Bank Marketing Analysis Report
Part 1

Gabriela Almeida Monteiro - s5198626
Jason Dias - s5216366
Julio Pimentel Albores – s5172620

Griffith University
7031ICT Applied Data Mining
Lecturer: Dr Can Wang
Assessment item: Written Assignment

Word count: 2,147
Due date: 03/09/2021

1. INTRODUCTION

1.1. Brief description of the problem

A Portuguese bank was experiencing a decline in its revenues, and after some investigation, it was found that customers were not making enough term deposits (Maity, 2020). Financial institutions' significant income source comes from term deposits, which are fixed-term investments where investors can withdraw their money only after the term ends (Investopedia, 2021). Term deposits allow banks to use the capital for other investments and make a profit.

Following the findings, the Portuguese bank decided to perform direct marketing campaigns by phone calls to persuade their customers to subscribe to a term deposit (Moro et al., 2014). This project aims to build a classifier that can help the bank correctly predict whether a customer will subscribe to a term deposit based on some customer's features. This information can be helpful during the marketing campaign to target new customers that could potentially have a 'yes' response. On the other hand, it can also be beneficial to explore strategies to target the customers who answered 'no'.

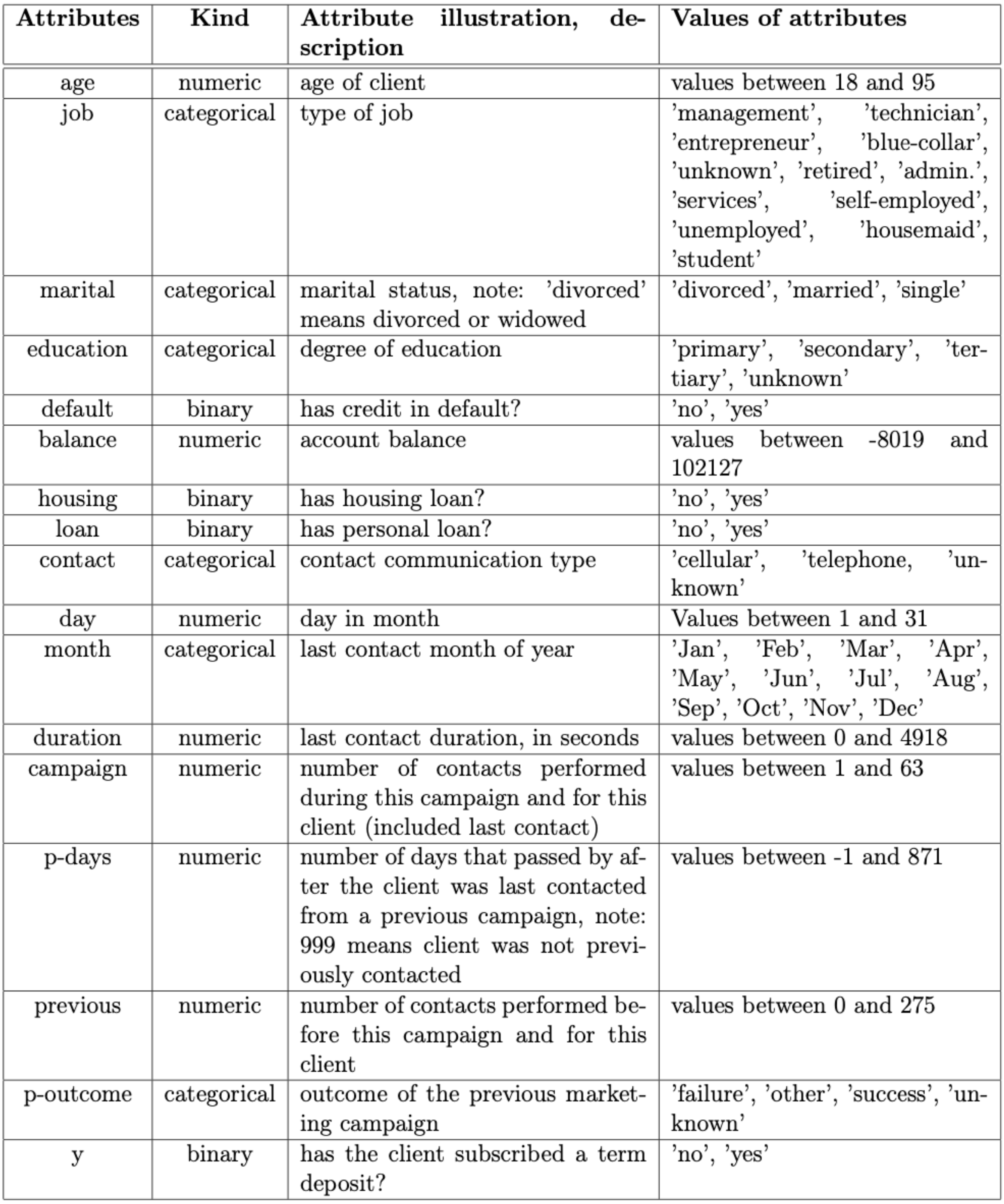
1.2. What data will you use, and where will you get it?

The data was extracted from the UCI Machine Learning Repository[1]. There were four datasets available in this repository, and the one chosen for this analysis was the "bank-full.csv". This dataset was chosen because it is like the older version but with fewer columns.

Table 1 shows all attributes present in the data as well as their definition. For example, the attribute 'y' is the customer's response (yes or no) to a term deposit. The project's hypothesis is that attribute 'y' can be predicted based on the other 16 attributes in the dataset. Therefore, in this analysis, we will treat attribute 'y' as the target attribute, whereas the other 16 features are assumed to be the input variables (or predictors of 'y').

[1] The dataset can be accessed here: https://archive.ics.uci.edu/ml/datasets/bank+marketing

Table 1. Dataset attributes and description.

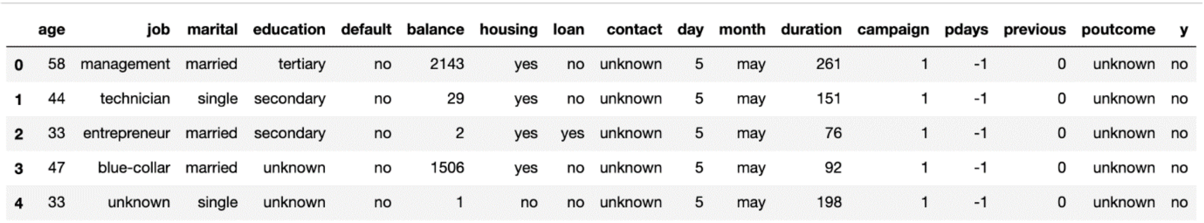


Attributes	Kind	Attribute illustration, description	Values of attributes
age	numeric	age of client	values between 18 and 95
job	categorical	type of job	’management’, ’technician’, ’entrepreneur’, ’blue-collar’, ’unknown’, ’retired’, ’admin.’, ’services’, ’self-employed’, ’unemployed’, ’housemaid’, ’student’
marital	categorical	marital status, note: ’divorced’ means divorced or widowed	’divorced’, ’married’, ’single’
education	categorical	degree of education	’primary’, ’secondary’, ’tertiary’, ’unknown’
default	binary	has credit in default?	’no’, ’yes’
balance	numeric	account balance	values between -8019 and 102127
housing	binary	has housing loan?	’no’, ’yes’
loan	binary	has personal loan?	’no’, ’yes’
contact	categorical	contact communication type	’cellular’, ’telephone, ’unknown’
day	numeric	day in month	Values between 1 and 31
month	categorical	last contact month of year	’Jan’, ’Feb’, ’Mar’, ’Apr’, ’May’, ’Jun’, ’Jul’, ’Aug’, ’Sep’, ’Oct’, ’Nov’, ’Dec’
duration	numeric	last contact duration, in seconds	values between 0 and 4918
campaign	numeric	number of contacts performed during this campaign and for this client (included last contact)	values between 1 and 63
p-days	numeric	number of days that passed by after the client was last contacted from a previous campaign, note: 999 means client was not previously contacted	values between -1 and 871
previous	numeric	number of contacts performed before this campaign and for this client	values between 0 and 275
p-outcome	categorical	outcome of the previous marketing campaign	’failure’, ’other’, ’success’, ’unknown’
y	binary	has the client subscribed a term deposit?	’no’, ’yes’

Source: Włodarczyk and Ikani, 2020.

Table 2 shows the output for the head method, the first 5 observations of the datasets, including the values for all columns.

Table 2. 17 features and top 5 observations



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

2. METHODOLOGY

2.1. Proposed algorithms and techniques

For this case of study, there is a clear associated response (called target feature), and this response is categorical because it can be ‘yes’ (the customer will subscribe to the term deposit) or ‘no’ (the customer will not subscribe). According to James et al. (2013), problems where “for each observation of the predictor measurement(s) (…) there is an associated response to the predictors” (p. 26) are considered supervised statistical learning problems. They also state that problems with “qualitative response are often referred to as classification problems” (James et al., 2013, p. 28). Therefore, this statistical learning problem falls into supervised (clear output) and classification (categorical output). This project is a binary classification problem because the observations belong to two classes: yes or no.

The algorithms are chosen to deal with this type of statistical learning problem and help predict whether a customer will subscribe to a term deposit or not are logistic regression, random forest, and support vector machine.

2.1.1. Logistic Regression:

Logistic regression is a Machine Learning algorithm to solve binary classification problems. It is implemented to predict the probability of the dependent variable. In this algorithm, the dependent variable is a binary variable that contains data as 1 (Yes) and 0 (No). Therefore, logistic regression predicts P(Y=1) as a function of X. There should only be two classes, considering the problem of predicting the customer accepting the deposit scheme. The variables are dichotomous. Logistic regression is a particular case of linear regression where the target variable is categorical in nature. The mathematical breakdown for logistic regression is as follows:

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots.. + \beta nXn$$

Where ‘Y’ is the dependent variable and X1, X2 ... and Xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-y}$$

Apply Sigmoid function on linear regression:

$$p = 1/1 + e^{-(\beta 0+\beta 1X1+\beta 2X2.....\beta nXn)}$$

The following steps should be followed to get the required output in the Logistic Regression: (1) Feature selection; (2) Data splitting; (3) Model Development; and (4) Prediction.

2.1.2. Random Forrest:

The Random Forest classifier model is a combination of many decision trees. Unlike decision tree that predicts value on average, random forest uses two key concepts (Koehrsen, 2018), which are:

- Random sampling of training data points.
- Random subsets of features while splitting the data set.

While training the data set, random forest selects knowledge from randomly sampled data points. The samples drawn could be used multiple times in a single tree. The reason being that the entire forest will have lower variance. During testing, the predictions are made by average the prediction of each decision tree. Bagging is the process that is used to train every individual learner on a different subset of data.

An alternate method is that only a particular subset of features is considered for the splitting of nodes. The square root of n features is used to classify subsets. For example, if there are 9 features at each node, only 3 random features will be considered for splitting the node.

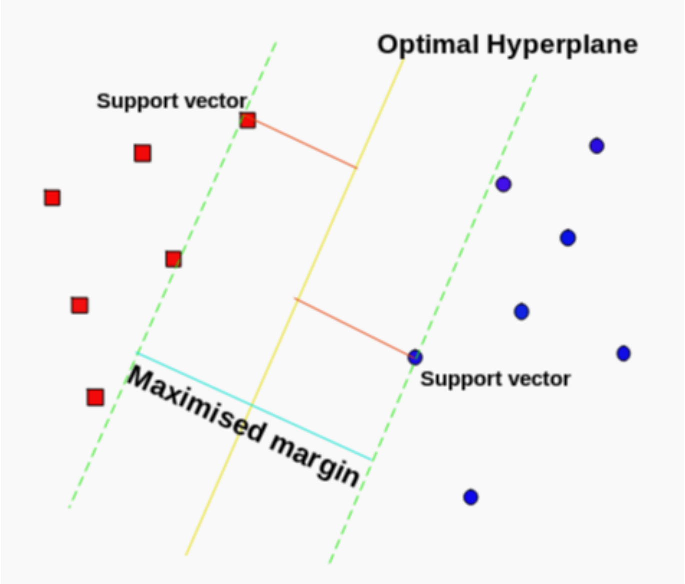
To summarise, the random forest implements the combination of hundreds of decision trees and trains each tree on a set of observations, splitting nodes in each tree considering a limited number of features. Finally, the last random forest is made by the average prediction of each tree.

2.1.3. Support Vector Machine (SVM)

SVM is used for classification and regression problems. SVM overcomes the issue of non-linear issues that are seen in linear regression. SVM can solve linear as well as non-linear problems. In addition, it works for practical cases as well. This algorithm is also being used to find outliers in a data set.

The SVM model results in a line or hyperplane which separates the data into different classes. A hyperplane is a flat, n-1 dimensional subset of space divided into two different but disconnected parts.

Figure 1. Example of SVM implementation.



Source: (Pupale, 2018)

Looking at the above figure, we can see that the algorithm tries to decide the line so that the separation between the groups/classes is as wide as possible.

The advantages of SVM are:

- They are highly effective in a high dimensional space.
- Makes use of a subset of training points in the decision function, which results in better memory efficiency.
- Give a possibility of specifying custom kernel functions for decision functions.

2.2. Measurements to evaluate the results

In case of performance issues, while using ensembling techniques, a data reduction of the dataset will be performed to create a representative sample. A confusion matrix will be created to visualise the performance of supervised learning tasks. Recall, precision, and F1 score will evaluate the result of the models.

Identifying customers that are likely to subscribe to a term deposit is an imbalanced classification problem because the “no” customers vastly outnumber the “yes” customers. In these cases, accuracy is not a good way to measure the model performance because even if our model is biased and predicts “no” in all cases, it would still get a very high accuracy score. This situation happens because the accuracy score looks at the ratio of correctly predicted cases divided by the total number of cases, disregarding the data distribution.

A good measure in this case of imbalanced data is precision and recall because their formula considers the minority datapoints and penalises the model for incorrect predictions in these cases.

The recall function identifies the minority class in a dataset and finds the accuracy for that class (Koehrsen, 2021). For instance, it shows the ratio of how many cases should be predicted as ‘yes’ customers and how many were predicted as ‘yes’ customers.

Recall function:

$$Recall = \frac{TP}{TP + FN}$$

$$\frac{'yes'\ customers\ correctly\ identified}{'yes'customers\ correctly\ identified + 'yes'customers\ incorrectly\ labeled\ as\ 'no'customers}$$

On the other hand, the precision function shows the ratio between how many ‘yes’ customers were correctly identified divided by how many cases were predicted as ‘yes’ customers.

Precision Function:

$$Precision = \frac{TP}{TP + FP}$$

$$\frac{'yes'customers\ correctly\ identified}{'yes'customers\ correctly\ identified + 'no'\ customers\ incorrectly\ labeled\ as\ 'yes'customers}$$

If the model is biased and predicts ‘no’ for all customers, both the recall and the precision function would have a zero score, pointing out an inferior performance to predict the minority group of ‘yes’ customers.

Using two scores can be challenging to compare different models. The best alternative is to use the F1 score formula, which considers both equations.

F1 score formula:

$$F1\ score = \frac{2 * precision * recall}{precision + recall}$$

In the analysis, recall, precision and F1 score will be used to evaluate the result of the models and compare them to see which one offers better predictions. These metrics will be visualised using the scikit learn’s *classification_report()* function.

3. PRELIMINARY RESULTS

3.1. Data exploration with data visualisation

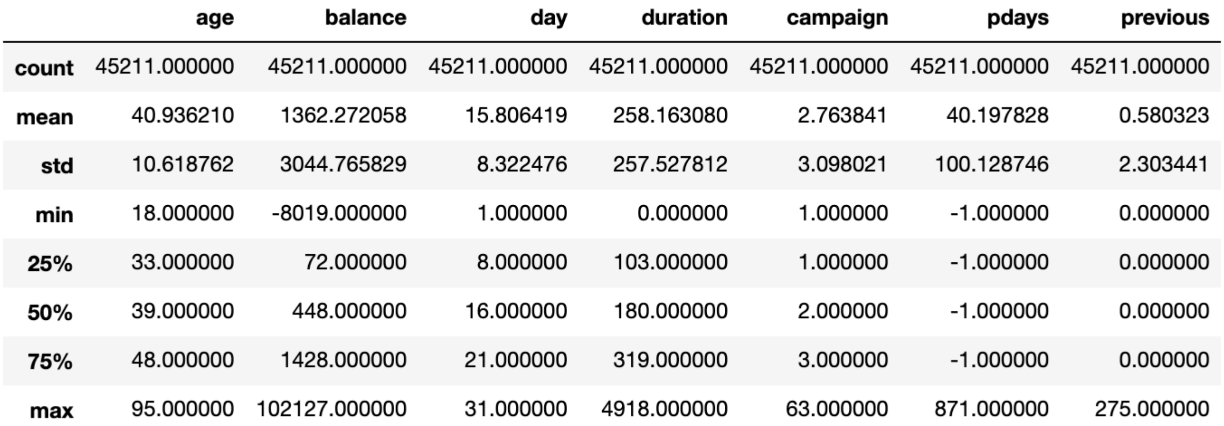
This dataset comprises 45,211 records and 17 columns, with 7 numerical variables and 10 categorical variables. The dataset covers campaign data from May 2008 to November 2010 (Moro, Cortez, & Rita, 2014). The info method (Table 3) shows that there are no null cells. However, this happens because cells are populated with the word “unknown” instead of null values. Therefore, we will treat these values as missing values when doing the cleaning.

Table 3. Number of data samples and types of the dataset.

```
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
#   Column     Non-Null Count  Dtype
---  ------     --------------  -----
0   age        45211 non-null  int64
1   job        45211 non-null  object
2   marital    45211 non-null  object
3   education  45211 non-null  object
4   default    45211 non-null  object
5   balance    45211 non-null  int64
6   housing    45211 non-null  object
7   loan       45211 non-null  object
8   contact    45211 non-null  object
9   day        45211 non-null  int64
10  month      45211 non-null  object
11  duration   45211 non-null  int64
12  campaign   45211 non-null  int64
13  pdays      45211 non-null  int64
14  previous   45211 non-null  int64
15  poutcome   45211 non-null  object
16  y          45211 non-null  object
```

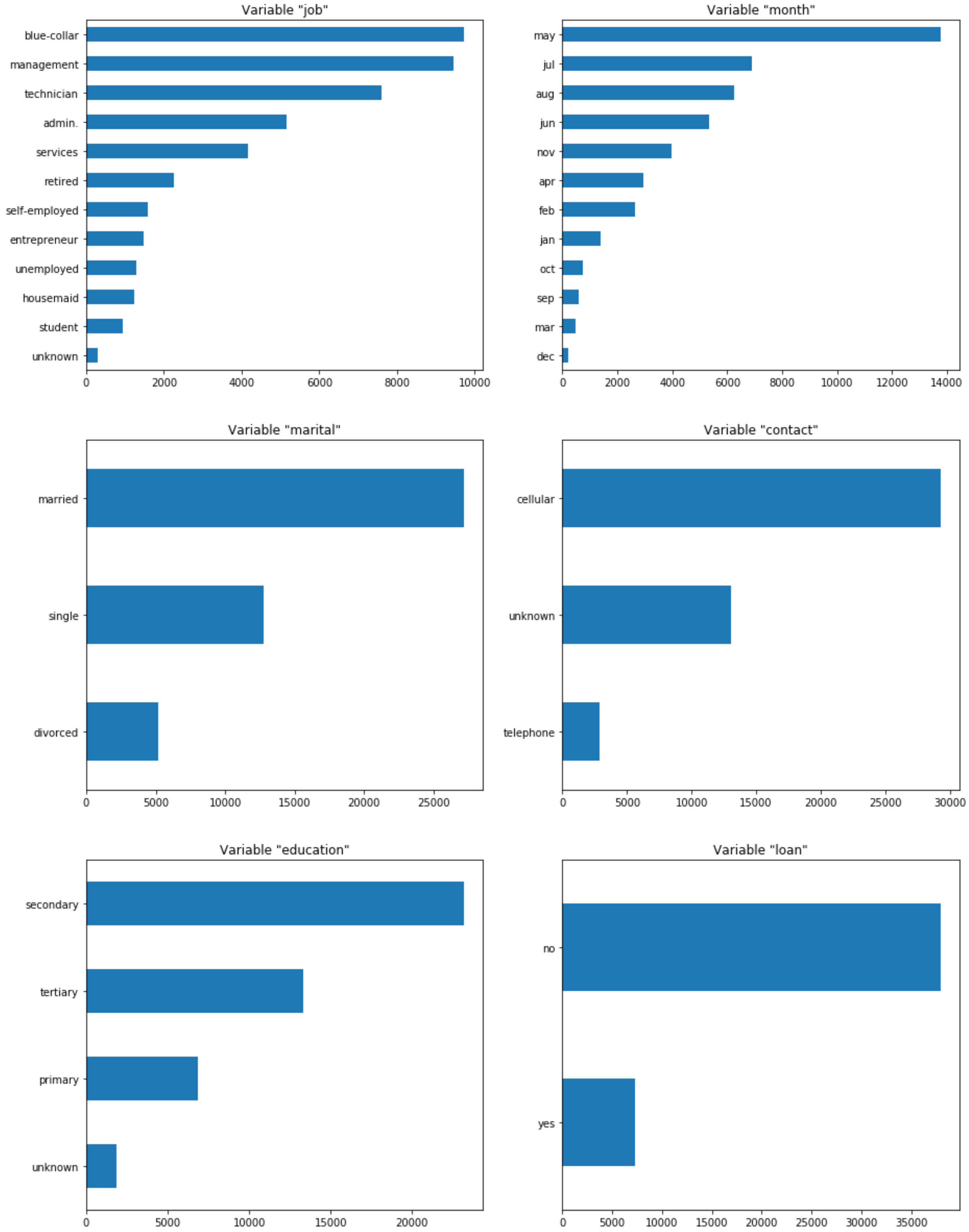
The describe method shows information for the numerical variables. It is possible to see, for instance, that the minimum age is 18 and the maximum age is 95. However, 75% of the customers are 48 years old, and 25% is 33 years. This insight suggests that 18-year-old customers are outliers.

Table 4. Statistical information of the dataset.



	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

Figure 2. Frequency count of dataset's categorical variables.



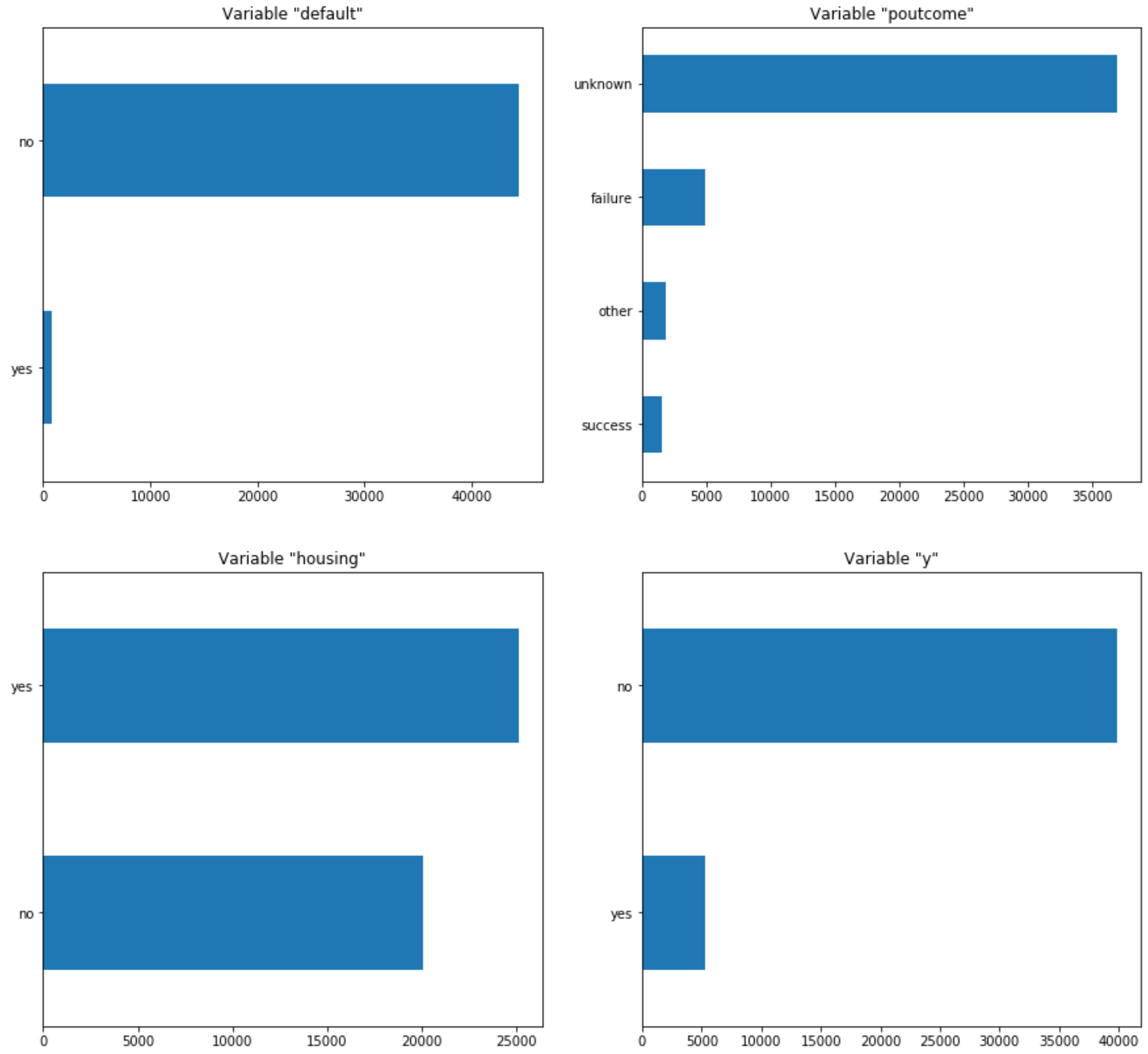


Figure 2 shows a further exploration of the unique values of some categorical columns. It is important to mention that the ‘y’ column (which happens to be the target variable for this analysis) is highly unbalanced. It contains 39,922 “no” and 5,289 “yes”. It may cause bias when training our models, and therefore, special attention will have to be paid to making sure the data is balanced when fitting our model to the data.

Figure 3. Pair plot of dataset's numerical variables.

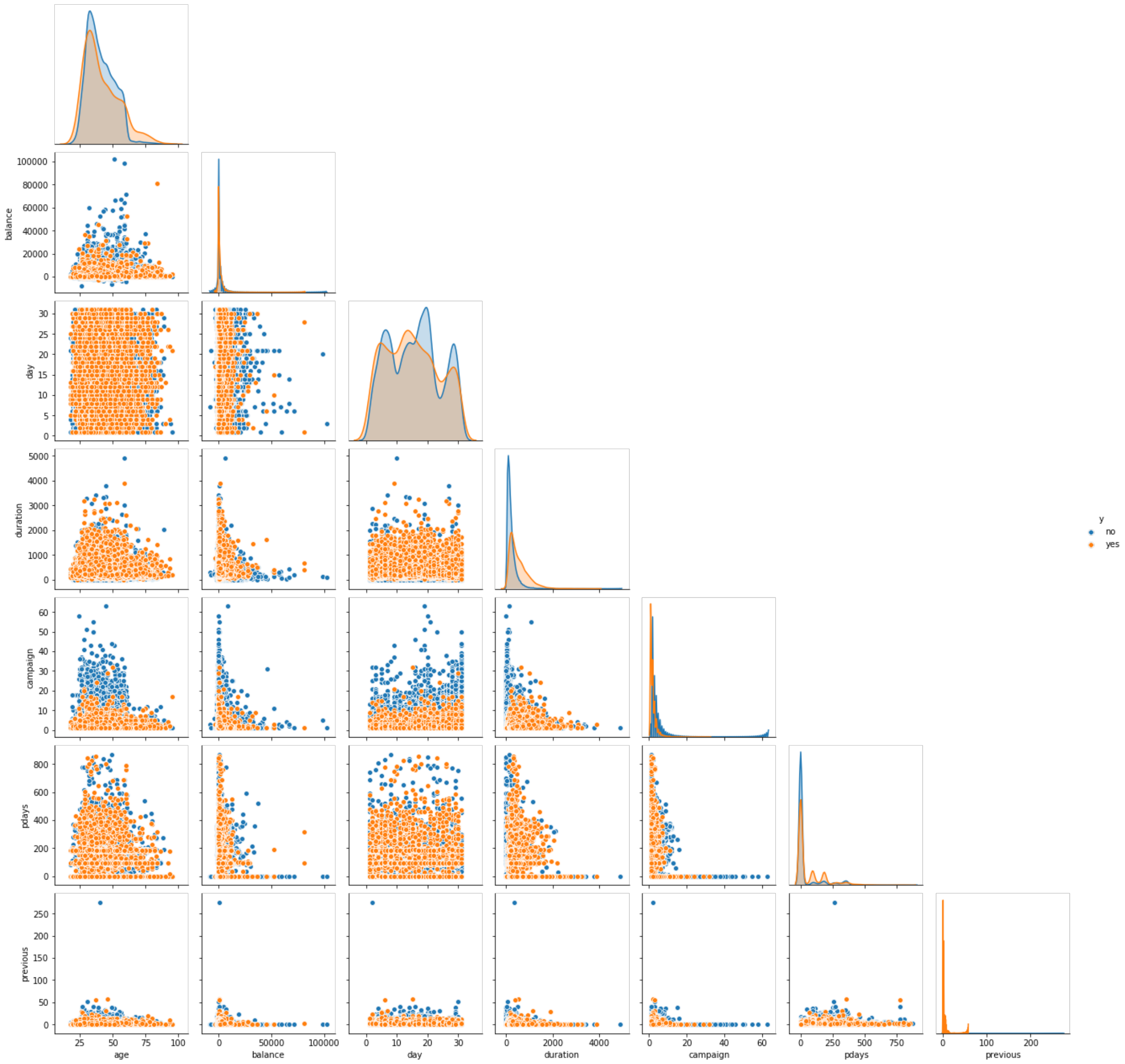


Figure 3 shows the pair plot of the numerical variables. Again, it is possible to see correlations between numerical variables that would be worth exploring. As a result, it seems that customers between 25 and 75 years old, who have a balance higher than 20,000, are more likely to say 'no' to the term deposit than customers in the same age range but with a lower balance.

3.2. Data pre-processing

The following techniques will be implemented for pre-processing:

3.2.1. Null (unknown) values

The dataset that is being analysed does not show any explicit NULL values, however, there are many cells populated with the value “unknown” (categorical), which for the purpose of this analysis is considered as NULL value. These types of values will be treated as follows:

- Columns with a large number of “unknown” values will be dropped. This is the case of the “contact” and “poutcome” columns.
- Unknown values in the age attribute will be replaced by the average age of the customers.
- Rows with unknown values in education will be dropped.
- For the “Unknown” values in the “job” column, we will try to predict them using the attributes of “education”, “default”, “balance”, “housing” and “loan”.

3.2.2. Categorising age attribute using equal-width binning:

Equal width binning is the division of data set into K number of bins of equal width.
Because the dataset has many observations, we will divide the age attribute into equal bins in order to facilitate the analysis.

3.2.3. Z-score Normalisation:

Z-score is a normalisation technique that would normalise our data and give an idea of how far a particular data point is from the mean value. Implementation of the Z-score technique would normalise the variation of units of data present in the data set. Thus, this technique would help with faster processing of data and achieving higher efficiency results.

Z-score formula: $\mathbf{z = (x - \mu) / \sigma}$

3.2.4. One-hot encoding

One-hot encoding will be applied to all categorical variables such as job, marital, education, default, housing and loan. This technique converts any categorical data into numerical variable so that it can be used for machine learning algorithm to implement a better prediction. For this conversion, we will use a method in python called “get_dummies()”.

4. REFERENCES

Chen, J. (2021, April 29). *Investopedia*. Retrieved from Term Deposit: https://www.investopedia.com/terms/t/termdeposit.asp

Fawcett, A. (2021, February 11). *Data Science in 5 Minutes: What is One Hot Encoding?* Retrieved from educative: https://www.educative.io/blog/one-hot-encoding#how

Glen, S. (2018). *Z-Score: Definition, Formula and Calculation*. Retrieved from Elementary Statistics for the rest of us!: https://www.statisticshowto.com/probability-and-statistics/z-score/#Whatisazscore

James, G., Witten, D., Hastie, T., & & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. *Springer*.

Kaggle. (2019). *Portuguese Bank Marketing Data Set: Telemarketing campaign about term deposits*. Retrieved from https://www.kaggle.com/yufengsui/portuguese-bank-marketing-data-set

Koehrsen, W. (2021, August 3). *When Accuracy Isn't Enough, Use Precision and Recall to Evaluate Your Classification Model*. Retrieved from Builtin: https://builtin.com/data-science/precision-and-recall

Maity, A. (2020, March 22). *Predict if the client will subscribe a term deposit or not, using 'Machine learning'* . Retrieved from https://medium.com/@ashim.maity8/predict-if-the-client-will-subscribe-a-term-deposit-or-not-using-machine-learning-c6e4024c7028

Moro, S., Cortez, P., & Rita, P. (2014, June). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, 62:22-31.

Pupale, R. (2018, June 17). *Support Vector Machines(SVM) — An Overview*. Retrieved from Towards Data Science: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Włodarczyk, K., & Ikani, K. S. (2020, March). *Data Analysis of a Portuguese Marketing Campaign using Bank Marketing data Set.* Retrieved from ResearchGate: https://www.researchgate.net/publication/339988208_Data_Analysis_of_a_Portuguese_Marketing_Campaign_using_Bank_Marketing_data_Set

Zach. (2020, July 7). *Equal Frequency Binning in Python*. Retrieved from statology: https://www.statology.org/equal-frequency-binning-python/

5. INDIVIDUAL CONTRIBUTION

Gabriela:
I have contributed with the following tasks:

- section 1.1. (brief description of the problem)
- section 1.2. (what data we will use and where we will get it)
- introduction of section 2.1. (proposed algorithm and techniques)
- section 2.2. (measurements to evaluate the results)
- the written part of section 3.1. (data exploration with visualisation)
- section 3.2. (Data pre-processing)