Customer Churn Prediction Using Machine Learning

Submitted by:
Shashank (102317107)
Manikanta (102317292)

BE Third Year

CSE

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 one of the most important challenges faced by subscription-based businesses such as telecom companies, streaming services, and internet service providers. Churn represents the number of customers who stop using a service during a given period. Since acquiring new customers is much more expensive than retaining existing ones, companies focus heavily on predicting churn in advance, so customer retention strategies can be applied.

In this project, we aim to build a **machine learning-based churn prediction system** using the **Telco Customer Churn Dataset** from Kaggle. This dataset contains customer demographics, account details, and service usage patterns. The goal is to analyze the data, understand factors leading to churn, and build a model capable of predicting whether a customer will leave or stay. The project includes detailed data preprocessing, exploratory data analysis (EDA), training and comparison of three machine learning models, and finally a simple Flask-based web demonstration for real-time predictions.

Problem Statement

Telecom industries lose a significant portion of revenue each year due to customer churn. Most customers churn due to service dissatisfaction, billing issues, or switching to competitors. Predicting churn helps companies take preventive actions such as personalized offers, proactive customer support, and service improvements.

The problem addressed in this project can be stated as:

> **"Given customer information such as demographics, services used, billing patterns, and account tenure, predict whether the customer will churn (Yes/No)."**

This is a **binary classification problem**.
Our goal is to create an accurate and reliable machine learning model that can help identify customers at high risk of churn.

Overview of the Dataset used

The dataset used is the **Telco Customer Churn Dataset** provided on Kaggle. It contains **7043 rows** representing different customers and **21 columns** representing customer details. The dataset includes:

Demographic Features

- Gender
- Senior Citizen
- Partner
- Dependents

Account Information

- Customer Tenure
- Contract Type (Month-to-Month, One-Year, Two-Year)
- Paperless Billing
- Payment Method

Service Details

- Phone Service
- Internet Service
- Streaming Movies
- Streaming TV
- Online Security
- Tech Support

Billing Information

- Monthly Charges
- Total Charges

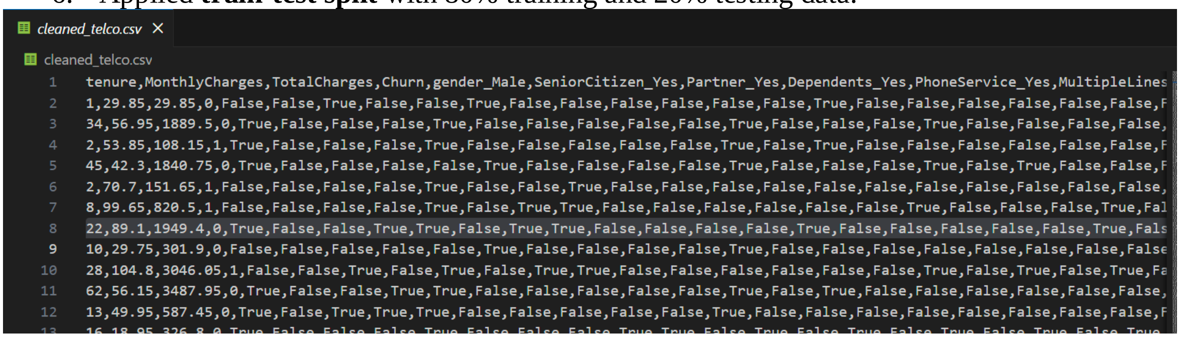
Target Variable

- **Churn** (Yes = 1, No = 0)

Data Preprocessing Performed

To prepare the dataset for model training, the following preprocessing steps were applied:

1. Converted the TotalCharges column from string to numeric.
2. Removed rows with missing or blank values.
3. Dropped the customerID column (not required for prediction).
4. Converted categorical variables into numerical format using **one-hot encoding**.
5. Normalized and prepared the dataset for model training.
6. Applied **train-test split** with 80% training and 20% testing data.




```
tenure,MonthlyCharges,TotalCharges,Churn,gender_Male,SeniorCitizen_Yes,Partner_Yes,Dependents_Yes,PhoneService_Yes,MultipleLines
1,29.85,29.85,0,False,False,True,False,False,True,False,False,False,False,False,False,True,False,False,False,False,False,False,F
34,56.95,1889.5,0,True,False,False,False,True,False,False,False,False,False,True,False,False,False,True,False,False,False,False,
2,53.85,108.15,1,True,False,False,False,True,False,False,False,False,False,True,False,True,False,False,False,False,False,False,F
45,42.3,1840.75,0,True,False,False,False,False,True,False,False,False,False,True,False,False,False,True,False,True,False,False,F
2,70.7,151.65,1,False,False,False,False,True,False,False,True,False,False,False,False,False,False,False,False,False,False,False,
8,99.65,820.5,1,False,False,False,False,True,False,True,True,False,False,False,False,False,False,True,False,False,False,True,Fal
22,89.1,1949.4,0,True,False,False,True,True,False,True,True,False,False,False,False,True,False,False,False,False,False,True,Fals
10,29.75,301.9,0,False,False,False,False,False,True,False,False,False,False,True,False,False,False,False,False,False,False,False
28,104.8,3046.05,1,False,False,True,False,True,False,True,True,False,False,False,False,False,False,True,False,True,False,True,Fa
62,56.15,3487.95,0,True,False,False,True,True,False,False,False,False,False,True,False,True,False,False,False,False,False,False,
13,49.95,587.45,0,True,False,True,True,True,False,False,False,False,False,True,False,False,False,False,False,False,False,False,F
```

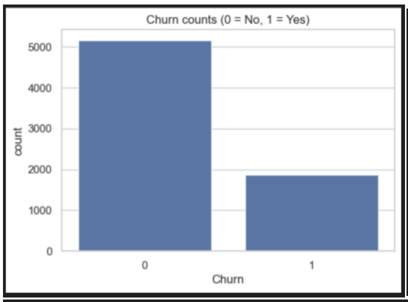
Project Workflow

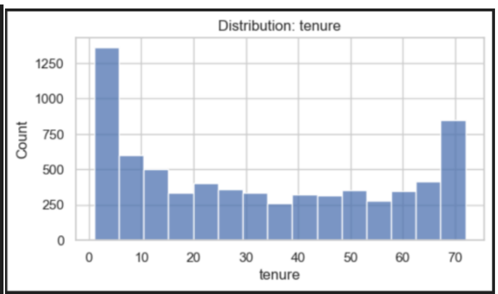
The project follows a structured and systematic pipeline that is commonly used in machine learning:

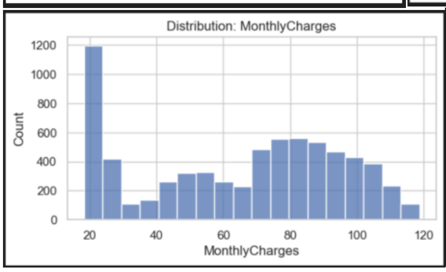
Step 1: Data Loading
Load the raw dataset using pandas.
Step 2: Data Cleaning
Convert datatypes, remove missing values, and clean inconsistencies.
Step 3: Exploratory Data Analysis (EDA)
Visualize distributions, relationships, and understand churn patterns.
Step 4: Feature Engineering
Convert categorical features using one-hot encoding.
Step 5: Model Training
Train three different machine learning models and compare the results.
Step 6: Model Evaluation
Evaluate each model using Accuracy, Precision, Recall, and F1-score.
Step 7: Deployment Demo
Create a simple Flask app that loads a sample customer row and predicts churn.
This workflow ensures a complete end-to-end churn prediction pipeline.

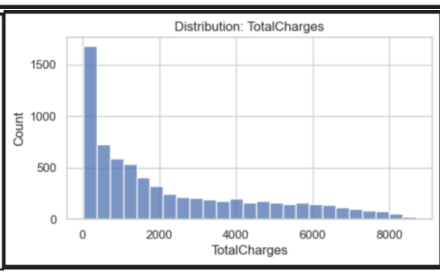
Exploratory Data Analysis (EDA)

EDA plays an important role in understanding how different features behave and how they affect churn. We visualized the distribution of churn values and several numerical features.









Results

Three machine learning models were trained and evaluated:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **XGBoost Classifier**

Each model was evaluated on:

- Accuracy
- Precision
- Recall
- F1-score

These metrics help understand both correctness and reliability of the model, especially since churn prediction is an **imbalanced dataset problem**.

Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	**0.7981**	**0.6355**	**0.5641**	**0.5977**
Random Forest	0.7882	0.6301	0.4919	0.5525
XGBoost	0.7782	0.5890	0.5481	0.5678

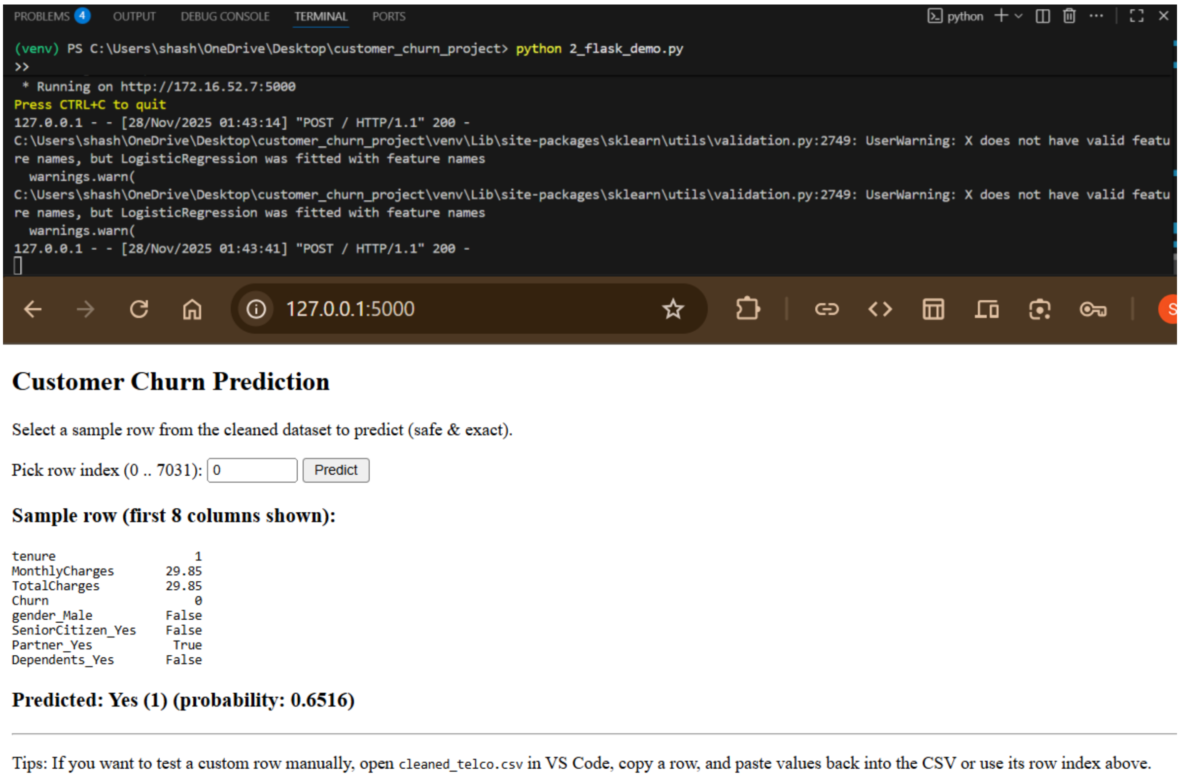
Conclusion From Results

- Logistic Regression performed the best.
- Its F1-score (0.5977) is the highest.
- It balances precision & recall better than other models.
- Simpler model but strong performance → good for real-world deployment.

Flask Web Application Demo

As the final stage, a simple Flask application was built to demonstrate how the trained model could be used in a real-time environment. The app loads a random customer row and predicts whether that customer will churn.

This demonstrates how machine learning models can be integrated with websites and dashboards.



Conclusion

In this project, a complete machine learning pipeline was developed to predict customer churn using the Telco dataset. The preprocessing and EDA steps helped reveal important churn patterns. Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were trained and compared.

Among these, Logistic Regression achieved the best overall performance, offering the highest F1-score, making it suitable due to its simplicity and interpretability.
The Flask demo successfully shows how a prediction model can be integrated into a web application for real-time decision-making.

Future enhancements may include hyperparameter tuning, handling dataset imbalance using SMOTE, feature importance using SHAP, and deploying the model on a cloud platform.