Customer Churn Prediction using Machine Learning

Student Name :Balasri.A

Register Number :732323106005

Institution : SSM College Of Engineering

Department :B.E.,ECE

Date of Submission :23.04.2025

Project Idea

- This project focuses on analyzing customer behavior and identifying patterns that lead to churn. Using machine learning, the goal is to proactively detect churn risks and support
- business decisions aimed at improving customer retention.

1.Problem Statement

- Customer churn significantly impacts the revenue and growth of subscription-based and service -oriented businesses. Traditional methods often fail to identify subtle behavioral patterns or risk
- factors that indicate customer dissatisfaction. This project aims to develop a machine learning model that accurately predicts customer churn by analyzing historical data and uncovering hidden patterns.

2.Objectives of the Project

- 1. Analyze historical customer data and identify factors influencing churn.
- 2. Design and train machine learning models for accurate churn prediction.
- 3. Evaluate and compare model performance using appropriate metrics.
- 4. Uncover actionable insights and hidden patterns in customer behavior.
- 5. Provide a basic interactive tool (optional) for real-time churn prediction.

3.Scope of the Project

In Scope:

1. Data Collection and Integration:

Use publicly available datasets (e.g., Telco Customer Churn from Kaggle).
Integrate various data attributes like customer demographics, services subscribed, billing info, and tenure.

2. Data Preprocessing and Cleaning:
Handle missing values, outliers, and data inconsistencies. Encode categorical variables and scale numerical data.

3. Exploratory Data Analysis (EDA):
Analyze feature distributions and relationships.
Identify patterns and trends that correlate with churn.

4. Feature Engineering:
Create derived variables that improve model performance.
Use domain knowledge to transform or combine existing features.

5. Model Development:
Apply machine learning models such as Logistic Regression, Decision Trees, Random Forest, XGBoost, etc.
Optimize models using cross-validation and hyperparameter tuning.

6. Model Evaluation:
Evaluate using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Compare models to select the best-performing one.

7. Insights and Recommendations:
Identify key factors driving churn
Provide actionable business recommendations to reduce churn.

8. (Optional) Deployment and Tool Development:
Build a user interface using Streamlit or Flask for input and churn prediction. Present results and predictions through visual dashboards.

Out of Scope:
Real-time streaming data or live production deployment.
Direct implementation of business strategies or campaigns to retain customers.
Integration with customer relationship management (CRM) systems. Handling of very large-scale or proprietary telecom datasets.

4. Data Sources

Primary Dataset:

- **Name:** Telco Customer Churn Dataset

- **Source:** Available on Kaggle

- **Format:** CSV file
- **Records:** Approximately 7,000 customer entries
- **License:** Public dataset (free for educational and research use)

Key Features in the Dataset:

Supplementary Data (Optional for Enhancement):

If you want to enhance your model or simulate real-world conditions, you could consider the following additional sources:

Customer Support Interactions (simulated or public samples): To include sentiment analysis or ticket volumes.

Customer Feedback/Survey Data: For subjective feedback and satisfaction levels.

Website/App Usage Data: Frequency of logins, time spent, features used.

External Demographic Data: Regional income levels, urban/rural classification, etc.

Data Preparation Tasks:

Convert TotalCharges to numerical format (often appears as string due to missing values).

Handle missing values in numerical and categorical fields.

Encode categorical variables using Label Encoding or One-Hot Encoding.

Scale numerical features such as MonthlyCharges, TotalCharges, and tenure.

High-Level Methodology – Summary

1. Data Collection
Import the customer dataset from Kaggle or other sources.
Gather relevant features such as demographics, service usage, billing, and churn labels.

2. Data Cleaning
Handle missing values, incorrect data types, and outliers.
Ensure the dataset is ready for analysis and modeling.

3. Exploratory Data Analysis (EDA)

Use visualizations and statistics to explore data patterns.

Understand relationships between features and the churn variable.

4. Feature Engineering

Create new meaningful features (e.g., tenure groupings).

Encode categorical variables and scale numeric features.

5. Model Building

Train multiple machine learning models (e.g., Logistic Regression, Random Forest, XGBoost).

Use grid search or cross-validation for hyperparameter tuning.

6. Model Evaluation

Evaluate model performance using metrics like accuracy, precision, recall, F1-score, and ROCAUC.

Select the best model based on both performance and interpretability.

7. Visualization & Interpretation

Visualize key insights and feature importances.

Interpret model outputs to support business understanding and decision-making.

8. Deployment (Optional)

Build a user-friendly web app using Streamlit or Flask.

Allow users to input data and get churn predictions in real-time

Tools and Technologies:

- Programming Language: Python

- Libraries: pandas, NumPy, matplotlib, seaborn, scikit-learn, XGBoost, Streamlit -

Environment: Jupyter Notebook / Google Colab / VS Code   - Version Control: Git/GitHub

Name	Responsibilites
Arasu.R	• Oversees project work flow,manages timeline,and coordinates tasks.
Ashwini.M	• Perform EDA,visualizies trends and extracts statistical insights.
Balasri.A	• Handles data preprocessing , model building,and model optimization.
Deepansri.B	• Prepares reports, presentations, and works on UI & tool deployment (Streamlit & Flask).