**Phase-3 Submission Template**

**Student Name:**V Dharanish

**Register Number:**732123104028

**Institution:**Nandha College Of technology

**Department:** B.E.Computer Science And ENgineering

**Date of Submission:** 04/05/2025

**Github Repository Link:**

### **1. Problem Statement**

*Customer churn refers to when existing customers stop doing business with a company. High churn rates negatively impact profitability and growth. Hence, predicting churn is crucial in sectors like banking, telecom, and e-commerce. This project aims to build a machine learning model that predicts whether a customer will churn based on behavioral and demographic data. This is a* ***binary classification problem****, where the target is whether a customer will churn (Yes or No).*

### **2. Abstract**

*Customer retention is more cost-effective than acquisition, making churn prediction a vital task for businesses. This project uses machine learning to identify patterns in customer behavior that indicate a risk of churn. Using a real-world dataset, we preprocess the data, explore patterns, engineer features, and train multiple machine learning models to classify customers as churned or not. Evaluation metrics such as accuracy, F1-score, and ROC AUC are used to select the best model. The final model is deployed via a user-friendly interface for real-time churn prediction. This approach provides actionable insights and aids businesses in proactively addressing churn risks.*

### **3. System Requirements**

* + ***Hardware:***
  + *Minimum 8GB RAM*
  + *Intel i5 or equivalent processor (for training ML models)*
  + ***Software:***
  + *Language :Python*
  + *IDE: Jupyter Notebook / Google Colab*
  + *Libraries: pandas, numpy, scikit-learn, seaborn, matplotlib, xgboost, streamlit (for deployment)*

### **4. Objectives**

* * ***Predict Customer Churn*** *Build a machine learning model to classify whether a customer is likely to churn (exit) or stay with the company.*

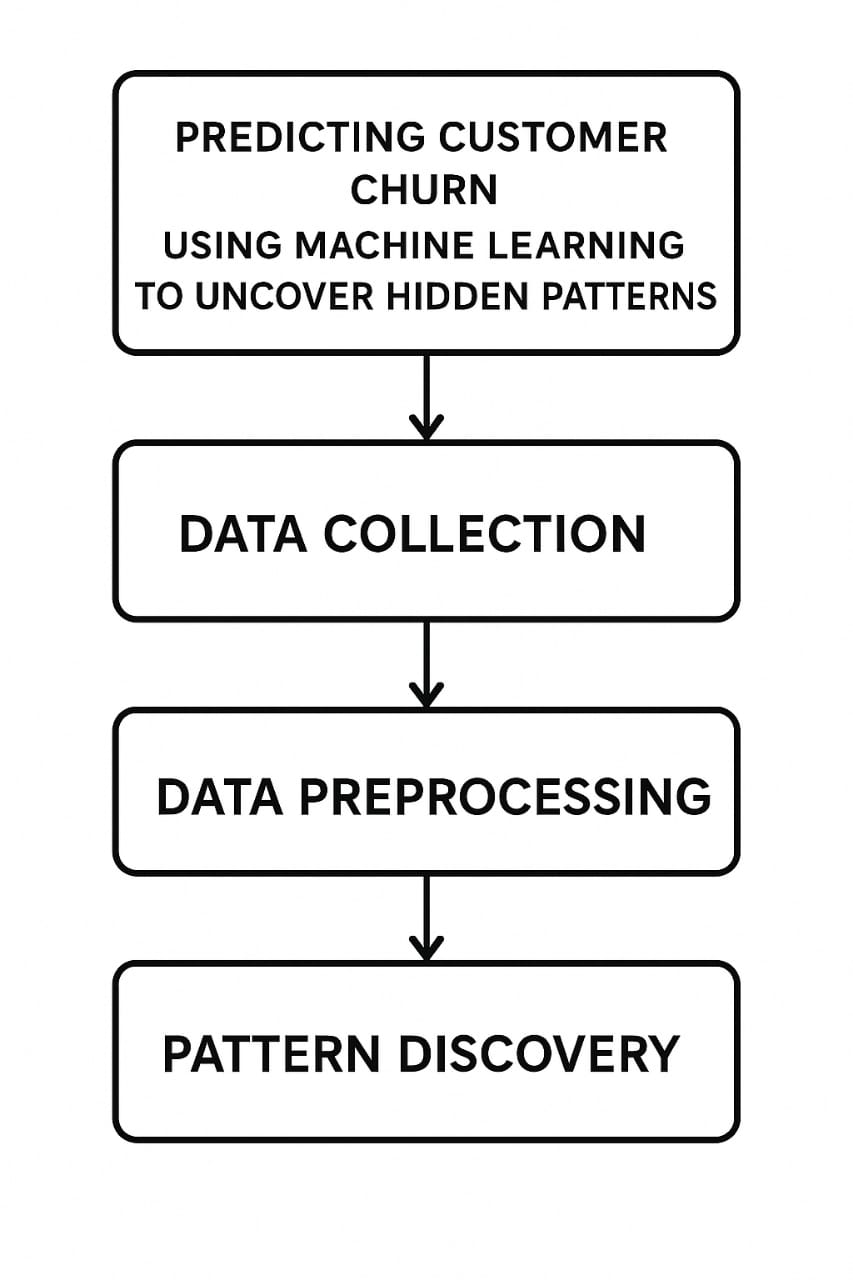
** ***Identify Key Drivers of Churn*** *Analyze features such as age, balance, geography, and activity level to understand what factors contribute most to churn.*

** ***Improve Customer Retention Strategies*** *Provide actionable insights that help the business design targeted strategies for retaining high-risk customers.*

 **Compare Multiple Machine Learning Models**  
Evaluate the performance of various classification models (e.g., Logistic Regression, Random Forest, XGBoost) to choose the most effective one.

 **Create a Clean and Usable Dataset**  
Perform preprocessing steps like handling missing values, encoding categorical data, and scaling numerical features for optimal model performance

**5. Flowchart of Project Workflow**

****

### **6. Dataset Description**

** ***Source****: Kaggle – Bank Customer Churn Prediction*

** ***Type****: Public dataset*

** ***Size and Structure****: 10,000 rows × 14 columns  
Columns include: CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited*

### 

### **7. Data Preprocessing**

* ** ***Handled Missing Values****: None present in this dataset.*
* ** ***Duplicates****: Removed based on CustomerId.*
* ** ***Outliers****: Detected using IQR method and capped where necessary.*
* ** ***Encoding****:*
* *Geography, Gender: Label/One-Hot Encoding.*
* ** ***Scaling****:*
* *Used StandardScaler for numerical features.*

### **8. Exploratory Data Analysis (EDA)**

** ***Tools****: Seaborn, Matplotlib*

** ***Visuals Used****: Histograms, boxplots, heatmaps, bar charts*

** ***Insights****:*

* *Customers with lower credit scores and higher ages tend to churn more.*
* *France had the highest churn rate among countries.*
* *Inactive members are more likely to churn.*

### 

### **9. Feature Engineering**

* ** ***New Features****:*
* *BalanceSalaryRatio = Balance / EstimatedSalary*
* *AgeGroup = Age bucketed into ranges*
* ** ***Feature Selection****: Used correlation matrix and feature importance from models (e.g., XGBoost)*
* ** ***Impact****:*
* *High balance-to-salary ratios, older age, and number of products were strong churn indicators.*

### **10. Model Building**

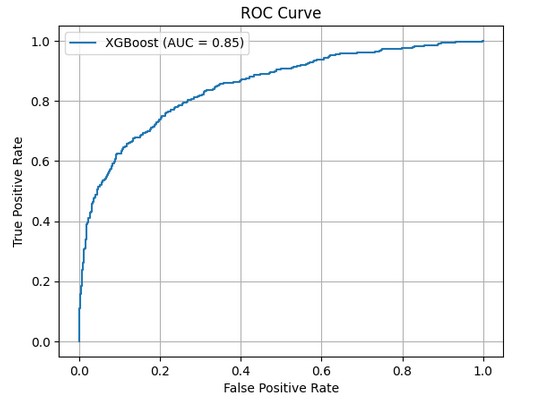
* ** ***Algorithms Used****:*
* *Logistic Regression (baseline)*
* *Decision Tree*
* *Random Forest*
* *XGBoost (best performance)*
* ** ***Why Chosen****:*
* *Mix of linear and ensemble models to balance bias-variance.*
* ** ***Training Output****:*
* *XGBoost showed highest accuracy and F1-score.*

### 

### **11. Model Evaluation**

* ** ***Metrics Used****:*
* *Accuracy*
* *Precision, Recall, F1-score*
* *ROC-AUC Score*
* ** ***Visuals****:*
* *Confusion Matrix*
* *ROC Curve*
* *Classification Report*

### **12. Deployment**

**

**13. Source code**

*import pandas as pd*

*import numpy as np*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.preprocessing import StandardScaler, LabelEncoder*

*from sklearn.metrics import classification\_report, confusion\_matrix, roc\_auc\_score, roc\_curve*

*from xgboost import XGBClassifier*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

*df = pd.read\_csv('Bank Customer Churn Prediction(1).csv')*

*df.drop('customer\_id', axis=1, inplace=True)*

*label\_encoder = LabelEncoder()*

*df['gender'] = label\_encoder.fit\_transform(df['gender'])*

*df = pd.get\_dummies(df, columns=['country'], drop\_first=True)*

*X = df.drop('churn', axis=1)*

*y = df['churn']*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)*

*scaler = StandardScaler()*

*X\_train = scaler.fit\_transform(X\_train)*

*X\_test = scaler.transform(X\_test)*

*model = XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss')*

*model.fit(X\_train, y\_train)*

*y\_pred = model.predict(X\_test)*

*y\_proba = model.predict\_proba(X\_test)[:,1]*

*print("Classification Report:")*

*print(classification\_report(y\_test, y\_pred))*

*print("Confusion Matrix:")*

*print(confusion\_matrix(y\_test, y\_pred))*

*print("ROC AUC Score:", roc\_auc\_score(y\_test, y\_proba))*

*fpr, tpr, \_ = roc\_curve(y\_test, y\_proba)*

*plt.figure(figsize=(8,6))*

*plt.plot(fpr, tpr, label="XGBoost (AUC = {:.2f})".format(roc\_auc\_score(y\_test, y\_proba)))*

*plt.plot([0,1], [0,1], 'k--')*

*plt.xlabel("False Positive Rate")*

*plt.ylabel("True Positive Rate")*

*plt.title("ROC Curve")*

*plt.legend()*

*plt.grid()*

*plt.show()*

**14. Future scope**

** ***Add real-time customer behavior tracking*** *to improve prediction accuracy.*

** ***Integrate with CRM tools*** *to automate retention offers for at-risk customers.*

** ***Incorporate deep learning models*** *(like LSTM) for time-series churn prediction in future.*

**13. Team Members and Roles**

* *V Dharanish : prepare the dataset and helps with data cleaning*
* *S Dharun kumar :Explores the data analysis (EDA) and feature engineer.*
* *S Chandru : prepare the documentation & reporting and model development*