**CHURN MODELLING**

**FOR**

**TELECOM COMPANY**

**1.1 INTRODUCTION:**

Python is the most popular programming language used for Machine Learning and Artificial Intelligence.

Machine Learning and Artificial Intelligence requires continuous data processing, and Python’s libraries let you access, handle and transform data. These are some of the most widespread libraries you can use for ML and AI:

1. Scikit-learn
2. Pandas
3. Keras etc.

We use python because:

1. Python for Machine learning and Artificial Intelligence is a great choice, as this language is very flexible.
2. It offers an option to choose either to use OOPs or scripting.
3. There’s also no need to recompile the source code, developers can implement any changes and quickly see the results.
4. Programmers can combine Python and other languages to reach their goals.

**1.2 OBJECTIVES OF RESEARCH:**

* Main objective of this model is to reduce the loss of customers to the company.
* This can be done by knowing the reason for loosing costumers and in return providing offers to attract the costumer.
* Company should focus on inconvenience found in customers and should able to reduce the inconvenience as soon as possible.

**1.3 PROBLEM STATEMENT:**

* Main disadvantage of any company or business is losing their customers.
* Due to increase in churn rate, the company may lose its profits or may shut down in the worst scenario.
* With the help of churn rate, company can know the number of customers leaving. So, company can know the reasons for losing customers and improve their services/ products accordingly.
* For example, if Netflix knew a segment of customers who were at risk of churning they could proactively engage them with special offers instead of simply losing them.

**2 REVIEW OF LITERATURE:**

These are the existing solutions and our solution is given in the conclusion.

* Analyse and determine why churn occur.
* Use co-browsing to offer personalised customer service.
* Improve user on-boarding process.
* User pro-active approach.
* Make it hard for your customers to ignore you.
* Create a community around a product.
* Remind your customers how much value you provide them.
* Increase customer engagement.
* Surprise and delight your customers.
* Provide additional services.

**3 DATA COLLECTION:**

## A dataset is collected regarding Telecom Customer Churn. Each row represents a customer, each column contains customer’s attributes described on the column Metadata. The raw data contains 7043 rows (customers) and 17 columns (16 features+1 output). The “Churn” column is our target.

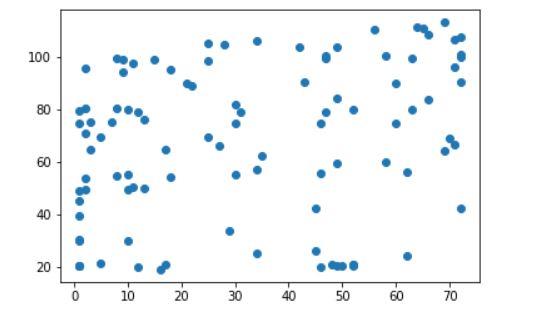
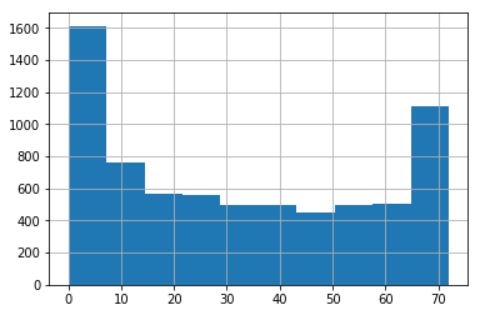
## Columns are as follows:

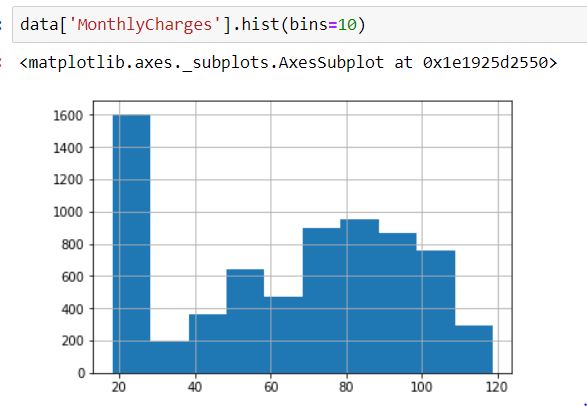
* gender: Whether the customer is a male or a female
* SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
* Partner: Whether the customer has a partner or not (Yes, No)
* Dependents: Whether the customer has dependents or not (Yes, No)
* tenure: Number of months the customer has stayed with the company
* PhoneService: Whether the customer has a phone service or not (Yes, No)
* MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
* OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
* DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
* StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
* StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract: The contract term of the customer (Month-to-month, One year, Two year)
* PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
* MonthlyCharges: The amount charged to the customer monthly
* Churn: Whether the customer churned or not (Yes or No)

1. **METHODOLOGY:**
   1. **Exploratory Data Analysis ( EDA ):**
      1. **Figures and Tables:**

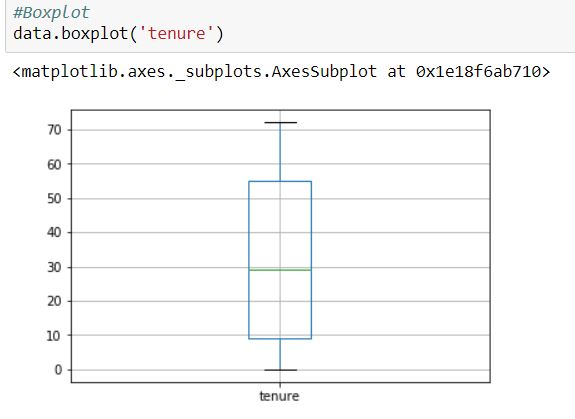
* Highest Positive correlation is for StreamingMovies and MonthlyCharges i.e., 0.629603
* Highest Negative correlation is for tenure and churn i.e., -0.352228
* There are no null values in this dataset
* The Contract column has been encoded to float type from String type using LabelEncoder and OneHotEncoder importing from sklearn.
* 73% of customers from the dataset are willing to continue to get the services of this company (who are not churns) i.e., 5174 people out of 7043.
* While the rest 27% are willing to discontinue their services of this company (who are churns) i.e., 1869 out of 7043.

**PLOTTING:**

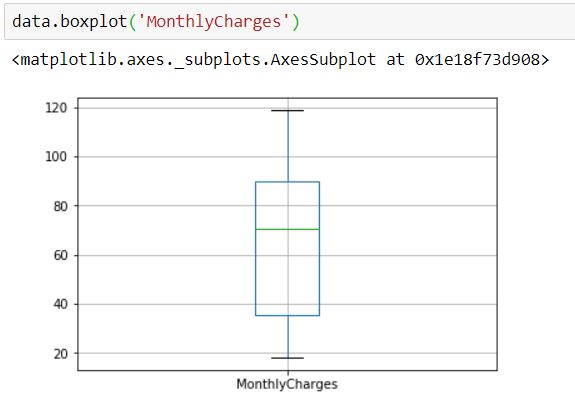
* Scatterplot for first 100 rows or instances between tenure and MonthlyCharges columns:
  + 
* Histogram for tenure column:
  + - Majority of the customers have a tenure of below 7 years (1600) and the least are of tenure 43 to 51 years(450). The maximum tenure is of 72 years.
  + 
* Histogram for MonthlyCharges column:
* Majority of the customers have a tenure of below 7 years (1600) and the least are of tenure 43 to 51 years (450). The maximum tenure is of 72 years.



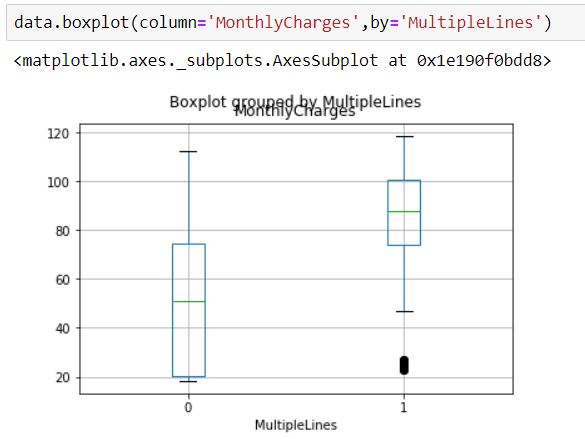
* Boxplot for tenure column:
  + Minimum tenure is 0 years and maximum is 72 years and average is 29.



* Boxplot for MonthlyCharges column:
* Minimum MonthlyCharges is 19 and maximum is 119 and average is 70.

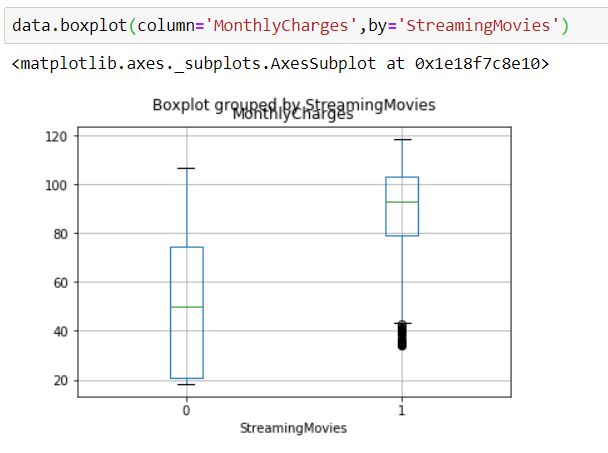


* Boxplot for Monthly Charges by Multiple Lines:



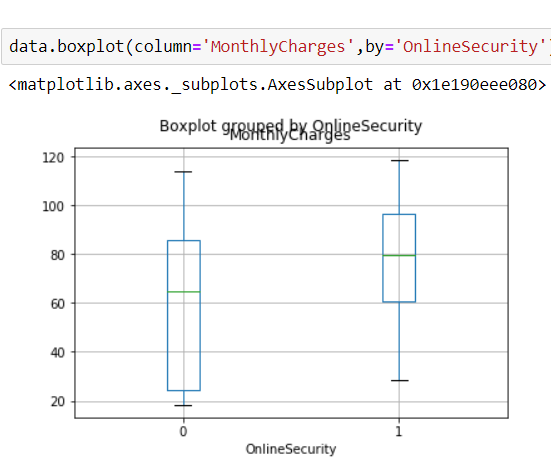
Customers with multiple lines are having high monthly charges. Few customers with multiple lines (outliers) are having low monthly charges.

* Boxplot for Monthly Charges with Streaming Movies:



Customers who are streaming movies are paying more monthly charges and vice versa. Few customers who are streaming movies (outliers) are having low monthly charges.

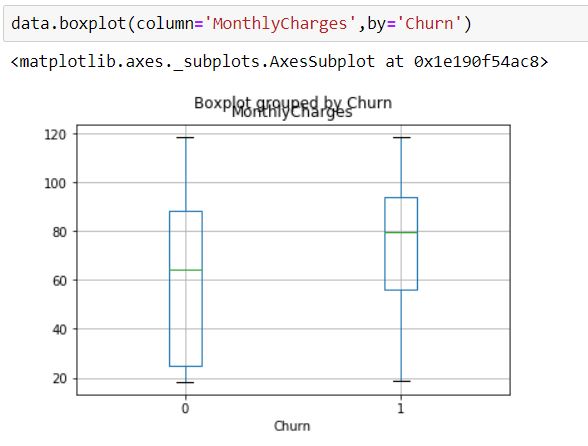
* Boxplot for Monthly Charges by Online Security:



Customers who are having online security are paying more monthly charges and vice versa.

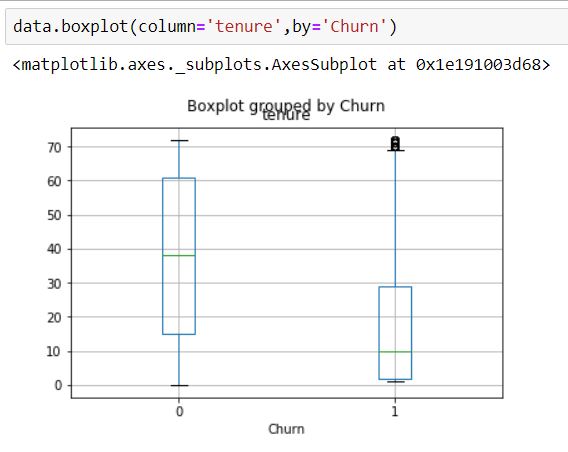
* Boxplot for MonthlyCharges column by Churn:

Customers with high monthly charges are becoming churns i.e., discontinue their services with the company.



* Boxplot for tenure column by Churn:

Customers with low tenure are becoming churns i.e., discontinuing their services of the company. And also we have an outlier here i.e., customers even with very high tenure are becoming churns.



* 1. **Data Modelling:**

**Decision Tree:**

Accuracy score: 73.937677

Recall score: 57.446809

ROC score: 68.684794

[[207 52]

[ 40 54]]

Accuracy score is 73.937. This indicates that 74% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 261 and total is 353 so accuracy score would be 261/353 i.e., 74%

Recall score is sensitivity that is 57%. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high.

ROC Score: Receiver Operating Characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.

Confusion matrix:

1. 207 out of 353 customers are detected by the model as churns and they are actually churns
2. 52 out of 353 customers are detected by the model as they churns but they are actually not churns
3. 40 out of 353 customers are detected by the model as they are not churns but those customers are churns actually
4. 54 out of 353 customers are detected by the model as they are not churns and the customers are also not churns in reality

**Random Forest:**

Accuracy score: 81.586402

Recall score : 57.446809

ROC score : 73.897149

[[234 25]

[ 40 54]]

Accuracy score is 81.58. This indicates that 81.6% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 288 and total is 353 so accuracy score would be 288/353 i.e., 81%

Recall score is sensitivity that is 57%. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high.

Roc Score: receiver operating characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.

Confusion matrix:

1. 234 out of 353 customers are detected by the model as churns and they are actually churns
2. 25 out of 353 customers are detected by the model as they churns but they are actually not churns
3. 40 out of 353 customers are detected by the model as they are not churns but those customers are churns actually
4. 54 out of 353 customers are detected by the model as they are not churns and the customers are also not churns in reality

**Naive Bayes Classification:**

Accuracy score: 77.337110

Recall score : 79.787234

ROC score : 78.117555

[[198 61]

[ 19 75]]

Accuracy score is 77.33. This indicates that 77% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 273 and total is 353 so accuracy score would be 273/353 i.e., 77%

Recall score is 79.78. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high. Svc classifier gives high accuracy for training data but fails for testing data

Roc Score: receiver operating characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.

Confusion matrix:

1. 198 out of 353 customers are detected by the model as churns and they are actually churns
2. 61 out of 353 customers are detected by the model as they churns but they are actually not churns
3. 19 out of 353 customers are detected by the model as they are not churns but those customers are churns actually
4. 75 out of 353 customers are detected by the model as they are not churns and the customers are also not churns in reality

**Logistic Regression:**

Accuracy score: 82.436261

Recall score : 58.510638

ROC score : 74.815165

[[236 23]

[ 39 55]]

Accuracy score is 82.43. This indicates that 82% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 291 and total is 353 so accuracy score would be 291/353 i.e., 82%

Recall score is 58.5. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high. Svc classifier gives high accuracy for training data but fails for testing data

Roc Score: receiver operating characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.

Confusion matrix:

1. 236 out of 353 customers are detected by the model as churns and they are actually churns
2. 23 out of 353 customers are detected by the model as they churns but they are actually not churns
3. 39 out of 353 customers are detected by the model as they are not churns but those customers are churns actually
4. 55 out of 353 customers are detected by the model as they are not churns and the customers are also not churns in reality

**Support Vector Classification:**

Accuracy score: 78.470255

Recall score : 58.510638

ROC score : 72.112462

[[222 37]

[ 39 55]]

Accuracy score is 78.47. This indicates that 78% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 277 and total is 353 so accuracy score would be 277/353 i.e., 78%

Recall score is sensitivity that is 58.5%. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high.

Roc Score: receiver operating characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.

Confusion matrix:

1. 222 out of 353 customers are detected by the model as churns and they are actually churns
2. 37 out of 353 customers are detected by the model as they churns but they are actually not churns
3. 39 out of 353 customers are detected by the model as they are not churns but those customers are churns actually
4. 55 out of 353 customers are detected by the model as they are not churns and the customers are also not churns in reality

**KNN Classifier:**

Accuracy score: 77.337110

Recall score: 54.166667

ROC score: 70.079442

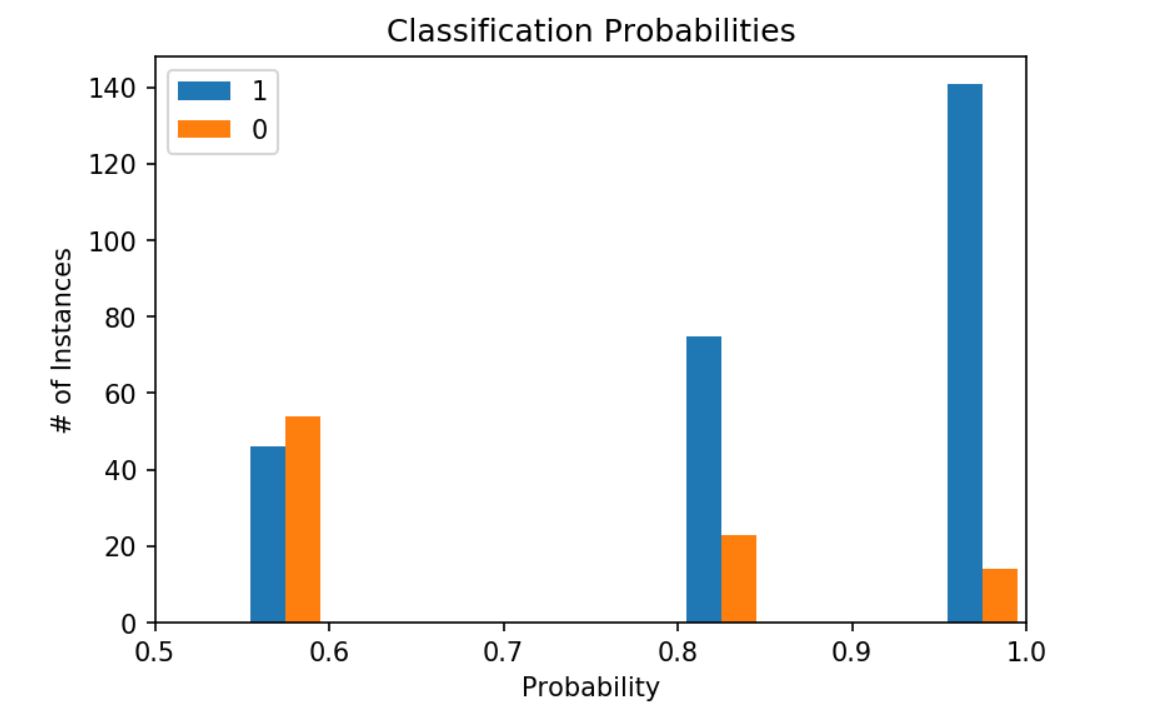
[[221 36]

[44 52]]

Accuracy score is 77.33. This indicates that 77% of predicted values matched with actual values. The sum of true positives and true negatives from the confusion matrix is 273 and total is 353 so accuracy score would be 273/353 i.e., 77%

Recall score is sensitivity that is 54%. Sensitivity is ratio of true positive to sum of true positive and false negative. For a good model sensitivity should always be high.

ROC Score: receiver operating characteristic curve. It is plotted between tpr on x-axis and fpr on y-axis.it is 1-specificity.



**So, logistic regression classifier best fits this churn model.**

1. **FINDINGS AND SUGGESTIONS:**

* Customers with high monthly charges are becoming churns i.e., discontinue their services with the company. Hence to reduce the loss of customers, the company should decease the monthly charges.
* Since customers who are streaming movies and having online security are paying high monthly charges, by reducing the charges on streaming movies, there is a chance of reducing loss of customers.

1. **CONCLUSION:**

* Churn Modelling helps to find the number of customer losing in a time period.
* Logistic Regression is the best model so far with highest Accuracy score among all the other classification models.
* So, this can be used for predicting the churns of the company and so by said to be the solution of their problem.