Customer Churn Prediction System

Prepared by: G.Anish

Domain: Telecom / SaaS / Banking

Tools: Python, Pandas, Scikit-learn, XGBoost, Matplotlib

Dataset: Telco Customer Churn (Kaggle)

1. Executive Summary

Customer churn occurs when users discontinue a service, directly affecting revenue and long-term business growth. Since acquiring new customers is significantly more expensive than retaining existing ones, churn prediction is a critical business problem across telecom, SaaS, and banking industries.

This project develops an end-to-end machine learning–based Churn Prediction System that identifies customers who are likely to churn and estimates their churn probability. The system enables businesses to proactively intervene with targeted retention strategies.

2. Problem Statement

The objective of this project is to predict whether a customer will churn based on historical customer data, including contract details, billing information, service usage, and tenure. This is formulated as a binary classification problem where churned customers are labeled as 1 and retained customers as 0.

3. Dataset Overview

The Telco Customer Churn dataset contains approximately 7,000 customer records with attributes covering demographics, subscribed services, billing information, and contract types.

Target Variable: Churn

Key Features: Tenure, MonthlyCharges, TotalCharges, Contract Type, Payment Method, Service Subscriptions

4. Data Cleaning and Preprocessing

Several preprocessing steps were applied to ensure data quality and model readiness:

- Converted TotalCharges from text to numeric format
- Removed rows containing missing values
- Dropped non-informative identifier (customerID)
- Encoded categorical variables using one-hot encoding
- Converted churn labels to binary values

5. Feature Engineering

Feature engineering focused on transforming raw customer attributes into meaningful numerical representations. Categorical features such as contract type and payment method were encoded, enabling the model to capture behavioral patterns associated with customer churn.

6. Modeling Approach

Three classification models were trained and evaluated to predict customer churn:

- **Logistic Regression:** Interpretable baseline model
- **Random Forest:** Ensemble model capturing non-linear relationships
- **XGBoost:** Gradient boosting model optimized for performance

7. Model Evaluation

Models were evaluated using Accuracy, Precision, Recall, Confusion Matrix, and ROC-AUC score. ROC-AUC was chosen as the primary metric as it measures the model’s ability to distinguish between churned and retained customers across all classification thresholds.

8. Model Performance Comparison

The following ROC-AUC scores were obtained:

- Logistic Regression: 0.832
- Random Forest: 0.816
- XGBoost: 0.824

Logistic Regression achieved the highest ROC-AUC score and was selected as the final model due to its strong performance and interpretability.

9. Key Churn Drivers and Insights

- Month-to-month contracts significantly increase churn risk
- High monthly charges correlate strongly with churn
- Customers with shorter tenure are more likely to churn
- Automatic payment methods reduce churn probability

10. Churn Risk Segmentation

Customers were segmented into low, medium, and high-risk groups based on predicted churn probability. This segmentation allows businesses to apply targeted retention strategies instead of uniform incentives.

11. Business Recommendations

- Encourage long-term contracts through discounts and loyalty programs
- Offer personalized retention offers to high-risk customers
- Improve onboarding for new customers
- Promote automatic payment options

12. Conclusion

This project demonstrates a complete churn prediction system combining data preprocessing, machine learning, model evaluation, visualization, and business interpretation. The approach is applicable across multiple subscription-based industries and supports data-driven decision-making.