Project Overview

This repository contains a comprehensive study and implementation of machine learning models designed to predict customer churn for banking institutions. Customer retention is a critical strategic priority, as the cost of acquiring a new customer is significantly higher than retaining an existing one.

The project evaluates five classification algorithms—Logistic Regression, Random Forest, XGBoost, LightGBM, and Gradient Boosting—specifically focusing on addressing class imbalance (20.4% churn rate) using SMOTE, Class Weight Adjustment, and SMOTE+Tomek Links.

Key Objectives

- Recall Optimization: Prioritizing the identification of the maximum proportion of at-risk customers to minimize the financial impact of missed churners.
- Imbalance Mitigation: Evaluating the effectiveness of synthetic oversampling versus model-level weight adjustments.
- Business Impact Modeling: Translating statistical metrics into financial outcomes through a synthetic cost-benefit simulation.

Methodology

1. Exploratory Data Analysis (EDA): Identification of key predictors such as Age, Number of Products, and Active Membership status.
2. Data Preprocessing: Handling categorical variables via one-hot encoding and standardized scaling for numerical features.
3. Model Training & Validation: Implementation of 15 unique model-technique configurations with 5-fold cross-validation to ensure stability and detect overfitting.
4. Feature Importance: Analyzing which customer attributes contribute most significantly to churn behavior.

Performance Evaluation

The experimental analysis concludes that ensemble methods, particularly when paired with class-weighting techniques, provide the highest recall.

Model	**Technique**	**Recall**	**Precision**	**F1-Score**	**Accuracy**
XGBoost	Class Weight	**0.7494**	0.5161	0.6112	0.8060
LightGBM	Class Weight	0.7420	0.5119	0.6058	0.8035
Random Forest	Class Weight	0.6855	0.5753	0.6256	0.8330
Gradient Boosting	SMOTE+Tomek	0.5405	0.7074	0.6128	0.8610

Synthetic Business Cost Simulation

To assess the practical utility of the models, a business-impact scenario was simulated based on industry-standard assumptions.

Framework Assumptions

- Cost of Attrition (False Negative): $100 per customer (representing lost Lifetime Value).
- Cost of Retention (False Positive): $10 per customer (representing intervention/administrative costs).
- Optimization Goal: Minimizing the Total Business Cost (TBC).

Financial Findings

- Cost Efficiency: The Gradient Boosting with SMOTE configuration achieved the lowest simulated TBC of $15,690, representing a 24.7% reduction in costs ($5,140 in savings) compared to baseline models.

- Linear Model Improvement: Logistic Regression showed a massive 49.9% cost reduction when paired with SMOTE, indicating significant sensitivity to data balancing.

Note: This analysis is a synthetic simulation designed for model comparison. Actual financial impact may vary based on specific institutional cost structures