Comprehensive Report: Predicting Term Deposit Subscription

This report summarizes the entire data analysis project aimed at predicting whether a client will subscribe to a term deposit at a banking institution. It covers the methodology, key findings, and actionable insights derived from the data and the predictive model.

1. Introduction

The objective of this project was to leverage a provided dataset of bank marketing campaigns to build a predictive model. The model's purpose is to identify clients most likely to subscribe to a term deposit, thereby enabling the marketing team to optimize their direct marketing efforts, reduce costs, and increase subscription rates.

2. Data Exploration and Preprocessing

The analysis began with a comprehensive Exploratory Data Analysis (EDA) of the `bank-full.csv` dataset, which contained 45,211 entries across 17 attributes.

- **Initial Data Assessment:** No missing values were identified, ensuring data completeness. The dataset comprised both numerical and categorical features.
- **Class Imbalance Handling:** A significant imbalance was observed in the target variable `y` (term deposit subscription), with a large majority of clients not subscribing ('no') compared to those who did ('yes'). To address this, the minority class ('yes') was **oversampled** to match the count of the majority class, resulting in a balanced dataset of 79,844 entries (39,922 'yes' and 39,922 'no'). This crucial step prevents the model from being biased towards the more frequent outcome.
- **Numerical Feature Analysis:**
    - Features like `balance`, `campaign`, `pdays` (days since last contact), and `previous` (number of previous contacts) exhibited highly skewed distributions and contained numerous outliers.
    - **Preprocessing Action:** To mitigate the impact of varying scales and outliers, all numerical features (`age`, `balance`, `day`, `campaign`, `pdays`, `previous`) were scaled using `StandardScaler`. This standardization ensures that features with larger value ranges do not disproportionately influence the model.

3. Feature Engineering

Strategic feature engineering was performed to prepare the data for effective modeling:

- **`duration_zero` Feature:** The original `duration` of the last contact is a strong predictor but poses a data leakage risk (it's known *after* the call). To address this, a new binary feature, `duration_zero`, was created. This feature indicates whether the contact duration was 0 seconds, capturing instances of no effective contact. The original `duration` column was then removed.
- **Categorical Feature Encoding:**
    - Binary categorical features (`default`, `housing`, `loan`, and the target `y`) were converted to numerical `1`s and `0`s.
    - Multi-category nominal features (`job`, `marital`, `education`, `contact`, `month`, `poutcome`) were transformed into numerical format using **one-hot encoding** (`pd.get_dummies`). This creates separate binary columns for each category, allowing the model to process them without assuming any ordinal relationship.

4. Predictive Model Building

A **Logistic Regression** model was chosen for its interpretability and robust performance in binary classification tasks.

- **Data Splitting:** The balanced and preprocessed dataset was split into training (80%) and testing (20%) sets. A **stratified split** was applied to maintain the proportional representation of 'yes' and 'no' subscriptions in both subsets.
- **Model Training:** The Logistic Regression algorithm was trained on the prepared training data to learn the intricate relationships between client features and their likelihood of subscribing to a term deposit.

5. Model Performance Evaluation

The model's effectiveness was rigorously assessed on the unseen test dataset using a suite of standard metrics:

- **Accuracy:** The model achieved an overall accuracy of **70%**, indicating that it correctly predicted the subscription outcome for 7 out of 10 clients.
- **Classification Report:**

    - **Precision (Class 'yes'):** 0.73. This is a critical metric for marketing, indicating that when the model predicts a client will subscribe, it is correct **73% of the time**. This helps minimize wasted marketing efforts.
    - **Recall (Class 'yes'):** 0.63. The model successfully identified 63% of all actual subscribers in the test set. While good, this suggests there's still a portion of potential subscribers that the model might miss.
    - **F1-Score (Class 'yes'):** 0.68. The F1-score, a harmonic mean of precision and recall, indicates a balanced performance for the 'yes' class.
    - The metrics for Class 'no' were also strong (Precision: 0.67, Recall: 0.77, F1-Score: 0.72), demonstrating the model's ability to identify non-subscribers effectively.
- **ROC-AUC Score:** The model achieved a **ROC-AUC score of 0.77**. This value indicates good discriminatory power, meaning the model is reasonably effective at distinguishing between clients who will subscribe and those who will not.

6. Key Findings and Actionable Recommendations

The analysis of the Logistic Regression model's coefficients provided crucial insights into which features most strongly influence subscription outcomes, leading to actionable recommendations for the marketing team:

- **Past Success is Key:** The poutcome_success feature was the most impactful predictor. Clients who previously subscribed to a term deposit are significantly more likely to subscribe again.
    - **Recommendation:** Prioritize and aggressively re-target clients with a history of successful campaign outcomes. They represent the highest-potential leads.
- **Optimize Campaign Timing:**
    - **Recommendation:** Concentrate marketing efforts during **March, October, and September**, as these months exhibit the highest subscription rates.
    - **Recommendation:** Re-evaluate or reduce campaign intensity for term deposits during **January, August, November, and July**, as these months show significantly lower conversion rates.
- **Refine Contact Strategy:**
    - **Recommendation:** Strongly favor cellular contact methods, which proved to be the most effective.
    - **Recommendation:** Improve data collection to ensure the contact type is always known; avoid campaigns where the contact method is unknown, as these are largely ineffective.
- **Segment by Job, Marital Status, and Education:**
    - **Recommendation:** Tailor specific marketing messages and offers for students and retired individuals, as they are more receptive to term deposits. Similarly, target single clients and those with tertiary education.

- **Consider Loan Status:**
    - **Recommendation:** Be aware that clients with housing loans are less likely to subscribe. While not a reason to exclude them, consider different product pitches or allocate fewer resources to this segment for term deposit campaigns.

7. Conclusion

The developed Logistic Regression model offers valuable predictive capabilities for the banking institution's direct marketing campaigns. By leveraging the insights into key client characteristics and optimal campaign timing, the marketing team can significantly enhance their targeting efficiency, leading to higher subscription rates for term deposits and a more effective allocation of marketing resources. Continuous monitoring and periodic retraining of the model with new data will ensure its ongoing relevance and performance.