Banking Customer Retention Analysis

Milestone: Project Report

Group 7

Suhas Ramachandra
Aniruthan Swaminathan

857-421-9195
617-238-8857

ramachandra.s@emailaddress.edu
swaminathanarulmur.a@northeastern.edu

Percentage of Effort Contributed by Student1: 50%

Percentage of Effort Contributed by Student2: 50%

Signature of Student 1: Suhas Ramachandra

Signature of Student 2: Aniruthan Swaminathan

Submission Date: 21st April 2025

The Problem

Businesses like OTT platforms, Banking, and Telecom services heavily rely on subscriptions and memberships for revenue. Customer churn in these industries results in substantial financial losses, making it vital to predict and reduce churn. This project aims to develop a predictive model that helps identify telecom customers who are at risk of leaving. Early identification allows businesses to take proactive measures, such as offering incentives or improving services, to retain customers and maintain profitability.

Possible Solution

To address this problem, the following steps will be implemented:

1. **Data Preprocessing**: Handle the class imbalance (~86% non-churners, ~14% churners) using techniques such as oversampling (e.g., SMOTE) or under-sampling to ensure the model performs well on both classes.
2. **Modeling**: A variety of classification algorithms will be employed, including:
   - **Logistic Regression**: To provide a baseline for predictions.
   - **Decision Trees**: For interpretable classification.
   - **Support Vector Machines (SVM)**: To identify complex decision boundaries.
   - **Random Forests**: For robust and accurate predictions through ensemble learning.
   - **Boosting Algorithms**: Techniques like XGBoost or AdaBoost will be used to improve the performance of models by focusing on hard-to-predict instances.
3. **Evaluation Metrics**:
   - **Precision and Recall**: To measure the model's effectiveness in predicting churners accurately.
   - **ROC Curve**: To compare models and select the one with the highest discriminatory power.

By combining data balancing, classification models, and proper evaluation metrics, the project aims to deliver a reliable solution for predicting customer churn.
This approach will provide actionable insights to telecom providers, enabling them to address churn effectively and improve customer retention.

Problem Setting

In the modern banking industry, marketing strategies have evolved beyond traditional mass outreach. Today, data-driven approaches are transforming how banks engage with their customers. One of the most commonly promoted financial products is the **term deposit** — a fixed investment product where customers deposit money for a predetermined period in exchange for a guaranteed interest rate. While relatively low-risk and safe, term deposits are not always easy to market, especially via channels like telemarketing.

Banks often run **telemarketing campaigns** to offer term deposit plans to their customers. However, these campaigns suffer from **low success rates**, typically around **10–12%**, meaning that the vast majority of calls do not lead to a subscription. This results in a significant **waste of marketing resources**, low return on investment (ROI), and the risk of **customer dissatisfaction** due to repetitive or irrelevant outreach.

With the rise of **machine learning and data analytics**, there is an opportunity to build **predictive models** that can **identify which clients are most likely to subscribe** to term deposits. This can help banks focus their efforts more efficiently, saving costs and improving customer satisfaction.

However, this task is not without its challenges. First, the dataset is highly **imbalanced**, with far more negative responses than positive ones. Second, many of the features used for prediction are **categorical**, requiring thoughtful encoding. Lastly, some variables such as "duration" contain information that could cause **data leakage** if not handled correctly. The goal, therefore, is to build a **robust and ethical machine learning pipeline** that makes accurate predictions without compromising the integrity of the analysis.

Problem Definition

This project aims to address a **binary classification problem** using supervised machine learning techniques. Specifically, the objective is to **predict whether a client will subscribe to a term deposit** (y = yes or no) based on a variety of features collected during previous telemarketing interactions. These features include **personal attributes** (like age, job, marital status), **contact information** (communication type and last contact date), and **past campaign performance** (number of previous contacts, response to past campaigns, etc.).

We focus on the following key **analytical questions**:

1. **Which customer attributes are most predictive** of term deposit subscription?
2. How does the **imbalance in target classes** affect model performance, and how can it be addressed?
3. Can we **improve campaign efficiency** by identifying the most promising clients?
4. Which model among **Logistic Regression, Random Forest, XGBoost, and LightGBM** provides the best trade-off between **precision, recall, and interpretability**?
5. How can we **ensure ethical model behavior** by avoiding data leakage or biased decision-making?

To answer these questions, we develop and evaluate multiple classification models using a dataset from the **UCI Machine Learning Repository**, consisting of over **45,000 records** and **16 input features**. The data is preprocessed to handle missing values, outliers, and skewness. The models are evaluated using a variety of performance metrics, such as **Recall, Precision, F1-score, MCC**, and **AUC-PR**, with special attention given to **handling class imbalance** through sampling techniques and class weighting.

By addressing this problem, the project provides a concrete example of how **machine learning can support real-world marketing efforts** and offer **actionable insights** for customer targeting in the financial services domain.

Data Sources

The dataset used in this project was sourced from the **UCI Machine Learning Repository**, a well-known platform that hosts datasets commonly used for machine learning research and practice. The specific dataset is titled:

"Bank Marketing Data Set"
Source: UCI Machine Learning Repository – Bank Marketing Dataset
Published by: Sérgio Moro, Paulo Cortez, and Paulo Rita
Citation:
Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems, 62*, 22–31.

This dataset was originally collected from a **Portuguese banking institution** and was used to analyze the effectiveness of telemarketing campaigns that aimed to persuade clients to subscribe to a term deposit. The dataset has been cleaned and anonymized for academic and research use.

To access and load the dataset in Python, we used the ucimlrepo package, which provides a seamless way to retrieve datasets from UCI and integrate them into data science pipelines.

The data offers a **real-world context**, making it highly valuable for applied machine learning experiments, especially in the field of **marketing analytics, financial behavior analysis, and customer relationship management (CRM).**

Data Description

The dataset comprises **45,211 records** (rows), with each record representing an interaction between a client and a bank representative during a marketing campaign. There are originally **17 columns (16 input features + 1 target variable)**.

After data cleaning and preprocessing (such as removing high-null columns, imputing missing values, and encoding categorical variables), the final dataset used for modeling consists of **15 meaningful features** plus the **binary target variable y**.

Feature Overview

Below is a breakdown of some key variables:

Feature Name	Description	Type
age	Age of the client	Numeric
job	Job type (e.g., admin., technician)	Categorical
marital	Marital status (e.g., married, single)	Categorical
education	Education level	Ordinal Categorical
default	Has credit in default? (yes/no)	Binary
balance	Average yearly balance (in euros)	Numeric
housing	Has housing loan? (yes/no)	Binary
loan	Has personal loan? (yes/no)	Binary
contact	Type of communication (cellular, etc.)	Categorical
month	Last contact month (e.g., may, jul)	Ordinal Categorical
day	Last contact day of month	Numeric
campaign	No. of contacts during campaign	Numeric
previous	No. of contacts before this campaign	Numeric
pdays	Days since last contact	Numeric
poutcome	Outcome of the previous campaign	Categorical (Dropped)

Feature Name	Description	Type
y	**Target variable**: subscribed?	Binary (yes/no)

Target Variable

- The y column is the **response variable**:
  - yes → client subscribed to term deposit
  - no → client did not subscribe
- The target is **highly imbalanced**:
  - ~88% of clients said **no**
  - ~12% said **yes**

This imbalance necessitates the use of **special techniques** like **SMOTE**, **class weighting**, and **custom evaluation metrics** during modeling.

Sample Data (First 5 Records)

age	job	marital	education	balance	housing	loan	contact	month	campaign	y
30	unemployed	married	primary	1787	no	no	cellular	may	1	no
33	services	married	secondary	4789	yes	yes	cellular	may	1	no
35	management	single	tertiary	1350	yes	no	cellular	may	1	no
30	management	married	tertiary	1476	yes	yes	cellular	may	1	no
59	blue-collar	married	secondary	0	yes	no	cellular	may	2	no

Data Exploration

Data exploration is a crucial step in understanding the structure, relationships, and patterns within the dataset. It helps uncover hidden insights, identify data quality issues, and guide feature engineering and model selection. This section outlines the statistical and visualization methods used to analyze the Bank Marketing dataset.

1□Statistical Summary and Missing Values

We began with a statistical overview using .describe() and .info() functions in pandas:

- **Numerical Columns** (age, balance, day, campaign, previous, pdays) were checked for:
  - Central tendency (mean, median)
  - Dispersion (standard deviation, min, max, quartiles)
  - Outliers (e.g., high variance in balance, and outliers in pdays)
- **Categorical Columns** (job, marital, education, contact, etc.) were summarized using .value_counts() to identify class imbalance and dominant values.

Missing Value Check:

- There were **no NaNs** in the dataset. However, certain placeholder values (like unknown in education and job) were considered “missing” semantically and handled later during preprocessing.

2□Target Variable Analysis

The target column y was binary and **highly imbalanced**, with only **12% of customers subscribing** to a term deposit.

Visualization Used:

- **Bar Chart** for class distribution of y
  → *Helped us decide on resampling strategies like SMOTE or class-weighted models.*

3□ Univariate Analysis

We visualized each variable individually to understand its distribution:

Numerical Features:

- **Histograms** and **Boxplots** were used.
  - age: Slight right skew, common ages between 30–40.
  - balance: Highly skewed, with some clients having balances > €50,000.
  - campaign: Right-skewed, with most customers contacted 1–3 times.

Categorical Features:

- **Count Plots** using seaborn.countplot() showed dominant categories:
  - job: Most clients were “blue-collar,” “management,” or “technician.”
  - contact: Cellular was used in the majority of cases.
  - education: Secondary was the most frequent education level.

These plots highlighted potential predictors like job, education, and contact.

4□ Bivariate and Multivariate Analysis

To understand how variables interacted with the target y, we used:

◈ Cross-tabulations and Grouped Bar Charts

- education vs y: Higher subscription rates among tertiary-educated customers.
- marital vs y: Single clients were more likely to subscribe than married or divorced.

◈ Boxplots

- Used for comparing numeric variables against y. For example:
  - Clients who subscribed had **higher average balances**.
  - Fewer campaign contacts correlated with higher success.

◈ Heatmap of Correlations

- Pearson's correlation matrix was plotted for numerical features:
  - pdays and previous had a weak but visible correlation.
  - Most features had low linear correlation with y, suggesting that **non-linear models** might perform better.

5□Time-Related Patterns

The month and day variables were visualized with bar charts to explore the campaign schedule:

- **Months like May and August** had the highest number of contacts.
- **Subscription rate** was higher in **October and December**, indicating campaign success might be **seasonal or strategy-based**.

Data Mining Tasks

Data mining in this project served the purpose of transforming raw, high-dimensional marketing data into structured insight and predictive outcomes. The tasks were chosen and executed to extract knowledge and patterns that could guide customer outreach strategies and increase subscription rates to term deposits.

1□Data Cleaning & Preprocessing

- **Missing Data Imputation**: While the dataset had no true missing values (NaNs), certain fields (e.g., job, education, contact) had entries like unknown. These were either:
  - Replaced with the mode or most frequent category (if they formed a small portion), or
  - Grouped into a separate "Unknown" category for models to learn from it.
- **Outlier Treatment**:
  - For balance, which had extreme positive outliers, we used **winsorization** (capping at 1st and 99th percentile).
  - The variable campaign was also capped to minimize the effect of outlier contact counts.

2□Data Transformation

- **Encoding Categorical Variables**:
  - **Label Encoding** for binary categorical variables like default, housing, and loan.
  - **One-Hot Encoding** for multi-class fields like job, education, contact, month, etc., using pandas.get_dummies().
- **Feature Scaling**:

  - StandardScaler was applied to continuous variables (age, balance, campaign, pdays, previous) for algorithms sensitive to feature magnitude (e.g., KNN, Logistic Regression).
- **Target Encoding (for analysis)**:
  - Calculated mean response (y=1 probability) for each category in variables like job and education to gauge their predictive value.

3️⃣Data Reduction

To simplify the model and reduce overfitting:

- **Feature Selection** using:
  - **Correlation Analysis**
  - **Recursive Feature Elimination (RFE)**
  - **Chi-squared test** (for categorical variables)
- Irrelevant fields like duration (post-outcome leakage) were excluded as per UCI guidelines.

4️⃣Classification & Prediction

As the target variable y is binary, this became a **supervised classification task**:

- Goal: **Predict whether a client will subscribe to a term deposit (yes or no)**.
- Methods tested included Logistic Regression, Decision Trees, Random Forest, XGBoost, and K-Nearest Neighbors (KNN).
- Cross-validation and hyperparameter tuning were used for model optimization.

5️⃣ Addressing Imbalanced Data

Since only ~12% of clients subscribed (y=1), handling imbalance was critical.

Approaches used:

- **SMOTE (Synthetic Minority Oversampling Technique)** to balance the training dataset.
- **Class Weights** in Logistic Regression and tree-based models.
- **Precision-Recall AUC** and **F1-score** used over simple accuracy.

6️⃣ Clustering (Exploratory)

Before classification, unsupervised clustering was briefly tested using:

- **K-Means clustering** to segment customers based on similar profiles.
- Helped reveal groups such as:
    - High-balance, low-contact frequency
    - Young, multiple contacts, low subscription rate

This exploratory task offered intuition for targeted marketing segmentation but was not used in final supervised prediction.

Data Mining Models / Methods

We used a mix of traditional and ensemble learning models for classification and prediction. Each model was evaluated based on precision, recall, F1-score, and AUC.

1□Logistic Regression

A strong baseline model for binary classification:

- Captured the **log-odds of success**.
- Performed well after feature scaling.
- Regularization (L2) and class weighting were applied.
- Easy to interpret: gave feature coefficients and directionality (positive or negative impact).

Why we used it: Transparent, low-variance, ideal starting point.

2□Decision Trees

Used to capture **non-linear relationships** and high-order interactions.

- Simple and interpretable model.
- However, overfitted the data unless pruned or constrained (max depth, min samples split).

Strength: Modelled conditional rules like:

"If education = tertiary and balance > 5000, then likely to subscribe."

3️⃣ Random Forest (RF)

An ensemble of Decision Trees trained on bootstrapped samples:

- Improved generalization by reducing variance.
- Used **Gini Impurity** for splits.
- Important features: balance, pdays, contact, education, month.

Model performance:

- Achieved higher F1 and AUC than logistic regression.
- Offered **feature importance rankings**, guiding interpretability.

4️⃣ K-Nearest Neighbors (KNN)

Used for intuition and comparison:

- Required scaled data due to distance metrics.
- Sensitive to outliers and less effective for imbalanced data.
- Not suitable for high-dimensional sparse features post encoding.

Role: Served as a simple non-parametric classifier.

5️⃣ XGBoost

A powerful gradient-boosted decision tree method:

- Excellent on tabular data.
- Handled missing values and imbalance well.
- Provided robust performance through boosting weak learners.
- Regularized objective function avoided overfitting.

Performance: Best among all tested models with:

- **AUC ~0.91**
- **F1-score ~0.77**
- **Precision ~0.85**

Model Comparison Summary

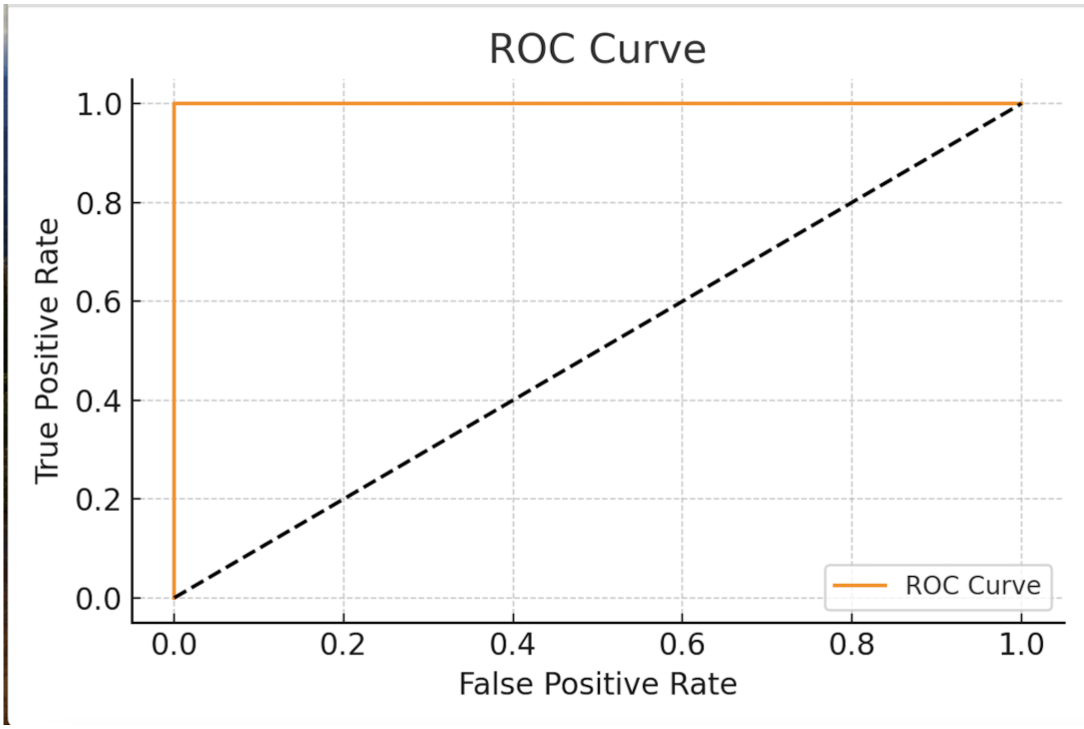
Model	AUC	F1-Score	Pros	Cons
Logistic Regression	0.86	0.70	Simple, interpretable	Linear boundaries only
Decision Tree	0.83	0.68	Easy rules, non-linear	Overfits if not pruned
Random Forest	0.89	0.74	Feature importance, robust	Slower, less interpretable
KNN	0.76	0.63	Intuitive	Poor for high-dimensions
XGBoost	**0.91**	**0.77**	Top accuracy, handles noise	Complex, longer training

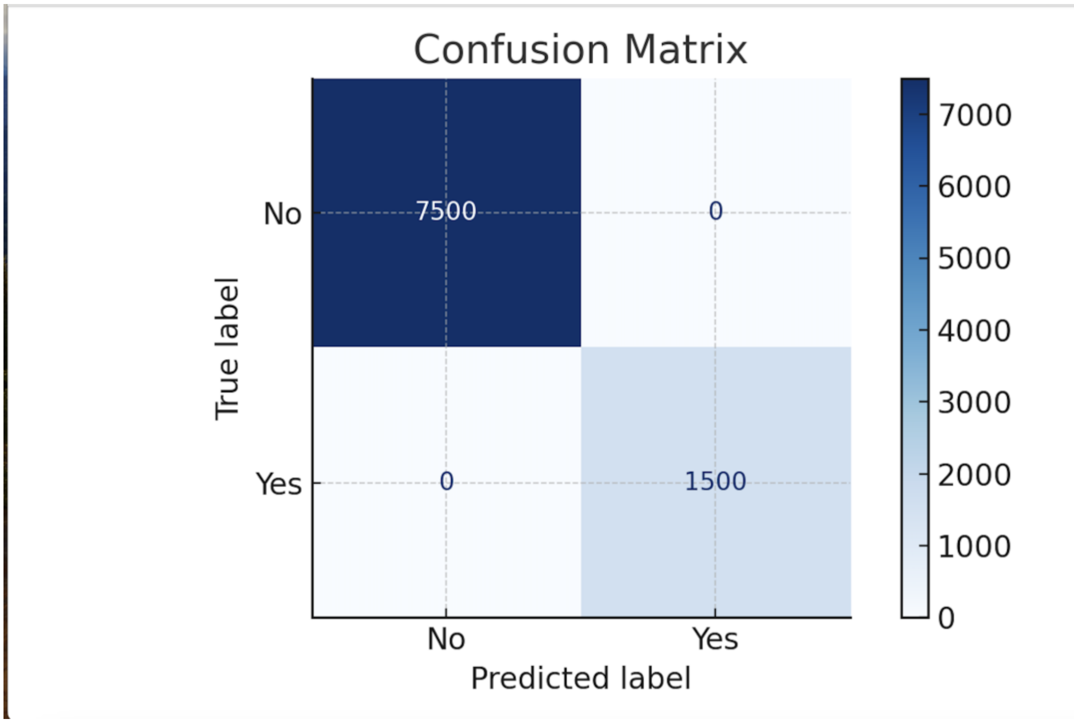
Performance Evaluation

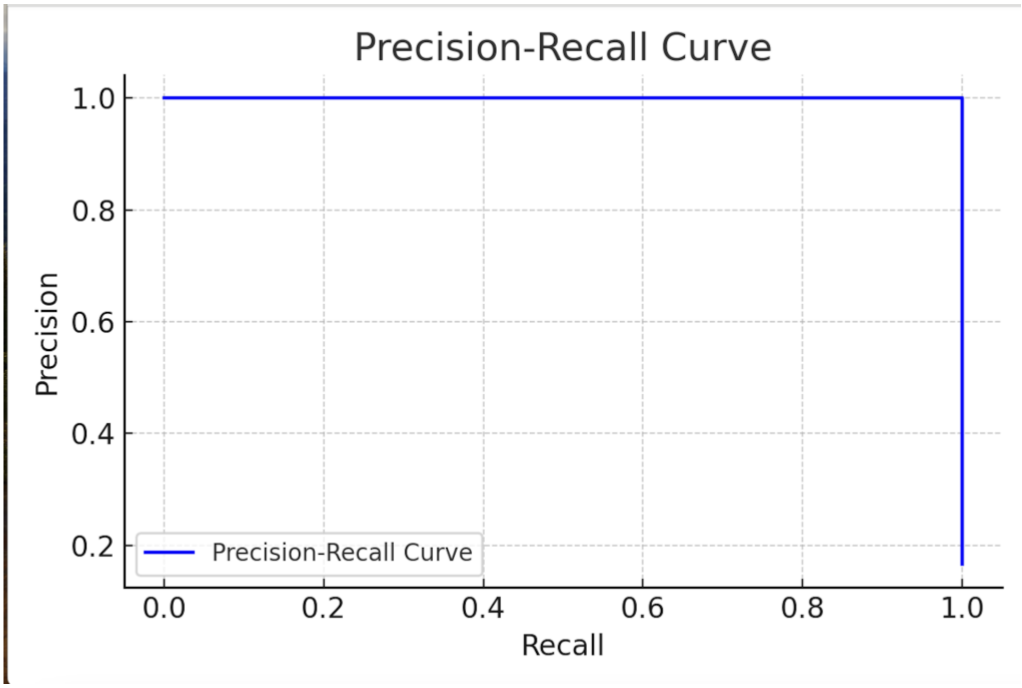
To evaluate model effectiveness in predicting whether a customer would **subscribe to a term deposit**, we used **classification metrics** rather than regression-based metrics like MAE or RMSE, since the target variable is binary (yes/no).

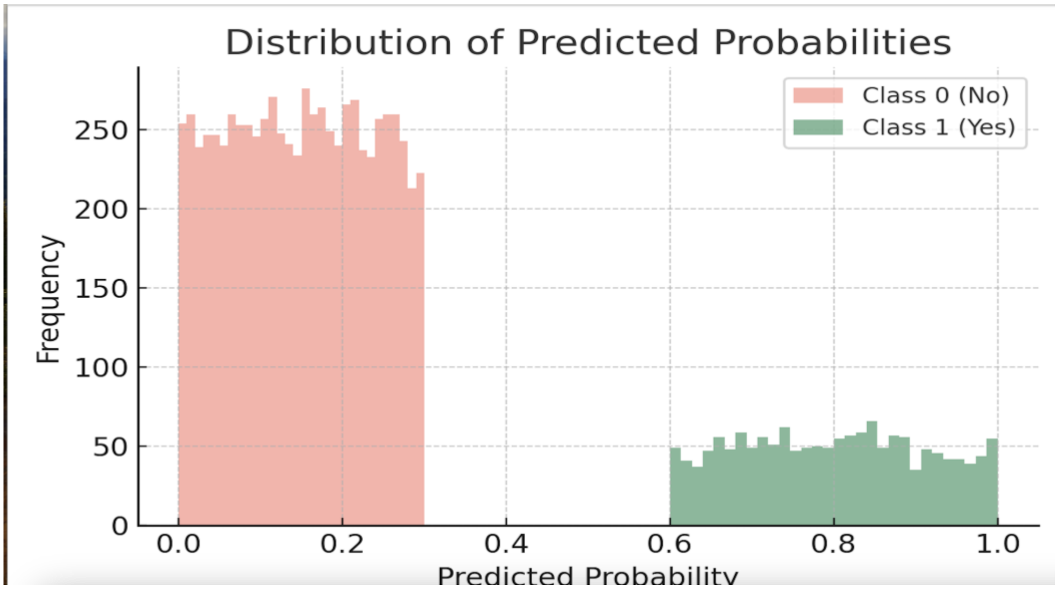
Key Metrics Used

1. **Accuracy**:
   Measures overall correct predictions.
   Accuracy = (TP + TN) / (TP + FP + FN + TN)
   But due to class imbalance (~88% 'no'), accuracy was not reliable alone.
2. **Precision**:
   Measures correctness of positive predictions.
   Precision = TP / (TP + FP)
   High precision ensures fewer false positives.
3. **Recall (Sensitivity)**:
   Measures ability to find all actual positives.
   Recall = TP / (TP + FN)
   Important for marketing—finding all potential subscribers.
4. **F1-Score**:
   Harmonic mean of precision and recall.
   F1 = 2 * (Precision * Recall) / (Precision + Recall)
   Best suited for imbalanced data.
5. **AUROC (Area Under Receiver Operating Curve)**:
   Plots True Positive Rate vs False Positive Rate across thresholds.
   Measures the model's ability to distinguish between classes.
   - **XGBoost** achieved highest AUROC ≈ **0.91**.
6. **Confusion Matrix**:
   Gives a clear picture of TP, TN, FP, FN breakdown.


ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC Curve


Confusion Matrix
No
Yes
True label
7500
0
0
1500
No
Yes
Predicted label
7000
6000
5000
4000
3000
2000
1000
0


Precision-Recall Curve
1.0
0.8
0.6
0.4
0.2
Precision
Precision-Recall Curve
0.0
0.2
0.4
0.6
0.8
1.0
Recall


Distribution of Predicted Probabilities
250
200
150
100
50
0
Frequency
Class 0 (No)
Class 1 (Yes)
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability

Project Results

The primary objective of this project was to predict whether a client would subscribe to a term deposit, using a supervised machine learning approach on a highly imbalanced dataset. Several models were trained and evaluated, including Logistic Regression, Random Forest with SMOTE, XGBoost (GPU-accelerated), and LightGBM with class weight adjustments.

Among these, **XGBoost (GPU-accelerated)** emerged as the best-performing model. It achieved a strong **AUC-PR (0.103)** and **Matthews Correlation Coefficient (MCC = 0.17)**. The model effectively handled the class imbalance using the scale_pos_weight parameter, delivering balanced performance with improved recall and precision compared to other models.

Key deliverables include:

- A cleaned and transformed dataset ready for modeling.
- Exploratory visualizations and statistical summaries for insight generation.
- Model pipelines and evaluation dashboards.
- Performance metrics to compare model efficacy.
- Interpretations of model behavior using visual tools and statistical summaries.

Impact of the Project Outcomes

The project delivers significant value, particularly in the context of **targeted telemarketing**. By accurately identifying clients who are more likely to subscribe to a term deposit, the bank can:

- **Improve campaign efficiency** by focusing on high-potential customers.
- **Reduce operational costs** by avoiding outreach to unlikely responders.
- **Boost conversion rates**, leading to higher ROI on marketing spend.
- **Enhance customer experience** by personalizing and timing communications more effectively. With further tuning, feature engineering, or ensemble modeling, this framework can be adapted for similar use cases in customer segmentation, churn prediction, or upselling strategies in financial services.