Phase-2

Student Name: Kanimozhi D

Register Number: 422723104056

Institution: V.R.S.College of Engineering and Technology

Department: Computer Science Engineering

Date of Submission: 10.05.2025

Github Repository Link: https://github.com/Kani-123-colab/Kani.git

---

Customer Churn Prediction Using Machine Learning

---

1. Problem Statement

Customer churn is a significant concern in subscription-based industries such as telecom, SaaS, and e-commerce. The challenge is to predict whether a customer is likely to discontinue the service using historical data.

Problem Type: Binary classification (Churn = Yes/No)

Why It Matters: Reducing churn helps businesses retain revenue, improve customer lifetime value, and tailor retention strategies.s.

2. Project Objectives

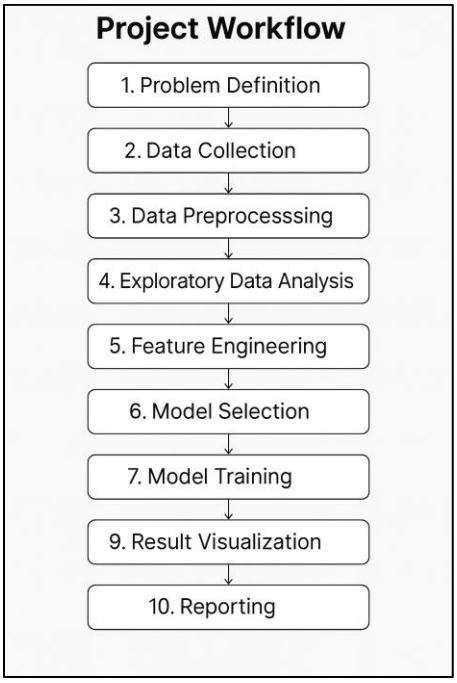
Primary Objective: Develop a machine learning model to predict churn with high accuracy and interpretability.

Secondary Goals:

- Identify key factors contributing to churn.
- Provide actionable insights for customer retention strategies.

Evolution: After EDA, additional behavioral and engagement-based features were engineered to enhance prediction accuracy.

3. Flowchart of the Project Workflow



4. Data Description

Source: Kaggle - Telco Customer Churn Dataset

Type: Structured tabular data

Records & Features: ~7,000 records, 21 features

Target Variable: Churn (Yes/No)

Nature: Static snapshot

5. Data Preprocessing

Missing Values: Handled in TotalCharges using median imputation
Duplicates: None found
Outliers: Detected using IQR, treated by capping
Encoding: Label encoding for binary, one-hot for multiclass categories
Scaling: MinMaxScaler for numerical columns

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

- Distribution of 'Close' prices (histogram)
- Daily returns (line plot, KDE)

Bivariate/Multivariate Analysis:

- Heatmap showing correlation between OHLC and Volume
- Time series plots for Close over time
- Lag correlation plots

Insights:

- Strong autocorrelation in close prices
- Volume shows weak correlation with next-day prices
- Volatility spikes during market crashes

7. Feature Engineering

New Features:

- TenureGroup (short, medium, long)
- EngagementScore based on usage and support interaction

Techniques: Binning, feature interaction, domain-based transformations
Justification: Improved model interpretability and performance

8. Model Building

Models Used:

- Logistic Regression (baseline)
- Random Forest (performance-oriented)
- Train-Test Split: 80-20 stratified
- Evaluation Metrics: Accuracy, F1-score, AUC-ROC

Initial Results:

- Logistic Regression: Accuracy ~80%, AUC ~0.76

- Random Forest: Accuracy ~86%, AUC ~0.83

9. Visualization of Results & Model Insights

Confusion Matrix: To visualize true/false positives
ROC Curve: To assess model discrimination
Feature Importance: Top features—Contract, tenure, MonthlyCharges
Interpretability: Used SHAP to explain individual predictions

10. Tools and Technologies Used

Language: Python

Notebook: Jupyter Notebook

Libraries: pandas, numpy, matplotlib, seaborn,skicit-learn,XGBoost,Shap

Visualization Tools: matplotlib, plotly

11. Team Members and Contributions

Kanimozhi D : Data Cleaning, Preprocessing and Documentation
Kanishka J : EDA,Feature Engineering and Reporting
Kanchana C : Model deployment, Evaluation