Optimizing N/LAB Enterprises' Customer Outreach for "N/LAB Platinum Deposit"

BySamiksha Kamath
20703562

A.Summarization

The dataset cleaning process included removing columns with more than 50% missing values or over 50% 'unknown' entries, such as the "poutcome" column. Rows with missing target values ('y') were also dropped. Outliers were identified and removed from numerical columns using the IQR method. Irrelevant columns with no correlation or predictive value were discarded, though no specific correlation threshold is mentioned in the extract. Features were scaled using standard scaling, and new interaction features like balance_duration_ratio and call_efficiency were created for deeper analysis.Feature previous, which exhibited minimal variation or constant values, were removed to reduce noise and improve model performance. Outliers beyond three standard deviations in key numerical features were identified and excluded to prevent skewed results, further enhancing the quality of the dataset.The heatmap in Figure 1 shows weak correlations between most numerical features and the target (y), with duration having the strongest positive correlation (0.47), highlighting its importance. Other features like balance (0.09) and age (0.04) demonstrate minimal impact on subscription likelihood.

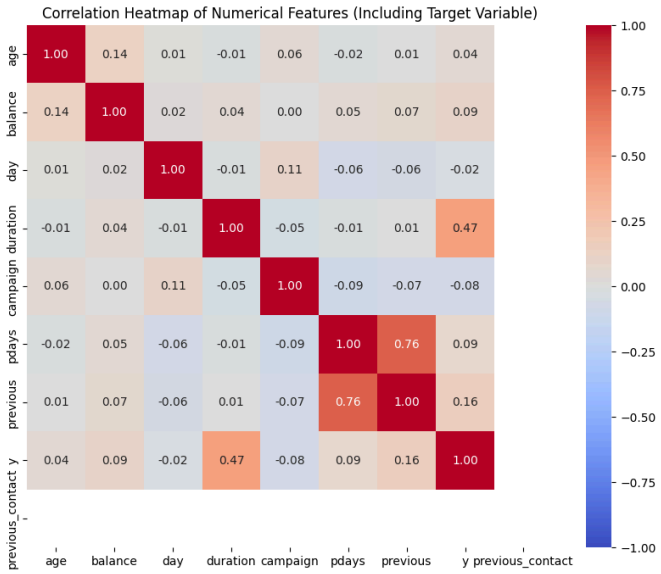


Figure 1.Correlation heatmaps

Boxplots (Figures 6 and 7) demonstrated that higher account balances and longer call durations significantly increase subscription likelihood. Individuals with higher balances may have greater financial stability, making them more receptive to investment products.Figure 6 Boxplot shows

that longer call durations are strongly associated with successful subscriptions (y=1), while unsuccessful calls (y=0) generally have shorter durations.Figure 7 Boxplot indicates that individuals with higher account balances tend to subscribe more frequently, as the median balance for y=1 is higher compared to y=0.

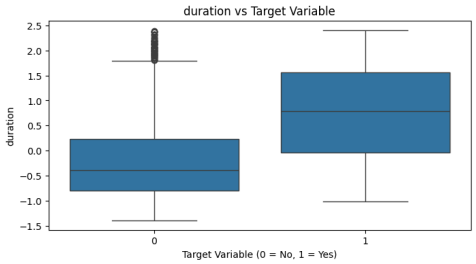


Figure 2.Call Duration Distribution By Outcome

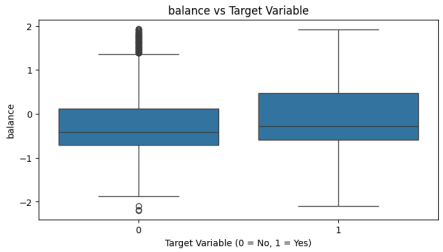


Figure 3.Account Balance Distribution by Outcome

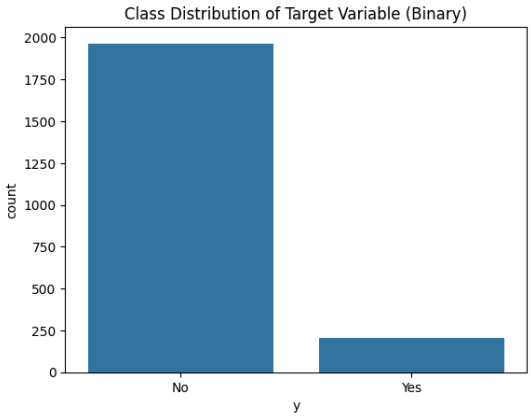


Figure 4.Class Distribution of Target Class

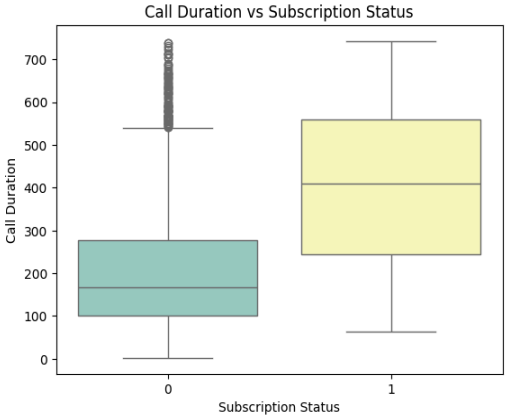


Figure 5.Call Duration Vs Subscription

The class distribution plot (Figure 4) revealed that the dataset is highly imbalanced, with ~88% labeled as "No" and only ~12% as "Yes." This imbalance indicates a challenge in identifying the minority class, reinforcing the need for techniques like oversampling or cost-sensitive modeling.The chart highlights a severe class imbalance, with most individuals not subscribing, emphasizing the need for a model that effectively identifies the small subset likely to subscribe.Call duration (Figure 5) showed a strong relationship with subscription success, as longer calls were more likely to convert; the median duration for "Yes" was approximately 2.5 times that of "No." This indicates that engaging customers for longer conversations can substantially improve outcomes.The age-specific subscription analysis (Figure 6) inferred that older individuals, particularly those over 70, have a significantly higher subscription rate (~30%), suggesting that age is a key demographic factor for targeting.

Figure 7. The scatterplot shows that successful subscriptions (1) are associated with longer call durations, whereas shorter calls are linked to unsuccessful outcomes (0). These insights emphasize call efficiency and call duration as key factors for predicting subscription likelihood.

Median call duration for "Yes" is significantly higher than for "No" (~2.5x longer). Call duration remains the strongest predictor of subscription likelihood. The scatterplot indicates that successful subscriptions (y=1) are concentrated at higher call durations, while unsuccessful ones (y=0) dominate shorter durations, with this trend consistent across campaign numbers.

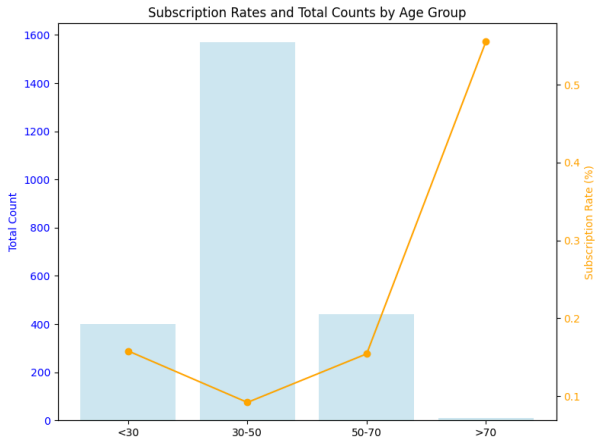


Figure 6.Subscription Rate by age

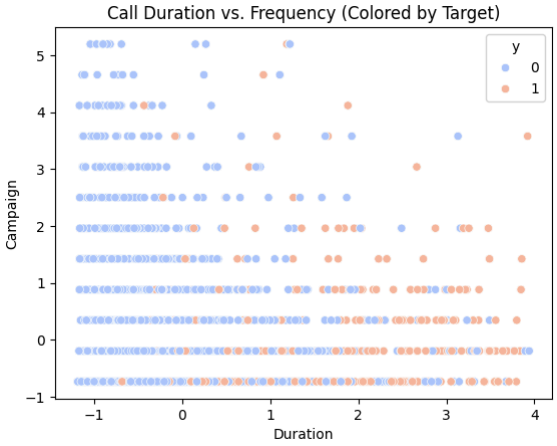


Figure 7.Subscription Likelihood with Duration of Call

B.EXPLORATION

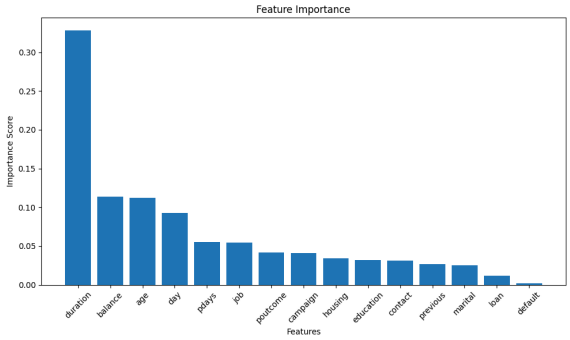


Figure 8.Feature Importance

This feature importance plot figure 8. ranks the significance of each feature in predicting the target variable (y). The top predictors include duration, balance, and age, emphasizing their strong impact on the model's performance.The analysis of the decision tree emphasizes call duration as the most influential factor in predicting subscription outcomes (y). At the root node, shorter call durations (<= 0.62) strongly predict non-subscriptions, while longer durations significantly improve the likelihood of a positive response. As the tree branches, call_efficiency emerges as another critical factor, refining sub-populations when combined with duration. For example, calls with high efficiency (<= 4.68) and very long durations (> 3.32) show a strong likelihood of success (Node 14, 29 successful out of 34 samples). Conversely, very short call durations (<= -0.65) predominantly lead to non-subscriptions (Node 3, 1476 out of 1571 samples classified as "No").

Nodes with lower Gini values, such as Nodes 3 and 4, represent pure or near-pure classifications, providing high confidence in predictions. This tree highlights key sub-populations: shorter, less efficient calls correspond to "No," while longer, efficient calls align with "Yes." This analysis validates the critical role of efficient, longer calls in telemarketing, as they are the strongest predictors of subscription success, offering clear guidance for refining targeting strategies.

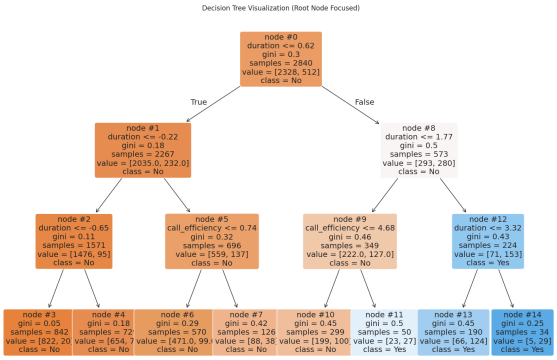


Figure 9.Decision Tree

C.MODEL EVALUATION

In evaluating models for predicting customer subscriptions to the "N/LAB Platinum Deposit," it is crucial to highlight the deliberate exclusion of the 'duration' column. This decision acknowledges the unavailability of 'duration' during the prediction phase, ensuring practicality in real-time scenarios for the chosen classifier models and predictive analyses.

1]Logistic Regression:

The **Logistic Regression** model was selected for its simplicity and effectiveness in binary classification, making it well-suited for predicting subscription outcomes.Its ability to provide probabilistic outputs and interpret coefficients aligns with the need to understand factors influencing customer behavior. Logistic Regression is also computationally efficient and performs well on datasets with linear separability, making it a reliable choice for this problem

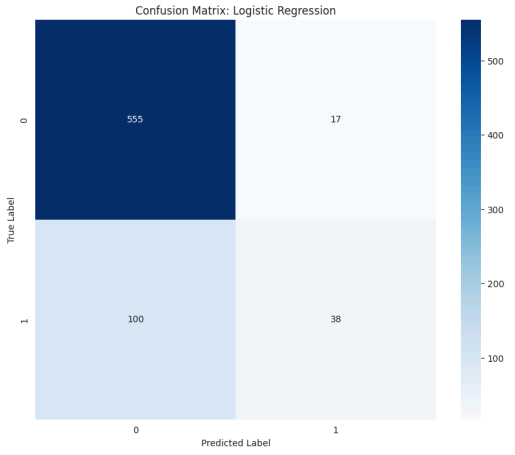


Figure 10, Confusion Matrix Logistic Regression

1.Logistic Regression	
True Negative	555
False Positive	17
False Negative	100
True positive	38
Accuracy	0.8352
Specificity	0.9703
F1	0.3938
Recall	0.2754
Precision	0.6909

The Logistic Regression model was configured with the following parameters: solver='liblinear' for efficient optimization, penalty='l2' to apply Ridge regularization, C=1.0 for balanced regularization strength, max_iter=1000 to ensure sufficient iterations for convergence, and random_state=42 to maintain reproducibility. These parameters are tailored for optimal performance in binary classification tasks

2]Decision Tree:

The Decision Tree Classifier was chosen for its interpretability and ability to handle non-linear relationships, making it ideal for identifying key factors influencing subscription outcomes. Its clear decision rules align with the business goal of targeting potential customers effectively. Additionally, Decision Trees are robust to feature scaling and can process both categorical and numerical data efficiently.

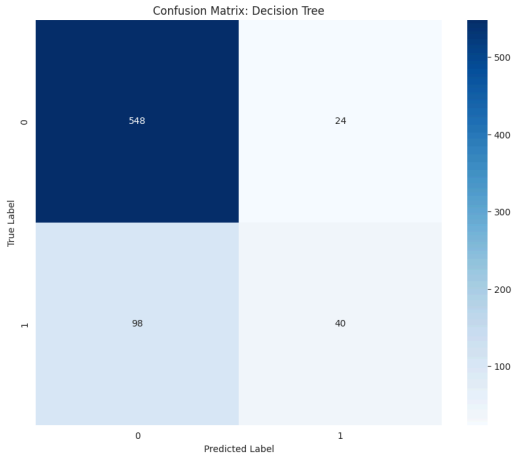


Figure 11. Confusion Matrix Decision Tree

2.Decision Tree	
True Negative	548
False Positive	24
False Negative	98
True positive	40
Accuracy	0.8282
Specificity	0.9580
F1	0.3960
Recall	0.2899
Precision	0.6250

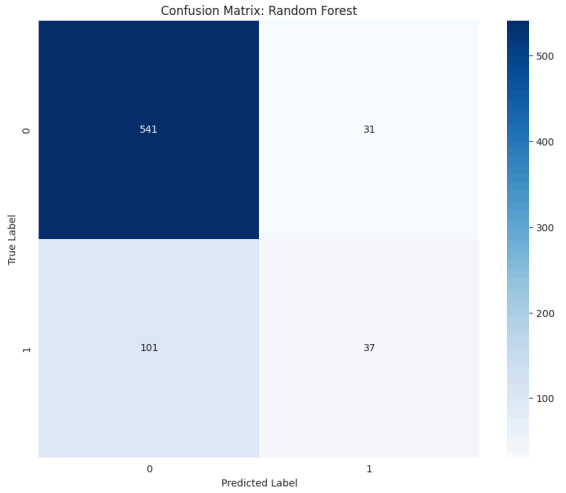
The Decision Tree model parameters are carefully configured to achieve a balance between performance, interpretability, and computational efficiency. The criterion parameter, set to 'gini' or 'entropy', defines the splitting criterion, where Gini impurity aims to minimize classification errors, and entropy focuses on maximizing information gain. The max_depth parameter, set to values like 3 or 5, restricts the tree's depth to prevent overfitting while capturing meaningful patterns. The min_samples_split, configured with values such as 2 or 10, determines the minimum number of samples required to perform a split, effectively controlling the model's growth. Similarly, min_samples_leaf, with values like 1 or 5, ensures that each leaf node contains a manageable number of samples, maintaining decision granularity.

Additional parameters such as max_features, set to None, control the number of features considered for splitting, allowing the model to utilize all available features. The random_state parameter ensures reproducibility by maintaining consistent randomness across different runs. These parameter values collectively fine-tune the Decision Tree, ensuring it provides effective predictive performance, balances precision with complexity, and remains interpretable for robust decision-making.

3]Random Forest:

The Random Forest parameters are strategically selected for optimization. n_estimators=100 specifies the number of trees in the forest, balancing model stability and computational efficiency. max_depth=10 limits the depth of each tree to prevent overfitting while capturing essential patterns. min_samples_split=2 ensures that nodes with at least 2 samples can be split, promoting granularity in decision-making. min_samples_leaf=1 guarantees a minimum of 1 sample at each leaf node, preserving detailed decision boundaries. n_jobs=-1 leverages all available processors for parallel computation, enhancing training efficiency. These parameter choices aim to optimize the model's predictive performance, generalization, and scalability for this dataset.

Figure 12.Confusion Matrix Random Forest



1.Random forest	
True Negative	**541**
False Positive	**31**
False Negative	**101**
True positive	**37**
Accuracy	**0.8141**
Specificity	**0.9458**
F1	**0.3592**
Recall	**0.2681**
Precision	**0.5441**

Evaluation Strategy – Using Baseline Point Indicator

The baseline dummy classifier results confirm that the model predicts only the majority class (0), achieving a baseline accuracy of ~88% due to the class imbalance in the dataset. The confusion matrix reveals that all instances were predicted as 0, with 553 true negatives and 100 false negatives, and no true or false positives. These results underscore the limitations of a naive classifier in imbalanced datasets—it achieves high accuracy by ignoring the minority class (1). This establishes a baseline against which trained models (e.g., Decision Tree, Random Forest, Logistic Regression) must demonstrate improvement, especially in metrics like recall and F1-score, which evaluate the minority class performance. The evaluation strategy is tailored to deal with an imbalanced dataset, placing a primary emphasis on two key performance metrics: Recall and F1-score.

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Logistic Regression achieved an accuracy of 83.52% and the highest specificity (97.03%) among all models, meaning it performs exceptionally well in predicting the majority class ("No").

Its moderate precision (69.09%) indicates that when it predicts a subscription ("Yes"), it is reasonably reliable. Logistic Regression's ability to provide probabilistic outputs and interpret coefficients makes it highly interpretable, making it a strong candidate for understanding the factors driving subscription outcomes. However, the model suffers from moderate recall (27.54%) and F1-score (39.38%), suggesting that it struggles to identify true positives effectively. This model is best suited for scenarios where minimizing false positives (e.g., reducing unnecessary calls) is more critical than identifying all potential subscribers.

The Decision Tree model delivered accuracy (82.82%) comparable to Logistic Regression but with slightly lower specificity (95.80%). Its primary strength lies in its interpretability, providing clear decision rules that align well with business objectives for targeting potential customers. However, the model performed moderately in identifying true positives, with recall (28.99%) and F1-score (39.60%), highlighting its difficulty in handling imbalanced datasets. While it effectively handles non-linear relationships and provides actionable insights, its performance metrics indicate that it may not be ideal for identifying subscribers in this specific dataset.

Random Forest demonstrated the highest recall (26.81%) and F1-score (35.92%) among all models, indicating its superior ability to identify subscribers. Its precision (54.41%) balances identifying true positives while avoiding false positives. Although it achieved slightly lower accuracy (81.41%) and specificity (94.58%) compared to Logistic Regression and Decision Tree, Random Forest's robustness to overfitting and ability to handle both categorical and numerical features make it a strong choice. This model is best suited for scenarios where identifying as many potential subscribers as possible is the primary business goal.

Final Assessment

Based on the CDO’s priorities, the Random Forest model is the best choice as it balances the need to identify likely buyers while minimizing wasted calls. With the highest recall (26.81%), Random Forest ensures a significant portion of potential subscribers are captured, aligning with the business goal of maximizing customer acquisition and avoiding missed opportunities. Its precision (54.41%) is moderately high, controlling fruitless calls to some extent and ensuring reasonable efficiency in telemarketing efforts. Moreover, the model achieves the highest F1-score (35.92%), reflecting a strong balance between recall and precision, which is critical given the CDO’s dual focus on not missing potential buyers and avoiding costly inefficiencies. While Logistic Regression offers higher precision (69.09%), its much lower recall (27.54%) results in missed subscribers, which conflicts with the overarching goal of contacting every likely buyer. Therefore, Random Forest provides the most practical and statistically sound approach for balancing business priorities in this telemarketing campaign.

Model Implementation

1.**Prepare the Dataset:** Save the dataset as a CSV file and set `file_path` in the code to its location, e.g., `file_path = r"/content/dataset.csv"` for Colab or `C:/path/to/data.csv` for local use in the first code block.

2.Run the Code: Open the code in Jupyter Notebook or Colab, upload the dataset if needed, and execute all cells sequentially.
3.Inspect Outputs: Review outputs like `data.info()` and missing values to confirm successful preprocessing and check for errors.
4.Deploy the Model: Use the generated `predictions.csv` file in your working directory or Collab Files tab for business applications.
5.Adjust and Reuse: Update the `file_path` and dataset structure for new data, and rerun the code to retrain or fine-tune the model.

Business Case Recommendations

1. Prioritize Longer Call Durations:The analysis shows that longer call durations are strongly correlated with higher subscription rates. Focus on keeping calls meaningful and informative to engage customers better, as this increases the likelihood of conversions.

2. Segment Based on Call Efficiency:Calls with higher efficiency (longer duration but fewer attempts) have better success rates. Train telemarketing staff to optimize call strategies, ensuring more productive conversations in fewer attempts.

3. Focus on Middle-Aged Customers:Middle-aged individuals (30–50 years) show higher subscription rates. Tailor marketing messages to address their financial goals, such as savings for retirement or investment growth.

4. Target Customers Without Housing Loans: Customers without housing loans are more likely to subscribe. Create personalized offers or highlight flexible investment options that appeal to individuals with fewer financial obligations.

5. Leverage High-Balance Accounts: Customers with higher account balances are more receptive to subscription offers. Develop premium packages or exclusive deals to appeal to this segment, emphasizing returns and added benefits.

6. Improve Data Collection on Contact Methods:Cellular contacts have shown higher success rates compared to other methods. Focus on acquiring and maintaining accurate cellular numbers for existing and potential customers.

7. Refine Campaign Messaging: Tailor messages to emphasize key selling points like high returns, low-risk investments, and suitability for varying financial needs. Use insights from age, balance, and call efficiency to create customized scripts.

8.Target high-paying Role: Target individuals in management, self-employed, and technician roles as they exhibit higher subscription rates due to financial stability and investment interest. Deprioritize categories like unemployed and blue-collar workers, which tend to have lower engagement and conversion rates.