Customer Churn Prediction Project Report

Introduction:

In this report, we present our analysis and modeling efforts aimed at predicting customer churn for a large dataset. Churn prediction is a critical task for businesses, as retaining existing customers is often more cost-effective than acquiring new ones. To accomplish this, we performed data preprocessing, exploratory data analysis (EDA), and prepared the data for modeling.

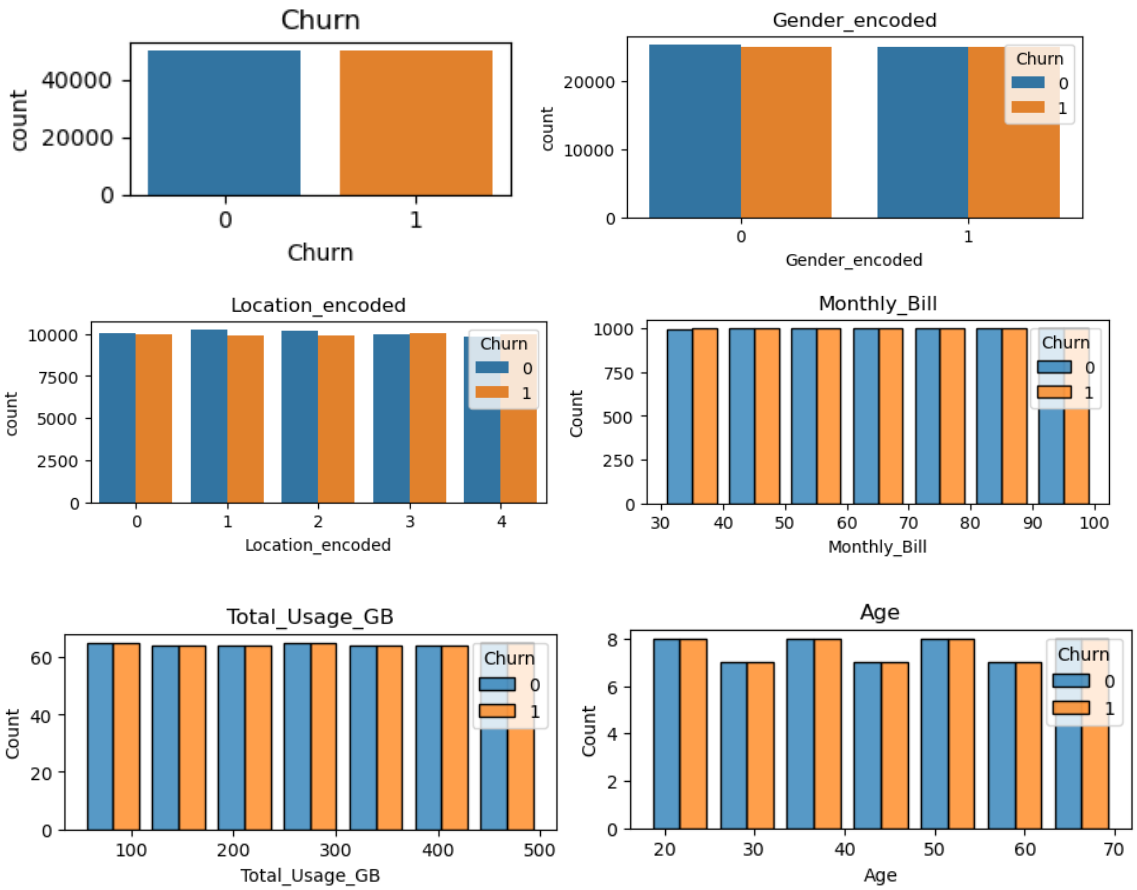
Data Processing:

- **Checking for Missing Values and Duplicated Rows**: We conducted checks for missing values and duplicated rows in the dataset. Fortunately, our data was found to be clean, with no missing values or duplicate rows.

- **Exploratory Data Analysis (EDA)**: In our initial EDA, we visualized the class distribution of the target variable, "Churn." We observed a slight class imbalance, with approximately 50,221 non-churned and 49,779 churned customers.

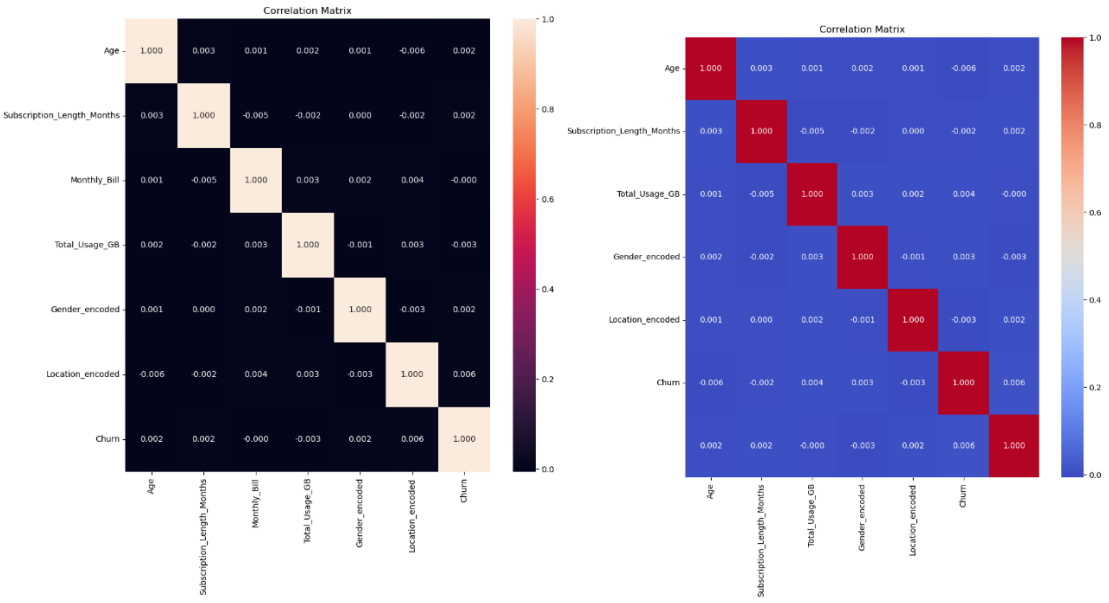
    - ❖ Our findings indicated that males had a slightly higher churn rate compared to females.
    - ❖ Customers from Houston and Los Angeles exhibited lower churn rates, while Chicago and New York had higher churn rates.
    - ❖ On visualizing the distribution of continuous variables—age, monthly bill, and total usage—in relation to churn. However, no significant trends or patterns emerged.

P.T.O



- **Correlation Martix before and after Normalization:**

  - ❖ We calculated and visualized the correlation matrix among the features before and after Noramalization. The analysis revealed no strong correlations among the features although the correlation was slightly better after normalization



Model Development and Evaluation

- ❖ We took two datasets for our model development:
  - noramlized_data
  - df
- ❖ We ran 4 algorithms for our models, Logistic Regression, Random Forest, Decision Trees and XGBoost
- ❖ We found the Logistic regression model to be with highest accuracy

	Precision	Recall	F1-Score	Support
0	0.50	0.64	0.57	12578
1	0.50	0.36	0.42	12422
Accuracy			0.50	25000
Macro Avg	0.50	0.50	0.49	25000
Weighted Avg	0.50	0.50	0.49	25000

Summary:

In summary, we trained and evaluated multiple machine learning models on our dataset, including logistic regression, random forest, decision tree, and XGBoost. We observed that the logistic regression model achieved an accuracy of approximately 50.31% on our dataset, and this model was serialized for future deployment.