DATA MINING PROJECT REPORT

BAN 620

INSTRUCTOR: BALARAMAN RAJAN

DATASET: TELCO

SUBMITTED BY

SURABHI PANDEY

RAJASHREE PRUSTY

UPMA PATIL

VANDANA HINGU

NIDHI PATEL

Project Objective

To focus on methods/algorithms which will help us in analyzing the selected dataset and deriving some logical conclusions and predictions for future use.

Data Source

Telco Customer Churn: https://www.kaggle.com/blastchar/telco-customer-churn.

- **Churn Analysis:** Churn Analysis is one of the worldwide used analysis on Subscription Oriented Industries through which they study the customer behaviors to predict the customers who are about to leave the service agreement from a company. As it is based on Data Mining methods and algorithms, companies in today's commercial conditions are concluding that gaining a new customer's cost is more than retaining the existing ones.

- **Telco Customer Churn Purpose:**
  - This analysis focuses on the behavior of telecom customers who are more likely to leave the platform. We intend to find out the most striking behavior of customers through Data Mining Algorithms using R language and eventually use some of the predictive analytics techniques to determine the customers who are most likely to churn.

  The objectives of this project are:

  - What factors contribute towards customer loyalty for Telco database?
  - Determining different data mining methods for churn analysis.
  - Shedding a light on methods that are used for forecasting reasons on why customers are churning out of the company.
  - Using the analysis to predict behavior of future customers.

We aim at predicting customers who are going to stop using a product or service with the Telco. And, the customer churn analysis will extract these possibilities. Today's cutthroat competitional market led to numerous companies selling the same product at a similar service and product quality. In the midst of all these, the cost of gaining new customers is more than retaining the existing customers. For this reason, existing customers are valuable for a company.

Our objective is to precisely predict the probable number of customers who are going to stop using services or products. This analysis is performed according to customer segments created and finding out the factors contributing.

Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be formulated basing on churn rate or customer attrition. In this way, profitability can be increased significantly for Telco or the possible damage due to customer loss can be reduced at a similar rate.

Motivation for Analysis

Analyzing customer- based datasets for businesses have innumerous advantages such as:

- ✓ Knowing Most and Least Profitable Customers
- ✓ How to Improve Customer Service
- ✓ Having a targeted Marketing approach
- ✓ Packaging products differently
- ✓ Building Loyal Relationships

Why Chose Telco Database-?

- ✓ Telecommunication advancements have greatly impacted the way people interact with one another at the global level.
- ✓ This is now an important tool not just for entertainment but also to communicate, to explore, to learn thereby helping many to improve productivity and save time.
- ✓ This Telco Customer Database has information related to most of the people using TV streaming, Movie Streaming, Internet Services, etc on daily basis.
- ✓ We ourselves being using similar internet/TV services, got interested to analyze this customer-business quotient and what factors contribute to the behavior of "loyal/not so loyal" customers for Telco industry.

Scope of the Project:

- Analysis of Customer Behavior: What factors contribute towards Customer Loyalty for Telco Database?
- Predictive Behavior for Future Customers: Using the Analysis predict behavior of future customers.

Statistical Tools used for Analysis:

- **Logistic Regression:** Our output variable being Churn which is a categorical Variable, Logistic regression was one of the most obvious choices to execute Churn Analysis. Benefits of using logistic regression for Telco dataset is it gives propensity (rankings), which helps in identifying the customers who have higher probability of leaving the platform. Hence, we can target those customers which needs to be better served by offering them with certain better promotional rates which prevents them from leaving the platform.

- **Classification Trees: -** Choosing Trees for an advantage of easy interpretability. We can also formulate Association Rules with trees, indicating the impact of various predictors on Churn Outcome Variable.
- **Random Forests: -** RF are known for improving prediction power and thus this quality of RF Will help us in predicting the customers which are likely to churn out of platform, thereby helping the Telco company to take some risk mitigation steps in advance.

STEPS INVOLVED in Analysis

- Obtaining the data & Understanding Purpose
- Data Pre-processing and Exploration
- Building and Comparing Models
- Deploying the best model
- Interpretations, Predictions and Recommendations

Step1 (a): Data Overview

- Dimensions: 7043 by 21.
- This has 17 categorical variables and 3 numerical variables.
- Column Churn: (Outcome Variable) Customers who left within the last month
- 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

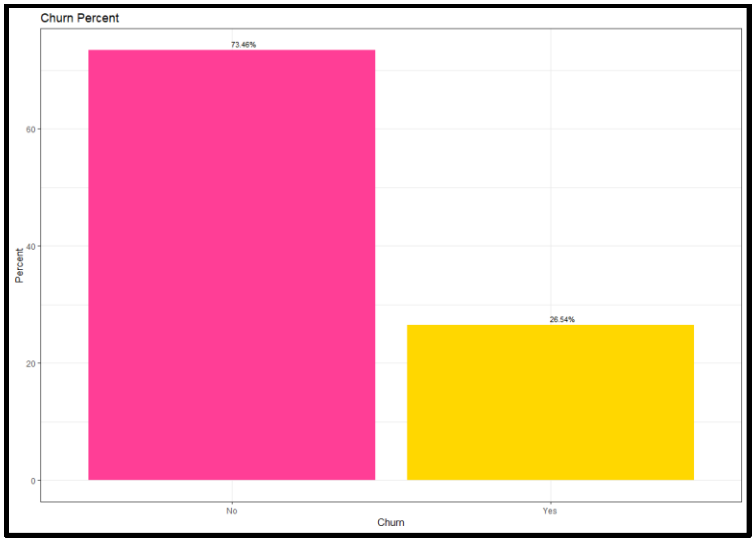
Step 1 (b): Variable Summary:

Sr. No	Variable Name	Variable Description
1	Customer ID	Customer ID
2	Gender	Whether the customer is Male/Female
3	Senior Citizen	Whether the customer is a senior citizen or not (1, 0)

4	Partner	Whether the customer has a partner or not (Yes, No)
5	Dependents	Whether the customer has dependents or not (Yes, No)
6	Tenure	Number of months the customer has stayed with the company
7	Phone Service	Whether the customer has a phone service or not (Yes, No)
8	Multiple Lines	Whether the customer has multiple lines or not (Yes, No, No phone service)
9	Internet Service	Customer's internet service provider (DSL, Fiber optic, No)
10	Online Security	Whether the customer has online security or not (Yes, No, No internet service)
11	Online Backup	Whether the customer has online backup or not (Yes, No, No internet service)
12	Device Protection	Whether the customer has device protection or not (Yes, No, No internet service)
13	Tech Support	Whether the customer has tech support or not (Yes, No, No internet service)
14	Streaming TV	Whether the customer has streaming TV or not (Yes, No, No internet service)
15	Streaming Movies	Whether the customer has streaming movies or not (Yes, No, No internet service)
16	Contract	The contract term of the customer (Month-to-month, One year, Two year)
17	Paperless Billing	Whether the customer has paperless billing or not (Yes, No)
18	Payment Method	The customer's payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic)
19	Monthly Charges	The amount charged to the customer monthly
20	Total Charges	The total amount charged to the customer
21	Churn	Whether the customer churned or not (Yes or No)

Step 2: Data Pre-Processing and Exploration:

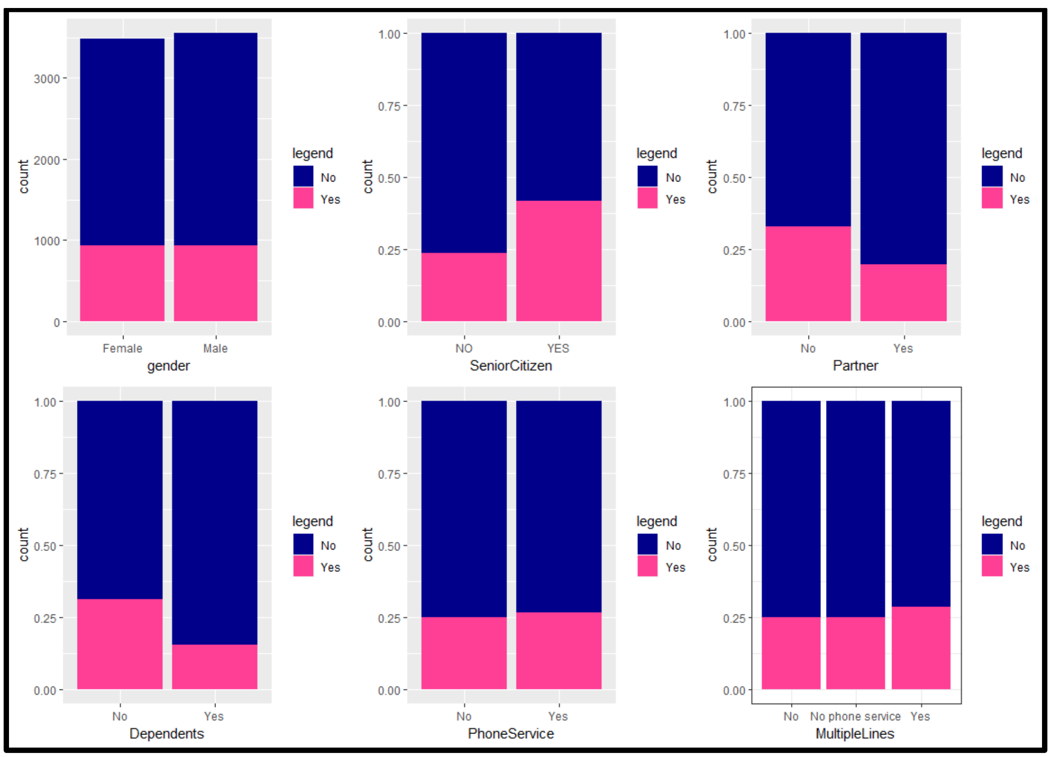
- **Understanding Outcome Variable Churn:**



Observations: 26.54% of the Customers have churned out of the Telco platform in a month's time. This is a significant loss to this company and needs to be dealt with. Our analysis will help this company to identify certain factors which can help this Telco company to increase customer loyalty.

- **Data Exploration: Categorical Variable v/s Outcome Variable Churn:**

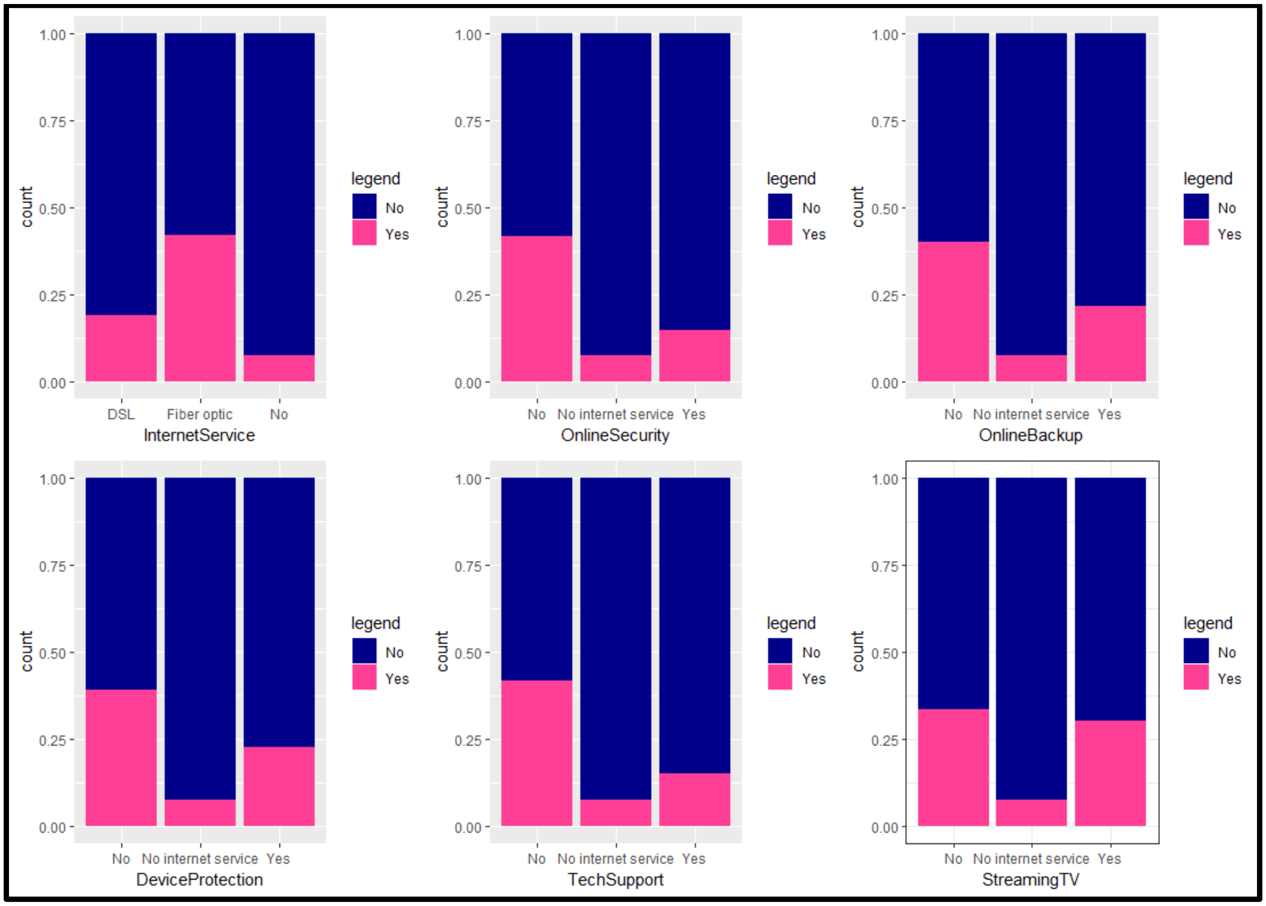
- **Set 1: Gender, Senior Citizen, Partner, Dependents, Phone Service, Multiple Lines:**



Observations:

Sr. No.	Variable v/s Churn	Observations
1	Gender	The Churn Percent is almost Equal for both the genders
2	Senior Citizen	Churn Percentage is higher for Senior Citizens
3	Partner	Churn Percentage is less for customers with Partners
4	Dependents	Churn Percentage is less for Customer with Dependents
5	Phone Service	Churn Percentage slightly higher for Customers with Phone Service
6	Multiple Lines	Couple of inconsistencies observed with Multiple Lines. Either it should be yes or no. We can see some duplication with No and No Phone service, each indicating similar answers. So will need cleaning of data. Churn Percentage is higher with customers who have Multiple Lines.

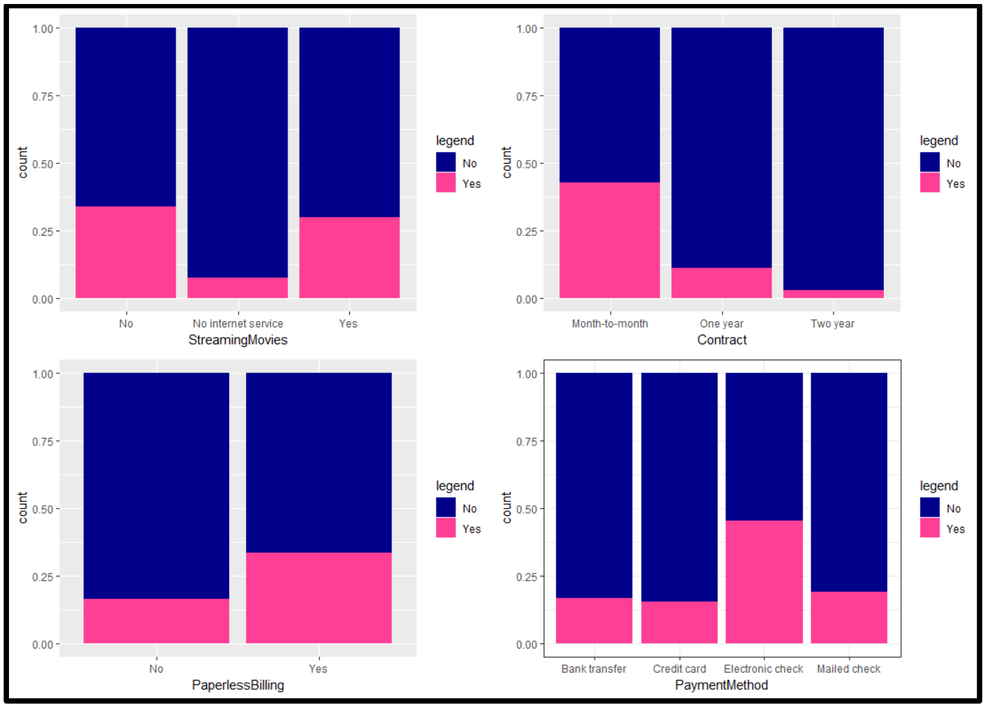
- **Set 2: Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV:**



- **Observations:**

Sr. No.	Variable v/s Churn	Observations
1	Internet Service	Churn Percentage is higher for Fiber Optic Internet Service
2	Online Security	Churn Percentage is higher for customers with No Online Security
3	Online Back up	Churn Percentage is higher for customers with No Online Back up
4	Device Protection	Churn Percentage is higher for customers with No Device Protection
5	Tech Support	Churn Percentage is higher for customers with No Tech Support
6	Streaming TV	Churn Percentage is higher for customers who do not have Streaming TV services

- **Set 3: Streaming Movies, Contract, Paperless Billing, Payment Method:**

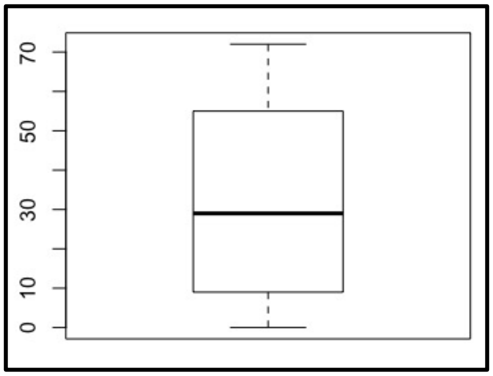


- **Observations:**

Sr. No.	Variable v/s Churn	Observations
1	Streaming Movies	Churn Percentage is higher for customers who don't have Streaming Movie Services
2	Contract	Churn Percentage is higher for customers who have monthly billing as compared to yearly contracts
3	Paperless Billing	Churn Percentage is higher for customers with Paperless Billing
4	Payment Method	Churn Percentage is higher for customer with Electronic Check Payment services

➢ Checking for Outliers :

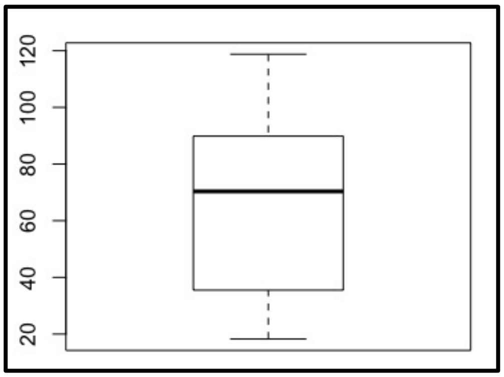
- **Set 1: Boxplot for Tenure:**



Observations:

- Boxplot indicates there is no outliers for the variable Tenure.
- The database shows good amount of variance from min to max (0 to 70 months) for tenure.

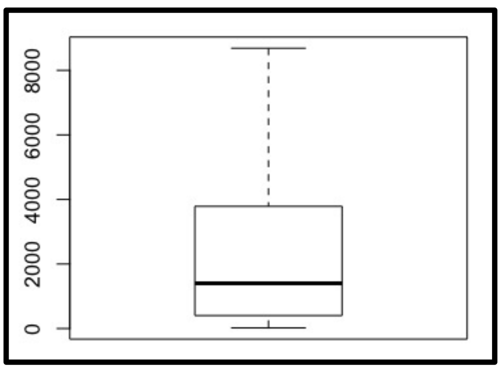
- **Set 2: Boxplot for Monthly Charges:**



Observations:

- Boxplot shows there is no outliers for monthly charges.
- Monthly charges for all the customers are in the range of 20 USD to 120 USD.
- The median for the variable Monthly charges is 70 USD.

- **Set 3: Boxplot for Total charges:**

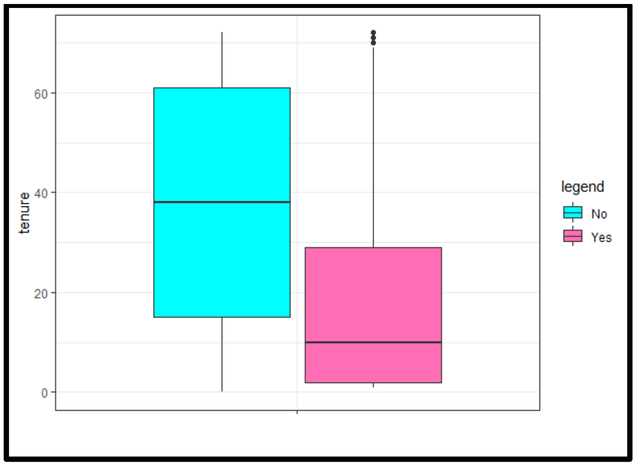


Observations:

- Boxplot shows there are no outliers for the total charges.
- The approximate median value for variable Total Charges is 1500 USD.

➢ Data Exploration: Numerical Variable v/s Outcome Variable Churn:

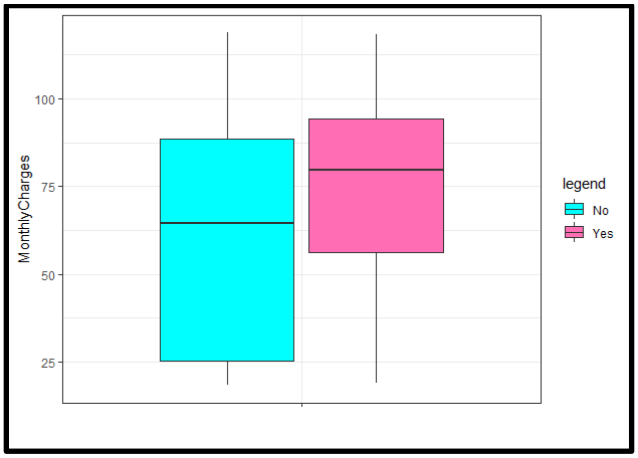
- **Set 1: Tenure V/s Churn:**



Observations:

- Customers who have churned out have median of 10 months
- While Customers who have stayed longer on the platform have lower churn rate

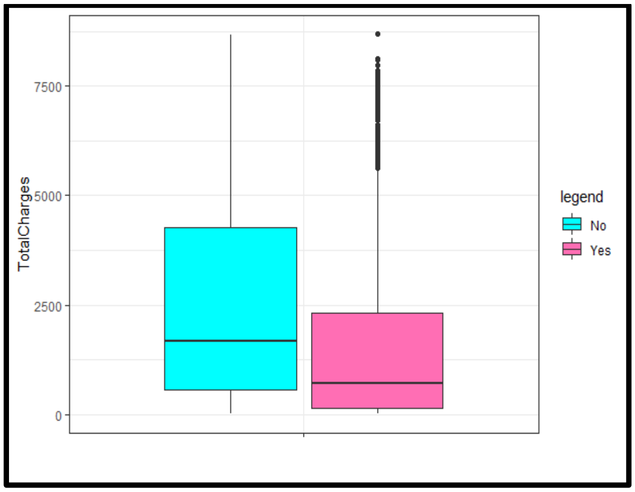
- **Set2: Monthly Charges V/s Churn:**



Observations:

- Customers who have churned out have high monthly charges. Median is above 75
- While Customers who have lower monthly charges have stayed longer on the platform.

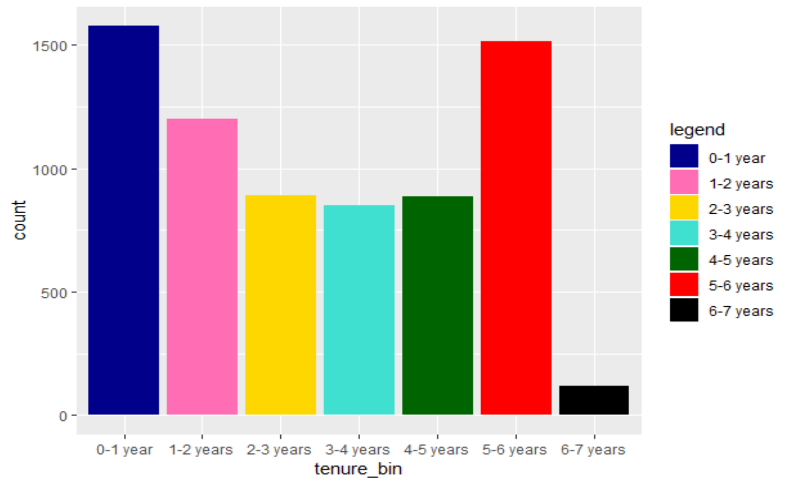
- **Set 3: Total Charges V/s Churn:**



Observations:

- Customers who have churned out have lower median of the total charges.
- We see some outliers as well at the upper end.

➢ Data Exploration: Binning the Tenure Column from Months to Years:



Observations:

- Most of the customers have been with Telco either for 0-1 year or 5-6 years
- Only 1 data point for beyond 6 years. So, removing that value for further analysis.

Final Pre-processing Steps Accomplished:

1. **Cleaning Data & Removing Redundancy, Duplications.**
    a. Data Set had 11 Missing values in Total Charges column. Replaced the missing values with Median Value of the column
    b. Certain Variables had Data Redundancy with options like No, No internet services and Yes, of these No and no internet services means NOT HAVING THE SERVICE. So replaced the duplication with No alone and changed the options to YES and NO only.
    c. Replaced Senior Citizen Column with YES and NO only.

2. **Dummy Creation:**
    a. Converted all the 17 factor variables into binary for ease of calculations.

3. **Standardizing the Continuous Variables:**
    a. Standardized 3 continuous variables like Tenure, Monthly Charges and Total Charges
4. **Processing the Final Data Set:**
    a. Created a final Data set to be used for further analysis – "Telco Final", which comprises of cleaned binary variables and standardized continuous variables.
5. **Partitioning:**
    a. Created 3 subsets of Telco Final as Train data, Valid Data and Test Data.
    b. Training data was used to train the models
    c. Validation data was used to validate the performance of each model and comparison for best model
    d. Test data was used for Predictions on the best model found.

Step 3: Building Models on Various Algorithms: -

Method 1: Logistic Regression:

- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more explanatory variables by estimating the probabilities using a logistic function.
- Regression analysis helped us to understand how the value of the dependent variable changes with change in independent variable keeping all others constant.
- Logistic Regression helps us to understand output in terms of odds & probabilities.

Building Logistic Regression Model

- We start with a Logistic Regression Model, to understand correlation between different predictors and outcome variable ‘Churn’.

- Beforehand we had cleaned dataset, converted all the non-numerical variables into factors.

```
> summary(glm.step)
Call:
glm(formula = Churn ~ tenure + MonthlyCharges + gender + Dependents + InternetService.xFibe
r.optic + InternetService.xNo + OnlineSecurity +  OnlineBackup + DeviceProtection + TechSup
port + StreamingTV + StreamingMovies + Contract.xOne.year + Contract.xTwo.year + PaperlessB
illing + PaymentMethod.xElectronic.check + tenure_bin.x1.2.years + tenure_bin.x5.6.years +
tenure_bin.x6.7.years, family = "binomial", data = train.data)

Deviance Residuals:
Min       1Q    Median      3Q      Max
-1.8216  -0.6952  -0.2953   0.7538   3.1299

Coefficients:
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)         -0.70278    0.14743  -4.767    0.000001870424265900 ***
tenure              -0.43124    0.07222  -5.971    0.000000002360041168 ***
MonthlyCharges      -0.10409    0.05172  -2.012              0.04418 *
gender              -0.15905    0.09039  -1.760              0.07846 .
Dependents          -0.28774    0.11128  -2.586              0.00972 **
InternetService.xFiber.o 0.74896 0.11460   6.535    0.000000000063466272 ***
InternetService.xNo -1.30881    0.18904  -6.923    0.000000000004409039 ***
OnlineSecurity      -0.52019    0.11465  -4.537    0.000005705167222105 ***
OnlineBackup        -0.49537    0.10486  -4.724    0.000002312116408877 ***
DeviceProtection    -0.20875    0.10829  -1.928    0.05391 .
TechSupport         -0.47943    0.11926  -4.020    0.000058172652494034 ***
StreamingTV          0.22030    0.11162   1.974    0.04842 *
StreamingMovies      0.21071    0.11054   1.906    0.05663 .
Contract.xOne.year -1.12039    0.13993  -8.007    0.000000000000001176 ***
Contract.xTwo.year -1.95360    0.24239  -8.060    0.000000000000000764 ***
PaperlessBilling     0.27604    0.10365   2.663    0.00774 **
PaymentMethod.xElectronic 0.37394 0.09576 3.905    0.000094183790189218 ***
tenure_bin.x1.2.years  -0.27506 0.11577  -2.376    0.01751 *
tenure_bin.x5.6.years 0.32116   0.19202   1.673    0.09442 .
tenure_bin.x6.7.years 0.56664   0.35578   1.593   0.11124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 4100.0  on 3520  degrees of freedom
Residual deviance: 3004.2  on 3501  degrees of freedom
AIC: 3044.2

Number of Fisher Scoring iterations: 6
```

Output Interpretation:

- Top Significant Predictors are:
    - ✓ Tenure
    - ✓ Internet Services
    - ✓ Online Security
    - ✓ Online Back Up
    - ✓ Tech Support
    - ✓ Contracts

- ✓ Electronic Payments

- The **negative coefficients** for categorical variables like ***Online backup, Device Protection, Online Security, Tech Support*** indicate that ***NOT HAVING THESE SERVICES*** would lead to higher probabilities of customers leaving the platform.

- The **positive coefficients** for categorical variables like ***Internet Services (Fiber Optic), Electronic Payment Methods*** indicate that HAVING THESE SERVICES would lead to higher probabilities of customers NOT leaving the platform.

- The **negative coefficients** for numerical variables like ***Tenure*** Represent that higher values of TENURE indicate lower chances of customers churning out of the Telco platform.

Validation Data Accuracy: (Confusion Matrix output displayed below)

```
> confusionMatrix(as.factor(ifelse(logit.reg.pr
ed > 0.5, 1, 0)), as.factor(valid.data$Churn))
Confusion Matrix and Statistics

Reference
Prediction    0    1
0 1360  257
1  202  293

Accuracy : 0.7827
95% CI : (0.7645, 0.8001)
No Information Rate : 0.7396
P-Value [Acc > NIR] : 0.000002433

Kappa : 0.4169

Mcnemar's Test P-Value : 0.01172

Sensitivity : 0.8707
Specificity : 0.5327
Pos Pred Value : 0.8411
Neg Pred Value : 0.5919
Prevalence : 0.7396
Detection Rate : 0.6439
Detection Prevalence : 0.7656
Balanced Accuracy : 0.7017

'Positive' Class : 0
```

Analysis:

- The accuracy obtained with confusion matrix on validation data is 78.27%.
- Sensitivity = 87.13%
- Specificity = 53.09%
- From confusion matrix, we observed that our model shows positive class 0 (non-churning customer) and negative class is 1 (churning customers). As a result, our model has the power

of predicting the customers churning out(specificity) from telco platform is 53.09% and prediction power of customers who are loyal and not churning(sensitivity) is 87.13%.

Method 2: Classification Trees:

- Decision Trees work best for Classifying outcome. The best feature of using trees are its easy interpretability of rules which is very useful for analytics and predictions.
- Decision trees implicitly perform variable screening or feature selection so no need to apply separate techniques to reduce number of predictors
- Decision trees require relatively little effort from users for data pre-processing
- Nonlinear relationships between parameters do not affect tree performance

Steps in Executing Classification Trees:

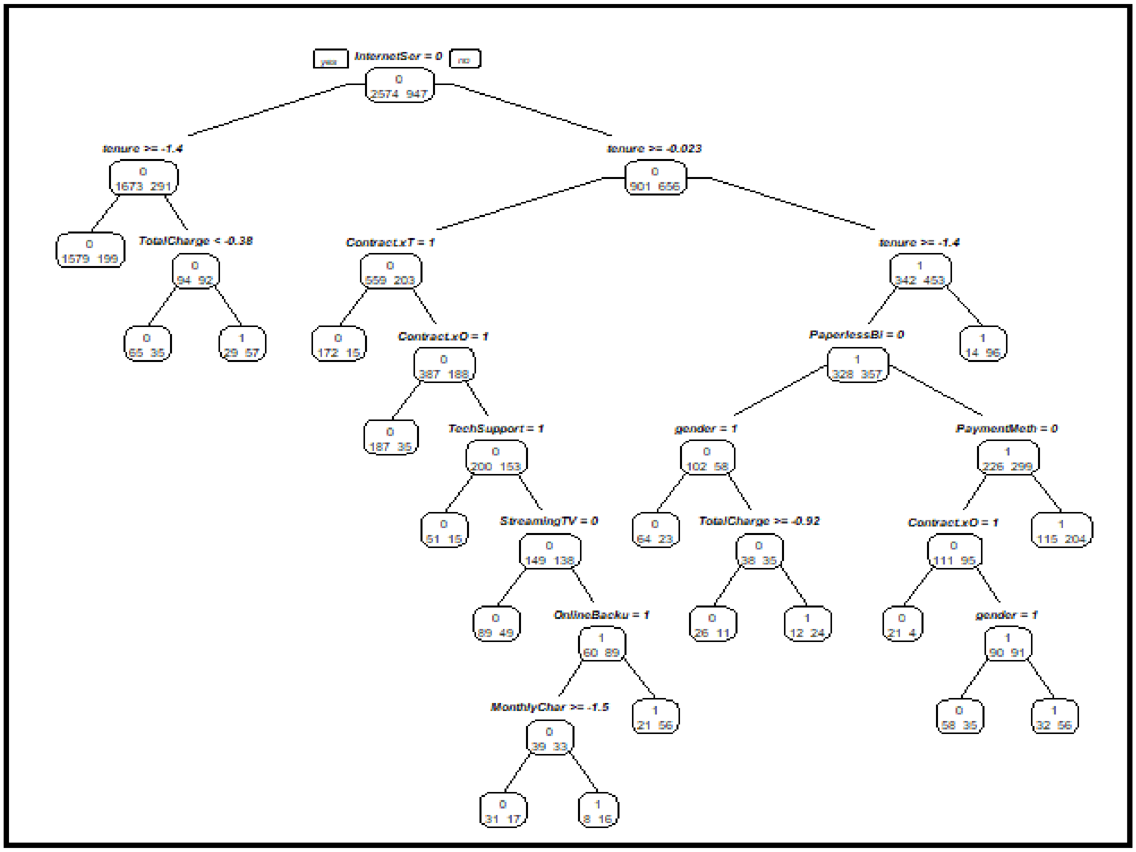
- Build a Classification Tree on the entire database with all the variables of Telco Final. Full grown Tree with minimum cp. (complexity parameter).
- Prune the Full-grown Tree based on minimum xerror obtained from the cp table. Full grown tree overfits the data and while it reduces error in training data, it reduces accuracy for validation data. Hence to increase the accuracy of the validation data, we prune the tree.
- ***Without compromising on the accuracy, we chose the BEST Pruned Tree(with cp corresponding to min xerror + 1 std) which was much interpretable and worked best on parsimonious property.***
- Tested the best pruned Tree model onto Validation Data. Noting the accuracy obtained from the validation data.

Output Interpretation:

- **Accuracies and splits of three trees built in the model:**

	Accuracy on validation data	Split
Deeper Tree	0.7182765 (71%)	678
Pruned Tree	0.7741477 (77%)	18
CP+1std error tree	0.7722538(77%)	10

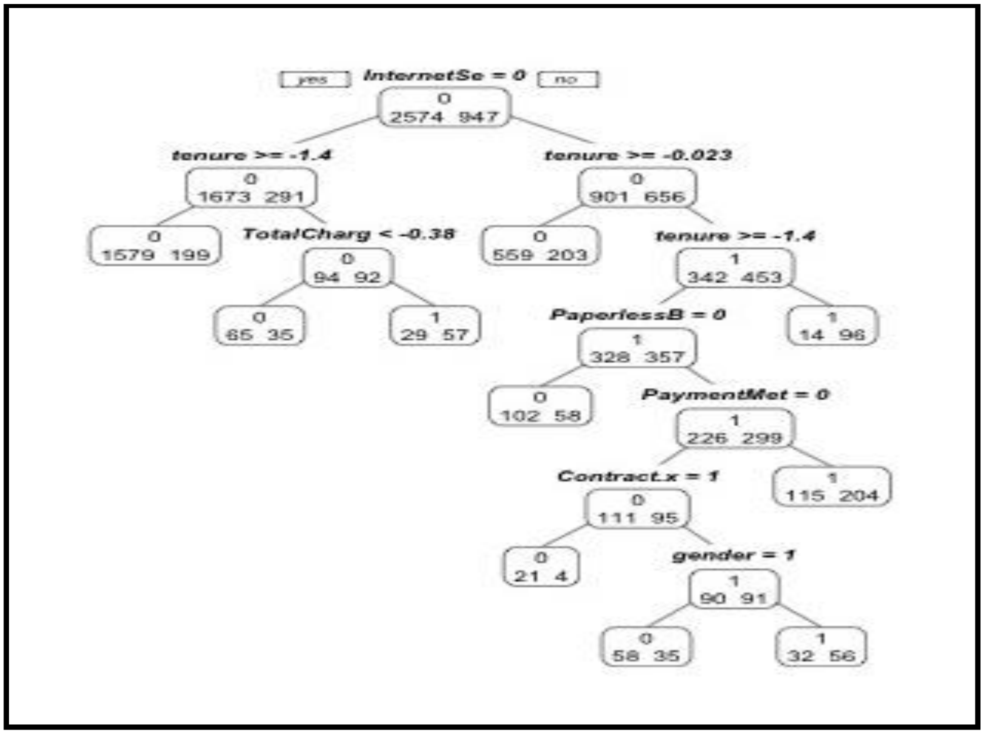
- **<u>Pruned Tree: (*18 Terminal Nodes and lowest cp*)</u>**



Observations:

- **<u>Top Significant Predictors:</u>**
  - Internet Services
  - Tenure
  - Total Charges
  - Contract
  - Tech Support
  - Payment Method

Best Pruned Tree: (10 Terminal Nodes and cp ~ (min xerror+1std))



Observations:

- **<u>Top Significant Predictors:</u>** ***<u>Variables contributing towards CHURN:</u>***
  - Not having internet Service
  - Smaller tenures of customers with Telco Company
  - Having Paperless Billing, which provides as easy access to customers for quitting
  - Electronic payment methods
  - Not having yearly contracts but monthly billing plans.
- **<u>Association Rules:</u>** <u>(</u>***values of Tenure are standardized values displayed on the tree)***
  - Customers having: **Internet service and Tenure < -0.023 and Tenure < -1.4 will be customers churning out of Telco platform.**
  - **Also, IF (Internet Service=0) AND (tenure<=-1.4) AND(Totalcharge<-0.38) THEN Class=0**
- **<u>Validation Data Accuracy: (Confusion Matrix output displayed below)</u>**

```
confusionMatrix(pruned.ct.point.pred.valid2, as.factor(val
id.data$Churn))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1417  336
         1  145  214

               Accuracy : 0.7723
                 95% CI : (0.7538, 0.79)
    No Information Rate : 0.7396
    P-Value [Acc > NIR] : 0.0002896

                  Kappa : 0.3338

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9072
            Specificity : 0.3891
         Pos Pred Value : 0.8083
         Neg Pred Value : 0.5961
             Prevalence : 0.7396
         Detection Rate : 0.6709
   Detection Prevalence : 0.8300
      Balanced Accuracy : 0.6481

       'Positive' Class : 0
```

Method 3: Random Forest:

- RF helps to improve predictive performance by averaging or voting the several trees grown from random samples of original data.
- To overcome co-relation of predictors, RF works the best. It runs parallel trees on random samples of training data with random subset of original predictors.
- Var Importance plot, will help us visualize the importance of each predictor, impacting the churning of the customer churn with this Telco company.
- The importance score is computed by summing up the decrease in the Gini index for that predictor over all the trees in the forest.
- It does not require extensive handling of missing values and outliers in dataset. Algorithm takes care of it internally in effective manner.

Steps in Executing Random Forest:

- Build Random forest of 1000 trees on training data.
- Tested the accuracy of the forest on Validation data
- Used Variable Importance plot to signify the top predictors.

➢ **Output Interpretations:**

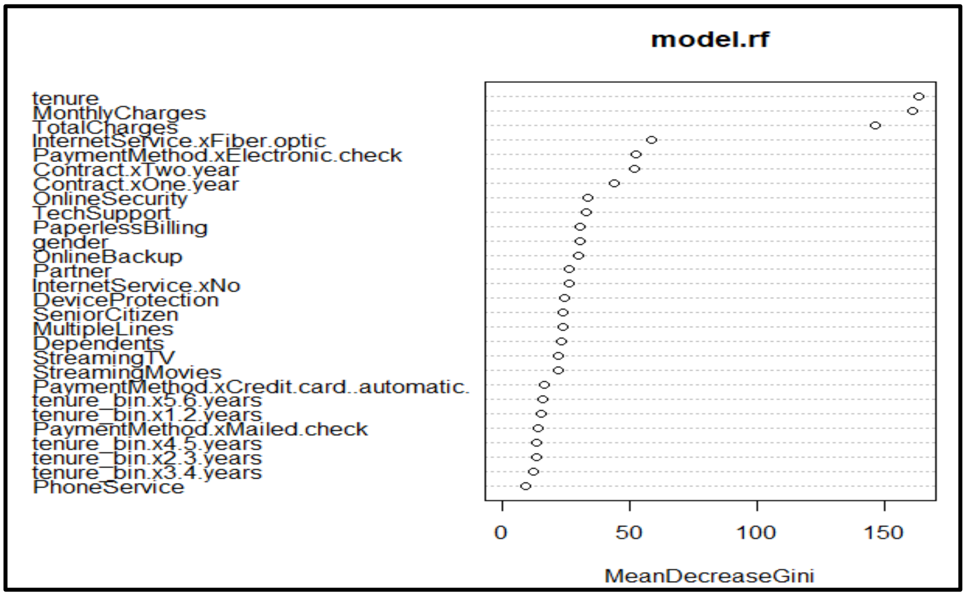
Confusion matrix on validation data

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1386  292
         1  176  258

               Accuracy : 0.7784
                 95% CI : (0.7601, 0.796)
    No Information Rate : 0.7396
    P-Value [Acc > NIR] : 2.004e-05

                  Kappa : 0.3826

 Mcnemar's Test P-Value : 1.061e-07

            Sensitivity : 0.8873
            Specificity : 0.4691
         Pos Pred Value : 0.8260
         Neg Pred Value : 0.5945
             Prevalence : 0.7396
         Detection Rate : 0.6562
   Detection Prevalence : 0.7945
      Balanced Accuracy : 0.6782

       'Positive' Class : 0
```

Observations:

- Accuracy obtained with Confusion Matrix on validation data is 77.84%.
- Sensitivity = 88.73%
- Specificity = 46.91%.
- Output of confusion matrix shows positive class is 0(Non-churning customers) and Negative class is 1(churning of customers), Which means that our model has the power of predicting the Churning customers (specificity) of 46.91%.

Variable Importance Plot



Observations:

- **Variable Importance plot measures the relative importance of each variable towards the outcome. IT measures the importance by summing the decrease in the GINI INDEX across all the trees in the forest.**
- From the variable importance plot, we observe that variable ***tenure, monthly charges and total charges*** have highest impact on the Churning of customers on the platform.
- Thus, from Gini impurity it can be said that variables tenure, monthly charges and total charges contributes maximum towards increasing the node purity.

Step 4: Comparing Models based on Validation Data Accuracy: -

Logistic Regression	Classification Trees	Random Forest
78.27%	77.22%	77.6%

- **Model Comparison Interpretations:**
  - Logistic Regression supersede in terms of accuracy on Validation Data
  - As customer churn prediction is a significant problem in customer relationship management, it would be the right choice to make predictions on future behavior by running the best model obtained on test data in order to get an unbiased estimate of how well the model will perform with new data.
  - Thus, we will progress further for testing our TEST Data using logistic model.

Step5: TEST DATA USING LOGISTIC REGRESSION MODEL: -

- **Confusion Matrix Accuracy for Test Data:**
  - Accuracy = 79.43%
  - Sensitivity = 88.34%
  - Specificity = 54.47%

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 917 169
         1 121 203

               Accuracy : 0.7943
                 95% CI : (0.7723, 0.815
    No Information Rate : 0.7362
    P-Value [Acc > NIR] : 0.0000002201

                  Kappa : 0.4477

 Mcnemar's Test P-Value : 0.005781

            Sensitivity : 0.8834
            Specificity : 0.5457
         Pos Pred Value : 0.8444
         Neg Pred Value : 0.6265
             Prevalence : 0.7362
         Detection Rate : 0.6504
   Detection Prevalence : 0.7702
      Balanced Accuracy : 0.7146

       'Positive' Class : 0
```

Observations:

- Our Logistic Model performed better on the Test data as compared to Validation Data. ***Our accuracy increased from 78.27% on Validation data to 79.43% on Test Data.***
- Also, sensitivity and specificity seem to be increased, as a result of which power of predicting loyal/non-loyal customers also increases.

Step6: Other Options Explored: -

- **<u>Option1</u>**: We tried Making **Tenure** as our Output Variable instead of CHURN.
  - We observed that Tenure depends on Predictors like Total Charges, Monthly Charges and Gender.
  - Churn has NO INFLUENCE ON Tenure
  - Also Accuracy of Tenure v/s Other predictors using Logistic Model was only 32%
  - Thus, we dropped this variable as outcome variable and continued with CHURN
- **<u>Option2</u>**: Use of Interaction Variables to obtain some interesting product bundling combinations for retaining the customers.
  - Unfortunately, none of the interaction variables that we obtained from the system had product combinations. Most of them were in combinations like (Tenure * Gender), (Monthly charges * Tenure), etc.
  - As well, feeding interaction variables into our Logit model was not increasing our accuracy on validation data in any way.
  - Thus, following the principle of parsimony, we continued with original variables, which gave us the good results and decent accuracy on test data.

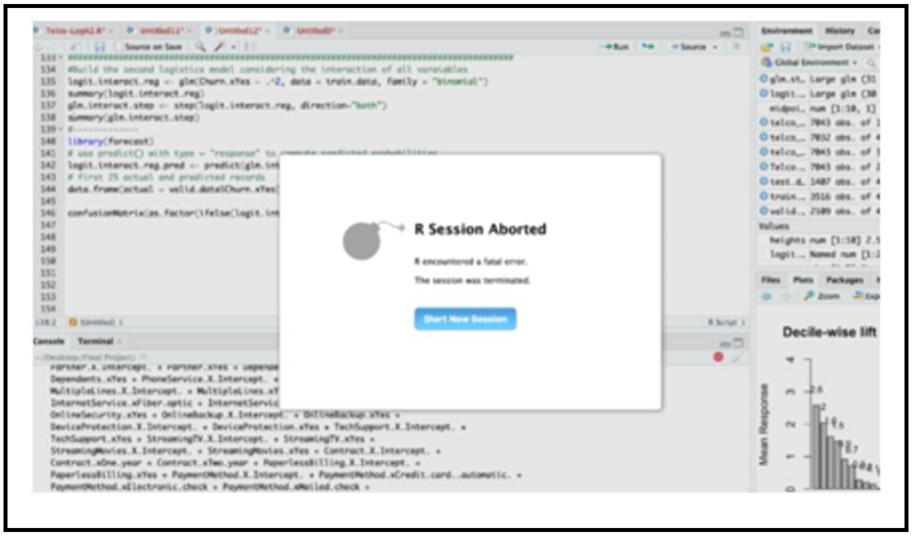
Step6: Challenges faced

- As interaction effects are common in Regression analysis, so we thought to test this concept.
- We tried to create two-degree relationships between the predictors to see how differently the model behaves.
- We pick the interaction variables with 3-star significance and lower p-values.

```
logit.interact.reg <- glm(Churn ~ (.)^2, data = train.data, family = "binomial")
summary(logit.interact.reg)

Contract.xOne.year                  0.001936 **
tenure_bin.x1.2.years               0.000363 ***
tenure_bin.x2.3.years               0.009850 **
tenure_bin.x6.7.years               0.026807 *
tenure:Contract.xOne.year           0.007760 **
tenure:tenure_bin.x1.2.years 0.00000116 ***
tenure:tenure_bin.x2.3.years 0.00000280 ***
tenure:tenure_bin.x3.4.years 0.00000280 ***
tenure:tenure_bin.x4.5.years 0.010345 *
tenure:tenure_bin.x5.6.years 0.003036 **
MonthlyCharges:gender               0.023321 *
MonthlyCharges:Contract.xOne.year   0.032760 *
gender:TechSupport                  0.005505 **
gender:Contract.xOne.year           0.000248 ***
SeniorCitizen:PaymentMethod.xCredit.card..automatic.  0.018837 *
Partner:PhoneService                0.001078 **
Partner:InternetService.xFiber.optic 0.013013 *
Partner:Contract.xTwo.year          0.038481 *
Dependents:PhoneService             0.000891 ***
Dependents:InternetService.xFiber.optic 0.035516 *
Dependents:InternetService.xNo  0.022722 *
Dependents:OnlineBackup         0.029908 *
Dependents:PaymentMethod.xMailed.check  0.039020 *
```

- After many trials, we choose three best interacting variables for our model. We tested them along with all the predictors and not with the subset of variables obtained from the step method. The motive behind this approach was to identify all possible significant relationships.

- But once we tried using step method for choosing the best subset of behavior, the R file was running for more than 6 hours and ultimately giving the below output.



- Hence, we choose below combination of variables to perform a stepwise method on them and obtain the best subset to run along with other predictors.

```
Call:
glm(formula = Churn ~ tenure + Contract.xOne.year + MonthlyCharges +     gend
er + TechSupport + Partner + PhoneService + InternetService.xFiber.optic +
Contract.xTwo.year + Dependents + InternetService.xNo + OnlineBackup +
tenure:Contract.xOne.year + Contract.xOne.year:MonthlyCharges +
MonthlyCharges:gender + gender:TechSupport + Contract.xOne.year:gender +
Partner:PhoneService + PhoneService:Dependents + Dependents:InternetService.x
No +Dependents:OnlineBackup, family = "binomial", data = train.data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9828  -0.6822  -0.2897   0.7185   3.1771

Coefficients:
                         Estimate Std.Error z value             Pr(>|z|)
(Intercept)              -0.83970  0.23232  -3.614             0.000301 ***
tenure                   -0.38215  0.06683  -5.718 0.0000000107564655 ***
MonthlyCharges            0.01539  0.06947   0.221             0.824707
gender                   -0.22064  0.10076  -2.190             0.028536 *
SeniorCitizen             0.16780  0.11814   1.420             0.155526
Partner                   0.95702  0.32672   2.929             0.003399 **
Dependents               -1.14507  0.40004  -2.862             0.004204 **
PhoneService              0.10670  0.22228   0.480             0.631222
InternetService.xFiber.opti0.81801 0.13067   6.260 0.00000000038415637 ***
InternetService.xNo      -1.32451  0.20188  -6.561 0.00000000005353790 ***
OnlineSecurity           -0.52526  0.11559  -4.544 0.00000551476055769 ***
OnlineBackup             -0.52220  0.10567  -4.942 0.00000077321658921 ***
DeviceProtection         -0.19893  0.10951  -1.817             0.069286 .
TechSupport              -0.47125  0.12038  -3.915 0.00009047164944057 ***
StreamingTV               0.20168  0.11297   1.785             0.074226 .
StreamingMovies           0.23612  0.11174   2.113             0.034592 *
Contract.xOne.year       -1.44420  0.20151  -7.167 0.00000000000076668 ***
Contract.xTwo.year       -1.96899  0.24328  -8.093 0.00000000000000058 ***
PaperlessBilling          0.27331  0.10452   2.615             0.008921 **
PaymentMethod.xElectronic0.41200   0.10596   3.888             0.000101 ***
PaymentMethod.xMailed.    0.20518  0.14096   1.456             0.145490
tenure_bin.x1.2.years    -0.27676  0.11680  -2.369             0.017814 *
tenure_bin.x5.6.years     0.27189  0.18565   1.465             0.143053
Partner:PhoneService     -1.17579  0.34192  -3.439             0.000584 ***
Dependents:PhoneService   1.03043  0.41940   2.457             0.014013 *
gender:Contract.xOne.year 0.64274  0.26057 2.467               0.013638 *
MonthlyCharges:gender    -0.23224   0.08963  -2.591             0.009571 **
```

- But ultimately, when testing the above model with validation data, we did not get any increase in accuracy from the confusion matrix.

```
Confusion Matrix and Statistics
          Reference
Prediction    0    1
         0 1361  258
         1  201  292

               Accuracy : 0.7827
                 95% CI : (0.7645, 0.8001)
    No Information Rate : 0.7396
    P-Value [Acc > NIR] : 0.000002433

                  Kappa : 0.4162

 Mcnemar's Test P-Value : 0.008953

            Sensitivity : 0.8713
            Specificity : 0.5309
         Pos Pred Value : 0.8406
         Neg Pred Value : 0.5923
             Prevalence : 0.7396
         Detection Rate : 0.6444
   Detection Prevalence : 0.7666
      Balanced Accuracy : 0.7011

       'Positive' Class : 0
```

- So finally, we thought to drop the idea of introducing the interaction variables in our model.

Step7: Recommendations:

Demographic

- We observe that individuals without partners/dependents are more likely to leave the company. Thereby, individuals 'with' partners and dependents are more likely to remain loyal with Telco as most telecom companies offer exclusive deals for families and couples. As a recommendation, we would recommend the Telco company to run promotional campaigns targeted towards individuals with partners/dependents, so that they can have bulk business and more loyalty.

Service-Specific

- Internet Services seem to have utmost importance on Customer Retainability. Thus we recommend Telco Company to make sure that their customers are well aware of their Internet Services and customers are being offered this service for long term relationship with this company.
- On the other hand, many customers that are leaving the company seems not registered for support-like services (i.e., Online Tech Support, Device Protection, Internet Services). We would recommend the company to start offering the above-listed value-added services bundled with other main products to the customers. This little change in strategy would help the company to retain more loyalty.
- Yearly contracts should be promoted to bind the customers for a longer time with the company. Increased Tenure has shown more retainability of the customers.
- Understanding Customer Budgets and offering affordable monthly packages can be another strategy to retain customers for a longer time.

Step8: Conclusions: -

After going through various prefatory steps, including data and library loading, preprocessing. We carried out three statistical classification methods conventional in churn analysis. We identified several important churn predictor variables from these models and compared these models on accuracy.

Here is a summary of our findings:

- Customers with month-to-month contracts are more likely to churn.
- Customers with internet service, in particular, fiber optic service, are more likely to churn.
- Customers who have been with the company longer or have paid more in total are less likely to churn.
- Logistic regression supersedes over other data mining methods like classification tree and random forest analysis.

Step 9: References: -

References:

1. *https://www.kaggle.com/blastchar/telco-customer-churn*
2. *https://www.displayr.com/how-to-interpret-logistic-regression-coefficients/*
3. *https://medium.com/@rohitlal/customer-churn-prediction-model-using-logistic-regression-490525a78074*
4. *https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/*
5. *https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/*
6. *https://ja.exploratory.io/note/aRr1gmQ0BX/Telco-Customer-Churn-ylV6EZY8nm*
7. *Data mining with R- Business Analytics book*
8. https://stackoverflow.com/