CLASSIFICATION MODELS ON THE "BANK MARKETING DATA SET"

Ritirupa Dey (MDS202136) , Ankush Dey (MDS202108)

- INTRODUCTION:

In this project, we have been provided a secondary dataset given - "BANK MARKETING DATA SET" obtained from the UCI Machine Learning Repository. It is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

We have built three classifiers for this data set: a decision tree, a naïve Bayes classifier, *and* a random forest and done a comparative study of their performances using F-score as the primary evaluation metric since the dataset is highly skewed.

- EXPLORATORY DATA ANALYSIS:

  - We have used barplots to visualize the categorical variables (including the binary outcome variable 'y') of the dataset and histograms to visualize the numerical variables.
  - From the barplot for 'y' we find that the data is highly imbalanced since the number of 'yes's is much lesser than the number of 'no's. Hence, F1-score,which balances the precision-recall trade-off, will be a much better evaluation metric than Accuracy.
  - We have also found that the variables 'default' and 'pdays' as the former contained high number of unknown samples and the latter contained high number of null samples.

- DATA PRE-PROCESSING:

  For Decision Tree and Random Forest classifiers,

  - Since we have visualized that the variables ‘default’ and ‘pdays’ did not play such an important role in predicting the outcome variable ‘y’, we have dropped them from our training and test sets.
  - We have removed the ‘unknown’s in the categorical variables with the maximum value of that feature variable.
  - We have labelled the ordinal features using a labelling dictionary.
  - We have used One Hot Encoding to label the nominal features.
  - The numerical variables have also been normalized using MinMaxScaler.

  For,Naïve Bayes,

  - We have performed outlier detection using EDA and replaced the outliers of the corresponding columns(‘campaign’, ’duration’ and ‘previous’) with the means of the respective features.
  - All the other pre-processing steps are same, except instead of labelling the ordinal and nominal features separately, we have used get_dummies() of pandas to label the categorical variables.

- COMPARITIVE STUDY OF THE THREE CLASSIFIERS:

	Decision Tree	Naïve Bayes	Random Forest
Accuracy	89.71%	86.38%	90.51%
Precision	50.09%	41.90%	56.13%
Recall	74.14%	54.72%	71.98%
F1-score	90.42%	87.15%	91.01%
Time taken	6.71 sec	2.23 sec	5.10 sec
Memory space required	4.06 MiB	15.92 MiB	49.50 MiB

- CONCLUSION:

  - From the above table, we can conclude that Random Forest has the highest F1-score among the three classifiers and hence, is the best classifier among these three. This is indeed in line with the theoretical performance of the classifiers since random forest is built using sampling over various features of the training set.
  - Also, inspite of not having performed outlier detection and imputation in Decision Tree and Random Forest Classifiers, those models performed substantially well.