DMML:ASSIGNMENT-1

Bharath Kumar Ravilla(MDS202133)
Reewa Malik (MDS202134)

March 25, 2022

1 Task

The task is to build three classification Models i.e Random Forest, Decision Tree, Naive Bayes for the Bank Marketing data set,which can predict whether a given bank customer will subscribe to the term deposit or not.

2 Solving the Task

In order to build the classification models we followed the following approach:
1. We analysed the data by checking how many categorical and numerical variables are there.
2. We used visualisation techniques to plot both Numerical and categorical variables in order to get something significant about the variables
3. The main observation picked from point 2 is that data is highly imbalanced(no = 36548, yes = 4640).
4. We converted all the categorical variables to numerical using **OneHotEncoding**.
5. We found corrrelation between the numerical variables and found out that *euribor3m, nr.employed, cons.price.idx, emp.var.rate* are positively correlated, so we used factor analysis and created another column X.factor and dropped these.
6. We have dropped duration attribute to get realistic predictive model.
7. We then used **Minmaxscale** to scale numerical variables *age, cons.conf.idx, X.factor* and also converted *pdays* variable range to 0 and 1.
8. After that we splitted the data using the technique **SMOTE** because the data is unbalanced.
9. Then we fitted the models and observed the following:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.928591	0.933934	0.922435	0.928149
Decision Tree	0.900616	0.890415	0.9133680	0.901897
Naive Bayes	0.691108	0.648902	0.832832	0.729451

10. Based on the metrics given in point 8, we found out that Random Forest is working better than Decision Tree and Naive Bayes in terms of accuracy.