Data Mining & Machine Learning

Data:

The "Bank Marketing Data Set" from the UCI Machine Learning Repository is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

Objective:

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

Classification Models on Bank Marketing dataset:

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

1) a decision tree,

2) a naïve Bayes classifier

3) a random forest.

4) Boosting with decision stumps

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	Boosting
Accuracy	0.83	0.85	0.84	0.83
Precision	0.35	0.37	0.38	0.36
Recall	0.58	0.47	0.60	0.62
Time Taken(s)	0.1	0.06	0.8	10.5
Space required(Mb)	323	324	418	392

Observation during fitting the Models:

Data cleaning:

- After importing the data, we plotted histograms of the numerical and categorical variables separately to understand the relationship between the input variables and output variable. We found that the attribute 'duration' 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. So,we drop this attribute'.

- We then used one-hot encoding to substitute dummy values for the categorical variables such as 'marital', 'loan', 'poutcome' , 'contact', 'job', 'education' ,'day of the week', 'month'.

- We used train test split to split the data into training and testing data in the ratio 80:20

Applying different classifiers:

- Our goal is to predict whether given the various information about marketing campaigns of a bank, a customer will subscribe for a term deposit at the bank or not. Our target hence is to find all the cases when a customer actually opened a term deposit at the bank, which is the true positive case. Hence, we decided to maximise the metric **Recall**. Keeping this in mind we proceeded as follows with each of the classifiers:

  - **Random forest classifier**: We used RandomizedSearchCV to find the best value of the parameters n_estimator(no. of decision trees in the random forest) and min_samples_leaf (The minimum number of samples required to be at a leaf node).Using those values we fit the model to the training data and evaluated its performance based on the test data.

  - **Boosting with decision stumps**: We used GridSearchCV to find the best value of the parameter n_estimator(The maximum number of estimators at which boosting is terminated). Using those values, we fit the model to the training data and evaluated its performance based on the test data.

  - **Decision Tree :** We used RandomizedSearchCV to find the best value of the parameters max_depth and max_leaf_nodes(The maximum number of leaf nodes in decision tree).Using those values we fit the model to the training data and evaluated its performance based on the test data.

- **Naïve Bayes:** We used Gausian NB as our classifier to fit the training the data

- Based on recall, we can say that Adaboost classifier with decision stumps is the best classifier among all four in classifiying whether a person will make a term deposit or not. We can see that the memory usage of adaboost and random forest is comparatively higher than decision tree and naïve bayes. On the other hand, Adaboost has a much higher running time than the other three and random forest has the highest memory used.