**Contents**

[1. Introduction 1](#_Toc75013461)

[2. Analysis Section 3](#_Toc75013462)

[2.1 About the data 3](#_Toc75013463)

[2.2 Data cleaning with VIS 4](#_Toc75013464)

[1) Gender 4](#_Toc75013465)

[2) Partner 5](#_Toc75013466)

[3) Tenure 5](#_Toc75013467)

[4) Phone service 5](#_Toc75013468)

[5) Internet service 6](#_Toc75013469)

[6) Payment method 7](#_Toc75013470)

[7) Monthly charges 7](#_Toc75013471)

[8) Total charges 8](#_Toc75013472)

[9) Churn 8](#_Toc75013473)

[2.3 Data EDA with VIS 9](#_Toc75013474)

[2.4 ML models used and how they work 13](#_Toc75013475)

[2.4.1 Logistic Regression 13](#_Toc75013476)

[2.4.2 Decision Tree 14](#_Toc75013477)

[2.4.3 Random Forest 15](#_Toc75013478)

[3. Results 17](#_Toc75013479)

[3.1 Logistic Regression 17](#_Toc75013480)

[3.2 Decision Tree 19](#_Toc75013481)

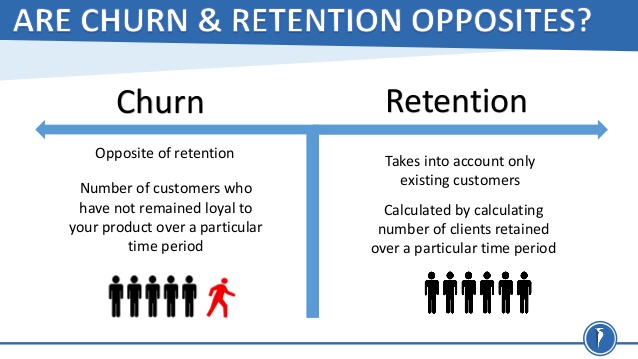
[3.3 Random Forest 20](#_Toc75013482)

[3.4 Comparison of three models 22](#_Toc75013483)

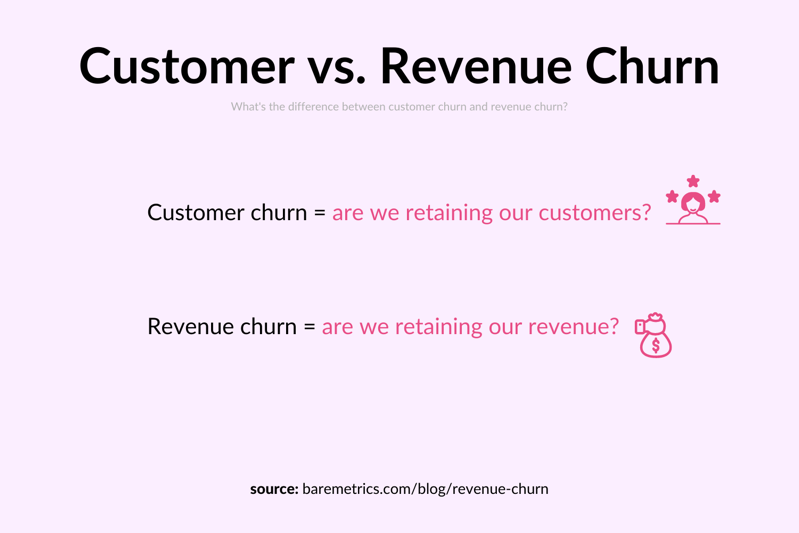
[4. Conclusions 22](#_Toc75013484)

1. Introduction

Churn analysis is the process of evaluating the data to discover where in the customer journey people stop using the product or service, and why. Churn and retention have an inverse relationship: When churn goes down, retention goes up, and vice versa. But it’s not a 1:1 relationship. The point of churn analysis is to figure out how to fix the problems that cause people to leave, while retention analysis includes both reducing churn and finding ways to add value to keep users.



There are two types of churn: customer churn and revenue churn. Customer churn is a measure of the loss of customers. Revenue churn is a measure of the loss of revenue. Analyzing customer churn is important for getting a picture of the overall health of the product and customer base because it shows how people are behaving during the service. Knowing what experiences or friction points drive users out of the service helps effectively prioritize improvements. Analyzing revenue churn recognizes that some of customers are more valuable than others based on what they spend on the product or service. Reducing revenue churn will have a direct impact on a company’s bottom line and profitability.



In order to build a sustainable SaaS business, companies need to focus their efforts on reducing customer churn. By being aware of and monitoring churn rate, companies are equipped to determine their customer retention success rates and identify strategies for improvement. According to the authors of “Leading on the Edge of Chaos”, reducing customer churn by 5% can increase profits 25-125%. Therefore, to reduce churn, most SaaS companies perform customer churn analysis.

There are five main reasons that cause customers to churn.

1. Competitors have more to offer. If a competitor has a product similar in quality and offerings but it’s set at a lower price, customers cannot be blamed for going with the more affordable option.



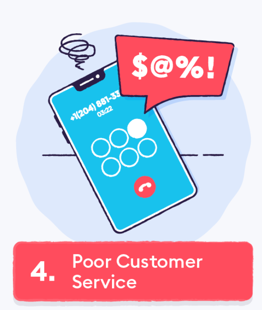
1. The product doesn’t meet customer needs. If the product doesn’t live up to customer expectations, they are going to be disappointed with their purchase and churn. The reason a product doesn’t meet customer needs can come from a variety of factors and doesn’t necessarily mean that the product isn’t up to snuff. Sometimes a customer may not understand how to use the product properly, leading them to think it’s low-quality, when in fact it’s the instructions and on-boarding process that need improvement.



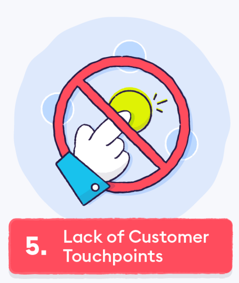
1. The customer was not a good fit for the product. This can be a particularly frustrating situation as the marketing resources required to reach new customers, such as social advertising and email campaigns, can be quite costly and end up being a waste when the customer immediately churns. This is why it is so important to conduct in-depth market research about the ideal customer base ahead of time.



1. Poor customer service is another one of the obvious reasons that come to mind for why customers leave.



1. Lack of touchpoints after conversion. Maybe customers aren’t dissatisfied with the product or customer service. They may just forget to use them. After someone converts, the company can’t assume they’ll remain a loyal customer until the end of time. It can follow up with customer touchpoints in a variety of ways, from free birthday gifts to emailed discounts.



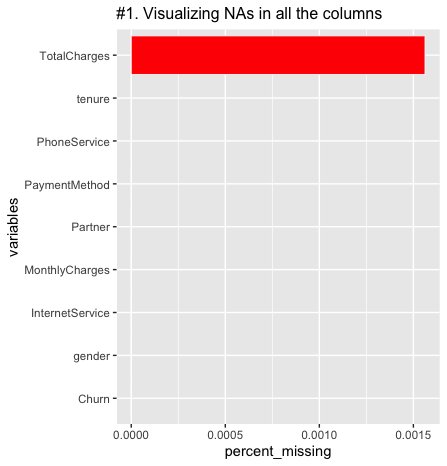
1. Analysis Section
   1. About the data

The dataset is an IBM sample data set, downloaded from Kaggle. After simplified, the raw data contains 7043 rows (customers) and 9 columns (features). The variables include:

|  |  |
| --- | --- |
| Gender | whether the customer is a male or a female |
| Partner | whether the customer has a partner or not (Yes, No) |
| Tenure | number of months the customer has stayed with the company |
| Phone Service | whether the customer has a phone service or not (Yes, No) |
| Internet Service | customer’s internet service provider (DSL, Fiber optic, No) |
| Payment Method | the customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)) |
| Monthly Charges | the amount charged to the customer monthly |
| Total Charges | the total amount charged to the customer |
| Churn | the customer churned or not (Yes or No) |

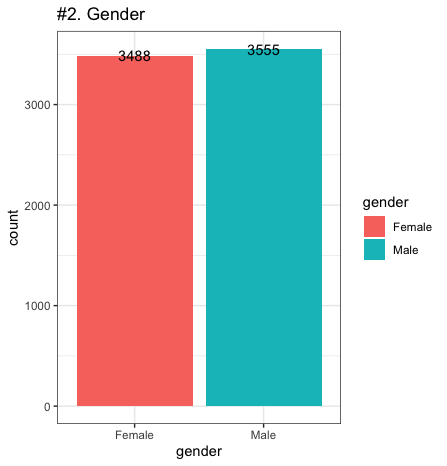
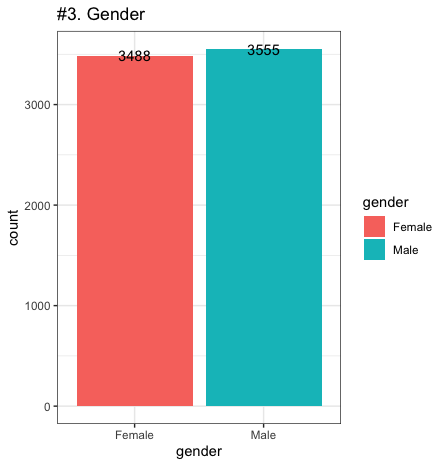
The objective is to predict churn or no churn in order to retain customers.

* 1. Data cleaning with VIS



The picture shows that there are no NA values in the columns except Tota lCharges. It contains 11 missing data, so what is required is to get rid of those rows from the dataset. After that, the dataset will be no missing values.

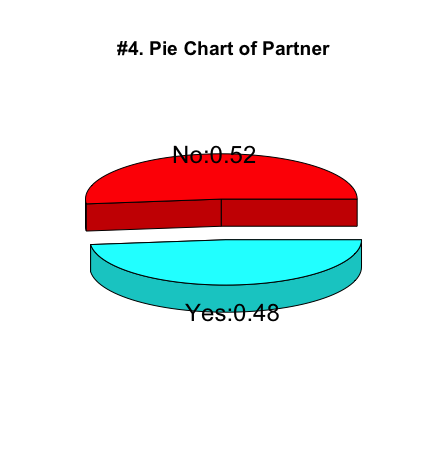
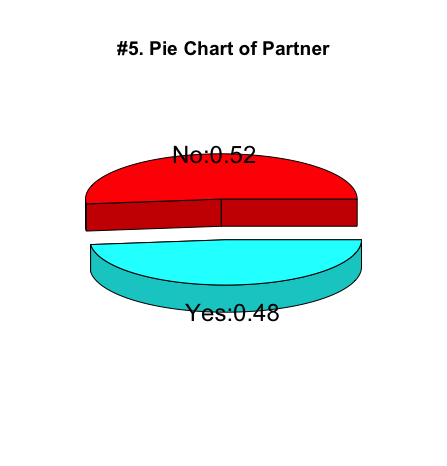
1. Gender

The first bar chart shows that there are no incorrect values in the gender column as it only shows “female” and “male” as the legend. 2488+3555=7043, which equals to the total rows, so there are no missing values neither.

The second bar chart is the same as the first chart, which can also demonstrate that the column gender is clean. It also displays that the company has near 1:1 ratio of male and female customers, which means that the service or product is both for males and females.

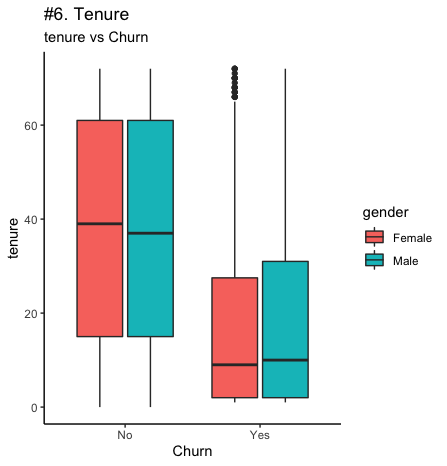
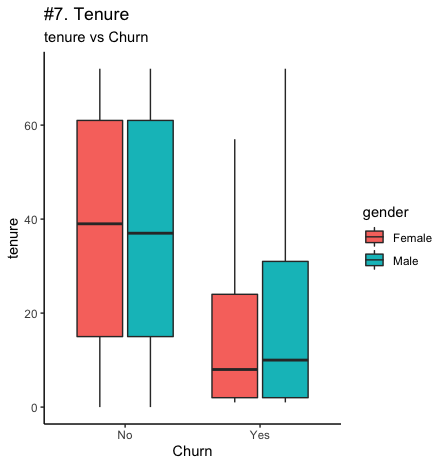
1. Partner

The first pie chart shows that more than half of the customers have no partners and 48% have. There are no incorrect values in the partner column, as it only shows red and blue two parts. 0.52+0.48 = 1, so there are no missing values neither.

The second pie chart is the same as the first one, which can demonstrate that the column partner is clean now. It also displays the same ratio of having or not having partners as the first pie chart. Most of customers (about 52%) have no partner.

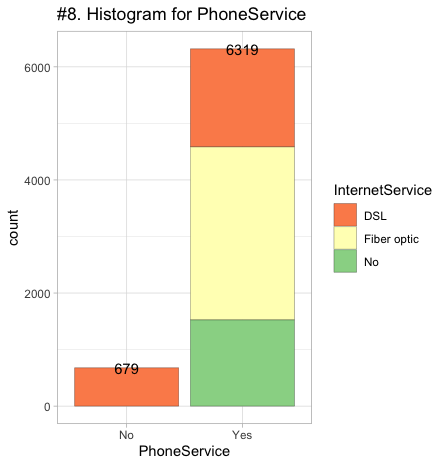
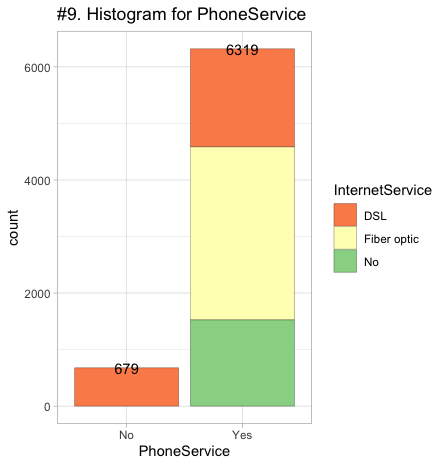
1. Tenure

The first boxplot shows that there are some outliers in the column tenure and these outliers belong to the rows where churn = “Yes” and gender = “Female”. The customers who left the platform have less tenures than those who chose to stay.

The second boxplot shows that the outliers have been removed and the female customers who have lower than 10 months tenure are more likely to leave. Therefore, the company mainly lose short-term customers.

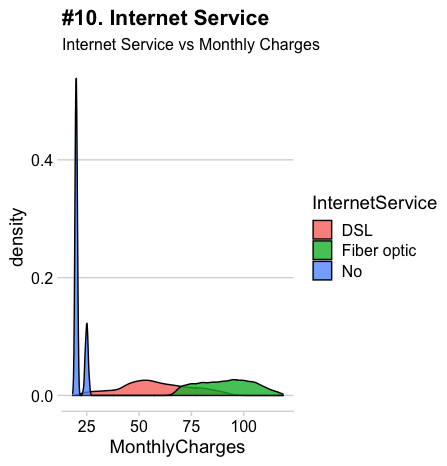
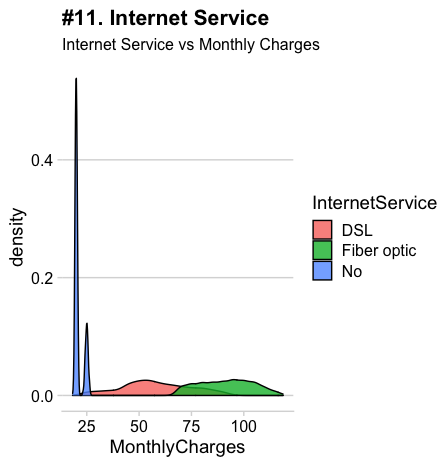
1. Phone service

The first plot shows that there are no incorrect values in the phone service column as x-axis only has “Yes” and “No” options. 6319+679=6998, which equals to the total rows minus outliers, so there are no missing values neither. Besides, most of customers have phone service.

The second plot is the same as the first one, which can also demonstrate that the column phone service is clean. It also displays that most of customers have both phone service and internet service and Fiber optic is the most favored internet service.

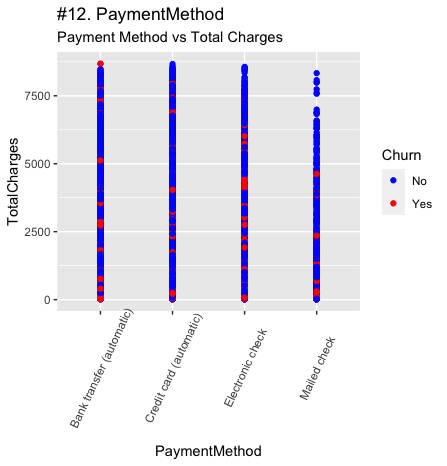
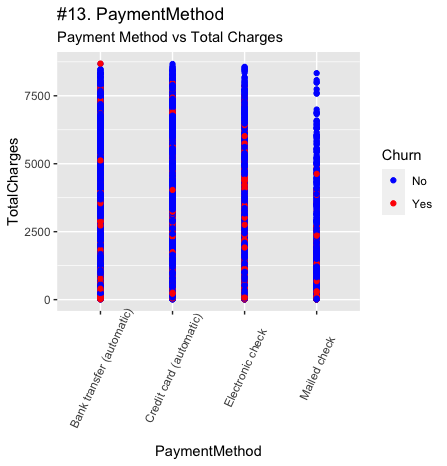
1. Internet service

The first density plot shows that there are no incorrect values, which do not belong to the three internet service options in the internet service column. Customers using fiber optic internet service have the largest monthly charges over than 60 dollars per month.

The second density plot is the same as the first one, which can also demonstrate that the column internet service is clean. It also displays that the monthly charges of customers who have no internet service has a high density.

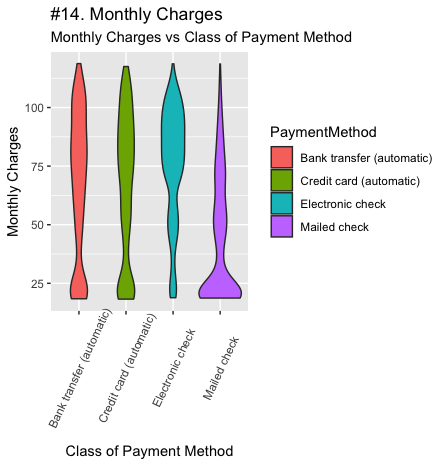
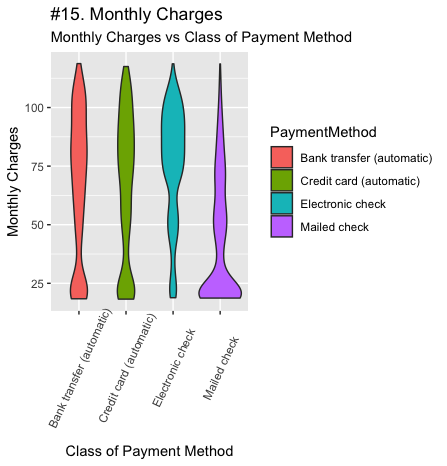
1. Payment method

The first point plot shows that there are no incorrect values, which do not belong to the four payment methods in the payment method column. And there are no outliers in the total charges column. The mailed check is mostly used for payment less than 7500 dollars.

The second plot is the same as the first one. It also shows the similar result that for total charges more than 7500 dollars, customers are less likely to pay by mailed check. Bank transfer and credit card are the most favored payment methods.

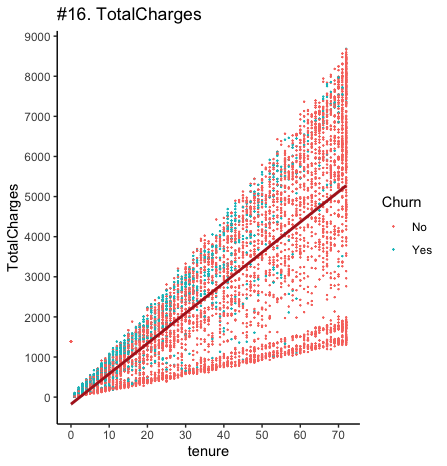
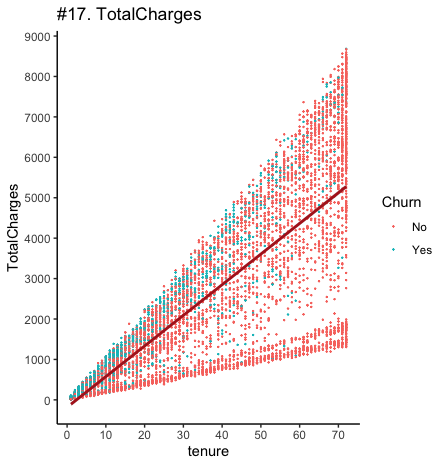
1. Monthly charges

The first violin plot shows that there are no outliers in the monthly charges column. There are also no incorrect values in the payment method column as the legend only has four options. But the shapes for four methods are different. Mailed check is mainly for monthly charges lower than 40 dollars. On the contrary, electronic check is manly for monthly charges over than 60 dollars.

The second violin plot is the same as the first one, which can also demonstrate that the column monthly charges is clean. It displays that for monthly charges more than 50 dollars, most of customers will pay by bank transfer, credit card, or electronic check, while for monthly charges less than 37.5 dollars, customers will prefer to pay by mailed check.

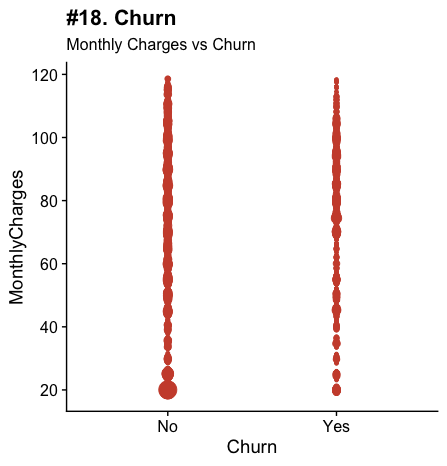
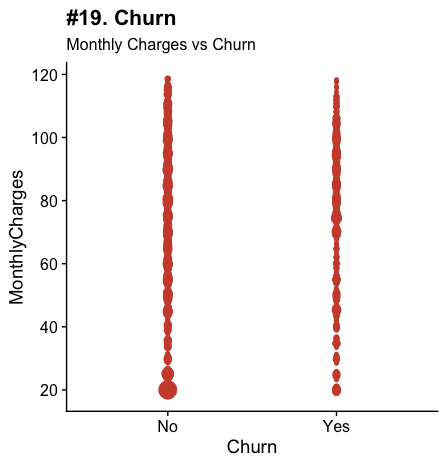
1. Total charges

At the beginning, the first plot tells 11 missing values in the total charges column. After removing them, the scatter point plot shows that there is still one incorrect value in this column as the tenure is 0 whereas the total charge is more than 1000 dollars. So what is required to do is find and remove this value.

The second scatter point plot shows that the incorrect value is removed. Customers who pay total charges less than 3500 dollars and whose tenure is less than 30 months are more likely to leave. Therefore, charge is an important factor that causes customers to churn.

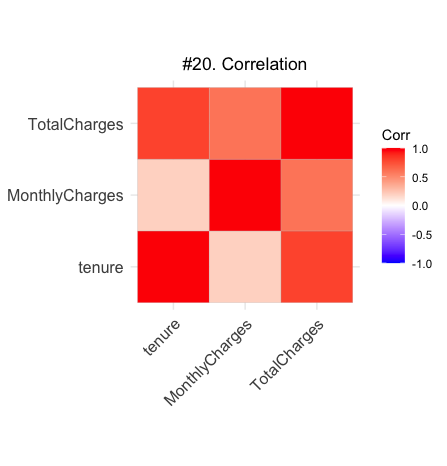
1. Churn

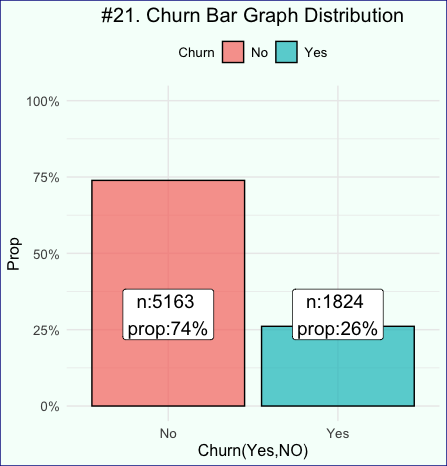
The first plot shows that there are no incorrect values in the column churn, as x-axis only has “Yes” and “No” choices. There are no outliers either because all the data are in two straight lines. With the conclusion before, the column has no missing values neither, so it is clean now.

The second plot is the same as the first one, which can also demonstrate that the column churn is clean. Besides, it displays that the customers who pay monthly charges more than 70 dollars or less than 30 dollars are more likely to leave.

* 1. Data EDA with VIS

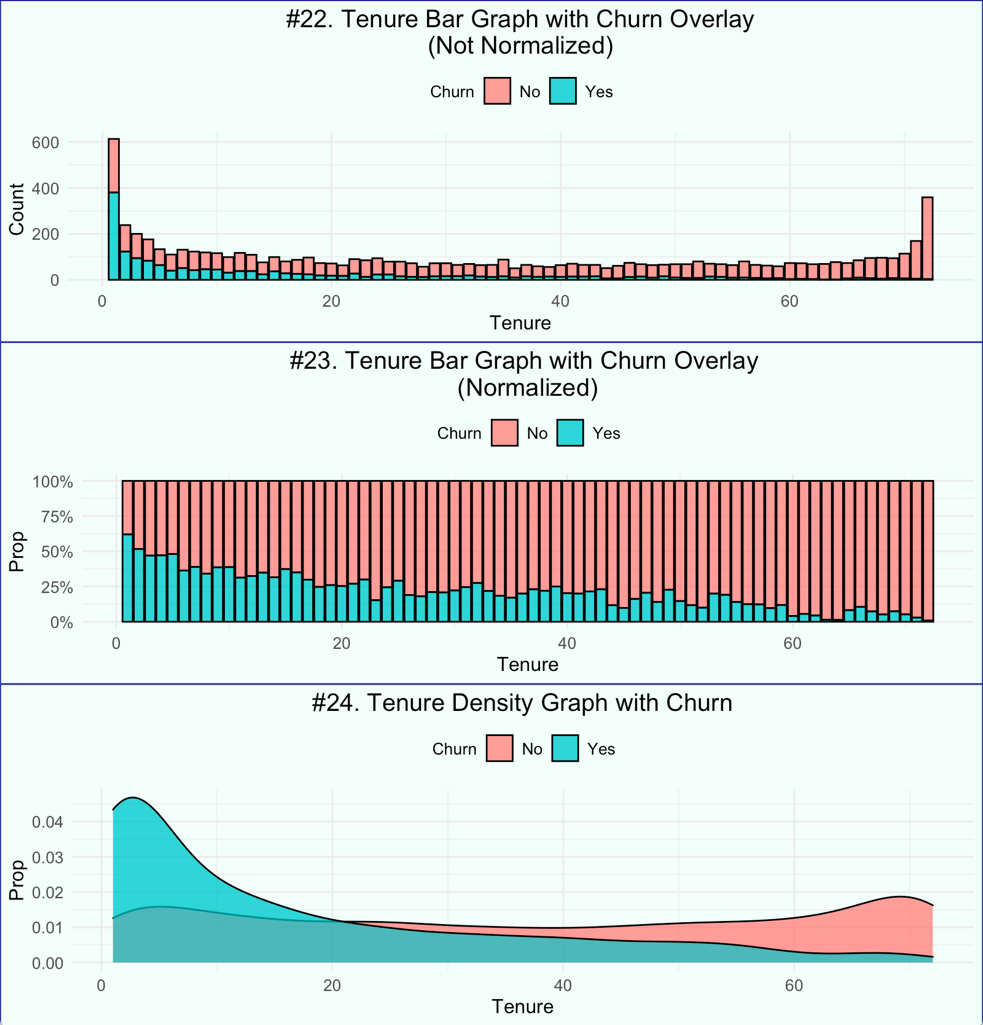


The correlation plot shows a strong positive correlation between tenure and total charges, a weak positive correlation between tenure and monthly charges, and a medium to strong positive correlation between monthly charges and total charges.

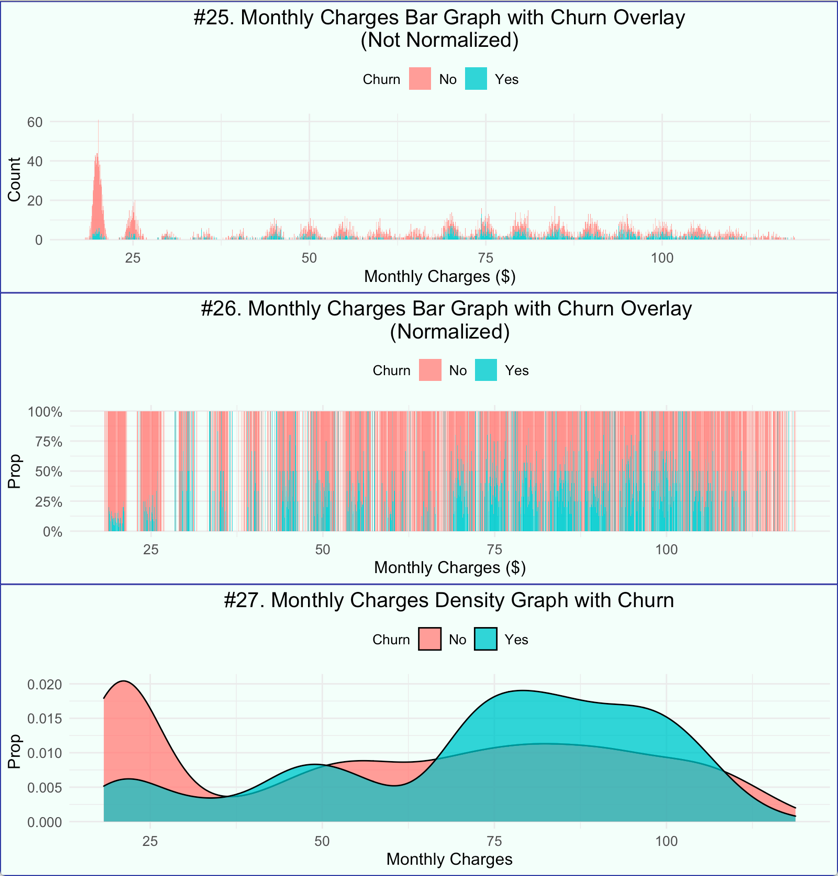


Based on the data, 74% of the customers stopped using the service. Only 26% are still active. Therefore, it is an important point for the company to perform the churn analysis to find the cause that make most of its customers leave.

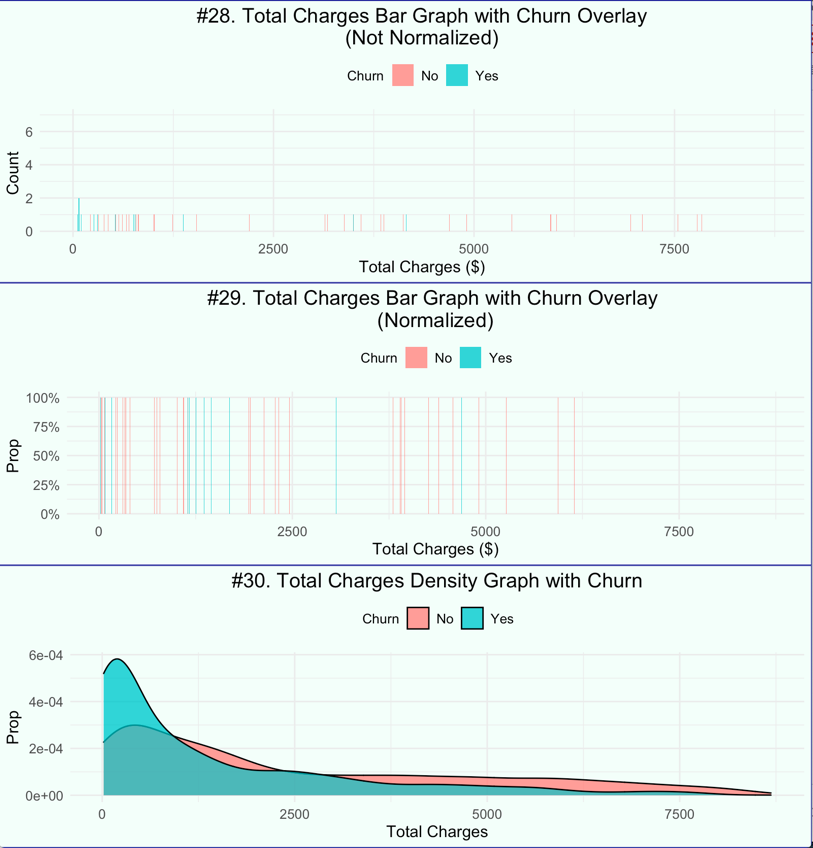
First, explore the quantitative variables: tenure VS. churn, monthly charges VS. churn, and total charges VS. churn.



The majority of customers who tend to leave have a tenure less than 20 months, which reflects that the company’s service does not cater to short-term customers. The cause may be poor customer service or lack of customer touchpoints. The company had better make its short-term program more attractive so as to reduce the churn rate for short-term customers.

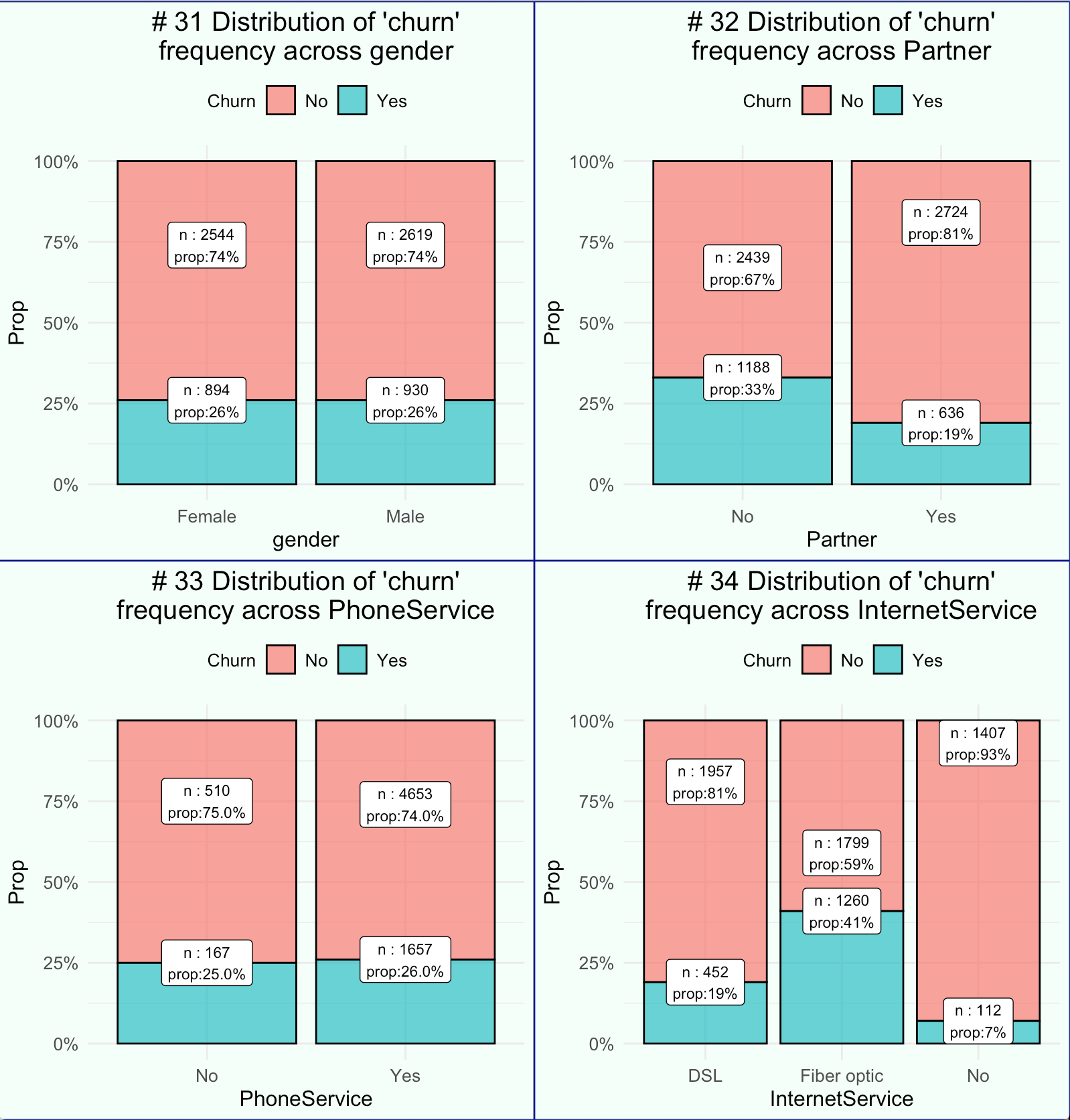


The majority of customers has low monthly charges. Those with the highest proportion of positive churn (left the platform) are the ones with high monthly charges (between ‘70’ to ‘112’$ / month). As what has analyzed before, charge is an essential factor that affects the customers’ choice to leave or stay. Therefore, the company need to adjust its monthly charges to decrease the churn rate.



The plots clearly shows that the highest proportion of positive churn are customers who has low total charges (from ‘0’ to ‘2000’ dollars). It seems a contradictory result versus the monthly charges, but it is reasonable. Higher the total charge means an expensive program, which, in other words, is a better and more considerable program so that the customers would like to stay. However, the plot reflects that the customers were not satisfied with the cheap program.

Then, explore the categorical variables VS. churn:

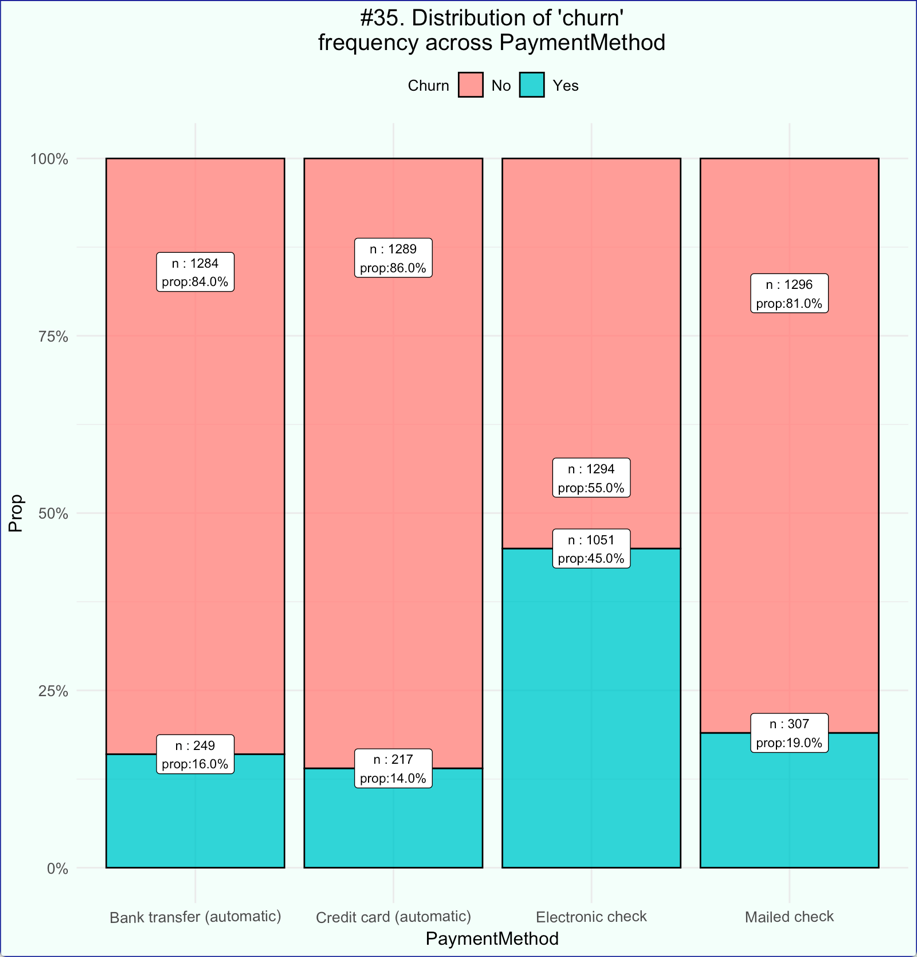


There is no insight that the company could get from gender as both female and male have almost the same proportion with regard of the churn variable.

33% of the customers with no partner left the platform. 81% of customers with partner didn’t leave.

Both customers with phone service and without share the same proportion. Therefore, phone service does not affect the churn rate greatly.

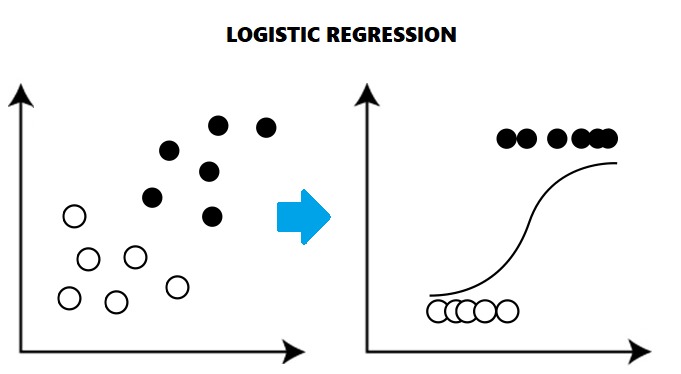
41% customers of fiber optic services left the platform. 93% of customers with no internet services did not leave. 81% of the customers with DSL services didn’t leave. Therefore, the fiber optic service need improvement.



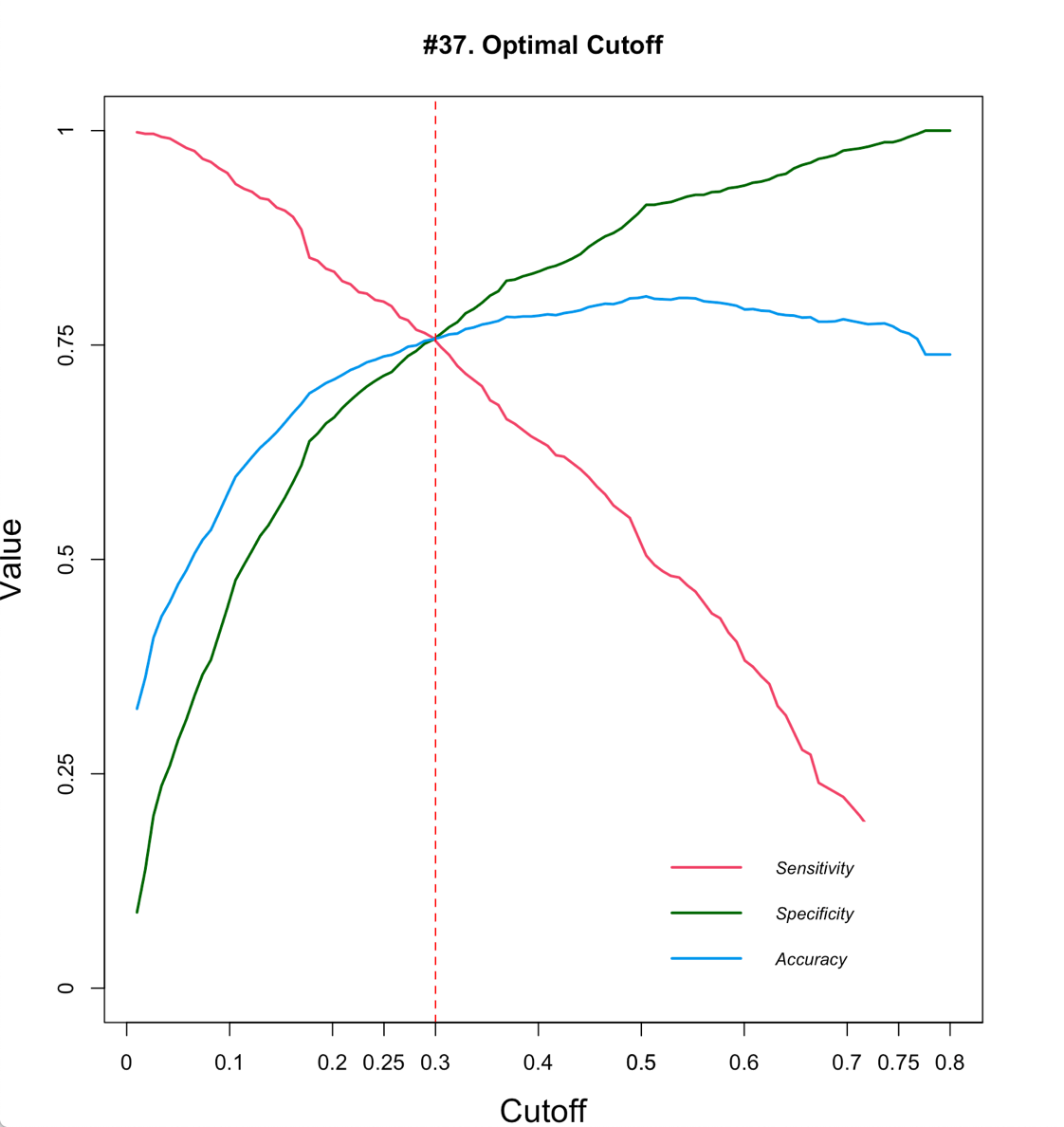
45% of customers with electronic check as payment method left the platform. 86% of customers using credit card as payment method stayed. What’s more, customers with bank transfer and mailed check as payment method also keep a high retention rate. So what the company need to focus on is electronic check payment service.

* 1. ML models used and how they work
     1. Logistic Regression

**Logistic Regression** – Logistic regression is used to predict a discrete outcome based on variables which may be discrete, continuous or mixed. Thus, when the dependent variable has two or more discrete outcomes, logistic regression is a commonly used technique. The outcome could be in the form of Yes / No, 1 / 0, True / False, High/Low, given a set of independent variables.

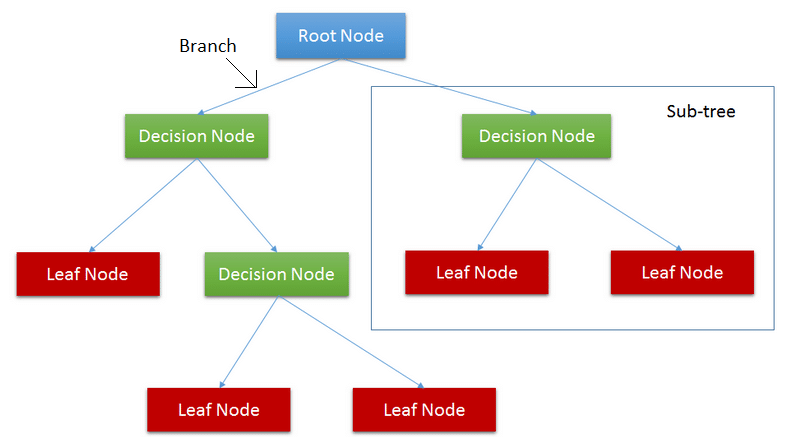


In the analysis, Logistic regression will be used to predict the variable Churn (Yes or No). At first, all predictors will be included to build the model1. Then stepAIC will be used to perform feature selection, which is an iterative process of adding or removing variables, in order to get a subset of variables that gives the best performing model. With significant variables, optimal cutoff will be chosen as the plot shows below.

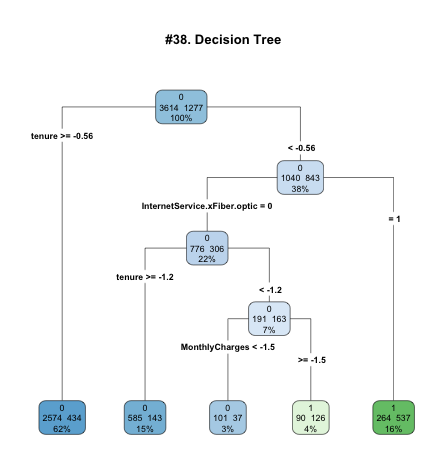


2.4.2 Decision Tree

**Decision Tree** - Splits the data into multiple sets and each set is further split into subsets to arrive at a tree like structure and make a decision. Homogeneity is the basic concept that helps to determine the attribute on which a split should be made. A split that results into the most homogenous subset is often considered better and step by step each attribute is chosen that maximizes the homogeneity of each subset. Further, this homogeneity is measured using different ways such as Gini Index, Entropy and Information Gain.

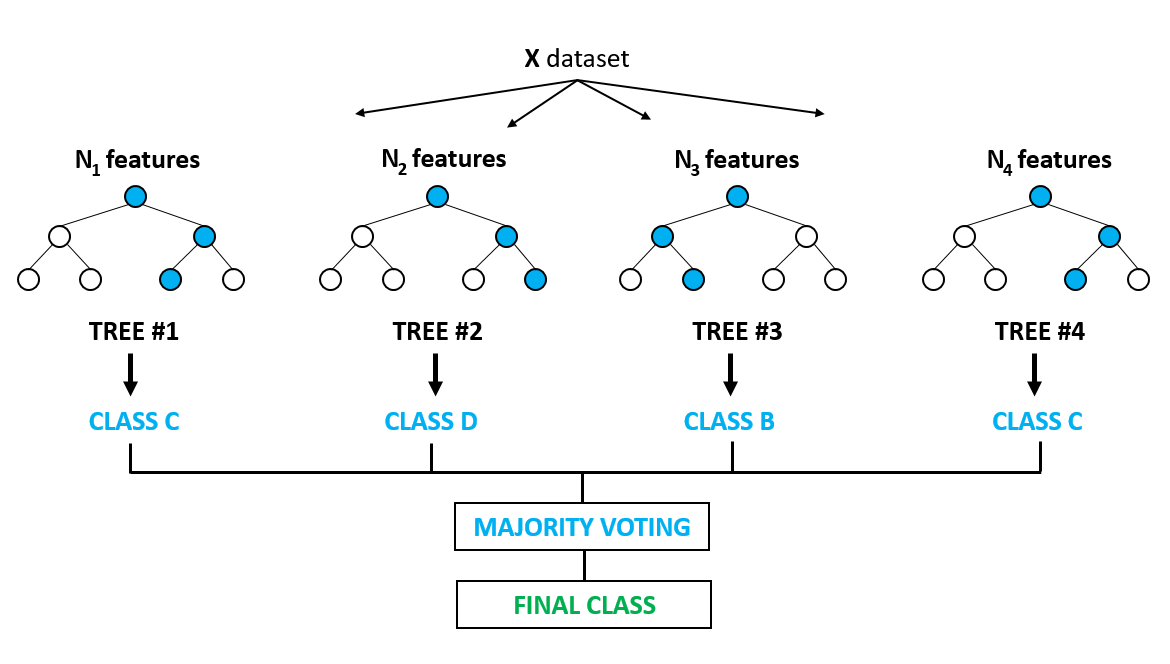


In the analysis, decision tree model will be trained using all variables and predict in the validation data. The result will be tree-structured, according to which, the company can predict whether a customer will leave according to the result of the predictors.

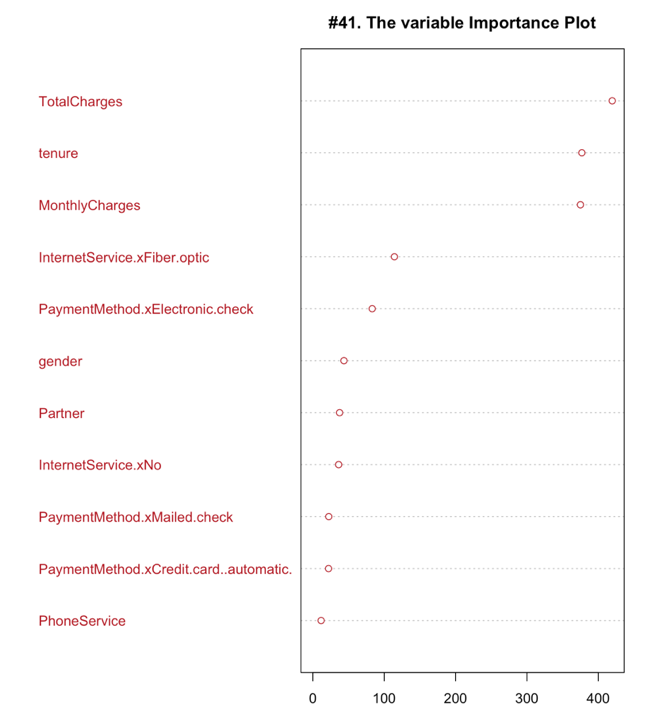


2.4.3 Random Forest

**Random Forest** - Often known as an ensemble of a large number of Decision Trees, that uses bootstrapped aggregation technique to choose random samples from a dataset to train each tree in the forest. The final prediction in a Random Forest is an aggregation of prediction of individual trees. One of the advantages of Random Forest is that, it gives out-of-bag (OOB) error estimates, which is the mean prediction error on a training sample, using the trees that do not have that training sample in their bootstrap sample. It may act as a cross validation error and eliminate the need of using test/validation data, thereby increasing the training the data.



In the analysis, random forest model will also be trained using all variables and predict in the validation data. After building the model Variable Importance Plot will be plotted to show the most significant attribute in decreasing order by mean decrease in Gini. The Mean decrease Gini measures how pure the nodes are at the end of the tree. Higher the Gini Index, better is the homogeneity.

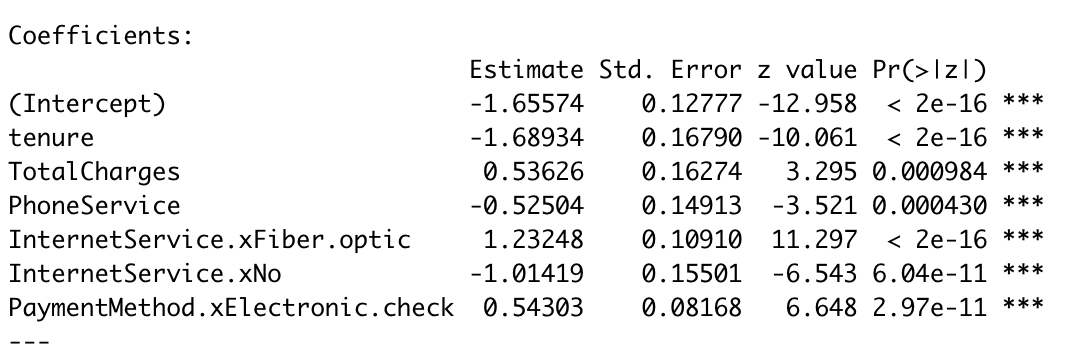


1. Results

Before building the model, the dataset will be prepared by standardizing continuous features, creating dummy variable, and splitting the data into training set and test set.

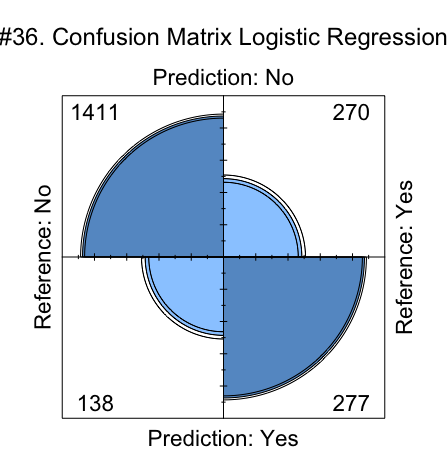
* 1. Logistic Regression

First, build the model 1 with all variables and then use stepAIC for variable selection. With the selected subset of variables, the best performing model can be got.



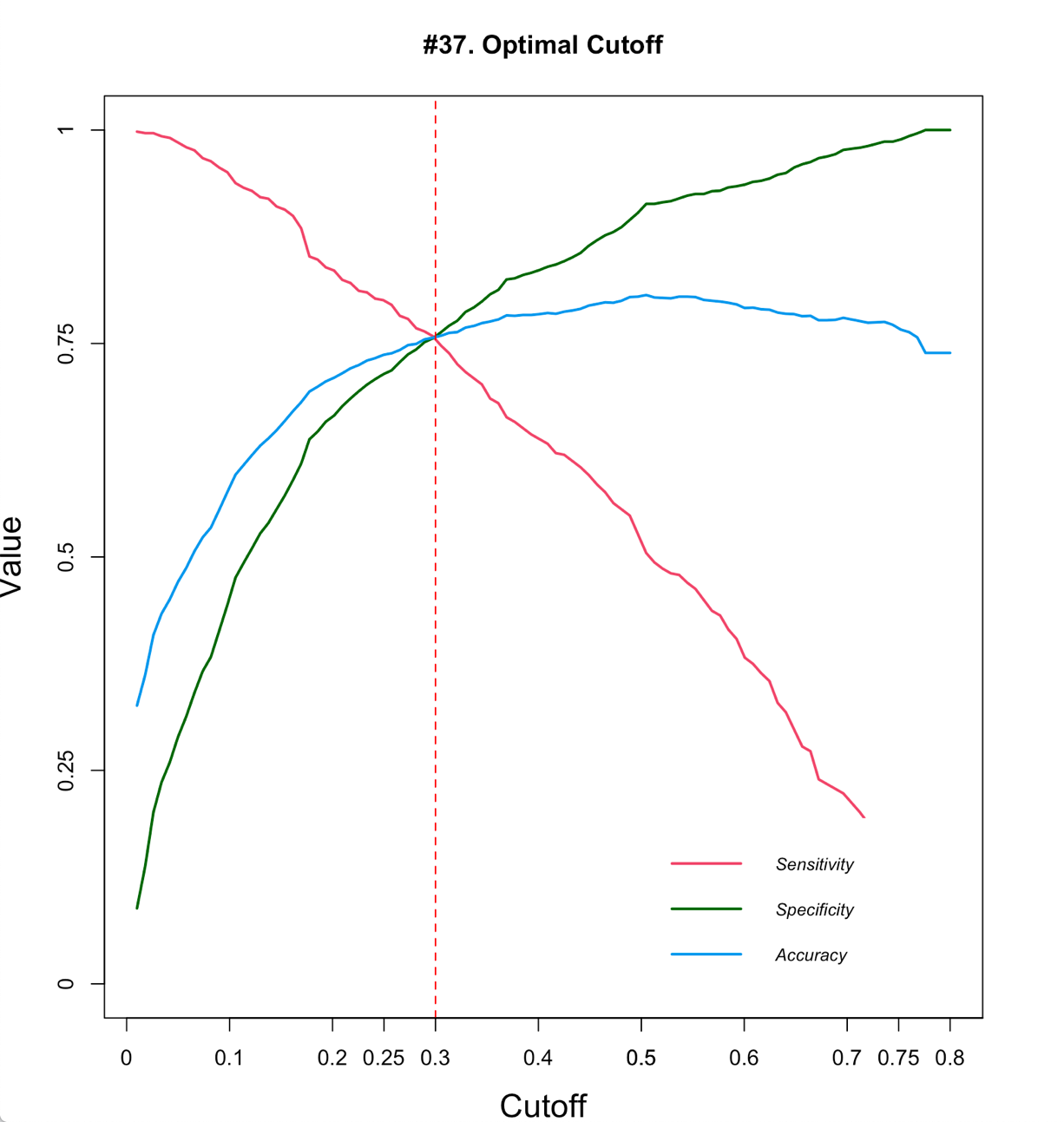
* This new fitted model displays that, holding all other features constant, the odds of churn for customers with phone service over those without is exp (-0.52504) = 0.5915317. In terms of percent change, the odds for customers with phone service are 40.8% lower than the odds for customers without.
* The coefficient for tenure says that, holding all others constant, the company will see 80.9% decrease in the odds of churn for a one month increase in tenure since exp (-1.655745) = 0.1909497.
* Similarly, with one more dollar that the customer has to pay for, there is a 71% increase in the odds of churn.

As for the model evaluation, the confusion matrix will be plotted.

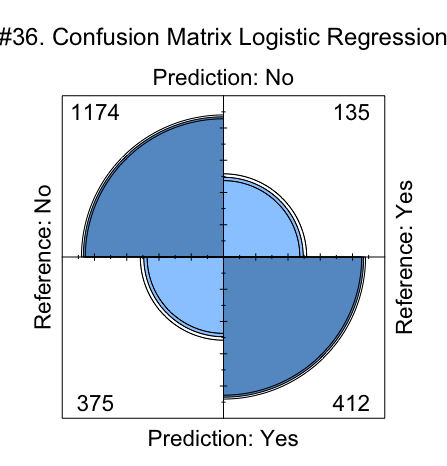


The confusion matrix shows the predicted probabilities of No (not leave) and Yes (leave the platform) for the test data with a default classification threshold of 0.5.

Since the model is predicting churn too many Type II errors is not advisable. A False Negative (ignoring the probability of churn = No when there actually is No) is more dangerous than a False Positive in this case. As the company care more about the people who have the willingness to leave, threshold can be lowered to increase the sensitivity.

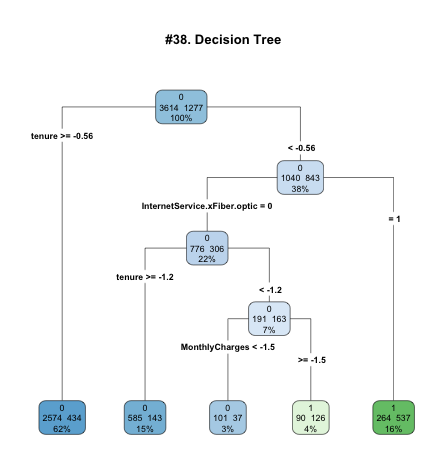


As the plot shows, 0.3 threshold is optimal because it is where the three curves for accuracy, specificity and sensitivity meet. With 0.3 threshold, the new confusion matrix shows below. Accuracy for those who will leave the platform with threshold 0.3 is 75.67%.



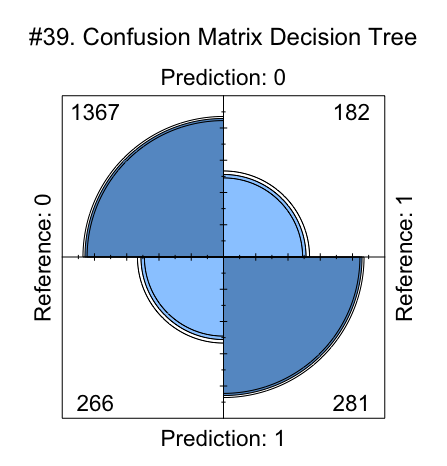
* 1. Decision Tree

Decision tree model is trained using all variables with the method “class” and the result shows as below.



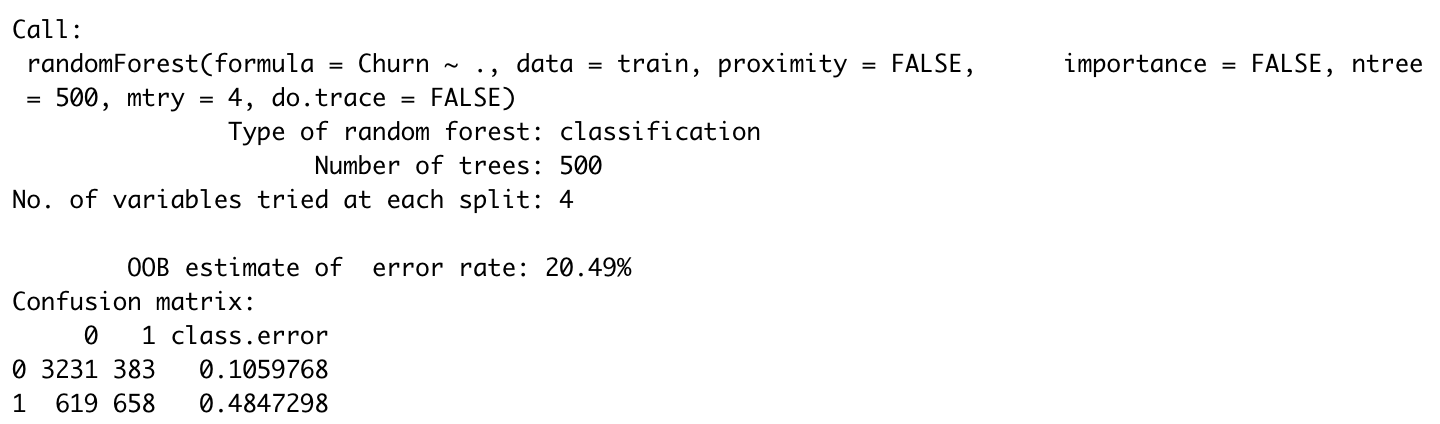
* 62% customers whose tenure is over than -0.56 (after standardization) will not leave the platform.
* 15% customers whose tenure is more than -1.2 and less than -0.56 (after standardization) and who do not use fiber optic internet service will not give up the service.
* 3% customers whose tenure is less than -1.2 (after standardization), monthly charge is less then -1.5 (after standardization) and who do not use fiber optic internet service will stay.
* However, 16% customers whose tenure is less than -0.56 (after standardization) and who use fiber optic internet service will leave the platform.
* 4% customers whose tenure is less than -1.2 (after standardization), monthly charge is more then -1.5 (after standardization) and who do not use fiber optic internet service will also drop off the service.

The decision tree model (accuracy 78.63%) gives slightly better accuracy with respect to the logistic regression model (accuracy 75.67%). The sensitivity is also better in case of Decision tree which is 83.71%. However, the specificity has decreased to 60.69% in case of Decision Tree as compared to logistic regression model.



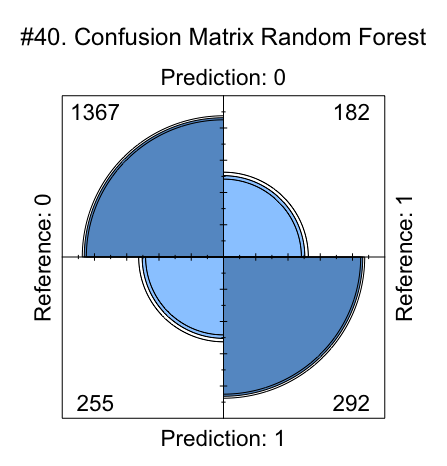
* 1. Random Forest

Random forest model is also trained using all variables with the method “class”. Set ntree (number of trees to grow) = 500 and mtry (number of variables randomly sampled as candidates at each split) = 4. The result shows as below.

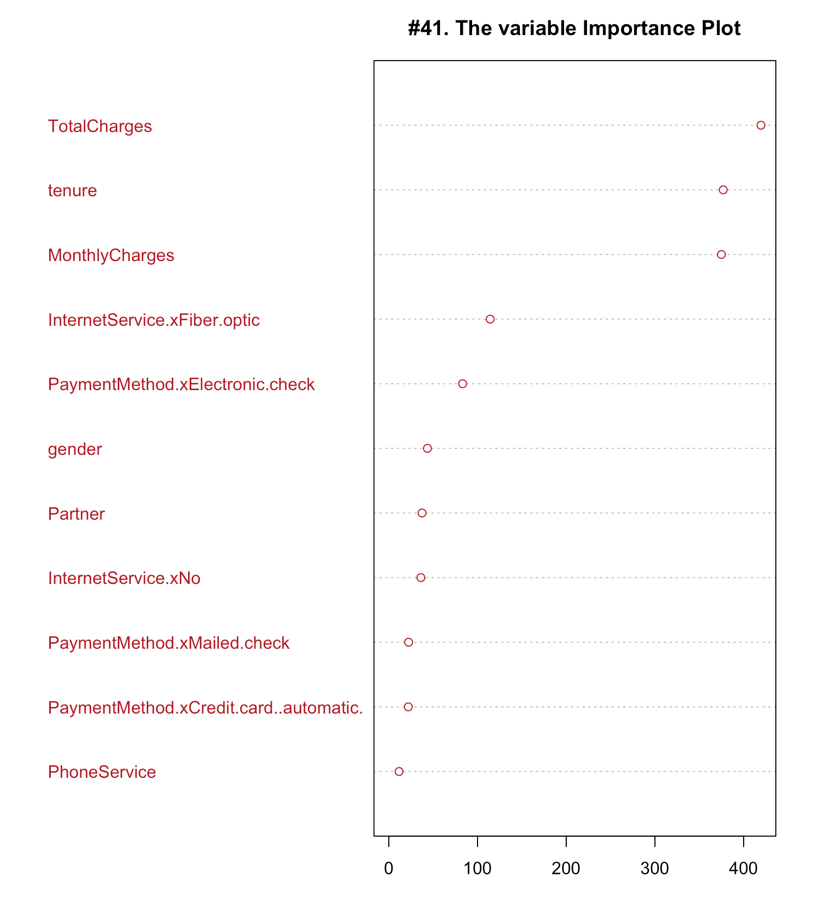


The OOB error estimate comes to around 20.49%, so the model has around 79.51% out of sample accuracy for the training set. Next, check the prediction and accuracy on the test data.

The basic random forest model gives an accuracy of 79.15% (almost close enough to the OOB estimate), Sensitivity 84.28% and Specificity 61.60%.

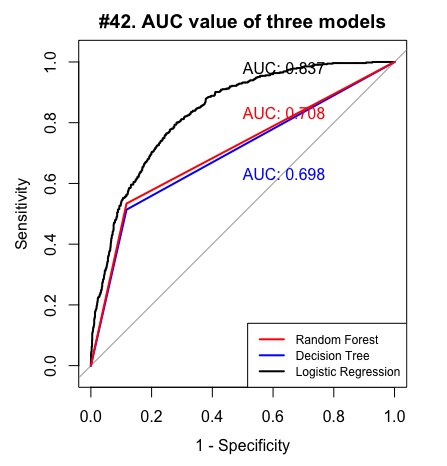


Below is the variable importance plot, which shows the most significant attribute in decreasing order by mean decrease in Gini. The Mean decrease Gini measures how pure the nodes are at the end of the tree. Higher the Gini Index, better is the homogeneity.

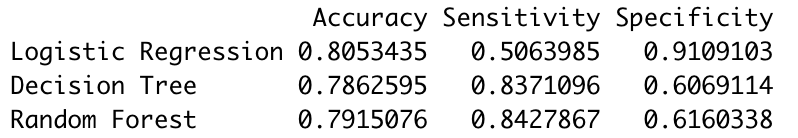


It displays that total charges, monthly charges, and tenure are the three most important attributes in the random forest model.

* 1. Comparison of three models



The ROC curves show that logistic regression model has the largest AUC value (equals to 0.837), followed by the random forest model with 0.708. Decision tree model has the least AUC value of 0.698. Therefore, in terms of AUC value, logistic regression model performs the best.



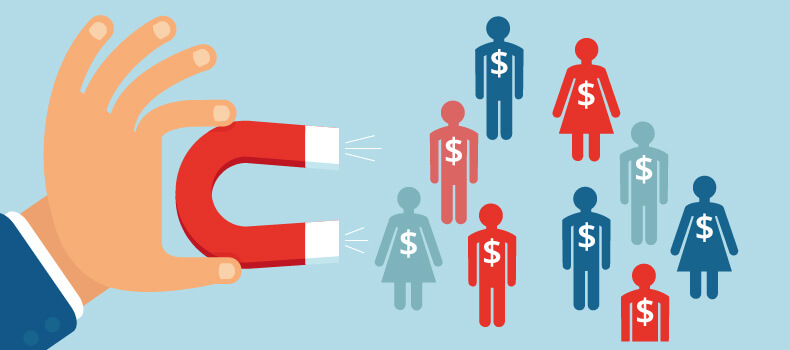
The table compares the accuracy, sensitivity, and specificity from three models, when logistic regression model with 0.5 threshold at this time. It shows that logistic regression model has the largest accuracy and specificity, while random forest model has the largest sensitivity. The three evaluation indexes of decision tree model are close to those of random forest model.

1. Conclusions

For the company, 74% of its customers stopped using the service. Only 26% are still active. Therefore, it is an important point to perform the churn analysis to find the causes that make most of its customers leave. Tenure, monthly charges, total charges, whether the customer has a partner, internet service, and payment methods are important factors that need to be taken into consideration.

For tenure, the customers who have less than 20 months tenure are more likely to leave, which reflects that the company’s service does not cater to long-term customers. The cause may be poor customer service or lack of touchpoints after conversion so that customers lose the interest to use it again.

**Suggestion 1:** The company had better improve its long-term customer service and follow up with customer touchpoints in a variety of ways, from free birthday gifts to emailed discounts.



For charges, most of its customers have low monthly charges. Customers whose monthly charges is between 70 to 112 dollars and total charges between 0 to 2000 dollars are highly possibly to leave the platform. Therefore, the company can assume that the problem may lie on the cheap program, which fails to satisfy customers’ needs or live up to customer expectations so that they choose to leave.

**Suggestion 2:** The company can put more efforts to improve its cheap service program by doing more marketing research on the customers whose monthly charges is lower than 112 dollars or total charges is less than 2000 dollars.

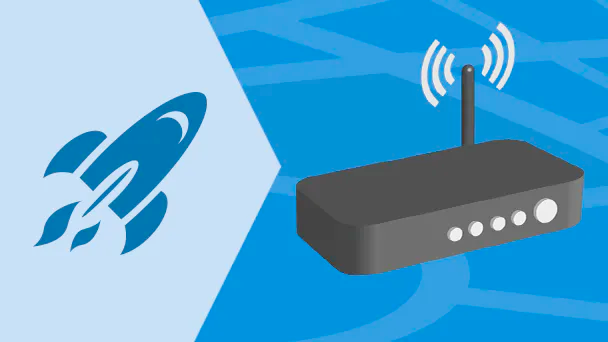
For partner, 33% of the customers with no partner left the platform, while 81% of customers with partner didn’t leave.

**Suggestion 3:** The founding gives the hint that the company can provide the service which is more convenient for singles or help them get rid of loneliness.



For internet service, 41% customers who used fiber optic services left the platform. 93% of customers with no internet services did not leave. 81% of the customers with DSL services didn’t leave. Therefore, the fiber optic service need improvement. The situation means that internet service does not have a positive impact on attracting customers, especially the fiber optic service.

**Suggestion 4:** The company should update the fiber optic service significantly or offer more internet service options like cable service for its customers to remedy this problem.



Finally, for payment methods, 45% of customers with electronic check as payment method left the platform. 86% of customers using credit card stayed. Customers with bank transfer and mailed check also keep a high retention rate. So, the problem lies on the electronic check payment service.

**Suggestion 5:** The company need to improve its electronic check payment service by increasing the efficiency or making it more reliable.

