Bank Telemarketing Data:

Cheung Yu Jeng Ethan

A0239399H

Introduction

Banks sell financial products aiming to maximise their success rate and minimise the costs of contacting clients. An easy and efficient way to do this is by analysing available data on customers and forming models to predict how likely each client is to buy their product. By employing machine learning to perform these predictions, banks automate the selection process for telemarketing campaigns. This report aims to analyse a particular data set provided by an anonymous banking institution and discuss the performance, advantages and shortcomings of different methods used to form predictive models.

The task of predicting successes for future telemarketing campaigns is a classification problem. For this, the main parametric method employed was logistic regression (LR), which served as a standard for the predictive power of models. LR models using different variables are also compared with each other for variable selection. As explained later, LR_1 includes variables selected using domain knowledge and economic intuition, while LR_2 is selected by minimising the Akaike Information Criteria (AIC). Other models explored also include K-Nearest-Neighbour (KNN), Naïve Bayes Classifier (NBC) and regression trees (DT). Principal component analysis gave insight on more important variables and principal component regression (PCR) was attempted. The effectiveness of each method is measured by partitioning the dataset into training and testing sets. The models are then trained and tested using the split data, and the areas under each Receiver Operator Curve (AUC) produced, as well as the mean squared errors (MSE) of the models are compared. The results, shown in figure 1 below, reflect that LR_2 performed the best with an AUC of 0.737, however the other models performed comparably well. in the following sections, we will discuss findings from the dataset, analysis of each model as well as applications and limitations of this study.

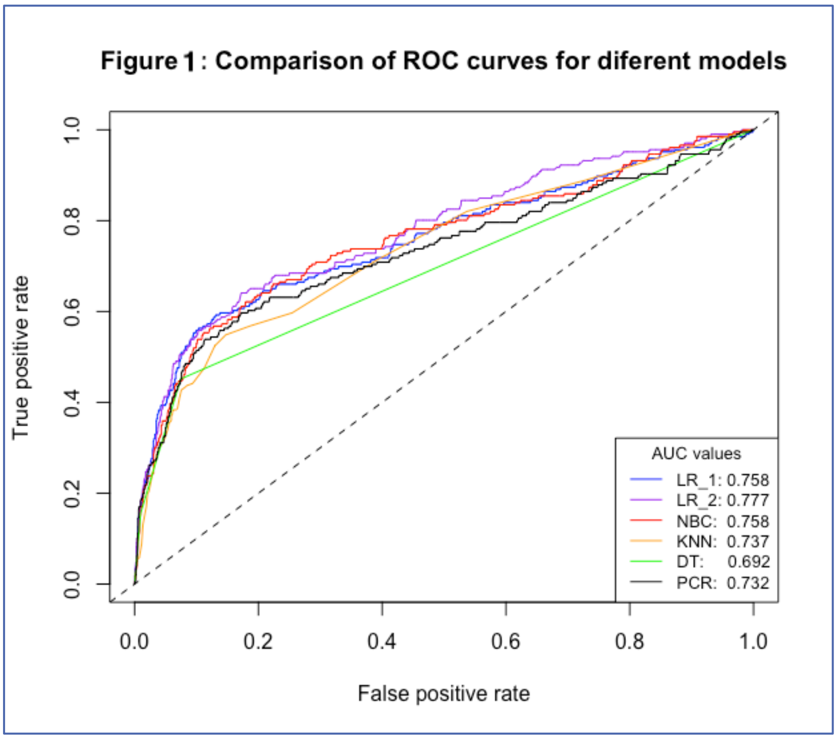


Figure 2: Comparison of models using area under ROC (AUC), mean squared error (MSE)

Method	PCR	LR_1	**LR_2**	KNN	NBC	DT
AUC	0.732	0.758	**0.777**	0.737	0.758	0.692
MSE	0.0775	0.0743	**0.0737**	0.0788	0.107	0.0774

The data

First, we clean the data. Some categorical predictors contained some missing values labelled “unknown”. We may choose to remove these unknowns with a simplifying assumption that future entries of client information is perfect, but in reality this is not true. Thus, we chose to include an “unknown” class for these predictors so that modelling is more lenient on future data gathering. Additionally, to remove the entries with unknowns will reduce our sample size from 4119 to 3085 which will reduce the effectiveness of model training.

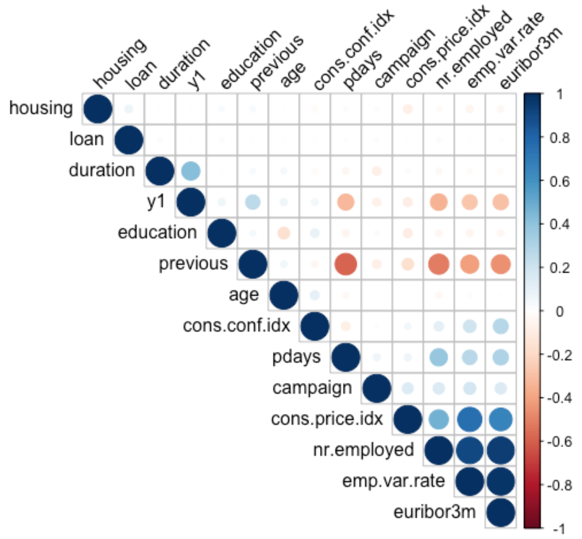
Next, dependent variable “Y” represents success in selling bank term deposit, which contains factors “yes” and “no”. We will convert this into dummy variable y1, 1 for “yes” and 0 for “no”.

The predictors “education” and “default” each has a class that only appears in one row. There will be a chance that the testing set contains the class, but the training set does not, which returns an error. For education, the class “illiterate” appears once and we remove the row. However, the class “yes” in default only appears once. The predictor default is of limited use if it is trained without any “yes”, thus we removed the entire predictor from the dataset.

We can transform categorical variables “housing” and “loan” into dummy variables, with 1 for “yes” and 0 for “no”. This allows us to resolve any missing entries by inputting the mean of all existing entries within the predictor. For “education”, we order the classes by level of education, from 4 years to university level, and convert and scale to produce numeric values. The assumption is that the “distance” between classes is constant, eg. The difference between “4 years” and “6 years” is equal to the difference between “professional course” and “university degree”. This allows models using “education” to accept unique entries outside of the original categories, for example “5 years” will be a number between “4 years” and “6 years”. We can take “illiterate” to be 0.

Using the package "corrplot" we can visualise the correlation between numeric variables, shown in figure 2. There is high negative correlation between "previous" and pdays" of -0.59. This makes sense as given a scenario where the client has no prior contact, previous will always register as "0" and pdays "999". We choose to keep pdays as it has a higher correlation with the dependent variable y1.

Figure 3:Correlation plot of numeric variables



"Euribor3m" refers to the 3-month Euro interbank Offer rate, this correlates strongly and positively with economic indicators such as CPI or consumer confidence, as lower interest rates stimulate consumption expenditure and investments, which will eventually positively affect employment. Therefore, to minimise the effect of multicollinearity we remove "Euribor3m".

The predictor "duration" must also be excluded. It is a strong predictor of y1, as seen in the correlation plot. However, it is not practical to include it as the duration of the call is unknown before the contact is made.

Finally, the data set partitioned into training and testing sets using a random 50-50 split.

Principal Component Analysis

PCA applied on numeric factors was unsuccessful in meaningfully reducing dimensionality of the dataset. Figure 4 below shows a scree plot, which showed that the proportion of variance captured in the first few principal components was not large enough, thus a gradual downwards slope. This may be explained by a low degree of correlation between variables. The first PC did manage to explain 28.7% of variance. From the biplot (Figure 5) below, we see that the highest loadings for PC1 were on the group of economic factors, including "euribor3m" and "cons.price.idx". This group of variables will continue to be among the more significant predictors used in later models. Modelling a linear regression on the produced PCs, we were able to achieve an AUC of 0.732, which sets a benchmark for the following supervised methods. Perhaps a more appropriate model would be to use the principal components on a logistic regression model.

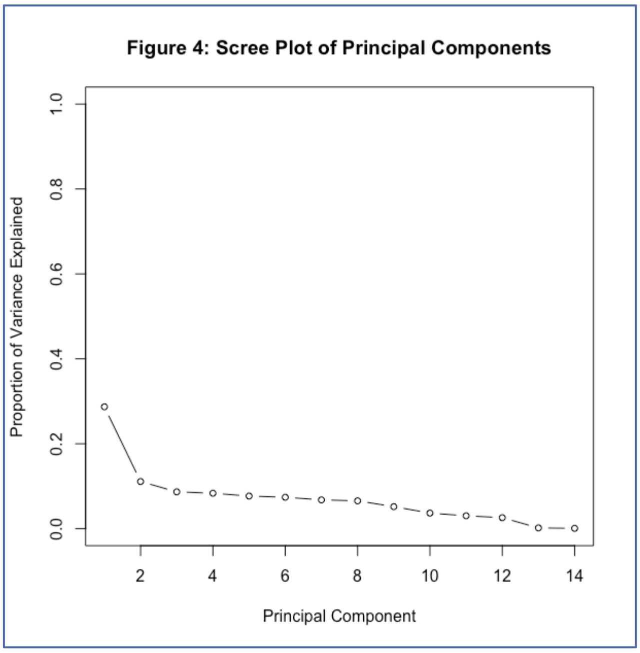


Figure 4: Scree Plot of Principal Components

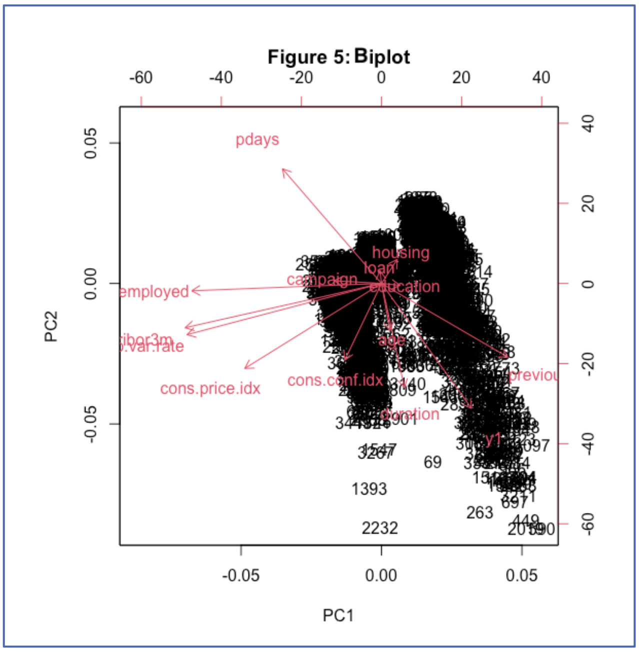


Figure 5: Biplot

Logistic regression

LR_1 uses logistic regression on all the variables, excluding the above-mentioned predictors: “duration”, “euribor3m”, and “previous”. Removing these variables reduced multicollinearity. Feature selection can also be employed to further improve the model, as seen in LR_2. LR_2 used the function “stepAIC” from the package “MASS”, which performs stepwise backwards selection. The Akaike Information Criteria (AIC) is a function of deviance which penalises on model complexity. At every step, the variable that will be removed will cause the greatest reduction to the AIC value. LR_2 performed the best out of all the models, with highest AUC of 0.777 and lowest MSE of 0.0737.

Using LR_2, the most significant predictors (highest Z-values) included mode of contact, outcome of previous contact, and economic factors such as CPI and employment variation rate. Logistic regression is one of the easiest models to interpret and modify. This is useful for banks as bankers with domain expertise can specify which predictors are to be included.

Naïve Bayes Classifier

The NBC performs relatively well at AUC of 0.758, despite a high degree of joint probabilities between variables not accounted for. NBC performs better when data has relatively many predictors, as these joint densities will get harder to estimate. Perhaps if more variables are found to be relevant in predicting y1, NBC will become a stronger model for banks to use.

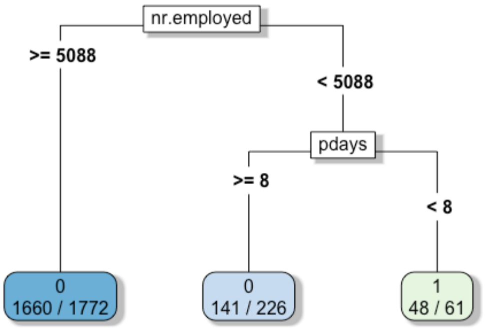
K-Nearest Neighbours

The model using KNN was able to achieve an AUC of 0.737. Using Leave-one-out cross validation, we determine the optimal number of neighbours to be 64. As KNN measures Euclidean distances between points, we are unable to include nominal variables or dummy variables that were deemed important for prediction in linear regression, such as mode and outcome of previous contact, thus a slightly worse relative performance. Additionally, the entire training set needs to be iterated through per datapoint. As the size of the dataset grows over time, this method will begin to experience computational inefficiencies.

Decision Tree

A regression tree is modelled by feeding all variables except for “duration”, using the package “rpart”. The resulting tree is simple, with only three terminal nodes. “pdays” and “nr.employed” are selected, thus only clients with number of days passed since last contact being less than 8 are identified, this during periods where the quarterly average number of employed citizens is less than 5088 thousand.

Figure 6: Regression Tree



For such a simple model, the tree performs relatively well with an AUC of 0.692. Trees provide good visual models for sales agents to follow if they are to manually select clients, however most banks may prefer to automate this process, and they lose out to other models in terms of predictive ability. An overly simplistic model may be problematic, for example in Figure 6, our tree predicts that if a national employment statistic is above a certain level, we should not perform any contacts at all. Such a generalised rule is not realistic. Still, trees may help elucidate more important predictors to be included in another model.

Concluding remarks

An important consideration would be the goal of classification. For standardisation purposes, this report uses a posterior probability threshold of 0.5, above which we identify a potential success. For each model, the accuracy maximising cut-off can be identified to further improve classification. For example, LR_2 has a maximum accuracy of 91.0%. This assumes the relative costs of false positive and false negative classifications are equal. In reality, the opportunity cost of a false negative, which is a missed sale, far outweighs the cost of false positive, which includes resources spent calling the client. Banks are likely to require a model which prioritises sensitivity, and additional information on relative costs is required to determine the optimal cut-off point. There may also be additional synergistic effects between variables, for example a retired person is less likely to commit to a term deposit, and even more so if he/she is also single without any dependents. This additive effect requires more domain expertise to identify.

Overall, it is likely that banks will prefer LR over non-parametric methods, for its high performance, customisability of features, as well as the ability to assess the relationship between y1 and its predictors.