Report on Customer Churn Prediction

SubaseData Assessment By - Anshika Singh

This Dataset is consist of customers basic information

CustomerID: A unique identifier for each customer.

Name: The name of the customer.

Age: The age of the customer. Gender: The gender of the customer (male or female).

Location: The geographic location where the customer is based.

Subscription_Length_Months: The number of months the customer has been subscribed to the service.

Monthly_Bill: The monthly bill amount for the customer.

Total_Usage_GB: The total usage of the service in gigabytes.

Churn: A binary indicator (1 or 0) representing whether the customer has churned (1) or not (0).

Importing libraries

Pandas: Pandas is a fundamental library for data manipulation and analysis, making it an essential tool for EDA. It provides data structures like DataFrames that facilitate data exploration and transformation.

Matplotlib: Matplotlib is a popular data visualization library for creating static, animated, or interactive plots and charts. It's often used to visualize data distributions, relationships, and trends.

Seaborn: Seaborn is built on top of Matplotlib and provides a high-level interface for creating aesthetically pleasing statistical graphics. It simplifies the process of creating complex visualizations.

NumPy: NumPy is used for numerical computing in Python. It provides support for handling arrays and matrices, which is crucial for performing mathematical operations on data.

Information we got after performing EDA

- **Shape of the dataset** --> (100000,9)
- **Null-Values** --> 0

- **Numerical** --> Index(['CustomerID', 'Age', 'Subscription_Length_Months', 'Monthly_Bill', 'Total_Usage_GB', 'Churn']
- **Categorical** --> Index(['Gender', 'Location']

- **Outliers** --> No outliers

- **Dropping Unnecessary columns** --> CustomerID, Name ***(for this specific dataset these were unnecessary.)***
  Creating ones, and zeros from categorical variables.

- **CHURN (count)** **0 50221**
  1 49779

Statistical Information of dataset

AGE - Average age of the customers is approximately 44(Years). Youngest person is 18 (Years) old. Oldest person is 70 (years) old.

Subscription_Length_Months - On an average, customers have a subscription length of around 12.5 months. The majority of customers have subscription lengths from 6 to 19 months.

Monthly_Bill - Average monthly bill is approx. $65. Customers pay between approximately 47.50 Dollar to 82.64 Dollar per month.

Churn - Churn values are binary - 1 and 0, indicating whether a customer has churned or not. Churn rate is evenly distributed due to a mean close to around 0.5.

Total_Usage_GB - The average total usage is about 274.4 GB. Total usage varies between 50 GB and 500 GB. Most customers have total usage between 161 GB and 387 GB.

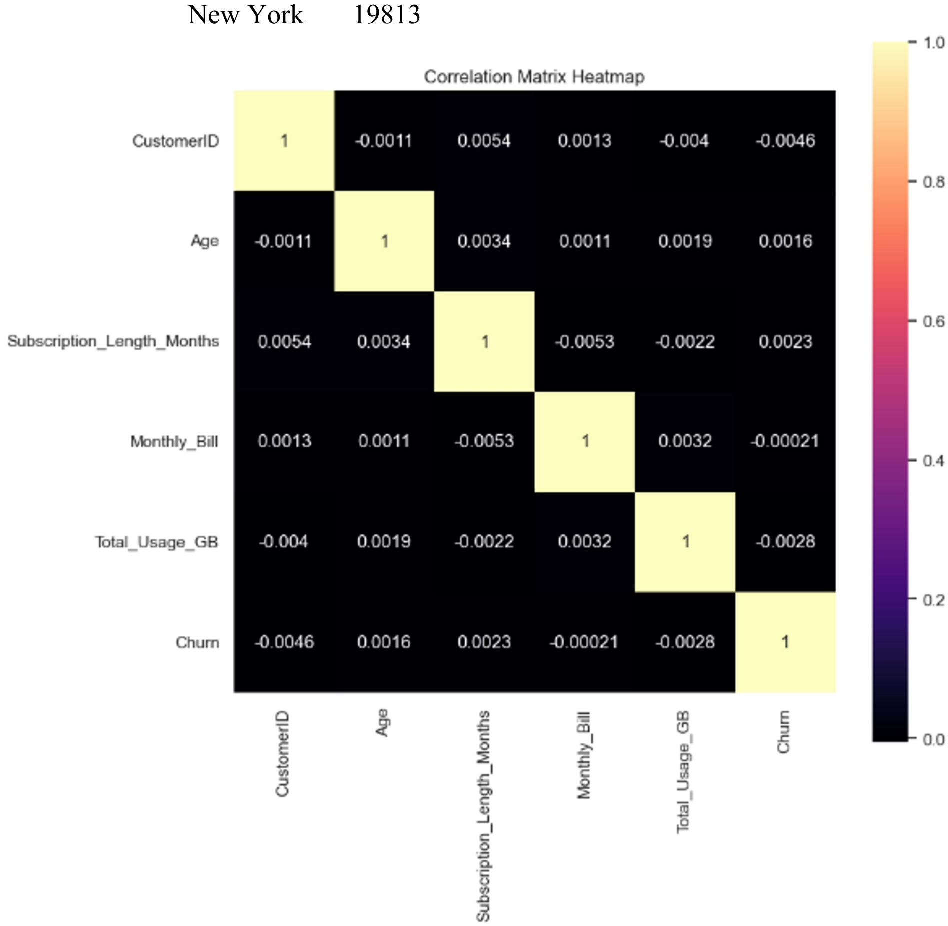
Count of categorical variables

Name: Gender

Female	50216
Male	49784

Name: Location

Houston	20157
Los Angeles	20041
Miami	20031
Chicago	19958
New York	19813



Heatmap Explanation

The darker the color in the heatmap, the stronger the correlation between the two variables. For example, the correlation between Subscription_Length_Months and Churn is very strong, as indicated by the dark blue color in the heatmap. This means that customers who have been subscribed for a longer period of time are less likely to churn.

- A correlation of 1 indicates a perfect positive correlation, while a correlation of -1 indicates a perfect negative correlation. A correlation of 0 indicates no correlation.
- A strong positive correlation between Monthly_Bill and Total_Usage_GB. This means that customers who have a higher monthly bill tend to use more of the service.
- A weak negative correlation between Age and Churn. This means that older customers are slightly more likely to churn than younger customers.
- A weak positive correlation between Gender and Churn. This means that female customers are slightly more likely to churn than male customers.

Results

	Algorithm	Accuracy	Precision	Recall	F1-score
0	LogisticRegression	0.505067	0.255092	0.505067	0.338978
1	DecisionTreeClassifier	0.499833	0.499863	0.499833	0.499844
2	KNeighborsClassifier	0.499567	0.499606	0.499567	0.499579
3	GradientBoostingClassifier	0.503200	0.502857	0.503200	0.502398
4	RandomForestClassifier	0.498933	0.498712	0.498933	0.498591
5	SVC	0.506500	0.506308	0.506500	0.422036

Final Model Selection

	Accuracy	Precision	Recall	F1 Score	Building Time
Gradient Boosting	0.529529	0.530200	0.529529	0.526217	5.672780
XGBoost	0.634829	0.634875	0.634829	0.634779	0.394677

- *XGBoost performed quite better.*

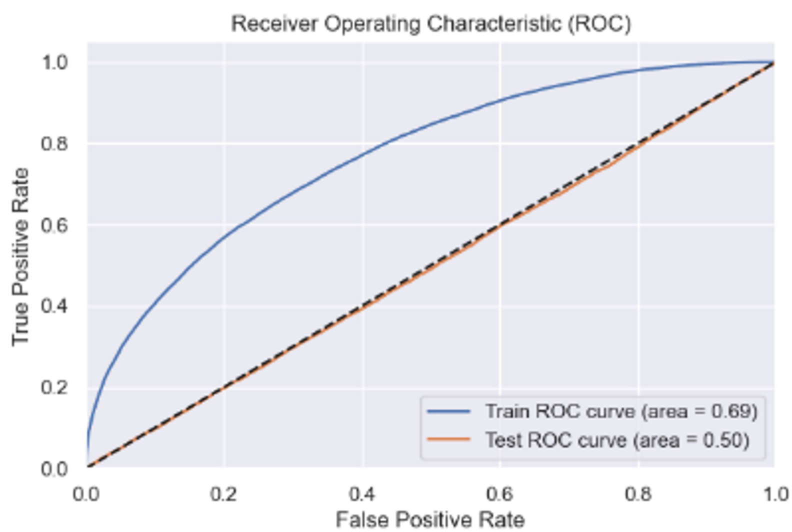
Performance metric

	Metric	Train	Test
0	Accuracy	0.689186	0.495767
1	Precision	0.691382	0.490442
2	Recall	0.681229	0.482085
3	F1-Score	0.686268	0.486228

Dataset	Accuracy	Precision	Recall	F1-score	
0	Train	0.689186	0.689222	0.689170	0.689159
1	Test	0.495767	0.495627	0.495629	0.495593

<u>Confusion Matrix</u>

	Training Set	Test Set
True Positive (%)	34.924286	25.716667
True Negative (%)	15.174286	24.790000
False Positive (%)	15.907143	25.633333
False Negative (%)	33.994286	23.860000



Final ROC Curve

['Customer_Churn_prediction_model.pkl'] -- > **Our Final Saved Model**