Bank Customer Churn Prediction Challenge: Project Requirement Document

Problem Statement

Customer retention is critical for banks and financial institutions. Acquiring new customers costs significantly more than retaining existing ones. In this project, your task is to build a machine learning model that predicts whether a bank customer will churn (leave the bank) based on their demographic information, account details, and banking behavior.

Understanding the factors that lead to customer churn helps banks take proactive measures to retain valuable customers.

Your goal is to: - Analyze customer data and identify churn patterns - Build a classification model to predict customer churn - Provide actionable insights for customer retention

Dataset Information

Dataset Specifications

- **File:** `dataset/Churn_Modelling.csv`
- **Rows:** 10,000 customers
- **Columns:** 14 features (including target)
- **File Size:** ~1 MB
- **Missing Values:** None (clean dataset)
- **Class Distribution:** ~7,963 retained vs ~2,037 churned (~20% churn rate)

Column Descriptions

Column Name	Description
RowNumber	Row index (drop this)
CustomerId	Unique customer identifier (drop this)
Surname	Customer surname (drop this)
CreditScore	Customer’s credit score
Geography	Country of residence (France, Germany, Spain)
Gender	Customer gender (Male, Female)
Age	Customer age in years

Column Name	Description
Tenure	Number of years as a bank customer
Balance	Account balance
NumOfProducts	Number of bank products used
HasCrCard	Has credit card (1 = Yes, 0 = No)
IsActiveMember	Active member status (1 = Yes, 0 = No)
EstimatedSalary	Estimated annual salary
Exited	**Target variable** - Churned (1 = Yes, 0 = No)

Tasks to Complete

1. Data Exploration & Preprocessing

- Load and explore the dataset
- Drop irrelevant columns (RowNumber, CustomerId, Surname)
- Analyze feature distributions
- Investigate class imbalance (20% churn rate)
- Visualize churn patterns across different segments

2. Feature Engineering

- Encode categorical variables (Geography, Gender)
- Scale numerical features
- Consider creating derived features (e.g., Balance per Product, Age groups)
- Handle class imbalance (SMOTE, class weights, or threshold tuning)

3. Model Building

- Build a classification model (suggested: Random Forest, XGBoost, Logistic Regression)
- Perform train-test split (e.g., 80-20)
- **Important:** Address class imbalance appropriately
- Evaluate using F1-Score, AUC-ROC, Precision, Recall
- Compare at least 2 different models

4. Business Insights

Provide a summary of: - Key factors driving customer churn - Customer segments at highest risk - Actionable recommendations for retention - Model performance and limitations

Deliverables

GitHub repository containing: - `notebooks/solution.ipynb` or `.py` script - `README.md` explaining approach, findings, and business recommendations - Any additional plots and outputs

Example Repository Structure

```
.
├── README.md
├── notebooks/
│   └── solution.ipynb
├── data/
│   └── Churn_Modelling.csv
├── outputs/
│   ├── churn_analysis.png
│   ├── feature_importance.png
│   └── confusion_matrix.png
```

Time Limit

Suggested time to complete: Maximum 2 hours

Focus on handling class imbalance effectively and providing business insights.

Evaluation Criteria

Criteria	Weightage
Data Exploration and Analysis	20%
Handling Class Imbalance	20%
Modeling and Evaluation (F1, AUC)	25%
Business Insights and Recommendations	20%
Code Quality and Readability	10%
Submission Format (Proper Repo)	5%

Notes

- You are free to use any ML library
- **Handling class imbalance is critical** - solutions ignoring this will score lower
- Bonus points for:
    - Customer segmentation analysis
    - Cost-sensitive evaluation (false negatives are costly)
    - SHAP or other explainability methods
    - Creative feature engineering