Johnny Rodriguez
Data 622

Analysis of Bank Marketing Data

Introduction

The central objective is to predict whether a bank client will subscribe to a term deposit. By accurately identifying clients with a higher propensity to subscribe, financial institutions can increase their return on investment and focus their marketing efforts more effectively. The approach involves preparing this data for modeling, evaluating and comparing a Logistic Regression and a Decision Tree model, and then using a robust evaluation strategy to select the superior algorithm for this task.

Data Preprocessing

The dataset contains a mix of continuous variables and categorical variables. To prepare this data for modeling, categorical data will need to be converted into a numerical format using one-hot encoding, and continuous data will be standardized using feature scaling.

This two-part approach is optimal for this dataset. One-hot encoding is essential because it creates new binary columns for each category, preventing algorithms from assuming an incorrect natural order between categories. Feature scaling is critical for the Logistic Regression model, as it ensures that continuous features with wide ranges do not disproportionately influence the model's predictions over features with smaller ranges.

Algorithm Selection: Logistic Regression vs. Decision Tree

Two models are selected to compare distinct modeling approaches: Logistic Regression and a Decision Tree. This selection allows for a direct comparison between an algorithm that assumes a linear relationship in the data and one that can capture complex, non-linear patterns.

- Logistic Regression finds a single linear boundary to separate classes, making its results highly interpretable through model coefficients. However, it requires data preparation, including scaling and encoding, and may miss complex interactions. In contrast, a Decision Tree creates a series of rule-based splits, allowing it to model non-linear relationships and interactions automatically. It handles categorical and continuous data natively without requiring feature scaling.
- Handling Data and Imbalance: For this dataset's mixed data types, the Decision Tree is more flexible. However, the primary challenge is the severe class imbalance. The Decision Tree's structure of partitioning data into segments also makes it naturally suited for isolating the small, distinct groups of clients who are likely to subscribe.

Modeling and Evaluation Strategy

The class imbalance will be handled directly during model training. The technique adjusts the learning algorithm to give more importance to the minority class, preventing the model from simply defaulting to the majority 'no' prediction and ensuring it learns the patterns associated with subscribers.

Because of this imbalance, accuracy is a misleading evaluation metric. Instead, model performance will be assessed using the following:

- Precision and Recall: To understand the trade-off between the correctness of positive predictions and the model's ability to find all actual positive cases.
- F1-Score: To find a balance between precision and recall, providing a single, robust measure of the model's effectiveness on the minority class.
- ROC-AUC Score: To evaluate the model's overall ability to discriminate between the 'yes' and 'no' classes, regardless of the classification threshold.

Final Algorithm Recommendation

The recommendation is to use the Decision Tree model. This algorithm is recommended because it is better equipped to solve the core business problem of identifying potential subscribers.

A client's decision to subscribe is influenced by a complex interplay of factors that are not linear. The Decision Tree's key advantage is its ability to automatically discover these non-linear relationships and feature interactions. For example, it can identify specific niches, such as "clients over 50 in management roles with no existing housing loan," that have a high propensity to subscribe—a pattern a linear model would likely miss. The Decision Tree provides a more accurate and actionable tool for targeted marketing, directly addressing the primary objective.