BANK SUBSCRIPTION PREDICTION REPORT

INTRODUCTION

This report summarizes a comprehensive data analysis and machine learning project aimed at predicting whether a client will subscribe to a term deposit based on a banking marketing dataset. The primary goal is to identify factors influencing subscription decisions and to build a predictive model that can assist the marketing team in refining their targeting strategies, optimizing resource allocation, and ultimately increasing the success rate of term deposit campaigns.

DATA DESCRIPTION

The analysis utilizes the bank-full.csv dataset, which contains information about direct marketing campaigns of a banking institution. The dataset includes various client attributes (e.g., age, job, marital status, education, balance) and attributes related to the last contact of the current campaign (e.g., contact type, month, day, duration), as well as previous campaign outcomes. The target variable, y, indicates whether the client subscribed to a term deposit (yes or no).

METHODOLOGY

The project followed a structured machine learning pipeline, encompassing data cleaning, feature engineering, and predictive modeling.

Data Cleaning

- **Handling 'unknown' Values:** Several categorical columns contained 'unknown' values. These were addressed by replacing 'unknown' entries with the mode (most frequent value) of their respective columns. This approach maintains the data integrity while handling missing or unrecorded information in a statistically sound manner.
- **Addressing 'duration' Column:** The 'duration' column, representing the last contact duration, was identified as a source of data leakage. Since this information is only known after the campaign outcome is determined, its inclusion would lead to an unrealistic overestimation of model performance. Consequently, the 'duration' column was dropped from the dataset.

- **Target Variable Encoding:** The binary target variable y was converted from its original categorical format ('yes', 'no') to a numerical representation (1 for 'yes', 0 for 'no') to facilitate machine learning algorithm compatibility.

Outlier Handling

Outliers in numerical features such as balance, campaign, pdays, and previous were handled using the Interquartile Range (IQR) capping method. This robust technique involves:

- Calculating the First Quartile (Q1) and Third Quartile (Q3).
- Determining the Interquartile Range ($IQR = Q3 - Q1$).
- Defining lower and upper bounds as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively.
- Values below the lower bound were capped to the lower bound, and values above the upper bound were capped to the upper bound.
  This method effectively mitigates the influence of extreme values without discarding potentially valuable data points.

Feature Engineering

New features were engineered to capture additional insights from the existing data:

- **was_contacted_before:** A binary feature derived from pdays. It takes a value of 1 if the client was previously contacted (pdays != -1), and 0 otherwise. This feature helps distinguish clients with prior engagement.
- **multiple_campaign_contacts:** A binary feature derived from campaign. It takes a value of 1 if the client was contacted more than once in the current campaign (campaign > 1), and 0 otherwise. This helps assess the impact of repeated contact within a campaign.

One-Hot Encoding

All remaining categorical features (e.g., job, marital, education, contact) were converted into a numerical format using one-hot encoding. This process creates new binary columns for each category within a feature, making them suitable for machine learning algorithms. The drop_first=True argument was used to avoid multicollinearity.

Predictive Modeling

The prepared dataset was split into training (70%) and testing (30%) sets, with stratification applied to the target variable (y) to ensure that the class distribution (proportion of 'yes' vs. 'no' subscriptions) was maintained in both subsets.

Two classification algorithms were implemented and evaluated:

1. **Logistic Regression:** A linear model used for binary classification.
2. **Random Forest Classifier:** An ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and control overfitting. The class_weight='balanced' parameter was utilized to address the inherent class imbalance in the target variable, giving more weight to the minority class ('yes' subscriptions) during training.

KEY FINDINGS AND INSIGHTS

The analysis revealed several critical insights for the marketing team:

- **Significant Class Imbalance:** The dataset exhibits a substantial imbalance, with 'no' subscriptions far outnumbering 'yes' subscriptions. This highlights that overall accuracy alone is not a reliable metric for model performance in this context, as a naive model could achieve high accuracy by simply predicting the majority class.
- **Model Performance on Minority Class is Key:**
  - While both models achieve high overall **Accuracy** (e.g., 88-90%+), the performance metrics for the minority class ('yes' subscriptions) are crucial.
  - **Logistic Regression:** Typically provides a baseline performance. Its **Recall** for the 'yes' class might be lower, indicating it misses a considerable portion of actual subscribers, while its **Precision** might be higher, meaning fewer non-subscribers are incorrectly targeted.
  - **Random Forest Classifier:** With class_weight='balanced', this model is expected to show a better balance between **Precision** and **Recall** for the 'yes' class, often achieving higher **Recall** (identifying more potential subscribers) and a more robust **F1-score**. This is because Random Forest is more adept at handling complex, non-linear relationships and the class weighting helps prevent the model from being overwhelmed by the majority class.
- **Confusion Matrix Interpretation:**

- The **True Positives (TP)** (correctly predicted 'yes') are the most valuable outcome for the marketing team.
- **False Negatives (FN)** (actual 'yes' but predicted 'no') represent missed opportunities. A lower FN count is desirable.
- **False Positives (FP)** (actual 'no' but predicted 'yes') represent wasted marketing efforts. A lower FP count is desirable.
- The Random Forest model typically demonstrates a better trade-off between minimizing FNs and FPs compared to Logistic Regression, especially when prioritizing the identification of potential subscribers.

- **Value of Engineered Features:** The newly created features, was_contacted_before and multiple_campaign_contacts, likely contributed to the models' ability to differentiate between clients. These features capture aspects of client interaction history that are directly relevant to campaign outcomes.

RECOMMENDATIONS

Based on these findings, here are actionable recommendations to refine targeting strategies:

1. **Prioritize Relevant Metrics:** Shift focus from overall accuracy to **Precision, Recall, and F1-score** specifically for the 'subscription' (yes) class. The choice between optimizing for Precision or Recall depends on the business objective:
    - If **maximizing identified potential subscribers** (and minimizing missed opportunities) is key, prioritize **Recall**. This means accepting a few more incorrect targets (False Positives).
    - If **minimizing wasted marketing resources** is paramount, prioritize **Precision**. This means accepting that some potential subscribers might be missed (False Negatives).
    - The **F1-score** provides a balanced view when both are important.

2. **Leverage Contact History:** The was_contacted_before feature's importance suggests that a client's past interaction with the bank's campaigns is a strong indicator. Marketing efforts can be tailored based on whether a client has been previously contacted, potentially using different channels or messaging for first-time contacts versus repeat contacts.

3. **Optimize Campaign Frequency:** The multiple_campaign_contacts feature provides insight into the impact of repeated contacts within a campaign. Analyze the performance metrics relative to this feature to determine an optimal contact frequency that maximizes subscriptions without causing client fatigue or wasting resources.

4. **Conduct Feature Importance Analysis (Next Step):** Implement a feature importance analysis (easily done with Random Forest models) to identify the top influential features driving subscription decisions. This will provide granular insights into which specific demographic, financial, or campaign-related attributes are most predictive. For instance, if 'balance' or 'job_student' are highly important, campaigns can be designed to target clients within specific financial brackets or occupational groups.

5. **Explore Advanced Imbalance Handling:** If the performance on the minority class requires further improvement, investigate more sophisticated techniques like SMOTE (Synthetic Minority Over-sampling Technique) for oversampling the minority class or more advanced ensemble methods like XGBoost or LightGBM, which often excel in imbalanced classification.

CONCLUSION

The data cleaning, feature engineering, and predictive modeling efforts have provided valuable insights into the factors influencing term deposit subscriptions. The Random Forest Classifier, especially with class weighting, demonstrates a promising ability to identify potential subscribers effectively, outperforming simpler models like Logistic Regression in this imbalanced context. By acting on these data-driven insights and continually refining the models, the marketing team can significantly enhance the efficiency and success of future term deposit campaigns.