**DSE 1101 INDIVIDUAL PROJECT**:

Analysis of Bank Telemarketing Dataset

Name: She Chee Yee

Matriculation Number: A0240383L

**1 Introduction**

Telemarketing calls are a common occurrence in our daily lives, often done to persuade consumers on the end of the phone to subscribe to some kind of financial product or plans. This form of marketing is present and utilised by the majority industries, and it is often thought as one of the most effective ways to drive sales. In recent years, rumours of telemarketing’s death are prevalent. Many argue that scam calls and the evolution of internet will render this type of marketing ineffective, however what many do not know is that telemarketing is evolving with the internet and it continues to do well today.

This project utilizes machine learning algorithms and data analysis techniques in an attempt to solve the million-dollar question in the telemarketing field, “What makes the customer bite?”. This dataset is obtained from Moro et al. (2014) with data on a telemarketing campaign by a Portuguese bank offering term deposit subscriptions. What will be attempted to predict is whether a customer subscribes to the plan, hence this is a binary response variable and binary classification problem. Various models will be used to and they will be evaluated by using the AUC score.

**2 Analysis / Data-Preparation**

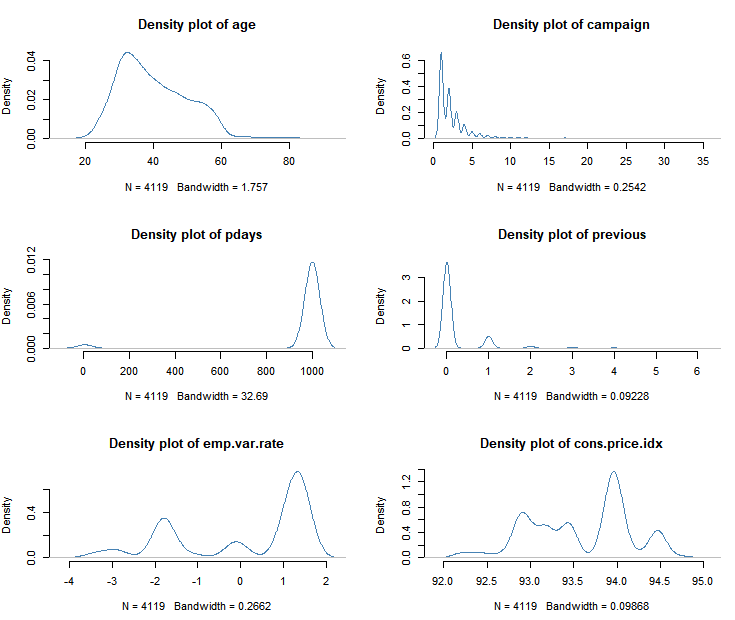
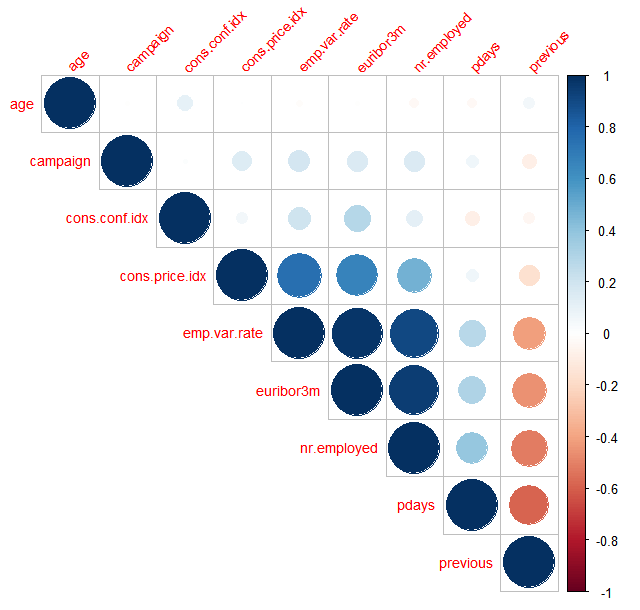
The data analysis started out by visualising and cleaning all the categorical predictor variables. Predictors “days\_of\_week” is removed as it is extremely unlikely that day of the week is able to influence whether one will purchase a product, predictor “duration” is also removed, this variable is not known before a call is performed and it highly influences the result, therefore it was removed. Some outliers like “illiterate” in education and “yes” in default are removed due to the very small sample and it is likely to be an outlier. There is a total of 9 categorical predictor variables ,9 nominal predictor variables and a response variable “y”.

To prepare the data, the response variable “y” is turned into a binary variable, with 1 representing “yes” and 0 for “no”. A subset of the dataset is also created for models that only takes in numeric inputs. The data is then split into train and test set (seed 103), 3200 observations are used for training set and the remaining for testing set for validation and evaluating the prediction. As the dataset is incredibly imbalanced (89% of “y” is no) measure of accuracy will not work, we will be using AUC, which is a measure of the performance of classification of the models.

Chart, bar chart, histogram, waterfall chart

Description automatically generated

**Figure 1: Boxplot of categorial variables with outliers**

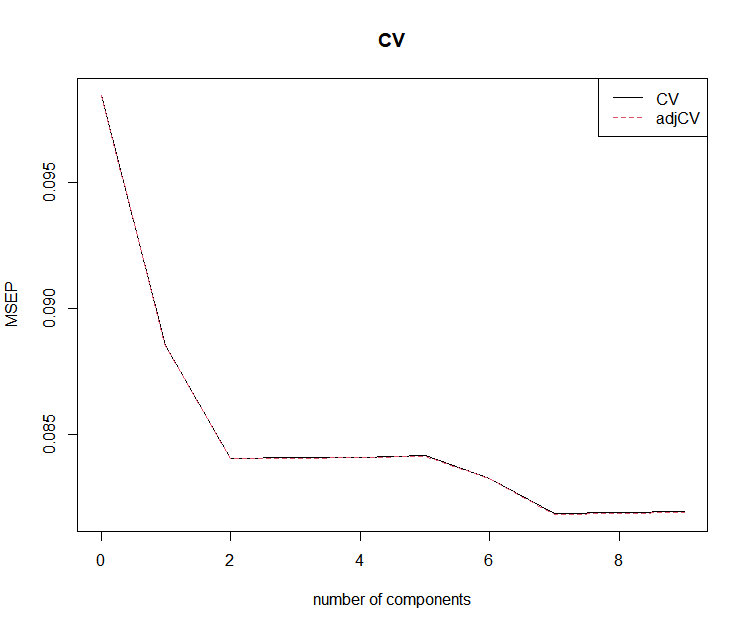


**Figure 2: Correlation plot for numerical predictor variables Figure 3: Density plot for numerical predictor variables**

From the correlation matrix, we can see variables like “euribor3m”, “nr.employed”, “pdays”, “emp.var.rate” and “cons.price.idx” are highly correlated to each other. From the plots, the distribution for most of the numerical variables are not normal as well. This primarily analysis revealed some information that will be crucial to the performance of some model later on.

**2.1 Principal Component Analysis**

Principal Component Analysis (PCA) is a type of unsupervised method that is used to summarize variables and group them together into uncorrelated components by taking linear combinations of predictors that accounts for highest variance. In this case, as the dataset has quite a number of correlated variables, we are attempting to use PCA to perform dimension reduction and to explore the dataset. This technique can be used to find common features of dataset with the biplot. The components can be further used for prediction and analysis.

* **2.1.1 Principal Component Regression** Principal Component Regression(PCR) is type of regression technique that uses the components from PCA to do regression analysis. In this case, not much dimension reduction was donefrom the 10 numerical variables and the first 7 components are chosen for the regression through cross validation. After training the PCR model on the train set and making predictions on the test set, we are able to achieve an AUC score of 0.77.  ***Figure 4: CV for number of principal components.***
* **2.1.2 Logistic PCA** Similar to the logic of PCR, the components from PCA will be used to do a regression analysis, however for this case we will be running a logistic regression analysis on the components instead, results are expected to be similar or better than PCR since the response variable is binary and logistic regression is more catered for binary variables. After training the LOGPCA model on the train set and making predictions on the test set, we are able to achieve an AUC score of 0.77 which is the same as PCA.

**2.2 Decision Tree**

The idea of decision tree (DT) is regression tree that splits covariate spaces into set of rectangles based on conditions and fit simple models in each of them. In this case, the tree was allowed to grow big and pruned back to minimize loss and to obtain best prediction performance. K-fold cross validation was used to find the best cp (penalty parameter that performs regularization). After training the DT model on the train set and making predictions on the test set, we are able to achieve an AUC score of 0.70.

**2.3 Naïve Bayes**

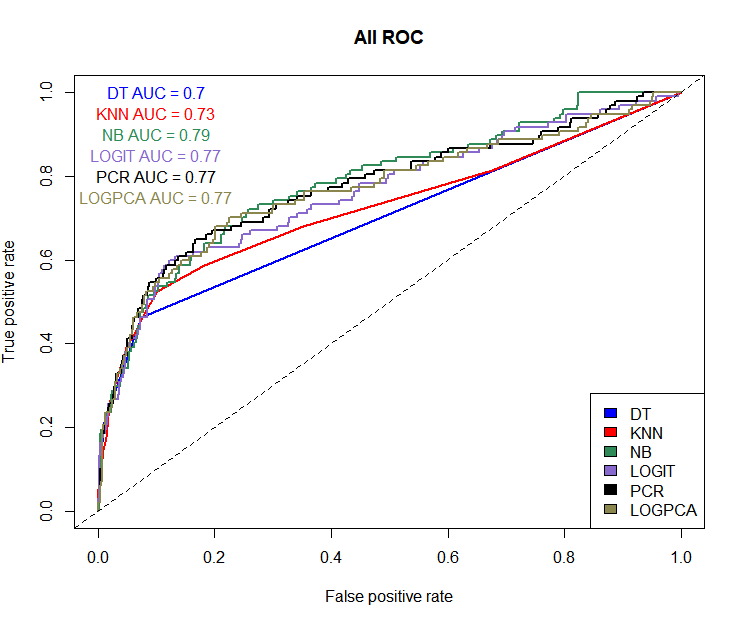
Naïve Bayes (NB) is based on Bayes Theorem, this model assumes that all predictors variables are independent with each other. However, this model is prone to introducing bias, but reducing variance at the same time. This model eliminates the problem of estimating joint distribution of predictors when the number of predictors get large, hence it often produces a strong model with little tendency to overfit. After training the NB model on the train set and making predictions on the test set, we are able to achieve an AUC score of 0.79.

**2.4 Logistic Regression**

Logistic regression (logit) is a simple model for a classification problem, it is used to model binary outcome variables, “y” in this case by modelling the log odds of the outcome as a linear combination of the other predictors. Firstly, by using all the predictors, training the logit model on the train set and testing it on the test set, we are able to achieve an AUC score of 0.76, by using the predictors (“nr.employed”, “pdays”) that the DT have picked out and another predictor “job” which I thought will be quite important to determining whether someone buys a product, we are able to get a slightly higher AUC of 0.77.

**2.5 K-nearest Neighbour**

K-nearest Neighbour (KNN), works by classifying inputs by the Euclidean distance to every other point in the dataset, the most represented class in the neighbour will decide the class of the input. However, Euclidean distance does not make sense when applied to categorial predictors and the distance between each category does not provide any proper information. Therefore, we will only be using nominal predictors in this case. After training the KNN model on the train set and making predictions on the test set, we are able to achieve an AUC score of 0.73.



***Figure 5: Plot of all ROC curves and AUC for each model***

**4 Evaluation and Conclusion**

Out of all the models, the Naïve Bayes model performed the best despite the assumption of independence not holding, as seen in the correlation matrix, there are a few variable predictors that have a high correlation with each other. In addition, most of the numerical variable predictors like age are not normally distributed as well, which is another assumption that is not holding. This shows how the naïve bayes model is able to perform so well despite its basic assumptions not holding, it ends up being a simple method with little tendency to overfit which is be useful in the real world, it proves to be a good benchmark to beat when producing models.

Decision tree model did the worst out of all the models, this might be attributed to the limitation of decision trees, that it is prone to overfitting. Due to the nature of decision tree, it is able to learn the training data to the very small details, this causes a problem as when these conditions are applied to other datasets, in this case the testing dataset, many will be wrongly classified as the model is so tailored to the training data. Therefore, although decision trees are a great method due to the easy interpretations and visualizations as well as the ability to handle both numerical and categorical datasets, it is too prone to overfitting and might not work well in this case. One thing the tree did pick out were some predictors that proved to be quite useful, namely the predictors, “nr.employed” and “pdays”, both which allowed the logistic regression model to yield decent results. Something worth looking into will be the random forest algorithm, it is able to take away the bad parts of decision tree, overfitting, while producing great result. Random forest algorithm is based on decision tree, but instead of only one tree, it utilises multiple randomly created trees and combines the output of them to produce a final output. This will reduce overfitting while also prevents an increase in error due to biasness, therefore it is a good improvement to decision trees that is worth looking into.

Finally, the logistic regression model, what will appear in most people’s mind when it comes to fundament binary classification. The logistic regression model did do decently well in this case, however what did not do as well as expected is the logistic PCA, this is attributed to the fact that we did not really manage to reduce the multicollinearity and dimensions of the dataset, which unfortunately resulted in the logistic PCA achieving the same result as the basic logistic regression model, which is by no means bad. This method should fair better in settings where many variable predictors are highly correlated to one another, say datasets like bond market datasets, where a change in bond price for say a short-term bond will greatly affect the price of other bonds with different maturity.

**Bibliography**

[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, In press, <http://dx.doi.org/10.1016/j.dss.2014.03.001>

Andrew J. Landgraf (2016) An Introduction to the logisticPCA R Package

<https://cran.r-project.org/web/packages/logisticPCA/vignettes/logisticPCA.html>

ISLR Ch. 10.2, 10, 8.1, 6.3, 4.4.4