DSE1101 Final Project

Koh Cai Ling, A0301000R, 17/11/2024

**Introduction**

This is an executive summary of a telemarketing campaign by a Portuguese bank offering term deposit subscriptions. Using unsupervised and supervised learning methods, the goal is to build a model which best predicts whether a customer will subscribe to a term deposit.

**Dataset**

This dataset is created by Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL) in 2014. Based on "Bank Marketing" UCI dataset, the data is enriched by the addition of five new social and economic features published by the Banco de Portugal.

While the dataset is fairly robust, it does have several issues.

1. The input variable ‘duration’(last contact duration, in seconds). This attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, I discarded this input for a more realistic model.

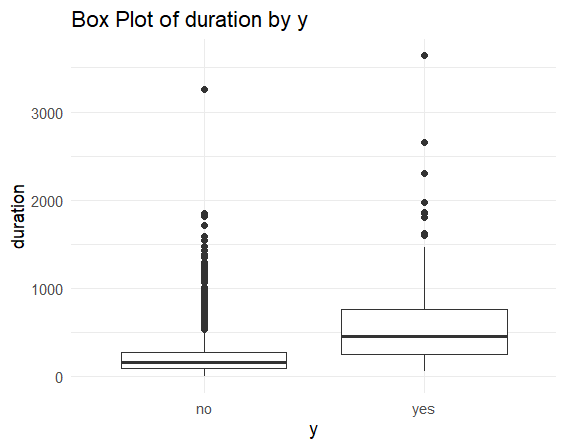


Figure 1: Box Plot of duration by y.

Figure 1 shows that as duration increases, the likelihood of y="yes" also rises. This could be due to the fact that, with more time, the telesales agent has a greater opportunity to persuade the customer to subscribe to the term deposit.

1. Under input ‘education’, there is only one level ‘illiterate’. Similarly under input ‘default’, only one level ‘yes’. Both led to y="no"(i.e., customer did not subscribe). To ensure consistency in levels across both train and test sets, these observations are omitted.
2. As 25% of observations contain missing values, these missing values are treated as a possible class label to retain most information.

A number of numbers on a white background

Description automatically generatedFigure 2: Correlation matrix for social and economic context attributes

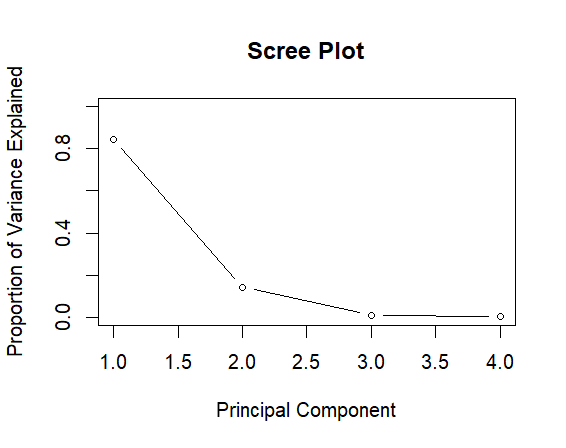
1. By summarising the correlation of the five social and economic features, it can be seen that emp.var.rate, cons.price.idx, euribor3m and nr.employed are moderately/highly correlated. Thus, to reduce dimension for improving the efficiency, accuracy, and interpretability of data for the upcoming data analysis, Principal Component Analysis(PCA) is used. From Figure 3, since PC1 and PC2 explains 98.6% of the variance in the data, only these two are updated in the new data frame.

Figure 3: Scree Plot

After the aforementioned transformation, the data frame now has 4117 rows and 18 columns. The data frame is then split to 30% for testing and the remaining 70% to train.

**Models**

**Logistic Regression:** Backward selection was implemented to remove variables that were not statistically significant (job, marital, education, default, housing, loan, day\_of\_week, pdays).

**K nearest neighbours & Leave-one-out Cross Validation:** All nominal categorical input variables (job, marital, default, housing, loan, contact, month, day\_of\_week, poutcome) were removed.

**Naïve Bayes Classifier:** Ignores associations between predictors.

**Decision Trees:** The tree is pruned to 5 leaves with Gini as criterion. Figure 4 shows that the most important features are PC1 and pdays .

A diagram of a number

Description automatically generated

Figure 4: Cross validation minimising Gini tree (with PCA)

**Performance**

The overall performance of the models is evaluated using the area under the Receiver-operating characteristic curve(ROC); AUC.

It is noted that the AUC of each predictive model after performing PCA is the same/lower compared to when PCA is not performed. This may be because the relationships between the economic and social features are not linear but complex, and these were not well-captured by the PCA which decomposes the variables to its linear combinations.

Meanwhile, though the AUC for logit regression with and without backward selection is the same, the sensitivity for backward selection is a little higher (98.1% to 98.6%).

|  |  |  |
| --- | --- | --- |
| Logistic Regression |  | AUC = 0.76  Second best |
| K nearest neighbours & Leave-one-out Cross Validation |  | AUC = 0.74  Third best |
| Naïve Bayes Classifier |  | AUC = 0.77  Best |
| Decision Trees |  | AUC = 0.68  Worst |

The best model for prediction is Naïve Bayes Classifier. This may be because this data has many categorical features, high-dimensional and features have low correlations between each other (except for the socioeconomic features that have been reduced dimensionally via PCA).

**Limitations**

However, the AUC for all models are still relatively low, and all four predictive models are not very accurate and reliable in predicting whether a customer will subscribe to a term deposit.

This may be due to the high number of categorical features that causes the use of PCA and clustering to be not appropriate. A more appropriate specialised model would be CATPCA.

**Conclusion**

Based on the above analysis, I think my model will not be useful in a real-world scenario. Advanced machine learning methods and basic ones work similarly based on AUC value, and it is not worth to go beyond such methods.