CMI/2022/Assignment1 Deadline: March 25,2022

Data Mining & Machine Learning

***Group Members*:**

1. Avik Das - MDS202112
2. Meghna Mondal – MDS202123

Classification Models on Bank Marketing dataset:

In this assignment we have built three classifiers for this data set:

1) a decision tree,

2) a naïve Bayes classifier, and

3) a random forest.

Then we used suitable evaluation metric to compare the performance of the three classifiers.

Table for Comparing the classifier models:

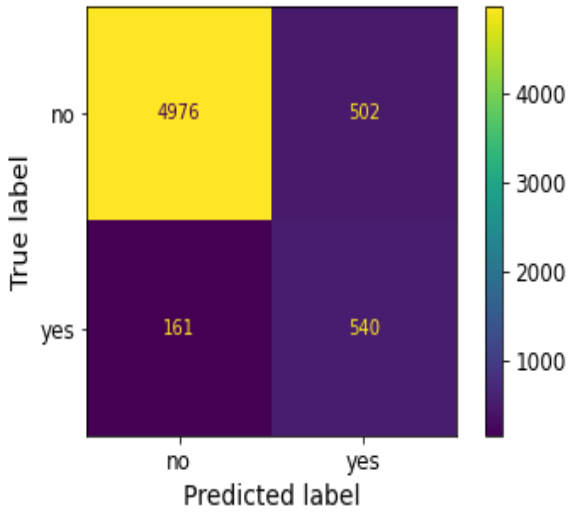
Performance Measure	Decision Tree	Naïve Bayes Classifier	Random Forest
Accuracy	**0.89**	**0.87**	**0.90**
Precision	**0.52**	**0.42**	**0.54**
Recall	**0.77**	**0.52**	**0.72**
Time Taken(s)	**14.01**	**14.44**	**27.51**
Space required(MiB)	**7.05**	**37.03**	**6.85**

Observation during fitting the Models:

- At first, we observe that in the Bank Marketing dataset there are 20 input variables out of which 10 are categorical and 10 are numerical variables and there is no missing value.
- The output variable is highly imbalanced with 88.7% 'no' and the rest is 'yes'.
- We split the dataset into train set, validation set and test set in case of Decision Tree and Random Forest. We considered validation set here for deciding the *max_depth* of the tree.
- Since there was very few "yes" and a large number of "unknown" in the 'default' column. So, we dropped this column.
- Due to having imbalanced data in the output variable we used class weight as a parameter for the Decision Tree and Random Forest Classifier placing greater weight on the 'yes' target variable.
    - We have tried using different values of parameter of Random Forest Classifier and taken *n_estimators* = 100, *class_weight*={'no':11.3,'yes':88.7} as they are producing better result.
    - And the *class_weight* = {'no':1,'yes':3} in case of Decision Tree Classifier.
- We have computed the Mean Absolute Error for different *max_depth* of the tree (used is Decision Tree and Random Forest Classifier) using validation set and chose the optimal *max_depth* for which MAE is minimum.

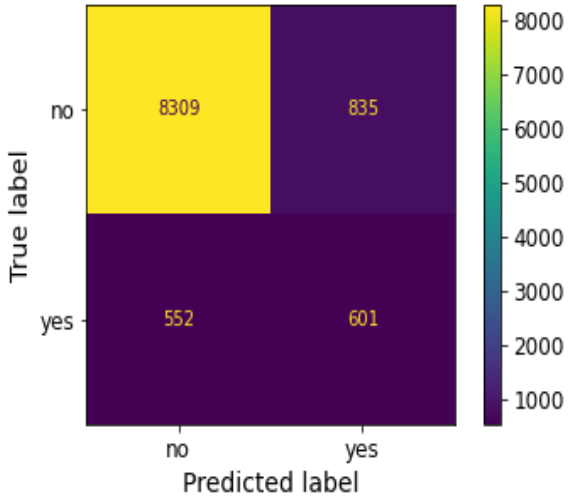
  Following this method, we get
    - 9 as *max_depth* for Decision Tree
    - 16 as *max_depth* for Random Forest
- Since Naïve Bayes Classifier is sensitive to outliers, we replaced the outliers belong to the numerical variables "Campaign" and "Duration" with their respective mean.
- As we consider minimum MAE for building the Decision Tree and Random Forest, so, we gain the optimal accuracy for the corresponding *max_depth* of the tree.
- For the Decision Tree and Random Forest, we also had to perform the precision-recall trade-off. We tried to maximise recall without drastic decrease in the precision or the accuracy.
- Theoretically we should gain more accuracy while using Random Forest than the Decision Tree, and we got the same practically also. And as usual, we have achieved lower accuracy in Naïve Bayes Classifier.

Comparison among Confusion Matrix:



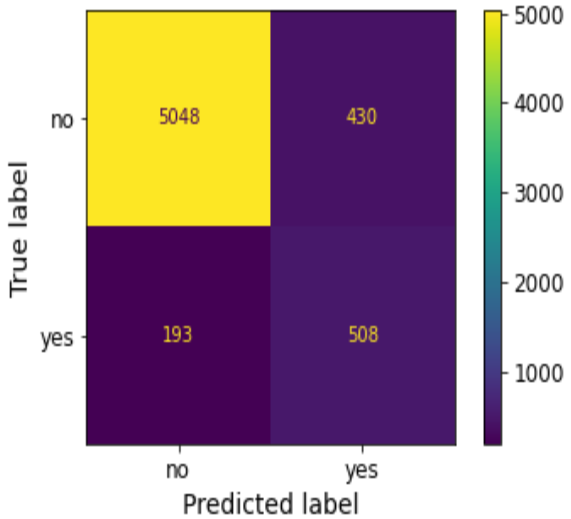
Model-I

Decision Tree



Model-II

Naïve Bayes Classifier



Model-III

Random Forest