Individual Project Submission

Sideek Headlie

EdX

Data Science: Capstone

Contents

Introduction

This report aims to analyze the ‘Telco Customer Churn’ through the application of data science to derive a greater understanding of the information within the data, as well as generate actionable insights. This dataset is owned and maintained by BlastChar based in Lisbon Portugal.

Each row represents a customer, each column contains customer’s attributes described on the column Metadata. 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, the primary target, included categorical information (i.e. Yes / No). 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

The aim of this project was to predict behavior to retain customers using customer data. All relevant customer data was analyzed and used to develop focused customer retention programs. The level of accuracy was used to determine the performance of the model developed.
N.B. An accuracy of 100% is the ideal performance of any predictive model. Best practices suggest an accuracy above 75% to be acceptable.

The level of accuracy generated by this predictive model was found to be 80.4%.
An accuracy level of 80.4% (i.e. greater that 75%) suggests that the model developed is fit for purpose – the identification of the most impactful customer attributes and behaviours that contribute to churn, as well as the robustness of the model performance on unseen data.

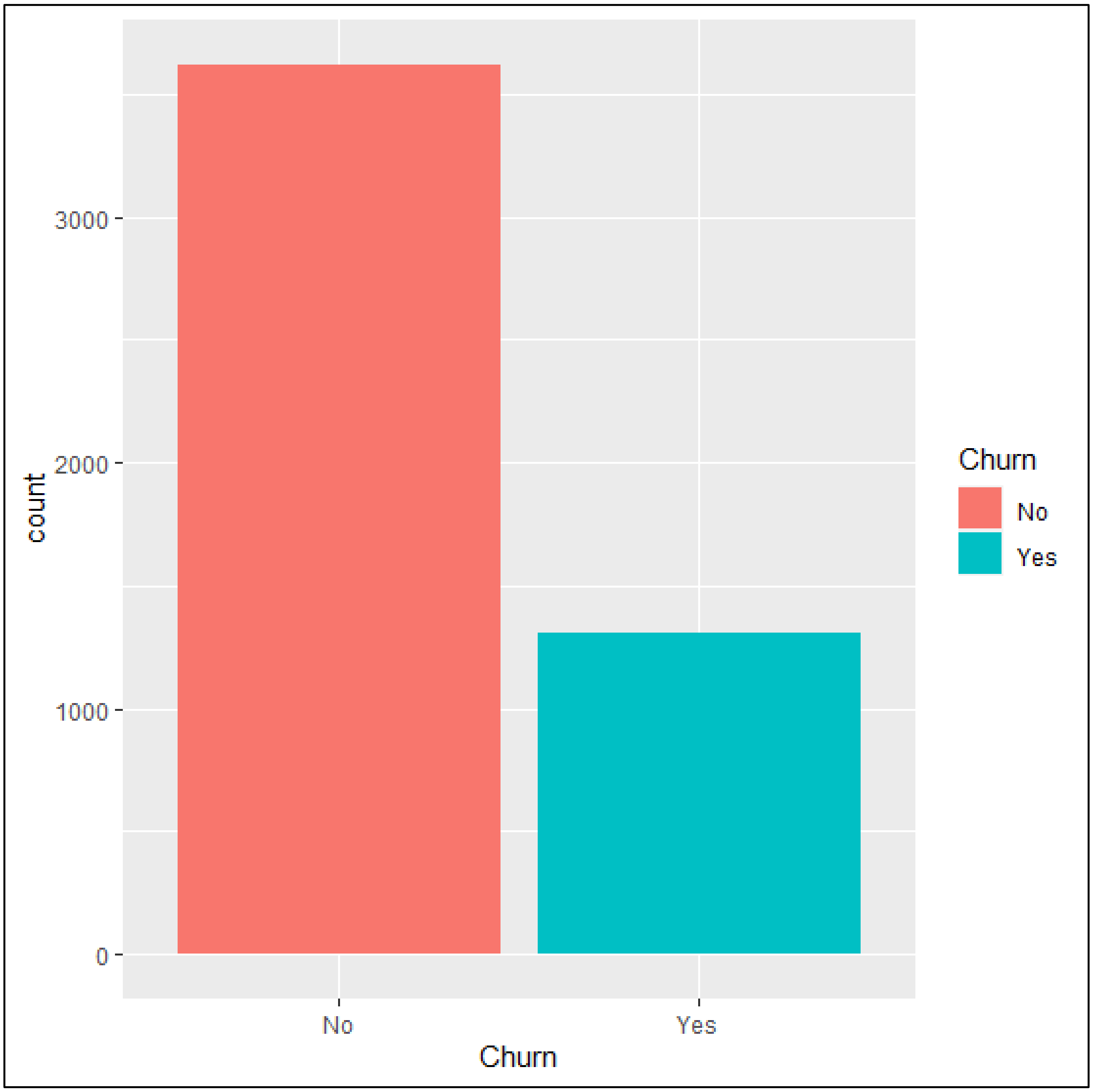
The report reviews the steps undertaken to arrive at the prediction of customer churn based on the various attributes provided in the data, provided as well as provide an overview of the rationale.

Key data science activities included;

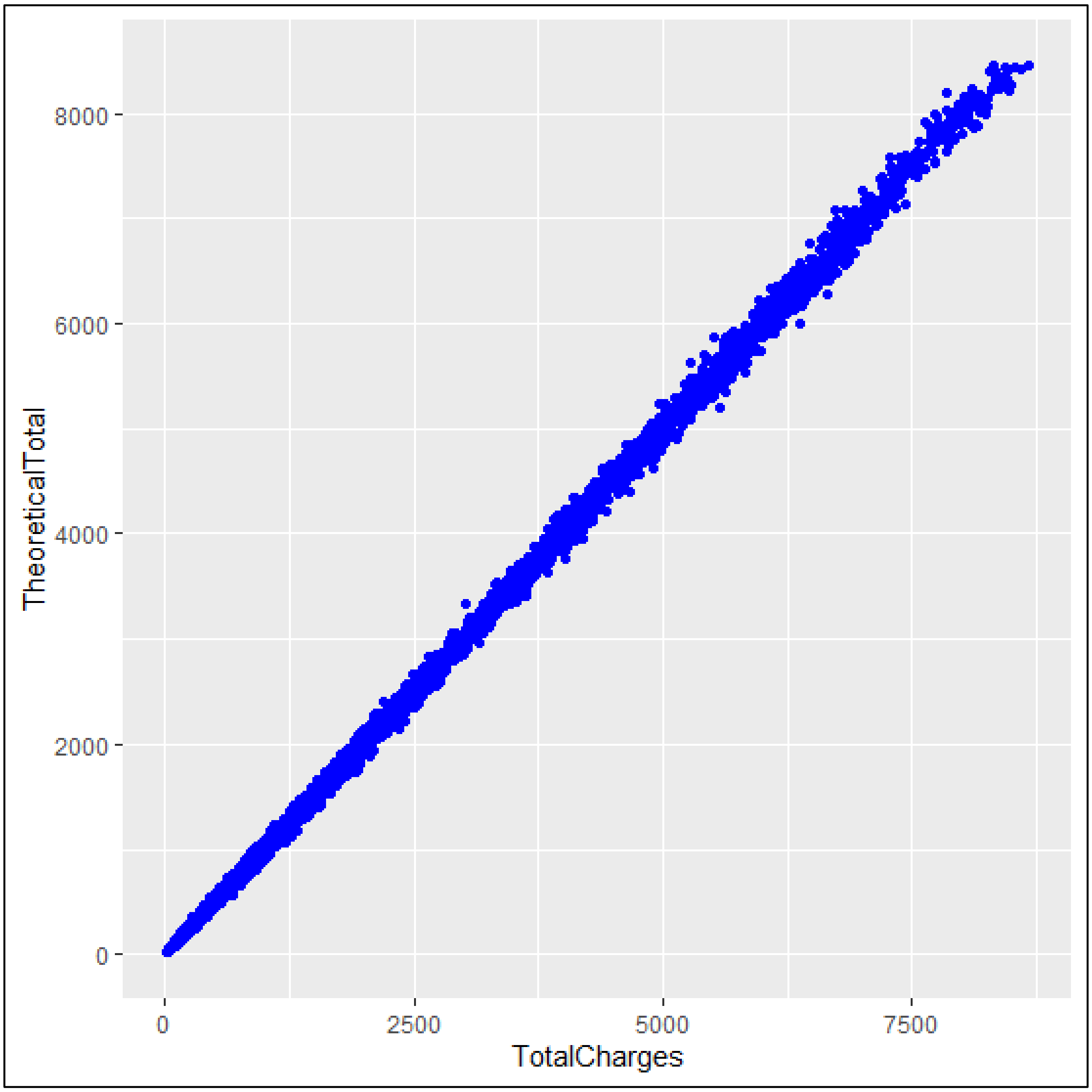
- Data Cleansing
- Data Extraction
- Data Visualization and Inferencing
- Data Prediction

Methods and Analysis

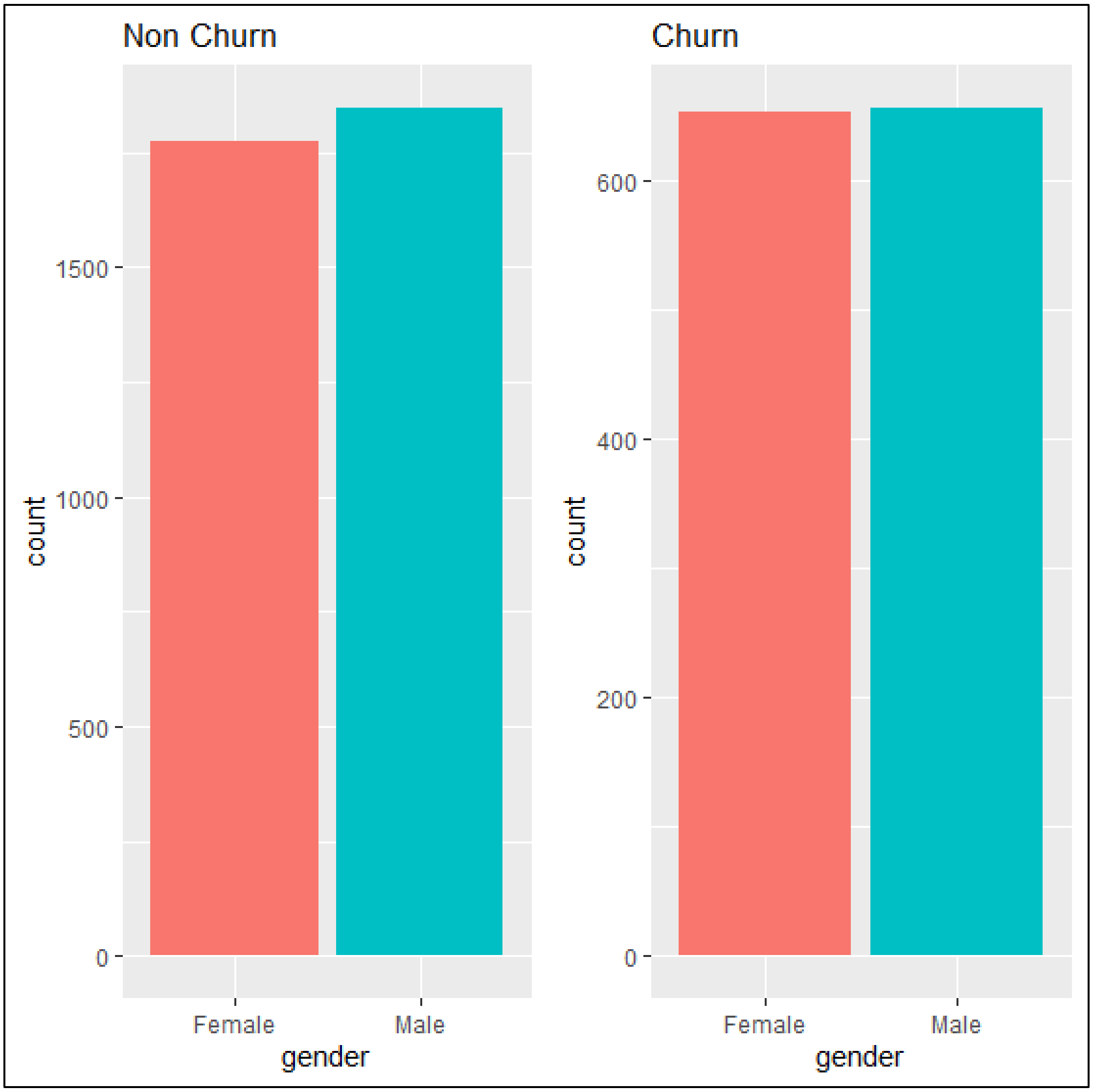
Firstly, preprocess the data and do data exploration. Data preprocessing mainly involved correcting the format of features. Data visualization was done to tease out relevant patterns and trends that may be useful for predicting churn. Data was first split into a training and test set, and model building was done on the training set using repeated 10-fold cross-validation to avoid overfitting on the training set. The models obtained were then used to predict churn behavior for customers in the test set. I then looked at the most important factors that contributed to prediction of churn.



The graph shows the distribution in the churn of telecommunication customers. As shown, a larger portion of customers do not churn (i.e. No). In effort to assess the attributes and behaviours of customers that churn, the ‘Yes – population’ would be further interrogated.

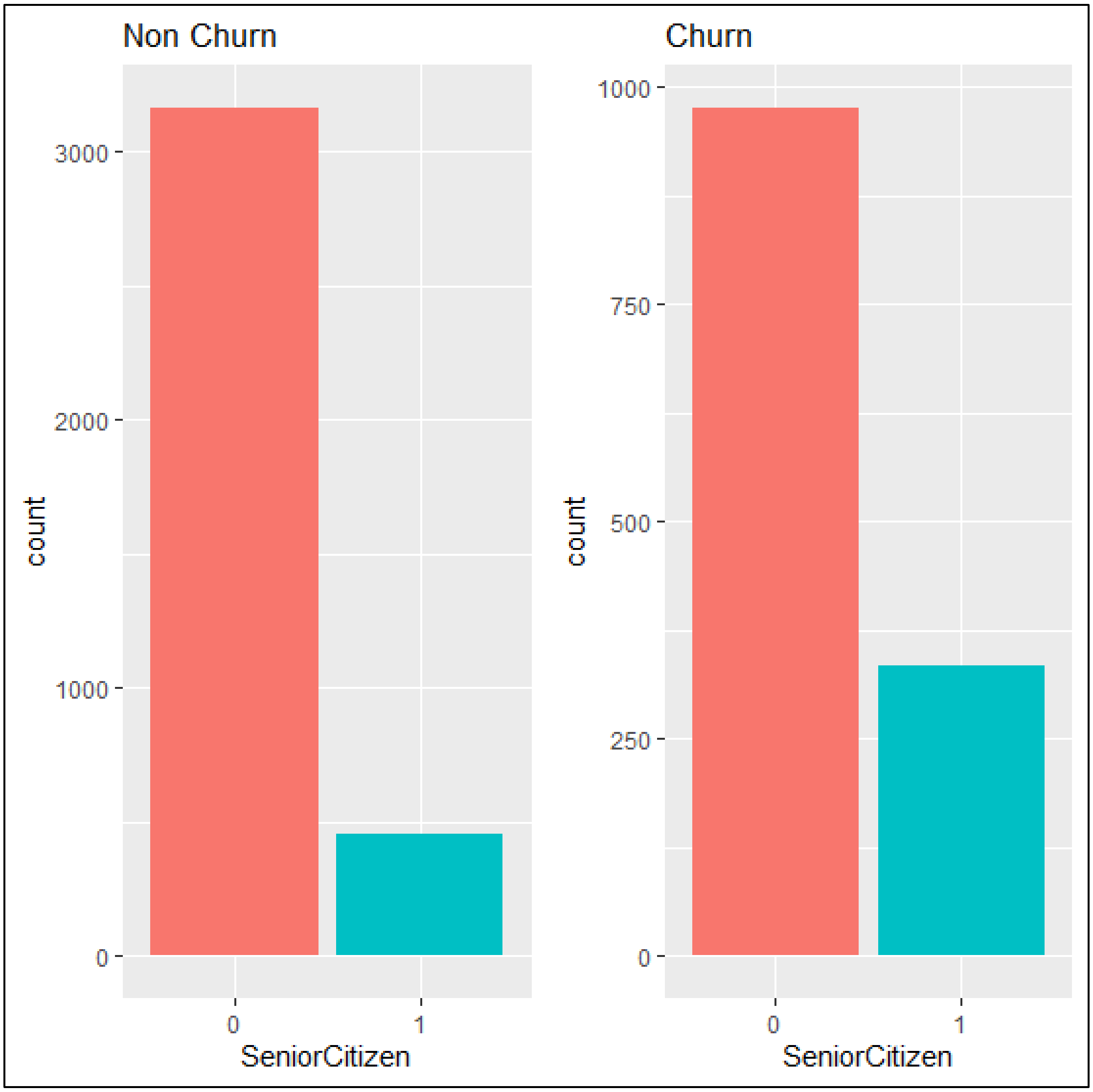


This chart shows that there is very strong correlation between the TotalCharges variable and the Theoretical Total Charge which was derived from the MonthlyCharges variable and the Tenure variable. Hence TotalCharges was a variable that was derived from existing features in the dataset and was therefore redundant. It was hence removed from the training set.



Based on the graph above there is no real variation between Churn and Non churn customers for gender disparity. Furthermore, for Churn customers, the male and female populations are near-equal.

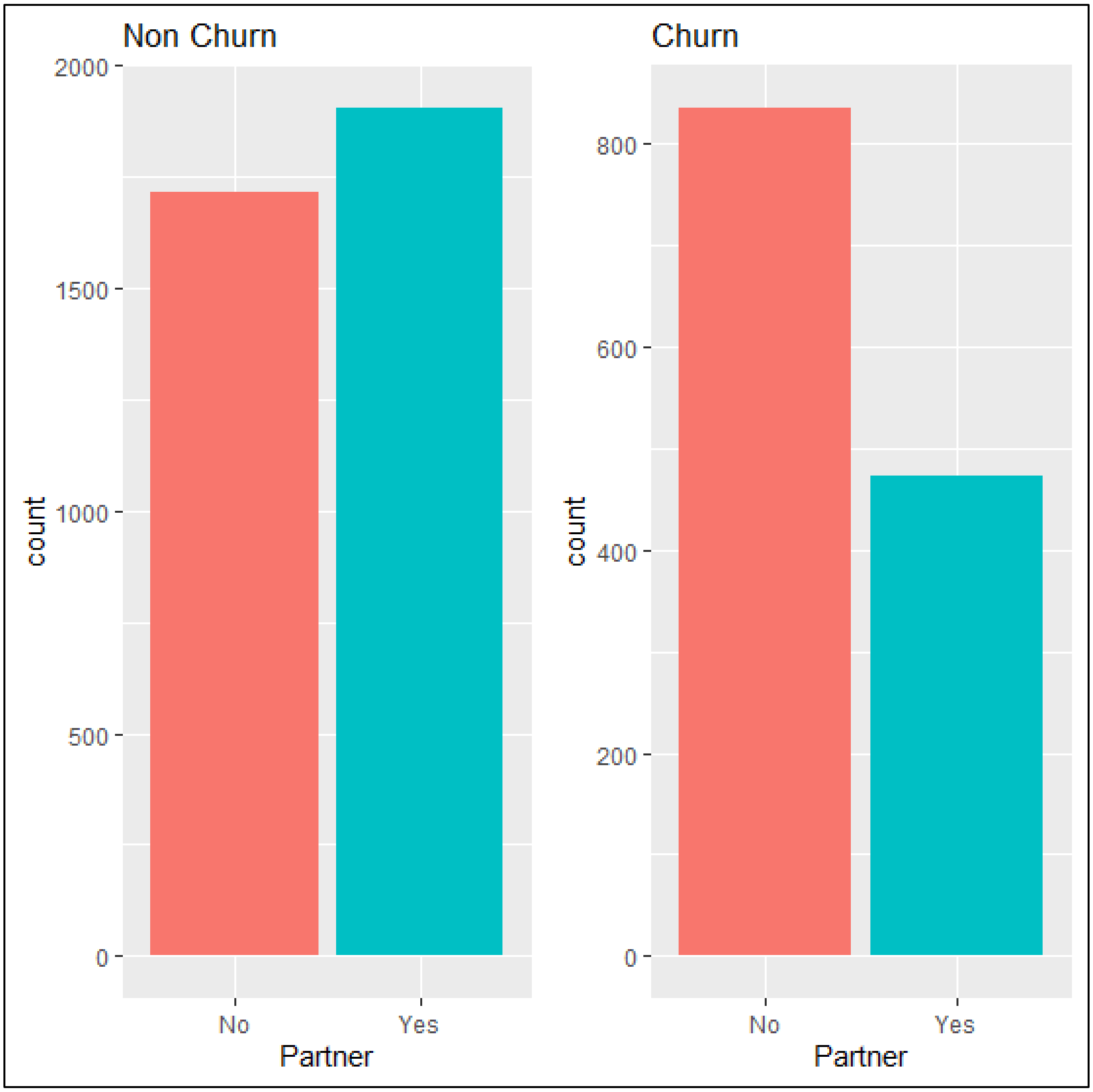
As such, ‘Gender’ is not a desirable customer attribute to be used for Churn prediction.



1 = Senior Citizen
0 = Non-Senior Citizen

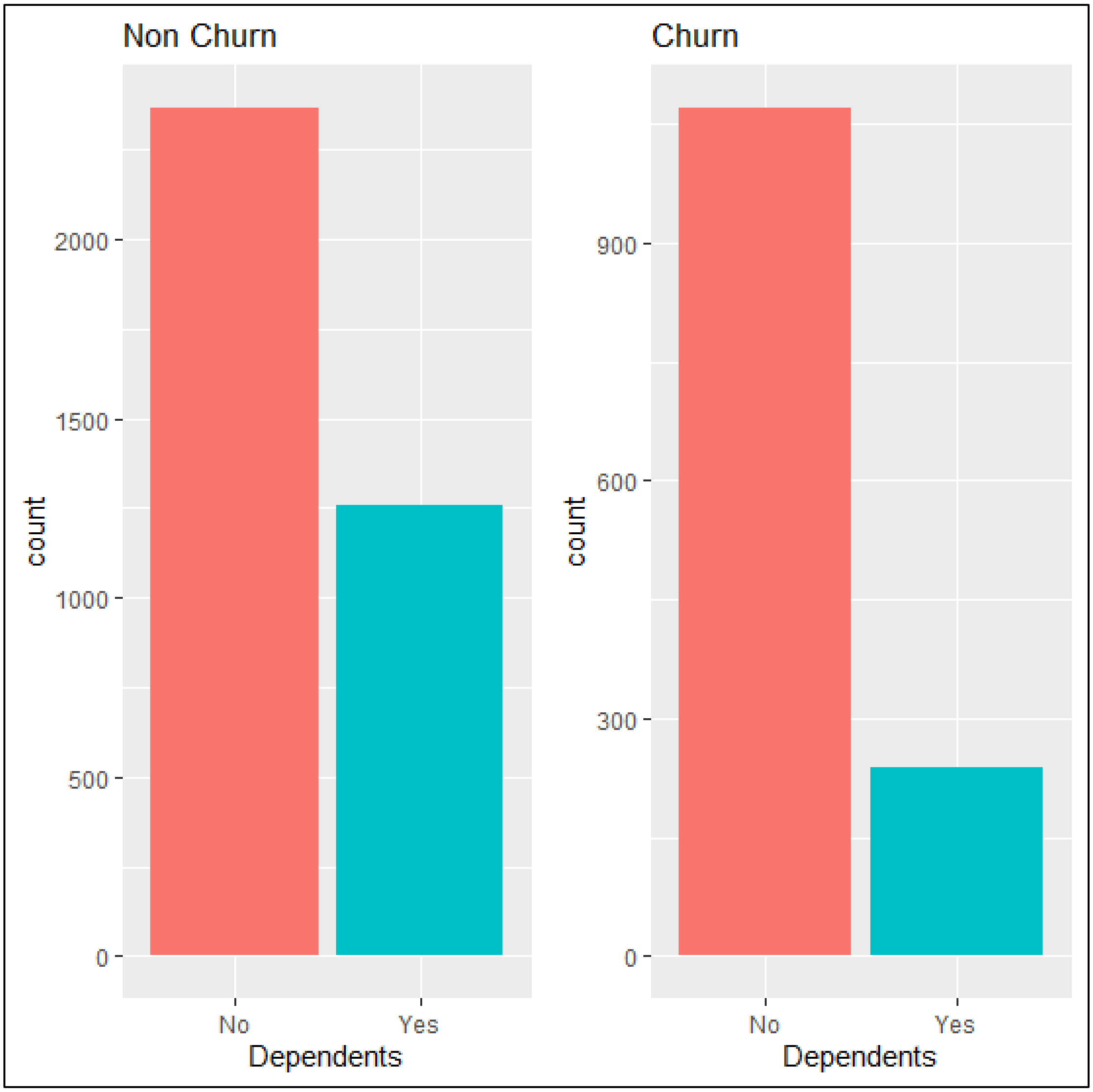
Based on the graph above there is some variation between Churn and Non churn customers for age (Senior Citizen/ Non-Senior Citizen). Furthermore, for Churn customers, more a larger population was categorized as No (i.e. Non-Senior Citizen).

As such, ‘Senior Citizen’ is a desirable customer attribute to be used for Churn prediction.



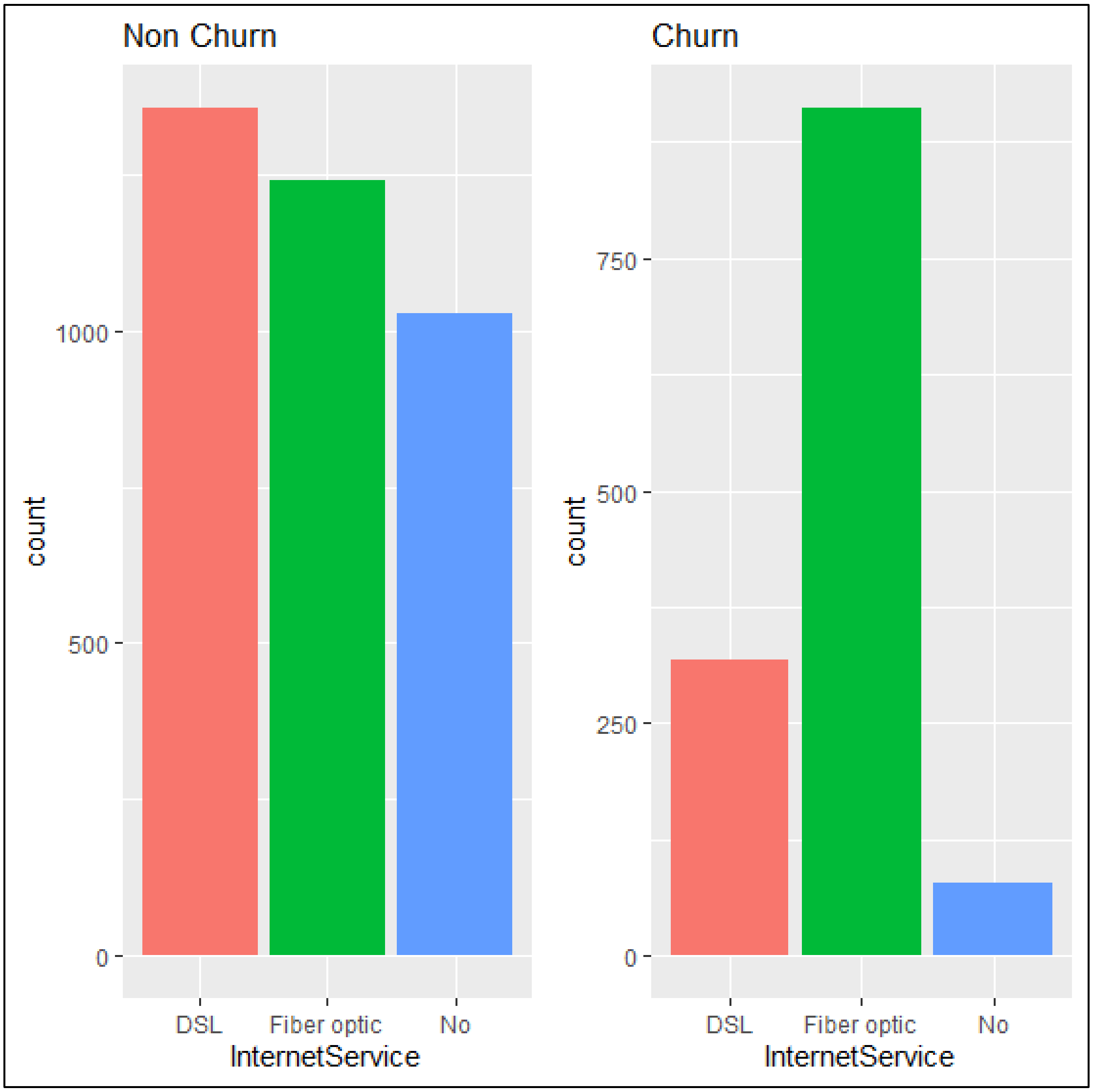
Customers without partners have a high propensity to churn based on the visualization above.

As such, ‘Partners’ is a desirable customer attribute to be used for Churn prediction.



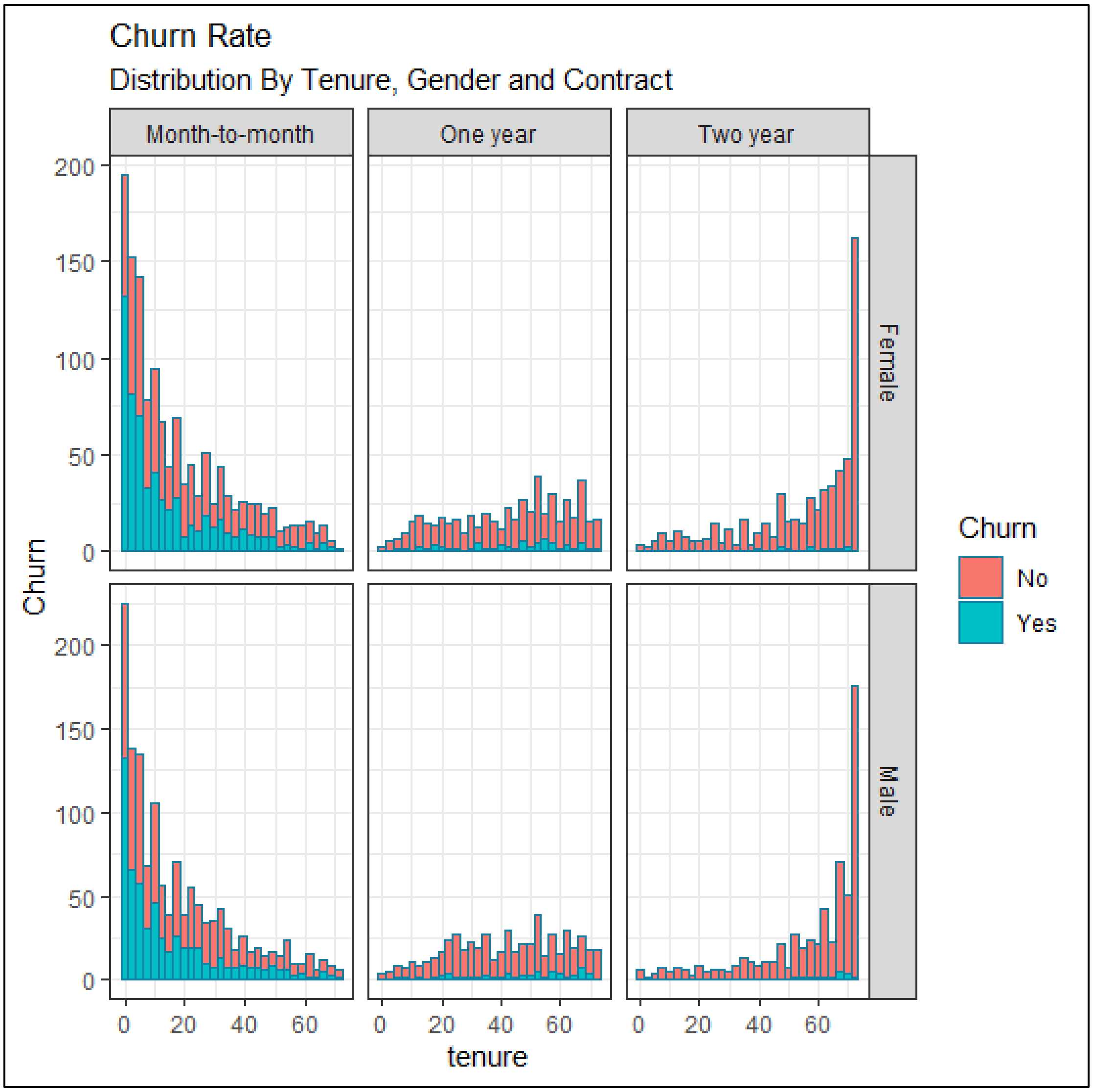
Customers without dependents are more likely to churn, as opposed to customer with dependents, based on the graph above.

As such, ‘Dependents’ is a desirable customer attribute to be used for Churn prediction.



Assessing the customers by Internet Service for churn, shows that customer that churn primarily use Fiber Optics, in contrast to DSL or not having any internet service at all.

As such, ‘Internet Service’ is a desirable customer attribute to be used for Churn prediction.

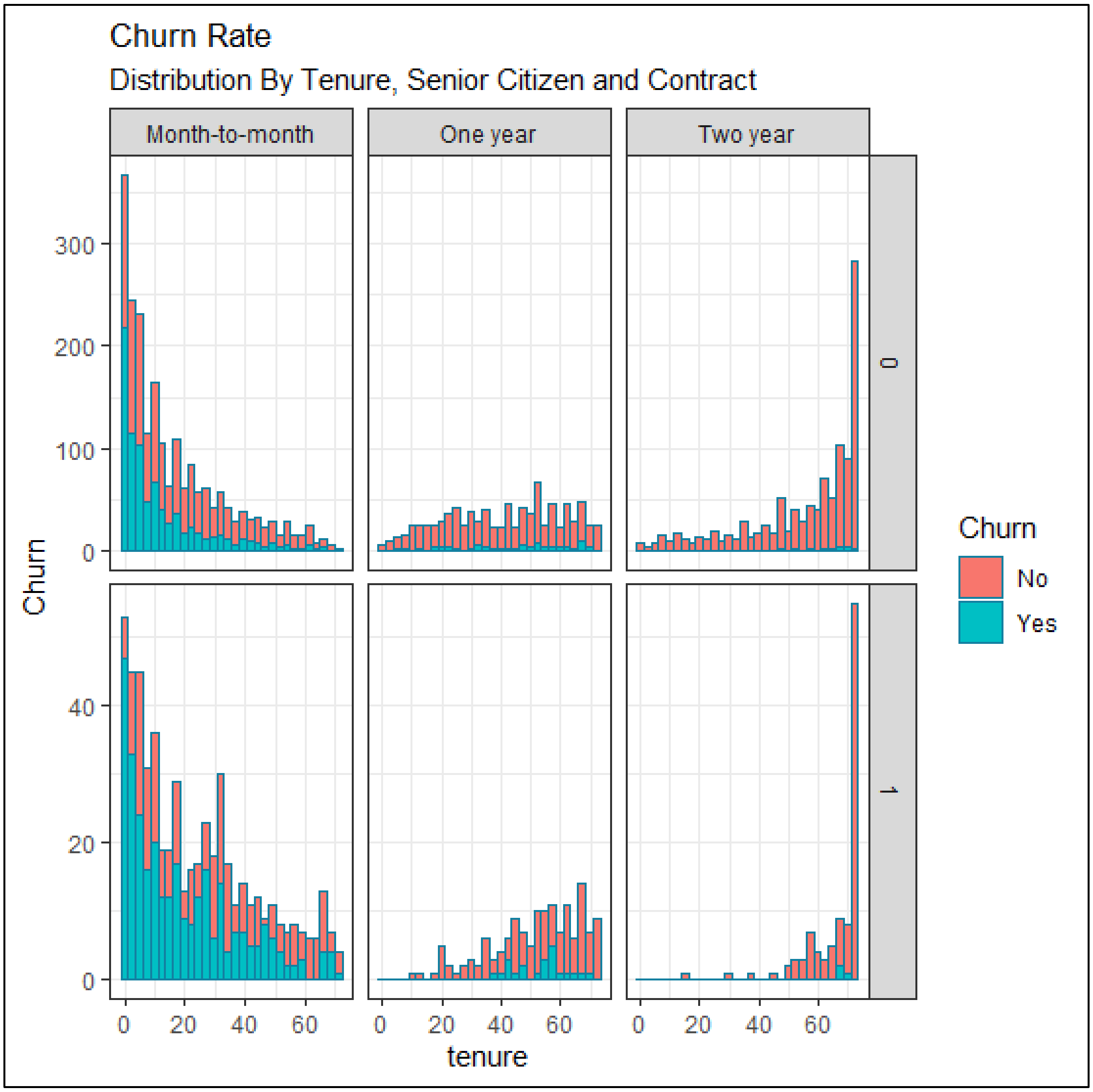


The chart above assessed Churn Rate by Tenure, Gender and Contract.

In blue (i.e. customers that churn) are greatest for month-to-month.

Contract (type) is a desirable customer attribute to be used for Churn prediction.

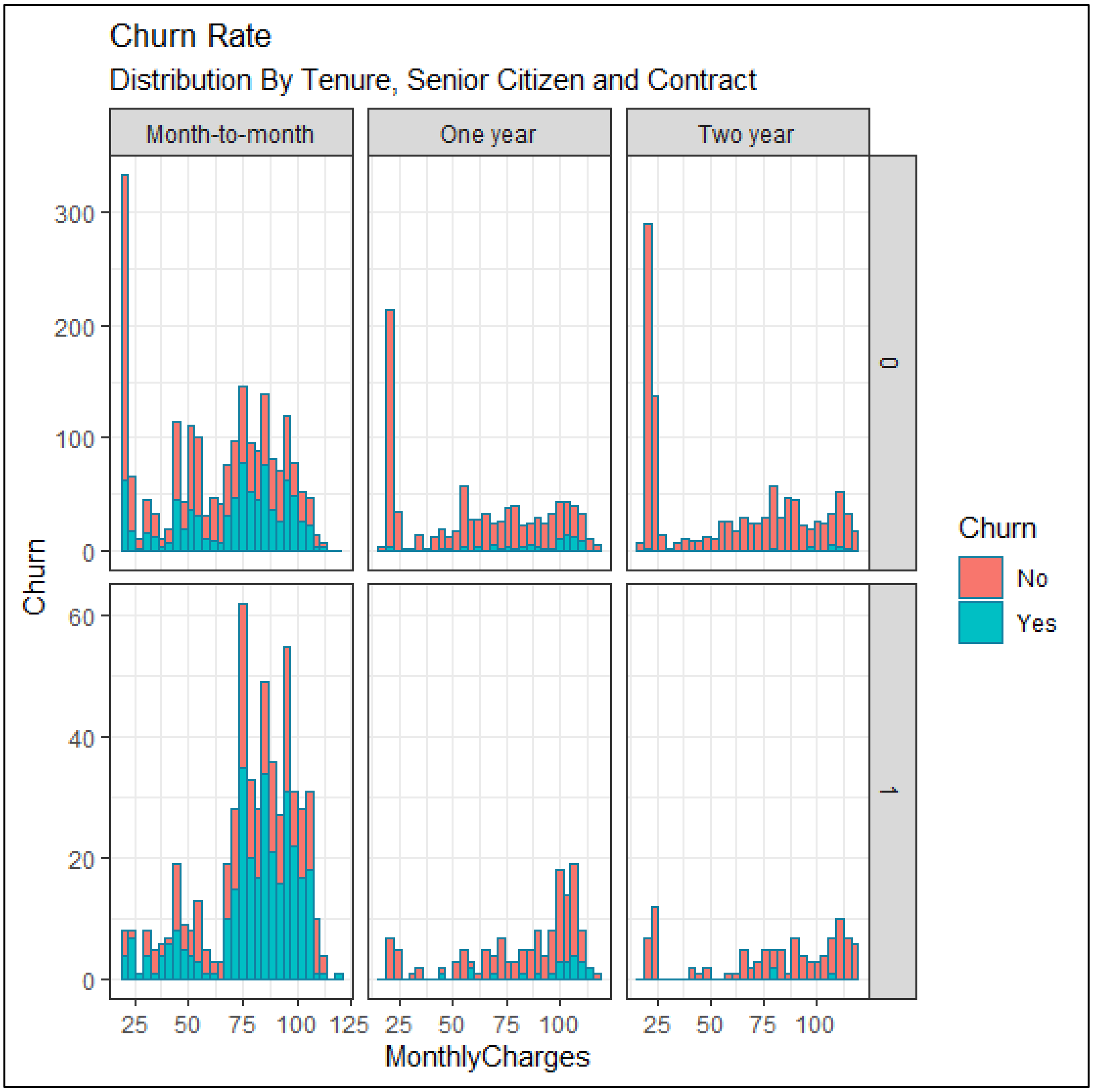
Here the customer profile was aggregated so that that is most likely to churn.



The chart above assessed Churn Rate by Tenure, Senior Citizen and Contract.

In blue (i.e. customers that churn) are greatest for month-to-month.

A higher degree of churn was noted in for Non-senior citizens with contracts that span 1 year, in contrast to Senior Citizens with similar contract types.

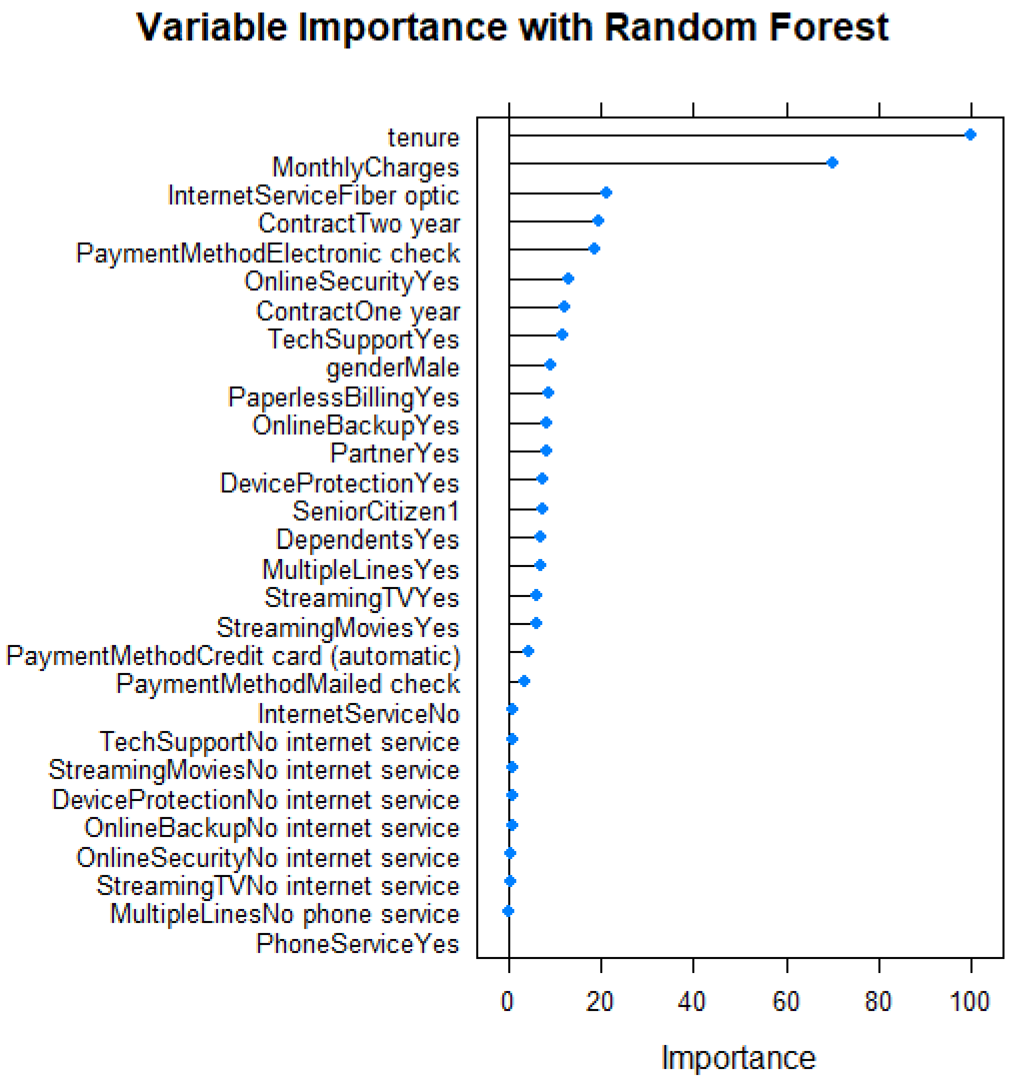


Assessing the same information as previous, with a substitution of ‘Tenure’ for ‘Monthly Charges’.

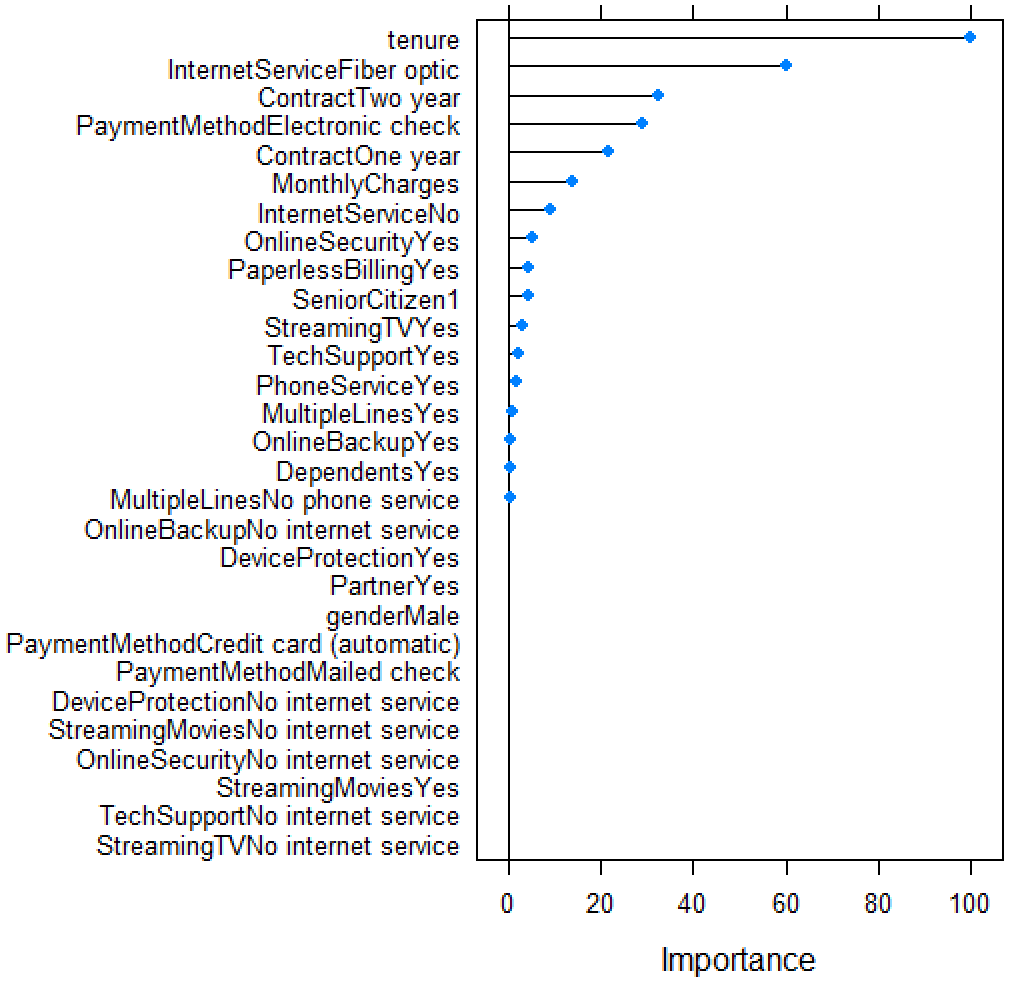
Significant variation of is noted between the Senior Citizens and Non-Senior Citizens for Month-to-month contract between monthly charges between 70 and 110.

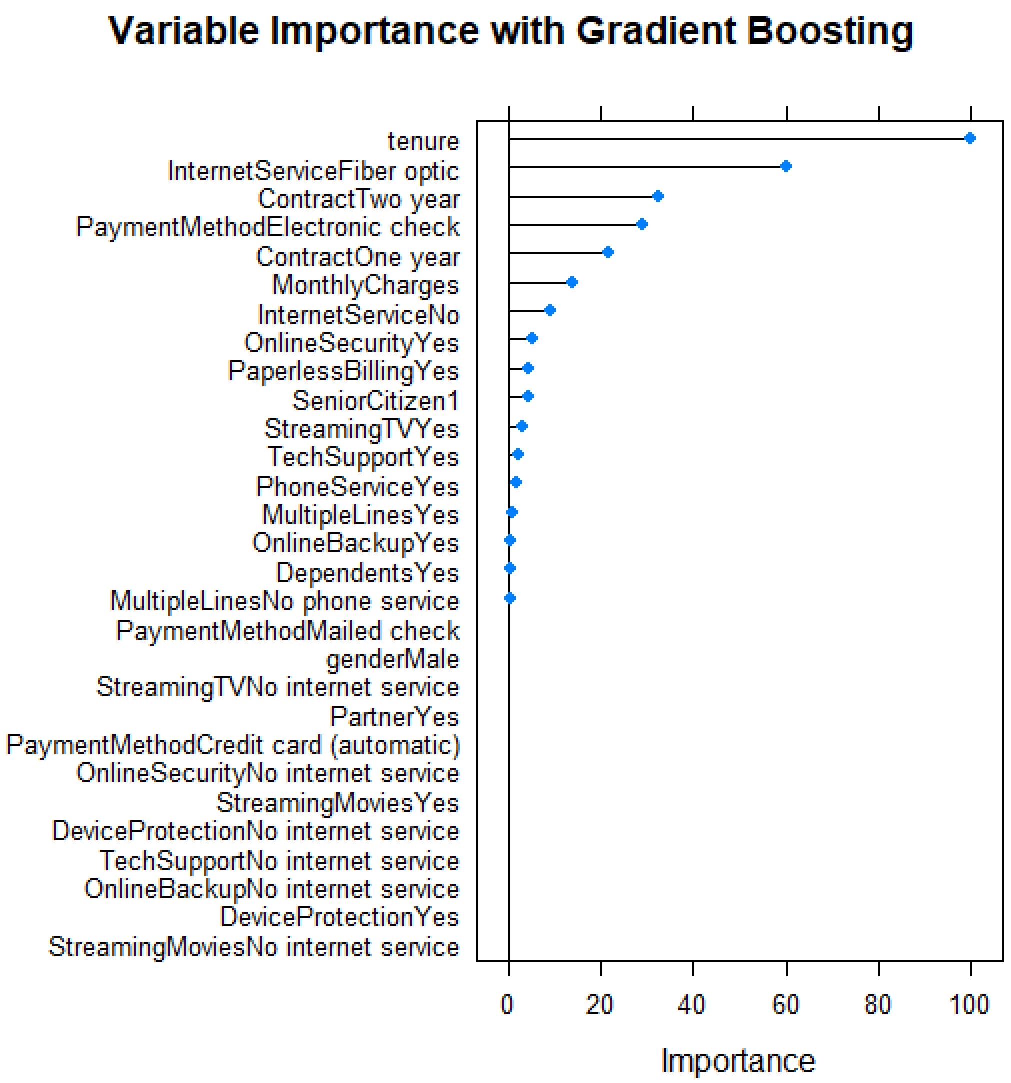
As such, ‘Monthly Charges’ is a desirable customer attribute to be used for Churn prediction.

Model Results



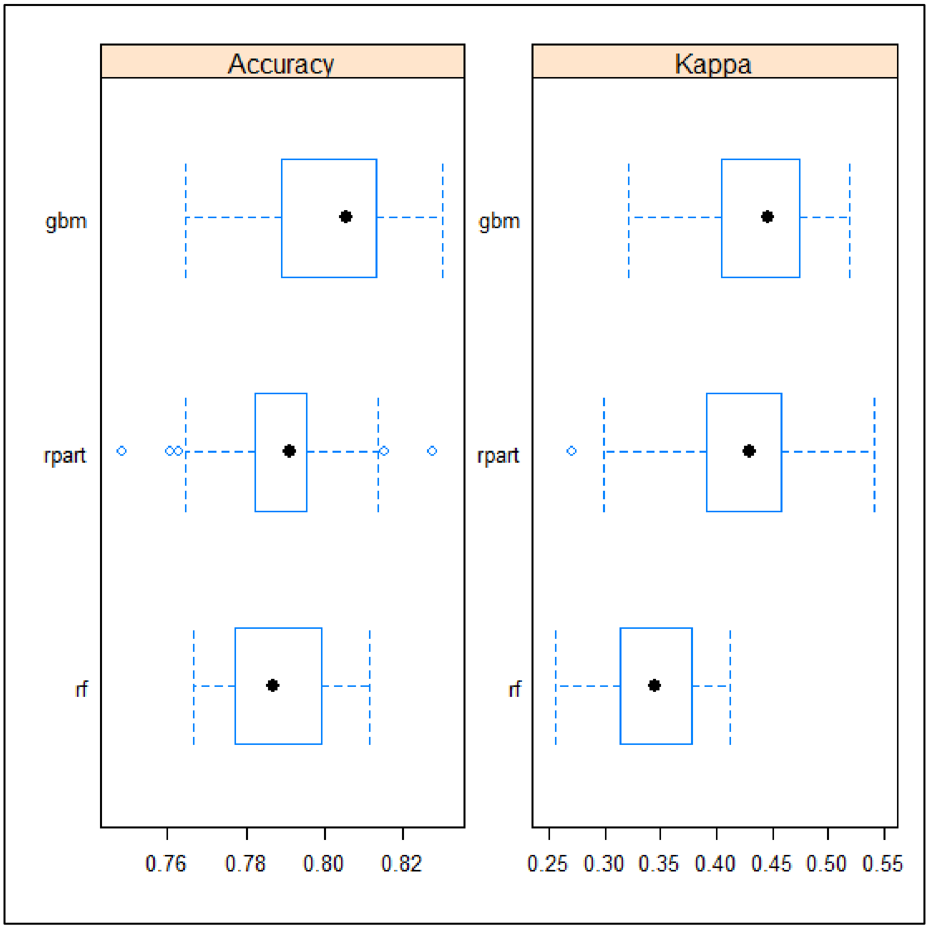
Variable Importance with Gradient Boosting





Variable Importance with Gradient Boosting

The three importance plots above show the most important features that contribute to each models' predictive value.



This figure shows a comparison of accuracy for the three models built. It is observed that the gradient boosting method has the highest accuracy when compared to the other two models. Additionally, the confidence interval for the accuracy is relatively narrow, but the decision tree model (rpart) has a narrower confidence interval.

The training accuracy of the gradient boosting model was found to be 0.8028 while the testing accuracy was found to be 0.8040.

Conclusion

Based on the model, a relatively good accuracy was achieved for predicting churn of telecommunication customers. The most important features that contribute to each models’ predictive value was derived using this model.

A limitation includes a relatively low specificity of approximately 50% which can be improved.

Future applications of this work includes predicting the ideal customer for upselling as well as improved management of the customer during the customer lifecycle (i.e. informed by data).