Predictive Analytics for N/LAB Platinum Deposit Marketing Campaign

A Classification Approach to Customer Targeting

Student ID: 20811152

Module Code: BUSI4371 – Foundational Business Analytics

Academic Year: 2025-2026

Submission Date: January 8, 2026

Word Count: 3152

AI Declaration
I declare that generative AI tools were used in the preparation of this coursework as follows:

Tools Used: Claude AI (Anthropic) and ChatGPT were used for code debugging, structuring report sections, and proofreading.

Nature of Use: AI assistance was used to refine code syntax and error handling, suggest visualisation improvements and provide feedback on report clarity. All analyses, model selection, interpretation, and business recommendations represent my own work and understanding.

Declaration: I understand that this work will be assessed as my own, and I take full responsibility for the content, conclusions, and recommendations presented herein.

EXECUTIVE SUMMARY

With the Platinum Deposit product offering existing clients who are ready to lock up money for a year a tempting fixed rate of interest, N/LAB Enterprises is hoping to expand its banking services. The primary means of acquisition has been determined to be telemarketing, although only a tiny percentage of the clients pursued end up subscribing. Thus, indiscriminate cold calling is incurring a high cost of operation for little actual gain. The purpose of the project will be to use the acquired data to devise a classification scheme which will assist in identifying customers likely to subscribe to minimize marketing efforts.

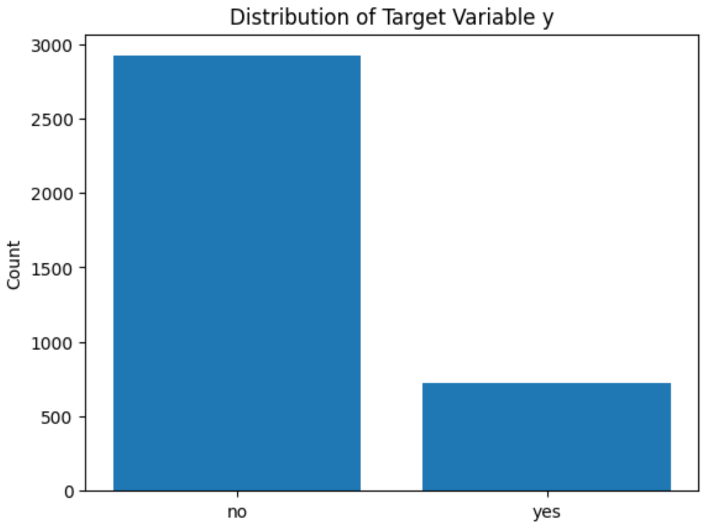
For this study, a dataset consisting of approximately 4,000 telemarketing transactions from a previous marketing campaign for a similar product is considered. In addition to a binary response variable for subscription success, it contains such variables as demographics, financial information, contact information, and information about previous contact with this marketing campaign. It is demonstrated that a highly imbalanced target response is represented, with roughly 80% for non-subscribers from initial statistical analysis. Due to such imbalance, performance assessment for models to be developed is essential to be done through different evaluation metrics than accuracy, such as recall and precision.

Stratified cross-validation and confusion-matrix-based metrics were used to evaluate several classification models: Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and k-Nearest Neighbours – are assessed through stratified cross-validation. In consideration of the business-related cost of unwanted telemarking calls, precision-specific measures of performance, in addition to conventional measures of recall and accuracy, are employed. Of the algorithms employed, it is found that the Decision Tree Classifier has the greatest compliance with business needs, in terms of precision-specific performance, having attained the highest possible F0.5 value.

SECTION A — Summarization

Dataset Overview

The data comes from historical records of telemarketing attempts made to promote a fixed term deposit scheme. A customer can appear more than once in the data, due to the number of attempts made to contact them, but every datum is for a separate attempt made on a potential customer. This is because, in the campaign viewpoint presented for the problem, every attempt is a separate datum since there is no distinct customer identity provided.



The dependent or target variable, y, is a dummy variable which is 1 if the customer enrols with the term deposit scheme after the call, and 0 otherwise. There are four major groups of predictor variables: customer information (age, employment, marital status, education level), customer financial information (account balance, housing loan, personal loan, default status), other information such as type of contact and day of the call, and previous call information which includes the number of calls, the outcome of the previous call, and the time of the previous call.

This extensive set of features allows for an in-depth assessment of client characteristics and behaviour, providing a strong foundation for predictive modelling.

Target Distribution

Analysis of the target variable denotes a very severe class imbalance where about 80.1% are non-subscribers and 19.9% subscribers. This will have important implications for model evaluation. A very simple classifier-a naive one-that predicts "no subscription" for every observation would achieve high accuracy but still would fail completely to identify the potential customers. Accuracy alone, therefore, is a poor performance measure in this context.

(Insert class imbalance bar chart here)

Numerical Feature Analysis

There are some numeric fields that have skewness and outliers, and these fields include account balance and the number of contacts for a given campaign. It can be observed that there is strong right skewness for the account balance variable; a few individuals have large account balances, and most people have balanced accounts themselves or have a negative balance.

The number of campaign contacts campaign_dma: The median value is low, inferring that a significant number of customers have had either one or two contact instances, and a few have had more than two contacts. The variables pertaining to the preceding contact (previous and pdays) show that a significant number of customers had not had a preceding contact, which underscores the prevalence of first-time contact instances.

Call Duration and Data Leakage

Call duration: Call duration shows a strong relationship with subscription outcome: successful calls tend to last significantly longer. This fact will likely relate to the higher levels of engagement of those customers who go on to subscribe. However, call duration is a variable that can only be measured after the call, and so it cannot be used for call target variables. Including call duration in any of the models would comprise data leakage. Call duration fails the first test of the benchmark report.

Categorical Feature Relationships

Subscription rates significantly vary across the different variables. For instance, customers who have pursued tertiary education have higher subscription rates compared to those who have only had primary or secondary education. This indicates that the level of education may act as a proxy variable for financial or investment knowledge. Additionally, the type of job also varies. Retired customers and students have higher subscription rates compared to others.

Indicators of financial commitment are also used as additional customer segmentation. People without residences or personal loans are more probable to consent to taking part in the survey, which means that customers with limited to no financial commitment are inclined to invest. A form of contact is also used because having cellular contact is generally correlated with a success rate that is more probable in unknown or unreliable forms of contacts.

Coupled with summarization, the phase suggests customer prior involvement, budgetary feasibility, and customer profiling as the central determinants of the subscription process.

Inter-Feature Relationships and Analytical Considerations

However, apart from checking for the nature of the relationship between individual predictor variables and the subscription variable, it is also important to consider interactions and dependencies of predictor variables. This is since some of the variables within the dataset are not independent. This is indicated by the fact that variables such as education level are related to employment type, and variables such as housing loan and personal loan approval are related to account balance and age.

The correlation analysis amongst numeric variables suggests that though no extreme multicollinearity exists, there are strong to moderate correlations between the campaign-related variables: campaign, previous, and pdays. Jointly, these features are supposed to describe the intensity and recency of marketing contact and hence convey complementary information, not redundant signals. These dependencies are worth noticing while choosing the modelling approach because linear models cannot handle such interactions without explicit feature engineering.

From a commercial perspective, these cross-correlations between features indicate the complexity whereby "responsiveness to the customer depends not simply on who the customer is, but also how they have behaved as a customer to the bank." This underscores the need for modelling solutions which address these potentially complex, non-linear, and cross-dependent phenomena, giving rise to an investigation into tree models and methodologies within the subsequent sections.

SECTION B — Exploration (Decision Tree)

A Decision Tree classifier is followed as an exploratory analytical tool to explore non-linear relationships and present interpretable customer segments. The key goal of the following analysis is not predictive accuracy but interpretability and the generation of insights.

The tree deliberately is constrained in-depth and minimum leaf size to avoid over-fitting, ensuring the resultant structure remains interpretable. Only variables available a priori before contacting a customer are included; call duration is excluded to avoid leakage.

Variable Importance and Structure

Analysis of the importance of features shows that variables with information regarding past interactions with campaigns are most prominent in decision-making. Specifically, information about previous campaigns (poutcome) repeatedly appears at or close to the root of decision trees, emphasizing its high importance for customer subscription. This confirms results from Section A, and again, customer engagement in the past has proved to be the most reliable predictor of customer behaviour.

Other significant variables are age, type of contact, and balance. Their corresponding variables are present in higher splits, and they offer significant reduction in impurity, which indicates they have an important role to play as far as separation between subscribers and non-subscribers is concerned. Some variables, like marital status and job class, have insignificant significance and, hence, insignificant discriminative capability as standalone variables.

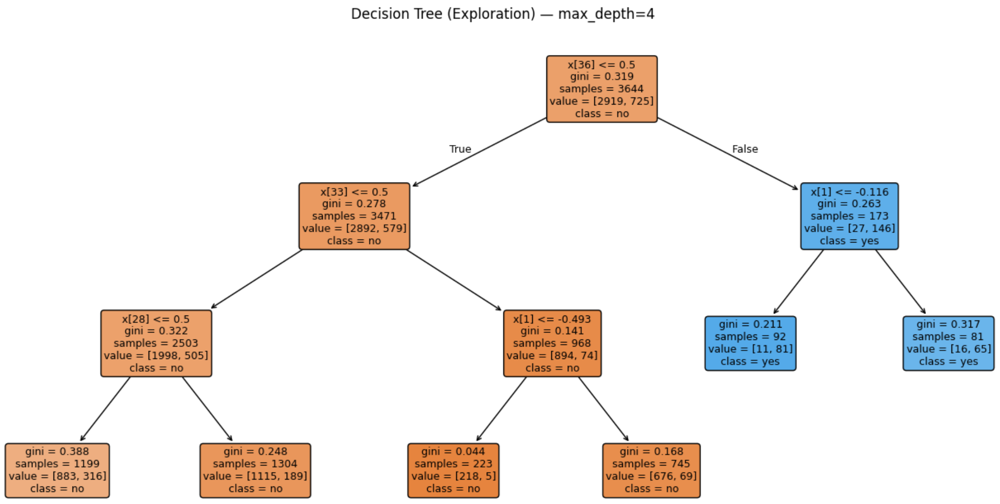
(Insert feature importance bar chart here)

Sub-Population Insights

The structure of the decision tree shows that there are different sub-populations of customers. The high-probability segment consists of customers who have had a successful campaign in the past. This segment needs very few additional conditions to predict a response. This is also an opportunity to market to them.

With the unconverted customers, the longer the customer the more valuable the customer information. Certain groups of working-age individuals have a greater chance of subscription using valid methods of contact, whereas the young or elderly customers with unknown contacts fall into the low-probability branches. The account balance distinguishes the customers with success and shows that financial capability plays a role in the chance of response even for the converted.

These rule-generating insights are rule-based and easy to understand, and they work in conjunction with statistical analysis to help build more complex prediction models.

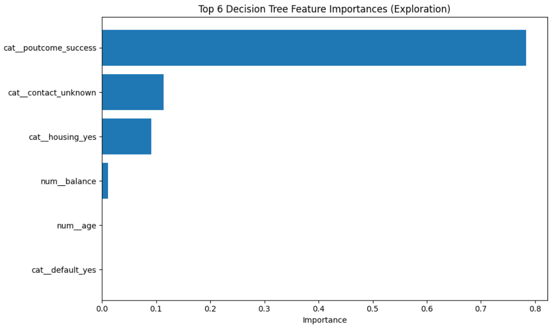


Limitations of the Decision Tree

Although the decision tree provides strong interpretability, there are some limitations in the decision tree classification method. The splitting criteria values are data specific and therefore lack generalizability. Moreover, the instability in single decision trees is high in the case of imbalanced datasets. Therefore, the decision tree method is only employed for exploratory purposes in this study. The predictive accuracy is then checked by other models in Section C.

Interpretability and Managerial Insight from Tree-Based Models

A major strength that exploring with decision trees uses is its interpretability. This is especially so because decision trees make decisions based on simple fractional or percentage criteria that are easy to communicate to nonspecialist audiences. An example is that "customers with prior successful outcomes on a campaign and fewer than two contacts in past two months are very likely to subscribe." This corresponds directly to intuitive models of consumer behaviour.



These rules can serve to test existing presumptions by managers against the patterns identified. Thus, for instance, the finding that persistent contact attempts are linked to lower rates of success for selected customer segments should alert managers to the possibility not merely that persistence improves outcomes but that disengagement with the service is taking place. Likewise, the interaction between the financial and prior engagement variables indicates that targeting must consider willingness and ability to invest.

It is significant to note, however, that these sets of rules are based on historical data and do not necessarily indicate determinism. Instead, these sets of rules form a guiding framework on which the customer behaviours can be interpreted and hypothesized. The reason why this is significant is for the purposes of making better-informed decisions in the real world, especially in the world of business.

SECTION C — Model Evaluation

Models and Rationale

Five machine learning classifiers will be tested against a baseline model, namely Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and k-Nearest Neighbors, commonly referred to as k-Nearest Neighbors (KNN). These machine learning classifiers were chosen for the sake of analytical diversity, testing a wide array of assumptions with distinct modes of learning.

Evaluation Strategy

Models will be evaluated using stratified 10-fold cross-validation, maintaining the same distribution of the original classes in each fold. It generalizes well for performance estimation. Confusion matrices will be drawn for each model, and then the performance will be gauged on precision, recall, and F1-score for the positive class ("yes"). Recall will be privileged, given the business objective of not missing subscribers; precision will also have attention to avoid too much unnecessary outreach.

Comparative Performance Analysis

The baseline model is not very good, and the high accuracy it scores is because of the class imbalance issue but otherwise completely fails to identify the subscribing customers. Logistic Regression performs better than the baseline model and scores better precision but lower recall value.

Decision Trees optimize the recall by recognizing the non-linear relations but incur more false positives, causing a reduction in precision. Naive Bayes classifier has moderate recall but has too many assumptions about independence, resulting in volatility related to probabilities with dependent variables. KNN classifier performs decently for the dominant class but fails to classify the subscribers within the high-dimensional spaces formed by the encoding of the categorical features.

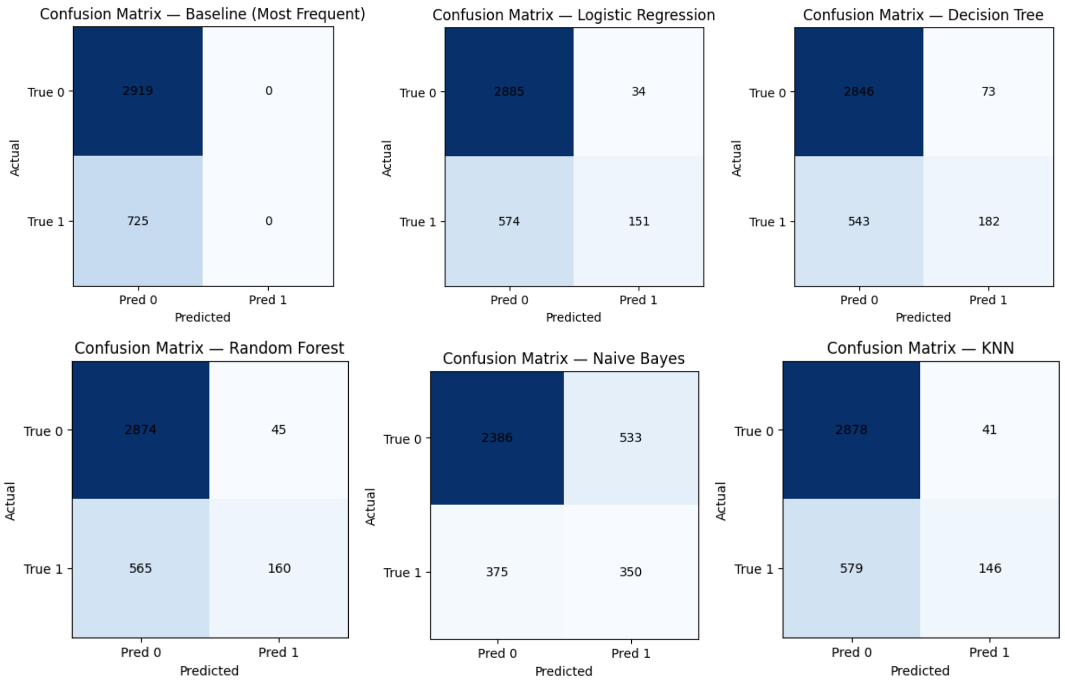
Although the Naive Bayes classifier had the best recall and F1-score, it also yielded a far larger number of false positives, which leads to many unproductive calls for telemarketing. Considering the additional Fβ metrics incorporates business cost explicitly. By giving a high priority to precision with the F0.5 score, the Decision Tree classifier achieved the best score, which indicates a better balance between identifying genuine subscribers and reducing unnecessary outreach to non-subscribers.

Explanation of Evaluation Metrics Justification

The choice of evaluation metrics is dependent on a function of a marketing campaign's purpose. The issue of accuracy is of little concern given the extent of class imbalance. There is a disregard for accuracy since it does not indicate how good the model is at pointing out prospective clients. Preference is given to recall.

The issue of precision is also taken into consideration since a high number of false positives would result in wasted marketing resources and increased operating costs. The F1 score is a measure of both requirements and will be used as the main evaluation metric for comparing the models. The confusion matrices will be analysed for the models to ensure a detail understanding of the misclassifications.

Though Naive Bayes had better recall and F1 measurement, it is to be noted that it had a larger number of false alarms. Regarding consideration of business costs, other options, which are Fβ scores, are to be evaluated. Stressing precision through F0.5 scores improved performance of the Decision Tree classifier.



SECTION D — Final Assessment

Based on performance metrics for cross-validation and specific focus on business expenses, the Decision Tree classifier emerges as the final solution. Although Naive Bayes performed better in recall value and F1 score, it was producing a much higher number of false positives, thus incurring a higher expense for unproductive telemarketing calls.

In consideration of the cost of telemarketing efforts and the need to allocate them effectively, a precision-weighted evaluation employing the F0.5 measure was used. In the context of the F0.5 measure, the Decision Tree model performed the best since it had the highest F0.5 measure, the highest precision, and the lowest number of missed subscribers.

This trade-off corresponds very well to the objectives N/LAB Enterprises would like to meet, since minimizing wasted outreach and customer fatigue is every bit as valuable to them as obtaining additional subscriptions. The interpretability and ease of implementation of the Decision Tree model are other reasons why it would be a good choice.

SECTION E — Model Implementation

After model selection, the Decision Tree classifier is then trained on the complete datasets (without calling duration) to maximize learning before deployment. The trained model is then saved as:

20811152_BUSI4371_2526_final_model. joblib

Detailed usage instructions are given to facilitate predictions on a new customer data set. The structure of a new data set will be identical to that of the training data, and the target and call duration fields will be removed from it if they exist. Predictions will be done through a predict_new() function that will provide binary predictions and probability scores.

SECTION F — Business Case Recommendations

Strategic Recommendations

N/LAB Enterprises would target its clientele through a multi-level targeting process that is tied to the possibilities that may exist within their subscriptions. The first people that N/LAB Enterprises would target would be those who have had a successful outcome from previous marketing processes and have a higher probability, and then they would selectively market to those who have a medium probability, and finally, they would eliminate those who have a low probability.

The number of contacts needs to be restricted to avoid fatigue in the campaigns, especially in the case of customers who have never been contacted before. Cellular contact methods need to be emphasized as much as is feasible due to the success rate associated with them.

Limitations and Future Opportunities

The analysis could be improved by its historical nature in that it may not capture current market dynamics and customer preferences. The analysis could also benefit from cost-sensitive analysis, threshold optimization, among other analysis types that could use customer transaction records, as well as credit details from other sources.

Operational Integration and Decision Support

In terms of deployment, it would be beneficial to integrate the predictive model into the existing process used by N/LAB Enterprises to manage campaigns. This would enable them to rank customers even before the start of each campaign. This would give them the ability to assign the contact process depending on the probability and return they are going to get. This would enable them to give priority to those with higher probabilities.

In particular, the threshold levels of the decision could be varied according to the type of campaigns as well as the resource capacity. In situations where the availability of human resource personnel is high, a low threshold level could be utilized to reach the highest possible numbers of students and patients. However, in situations where the resource capacity is low, a high threshold level might be necessary.

The forecasts produced from these models can also be monitored for drift over time with actual results to retrain these models. Likewise, this creates a cycle of how analytics contributes to optimization constantly.

In addition to the predictive power, other important aspects need to be pointed out. Too much focus must be given to the vulnerable and disengaged groups, which may cause dissatisfaction and reputation harm. In the future, the focus may be given to fair methods for this predictive task with data regarding the customer consent and preferences. Additionally, it would be feasible to maximize the profitability with a focus on the long-run horizon.

Using a precision-minded decision tree model enables it to reduce unwanted customer interactions, and that is very helpful for cost management and efficiency of campaigns.