DATA102: Data Mining

T3 2023-2024 Project Specifications

Project Timeline:

- ➢ **Deadline of deliverables on August 1, 11:59 PM**
- ➢ **Final Project Presentation (online) on August 3 to 7.**
    - There will be a GoogleSheet for the final project demo. First come first serve basis on the timeslots. Time slots for each group would be 30 mins max. 20 mins for presentation and demo and 10 mins for Q&A.
    - All group members must be present during the presentation. The camera of the current presenter should be open.

- ➢ **Chosen topic and dataset by July 5, 2024 at 11:59pm -** post your chosen topic and dataset in the Animo Space discussion thread.
    - Topic and dataset should be unique per group. I will comment in your post to confirm that you may proceed.
    - You may change topic and dataset along the way, but make sure to post in the AnimoSpace discussion thread again and submit exploratory data analysis again.

- ➢ **Project update (synchronous or asynchronous) from July 20 to 24.**
    - The group is expected to at least be doing exploratory data analysis.
    - Should this be synchronous, I will post a GoogleSheet for the schedule of the 15 to 20 minute time slots for the project update.

Objectives:

- ➢ Apply data mining concepts and methodologies to a chosen relevant topic
- ➢ Identify the problem statement and formulate data-driven recommendations

Instructions:

1. Choose one (1) dataset from the list of datasets provided by your instructor.
2. Go understand the dataset and its relevance to the overall objective, and identify critical insights using data mining techniques to form recommendations.
    a. Devise a problem statement containing one (1) business objective/research question supported by at least one (1) data mining objective. Provide a data dictionary or data documentation as well.
    b. **Exploratory Data Analysis** - provide a explore and get a high-level understanding of your dataset using OLAP operations and descriptive statistics
    c. **Modeling** - generate additional insights which are not found in the data understanding but may support the problem statement using different modeling techniques. Implement at least two models with different experiments (e.g test different parameters).
3. Present your understanding of the problem, analysis and findings, and recommendations. Make sure that the presentation is addressed to a general audience that may have varying levels of technical expertise.

Tips:

- Do not limit yourselves to just business datasets.
- Be creative! Almost all fields involve data analysis today. You can devise a problem statement whether it is business-related or not.
- Make sure you have a good understanding of your data.
- Document every step of the process. Provide screenshots, graphs, related literature etc. of the analysis as evidence for your conclusions.
- As emphasized in class, go back to the context. Make sure that your conclusion responds to the problem statement you created.

Suggested Topics and Datasets:

- Kaggle
- UCI
- You may consult the use of other datasets but seek approval first.

** Make sure to properly cite the source of the dataset in your deliverables

Grading Criteria

- Exploratory Data Analysis, Effective Use of Data Preprocessing and Visualization - 35%
    - Well documented and effective use data preprocessing techniques
    - Effective Exploratory Data Analysis by understanding initial patterns within the data, detecting outliers or anomalous events, finding interesting relations among the variables, basic or descriptive statistics.
    - Clear, assured delivery with suitable use of visual aid
    - Most suitable charts used
    - All of the data in the dashboard are accurate and not manipulated
    - Credible references (complete links) to datasets, methods/techniques are cited
    - Data source presented for each image/chart is preferred

- Model Based Analysis and Insights - 35%
    - Implement at least two models with different experiments (e.g test different parameters). If you are only able to implement one model with various experiments, then your group may not be able to get the full points for this criteria.
    - Well documented experiments and effective use of data mining or machine learning techniques.
    - Selection and explanation of performance metrics.
    - The problem statement is clear
    - Capability to tell a compelling and engaging story that is logical and critical
    - Use of clear, focused, and quality visual analysis, compelling charts, and graphs to depict the datasets and give viewers meaningful insights
    - Graph(s) and storyboard are well-structured and organized
    - Related literature, documentation, references etc.
- Recommendations - 10%
    - Recommendations are proposed on the basis of data analysis
    - Recommendations are creative, feasible, specific, sustainable, original, and impactful
    - There is clear information on relevant stakeholders involved in the realization of the recommendations/solutions
- Technical Report/Oral Presentation/Defense - 20%
    - Well written technical report in ACM/IEEE format.
    - Must be kept within the allowed duration
    - Should be addressed to a general audience - engaging, comprehensive, with sufficient technical details

    - Should be able to answer follow-up questions or deflect unanswerable ones effectively

List of deliverables

- Signed certificate of authorship
- Technical Report in ACM/IEEE format. Sample template in our AnimoSpace coursepage.
- IPython Notebook
    - Must be properly working well documented
- Presentation Deck
    - Make sure it's comprehensive but still presentable. Include what is essential only!
- Appendix Document
    - Contains supporting screenshots and documentation
    - Redirect your reader to specific pages or figures in your appendix.