Abstract

The focus of this report aims at building a classification prediction model, capable of being able to predict whether an individual would subscribe to a term deposit. The bank marketing dataset used spans across 3 years, ranging from 2008 to 2010. Data analysis and feature selection allowed for the most important variables to be selected; 7 continuous, 3 categorical. The dataset was then split on 70% training and 30% testing. This report utilised the K-Nearest Neighbour algorithm, and generated the training model to find the most suitable K-value through a 10-fold cross validation. Through the Minkowski distance with p=2, it was found that k=6 output the most ideal results. The training model had an accuracy of 70.3%, whereas the final model using the testing set at k=6, had an accuracy of 71.4%. The sensitivity and specificity of the final model were found to be 79.4% and 66.7%, respectively.

Introduction

The marketing of a company is one of the most important things, as it allows the company to reach not only old customers, but potentially new ones. Often at times, this effort is carried out by specific branches, such as telemarketing. Usually, it falls down to the telemarketer whether they can convince the person to become a customer. If they succeed, the company prospers with a new client. However, if they do not succeed, that is time wasted which leads to an ineffective effort. Therefore, before a call is made, there should be some insight into who is being called, and the likelihood that they would become a customer. Not only does this decrease the number of resources used, but additionally offers higher profits. Such targeted marketing can generate insight from a prediction model, which provides groups of people that would more likely be convinced. For this report, the focus is on banking information, and how certain information can help generate such a model. Specifically, through the use of a dataset, a classification model will be generated that allows the bank to determine whether someone (within targeted marketing) would open an account with them.

Dataset Description

The dataset which the classification prediction model will be based off is a real-life dataset called “Bank Marketing Data Set” (Moro et al., 2014) found from UCI’s Machine Learning Repository from 2014. With its focus being on business and marketing. The data collected is from a Portuguese bank, and is a result of the several marketing campaigns carried out throughout the course of 3 years, ranging from early 2008 to late 2010. Potential customers were contacted via phone calls, either being cellular or telephone. Phone calls were made to contact a potential customer, and see whether they would open a bank account and subscribe to a term deposit. The dataset has over 40,000 entries with 20 attributes. Such variables can be used to create a classification prediction model.

The variables in the dataset range from categorical to numerical. A large portion of the data is revolved around the banking information of the person, including age, marital status, occupation, education and other factors such as credit status and loans. Furthermore, the dataset contains various pieces of current campaign (at the time of the dataset) information from the last time that the client was contacted. Meaning what month, and weekday, duration of call and how they were contacted. Other miscellaneous pieces of data consist of how many times they were contacted as of current campaign, how many times in the last campaign, number of contacts from last campaign and duration since last contact from last campaign. Lastly, the dataset provides various less individual-based factors of

Portugal at the time, describing the state of Portugal in terms of socio-economic perspectives. These were factors generated from a national standpoint by Banco de Portugal and include rates of variance in employment, the consumer price index and confidence index, Euribor rate based on 3 months and finally the number of employees. There are several attributes here, and there needs to be a selection for input to the model. Further insight must be required, to find out which variables are most relevant.

Dataset Analysis

A classification prediction model will be created, from the variables mentioned above. It is hoped that a combination of continuous and categorical variables may be applied to attain the most ideal result. Hence, the dataset analysis section refers to both types of variables, which would be deemed most suitable. The first step taken was towards the analysis of the continuous variables. The continuous variables were extracted, followed by normalizing them and inputting them to a heat map. Through this heatmap, correlations may be looked into to find similarity. To note for variable understanding: campaign – number of calls throughout the campaign, previous – number of calls during last campaign, emp.var.rate – variation in employment (quarterly), cons.price.index – monthly value on the consumer price index, cons.conf.idx – monthly value on the consumer confidence index, euribor3m – Offered rate every 3 months by Euro Interbank and lastly, nr.employed – quarterly monitored amount of employees.

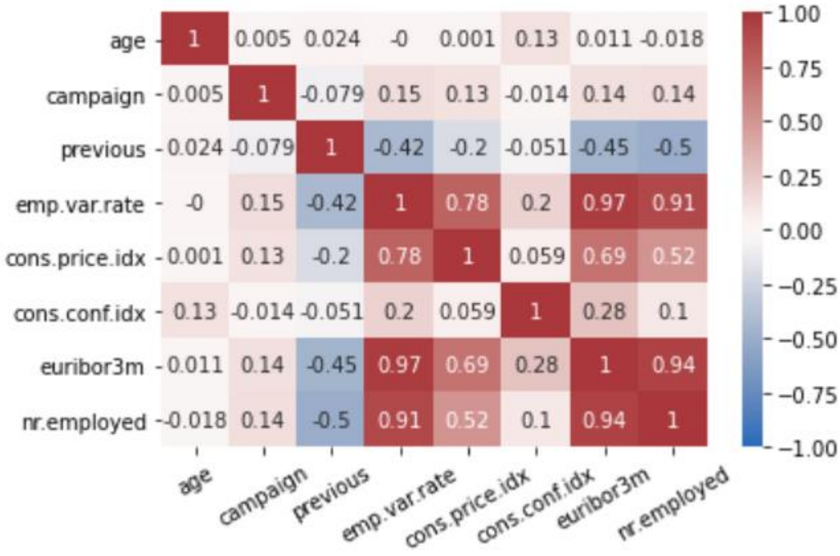

Figure 1: Heatmap of all continuous variables against each other

The heatmap utilised Pearson's correlation, as all the data had been normalized. From there results, it can be seen that there are no strong correlations outside of the socio-economic factors. The high-value correlations can be attributed to the fact that they are not a result of particular individuals. These are economic and social measurements of an entire nation. The "euribor3m" variable stands out the most on this chart because it has some of the highest correlations (0.97, 0.69, 0.94) with other continuous variables. Hence, it is believed this is a very influential factor, however due to its high correlation (0.97) with emp.var.rate it is almost exactly the same. Meaning that having both variables in the model would be redundant. Therefore, the emp.var.rate was removed. The same could be applied to the "nr.employed" variable (0.94 correlation), however since it was lower it was kept. It was also chosen to remain so that the socio-economic factor overall would be most accurate, to better determine when to make calls.

Other numeric variables which belong to specific individuals were age, campaign and previous. It is important to look into these because they provide relevant information on how old are most clients and how many times were they contacted on average. For example, age may indicate when it would be best deemed in a person's lifetime to call them about term deposits. This category may even be split into various groups, to include a range instead of specific numbers. However, for the purpose of this report it was kept continuous. It was found that the mean age in general for contacting individuals was 40.2, and the average age of people who said yes to a term deposit was 40.9. Interestingly, the highest number of calls were made to individuals between the age of 30 and 40, which demonstrates that age certainly has an impact on who is most likely to say yes. Hence, this variable was deemed appropriate to include in the model.

The "campaign" and "previous" variables were also looked into. Questions applying to these could include does calling more often help convince a potential customer? Or is it deemed less successful? On average at least 2 calls were made to every individual, with no prior calls being done from the previous campaign. It was further investigated with individuals that said yes to the campaign. It was found that half of the first calls resulted with a subscribed deposit, with 88% of deposits being subscribed by the 3rd call. As the number of calls increased, more people were convinced, however the rate begins to slowly come to a decrease (more calls – less numbers of total people). The same was found with the calls from the previous campaigns. Most people (3141 out of 4640) weren't contacted during the last campaign, however they were convinced. With 96.1% of subscribed people being convinced with at least 2 calls from the previous campaign. Taking this into account, these two variables were deemed important to include in the model, because it would allow to see the success rate with prior contact being taken into account.

There were 7 continuous variables chosen to be included; age, current campaign calls, previous campaign calls, employment variation rate, customer confidence and price index, the Euro Interbank rate, and number of employees. Giving a fair divide between socio-economic factors (4) and individual based (3). However, there are other potential variables that can influence the data and could make the model more accurate. These potential customers belong to specific groups, and these can be analysed to see if they impact the effect of attracting a client. Questions appear such as does education matter, which job occupations were most prominent, how does the financial situation impact a customer; among others were deemed important to gathering information which is crucial to getting a more accurate model. First, occupation was looked at.

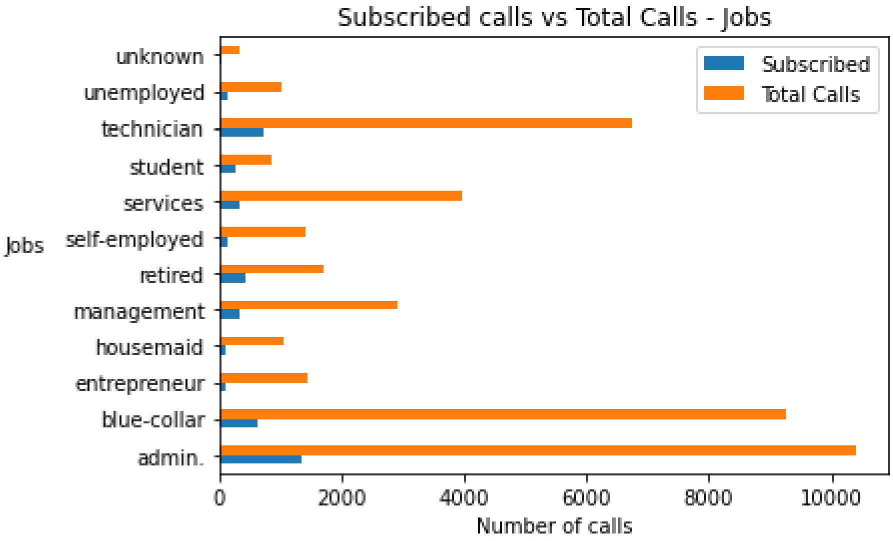

Figure 2: Number of calls (total and subscribed) vs job

Two things were noted down. The number of people who subscribed a term deposit when called, and the total number of calls. This way, it is not only possible to see which occupation had the most yes answers, it is also possible to see the rate. The rate in this case is important as it allows to make a more concrete decision of whether the job variable should be included. If a large enough difference may be seen between specific jobs, then the general consensus can be that the data supports the premise that jobs matter in the final model. In this case, the admin job had the most subscribers (1352), however it also had the greatest number of calls (10422). This means roughly 13% of admin workers, subscribed. See table below of results.

Jobs	Subscribed	Total	(%) of total subscribed
Admin	1352	10422	12.97%
Blue-collar	638	9254	6.89%
Entrepreneur	124	1456	8.52%
Housemaid	106	1060	10%
Management	328	2924	11.22%
Retired	434	1720	25.23%
Self-Employed	149	1421	10.49%
Services	323	3969	8.14%
Student	275	875	31.43%
Technician	730	6743	14.2%
Unemployed	144	1014	11.21%

Table 1: Job summary of results, showing the percentage rate of successful calls compared against total number of calls

Although the admin job had the highest number of calls, it didn't have the highest rate of success. Meaning that simply the number of successful calls alone cannot be taken into account. The group with the highest success rate were students, with almost a third (31.43%) of all calls being successful. Whereas the blue-collar job group had the lowest success rate (6.89%) despite having the second largest number of calls. The data in the table suggests that a future campaign should further investigate the importance of students. The average age of a student was found to be 25, which could imply that the reason the success rate is high is because most of the people in this group have not subscribed a term deposit before. It could be their first, and are convinced it is an efficient way to saved money, and let it grow over time. Overall, what the data gathered supports is that the difference between groups (especially min and max) are valid indicators that occupation does play a role in whether a customer subscribes a term deposit. Hence, the job variable was kept. Education was looked at next.

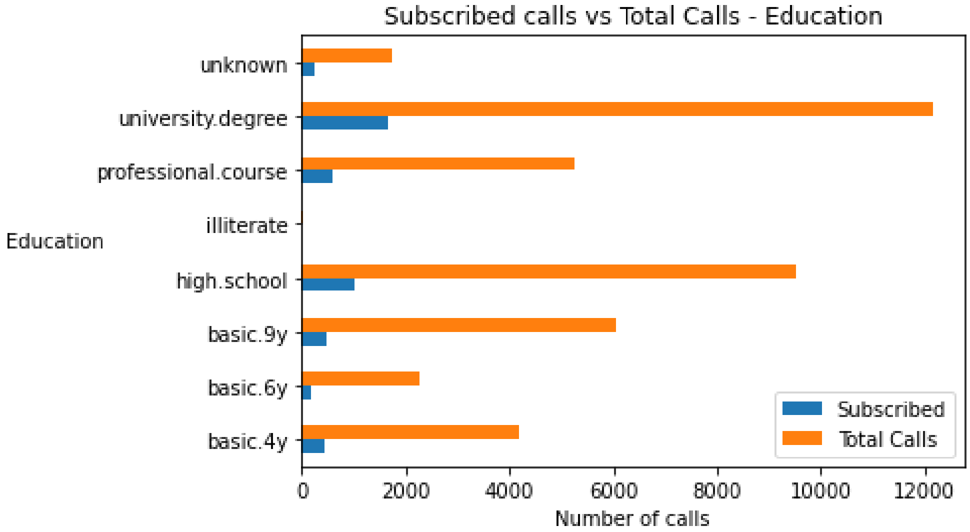

Figure 3: Number of calls (total and subscribed) vs education

Education	Subscribed	Total	(%) of total subscribed
Basic – 4 years	428	4176	10.25%
Basic – 6 years	188	2292	8.20%
Basic – 9 years	473	6045	7.83%
High School	1031	9515	10.84%
Illiterate	4	18	22.22%
Professional Course	595	5243	11.35%
University Degree	1670	12168	13.73%

Table 2: Education summary of results, showing the percentage rate of successful calls compared against total number of calls

In the education sector, it was found that the illiterate group had the highest success rate (22.2%), however due to its extremely small sample (4 successful, 18 total calls) it was deemed that there isn't enough data with this group to come to any concrete conclusion. Excluding it from the overall group. Interestingly, it was found that the university degree education group had the highest success rate (13.73%). In this case, the university degree group had the highest number of successful calls (1670), while also being the group that was called the most (Total calls: 12168). Meaning that the targeted audience was correctly aimed, and the success rate was likely due to more educated people knowing what was being sold to them, and how they feel it can benefit them. However, the difference was not as large this time. The lowest group, being the basic 9 years education group, had a success rate of 7.83%, which is barely 5% difference. Such information doesn't necessarily state that education plays a big role, however it can be combined with the job (since education level is often relevant to what job one holds) in order to gain a better insight. Another table was generated, combining both job and education.

Education / **Job**	Basic – 4 Years	Basic – 6 Years	Basic – 9 Years	High School	Illiterate	Professional Degree	University Degree
Admin	10	8	42	382	0	49	823
Blue-Collar	123	107	240	94	0	41	9
Entrepreneur	7	9	12	16	1	9	66
Housemaid	51	5	3	14	0	11	17
Management	5	10	11	17	0	8	257
Retired	185	10	19	62	2	57	66
Self-employed	3	1	18	8	1	20	96
Services	7	20	29	203	0	19	26
Student	8	7	35	114	0	17	35
Technician	9	6	37	85	0	343	225
Unemployed	16	4	26	34	0	20	39

Table 3: Education and Job combined showing numbers of successful calls in the categories

Generally, what can be gathered from table 3 when viewing specific jobs such as admin, is that it is possible to see that most people (60.9%) with that occupation have a university degree. The data is not distributed evenly, as was originally thought from looking at education only. While some jobs do have more similar numbers, it can be seen that in most there are specific education groups which stand out and have greater amounts. This can be visible in Blue-Collar and the basic education groups. Technicians with professional and university degrees. Services with mostly an education from high school. The data from the table reveals that education when combined with job, shows a difference is visible between different education groups. Hence, the education variable can be used with the prediction model, however it should be used with caution and checked for further effect on the subscribed calls. Continuing, financial status of an individual was looked at.

There were three categorical variables that may describe the financial situation of an individual from a dataset. Firstly, "default" which determines if an individual contains credit in default. Then a housing and personal loan. These would be checked to see if and how a financial situation may affect whether a person subscribes a deposit or not. It was found that every individual that subscribed a deposit had no credit in default. Whereas 2098 people had a housing loan, 274 had a personal loan and 409 had both. A total of 1569 people that said yes to the term deposit had neither a housing loan or a personal loan. These results indicate that default is not an ideal variable to use, as there are only "no" answers. Considering that housing had a more similar score, roughly half said yes, it would be harder to determine whether this has an impact on the answer. Hence, with a less even result with the loan variable, it was chosen as one of the categorical variables. The last factor that was looked into detail was marital status.

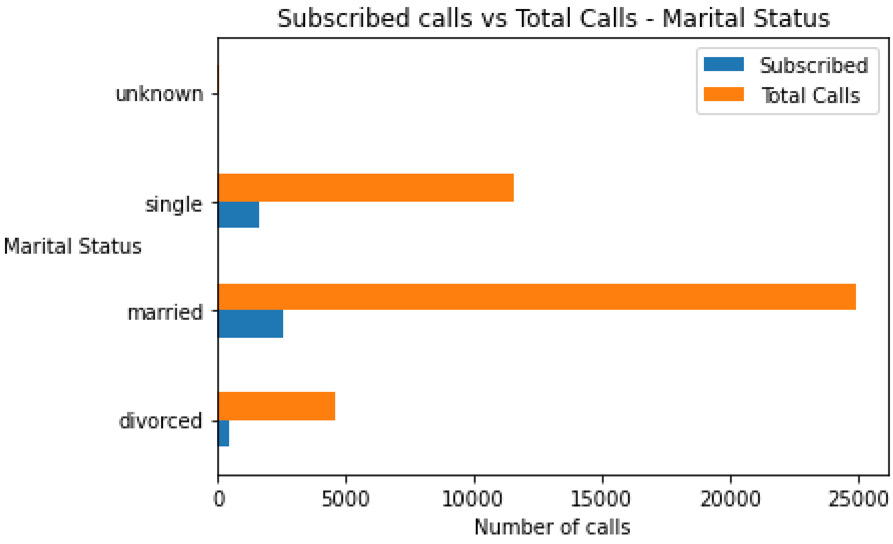

Figure 3: Number of calls (total and subscribed) vs marital status

Marital Status	Subscribed	Total	(%) of total subscribed
Divorced	476	4612	10.32%
Married	2532	24928	10.16%
Single	1620	11568	14%

Table 3: Marital Status summary of results, showing the percentage rate of successful calls compared against total number of calls

In regards to marital status, there did not seem to be any large difference. The largest difference (between divorced and single) was even smaller than that of education. The success rate of married people was 10.16%, and they were the group that was contacted the most. 2532 subscribed to a term deposit, out of 24928. The target audience for the campaign may need refinement. From the data present, mothing in particular sticks out in terms of which group should specifically be focused. Further research would be required, with more samples to have any conclusions. However, as this report refers to the present dataset, the findings from the marital status deemed it unsatisfactory for the prediction model.

Other factors were looked into such as time. There was not much change between the different years. It was specifically looked into what days would be most suitable to make calls. However, all days showed similar numbers and as such was considered unsatisfactory. At least, there was insufficient evidence (Mondays – lowest at 847 and Thursday – highest at 1045). The variables which were thought to have some relevance to the outcome were gone over. The data analysis looked into a variety of continuous and categorical variables; picking out those that were deemed most satisfactory and suitable to be used for the prediction model. The final variables remaining were: (continuous) age, current campaign calls, previous campaigns calls, employment variation rate, customer confidence and price index, the Euro Interbank rate, number of employees, (categorical) job, education and loan. However, the importance of these variables may still be subject to change. The data analysis section brought forth variables that seemed viable for the model, and may still be narrowed down further.

Feature Selection

The feature selection section narrows down the most important features and isolates them, to use for the prediction model. Albeit a number of variables were found which were deemed important through the data analysis section, the feature selection further checks the chosen variables' effect on the data. Through this, the model may be better built and only include features that are most impactful. However, before this was done the data needed to be adjusted. If "unknown" was found, it was removed from the dataset, as it was not useful. To gain the most fair and efficient results, the dataset was made that there was an equal number of "yes" and "no" answers to the subscribed deposit (output variable). In the end there were 8516 rows of data. The last part of the data modification, occurred by splitting the categorical variables, into numerical by making them binary values. The features were then standardized to be ready for use. The importance level of every feature was calculated, and plotted on a graph.

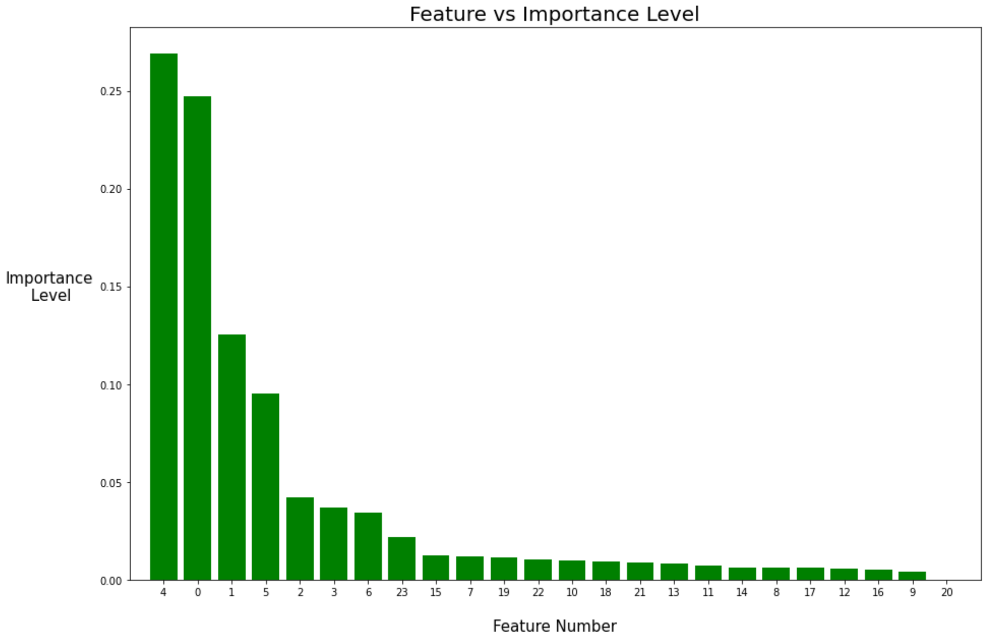

Figure 4: Importance Level of every variable/feature selected

A total of 24 features were input, and their importance calculated. Without looking at the labels, it can already be possible to see that two variables stick out. The highest values at 0.264 and 0.244. These are the features which are deemed most important to having an effect on the data. Table 4 contains the information of every feature, its importance level and ranking. What was found was that the continuous variables were the ones who had the highest importance, whereas the categorical variables only had a fraction of the importance. This process is important to follow, because as was mentioned above there should be some scepticism for the education group for example. The table shows that the highest-ranking education had was 11, with a High School education, and with an importance level of 0.0116. The data shows that the education group is not important enough to include. And the same premise goes for most of the job categories. Due to the pattern in the graph, the ending point was decided to be an important ranking of 11, as that is when the graph begins to steady out. 10 features would then be kept for the prediction model to avoid a more complex model.

Importance Ranking	Feature Number	Variable Name	Importance Level
1	4	euribor3m	0.269
2	0	Age	0.247
3	1	Campaign	0.126
4	5	Nr. employed	0.0952
5	2	Cons.conf.idx	0.0424
6	3	Cons.price.idx	0.0371
7	6	Previous	0.0343
8	23	Loan	0.0223
9	15	Job_ Technician	0.0129
10	7	Job_ Blue-collar	0.0121
11	19	Edu_High school	0.0115
12	22	Edu_University degree	0.0108
13	10	Job_Management	0.0099
14	18	Edu_Basic – 9 Years	0.0095
15	21	Edu_Professional Course	0.0090
16	13	Job_Services	0.0083
17	11	Job_Retired	0.0074
18	14	Job_Student	0.0063
19	8	Job_Entrepreneur	0.0063
20	17	Edu_Basic – 6 Years	0.0063
21	12	Job_Self-employed	0.0061
22	16	Job_Unemployed	0.0056
23	9	Job_Housemaid	0.0045
24	20	Edu_Illiterate	0.0004

Table 4: Importance level of all features previously selected, and their importance level

The following 10 features were kept for the classification prediction model: euribor3m, age, campaign, cons.conf.idx, cons.price.idx, previous, loan, job_blue-collar, job_technician. The rest of the variables were omitted, as the data found from the importance level graph revealed these 10 mentioned variables to have the most relevance and informative factors in regards to the output variable. An important factor to highlight, when viewing the combination of data analysis and the feature selection sections; the findings revealed that the most influential features were the continuous variables. The higher ones being the socio-economic factors. Albeit age, and campaign were high, such information can be interpreted to mean that whether an individual will subscribe or not is not necessarily tied to their individual qualities, but more economic and social situations and status at the time. Considering that this dataset takes place during and after the financial crisis of 2008, it means that the collapse of a market definitely influenced that data this report explored. Should such a report be repeated in the future, it would be interesting to see to what extent the environment had changed to see shifts between the importance of continuous and categorical variables. However, since this report focuses on results between 2008 and 2010 and takes these into account, there should be a greater focus on the socio-economic atmosphere while still taking some portion of an individual's life-qualities into account.

Classification Prediction Model

Finally, now that the variables had been selected it would be possible to create a prediction model. There are several different methods to creating a classification prediction model, however this report focuses on using one in particular. This report utilises the K-nearest neighbours algorithm, which focuses on the (k) amount of training datapoints closest to the new point and was chosen due to the amount of actual data available. Two sets would be created, one for training and one for testing. The split was done on a 70% (training) and 30% basis in order to achieve the best results. The training model would utilise the training set, and the final model would utilise the test set to make a comparison.

The training model in this case would help build the final model. The k value is not yet known; at least which value would be deemed best to use. This analysis uses a k-fold cross validation method to find different accuracy scores for different numbers of nearest neighbours. The training model with look into k-nearest neighbour accuracy values from 1-10, while using the training set. The training set will be further split through k-fold cross validation. The number of splits utilised was 10, and furthermore the nearest neighbours were calculated through the Minkowski metric, and sets the value of p=2, meaning it will represent a generalisation of the Euclidean Distance. Results of the training model can be found in the figure below.

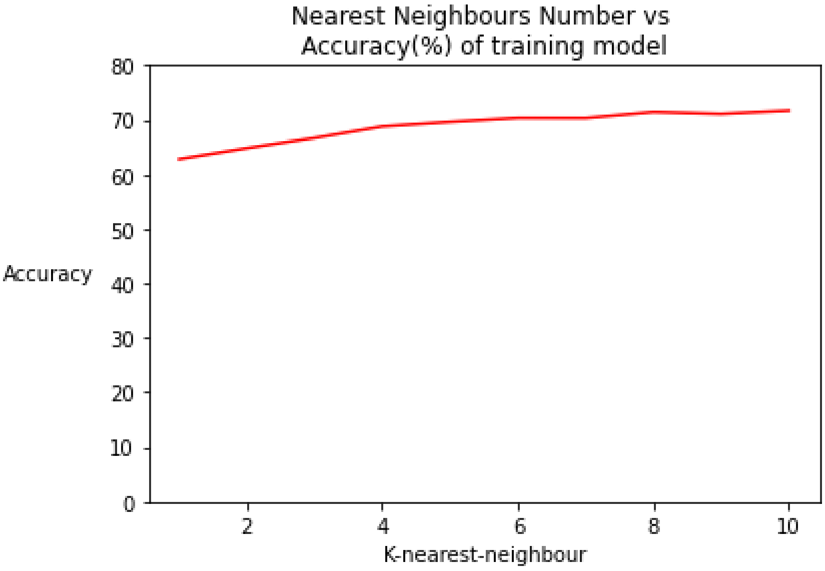

Figure 5: Graph demonstrating accuracy as model becomes increasingly complex

Overall, there wasn't any drastic change in accuracy as the number of nearest neighbours increased. However, what can be seen is that the largest difference occurs between K=1 (62.8%) and K=6 (70.3%). Indicating that the accuracy was not stabilizing until the K value was at around 6 or more. At that point the accuracy barely changes as at a K value of 10, the accuracy was 71.6%. Taking these factors into account, it was then chosen for the final model to use the nearest 6 neighbours to determine the result of the output.

When inputting 6 as the nearest neighbour value the final model, which was now working on the test dataset (30% of the data), it was found that the accuracy of the final model showed a promising result of 71.4%. Compared to the training set value at K=6, at 70.3% accuracy, it was a minor improvement. Albeit the results could still improve in terms of overall accuracy, these were still sufficient for a prediction model. Perhaps a more thorough data analysis would allow for better variables to be selected, or to combine those that were more relevant to allow more categorical variables to be used. Either way, the misclassification rate of the final model was 28.6%. These results imply that the final model would predict 71.4% of the observed results correctly. The results of the K-NN final model are visible below in a confusion matrix.

[redacted]		Observed	
		No	Yes
Predicted	No	1062	212
	Yes	514	767

Table 5: Confusion matrix showing observed and predicted values of "yes" and "no" based on output

From this information, two things can be further calculated. The sensitivity and specificity of the model. Sensitivity checks the rate at which the model can accurately predict whether a client would subscribe to a term deposit (True Positives), whereas specificity checks the rate at which the model can accurately predict that a client would not subscribe to a term deposit (True Negatives). The final model in this report calculated the sensitivity to be 79.4%, and the specificity to be 66.7%. The rates found are desirable in the case of the bank campaign, as the primary goal is to get customers to subscribe term deposits. False negatives (meaning a customer did want to subscribe, but the model thought they would say no) damage the campaign more than false positives. Each false negative prediction is money lost from a customer who would've subscribed. Whereas the same can be said for the other scenario, false positives (customer did not want to subscribe, but the model predicted they would) can waste time and other resources. However, considering this case scenario with the bank, this analytical report deems it more crucial to reduce false negatives, as they would be seen as more costly.

In essence, an end result of having sensitivity at 79.4% means that the classification prediction model created, based on all data collected, is a good indicator and setting for the bank to utilise. While accuracy was found to be lower, the higher sensitivity makes the model optimal for it to be deployed, however further revisions may still be warranted. It should be noted that the model will not get predict every client's answer correctly. Therefore, it should also be considered by those making the call to look into the factors (especially socio-economic) themselves and use the model's prediction as a guideline to attain the most ideal results. While the model may be improved and benefited from human interaction, it still demonstrates the model's learning capabilities are sufficient and outputs desirable results.

Limitations

Although conclusions were reached, which were supported by the data found during the course of this report, there are some limitations that should be taken into account. The data analysis section and feature selection show the process of how specific variables and categories within a specific variable were chosen. While the feature selection presented a more solid and mathematical approach to the importance of the variables, the initial variables selected during data analysis were solely based off interpretation of findings. Albeit, it was attempted to be done as detailed as possible, there could have still been something missed. This should not be ignored as missing something of importance may shift the results of the model entirely. Certain assumptions that were made may bring inaccuracy to the final model. For example, in the variable "campaign", it was not certain whether a value of 1 implied that there was 1 call before, or that it implied it was the first call. The assumption for this report was that it was the first call, however if different then interpretations and conclusions may be subject to speculation.

Furthermore, it was deemed important to have an equal amount of data for both "yes" and "no" answers for the output, to generate the most accurate model. However, when the "No" answers were limited, a random sample would be taken, hence the values would be changing leading to several different results for the model. Sometimes more accurate, sometimes less accurate. A total of 10 or even more runs should have been performed, to note down an average accuracy, misclassification rate, sensitivity and specificity. Such results would provide further credibility to the prediction model and its results.

Conclusion

To give an overview of this report, the general premise followed the process of getting data prepared, finding relevant pieces of information as inputs and then creating a reliable prediction model based on classification. The data analysis section of the report aimed towards narrowing down the number of variables that were deemed informative overall, however feature selection allowed for further limitation of input variables to the model through viewing their importance. In the end it was found that continuous variables weighed the most importance, including the age of the client, number of contacts during the concurrent campaign and the previous, the Euro Interbank 3 month offer rate, the consumer price and confidence indexes. Three categorical variables were found to be important, these being whether the client has a personal loan, and the jobs of most relevance were blue-collar and technician. These input variables were split into training and test sets. The training set was used for a training model in which the most ideal number of nearest neighbours was found through the use of k-fold cross validation. The classification final model slightly improved on the data gained from the test model, however the final model demonstrated a promising sensitivity rate. A high sensitivity rate was seen desirable to maximise profits and would be less costly. The overall consensus on the classification model was that it was deemed efficient enough for use and would provide an informative insight into which clients should be contacted, and which should not, to get the largest amount of subscribed term deposits.

References

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Banco de Portugal. (n.d.). estatisticasweb. Retrieved February 27, 2022, from https://www.bportugal.pt/