Honours Assignment for DSA822S Data Science & Analytics: 2022-23
40% Contribution: Deadline Saturday, September 2nd, 2023

BASIC REQUIREMENTS

- ☐ This assignment is not an essay assignment but your ipython notebook should be well commented with both text, equations and explanations.
- ☐ Your submitted notebook should be able to run to completion successfully.

PART 1 - VISUALISATION

Question 1
A survey was conducted to gauge audience interest in different data science topics, namely:

Big Data (Spark / Hadoop)
Data Analysis / Statistics
Data Journalism
Data Visualization
Deep Learning
Machine Learning

The participants had three options for each topic: Very Interested, Somewhat interested, and Not interested. 2,233 respondents completed the survey.

If you examine the csv file, you will find that the first column represents the data science topics, and the first row represents the choices for each topic.

Question 2
Use the artist layer of Matplotlib to plot the bar chart to visualise the percentage of the respondent's interest in the different data science topics surveyed.
To create this bar chart, you can follow the following steps:

1. Sort the dataframe in descending order of Very interested.
2. Convert the numbers into percentages of the total number of respondents.

Recall that 2,233 respondents completed the survey. Round percentages to 2 decimal places.

As for the chart:

1. Use a figure size of (20, 8),
2. Bar width of 0.8,

3. Use colour **#5cb85c** for the Very interested bars, colour **#5bc0de** for the Somewhat interested bars, and colour **#d9534f** for the Not interested bars.
4. Use font size 14 for the bar labels, percentages, and legend.
5. Use font size 16 for the title, and
6. Display the percentages above the bars and remove the left, top, and right borders.

MARK SCHEME: Part 1 - Visualisation

Reading and opening the file	**Sorting of dataframe and converting to percentages**	**Plotting of the graph with the required format of the graph**	**TOTAL**
(ex 5)	(ex 10)	(ex 15)	(ex 30)

Part 2 - Supervised Learning Using Tree-Based Model

Using the “**ClaimsData.csv**” vehicle insurance data file provided, you can explore the data and fit a tree-based model to the vehicle insurance data. The risk departments in insurance companies decide the client’s premium by looking at their risk profile. Clients who are likely to claim large amounts are given higher quotes for premiums than low-risk clients.

You may do an exploratory data analysis, Several graphs and calculations can be extracted from the data provided: two scatterplots (*continuous variables*), two bar graphs (*mean claim amount for categorical variables*), and two calculations without graphs (*mean claim amount of vehicles in each category* and *mean claim amount of vehicles for each model year)*. What insight can you draw from these? In your answer, give a short interpretation of two graphs or calculations. For example, are there cars with a particular mileage that have larger claims? Are the mean claim amounts in any of the categories different? Ensure that you provide possible reasons for any differences you observe.

Insight Questions from a Fitted Tree-Based Model

1. You were asked to fit a tree-based model to this data set. Why do you think a tree-based model is more appropriate than a neural network in this scenario?

2. Use the decision tree image you generated to gain insight into the factors that affect the claim amount. Start by looking at the first split of the decision tree, and recommending a course of action with regard to the age of the driver. Choose two more nodes to interpret, and comment on the impact the output could have on premiums.

NOTE: Remember to set the `random_state=0` for reproducibility.

MARK SCHEME: Supervised Learning Using Tree-Based Model

Overview of the data	**Plotting of the scatter plots**	**Plotting of bar graphs**	**Calculations without graphs**	**Insights drawn from graphs**	**Implementation of a tree-based model**	**Evaluation of ML performance**	**Style of Presentation & Insights from Fitted Tree-based Model**	**TOTAL**
(ex 5)	(ex 10)	(ex 10)	(ex 5)	(ex 15)	(ex 10)	(ex 5)	(ex 10)	(ex 50)

The annotated notebook is worth 40% of the total of 100% credits for your CAS, and therefore evidence for a significant amount of work, over and above the repackaging of lecture material, is required. It should demonstrate your own calculations/estimates of relevant quantities. The level should be suitable for reading by a fellow honours student. You should properly reference your academic sources, which may include websites.

A PDF OF THE NOTEBOOK* INCLUDING A LINK TO THE EXECUTABLE VERSION ON BINDER** MUST BE SUBMITTED **VIA EMAIL** BY 23:59 02 SARTUDAY 2023

* https://nbconvert.readthedocs.io/en/latest/usage.html

** https://mybinder.org