Data Mining and Machine Learning
Assignment 1

Group Members: Aniket Santra - MDS202106
Soham Biswas - MDS202147

Dataset: The dataset used in this assignment is the "Bank Marketing Data Set" from the UCI Machine Learning Repository. It is based on a direct marketing campaign (via phone calls) of a Portuguese banking institution. The target of this marketing campaign is to get their clients to subscribe for a term deposit.

Observations from the Exploratory Data Analysis:

- The data consists of both categorical and numerical factors.
    - The categorical factors include:
        - job: type of job (admin., bluecollar, entrepreneur, housemaid, management, retired, self employed, services, student, technician, unemployed, unknown)
        - marital : marital status (divorced, married, single, unknown)
        - education: (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)
        - default: has credit in default? (no, yes, unknown)
        - housing: has housing loan? (no, yes, unknown)
        - loan: has personal loan? (no, yes, unknown)
        - contact: contact communication type (cellular, telephone)
        - month: last contact month of year (jan, feb, mar, …, nov, dec)
        - day of week: last contact day of the week (mon, tue, wed, thu, fri)
        - poutcome: outcome of the previous marketing campaign(failure, nonexistent, success)
    - The numerical features include:
        - age
        - day: last contact day of the month (1 to 31)
        - duration: last contact duration, in seconds (numeric)
        - campaign: number of contacts performed during this campaign and for this client (includes last contact)
        - pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
        - previous: number of contacts performed before this campaign and for this client
        - emp.var.rate: employment variation rate - quarterly indicator
        - cons.price.idx: consumer price index - monthly indicator
        - cons.conf.idx: consumer confidence index - monthly indicator
        - euribor3m: euribor 3 month rate - daily indicator
        - nr.employed: number of employees - quarterly indicator
    - The target variable is ‘subscribed’: has the client subscribed a term deposit? (it is of binary nature: yes, no)

- The categorical variables ‘job’, ‘education’, ‘housing’, ‘loan’, ‘default’ and ‘marital’ had records with unknown values. Therefore, all such records have been dropped.
- Since nearly all categorical variables are nominal in nature, we have performed one hot encoding on these variables to convert them to numerical values.
- The binary target variable ‘subscribed’, renamed as ‘y’ has been encoded as follows: (1 if yes, 0 if no)
- We observe that the data is highly imbalance as the target variable has approximately 8 times the value ‘no’ with respect to the value ‘yes’.

Procedure:

- We perform feature selection to select the best features to predict our target variable. The process is as follows:
    - We’ve estimated mutual information for the discrete target variable in feature selection process.
    - Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. It is used for univariate feature selection.
    - The mutual information between two random variables X and Y can be stated formally as follows:
        - I(X ; Y) = H(X) – H(X | Y) Where I(X ; Y) is the mutual information for X and Y, H(X) is the entropy for X and H(X | Y) is the conditional entropy for X given Y.
    - After encoding the categorical features we’ve performed SelectKBest feature selection using mutual information classifier and drop those features which have feature_scores less than equal to 0.001.
- We have split the data into training set and testing set in the ratio of 80:20.
- We have chosen the accuracy, precision, recall and F1 score as performance measure and calculated them for each of the 3 models.
- To counter the high imbalance in the data, we have used class weights in the decision tree classifier and random forest classifier.
- We have also tried to choose an appropriate max_depth of these 2 models with regards to the performance measures.
- In the above cases we have tried to select these attribute values as such that the recall is maximized without too much of a fall in the accuracy and precision.
- We have also calculated the time taken from fitting to prediction for each model.

Observation:

Performance Measures	Decision Tree	Random Forest	Naive Bayes
Accuracy	87.81%	87.34%	84.80%
Precision	52.05%	50.82%	42.75%
Recall	81.36%	85.39%	49.37%
F1 Score	63.49%	63.72%	45.82%
Time	0.26s	3.09s	0.14s

- Due to the highly imbalanced nature of the data, accuracy and precision is more or less close for both decision tree classifier model and the random forest classifier model. It is slightly lesser for the naive bayes classifier model.
- The recall and F1 score if used to compare the performance of the classifier we find that the random forest model seems to be slightly better than the decision tree model and the naive bayes model is worse than them.
- The random forest model as expected needs the most amount of time (by quite an amount) to complete the process from fitting to prediction, while the naive bayes model requires the least amount of time.