

# Bank Marketing Data during a Financial Crisis

## DATA 5000 Project Report

Norbert Eke

School of Computer Science  
Carleton University  
Ottawa, ON  
norbert.eke@carleton.ca

Olivia Faria

School of Journalism and Communications  
Carleton University  
Ottawa, ON  
olivia.faria@carleton.ca

### ABSTRACT

This project proposes to use a variety of statistical methods including clustering of bank telemarketing customer data to predict if a client will subscribe a term deposit based on demographic information in the context of a financial crisis, where nefarious activities are often deployed by financial institutions at the expense of vulnerable populations. The technical analysis will be complimented and contextualized by critical data studies in order to highlight how data are inherently political.

### KEYWORDS

Data Mining, Data Analysis, Statistical Machine Learning, Telemarketing, Targeted Advertising, Classification, Clustering, Supervised and Unsupervised Learning, Banking

#### ACM Reference Format:

Norbert Eke and Olivia Faria. . Bank Marketing Data during a Financial Crisis: DATA 5000 Project Report. In *Proceedings of* . ACM, New York, NY, USA, 11 pages.

## 1 INTRODUCTION

A good marketing strategy is one of the essential tools required for any successful business. In order to develop said marketing strategy, it is critical to determine the audience or demographic for a specific product. Who will this product appeal to? Who is most likely to buy this product? Who is the least likely to buy this product? What will the success rate of a specific advertising medium (commercial, telemarketing campaign, web advertisement) be? [6] All of these questions are important in informing a business' marketing strategy.

However, another important aspect of marketing is the social, economic and political environment at the time of the campaign, as these factors can greatly affect the outcome. Furthermore, marketing strategies can also have nefarious intentions embedded within them that do not have the consumer's best interests in mind. For example, the financial industry (primarily banks) have been exposed for various practices such as fraud, corruption and targeting of consumers leading to the Great Recession of 2008.

For this interdisciplinary research project, data from a Portuguese national bank's 2008-2013 telemarketing campaign on term deposit subscriptions will be analyzed socially and technically in order to answer the following questions: 1) Which attributes help to predict if a client will subscribe to a term deposit? 2) Which (if any) demographics respond well/poorly to telemarketing? and 3) Are there any hidden relationships within the dataset that have not been previously discovered through methods previously used?. The following section will provide additional background to the historical context, information on the dataset and a small literature review.

## 2 BACKGROUND AND CONTEXT

### 2.1 Dataset

The dataset used for this project is titled *Direct Marketing Campaigns of a Portuguese Banking Institution*. It was made available via a donation by the authours to the University of California Irvine's Machine Learning Repository. The authors of the dataset are SÃlrgio Moro (University Institute of Lisbon) , Paulo Cortez

(University of Minho) and Paulo Rita (University Institute of Lisbon). An interesting aspect of this dataset is that the authors are of different disciplines, with Moro and Cortez being of technical backgrounds (Computer Engineering and Information Systems respectively) and Cortez of a marketing background.

The dataset is multivariate, contains 17 attributes, no missing values and contains just over 45,000 instances (45,211) [6]. The variables include bank client data (ie. demographic information such as Age, Occupation, Education, Loan History), campaign details (ie. cellular versus land line telephone, day of the week, duration of call) and whether the client subscribed to a term deposit or not. While it is not described in the dataset, it is important to define what a term deposit is. A term deposit is a "fixed term deposit" where once purchased, the client "understands that the money can only be withdrawn after the term has ended" or by giving notice as per the financial institution's policy [2].

## 2.2 Dataset Limitations

There are three main limitations to the dataset used in the project. First, the data was collected about one unnamed Portuguese banking institution. Had the bank been named, a deep social analysis could have been conducted, but this does not necessarily constrain the project too drastically. Second, there are a limited number of variables, such as no variable for gender or year, which could have potentially yielded some interesting results. Lastly, it is important to remember that this dataset only contains telemarketing data, which is only one form of marketing that is becoming increasingly outdated as less people have a landline at home and are not likely to answer calls on their cellphones from unrecognized numbers.

## 2.3 Background and Literature Review

At first glance, this dataset may seem fairly innocuous and suitable for marketing analysis. The authors of the dataset have used it in a variety of papers, with the most recent being in 2017. In the first paper, the authors proposed a data mining approach to "predict the success of telemarketing calls for selling bank long-term deposits" [6]. The most recent article elaborates on the original paper with the goal of "further enhancing

the classification model by presenting a divide-and-conquer strategy" which entails splitting problems into sub-problems [7]. The authors used feature reduction to narrow the dataset from 25 to 10 features and chose to focus on inbound calls as a "sub-problem". The results of the model confirmed the "inbound optimized model" was the best solution as it outperformed the baseline in addition to achieving "ideal lift performance (ie. reaching all potential buyers" without needing a large sample like the baseline model [7]. While the most recent paper purports to be the first study to address inbound banking telemarketing, a major research gap that the authors do not sufficiently address in either paper are the impacts of the financial crisis on the citizens. Rather, the research is much more institutionally focused on the financial industry and marketing campaign managers. Despite having variables that account for the financial crisis, the authors do not really use them to do any meaningful social analysis, even suggesting that future work may need to be done on the dataset to split the time period of 2008-2013 between the peak of the financial crisis and the slow recovery to achieve more meaningful results [6].

The context of this dataset is extremely important to consider as the time period it was collected greatly affects any results that will be yielded, especially since the product is a term deposit from a bank. During the time this data was collected, Portugal was experiencing a financial crisis that at some points was worse than the Great Depression of the United States and Japan's Lost Decade as seen in figure 1 [11]. This crisis was a result of both local and global trends in financial mismanagement. The catalyst of Portugal's economic woes is often cited as being the Great Recession of 2008, which began in the United States as a result of the subprime mortgage crisis and other predatory lending practices by American banks in order to make more money for themselves and their shareholders at the expense of everyday people. There is a wide variety of literature in the social sciences and economics on this nefarious activity, but it often does not cross over to more technical disciplines. The effect of the Great Recession on this data cannot be understated as "out of the twenty-five top subprime mortgage lenders, twenty-one were either owned or financed by major Wall Street or European banks" [8]. Furthermore, tactics on customer profiling and algorithmic targeting were becoming more advanced during this time. Closer to Portugal, 2010 was the peak year

Figure 1. Lost Decades: Portugal, Japan, and the United States

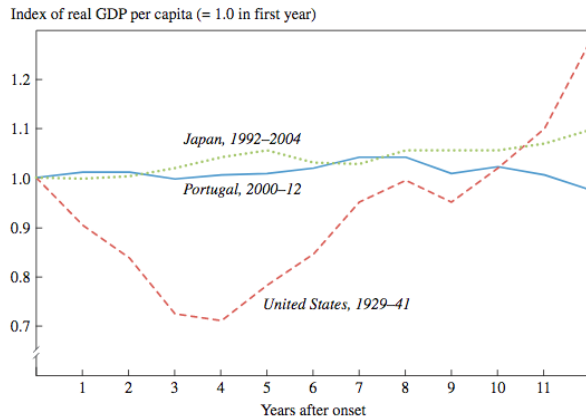


Figure 1: Portuguese Financial Crisis in comparison to Japan and the United States

for the Eurozone Crisis, where a variety of countries (Portugal, Ireland, Iceland, Greece and Spain) in the Eurozone, which is comprised of countries who use the Euro currency, were unable to pay back government debts. Within Portugal, the collapse of two banks (Banco Português de Negócios and Banco Privado Português) due to fraud, corruption and money laundering greatly impacted consumer confidence in the country's financial system [10] [1]. Clearly, a dataset on the success of term deposit description will be impacted by these social, economic and political factors as people will likely have less income to put towards a term deposit as well as have decreased trust in the government and financial industry during times of financial crisis. This social science lens allowed the project to frame the questions in a way that would address consumer concerns and to potentially uncover if the banks were targeting vulnerable populations for term deposits.

### 3 METHODOLOGY

#### 3.1 Tools

The project uses the the statistical language R (version 3.3) paired with RStudio (version 1.0.143) for statistical analysis, while Tableau was also used for some additional data visualizations. R packages used include MASS, plyr, tree, e1071, AUC, caret, naivebayes, randomForest, fastAdaboost, bnlearn, readr, graphics, ggplot2, lattice, Hmisc, corrplot and cluster.

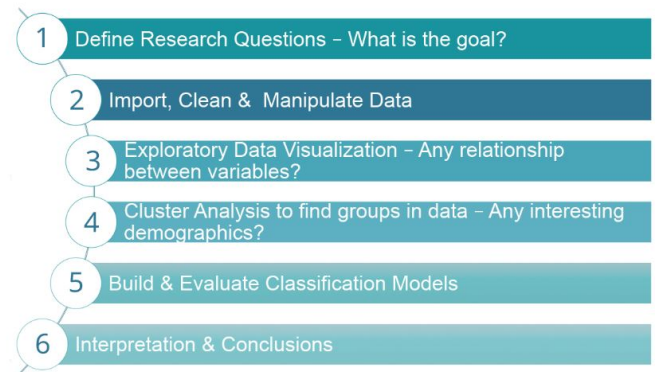


Figure 2: Data Science Process

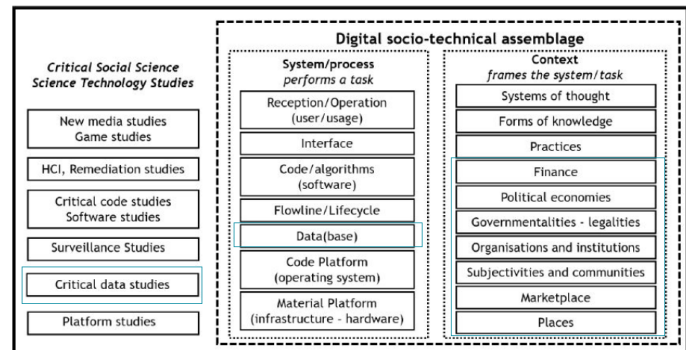


Figure 3: Rob Kitchin's socio-technical data assemblage (2014)

#### 3.2 Data Science Process

Figure 2 shows the Data Science Process used throughout this project. The first step was defining the research questions as to meet the requirements for the project. In the case of this project, no data collection was deployed as a pre-made dataset was used. Next, the dataset chosen was imported, cleaned and manipulated in preparation for analysis. Exploratory data visualization was performed in order to see distribution patterns and relationships between the variables of the dataset. Next, clustering analysis was used to find any potential groups within the dataset. Finally, the largest portion of the process was spent on building and evaluating classification models for the classification task. The last step of the process is to interpret the results and formulate conclusions.

Figure 3 illustrates the critical data studies framework used throughout the project, primarily in the interpretation of the results and background information. The socio-technical data assemblage is a tool used by scholars in disciplines such as critical data studies, human computer interaction and surveillance studies to examine how seemingly neutral systems or processes such as code, algorithms and material infrastructure are actually framed and directly influenced by contexts such as systems of thought, political economy, institutions and governments [5]. In the context of this dataset, the dataset itself would be considered the system/process and the contexts that frame it would include finance, marketplace, governmentalities and institutions, political economies, subjectivities and communities and places. This plays an important part in interpreting the results as the various framings play a major role in how the data can be used. For example, due to the unique placement of the dataset within Portugal during the financial crisis of 2008-2013, the results may not be easily extrapolated to the context of another country with different economic systems, history or governmentalities.

### 3.3 Data Cleaning

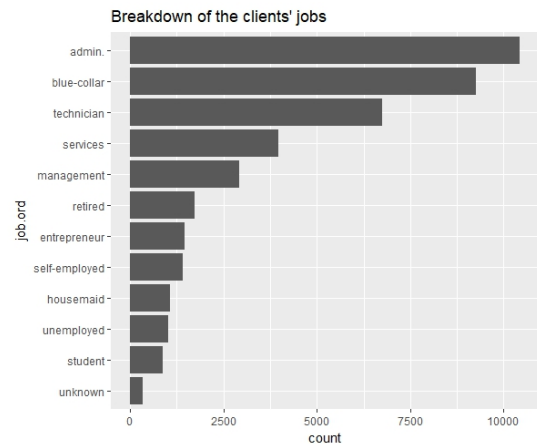
The dataset was fairly clean and did not require rigorous cleaning in order to prepare it for analysis. There were no missing values, but lots of categorical variables were present in the dataset, thus all these variables were converted into factors in R. Other data manipulations necessary included converting numerical strings to integer and "Yes"/ "No" responses to binary 0/1 values.

### 3.4 Exploratory Data Visualization

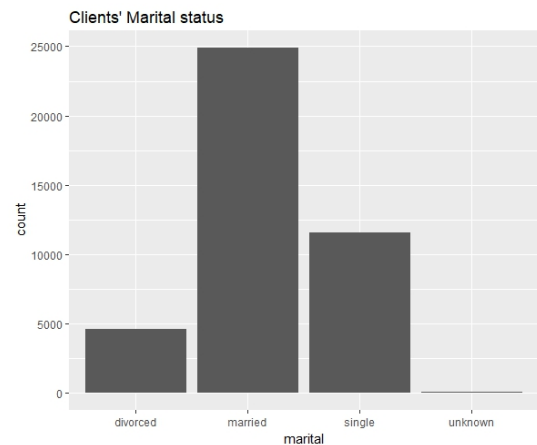
Exploratory data visualization was used in order to better understand the distributions of each explanatory variable. A co-linearity check was also performed on the variables.

Figure 4 shows the distribution of jobs that the clients have. Most clients are in administrative positions, then blue-collar, technical and services jobs are also popular in the dataset. The rest of the jobs include management, entrepreneur, housemaid, and there are plenty of self-employed, unemployed and retired people, additionally some students as well.

Figure 5 shows the distribution of the marital status of the clients. Most clients are married, then single, and the minority is divorced.



**Figure 4: Histogram showing the distribution of jobs that clients have**

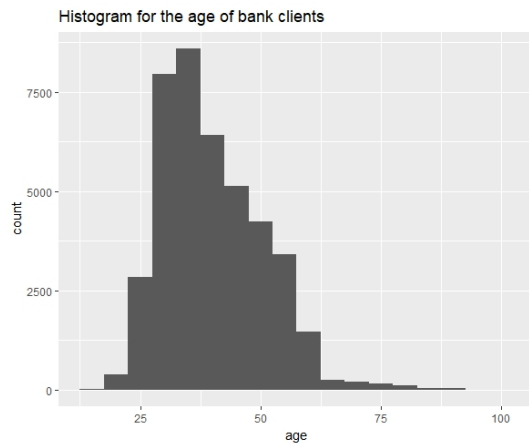


**Figure 5: Histogram showing the distribution of clients' marital status**

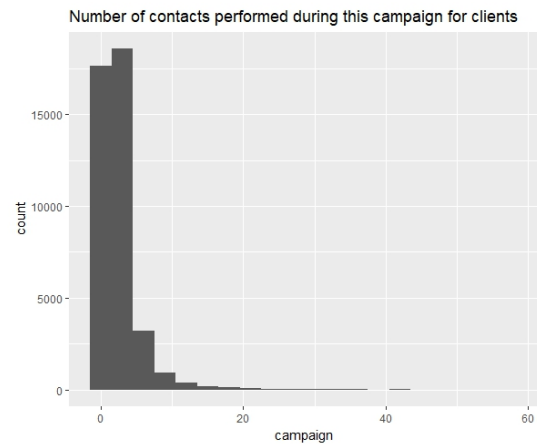
In terms of age, figure 6 shows that the vast majority of the clients are in their forties, while 25 to 40 is also a popular age group within the dataset. The older the client is, the less percentage of the population their age group represents.

The education levels of the clients are significantly high, as figure 7 show that the majority have university degrees, but some only have high school degrees or less, with a small minority having completed professional courses (similar to college or trade school).

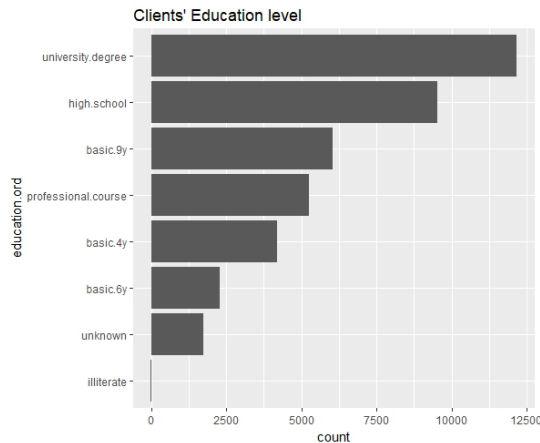
In terms of the number of contacts made by the bank during the campaign, figure 8 illustrated that most clients have received between 1 and 5 calls, while occasionally some clients got up to 10-20 calls.



**Figure 6: Histogram showing the distribution of clients' age**



**Figure 8: Histogram showing the number of times clients got contacted during the campaign**



**Figure 7: Histogram showing the distribution of clients' education levels**

Other observations from data include clients having housing loans is close to evenly split between yes and no, while 82.4 % of clients do not have personal loans, 15.17 % do have personal loans and the rest are unknown. Most people have credit in default, likely due to the economic environment at the time and campaign calls were made pretty evenly during the week, close to the same amount of calls per day.

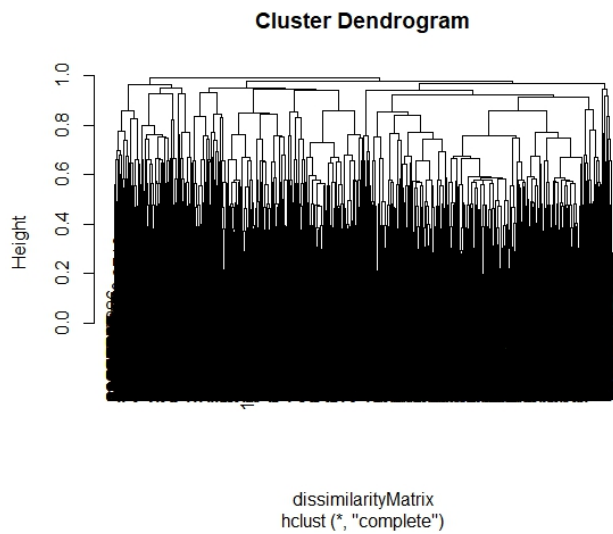
### 3.5 Cluster Analysis

Cluster analysis was an important step in the data analysis process, as the goal was to discover potential demographic groups in the campaign's client list. There

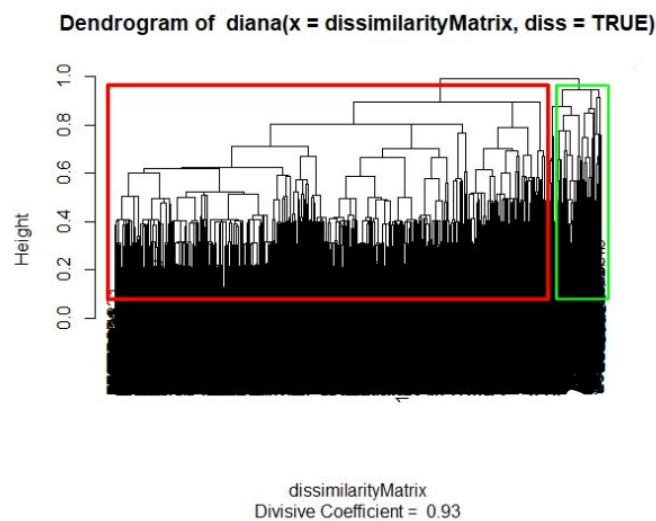
were over 40000 client observations in the dataset. Computing a dissimilarity matrix with multiple categorical levels for most variables was not computationally feasible, as running out of memory was an issue. Not having a computer with larger memory, the solution was to randomly select a sample size of 10000 clients for the analysis. The attributes included into the cluster analysis were all related to demographics, and additionally some attributes related to last contact of the current campaign was also added.

It is important to note that Gower's distance was the metrics used when creating the dissimilarity matrix, as the usual Euclidean distance could not be used due to the presence of categorical variables. There are two different and opposite approaches to clustering analysis, bottom up and top down. Bottom up approaches start out with each observation being one cluster, then individual observations get merged into larger clusters using an iterate process. Top down clustering approaches start out with one large cluster and use different algorithms for splitting the original cluster into multiple smaller clusters.

First, hierarchical agglomerate clustering, a bottom up approach was applied to the dissimilarity matrix. The dendrogram could be seen in figure 9. One can see how there are too many clusters to explore and a further look into the clusters reveals that this bottom up clustering approach does not distinguish between clients who subscribed to a term deposit. After noticing that a bottom up clustering approach does not work, naturally a top down clustering should be applied. Divisive

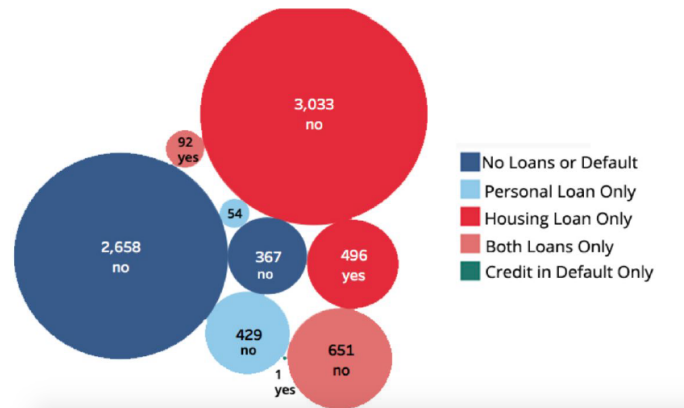


**Figure 9: Dendrogram from Hierarchical Cluster Analysis**



**Figure 10: Dendrogram from Divisive Cluster Analysis**

clustering is such an approach, and judging from the dendrogram from figure 9, it was expected to produce a more interpretive output. Figure 10 shows the dendrogram from divisive cluster analysis. Not only that the cluster memberships distinguish between clients who subscribed to a term deposit, but divisive clustering also computes a divisive coefficient, which is the algorithmically determined height where the dendrogram tree



**Figure 11: Subscription (Yes/No) based on loan history**

structure should be cut. Here, the divisive coefficient is 0.93 and cutting the dendrogram at this height would result in 4 clusters. The large cluster colored in red represents all clients not subscribing to a term deposit, while the 3 smaller clusters colored in green are the clients who subscribed. These clustering results fitted our task in a better way, as it created an opportunity to visualize some potential demographic groups using Tableau.

This visualization in figure 11 was created in order to explore any potential relationship between client subscriptions and loan history under the assumption that it may mirror trends in predatory lending, where lower-income and indebted people are targeted for financial products. The visualization demonstrated that regardless of loan history, a majority of the sample declined a term deposit subscription, with the majority of clients agreeing to a term deposit only having a housing loan, which is fairly common.

In a similar vein, the visualization in figure 12 sought to find a relationship between subscription success rate based on occupation, with the assumption that it would be most successful amongst people with lower-income jobs as a result of predatory banking practices. In the sample, a majority of clients across occupations said no to a term deposit, with the largest amount of "yes" responses being found in middle income occupations such as administration and blue collar work, who are likely concerned with saving money but also have some money available to be put into a term deposit that cannot be withdrawn from for a certain period of time.



Y	Admin	Blue Collar	Entrepre..	Maid	Management	Retired	Self Employ..	Services	Student	Technician	Unemployed
no	2,189 21.89%	2,103 21.03%	295 2.95%	224 2.24%	640 6.40%	308 3.08%	355 3.55%	869 8.69%	137 1.37%	1,455 14.55%	214 2.14%
yes	335 3.35%	150 1.50%	34 0.34%	34 0.34%	75 0.75%	100 1.00%	31 0.31%	80 0.80%	76 0.76%	190 1.90%	24 0.24%

**Figure 12: Subscription (Yes/No) based on occupation**

### 3.6 Classification

For the task of classification, the main goal was to be able to fit a classification model that could predict whether or not someone would sign up for a term deposit. This project sought to also find out which attributes are considered important when a potentially successful model would predict subscription success rates. Based on the mixture of categorical and numerical attributes, two main groups of classification algorithms could be fitted. One of the groups used all available variables, thus both categorical and numeric variables, and the other group of algorithms was only able to use numerical values, as some algorithms cannot deal with categorical variables.

The first group of algorithms included tree based models such as random forest, bagging, adaptive boosting, gradient boosting and classification (decision) trees, a probabilistic approach called Naive Bayes, and a regression based approach, logistic regression.

**3.6.1 Classification (or Decision) Trees.** This method uses binary splits to grow a decision tree, making a decision at each branch based on threshold values of predictors. In a classification tree model, each observation is predicted to belong to "the most commonly occurring class of training observations in the region to which it belongs" [4].

**3.6.2 The bagging.** This method applies re-sampling with replacement (bootstrapping) to the training set, thus creating multiple training sets that the model can be trained on. The predictions of the models are averaged over all bootstrap samples, thus obtaining a stronger classifier, which is known for reducing the variance of a predictive model [4].

**3.6.3 The random forest.** Random Forest is similar to bagging, as it uses random sampling from the original dataset, and only considers the a random subset of all predictor variables when building the decision trees. In this method, numerous decision trees are fitted, each based on a different re-sampling of the original training data. In the random forest algorithm, there are 2 major assumptions, namely that most of the trees can provide correct prediction of the class for most of the data, and that the decision trees are making mistakes at different branches. These 2 assumptions help the model to decide what class to predict based on what the majority of the decision trees are predicting.

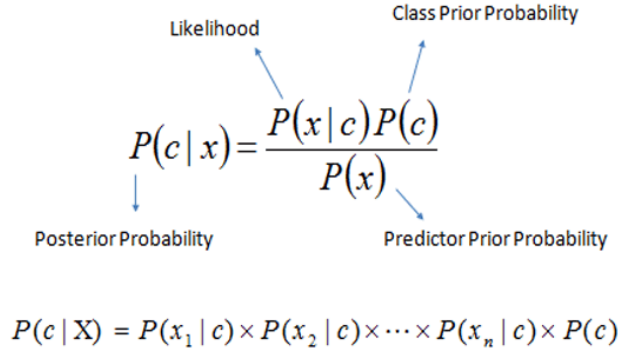
**3.6.4 Adaptive Boosting.** Adaptive Boosting is an iterative tree based algorithm, and it is a simple variation on the bagging algorithm. It is known for improving the learning process where the system is not performing well. The main concept behind the algorithm is that it iteratively learns from misclassifications of previously fitted models. In adaptive boosting each tree within the model (called weak classifier) is fitted on random subsets of the original training set, without bootstrapping, then finally the iterative models are added up to create the final model (called the strong classifier) [4].

**3.6.5 Gradient Boosting.** This method involves the introduction of loss functions and a process similar to gradient descent into the general boosting algorithm. Here, the objective is to minimize the loss of the model by adding weak classifiers using a gradient descent like procedure. The weak classifiers are still decision trees, and the loss function used is binary cross entropy loss (negative log likelihood loss). Tree constraints to number of trees, and depth of the trees can be added to combat over-fitting.

**3.6.6 Naive Bayes.** Naive Bayes is a classifier model that uses Bayes' rule to determine the probability that an observation belongs to a class, given some observed values for the predictor variables.

In figure 13 the posterior probability of a class given some observation is calculated using the product of the predictor observation's likelihood given the class and the class' prior probability, then it is divided by the predictor observation's prior probability.

**3.6.7 Logistic Regression.** Logistic Regression is a regression model used for binary classification. This model calculates the probability that an observation belongs



The diagram shows the Naive Bayes classifier formula  $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$  with arrows pointing from the terms to their labels:  $P(x|c)$  is Likelihood,  $P(c)$  is Class Prior Probability,  $P(c|x)$  is Posterior Probability, and  $P(x)$  is Predictor Prior Probability. Below this, the joint probability formula is given:  $P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$ .

**Figure 13: Naive Bayes classifier**

to a particular class. It uses the logistic function in equation (1) [4]

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}} \quad (1)$$

to force the output to a probability, then it can be converted into a 0/1 binary output representing the two classes.

The second group of algorithms, which were only used with numeric attributes were Linear and Quadratic Discriminant Analysis, L1-norm and L2-norm regularized regression models: Lasso and Ridge regression and other methods like support vector machine and k-nearest neighbour.

**3.6.8 In Linear and Quadratic Discriminant Analysis .** the discriminant function is the driving force behind the predictive power of the model. Equation (2) [4] shows the calculation of the discriminant function,  $\hat{\delta}_k$ , which decides to what group does an observation belong to. There are multiple parameters for the discriminant function:  $\hat{\mu}_k$  is the mean of the group  $k$ , while  $\hat{\sigma}_k$  is the variance of the group  $k$ , and  $\hat{\pi}_k$  is the prior class membership probability of a group  $k$ .

$$\hat{\delta}_k(x_i) = x_i \cdot \frac{\hat{\mu}_k}{\hat{\sigma}_k^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}_k^2} + \log(\hat{\pi}_k) \quad (2)$$

The linear discriminant classifier creates a linear decision boundary between  $k$  classes by estimating the mean  $\hat{\mu}_k$  and variance  $\hat{\sigma}_k$  of a group  $k$ , then it requires to know or estimates the prior class membership probability,  $\hat{\pi}_k$ . After the parameters have been estimated, the classifier plugs in the estimates for  $\hat{\mu}_k$ ,  $\hat{\sigma}_k$  and  $\hat{\pi}_k$

into equation (2), and assigns an observation  $X = x_i$  to the class for which  $\hat{\delta}_k$  is the largest.

The quadratic discriminant classifier is similar to the linear discriminant model, except that it creates a quadratic decision boundary by assuming that each class  $k$  has its own covariance matrix.

**3.6.9 Ridge Regression .** is similar to least squares regression, except that this model includes a penalty in the estimation process by adding an extra term containing a tuning parameter  $\lambda$ . Ridge regression is also known as L2-norm regularized regression, or shrinkage method.

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (3)$$

In equation (3) [4] the first term is the least squares estimation, while the second term is called the shrinkage penalty, and it is known for the effect of taking estimates of  $\beta_j$  close to zero and shrinking them towards zero. This process gets rid of unnecessary, useless parameters and creates an overall better, (hopefully) more accurate model.

**3.6.10 LASSO .** LASSO stands for 'least absolute shrinkage and selection operator', and it is the L1-norm regularized regression. It is very similar the previous shrinkage method, except that has an absolute shrinkage penalty, which forces some  $\beta_j$  coefficients to zero, whereas ridge regression will only force them to close to 0. In equation (4) [4] one can see the absolute value being applied to the  $\beta_j$  coefficient inside the shrinkage penalty term.

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (4)$$

**3.6.11 Support Vector Machine.** Support Vector Machine is a separating hyper-plane classifier, where the most important training points become the "support vectors" and they create a margin between the two classes. This model's objective is to maximize the margin between the training points, and that requires the maximization of the separation between the two classes, which is usually using a hyper plane, not linear space. Support vector machines can be linear or non-linear. The linear model operates as described above, while



**Table 1: Performance of models fit on categorical and numerical predictors**

Model	F1 Score	Precision	Recall
<b>Logistic Regression</b>	0.1717	0.7974	0.0962
<b>Random forest</b>	0.3522	0.2526	0.5817
<b>Bagging</b>	0.3543	0.2693	0.5180
<b>Adaptive Boosting</b>	0.3410	0.2609	0.4921
<b>Gradient Boosting</b>	0.3229	0.2171	0.6303
<b>Naive Bayes</b>	0.4264	0.6116	0.3273
<b>Decision Trees</b>	0.2696	0.1649	0.7383

the non-linear model uses feature space mapping using kernel functions.

**3.6.12 *k*-nearest neighbour.** This method is a simple machine learning technique that measures pairwise distances between data points, then classifies a new observations to the class which the majority of *k* nearest neighbours belong to, given that *k* is a positive integer between 1 and number of observations.

## 4 RESULTS

When evaluating the classification models, it was decided that precision is the measurement that matters the most for our project. Precision looks at the situation when the model predicts "YES" (meaning that the client will subscribe to a term deposit) how often does it predict correctly, meaning that it measures how well the model avoids false positives.

Most classification models performed poorly at predicting if a client will subscribe to a term deposit. Table 1 and 2 show the performance of the classification models. Logistic regression has a very high precision score, but it's recall value is really low, as it turns out this model predicts that a client will subscribe to a term deposit in over 90% of the cases. This model can be discarded. Tree based models like simple decision trees, random forest, boosting and bagging have not worked either, as their precision is in the range of 0.25 - 0.26, while the F1 scores are around 0.34 - 0.35. Tree based models seemed to not work for our data set. The naive Bayes classifier seemed to be the only classifier that could achieve a decent precision score of 0.6116, but unfortunately it suffers of low recall score. This probabilistic approach seems to work because the assumption of independence seems to hold, and the unbalanced prior

**Table 2: Performance of models fit on numerical predictors only**

Model	F1 Score	Precision	Recall	Specificity
<b>LDA</b>	0.4040	0.3340	0.5111	0.9161
<b>QDA</b>	0.4673	0.4780	0.4570	0.9308
<b>SVM</b>	0.2243	0.1440	0.5073	0.8970
<b>k-NN</b>	0.2355	0.1586	0.4570	0.9308
<b>Lasso</b>	0.1687	0.0981	0.6025	0.8930
<b>Ridge</b>	0.1760	0.1023	0.6282	0.8936

class probability is influencing the posterior distribution in a positive way, making the Naive Bayes model to output more accurate predictions.

For other models, linear and quadratic discriminant classifiers seemed to work better than any other method tried out in Table 2, as their F1 score was larger than Lasso, Ridge SVM and k-NN models. Lasso and ridge regression seems to work better for recall, but does very poorly on precision, so those models can be discarded. All models in general have high values for specificity, which is measuring how accurately can the model predict if someone will not subscribe. This is not a useful attribute to consider for our models. Surprisingly the both the SVM and k-NN model have similar precision scores, in the range of 0.14 - 0.15, which is really poor.

An observation worth noting is that the failure of close to a dozen classification models hints that the explanatory (predictor) variables do not explain the output variable. The overall conclusion drawn from the data is that from an analysis prospective, it is difficult to predict the outcome of an event when the attributes explaining what factors are impacting the outcome of the event are left out of the technical analysis.

## 5 DISCUSSION AND IMPLICATIONS

The following section will discuss the implications of this project as well as the need for future work in this area. In particular, this project seeks to show that both technical and social analysis must be done in order to achieve meaningful and well-rounded results.

**5.0.1 'Reverse-Engineering' Targeting.** In terms of future work, it is critical to re-examine the role of the data science researcher from an ethical perspective. As the gaps in the literature review showed, oftentimes

this kind of work is done in the direct benefit of institutions or companies due to funding and resources. As data scientists continue to be in high demand [3], it is critical that data scientists' increasing power is used for good rather than nefarious purposes. This is becoming a major issue in areas such as predictive policing, where fears of algorithms that perpetuate racial profiling and heat-mapping of low-income communities pose serious threats to vulnerable populations who are at risk of being predicted to be a criminal due to where they live or their race. [9]. In the context of this project, if the classification algorithm was able to successfully identify demographic groups, how can researchers use this information for good rather than predatory purposes? Especially if the algorithm was able to identify groups of vulnerable people (low-income, low education, multiple sources of debt, racialized or gender), this information could potentially be used to empower these groups through financial literacy campaigns and awareness.

## 5.1 How Data are Inherently Political

This bank marketing data, as well as this project, provide an excellent case study for why and how data are inherently political. Seemingly neutral bank data have politics embedded within them and can be easily weaponized in the wrong hands, as seen by the greedy actions of financial institutions during the financial crisis. Furthermore, this project opens the discussion that critical data studies, science and technology studies and communication scholars are having about data to a wider technical audience who ought to be paying attention and cognizant to these issues in order to build better, more equitable systems. Future work in data science should be conducted with both technical and non-technical scholars in order to achieve this balance.

## 6 CONCLUSION

In conclusion, this project sought to use a variety of technical methods to classify and cluster bank marketing data in order to find any relevant demographic groups that may respond well or poorly to telemarketing and to see if any of the trends of the financial crisis such as predatory lending were mirrored in the bank's telemarketing strategy. The project used technical methods such as Naive Bayes, Logistic Regression, Adaptive

Boosting, LASSO and Support Vector Machine amongst others to attempt to predict which clients would say yes to a term deposit. These findings were contextualized into the financial crisis of Portugal through critical data studies analysis.

Ultimately, the technical results were inconclusive due to the technical analysis being constrained to a sample size that was not representative of the dataset at large. In the future, it is vital for the technical component to have stronger computing power in order to cluster the entire dataset or potentially find a more representative sample. For example, the sample size was not representative of the amount of people contacted who had credit in default, which made up a significant amount of the clients. Had the sample size been more representative, the project could have better examined if these vulnerable people were being preyed upon to subscribe to term deposits when they were already likely experiencing financial troubles. As stated in the discussion section, future work is to be done in this area both technically and socially to see if the capability to build a similar algorithm to what banks already deploy is possible.

However, what this project has been able to achieve is to show the need for data science to work with other disciplines in order to fully take into account all of the aspects of the dataset, such as the politics embedded within them. For example, when embarking on this project, a technical scholar may not immediately notice the significance of the dataset's date and location of Portugal between 2008-2013. Without delving into this important historical context, wrong conclusions are likely to be made and the analysis will be missing a lot of key components that make the research relevant to scholars outside of technical disciplines. This focus on interdisciplinarity is important for data science as a fairly interdisciplinary field. The more that data science becomes palatable, intelligible and welcoming to less technical audiences will further enrich its findings. Furthermore, another important lesson from this project is that data about unique historical circumstances such as a financial crisis cannot and should not be used to make useful predictive algorithms as they are hyper-specific to a period of time that cannot always be extrapolated to other periods of time. Overall, data cannot be separated from the social, political and economic contexts that inform them and the future of data science will require "both technical and philosophical research" in

order to make sense of this ever expanding and highly in-demand field, hopefully for the benefit of society at large[5].

## REFERENCES

- [1] 2011. BPN: Oliveira Costa vendeu a Cavaco e filha 250 mil acoes da SLN. *Expresso* (2011). <https://expresso.sapo.pt/economia/bpn-oliveira-costa-vendeu-a-cavaco-e-filha-250-mil-acoes-da-sln=f643506#gs.U094mKA>
- [2] James Chen. 2017. Term Deposit. (Dec 2017). <https://www.investopedia.com/terms/t/termdeposit.asp>
- [3] Thomas H. Davenport and D.J. Patil. 2012. Data scientist: the sexiest job of the 21st century. *Harvard Business Review* 90, 10 (Oct 2012). <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>
- [4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. [n. d.]. *An introduction to statistical learning*. Vol. 112. Springer.
- [5] Rob Kitchin. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.
- [6] Sergio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.
- [7] Sergio Moro, Paulo Cortez, and Paulo Rita. 2018. A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems* 35, 3 (2018), e12253.
- [8] Scott Patterson. 2010. *The quants: How a new breed of math whizzes conquered wall street and nearly destroyed it*. Crown Business.
- [9] Walt L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, and John S. Hollywood. 2013. *Predictive policing: the role of crime forecasting in law enforcement operations*. RAND Corporation.
- [10] João Pedro Pincha. 2015. Caso BPP. Bens no valor de 4,7 milhoes de euros arrestados a Joao Rendeiro. *Observador* (May 2015). <https://observador.pt/2015/05/06/caso-bpp-bens-no-valor-47-milhoes-euros-arrestados-joao-rendeiro/>
- [11] Ricardo Reis. 2013. *The Portuguese slump and crash and the euro crisis*. Technical Report. National Bureau of Economic Research.