**CUSTOMER CHURN PREDICTION**

Customer Churn is a challenging problem in the Telecom industry. Due to the direct effect on the revenues of the company, companies are looking for robust models to predict churn of customers. Building a strategy targeting the section of customers who would churn will help the companies grow. Churn prediction model studies the customer activity from the past and probabilistically identifies the stage at which customers will leave the company. So, in this problem, based on the customer behavior data from past a churn prediction model is developed. Different machine learning models are applied for a comparative study and the best prediction model is finalized.

Telecom customer churn dataset from IBM Sample Datasets is used for this study.

**Data Analysis**

There are 7043 rows and 21 columns in this dataset. 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents’, 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn' are the different columns. ‘Churn’ is the target column and the others are features. This is a Binary Classification problem.

On a first look, when we check for null values, there is no missing data. On checking the datatypes of the columns, it is found that ‘TotalCharges’ is an object datatype though it has float values. Further analysis led to the fact that it has empty spaces making the datatype object.



Also ‘Tenure’ has zero values corresponding to empty spaces in ‘TotalCharges’ even though ‘MonthlyCharges’ has values.

Convert the datatype of ‘TotalCharges’ to float and remove the rows with missing values.



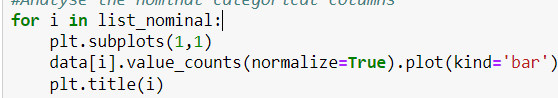
Columns are to be divided into continuous, nominal categorical and ordinal categorical.

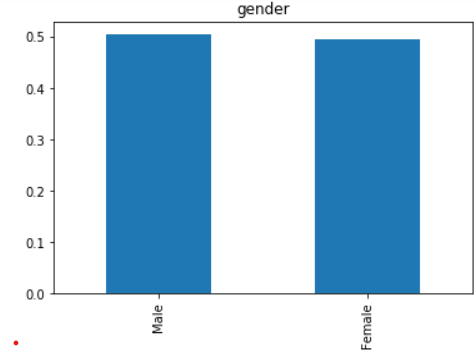
Nominal Categorical- 'gender', 'SeniorCitizen', 'Partner', 'Dependents', ‘PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'Churn'.

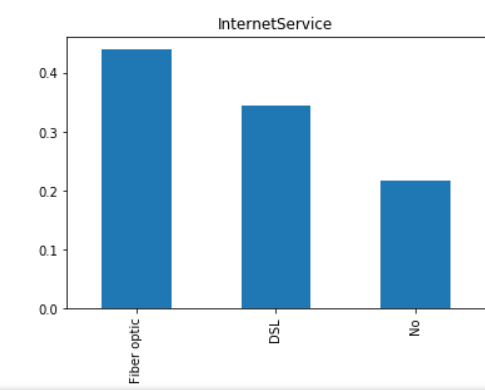
Continous Columns- ‘tenure’, ‘TotalCharges’, ‘MonthlyCharges’

**Data visualisation** is done both in terms of univariate and bivariate analysis.

**Univariate analysis**

****

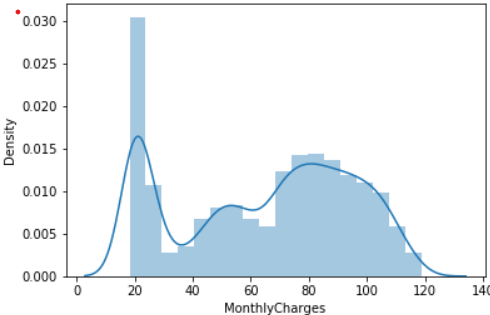


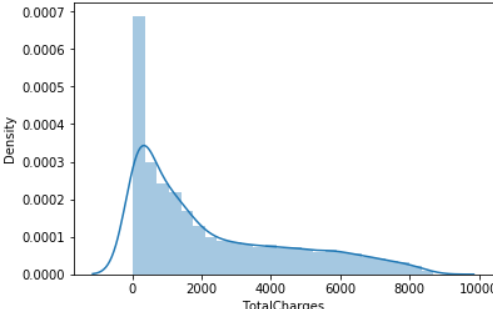


Univariate analysis indicates the following points

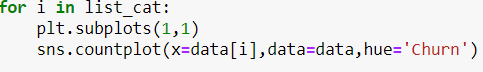
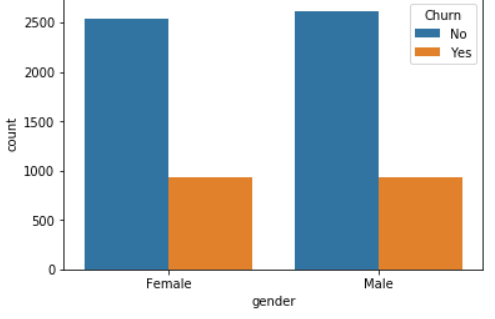
* Dataset has almost equal proportion of male and female customers.
* There is younger population in the dataset.
* Half of the customers have partner and not many customers seem to have dependents
* Most of the customers have phone service. Almost half of the customers have multiple lines.
* Majority of the customers opted for fibre optic internet service
* Majority of the customers are on a month-to-month contract.
* Most of them opted for paperless billing.

Distribution of the continuous columns is checked which indicates data is skewed.





**Bivariate Analysis**

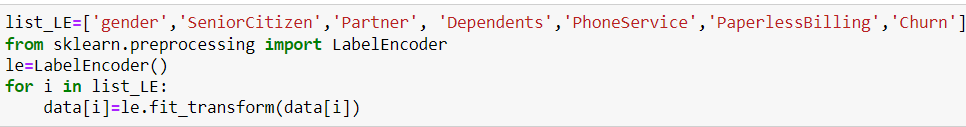
****

In this visualisation, effect of different features on churn is studied and it is found that features seem to have an effect on target variable.

**Checking For Outliers-** Box plot of continuous columns is done to check for outliers. It is observed that no outliers are present.

**Encoding-** Machine Learning Models can interpret only numerical data, so all the categorical data is to be encoded.

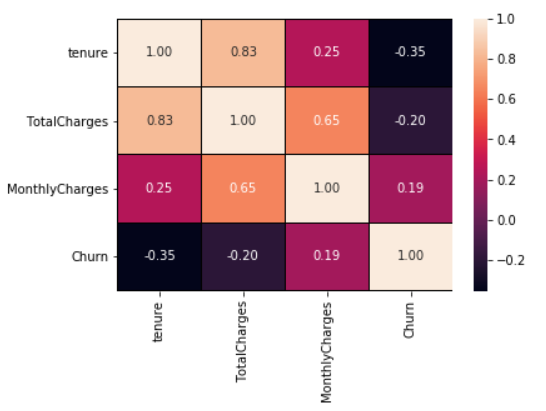
Target variable 'Churn' and the categorical features that have 2 unique categories can be Label Encoded.



The other categorical columns are OneHot Encoded using get\_dummies method.

**Correlation** and the corresponding **Heatmap** of the continuous data can be checked.





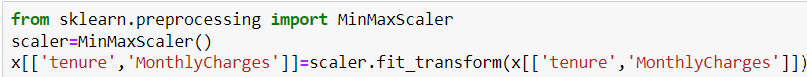
High correlation between Tenure and TotalCharges is understandable. But the high correlation between TotalCharges and MonthlyCharges may bring in multicollinearity.

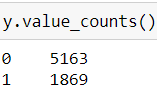
Infact tenure\*MonthlyCharges=TotalCharges. So TotalCharges can be removed. Also, customerID is of no use and can be removed.



Split the dataset into x and y. Check the statistics of features dataset i.e x. There is difference in range of values with respect to tenure and MonthlyCharges. So, Scaling has to be done.

**Scaling –** MinMax Scaler is used.

**Imbalance-** Target column has imbalance interms of count of ‘churn’ and ‘no churn’.



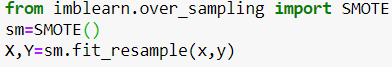
Class Imbalance can hamper the model accuracy. Almost all the machine learning models work best when there are equal number of samples in all the classes. So it has to be treated.

Class Imbalance can be treated in two ways –

1.Balncing the classes- Upsampling(SMOTE), Downsampling(Near Miss)

2.Not Balancing the classes- Use f1 Score /ROC Score to choose the best model.

In this problem, Imbalance is treated with SMOTE.

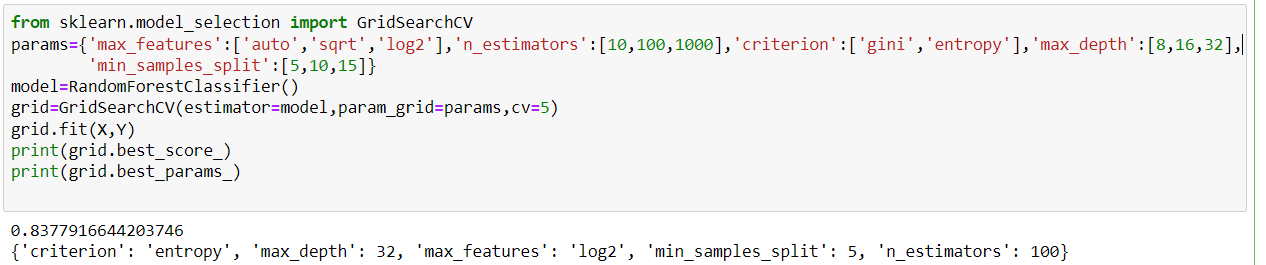


**Fitting Models**

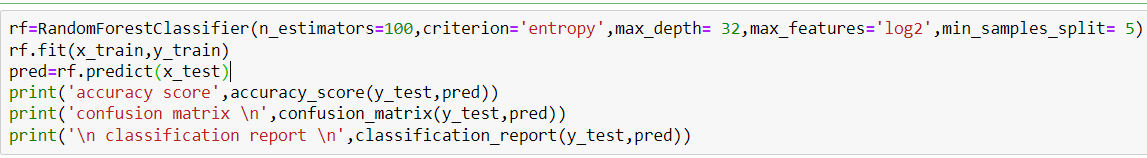
Logistic Regression, Decision Tree Classifier, KNN Classifier, Support Vector Machine Classifier, Random Forest Classifier and Gradient Boosting Models are applied. Cross Validation is done for each of the models. Based on the accuracy score and difference between accuracy score and CV score (Minimum Difference and maximum accuracy) Random Forest Classifier is chosen as the best model.

**Hyper Parameter Tuning** of Random Forest Classifier is done using GridSearchCV method.

Hyper Parameter Tuning is the process of determining the right combination of hyperparameters of a model which will maximise the performance of the model.

 found to be the best combination of parameters

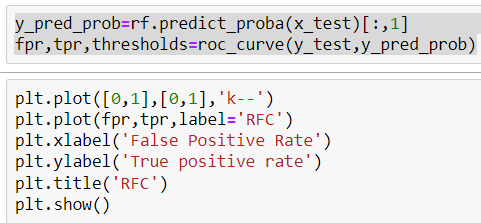
Using these hyperparameters, Random Forest Model is again fit to the data

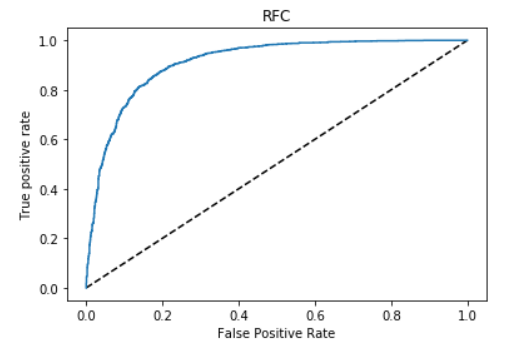
An accuracy score of 84% is achieved.

**ROC Curve**

Reciever Operating Characteristic Curve is a good tool for predicting the probability of a binary outcome. It is plot of the false positive rate on x-axis versus the true positive rate on y-axis for a number of different threshold values between 0 and 1.

Roc curve is plotted for the model





AUC Score is calculated.



AUC represents the area under the ROC Curve. Higher the AUC, the better the model at accurately classifying the classes. Ideally the ROC curve should extend to the top left most corner in which case AUC is 1. In the present case i.e on fitting Random Forest Classifier with hypertuned parameters, AUC is found to be 0.84.

A more in-depth Feature Engineering may improve the model performance. Also, more advanced machine learning models might give improved accuracy but at the cost of more data and computing power.