A. Summarisation

A.1 Categorical input features

Categorical input variables were plotted in terms of outcome distribution to detect potential variations in outcomes across different groups.

Older people and young people aged 20-30 are more likely to buy the product. The same notion was also reflected in the target distribution by job, where students and retired people have significantly higher response rates. This is understandable as older people tend to make secure and predictable investments, while younger individuals may prefer term deposits due to limited knowledge about alternative investment options.

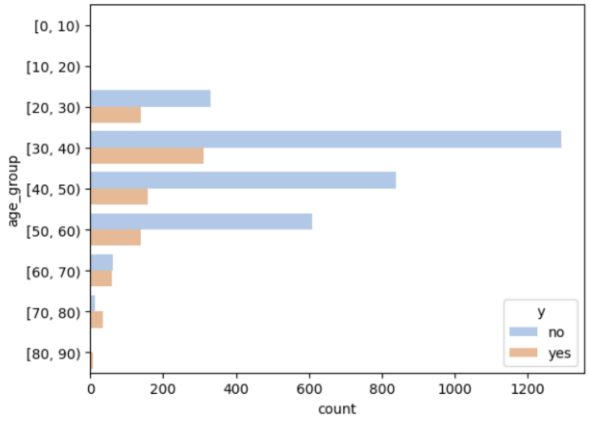

Figure 1: Target distribution by age

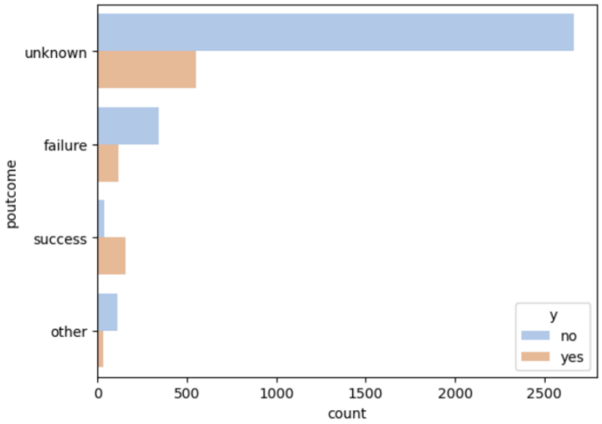

Figure 2: Target distribution by previous outcome

Default is also a notable factor in determining the outcome, as those with default credit seem to be less likely to buy the product. In addition, previous outcomes could be an important determinant of the target class. A call is more likely to be successful if the customer has said yes in a previous campaign, and vice versa.

On the other hand, the distribution of the target class seems to be the same across different marital status, education levels and loan.

A.2 Numerical input features

Numerical variables were plotted using a correlation matrix to examine whether there is any notable relationship between the input variables and output variable. Numeric features that are the most correlated with y are shown below:

Table 1: Features' correlation with y

Feature	Corr with y
duration	0.49
previous	0.13
pdays	0.12
campaign	-0.095
balance	0.068

Call duration has the most impact on whether a sale happens, which makes sense intuitively as the longer the call duration, the more likely a sale is. This information could be explored more in-depth later, for example through a decision tree to figure out a duration threshold, which if a call exceeds there is higher chance of making the sale. Other features with high correlation are previous, pdays and campaign. This suggests that actively following up with customers of the prior campaigns increases the likelihood of conversion.

B. Exploration

Two decision trees with different hyperparameters were applied to the dataset to examine influencing factors in the data.

In the first decision tree, duration offers the greatest predictive power in determining the likelihood of a successful sale. This observation aligns with intuitive expectations and is supported by the results in Section A, where duration exhibits the highest correlation with the outcome variable. In detail, the decision tree assigns a 55% chance of success to a sales call if the call exceeds 650 seconds. Meanwhile, calls lasting less than 396 seconds exhibit a higher likelihood of failure, with a 79% chance of an unsuccessful outcome.

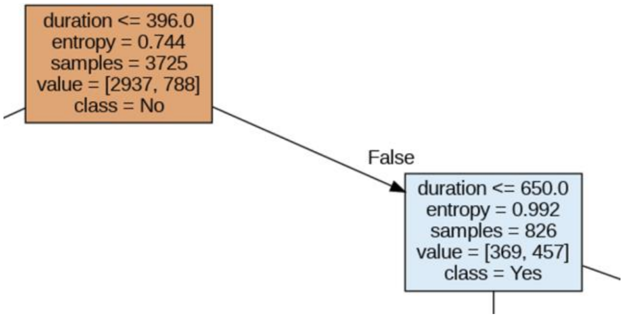

Figure 3: Snippet of the first decision tree

Previous contacts and age are also influential predictors. Notably, a moderate call duration (between 396 and 650 seconds), coupled with a recent contact (pdays <=

188.5), enhances the probability of success. In instances where no prior contact has been made, age emerges as a significant predictor of success.

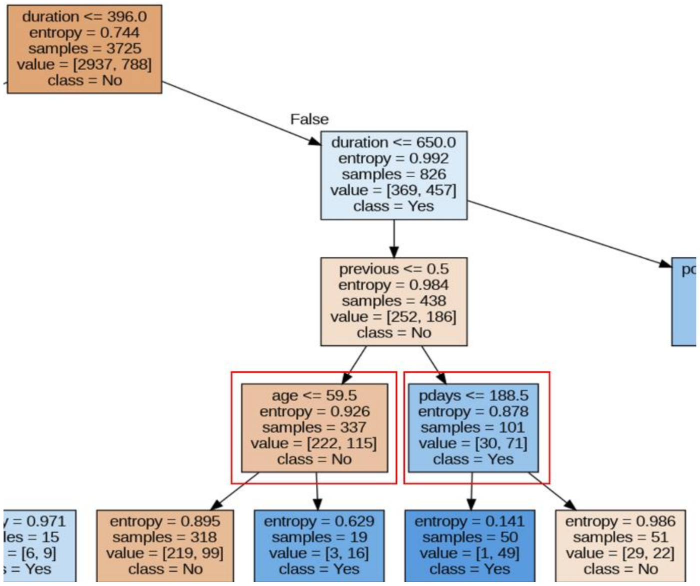

Figure 4: Snippet of the first decision tree

It is noteworthy that certain variables with relatively high correlations with the outcome variable, such as campaign and balance as well as certain job categories with high success rates, such as retired and students were not featured in the decision tree, which might suggest that these factors may have less importance in determining call outcomes compared to other variables. Similarly, variables such as marital, default, loan, and day are absent from the decision tree, implying their limited impact on predicting call success.

In the second decision tree, only duration and poutcome were featured. It is also very similar to the first second tree, with a longer call duration, or a moderate duration coupled with a successful outcome from the previous campaign increasing the likelihood of success in the current call.

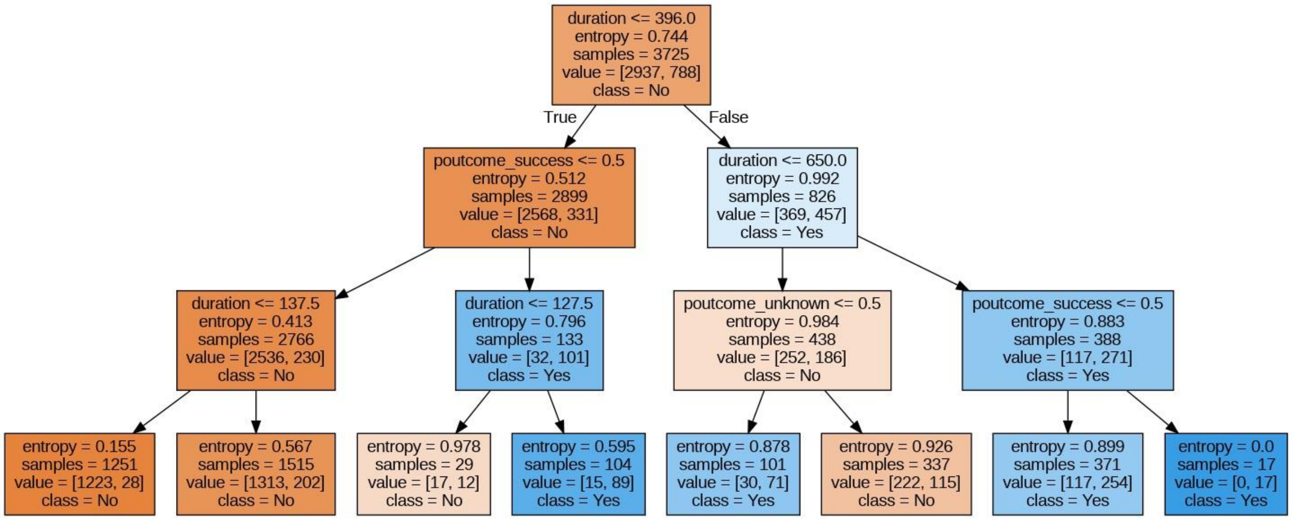

Figure 5: The second decision tree

The feature importance ranking for both trees also pointed out the importance of duration in cold calls for the new products, with its importance score being the highest, followed by poutcome.

The feature importance rankings for both decision trees again underscore the significance of the call duration. Duration consistently holds the highest importance score, indicating its substantial impact on the outcomes. Following closely in importance is the outcome of previous calls.

C. Model evaluation

The model evaluation workflow will be as follows:

1. Train-test split: Splitting 80% of the dataset to a training set and the remaining 20% to a testing set. As the dataset is highly imbalanced, a stratified train-test split was chosen to ensure that the train and testing sets have approximately the same proportion of each target class as the complete set.
2. Baseline model training: A dummy classifier is trained against the training set. It will classify everything as the majority class of the training data and act as a baseline for comparison with more sophisticated models.
3. Model deployment and tuning: Three classification models are deployed, namely Logistic Regression, Decision Tree and Random Forest. Hyperparameter tuning is performed on each model using the training data and Grid Search technique with 5-fold cross-validation (except for Random Forest with 4 folds for better computationality).

4. Model testing: The best-performing set of hyperparameters for each model, identified through Grid Search, will be tested against the hold-out testing set. This step ensures an unbiased assessment of model performance, preventing over-optimistic evaluations that may result from testing on the same dataset used for training and tuning.
5. Model evaluation: Different performance scores and confusion matrix will be computed and used to compare models. As N/LAB is more concerned about mistakenly contacting users who are not interested (False Positives) than missing potential customers (True Positives), it is crucial that when the model predicts a positive outcome, it is correct. Therefore, Precision was chosen as the main performance measure to compare models. However, models are also being compared against the dummy classifier in terms of its Accuracy score, to ensure a balance between minimising False Positives and maintaining overall prediction accuracy.
6. The best performing model will be deployed on the whole dataset

Each model's performance will be discussed more in-depth in the following sub-sections.

C.1 Dummy Classifier

When testing on the hold-out set, the Dummy Classifier has an Accuracy of 78.8%, Balanced Accuracy of 50% and Precision and Recall of 0%. This is because it classifies everything as the majority class, which is the unsuccessful outcome in this case.

C.2 Logistic Regression

This model was chosen due to its simplicity and computational efficiency. It was designed for binary classification and, therefore, goes well with the dataset, consisting of two different outcomes.

When testing on the hold-out set, the logistic regression classifier has an Accuracy of 84.3%, Precision of 69.5%, Recall of 46.2% and Balanced Accuracy of 70.4%. The logistic regression model is more accurate than the dummy classifier in predicting the outcome with a higher Accuracy score.

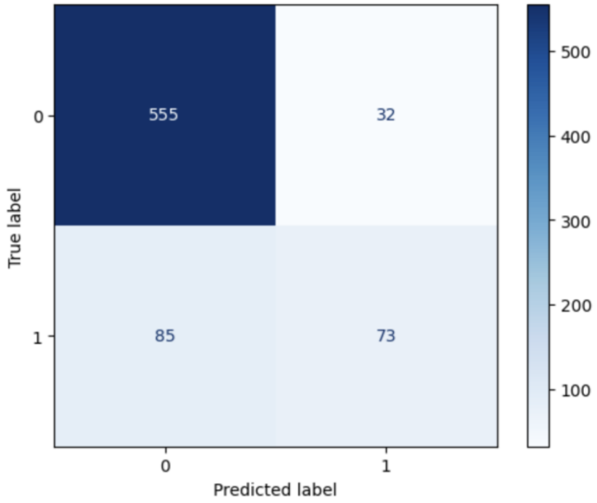

Figure 6: Confusion Matrix of Logistic Regression model

However, the logistic regression model still produces a considerable number of false positives (32 out of 105 positive target class predictions).

C.3 Decision Tree

Decision trees are easy to understand and interpret, and with proper pruning the model can overcome the problem of overfitting.

Upon testing on the hold-out set, the decision tree classifier has an Accuracy of 83.8%, Precision of 69.5%, Recall of 41.8% and Balanced Accuracy of 68.4%. Similar to the logistic regression model, this model is also more accurate than the dummy classifier with a higher Accuracy score.

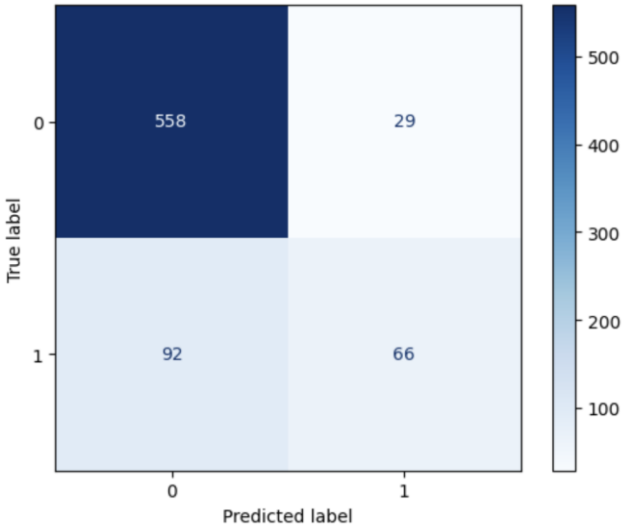

Figure 7: Confusion Matrix of Decision Tree model

C.4 Random Forest

Random Forests were chosen because they can model complex decision boundaries, while reducing overfitting in decision trees.

In the evaluation on the hold-out set, the random forest classifier has an Accuracy of 81.1%, Precision of 79.3%, Recall of 14.6% and Balanced Accuracy of 56.8%. Its Accuracy score is higher than that of the dummy classifier, showing its accuracy in predicting the target class.

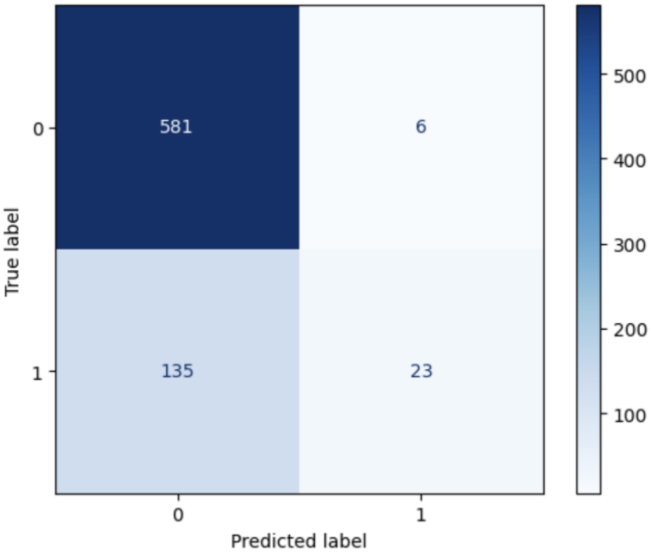

Figure 8: Confusion Matrix of Random Forest model

The recall score for this model is significantly lower than those of other models since it tends to predict y to be negative. As can be seen from the confusion matrix, the model predicted only 29 positive instances for y.

D. Final assessment

Random forest is the best performing model. Among all algorithms, random forest achieved the highest precision of about 79.3%, suggesting its strength in minimising false positives. This model has a drawback as it only predicts a small number of positive outputs. However, a large proportion of them are correct predictions, which meet the main goal of avoiding fruitless calls (false positives) of the classification task.

E. Model implementation

After being picked as the best performing model, the random forest model is then retrained on the whole dataset.

For deployment on new data, the entire code file should be rerun again, excluding certain sections related to statistical summary or exploration. Subsequently, the new hidden dataset can be uploaded, cleaned, encoded, classified, and evaluated by executing the pre-written code in the "Model Implementation" section of the file.

F. Business recommendations

Random forest was chosen as the most effective model after implementing and evaluating three distinct classification models on the cold call dataset for fixed-term deposits. Upon tuning, this model can help minimise contacts to customers who are not interested, avoid fruitless calls, and save cost and time for N/LAB. Using this model, the bank will be able to predict a customer's response to its telemarketing campaign before making a call. This strategic approach allows for a more targeted allocation of marketing efforts, concentrating on clients who are more likely to accept term deposits and reducing outreach to those less inclined.

To further enhance the effectiveness of the telemarketing campaign, N/LAB banking should implement the following:

1. Focus on younger (20 – 30 years old) and older (over 60 years old) age groups. Currently, N/LAB focuses most of its telemarketing effort on the middle age group (30-50 years old), which does not perform well compared to the younger and older age group. The older age group holds enormous potential and could be easier to convert as they most likely are already interested in this type of investment. Meanwhile, the younger age group might require persistent follow-ups for conversion.
2. Focus on customers who have been contacted from previous campaign. As seen from previous analysis, a customer is more likely to be converted if they have already been contacted from previous campaign. In addition, the more recent the contact is, the more likely of a sale. Therefore, the bank should prioritise customers contacted from previous campaigns and avoid prolonged gaps between contacts to maintain engagement.
3. Consider call duration. Call duration is the most important factor in determining whether a call is successful or not. The telemarketers should engage customers in longer calls, aiming for durations exceeding 10 minutes to foster meaningful interactions and potentially increase conversion rates.