**Introduction**

Banks have thousands of customers, but it would be unwise to call up every customer whenever a new product is launched due to manpower and time constraints. The aim of this report is to create the best machine learning (ML) model that can predict who, when and how do we sell our products that would maximize our success rate and minimize resources spent on “wasted” calls. The goal is to establish an automated predictive model that, given customer data and current economic factors, can identify the customers most likely to make a purchase under specific economic conditions. A secondary goal is to generate interpretable insights which can help management to make better business decisions.

This is a binary classification problem. After cleaning the data, the dataset is split into a test set and training set using a random 50/50 split. We first perform unsupervised ML techniques to check for multicollinearity, and thereafter reduce the dataset’s dimensionality. Then, we will fit various ML models with the training set and evaluate each model’s effectiveness on the test set using the Accuracy Under Curve (AUC) value. The loss function used is the mean squared error (MSE), where the error of correctly identifying the class is 0 and incorrectly identifying is 1.

The results below show that the Logistic Regression model formed using hand selected variables performed the best, as shown by the highest AUC value and lowest MSE value.

**Figure 1: Performance of ML Models**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ML Model | LRM (selected variables) | Naïve Bayes | PCR | KNN | LRM (all variables) | DT |
| AUC | **0.791** | 0.778 | 0.766 | 0.760 | 0.758 | 0.723 |
| MSE | **0.0762** | 0.114 | 0.0786 | 0.0773 | 0.0782 | 0.0774 |

**Data Cleaning and Exploratory Data Analysis (EDA)**

Firstly, we convert the dependent variable "Y" into a binary variable “y1”, where 1 represents "yes" and 0 represents "no". Next, we address the missing values in some categorical predictors marked as "unknown." We chose to retain these unknowns, as removing entries with “unknown” would diminish the sample size from 4119 to 3085, potentially compromising the effectiveness of model training. Thus, we will count "unknown" as a separate class for categorical predictors “job” and “marital”.

We convert categorical variables "housing" and "loan” into binary variables. We assume that the missing values are missing completely at random or missing at random (likelihood of a value being missing is unrelated to the true value itself) and resolve them by inputting the mean of all existing entries within each predictor.

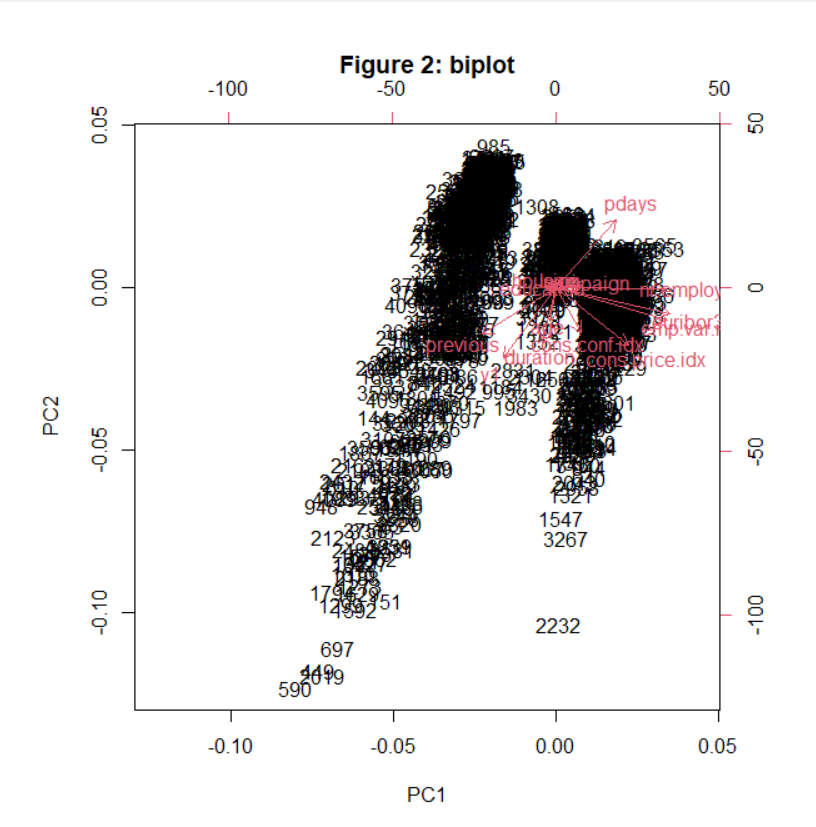
In the predictor "education," the class "illiterate" appears once, prompting the removal of the corresponding row. In the predictor "default," where the class "yes" appears only once, the entire predictor is removed due to the limited utility of the "default" predictor if trained without any instances of the "yes" class, potentially causing issues during testing.

For the "education" variable, we reorder the level of education from 4 years to university level. The classes are then converted and scaled to produce numeric values. The assumption underlying this transformation is that the "distance" between classes remains constant. For example, the difference between "basic.4yr" and "basic.6yr" is considered equal to the difference between “high.school” and "professional.course".

Lastly, the variable "duration" is not useful for analysis because it is an outcome of Y rather than a predictor. Thus, we will exclude "duration" to ensure a more realistic predictive model such that the model is based on information available at the time of decision-making, aligning with real-world scenarios in telemarketing campaigns.

**Principal Component Analysis (PCA) and Principal Component Regression (PCR)**

**A graph of a function

Description automatically generated**

Before applying PCA, we first need to extract out just the numeric variables. The model did not result in a meaningful reduction of the dataset's dimensionality. The scree plot presented in Figure 2 illustrates a gradual downward slope, suggesting that the proportion of variance captured by the initial principal components is inadequate, possibly due to a low degree of correlation between variables. The first principal component only explained 28.7% of the variance in our predictors. However, Figure 3's biplot does show us that PC1's highest loadings are on economic factors like "euribor3m", "emp.var.rate" and “nr.employed” , indicating the high level of correlation.

The PCR performs well and sets a benchmark for future models, with an AUC of **0.766**.

**Logistic Regression Model**

2 LR models were built, one with all the variables except “duration” and “euribor3m”, and another with hand selected variables based on economic knowledge and intuition. The later performed better, with an AUC of **0.791**, while the former have an AUC of 0.**758**. LR models prove to be the most effective model in predicting potential successes as the bank can utilize its domain knowledge to hand-select variables that will optimize the outcome even more.

**Decision Trees (DT)**

**Figure 3 and 4: Decision Tree generated by excluded different variables**

A diagram of a tree

Description automatically generatedA diagram of a diagram

Description automatically generated

The tree constructed using all the variables except “duration” (Figure 3) suggests that when employment is below 5088 thousand, the bank should focus on clients contacted within the past 8 days from a previous campaign. However, this finding poses limited practical utility as employment rates are beyond the bank's control. Moreover, advising the bank to selectively target clients based on previous campaign contact will hinder their ability to expand and engage with new clients.

Even though the model has a AUC of **0.72** and fails to outperform the other models, an advantage of DT is it can be used to extract interpretable insights for the management’s use by stepwise removal of variables. Upon removing “eurobor3m”, “nr.employed” and “pdays”, the tree(figure 4) implies that a successful outcome in the previous campaign correlates with a higher likelihood of success in the current campaign, compared to unsuccessful or nonexistent outcomes. Even though the AUC is now lower at 0.59, the result underscores the managerial imperative for banks to maintain consistent contact with clients. Subsequent pruning led to single-node trees, signifying the absence of discernible patterns for targeting clients based on specific attributes such as age group or occupation.

**K-Nearest Neighbours (KNN)**

To perform KNN, we remove the binary variables and “euribor3m” to reduce effect of multicollinearity, focusing on numeric and scaled ordinal variables. The AUC of the model is **0.76**, indicating relatively strong performance.

However, practical limitations arise as 4 out of the 9 variables used pertains to economic factors. Thus, the model heavily relies on the 2 customer data variables and 3 campaign variables for predictions due to the inherent nature of economic variables being fixed at any given time. This limitation is inherent to KNN due to its exclusion of binary variables, impacting the model's practical applicability.

**Naïve Bayes**   
The model was fitted with all variables except “duration” and “euribor3m”. AUC of model is **0.78**, indicating strong performance. Sensitivity was low at 43%, indicating a challenge in correctly identifying positive instances. We can consider reducing the probability threshold to increase sensitivity.

**Conclusion**

In conclusion, the LR model with hand selected variables proves to be the most effective. However, the results show that this is not an easy classification problem to solve and more optimization is needed to improve the accuracy of the model. One must also consider the computation efficiency of the model as KNN would be computationally more expensive as the dataset grows larger.

A common shortcoming of all the models is the low sensitivity (True Positive Rate) based on the probability threshold p used of 0.5. We can adjust p to 0.1 or 0.05 based on the bank’s available resources. In the Naïve Bayes example, if the bank has fewer resources, it can set p=0.5 and make fewer total calls (31). Although the bank is guaranteed a higher success rate (62%), the large number of missed sales (199 false negatives) will incur a huge cost. If the bank has plentiful of resources, it can set p=0.05 and make 1218 calls. Even though only 14% of the calls succeed and there are many “wasted” calls, the bank achieves a greater number of successful calls (181) and minimises the number of missed sales (37).

|  |  |  |  |
| --- | --- | --- | --- |
| Probability Threshold | 0.5 | 0.1 | 0.05 |
| Confusion Matrix of Naïve Bayes Model | A number on a white background  Description automatically generated | A number and numbers on a white background  Description automatically generated | A number on a white background  Description automatically generated |
| No. of estimated success and hence calls to make | 31 | 560 | 1218 |
| Success Rate | 62% | 24% | 14% |

Another limitation is that the loss function MSE used assumes that the loss due to a false positive is and false negative are equal when in reality, a missed sale is more costly than a “wasted” call. The loss function can be altered to penalize a false negative more.