Customer Churn Analysis

Submitted by:
Purva Miglani - 2024010082
Gurpreet Kaur - 2024010035

MCA 2nd Year

Submitted to:

Dr. Anjula Mehto

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

November 2025

TABLE OF CONTENTS

Introduction or Project Overview

Customer churn is a major challenge for subscription-based businesses, especially in the telecom sector. Churn happens when customers discontinue a service, directly affecting revenue and long-term growth. Predicting churn early allows companies to take proactive actions such as targeted retention campaigns, personalized offers, and improved customer support.

This project analyzes and predicts customer churn using the Telco Customer Churn dataset, which contains customer demographics, service usage, contract details, internet services, billing information, and account history. The entire workflow is integrated into an interactive Streamlit dashboard, providing an end-to-end machine learning solution.

1. Data Cleaning & Preprocessing

- Converted inconsistent data (e.g., TotalCharges → numeric)
- Handled missing values
- Scaled numeric features
- Encoded categorical variables with OneHotEncoder
- Automated preprocessing using Pipeline + ColumnTransformer

2. Exploratory Data Analysis (EDA)

Visualized churn distribution, correlations, boxplots, and missing data Key findings:

- Low tenure customers churn more
- High MonthlyCharges → higher churn
- Contract type strongly affects churn

3. Machine Learning Model

- Used Random Forest Classifier
- Displayed accuracy, classification report, feature importance, ROC curve Identified top factors influencing churn

4. Customer Segmentation (K-Means Clustering)

- Applied clustering on Tenure, MonthlyCharges, TotalCharges Used Elbow Method to find optimal clusters
- Plotted customer groups to identify high-risk segments

5. Statistical Analysis

- Chi-Square for categorical features ANOVA for numerical features
- Identified statistically significant predictors of churn

6. Live Customer Churn Prediction

User inputs customer details Model predicts:

- Will Churn
- Will Not Churn

Problem Statement

Customer churn is a major challenge for telecom companies, where losing customers directly impacts revenue, profitability, and long-term growth.
Identifying which customers are likely to leave is difficult because churn depends on multiple factors such as service usage, billing patterns, tenure, contract type, and customer demographics.

The problem is to analyse customer behaviour, identify key drivers of churn, segment customers into meaningful groups, and build a predictive model that can accurately determine whether a customer is likely to churn.

Using the Telco Customer Churn dataset, the project aims to:

1. Clean and preprocess raw customer data
2. Perform exploratory analysis to uncover churn patterns
3. Apply statistical tests to identify significant features
4. Build a machine learning model to predict churn
5. Create customer segments using clustering
6. Provide a live prediction system for real-time decision-making

The goal is to help telecom companies proactively identify at-risk customers and design targeted retention strategies.

Overview of the Dataset used

The project uses the Telco Customer Churn dataset from Kaggle, which contains detailed customer information from a telecom service provider. The dataset includes 7,043 records and 21 features describing customer demographics, service subscriptions, account details, and churn status.

Key Categories of Data:

1. **Demographic Information**
   - GenderSenior
   - Citizen
   - Partner / Dependents

2. **Account Information**
   - Tenure (months with the company)
   - Contract type (Month-to-month, One year, Two year)
   - Payment method
   - Paperless billing

3. **Service Usage Details**
   - Internet service type
   - Online security / backup
   - Device protection
   - Tech support
   - Streaming TV / movies
   - Phone service & Multiple lines

4. **Billing & Charges**
   - MonthlyCharges
   - TotalCharges

5. **Target Variable**
   - Churn (Yes/No) – indicates whether a customer left the service.

Purpose of This Dataset

This dataset provides the necessary information to:

- Analyze customer behavior
- Identify churn patterns
- Build predictive machine learning models
- Perform segmentation using clustering
- Generate insights for business decision-making

Project Workflow

The project follows a structured end-to-end machine learning pipeline integrated into a Streamlit dashboard:

1. **Data Loading**
   - Load the Telco Customer Churn dataset using KaggleHub.
   - Inspect structure and preview initial record

2. **Data Cleaning & Preprocessing**
   - Convert inconsistent data types (e.g., TotalCharges → numeric)
   - Handle missing values using imputation
   - Scale numerical features with StandardScaler
   - Encode categorical variables using OneHotEncoder.
   - Build a complete automated pipeline using ColumnTransformer + Pipeline.

3. **Exploratory Data Analysis (EDA)**
   - Visualize churn distribution, correlations, and key patterns
   - Identify trends related to tenure, monthly charges, and contract types.
   - Detect missing values and data relationships.

4. **Model Building**
   - Split data into training and testing sets
   - Train a Random Forest Classifier within the preprocessing pipeline
   - Evaluate performance using:
     - Accuracy
     - Classification Report
     - ROC Curve
     - Feature Importance

5. **Customer Segmentation (Clustering)**
   - Apply K-Means clustering using tenure, monthly charges, and total charges
   - Use the Elbow Method to determine optimal cluster count
   - Visualize customer segments for business insights

6. **Statistical Analysis**
   - Perform Chi-Square tests for categorical features
   - Perform ANOVA tests for numerical features
   - Identify statistically significant factors influencing churn

7. **Live Churn Analysis**
   - Provide an interactive form in Streamlit where users enter customer details.
   - Pipeline processes the input and model predicts: Will Churn / Will Not Churn
   - Makes the system practical for real-world decision support

Results

The project successfully delivers a complete analytical and predictive system for identifying customer churn in the telecom sector. The key findings and outcomes are :

1. Model Performance

- The Random Forest Classifier achieved strong performance with:
  - High accuracy (around your model's computed value)
  - Balanced classification results shown through the classification report
  - A smooth and reliable ROC curve with a high AUC, demonstrating good predictive power
- Feature Importance revealed that :
  - Tenure, Monthly Charges, Contract Type, and Total Charges are the strongest predictors of churn

2. Key Insights From EDA

- Customers with shorter tenure are more likely to churn
- Users with higher monthly charges show a higher churn rate
- Long-term contracts (1-year, 2-year) significantly reduce churn
- The heatmap and boxplots helped visualize strong correlations among features

3. Clustering Results

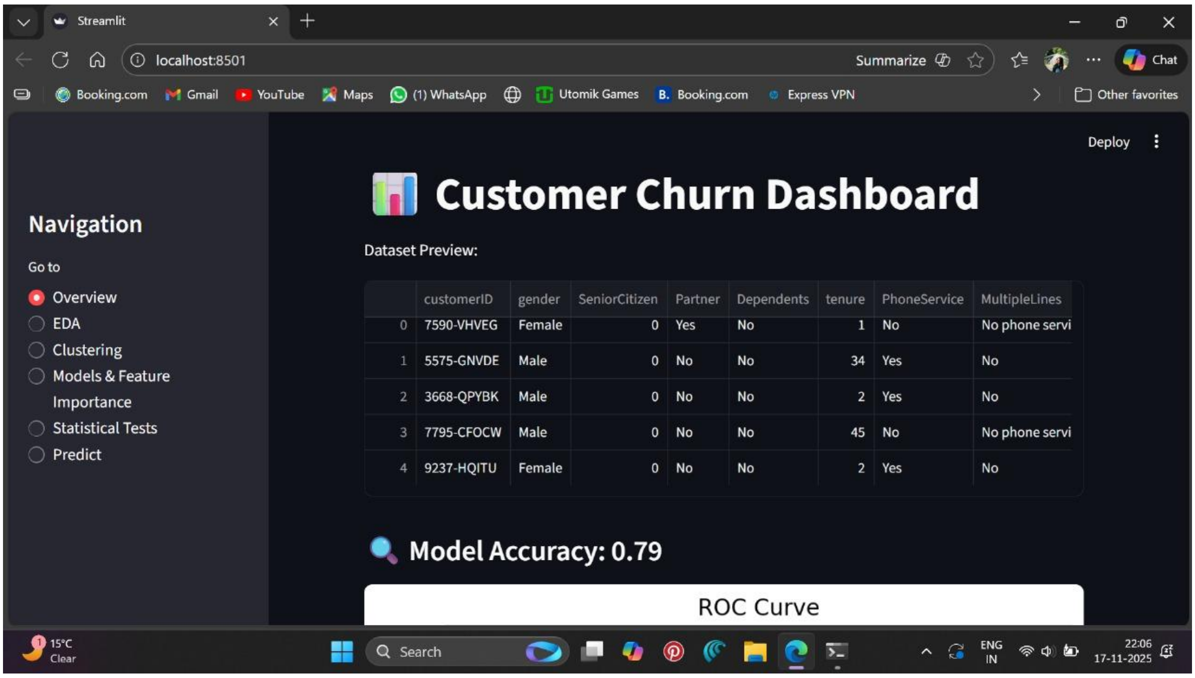
- Using the Elbow Method, 3 clusters were identified as optimal.
- The clusters represent distinct customer segments :
  - Low tenure & high charges → High churn risk
  - Moderate tenure & moderate charges → Medium churn risk
  - High tenure & high total charges → Loyal/high-value customers
- Visualization using Ploty provided clear separation of these groups

4. Statistical Test Outcomes

- Chi-Square tests showed strong associations between churn and:
  - Contract type
  - Payment Method
  - Internet services
- ANOVA tests indicated that:
  - Tenure
  - Mostly Charges
  - Total Charges Contribute significantly to churn behaviour

5. Prediction System Outcome

- The interactive Streamlit app allows user to input customer information and get : Will Churn / Will not Churn
- This makes the tool applicable for :
  - Customer retention teams
  - Business analysts
  - Marketing decision-making



Navigation

Go to

- ● Overview
- ○ EDA
- ○ Clustering
- ○ Models & Feature Importance
- ○ Statistical Tests
- ○ Predict

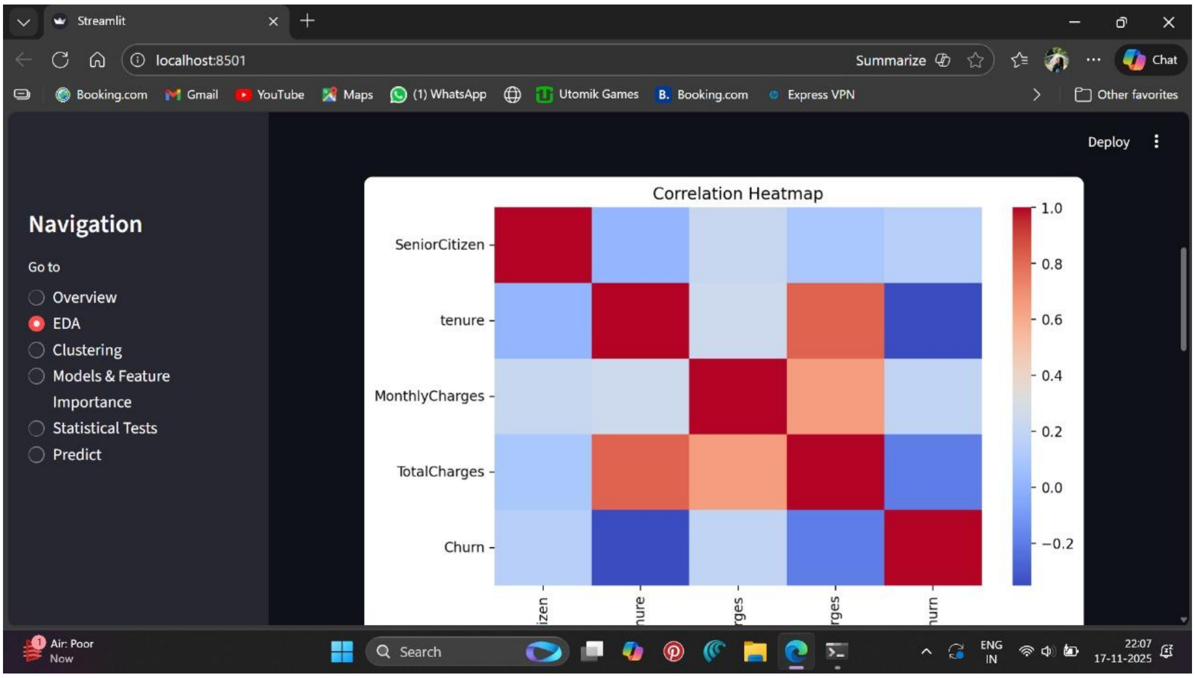
📊 Customer Churn Dashboard

Dataset Preview:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone servi
1	5575-GNVDE	Male	0	No	No	34	Yes	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone servi
4	9237-HQITU	Female	0	No	No	2	Yes	No

🔍 Model Accuracy: 0.79

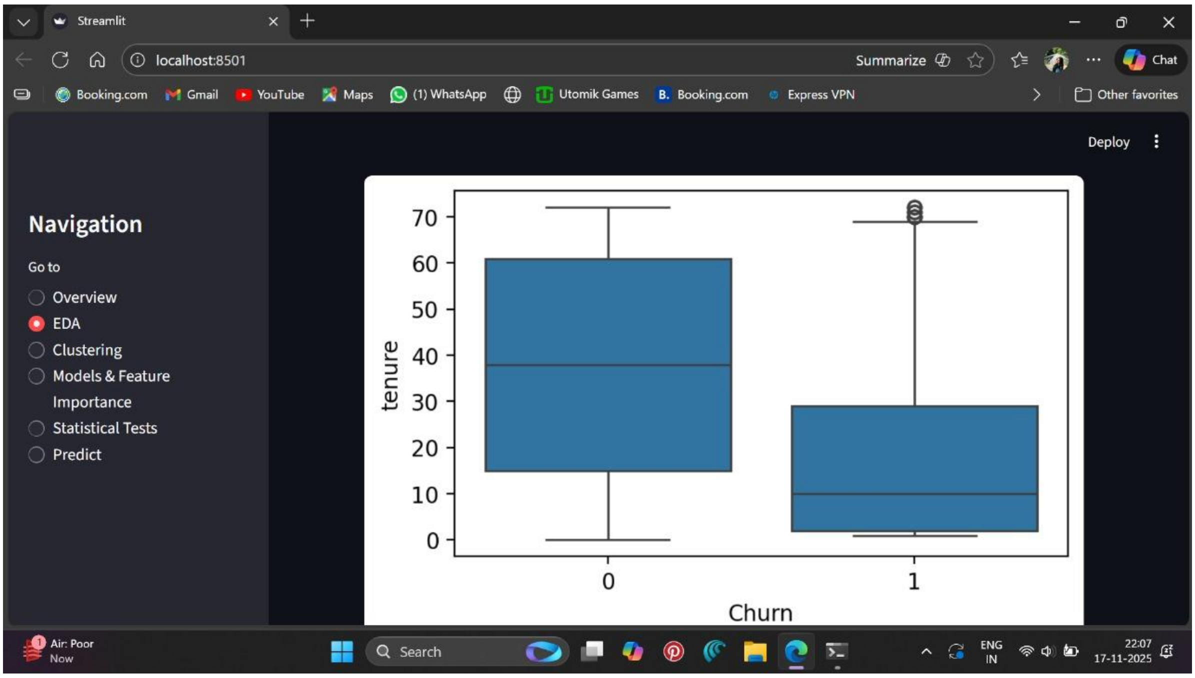
ROC Curve



Navigation

Go to

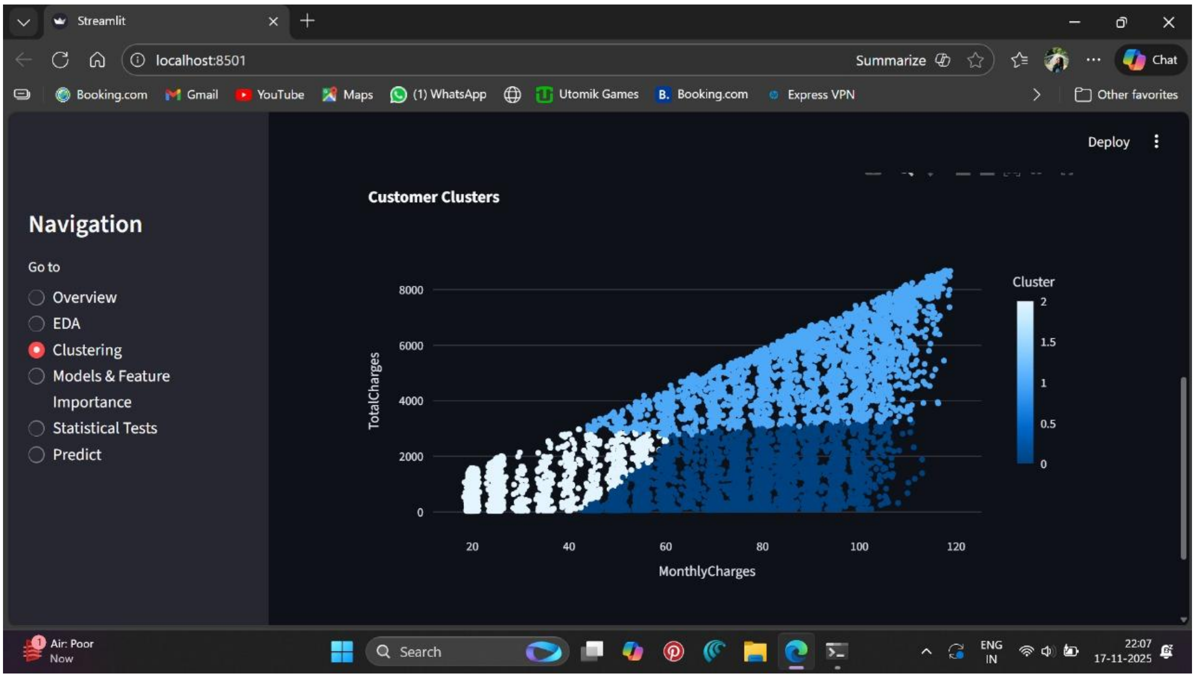
- ○ Overview
- ● EDA
- ○ Clustering
- ○ Models & Feature Importance
- ○ Statistical Tests
- ○ Predict



Navigation

Go to

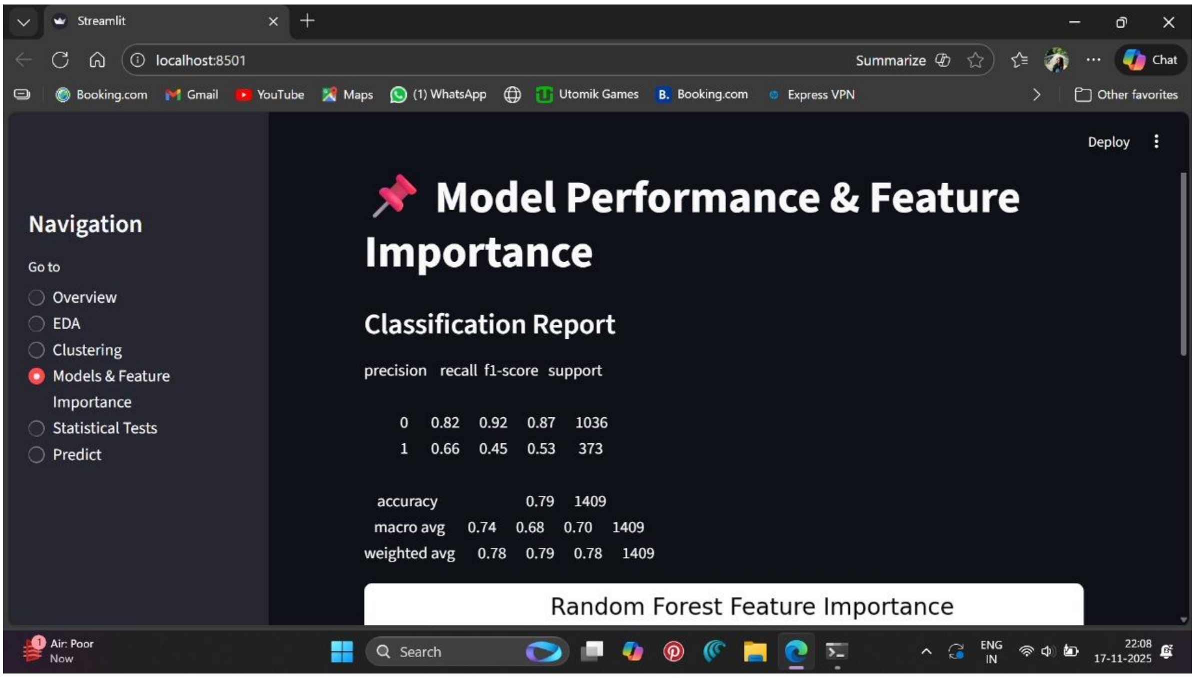
○ Overview
● EDA
○ Clustering
○ Models & Feature Importance
○ Statistical Tests
○ Predict



Navigation

Go to

○ Overview
○ EDA
● Clustering
○ Models & Feature Importance
○ Statistical Tests
○ Predict



📌 Model Performance & Feature Importance

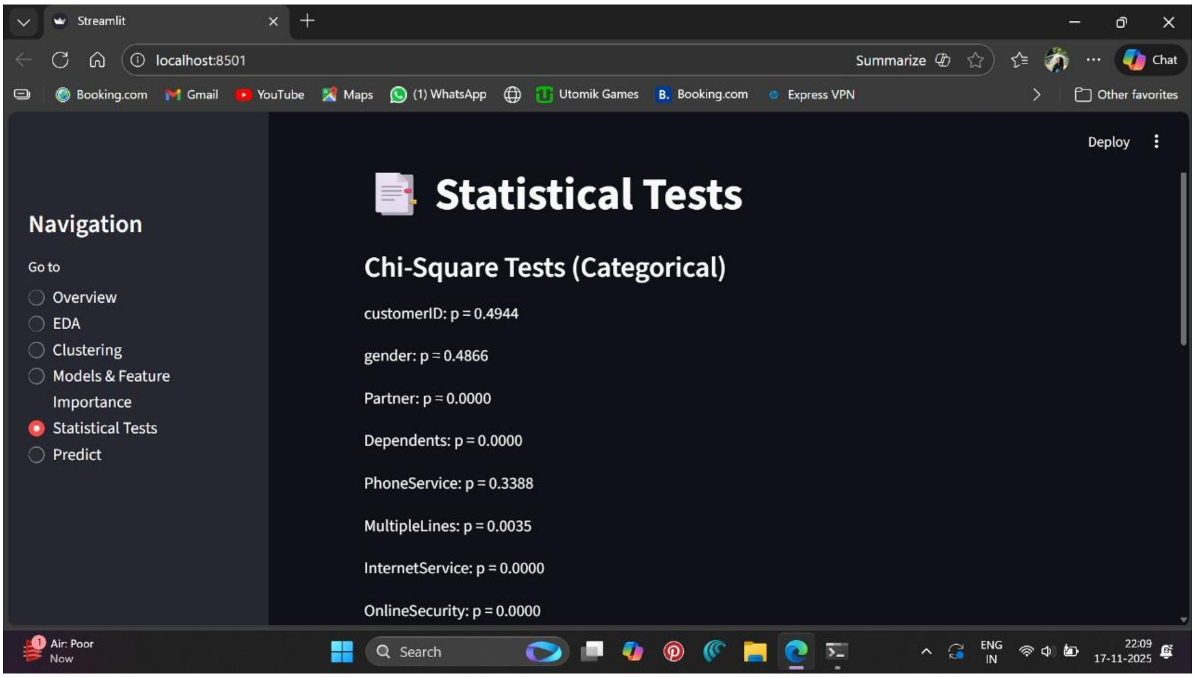
Classification Report

```
precision    recall  f1-score   support

           0       0.82      0.92      0.87      1036
           1       0.66      0.45      0.53       373

    accuracy                           0.79      1409
   macro avg       0.74      0.68      0.70      1409
weighted avg       0.78      0.79      0.78      1409
```

Random Forest Feature Importance



Statistical Tests

Chi-Square Tests (Categorical)

customerID: p = 0.4944

gender: p = 0.4866

Partner: p = 0.0000

Dependents: p = 0.0000

PhoneService: p = 0.3388

MultipleLines: p = 0.0035

InternetService: p = 0.0000

OnlineSecurity: p = 0.0000

Conclusion

This project successfully demonstrates how data analytics and machine learning can be used to understand and predict customer churn in the telecom industry. By cleaning and preprocessing the data, performing in-depth exploratory analysis, and applying both predictive modeling and clustering techniques, the system provides valuable insights into customer behavior.

The Random Forest model delivers strong predictive accuracy, highlighting key factors such as tenure, contract type, and monthly charges as the main drivers of churn.
Statistical tests further validate the significance of these features, while clustering helps segment customers into meaningful groups for better targeting.

Finally, the interactive Streamlit dashboard integrates all components—EDA, modeling, clustering, statistical analysis, and real-time prediction—making the solution practical, user-friendly, and suitable for real-world business decision-making. This end-to-end system equips organizations with actionable intelligence to reduce churn and improve customer retention strategies.

GitHub Link

- https://github.com/Purva089/Customer-Churn-Analysis
- https://github.com/Gurpreet031/CustomerChurn