Predicting Customer Response in Portuguese Banking Telemarketing Campaigns

(COMP3125 Individual Project)

Qranfal  
*Wentworth Institute of Technology*

# Introduction (*Heading 1*)

Portuguese banking institutions want to identify the types of customers most likely to respond to their marketing campaigns, meaning those more likely to deposit, to optimize spending on campaigns. It is interesting to see what leads people to respond to such marketing campaigns more often, as it could allow for more profits using targeted campaigns. The current research in this area focuses on predicting the success of bank telemarketing campaigns by evaluating the importance of various features that contribute to their success.

# Datasets

## Source of dataset (Heading 2)

The dataset was downloaded from the 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 the success of the campaign.

## Character of the datasets

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

# Methodology

## K-Means Cluster

K-Means Cluster works in a way that it initially picks random points within the data, then finds the closest values nearest to that point, averages 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 number of iterations. These clusters include groups of similar people within those features and can be helpful in understanding different groups and how they can be targeted.

## Decision Trees Classification

Decision trees select the feature with the highest information gain as the root, calculated using metrics like the Gini index or entropy. Then, it splits the data recursively, using subsets of the data based on the root feature, and does so until a certain amount of max depth is stated or when all data points belong to one class in the leaf node, essentially ending when there is no more significant information gained by splitting anymore. These trees are easy to interpret and can provide a good starting point to understand how important the features are in predicting customers' responses.

## Random Forest Classification

Similar to decision trees, this method creates multiple trees using randomly chosen subsets of the data or features with replacement. The prediction from such a method is determined by majority votes of all trees, leading to a more accurate model that captures more of the patterns within the data.

# Results

## Calls response rate per month

Most calls were made in May, which resulted in the lowest response rate of yes. In months with a lower number of calls, there was a general trend of higher response rates, indicating that the more calls made in a month, the lower the response rate, suggesting that there could be many repeat calls or a 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, 43.81%, then cluster 3 with a 16.19% response rate, and then cluster 4 with 15.91%. The average age of cluster 5 is 72, with an average balance of $1875.90. The average age of cluster 3 is 35, and the average balance of $5514.14. The average age of cluster 4 is 51, with an average balance of $50199.00. The reason for this is that those with either a very high average age, very high average balance, or a balance of both tend to respond and deposit to Portuguese campaigns in response to campaigns. This makes sense because older people or those with much more money are 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 the call duration and the previous number of calls received as the features, created a model with 89% accuracy in predicting response rate. The model predicts better that people will not respond with 90% accuracy than if they will respond yes with 57%. Duration has a feature importance of around 0.84, meaning it's the major factor in understanding and predicting a person's response rate to the campaign compared to the previous number of calls received from a client. We see that the average response time for clients 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 percentage of clients contacted via cellular is 64.8%, and by telephone is 6.4%, the rest being unknown. Our analysis shows that of those contacted by cellular, about 85% responded no to the campaign, and 15% responded yes. Of those contacted by telephone, about 86.6% responded no, and 13.4% responded yes to the campaign. Those contacted by cellular have a slightly improved response rate, which can be explained by 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 for predicting the response rate to a campaign are the duration of the call, the response to the campaign from the last contact, the number of days that passed since the last contact, and information about a client's housing loan. The random forest model with such features explained 74.7% of the variance in the model's prediction, indicating much importance based on these factors toward predicting the response rate.

# Discussion

K-Means' weakness is that even though eight clusters best separated the data into groups, many of the groups are very similar, and there is no information that makes them distinct from each other in terms of responsive rate, except for a few of the top three clusters. The decision tree model's weakness is that it tends to overfit and often struggles when the data is not balanced 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 forests have the same issue of overfitting, but they are more resistant as they create multiple trees and result in higher accuracy when predicting the response rate for those who responded yes. In future works, for K-Means Clustering, we can look more into the optimal number of clusters by understanding patterns within the selected features and, therefore, choose the number of clusters that best separate the groups. For decision trees and random forest classifiers, we can choose better hyperparameters through the use of a grid search that looks through many possibilities for the best result, therefore optimizing the models.

# Conclusion

The first key finding is that more calls within a month lead to a lower response rate, likely due to the volume of calls being repeated or being ignored. Clustering shows that older people, those with high balances, and those who are young and have medium average balances had the highest response rates to the campaign. Call duration was the most crucial feature in predicting response rates, as longer calls lead to higher response rates. However, those who responded no to the campaign did so in fewer days than those who answered yes, indicating that more thought was put in when considering whether to deposit in the bank. We concluded with a random forest model that the top four important factors were call duration, previous responses, time since the last contact, and housing loan information. In the real world, we can target these demographics of people, such as older people or those with higher balances, along with ensuring longer calls to ensure they make a deposit and respond yes to the campaign. Additionally, we can focus on looking into the four main important features when deciding to contact someone to be efficient and save 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