Telco Churn Analysis

- Data-Processing
  - The dataset has 7043 rows and 21 columns
  - 11 rows have missing values and they constitute to 0.1% of the dataset [7/7043]
  - Since they are negligible they were removed from the dataset
  - Now the dataset has 7032 rows
  - I am splitting the data train [70%] and test [30%]

- Attributing Importance to variables

```
> str(train)
'data.frame':   4931 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 5536 6552 1003 4771 5605 4535 6872 5752 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 1 2 2 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 2 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 2 1 ...
 $ tenure          : int  1 34 45 8 22 10 28 62 13 58 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 2 3 3 2 3 1 1 3 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 2 2 1 2 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 1 1 3 1 3 3 1 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 1 1 3 1 1 3 1 1 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 3 1 1 3 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 3 1 1 1 3 1 1 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 3 1 3 1 1 3 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 3 1 1 3 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 2 1 1 1 2 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 1 3 2 4 3 1 4 2 ...
 $ MonthlyCharges  : num  29.9 57 42.3 99.7 89.1 ...
 $ TotalCharges    : num  29.9 1889.5 1840.8 820.5 1949.4 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
```

  - Churn is our dependent variable.
  - Remaining 20 columns are predictor variables
  - CustomerID column has random unique numbers and is irrelevant to our analysis therefore it can be ignored.
  - Variables with 3 levels
    - MultipleLines – OnlineSecurity-Onlinebackup-Deviceprotection-TechSupport-StreamingTV-StreamingMovies
    - Can get reduced to 2 levels if efficient content mapping is done.
    - When "No internet service" / "No phone service" is replaced by "No" 3 levels get reduced to 2 levels without changing the meaning in the above mentioned 7 variables.
  - According to R Senior Citizen Datatype is int
    - Senior Citizen column can be changed to a factor with 2 levels
    - 0 = No 1= Yes
  - Tenure column datatype is also a int
    - Range of tenure column is 0 thru 72 Months
    - They can be grouped by year
  - There are two num variables in the dataset [TotalCharges & MonthlyCharges]
    - Corrplot shows that they are correlated
    - So one can be removed
  - Following Screenshot shows the modified dataset

```
> str(train)
'data.frame':   4931 obs. of  19 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 1 2 2 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 2 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 2 1 ...
 $ tenure          : Factor w/ 6 levels "0 thru 12","13 thru 24",..: 1 3 4 1 2 1 3 6 2 5 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 2 3 3 2 3 1 1 3 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 2 2 1 2 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 1 1 3 1 3 3 1 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 1 1 3 1 1 3 1 1 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 3 1 1 3 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 3 1 3 1 1 3 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 3 1 1 3 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 2 1 1 1 1 2 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 1 3 2 4 3 1 4 2 ...
 $ MonthlyCharges  : num  29.9 57 42.3 99.7 89.1 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 1 ...
```

- Logistic Regression
    - Since Churn is a categorical variable logistic regression will help us understand the importance of each column and their weightage in contributing to Churn

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                              0.21899    0.98381   0.223   0.8238
genderMale                              -0.05947    0.07771  -0.765   0.4441
SeniorCitizenYes                         0.18462    0.10264   1.799   0.0721 .
PartnerYes                              -0.04003    0.09299  -0.431   0.6668
DependentsYes                           -0.10308    0.10757  -0.958   0.3379
tenure13 thru 24                        -0.87683    0.11546  -7.594 3.10e-14 ***
tenure24 thru 36                        -1.32026    0.13981  -9.443  < 2e-16 ***
tenure37 thru 48                        -1.34217    0.15829  -8.479  < 2e-16 ***
tenure49 thru 60                        -1.49152    0.16850  -8.852  < 2e-16 ***
tenure61 thru 72                        -1.78883    0.20572  -8.696  < 2e-16 ***
PhoneServiceYes                         -0.18262    0.78756  -0.232   0.8166
MultipleLinesYes                         0.29438    0.21466   1.371   0.1703
InternetServiceFiber optic               1.38275    0.97074   1.424   0.1543
InternetServiceNo                       -1.21218    0.98147  -1.235   0.2168
OnlineSecurityYes                       -0.25975    0.21834  -1.190   0.2342
OnlineBackupYes                         -0.08717    0.21212  -0.411   0.6811
DeviceProtectionYes                      0.11378    0.21443   0.531   0.5957
TechSupportYes                          -0.29215    0.21772  -1.342   0.1797
StreamingTVYes                           0.38199    0.39798   0.960   0.3371
StreamingMoviesYes                       0.47670    0.39567   1.205   0.2283
ContractOne year                        -0.70994    0.12925  -5.493 3.96e-08 ***
ContractTwo year                        -1.49546    0.20901  -7.155 8.38e-13 ***
PaperlessBillingYes                      0.38717    0.08939   4.331 1.48e-05 ***
PaymentMethodCredit card (automatic)    -0.16430    0.13624  -1.206   0.2278
PaymentMethodElectronic check            0.27621    0.11149   2.477   0.0132 *
PaymentMethodMailed check               -0.02339    0.13598  -0.172   0.8634
MonthlyCharges                          -0.01649    0.03858  -0.427   0.6691
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Columns with statistical significance are
    - PaymentMethod – Contract - paperlessBilling – tenure

```
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                              4922     5694.0
gender            1     2.28      4921     5691.7 0.1308246
SeniorCitizen     1    98.75      4920     5593.0 < 2.2e-16 ***
Partner           1   118.58      4919     5474.4 < 2.2e-16 ***
Dependents        1    33.64      4918     5440.8 6.643e-09 ***
tenure            5   556.48      4913     4884.3 < 2.2e-16 ***
PhoneService      1     1.03      4912     4883.2 0.3100353
MultipleLines     1   106.84      4911     4776.4 < 2.2e-16 ***
InternetService   2   466.21      4909     4310.2 < 2.2e-16 ***
OnlineSecurity    1    33.08      4908     4277.1 8.829e-09 ***
OnlineBackup      1     3.74      4907     4273.4 0.0531918 .
DeviceProtection  1     0.43      4906     4272.9 0.5118753
TechSupport       1    28.64      4905     4244.3 8.715e-08 ***
StreamingTV       1    13.83      4904     4230.5 0.0001997 ***
StreamingMovies   1    11.09      4903     4219.4 0.0008690 ***
Contract          2    84.23      4901     4135.2 < 2.2e-16 ***
PaperlessBilling  1    20.40      4900     4114.8 6.297e-06 ***
PaymentMethod     3    17.57      4897     4097.2 0.0005397 ***
MonthlyCharges    1     0.18      4896     4097.0 0.6690575
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Tenure – MultipleLines – InternetService have worthy impact on decreasing the deviance
- Testing Logistic Regression with Test Data

```
> Fit_Output <- predict(Logistic_regression,newdata=test,type='response')
> fitted.results <- ifelse(Fit_Output > 0.5,1,0)
> Fit_Output <- ifelse(Fit_Output > 0.5,1,0)
> mis_Classification <- mean(Fit_Output != test$Churn)
> print(paste('Logistic Regression Accuracy',1-mis_Classification))
[1] "Logistic Regression Accuracy 0.807017543859649"
>
```

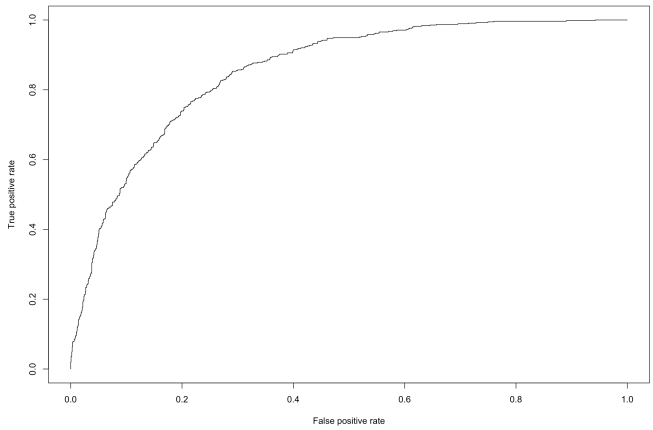
- Logistic Regression Confusion Matrix

```
table(test$Churn, fitted.results > 0.5)

  FALSE TRUE
0  1415  142
1   265  287
```

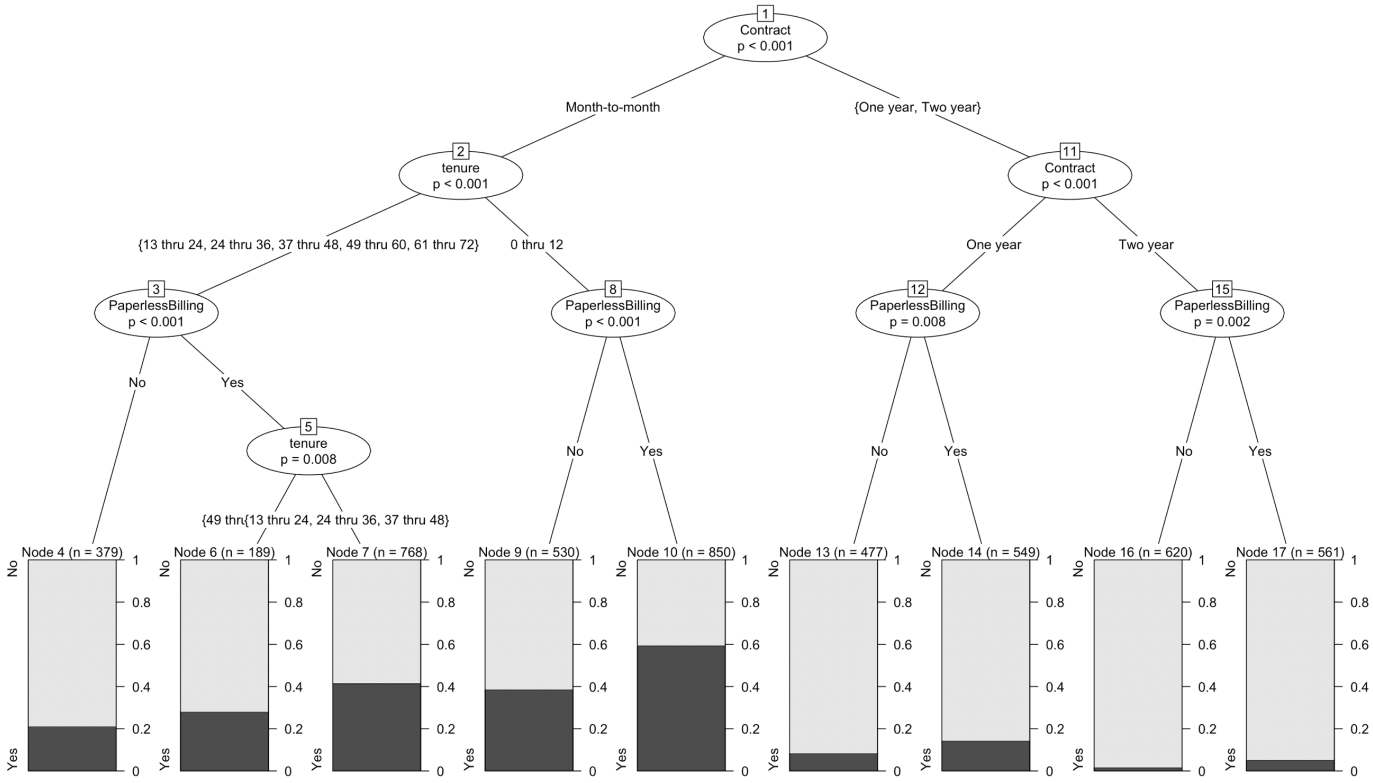
- Accuracy = (TP+TN)/Total = 287+1415/(287+1415+265+142)
  - .80
- Misclassification Rate = (FP+FN)/Total = 142+265/(287+1415+265+142)
  - .19
- Specificity =TN/actual no=1415/(1415+142)
  - .90
- Roc Curve

```
> Fit_Output <- predict(Logistic_regression,newdata=test,type='response')
> pr<-prediction(Fit_Output,test$Churn)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.8538275
>
```



- Roc Curve close to 1 is a great sign

- Decision Tree



- Contract column plays a vital role in predicting Churn
- Folks who are on long term and receive bill in paper are less likely to Churn
- Folks on short term contracts [month-month basis] are more likely to churn

- Decision Tree Confusion Matrix

```
> predicting_tree <- predict(decision_tree, test)
> table(Predicted = predicting_tree, Actual = test$Churn)
         Actual
Predicted    0    1
      No  1404  323
      Yes  153  229
> 
```

- Decision Tree Accuracy

```
> ptree<-predict(decision_tree,test)
> ptree_train<-predict(decision_tree,train)
> tree1<-table(Predicted=ptree_train,Actual=train$Churn)
> tree2<-table(Predicted=ptree,Actual=test$Churn)
> sum(diag(tree2)/sum(tree2))
[1] 0.7743006
>
```

- Proactive incentives has to be offered to folks who are on Month-Month basis
    - Because they are more likely to Churn