**Customer Churn Prediction with R**

Customer churn is the rate at which a business loses customers. Customer churn occurs when customers or subscribers stop doing business with a company or service, also known as customer attrition or customer defection. It is also referred as loss of clients or customers. Customer churn is important because it costs more to acquire new customers than to sell to existing customers. This is the metric that determines the success or failure of a business.

**Dataset:**

* The raw data consists of 7043 rows and 21 columns.
* Each row represents a customer and each column contains that customer’s attributes.
* Churn is the dependent (target) variable that represents customers who left within the last month, and the rest other variables are independent variables.
* Demographic information of customers is given in the variables gender, age, partners and dependents variables.
* Customer account information is given in the variables contract, payment method, Monthly charges, total charges and tenure.
* Services opted by customers is given in the variables Phone service, Multiple lines, Online backup, Tech support, Streaming Movies/TV.

**Objective:**

The main objective is to predict the customer churn and analyse the relevant customer data in R.

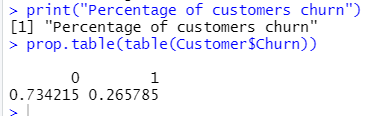
**Data importing and preprocessing:**

* All the required libraries and the dataset is imported in R.
* The .str() function gives the datatype of each variable.

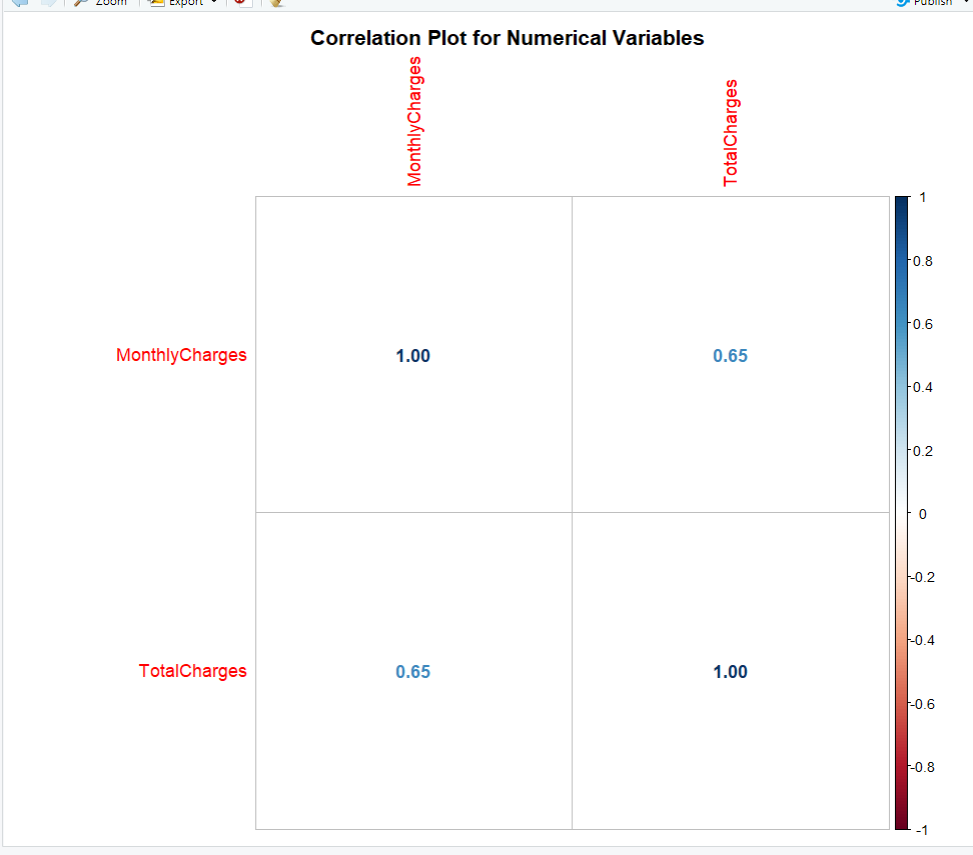


* The missing values in each column are checked with the help of sapply. 11 missing values are found in the column “TotalCharges”. The rows with the missing values are removed.
* “No internet service” is replaced with “No” for six columns, which are “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies”. “No phone service” is replaced with “No” for the column “MultipleLines”.
* The minimum tenure is 1 month and maximum tenure is 72 months. Hence, we can group them into five tenure groups: 0–12 Month, 12–24 Month, 24–48 Months, 48–60 Month, > 60 Month.
* The columns that are not needed for analysis are removed.

**Exploratory Data Analysis:**

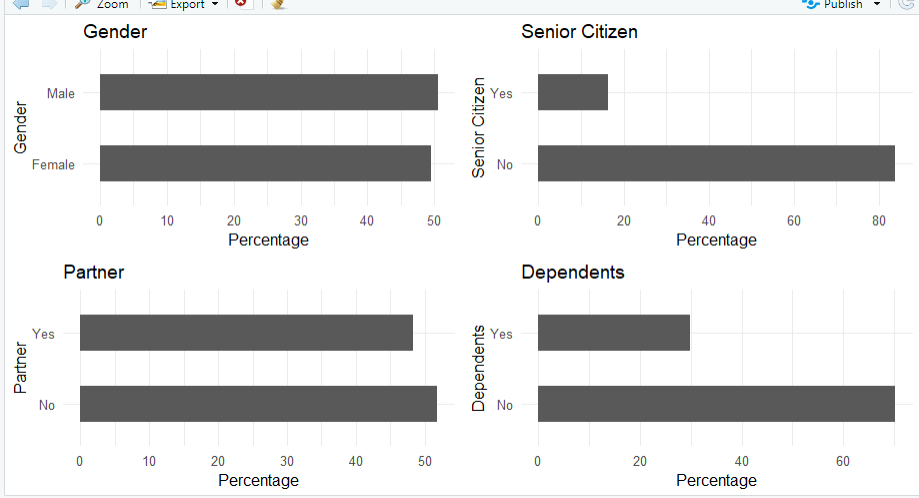


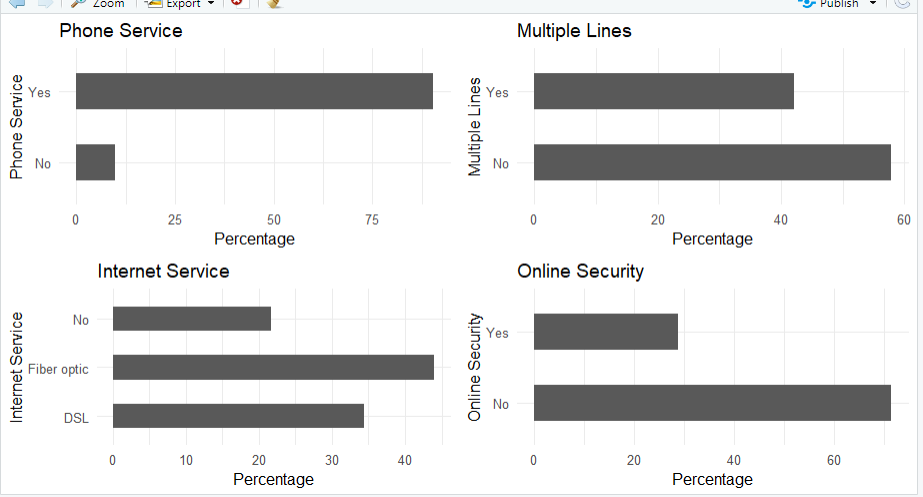
* From the table above, we notice that 73.4% of the customers did not churn. This can server as our baseline model i.e. if we predict every customer to not churn, we will be right on average 73.4% of the times.
* Correlation between numeric variables:

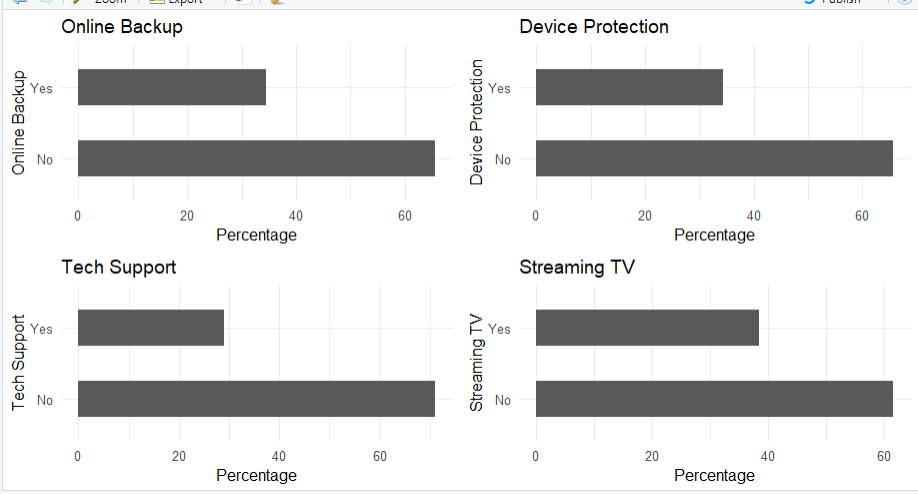


The Monthly Charges and Total Charges are correlated. So, one of them will be removed from the model. We remove Total Charges.

* Bar plots of categorical variables:

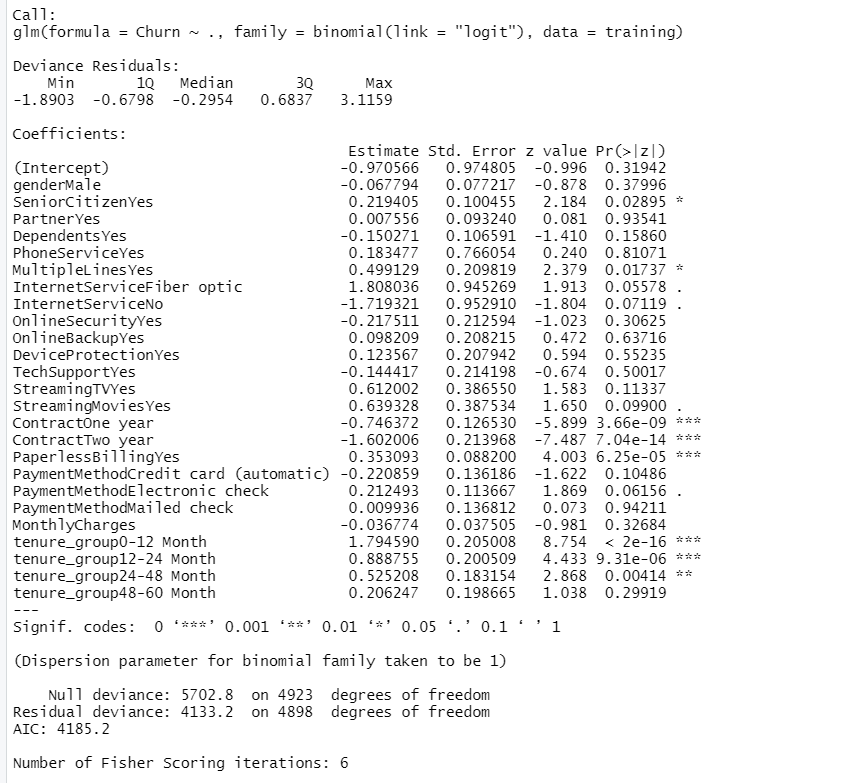
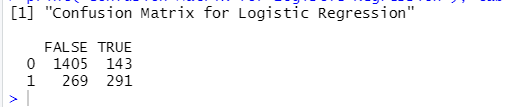






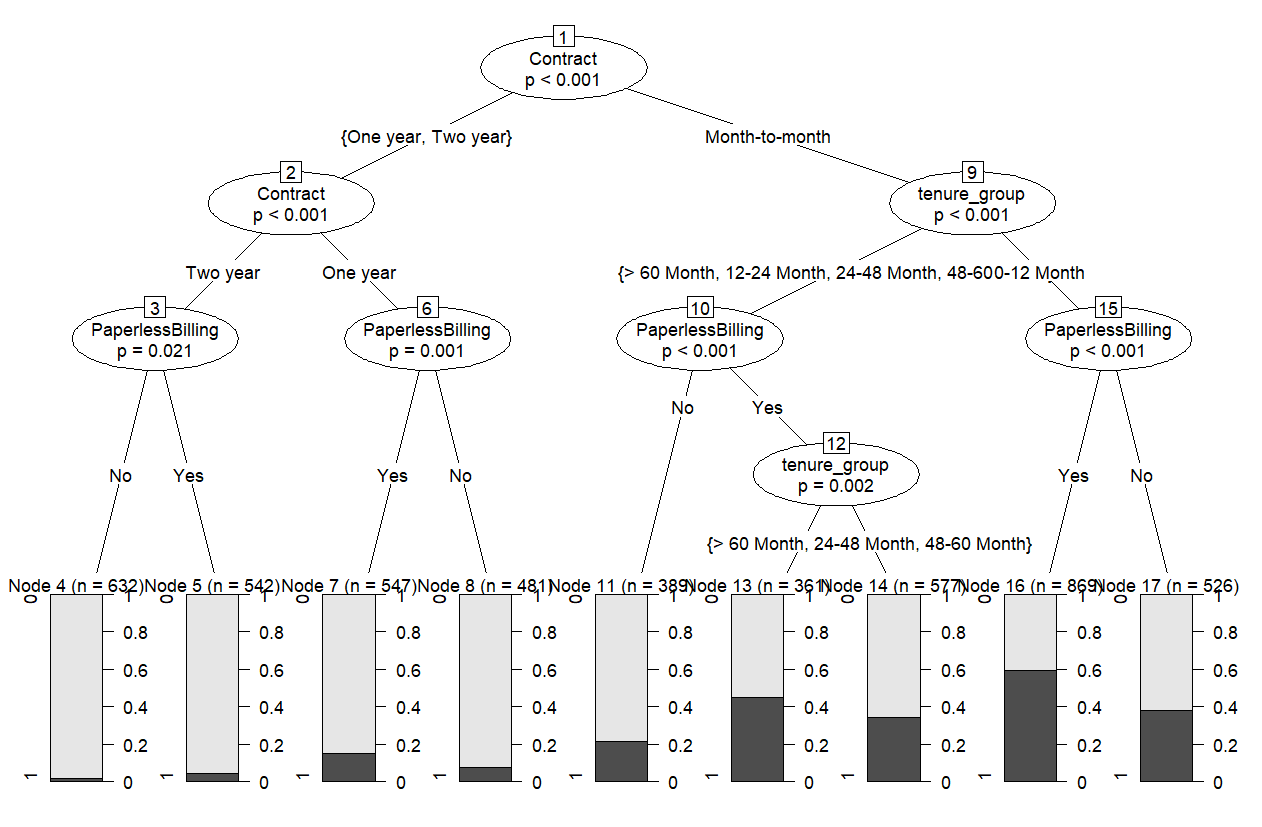


**Logistic Regression:**

* First, we split the data into training and testing sets.
* Fitting the Logistic Regression Model:
* Logistic Regression Accuracy: 
* Logistic Regression Confusion Matrix: 

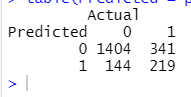
**Decision Tree:**

* Decision Tree visualization:

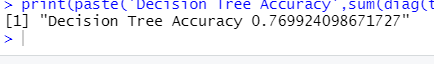
For illustration purpose, we will be using only three variables for plotting the decision tree - “Contract”, “tenure\_group” and “PaperlessBilling”. 

* Out of three variables used, Contract is the most important variable to predict customer churn or not churn.
* If a customer in a one-year or two-year contract, no matter he/she has PapelessBilling or not, he/she is less likely to churn.
* On the other hand, if a customer is in a month-to-month contract, and in the tenure group of 0–12 month, and using PaperlessBilling, then this customer is more likely to churn.
* Decision Tree Confusion Matrix:

All the variables are used to produce the confusion matrix and to make predictions.

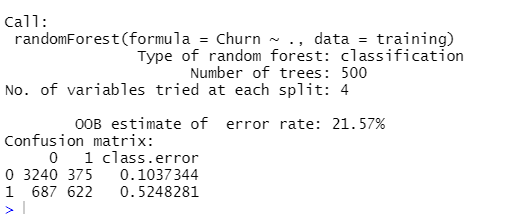


* Decision Tree Accuracy:



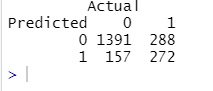
**Random Forest:**

* Random Forest Initial Model:



The error rate is relatively low when predicting “No” (0), and the error rate is much higher when predicting “Yes” (1).

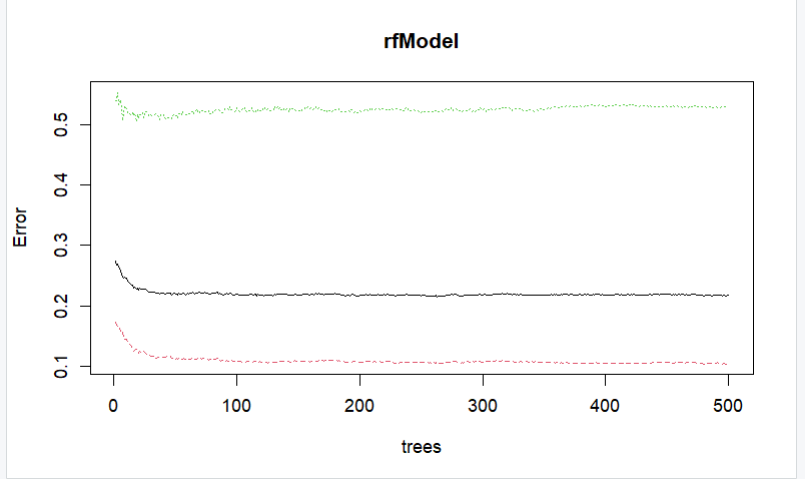
* Random Forest Confusion Matrix:

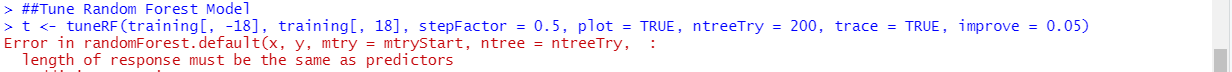
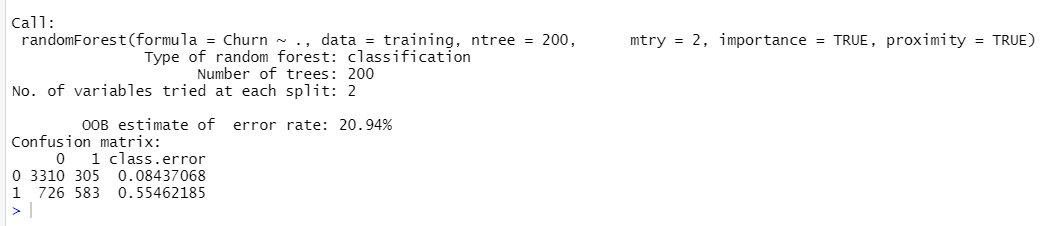


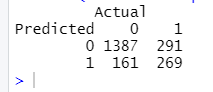
* Random forest Accuracy:

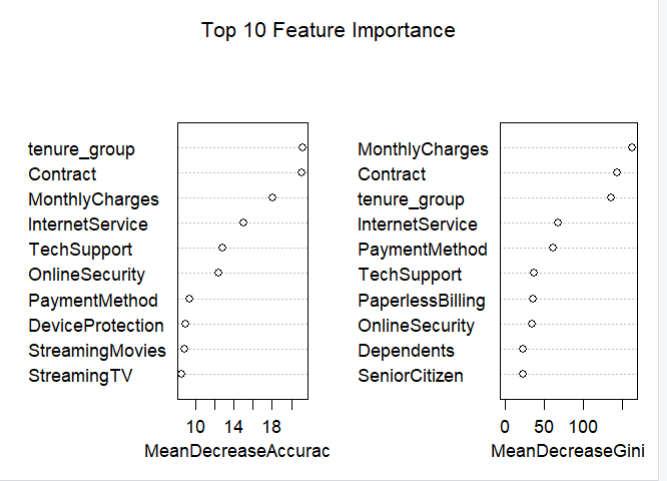


* Random Forest Error Rate:



* Tune Random Forest Model: 
* Fitting the Random Forest Model After Tuning: 
* Random Forest Confusion Matrix After Tuning:



* Random Forest Feature Importance: 

**Conclusion:**

* For this particular dataset, we can see that Logistic Regression, Decision Tree and Random Forest can be used for customer churn analysis.
* From the analysis done, we can conclude that
* The features such as tenure\_group, Contract, PaperlessBilling, MonthlyCharges and InternetService appear to play a role in customer churn.
* There does not seem to be a relationship between gender and churn.
* Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are more likely to churn.

On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.

* Also, Telco should be careful on collecting and using the customer data. Data security is an important concern. Customers won’t want their personal information, like partner, dependents, to be analysed.