Report Title\*

(COMP3125 Individual Project)

Qranfal  
*dept. name of organization*

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*provide a short abstract*)

Keywords—example1, example2, example3, example 4, example 5 (provide 3-5 keywords)

# Introduction (*Heading 1*)

Portuguese banking institution wants to identify the types of customers that are most likely to respond to their marketing campaigns (those that are more likely to deposit) to optimize spending on campaigns. It is interesting to see what leads to people responding to such marketing campaigns more often, as it could allow for more profits using targeted campaigns. The current research/results in this area are predictions on success of the bank telemarketing, checking relevance of the many features towards the success of the campaign.

# Datasets

## Source of dataset (Heading 2)

The dataset was downloaded from UCI Machine Learning Repository. It is a credible source maintained by many universities, and the datasets were generated by researchers that recorded information given by these Portuguese banks with information on their clients and success of campaign.

## Character of the datasets

Example: The dataset is in CSV format, with size of 45211 rows, with 17 features, and one target variable, a total of 18 columns. The dataset was checked for any missing values and converted the units of the target variable and categorical variables into those that are applicable with machine learning models. The rule applied for the target variable was encoding, 0 for no, and 1 for yes. For the categorical variables they were one-hot encoded, such that there was a column generated for each possible value in the category. There was no combination of any datasets as the bank dataset downloaded provided the full data.

# Methodology

## K-Means Cluster

K-Means Cluster work in a way that it initially picks random points within the data, then finds the closest values nearest to that point, average them out, and uses that as the center. It continuously repeats this process until the center/cluster’s values of the previous and next fall under a certain threshold or after a certain amount of iterations. These clusters include group of similar people within those features, and can be useful to understand different groups and how they can be targeted.

## Decision Trees Classification

Decisions trees work by finding the feature with highest information gain as the root, which can be calculated using the Gini index, entropy, or other metrics. Then, it does this from the root, using a subset of the data using the root feature data, and does this until a certain amount of max depth that is stated. These trees are easy to interpret and can provide a good starting point to understand how important the features are towards predicting customers’ response.

## Random Forest Classification

Similar to decision trees, except multiple trees are created using subsets of the data or features that are randomly chosen with replacement, allowing for a more accurate representation of the data along with higher accuracy, as a vote is taken from all the trees to decide on the most accurate prediction.

# Results

In this part, you need to select a reasonable way to deliver the result of your topic. For example, equation or numerical results, or visualization of your result. You also need to provide a clear explanation of all results and how to understand the results. If there exist any unexpected results, please explain why or possible cause of this special result. You can use subsection A. B. to separate your results.

Example: After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Calls response rate per month

Most calls were made in May, which resulted in the lowest response rate of yes. In months with lower number of calls, there were a general trend of higher response rates, indicating that the more calls made in a month, the less response rate, indicating that there could be many repeat calls or there could be spread of the calls such that people do not bother responding.

## Clusters

With the eight clusters, using age and balance as the features, we see that cluster 5 has the highest response rate, of 43.81%, then cluster 3 with 16.19% response rate, and then cluster 4 with 15.91%. The average age of cluster 5 is 72, with average balance of $1875.90. The average age of cluster 3 is 35, and average balance of $5514.14. Average age of cluster 4 is 51, with average balance of $50199.00. The reason for this seems to be that those with either very high average age, or very high average balance, or a balance of both tends to respond and deposit to Portuguese campaigns in response to campaigns. This makes sense because older people or those that have a lot more money are more often willing to open a bank for their savings.

## Duration of call impact on response rate, and how long until they respond

Using a decision tree classifier model, with duration of call and previous number of calls received as feature, created a model with 89% accuracy towards predicting response rate. The model predicts better that people will not respond at 90% accuracy, compared to predicting if they will respond yes with 57%. Duration has a feature importance of around 0.84, meaning it’s the major factor towards understanding and predicting a person response rate to the campaign. We see that average respond time for client that responded yes is 69 days, while for responding no it’s 37 days, showing those who respond negatively do it within a quicker amount of time.

## Response rate for cellular and telephone contact

The percent of clients contacted via cellular is 64.8%, and by telephone is 6.4%, the rest being unknown. Our analysis shows that of the those contacted by cellular, about 85% respond no to campaign and 15% respond yes. Of those contacted by telephone, about 86.6% respond no, and 13.4% respond yes to the campaign. Those contacted by cellular have slight improvement in response rate, which can be explained by either the larger quantity of calls done via cellular or it’s a preferred contact method for clients.

## Main factors towards predicting response rate

The four main factors towards predicting response rate to a campaign are duration of the call, the response to campaign from last contact, the number of days passed since last contact, and information of a clients housing loan. The random forest model with such features explained for 74.7% of the variance in the models prediction, indicating a lot of importance based of these factors toward predicting the response rate.

# Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

Example: K-Means weakness is that even though eight clusters best separated the data into groups, many of the group are very similar that there is no information that make them distinct from each other in terms of responsive rate, except for a few of the top three clusters. The decision tree model weakness is that it tends to overfit, and often struggles when the data is not balanced as in this case more people responded no compared to yes, which resulted in a higher accuracy when predicting the response of no, compared to yes. Finally, random forest have the same issue of overfitting, but it’s more resistant as it creates multiple trees, and resulted in a higher accuracy when predicting response rate for those who responded yes.

# Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what’s the effect of this result on the real-world?

Example: The first key finding is that more call withins a month lead to lower response rate, likely due to volume of calls largely being repeats or being ignored. Clustering shows older people, those with a high balances, and those who are young and medium average balances had the highest response rates. Call duration was the most important feature towards predicting response rate, as longer calls lead to higher response rates, although those who responded no did so in a quicker manner. With a random forest model, we concluded that the top four important factors were call duration, previous responses, time since last contact, and housing loan information. In the real world, we can target these demographics of people, such as those who are older or with higher balances, along with ensuring a longer calls to ensure they make a deposit and respond yes to campaign. Along with that, we can focus on looking into the four main important features when deciding to contact someone in order to be efficient and saving time with calls.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...” Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.