**Customer Churn Analysis**

1. Introduction

Customer retention can be defined as the ability of an company to procure their customers as regular consumers and prevent them from switiching to a comptitor company. Customer churn is when a customer start losing interest and switch to compititor companies. Both can indicate whether the quality of your product or service is sufficient to please your existing customers as well as to identify business growth. It’s also the lifeblood of most subscription-based companies and service providers.

Customer retention strategies are the processes and initiatives businesses put in place to build customer loyalty and improve customer lifetime value. Keeping current customers happy is generally more cost-effective than acquiring first-time customers. According to the Harvard Business Review, acquiring a new customer can be five to 25 times more expensive than holding on to an existing one.

With customer churn rates as high as 30 percent per year in some global markets, identifying and retaining at-risk customers remains a top priority for communications executives. Markets are saturated, unhappy customers defect or downsize their service usage, and a class of professional churners is beginning to emerge.

1. PROBLEM DEFINITION

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low. The dataset is from IBM Telco's sample dataset and is easily available on kaggle and other platforms.

The data set includes information about:

Customers who left within the last month – the column is called Churn

Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges

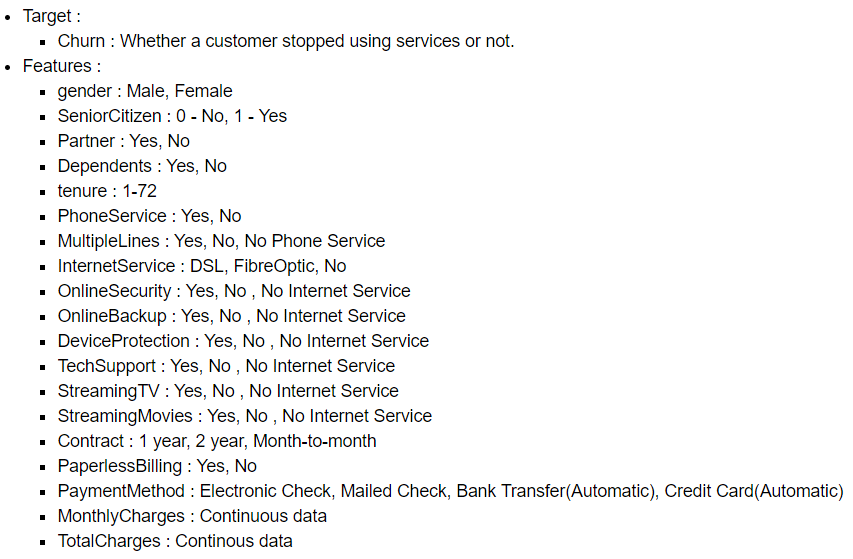
Demographic info about customers – gender, age range, and if they have partners and dependents

Main goal is to build a binary classification model to identify customer churn on the basis of given variables.

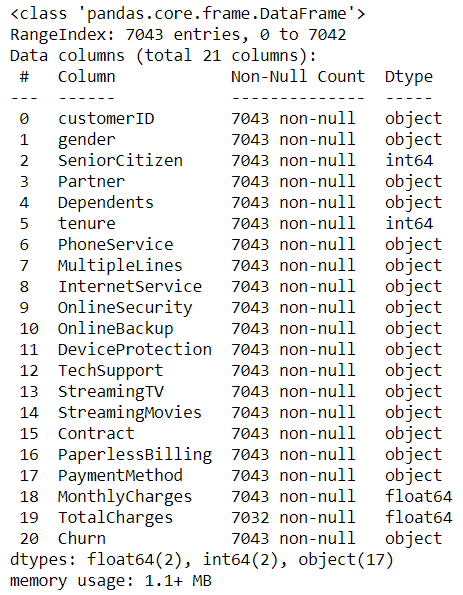
1. EXPLORATORY DATA ANALYSIS

In this section, we'll be analyzing our dataset and gathering insights on dataset shape, null - values, duplicates, etc.

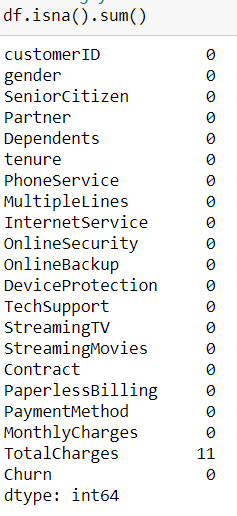
Our Dataset has the following attributes:



Our dataset consists of 17 categorical and 4 continuous attributes.

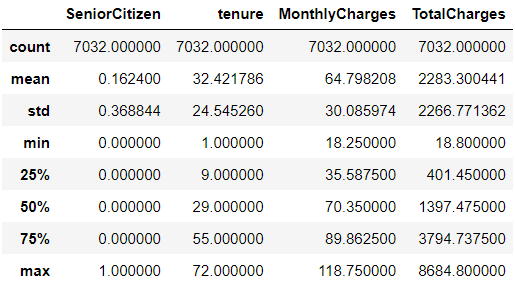


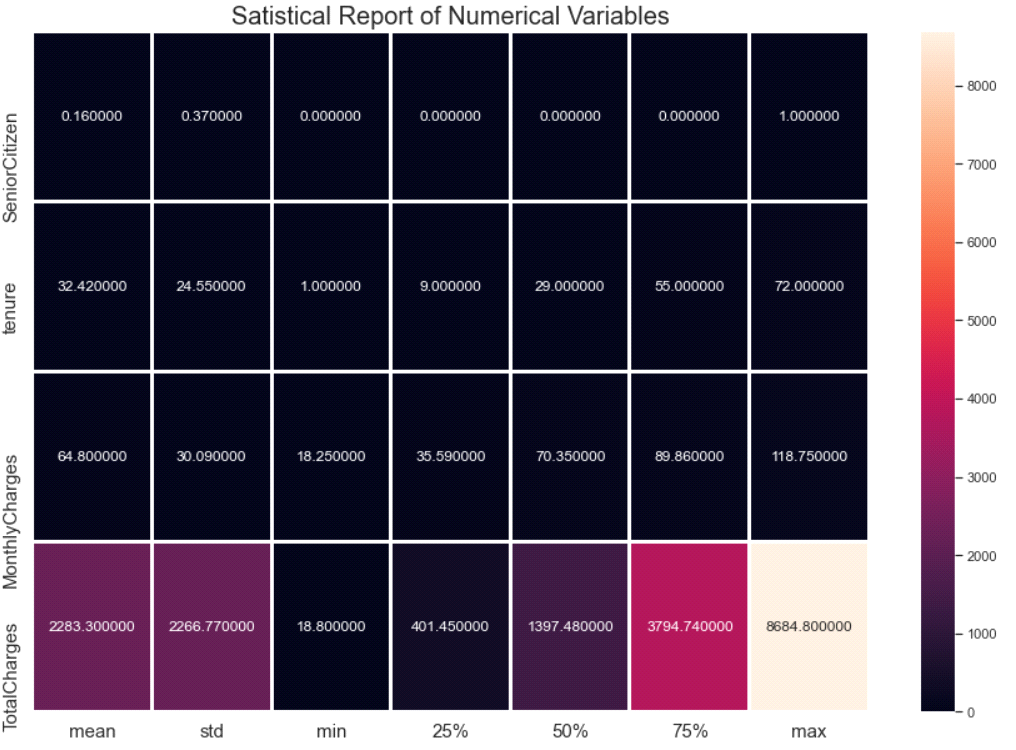
Next I've checked for nulls and duplicates. Our dataset has 0 duplicate records but 11 null fields in Total charges column. Since amount of nulls present are very minimal instead of imputing them I've just dropped all the null fields.



Getting Statistical Description of Continious variables:

Note that "SeniorCitizen" is categorical but already encoded.

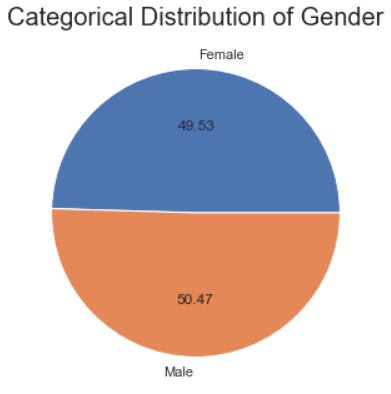
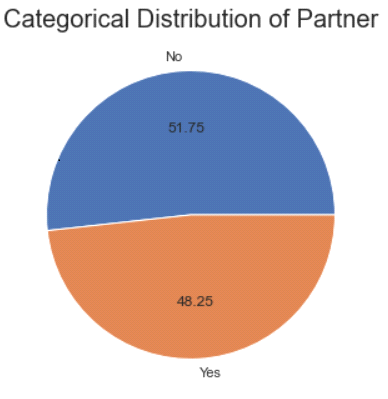


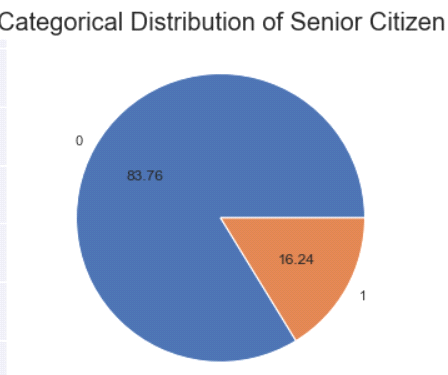
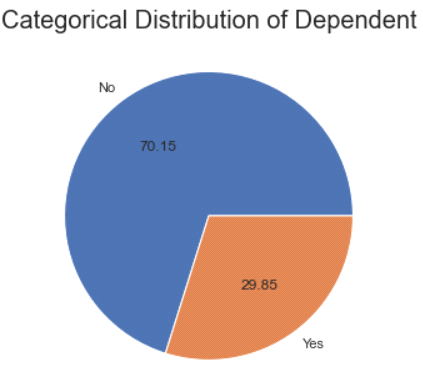


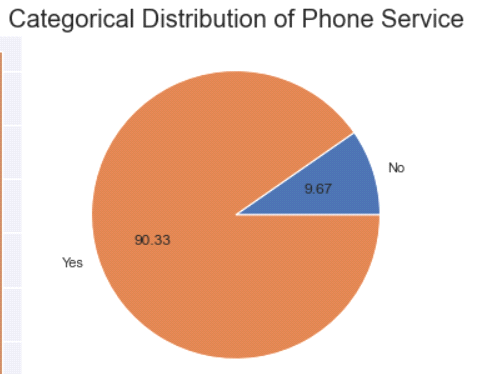
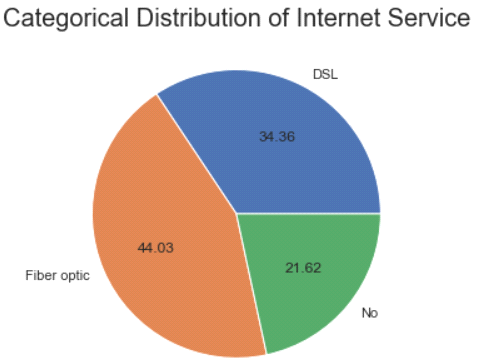
3.1 Visualizations:

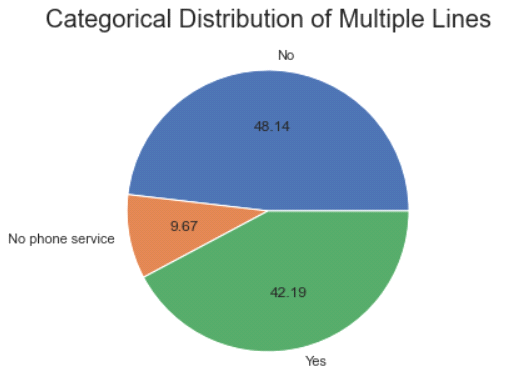
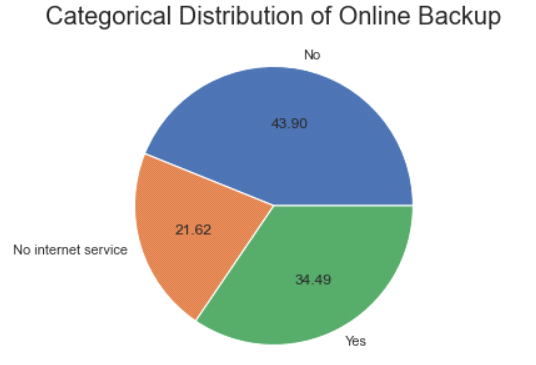
**Univariate Analysis of Variables**: To visualize data distribution of each individual attribute. I've used seaborn and matplotlib's library for visualizations.

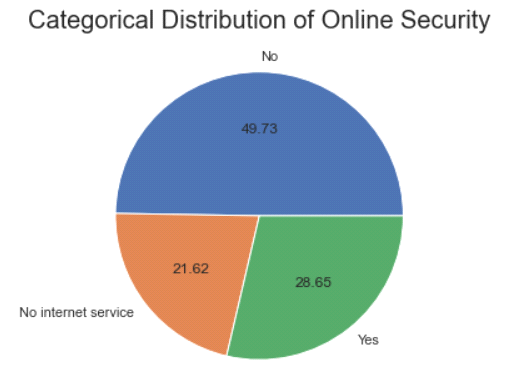
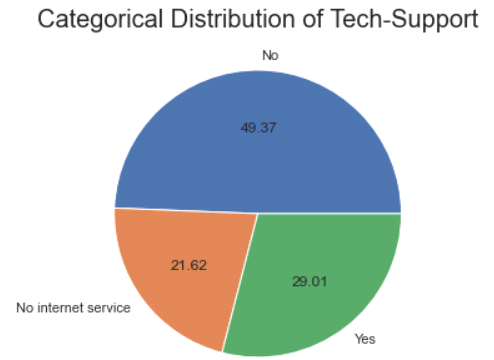
Categorical variables:

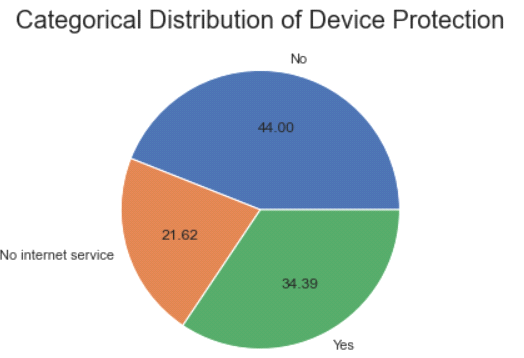
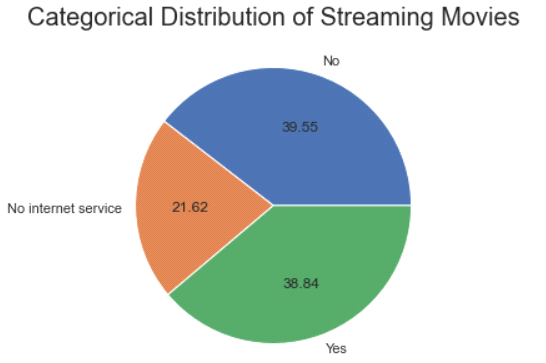
 

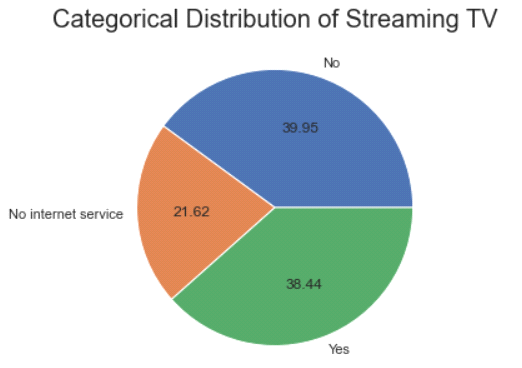
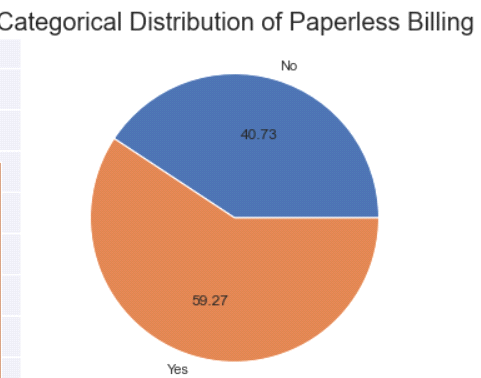
 

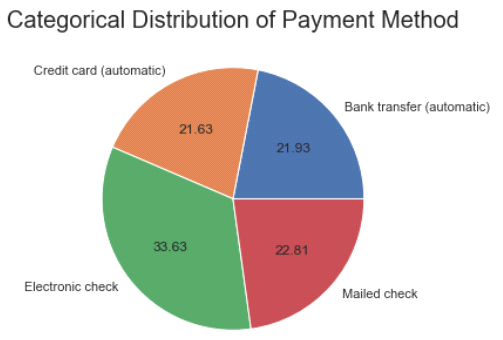
 

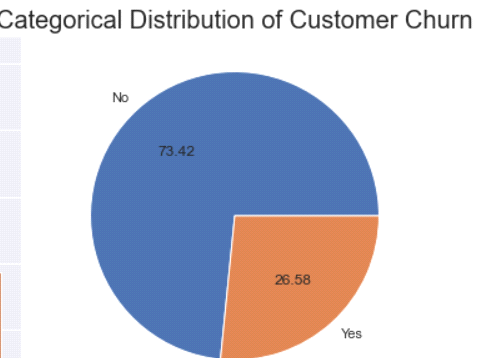
 

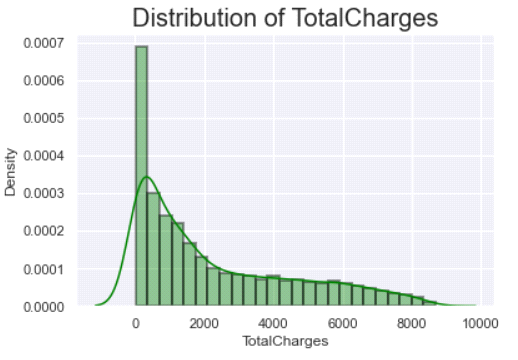
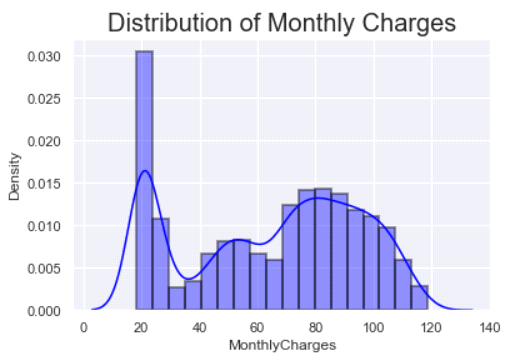
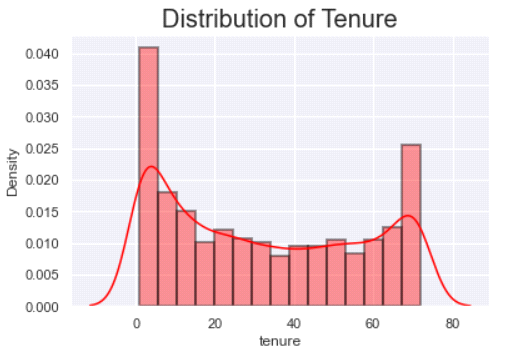
 



Insights:

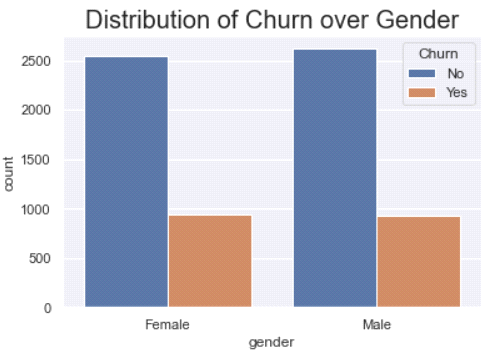
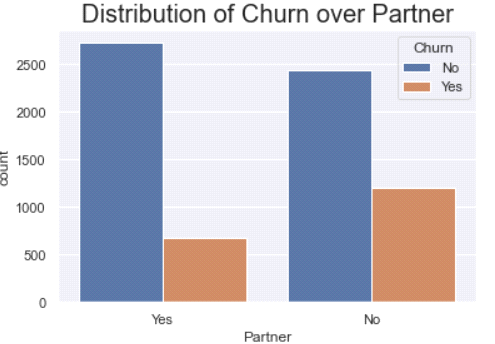
* We've almost 50% distribution of both genders.
* Only 16.21% of customers are senior citizens.
* Almost equal distribution of customers with partners and customers without partners.
* Only 29.96% customers have dependents.
* 90% of customers have phone services.
* Majority(48.13%) of customers don't have multiple lines.
* Majority of customers(43.96%) use Fibre Optic Internet Services.
* almost 50% of customers don't have internet security.
* 43.84% of customers don't have any Online Backup.
* 43.94% of customers don't have device protection.
* 49.31% of customers don't have any Tech Support.
* 39.90% of customers don't have Streaming TV.
* 39.54% of customers don't have Streaming Movies.
* more than 50% of customers opt for Month-to-month contract.
* Almost 60% of customers prefer paperless Billing.
* Electronic Check is the most preferred payment option and credit card is least preferred.
* There is a class imbalance in our target variable.

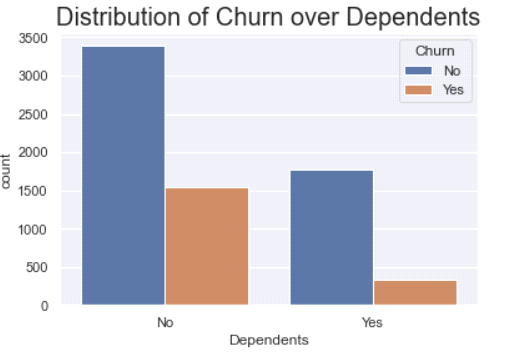
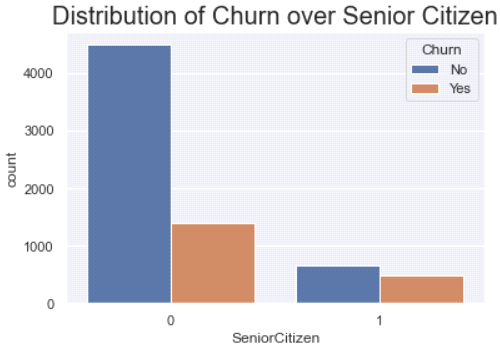
Distribution of continuous variables using seaborn's distribution plot.

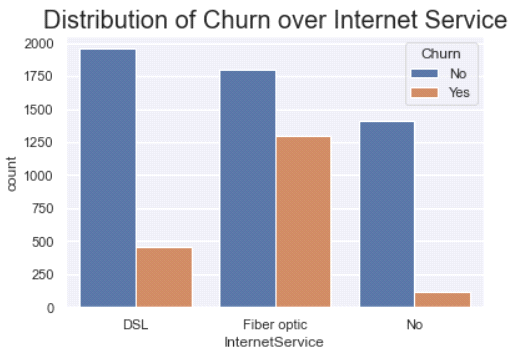
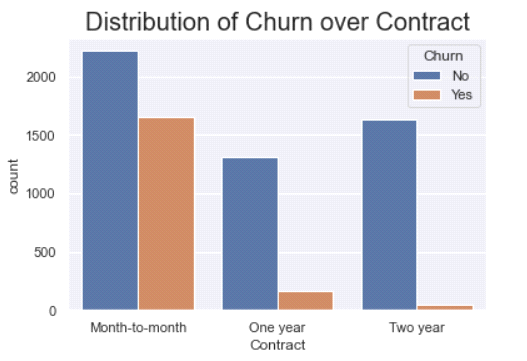


* We can see there's skewness in "TotalCharges". Data is right-skewed.

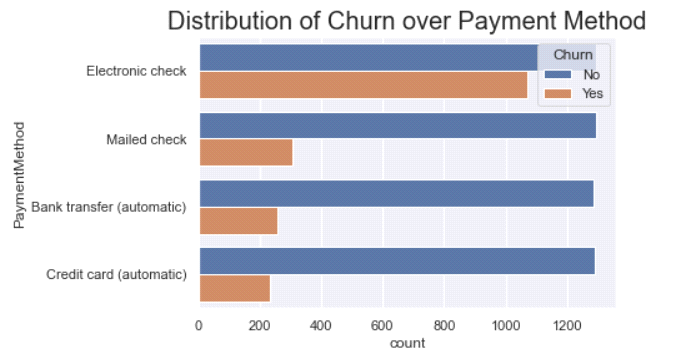
**Bivariate Analysis**: To visualize how our features are affected by our target variable. Also, to analyze correlation between continuous variables.

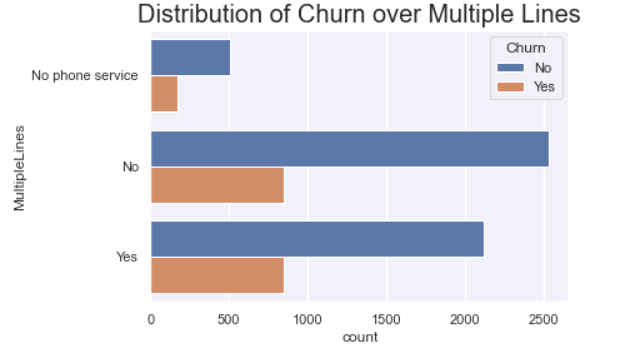
 

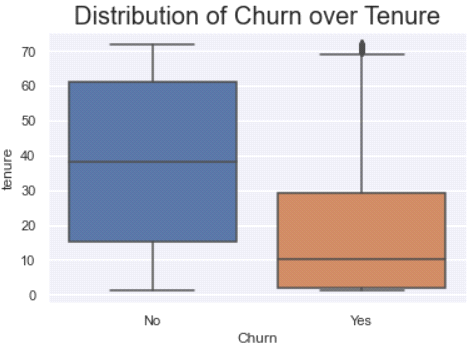
 

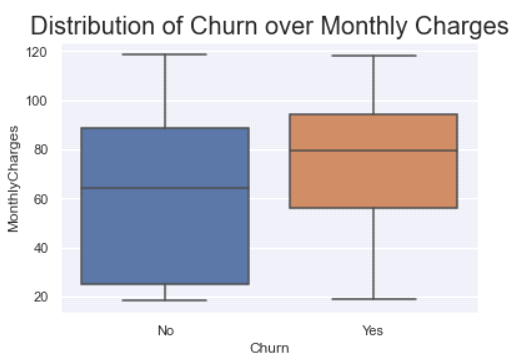
* Churn is likely to be equal in both genders.
* Customers with no partners more likely to cause customer Churn.
* Customers who are not Senior Citizens are more likely to cause customer Churn.
* Customers with no dependents more likely to cause customer Churn.
* Customers with Fibre Optic Internet Services are more likely to cause customer Churn.
* Customers with Month-to-month contract are more likely to cause customer Churn.

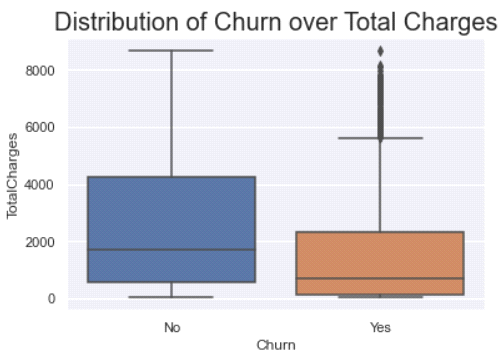




* Customers choosing Electronic Check as payment mode are more likely to cause customer Churn.
* Customers with no Multiple lines are more likely to cause customer Churn.

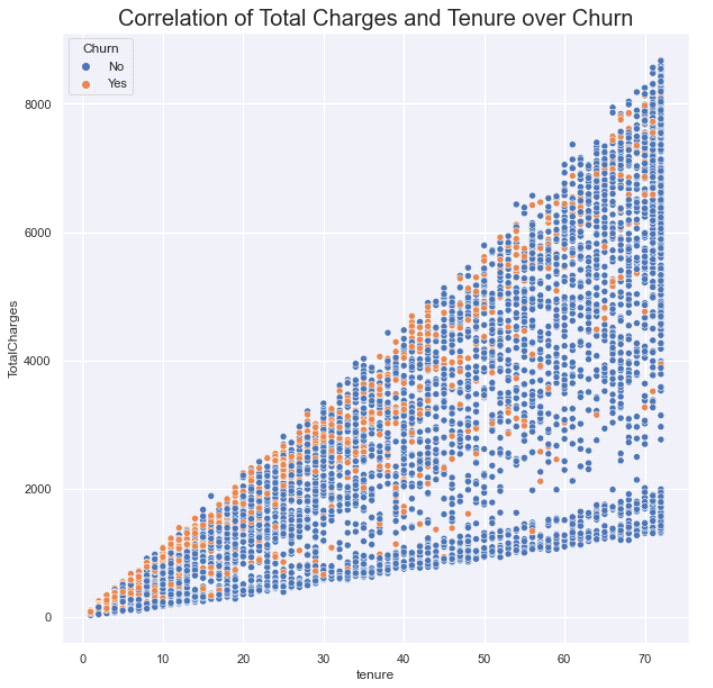


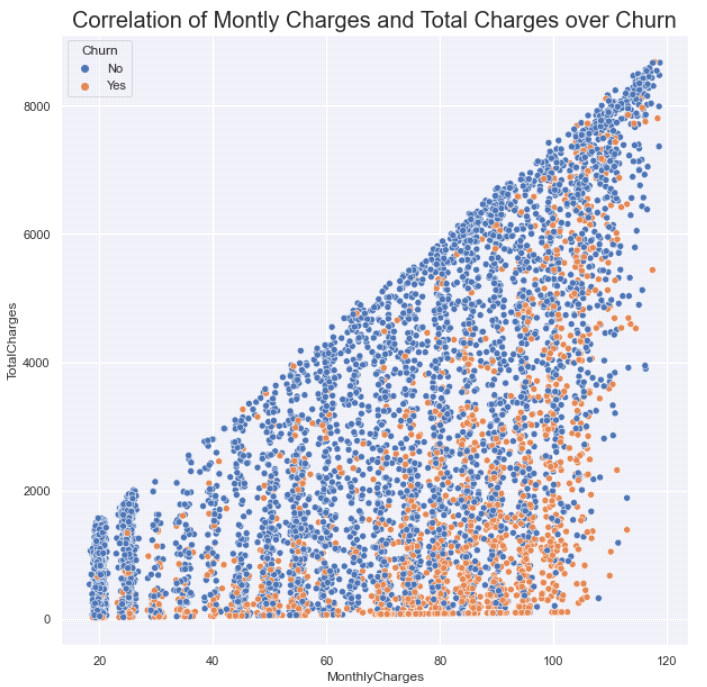




* Customers with Monthly charges between 55 - 85 are more likely to cause customer Churn.
* Customers with Total charges between 0 - 2200 are more likely to cause customer Churn.

**Correlation Analysis** :

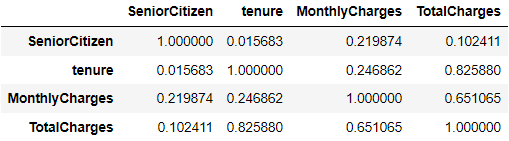


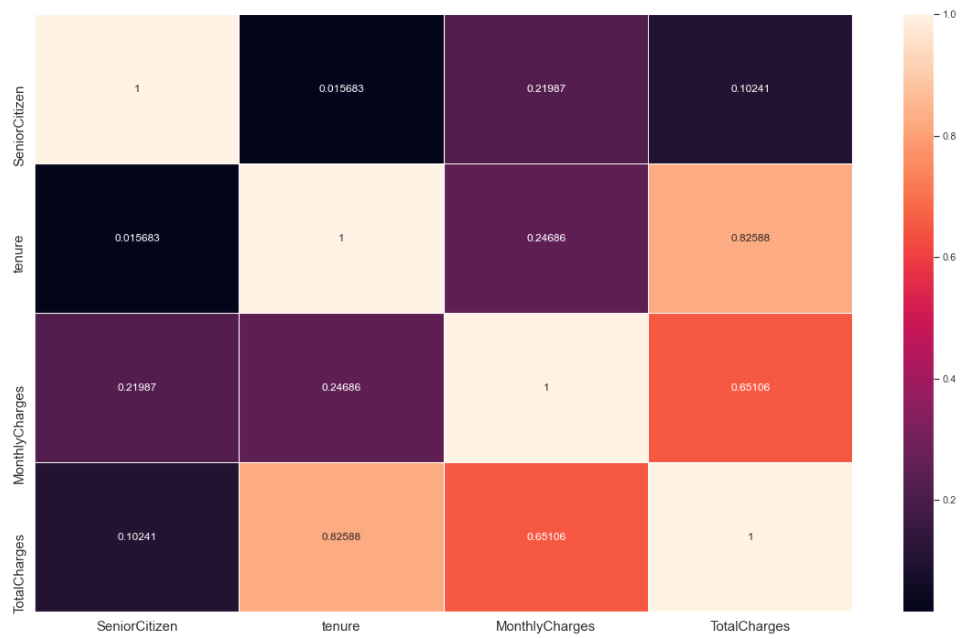


As we can see,

* Churn is more when both tenure and total charges is less. Churn reduces as both Tenure and Total Charges increases.
* Positive Correlation between tenure and Total charges

**Correlation Matrix**



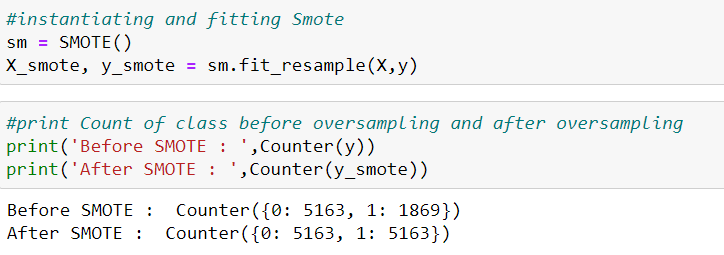


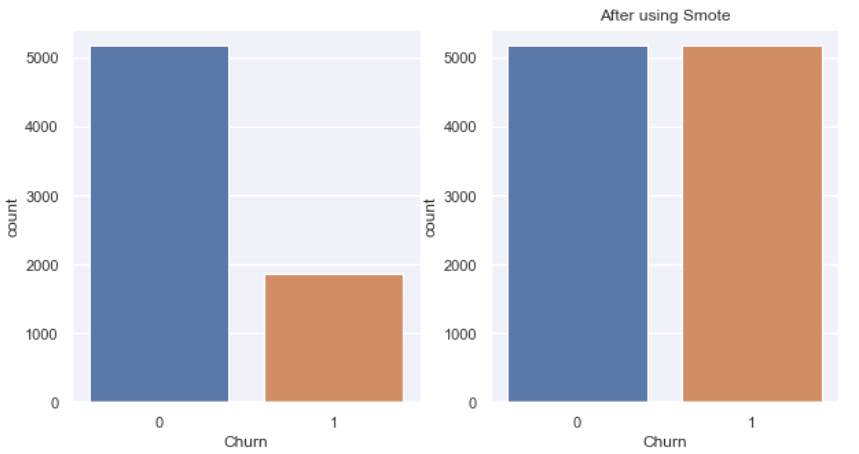
* Monthly Charges and tenure have high positive correlation of 0.82588.
* Monthly Charges and Total Charges have moderate positive correlation of 0.65106.

1. DATA PREPROCESSING

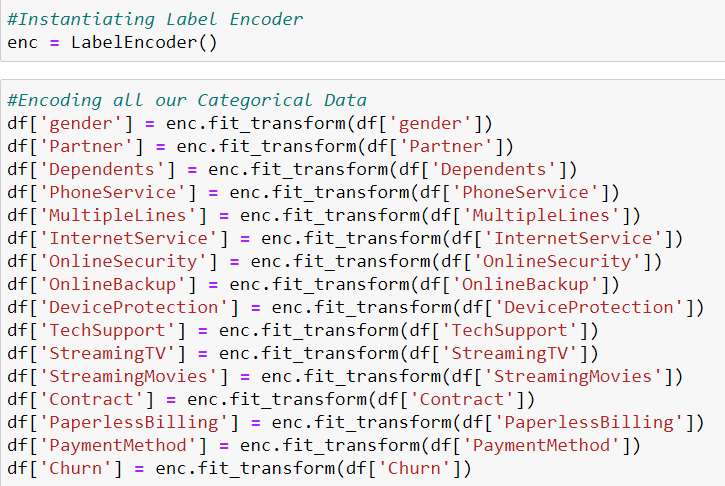
**4.1 Tackling Data Imbalance using SMOTE**

The dataset is imbalanced. It was encountered in our target variable ‘0’(No Churn) has approximately 73.42% records, while, churn ‘1’(Churn) has approximately 26.58% records. In such a case, the danger is that any model fitted to the data might end up predicting the majority class all the time, even though the model diagnostics show that the model is good. To address this class imbalance condition, we adopted the machine learning synthetic minority over-sampling technique (SMOTE). To over-sample the minority classes to achieve fair representation for all classes in the data set.





**4.2 Encoding our categorical variables using Label Encoder**



* 1. Training–Test Set Split

For all the models fitted in this article, I've split the balanced data into 80% for the

training set and 20% for the testing set (validation).

1. Model Building

Scikit-Learn library is used to deploy various machine learning model such as Logistic Regression, KNN, Decision tree Classifier and other ensemble models.

1. MODEL EVALUATION

6.1 Prediction Accuracy

The idea here was to determine which model performs best with our data, and as a

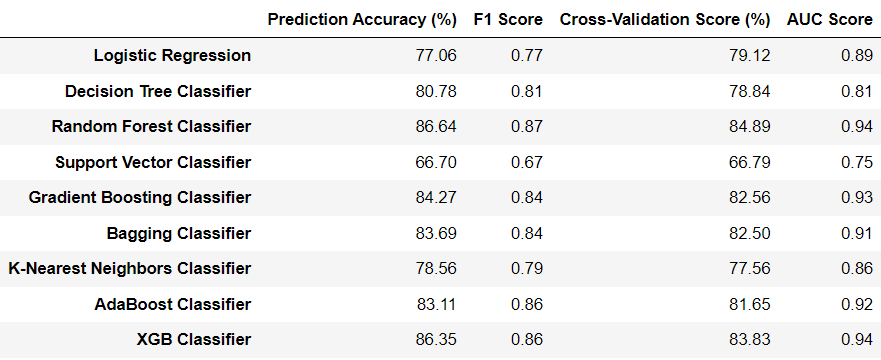
first step, we considered each model’s overall out-of-sample prediction accuracy on the

test set. The least performing model in terms of prediction accuracy was the Logistic Regression and KNN Classifier model.However, note that the best performing models were the machine learning ensemble classifiers (random forest, XGBoost, and Gradient Boost). XGBoost and Random Forest slightly outperformed the Gradient classifier.

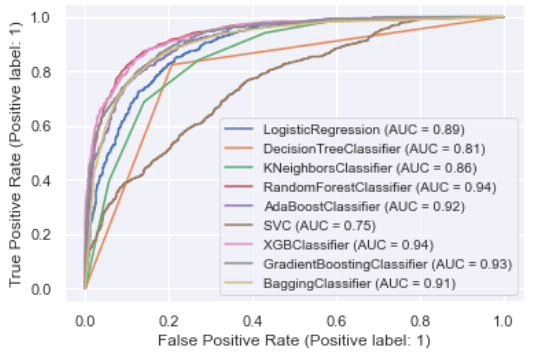
Note that apart from the out-of-sample prediction accuracy on the validation set, other

model diagnostic metrics such as confusion matrix, receiver operating characteristic (ROC) curve, area under the curve (AUC), f1-score, precsion, and recall showed that the ensemble classifiers performed better with our data set than the rest of the models.

***Model Evaluation Metrics***

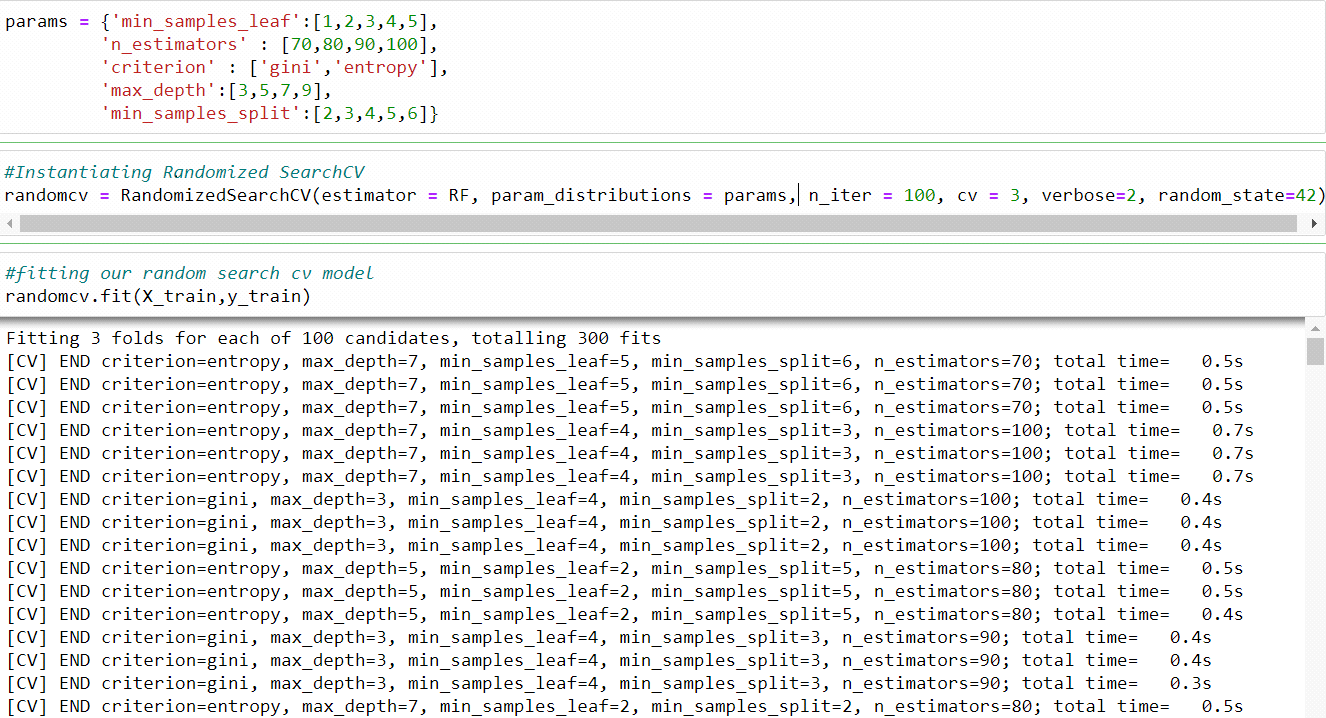


***ROC – AUC CURVE***



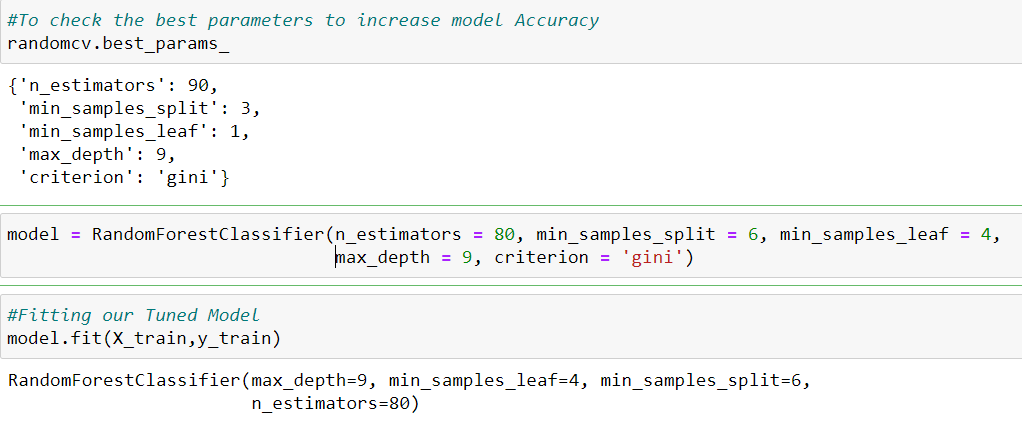
1. Hyper Parameter Tuning :

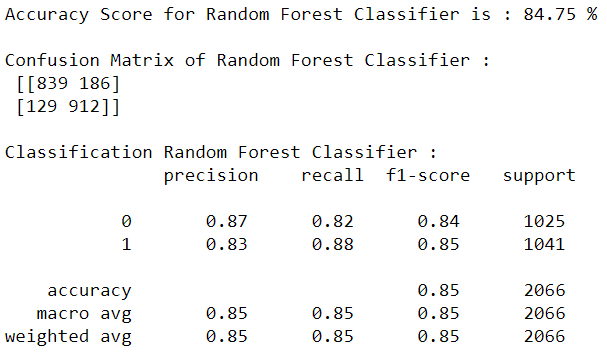
In this subsection, we present the optimal hyper parameters to tune our best model i.e. Random Forest Classifier model with the following hyper parameters:

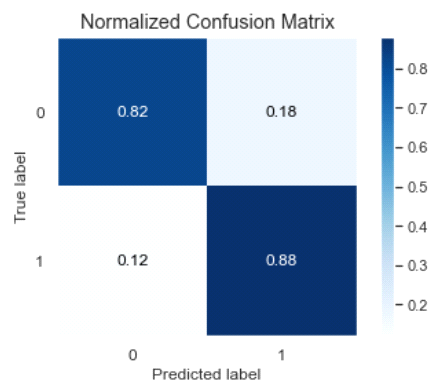


We've deployed Randomized Search CV for Hyper parameter Tuning of our model with 3 cross validations and 100 iterations each which is equivalent to 300 fits.

It was found that our model was over fitted and the final tuned model gave us the prediction accuracy of 84.75%.







1. Conclusion

This article evaluated customer churn status in a telecom industry using various machine learning techniques. Using IBM's sample data, we compared the various machine learning algorithms’ accuracy by performing detailed experimental analysis while classifying individuals' into categories of causing customer churn or not.

Generally, tree-based machine learning algorithms have shown a better performance with our dataset than others, and the most performing models are all ensemble classifiers. It was found that the Random Forest classifier generated the most accurate prediction.

I hope this Article helped you grasp some of the applied machine learning techniques in Data cleaning and preprocessing, applying machine learning techniques, identifying an evaluation metric, hyper parameter and to sum up building a machine learning binary classification model from scratch.

For any further queries please feel free to reach out!

Thank you and happy learning!