MACHINE LEARNING (EXTENDED PROJECT)

Dear Participants,

Please find below the Machine Learning Extended Project instructions:

- You have to submit 2 files:

  Business Report: In this, you need to submit all the answers to all the questions in a sequential manner. Your answer should include detailed explanations & inferences to all the questions. Your report should not be filled with codes. You will be evaluated based on the business report.

  Jupyter Notebook file: This is a must and will be used for reference while evaluating.

- Any assignment found copied/ plagiarized with another person will not be graded and marked as zero.
- Please ensure timely submission as a post-deadline assignment will not be accepted.

Part 1: Machine Learning Models

You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters like age, salary, work experience etc. given in the data set 'Transport.csv', you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalised.

Data Dictionary

Age : Age of the Employee in Years

Gender : Gender of the Employee

Engineer : For Engineer =1 , Non Engineer =0

MBA : For MBA =1 , Non MBA =0

Work Exp : Experience in years

Salary : Salary in Lakhs per Annum

Distance : Distance in Kms from Home to Office

license : If Employee has Driving Licence -1, If not, then 0

Transport : Mode of Transport

The objective is to build various Machine Learning models on this data set and based on the accuracy metrics decide which model is to be finalised for finally predicting the mode of transport chosen by the employee.

Questions:

1. Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.
2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not?

3. Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.:
    a. Logistic Regression Model
    b. Linear Discriminant Analysis
    c. Decision Tree Classifier – CART model
    d. Naïve Bayes Model
    e. KNN Model
    f. Random Forest Model
    g. Boosting Classifier Model using Gradient boost.
4. Which model performs the best?
5. What are your business insights?

Part 2: Text Mining

A dataset of Shark Tank episodes is made available. It contains 495 entrepreneurs making their pitch to the VC sharks.

You will ONLY use "Description" column for the initial text mining exercise.

1. Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.
2. Create two corpora, one with those who secured a Deal, the other with those who did not secure a deal.
3. The following exercise is to be done for both the corpora:
    a) Find the number of characters for both the corpuses.
    b) Remove Stop Words from the corpora. (Words like 'also', 'made', 'makes', 'like', 'this', 'even' and 'company' are to be removed)
    c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)?
    d) Plot the Word Cloud for both the corpora.
4. Refer to both the word clouds. What do you infer?
5. Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

RUBRIC

Criteria	Pts
1.1 Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.	5.0 pts
1.2 Split the data into train and test in the ratio 70:30. Is scaling necessary or not?	4.0 pts
1.3 Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.: a. Logistic Regression Model b. Linear Discriminant Analysis c. Decision Tree Classifier – CART model d. Naïve Bayes Model e. KNN Model	25.0 pts

Criteria	Pts
f. Random Forest Model g. Boosting Classifier Model using Gradient boost.	
1.4 Which model performs the best?	3.0 pts
1.5 What are your business insights?	3.0 pts
2.1 Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.	2.0 pts
2.2 Create two corpora, one for those who secured a Deal, the other for those who did not secure a deal.	2.0 pts
2.3 The following exercise is to be done for both the corpora: a) Find the number of characters for both the corpuses. b) Remove Stop Words from the corpora. (Words like 'also', 'made', 'makes', 'like', 'this', 'even' and 'company' are to be removed) c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)? d) Plot the Word Cloud for both the corpora.	10 pts
2.4 Refer to both the word clouds. What do you infer?	3 pts
2.5 Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?	3 pts
Total Points	60 pts