Levesque Marie
March 24, 2025

Predictive Modelling of Long-Term Bank Deposits

An Exploratory and Statistical Analysis

1. Introduction

The competitive nature of the banking sector has intensified the pursuit of effective marketing strategies to retain existing customers and attract new clients. One prevalent approach involves telemarketing campaigns that encourage customers to subscribe to long-term deposit accounts. This study investigates whether data-driven techniques can accurately predict an individual's likelihood of accepting a long-term deposit offer, thereby enabling more efficient and cost-effective telemarketing strategies. The data used in this analysis was collected from a Portuguese bank's telemarketing campaigns between 2008 and 2013, as documented by Moro, Cortez, and Rita (2014). Each observation comprises nineteen features that capture demographic information, past campaign results, and macroeconomic conditions, alongside a binary target variable indicating whether the customer eventually subscribed to a deposit. The primary objective is to identify a robust predictive model that will facilitate effective resource allocation during telemarketing campaigns.

2. Data Preprocessing and Exploratory Analysis

The dataset was imported into a Python environment using pandas, and an initial inspection of the data structure was conducted. These steps confirmed that the dataset contained no blank or null values in the traditional sense, although the string "unknown" was employed in several categorical fields to denote missing or unrecorded information. This approach to missing data aligns with observations noted by Moro et al. (2014), who suggested that the banking dataset frequently uses placeholders instead of standard null values. Consequently, all "unknown" entries were replaced with actual NaN values to facilitate appropriate handling of *missingness*, an approach that is consistent with standard practices in data cleaning as described by Witten, Frank, and Hall (2011).

After identifying the columns that contained missing data, each numerical column was imputed using the median value, whereas each categorical column introduced a new category labelled "missing." This method preserves the distribution of the data and prevents unnecessary row deletions that might reduce the sample size. Exploratory Data

Analysis (EDA) followed these preprocessing steps to reveal patterns in the data and to examine the distribution of each variable (James, Witten, Hastie, & Tibshirani, 2013).

Several insights emerged from the EDA. A histogram of the age variable indicated that most customers were in their thirties and early forties, with a small portion above the age of 60. This distribution exhibits a mild right skew, suggesting that younger clients are less common in the dataset. The most frequent job categories were “admin.” and “blue-collar,” while the predominant marital status was “married.” The education variable revealed that “university.degree” was the most frequently recorded level, although several observations had originally been labelled “unknown” or “missing,” highlighting some degree of uncertainty about clients’ educational backgrounds. The outcome of previous marketing campaigns, stored in the poutcome variable, was predominantly labelled “nonexistent,” which indicates that a substantial proportion of clients had not been contacted in prior campaigns. This observation limits the extent to which one can rely on historical success or failure for future predictions, as also noted by Moro et al. (2014).

The correlation matrix generated for the numerical variables, including employment variation rate (emp.var.rate), consumer confidence index (cons.conf.idx), and euribor3m, highlighted moderate relationships among certain macroeconomic indicators. For instance, euribor3m showed a moderate positive correlation with nr.employed, which is consistent with broader economic trends in labor and interest rates (Friedman, 2001). Nonetheless, no exceptionally high correlations were observed that might pose a concern for multicollinearity in modelling. Furthermore, an analysis of missing data, visualised through a heatmap, demonstrated that missing values were concentrated in certain features, especially job, education. This pattern informed the decision to impute and to treat *missingness* as a new category for the affected features, a technique recommended in scenarios where missing data can hold meaningful information (Bishop, 2006).

3. Model Building and Methodology

Three classification models were chosen to predict whether a client would subscribe to a long-term deposit. The first model, Logistic Regression, was selected for its interpretability and widespread adoption in binary classification tasks (James et al., 2013). Logistic Regression uses a linear function of predictors and applies the logistic function to estimate the probability of a positive outcome. This approach makes it straightforward to interpret coefficient signs and magnitudes, although it assumes a log-odds linearity that might not always hold in complex real-world datasets.

The second model was a Decision Tree, which splits the data into subgroups based on the most informative features until reaching leaf nodes. Decision Trees are valuable for their transparency, as they provide explicit if-then decision rules that can be readily understood by non-technical stakeholders (Witten et al., 2011). However, they can over-fit easily, especially if grown to excessive depth without proper regularisation. Over-fitting occurs when the model captures noise rather than signal, leading to diminished performance on unseen data (Friedman, 2001).

The third model was a Random Forest, which is an ensemble of Decision Trees that aggregates predictions through majority voting. This ensemble approach often enhances generalisation by reducing variance compared to a single tree (Bishop, 2006). Each tree in the forest is trained on a bootstrap sample of the data, and a subset of features is randomly chosen at each split, further encouraging diversity among trees. This procedure typically yields higher accuracy than a single Decision Tree, though it can be more computationally expensive (James et al., 2013).

Before training, the dataset was divided into training (80%) and testing (20%) subsets. Stratification ensured that the ratio of positive to negative target outcomes was preserved in both sets, thus reflecting the original distribution of “yes” and “no” responses. One-hot encoding was applied to categorical variables to convert them into a series of binary indicators. This encoding approach is recommended by scikit-learn (scikit-learn developers, 2023) because it allows models to handle categorical data in a numerically consistent manner. Numerical variables were standardised using StandardScaler to ensure that features with larger numerical ranges would not dominate the learning process.

4. Results and Analysis

The Logistic Regression model achieved an accuracy of approximately 91 percent on the test set and an Area Under the ROC Curve (AUC) of 0.94. The confusion matrix revealed a relatively balanced classification, with few false positives and false negatives. The Decision Tree exhibited an accuracy of around 89 percent, accompanied by an AUC of 0.71, indicating that it did not discriminate as effectively between customers who accepted the deposit and those who declined. This result is consistent with prior observations that a single Decision Tree can over-fit when not sufficiently pruned (Bishop, 2006). The Random Forest classifier also reached an accuracy of roughly 91 percent and matched Logistic Regression with an AUC of 0.94. These findings underscore the advantage of ensemble methods in improving predictive performance compared to a single-tree approach.

The ROC curves for the three models showed that both Logistic Regression and Random Forest remained well above the diagonal line representing random guessing, while the Decision Tree’s curve stayed closer to the diagonal in certain regions. The classification reports, including precision, recall, and F1-score, confirmed that Random Forest and Logistic Regression were particularly adept at correctly identifying customers who did not subscribe, while still capturing a substantial fraction of those who did. The Decision Tree struggled more with false positives and false negatives, suggesting over-fitting to the training data.

5. Discussion

These results demonstrate that Random Forest and Logistic Regression are both effective for predicting the likelihood of deposit subscription, with comparable AUC scores of 0.94. The Random Forest model benefits from reduced variance through the ensemble of Decision Trees, which is a common outcome of bagging methods (James et al., 2013). Logistic Regression's competitive performance can be attributed to its suitability for binary classification and the relatively linear relationships within the dataset once appropriate transformations and imputations are performed. The Decision Tree's comparatively lower performance highlights its tendency to memorise training data without proper constraints.

From a practical standpoint, banks seeking to optimise telemarketing resources may prioritise customers who have higher predicted probabilities of subscription. Both Logistic Regression and Random Forest could effectively serve this purpose, with Random Forest often offering a slight performance edge and Logistic Regression delivering superior interpretability for stakeholders interested in understanding the effects of specific predictors. However, additional model tuning, including hyper-parameter optimisation, might further enhance performance. Methods such as cross-validation could also confirm the reliability of the models and mitigate issues of over-fitting (Friedman, 2001). Feature engineering, for example integrating external economic indicators or additional customer behavioural data, could increase the predictive power of the models. Furthermore, adopting cost-sensitive learning might be beneficial if the cost of a false positive (calling an uninterested customer) differs significantly from the cost of a false negative (failing to call a potentially interested customer).

6. Conclusion

The objective of this analysis was to develop a data-driven approach to predicting whether a customer would subscribe to a long-term bank deposit. Through rigorous data preprocessing and exploratory analysis, it was evident that age, job type, previous campaign outcomes, and macroeconomic features provided significant insights into customer behaviour. The modelling phase demonstrated that both Logistic Regression and

Random Forest achieved high predictive performance, each with an AUC of 0.94, whereas a single Decision Tree performed less favourably, consistent with well-known tendencies toward over-fitting. The insights gained underscore the importance of employing ensemble methods or robust regularisation strategies to manage complex data. From an applied perspective, banks could leverage these models to enhance telemarketing campaigns by focusing resources on customers with a higher probability of accepting deposit offers, thus minimizing operational costs and improving overall efficiency. Future research could explore more advanced methods, such as gradient boosting or deep learning architectures, alongside hyper-parameter tuning, to refine model performance and maintain adaptability to changing market conditions. Continuous monitoring of the models' predictive power is crucial, as shifts in customer behaviour or economic trends can alter the underlying relationships upon which these predictions are based. This study contributes to the growing body of literature that illustrates how data analytics and machine learning can support effective marketing strategies within the banking sector.

References

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Friedman, J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), pp.1189–1232.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Moro, S., Cortez, P. and Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22–31.

scikit-learn developers, 2023. *scikit-learn documentation*. [online] Available at: https://scikit-learn.org/stable/ [Accessed 24 March 2025].

UCI Machine Learning Repository, 2023. *Bank Marketing Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing [Accessed 24 March 2025].

Witten, I.H., Frank, E. and Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco: Morgan Kaufmann.

7. Annexes

7.a) Exploratory Data Analysis

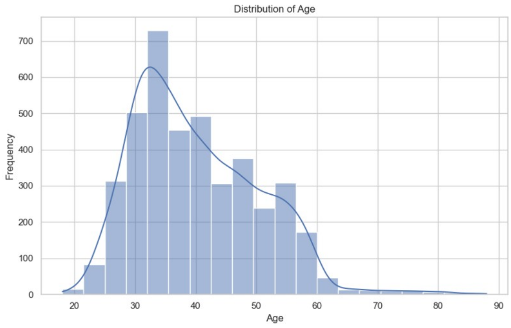


Figure 1: Age distribution

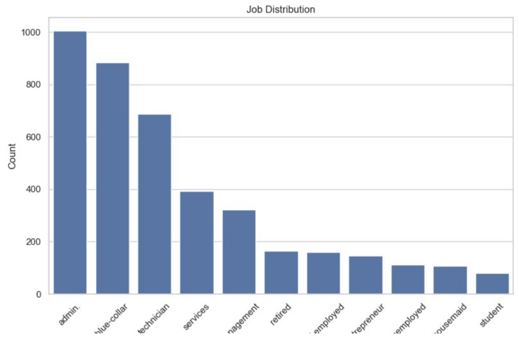


Figure 2: Job distribution

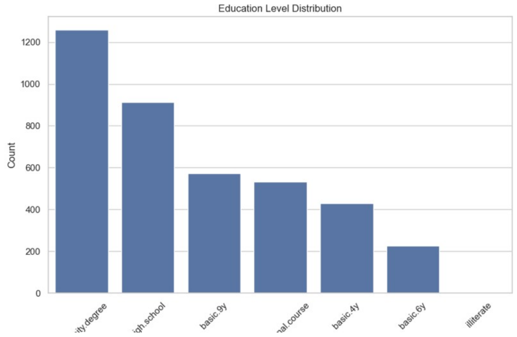


Figure 3: Education level distribution

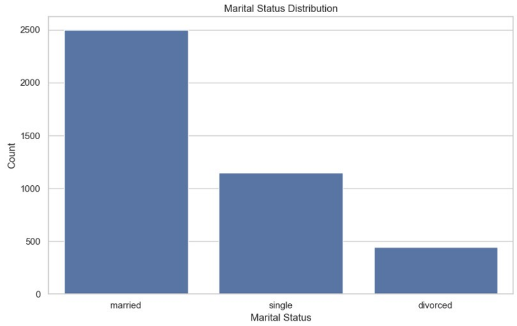


Figure 4: Marital status distribution

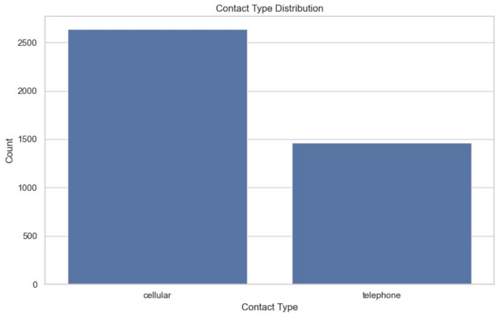


Figure 5: Contact type distribution

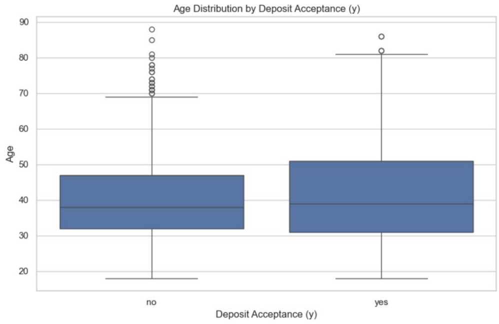


Figure 6: Age distribution by Deposit Acceptance

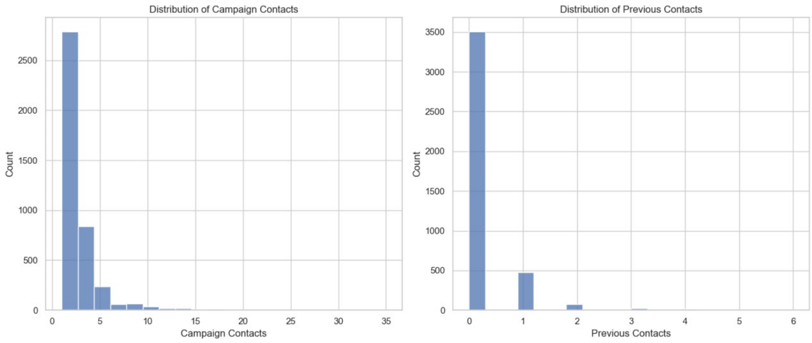


Figure 7: Distributions of Campaign Contacts and Previous Contacts

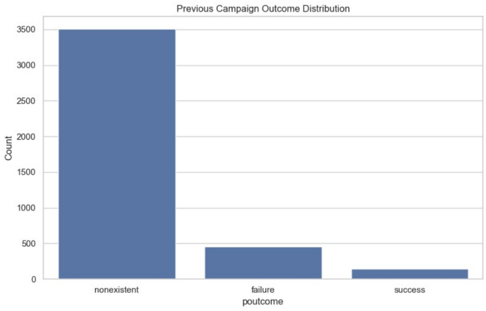


Figure 8: Previous Campaigns Outcome distribution

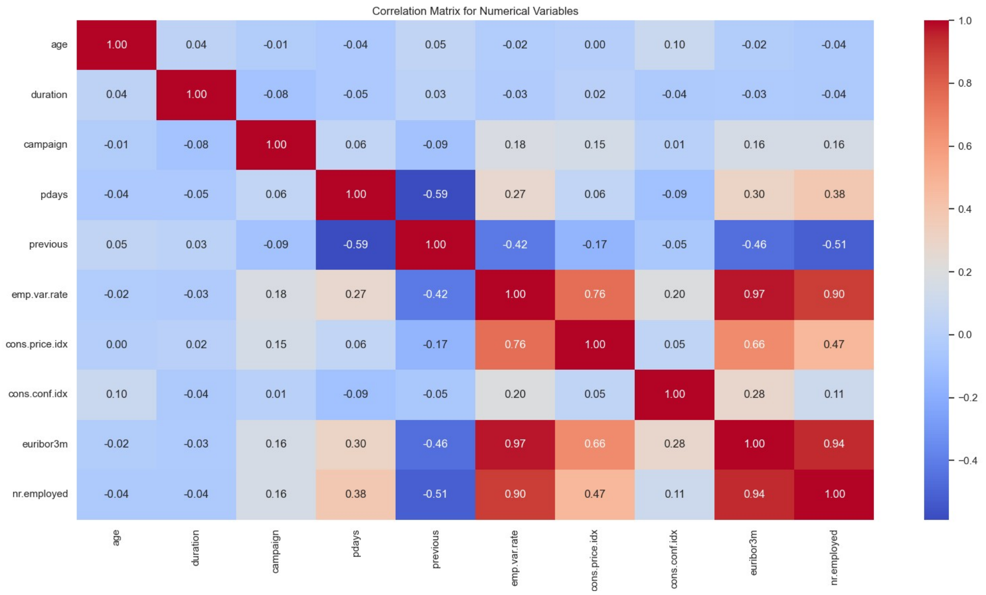


Figure 9: Correlation Matrix for Numerical Variables

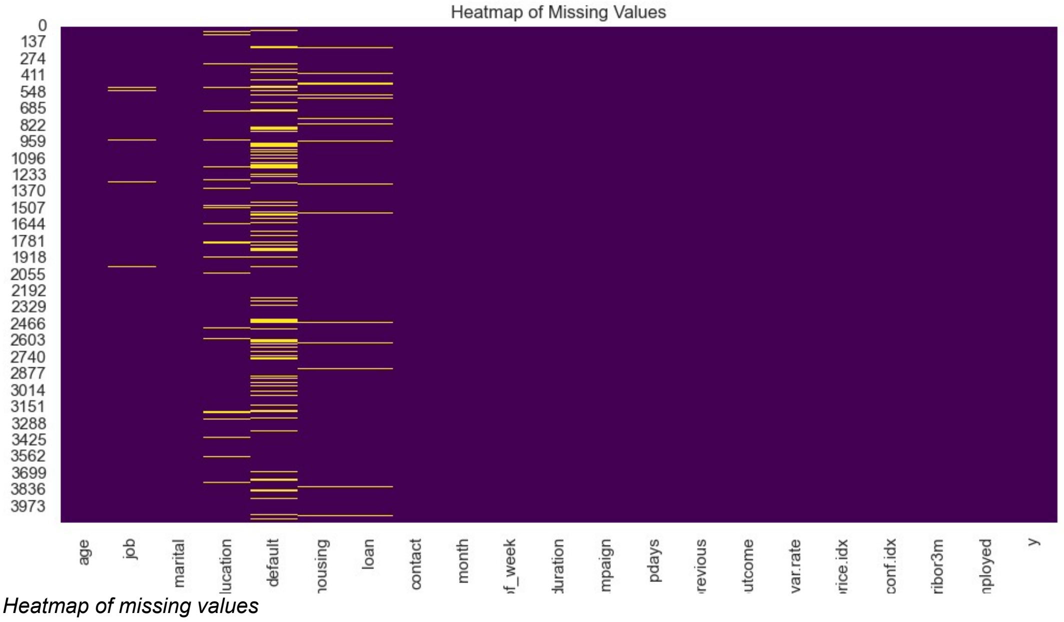


Figure 10: Heatmap of missing values

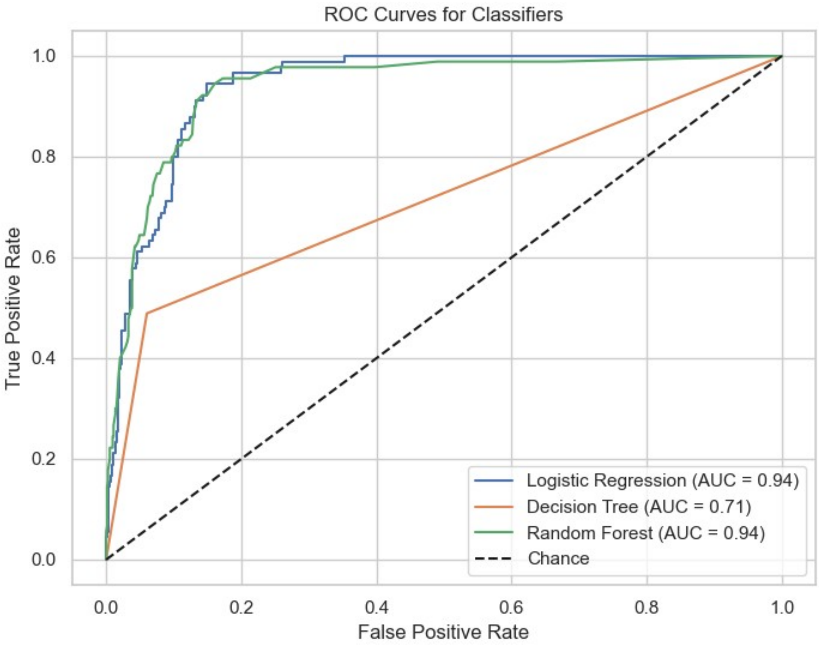
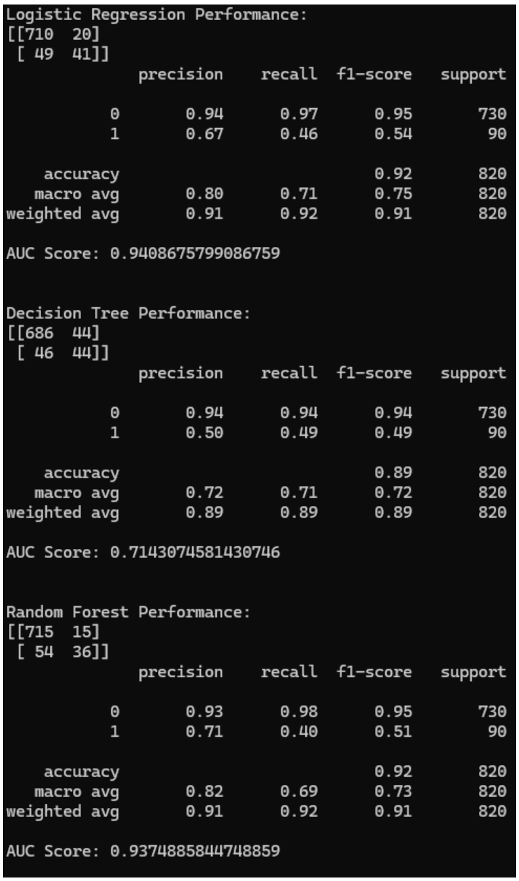


Figure 11: ROC Curves for Classifiers



```
Logistic Regression Performance:
[[710  20]
 [ 49  41]]
              precision    recall  f1-score   support

           0       0.94      0.97      0.95       730
           1       0.67      0.46      0.54        90

    accuracy                           0.92       820
   macro avg       0.80      0.71      0.75       820
weighted avg       0.91      0.92      0.91       820

AUC Score: 0.9408675799086759


Decision Tree Performance:
[[686  44]
 [ 46  44]]
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       730
           1       0.50      0.49      0.49        90

    accuracy                           0.89       820
   macro avg       0.72      0.71      0.72       820
weighted avg       0.89      0.89      0.89       820

AUC Score: 0.7143074581430746


Random Forest Performance:
[[715  15]
 [ 54  36]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95       730
           1       0.71      0.40      0.51        90

    accuracy                           0.92       820
   macro avg       0.82      0.69      0.73       820
weighted avg       0.91      0.92      0.91       820

AUC Score: 0.9374885844748859
```

Figure 12: Performances of the three chosen models