Seth Harris

Joaquin Ramirez

Collin Real

Leonel Salazar

Professor Roy

DA 6813-901

10 September 2024

**Bank Marketing Case Study**

**I. Executive Summary**

This case study aims to analyze a Portuguese banking institution’s marketing campaign data and build a predictive model to determine whether a client will subscribe to a term deposit. Using the bank-additional.csv dataset, a logistic regression model, a decision tree, and a random forest were compared. The random forest model provided the best performance for predicting term deposit subscriptions, achieving an accuracy rate of approximately 89%.

**II. Introduction**

The dataset contains data on direct marketing campaigns conducted through phone calls. These campaigns are designed to promote a bank’s term deposits. Multiple contacts with a client may be required to assess whether they will subscribe to a term deposit (binary classification: ‘yes’ or ‘no’). This study uses various input variables such as the client’s age, job, and other social and economic factors to predict the likelihood of subscription. The dataset’s goal is to improve the efficiency of future campaigns by understanding which factors are most predictive of success.

**III. Review of Related Literature**

Predictive models such as logistic regression, decision trees, and random forests are widely used in marketing and customer acquisition studies. According to Breiman (2001), random forests have emerged as a leading method for classification due to their ability to handle high-dimensional data and interactions without requiring extensive feature engineering. This study compares random forests with logistic regression and decision tree models to identify the best model for predicting bank term deposit subscriptions.

**IV. Methodology**

The dataset consists of 21 variables, including social, economic, and campaign-related features. Key variables include age, job type, marital status, education, and duration of the call. In terms of data pre-processing, missing values were handled by removing instances where the ‘unknown’ category occurred in significant variables like job, education, and default status. The duration variable was excluded as its inclusion would not create a realistic model since it would be known only after a campaign call. Our modeling approach used

• **Logistic Regression**: Used to predict the binary response variable based on a linear combination of predictor variables.

• **Decision Tree**: A non-parametric model that splits the data based on the most important predictor variables.

• **Random Forest**: An ensemble method that uses multiple decision trees to improve classification accuracy by reducing overfitting.

The models were evaluated based on accuracy, precision, recall, and F1-score to assess performance.

**V. Data Exploration**

A preliminary analysis of the data showed the following insights:

• **Age Distribution**: Clients aged 30–40 were more frequently contacted, and younger clients tended to subscribe more.

• **Job and Marital Status**: Clients working in ‘management’ and ‘services’ were more likely to subscribe. Married individuals had a slightly higher subscription rate than single individuals.

• **Education**: Clients with higher education (university degree) were more likely to subscribe.

• **Campaign Factors**: The number of contacts, month of contact, and the outcome of the previous campaign (poutcome) were crucial variables. Clients contacted more than 5 times were less likely to subscribe.

**VI. Data Cleaning**

The following data cleaning steps were taken:

1. Removal of the duration variable due to its high influence on the target variable.

2. Replacement of ‘unknown’ values in job, education, and default with the most frequent category.

3. Standardization of numeric variables such as age, campaign, and pdays to improve model convergence.

**VII. Data Analysis**

1. **Logistic Regression**: This model showed moderate accuracy in predicting term deposit subscriptions, with age, job, marital status, and education as the most significant predictors. However, it struggled with non-linear relationships and interactions between variables.

2. **Decision Tree**: The decision tree model performed slightly better, especially in capturing interactions between variables such as age and job type. However, it was prone to overfitting, which reduced its generalizability.

3. **Random Forest**: The random forest model performed the best, achieving an accuracy of 89%. It effectively handled interactions and non-linear relationships. The most important features identified were age, poutcome (previous campaign outcome), and number of contacts during the campaign.

**VIII. Discussion of Results**

The random forest model’s superior performance can be attributed to its ability to handle complex interactions between variables without explicit modeling. Variables like the number of contacts, previous campaign outcome, and age were the most important predictors. Clients with a successful outcome from a previous campaign were more likely to subscribe in the current campaign.

**IX. Interpretation and Conclusion**

The random forest model is the most effective for predicting whether a client will subscribe to a term deposit. It captures interactions between variables without requiring complex feature engineering. The model’s ability to handle both categorical and numeric variables makes it suitable for marketing campaigns. Based on these results, the bank should focus its marketing efforts on clients who had a positive outcome in previous campaigns and use a targeted approach based on age, job, and number of contacts.

Future studies could explore deeper feature engineering techniques to improve the performance of logistic regression and decision tree models.

**Appendix: Code**