PROJECT REPORT – Customer Churn Prediction

1. Introduction

This project focuses on predicting customer churn using the IBM Telco Customer Attrition dataset. In the telecom industry, churn refers to customers discontinuing their service. Since customer loss directly impacts revenue, companies rely on predictive models to identify at-risk users early.

Our objective was to build an end-to-end machine learning pipeline that covers data understanding, preprocessing, model development, evaluation, and threshold analysis. The idea was to create a workflow that is both practical and easy to reproduce.

2. Problem Statement

The task is framed as a binary classification problem:
0 → Customer does not churn
1 → Customer churns

The dataset includes demographic details, service usage patterns, billing information, and contract-related fields. It also consists of mixed data types and contains missing or invalid entries, which makes preprocessing a critical step. Another factor is the class imbalance, as churned customers form a smaller portion of the data. The goal was to design a model that can reliably identify churn cases and provide a clear understanding of the underlying patterns.

3. Project Deliverables & Scope

The project was structured according to the required deliverables:

- A well-organized Git repository containing data, notebooks, source code, figures, and documentation.

- A reproducible setup using run.sh and requirements.txt.

- A comprehensive report covering the problem definition, EDA, methodology, results, error analysis, and conclusions.
  Overall, the work executed reflects each stage of a standard machine learning workflow from start to finish.

4. Dataset Overview

The IBM Telco Customer Churn dataset contains records for around 7,043 customers and includes more than 20 features, such as:

- Demographic information
- Tenure
- Subscription details for phone and internet services
- Contract and payment method
- Monthly and total charges
- Target variable: Churn

The dataset includes numerical, categorical, and boolean fields. Some columns, like “TotalCharges,” include non-numeric values that need cleaning. The imbalance between churn and non-churn users also influences how the model performs and how metrics should be interpreted.

5.Methodology

5.1 Exploratory Data Analysis (EDA)

EDA was carried out to understand the data distribution, spot anomalies, identify missing values, and observe correlations.
Major takeaways:

- Tenure shows a strong link with churn.

- Month-to-month contract users churn at a higher rate.
- Customers who pay through electronic checks tend to churn more.
- “Total Charges” contains incorrect string values that require fixing.

5.2 Data Preprocessing

Key preprocessing steps included:

- Handling missing values
- Converting string-type numeric fields into valid numerical values
- Encoding categorical features appropriately
- Scaling numerical variables if needed
- Splitting data into training and test sets (80:20)

These steps ensured the dataset was clean and well-structured before training the model.

5.3 Model Development

Logistic Regression was chosen as the baseline model due to its simplicity, interpretability, and suitability for binary classification. The model was trained on the processed data to learn patterns associated with churn.

5.4 Model Evaluation

The model was assessed using standard performance metrics:

- Accuracy
- Precision

- Recall
- F1-score
- Confusion matrix
- ROC curve
- AUC

These metrics helped evaluate how well the model identifies churners and highlighted areas where it performs well or struggles.

5.5 Threshold Tuning

The default probability threshold of 0.5 was adjusted to study its effect on different metrics.
Key observations:

- Lowering the threshold increases recall, helping the model catch more churn cases.
- Increasing the threshold improves precision but reduces recall.
- Choosing the threshold depends on what the business prioritizes—either reducing false positives or catching as many churners as possible.

6. Results & Findings

The Logistic Regression baseline model performed consistently across evaluation metrics.
Important insights:

- The model is naturally more inclined toward predicting non-churn due to class imbalance.

- Churn cases become harder to detect at higher thresholds.
- ROC-AUC scores confirm that the model can distinguish fairly well between churn and non-churn classes.
- Features like contract type, tenure, payment method, and charges strongly influence churn predictions.

These results align with typical customer behavior trends seen in telecom datasets.

7. Challenges

Several challenges were encountered throughout the workflow:

- Managing mixed feature types and fixing invalid or missing values
- Encoding categorical variables with many unique entries
- Addressing the class imbalance
- Ensuring the preprocessing pipeline stays consistent during training and evaluation
- Maintaining reproducibility with proper repository organization and environment setup

These issues guided the design of the final pipeline and helped strengthen the overall structure of the project.