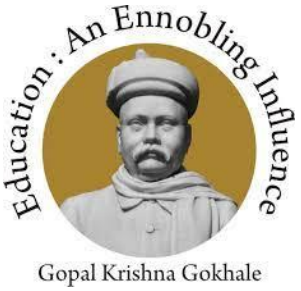
GOKHALE INSTITUTE OF POLITICS & ECONOMICS



SUBJECT: BUSIENSS ANALYTICS

TOPIC – CHURN PREDICTION OF A TELECOM COMPANY

ADITYA JORI [AE2017]

DATE: 26th July, 2021

Business Analytics Project

"Telecom Customer Churn Analysis"

By

Branch: Agri-Business

PROBLEM STATEMENT

Using the data provided, this paper aims to analyze the data to determine what variables are correlated with customer churn, if any. Additionally, prediction models, to identify the people that might churn, will also be built. To build a prediction model, we will make different models using techniques such as Logistic Regression, Decision Tree Classifier, K-Nearest Neighbour Classifier and Naïve Bayes, Support Vector Classifier. These models will then be compared on the number of parameters obtained and the model optimized for final use. After the churn rate, we will also identify a subset of customers who will be offered retention plans.

Our problem is based on the telecom customer churn prediction of a company in the Telecom industry.

The Telecom industry has a strong presence across the country. It offers Internet services which are necessary for every citizen. The industry is the foothold of this country which is responsible for our everyday communication. India is the world's second-largest telecommunications market globally. The revenue generated by this industry has a large portion of contribution in our Nation's GDP.

DATASET DETAILS

- The data which was given to us provided information about the customer's characteristics who use the telecom service.
- There are a total of 15 features which give us various kinds of detail about each client.
- Out of the 15 features *Senior Citizen, Tenure, Monthly Charges and Total Charges* are Numerical type of features. Remaining features are of the Categorical type. (Barring Customer ID)
- There are more than 7000 data samples in the dataset.
- The Output Label takes value of 1 if the customer discontinues the telecom service and 0 if they continue to use the service.
- Out of the given features, namely *Senior citizen, Service, Monthly Charges* are comparatively more significant in the churn prediction.

	customerID	gender	SeniorCitizen	Partner	Dependents	...	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	...	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	Male	0	No	No	...	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	Male	0	No	No	...	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	Male	0	No	No	...	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	Female	0	No	No	...	Yes	Electronic check	70.70	151.65	Yes
5	9305-CDSKC	Female	0	No	No	...	Yes	Electronic check	99.65	820.5	Yes
6	1452-KIOVK	Male	0	No	Yes	...	Yes	Credit card (automatic)	89.10	1949.4	No
7	6713-OKOMC	Female	0	No	No	...	No	Mailed check	29.75	301.9	No
8	7892-POOKP	Female	0	Yes	No	...	Yes	Electronic check	104.80	3046.05	Yes
9	6388-TABGU	Male	0	No	Yes	...	No	Bank transfer (automatic)	56.15	3487.95	No

10 rows × 16 columns

GOAL

"To build a Machine Learning Model using the Training dataset and precisely predict the customer churn on the Testing Dataset"

LIBRARIES

- To carry out the Exploratory Data Analysis, certain libraries were imported to execute their respective tasks
- For Data Manipulation and Data Visualization *Numpy, Pandas, Matplotlib, Seaborn* were used.
- For Data processing and Feature importance *Scikit Learn* was used.
- For Machine Learning model *Scikit Learn* and *XGBoost* were used.
- We also engineered two features using the Pandas data frame functions and basic operations on them

EXPLORATORY DATA ANALYSIS

- The data was found to be Unbalanced or Skewed given the small size of the Dataset.

```
Total Percentage of People which did not change company 73.46301292063042%
No     5174
Yes    1869
Name: Churn, dtype: int64
```

- After the identification of the Data types of the given features, *Senior Citizen* and Tenure were found to be of Integer Data type, *Monthly Charges* was of Float Data type. The rest of the features were of Object type.

```
customerID          object
gender              object
SeniorCitizen       int64
Partner             object
Dependents          object
tenure              int64
PhoneService        object
MultipleLines       object
InternetService     object
StreamingService    object
Contract            object
PaperlessBilling    object
PaymentMethod       object
MonthlyCharges      float64
TotalCharges        object
Churn               object
dtype: object
```

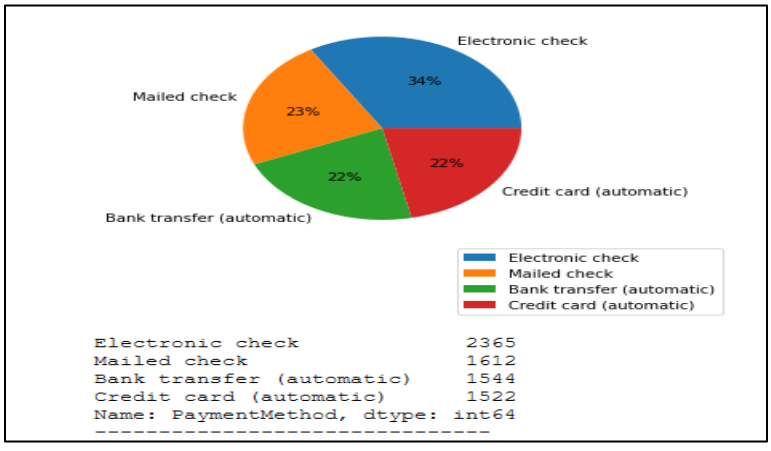
- Even though we can see numerical data in the *TotalCharges* column, its data type is object. It implies that the column consists of Strings as well.

- For each unique item in that column, we checked whether it was of type float. We were able to see that the column consists of 11 instances of single space strings i.e. “ ”.
- We change this space into a NULL value which can be called using the Numpy np.NaN directly.
- By setting them to NULL we were able to convert the data type of the column to float and finally these NULL values were imputed using the column mean.
- The Gender column was converted to Integer data type by giving the values of 0 and 1 to the characteristics in the feature. (Male- 1, Female- 0). This is easily done using a Label Encoder.
- We also list out the unique values present in each column to get an idea of the cardinality of the categorical features.
- Highly cardinal features make the Model more complex. However, we do not have any highly cardinal features in this dataset.

DATA VISUALIZATION

Pie Chart

To understand the categorical columns and their underlying distribution of the categories we use the Pie chart.

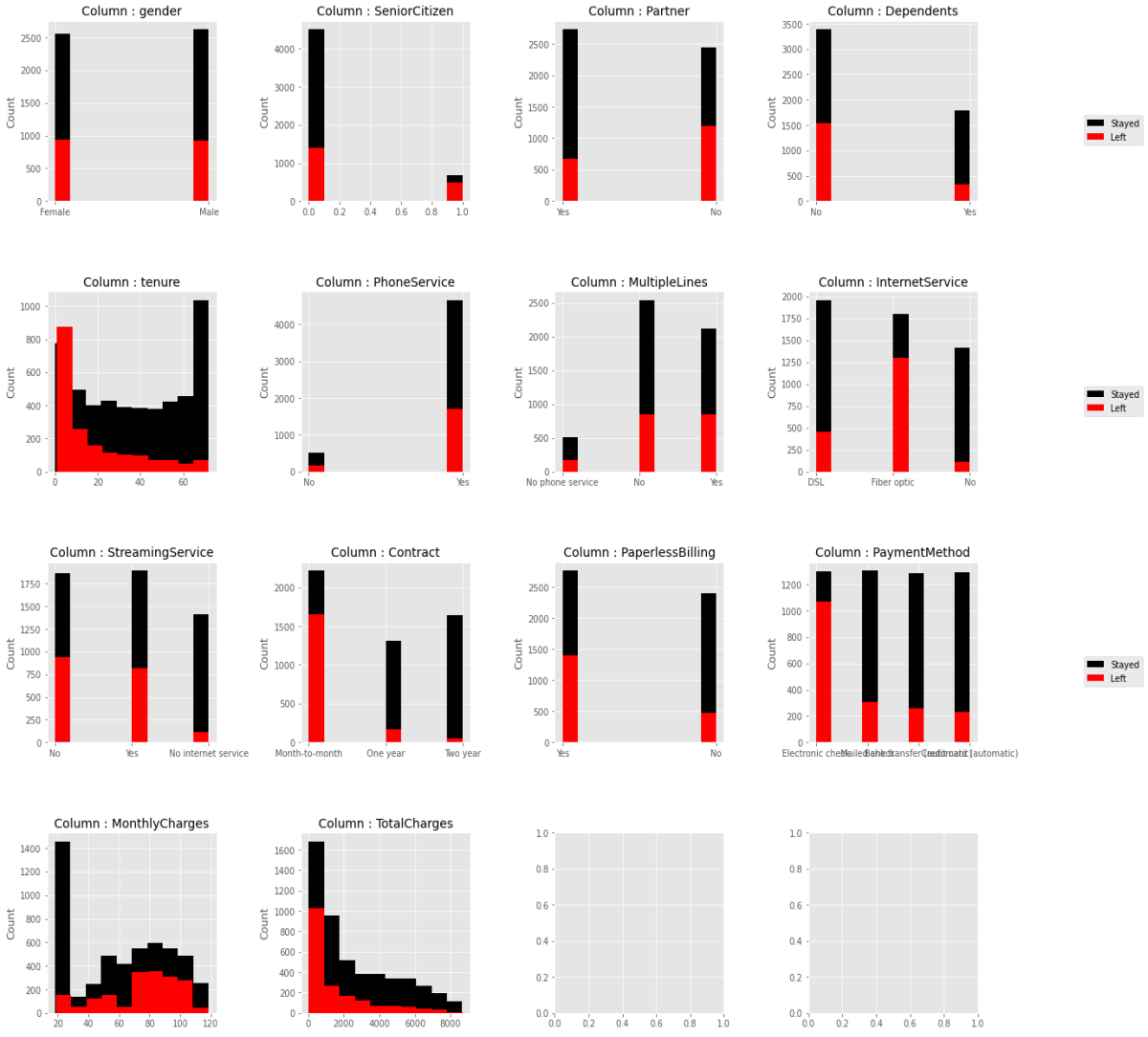



```
Electronic check                2365
Mailed check                    1612
Bank transfer (automatic)       1544
Credit card (automatic)         1522
Name: PaymentMethod, dtype: int64
--------------------------------
```

We can easily figure out that the most used method of payment is Electronic check. We can also call it as the mode. We generally impute categorical features using the mode since mean does not exist in this case.

Histogram

To understand the importance of features, it is necessary that we plot the histogram of features based on the outcome of churn. So we plot the data where churn is 1 and data where churn is 0 for each feature to analyze each feature in depth.

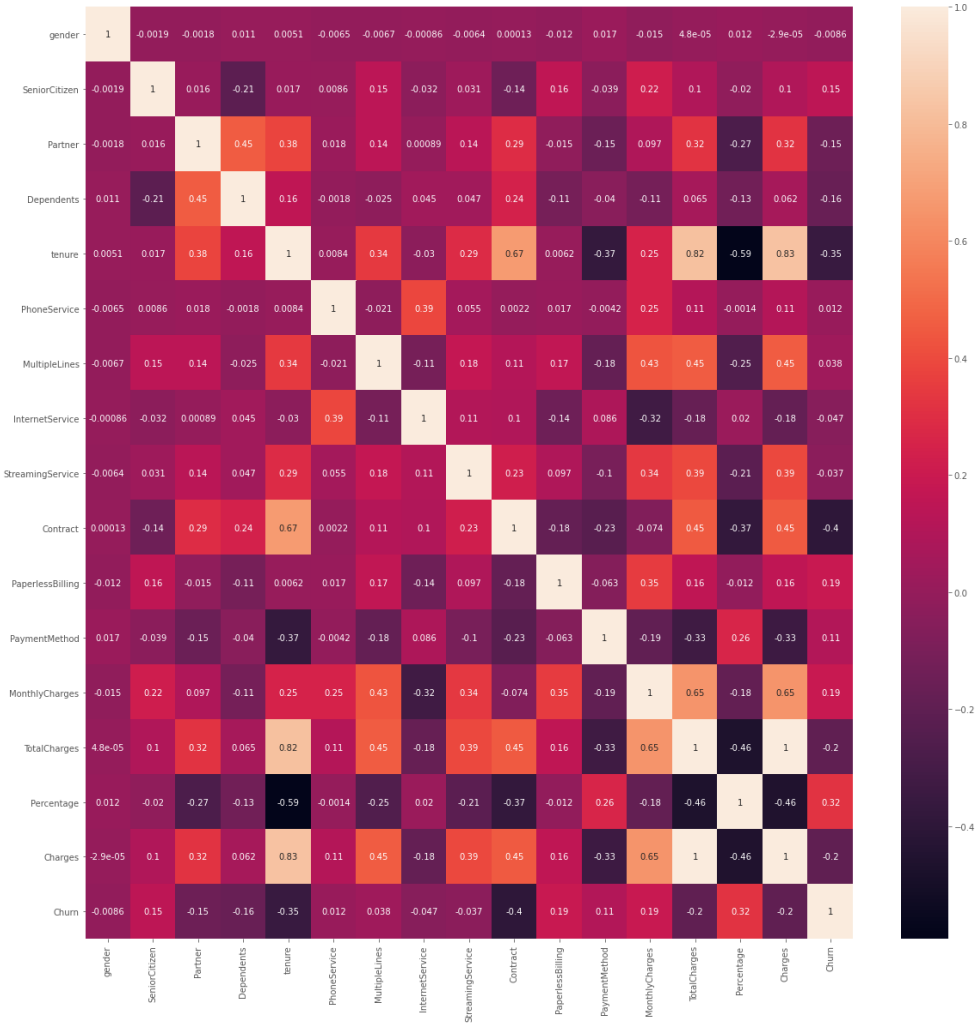


We see that the Tenure feature is quite important as most of the customers with low tenure are the ones to leave the most. Retaining the customers should be prioritized and the customer-care policies should ensure that the customer has a satisfied experience over the years. We also see that month-to-month subscribers also tend to discontinue the services a lot more than the customers having yearly subscriptions. To avoid this, the company needs to come up with more lucrative offers to attract the customers and choose yearly subscriptions over monthly subscriptions. Fiber Optic users are the major contributors to the churn as well. Customers having No Internet services tend to keep using the service while those who do not require the company's service. Hence, the Internet service which is provided should be improved. Customers who prefer paperless billing are more familiar with technology and tend to change their serviceThe customers whose payment is done automatically through e-banking/credit cards have higher probability of staying. So the Manual Payment methods should be reduced.

Heat Map

This heat map shows the correlation coefficient between two features. Highly correlated features do not add new information. Good features are those which

have reasonably high correlation with the Churn Column.



FEATURE ENGINEERING

- Two features namely Percentage and Charges.
- The Percentage feature is basically the Ratio of Monthly to Total Charges.
- The Charges feature is the multiplication of Tenure with Monthly Charges.

FEATURE SCALING

Although feature scaling is not needed for Trees based classifiers, we do the scaling using the Standard Scalar because we are going to compare different classifiers. Linear Classifiers require feature scaling as it makes the gradient descent much easier to converge.

DATA SPILTING

In the process of Model Building the data was split into training data and testing data. The size of the testing data was taken to be 0.3 and the size of the training data to be 0.7. It means that 30% of the data will be tested and the remaining 70% data will be used for training the models.

As the dataset is skewed, we must use the Stratify option while splitting the data so that the distribution of Labels is same in the training and testing samples.

MACHINE LEARNING MODELS

The following ML Models were used and were compared based on the F1 Score and Accuracy.

- **Random Forest Classifier**

  Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. The low correlation

between models is the key. The attributes of the Random Forest Classifier such as 'n_estimators', 'criterion' and 'random state' were added to the model.

- **Logistic Regression Model**

  In logistic regression, the dependent variable is a binary variable that contains data coded as 1 or 0. In other words, the logistic regression model predicts p(Y=1) as a function of X.

- **Support Vector Classifier**

  Also called as “Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. In the SVM algorithm, each data item was plotted as a point in N-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. The attributes of the SVC such as 'n_estimators', 'criterion', 'random_state' were added to the model.

- **K-Nearest Neighbors Classifier**

  K Nearest Neighbor (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms.

  The distance between the data sample and every other sample is calculated with the help of a method such as Euclidean. These values of distances are sorted in ascending order. The top K values from the sorted distances are chosen. The class which is assigned to the sample is based on the most frequent class in the above K values.

- **Decision Tree Classifier**

  The model uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

- **MLP Classifier**

  Multi-layer Perceptron classifier connects to a Neural Network. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLP Classifier relies on an underlying Neural Network to perform the task of classification. A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses back propagation for training the network.

- **XGB Classifier**

  XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modeling problems involving unstructured data (images, text, etc.).

- **Naïve Bayes**

  Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts

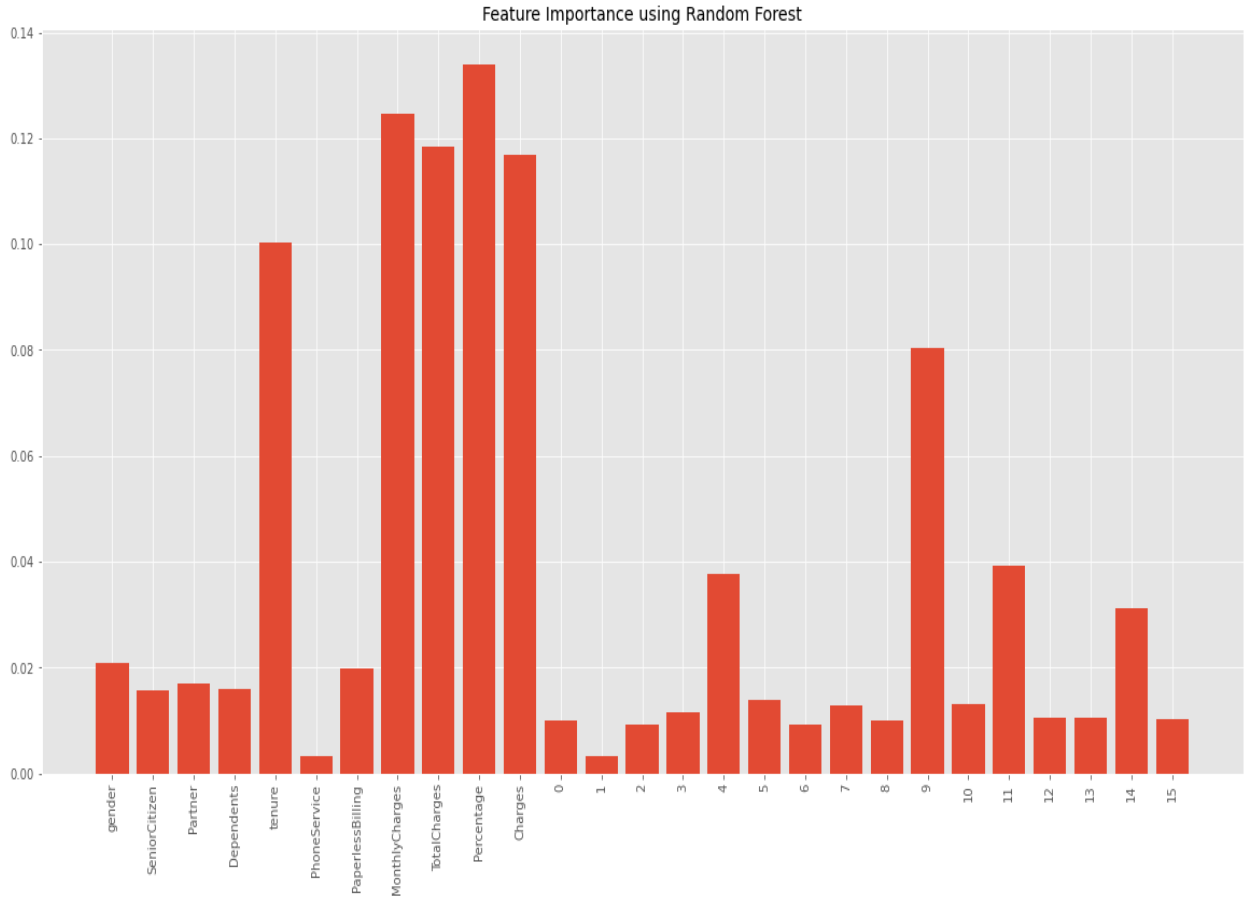
on the basis of the probability of an object. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

RESULTS

	CLASSIFIER	ACCURACY	F1_SCORE
1	Naive Bayes	0.7434926644581165	0.6150568181818181
2	Logistic Regression	0.8021769995267393	0.5777777777777778
3	Support Vector Machine	0.7889256980596309	0.534446764091858
4	Decision Tree	0.7359204921911974	0.49364791288566245
5	K Nearest Neighbours	0.7699952673923331	0.5281553398058253
6	Random Forest Classifier	0.7860861334595362	0.5349794238683128
7	XGBoost Classifier	0.7444391859914813	0.6041055718475073
8	MLP Classifier	0.7969711310932324	0.5671039354187689

FEATURE IMPORTANCE

Using the Random Forest classifier trained earlier, we plot the importance of features using scores provided by the RF classifier.



CONCLUSION

- The given data set is very small and is unbalanced. Further, some of the features like Gender are of no use as they are evenly distributed between the class labels.
- Pandas Profiling Report provides some quick Data Insights. It also shows various correlation coefficients and acts as good starting point for the EDA.
- We see that features related to Charges are quite important followed by features like tenure and connection type.
- The most reliable Model turns out to be the Logistic Regression Model not only because of the high accuracy but also because of the reasonably high F1 score.
- Models such as XGB and Naïve Bayes have the best F1 score but they cannot exceed the 75% accuracy mark. (Since Class 0 labels are almost 73.5%, classifying all test data as 0 will also result in 73.5% accuracy)
- Feature Scaling is important while training SVM. Otherwise, it gives 0 F1 score.
- In order to increase the accuracy and the F1 Score of the Models, we must use proper feature selection techniques, Data augmentation methods like SMOTE followed by Hyper-parameter tuning and Cross Validation during Training.
- An ensemble of the top four classifiers might perform better as well.
- Finally, the Prediction of Customer Churn is best achieved by the Logistic Regression Model.
- The company should focus mainly on Customer Retention rather than focusing on bringing in new Customers. It is observed that the customer becomes more valuable with the passage of time. Hence, a strong Retention Strategy should be opted by the company.