Customer Churn Prediction

Rishabh Khorana (22768807)

The University of Western Australia

Executive Summary

The purpose of the report is to find the reasons which influence the churning of the customers for Flash Telecom services. This report explores the data set that contains 7043 rows and has 20 variables. The data contains 3 Quantitative variables and 17 Qualitative variables. The insight provided by the data analysis targeted contract duration, tenure of the customer in the company and the services provided by the company played an important role influencing customer Churn.

At the end of the report my recommendation to "Flash telecom services" would be to increase their engagement with the customers by working on their marketing strategy and implementing a customer loyalty program.

TABLE OF CONTENTS

1. INTRODUCTION

As a private data analyst, I have been approached by our client "**Flash Telecom services".** Client wants to know what they can do to reduce their Customers switching over to their competitors. For the analysis we asked for their customer data which contained the entries of last 6 year or 72 months. This data would be essential to understand the Churn of the Customers whether they leave the company or not.

Our aim is to find the factors which influence the Churn of the Customers. This would give an insight about their customers and which factors led them to change to their competitors in the market. Moreover, using the analysis from the data this would help the Flash Telecom Services plan their strategy to retain their existing customers and potentially increase their customer base.

The results from the analysis would help the company in marketing and approach the right customers.

2.METHOD

Data set contains 20 variables which has 17 categorical variables and 3 Qualitative variables (Numerical). The 17 categorical variables these can classified as Nominal variables(N) and Dichotomous Variables(D).

1. **Churn(N)(D)** – Did the customer leave the company Yes or no. (Target Variable)
2. **Payment Method(N)** – Mode of payment customer Chooses to pay the bill.
3. **Paperless Billing(N)(D)** – Customer opted for paperless bill yes or no.
4. **Contract(N)-** Type of contract customer has taken with company. Monthly, 1yearly, 2 yearly.
5. **Streaming Movies(N)(D)-** If customer uses internet services to stream movies from a third-party provider. Yes or no.
6. **Streaming TV(N)(D)-** If customer uses internet services to stream TV program from a third-party provider. Yes or no.
7. **Tech Support(N)(D)-** If Customers subscribes to additional technical support plan from the company with reduced wait times. Yes or No.
8. **Device Protection(N)(D)-** If Customers subscribes to additional Protection plan provided for the internet equipment provided by the company. Yes or No.
9. **Online Backup(N)(D)-** If Customers subscribes to additional online backup service provided by the company. Yes or No.
10. **Online Security(N)(D)-** If Customers subscribes to additional online security service provided by the company. Yes or No.
11. **Internet Service(N)(D)-** If Customers subscribes to internet service with the company. Yes or No.
12. **Multiple Lines(N)(D)-** If Customers subscribes to multiple telephone lines with company. Yes or No.
13. **Phone Services(N)(D)** – Customer has opted for phone services Yes or No.
14. **Dependents(N)(D)** – Does live with Children, parents, grandparent yes or No.
15. **Partner(N)(D**) – Married or not.
16. **Senior Citizen(N) (D)-** Customer over 65 or not.
17. **Gender (N) (D)** – Male or Female.

Qualitative variables

18. **Tenure** – Total amount of months the customer has been with the company.
19. **Monthly Charges**- Total monthly Charge for all the services for each customer.
20. **Total Charges**- Total Charge for all the services for each customer, Calculated as product of tenure and Monthly Charges.

Since we are trying to predict customer churn which has Values “Yes” or “No”. This a binary variable with only two outcomes. For this problem we need logistic regression which can predict the likelihood of churn of the customer. “Will the customer leave the company or not leave the company?”

Data pre-processing

The process starts with pre-processing of the data. Before we move to the modelling, we need to make sure the data is free of value errors, blank values, the data should have the correct information under each column, to check for any duplicate values.

In Fig1, Fig2 and Fig3 we have done data pre-processing. We have removed the unnecessary data from the file before loading and transforming the data in R.

Concept of Logistic Regression

The general assumption about the logit transformation of the outcome variable is has liner relationship with the predictor variables. The odds of Success of an event are:

- Odds of Success = probability of success / probability of failure

The Range of probability is 0 to 1(max) and odds is 0 to positive infinity. This means as the probability increase the odds of success or failure increases. So, Odds ∝Probability. Transforming odds to log odds is know as the log transformation or logit transformation. This transformation helps to remove the problem of range restriction as we have with probability that is a range of 0 to 1 and this helps in creating a model with a range of negative infinity to positive infinity.

A standard logit equation for a binary outcome variable and other predictor variables with failure = 0 and Success = 1 and P = probability of success for an event y = 1. Predictor variables be x1,x2,x3,…xn and their coefficient of the predictors be B0,B1,B2….Bn. The equation we get is.

$$\text{Logit(P)} = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \dots \beta_n x_n$$

Solving the equation by exponentiate and inverse multiplication we get.

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}$$

In the end we get:

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}$$

Splitting the data into Test set and Training set

We split the data into test set and training set so that we can create our models on training and validate the results of the model on test set. When we split the model between training and test set, we must be sure of that our training set is not so small otherwise the model will have less accuracy and would have a higher entropy (Disorder). For the logistic model, the data has been split into 70% training set and 30% test set. For the decision tree the data has been split into 80% training set and 20% test set.

Splitting the data into test and training sets prevents the model going into overfitting (good performance on training data but poor performance on other data) and Underfitting (Where the model performs poorly on training and other data set.)

Finding relationship between the Variables

To find the relationship between our target variable churn and other predicted variables we conducted test of independence. For Qualitative variables we conducted a **Chi-square test for independence** and Quantitative tests we conducted **independent T-test.** We check if there is a significance between two variables (Dependent and predictor). For the test we have taken p-value if it is less than 0.05 then it means there is a relationship between the Dependent and the independent variable.

Ho: there is no relationship between dependent (Customer Churn) and Independent variable

Ha: there is a level of significance between dependent (Customer Churn) and independent variable.

We are not conducting a Fisher's test as all the observations for our chi-square test are above 5 for the all the variables.

Multicollinearity and VIF (Variance inflation factor)

Multicollinearity means if the predicting variables are related to each other. This will affect your model as it will increase standard errors of the coefficients and it can influence your predictors to be statistically insignificant. VIF or Variance Inflation factor is measure of variance for each coefficient inflated to due to Multicollinearity.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R-square is known as a pseudo R2 or McFadden's R square. The VIF value is calculated for each predictor against all the other predictors in the model. This is true when the variables have **1df** but for variables which more than **1df** there is a need of corrected VIF or Generalised VIF

$$GVIF = VIF^{\left(\frac{1}{(2*df)}\right)}$$

GVIF more than 5 indicates a strong collinearity between the variables. GVIF of 10 is considered a critical value of GVIF. So, a VIF 1.9 for a particular variable would have a coefficient 90% bigger than one without no multicollinearity.

Test for overall significance (Anova)

This test is carried out to check the significance of the predictors in the model and Conducting an "LRT" test compares the nested the model with the full model by adding each predictor at a time. This test uses the likelihood ratio to compare the models.

3.Descriptive Statistics and Preliminary Correlation Analysis

The data has 7,043 observations and has 20 variables. Since our data contains qualitative data and must be converted into factors before we start the analysis.

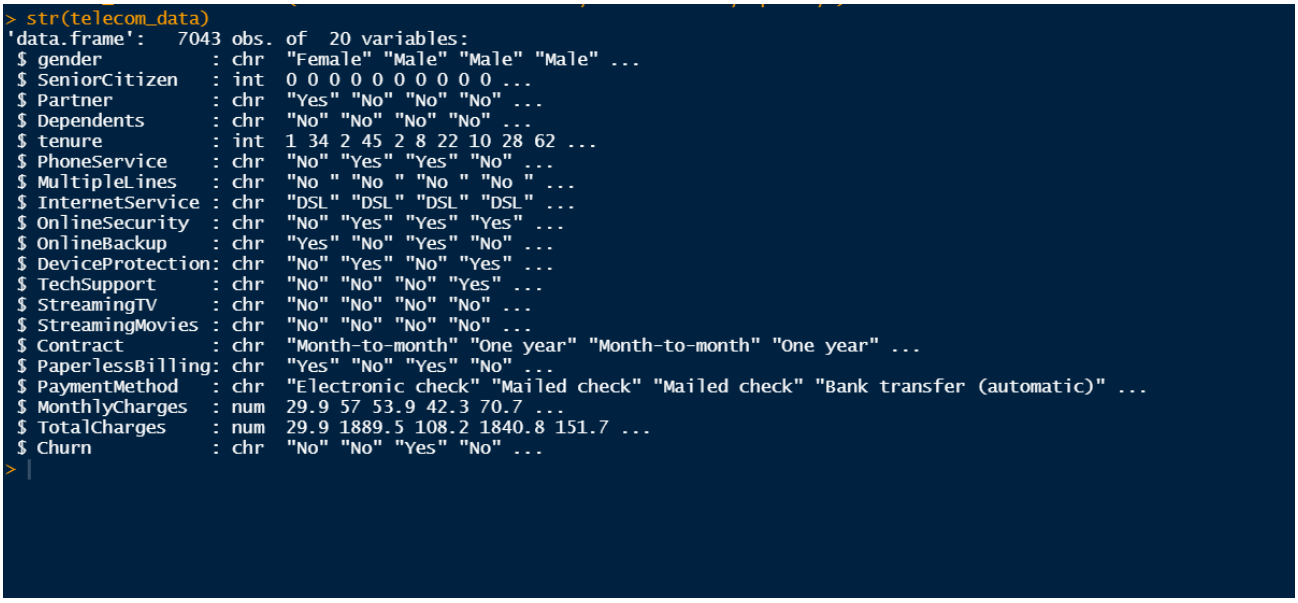


Fig: Structure of the raw data

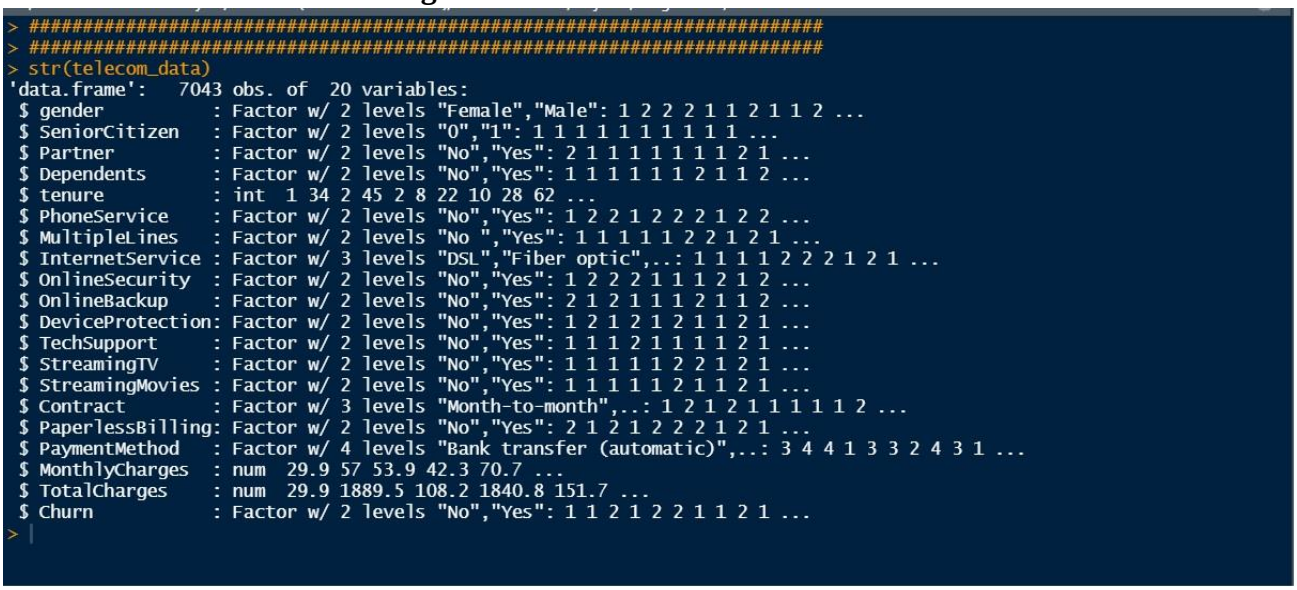


Fig: Structure of the Telecom Data after converting the variables to factors.

After converting the variables in the data set to factors and to get an insight use the **summary (Telecom_data)** to see the summary statistics. The data contains 49.52% of Females and50.47% Males. But this data set has not been split yet into training set and test set.

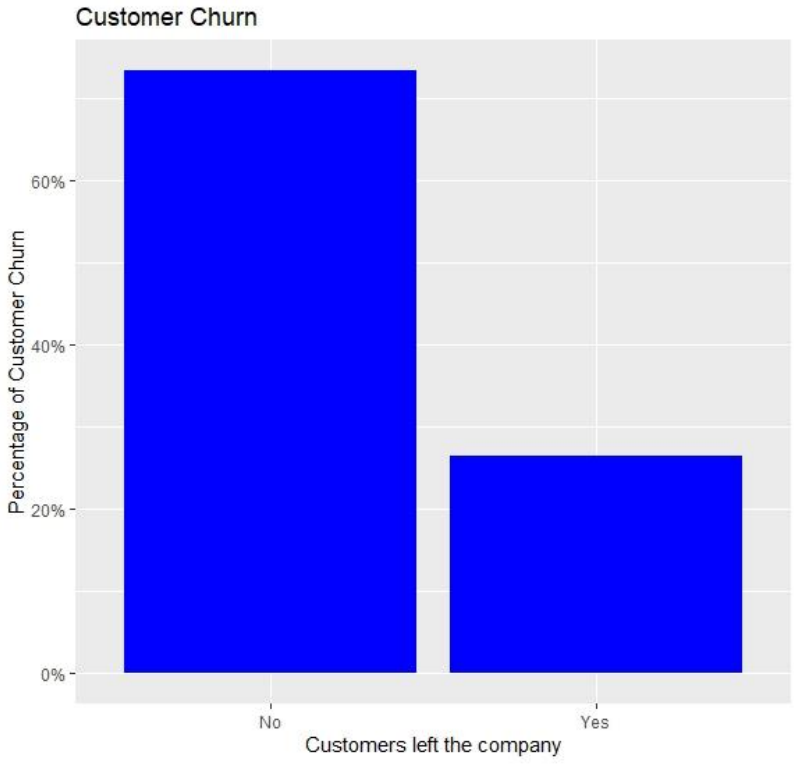
Training set is to create the model and test set is for testing the model for accuracy of predictions.

```
###creating test and training data sets######
set.seed(100)
train_ind <- telecom_data$Churn %>%
  createDataPartition(p = 0.70, list = FALSE) #splitting data into 70% for training and 30%  testing sets

Telecom_train <- telecom_data[train_ind,]
Telecom_test <- telecom_data[-train_ind,]

table(Telecom_train$Churn)

#### EXPLORING THE DATA SET ############
```

Fig: Splitting the data into Test set (30%) and Training set (70%)

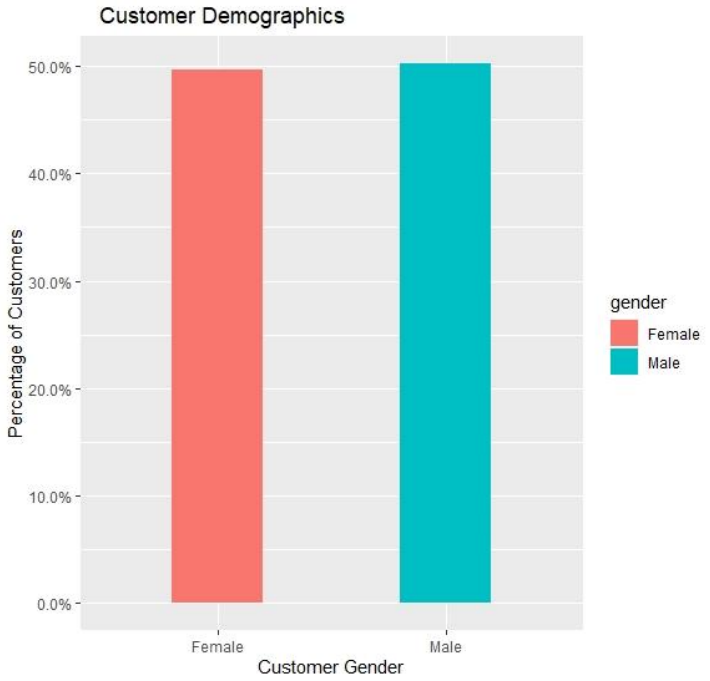
Creating plots to get the Insight from the data we get the following graphs.

1. **Customer Churn Ratio**



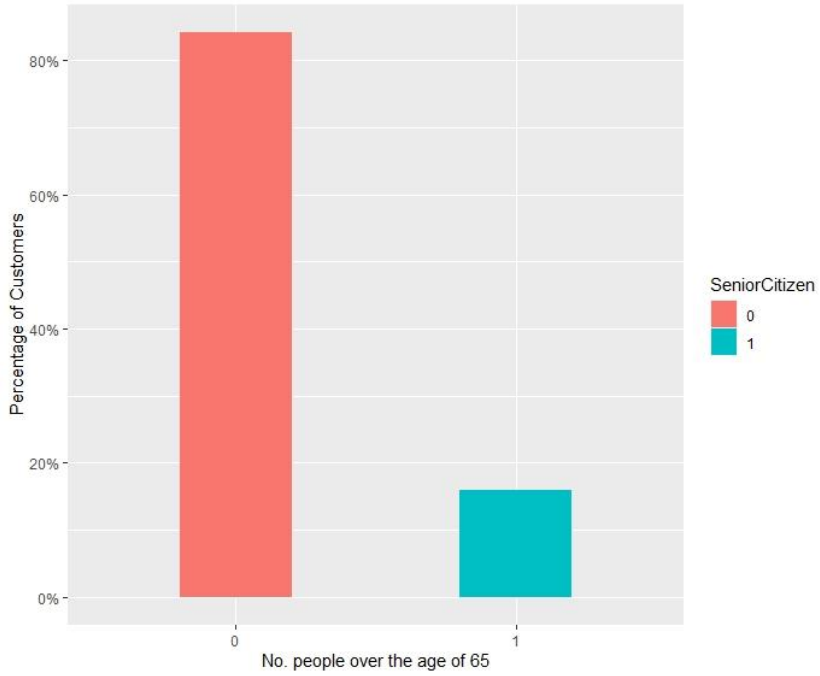
74% of the Customers did not leave the company but 27% customers did leave the company. There is a need of further investigation to figure out why these customers leave the company and which factors influenced these customers to change the company.

2. **No. of Male and female Customers in the company**



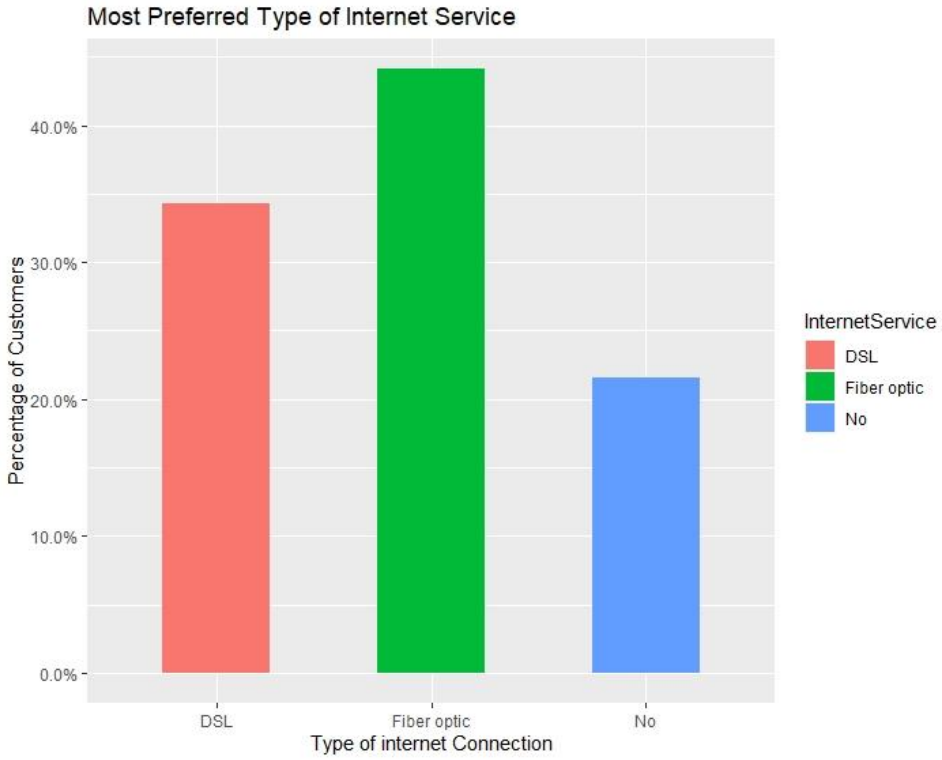
The data shows that there are 50% Males and 49% females Customers in the company.

3. **Customers over 65 years of age.**



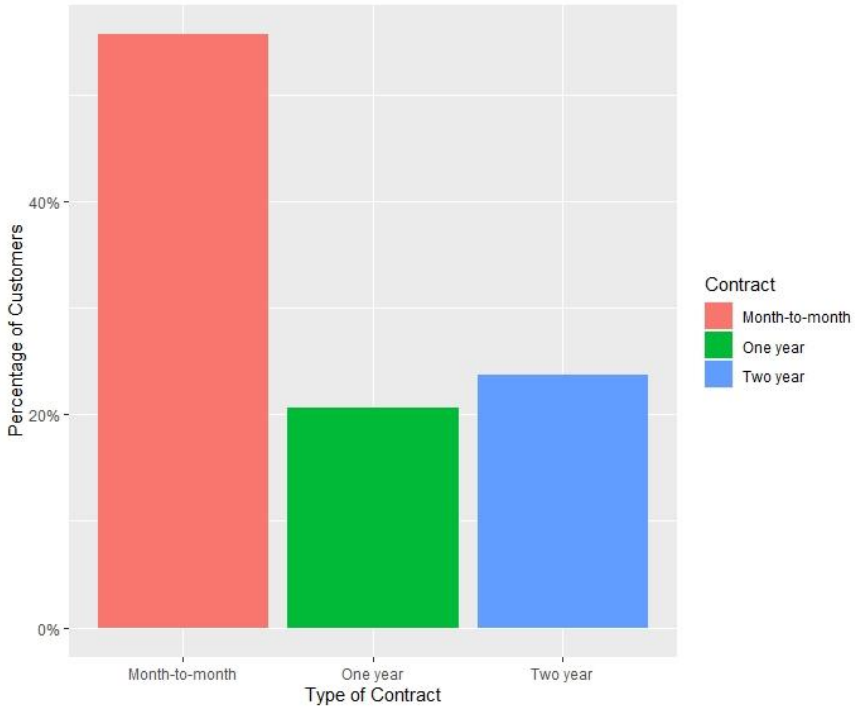
The data shows 85% of the customers of the company are below the age of 65 and 17% of the customers are senior Citizen.

4. **Most Preferred type of Internet Connection.**



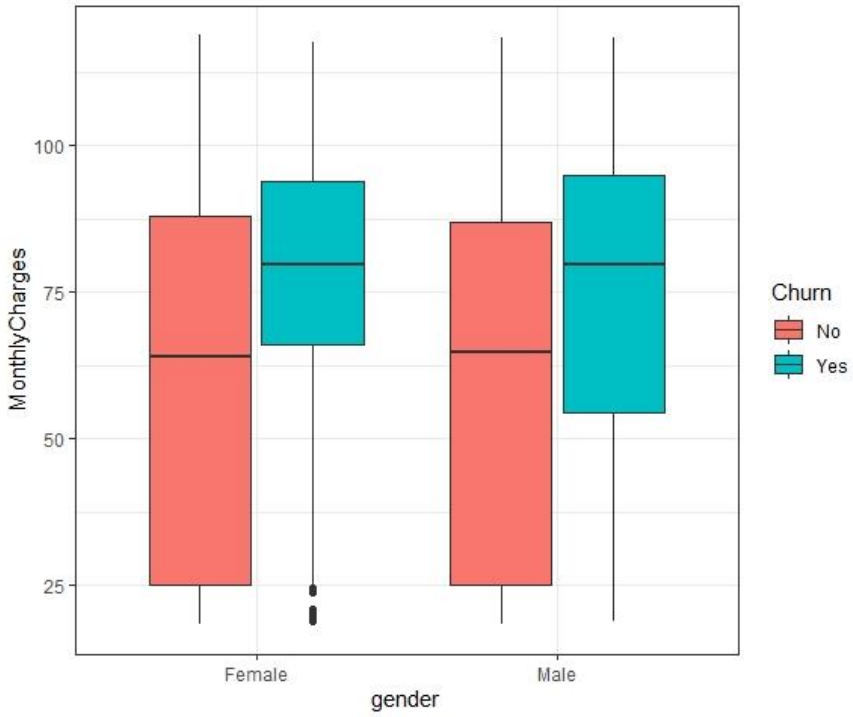
The most prefered internet connection type is Optic Fiber which is taken by 44 % and 34% have taken a DSL connection and 22% customer does not have an internent connection.

5. **Type of Contracts the Customer take.**



Most Customers are in a Monthly Contract accounting to 55% of the customers 21% Customers are on a one-year Contract and 24% Customers are on a two-year Contract.

6. **Gender Based Monthly Charges with Churn**



From the boxplot above we can see that median monthly charges for Customers who did not leave the company are same for both males and females. With first quartile Q1: 25$/month and third quartile: 87.5$/month and median charges being 62.5$/month. The difference is evident in Customers that have left the company where Males have churned more than females. Churned Males and females have a median monthly charge of 81.25$/month (approx.). Churned Males have a First quartile Q1: 56.25$/month (approx.) and males Q3:93.75$/month (approx.). Females have Q1: 64$/monthly (approx.) and Q2: Q3:93.75$/month.

Tests for independence Chi-square test and independent T test.

Ho: there is no relationship between churn and other variables

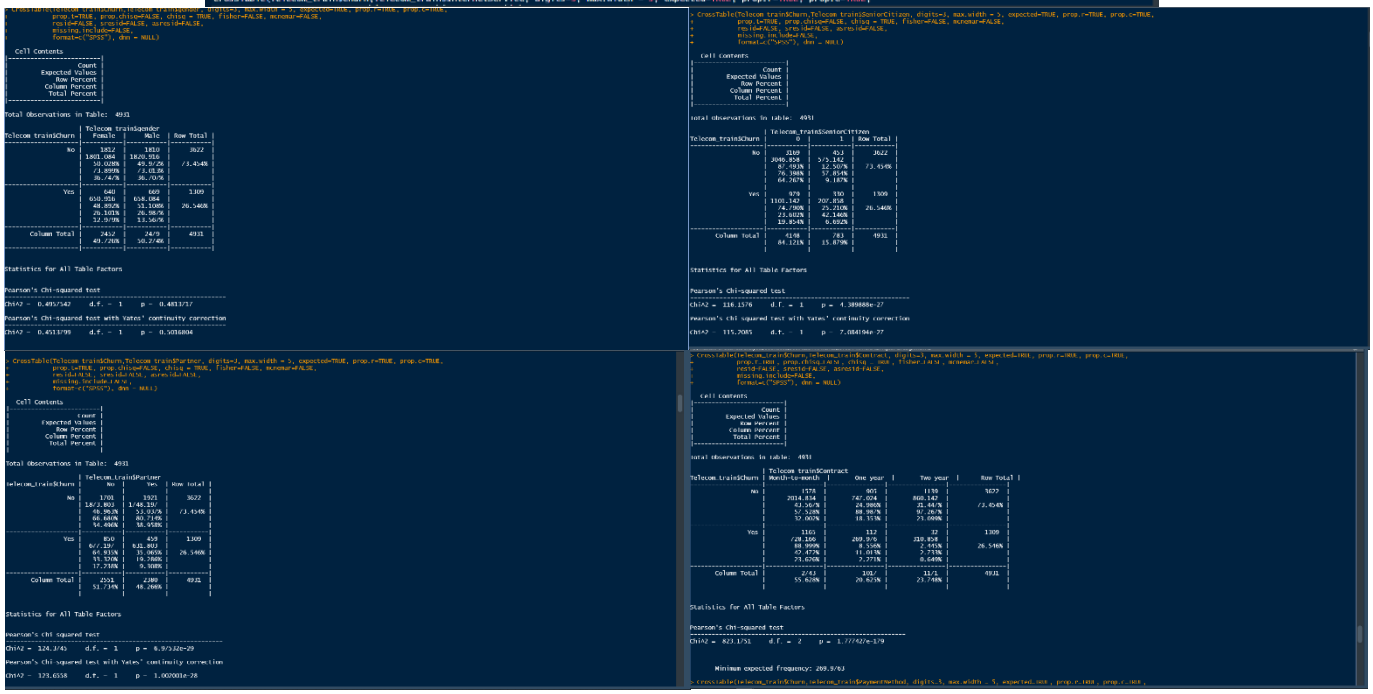
Ha: there is a level of significance between dependent variable and at least one predictor variable

We have used Cross tabulation which can perform Chi-square test, fisher's test and McNemar test for the nominal variables we have only conducted a Chi-square test. From the pre modelling analysis we compared the p-value at 95% confidence interval whether the variable would be significant or not.

Variables that should be able to explain customer chur: Senior Citizen, partner, Dependents, Multiple lines, Internet Service, Online security, Online backup, Device protection, tech support, Streaming TV, Streaming Movies, Contract, Payment method, paperless Billing, Tenure, Monthly Charges, Total Charges.

Variables that are not able to explain customer churn: Gender, Phone service,

```
##########################################################################
### conducting a CHI square test to find the relationship between dependents and Independents###
### Chi square test of independence#####
###this would tell us weather there is a relationship between two categorical variable###
### our target variable is Churned###
### Ho: there is no relationship between churn and other variables###
### Ha: there is a level of significance between both the variables###
### using CrossTable####
#### for all the variables below we will just consider the Chi square test as all the nominal have frequency more than 5 which the sufficient
##   to validate the Chi Square test
#### this test this not explain how strongly the variables are related but it only tells whether there is an association or not.

CrossTable(Telecom_train$Churn,Telecom_train$gender, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)


CrossTable(Telecom_train$Churn,Telecom_train$SeniorCitizen, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)

CrossTable(Telecom_train$Churn,Telecom_train$Partner, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)

CrossTable(Telecom_train$Churn,Telecom_train$Dependents, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)

CrossTable(Telecom_train$Churn,Telecom_train$PhoneService, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)

CrossTable(Telecom_train$Churn,Telecom_train$MultipleLines, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
           prop.t=TRUE, prop.chisq=FALSE, chisq = TRUE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,
           missing.include=FALSE,
           format=c("SPSS"), dnn = NULL)

CrossTable(Telecom_train$Churn,Telecom_train$InternetService, digits=4, max.width = 5, expected=TRUE, prop.r=TRUE, prop.c=TRUE,
```



These are the codes and results from the Chi-square tests.

There are the codes and the result from the T-tests for the Quantitative

```
######### independent T test for numerical data to find a relationship between the variables#####
###Considering equal varinces here###
t.test(Telecom_train$tenure~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level = 0.95,  var.equal = TRUE)
t.test(Telecom_train$MonthlyCharges~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level = 0.95,  var.equal = TRUE)
t.test(Telecom_train$TotalCharges~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level =0.95, var.equal= TRUE)
```

```
> ######### independent T test for numerical data to find a relationship between the variables#####
> ###Considering equal varinces here###
> t.test(Telecom_train$tenure~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level = 0.95,  var.equal = TRUE)

        Two Sample t-test

data:  Telecom_train$tenure by Telecom_train$Churn
t = 26.354, df = 4929, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.09922 21.00838
sample estimates:
 mean in group No mean in group Yes
         37.42546          17.87166

> t.test(Telecom_train$MonthlyCharges~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level = 0.95,  var.equal = TRUE)

        Two Sample t-test

data:  Telecom_train$MonthlyCharges by Telecom_train$Churn
t = -14.446, df = 4929, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.56133 -11.84248
sample estimates:
 mean in group No mean in group Yes
         61.09232          74.79423

> t.test(Telecom_train$TotalCharges~Telecom_train$Churn,mu=0,alternative = "two.sided",conf.level =0.95, var.equal= TRUE)

        Two Sample t-test

data:  Telecom_train$TotalCharges by Telecom_train$Churn
t = 13.973, df = 4929, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  861.6381 1142.8770
sample estimates:
 mean in group No mean in group Yes
         2532.561          1530.304
```

Multicollinearity and Variance Inflation Factor

For the model we checked for multicollinearity among various independent variables for that we fitted all the variables in the model to see the GVIF.

As test model has all the predictors fitted in the model, we can see that GVIF value for monthly charges is 711.15 which is very high and influences the value of Total charges. A VIF of 10 is considered very high. So, we removed the **monthly charges** which reduces the GVIF^(1/(2*df)) for internet services as it has 2 df and for Streaming movies and streaming TV. In test_model_1 the VIF values have reduced drastically for the all the other factors except Total Charges and Tenure. By removing **total charges** in test_model_2 we have reduced the **VIF for all the variables and are below 1.5 which is acceptable.**

```
#### checking Multicollinearity###
test_model <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
        + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod+MonthlyCharges+TotalCharges, data = Telecom_train, family = "binomial")
car::vif(test_model)
## there is high mulitcollinearity between tenure, Phone service, Internet Serivces ,Streaming TV ,Streaming Movies, Monthly Charges ,Total Charges
### after removing monthly charges
test_model1 <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
            + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod+TotalCharges, data = Telecom_train, family = "binomial")
vif(test_model1)
#### after removing Total Charges
test_model2 <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
                + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod, data = Telecom_train, family = "binomial")
vif(test_model2)
> #### checking Multicollinearity###
> test_model <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
+         + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod+MonthlyCharges+TotalCharges, data = Telecom_train, family = "binomial")
> car::vif(test_model)
                    GVIF Df GVIF^(1/(2*Df))
gender              1.006325  1        1.003158
SeniorCitizen       1.134401  1        1.065082
Partner             1.375777  1        1.172935
Dependents          1.278724  1        1.130807
tenure             16.507213  1        4.062907
PhoneService       37.231985  1        6.101802
MultipleLines       7.706332  1        2.776028
InternetService   391.508261  2        4.448209
OnlineSecurity      5.172779  1        2.274375
OnlineBackup        6.576393  1        2.564448
DeviceProtection    6.583015  1        2.565739
TechSupport         5.330912  1        2.308877
StreamingTV        25.702576  1        5.069771
StreamingMovies    25.798896  1        5.079261
Contract            1.597281  2        1.124205
PaperlessBilling    1.136041  1        1.065852
PaymentMethod       1.370824  3        1.053975
MonthlyCharges    711.155969  1       26.667508
TotalCharges       21.245269  1        4.609259
> ## there is high mulitcollinearity between tenure, Phone service, Internet Serivces ,Streaming TV ,Streaming Movies, Monthly Charges ,Total Charges
> ### after removing monthly charges
> test_model1 <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
+             + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod+TotalCharges, data = Telecom_train, family = "binomial")
> vif(test_model1)
                    GVIF Df GVIF^(1/(2*Df))
gender              1.005555  1        1.002773
SeniorCitizen       1.134302  1        1.065036
Partner             1.374382  1        1.172340
Dependents          1.278743  1        1.130815
tenure             16.376565  1        4.046797
PhoneService        1.511746  1        1.229531
MultipleLines       1.486504  1        1.219223
InternetService     2.674825  2        1.278863
OnlineSecurity      1.134103  1        1.064943
OnlineBackup        1.237223  1        1.112305
DeviceProtection    1.305530  1        1.142598
TechSupport         1.189810  1        1.090784
StreamingTV         1.540525  1        1.241179
StreamingMovies     1.541962  1        1.241757
Contract            1.596749  2        1.124111
PaperlessBilling    1.135806  1        1.065742
PaymentMethod       1.369428  3        1.053796
TotalCharges       21.088149  1        4.592183
> #### after removing Total Charges
> #### after removing Total Charges
> test_model2 <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+tenure+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup
+                 + DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod, data = Telecom_train, family = "binomial")
> vif(test_model2)
                    GVIF Df GVIF^(1/(2*Df))
gender              1.005396  1        1.002695
SeniorCitizen       1.135291  1        1.065501
Partner             1.368966  1        1.170028
Dependents          1.281906  1        1.132213
tenure              2.200108  1        1.483276
PhoneService        1.451105  1        1.204618
MultipleLines       1.435922  1        1.198300
InternetService     2.371567  2        1.240963
OnlineSecurity      1.118700  1        1.057686
OnlineBackup        1.188317  1        1.090100
DeviceProtection    1.268087  1        1.126094
TechSupport         1.171434  1        1.082328
StreamingTV         1.451887  1        1.204943
StreamingMovies     1.445030  1        1.202094
Contract            1.564387  2        1.118371
PaperlessBilling    1.133504  1        1.064662
PaymentMethod       1.345704  3        1.050731
>
```

Fig. VIF Models: Test model, Test_model1, Test_model2

4. ANALYTICS

In model 1 we have all the covariates but Monthly and total charges we removed to remove the multicollinearity.

Deviance is measure of goodness of fit for generalised liner model. **Null deviance** shows how well the target variable is predicted by the model which includes intercept.

Residual Deviance show how well the target variable is predicted by the covariates or the predictors including the intercept in the model. Lower the residual deviance better the variables explain the target variable in the model.

Akakine information criterion (AIC) is calculated from 2K – 2(L) where k is the independent variable and L is the loglikelihood of the model. Lower the AIC better the model fits the data. If the more variables are explaining the model the AIC would increase indicating an overfitting model.

McFadden's R square is given by the formula 1- deviance(model)/Null deviance. The r-Square gives an insight of the explanatory power of the model. Higher the better the model is explained by the covariant.

- Model 1 has a residual deviance of 4068.5 and an aic value of 4112.5 this means the model can be improved further by removing one variable based on the p-value. Device Protection has a p-value of 0.718068 which is highly insignificant in model at 5% significance level.
- Model 2 we have removed the Device protection; thus, we have a residual deviance of 4068.7 and an aic value of "4110.7". There is not much difference between the residual deviance of model 1 and model 2 so we are taking the second parameter aic value to choose the best fitted model. Partner has a p-value of 0.635687 which is highly insignificant.
- Model 3 we have removed partner and we have residual deviance of 4068.9 and an aic value of 4108.9 which means our model has improved drastically from model 1 to model5**. Payment Method** (Mailed check and electronic card) has a p-value of 0.483648, 0.376432 which are highly insignificant. To overcome this, we did data transformation using the code below. This would remove Mailed check and electronic card.

```
## our target is Electronic check
elecom_train$PaymentMethod <- ifelse(Telecom_train$PaymentMethod == "Electronic check" ,'yes','no')
elecom_test$PaymentMethod <- ifelse(Telecom_test$PaymentMethod == "Electronic check" ,'yes','no')
elecom_train$PaymentMethod <- as.factor(Telecom_train$PaymentMethod)
elecom_test$PaymentMethod<- as.factor(Telecom_test$PaymentMethod)
```

- Model 4 we get a residual deviance of 4069.8 but has an aic value of 4105.8 and is better model from the previous model has better predictors in the model. To improve the model the further we can remove gender.
- Model 5 after removing gender we get a residual deviance of 4070.5 and an aic 4104.5. If further remove the variables which is evident in model 6 , model7 and model8. These models tend to show overfitting. So, model 5 is our best model.

McFadden's R square

```
> McFaddenModel1
[1] 0.2871057
> McFaddenModel2
[1] 0.2870829
> McFaddenModel3
[1] 0.2870436
> McFaddenModel4
[1] 0.2868904
> McFaddenModel5
[1] 0.2867695
> McFaddenModel6
[1] 0.2862927
> McFaddenModel7
[1] 0.2857216
> McFaddenModel8
[1] 0.2849088
```

We can see that model5 and model4 are the best models with least number of variables and lowest aic values. **Model 5 is best model** with r2 value almost the same as model 4 but with a lower aic value.

Test for overall significance (Anova)

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             4930     5707.1
gender           1     0.50      4929     5706.6 0.4813553
SeniorCitizen    1   107.68      4928     5598.9 < 2.2e-16 ***
Partner          1   131.96      4927     5466.9 < 2.2e-16 ***
Dependents       1    32.86      4926     5434.1 9.900e-09 ***
tenure           1   571.79      4925     4862.3 < 2.2e-16 ***
PhoneService     1     1.16      4924     4861.1 0.2824317
MultipleLines    1   130.72      4923     4730.4 < 2.2e-16 ***
InternetService  2   466.91      4921     4263.5 < 2.2e-16 ***
OnlineSecurity   1    36.65      4920     4226.9 1.410e-09 ***
OnlineBackup     1     3.16      4919     4223.7 0.0755400 .
DeviceProtection 1     0.02      4918     4223.7 0.8866606
TechSupport      1    17.22      4917     4206.5 3.333e-05 ***
StreamingTV      1    20.51      4916     4186.0 5.939e-06 ***
StreamingMovies  1    12.15      4915     4173.8 0.0004921 ***
Contract         2    71.49      4913     4102.3 2.990e-16 ***
PaperlessBilling 1    13.77      4912     4088.5 0.0002066 ***
PaymentMethod    3    20.00      4909     4068.5 0.0001695 ***
```

Model 1

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             4930     5707.1
SeniorCitizen    1   107.70      4929     5599.4 < 2.2e-16 ***
Dependents       1   104.35      4928     5495.0 < 2.2e-16 ***
tenure           1   631.32      4927     4863.7 < 2.2e-16 ***
PhoneService     1     1.17      4926     4862.5 0.2796779
MultipleLines    1   131.09      4925     4731.4 < 2.2e-16 ***
InternetService  2   466.64      4923     4264.8 < 2.2e-16 ***
OnlineSecurity   1    37.23      4922     4227.6 1.051e-09 ***
OnlineBackup     1     3.25      4921     4224.3 0.0715147 .
TechSupport      1    16.84      4920     4207.5 4.059e-05 ***
StreamingTV      1    20.74      4919     4186.7 5.254e-06 ***
StreamingMovies  1    11.90      4918     4174.8 0.0005606 ***
Contract         2    71.81      4916     4103.0 2.555e-16 ***
PaperlessBilling 1    13.62      4915     4089.4 0.0002243 ***
PaymentMethod    1    18.96      4914     4070.5 1.338e-05 ***
---
```

Model5

In the test we can compare model1 with all the variables and model5 with best fitted variables. Model 5 has less insignificant variables which provide a better fir for the model.

Running the model, we get the Confusion matrix for test data:

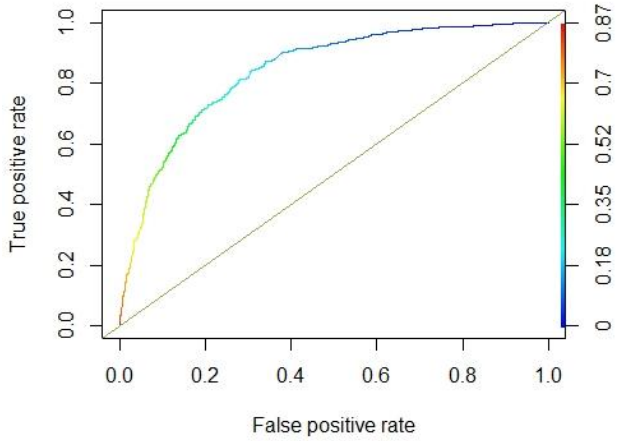
	False	True
NO	TN =1399	FP=153
Yes	FN=271	TP= 289

Precision of the model: TP / (TP + FP) = 0.6538

Recall of the model: TP/ (TP + FN) = 0.5160

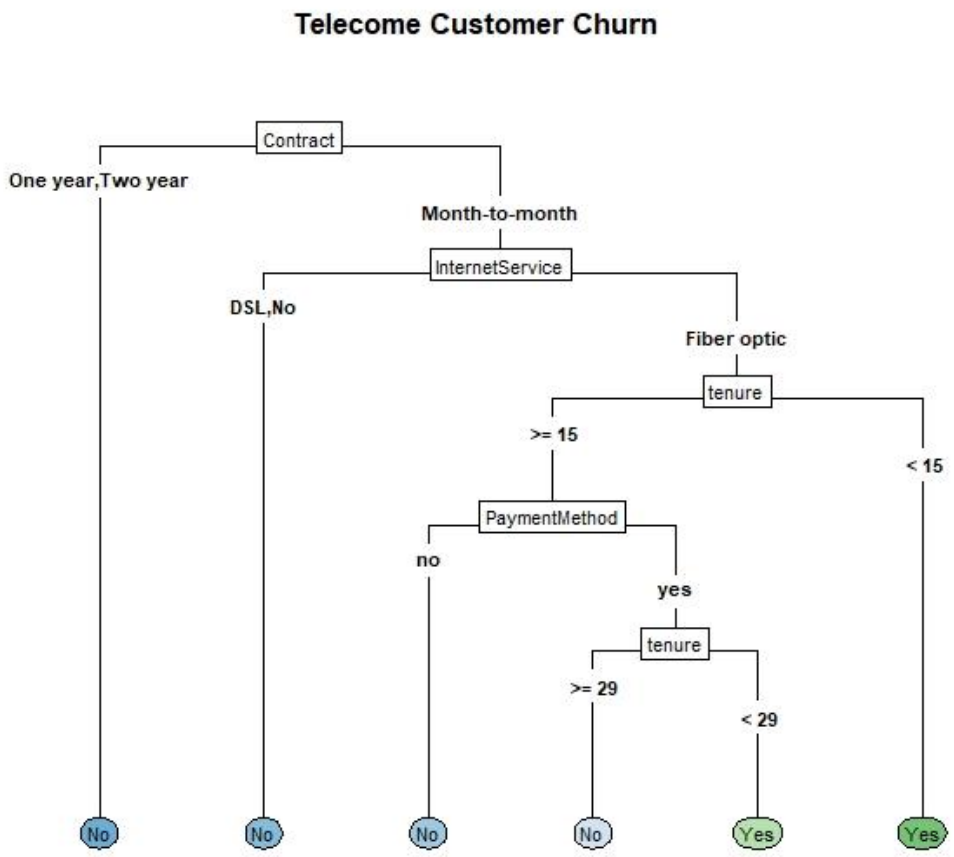
Accuracy of the model: (TP + TN) /(TP+FP+TN+FN) = 1399+289/ (1399+153+271+289) =0.7992

The model predicts the customer that have not left the company or have not Churn with 80% accuracy.65% of the corrected predicted cases turned out to be Customers who got churned or left the company. (positive cases) (precision).51% of the predicted positive cases which explains the customer who got churned or left the company were successfully predicted by the model. (recall)



The model has an Auc of 0.840 which is a good model. This means the model predicts the actual positive cases correctly and less no. of actual negative cases as positive. The model explains the customers who were churned or had yes for churn were predicted correctly than the customer which did not churn as churned customers (as false positive).

Decision Tree



We have split the variables based on the Gini coefficient. The models show how a decision is made whether the customer would leave the Flash telecom services or not. The decision tree shows us that the most important predictor of Customer Churn is the **contract (Root node).** If the customer is in one- or two-year contract, then it is less likely to churn **(fallen leaf, Churn = No).** If the customer is in Month-to Month contract, then it goes to the **internet service (Decision node).** If the **customer has a DSL or no internet service** then he would be less likely to be churned **(fallen leaf, Churn = No).** If Internet service is fibre optic, then it goes to **tenure (Decision node).** If the tenure of customer is less than 15 months then customer is likely to churned **(fallen leaf, Churn = Yes).** If the tenure of the customer is more than or equal to 15 months, then It moves to **payment method (Decision node).** If the Payment method is not "electronic check", then customer is less likely to churn **(fallen leaf, Churn = No).** If the payment method is "electronic check" then it branches out to **tenure (Decision node).** If the tenure is greater or equal to 29 months the customer is less likely to leave the company but if the tenure of customer is less than 29 months, then it is more likely the customer will leave the company. The overall accuracy for the decision tree is 80.9 %.

5.Recommendations & Conclusions

Our final model contains **Logit(P(Churn)) = (-0.26070) +0.188035(Senior Citizen)+(-0.171861) (Dependents Yes)+(-0.034270)(tenure) +(-0.528036)(Phone Service Yes) +(0.287644)(Multiple Lines Yes)+(0.975575)(Internet Service Fibre optic)+(-0.773786)(Internet Service No)+(-0.413111)(Online Security Yes)+(-0.150098)(Online Backup Yes)+(-0.289328)(Tech Support Yes)+(0.266155)(Streaming TV Yes)+(0.329091)(Streaming Movies Yes)+(-0.648072)(Contract One year)+(-1.372070)(Contract Two year)+(0.308419)(Paperless Billing Yes)+(0.362899)(Payment Method yes)**

- The chances of customer leaving the company is 43% without considering any predictors in the model.
- If a customer is a senior citizen(yes) with dependents(yes) and tenure (20 months) and has taken phone services(yes) but has no Multiple Lines but has Internet Service (Fibre optic) has taken Online Security and Online Backup and has Tech Support but does not Stream Tv and Movies and is in a two-year contract has not opted for (Paperless Billing) and does not use electronic check as payment method, **The chance of Customer leaving the company is 6%.**
- If a customer is a Senior citizen has no internet connection and is in a monthly-to-month contract and has been with company for 10 months and have taken paperless billing and payment method is electronic check. **There is 26% chance the customer will leave the company.**
  Please the see the R script for the Calculations.

From the analysis of above the suggestion for Flash Telecom Service should:

- Focus on converting their existing monthly customers to one- or two-years contract and longer the contract less likely the customer would leave the company.
- Consider segmented marketing and build a Direct relationship with each of their customers to provide them with customised services based on their needs.
- Implement a customer loyalty program to increase customer spend and to attract new customers. So, if a customer signs up for an internet connection along with the Phone service they get points which are redeemable for discounts on the bill.
- Offer Tv and movie streaming services as promotional for new and existing customers would reduce chances of customer leaving the company as they will not have to pay additional charges for purchasing the service from a third party.
- Focus on Customers with dependents, as they are less likely to leave the company (Churn). To retain them for a longer period in the company, they should be provided with better offers based on their feedback.
- Promote customers to sign-up for their services like online security and Online backup as these tend to reduce the like-hood of customer churn.

Hence, we can conclude that the classifiers for the model can have an affect to reduce the Churn of the customers in the company.

6. APPENDICES

Data pre-processing

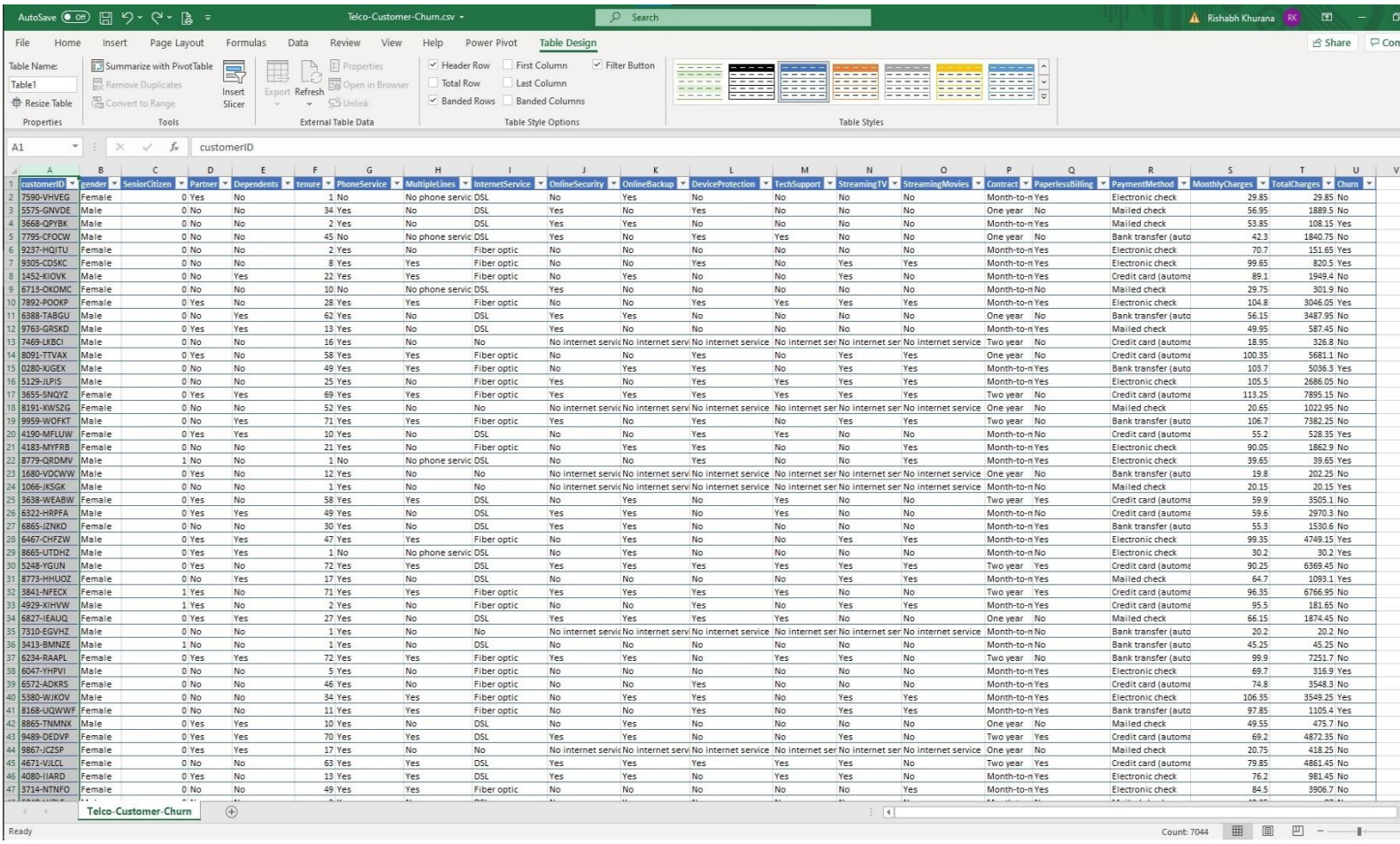


Fig1: Removing unwanted data from the file.

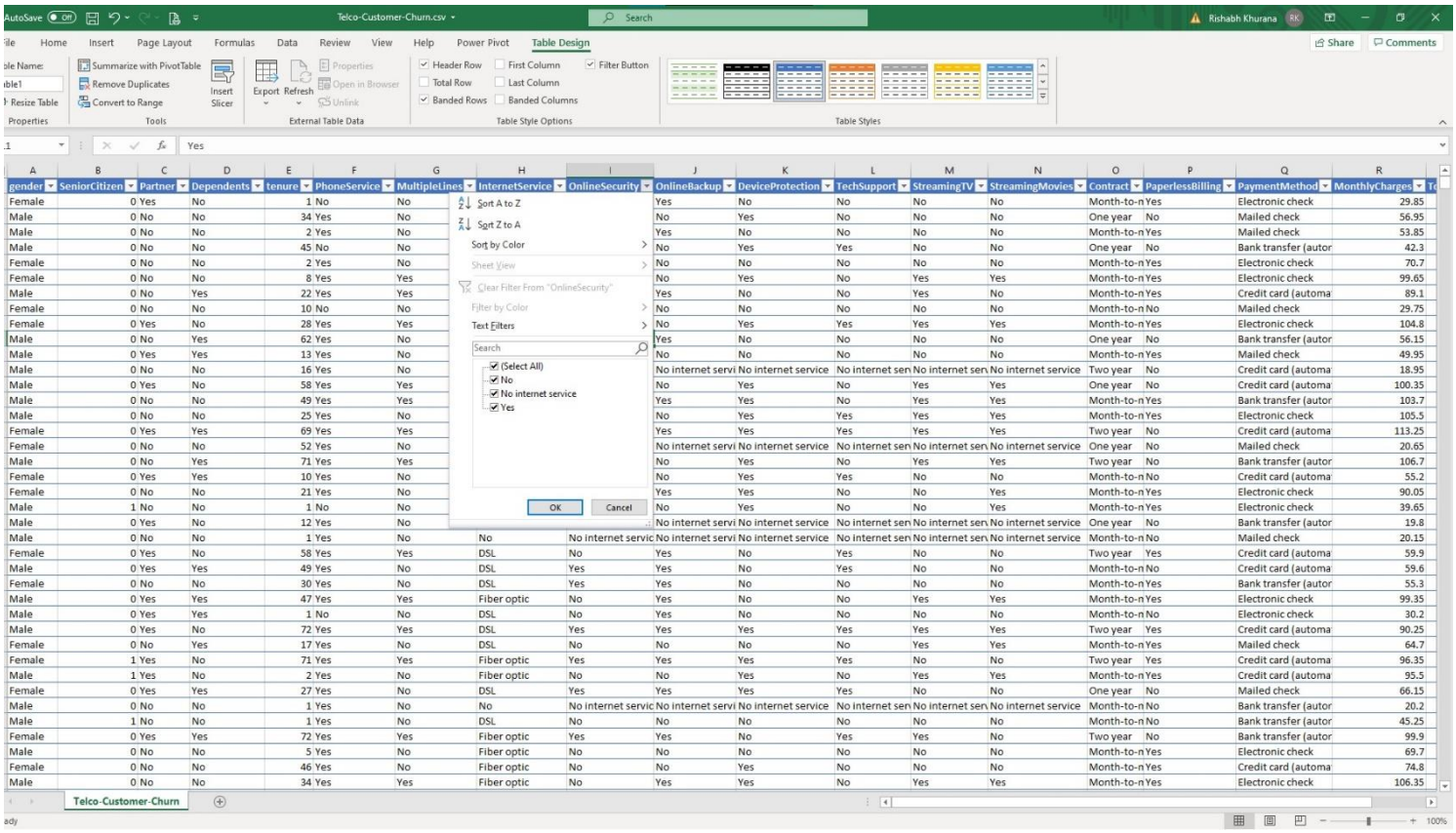


Fig 2: Converting the repeated values in the column for each factor and normalising the data.

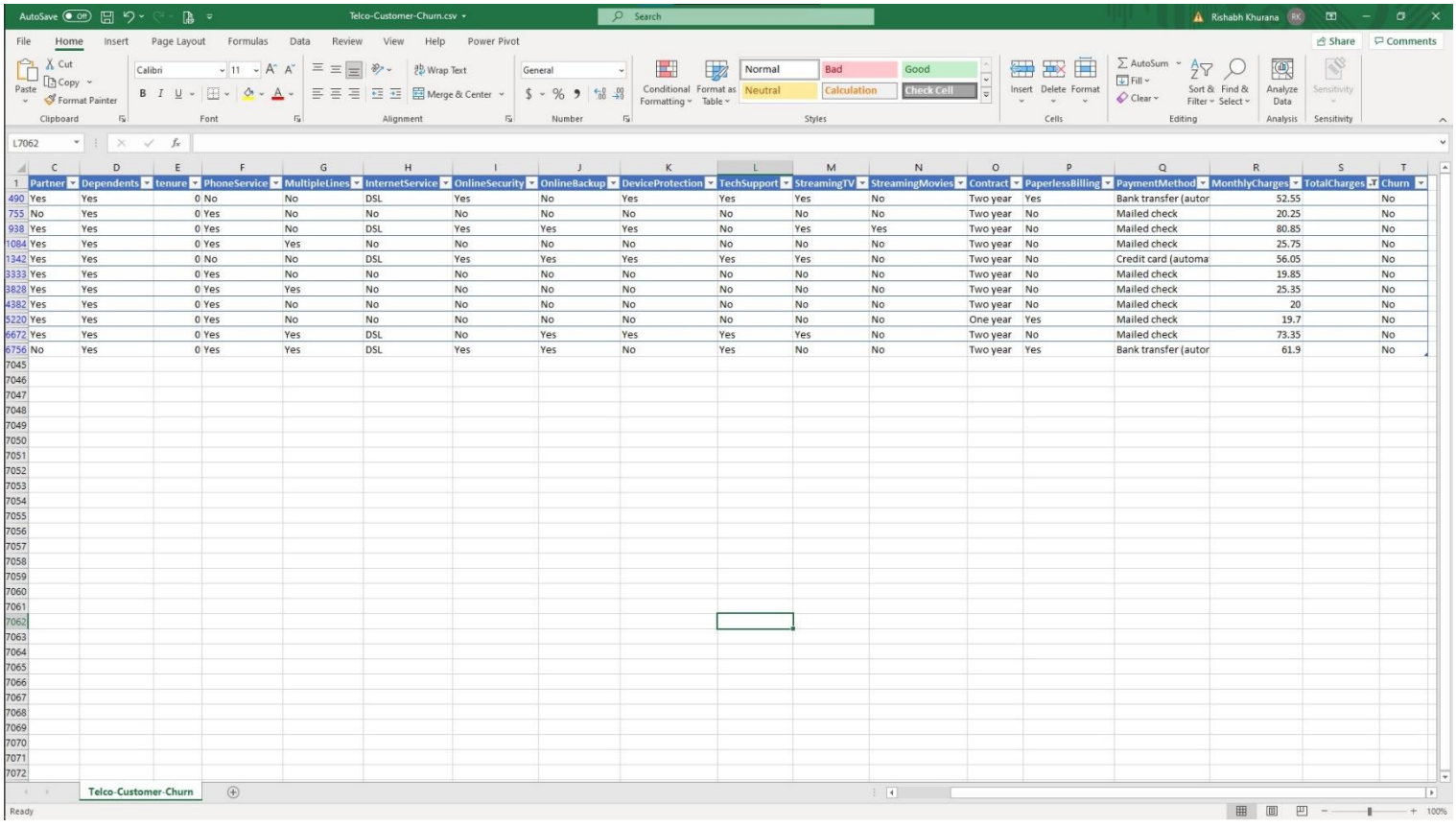


Fig 3: Converting the blank values in total charges columns based on monthly charges.

LOGISTIC MODELS

```
Call:
glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
    tenure + PhoneService + MultipleLines + InternetService +
    OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod, family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0027  -0.6660  -0.2893   0.6908   3.1885

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.212473   0.180807  -1.175 0.239940
genderMale                       0.065980   0.077938   0.847 0.397237
SeniorCitizen1                   0.190557   0.102021   1.868 0.061788 .
PartnerYes                      -0.045226   0.093259  -0.485 0.627716
DependentsYes                   -0.149000   0.106942  -1.393 0.163534
tenure                          -0.034391   0.002836 -12.125  < 2e-16 ***
PhoneServiceYes                 -0.534199   0.151180  -3.534 0.000410 ***
MultipleLinesYes                 0.289647   0.094385   3.069 0.002149 **
InternetServiceFiber optic       0.976658   0.110477   8.840  < 2e-16 ***
InternetServiceNo               -0.755801   0.167685  -4.507 6.57e-06 ***
OnlineSecurityYes               -0.407980   0.101779  -4.008 6.11e-05 ***
OnlineBackupYes                 -0.148072   0.091024  -1.627 0.103795
DeviceProtectionYes              0.034137   0.094552   0.361 0.718068
TechSupportYes                  -0.291579   0.103350  -2.821 0.004783 **
StreamingTVYes                   0.261976   0.095205   2.752 0.005928 **
StreamingMoviesYes               0.324160   0.094820   3.419 0.000629 ***
ContractOne year                -0.649880   0.130133  -4.994 5.92e-07 ***
ContractTwo year                -1.376206   0.212257  -6.484 8.95e-11 ***
PaperlessBillingYes              0.311234   0.089544   3.476 0.000509 ***
PaymentMethodCredit card (automatic) -0.125683   0.138002  -0.911 0.362436
PaymentMethodElectronic check    0.290421   0.113836   2.551 0.010735 *
PaymentMethodMailed check       -0.099808   0.138001  -0.723 0.469533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4068.5  on 4909  degrees of freedom
AIC: 4112.5

Number of Fisher Scoring iterations: 6
```

Fig: Model 1

```
Call:
glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
    tenure + PhoneService + MultipleLines + InternetService +
    OnlineSecurity + OnlineBackup + TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
    family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0076  -0.6660  -0.2893   0.6890   3.1922

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.208644   0.180489  -1.156 0.247685
genderMale                       0.066344   0.077933   0.851 0.394611
SeniorCitizen1                   0.191218   0.101998   1.875 0.060832 .
PartnerYes                      -0.044154   0.093204  -0.474 0.635687
DependentsYes                   -0.149731   0.106910  -1.401 0.161356
tenure                          -0.034242   0.002806 -12.205  < 2e-16 ***
PhoneServiceYes                 -0.534660   0.151194  -3.536 0.000406 ***
MultipleLinesYes                 0.289801   0.094380   3.071 0.002136 **
InternetServiceFiber optic       0.976695   0.110483   8.840  < 2e-16 ***
InternetServiceNo               -0.762157   0.166754  -4.571 4.86e-06 ***
OnlineSecurityYes               -0.408170   0.101790  -4.010 6.07e-05 ***
OnlineBackupYes                 -0.147646   0.091022  -1.622 0.104783
TechSupportYes                  -0.288908   0.103079  -2.803 0.005066 **
StreamingTVYes                   0.266128   0.094519   2.816 0.004869 **
StreamingMoviesYes               0.327499   0.094380   3.470 0.000520 ***
ContractOne year                -0.646688   0.129814  -4.982 6.30e-07 ***
ContractTwo year                -1.371922   0.211919  -6.474 9.56e-11 ***
PaperlessBillingYes              0.310538   0.089526   3.469 0.000523 ***
PaymentMethodCredit card (automatic) -0.125322   0.137980  -0.908 0.363743
PaymentMethodElectronic check    0.290099   0.113828   2.549 0.010817 *
PaymentMethodMailed check       -0.099621   0.137984  -0.722 0.470308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4068.7  on 4910  degrees of freedom
AIC: 4110.7

Number of Fisher Scoring iterations: 6
```

Fig: Model2

```
Call:
glm(formula = Churn ~ gender + SeniorCitizen + Dependents + tenure +
    PhoneService + MultipleLines + InternetService + OnlineSecurity +
    OnlineBackup + TechSupport + StreamingTV + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod, family = "binomial",
    data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0016  -0.6664  -0.2894   0.6902   3.1851

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.214394   0.180085  -1.191 0.233845
genderMale                       0.065741   0.077920   0.844 0.398837
SeniorCitizen1                   0.186005   0.101393   1.834 0.066582 .
DependentsYes                   -0.170607   0.097402  -1.752 0.079847 .
tenure                          -0.034502   0.002753 -12.535  < 2e-16 ***
PhoneServiceYes                 -0.533768   0.151198  -3.530 0.000415 ***
MultipleLinesYes                 0.288912   0.094354   3.062 0.002199 **
InternetServiceFiber optic       0.975130   0.110415   8.832  < 2e-16 ***
InternetServiceNo               -0.763121   0.166746  -4.577 4.73e-06 ***
OnlineSecurityYes               -0.409494   0.101758  -4.024 5.72e-05 ***
OnlineBackupYes                 -0.148114   0.091019  -1.627 0.103675
TechSupportYes                  -0.289026   0.103077  -2.804 0.005048 **
StreamingTVYes                   0.265085   0.094479   2.806 0.005020 **
StreamingMoviesYes               0.327016   0.094370   3.465 0.000530 ***
ContractOne year                -0.647173   0.129801  -4.986 6.17e-07 ***
ContractTwo year                -1.371435   0.211895  -6.472 9.66e-11 ***
PaperlessBillingYes              0.310732   0.089521   3.471 0.000518 ***
PaymentMethodCredit card (automatic) -0.121874   0.137790  -0.884 0.376432
PaymentMethodElectronic check    0.291376   0.113792   2.561 0.010449 *
PaymentMethodMailed check       -0.096533   0.137817  -0.700 0.483648
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4068.9  on 4911  degrees of freedom
AIC: 4108.9

Number of Fisher Scoring iterations: 6
```

Fig: Model 3

```
Call:
glm(formula = Churn ~ gender + SeniorCitizen + Dependents + tenure +
    PhoneService + MultipleLines + InternetService + OnlineSecurity +
    OnlineBackup + TechSupport + StreamingTV + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod, family = "binomial",
    data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0021  -0.6663  -0.2888   0.6910   3.1678

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.294822   0.154847  -1.904 0.056915 .
genderMale                  0.064658   0.077869   0.830 0.406346
SeniorCitizen1              0.188120   0.101347   1.856 0.063426 .
DependentsYes              -0.171849   0.097377  -1.765 0.077602 .
tenure                     -0.034334   0.002715 -12.645  < 2e-16 ***
PhoneServiceYes            -0.529738   0.151069  -3.507 0.000454 ***
MultipleLinesYes            0.287762   0.094300   3.052 0.002276 **
InternetServiceFiber optic  0.977883   0.110033   8.887  < 2e-16 ***
InternetServiceNo          -0.771256   0.165300  -4.666 3.07e-06 ***
OnlineSecurityYes          -0.409859   0.101740  -4.029 5.61e-05 ***
OnlineBackupYes            -0.148284   0.091017  -1.629 0.103273
TechSupportYes             -0.290003   0.103050  -2.814 0.004890 **
StreamingTVYes              0.267276   0.094464   2.829 0.004664 **
StreamingMoviesYes          0.328349   0.094349   3.480 0.000501 ***
ContractOne year           -0.647623   0.129774  -4.990 6.03e-07 ***
ContractTwo year           -1.370762   0.211863  -6.470 9.80e-11 ***
PaperlessBillingYes         0.309945   0.089401   3.467 0.000527 ***
PaymentMethodyes            0.362922   0.082997   4.373 1.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4069.8  on 4913  degrees of freedom
AIC: 4105.8

Number of Fisher Scoring iterations: 6
```

Fig: Model 4

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
    TechSupport + StreamingTV + StreamingMovies + Contract +
    PaperlessBilling + PaymentMethod, family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9877  -0.6667  -0.2916   0.6944   3.1564

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.260770   0.149355  -1.746 0.080816 .
SeniorCitizen1             0.188035   0.101358   1.855 0.063573 .
DependentsYes             -0.171861   0.097386  -1.765 0.077607 .
tenure                    -0.034270   0.002713 -12.632  < 2e-16 ***
PhoneServiceYes           -0.528036   0.151079  -3.495 0.000474 ***
MultipleLinesYes           0.287644   0.094286   3.051 0.002283 **
InternetServiceFiber optic 0.975575   0.109965   8.872  < 2e-16 ***
InternetServiceNo         -0.773786   0.165264  -4.682 2.84e-06 ***
OnlineSecurityYes         -0.413111   0.101668  -4.063 4.84e-05 ***
OnlineBackupYes           -0.150098   0.090963  -1.650 0.098924 .
TechSupportYes            -0.289328   0.103043  -2.808 0.004988 **
StreamingTVYes             0.266155   0.094471   2.817 0.004843 **
StreamingMoviesYes         0.329091   0.094357   3.488 0.000487 ***
ContractOne year          -0.648072   0.129765  -4.994 5.91e-07 ***
ContractTwo year          -1.372070   0.211827  -6.477 9.34e-11 ***
PaperlessBillingYes        0.308419   0.089380   3.451 0.000559 ***
PaymentMethodyes           0.362899   0.082995   4.373 1.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4070.5  on 4914  degrees of freedom
AIC: 4104.5

Number of Fisher Scoring iterations: 6
```

Fig: Model 5

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
    MultipleLines + InternetService + OnlineSecurity + +TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod, family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9737  -0.6657  -0.2892   0.6926   3.1894

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.283942   0.148667  -1.910 0.056143 .
SeniorCitizen1             0.185802   0.101245   1.835 0.066482 .
DependentsYes             -0.175029   0.097300  -1.799 0.072042 .
tenure                    -0.035338   0.002638 -13.396  < 2e-16 ***
PhoneServiceYes           -0.524220   0.150986  -3.472 0.000517 ***
MultipleLinesYes           0.285297   0.094234   3.028 0.002466 **
InternetServiceFiber optic 0.974759   0.109894   8.870  < 2e-16 ***
InternetServiceNo         -0.738031   0.163860  -4.504 6.67e-06 ***
OnlineSecurityYes         -0.414556   0.101568  -4.082 4.47e-05 ***
TechSupportYes            -0.297727   0.102877  -2.894 0.003804 **
StreamingTVYes             0.267111   0.094413   2.829 0.004667 **
StreamingMoviesYes         0.329108   0.094290   3.490 0.000482 ***
ContractOne year          -0.649847   0.129727  -5.009 5.46e-07 ***
ContractTwo year          -1.371195   0.211741  -6.476 9.43e-11 ***
PaperlessBillingYes        0.303341   0.089313   3.396 0.000683 ***
PaymentMethodyes           0.362569   0.082959   4.370 1.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4073.2  on 4915  degrees of freedom
AIC: 4105.2

Number of Fisher Scoring iterations: 6
```

Fig: Model 6

```
Call:
glm(formula = Churn ~ SeniorCitizen + tenure + PhoneService +
    MultipleLines + InternetService + OnlineSecurity + +TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod, family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9748  -0.6654  -0.2916   0.6995   3.2168

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.32162    0.14707  -2.187 0.028755 *
SeniorCitizen1              0.21564    0.10000   2.156 0.031055 *
tenure                     -0.03566    0.00263 -13.560  < 2e-16 ***
PhoneServiceYes            -0.53181    0.15077  -3.527 0.000420 ***
MultipleLinesYes            0.28648    0.09416   3.042 0.002346 **
InternetServiceFiber optic  0.98152    0.10977   8.942  < 2e-16 ***
InternetServiceNo          -0.74495    0.16375  -4.549 5.38e-06 ***
OnlineSecurityYes          -0.41900    0.10150  -4.128 3.66e-05 ***
TechSupportYes             -0.30106    0.10285  -2.927 0.003419 **
StreamingTVYes              0.26191    0.09429   2.778 0.005476 **
StreamingMoviesYes          0.33400    0.09418   3.546 0.000391 ***
ContractOne year           -0.66124    0.12947  -5.107 3.27e-07 ***
ContractTwo year           -1.38518    0.21135  -6.554 5.61e-11 ***
PaperlessBillingYes         0.31015    0.08921   3.477 0.000508 ***
PaymentMethodyes            0.36703    0.08290   4.427 9.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4076.4  on 4916  degrees of freedom
AIC: 4106.4

Number of Fisher Scoring iterations: 6
```

Fig: Model 7

```
Call:
glm(formula = Churn ~ tenure + PhoneService + MultipleLines +
    InternetService + OnlineSecurity + +TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod,
    family = "binomial", data = Telecom_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9095  -0.6683  -0.2900   0.7108   3.2144

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.289396   0.146109  -1.981 0.047628 *
tenure                     -0.035146   0.002615 -13.442  < 2e-16 ***
PhoneServiceYes            -0.552722   0.150326  -3.677 0.000236 ***
MultipleLinesYes            0.296623   0.093950   3.157 0.001593 **
InternetServiceFiber optic  1.007755   0.109119   9.235  < 2e-16 ***
InternetServiceNo          -0.756594   0.163615  -4.624 3.76e-06 ***
OnlineSecurityYes          -0.426488   0.101431  -4.205 2.61e-05 ***
TechSupportYes             -0.314401   0.102578  -3.065 0.002177 **
StreamingTVYes              0.259427   0.094217   2.754 0.005896 **
StreamingMoviesYes          0.335787   0.094127   3.567 0.000361 ***
ContractOne year           -0.682725   0.129049  -5.290 1.22e-07 ***
ContractTwo year           -1.413914   0.210915  -6.704 2.03e-11 ***
PaperlessBillingYes         0.316921   0.089112   3.556 0.000376 ***
PaymentMethodyes            0.377867   0.082674   4.571 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5707.1  on 4930  degrees of freedom
Residual deviance: 4081.1  on 4917  degrees of freedom
AIC: 4109.1

Number of Fisher Scoring iterations: 6
```

Fig: Model 8

7. REFERENCES

- HFAQ: HOW DO I INTERPRET ODDS RATIOS IN LOGISTIC REGRESSION? (n.d.). Retrieved May 10, 2021, from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/
- Practical Guide to Logistic Regression Analysis in R. (2020, August 20). Retrieved May 10, 2021, from https://www.hackerearth.com/blog/developers/practical-guide-logistic-regression-analysis-r/
- Kassambara, Thanos, Kassambara, & Sfd. (2018, March 11). Logistic Regression Essentials in R. Retrieved May 10, 2021, from http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/
- Interpreting Generalized Linear Models. (2018, November 09). Retrieved May 10, 2021, from https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/
- Editor, M. B. (n.d.). Enough Is Enough! Handling Multicollinearity in Regression Analysis. Retrieved May 10, 2021, from https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis
- Taralova, V. (2021). *Census Data Project – Part 3: Predictive Analysis*. Rstudio-pubs-static.s3.amazonaws.com. Retrieved 9 May 2021, from https://rstudio-pubs-static.s3.amazonaws.com/265202_278cf395891a4f55be22a9ac7c0ac9c3.html#1_introduction.
- Says:, S. A., Says:, V., Says:, S., Says:, M. E., Says:, D., Says:, B. K., . . . Says:, U. (2017, September 13). Basic evaluation measures from the confusion matrix. Retrieved May 10, 2021, from https://classeval.wordpress.com/introduction/basic-evaluation-measures
- Matt ReichenbachMatt Reichenbach 3, ChrisChris 70366 silver badges88 bronze badges, Jonathan BartlettJonathan Bartlett 66544 silver badges1010 bronze badges, & YiranYiran 6111 silver badge11 bronze badge. (1962, October 01). McFadden's Pseudo-R^2 Interpretation. Retrieved May 10, 2021, from https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation
- Aniruddha BhandariI am on a journey to becoming a data scientist. I love to unravel trends in data. (2020, April 17). Confusion Matrix for Machine Learning. Retrieved May 10, 2021, from https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/
- Steen, D. (2020, September 14). Understanding the ROC Curve and AUC. Retrieved May 10, 2021, from https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb
- Colman, R. (2018, September 14). Interpretation of the AUC: R-bloggers. Retrieved May 10, 2021, from https://www.r-bloggers.com/2018/09/interpretation-of-the-auc/
- Dalpiaz, D. (2020, October 28). R for Statistical Learning. Retrieved May 10, 2021, from https://daviddalpiaz.github.io/r4sl/logistic-regression.html#logistic-regression-with-glm
- Brownlee, J. (2019, August 12). Overfitting and Underfitting With Machine Learning Algorithms. Retrieved May 10, 2021, from https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
- Bhalla, D. (n.d.). Decision Tree in R : Step by Step Guide. Retrieved May 10, 2021, from https://www.listendata.com/2015/04/decision-tree-in-r.html

- Sid100158, Shuvayan, Abhishek1608, & Joe222. (2016, January 02). What are the packages required to plot a fancy Rpart plot in R. Retrieved May 10, 2021, from https://discuss.analyticsvidhya.com/t/what-are-the-packages-required-to-plot-a-fancy-rpart-plot-in-r/6776
- Rulphus, A. (2018, June 20). Decision Tree Rpart() Summary Interpretation. Retrieved May 10, 2021, from https://community.rstudio.com/t/decision-tree-rpart-summary-interpretation/9996
- Bevans, R. (2020, March 27). Akaike Information Criterion: When & How to Use It. Retrieved May 10, 2021, from https://www.scribbr.com/statistics/akaike-information-criterion/#:~:text=The AIC function is 2K,it is being compared to.
- Blog, B. (2018, April 13). How do I interpret the AIC: R-bloggers. Retrieved May 10, 2021, from https://www.r-bloggers.com/2018/04/how-do-i-interpret-the-aic/