DIPLOMA IN DATA ANALYTICS CO-OP

Capstone Project

Instructions

- This project is an individual assignment.
- Read each section carefully and provide answers.
- The following are the expected deliverables:
    - **Deliverable 1 - Project Plan (5%):** This is a formal project plan (2 pages) outlining the project purpose, work breakdown, milestones, schedule, and other requirements. This must be submitted within 1 week after the project kickoff meeting.
    - **Midterm Review (25%):** This is online class review to assess if student completed step 1 to 6 under **Requirements of Deliverables 1, 2 and 3-1**: (Data exploration (ERD) and data handling)
    - **Deliverable 2-1, 2-2 - Final Presentation (30%):** This is a technical presentation; Therefore, it must be professionally presented with adequate technical details. Each group must record their presentation using any collaboration tool and submit it along with the PPT file. Both the PPT file and the **video** (a link) are due before 11.59 PM on the set deadline.
    - **Deliverable 3-1 – Code File (25%):** Each group needs to submit a code file. The file must be professionally arranged, and each code segment must be clearly explained by using comments.
    - **Deliverable 3-2 – Individual Reflection (15%):** Each "individual" needs to submit their own professional reflection as per the guidelines provided in this project assignment file.
- Time allocated for the presentation: ***30 minutes*** for the recorded group presentation.
- Toronto School of Management (TSoM) requires students to maintain high standards of academic integrity. Students are responsible for conducting themselves honestly and ethically in all aspects of their academic career and for becoming familiar with this policy and abiding by all aspects of it. To support academic honesty at TSoM, all work submitted by students may be reviewed for authenticity. In submitting their own work to TSoM, students consent to their submissions undergoing such a review and being retained in a database for comparison with other work submitted by fellow students.

Project Requirements

Goal:

Synopsis:

The following are the key variables:

Your Task:

1. Identify a series of analytical objectives. These must include:
    a. At least 5 descriptive (statistical) questions.
    b. At least 3 predictive objectives (3 models).
    c. At least 2 cluster-based learning objectives.
2. Using a programming language, you need to achieve these objectives by following all the steps of the data handling process. Each group needs to submit ONE code file. The file must be professionally arranged, and each code segment must be clearly explained by using comments.
3. Design a PowerPoint presentation with "professional speaker notes" as per the requirements below. ***These notes should explain each slide in detail*** (You do not have to read out all the notes during the presentation). This is a technical presentation; Therefore, it must be professionally presented with adequate technical details. Each group must record their presentation using any collaboration tool and submit it along with the PPT file.
4. Submit your PowerPoint file, video presentation, code file, and individual reflection file before 11.59 PM on the deadline.
5. When submitting the video, please upload it in a shared folder and submit only the link to the video.

Requirements of Deliverables 1, 2 and 3-1:

The following sections must be clearly reflected in both deliverables 1 and 2. [Please refer to the attached rubrics for score breakdown]

1. Introduction
    a. Briefly introduce the case, data handling process, and presentation structure.
2. Analytical objectives. These must include:
    a. At least 5 descriptive (statistical) questions.
    b. At least 3 predictive objectives (3 regression models).
    c. At least 2 cluster-based learning objectives.
3. Assumptions
    a. List all the assumptions you had to make to complete the analysis here.
4. Data Analysis Process
    a. Briefly describe the analysis process (include a process map).
    b. You may refer to the subsequent sections. This section may outline the underlying theories. Please do not include any implementation steps or findings in this section.

5. Data Sanity Checks
    a. Importing Data.

       <u>For the following data handling steps write your code in a way that it automates data handling or cleaning part.</u>
    b. Variable Analysis
        i. Perform a variable summary.
        ii. Perform an initial visualization if required.
        iii. Check for the structural integrity of the dataset.
        iv. If any variable is not conforming, you may consider casting/conversion.
        v. Recognize and treat categorical variables.
    c. **Treating Duplicate** Entries
        i. Describe the process/ strategy.
        ii. Describe the findings/results.
    d. **Treating Missing** Values
        i. Describe the process/ strategy.
        ii. Describe the findings/results.
6. Descriptive Analysis and Findings: Data Summarization, Visualization, Statistical Analysis, **outliers' treatments** and Calculations
    a. Describe how the calculations were performed and why they had to be performed.
    b. Describe the use of visual artifacts in the analysis.
    c. Explain the findings.
7. Predictive Analysis and Findings
    a. Describe how **data partitioning** applied and how regression as a predictive analysis technique was used to achieve the objectives.
    b. Describe the use of visual artifacts in the analysis.
    c. Implementation of the **feature engineering, hyperparameter tuning, ensemble model**.
    d. Explain **model evaluation** and findings.
8. Cluster-based Analysis and Findings
    a. Describe how data partitioning applied and clustering was used to discover classes of market segments/ achieve the set objectives.
    b. Describe the use of visual artifacts in the analysis.
    c. Implement feature engineering, hyperparameter tuning, ensemble model.
    d. Explain model evaluation and findings.
9. References: Please follow the standard Microsoft APA referencing style.

Deliverable 3:

Individual Reflection and Peer Review. ***[20% of the final score | 100 points]***

1. Each student on your team must write their independent reflection on the project. Your reflection article should be no more than 2 pages (excluding the peer evaluation) and must be uploaded separately. Your task:
    a. ***[30 Marks]*** Choose a topic from the module that you found most interesting and explain it.
    b. ***[30 Marks]*** Explain how well the chosen topic was applied in the project.
    c. ***[20 Marks]*** Explain the strengths you recognized when executing the project.
    d. ***[20 Marks]*** Discuss the areas to improve when working on similar data analytics project

2. **Marking Rubric: Deliverable 1 – PowerPoint File and Live Presentation**

3.

Question	0-20%	20-50%	50-80%	80-100%	**Final Marks**
Sections	The answer does not meet the question's requirements. Invalid or incorrect answer.	A poorly structured and presented answer that demonstrates the lack of understanding of the subject matter. The section addresses some of the topic's objectives. The answer has major grammatical errors.	A well-structured and presented answer. The answer addresses most of the underlying subject matters of the question. The section addresses most of the topic's objectives. The answer has limited grammatical errors.	An exceptional answer that covers all the underlying subject matters of the question. The answer exerts critical thinking and includes examples. The section addresses all of the topic's objectives. The answer has no grammatical errors.	**Marks Out of 100**
Introduction	0-2	2-3	3-4	4-5	
Case & Analytical Objectives	0-2	2-5	5-8	8-10	
Assumptions	0-2	2-3	3-4	4-5	
Data Analysis Process	0-2	2-3	3-4	4-5	
Data Importation and Sanity Checks	0-2	3-10	11-15	16-20	
Descriptive Analysis	0-2	2-5	5-8	8-10	
Predictive Analysis and Visualization	0-3	3-7	7-12	12-15	
Cluster Analysis and Visualization	0-2	2-5	5-8	8-10	
PPT Presentation File	0-2: Poor quality	2-3: Structure is maintained. Include Notes.	3-4: Structure is maintained. Descriptive Notes. Accurate referencing.	4-5: Structure is maintained. Accurate referencing. Well-written Notes. High-quality presentation.	
Group Presentation (Audio/Visual)	0-3: Poor Quality Presentation	3-7: Satisfactory Presentation	7-12: Very Good Presentation	12-15: Exceptional Presentation.	
Total					

4. **Marking Rubric: Deliverable 2 – Code**

5.

Question	0-20%	20-50%	50-80%	80-100%	**Final Marks**
Sections	Invalid or incorrect answer. The code doesn't run, and the steps are incorrect.	The code works but the intermediate steps are not compatible with the answer. A correct answer with a poor approach. Some of the steps are correct, although the code doesn't work.	The code works and most of the steps meet the question's expectations. Some comments are available. Most of the steps are correct, although the code doesn't work.	The code works and all the steps meet the question's expectations. The code is well commented and well-structured.	**Marks Out of 100**
Importing Data	0-2	2-3	3-4	4-5	
Variable Analysis	0-2	2-5	5-8	8-10	
Treating Duplicate Entries	0-2	2-5	5-8	8-10	
Treating Missing Values	0-2	2-5	5-8	8-10	
Descriptive Analysis	0-2	2-5	5-8	8-10	
Predictive Analysis and Visualization	0-3	3-7	7-12	12-15	
Cluster Analysis and Visualization	0-2	2-5	5-8	8-10	
R code file quality	0-2: Poor quality	2-5: Structure is maintained. Some code is explained using comments.	5-8: Structure is maintained. Most of the important code sections are explained using comments.	8-10: Structure is maintained. All important code sections are explained using comments.	
Total					

6.