EGT 305

Assignment

Big Data Analytics and Comparison

Use Case 24S2

PLEASE TAKE NOTE OF THE FOLLOWING DETAILS:

1. Failure to meet the dateline would be subjected to *Guidelines for Late and Extension for Submission for Learner* **as stated in Module Overview on BrightSpace.**

2. Copying or asking someone else to do your work is an offence as stated in *NYP Academic Integrity Policy and Artificial Intelligence Policy* **as stated in Module Overview on BrightSpace.**

Business Context:

A Malaysian government organization would like to know the overall situation of the job market currently in the country. They have engaged a consultancy firm to conduct a survey and to report the findings to the government organization. The survey has been completed, and data collected.

Objective: To explore the data and provide findings to the government organization and build a ML model to predict the salary.

Dataset:

- Employee_dataset- Data of the people who participated in the survey.
- Employee_salaries- Salaries of the people who participated in the survey.

Dataset Description:

- Employee_dataset:
    - jobId: Unique job id for each person surveyed.
    - companyId: Company that there are employed at.
    - jobRole: Role of the person surveyed.
    - education: Highest education level
    - major: The specialization of the person surveyed.

  - Industry: The industry there are currently employed
  - yearsExperience: Number of years working in that industry.
  - distanceFromCBD: Distance of their house from the Central Business District in kilometres.
- Employee_salaries
  - jobId: Unique job id for each person surveyed.
  - salaryInThousands: Yearly income of the surveyed person thousands in Malaysian Rinngt.

Assignment Tasks:

You are to use PySpark for the entire assignment from part 1 to 3.

1. **Perform data cleaning (20 Marks)**
   - Remove duplicate or irrelevant observations.
   - Fix structural errors.
   - Fix unwanted outliners.
   - Handle missing data.
   - Validate and QA

   Please explain the reasons behind any decision that was made within a word limit of 100 words or less.
   (It is referring to a single decision, for example dropping of "jobId", the reasoning should be 100 words or less, it is not referring: 100 words or less for "Remove duplicate or irrelevant observations")

   Note: This portion is only for the initial data cleaning, if further data cleaning is required in the later parts of the assignment please do it accordingly.

2. **Exploratory data analysis (25 Marks):**
   In this section, please explore the dataset and answer the following questions with **reasons and justifications within a world limit of 100 words or less for each question**.
   - What is the JOBID that has the highest paying salary in the web industry?
   - Rank the top 5 jobs roles with the highest salary for all the industry?
   - Which industry has the highest average salary?
   - Which job role has the lowest pay?
   - Is there anything interesting in the data for the job role Janitor?
   - Given that the median salary per year is $98,000

        - Which industry has the highest percentage of people who are below the median salary?
        - What are the job roles that are above the median salary?
    - Determine if there is a relationship between job role and salary.
    - Is there a relationship between distance from CBD and job role?
    - Does the major they studied affect the salary?

3. **ML Model Development (20 Marks)**
    - Predict the salary of a person using an appropriate ML model and the features provided.
    - Explain the reasons why this model was chosen.
    - Keep it within 150 words

    Note: Do not compare more than 3 models for selection.

4. **Comparison between PySpark and non-PySpark workflow (15 Marks)**
    For this section you are to do part 1 to 3 with a non-PySpark workflow.
    (You can use any other non-PySpark workflow. Example: Pandas, SciKitLearn, Keras,etc)
    - Compare in terms of speed, ease of use, efficiency etc.
    - Provide a recommendation and justification of which workflow is more suited based on this use case.
    - Keep it withing 150 words

5. **Recommend strategies for how a person can chart their career (5 Marks).**
    - A 40-year-old mid-careerist who recently got retrenched from the Web (AI, Data Engineering) industry. You are to provide the following insights based on your findings and information available freely:
        - Which industry he or she should move into?
        - Give the reasons of why he or she should pursue that other than salary.
        - What are the skillsets needed for that industry?
        - How can he or she obtain the skillset?
    - The profile of the person is:
        - Years of experience in the following:
            - Web: 3
            - Service: 4
            - Education: 3
            - Oil:4

- Interest: Collecting blind boxes, 3D printing.
- Yearly salary: RM70,000
- Married with 10 cats, one goldfish and a dog.
- Owns a house with a monthly mortgage of RM2,400
- Monthly expenses: RM3,200

6. **Presentation (15 Marks)**
    - Share your findings with the government organization in 10 minutes or less.
    - The panel in the presentation consist of the following people:
        - Business user
        - Project sponsor
        - Chief Technology Officer

Submission Requirements (Please zip into a file and rename it with your admin number):

- Python files: Python Notebooks should have adequate documentation on the analysis using proper markdown and comments.
- Presentation slides
- Etc....

Submission dateline for Zip file on BS is:

- Week 17
- 9 Feb 2025, Sunday
- 2359

Presentation to be conducted on:

- Week 17
    - T1: Monday and Wednesday
    - T2: Monday and Thursday
- A link would be shared closer to the date to book your timeslot