Customer Churn Prediction using Machine Learning

Student Name: **Ashwini M**

Register number: **732323106004**

Institution: **SSM College of Engineering**

Department: **B.E/ECE**

Date of Submission: **29/04/2025**

Github Repository Link:https://github.com/ashwini2500/Predicting-customer-churn-using-machine-learning-to-uncover-hidden-patterns

PHASE-2

1. Problem Statement

Customer churn, the phenomenon where customers discontinue their service or subscription with a company, poses a serious challenge to the long-term profitability and sustainability of businesses, especially those operating in highly competitive and subscription-based markets such as telecommunications, SaaS, banking, and streaming services. Retaining existing customers is not only more cost-effective than acquiring new ones, but also critical for maintaining steady revenue growth.

Despite the availability of large volumes of customer data, traditional analytics techniques often fall short in detecting the subtle behavioral, transactional, and service usage patterns that precede churn. These methods typically rely on static thresholds or descriptive statistics that cannot adapt to complex customer dynamics or changes over time.

The goal of this project is to leverage the power of machine learning to proactively identify customers who are likely to churn by analyzing their historical data, including demographics, usage behavior, and billing history. By doing so, businesses can implement timely interventions such as targeted offers, improved support, or loyalty programs to improve retention rates.

This problem is addressed as a binary classification task where the model outputs a probability indicating whether a customer will churn or not. The solution must be accurate, interpretable, and scalable, and should offer actionable insights that decision-makers can trust

2. Project Objectives

1. Analyze Historical Data to Identify Churn Drivers:

Conduct a thorough analysis of past customer data to understand which features most significantly influence churn. These may include contract type, tenure, payment method, service usage, and demographic characteristics.

2. Design and Implement Robust Machine Learning Models:

Build multiple classification models such as Logistic Regression, Random Forest, and XGBoost to predict customer churn. These models should be capable of capturing both linear and non-linear relationships in the data.

3. Evaluate Models Using Reliable Performance Metrics:

Use industry-standard metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC to evaluate the effectiveness of each model. These metrics ensure the model is not only accurate but also balanced in identifying true churners and non-churners.

4. Extract Actionable Business Insights from the Model:

Go beyond prediction by interpreting feature importances and model outputs to provide strategic recommendations. For example, identifying that customers with month-to-month contracts and high monthly charges are at higher risk of churn can directly inform retention strategies.

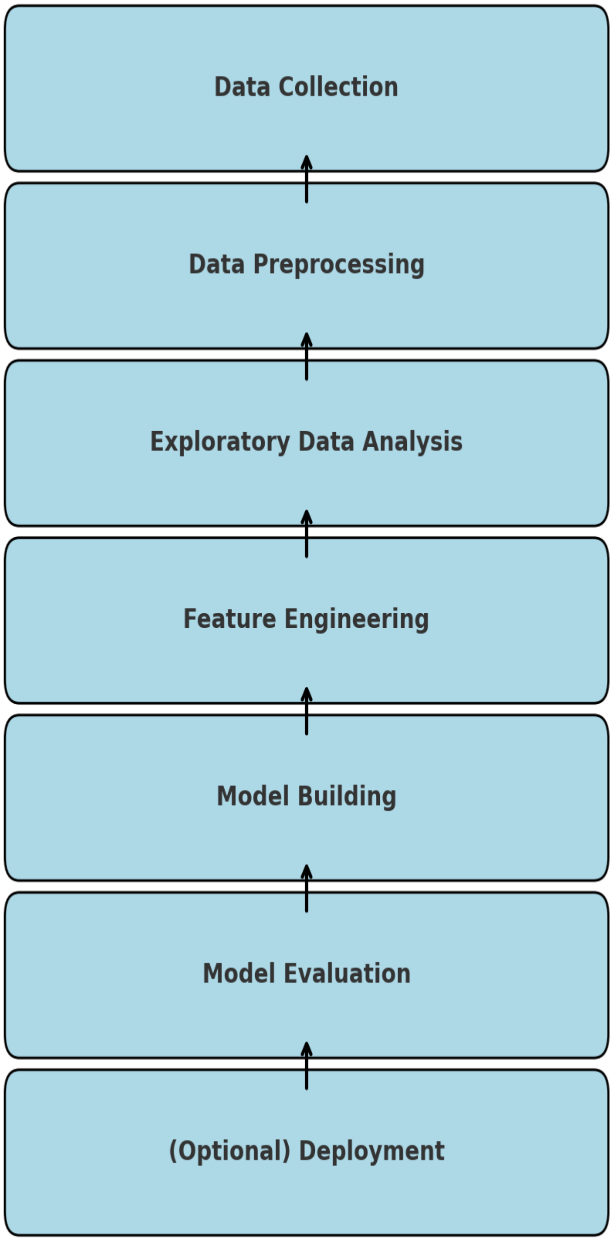
5. Develop a Basic Interactive Prediction Tool (Optional):

Build a user interface using Streamlit or Flask that allows non-technical stakeholders to input customer data and receive churn predictions. This tool could support frontline staff in customer service or marketing in real-time decision-making.

6. Ensure Model Interpretability and Explainability:

Use tools like SHAP (SHapley Additive exPlanations) or feature importance plots to make the model's decisions transparent and understandable, which is essential for trust and adoption in business contexts.

3. Flowchart of the Project Workflow



4. Data Description

- Dataset Name: Telco Customer Churn
- Source: Kaggle
- Format: CSV
- Records & Features: ~7,000 records, multiple categorical and numerical attributes
- Target Variable: Churn (Yes/No)

- Static or Dynamic: Static dataset
- Key Features: demographics, service usage, tenure, billing, and payment method

Data set Link 🔗 https://www.kaggle.com/datasets/blastchar/telco-customer-churn

5. Data Preprocessing

- Converted TotalCharges to numeric (handled missing values).
- Removed irrelevant or uniform columns (if any).
- Encoded categorical features using Label Encoding / One-Hot Encoding.
- Scaled numerical fields (MonthlyCharges, tenure, TotalCharges) using MinMaxScaler.
- Outliers checked via boxplots and IQR; handled as needed.

6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a critical phase in understanding the structure, trends, and anomalies in the dataset before applying any machine learning algorithms. It helps uncover patterns, detect outliers, and form hypotheses about relationships between variables.

6.1 Univariate Analysis

Focuses on analyzing the distribution of individual features:

Churn Distribution:

The target variable Churn is imbalanced, with a higher percentage of non-churned customers. This necessitates handling class imbalance during modeling.

Numerical Features (Histograms & Boxplots):

tenure, MonthlyCharges, and TotalCharges were analyzed to understand their spread and skewness.

Customers with lower tenure and higher monthly charges show higher churn rates.

Categorical Features (Count Plots):

Features like Contract, PaymentMethod, InternetService, and SeniorCitizen were plotted.

Most customers who churned had month-to-month contracts and used electronic check payments.

6.2 Bivariate Analysis

Examines the relationship between independent features and the target (Churn):

Tenure vs Churn:

Churn is inversely proportional to tenure. Long-term customers tend to stay, while new users churn more.

Contract Type vs Churn:

Month-to-month contracts showed the highest churn rate, whereas customers on two-year contracts had the lowest churn rate.

Payment Method vs Churn:

Customers using electronic check payment had significantly higher churn rates.

Service Usage Patterns:

Users without internet service showed the least churn, indicating simplicity in service may lead to better retention.

6.3 Multivariate Analysis

Evaluates how multiple variables interact together to influence churn:

Correlation Heatmap (for Numerical Features):

Displayed strong positive correlation between MonthlyCharges and TotalCharges, but weak correlation with Churn.

Grouped Bar Charts & Stacked Plots:

Visualized churn rates across combinations like InternetService & Contract, Gender & SeniorCitizen, etc.

Found that senior citizens with month-to-month contracts and fiber optic internet are more likely to churn.

6.4 Key Insights from EDA

Customers with month-to-month contracts, higher monthly charges, and electronic check payments are most likely to churn.

Churn probability is lower among long-tenured users and those with automatic payments.

Contract type is the most important categorical indicator, and tenure is the most critical numerical feature for predicting churn.

There’s no significant difference in churn between genders, indicating gender is not a strong predictive factor.

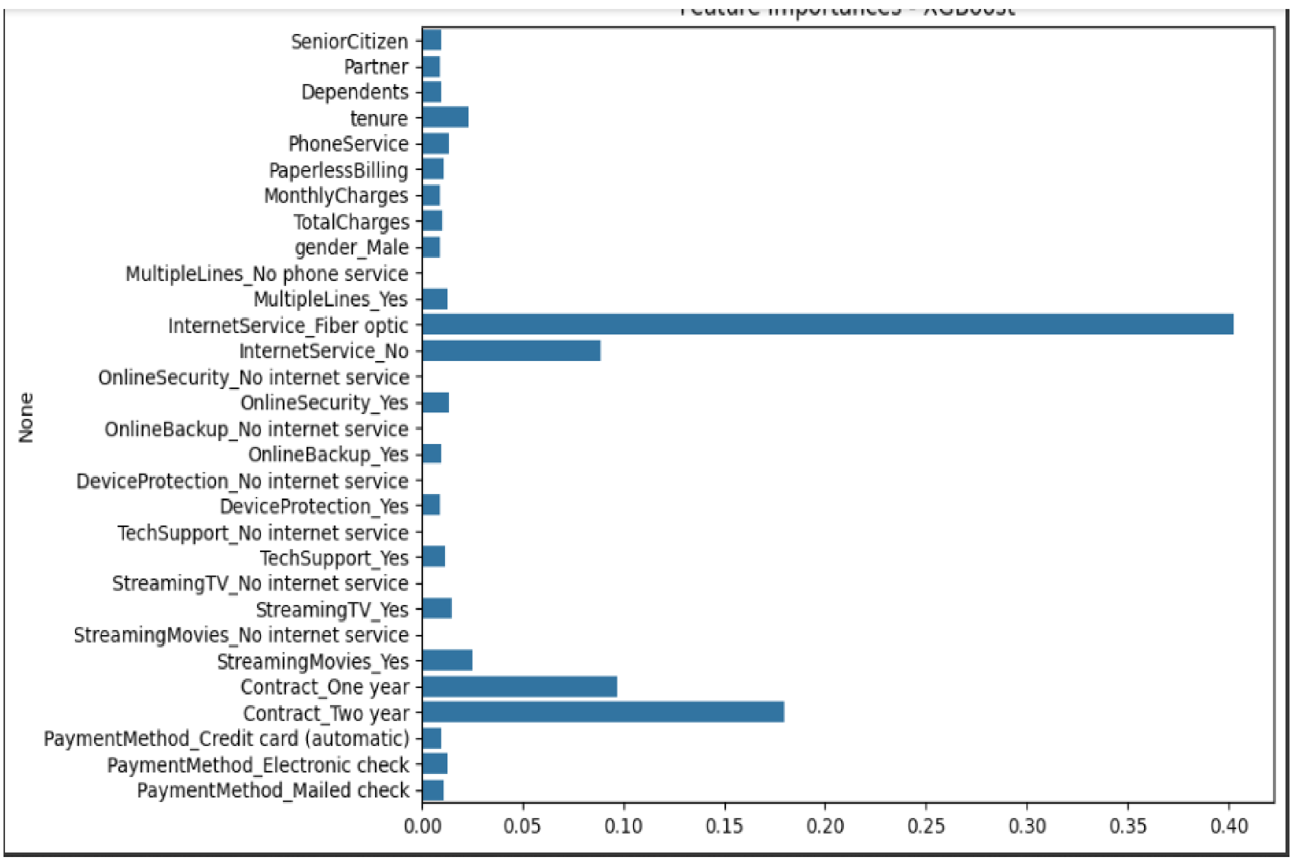
7. Feature Engineering

- Created tenure_group (e.g., 0–12, 13–24 months).
- Derived has_multiple_services based on service counts.
- Encoded binary features like Partner, Dependents.
- Removed multicollinear or redundant features.
- Applied scaling to numerical attributes.

8. Model Building

- Algorithms Used: Logistic Regression, Decision Tree, Random Forest, XGBoost.
- Model Selection Rationale: Interpretable and fast (Logistic); Tree-based for categorical and non-linear patterns.
- Train-Test Split: 70% train, 30% test.
- Validation: Cross-validation (5-fold).
- Tuning: GridSearchCV for hyperparameters.

9. Visualization of Results & Model Insights



Model Metrics Table:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.803318	0.652083	0.557932	0.601345	0.725060
Decision Tree	0.710900	0.459098	0.490196	0.474138	0.640514
Random Forest	0.782938	0.619490	0.475936	0.538306	0.685031
XGBoost	0.781517	0.606838	0.506239	0.551992	0.693726

Feature Importance visualized using bar plots. Top features include Contract Type, Tenure, MonthlyCharges.

ROC Curve and Confusion Matrix (visual to be added).

Streamlit-based web app developed for user interaction.

10. Tools and Technologies Used

- Programming Language: Python
- Notebook Environment: Google Colab / Jupyter
- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost, Streamlit / Flask

11. Team Members and Contributions

Name	Responsibilities
Arasu.R	Project lead, coordination, workflow management
Ashwini.M	EDA, visualizations, statistical insight generation
Balasri.A	Data preprocessing, model building, optimization
Deepansri.B	Documentation, reporting, deployment (UI design with Streamlit/Flask)