Humber BIA

Big Data 2

Prabhpreet Sidhu

Case Study 2

Assignment Instructions

This assignment contains two (2) questions. Each question has one or more tasks. There are three possible types of tasks: tasks that require you to write code; tasks that require you to write text responses; and tasks require you to write math and/or diagrams.

Bonus marks will be provided for using Machine Learning in Q2.

For tasks that require **code**:

- Use Python, Spark, Hive or R to complete the task.
- Submit code that runs without errors.
- Submit code that is reproducible. E.g., set random number seeds as appropriate. You should be able to run you code again and again and again, from the top of the file to the bottom of the file, and get the exact same results each time. I should be able to run your code, from scratch, on my machine, again and again, and get the exact same results that you get.
- Submit code that is organized. Make your code readable. Provide comments to describe what the code is doing and why. Don't leave "old" code laying around. Overall, if your code is clear and easy to read, then we will be happy. When we are happy, we give better marks.

For tasks that require **text responses**:

- Be clear about which task you are responding to.
- Use English. Use proper grammar, spelling, and punctuation. Be professional and clear. Be complete, but not overly-verbose.
- You may refer to your code. Please do so very clearly. E.g., "As can be seen in on line X of file Y…"

For tasks that require **math** or **diagrams**:

- Type or insert your response in the presentation document.
- Be clear about which task you are responding to.
- Where appropriate, please use PowerPoint's features such as Equations, Symbols, or Diagrams. You may also import diagrams created in other programs, as long as they are clear and easy to read.
- Please do not submit pictures/scans of hand-written diagrams, equations, or work of any kind.

Your assignment submission should contain exactly two files:

- Presentation (PPT, Canva, Google Slides, etc)
- Technical report with documented code

ELI5

PREAMBLE

"If you can't explain it simply, you don't understand it well enough." – Albert Einstein

Explaining technical concepts to a non-technical audience is an underappreciated skill; one which the Humber BIA program aims to give its students; and one that will truly set you apart in the job market. The only way to gain a skill is by practice, so here we go.

Answer each question below as though you were talking to a 5 year old (equivalently: a grandma, or a completely non-technical manager, or an Ivey grad). Use your own words. Use analogies where possible. Examples are better than theory. Keep it short, but be complete. Use simple, plain English. Do not use business ChatGPT lan or buzzwords like *actualize*, *empower*, *fungible*, *leverage*, or *synergize*. Do not use technical buzzwords that most people don't know like *model, agile, bandwidth, IoT, blockchain, AR, VR, actionable insights.* Inform the audience without going into too much technical detail. Your goal is to truly help them understand, not to give what you feel is a "technically precise" answer and move on (but they still don't understand!). Don't be that guy!

TASKS

A. [Text] What is "Big Data" and how is it different than "regular data"?
B. [Text] What is Hadoop? Hint: What problems in previous data storage and processing was Hadoop designed to solve? How did Hadoop accomplish that?
C. [Text] How does Big Data and the cloud help Machine Learning? Hint: Machine Learning and AI are often discussed in tandem with the trend towards leveraging big data and cloud computing. Provide some reasons why big data and cloud computing enable recent advancements in and adoption of machine learning.
D. [Text] What is NoSQL?
E. [Text] Name three ways topic modeling could help a bank.
F. [Text] What is Apache Spark, exactly, and what are its pros and cons?

1. Sentiment Analysis via ML-Based Approach

Preamble

Download the "Product Sentiment" dataset from the course portal: *sentiment_train.csv*.

Tasks

1. [Code] Perform analysis on the dataset using any analytics approach (BI, AI, ML) using Hadoop, Spark, Python or R .
    a. Load, clean, and preprocess the data as you find necessary.
    b. Present your analysis in a business presentation format.
    c. Bonus marks for building a ML based classification model.
2. Share 3 recommendations and insights based on your analysis.
3. **Machine Learning based solutions only - Show five example instances in which your model's predictions were incorrect. Describe why you think the model was wrong. Don't just guess dig deep to figure out the root cause.**