<—Go Back to Machine Learning

~ Course Content

Project - Machine Learning

Submission type	:	File Upload
Due Date	:	Sep 05, 11:59 PM
Total Score	:	60
Available from	:	Aug 20, 8:00 AM

Description

Dear Participants,

Please find below the Machine Learning Project instructions:

- You have to submit 2 files :

1. **Answer Report:** In this, you need to submit all the answers to all the questions in a sequential manner. **It should include a detailed explanation of the approach used, insights, inferences, all outputs of codes like graphs, tables, etc.** Your report should **not** be filled with codes. You will be evaluated based on the business report.
2. **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.

Problem 1:

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 11 marks
1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Data Preparation: 4 marks
1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)

Modeling: 22 marks
1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)
1.5 Apply KNN Model and Naive Bayes Model. Interpret the results. (4 marks)
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

Inference: 5 marks

1.8 Based on these predictions, what are the insights? (5 marks)

Problem 2:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

(Hint: use .words(), .raw(), .sent() for extracting counts)

2.1 Find the number of characters, words, and sentences for the mentioned documents. - 3 Marks

2.2 Remove all the stopwords from all three speeches. - 3 Marks

2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words, (after removing the stopwords) - 3 Marks

2.4 Plot the word cloud of each of the speeches of the variable, (after removing the stopwords) - 3 Marks
[refer to the End-to-End Case Study done in the Mentored Learning Session]

Code Snippet to extract the three speeches:

II

```
import nltk
nltk.downloadfinaugural')
from nltk.corpus import inaugural
inaugural.fileids()
inaugural.rawC1941-Roosevelt.txt')
inaugural.rawC1961-Kennedy.txt')
inaugural.raw('1973-Nixon.txt')
```

II

Important Note: Please reflect on all that you have learned while working on this project. This step is critical in cementing all your concepts and closing the loop. Please write down your thoughts here.

****

The Machine Learning project has been designed to give you a glimpse of real-world analytics projects. The scope of this project is quite exhaustive which focuses much on model performances with possible parameter alterations and you will be expected to bring all the knowledge you have learned so far in this program. You will get a chance to try out multiple algorithms and compare them on various performance metrics. **The intent is to help you understand and appreciate WHY under a given circumstance a particular algorithm works better than others.** This is an important skill set and will serve you well when you work on real-world data science projects.

****

Scoring guide (Rubric) - Machine Learning Project

Criteria	Points

Criteria	Points
1.1) Read the dataset Describe the data briefly. Interpret the inferences for each. Initial steps like head().info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.	4
1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.	7
1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.	4
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.5) Apply KNN Model and NaTve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.	7

Criteria	Points
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC.AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)	7
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.	5
2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words)), .raw(), .sent() for extracting counts)	3
2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.	3
2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words, (after removing the stopwords)	3
2.4) Plot the word cloud of each of the three speeches, (after removing the stopwords)	3
Quality of Business Report	6
Please reflect on all that you learnt and fill this reflection: https://docs.google.eom/forms/d/e/lFAIpQLSfqHHImJyUkniiBiejtudluRFk_TVCLe843wfX6lu3QNRPmng/viewform?usp=sf_link	0
Points	60

Submit Assignment

@ **Upload file**

Add comments

Submit Assignment

< Previous