CUSTOMER CHURN PREDICTION

Customer churn occurs when customers or subscribers stop doing business with a company or service. Customer churn impedes growth, so companies should have a defined method for calculating customer churn in a given period.

Project Report by
Ji Wu
Ram Swaroop

CUSTOMER CHURN PREDICTION

INDEX

CUSTOMER CHURN PREDICTION

INTRODUCTION

- **GOAL** - The goal of the project is to predict the customer churn by analyzing appropriate customer data.
- **About the Dataset** - The dataset is of a telecom company which has
  Customer data like their age, gender, number of dependents.
  The services used by the customer like phone, internet, tech support.
  Customers who left within the last month (churn).

- **Dataset Overview -** The dataset has 7043 rows and 21 columns.

  The variables are:

  customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity,OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges,

  Churn is a categorical variable with Yes or No value.
  MonthlyCharges and TotalCharges are Numerical variables.
  Rest of the variables are categorical.

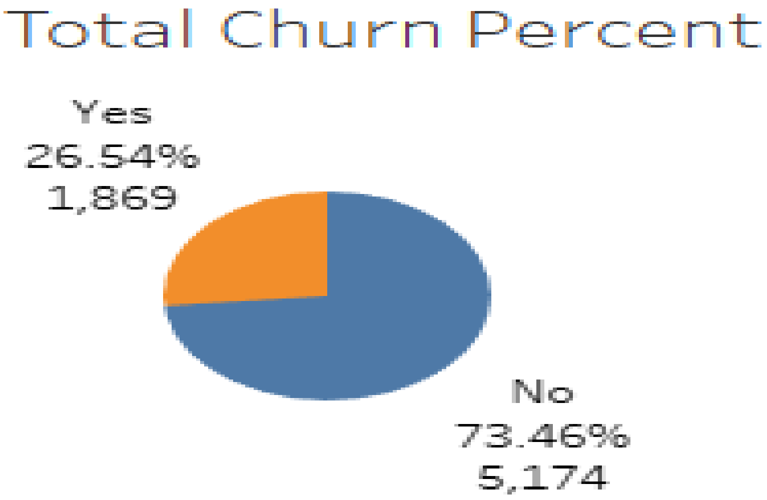
- **Approach** – Churn is categorical variable with two values (Yes or No) , we use classification models like logistic regression and Random Forest to predict the class of the variable.

- **Software and Tools** – Python 3 has been used for Data cleaning and modelling. Tableau has been used for Exploratory data analysis and R studio has been used for developing few plots.

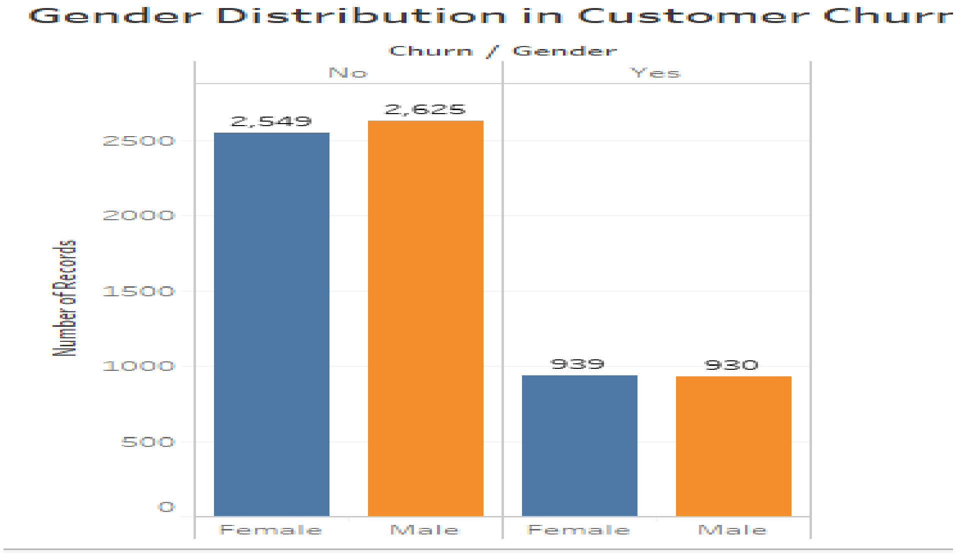
CUSTOMER CHURN PREDICTION

Exploratory Data Analysis

- There are total of 7043 customers.
- Out of which 1869 (26.54%) have left the service.

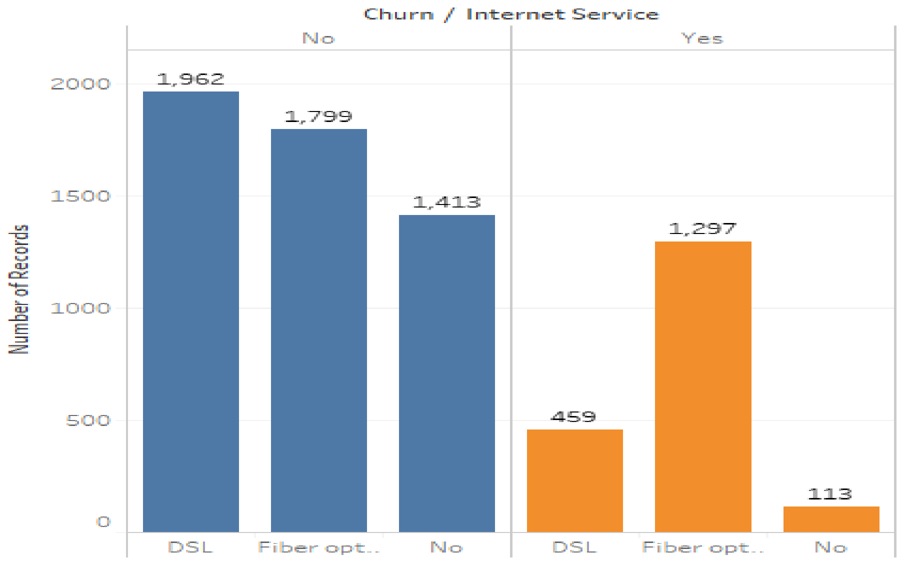


- The customer churn is almost uniform with respect to gender.
- Out of 1869 churn customers 930 are male and 939 are female.

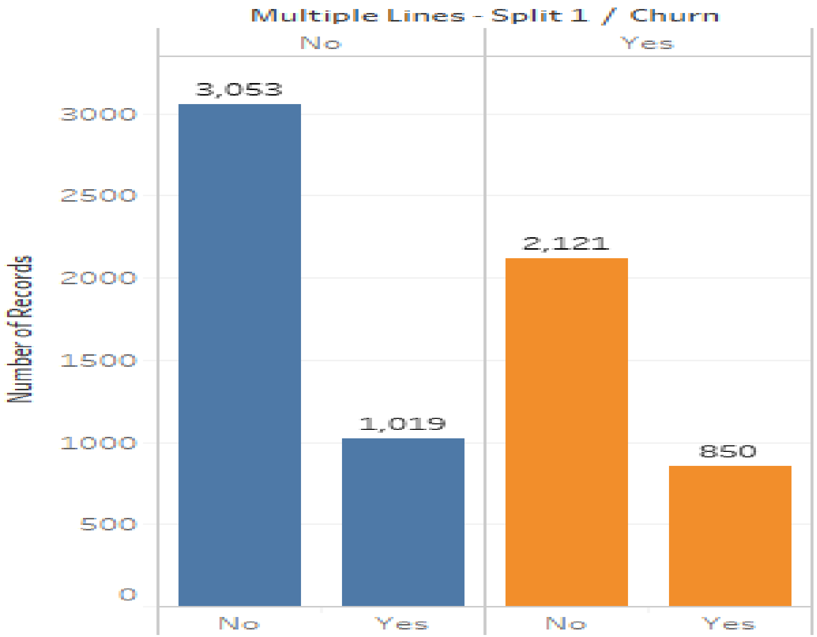


Exploratory Data Analysis

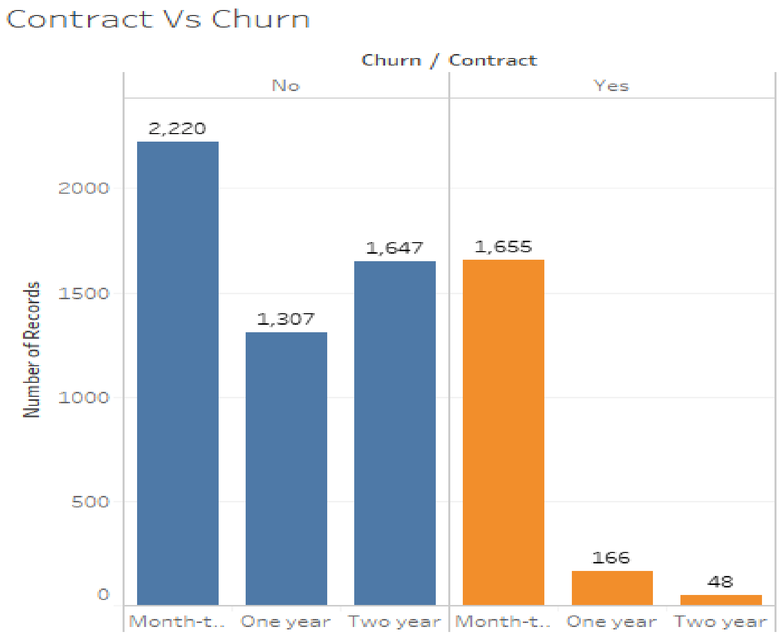
Internet Service Vs Churn Rate

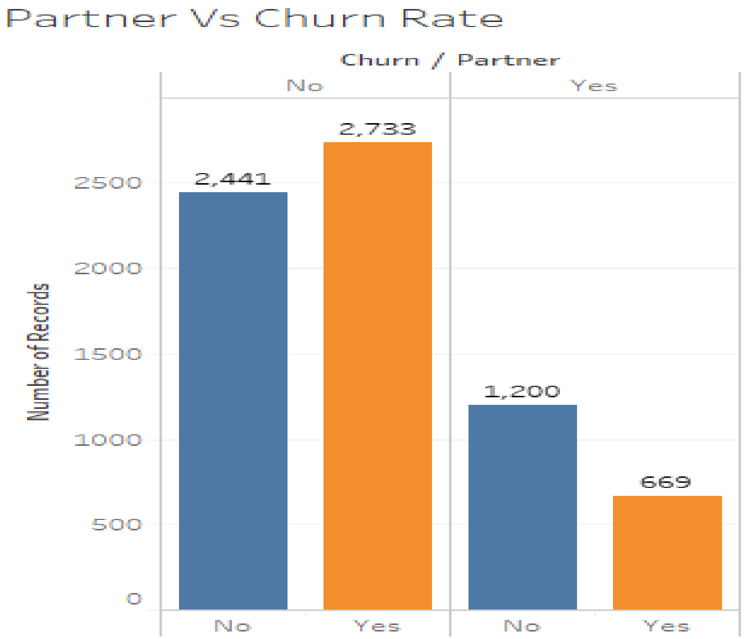


Multiple Lines Vs Churn



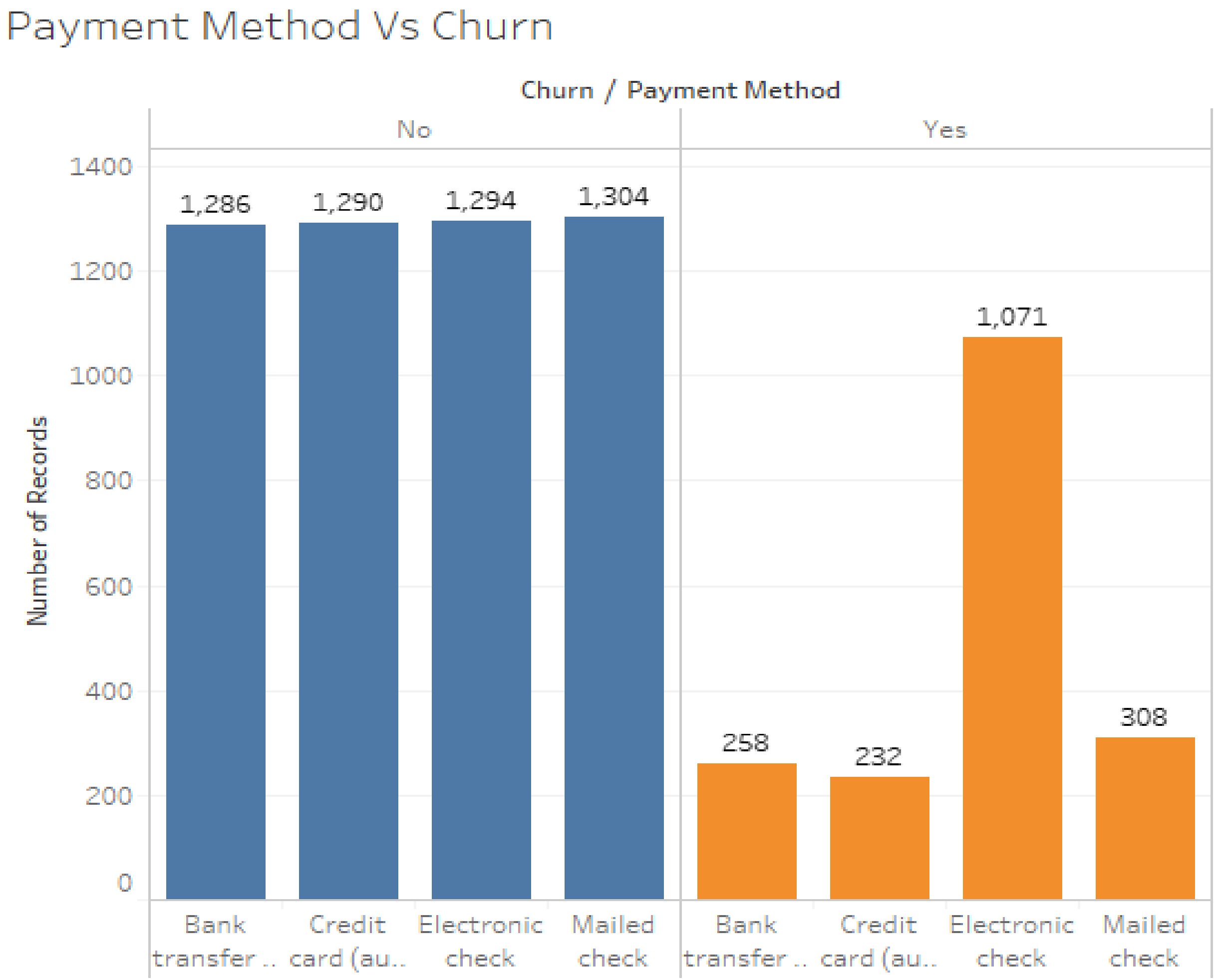
Exploratory Data Analysis





Exploratory Data Analysis

- More than 1000 people who used Electronic check have left the service.



CUSTOMER CHURN PREDICTION

Data Cleaning and Pre-Processing

- Data cleaning involves following steps:
  - To check if any missing values in the dataset.
  - Encoding the categorical variables with 2 or multiple values.
- Data Pre – Processing involves following steps:
  - Scaling the numerical values.
  - Dividing the dataset into dependent and independent variables.
  - Splitting the dataset into Train and Test.

To check if any missing values in the dataset:

- df_isnull = df.isnull().values
- df_isnull is a Boolean with True and False values. If there are Null values the value would be True, else value is false.

To replace missing values in TotalCharges with the mean

- There are no Nan values in the dataset. However there is a “ “ character for few rows in the TotalCharges column.
- The TotalCharges is first replaced with Nan and then Nan is replaced with the mean of the TotalCharges column.
- df['TotalCharges'] = df["TotalCharges"].replace(" ",np.nan)
df["TotalCharges"] = df["TotalCharges"].astype(float)
df['TotalCharges’]=df['TotalCharges'].fillna((df['TotalCharges'].mean()))

Data Cleaning and Pre-Processing

Assigning the variables into their respective category

- MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies have 3 categories each.
- Yes, No and No service.
- No service is equal to No. Hence No service should be replaced with No.
- Replacing No service with No.
  df['StreamingTV'] = df['StreamingTV'].replace(['No internet   service'], 'No’)
- All the variables have been replaced using the same method.

To Map the Variables

- Using the map function to map Yes and No values to 1 and 0
  d = {'Yes':1 , 'No': 0}
  df['Churn'].map(d)
- The above method is used for all the categorical variables with 2 values.

Encoding the categorical variables

- Encoding the columns with two or more values.
- Label Encoder class from sklearn library is used for columns with two categories.
- Pd.get_dummies method is used for columns with more than two categories.

CUSTOMER CHURN PREDICTION

Data Cleaning and Pre-Processing

Scaling the numerical columns

- To compare two variables ‘Total charges’ and ‘Monthly charges’ which are measured different scales.
- The variables are normalized into a range from -1 to +1.
- Scaling is done using Standard Scaler available in sklearn library.
- from sklearn.preprocessing import StandardScaler
  sc = StandardScaler()

Dividing the columns into dependent and independent variable

- Churn is the dependent variable and rest of the variables are independent.
- y is the dependent variable (churn).
- X had all the independent variables.
- To divide x and y from dataset
  y = df.iloc[:, 13].values
  x= df.drop('Churn’,1)

Splitting the data into Test and Train

- 80 % of the data is taken for training the model and rest 20% is used for testing.
- from sklearn.model_selection import train_test_split
  x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)

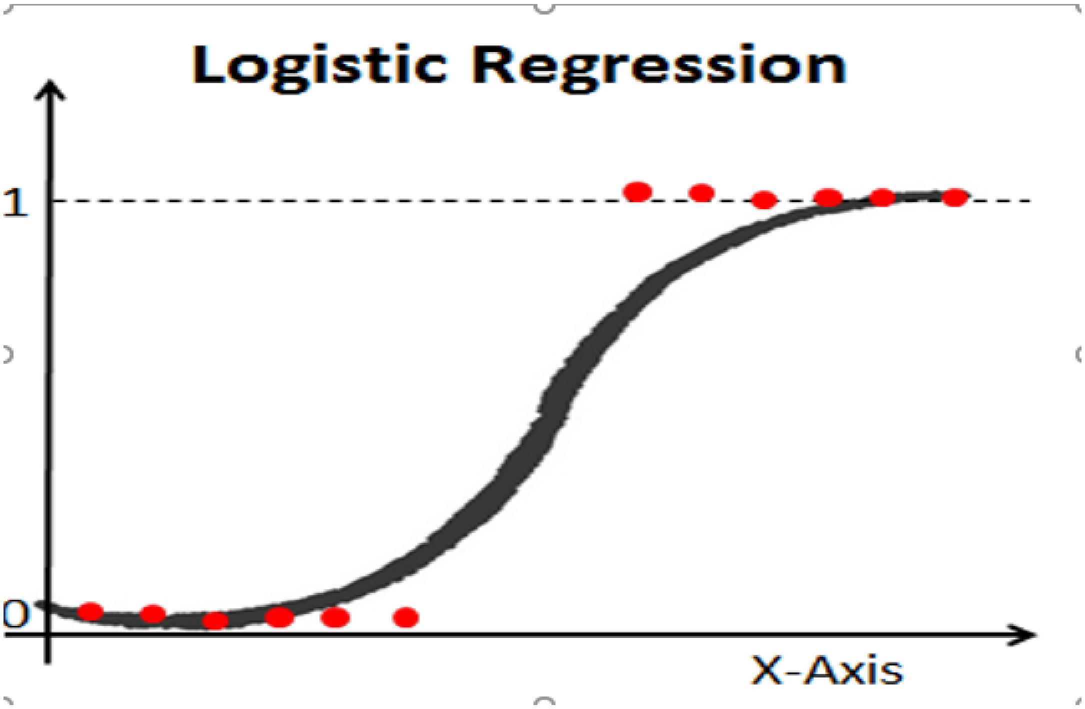
CUSTOMER CHURN PREDICTION

CLASSIFICATION

- Classification is the process of predicting a class of a datapoint.
- In our project customer churn is to be predicted which has two classes(Yes or No). Hence this is a binary classification problem.
- We have used Logistic Regression, Decision Tree and Random Forest models to classify the customer churn

Data Modelling

Logistic Regression



- The curve is the fit to the data to predict the probability a customer would churn or not.
- The Y axis is the probability of the dependent variable from 0 to 1.
- The X axis are the independent variables.
- The Y axis in logistic regression is transformed from the probability of churn to the log (odds of churn).
- Now Y axis can go from -Infinity to +Infinity (Similar to Regression).

- This is done using the logit function. Given by log(p/1-p).
- Maximum likelihood is used to find the best fitting line.

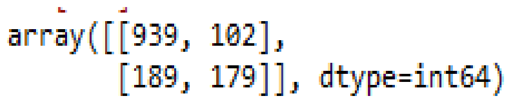
Data Modelling

Fitting the Model

- The code for logistic regression model
- from sklearn.linear_model import LogisticRegression
  classifier = LogisticRegression(random_state = 0)
  classifier.fit(x_train, y_train)
  Predicting the Test set results
  y_pred = classifier.predict(x_test)

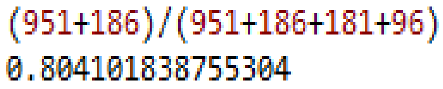
Model Evaluation

- The confusion matrix helps in finding the accuracy and precision of the model
- from sklearn.metrics import confusion_matrix
  cm = confusion_matrix(y_test, y_pred)
- Confusion matrix for Logistic Regression



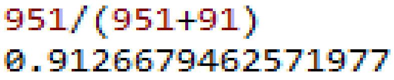
```
array([[939, 102],
       [189, 179]], dtype=int64)
```

- Accuracy is given by: True Positive + True Negative/ Total



```
(951+186)/(951+186+181+96)
0.804101838755304
```

- Precision is a measure that tells us what proportion of customers we analyzed as churn customers, left the company.



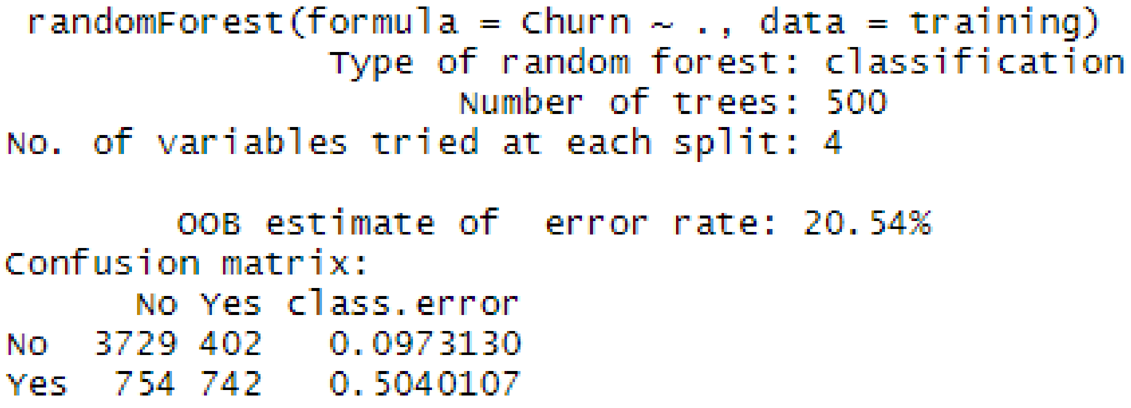
```
951/(951+91)
0.9126679462571977
```

Data Modelling

Random Forest and Decision Tree

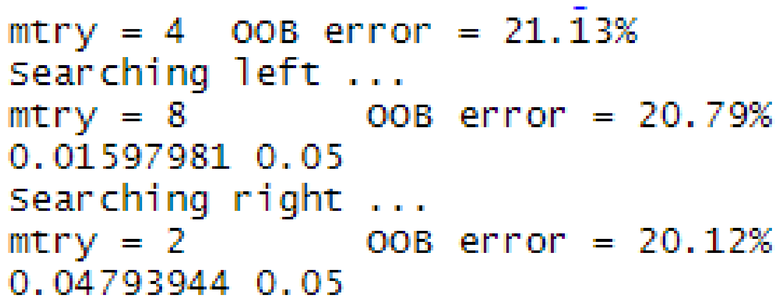
- Random forest consists of large number of individual decision trees that operate as an ensemble.
- Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Tuning the model



```
 randomForest(formula = Churn ~ ., data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 20.54%
Confusion matrix:
      No Yes class.error
No  3729 402   0.0973130
Yes  754 742   0.5040107
```

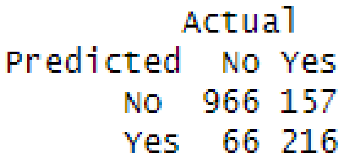


```
mtry = 4  OOB error = 21.13%
Searching left ...
mtry = 8        OOB error = 20.79%
0.01597981 0.05
Searching right ...
mtry = 2        OOB error = 20.12%
0.04793944 0.05
```

Data Modelling

Model Evaluation

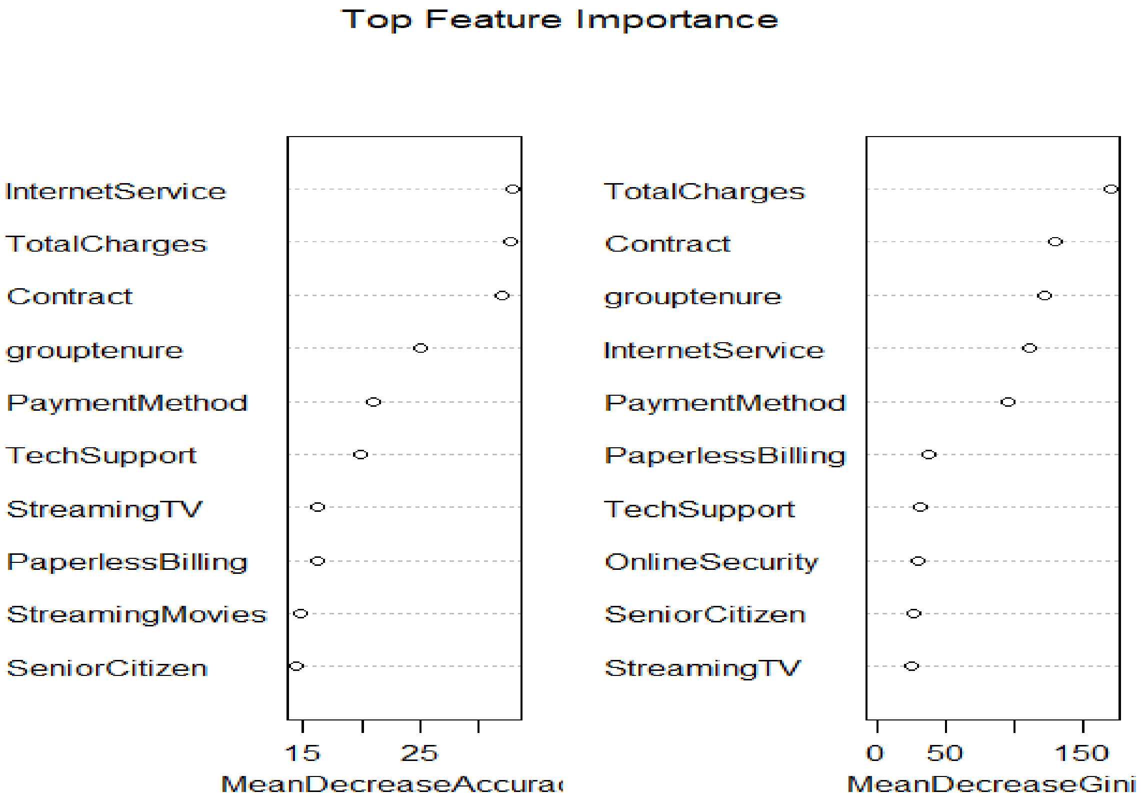
- Confusion Matrix for Random Forest



```
         Actual
Predicted  No Yes
      No  966 157
      Yes  66 216
```

- Accuracy of Random Forest model is 0.8412

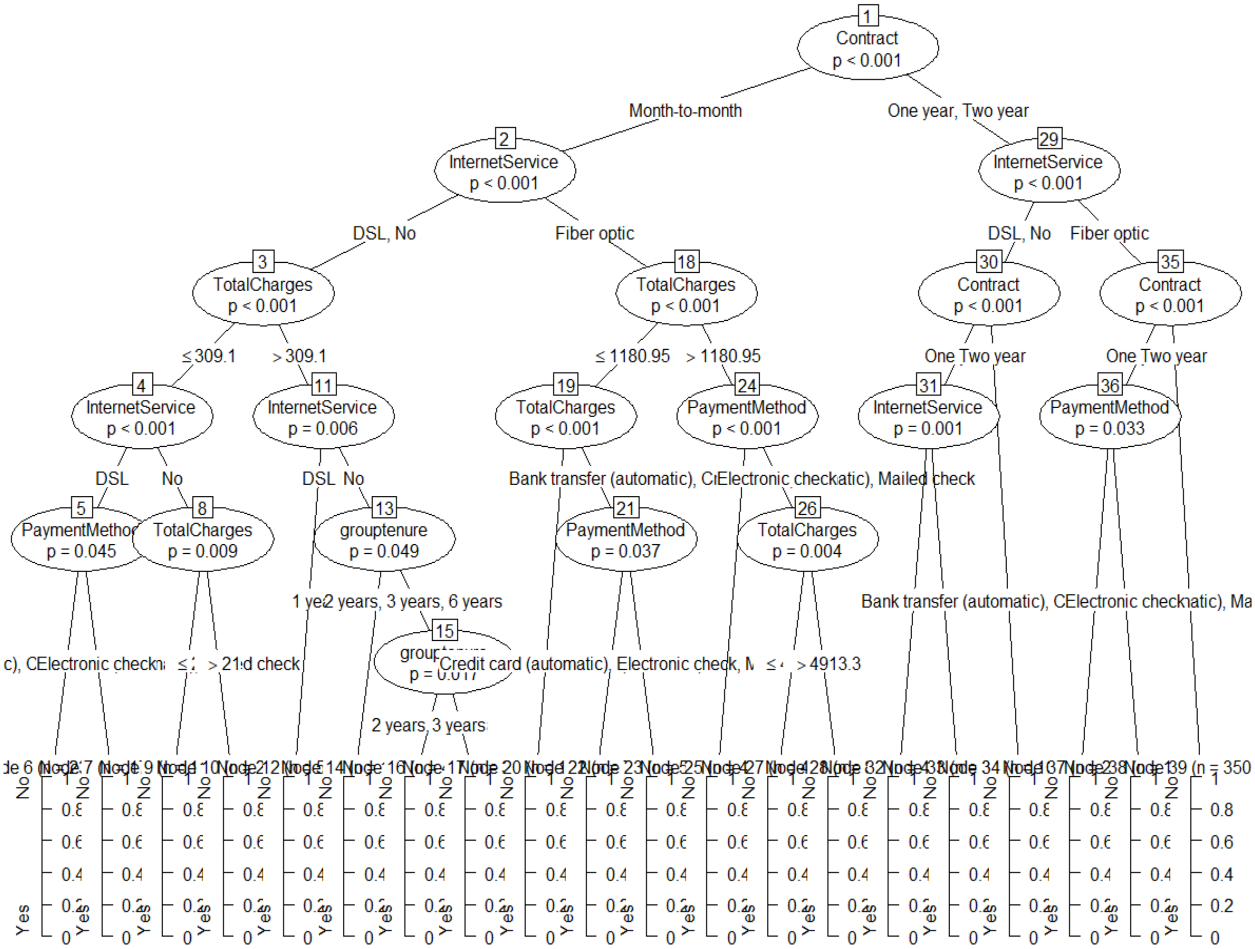
Total feature importance



CUSTOMER CHURN PREDICTION

Data Modelling

Decision Tree



Conclusion

Random Forest model has higher accuracy than Logistic Regression model.

CUSTOMER CHURN PREDICTION

Applications

- Customer Churn prediction is helpful in determining customer lifetime value.
- Can be useful in personalized marketing and CRM.
- Churn prediction can help in retaining customers.