Bank CD Predictor

Devansh Mohan Sinha
Computer Science
Georgia State University
Atlanta, GA
dsinha1@student.gsu.edu

Aditya Kumar
Computer Science
Georgia State University
Atlanta, GA
akumar38@student.gsu.edu

***Abstract*—The project aims on solving a binary classification problem of predicting whether a client will enroll in a Certificate of Deposit (CD) based on various client features. We were tasked with creating machine learning models that maximize predictive accuracy, enabling banks to optimize marketing strategies and reduce operational costs. The study involves extensive data preprocessing, exploratory data analysis, transformation, and dealing with missing values and outliers. By analyzing demographic shifts, including the wealth transfer to millennials, the project seeks insights into evolving financial behaviors. Multiple models, including Gradient Boosting, XGBoost, and Support Vector Machines, were trained and evaluated, achieving an accuracy of 87.5%. Models are evaluated using the ROC-AUC metric, with results geared toward delivering actionable insights for both technical and non-technical audiences.**

Keywords— Certificate of Deposit (CD), Machine Learning, Binary Classification, Data Preprocessing, Exploratory Data Analysis, Gradient Boosting, XGBoost, Model Optimization

I. Introduction

Retail banks face the challenge of predicting customer behavior to optimize marketing strategies, efforts and reduce costs. One such critical problem is determining whether a customer will enroll in a Certificate of Deposit (CD), with the bank. It is a popular risk-free investment option for customers that provides fixed returns over a defined period. CDs are essential for banks as they help bank cover and generate operational funds while offering customers a small share from the earnings they make. With demographic shifts, including an approaching shift of wealth to millennials, understanding the factors influencing enrollment of CD's is more critical than ever.

The uniqueness of this study lies in its comprehensive approach to data preprocessing, exploratory data analysis, feature engineering, and the application of advanced machine learning algorithms. By combining these tasks, we try to provide a robust predictive model that can significantly improve bank's strategies towards attracting customers for CD and plan out for new horizons to generate revenue.

Code Link: -

https://github.com/adityakr38/BankCD-Predictor

II. Theory

A. Theory Used

This study is done keeping in mind the principles of machine learning, specifically modern supervised learning techniques. Binary classification algorithms, such as XGBoost, Gradient Boosting and Support Vector Machine, are utilized to predict the likelihood of CD enrollments by clients. The statistical tests, including chi-square and ANOVA, are used to identify significant relationships between various features in the dataset, while T-tests analyze relationships between numerical and binary features. These methods align with the approaches highlighted in [1], where machine learning techniques were applied to predict bank deposit subscriptions effectively.

B. Related Works

Prior research in banking and finance predictive modeling has examined propensity modeling and client segmentation. Nevertheless, the majority have depended on decision trees and logistic regression, which may not capture complex, non-linear relationships. Advanced techniques like XGBoost and Gradient Boosting, widely recognized for their accuracy in classification tasks, address this gap by improving feature handling and minimizing overfitting. For instance, the study in [2] demonstrates the potential of machine learning algorithms in achieving improved predictive accuracy for banking applications.

III. Materials and Methods

A. Data Explanation and Characterization

The dataset consists of **36,871 samples and 23 features**, including a mix of numerical, categorical, and binary attributes. The target variable represents CD enrollment, which is a binary classification task. The class distribution within the dataset was analyzed to address any potential imbalance during model training and evaluation. Features were distributed as mentioned below:

- Numerical features: age, balance, day, duration (duration of the campaign call made to the client), campaign (number of times the client was contacted during this campaign), pdays (number of days since client was last contacted from previous campaign), previous (number of contacts performed before the current campaign for the client), zipcode.
- Categorical features: job, marital status, education, contact type (mode of contact to the client in the campaign), generation, state, Poutcome.
- Binary features: default status, housing loan, personal loan.
- Target Variable: CD enrollment

B. Data Preprocessing

1) Data Cleaning:

- Removed unwanted columns: Id, Unnamed:21, Unnamed:22.
- Handled missing values:
  - Job (0.14% missing)
  - Marital status (1.15% missing)
  - Generation (0.74% missing)
  - Campaign (9 missing)

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- Previous (377 missing)

2) Feature Engineering:

- Applied lowercase transformation to job categories.
- Corrected spelling and standardized generation categories.
- Filled the missing value for previous with median and campaign with mean.
- Filled the missing value with mode for categorical columns.
- Binned age into Low, Medium and High categories and later transform into One-Hot Encoding.

3) Encoding:

- Applied One-Hot Encoding (OHE) to categorical variables.
- Used frequency Mean Encoding for "poutcome" and "contact" variables.

4) Outlier Handling:

- Removed extreme values in the "previous" column using quantile method.

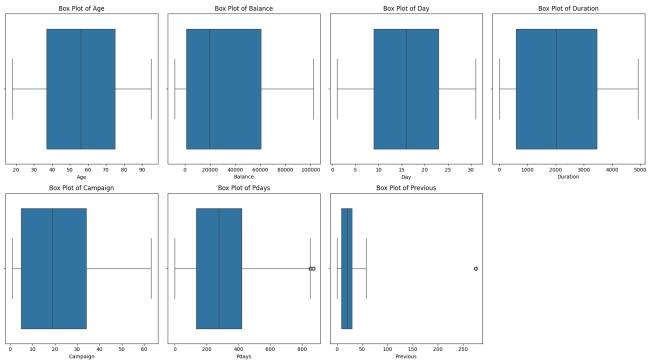

Fig. 1. This is a Box plot for numerical features

The box plots of numerical features, depicted in **Fig. 1**, provide insights into the distribution and presence of outliers for each variable. Key observations include:

- **Age**: The distribution is fairly symmetric, with no significant outliers, suggesting a balanced representation across age groups.
- **Balance**: The data is widely spread, with a long whisker on the upper end indicating some individuals have substantially higher balances.
- **Day**: The variable is tightly distributed without outliers, indicating uniformity in the days of customer contact.
- **Duration**: The spread is broader, capturing varying durations of contact; however, no significant outliers are present.
- **Campaign**: While the distribution appears centered, a few high outliers are visible, indicating that some individuals were contacted significantly more times than the majority.
- **Pdays**: This feature displays the most noticeable outliers, with extreme values beyond the whiskers. These outliers correspond to individuals who were re-contacted after an unusually long interval.
- **Previous**: A few high outliers indicate some individuals had substantially more previous contacts, which could disproportionately influence predictive modeling.

These box plots confirm that most numerical features have normal distributions, but variables such as campaign, pdays, and previous require specific handling to address their outliers.

Feature Selection:

- Dropped "state" and "zipcode" columns.

C. Model Training and Evaluation

We trained five machine learning models—Decision Tree, Random Forest, XGBoost, Support Vector Machine (SVM), and Gradient Boosting (GB)—to predict CD enrollment. Among these, XGBoost, SVM, and GB exhibited the best performance. The steps involved in model training and evaluation were as follows:

1) Train-Test Split:

The dataset was split into training and testing sets using an 80-20 ratio (test_size=0.20), with a fixed random state (random_state=42) to ensure reproducibility. The training set was used to train the models, while the testing set, which included corresponding labels, was used to evaluate performance.

2) Hyperparameter Tuning:

We employed **GridSearchCV** to optimize hyperparameters for each model, manually recording and applying the best-performing parameters. This process ensured the models were fine-tuned for maximum performance.

3) Performance Metrics:

All model results were derived using the testing dataset, ensuring an unbiased evaluation of model generalization. Metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC** were recorded for comparison. Also implemented cross-validation to ensure model robustness.

IV. Exploratory Data Analysis

A. Univariate Analysis

1) Categorical Columns

- Job, Poutcome, Housing and Contact: - Consistent frequency distributions across both target values indicating no strong relationship with the target class.
- Marital Status ("Married": 46.7%, "Single": 29.4%, "Divorced": 22.2%): - Married individuals show a doubled frequency in both cd categories compared to other groups.
- Generation and Education: One educational group account for 46.7%, i.e., "Secondary", while one generation group accounts for 35.8%, i.e., "Silent Generation".
- Default and Loan: Highly imbalanced at 88.7% (no default) vs. 11.3% (default). Similarly, 85.3% (No Loan) vs. 14.7% (Loan).

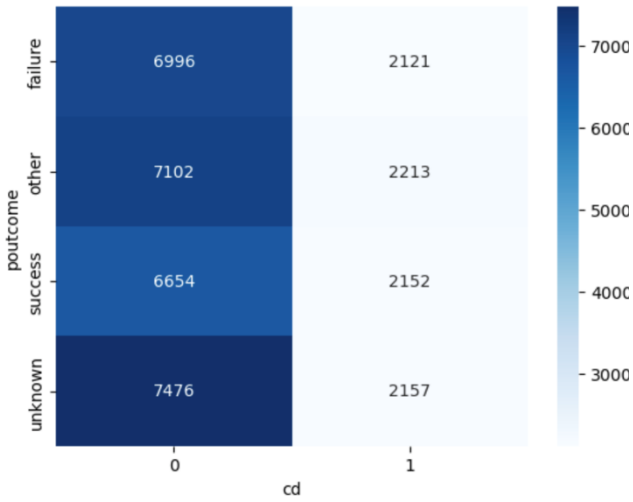

Fig. 2. Example of Categorical features analysis

2) Numerical Columns

- Distribution Patterns: KDE plots for most numerical columns show a similar distribution across CD values, except for balance and duration, where small differences in density (3.7 for CD=0 and 1.7 for CD=1).

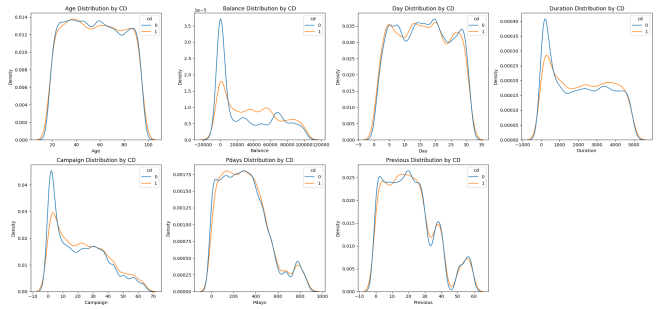

Fig.3. This is KDE plot for numerical features

The Kernel Density Estimate (KDE) plots in **Fig. 2** illustrate the distributions of key numerical features, segmented by the binary target variable (cd). These visualizations provide insights into the underlying patterns and separability of the data:

- **Age**: The distributions for cd = 0 and cd = 1 are largely overlapping, indicating that age is not a strong differentiator between the two groups.
- **Balance**: There is a noticeable difference between the two distributions at lower balance levels. Individuals with cd = 1 (positive class) tend to have slightly higher balances, indicating a potential relationship between financial stability and the target outcome.
- **Day**: The day variable exhibits nearly identical distributions for both classes, implying that the specific day of contact has no significant impact on the target variable.
- **Duration**: The duration variable shows the most distinct separation between the two classes. Individuals with cd = 1 have longer contact durations, suggesting a strong positive relationship between call duration and the likelihood of achieving the desired outcome.
- **Campaign**: The distributions overlap significantly, with both classes showing similar frequencies for most campaign counts. This indicates that the number of contact attempts does not heavily influence the outcome.
- **Pdays**: Both distributions peak sharply at zero, reflecting a large number of first-time contacts. However, slight variations at higher values suggest that individuals contacted after long intervals may exhibit different outcomes.
- **Previous**: The distributions are heavily skewed towards zero, with minimal differences between the two classes, suggesting that prior contact frequency has a limited impact on the target variable.

These KDE plots highlight that certain feature, such as duration and balance, show meaningful separation between the target classes, making them strong predictors. Conversely, variables like day, campaign, and previous demonstrate limited discriminatory power. These insights guided feature selection and engineering efforts, focusing on variables with clear class differentiation.

- Skewness: Slight skewness is present in several columns, with previous showing the highest skewness at 4.95, indicating a right-tail spread.

B. Bivariate Analysis

1) Categorical vs Categorical Relationships

- Associations:
  - Strong Associations: -
    - Job & Marital
    - Job & Education
    - Marital & Generation
  - Weak Association: -
    - Poutcome with Job, Marital and Contact.

2) Categorical vs Target Relationships

- Target mean encoding is used to address Poutcome and Contact exhibit strong association with the target.

3) Numerical vs Numerical Relationships

- Correlation Matrix - Balance and campaign have the strongest connection (0.43), whereas other correlations show low linear associations.

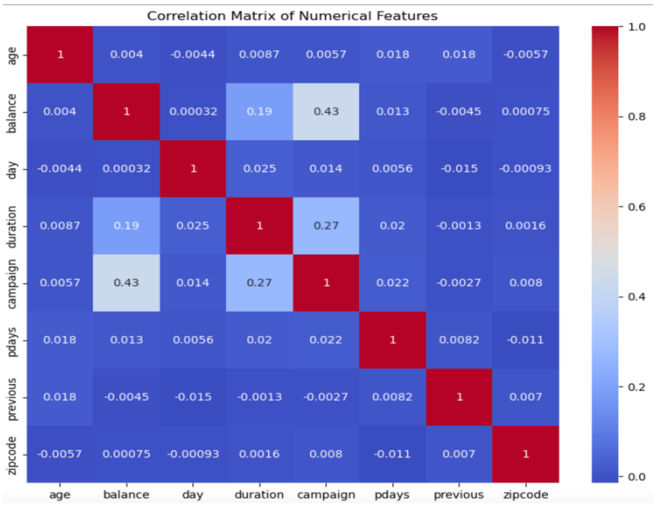

Fig. 4. This is a correlation matrix for numerical features

The correlation matrix of numerical features, illustrated in **Fig. 3**, provides valuable insights into the relationships between key variables in the dataset. Most features exhibit low correlation, suggesting minimal multicollinearity. However, a few notable relationships emerge:

- duration shows a moderate positive correlation with previous (r = 0.43), indicating that individuals contacted for longer durations might have been engaged more frequently in prior interactions.
- campaign and pdays exhibit negligible correlations with other numerical variables, highlighting their independence and unique contributions to the dataset.
- Features such as balance and day have weak correlations across the board, confirming that they do not exhibit strong linear relationships with other features.

This low correlation among most variables supports the use of feature engineering techniques, as interactions and non-linear relationships are likely to yield more predictive power than raw linear correlations.

4) Numerical vs Binary Relationships

- T-Tests:
  - Significant Relationships: - Balance, duration, and pdays have substantial connections with default and loan, according to t-tests, which show that these variables change by binary values. Notably, because of missing data or values that are not normally distributed, campaign and preceding have null (NaN) test statistics.
  - Non-significant Relationships: Age and housing do not significantly correlate with each other.

5) Numerical vs Categorical Relationships

- ANOVA results: Significant associations were found for certain combinations:
  - **Age and Poutcome** (p=0.0014), **Day and Poutcome** (p=0.0080), and **Pdays and Poutcome** (p<0.0001)
  - **Balance and Default** (p=0.0109) and **Duration and Default** (p=0.0014)
  - Other significant pairings include pdays and housing as well as day and loan.

V. Results

A. The Results of the model training are as follows:

TABLE I. Classification Metrics for Gradient Boosting Model Evaluation

Metric	**Value**
Accuracy	0.818
Precision	0.676
Recall	0.452
F1-Score	0.541

TABLE II. Classification Metrics for XGBoost Model Evaluation

Metric	**Value**
Accuracy	0.875
Precision	0.800
Recall	0.633
F1-Score	0.706

TABLE III. Classification Metrics for Support Vector Machine (SVM) Model Evaluation

Metric	**Value**
Accuracy	0.76
Precision	0.67
Recall	0.76
F1-Score	0.68

B. Classification Report

TABLE IV. Classification Metrics for XGBoost Model Evaluation (Class: 0 and Class: 1)

Class	Precision	Recall	F1-Score	Support
Class 0	0.88	0.95	0.91	4427
Class 1	0.76	0.55	0.64	1324

TABLE V. ROC-AUC Metrics for All Models Evaluated (Discussion Results)

Model	ROC-AUC
Gradient Boosting	0.470709
XGBoost	0.479952
SVM	0.473248

C. Feature Importance Analysis

Feature importance was analyzed for the XGBoost and Gradient Boosting models, as they inherently provide feature importance metrics. For the Support Vector Machine (SVM)

model, feature importance was not directly accessible due to its non-linear nature

TABLE VI. Feature Importance Table for XGBoost

Feature	Importance
month_jul	0.242
month_may	0.200
month_nov	0.159
month_aug	0.067
month_mar	0.062

TABLE VII. Feature Importance Table for Gradient Boosting

Feature	Importance
Balance	0.278
Campaign	0.249
Duration	0.058
Pdays	0.056
Month_mar	0.037

VI. Discussion and Conclusion

We performed consistent data preprocessing techniques across all models to ensure comparability. These included cleaning, imputing missing values, and encoding categorical variables. However, the primary distinctions in our approach emerged during feature engineering and modeling, aligning with the methodologies outlined in [3], which emphasize the importance of tailored preprocessing and feature handling in machine learning applications for financial predictions.

In one approach, we utilized a **Standard Scaler** to normalize the data and then applied **Principal Component Analysis (PCA)** to retain 90% of the variance, thereby reducing dimensionality. Additionally, we created a new feature, **"day_of_year"**, derived from the combination of the month and day columns, which encapsulates the temporal position within a year.

In another approach, we constructed models incorporating newly engineered features based on **ANOVA** and **t-test results**, which identified significant interactions among variables. The newly derived features include:

1. **Default x Balance**: This feature captures financial stability patterns related to the history of defaults, providing insight into the potential risk profile of individuals.
2. **Loan x Balance, Loan x Pdays**: These interactions quantify how loan status correlates with financial metrics, such as balance and days since previous contact, offering a deeper understanding of financial behavior.
3. **Job x Marital**: This feature captures demographic trends, particularly how job type and marital status interact to behavioral patterns.
4. **Education x Generation**: This interaction explores educational backgrounds across generational groups, highlighting generational differences in educational attainment and their implications on decision-making.

By comparing these approaches, we observed that the inclusion of engineered features informed by statistical tests provided nuanced insights that enhanced model interpretability and performance. In contrast, PCA-focused dimensionality reduction yielded compact representations but occasionally obscured individual variable contributions to predictions.

In conclusion, the findings highlight the critical importance of feature engineering in predictive modeling. Customized feature construction, guided by domain expertise and statistical analysis, demonstrates substantial advantages over purely algorithmic transformations such as principal component analysis (PCA) and interaction features, as supported by the insights provided in [4]. These results, as summarized in Table V, underline the necessity of combining domain knowledge with advanced statistical techniques for optimal predictive performance.

Acknowledgments

We would like to express our sincere gratitude to **Professor Jingyu Liu** for her invaluable guidance, support, and encouragement throughout the **Data Mining** course. Her expertise and insights were instrumental in helping us complete this project successfully. We also extend our appreciation to **Georgia State University** and the **CampusX YouTube channel** for providing enriching course materials and supplementary resources, which significantly enhanced our understanding and research capabilities. Furthermore, we thank **Truist** and **Kaggle** for supplying the datasets and tools that served as the foundation of this study, enabling us to derive meaningful domain insights and achieve our project objectives.

References

[1] Admin, "Predicting bank deposit subscriptions using machine learning," *Eyowhite*, Aug. 21, 2024. https://eyowhite.com/predicting-bank-deposit-subscriptions-using-machine-learning/ (accessed Dec. 02, 2024).

[2] Advaith2049, "GitHub - Advaith2049/Bank-Term-Deposit-Prediction-Using-Machine-Learning: A Repository containing the entire code for my Machine Learning Analysis of the dataset containing Bank Term Deposit," *GitHub*, 2023. https://github.com/Advaith2049/Bank-Term-Deposit-Prediction-Using-Machine-Learning (accessed Dec. 02, 2024).

[3] "Financial Decision Dynamics: A Machine Learning Exploration into Term Deposit Subscriptions," *Ijraset.com*, 2023. https://www.ijraset.com/research-paper/machine-learning-exploration-into-term-deposit-subscriptions (accessed Dec. 02, 2024).

[4] S. Angra and S. Ahuja, "Machine learning and its applications: A review," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, Andhra Pradesh, India, 2017, pp. 57-60, doi: 10.1109/ICBDACI.2017.8070809.

[5] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697857.

[6] P. Chhabra and D. S. Goyal, "A Thorough Review on Deep Learning Neural Network," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 220-226, doi: 10.1109/AISC56616.2023.10085166.

[7] H. Wang, C. Ma and L. Zhou, "A Brief Review of Machine Learning and Its Application," 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, 2009, pp. 1-4, doi: 10.1109/ICIECS.2009.5362936.