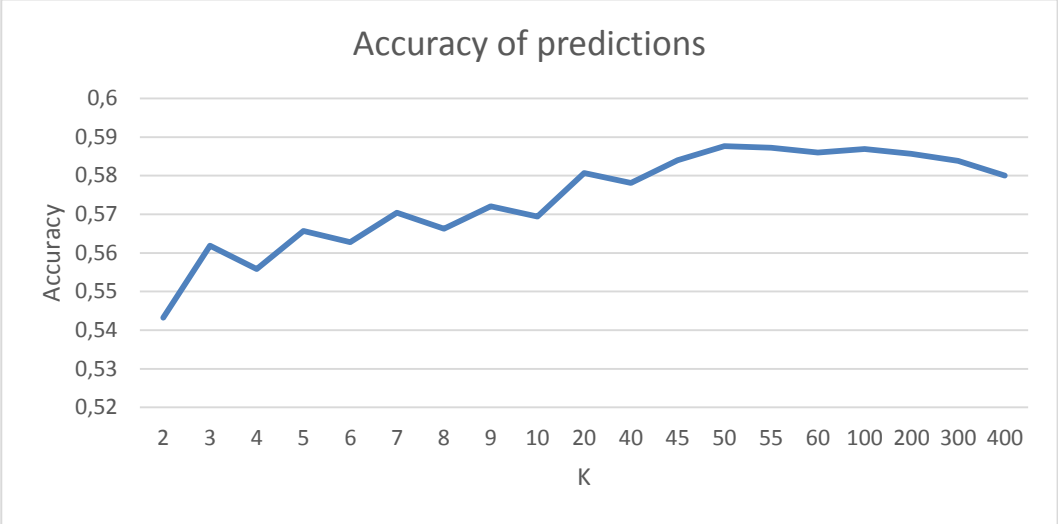
5.1.4. K-nearest neighbors

As explained in section 3.1.4, the k-nearest neighbor method computes distances to generate a prediction. Because the distance between character variables cannot be calculated, we transformed these variables using binary numbers and added dummy variables for each unique character value. This resulted in using 332 variables for the prediction. The runtime of creating the model was 52:12 minutes. We started the k-nearest neighbor method using k = 2, which resulted in the following confusion matrix with an accuracy of 0.5433.

	Predicted	
Actual	Churn	Retention
Churn	4135	4187
Retention	4948	6730

Table 15: Confusion matrix

It is visible from table 15 that this is not a very accurate prediction, so we try to improve this by changing the value of k.



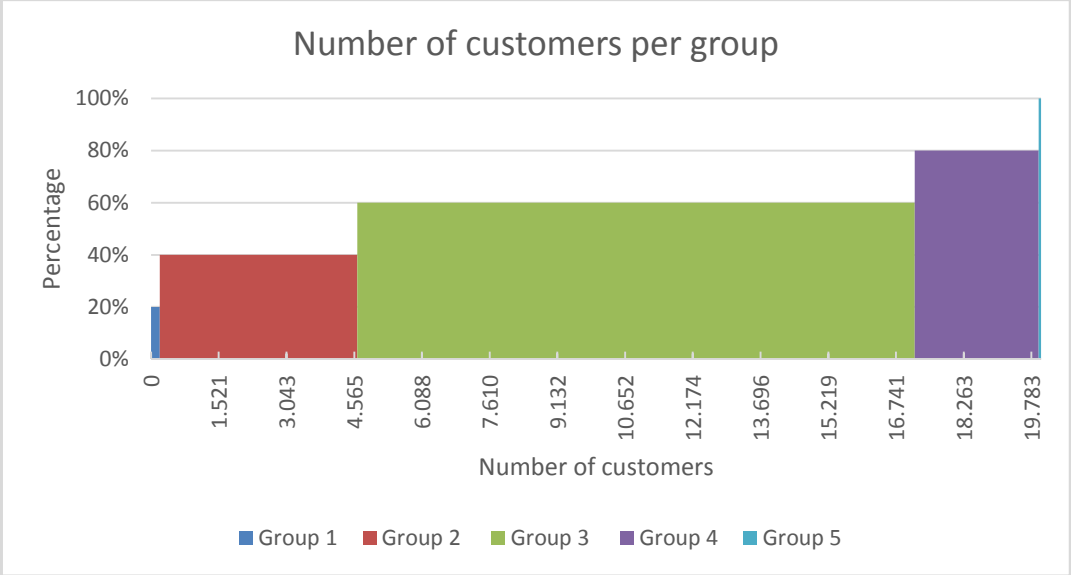
Graph 16: Accuracy of predictions with different k

Graph 16 shows that the accuracy increases, when k increases. We find the highest accuracy of 0.5877 when k = 50. Creating this model increased the runtime to 57:40 minutes. When we use k = 50 there are a bit less accurate predictions for the churned customers, this lowered from 4135 to 4064. However there are more loyal customers correctly classified, an increase from 6730 to 7689, as can be seen in table 16.

	Predicted	
Actual	Churn	Retention
Churn	4064	4258
Retention	3989	7689

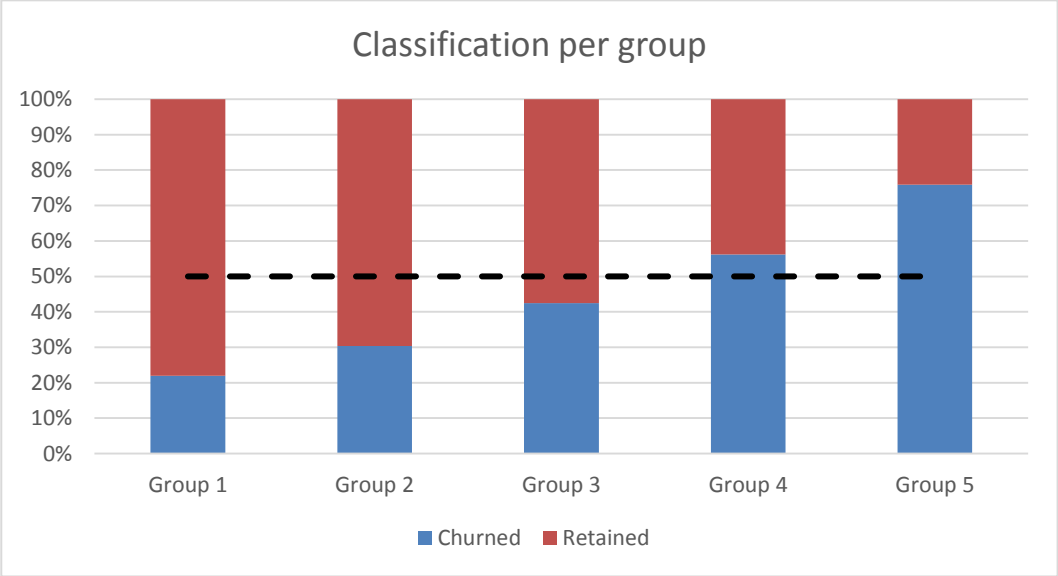
Table 16: Confusion matrix

When the customers are divided into groups we see in graph 17 that most of the customers have a probability between 0.4 and 0.6 to churn, which is more than half of the total 20000 customers.



Graph 17: Number of customers per groups

The percentage of correctly classified customers is the highest for group 1, followed by group 5, group 2 and group 4 as reported in graph 18. No conclusions can be made regarding group 3.



Graph 18: Classification per group

We applied 10 fold cross validation to the dataset and calculated a mean of 0.5672. This is lower than the accuracy of 0.5877 we found earlier. The confidence interval is between 0.5638 and 0.5706. We see that when using 90% for training and 10% for testing, the accuracy decreases. The mean of

the time creating the k-nearest neighbor model is 22:44 minutes and has a 95% confidence interval between 21:30 and 23:58 minutes. It is surprising that the time it takes to create a model using 90000 customers is lower than when using only 80000 customers.

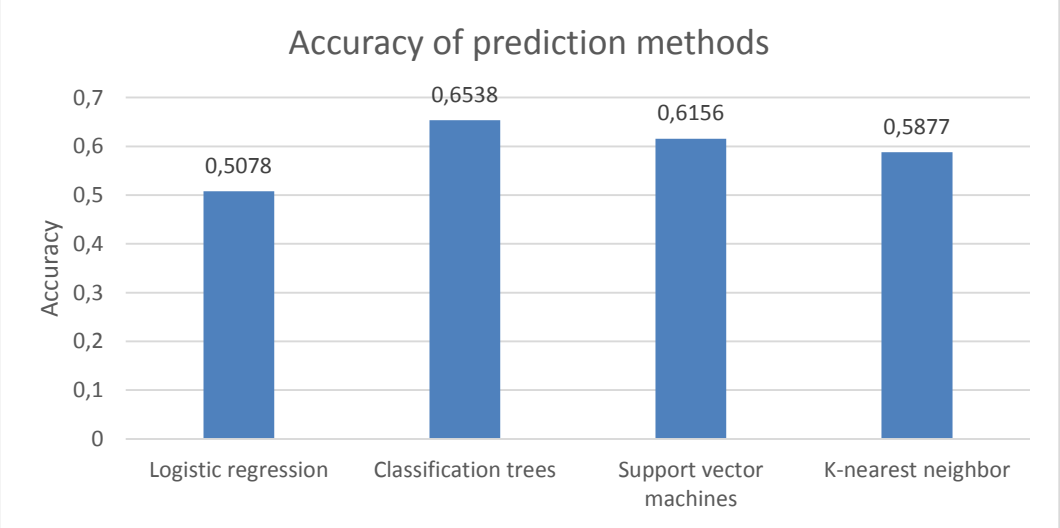
We are also interested to see what would happen to the accuracy of the k-nearest neighbor model if we removed some variables. Table 17 shows that this does not improve the accuracy of the model, but even decreases the accuracy a little.

	Accuracy
Used all variables	0.5877
Used variables that have less than 95000 missing values	0.5788
Used variables that have less than 50000 missing values	0.5795
Used variables that have less than 5000 missing values	0.5798

Table 17: Accuracy when using different amount of variables

5.1.5. Comparison of prediction models

When we compare the prediction models we see in graph 19 that the classification tree generates the highest accuracy of 0.6538 when using the 80/20 rule and only one run. The logistic regression prediction method performs the least accurate with an accuracy of only 0.5078.

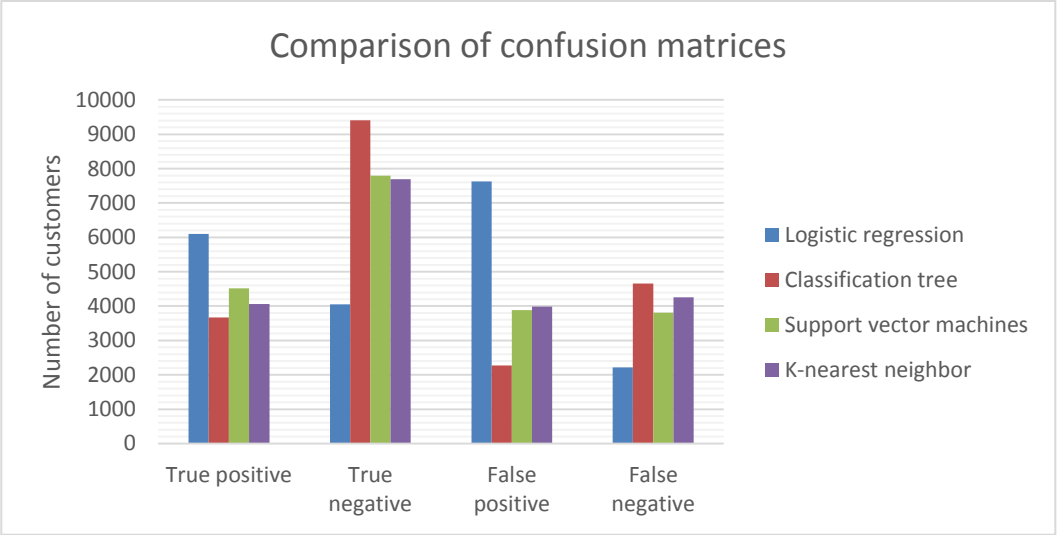


Graph 19: Accuracy of prediction models

Comparing the predictions of the different models, we see in graph 20 that the logistic regression model predicts the most churners correctly (true positive), but the least amount of loyal customers (true negative). It also predicts the most amount false positives, thus predicts the most amount of customers to churn when they in fact stay loyal.

The classification tree predicts the least amount of churners correctly but predicts the most amount of loyal people correctly. It also predicts the most amount of false negatives and the least amount of false positives. This can be explained by the fact that the classification tree method predicts most of the customers to retain as found in section 5.1.2.

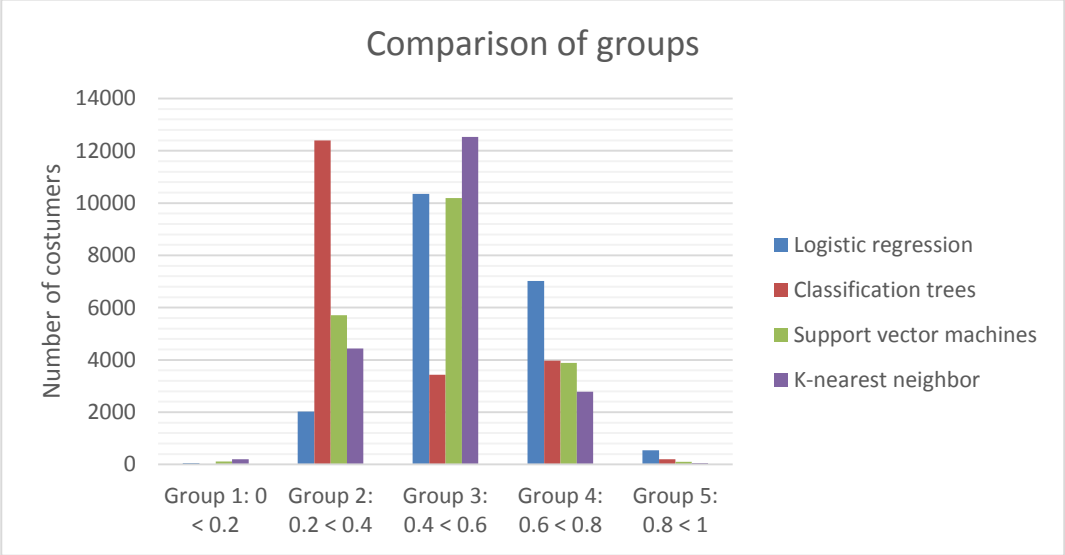
We see that the support vector machine and the k-nearest neighbor method perform very similar. They have almost an equal amount of customers classified in all four categories, but with the support vector machine method classifying more customers correctly.



Graph 20: Comparison of confusion matrices

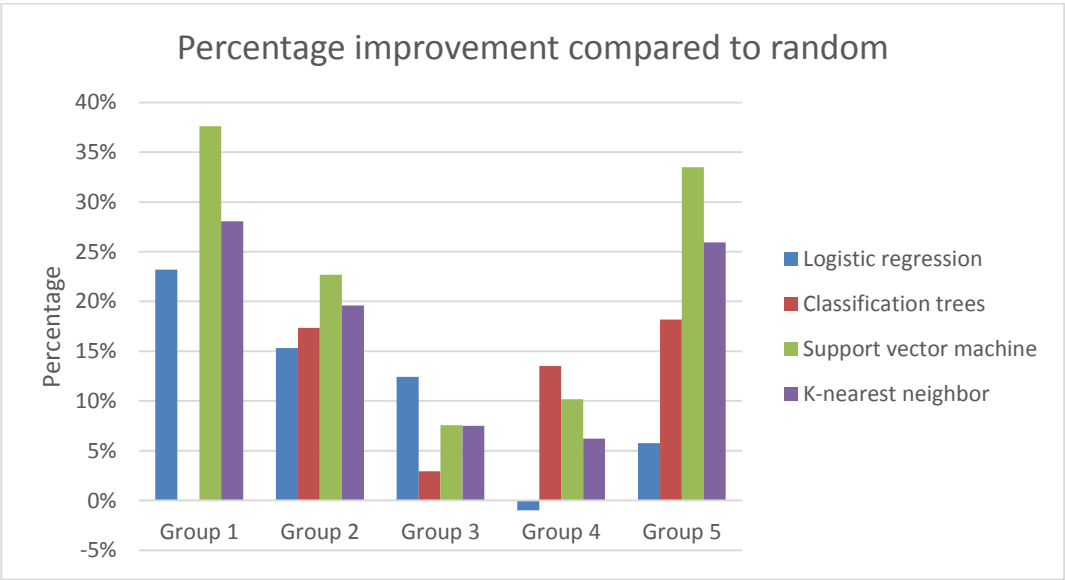
As can be seen from graph 21, the groups have different sizes per prediction method. The k-nearest neighbor method predicts the most customers with an accuracy between 0.4 and 0.6. The logistic regression method and the support vector machine also classify most of the customers in this group. The logistic regression method classifies more customers in group 2 and the support vector machine method classifies more customers in group 4. The classification tree is the only one that classifies over 12000 customers in group 2 and does not generate a probability between 0 and 0.2 to divide customers in group 1. The

other techniques classify very few customers in group 1 and the same applies for group 5.



Graph 21: Comparison of groups

Graph 22 shows the percentage improvement of the classification per group. Especially the support vector machine generates the most correctly classified customers per group. The logistic regression model has the least amount of improvement compared to random. This method even predicts more customers to retain than churn for group 4.



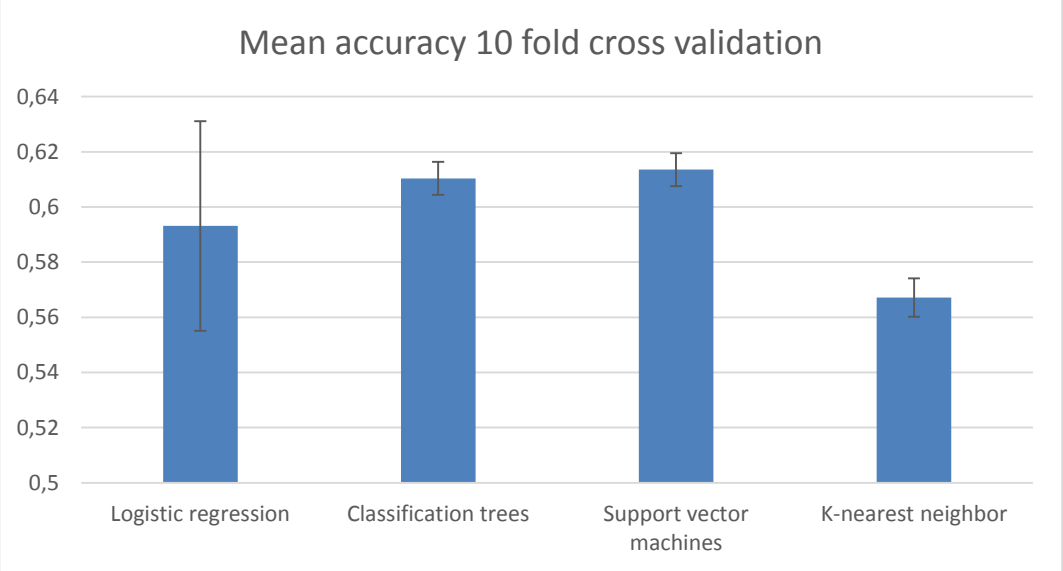
Graph 22: Percentage improvement compared to random

Looking at the runtime of the prediction methods we see that there is a large difference between the time it takes to create the prediction model. The logistic regression and the classification tree only take a few minutes. The k-nearest neighbor method takes around 20 minutes, but the support vector machine model takes hours to create as can be seen in table 18.

	Mean time creating model in minutes
Logistic regression	2:13
Classification trees	3:46
Support vector machines	252:13
K-nearest neighbor	22:44

Table 18: Comparison of mean time to create the model

When we look at the mean of the accuracy of the 10 fold cross validation, we see that the logistic regression method has a higher accuracy than with only a single run using 80% for training and 20% for testing. The accuracy increased from 0.5078 in graph 19 to 0.5931 in graph 23. The 10 fold cross validation also has a noticeable impact on the other methods. The accuracy of these three methods all decreased. Thus determining how many customers to use for the training and testing phase has an impact on the accuracy of all the models. However, because of this impact we see that the support vector machine now has a slightly higher accuracy than the classification tree.



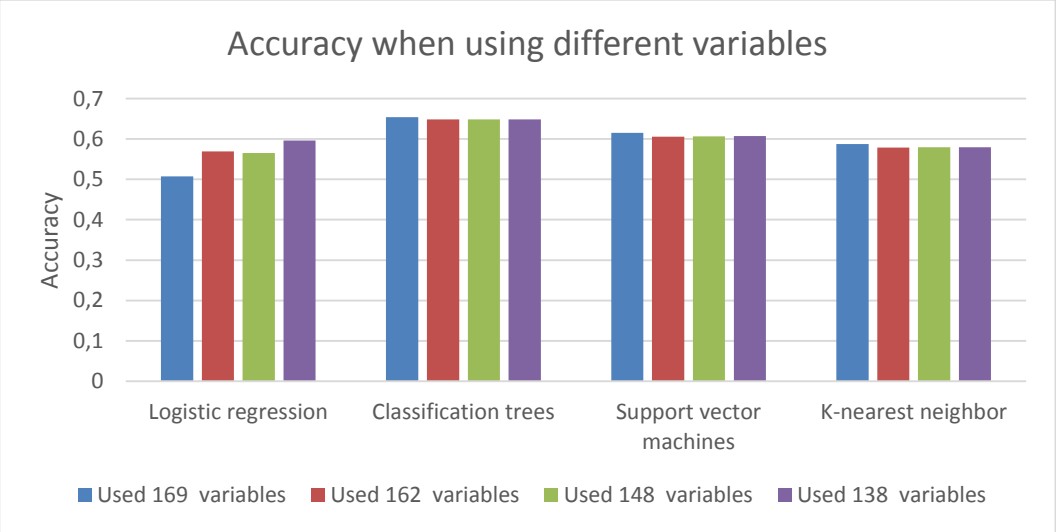
Graph 23: Comparison of the mean accuracy of 10 fold cross validation

There is not much difference when looking at the ranges of the 95% confidence interval of all four methods. The logistic regression method is the only one with a range above 0.007. Table 19 shows that the support vector machine and the classification tree have the highest confidence interval and thus perform the most accurate. There is a small but not significant difference between these two techniques.

	95% Confidence interval		Range
Logistic regression	0.5740	0.6120	0.038
Classification trees	0.6074	0.6134	0.006
Support vector machines	0,6107	0,6164	0,006
K-nearest neighbor	0,5638	0,5706	0,007

Table 19: 95% confidence interval

Graph 24 shows that removing the variables with a lot of missing values seems to greatly improve the accuracy of the logistic regression model, an increase of almost 10%. However the accuracy of the other techniques are slightly less when removing these variables from the data set. Removing variables with missing values from the dataset thus has an effect on all four prediction techniques.



Graph 24: Comparison of accuracy when using different variables

5.1.6. Additional analysis

Based on the results in the previous sections we decided to perform some additional analysis. Starting with applying bootstrap samples and 5 fold cross validation. Next we will provide an example when the group sizes are adjusted.

5.1.6.1. Bootstrap samples and 5 fold cross validation using the classification tree

Due to some of the remarkable results found earlier, we decided to perform some additional analysis. We saw in section 5.1.2 and 5.1.5 that the accuracy of the classification tree was higher when performing one run and using 80,000 customers for training than the mean accuracy when performing 10 runs and using 90,000 customers for training. We also saw that the first mentioned accuracy was not inside the confidence interval of the 10 fold cross validation. The same applied for the logistic regression and the k-nearest neighbor prediction technique. To see if we can explain this difference we started with applying 5 fold cross validation, thus using also 80,000 customers for training. In table 20 we see that the mean and the 95% confidence interval of the 5 fold cross validation is slightly lower than that of the 10 fold cross validation. The accuracy when only running the classification tree once is also outside the confidence interval of the 5 fold cross validation.

	Accuracy	95% Confidence interval	
One run	0.6538		
Mean of 10 fold cross validation	0.6104	0.6074	0.6134
Mean of 5 fold cross validation	0.6091	0.6064	0.6118
Mean using bootstrap samples (10 runs)	0.6571	0.6536	0.6605

Table 20: Comparison of accuracy for different number of runs

Next we used bootstrap samples for the test set. We started with running the classification tree one time, with 80,000 customers for training. Then we repeatedly draw samples from the test set, with replacing each sample after it was drawn. This is also known as bootstrap samples. We thus created a test set where the data of some customers can exist more than once. We ran the

prediction model 10 times. We see in table 20 that the mean accuracy is close to the accuracy of only one run. The last mentioned accuracy also lies in the confidence interval when using bootstrap samples.

We can conclude from this analysis that the training set used for creating the model has an influence on the accuracy of the predictions. Adjusting the data in the test set did not seem to have an effect on the accuracy. When creating the model multiple times the accuracy lowered and we can thus assume that the high accuracy was an outlier. In our situation the accuracy was higher with only one run, but it could also be the case that the accuracy is lower. Therefore we would advise to create the model more than once to generate a more reliable prediction model.

5.1.6.2. Adjusting group size

We chose the sizes on the groups based on what information we assume that a telecom company would like to receive. However, each company has a different strategy. It may be possible that a company does not want to target a group with a specific probability. If they have a predefined budget, they may want to target a certain amount of customers instead. We extended our research with this option. For example, if a telecom company would like to target exactly 4000 customers the probabilities of the groups would change. This is visible in table 21.

	Logistic regression	Classification tree	Support vector machine	K-nearest neighbor
Group 1	$0 < 0.455$	$0 <= 0.382$	$0 < 0.361$	$0 < 0.380$
Group 2	$0.455 < 0.530$	$0.382 <= 0.382$	$0.361 < 0.442$	$0.380 < 0.440$
Group 3	$0.530 < 0.593$	$0.382 <= 0.382$	$0.442 < 0.512$	$0.440 < 0.500$
Group 4	$0.593 < 0.666$	$0.382 < 0.602$	$0.512 < 0.599$	$0.500 < 0.560$
Group 5	$0.666 < 1$	$0.602 < 1$	$0.599 < 1$	$0.560 < 1$

Table 21: Probability interval based on group size of 4000

We see that the support vector machine and the k-nearest neighbor generate similar groups. The logistic regression technique is slightly different as the range in group 1 is higher, which leads to the range of group 5 being smaller. The results from the classification tree are surprising. This technique

divides the customers based on the branches in the tree, where each branch has a probability. Because some customers follow the same branch, they will receive the same probability. In this situation, over 12,000 customers receive a probability of 0.382 or lower. If a company would want to target 4000 customers, they can randomly select 4000 customers from these three groups.

This is only one example of how our research can be extended to meet the company's wishes. We will discuss other possibilities in the recommendations for future research section in the following chapter.

5.2 Risk Group Analysis

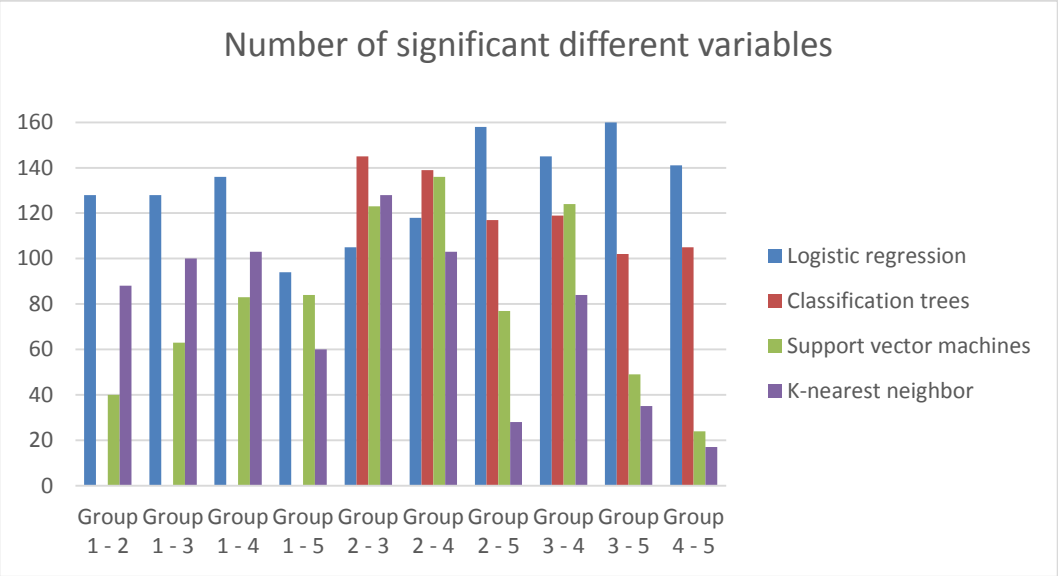
This statistical analysis was executed to test if there was any significant differences between the mean of each variable of the groups that were created with the before mentioned data mining techniques. To perform statistical analysis on the dataset, all variables have to be numeric. The character variables were transformed into binary variables by creating dummy variables. Using these variables we were able to perform the following tests on the dataset.

5.2.1. Independent t-tests

We started with applying independent t-tests on each variable in the dataset and applied this to all groups that were created with the data mining techniques. The results can be found in graph 25. This graph shows the number of variables that have a significant different mean when comparing two groups. As the classification tree did not divide customers into group 1, this group cannot be compared.

We assumed that there would be differences between the mean of the variables of the groups in order to divide the customers correctly. When there are no differences between the groups, it might be that the customers are not divided correctly and the prediction could be less accurate. From graph 25 we see that there is a lot of difference between the data mining techniques and also between the group comparisons. What stands out is that the logistic regression technique has the most number of variables with a significant different

mean in eight out of ten comparisons. It appears that when classifying customers into groups with this technique, this depends on a lot of variables. For example, when comparing groups 3 and 5 we see that 160 of the 169 variables have a significant different mean. The k-nearest neighbor and the support vector machine techniques have a lesser amount of variables that have a significant different mean.



Graph 25: Number of variables with a significant different mean using independent t-tests

Table 22 shows that the least amount of significant different variables are generated when compared to group 5. When the most amount of significant different variables are generated, groups 2 or 3 are involved in the comparison. These groups have more variables with a significant different mean than the other groups.

	Least amount of significant different variables	Most amount of significant different variables
Logistic regression	Group 1 - group 5	Group 3 - group 5
Classification trees	Group 3 - group 5	Group 2 - group 3
Support vector machines	Group 4 - group 5	Group 2 - group 4
K-nearest neighbor	Group 4 - group 5	Group 2 - group 3

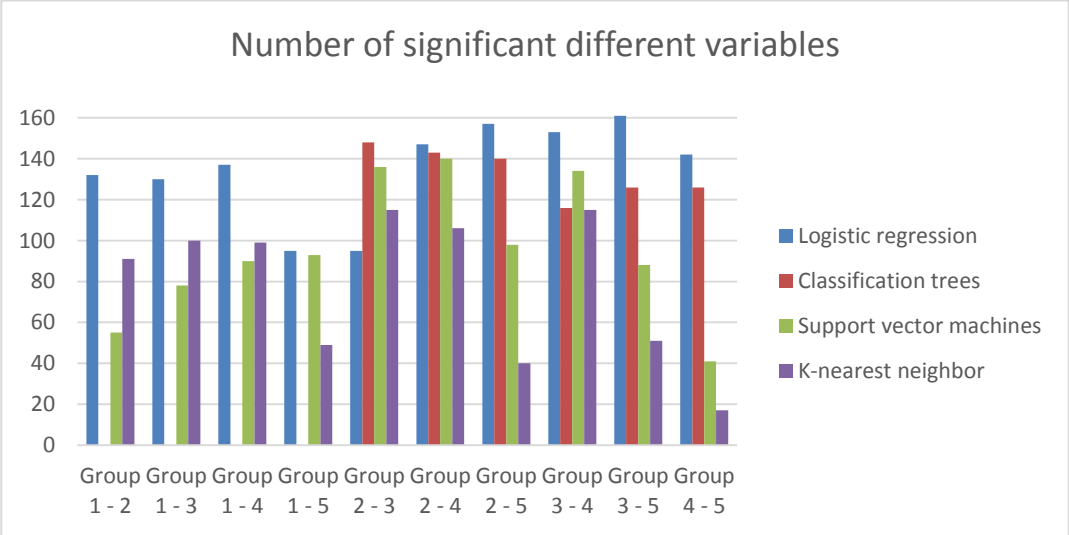
Table 22: Groups with least and most amount of significant different variables

Based on this information we cannot conclude if group 5 is the most similar to other groups. In addition, we cannot determine if a data mining technique groups the customers more or less accurate than another technique based on the number of variables that have a significant different mean. It could also be the case that some variables are more important than others and have a greater influence on the grouping. Due to the large amount of variables, we will not present the similarities and/or differences between all the variables that have a significant different mean for each technique.

What we can conclude from these tests is that the data mining techniques uses a lot of variables to divide the customers into groups, ranging from 17 till 160 variables.

5.2.2. Wilcoxon rank-sum test

To research if the results of the independent t-test would change when using a different statistical test, we also performed Wilcoxon rank-sum tests on the dataset.



Graph 26: Number of variables with a significant different mean using Wilcoxon rank-sum tests

It is visible from graph 26 that the results are not very different from the independent t-tests. In this graph we see that the logistic regression technique has a bit more significantly different variables than the classification tree when comparing groups 2 and 4. The least amount of significant different variables

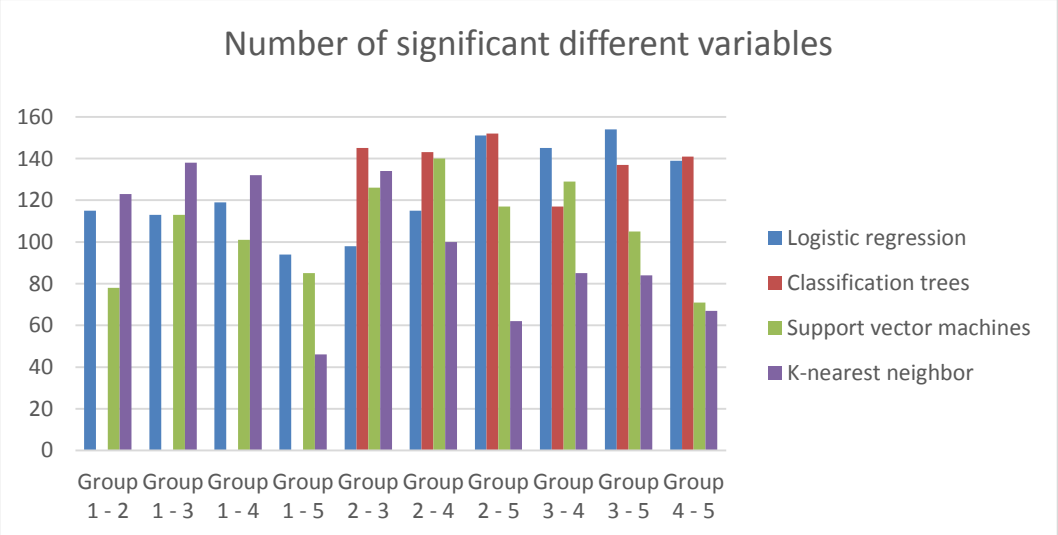
changed for the classification tree which is visible in table 23. With the independent t-tests this was the case when comparing groups 3 and 5. For the Wilcoxon rank-sum tests this was the case when comparing groups 3 and 4.

	Least amount of significant different variables	Most amount of significant different variables
Logistic regression	Group 1 - group 5	Group 3 - group 5
Classification trees	Group 3 - group 4	Group 2 - group 3
Support vector machines	Group 4 - group 5	Group 2 - group 4
K-nearest neighbor	Group 4 - group 5	Group 2 - group 3

Table 23: Groups with least and most amount of significant different variables

5.2.3. Welch's corrected unpaired t-test

Another statistical test that we applied is the Welch's corrected unpaired t-test. In graph 27 it is visible that the number of variables for each technique has changed. For example, the k-nearest neighbor technique has a higher amount of significantly different variables than the logistic regression when comparing groups 1 - 2, groups 1 - 3 and group 1 - 4. With the independent t-tests and the Wilcoxon rank-sum tests the logistic regression had a higher amount.



Graph 27: Number of variables with a significant different mean using Welch's corrected unpaired t-tests

These t-tests also affected the least and most amount of significantly different variables for the k-nearest neighbor technique in table 24. The least amount of variables are generated when comparing groups 1 and 5 and the most amount of variables are generated when comparing groups 1 and 3. The other techniques show the same results as the Wilcoxon rank-sum tests.

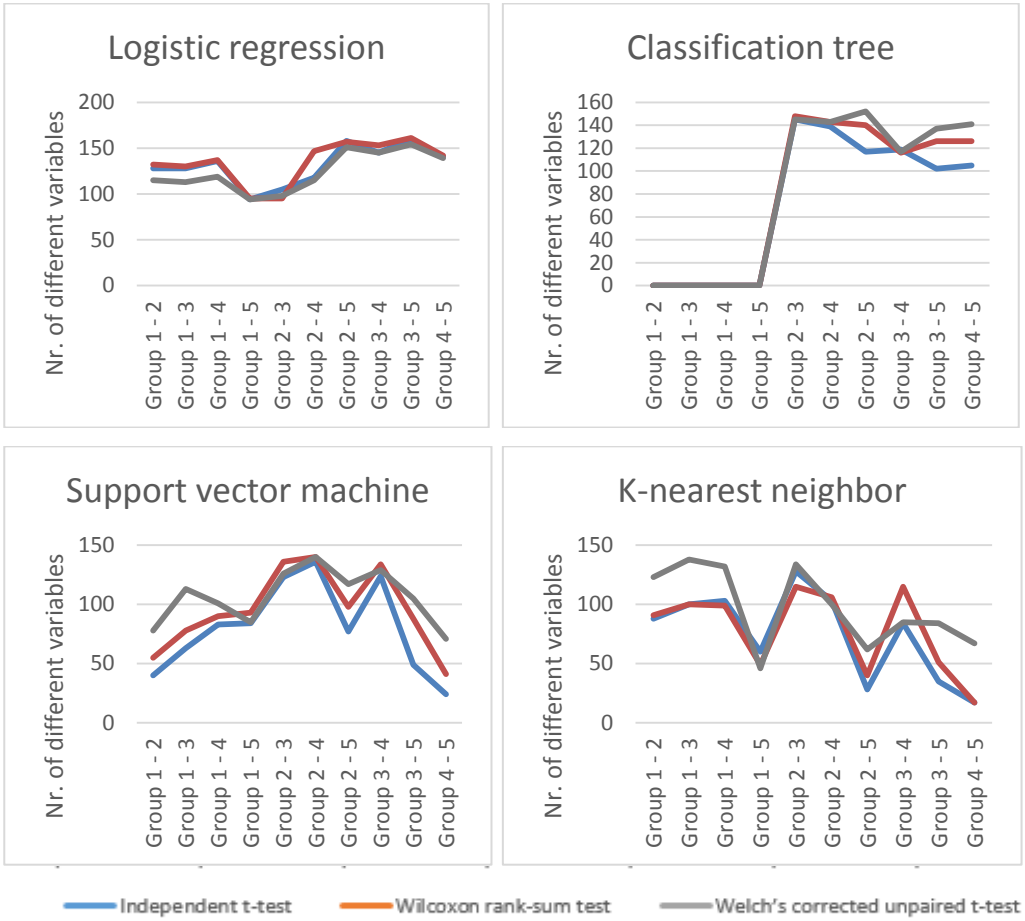
	Least amount of significant different variables	Most amount of significant different variables
Logistic regression	Group 1 - group 5	Group 3 - group 5
Classification trees	Group 3 - group 4	Group 2 - group 3
Support vector machines	Group 4 - group 5	Group 2 - group 4
K-nearest neighbor	Group 1 - group 5	Group 1 - group 3

Table 24: Groups with least and most amount of significant different variables

5.2.4. Comparison statistical analysis

For a clearer visual representation, we grouped the statistical tests in one graph 28 per data mining technique. The exact numbers can be found in the Appendix B. There are some group comparisons where all three statistical tests seem to generate the same amount of significantly different variables. For example, when looking at the classification tree, all techniques seem to generate the same amount when comparing groups 3 and 4. However, there are different results when comparing groups 4 and 5.

When comparing groups 1 and 5, we see an amount around 94 for the three possible data mining techniques, for all three statistical tests. It is possible that these results are so similar because these groups both have the highest probability to churn (group 5) or to retain (group 1).

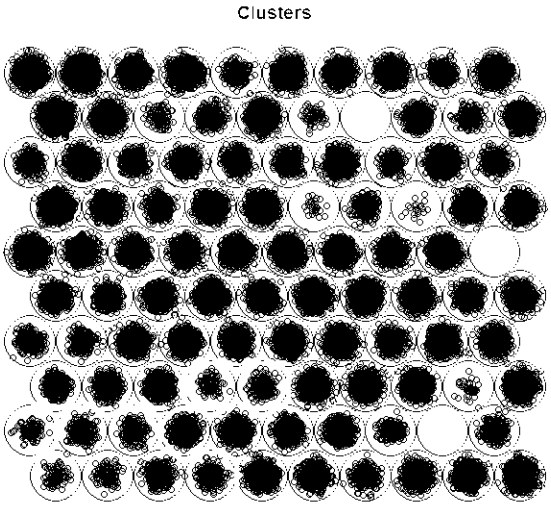


Graph 28 a-d: Comparison of statistical analysis for different data mining techniques

Because of the large amount (ranging up to 160 variables) we will not present and compare the different variables for each group and /or data mining technique. As mentioned in section 5.2.1 we can conclude that there are quite a few variables that are used when grouping the customers. In the previous chapter 4 we found that the accuracy of three of the four data mining techniques decreased slightly when removing variables from the data set. This analysis explains those results for three of the four data mining techniques, because a lot of variables are taken into account when grouping the customers. However, it does not explain why the logistic regression technique is the only one where the accuracy increases when removing variables for the dataset.

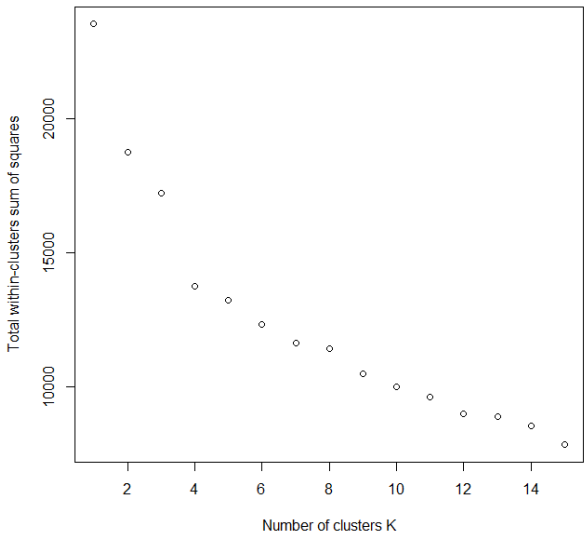
5.3 Segmentation

We conclude our results with the segmentation of the customers. Due to the many customers and variables we started with applying self-organizing maps to reduce the dimension.



Graph 29: Self organizing map

Before applying the hierarchical clustering method we first want to determine the optimal number of clusters using k-means clustering. As mentioned in section 3.3 we will calculate the total within sum of squares for each run of k-means clustering, which is visible in graph 30.



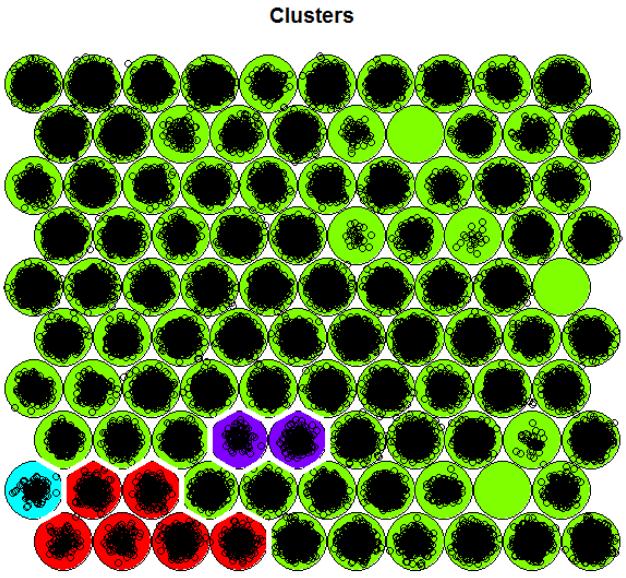
Graph 30: Total within sum of squares for different number of clusters

In this graph the 'elbow' is clearly visible at k = 4. Using 4 clusters we can now apply hierarchical clustering to the self-organizing map. This resulted in the following cluster sizes.

	Number of nodes
Cluster 1	6
Cluster 2	91
Cluster 3	1
Cluster 4	2

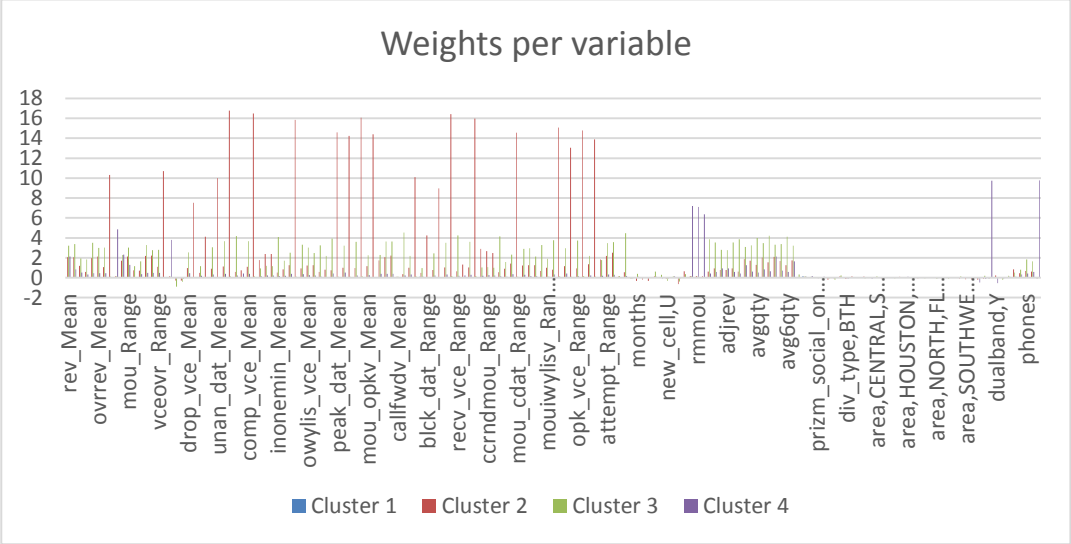
Table 25: Number of nodes in the self-organizing map per cluster

From table 25 we see that there is one very large cluster and three smaller ones. The self-organizing map is displayed with the clusters in graph 31.



Graph 31: Self-organizing map with clusters

Now that we created the clusters we would like to know what kind of customers are in each clusters. When we plot the results of the self-organizing map with the variables we find the following graph 32.



Graph 32: Weights per variable

Graph 32 shows that almost all of the character variables are of no influence to the clusters. Looking even closer we see in graph 33 that the clusters are very different. Almost all of cluster 1's weights are around 0. Cluster 2 has a lot of high weights. Cluster 3 has mostly weights between 2 and 4 and cluster 4 has some high weights for a certain of specific variables.

Graph 33: Weights per cluster

Comparing the clusters to the variables in the data set we saw a very distinct separation between the clusters. With this information we can divide the customers into the following categories.

	Category
Cluster 1	General customers
Cluster 2	Customers who primarily use data
Cluster 3	Customers who primarily call
Cluster 4	Customers who primarily use roaming

Table 26: Cluster categories

It was surprising that none of the character variables had an influence on the clusters. These clusters are created based on the usage of each customer. The retention team can use this information and provide offers based on what kind of type a customer is.

6 Conclusions and recommendations

6.1 Conclusions

The goal of this research was to generate a model that uses data mining techniques to increase customer retention at a telecom company. We analyzed the data set and we found that that there was little difference for a lot of variables between churners and non-churners. Because the data set was retrieved from a website, we did not possess information about the customers, variables or if the data set was already altered that might explain this randomness. Prediction models were created using the data mining techniques logistic regression, classification tree, support vector machine and k-nearest neighbor using 80% of the customers for training and 20% for testing. The classification tree method generates the highest accuracy of 0.654 and the k-nearest neighbor the lowest accuracy of 0.508. Next, we applied 10 fold cross validation using 90% for training and 10% for testing. The classification tree and the support vector machine generated a similar accuracy, with confidence intervals between 0.6074 and 0.6134 and between 0.6107 and 0.6164 respectively. It was surprising that the accuracy when using 80% for testing was very different then the accuracy that was generated when using 90% for training. After applying further analysis, we found that the creation of the model has to be run multiple times to create a reliable prediction model. The logistic regression and the classification tree methods generated a similar runtime when creating the models of around 3 minutes. The runtime of the support vector machine was on average over 4 hours. When removing variables with missing values from the data set the accuracy decreased for three of the four data mining techniques. The logistic regression method was the only method that generated a higher accuracy when variables with missing values were removed. The customers were also divided into one of five groups based on their probability to churn. These groups were compared using statistical analysis to test if there was any significant differences between the groups. We found that a lot of variables (ranging from 17 till 160) had a significant different mean when two groups were compared. Segmentation was applied to all the customers and four clusters were found based on the usage of the customers.

Unfortunately, we were not able to test our approach at an actual telecom company. Therefore, to answer our research question we cannot conclude that data mining can be used to optimize customer retention based on this research.
However, we did provide a model that when applied at an actual telecom company will answer our research question. This model provides the retention team of a telecom company with the information on who to target using the predictions from the data mining techniques and how to approach those customers using segmentation.

6.2 Recommendations

There are more data mining techniques besides the ones used in this research. A suggestion for future research could be to try out different techniques or perhaps combine these techniques to try and improve the accuracy of the prediction. The techniques could also be further researched by, for example, removing customers with a lot of missing values and tested to see if the accuracy and the confidence interval will increase.

In this research the dataset was obtained from a website. There was no information about the time period of the data. The retention team provided the customers with offers, but some customers churned. From our dataset it was not visible what the time period was between the offer and the time of leaving. This information can provide more insight to the results of the offers. It would be interesting to test our approach on a dataset of another company that has this information.

Based on our research we would recommend a telecom company to apply the support vector machine or the classification tree, depending on the company's desires. The classification tree has a lower runtime, but the support vector machine has a slightly higher confidence interval, although not significant. Using the results of this prediction, the telecom company can decide which customers to target. It would probably not be profitable to target the customers with the highest probabilities, because they are already most likely to churn. The customers that have a probability between 0.4 and 0.6 are probably more likely to be convinced to stay loyal. In our research we saw that this was a large group of customers. If the telecom company would like

to target a certain amount of customers, the group sizes can easily be adjusted to meet the company's wishes. The retention team can use the segmentation to make informed decisions on how to approach the customers.

We recommend the telecom company to use our approach and test if customer retention can be increased. .

7 Bibliography

Algina, J., Oshima, T. C., & Lin, W. Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics*, *19*(3), 275-291.

Anderson, B. J., Gross, D. S., Musicant, D. R., Ritz, A. M., Smith, T. G., & Steinberg, L. E. (2006, April). Adapting K-Medians to Generate Normalized Cluster Centers. In *SDM* (pp. 165-175).

Baldi, P., & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509-519.

Bronner, A. E. (2009) 10. Het voorspellen van switchgedrag in een markt met een lage mobiliteit: Een case study. *Ontwikkelingen in het marktonderzoek*, 167.

Chu, B. H., Tsai, M. S., & Ho, C. S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, *20*(8), 703-718.

Consultatie marktanalyse vaste telefonie (2016, July 18). Retrieved January 9, 2017, from https://www.acm.nl/nl/publicaties/publicatie/16052/Consultatie-marktanalyse-vaste-telefonie/

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, *16*(22), 10881-10890.

Davenport, T. H. (2006). Competing on analytics. *harvard business review*, *84*(1), 98.

Devroye, L., Györfi, L., & Lugosi, G. (2013). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science & Business Media.

Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, *12*(1), 1-22.

Grimm, E. C. (1987). CONISS: a FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Computers & Geosciences*, *13*(1), 13-35.

Hall, M., Witten, I., & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*.

Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Upper Saddle River, NJ, USA:: Pearson.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of management*, *21*(5), 967-988.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524.

Inc. IBP (2016). Netherlands Investment and Business Guide Volume 1 Strategic and Practical Information.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137-163.

Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Liu Peng, L. L. (2005). A review of missing data treatment methods. *Int. Journal of Intel. Inf. Manag. Syst. and Tech*, *1*(3).

Máša, P., & Kočka, T. (2006). Finding Optimal Decision Trees. In *Intelligent Information Processing and Web Mining* (pp. 173-181). Springer Berlin Heidelberg.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research*, *60*(6), 656-662.

Milovic, B., & Milovic, M. (2012). Prediction and decision making in health care using data mining. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *1*(12), 126.

Min, H. (1994). International supplier selection: a multi-attribute utility approach. *International Journal of Physical Distribution & Logistics Management*, *24*(5), 24-33.

Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European journal of operational research*, *174*(3), 1742-1759.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, *72*, 72-81.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, *11*(3), 690-696.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13-20.

Ontwerpbesluit marktanalyse vaste telefonie 2012 (2011, July 14). Retrieved January 9, 2017, from https://www.acm.nl/nl/publicaties/publicatie/10240/Ontwerpbesluit-marktanalyse-vaste-telefonie-2012/

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *arXiv preprint arXiv:1002.1144*.

Reichheld, F. F., Markey Jr, R. G., & Hopton, C. (2000). The loyalty effect-the relationship between loyalty and profits. *European Business Journal*, *12*(3), 134.

Reichheld, F. F., & Teal, T. (2001). The loyalty effect: The hidden force behind growth, profits, and lasting value. Harvard Business Press.

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198-208.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in society*, *24*(4), 483-502.

Salvador, S., & Chan, P. (2004, November). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on* (pp. 576-584). IEEE.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999-2006.

Sree Hari Rao, V., & Jonnalagedda, M. V. (2012). Insurance Dynamics–A Data Mining Approach for Customer Retention in Health Care Insurance Industry. *Cybernetics and Information Technologies*, *12*(1), 49-60.

Taking a next best action approach to strengthening telecom customer relationships (2015, June 30) Retrieved January 9, 2017, from http://www.1to1media.com/customer-engagement/taking-next-best-action-approach-strengthening-telecom-customer-relationships

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Verdu, S. V., García, M. O., Franco, F. J. G., Encinas, N., Marin, A. G., Molina, A., & Lazaro, E. G. (2004, October). Characterization and identification of electrical customers through the use of self-organizing maps and daily

load parameters. In *Power Systems Conference and Exposition, 2004. IEEE PES* (pp. 899-906). IEEE.

Viveros, M. S., Nearhos, J. P., & Rothman, M. J. (1996, September). Applying data mining techniques to a health insurance information system. In *VLDB* (pp. 286-294).

Walsh, G., Groth, M., & Wiedmann, K. P. (2005). An examination of consumers' motives to switch energy suppliers. *Journal of Marketing Management*, *21*(3-4), 421-440.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2005). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473-1480).

Wieringa, J. E., & Verhoef, P. C. (2007). Understanding customer switching behavior in a liberalizing service market an exploratory study. *Journal of Service Research*, *10*(2), 174-186.

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005, July). Customer churn prediction using improved one-class support vector machine. In *International Conference on Advanced Data Mining and Applications* (pp. 300-306). Springer Berlin Heidelberg.

8 Appendices

8.1 Appendix A: Tables of data mining techniques

Group based on probability	Predicted total customers
Group 1: 0 < 0.2	56
Group 2: 0.2 < 0.4	2027
Group 3: 0.4 < 0.6	10346
Group 4: 0.6 < 0.8	7024
Group 5: 0.8 <= 1	547

Table 1: Customers divided into groups with logistic regression

Group based on probability	Predicted total customers
Group 1: 0 < 0.2	0
Group 2: 0.2 < 0.4	12397
Group 3: 0.4 < 0.6	3437
Group 4: 0.6 < 0.8	3968
Group 5: 0.8 <= 1	198

Table 2: Customers divided into groups with classification tree

	Predicted total customers
Group 1: 0 < 0.2	113
Group 2: 0.2 < 0.4	5710
Group 3: 0.4 < 0.6	10197
Group 4: 0.6 < 0.8	3881
Group 5: 0.8 <= 1	99

Table 3: Customers divided into groups with support vector machine

	Predicted total customers
Group 1: 0 < 0.2	196
Group 2: 0.2 < 0.4	4439
Group 3: 0.4 < 0.6	12526
Group 4: 0.6 < 0.8	2785
Group 5: 0.8 < 1	54

Table 4: Customers divided into groups with k-nearest neighbor

8.2 Appendix B: Tables of comparison of data mining techniques

	True positive	True negative	False positive	False negative
Logistic regression	6103	4052	7626	2219
Classification tree	3667	9409	2269	4655
Support vector machines	4513	7798	3880	3809
K-nearest neighbor	4064	7689	3989	4258

Table 1: Comparison of confusion matrices

	Group 1: $0 < 0.2$	Group 2: $0.2 < 0.4$	Group 3: $0.4 < 0.6$	Group 4: $0.6 < 0.8$	Group 5: $0.8 < 1$
Logistic regression	56	2027	10346	7024	547
Classification trees	0	12397	3437	3968	198
Support vector machines	113	5710	10197	3881	99
K-nearest neighbor	196	4439	12526	2785	54

Table 2: Comparison of groups

	Mean accuracy 10 fold cross validation
Logistic regression	0,59313
Classification trees	0,61039
Support vector machines	0,61354
K-nearest neighbor	0,56716

Table 3: Comparison of 10 fold cross validation

	Used 169 variables	Used 162 variables	Used 148 variables	Used 138 variables
Logistic regression	0,50775	0,5695	0,56555	0,5966
Classification trees	0,6538	0,6487	0,6487	0,6489
Support vector machines	0.61555	0,6061	0.6066	0.60715
K-nearest neighbor	0,58765	0,57875	0,57945	0,5798

Table 4: Comparison of accuracy when using different variables

	Logistic regression	Classification trees	Support vector machines	K-nearest neighbor
Group 1 - 2	128	-	40	88
Group 1 - 3	128	-	63	100
Group 1 - 4	136	-	83	103
Group 1 - 5	94	-	84	60
Group 2 - 3	105	145	123	128
Group 2 - 4	118	139	136	103
Group 2 - 5	158	117	77	28
Group 3 - 4	145	119	124	84
Group 3 - 5	160	102	49	35
Group 4 - 5	141	105	24	17

Table 5: Number of variables with a significant different mean using independent t-tests

	Logistic regression	Classification trees	Support vector machines	K-nearest neighbor
Group 1 - 2	132	-	55	91
Group 1 - 3	130	-	78	100
Group 1 - 4	137	-	90	99
Group 1 - 5	95	-	93	49
Group 2 - 3	96	148	136	115
Group 2 - 4	147	143	140	106
Group 2 - 5	157	140	98	40
Group 3 - 4	153	116	134	114
Group 3 - 5	161	126	88	51
Group 4 - 5	142	126	41	17

Table 6: Number of variables with a significant different mean using Wilcoxon rank-sum tests

	Logistic regression	Classification trees	Support vector machines	K-nearest neighbor
Group 1 - 2	115	-	78	123
Group 1 - 3	113	-	113	138
Group 1 - 4	119	-	101	132
Group 1 - 5	94	-	85	46
Group 2 - 3	98	145	126	134
Group 2 - 4	115	143	140	100
Group 2 - 5	151	152	117	62
Group 3 - 4	145	117	129	85
Group 3 - 5	154	137	105	84
Group 4 - 5	139	141	71	67

Table 7: Number of variables with a significant different mean using Welch's corrected unpaired t-tests

8.3 Appendix C: Variable descriptions

ID	Column Name	Type	Description
1	rev_Mean	Num	CHARGE_AMT
2	mou_Mean	Num	MINUTE_QTY
3	totmrc_Mean	Num	total MRC, mean
4	da_Mean	Num	directory_assisted mean
5	ovrmou_Mean	Num	Overage minutes of use, mean
6	ovrrev_Mean	Num	Overage revenue, mean
7	vceovr_Mean	Num	voice overage, mean
8	datovr_Mean	Num	data overuse mean
9	roam_Mean	Num	roaming, mean
10	rev_Range	Num	CHARGE_AMT
11	mou_Range	Num	MINUTE_QTY
12	totmrc_Range	Num	total MRC, range
13	da_Range	Num	directory_assisted range
14	ovrmou_Range	Num	Overage minutes of use, range
15	ovrrev_Range	Num	Overage revenue, range
16	vceovr_Range	Num	voice overage, range
17	datovr_Range	Num	data overuse range
18	roam_Range	Num	roaming, range
19	change_mou	Num	% change of minutes of use
20	change_rev	Num	% change of revenue
21	drop_vce_Mean	Num	nbr_dropped_calls_voice (Failed Calls)
22	drop_dat_Mean	Num	nbr_dropped_calls_data (Failed Calls)
23	blck_vce_Mean	Num	nbr_blocked_calls_voice (Failed Calls)
24	blck_dat_Mean	Num	nbr_blocked_calls_data (Failed Calls)
25	unan_vce_Mean	Num	nbr_unanswered_calls_voice
26	unan_dat_Mean	Num	nbr_unanswered_calls_data
27	plcd_vce_Mean	Num	nbr_calls_placed_voice (attempts)
28	plcd_dat_Mean	Num	nbr_calls_placed_data(attempts)
29	recv_vce_Mean	Num	nbr_calls_received_voice
30	recv_sms_Mean	Num	nbr_calls_received_sms
31	comp_vce_Mean	Num	nbr_calls_completed_voice
32	comp_dat_Mean	Num	nbr_calls_completed_data
33	custcare_Mean	Num	Customer Care Calls
34	ccrndmou_Mean	Num	Customer Care Rounded MOU
35	cc_mou_Mean	Num	nbr_unrnd_MOU_cust_care_calls
36	inonemin_Mean	Num	Inbound Calls Less Than One Minute

37	threeway_Mean	Num	Three Way Calls
38	mou_cvce_Mean	Num	nbr_unrnd_mou_completed_voice_calls
39	mou_cdat_Mean	Num	nbr_unrnd_mou_completed_data_calls
40	mou_rvce_Mean	Num	nbr_unrnd_mou_received_voice_calls
41	owylis_vce_	Num	nbr_outbound_wylis2wylis_voice_calls Mean
42	mouowylisv_	Num	nbr_unrnd_mou_outbnd_wylis2wylis_voice_call mean
43	iwylis_vce_	Num	nbr_inbound_wylis2wylis_voice_call Mean
44	mouiwylisv_	Num	nbr_unrnd_mou_inbnd_wylis2wylis_voice_calls mean
45	peak_vce_Mean	Num	nbr_peak_voice_calls (inbnd & outbnd)
46	peak_dat_Mean	Num	nbr_peak_data_calls
47	mou_peav_Mean	Num	nbr_unrnd_mou_peak_voice_calls
48	mou_pead_Mean	Num	nbr_unrnd_mou_peak_data_calls
49	opk_vce_Mean	Num	nbr_off_peak_voice_calls
50	opk_dat_Mean	Num	nbr_off_peak_data_calls
51	mou_opkv_Mean	Num	nbr_unrnd_mou_off_peak_voice_calls
52	mou_opkd_Mean	Num	nbr_unrnd_mou_off_peak_data_calls
53	drop_blk_Mean	Num	Drop/Block Calls
54	attempt_Mean	Num	Attempted Calls
55	complete_Mean	Num	Completed Calls
56	callfwdv_Mean	Num	Call Forward Calls
57	callwait_Mean	Num	Call Wait Calls
58	drop_vce_Range	Num	nbr_dropped_calls_voice (Failed Calls)
59	drop_dat_Range	Num	nbr_dropped_calls_data (Failed Calls)
60	blck_vce_Range	Num	nbr_blocked_calls_voice (Failed Calls)
61	blck_dat_Range	Num	nbr_blocked_calls_data (Failed Calls)
62	unan_vce_Range	Num	nbr_unanswered_calls_voice
63	unan_dat_Range	Num	nbr_unanswered_calls_data
64	plcd_vce_Range	Num	nbr_calls_placed_voice (attempts)
65	plcd_dat_Range	Num	nbr_calls_placed_data(attempts)
66	recv_vce_Range	Num	nbr_calls_received_voice
67	recv_sms_Range	Num	nbr_calls_received_sms
68	comp_vce_Range	Num	nbr_calls_completed_voice
69	comp_dat_Range	Num	nbr_calls_completed_data
70	custcare_Range	Num	Customer Care Calls
71	ccrndmou_Range	Num	Customer Care Rounded MOU
72	cc_mou_Range	Num	nbr_unrnd_MOU_cust_care_calls
73	inonemin_Range	Num	Inbound Calls Less Than One Minute
74	threeway_Range	Num	Three Way Calls
75	mou_cvce_Range	Num	nbr_unrnd_mou_completed_voice_calls
76	mou_cdat_Range	Num	nbr_unrnd_mou_completed_data_calls
77	mou_rvce_Range	Num	nbr_unrnd_mou_received_voice_calls

78	owylis_vce_	Num	nbr_outbound_wylis2wylis_voice_calls Range
79	mouowylisv_	Num	nbr_unrnd_mou_outbnd_wylis2wylis_voice_call range
80	iwylis_vce_	Num	nbr_inbound_wylis2wylis_voice_call Range
81	mouiwylisv_	Num	nbr_unrnd_mou_inbnd_wylis2wylis_voice_calls range
82	peak_vce_Range	Num	nbr_peak_voice_calls (inbnd & outbnd)
83	peak_dat_Range	Num	nbr_peak_data_calls
84	mou_peav_Range	Num	nbr_unrnd_mou_peak_voice_calls
85	mou_pead_Range	Num	nbr_unrnd_mou_peak_data_calls
86	opk_vce_Range	Num	nbr_off_peak_voice_calls
87	opk_dat_Range	Num	nbr_off_peak_data_calls
88	mou_opkv_Range	Num	nbr_unrnd_mou_off_peak_voice_calls
89	mou_opkd_Range	Num	nbr_unrnd_mou_off_peak_data_calls
90	drop_blk_Range	Num	Drop/Block Calls
91	attempt_Range	Num	Attempted Calls
92	complete_Range	Num	Completed Calls
93	callfwdv_Range	Num	Call Forward Calls
94	callwait_Range	Num	Call Wait Calls
95	churn	Num	Dependent Variable: Churn between 31-60 days after
96	months	Num	Months in Service
97	uniqsubs	Num	Number of Uniq Subs
98	actvsubs	Num	Number of Active Subs
99	crtcount	Num	Number of Courtesy Credits
100	new_cell	Char	New Cell Phone User
101	crclscod	Char	Credit Class Code
102	asl_flag	Char	Account Spending Limits
103	rmcalls	Num	Roaming Calls
104	rmmou	Num	RMMOU
105	rmrev	Num	RMREV
106	totcalls	Num	TOTCALLS
107	totmou	Num	TOTMOU
108	totrev	Num	TOTREV
109	adjrev	Num	adjusted revenue
110	adjmou	Num	adjusted minutes of usage
111	adjqty	Num	adjusted quantity of calls
112	avgrev	Num	average revenue
113	avgmou	Num	average minutes of use
114	avgqty	Num	average number of calls
115	avg3mou	Num	Subs Last 3 months avg: MOU
116	avg3qty	Num	Subs Last 3 months avg: QTY
117	avg3rev	Num	Subs Last 3 months avg: REV
118	avg6mou	Num	Subs Last 6 months avg: MOU

119	avg6qty	Num	Subs Last 6 months avg: QTY
120	avg6rev	Num	Subs Last 6 months avg: REV
121	REF_QTY	Num	Total Number of Referrals
122	tot_ret	Num	Total Calls into Retention Team
123	tot_acpt	Num	Total Offers Accepted From Retention Team
124	prizm_social_	Char	Social Group Letter Only one
125	div_type	Char	Division Type Code
126	csa	Char	Communications Service Area
127	area	Char	Area
128	dualband	Char	Dualband
129	refurb_new	Char	Handset Refurb or New
130	hnd_price	Num	Handset Price
131	pre_hnd_price	Num	Previous Handset Price
132	phones	Num	# Handsets Issued
133	last_swap	Num	Date of Last Phone Swap
134	models	Num	# Models Issued
135	hnd_webcap	Char	Handset Web Capable
136	truck	Char	Truck Indicator
137	mtrcycle	Char	Motorcycle Indicator
138	rv	Char	RV Indicator
139	occu1	Char	Occupation for 1st Individual
140	ownrent	Char	Home Owner/Renter Status
141	lor	Char	Length of Residence
142	dwlltype	Char	Dwelling Unit Type
143	marital	Char	Marital Status
144	mailordr	Char	Mail Order Buyer
145	age1	Char	Age of 1st household member
146	age2	Char	Age of 2nd household member
147	wrkwoman	Char	Working Woman in HH
148	mailresp	Char	Mail Responder
149	children	Char	Children present in HH
150	adults	Char	Number of Adults in HH
151	infobase	Char	Infobase Match
152	income	Char	Estimated Income
153	numbcars	Char	Known Number of Vehicles
154	cartype	Char	Dominant Vehicle Lifestyle
155	HHstatin	Char	Prem Household Status Ind
156	mailflag	Char	DMA Do Not Mail Flag
157	solflag	Char	Infobase No Phone Sol Flag
158	dwllsize	Char	Dwelling Size
159	forgntvl	Char	Foreign Travel Dummy Var

160	educ1	Char	Education for 1st Individual
161	proptype	Char	Property Type Detail
162	pcowner	Char	PC Owner Dummy Var
163	ethnic	Char	Ethnicity Roll-Up Code
164	kid0_2	Char	Kid 0-2 years of age in HH
165	kid3_5	Char	Kid 3-5 years of age in HH
166	kid6_10	Char	Kid 6-10 years of age in HH
167	kid11_15	Char	Kid 11-15 years of age in HH
168	kid16_17	Char	Kid 16-17 years of age in HH
169	creditcd	Char	Credit Card Indicator
170	car_buy	Char	New or Used car buyer
171	retdays	Num	Number of days since last rentition call
172	eqpdays	Num	Number of days of the current equipment
173	Customer_ID	Num	Unique tournament specific customer ID