***Telco Customer Churn Prediction***

Santosh Kumar Biswal (M12988685)

University of Cincinnati – MS-BANA Capstone, Summer 2019

Readers:

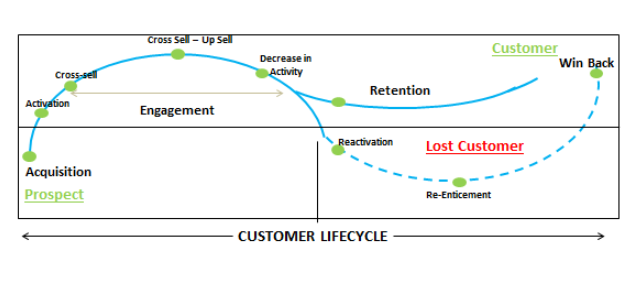
Asst. Professor Dungang Liu, University of Cincinnati

Asst. Professor Yichen Qin, University of Cincinnati

July 2019

**Abstract**

Customer churn is the loss of customers.The goal of the project is to predict the churn rate of the customers for one the telecommunication client. Knowing how churn rate varies by time of the week or month, product line can be modified according to customer response. A very methodological approach has been followed. we start with data cleaning and exploratory data analysis following which various machine learning algorithms like logistic regression, Decision trees, Random forest used to formulate an appropriate model giving out the best results, i.e lowest misclassification rates. Random Forest found to be best model to predict churn rate and important factor contributing to churn rate is “Monthly Charges”.



Contents

[Abstract 2](#_Toc13824714)

**1. Introduction……………………………………………………………….…………….…………………………………4**

[1.1 Objective 4](#_Toc13824716)

**1.2**  **Churn Rate Definition…………………………………………………………………………………………….4**

[1.3 Summary of Data 4](#_Toc13824717)

[1.4 Data Integrity](#_Toc13824718) 5

[2. Exploratory data analysis 5](#_Toc13824721)

[3. Data Mining methods and techniques](#_Toc13824721) 9

[3.1 Logistic Regression 9](#_Toc13824725)

[3.2 Decision Tree 11](#_Toc13824726)

[3.3 Random Forest 12](#_Toc13824727)

[3.4 Variable Importance 12](#_Toc13824728)

[4. Recommendations 13](#_Toc13824729)

[5. Conclusion 13](#_Toc13824730)

**6.** **References………………………………………………………………………………………………….13**

# **Introduction**

**1.1 Objective**

Churn prediction is one of the common applications of the classification in the business settings. In most of the churn prediction studies, the objective of classification is to minimize the prediction error and accordingly maximize the accuracy of the prediction.

**1.2 Churn Rate Definition**

Churn rate is calculated by dividing the number of customer cancellations within a period by the number of active customers at the start of that period. Very valuable insights can be gathered from this simple analysis — for example, the overall churn rate can provide a benchmark against which to measure the impact of a model.

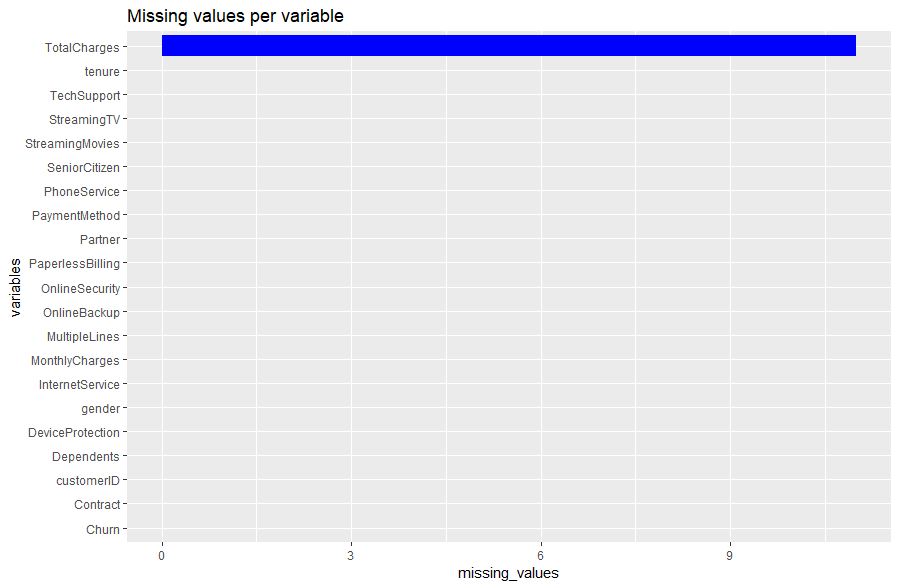
**1.3 Summary of Data**

* The data has been taken up from Kaggle.com under the Telco customer churn data.<https://www.kaggle.com/blastchar/telco-customer-churn>
* Each row represents a customer, each column contains customer’s attributes described on the column Metadata.
* The raw data contains 7043 rows (customers) and 21 columns (features).
* The “Churn” column is our target.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

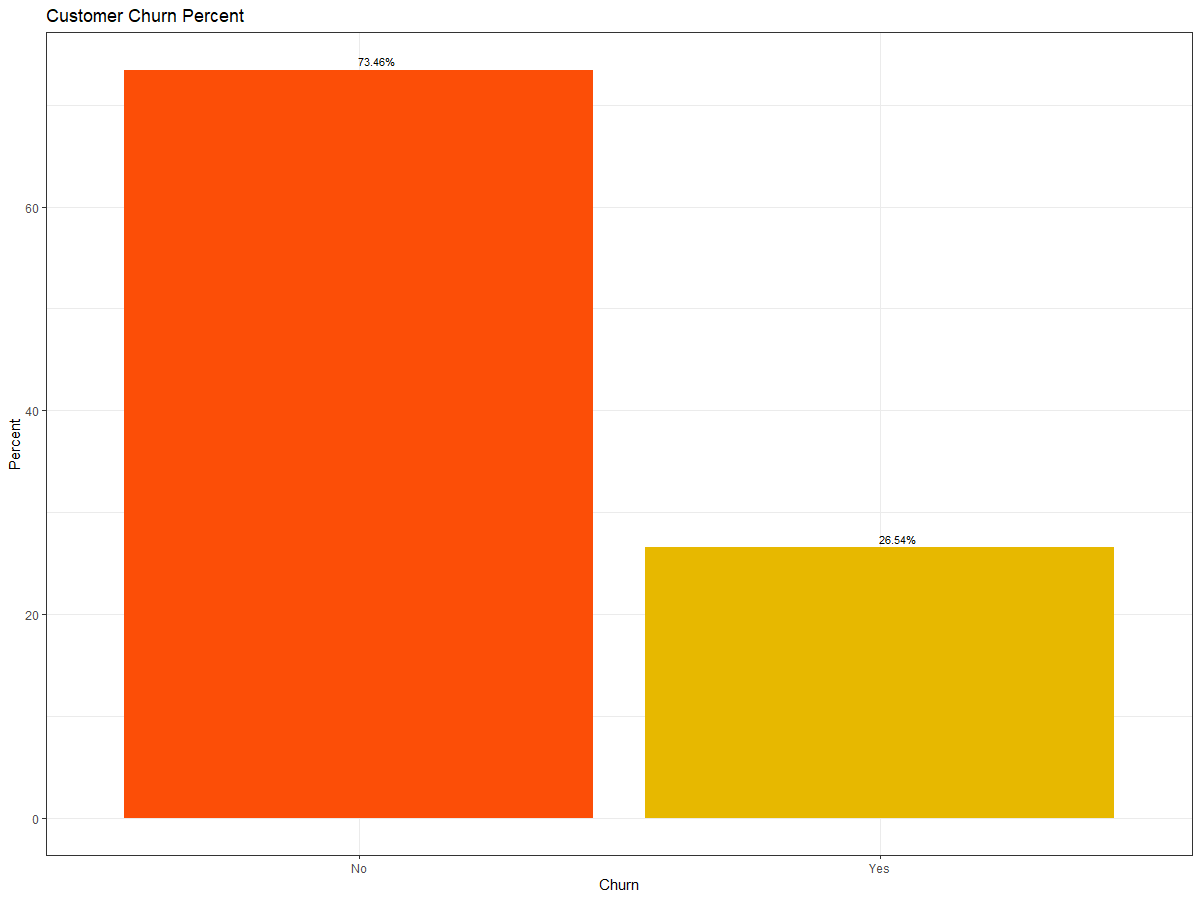
**1.4 Data Integrity**



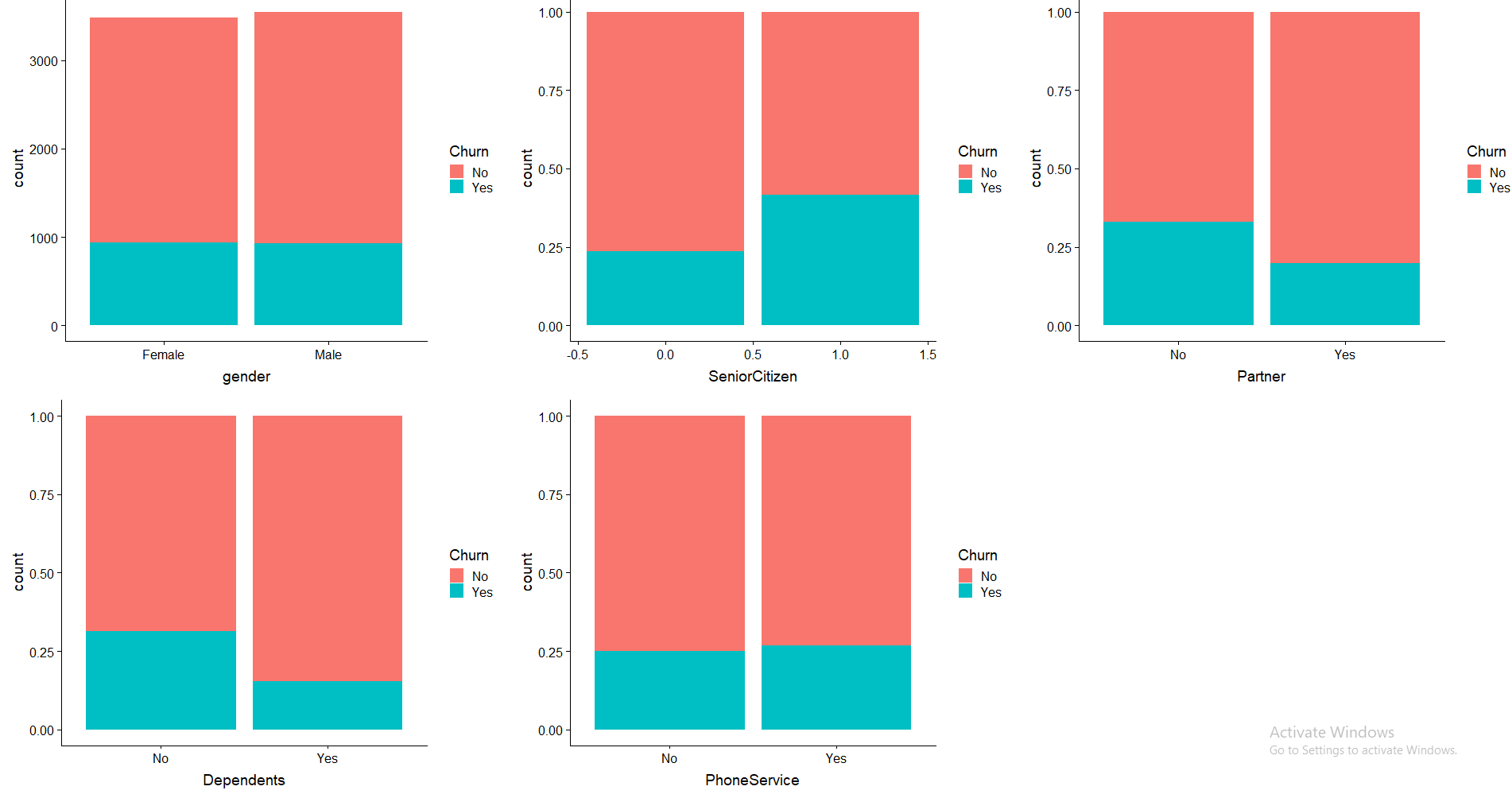
The variable total charges have some values missing which is handled in data cleaning process

1. **Exploratory Data Analysis:**

* In this section we see how individual factors affect the churn rate
* It can be seen from the below graph that customer likely to churn are nearly 27% of the total customers.



* Gender - The churn percent is almost equal in case of Male and Females
* The percent of churn is higher for senior citizens
* Customers taking phone services does not affect the churn rate.
* Customers with Partners and Dependents have lower churn rate as compared to those who don't have partners & Dependents.



*Figure: 2 Churn v/s demographic of customers*

* Churn rate is higher for the customer opting for the Fiber optic services
* Customers with no internet services are likely to leave than for customers with Internet services.
* Customers who opted for internet services and then availed services such as Online Security, Online Backup and device protection are more likely to leave.

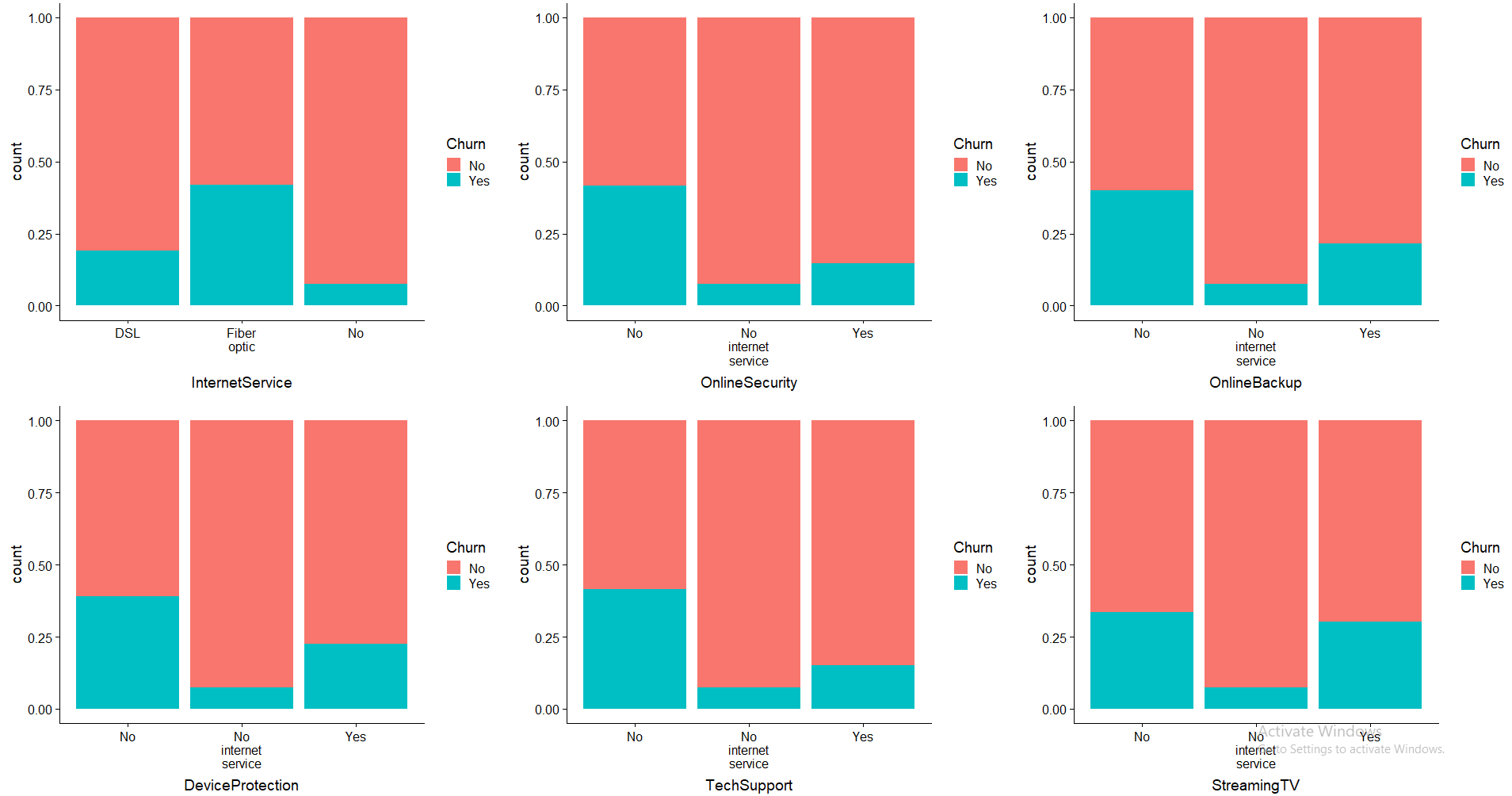


Figure: 3 Churn v/s services availed by the customers

* Tenure of customer with the company and count

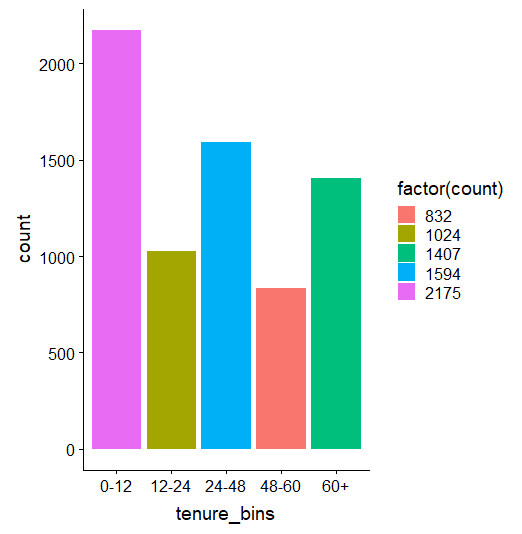
**

Figure: 4 Tenure of customer with the company and count

* It is evident from the below graph that the low paying customers are more likely to leave which is kind of good as high percentage of well-paying customers are likely to stay.

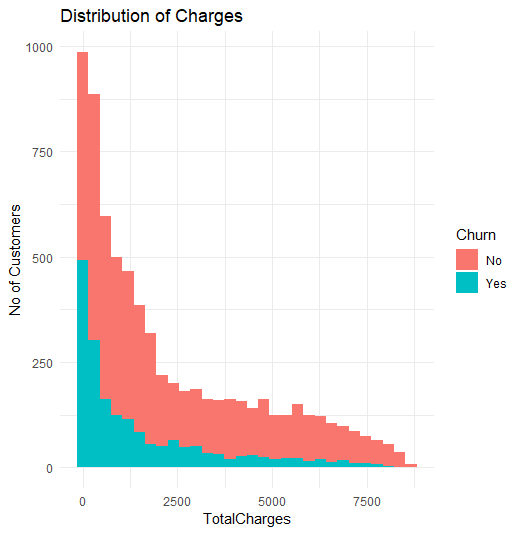
**

Figure: 5 Distribution of charges

* Contracts and Tenures are related to churn rate
* The longer a person stays with the company, the higher the total amount of charges.

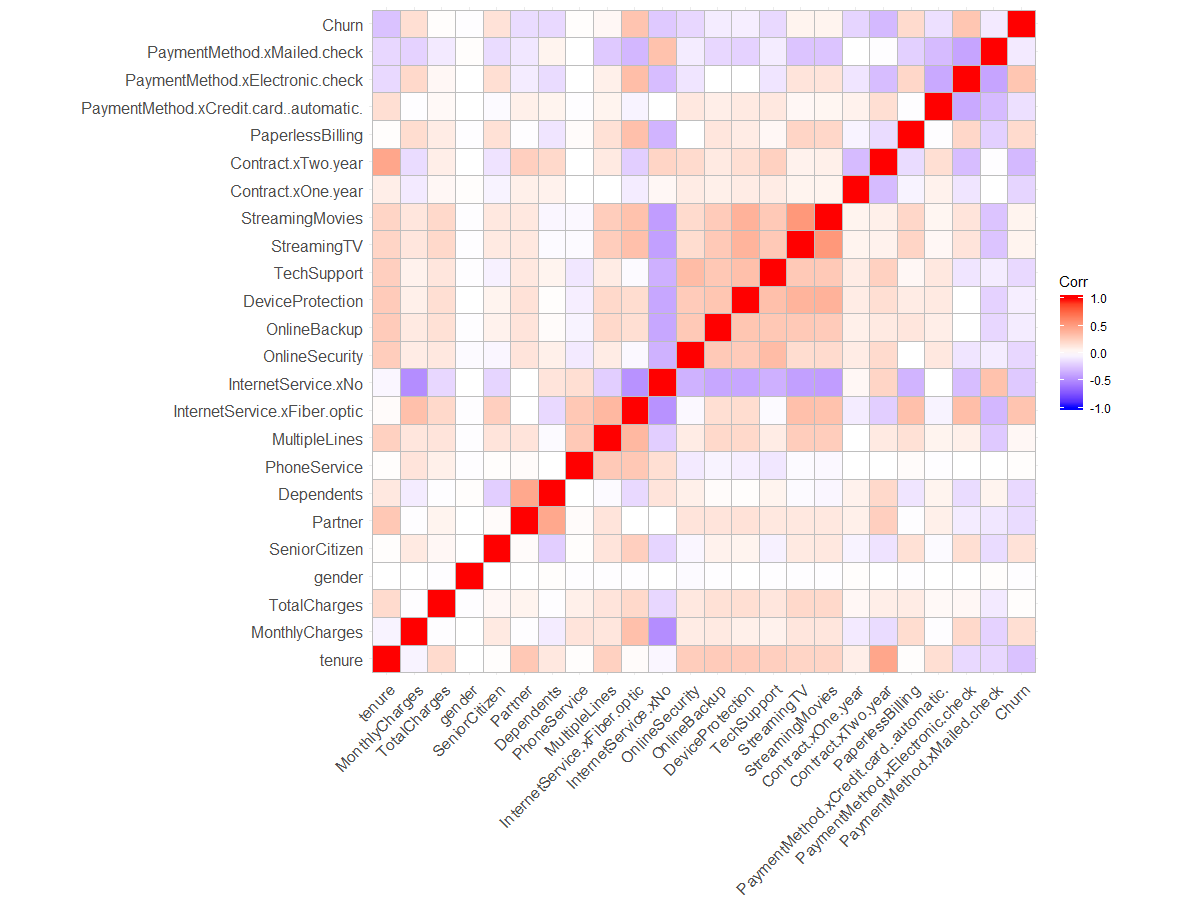
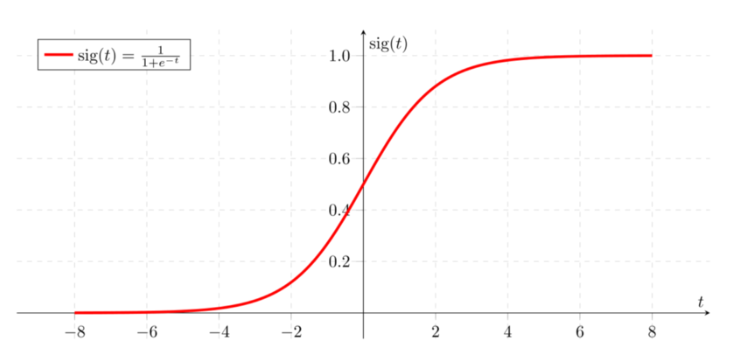


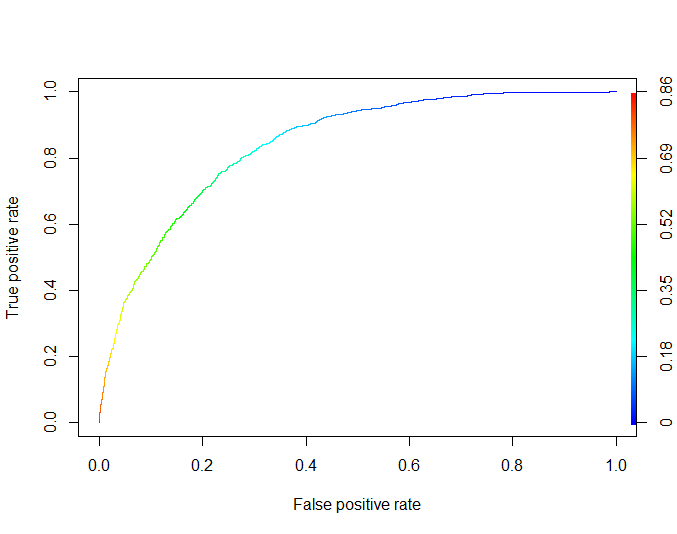
Figure: 6 Correlation of factors

1. **Data Mining Methods and Techniques**
   1. **Logistic Regression: -**

[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.



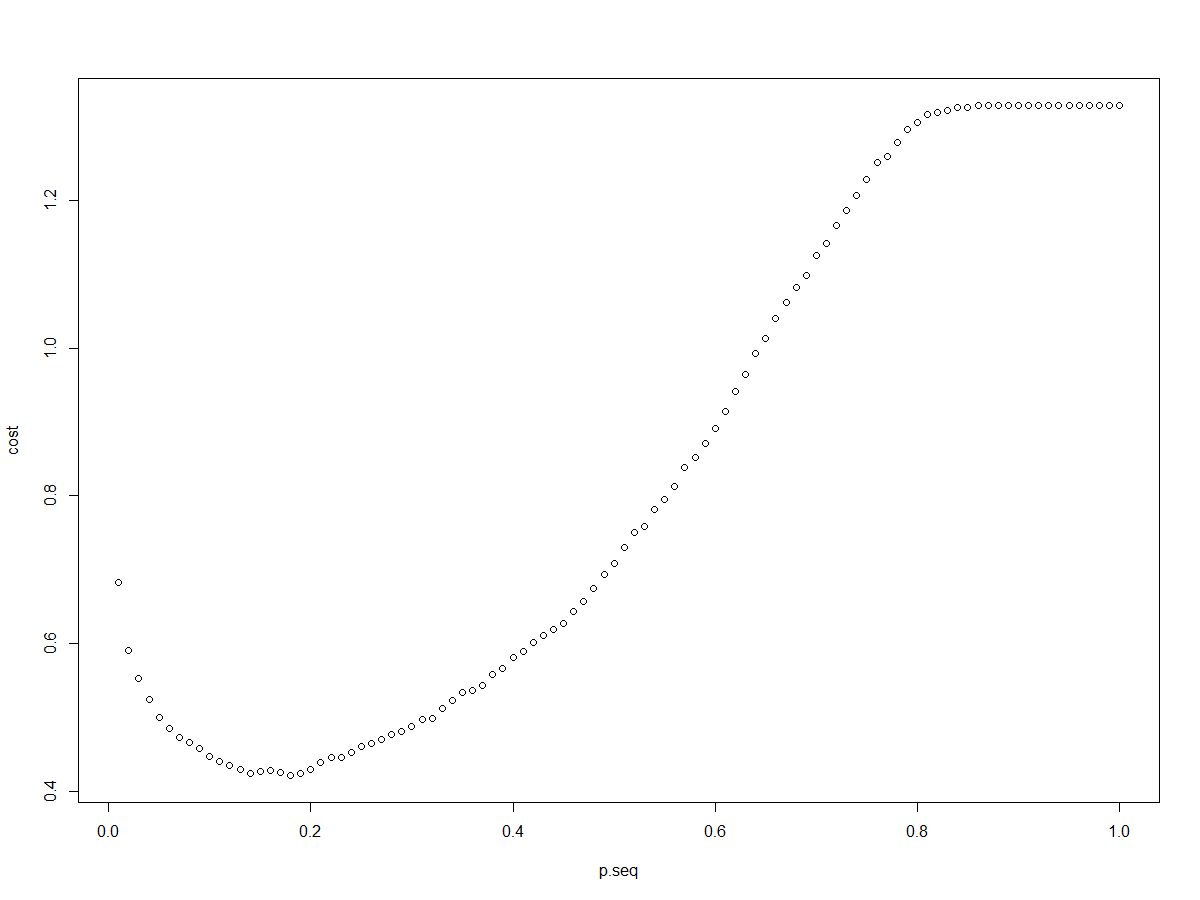
Sigmoid function curve



AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting Yes as Yes and No as No.

In sample AUC:-0.8315

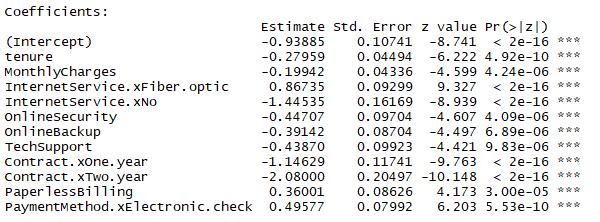
Optimal Cost Function:- The probability cut off value was determined to be 0.18.



**Stepwise Selection:-**

To reduce the complexity of the model and to eliminate the insignificant variables, a stepwise selection approach was used. An AIC Model and BIC model was run to determine the most efficient model.

Below were the results of the BIC model run



Accuracy - It determines the overall predicted accuracy of the model. It is calculated as

Accuracy = (True Positives + True Negatives)/ (True Positives + True Negatives + False Positives + False Negatives)

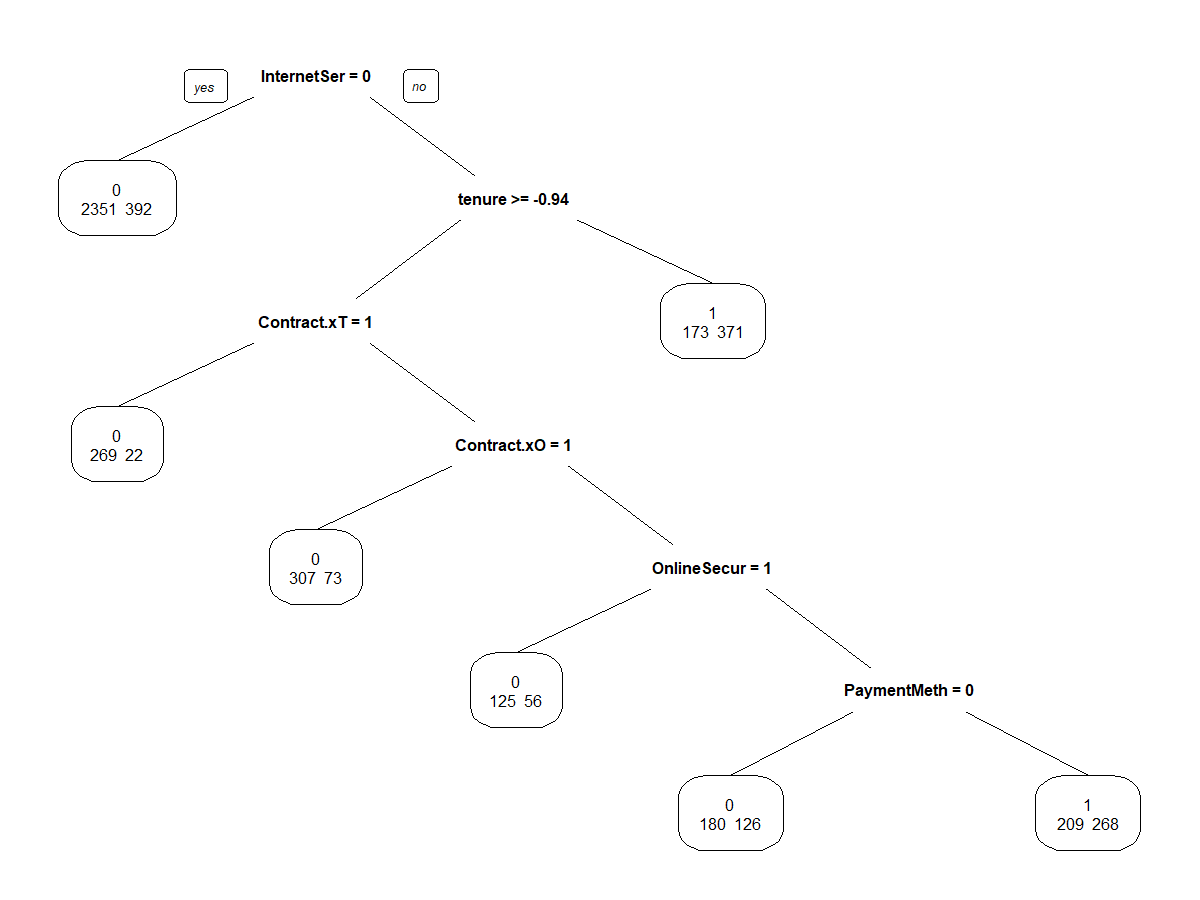
|  |  |  |
| --- | --- | --- |
|  | **Predicted(0)** | **Predicted(1)** |
| **True(0)** | 946 | 603 |
| **True(1)** | 73 | 488 |

Misclassification Rate: 0.3203

* 1. **Decision Tree**

Classification and regression trees are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree. Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values

Here as our response variable is binary i.e in the form of 0,1. We will use the classification tree for further analysis. First a default tree is constructed



**Misclassification Rate (Out-of-sample)**

|  |  |  |
| --- | --- | --- |
|  | **Predicted(0)** | **Predicted(1)** |
| **True(0)** | 1142 | 541 |
| **True(1)** | 189 | 372 |

In sample AUC:-0.8746

Misclassification Rate: 0.332710

* 1. **Random Forest**

Random Forest is a supervised learning algorithm. Like you can already see from it’s name, it creates a forest and makes it somehow random. The “forest“ it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

**Misclassification Rate (Out-of-sample)**

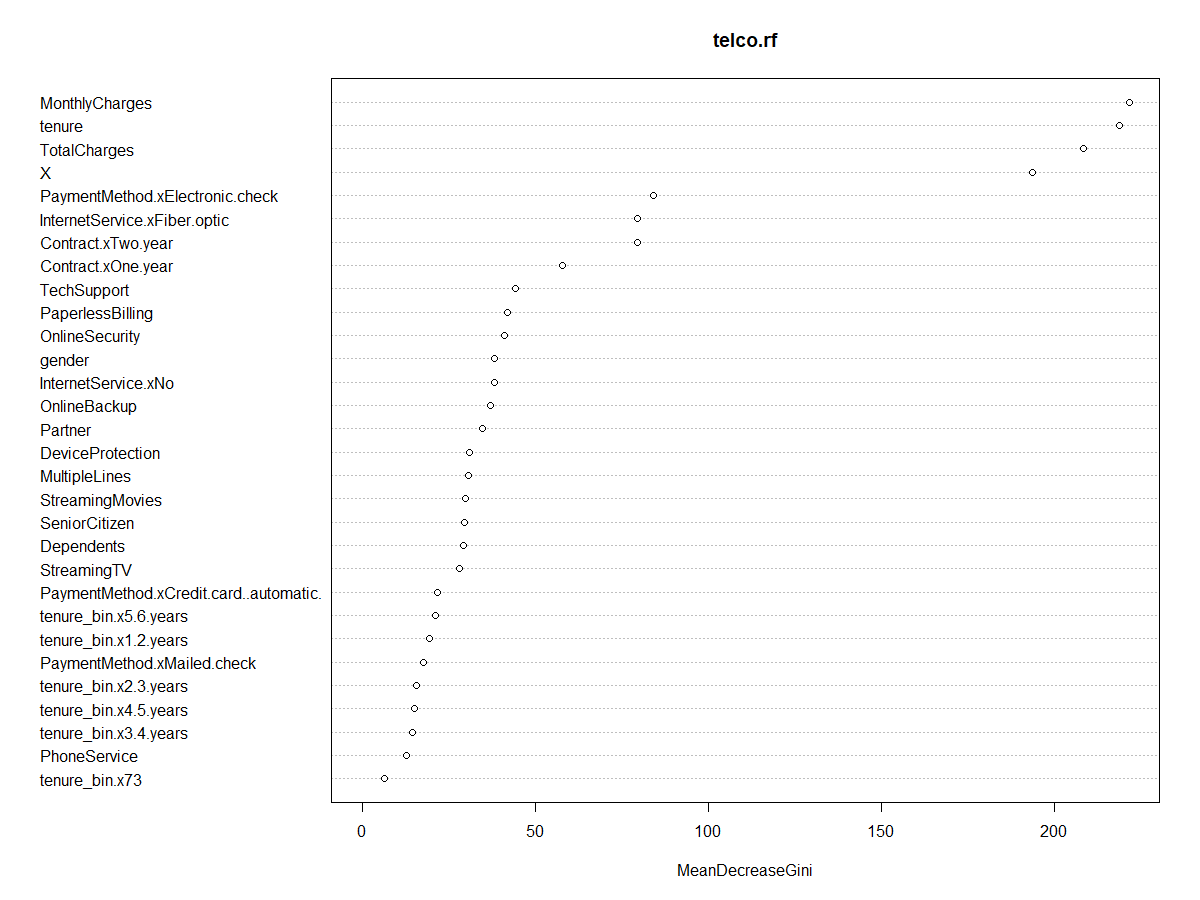
|  |  |  |
| --- | --- | --- |
|  | **Predicted(0)** | **Predicted(1)** |
| **True(0)** | 920 | 530 |
| **True(1)** | 65 | 488 |

In sample AUC:-0.8984

Misclassification Rate: 0.2824

**3.4 Variable Importance**

With many variables, we try to use Random Forest variable importance to see which variable is prominent in business decision making that effects the customer churn.



1. **Recommendation**

* Recommend the customer whose contract are month-to-month to change to one-year or two-year contract by offering the better promotions to changing.
* For customer who are currently owning month-to-month contract and using Fiber internet, the company should take care more
* Sex has no impact on churn.
* Churn ratio for senior citizens is significantly higher. Company must focus on some specific needs of senior citizens.
* People paying bills via Electronic check have higher churn ratio. Company must focus on them and ask them if they are facing any difficulties in paying bills via electronic check or not?

1. **Conclusion**

We have used Logistic Regression, Decision tree and Random Forest to predict the churn rate of the telecommunication data. Comparing all models, Random forest gives the best model with AUC (0.84), we believe this model can also be used to anticipate the churn probability well on future data. The important factors that affect the churn prediction rate are mentioned in RF variable importance section, which can be used for future strategy making.

1. **References**

* <https://www.kaggle.com/blastchar/telco-customer-churn/kernels?sortBy=hotness&group=everyone&pageSize=20&datasetId=13996&language=R>
* https://towardsdatascience.com/machine-learning-powered-churn-analysis-for-modern-day-business-leaders-ad2177e1cb0d
* Lecture Notes