

Customer Churn Prediction and Analysis Report

Team Members:

Mariam Yasser Saad Mohamed

Farah Alaa Hamoudy

Asmaa Elsayed

Lobna Hegazy Mohamed

Seif eldeen mohamed elbialy

Mahmoud Mohamed Mahmoud

Under the supervision of:

Eng : Eslam Adel

Contents

Table of Figure :-	3
1. Introduction	4
1.1 Objective	4
2. Data set	4
3. Data Analysis and Preprocessing	5
3.1 Data Loading	6
3.2 Data Cleaning	6
3.3 Exploratory Data Analysis (EDA)	7
4. Streamlit	11
.....	Error! Bookmark not defined.
5. Model Development	13
5.1 Data Preprocessing	14
5.2 Model Training and Evaluation	14
5.3 Model Evaluation	15
6. Mlops	16

Table of Figure :-

Figure 1 : there is no outliers (false)	7
Figure 2 : churn rates remain consistently high across all subscription types and contract lengths for customers with frequent support calls. ...	8
Figure 3 : the distribution of contract lengths, with 39.3% being annual, 39.1% quarterly, and 21.6% monthly.	9
Figure 4 : churn rate fluctuates over tenure, with a noticeable drop around the 20-month mark, followed by more consistent, lower churn rates after 25 months.....	9
Figure 5 : that total spend is higher for non-churned customers (both male and female) compared to churned customers, with a similar distribution pattern for both genders.	10
Figure 6 : the customer base is composed of 55% males and 45% females.	10
Figure 7 : the customer is likely to churn.....	12
Figure 8 : the customer is likely to stay	12
Figure 9 : the customer is likely to churn.....	13
Figure 10 : Receiver Operating Characteristic (ROC) Curve for GradientBoostingClassifier.....	15
Figure 11 : Receiver Operating Characteristic (ROC) Curve for LightGBM	16
Figure 12 : Receiver Operating Characteristic (ROC) Curve for AdaBoostClassifier	16
Figure 13 : Mlops 1.....	17
Figure 14 : Mlops 2.....	18

1. Introduction

In today's competitive market, understanding customer churn is vital for improving retention and optimizing services. This project analyzes factors influencing customer churn in a subscription-based service through data analysis and visualization. The goal is to uncover patterns that explain why customers leave and provide actionable insights to enhance retention strategies. By leveraging these insights, businesses can proactively address customer concerns and improve satisfaction.

1.1 Objective

- Analyze customer churn in a subscription-based service.
- Develop strategies to enhance customer retention.
- Increase overall customer satisfaction.

2. Data set

The dataset used in this analysis is a combination of various customer attributes, churn status, and behavioral metrics collected from a subscription service. This dataset is crucial for identifying the elements contributing to customer churn. The primary columns of interest include:

- Customer ID: Unique identifier for each customer.
- Gender: The gender of the customer (Male/Female).
- Age: Age of the customer.
- Total Spend: Total amount spent by the customer.
- Contract Length: Duration of the customer's contract with the service.
- Payment Delay: Number of days late in payment.
- Support Calls: Number of support calls made by the customer.
- Churn: Indicator of whether the customer has churned (1 = Churned, 0 = Not Churned).

CustomerID	Age	Gender	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
1.0	22.0	Female	25.0	14.0	4.0	27.0	Basic	Monthly	598.0	9.0	1.0
2.0	41.0	Female	28.0	28.0	7.0	13.0	Standard	Monthly	584.0	20.0	0.0
3.0	47.0	Male	27.0	10.0	2.0	29.0	Premium	Annual	757.0	21.0	0.0
4.0	35.0	Male	9.0	12.0	5.0	17.0	Premium	Quarterly	232.0	18.0	0.0
5.0	53.0	Female	58.0	24.0	9.0	2.0	Standard	Annual	533.0	18.0	0.0

Figure 1 : sample of dataset

3. Data Analysis and Preprocessing

This section outlines all tasks related to the dataset, including data preprocessing, cleaning, and analysis conducted in the Python notebook. It details how the data was prepared and transformed to ensure quality and consistency. Key processes like handling missing values and addressing inconsistencies are highlighted. The goal is to create a reliable dataset for accurate analysis and modeling.


3.1 Data Loading

In this analysis, we utilized two datasets: one containing 440,882 customer records and the other with 64,374 records. Both datasets shared the same 12 features. We combined these two files to create a single dataset with a total of 505,207 records and 12 features, ensuring consistency for further analysis.

3.2 Data Cleaning

- **Handling Missing Values:** Checked for missing data across all features and applied appropriate techniques (e.g., imputation or removal) to address any gaps.
- **Duplicate Removal:** Identified and removed duplicate records to ensure the integrity of the dataset.
- **Data Type Correction:** Converted data types where necessary, such as transforming numerical fields stored as strings and categorizing binary fields like 'Age' and 'Churn' columns to integer type

- **Outlier Detection and Treatment:** Detected and handled outliers that could skew the analysis, applying suitable transformations or exclusions.



	Age	Tenure	Usage	Frequency	Support	Calls	Payment	Delay	\
0	False	False		False		False		False	
1	False	False		False		False		False	
2	False	False		False		False		False	
3	False	False		False		False		False	
4	False	False		False		False		False	
...	
440828	False	False		False		False		False	
440829	False	False		False		False		False	
440830	False	False		False		False		False	
440831	False	False		False		False		False	
440832	False	False		False		False		False	

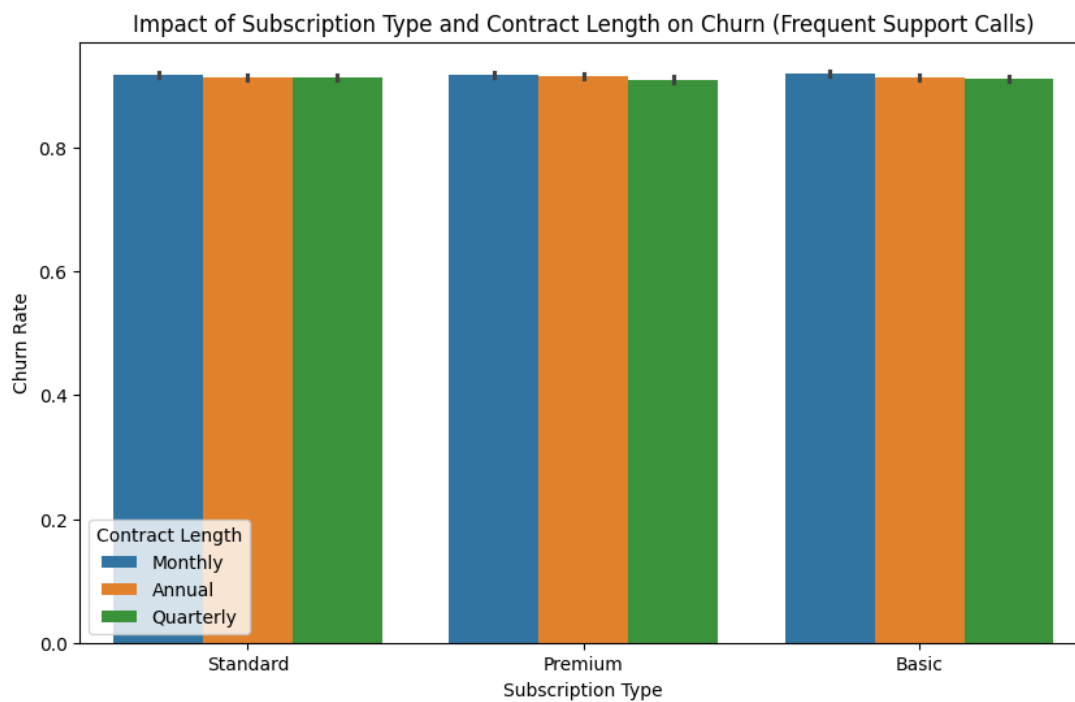
Figure 2 : there is no outliers (false)

- **Consistent Formatting:** Ensured consistent formatting across features, especially for categorical variables like gender and contract types.

3.3 Exploratory Data Analysis (EDA)

- **Churn Distribution Analysis:** Visualized the distribution of churned versus non-churned customers to understand the overall churn rate and detect any imbalances in the dataset.
- **Feature Correlation:** Analyzed the correlation between various features (e.g., total spend, contract length) and churn to identify key predictors of customer churn.

- **Visualization of Key Metrics:** Created box plots, bar charts, and scatter plots to identify trends and outliers that might impact churn, such as relationships between total spend and churn or the impact of payment delays.



- **Figure 3 :** churn rates remain consistently high across all subscription types and contract lengths for customers with frequent support calls.

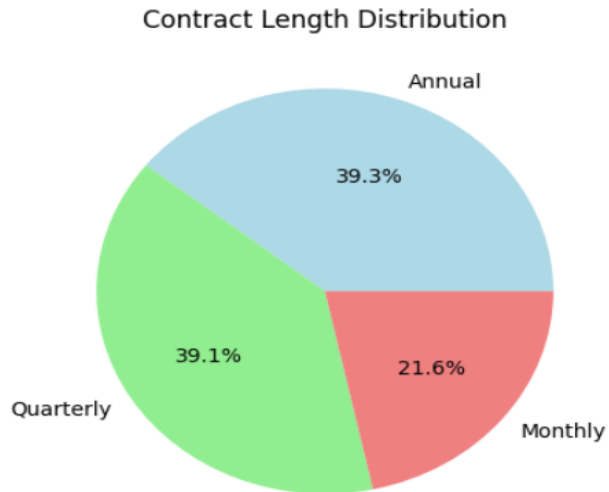


Figure 4 : the distribution of contract lengths, with 39.3% being annual, 39.1% quarterly, and 21.6% monthly.

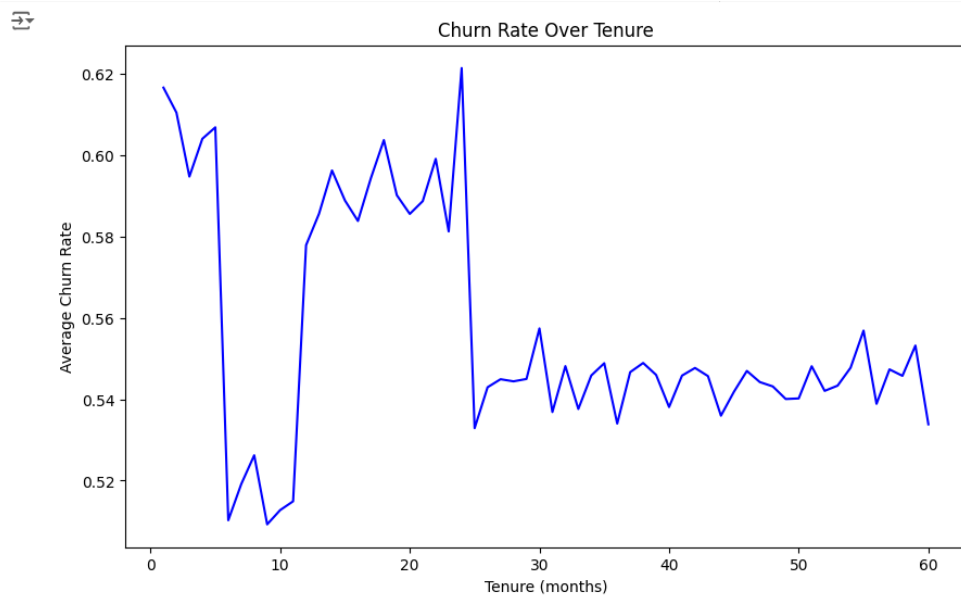


Figure 5 : churn rate fluctuates over tenure, with a noticeable drop around the 20-month mark, followed by more consistent, lower churn rates after 25 months.

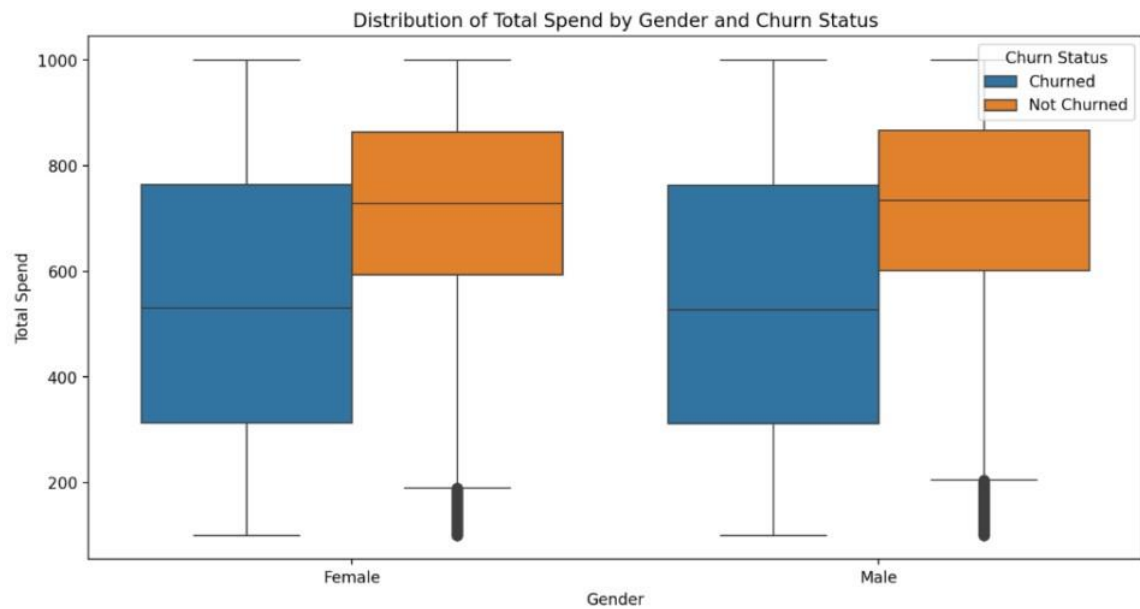


Figure 6 : that total spend is higher for non-churned customers (both male and female) compared to churned customers, with a similar distribution pattern for both genders.

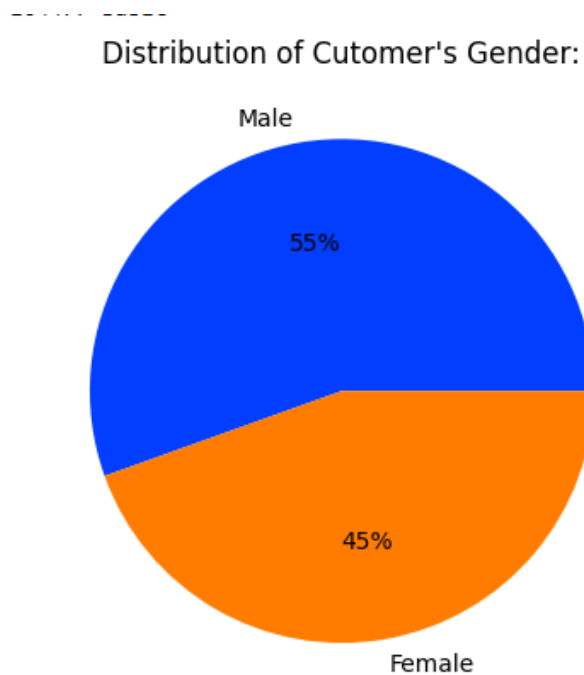


Figure 7 : the customer base is composed of 55% males and 45% females.

4. Streamlit

- **Interactive Dashboard:** Provides an easy-to-use interface for visualizing customer churn data.
- **Various Visualizations:** Includes charts such as pie charts, scatter plots, and heatmaps to explore relationships between variables.
- **Data Exploration:** Allows users to analyze key metrics like tenure, total spend, and churn status to understand customer behavior.
- **Churn Prediction:** Enables users to input customer data and predict whether they are likely to churn.

- **Customizable:** The app is highly flexible, allowing users to modify and customize the analysis as per their needs.

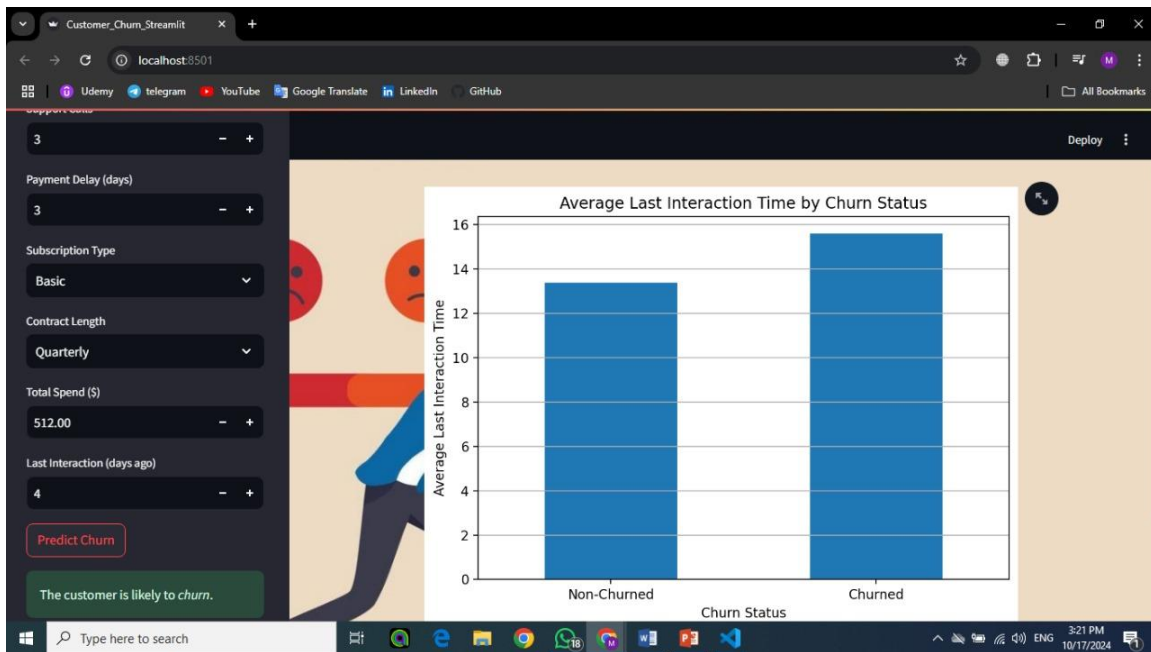


Figure 8 : the customer is likely to churn

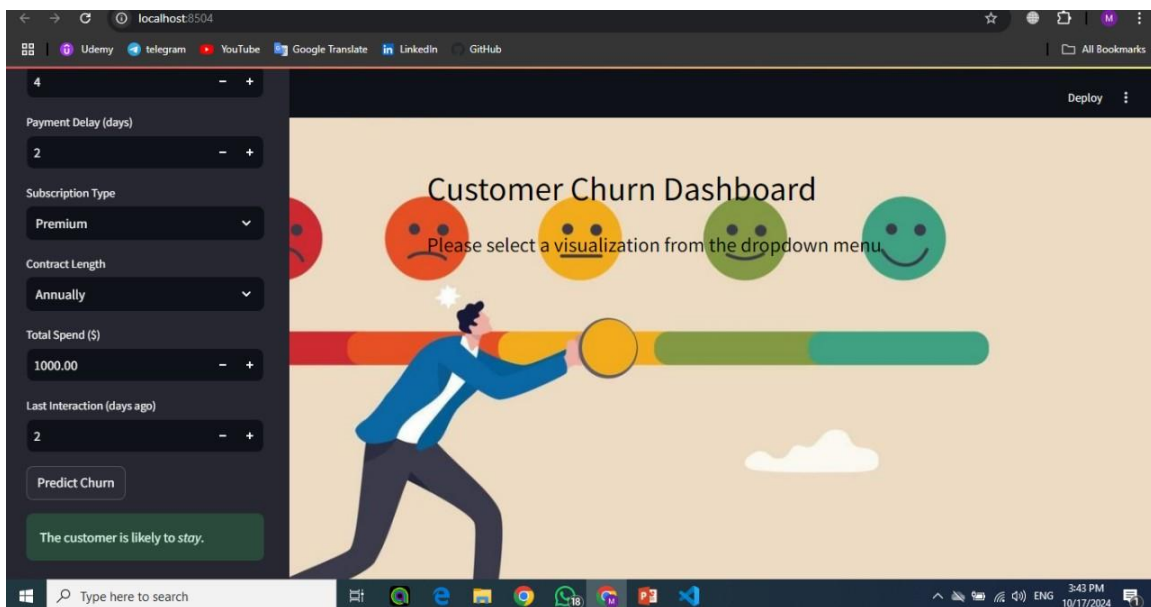


Figure 9 : the customer is likely to stay

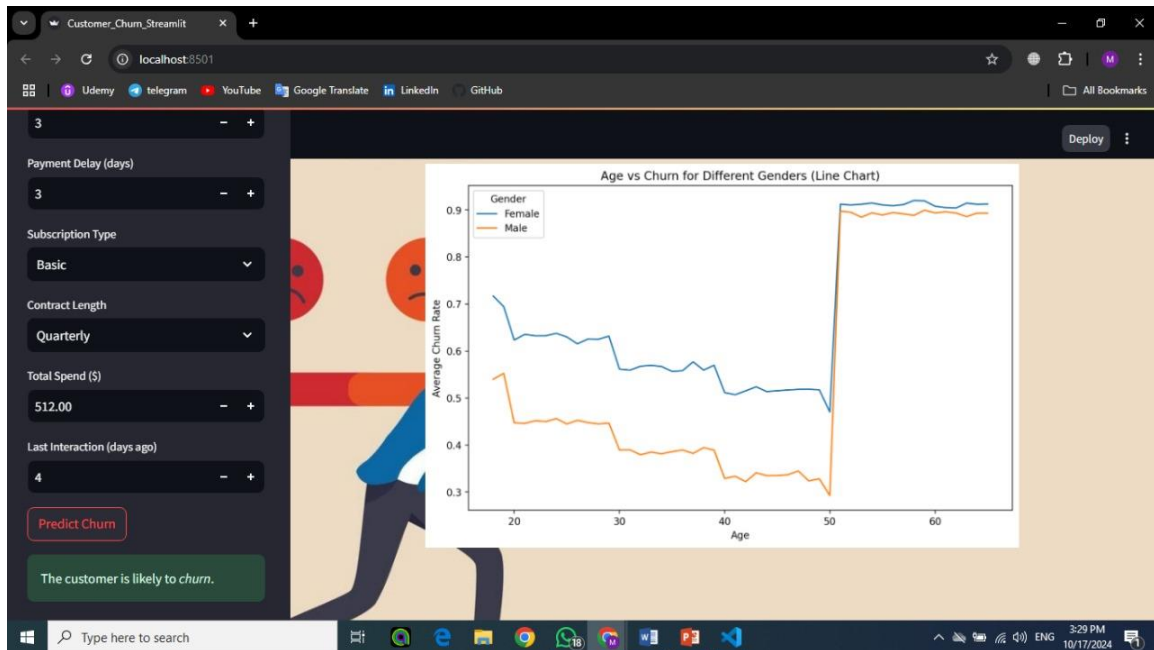


Figure 10 : the customer is likely to churn

5. Model Development

This section outlines the steps involved in developing the model. It covers data preprocessing to prepare the dataset for analysis, the selection of suitable machine learning models, and the process of tuning hyperparameters to optimize performance. Additionally, it describes how the model's effectiveness was evaluated to ensure accuracy and reliability.

5.1 Data Preprocessing

- **Handling Missing Values:** We manage any missing values by filling them with appropriate data or removing incomplete rows to ensure data integrity.
- **Encoding Categorical Variables:** We use `LabelEncoder` to convert categorical variables into numerical values for model compatibility.
- **Scaling Numerical Features:** We standardize the numerical features to ensure consistency and enhance model performance during training and testing.
- **Splitting the Dataset:** We split the data into training and testing sets, with 80% used for training and 20% for testing

5.2 Model Training and Evaluation

- **Model Training:** We trained both models using the training data. Each model was fitted to the features and corresponding target variable to learn the underlying patterns and relationships.
- **Evaluation:** After training, we evaluated both models on the testing data by calculating the accuracy, ROC AUC score, and other relevant metrics. These performance metrics help in comparing the effectiveness of both models.
- **Comparison:** Finally, we generated and compared the ROC curves for both models. The AUC (Area Under the Curve) score allows us to measure how well each model can distinguish between the positive and negative classes, providing a comprehensive view of their classification performance.

5.3 Model Evaluation

Metric	GradientBoostingClassifier	LightGBM	AdaBoost
Accuracy	0.93	0.94	0.93
ROCAUC Score	0.98	0.94	0.93

- **LGBMClassifier** would be my recommendation due to its balance of high accuracy, high ROC AUC, and efficiency, especially if you are working with larger datasets.
- **GradientBoostingClassifier** is ideal if precision is crucial, particularly for imbalanced datasets where distinguishing between classes is more important.
- **AdaBoostClassifier** is a good option for simpler problems or when interpretability is key, but it might not perform as well on more complex tasks compared to the other two models.

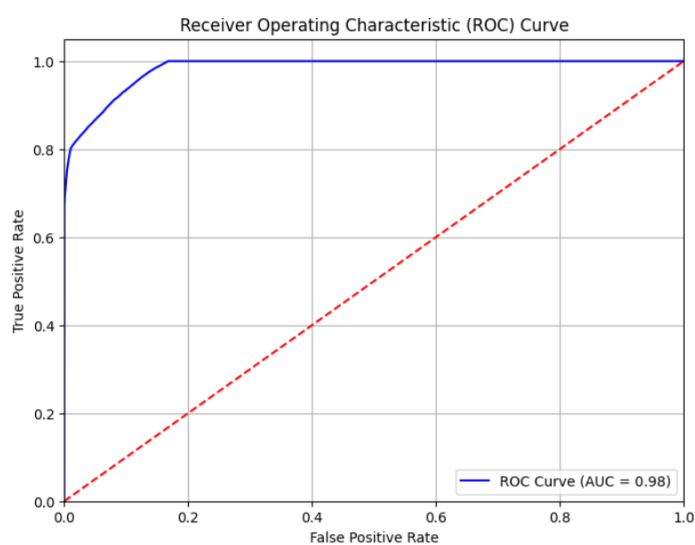


Figure 11 : Receiver Operating Characteristic (ROC) Curve for GradientBoostingClassifier

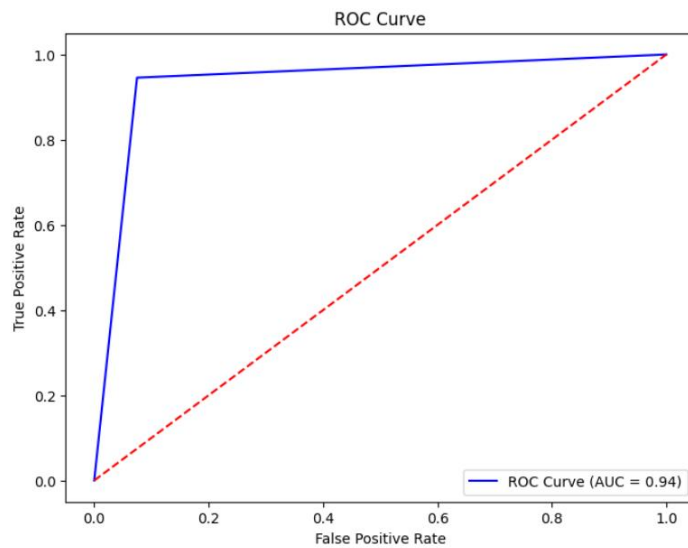


Figure 12 : Receiver Operating Characteristic (ROC) Curve for LightGBM

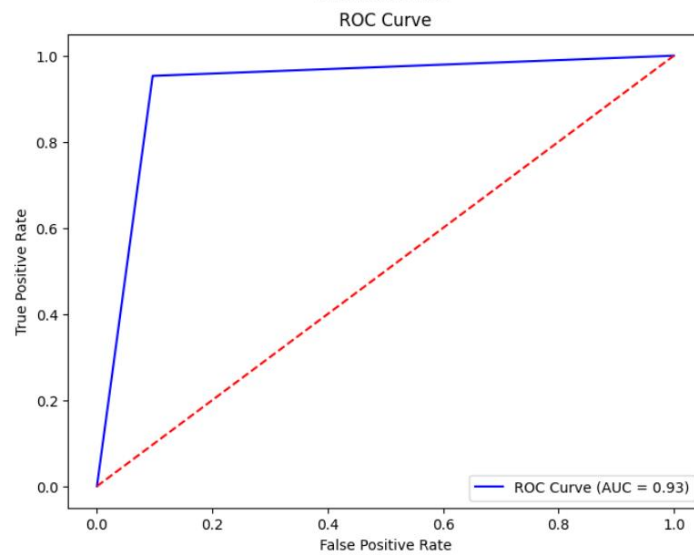


Figure 13 : Receiver Operating Characteristic (ROC) Curve for AdaBoostClassifier

6. Mlops

To effectively manage the machine learning pipeline, **MLflow** was utilized for:

- **Experiment tracking:** Capturing hyperparameters, performance metrics, and other key details to ensure reproducibility and comparison across experiments.

- **Model logging and registration:** Storing the final model version and registering it for seamless access and future deployment.

This ensures the model is versioned, monitored, and continuously improved as new data becomes available, facilitating efficient updates and maintaining optimal performance over time.

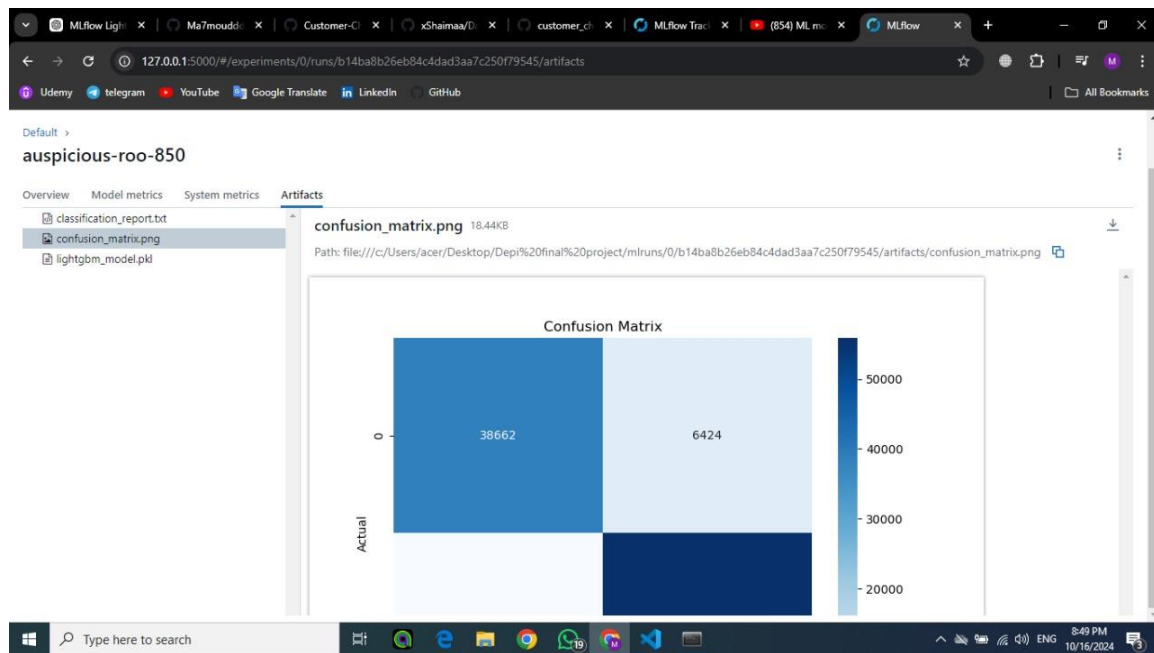


Figure 14 : Mlops 1

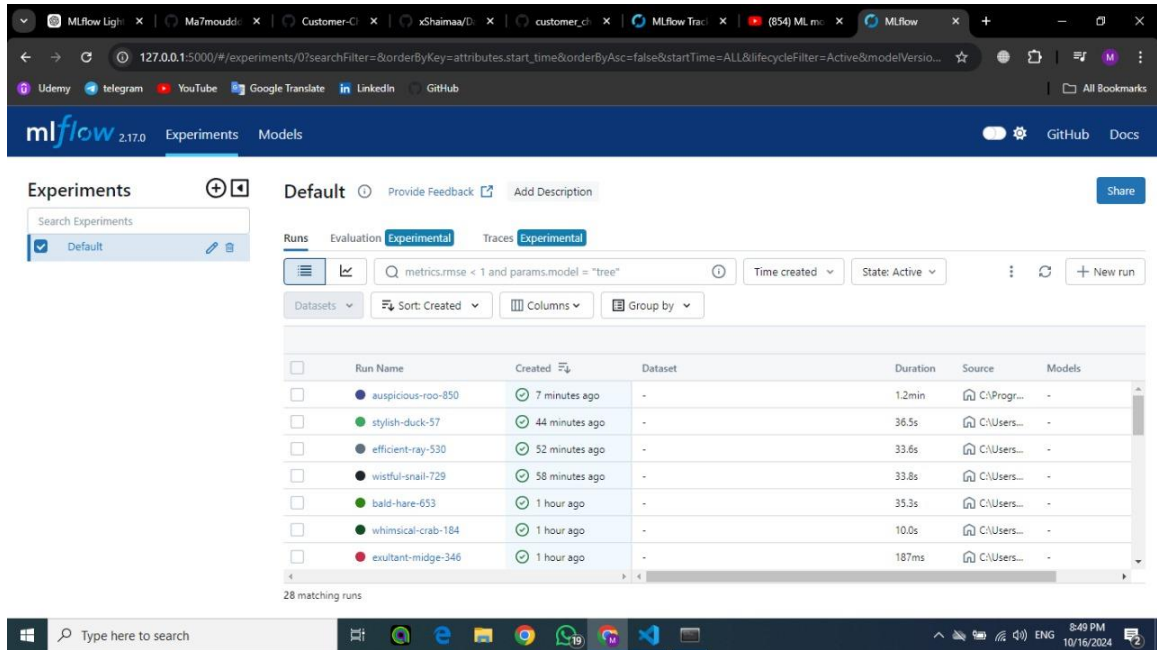


Figure 15 : Mlops 2

7. Conclusion

In this project, we successfully built a customer churn prediction model using multiple machine learning algorithms, including GradientBoostingClassifier, AdaBoost, and LightGBM. The LightGBM model achieved the highest accuracy of 94% and was integrated into a Streamlit app for real-time interaction and visualization. This tool enables telecom companies to analyze customer behavior, segment customers, and proactively address churn, leading to improved retention rates and business growth.