Churned Prediction System for Telecommunications Industry

Dhruv Dasadia (A20536411), Dharmik Patel (A20526771)

April 27th, 2024

1 Introduction

As the competition has become more intense, it has become more crucial to retain the existing customers then onboarding the new ones. As the demand of customer increases, so to meet the requirements the service provider makes innovative strategies to lower the customer churn. Through Machine Learning Algorithmic Models, it can analyze and visualize datasets to find the patterns which can be helpful, machine learning models can predict which customers is about to leave. By applying the training and validation process, aim to build a model which can help the telecom's o reduce the churn.

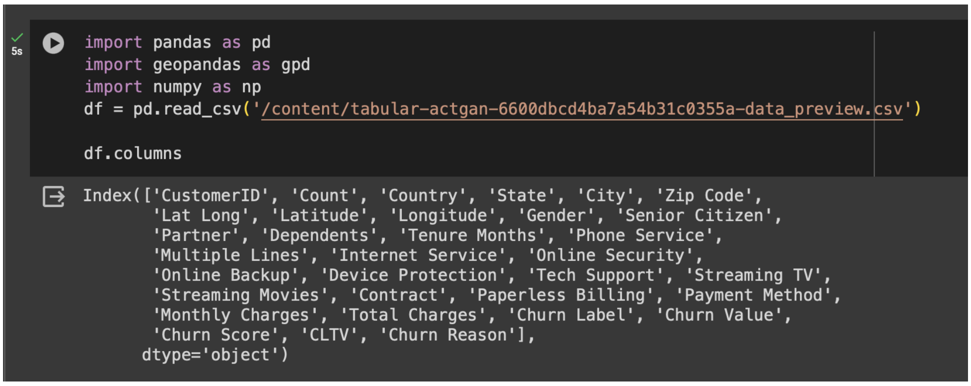
Churn Customers are the numbers of existing customers who may leave the current service provider in a span of time and likely to join other service provider. The main goal of churn is to predict the customers who are likely to churn earliest and to identify the reason for churning. By doing so this will rectify the problems faced by the customers.

The Objective of this project is to enhance the productivity of Machine Learning Algorithms to create a Churn Prediction System which identifies which customer are about to leave. The project will not be only focused on customers which are about to leave but will also be aimed at improving the retention rates and overall customer satisfactions.Through a diligent process of training and validation, the project aims to develop a model which help the telecom companies to decrease the churn. These type of models help telecom industry by making them profitable.

2 Dataset

The dataset used for Training the model contains of Telco Customer Churn dataset which is obtained from Kaggle website, which is of IBM. Each row in the dataset represent the customers and columns describes the attributes, each one of them has features. One of attribute contains two classes in which it is represented by 'T' which indicates that these customers are churning and rest of them as 'F' which are likely to stay with the service provider. The dataset comprises of 16 categorical and 5 numerical columns.

The dataset contains many independent variables which can be classified into three categories.



```
import pandas as pd
import geopandas as gpd
import numpy as np
df = pd.read_csv('/content/tabular-actgan-6600dbcd4ba7a54b31c0355a-data_preview.csv')

df.columns
```

```
Index(['CustomerID', 'Count', 'Country', 'State', 'City', 'Zip Code',
       'Lat Long', 'Latitude', 'Longitude', 'Gender', 'Senior Citizen',
       'Partner', 'Dependents', 'Tenure Months', 'Phone Service',
       'Multiple Lines', 'Internet Service', 'Online Security',
       'Online Backup', 'Device Protection', 'Tech Support', 'Streaming TV',
       'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method',
       'Monthly Charges', 'Total Charges', 'Churn Label', 'Churn Value',
       'Churn Score', 'CLTV', 'Churn Reason'],
      dtype='object')
```

Figure 1: Dataset Columns

- Demographic Information
  - Gender
  - Senior Citizen
  - Partner
  - Dependents

- Account Information
  - Tenure
  - Contract
  - Payment Method
  - Paperless Billing
- Service Information
  - Phone Service
  - Device Protection
  - Tech Support
  - Online Backups
- Above are the few one's which are put into the categories.

3 Methodologies

- **Data Processing**: In this phase, Handling of raw data, and implementing the methods for cleaning, normalization and transformation. Basically it covers the techniques for missing values, outlier identification and normalization to standardize the data, mitigating it for analysis. By doing so a high-quality of dataset which represents the attributes useful for predictive accuracy of the model.
- **Exploratory Data Analysis**: EDA is responsible for gaining insights from dataset's structure and identifying the patterns of churning the customers. During this it will highlight on the summaries, analysis and visualization methods to find the relationship.
- **Feature Selection**: It mainly focuses on identifying the attributes or variables which contributes on predicting the customers churn. It delves into the application of both filter and wrapper methods through assisting the impact on complexity of model.
- **Model Developing**: Main integral part of the project is developing the model which will predict the customers likely to churn, using the specific machine learning algorithms. It focuses on training process which includes techniques such as partitioning of data, cross validation and tuning of parameters, through which it optimizes the model performance.
- **Evaluation**: This module is for measuring the performance of the churn prediction model, by analysis of performance metrics. The evaluation extends beyond the accuracy to consider other metrics which offers a comprehensive view of model performance.
- **Validation**: Validation involves testing the model on a distinct dataset from the one used during training. These process measures the model's performance to ensure it's reliability and its ability to generalize to unseen data. It is designed to prevent the overfitting and to ensure model's robustness.
- **Comparison of Results**: An analysis of various model based on the performance metrics which highlights on Strengths and weakness of each model. The most effective model for prediction can be chose based on the evaluation.
- **Conclusion**: Concluding, our model predicting ability to identify customers at risk of leaving. Based on the findings the customer retention strategies can be improved and put into the consideration. It suggests the improvements in churn prediction methods.

4 Milestone

Below are the Milestones for the project.

1. **Data Preprocessing**
   - **Data Extraction:** First Step where data is gathered and loading the necessary data into a single format to present a comprehensive view.
   - **Data Augmentation:** Increasing the diversity of data by generating synthetic data or creating new features to enhance the robustness of Model.

- **Data Balancing:** To prevent the bias in the Model's Prediction, which in general favors the majority class. So techniques such as oversampling the minority or undersampling the majority class.

2. **Exploratory Data Analysis**
   - **Data Exploration with Visual:** Creating visualizations to understand the dataset, it is done by identifying the patterns, detecting the outliers
   - **Finding Correlation:** Plotting the correlation matrix and calculating the Coefficients and finding the relationship between the variable and identify which one has the strong relationship with customer churn.
   - **Feature Extraction:** Developing new features on the context of the existing one which can help the model's ability to predict outcomes.
   - **Preparing Final Dataset split for Modelling:** This step involves splitting the dataset into training, testing and validation. By doing so it helps the model to tune the parameters and evaluate the model's performance to determine how well it generalizes to unseen data.
3. **Modelling**
   - **Algorithm Selection:** Selecting the right ones which are best fitted for the model, which provides good interpretability and performance.
   - **Ensembling Multiple Models:** Integrating multiple models which significantly boost precision and reliability. Techniques are used to deal with the underfitting or overfitting by changing the variance and bias.
   - **Determining the best Model:** Once the development and integrating the models, next step is to assess their performance through the evaluation metrics.
4. **Model Evaluation**
   - Evaluation is not only judged on the overall accuracy of the model, but also how well it identifies True Positives and True Negatives; customers who are going to churn and the retained customers.
5. **Optimization**
   - Adjusting or Tuning the model's parameters to improve the accuracy of model on unseen data and generalizes it well by preventing the overfitting of data, making the model reliable.
6. **Conclusion**
   - It implicates all the steps from Data Preprocessing to Evaluation of Model by stating the predicitve performance and enhancing the Customer Retention Strategies.

5 Implementation

1. **Data Augmentation**
   - Due to the limitations of data we need to synthetically generate new data via different model which generates data based on the existing.
2. **Exploratory Data Analysis**
   - **Box Plot**
     - During the process we tried to understand the dataset and what patterns are implicate.
       We started it by plotting the box plot which is categorised by the Contract, where Customers are bound to service provider. The Contract is categorised into three types which are Two Year, One Year and Month-to-Month.
     - The distribution is calculated by calculating the charge by multiply of monthly charge and tenure months that the customer used the service, it is then subtracted by total charge to give the difference in charge.

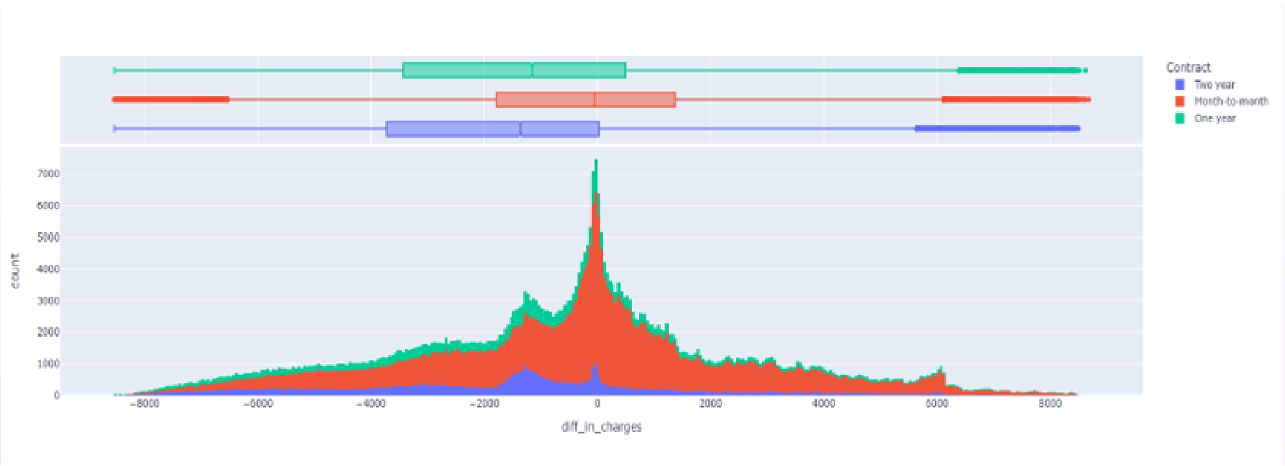


Figure 2: Box Plot: Distribution of Charges over the Contract Types

- **Churn Rate By Contract Type**
  - Chart display that the churn rate by contract type, which means that the customers are in the contract type of month-to-month have high churn rate then the other 2 contract types One year and Two Year. The other pie chart shows the churn rate for the customers whose churn label is equal to 'No'.

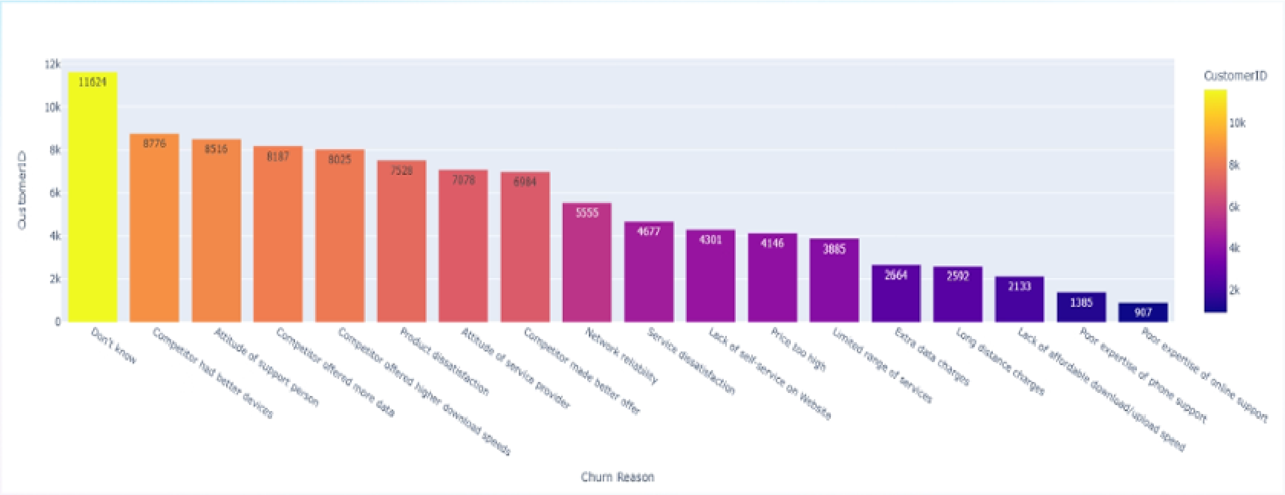


Figure 3: Bar Chart: Reasons for Customer Churn

- **Heat Map**
  - Heat Map Visualize the correlation between the the services used by customers and churn. The cells which are dark red are associated with high churn rates.
  - The Dark red cells indicate a strong positive correlation related to 1 and one with dark blue a negative correlation equal to -1.
  - Customers with no tech support are more likely to churn, then customer with online security are less likely to churn.

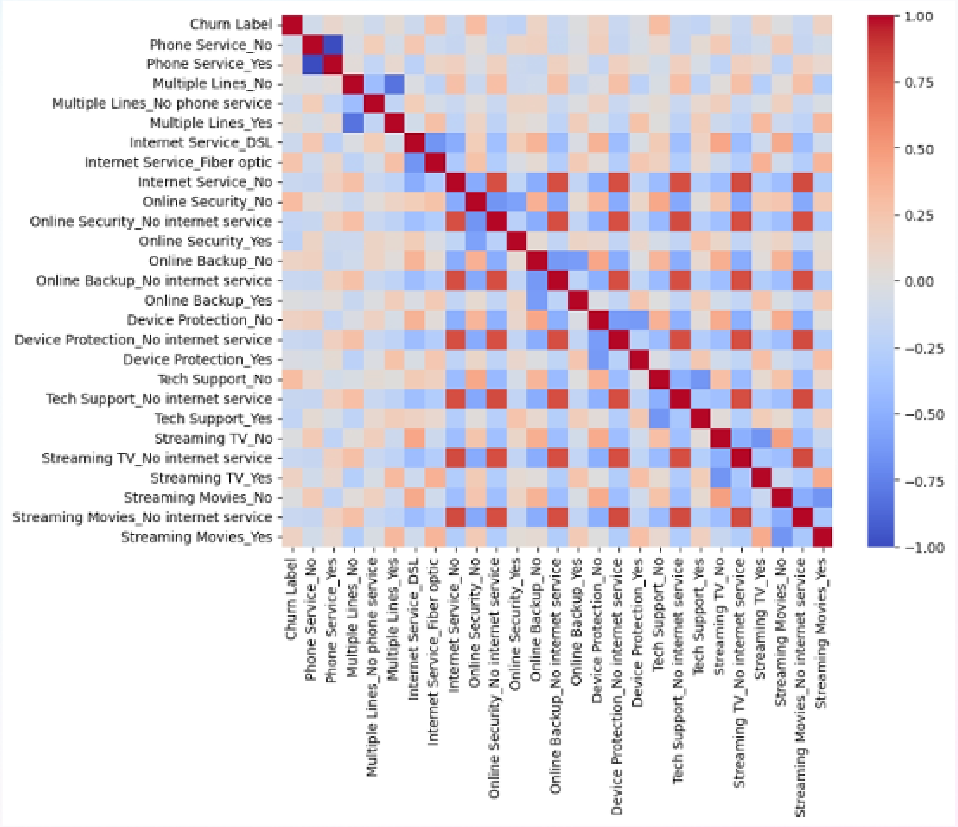


Figure 4: Heat Map: Features vs Churn

- **Latitude and Longitude Map**
  - The map in Figure visualizes customer churn rates which is represented geographically, which employs the hexagonal binning for spatial analysis. Hexagons are represented as color-coded according to gradient colors which will be ranging from green to orange and then to red, showcasing the churn rates from low to high.
  - Map is plotted on the latitude and longitude which is given in the dataset, and each hex bin represent the churn rate and customer count.

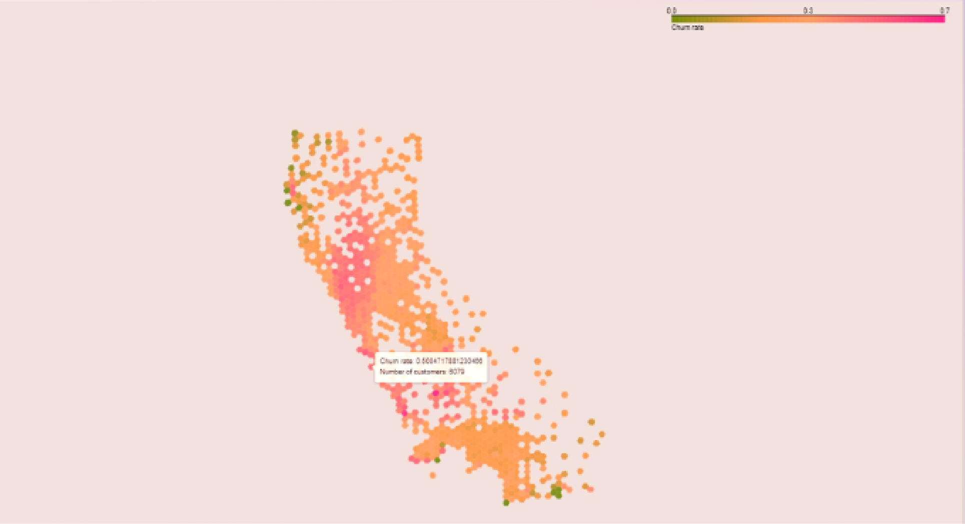


Figure 5: Scatter Plot: Probability of Churn

- **Customer Density Analysis Across Geographical Coordinate**
  - Map indicates the scatter plot of customer distribution with the help of Latitude and Longitude. The blue color represents the customer count in that coordinate.

– Map is plotted on the latitude and longitude which is given in the dataset, and each hex bin represent the churn rate and customer count.

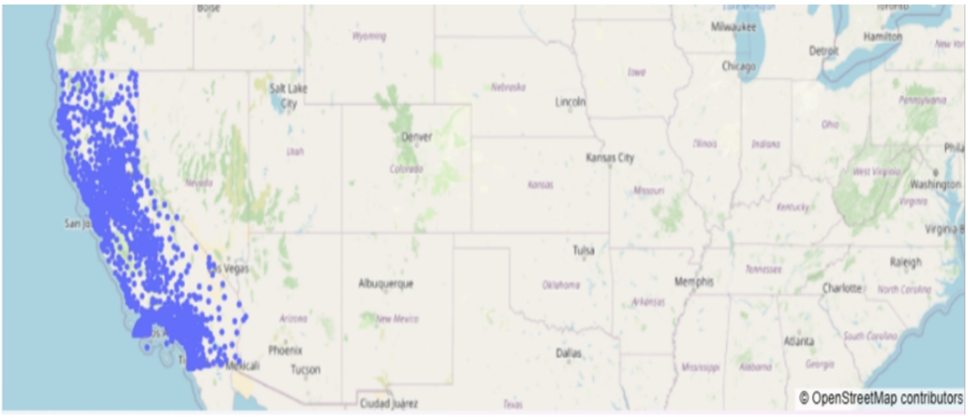
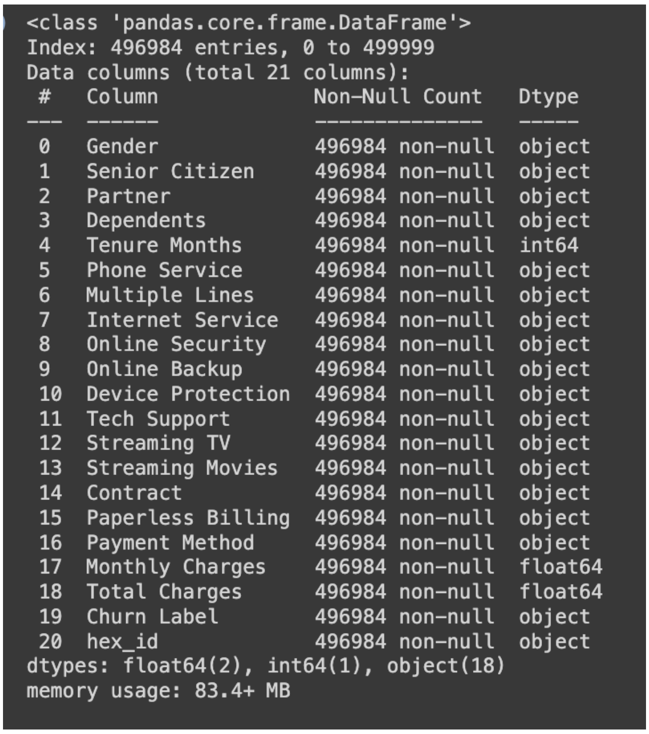


Figure 6: Geographic Distribution Map

3. **Feature Engineering**

- This process involves removing the columns from the dataset which are unnecessary for the prediction. Columns which we removed are 'Country', 'State', 'Count', 'ZipCode', 'City', 'Longitude', 'Latitude', 'Churn Score', 'Churn Value', 'Churn Reason', 'CustomerId'. By doing so it helps to reduce dimensionality which simplifies the model and improves the performance which focuses on the important attributes.
- Thus removing the columns aids in addressing the potential issues which might arise from including the unnecessary variables.

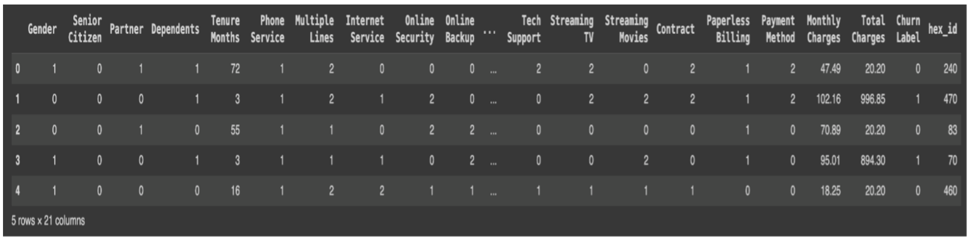


```
<class 'pandas.core.frame.DataFrame'>
Index: 496984 entries, 0 to 499999
Data columns (total 21 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   Gender             496984 non-null  object 
 1   Senior Citizen     496984 non-null  object 
 2   Partner            496984 non-null  object 
 3   Dependents         496984 non-null  object 
 4   Tenure Months      496984 non-null  int64  
 5   Phone Service      496984 non-null  object 
 6   Multiple Lines     496984 non-null  object 
 7   Internet Service   496984 non-null  object 
 8   Online Security    496984 non-null  object 
 9   Online Backup      496984 non-null  object 
 10  Device Protection  496984 non-null  object 
 11  Tech Support       496984 non-null  object 
 12  Streaming TV       496984 non-null  object 
 13  Streaming Movies   496984 non-null  object 
 14  Contract           496984 non-null  object 
 15  Paperless Billing  496984 non-null  object 
 16  Payment Method     496984 non-null  object 
 17  Monthly Charges    496984 non-null  float64
 18  Total Charges      496984 non-null  float64
 19  Churn Label        496984 non-null  object 
 20  hex_id             496984 non-null  object 
dtypes: float64(2), int64(1), object(18)
memory usage: 83.4+ MB
```

Figure 7: Column Description

- The feature engineering process outlined in the given code which mainly focuses on converting categorical variables into numerical for preparing the data for machine learning models. The column 'Churn Label' indicates whether a customer churned or not, which is represented in binary values. So, the binary values convert to 'Yes' and 'No' for '1' and '0'.
- A custom function named 'encode data' is defined to automate the encoding process for categorical variables. This function utilizes the Label Encoder from the scikit-learn library to transform categori-

cal variables into numeric labels. By encoding categorical to numerical values, the data becomes more compatible with the algorithms, which facilitates the training and evaluation of predictive models. Hence the Feature Engineering Steps ensures that it is ready for analysis and modeling.

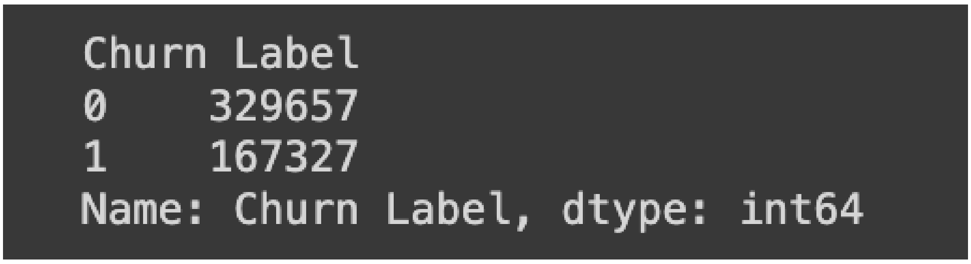


	Gender	Senior Citizen	Partner	Dependents	Tenure Months	Phone Service	Multiple Lines	Internet Service	Online Security	Online Backup	...	Tech Support	Streaming TV	Streaming Movies	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	hex_id
0	1	0	1	1	72	1	2	0	0	0	...	2	2	0	2	1	2	47.49	20.20	0	240
1	0	0	0	1	3	1	2	1	2	0	...	0	2	2	2	1	2	102.16	996.85	1	470
2	0	0	1	0	55	1	1	0	2	2	...	0	0	0	0	1	0	70.89	20.20	0	83
3	1	0	0	1	3	1	1	1	0	2	...	0	0	2	0	1	0	95.01	894.30	1	70
4	1	0	0	0	16	1	2	2	1	1	...	1	1	1	1	0	0	18.25	20.20	0	460

5 rows x 21 columns

Figure 8: Column Description

- Data Balancing process outlined in the code involving Synthetic Minority Over Sampling Techniques which is used to address class imbalance in the dataset. Class imbalance occurs generally when one class is underrepresented compared to the other class. SMOTE is a popular technique used to alleviate class imbalance which is generated through synthetic samples of the minority class. This ensures that the predictive model is trained on a more balanced dataset by mitigating the bias towards the majority class and help to improve the models ability on generalization on unseen data.
- The pre-processed data is then ready for training models which can effectively capture the underlying patterns in the dataset, which leads to accurate predictions mainly for the minority class.



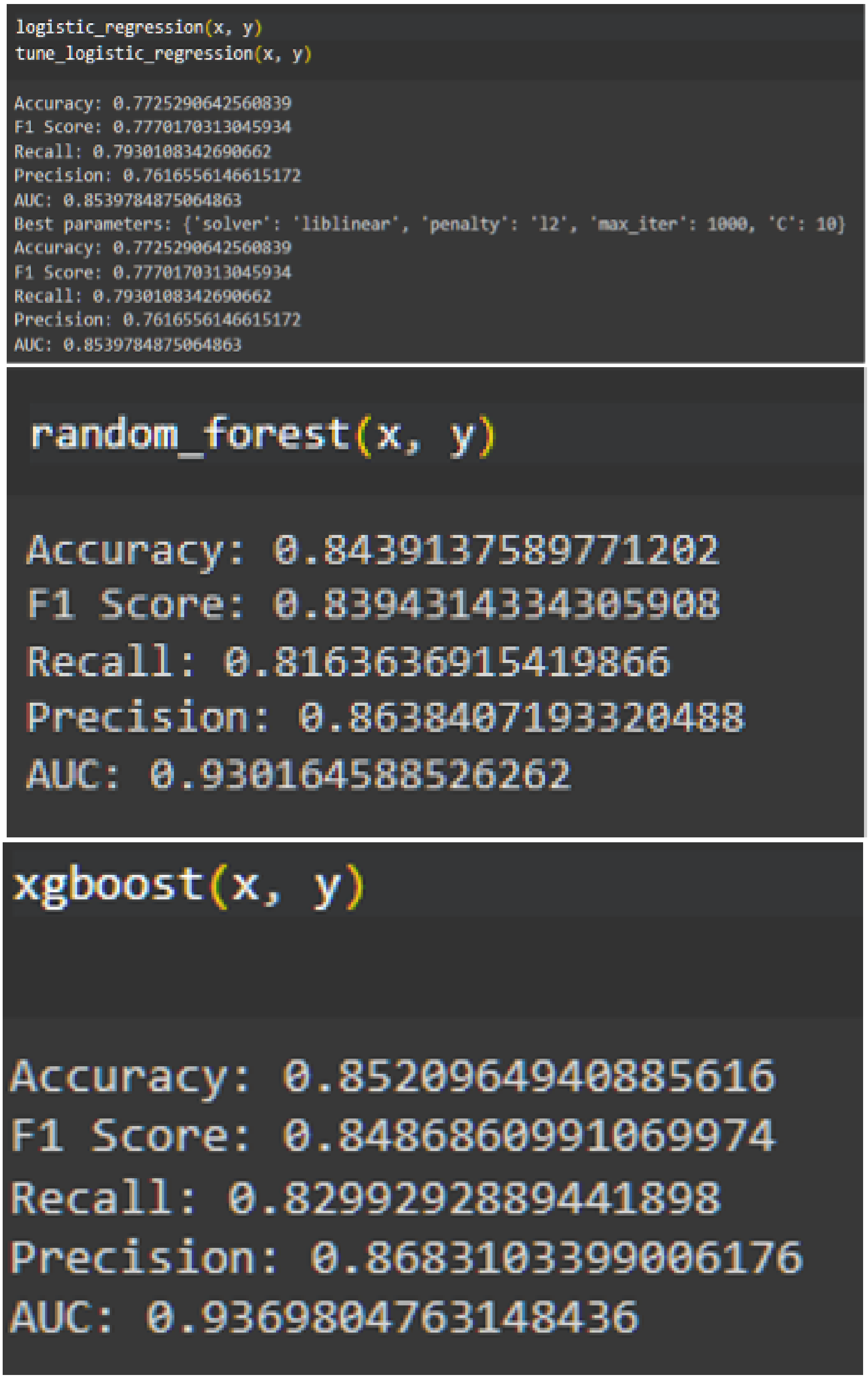
```
Churn Label
0    329657
1    167327
Name: Churn Label, dtype: int64
```

Figure 9: Churn Label

4. **Modeling and Tuning**

   - We implemented various tuning methods to optimize the performance of the models. The techniques are feature engineering, hyper-parameter tuning and handling the missing values. Analyzed the impact of each and every method on model performance.
   - **Models Used:**
     - Logistic Regression
     - Random Forest
     - XG-Boost
     - Each model is used based on its ability to analyse the various aspects of the data. Analyzing each models strengths and Weakness providing the insights into their performance.
   - **Model Evaluation**
     - Process of evaluation begins with the splitting the datasets into training and testing. Split is in the form of 80:20, which ensures that the models are trained on the particular portion of the dataset, and then evaluated on unseen data which enables the model's generalization capability.
     - Machine Learning Algorithms which we defined in the above step are selected for evaluation. For optimization of Model's Performance, hyper-parameter tuning is applied using RandomizedSearchCV, this technique helps to explore the range of values for finding the combination that maximizes performance metrics. With the help of hyper parameters, every model is trained on the training set which allows the model to learn the patterns and relationship within the data.
     - Models Performance is then evaluated through the multiple metrics, including accuracy, F1 score, recall, precision, and AUC, which provides the insights into various aspects of model performance, such as classification accuracy, balance between precision and recall.

- The process of model evaluation ensures that the algorithms are assessed and compared based on the performance on unseen data, making the decision on model selection and deployment.



```
logistic_regression(x, y)
tune_logistic_regression(x, y)

Accuracy: 0.7725290642560839
F1 Score: 0.7770170313045934
Recall: 0.7930108342690662
Precision: 0.7616556146615172
AUC: 0.8539784875064863
Best parameters: {'solver': 'liblinear', 'penalty': 'l2', 'max_iter': 1000, 'C': 10}
Accuracy: 0.7725290642560839
F1 Score: 0.7770170313045934
Recall: 0.7930108342690662
Precision: 0.7616556146615172
AUC: 0.8539784875064863
```

```
random_forest(x, y)

Accuracy: 0.8439137589771202
F1 Score: 0.8394314334305908
Recall: 0.8163636915419866
Precision: 0.8638407193320488
AUC: 0.930164588526262
```

```
xgboost(x, y)

Accuracy: 0.8520964940885616
F1 Score: 0.8486860991069974
Recall: 0.8299292889441898
Precision: 0.8683103399006176
AUC: 0.9369804763148436
```

Figure 10: Learning Models

- Three algorithms for Telcommunication
- Logistic Regression:
  * Simple and interpretable model.
  * Suitable for binary classification problems such as churn prediction.
  * Provides probabilities of churn, which can be useful for risk assessment.
- Random Forest:

    - ∗ Ensemble learning algorithm that combines multiple decision trees.
    - ∗ Robust to over-fitting which can handle non-linear relationships.
    - ∗ Provides feature importance scores, which can help identify key drivers of churn.
  - Extreme Gradient Boosting Classifier (XGBoost):
    - ∗ Gradient boosting algorithm that builds an ensemble of weak learners.
    - ∗ Highly accurate and efficient, often outperforming other algorithms in prediction.
    - ∗ Can handle both linear and non-linear relationships, making it suitable for churn datasets.

- **ROC Metric**
  - ROC Curves generated provides a comprehensive visualization of the performance of three models - Logistic Regression, Random Forest and XG-Boost. The curve represents the trade-off between the 'True Positive Rate' and 'False Positive Rate'. The ROC curve represent the model's ability between positive and negative instances and the AUC values higher indicating the better performance. Through the positions of ROC curves, one can assess the effectiveness of every model by differentiating between the churners and non-churners. The diagonal line represents the random guessing and the model lying above this line states predictive superiority.

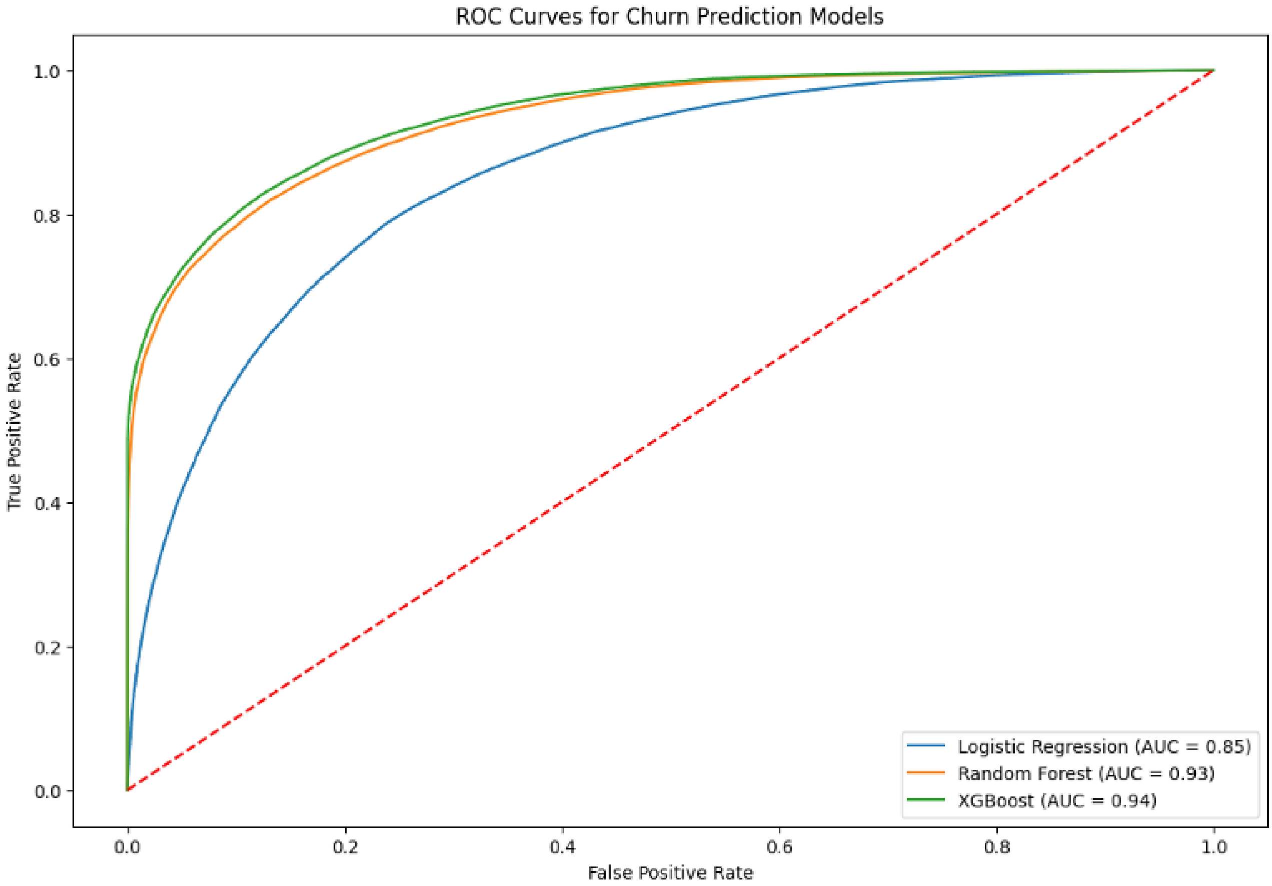


Figure 11: ROC Curve Comparision

5. **Model Comparision**

- Comparison model illustrates the performance of three models which we have implemented in the code for Model Learning and Evaluation, which are Logistic Regression, Random Forest and XG-Boost. Each model's performance is evaluated through the Evaluation Metrics such as Accuracy, F1 Score, Precision, Recall and Area under the Curve.
- The figure below elaborates the model performance on each of the evaluation metrics, which showcase their strengths and weakness. By doing so it helps to determine which model is best suited and has surpass all the metrics. Besides that it also helps to find which model is best suitable for the classification task, and appropriate for the scenario.

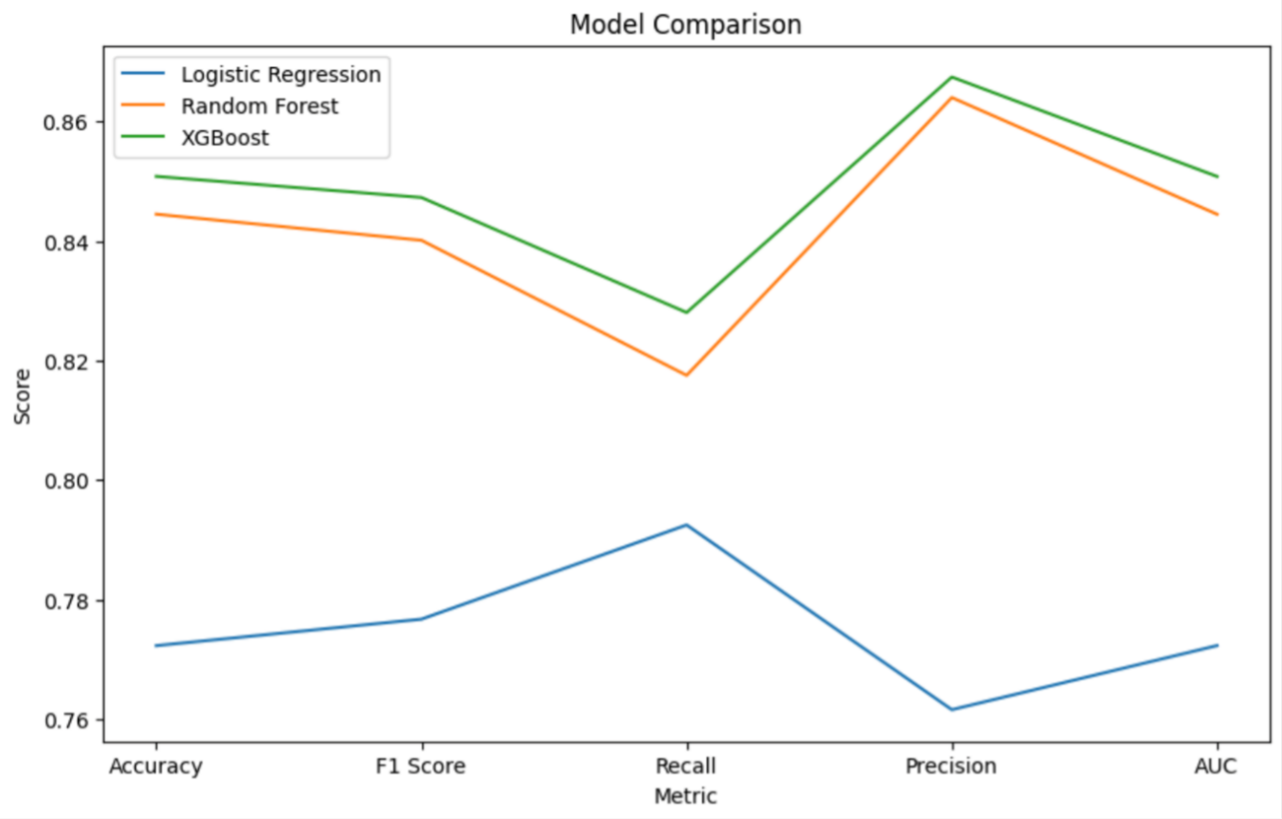


Figure 12: Model Performance Comparision

6. **Confusion Matrix**

   - Confusion Matrix represents the model's classification accuracy where each matrix displays the count of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative(FN).
   - In the below Figure the counts along the main diagonal indicates the accurate predictions and the one which is off diagonal shows the miss-classifications. Additionally we can identify the class imbalance between the classes, which helps to the adjustments and improve the model performance.

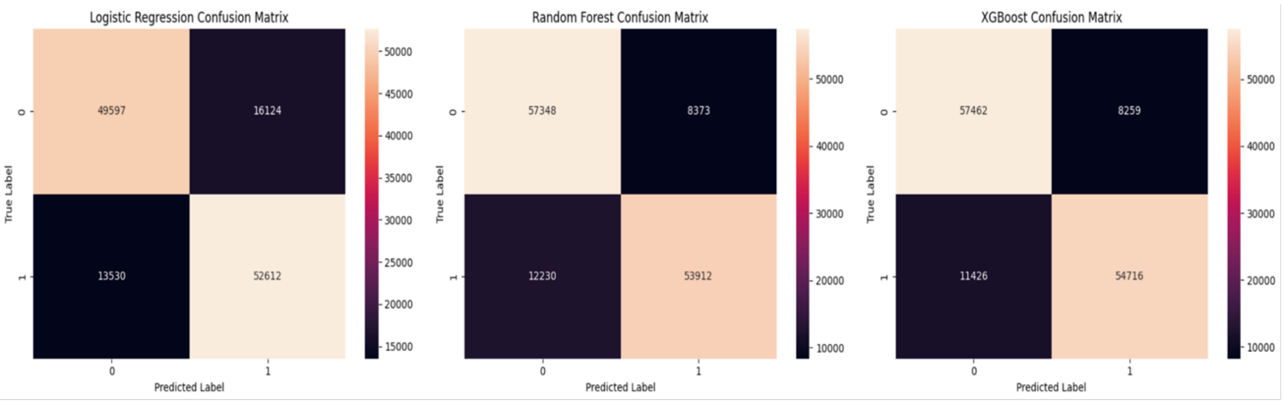


Figure 13: Confusion Matrix

7. **Error Analysis**

   - Over here, the error analysis helps to examine the types of errors the model is prone too, and identifying any patterns which lead to the errors. Using the Error Analyzing, we can exhibit the opportunities to process the model by tuning the parameters or implementing the ensemble methods to increase the generalization of the model on unseen data.

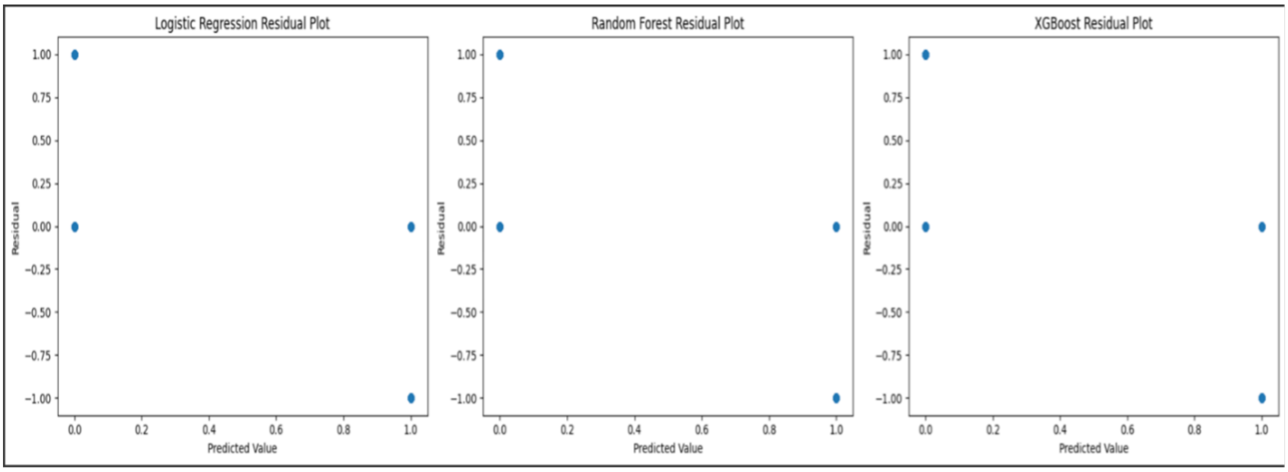


Figure 14: Residual Plot: For Prediction Models

6 Conclusion

- In this Project, we explored the Telecommunication Churn Prediction, where we started by analyzing the dataset, then performing Feature Engineering, and lastly comparison of three learning models to see which is the best fit for the dataset in terms of accuracy.
- Models used were Logistic Regression, Random Forest and XG-Boost out of three models XG-Boost outperformed others with Accuracy of 85.2 percentage, F1 Score of 85.3 percentage, and precision of 85.4 percentage and recall of 85.2 percentage. ROC curve illustrated that the XG-Boost has the highest AUC value of 0.92, which has a strong ability to distinguish between churn and non-churn customers.
- Error Analysis revealed that all the models showcased the similar patterns of errors, which suggests that the model may benefit from tuning or extra features o improve the accuracy for the customer segments.
- To Conclude, we can state that the XG-Boost Model showed better performance, which helped for churn prediction in telecommunication industry, and further it can be optimized to provide the valuable insights for Customer Retention Strategies.

7 References

1 SharmilaK.Wagh,AishwaryaA.Andhale,KishorS.Wagh, Jayshree R. Pansare, Sarita P. Ambadekar, S.H. Gawande, Customer churn prediction in telecom sector using machine learning techniques.

2 AbhishekGaur,RatneshDubey,PredictingCustomerChurn Prediction in Telecom Sector Using Various Machine Learning Techniques.

3 Kaggle Dataset Link

4 Gretel for generating Synthetic Data based on the dataset available.

5 Link to Github Reporitory

6 Google Drive Link for all Files