**Phase-2 Submission Template Student Name:** Rohith S

**Register Number:** 510623104086

**Institution:** CAHCET

**Department:** B.E.C.S.E

**Date of Submission:** 09-05-2025

**Github Repository Link: https://github.com/rohith0608/Predicting-customer-churn.git**

# Problem Statement

**Type of Problem**: Supervised classification (binary or multi-class, depending on the business use case)  
Customer churn refers to when customers stop using a product or service. Retaining customers is significantly more cost-effective than acquiring new ones. However, identifying churn-prone customers in time is challenging due to the complexity and volume of customer data. This project focuses on developing a churn prediction model using machine learning to detect underlying behavioral patterns and risk factors that lead to churn.

**Why It Matters**:

* 💸 Reduces revenue loss by identifying at-risk customers early
* 📊 Improves decision-making for marketing and retention strategies
* 🧠 Enhances understanding of customer behavior
* 📈 Optimizes customer lifetime value

# Project Objectives

**Technical Objectives**:

1. Build a predictive model to classify whether a customer is likely to churn.
2. Perform comprehensive data preprocessing, including cleaning and feature transformation.
3. Compare various machine learning models (e.g., Logistic Regression, Random Forest, XGBoost).
4. Evaluate model performance using classification metrics.
5. Visualize key features and churn trends for business insights.

**Model Goals**:

* Achieve an F1-score of at least 0.80
* Prioritize interpretability and business value
* Adapt objectives based on data insights (e.g., class imbalance handling)

# Flowchart of the Project Workflow

[Start]

↓

[Data Collection]

↓

[Preprocessing]

↓

[EDA]

↓

[Feature Engineering]

↓

[Model Selection & Training]

↓

[Evaluation]

↓

[Visualization & Insights]

↓

[Conclusion]

↓

[End]

# Data Description

**Dataset Name**: Telco Customer Churn or similar

**Source**: Kaggle or company CRM system

**Type**: Structured tabular data

**Records**: ~7,000–10,000 rows

**Features**: Demographics, services, usage behavior, tenure, contract type, billing

**Target Variable**: Churn (Yes/No)

**Dataset Nature**: Static

# Data Preprocessing

**Missing Values**: Remove rows with nulls or impute if minimal

**Duplicates**: Removed exact duplicates

**Outliers**: Truncated extreme usage or bill values

**Encoding**:

* + Categorical features: One-hot or Label Encoding

**Feature Scaling**:

* + Normalize/standardize numeric columns like tenure, charges

# 6.Exploratory Data Analysis (EDA)

* ***Univariate****: Countplots for churn, histograms for tenure/charges*
* ***Bivariate****: Boxplots, grouped bar charts (e.g., churn vs. contract type)*
* ***Multivariate****: Correlation matrix, pairplots*
* ***Key Insights****:*
  + *Higher churn in month-to-month contracts*
  + *Auto-payment customers churn less*
  + *Short tenure often correlates with churn*

# 7.Feature Engineering

Created binary features from service columns

Extracted tenure buckets

Created interaction terms (e.g., monthly charges × tenure)

Applied scaling to numeric fields

Removed uninformative features (e.g., customer ID)

# 8.Model Building

* ***Models Used****: Logistic Regression, Random Forest, XGBoost*
* ***Train-Test Split****: 80/20 with stratification*
* ***Performance Metrics****:*
  + *Accuracy, Precision, Recall, F1-score, ROC-AUC*
* ***Initial Findings****:*
  + *Random Forest and XGBoost outperform baseline Logistic Regression*

# 9.Visualization of Results & Model Insights

 **Confusion Matrix**: Shows churn detection performance

 **ROC Curve**: AUC > 0.85 for best models

 **Feature Importance**: Monthly charges, contract type, and tenure are top predictors

 **Insights**:

* Target auto-renewal users and new joiners for retention
* High charges with low tenure is a churn risk signal

# 

# 10.Tools and Technologies Used

 **Programming Language**: Python

 **IDE/Notebook**: Google Colab, Jupyter Notebook

 **Libraries**:

* pandas, numpy, matplotlib, seaborn
* scikit-learn, xgboost

 **Visualization**: Seaborn, Plotly

 **(Optional) Deployment Tools**: Streamlit or Flask

# 

# 11.Team Members and Contributions

| ***Name*** | ***Contribution*** |
| --- | --- |
| ***Salman Faris T S*** | *Model development, evaluation, visualization, deployment* |
| ***Rohith S*** | *Data collection, cleaning, EDA, feature engineering, report writing* |