NOTTINGHAM UNIVERSITY BUSINESS SCHOOL

Academic Year: 2020 – 2021

FOUNDATION BUSINESS ANALYTICS

Coursework

PREDICTING POTENTIAL CUSTOMERS FOR N/LAB PLATINUM PRODUCT

Student ID: 20243144

Jan 2021

CONTENT

A PROBLEM ANALYSIS

From the CEO message, there was 2 main business objectives to be translated into prediction parameters.

First, to address the concern regarding **missing potential customers** or l**osing business opportunity**, **'recall'** score, which represents the percentage of correctly detected subscribers among actual base should be taken into consideration.

Second, evaluating **'precision'** score that illustrates the proportion of actually accepted customers among the predicted subscribers will help to **avoid cost from fruitless calls** to uninterested people.

Due to the **imbalance between classes** in target output and the undefined cost for cold call and opportunity loss, it is important to separately consider 'precision and 'recall'. **'F1**', a summary indicator will also be analysed as indicator for trade-off cost between these two scores.

B SUMMARIZATION

To implement data exploration and analysis, all categorical features were converted into numerical form. (details in Section E: Implementation).

The transformed dataset then contained 35 input features, including 7 continuous variables and 28 discrete variables. Full converted elements are listed in **Table H.1 – Appendix.**

A heatmap based on features' relationship was plotted with key features selected in **Figure B.1** below (full variables Heatmap in **Figure H.2 - Appendix**).
Overall, there was **no strong correlation between input factors and the target outcome**, meaning that no single information from customer's background firmly decide whether they accept the product offer or not.

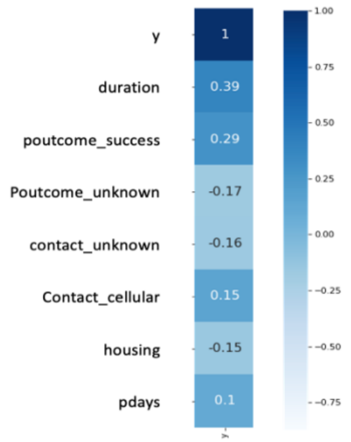


Figure B.1. Shortened heat map

However, slight to medium impact (correllation rate from 0.15 to 0.3) from input features towards the output were observed which can be translated into business insights:

- Longer **call duration** for this campaign indicated a higher chance that customers will accept the offer.
- Customers that had **accepted a previous offer** were more likely to accept this time.
- **Cellular** communication seemed to be quite effective in convincing customers, while **undefined contact** could lead to failure outcome.
- People with a **housing dept** were more likely to reject the offer.

Excluding the plain relationship between features extracted from same category (for example, 'contact_cellular' and 'contact_unknown', 'marital_single' and 'marital_married'), and between demographic variables (for example, higher education link to management job, age with retirement and marital status), **most input variables appeared to be independent and uncorrelated**.

An *exception was that the* ***failure result of last offer*** *('poutcome_failure') had no correlation with the* ***outcome*** *in this campaign ('y'), but highly correlated (0.71) with the* ***number of days from most recent contact*** *('pdays').*

? This requires validation on whether the sales team had a misfocused strategy, ignoring the rejectors from last campaign, or simply because halted communication led to unsuccessful offer?

C EXPLORATION

As the transformed dataset included a large number of variables, a decision tree was applied to understand which features have significant impact to output.

Tree's structure exploration

The two Decision trees (with and without 'duration' variable) are illustrated by **Figure H.3, H.4 in Appendix** respectively. To avoid overfit, samples' leaf that are smaller than 30 were not considered in analysis.
It can be seen that he information gained from Decision Tree were highly consistent with observation from Section B.
From **Figure H.3**, call duration ('duration') appeared to be the most important factor for classification. Customer having **call longer than 13 minutes** (787.5 seconds to be specific) were more potential target while **call under 3 minutes** (95.5 seconds) highly suggested offer's rejection (99.5% people declinced).
Excluding 'duration' from input data as this variable only comes after customer's contact, the critical features' combination that led to different classes were previous offer's acceptance ('poutcome_success'), and housing loan ('housing').

The **successful result of previous campaign** ('poutcome_success') was the key factor in classifying output feature.

Customers that had not accepted in previous campaign will have 90.5% chance rejecting this time:

- The **contact communication type** and below components in this branch were not striking elements for classification. However, customers with unknown contact ('contact_unknown') were more likely to reject the offer.

Among customers accepted in previous campaign:

- **Housing loan** ('housing') provided **further split** in this group, indicating that clients without a housing dept would be a better target group (70% accepted the offer, compared to 40% in group with housing loan).
- Next feature which is the number of days from last contact ('pdays') played insignificant role in splitting the classes.

Features' importance

Examining feature importance will provide a more factual and comprehensive look on the role of elements in classifying the result. This step can be done by calculating the Total reduction in Decision Tree's Node Impurity. Since unpruned tree could result in overfitting with unlimited depth and therefore giving bias to continuous variables as they had more chances to appear in split points, a tuning model (in this case, GridSearchCV) was applied in advance to detect the most optimal tree pattern.

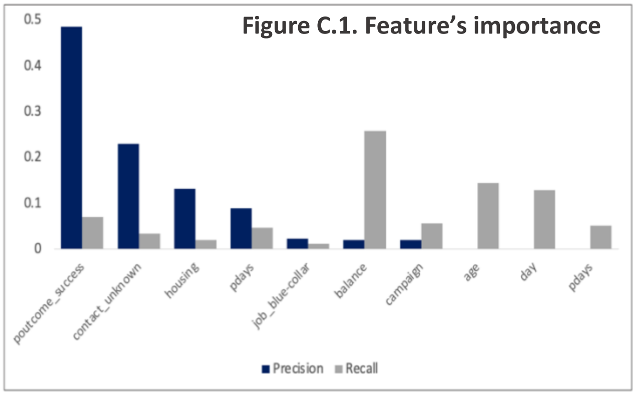


Figure C.1. Feature's importance

One of the key elements to consider before tuning is the optimisation target ('scoring'), which includes model evaluation metrics. Applying 2 considered parameters ('precision', 'recall') one by one results in below findings:

In utilising 'precision': Features' importance ranking **highly resembled** their position in **correlation map** (Section B), but with a more optimised approach that excluding highly correlated input variables (for example, using 'poutcome_success' and excluding 'poutcome_unknown'). However, versus the combination of 'poutcome_success' and 'housing', the presence of other features plays insignificant role in improving precision (0.705 versus 0.696) and deliver same accuracy (0.897%), which was similar with the observation from Tree structure.

In utilising 'recall': More features involved in classification if the objective is to find as many potential customers as possible. However, **expensive trade-off** in 'precision' and accuracy was required ('recall'=0.267 while 'precision'=0.252)

The details of tuning models and Feature ranking are included in **Table H.5, H.6 – Appendix**.

D MODEL EVALUATION

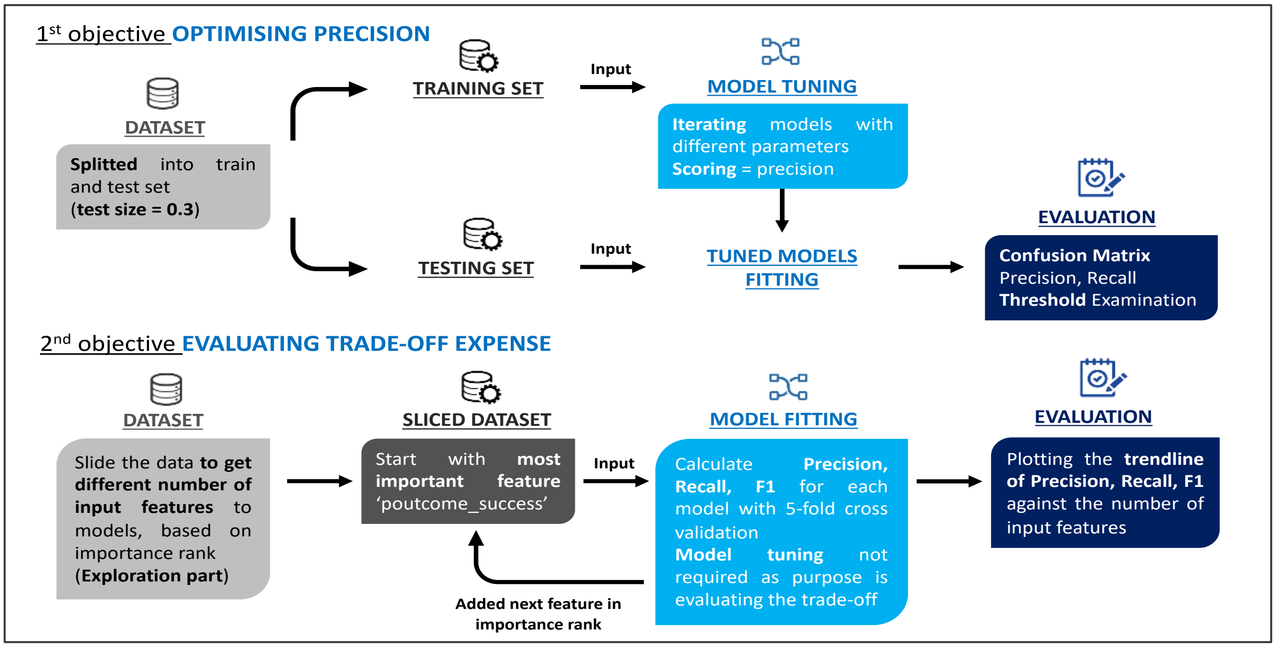
Evaluation strategy

From the Exploration, it can be seen that different prediction objectives result in different ways of building model. As such, there is no one-size-fit all approach to satisfy both business objectives at the same time.

Having said that, one strategy that could be useful in cost and benefit optimisation is to **implement a two-stage campaign**. The first one focuses on spending utilisation, customer exploration and most importantly the capital size gained from each subscriber. While in the second phase, after the cold-call and opportunity costs are defined, a specific target on the 'recall' and 'precision' rates can be set to maximise the profit (more details in Section F).

The evaluation strategy, therefore, will be set in accordance with above procedure, considering both the models' capacity to **achieve high 'precision' in the first stage**, and the flexibility to **deliver targeted 'recall' without losing significantly the 'precision' later on**. As 'recall' rate tent to increase when there are more input fed into modelling, an approach of increase variables as input (following their important ranking in precision utilisation) will be applied to observe the trade-off cost.

Flow chart representing the evaluation process can be found in Figure D.1 below.



Splitted into train and test set (**test size = 0.3**)

Iterating models with different parameters
Scoring = precision

Confusion Matrix
Precision, Recall
Threshold Examination

Slide the data **to get different number of input features** to models, based on importance rank (**Exploration part**)

Start with **most important feature** 'poutcome_success'

Calculate **Precision, Recall, F1** for each model with 5-fold cross validation
Model tuning not required as purpose is evaluating the trade-off

Plotting the **trendline of Precision, Recall, F1** against the number of input features

Figure D.1. Flow chart for evaluation process

Five different models including k-Nearest Neighbour (kNN), Random Forest (RF), Decision Tree (CT), Logistics Regression (LR) and Naïve Bayes (NB) were taken into consideration and comparison to find the best performed method. A Point Model (DM) using 'most frequent' strategy was also included as benchmark for classification performance due to the imbalance between classes in output feature.

Precision Optimisation

Input features for this stage includes **'poutcome_success'** and **'housing'** as they are the most important combination in utilising 'precision' as discussed in Section C.

The optimised parameters for each model using an exhaustive search (model iteration) method (GridSearchCV) are as follows:

- kNN: number of neighbours = 7
- RF: number of trees = 5, maximum depth = 2
- CT: maximum depth = 2
- NB: Categorical Naïve Bayes

Results of the Confusion Matrix from tuned models against the test set was illustrated in below figure.

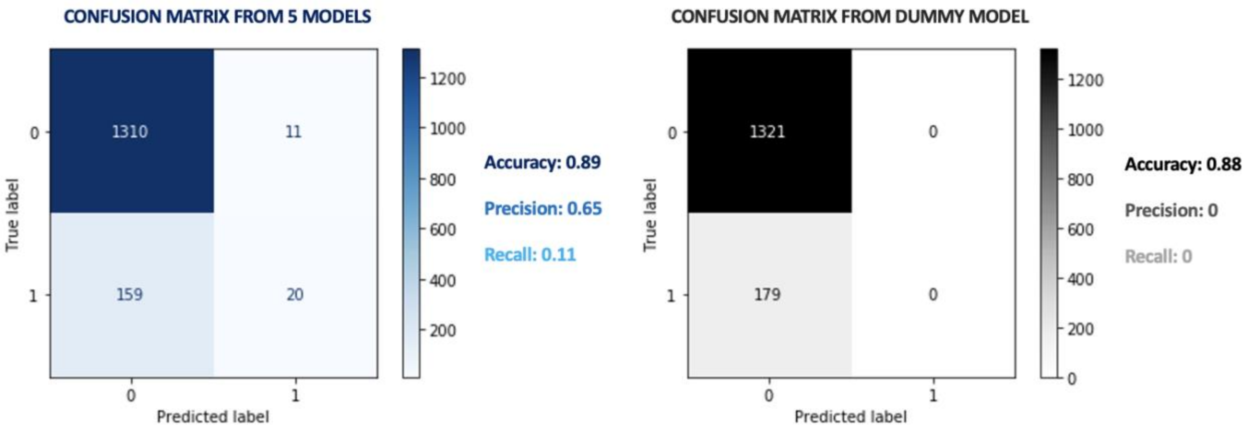
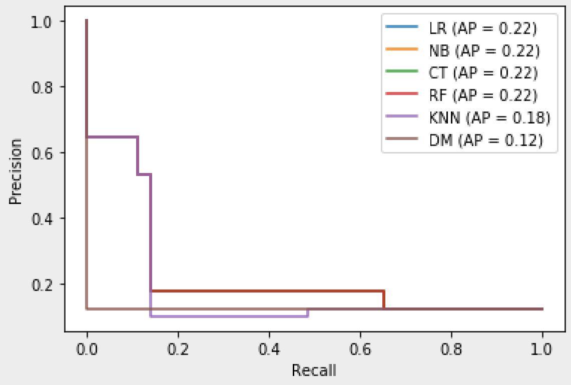


Figure D.2. Confusion Matrix for Test set prediction

Overall, it can be seen that with **proper feature deduction** and **parameters adjustment**, all models achieved **same classification performance** on test sample and overperformed benchmark (DM) model in all indicators. Therefore, the model selection would highly depend on the capacity to utilise trade-off cost.

Trade-off cost Evaluation

Figure D.3. Recall-Precision curve



Quick look on Trade-off cost with feature reduction

'Precision' and 'recall' curve by threshold's changes was also examined (Figure D.3) with the combination input of 'poutcome_success' and 'housing'. However, models tent to behave similarly in terms of trade-off cost, except for lower performance in terms of Average Precision from kNN (AP=0.18) and DM (AP=0.12).

Since kNN was the most computationally expensive estimator that requires model's adjustment and neighbors' number specified with every change in inputs, it was excluded in this trade-off cost evaluation. Another note is that NB model used in this section was changed to a different type (Gaussian) that can work with continuous variables.

Plotting the 'precision', 'recall' and 'f1' scores of each model **against the number of input features** that follow the importance ranking generated in Section B (full table in submitted file) helps open up findings as below:

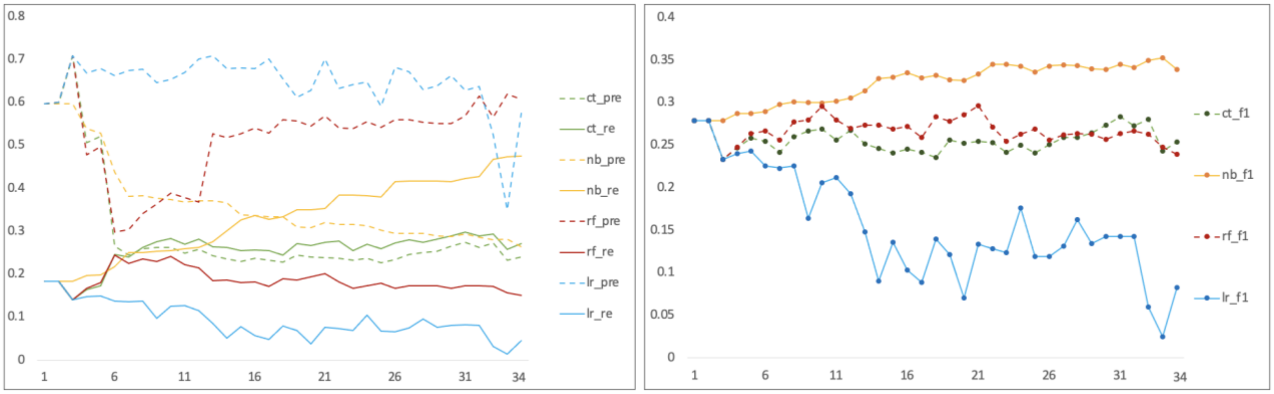


Figure D.4. 'Recall', 'Precision and 'f1' with increments in input number

NB appeared to deliver the most balanced scores and least opportunity cost between 'precision' and 'recall' (highest in 'f1'). It was also the model to achieve siginificantly higher recall rate when acquiring more variables to its pocket, hence was generally **better in covering the opportunity size**.

LR provided consistenly high 'precision' scores regardless neither the number of inputs nor the requirement for tuning. However, the 'recall' rate remains the lowest among all model, indicating higher chance for missing out potential customers.

CT and **RF** can be tuned (just one time and without requiring in advance feature reduction stage) to achieve high 'recall' or 'precision', but each option results in an expensive trade-off in the other parameter. Overall, **RF** showed more stable performance versus **CT** under large number of inputs.

E FINAL ASSESSMENT

Regarding technical side, as previously discussed, it can be seen that Naïve Bayes was the only estimator having recall rate improved without a significant trade-off in precision. This was because the input features from given dataset are mostly independent. Therefore, Naïve Bayes with its probability calculation method delivered balance between overfitting and underfitting, thus performed exceptionally well compared to other models.

Furthermore, the model required no complicated and time-consuming tuning method each time a new variable added, thus, can be easily adjusted to achieve targeted opportunity size and cost optimisation.

On business perspective, while the cost for cold call and benefit gained from submitted clients are currently unknown and could change by economy of scale, Naïve Bayes should be selected to allow flexibility in choosing the right model for each strategy stage.

F MODEL IMPLEMENTATION

The detailed instruction and execution steps in applying Naïve Bayes model to the whole dataset were included in the 'Final Model' file. Overall, the key steps include:

Data Loading and Processing

Categorical variables are transformed into numerical format:

- **Binary variables:** Assigning number to nominal values, with 'No' replaced by 0, 'Yes' replaced by 1.
- **Multi-class variables:** Using LabelBinarizer transform method to separate each class to a binary label with 1 means the sample belong to the class and 0 means the sample does not belong to the class.

Model fitting and prediction

The dataset is split based on requirement of each project stage before fitting to model:

- **Project stage 1:** Only include **'poutcome_success'** and **'housing'** in Input Dataframe to optimise 'precision'. The Input and Output data are then fitted to NB model to be ready for prediction and evaluation.
- **Project stage 2:** A trade-off table between precision and recall is created (detail in Jupyter file) with the increment of input features. The gained benefit can be calculated as:

Total gain = (Benefit per subscriber * True Positive) – (Cost per call * Predicted Positive)

The new option of features can be then re-input in Project stage 1 for prediction if the combination in Stage 1 is not the optimal.

G BUSINESS CASE RECOMMENDATION

Project execution

In the first stage of the project, a Naïve Bayes model with highest precision score will be applied to avoid fruitless cost and understand the benefit size. Only **a part of customer database** should be used in this stage using stratified random sampling method.

For **opportunity sizing purpose**, a new data field of the associated profit should be added for accepted clients. Depend on database's size, new classification labels such as 'super customers' or 'sizeable customers' could be created to provide more insights.

Follow-up diagnosis by economical surveying to understand the reasons behind 'false positive' batch (people accepting last offer and having no house loan but decline subscribing) can be helpful in finding other critical variables or improving the product.

In the latter stage, with the involvement of defined cost and benefit, **total gain calculation** can be applied to identify which level of Precision and Recall rate is the most profitable and relevant to business, followed by an appropriate Naïve Bayes model.

Other recommendation

The most likely people to subscribe for new product were those that had **previously accepted an offer** ('poutcome_success'), had **no house loan** ('housing') and had **been contacted** ('pdays') in last 3 months (96 days). It showed that while customers' background was the striking element to the offer result, **maintaining engagement** also played an important role in attracting customers, which can be done economically by texting or emailing.

96% of the acceptance cases received 5 calls or less for this campaign ('campaign'), indicating that people tend to accept if they have interest, without requiring numerous follow up. Therefore, **a cap on contact trial** can be applied to avoid time consuming.

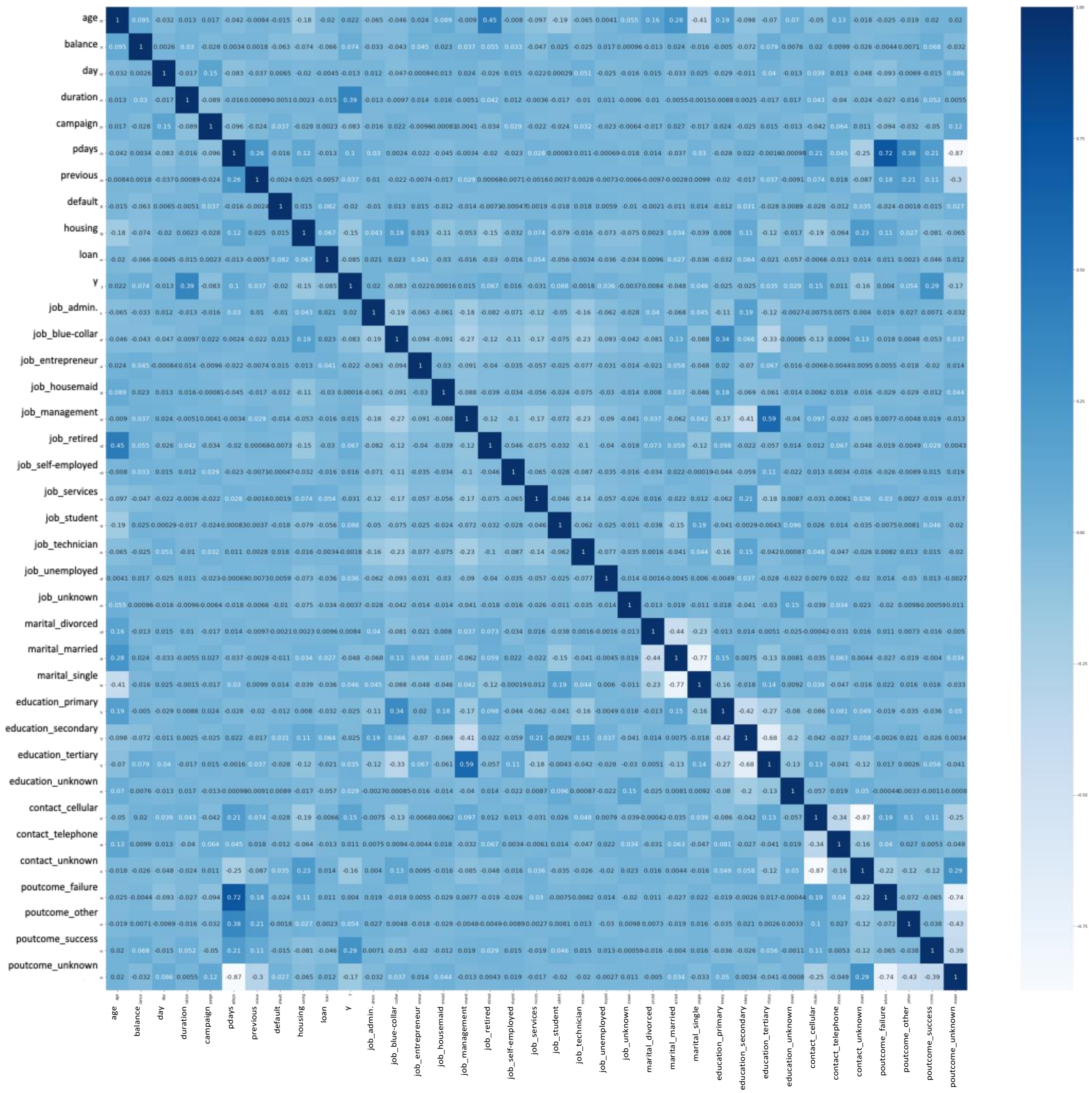
H APPENDIX

Table H.1. Variables after transformation

Type	Name	Meaning	Transformed
Output	‘y’	Offer result	‘y’ (binary)
Input	‘age’	Individuals’ age	'age' (continuous)
	'balance'	Individual’s current bank balance	'balance' (continuous)
	'day'	Day of the month from last contact	'day' (continuous)
	'duration'	Last contact duration	'duration' (continuous)
	'campaign'	Number of contact this campaign	'campaign' (continuous)
	'pdays'	No. of days from last contact	'pdays' (continuous)
	'previous'	Number of contacts before this campaign	'previous' (continuous)
	'loan'	Having personal loan or not	'loan' (binary)
	‘default’	Having default or not	'default' (binary)
	‘housing’	Having house loan or not	'housing' (binary)
	‘job’	Individuals’ job	'job_admin' (binary)
			'job_blue-collar' (binary)
			'job_entrepreneur' (binary)
			'job_housemaid' (binary)
			'job_management' (binary)
			'job_retired' (binary)
			'job_self-employed' (binary)
			'job_services' (binary)
			'job_student' (binary)

			'job_technician' (binary)
			'job_unemployed' (binary)
			'job_unknown' (binary)
	‘marital’	Individuals’ marital status	'marital_divorced' (binary)
			'marital_married' (binary)
			'marital_single' (binary)
	‘education	Individuals’ education	'education_primary' (binary)
			'education_secondary' (binary)
			'education_tertiary' (binary)
			'education_unknown' (binary)
	‘contact’	Type of communication	'contact_cellular' (binary)
			'contact_telephone' (binary)
			'contact_unknown' (binary)
	‘poutcome’	Result of previous campaign	'poutcome_failure' (binary)
			'poutcome_other' (binary)
			'poutcome_success' (binary)

Figure H.2. Full correlation table



Reading the tree:

- *For binary variables, <0.5 means =0 (not happen) and >0.5 mean =1 (happen)*
- *Class: classification result for that sample leaf*
- *Right node represents 'No', Left node represents 'Yes' answer for leaf's condition*

Figure H.3. Decision Tree plot – With 'duration'

(Depth reduced till different class split)

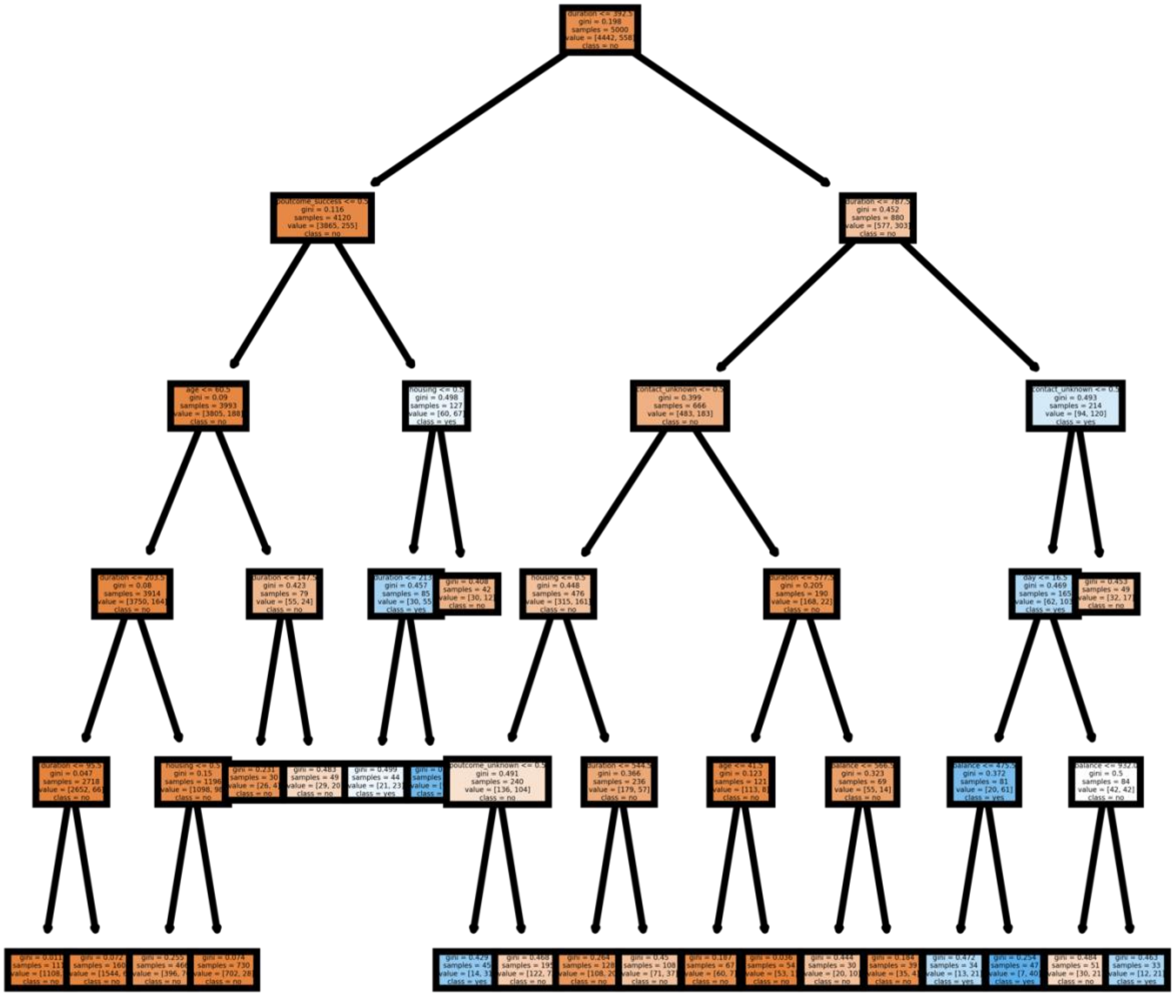


Figure H.4. Decision Tree plot – Without ‘duration’

(Depth reduced till different class split)

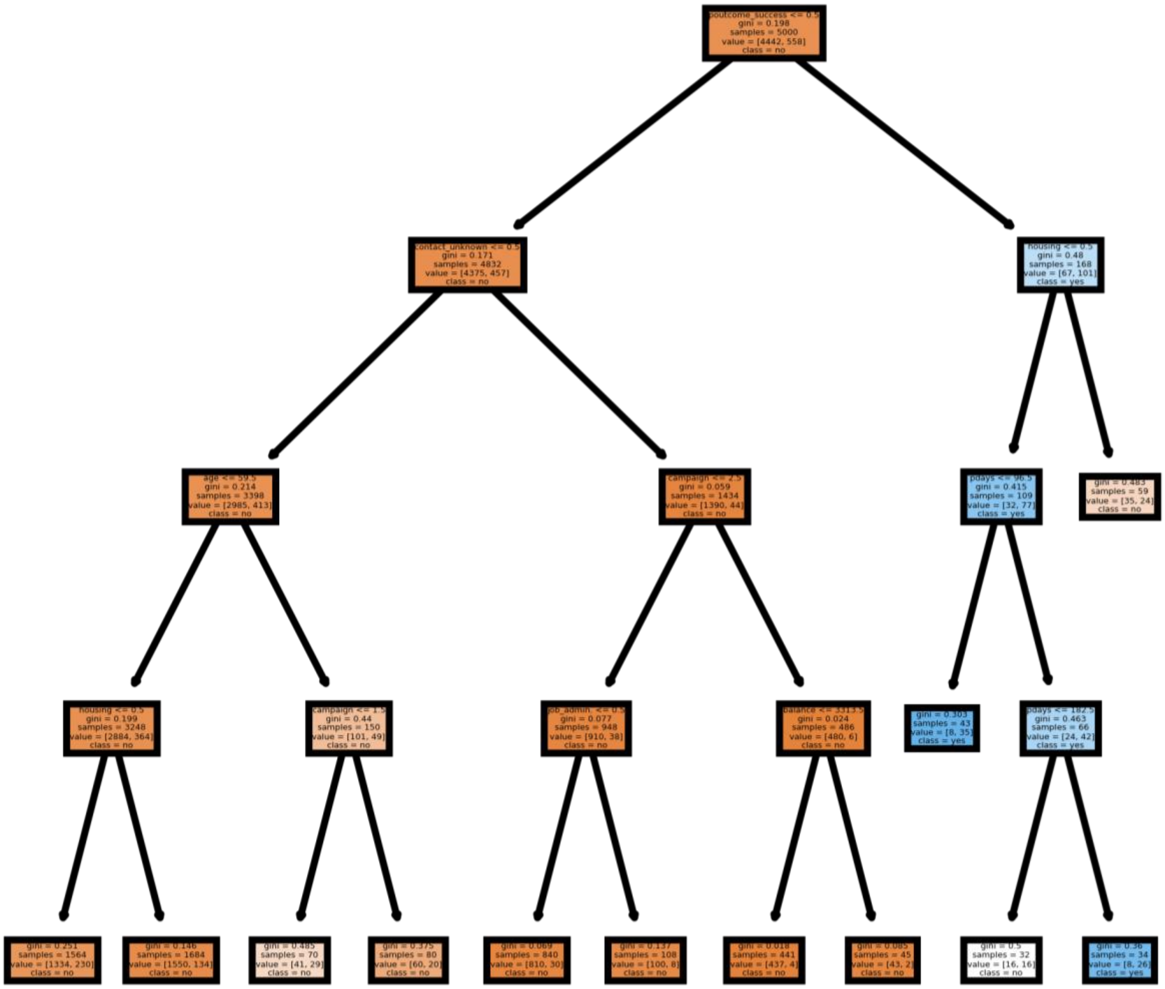


Table H.5. Tuned Decision Tree model specification

Optimising ‘precision ‘	{'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 26}
Optimising ‘recall’	{'criterion': 'entropy', 'max_depth': 28, 'min_samples_leaf': 1}

Figure H.6. Feature importance ranking

By recall	**%Importance**	**By precision**	**%Importance**
balance	0.258503	**poutcome_success**	0.485918
age	0.144725	**contact_unknown**	0.229976
day	0.129438	**housing**	0.131605
poutcome_success	0.071390	**pdays**	0.089199
campaign	0.056302	**job_blue-collar**	0.023103
pdays	0.051503	**balance**	0.020598
contact_unknown	0.033788	**campaign**	0.019599
housing	0.020679	**education_tertiary**	0.000000
education_tertiary	0.018390	**marital_married**	0.000000
previous	0.017779	**marital_single**	0.000000
job_admin.	0.017752	**education_primary**	0.000000
marital_single	0.016313	**education_secondary**	0.000000
marital_divorced	0.014595	**age**	0.000000
education_secondary	0.013958	**education_unknown**	0.000000
job_management	0.012529	**contact_cellular**	0.000000
job_services	0.012414	**job_unknown**	0.000000

education_primary	0.012366	**contact_telephone**	0.000000
job_technician	0.012332	**poutcome_failure**	0.000000
job_blue-collar	0.011666	**poutcome_other**	0.000000
job_self-employed	0.010617	**marital_divorced**	0.000000
loan	0.010116	**job_student**	0.000000
education_unknown	0.007456	**job_unemployed**	0.000000
job_student	0.007342	**job_technician**	0.000000
marital_married	0.006606	**job_services**	0.000000
poutcome_other	0.005036	**job_self-employed**	0.000000
default	0.004801	**job_retired**	0.000000
job_unemployed	0.004050	**job_management**	0.000000
job_retired	0.004045	**job_housemaid**	0.000000
job_housemaid	0.003684	**job_entrepreneur**	0.000000
contact_telephone	0.003192	**job_admin.**	0.000000
job_entrepreneur	0.003015	**loan**	0.000000
contact_cellular	0.002168	**default**	0.000000
poutcome_failure	0.000759	**previous**	0.000000
job_unknown	0.000690	**day**	0.000000
poutcome_unknown	0.000000	**poutcome_unknown**	0.000000