RMIT UNIVERSITY

Classification of Bank Telemarketing Success
COSC2670 Practical Data Science

Abhishekh Shankar s3652116
Usman Siraj s3652313

2nd June, 2019

Table of Contents

Abstract

The objective of this project was to determine the factors which lead to whether an individual would subscribe to a term deposit through the marketing campaign of an unnamed Portuguese banking institution. A data set was sourced from the UCI Machine Learning Repository and was segmented into different ratios of training and testing splits before the *K*- Nearest Neighbours and Decision Tree algorithms were fit and evaluated. The report concludes that the general public's tendency to subscribe to a term deposit is very low with the current state of the marketing campaign. It is recommended that the bank restructure the focus of their marketing campaign to target clients at times they are more likely to subscribe such as in autumn or winter and to target a younger generation of people (< 30 years). Furthermore, the marketing team should collect more data from clients that are older than 60 as there is promise for another target market, however a lack of data points for that age bracket results in inconclusive results. Finally, the bank should ensure to engage with clients during phone calls as there was a positive relationship between the call duration and subscription rates, and refrain from contacting clients with shorter call durations too many times as this saw very little subscription rates.

1 Introduction

An effective marketing strategy allows companies to engage with their audience, sustain their image to the public, and optimise their financial returns. Conversely, a poor marketing strategy risks unnecessary time and money being spent on a potentially ineffective target market. An unnamed Portuguese banking institution conducted a marketing campaign over phone calls to clients in an attempt to secure a subscription to a term deposit. This report covers two main tasks: firstly, a look into the results of this marketing campaign is conducted via descriptive statistics which offers insights into areas that can improve subscription rates, and secondly, data modelling by fitting and comparing two different classification algorithms: *k*-Nearest Neighbour (KNN) and Decision Trees (DT). Finally, recommendations to the bank's marketing team will be made so improvements can be made on any subsequent marketing campaigns.

2 Methodology

2.1 Data Set

The data set used in this report was sourced from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (Moro, Cortez, & Rita, 2014). The data set contains 41,188 instances with 21 variables including the target variable and is summarised below:

- age: continuous – age of client.
- job: categorical – type of job.
- marital: categorical – client's marital status.
- education: categorial – client's highest education level.
- default: categorical – whether the client has credit in default.
- housing: categorial – whether the client has a housing loan.
- loan: categorical – whether the client has a personal loan.
- contact: categorical – type of communication with client.
- month: categorial – the month that the client was contacted.
- day_of_week: categorical: the day of the week the client was contacted.
- duration: continuous – the last contact duration in seconds.
- campaign: continuous – the number of contacts performed during this campaign.
- pdays: continuous – number of days that passed since the last contact from a previous campaign.
- previous: continuous – number of contacts performed before the current campaign.
- poutcome: categorial – the outcome of the previous marketing campaign.

- emp.var.rate: continuous – employment variation rate, quarterly indicator.
- cons.price.idx: continuous – consumer price index, monthly indicator.
- cons.conf.idx: continuous – consumer confidence index, monthly indicator.
- euribor3m: continuous – Euribor 3 month rate – daily indicator.
- nr.employed: continuous – number of employees – quarterly indicator.
- **y**: **target variable** – binary: whether the client has subscribed to a term deposit.

2.3 Analytical Techniques

This report utilises two classification techniques: *k*-Nearest Neighbours and Decision Trees when classifying the target variable. When applying these algorithms, the data set is segmented into three different ratios of training and testing splits: 50/50, 60/40, and 80/20, with the confusion matrix (CM), classification error rate (CER), precision, recall, and F1-score all computed as a way to measure the accuracy of the models. The results of the precision, recall and F1-score are all summarised in Appendix One as classification reports.

Finally, the **duration** variable was removed for classification purposes as (Moro, Cortez, & Rita, 2014) suggest that since the duration is not known before the call is completed, it should be removed in order to have a realistic predictive model.

3 Results

3.1 Data Preprocessing

Prior to any analysis, the data set was thoroughly cleaned and processed in order to handle any erroneous values or missing information that may skew results. A look into the unique values present in each of the variables showed that there were zero instances of data being incorrectly encoded, defined outside the specified domain of that variable, or missing.

3.2 Data Exploration

Figure 1 below shows a bar graph of the distribution of the target variable in the data set, and it is observed that the majority of clients contacted did not subscribe to a term deposit. The number of clients that subscribed total to approximately 5000 out of the 41,188 instances which amounts to roughly 12%.

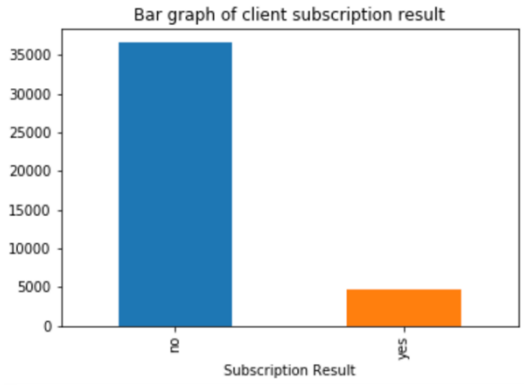

Figure 1: Bar graph of target variable distribution.

Figure 2 below shows a scatterplot of the last contact duration and the number of calls that were made during a campaign, while also labelling the client's subscription result. It can be observed that as the number of calls increased, the likelihood of a client subscribing to a term deposit decreased. Furthermore, as the duration of each call increased, clients tended to subscribe more often. As a result, it is clear that the marketing team should focus returning calls on clients who had longer call durations while also trying to engage with the clients and have longer calls.

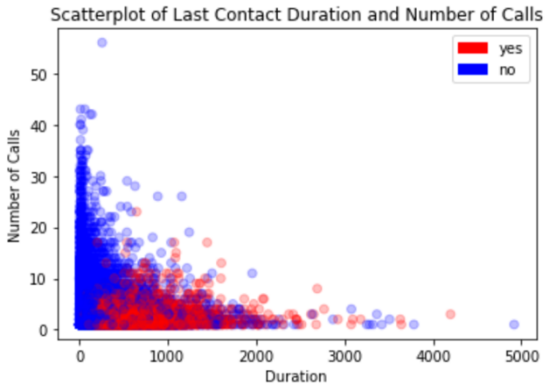

Figure 2: Scatter plot of the number of calls and the call duration by subscription result.

Figure 3 below shows a bar graph of the percentage of successful subscriptions at each age present in the data. It is observed from the plot that people aged below 30 and above 60 showed the largest percentage of clients who subscribed. From this, it can be said that the middle-aged population were the least likely to subscribe to a term deposit.

The density plot of the age variable is shown below in Figure 4 and it can be observed that the majority of people that was contacted were aged between 30 and 60 and those that are older were very rarely contacted. As a result, the clients who were aged 87, 89, and 98 who showed 100% subscription rates in Figure 3 is only a small number of cases and as such, is not a large enough sample to be able to draw conclusions from.

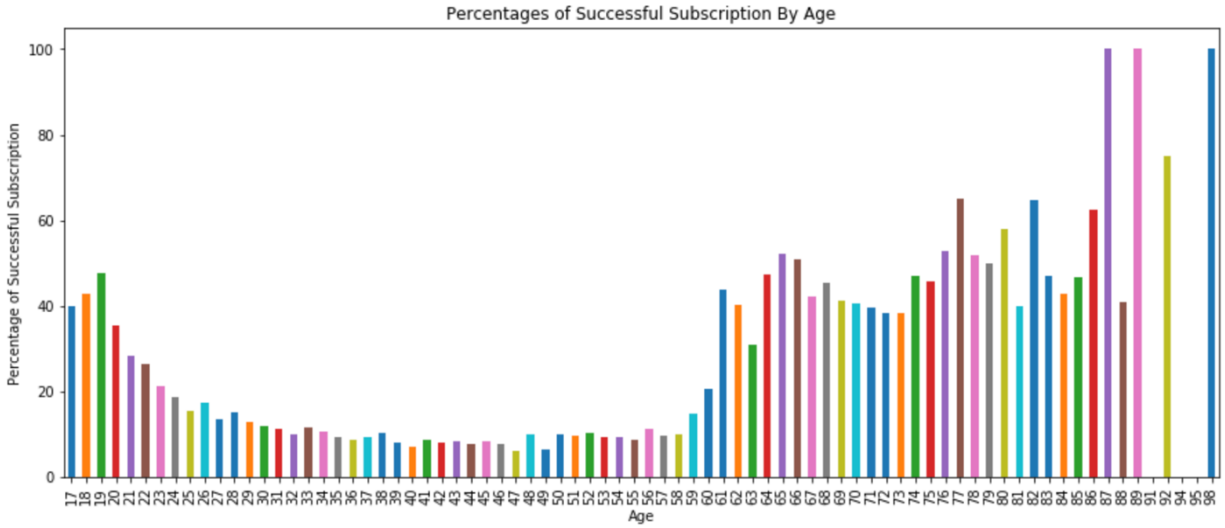

Figure 3: Bar graph of successful subscription percentage by age.

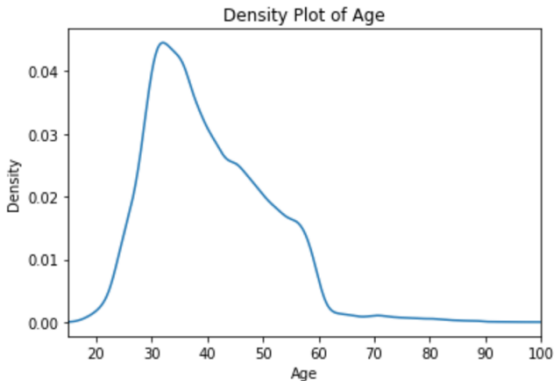

Figure 4: Density plot of the client ages.

Students and the retired were the two largest groups to subscribe to a term deposit with subscription rates exceeding 25%, as shown below in Figure 5. The clients in the remaining job titles all showed subscription rates no greater than 15% with blue collar workers being the least likely to subscribe with a rate of approximately 7%.

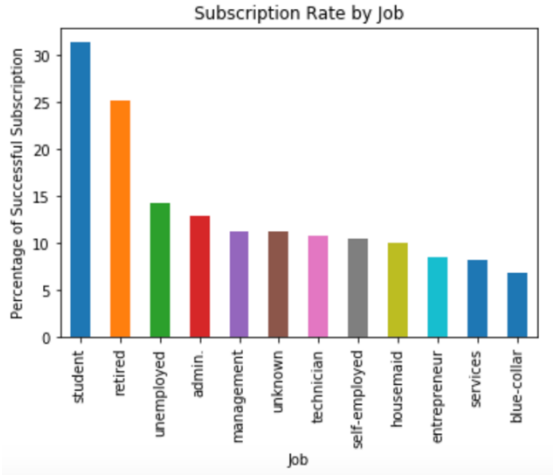

Figure 5: Successful subscription rate by job title.

The months of March, December, September, and October showed the successful subscription rates of 50.5%, 48.9%, 44.9%, and 43.9% respectively as shown below in Figure 6. All of the summer months (August, June, and July)[1] collectively showed some of the lowest successful subscription rates, with values around 10%. As a result, most clients are more likely to subscribe to a term deposit in the autumn months, and during December which marks the Portuguese fiscal year. Simultaneously, Figure 7 shows that the total number of calls are largest for the months that show the lowest percentage of subscribers, indicating inappropriate timing of calls during the marketing campaign.

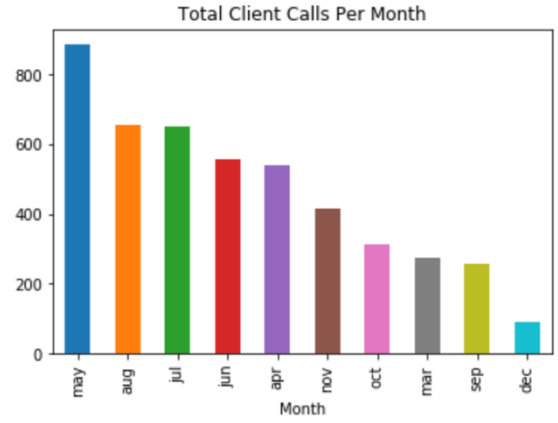

Figure 7: Total number of client calls per month.

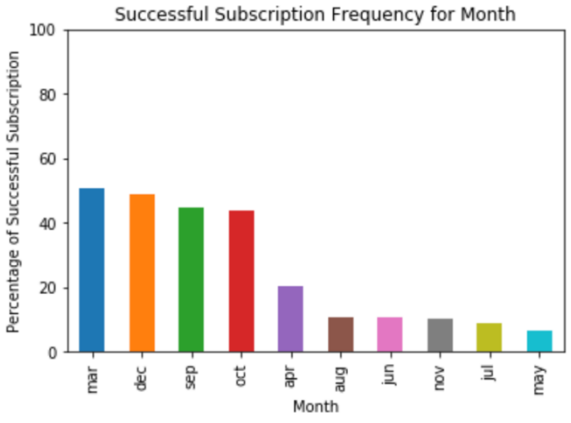

Figure 6: Successful subscription rate by month.

Figure 9 shows a bar plot of the successful subscription percentages per the number of days that passed since the client was last contacted. While the 999 encoded variable[2] showed the greatest density of instances as shown in the density plot in Figure 8, it recorded one of the lowest subscription rates indicating that the majority of people declined to subscribe when initially contacted. Subscription rates are shown to increase greatly amongst clients who were contacted multiple times

[1] Summer months in the northern hemisphere where the Portuguese banking institution is based.

[2] 999 encoded variable indicates that the client had not been contacted beforehand.

during the campaign, with rates fluctuating between 20% and 80%. There was seen to be several instances of 100% subscription rates when clients were contacted 21, 25, 26, and 27 days since the last contact, however this result may be the result of a small number of clients who were called on those days.

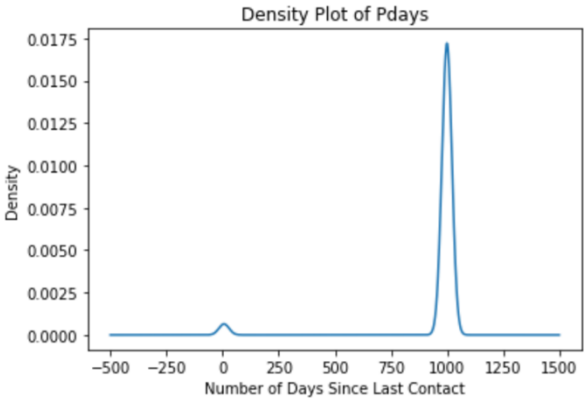

Figure 8: Density plot of pdays.

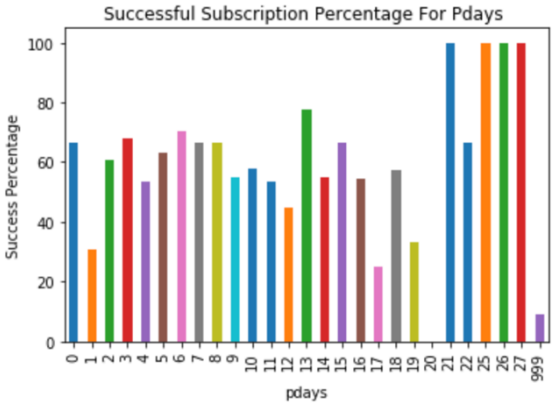

Figure 9: Successful subscription percentages by pdays.

3.2 Data Modelling

3.2.1 K Nearest Neighbours

The three plots below in Figure 10 show the testing accuracies for k values for $1 \leq k \leq 30$. While larger values were shown to have very slight increases in testing accuracy ($\leq 1\%$), selecting a k value that is too large may lead to underfitting in the model. From this, a value of $k = 16$ is deemed to be optimal for both the 50/50 and 60/40 split as shown in the left and middle plots in Figure 1, and $k = 15$ for the 80/20 split on the right.

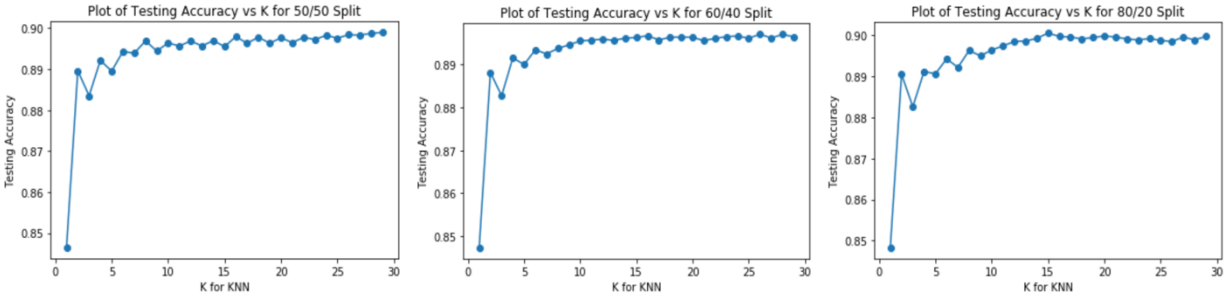

Figure 10: Plots of testing accuracy for different k values for 50/50 (left), 60/40 (center), and 80/20 (right) training/testing splits.

Table 1 below shows both the confusion matrices (CM) for each of the training/testing splits alongside their respective Classification Error Rates (CER) for the KNN classification algorithm. Furthermore, Appendix One shows the recall, precision, and F1-score for each of the two algorithms for each training/testing split. The 80/20 split showed the lowest CER with 9.94%, and the 60/40 split showed the greatest CER with 10.32%.

Each of the training/testing splits showed high values of precision and recall (0.91 and 0.98 respectively) when classifying clients who did not subscribe to a term deposit, as opposed to 0.62 and 0.23 respectively for clients who did subscribe.

Table 1: KNN Confusion Matrix and Error Rate for each split.

	50/50 Split	60/40 Split	80/20 Split
Confusion Matrix	$\begin{bmatrix} 17966 & 325 \\ 1776 & 527 \end{bmatrix}$	$\begin{bmatrix} 14394 & 212 \\ 1489 & 381 \end{bmatrix}$	$\begin{bmatrix} 7182 & 126 \\ 693 & 237 \end{bmatrix}$
Classification Error Rate	10.20%	10.32%	9.94%

3.2.2 Decision Trees

Table 2 below shows both the confusion matrices (CM) for each of the training/testing splits alongside their respective Classification Error Rates (CER) for the decision tree classification. The 50/50 split showed the lowest CER with 15.69% and the 60/40 split showed the greatest CER with 16.62%.

Each of the training/testing splits also showed high values of precision and recall (0.92 and 0.91 respectively) when classifying clients who did not subscribe to a term deposit, as opposed to 0.33 and 0.34 respectively for clients who did subscribe.

Table 2: Decision Tree Confusion Matrix and Error Rate for each split.

	50/50 Split	60/40 Split	80/20 Split
Confusion Matrix	$\begin{bmatrix} 16534 & 1705 \\ 1527 & 828 \end{bmatrix}$	$\begin{bmatrix} 13095 & 1511 \\ 1228 & 642 \end{bmatrix}$	$\begin{bmatrix} 6604 & 704 \\ 617 & 313 \end{bmatrix}$
Classification Error Rate	15.69%	16.62%	16.04%

4 Discussion

The bank's current marketing campaign saw an approximate 12% subscription rate. It was seen that the majority of clients contacted were people aged between 30 and 60 and this population showed the lowest subscription rates. Furthermore, students and those that are retired were shown to be the most likely to subscribe and as such, concentrating the marketing campaign to these individuals may see increased subscription rates. Conversely, the number of clients aged above 60 that were contacted were low and as a result, their high subscription rates may not be significant. As such, the banking institution should collect more data on this particular age bracket as it could result in another target market.

Investigating the months that clients were contacted showed that summer months showed very little rates of subscription while autumn and winter months showed the highest. The current marketing campaign however, showed the majority of calls during the summer, indicating the inappropriate timing of the campaigns that the bank currently has. As such, the marketing campaign should concentrate their calls during autumn and winter.

Finally, a look into the number of days since a client was last contacted was investigated. The vast number of calls were to clients who were not yet contacted and the subscription rate for this population was the lowest. When clients were contacted more than once, their subscription rates increased greatly, from approximately 10% to between 40-60%. As such, a refined marketing campaign should contact clients multiple times. Alongside this, clients were shown to be much more prone to subscribing when call durations were longer (indicating that engaging with clients results in a positive experience), however, calling too many times showed very low rates of subscription, thus, a

threshold should be defined for which the bank no longer contacts clients once the threshold has been reached. From this analysis, a suitable threshold sits around 10 calls to a client.

Regarding the two classification algorithms that were utilised, the *KNN* algorithm showed more accurate results in all cases (CM, CER, Recall, Precision, F1-Score) compared to the *DT* algorithm and because of this, it is the recommended algorithm to apply. Regardless, the testing and analysis of other classification algorithms (such as a Random Forest or Naïve Bayes) may be worth investigating alongside those tested in this report.

5 Conclusion

An analysis into the marketing campaign showed a number of areas where the bank may be able to concentrate their campaign in an attempt to increase the number of clients who subscribe to a term deposit. Analysis showed that the highest rate of subscription came from young adults and those under 30, and those over 60 (however more data is needed to be collected for this age bracket), and as such, the total number of contacts made to these clients should be made. Furthermore, students, and the retired population were shown to have the highest rates of subscription amongst all the job titles, while blue collar workers very rarely subscribed. The bank should also concentrate their campaign in winter and autumn months as opposed to the summer months and when contacting clients, the marketers should try to engage with clients and have longer call durations as this results in higher subscription rates. Finally, a threshold of 10 calls per client should be defined per campaign as subscription rates beyond 10 calls showed minimal subscription rates.

Applying both the *KNN* and *DT* algorithm to the data set showed promising results, however the *KNN* showed the best results. Investigations into other classification algorithms may be worth conducting as they may produce more accurate results, however the results of the *KNN* are deemed to be satisfactory.

Finally, investigations into other factors should be conducted as a means to determine the conditions a client is most likely to subscribe, such as the current economic state in the nation, the absence or presence of loans that a client may have, a look into their equity or available funds. As such, slight modifications to the current marketing campaign as outlined in this report may be able to increase the rate of subscription to a term deposit for this particular banking institution.

References

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 22-31.

Appendix 1 – Classification Reports for KNN and DT

$\underline{KNN}$

	precision	recall	f1-score	support
0	0.91	0.98	0.94	18291
1	0.62	0.23	0.33	2303
micro avg	0.90	0.90	0.90	20594
macro avg	0.76	0.61	0.64	20594
weighted avg	0.88	0.90	0.88	20594

Figure 11: KNN Classification Report for the 50/50 Training/Test Split.

	precision	recall	f1-score	support
0	0.91	0.99	0.94	14606
1	0.64	0.20	0.31	1870
micro avg	0.90	0.90	0.90	16476
macro avg	0.77	0.59	0.63	16476
weighted avg	0.88	0.90	0.87	16476

Figure 12: KNN Classification Report for the 60/40 Training/Test Split.

	precision	recall	f1-score	support
0	0.91	0.98	0.95	7308
1	0.65	0.25	0.37	930
micro avg	0.90	0.90	0.90	8238
macro avg	0.78	0.62	0.66	8238
weighted avg	0.88	0.90	0.88	8238

Figure 13: KNN Classification Report for the 80/20 Training/Test Split.

$\underline{DT}$

	precision	recall	f1-score	support
0	0.92	0.91	0.91	18239
1	0.33	0.35	0.34	2355
micro avg	0.84	0.84	0.84	20594
macro avg	0.62	0.63	0.62	20594
weighted avg	0.85	0.84	0.85	20594

Figure 14: DT Classification Report for the 50/50 Training/Test Split.

	precision	recall	f1-score	support
0	0.91	0.90	0.91	14606
1	0.30	0.34	0.32	1870
micro avg	0.83	0.83	0.83	16476
macro avg	0.61	0.62	0.61	16476
weighted avg	0.84	0.83	0.84	16476

Figure 16: DT Classification Report for the 60/40 Training/Test Split.

	precision	recall	f1-score	support
0	0.91	0.90	0.91	7308
1	0.31	0.34	0.32	930
micro avg	0.84	0.84	0.84	8238
macro avg	0.61	0.62	0.62	8238
weighted avg	0.85	0.84	0.84	8238

Figure 15: DT Classification Report for the 80/20 Training/Test Split.