PHASE-2 SUBMISSION

Predicting Customer Churn Using Machine Learning to Uncover Hidden Patterns

Student Name: Bejoymjose

Register Number: 513523104004

Institution: Annaimira college of engineering &technology

Department: Computer science

Date of Submission: 3-5-2025

Github Repository Link:
https://github.com/Bejoy2/phase2.git

1. Problem Statement:

Customer churn—when customers stop doing business with a company—represents a significant loss of revenue. Many companies struggle to understand why churn occurs and how to proactively retain customers. This project aims to develop a machine learning-based model that accurately predicts customer churn, enabling businesses to identify at-risk customers and address issues before it's too late.

2. Project Objectives :

- *Analyze historical customer data to uncover patterns leading to churn.*

- *Uncover hidden patterns in customer behavior and service usage that correlate with churn.*
- *Build and evaluate machine learning models to predict churn.*
- *Identify key drivers of customer churn through feature importance*
- *Visualize key churn drivers to enhance model interpretability for nontechnical decision-makers.Visualize insights to support business decisionmaking.*
- *Provide actionable insights to help business stakeholders implement targeted customer retention strategies.*
- *Create a reproducible workflow for churn prediction that can be adapted across industries.*

The objectives have evolved from simple prediction to also emphasizing interpretability and real-world application after exploring the data sources and observing their seasonal and geographic patterns.

3. Flowchart of the Project Workflow:

Imbalanced Customer Churn Data as Input

↓

Data Cleaning

↓

Exploratory Data Analysis (EDA)

↓

Feature Engineering(Feature Selection)

↓

Model Building (Training& Evaluation)

↓

Model Evaluation & Comparison

↓

Model Interpretation & Visualization

↓

Deployment or Reporting

4. Data Description :

The dataset includes the following customer attributes:

- ***Demographic:*** *Age, Gender, Location*
- ***Account Information:*** *Subscription Type, Tenure, Monthly Charges*
- ***Usage Metrics:*** *Total Calls, Internet Usage, Support Tickets Raised*
- ***Behavioral:*** *Payment Method, Contract Type*
- ***Target:*** *Churn (Yes/No)*

5. Data Preprocessing

- *Handled missing values via imputation*
- *Encoded categorical variables using One-Hot and Label Encoding*
- *Scaled numerical features using StandardScaler*
- *Removed duplicate entries and irrelevant columns*
- *Balanced the dataset using SMOTE (if imbalanced)*

6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in predicting customer churn. It helps you understand patterns, trends, and relationships in your dataset that could contribute to customers leaving. Here's a structured approach to EDA for customer churn prediction:

1. Understand the Dataset

Load and inspect the dataset (.head(), .info(), .describe())

Check for null or missing values

Understand column types (categorical vs. numerical)

2. Target Variable Analysis

Plot the churn distribution (e.g., bar plot of churned vs. retained customers)

Compute churn rate: churned customers / total customers

3. Univariate Analysis

Categorical features: Count plots / bar plots (e.g., gender, contract type, payment method)

Numerical features: Histograms / boxplots (e.g., tenure, monthly charges, total charges)

4. Bivariate Analysis

Compare features against churn:

Boxplots (e.g., MonthlyCharges vs. Churn)

Stacked bar charts (e.g., Contract type vs. Churn)

Grouped means or medians (e.g., average tenure by Churn)

7. Feature Engineering

Created tenure groups (e.g., new, medium, long-term)

Aggregated usage metrics into customer engagement scores

Derived features from timestamps and payment history

Performed feature selection using mutual information and tree-based importance

8. Model Building

- ***Trained various ML models:*** Logistic Regression, Decision Trees, Random Forest, XGBoost, and SVM
- Split data into training and test sets (e.g., 80/20)
- Evaluated models using metrics like accuracy, precision, recall, F1score, and ROC-AUC
- Chose the best-performing model based on both accuracy and interpretability

9. Visualization of Results & Model Insights

- *Confusion matrices for model evaluation*
- ***ROC** curves to visualize trade-offs*
- *Feature importance plots to interpret the model*
- ***SHAP** values or **LIME** for individual prediction explanations*
- *Dashboard-style visuals summarizing insights for stakeholders*

10. Tools and Technologies Used

Programming: *Python*

Libraries: *Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn, SHAP, LIME*

Data Handling: *Jupyter Notebook, Excel/CSV files*

Version Control: *GitHub*

Optional Deployment: *Streamlit / Flask*

11. Team Members and Contributions

ANUSHA : Data Collection and Integration: Responsible for sourcing datasets, connecting APIs, and preparing the initial dataset for analysis.

AASHIDA : Data Cleaning and EDA: Cleans and preprocesses data, performs exploratory analysis, and generates initial insights.

BALAJI : Feature Engineering and Modeling: Works on feature extraction and selection; develops and trains machine learning models.

BEJOYMJOSE : Evaluation and Optimization: Tunes hyperparameters, validates models, and documents performance metrics.

BRINDHA : Documentation and Presentation: Compiles reports, prepares visualizations, and handles presentation and optional deployment.