Bank Customer Churn Prediction – Detailed Project Report

1. Problem Statement

Customer churn is a major challenge in the banking sector, as acquiring new customers is significantly more expensive than retaining existing ones. The objective of this project is to build a **machine learning–based churn prediction system** that identifies customers who are likely to leave the bank. Early identification of such customers enables banks to take proactive retention measures and reduce revenue loss.

2. Dataset Overview

The dataset consists of **10,000 customer records** with demographic, financial, and behavioral attributes.

Target Variable

- **Churn = 0** → Customer stayed with the bank
- **Churn = 1** → Customer left the bank

Features Description

- Credit_score – Customer's credit score
- country – Customer's country
- gender – Customer's gender
- age – Customer's age
- tenure – Number of years with the bank
- balance – Account balance
- products_number – Number of bank products used
- credit_card – Credit card ownership (0/1)
- active_member – Activity status (0/1)
- estimated_salary – Estimated annual salary

The dataset contained **no missing values and no duplicate records**, making it suitable for direct modeling after preprocessing.

3. Exploratory Data Analysis (EDA)

Class Imbalance

The churn distribution showed clear imbalance:

- Non-churn customers: **7,963**
- Churn customers: **2,037**

This imbalance justified the use of **SMOTE (Synthetic Minority Over-sampling Technique)** during model training.

Key EDA Insights

- Churn probability increases with customer age.
- Inactive customers are more likely to churn.
- Customers using fewer banking products show higher churn rates.

4. Data Preprocessing

Steps Applied

- Dropped irrelevant column: customer_id
- Split features and target variable
- Train–test split with stratification to preserve class distribution
- Applied preprocessing using ColumnTransformer

Feature Engineering

- **Numerical Features** were scaled using **MinMaxScaler**
- **Categorical Features** (country, gender) were encoded using **OneHotEncoder**
- **SMOTE** was applied to handle class imbalance

All preprocessing and modeling steps were combined using **pipelines** to avoid data leakage.

5. Model Building

Multiple machine learning models were implemented and evaluated:

1. Logistic Regression
2. Random Forest Classifier
3. AdaBoost Classifier
4. Gradient Boosting Classifier

Each model was trained using the same preprocessing pipeline to ensure fair comparison.

6. Model Evaluation Metrics

Given the business objective of identifying churn customers, the following metrics were prioritized:

- **Recall** – Ability to identify actual churn customers
- **ROC-AUC** – Overall discrimination power of the model
- **F1-score** – Balance between precision and recall

Accuracy alone was not considered sufficient due to class imbalance.

7. Model Performance Summary

Logistic Regression

- Accuracy: ~72%
- Recall (Churn): High
- ROC-AUC: ~0.79

This model provided good interpretability but limited overall performance.

Random Forest Classifier

- Accuracy: ~84%
- ROC-AUC: ~0.86
- Improved balance between precision and recall

AdaBoost Classifier

- Accuracy: ~80%
- Recall (Churn): Relatively high
- ROC-AUC: ~0.85

Gradient Boosting Classifier

- Accuracy: ~83%
- ROC-AUC: ~0.87
- Best overall trade-off between recall, F1-score, and ROC-AUC

Among all models, **Gradient Boosting Classifier performed the best**.

8. Hyperparameter Tuning

Since Gradient Boosting showed the strongest baseline performance, **GridSearchCV** was applied to optimize:

- Number of estimators
- Learning rate
- Maximum tree depth

Tuned Model Performance

- Recall improved further
- ROC-AUC increased to **~0.87+**

This confirmed that hyperparameter tuning enhanced the model's ability to identify churn customers.

9. Feature Importance Analysis

Feature importance analysis from the Gradient Boosting model revealed:

- **Age** – Most influential feature
- **Number of products** – Strong inverse relationship with churn
- **Balance**, **active member status**, and **tenure** – Moderate influence

Age vs Churn Insight

- Churn is lowest among younger customers
- Churn increases steadily with age
- Peak churn observed in the **51–60 age group**
- Decline in churn after age 60

Products vs Churn Insight

Customers using **fewer bank products** are significantly more likely to churn.

10. ROC Curve Comparison

ROC curves were plotted for all models. The **Gradient Boosting Classifier** curve was closest to the top-left corner, indicating superior discrimination ability compared to other models.

11. Final Model Selection

Based on recall, ROC-AUC, and F1-score, the **tuned Gradient Boosting Classifier** was selected as the final model.

12. Business Recommendations

Based on model insights:

- Focus retention efforts on customers aged **51–60**
- Encourage customers to adopt **multiple banking products**
- Use churn predictions to trigger early interventions such as personalized offers, loyalty programs, and proactive customer engagement

This targeted strategy can significantly reduce churn and improve customer lifetime value.

13. Conclusion

This project demonstrates the practical application of machine learning in solving a real-world banking problem. By combining robust preprocessing, handling class imbalance, multiple model comparisons, and hyperparameter tuning, the final model provides reliable churn predictions that can directly support business decision-making and customer retention strategies.

Domain: Banking & Finance
Project Type: Supervised Machine Learning – Classification
Tools & Technologies: Python, Pandas, NumPy, Scikit-learn, Imbalanced-learn, Matplotlib, Seaborn