Final Project

Data Patterns and Representations

Introduction

Welcome to the culmination of your learning journey in this course! In lieu of a traditional final exam, we have an exciting and dynamic challenge for you. You and a partner will undertake a comprehensive project that synthesizes all the key concepts you have learned throughout the course. This is your opportunity to demonstrate your expertise as emerging data professionals in a practical, real-world scenario.

- Format: Live Presentation during an assigned time
- Duration: 15-20 minutes
- Partnership: Individual or Pairs

Your project will be a testament to your understanding, analysis, and application of the data science concepts covered in this course.

Info: To successfully complete the final project, it is essential to approach it in a phased manner. Although there is a certain degree of overlap between the stages, each phase incrementally builds upon the previous one, mirroring the structure of the course modules. To maintain your progress and ensure readiness for the final presentation, you will be required to submit various assignments at different points throughout the course. These interim submissions are crucial not only to keep you on track, but also to provide opportunities for feedback and improvement as you advance towards the final stage.

Info: The milestones provided represent a suggested timeline to help you break the project into manageable steps and stay on track throughout the semester. While these milestones are not graded, we will hold periodic check-ins with teams to discuss progress, provide guidance, and ensure that deliverables align with instructor expectations. This collaborative approach is designed to support your success and maintain the quality of your final presentation.

Phase 1: Topic and Dataset Selection & Exploratory Analysis

Topic Selection

The first step of your project is to choose a topic that genuinely interests you, something that connects to your personal passions, career goals, or an issue you care about. The topic should be broad enough to support multiple lines of inquiry, yet focused enough to yield a meaningful conclusion or recommendation by the end of the semester.

Warning: Avoid cliché or overused topics, such as those based on Kaggle competitions (e.g., Titanic, Iris, MNIST, Population). These datasets are well-worn and limit your opportunity to think critically or creatively. Your topic should allow you to tell a unique and relevant story, even if your final investigation is partial, your story should matter.

Tasks:

1. Write a short paragraph describing your proposed topic.
2. Explain why the topic is relevant or important (personally, professionally, or socially).
3. Suggest the types of questions or challenges you might explore through data.

Milestone: Submit your topic proposal and rationale by the end of Module 3. Feedback will be provided before you move to dataset selection.

Dataset Identification

Once your topic is approved, the next step is to identify a dataset (or multiple datasets) that support meaningful analysis of your topic. You should find data that offers depth and complexity—not just something to plug into a model. If your data has gaps, that's fine—as long as your analysis still points toward a valid recommendation or insight.

1. Search for a dataset (or combine multiple sources) that aligns with your chosen topic.
2. Submit a link to the dataset(s), along with a summary describing:
   - the source and structure of the data;
   - how it connects to your topic and questions;
   - any limitations or challenges you anticipate.
3. Propose a set of exploratory questions or hypotheses to guide your initial analysis.

Milestone: Dataset summary and exploratory questions, due by the end of Module 4. Once you've finalized your dataset, email the instructors a brief description and schedule your check-in when prompted.

Exploratory Data Analysis (EDA)

With your dataset in hand, you'll begin exploring its structure and content. This phase is about understanding what your data can (and can't) tell you, and beginning to shape your project scope.

Tasks:

1. Inspect numerical and categorical features.
2. Calculate basic statistics (e.g., mean, median, standard deviation).
3. Visualize key features and patterns using appropriate plots (e.g., histograms, box plots, scatter plots).
4. Refine your research questions or objectives based on what the data reveals.

Milestone: A Jupyter Notebook or equivalent file summarizing your findings, due by the end of Module 5.

Phase 2: Representation-Specific Analysis

Data Representation and Preprocessing

This phase focuses on analyzing your dataset based on its specific representations (e.g., time series, images, text). You will apply preprocessing techniques to prepare the data for modeling and analysis.

Tasks:

1. Apply representation-specific preprocessing techniques (e.g., scaling for numerical data, tokenization for text data).
2. Address missing values, outliers, and inconsistencies.
3. Engineer features to enhance the dataset's predictive power.

Milestone: Documented preprocessing code and a report describing the applied transformations, due by the end of Module 10.

Advanced Analysis and Modeling

You will conduct a more thorough analysis using modeling techniques relevant to your data representation.

Tasks:

1. Select an appropriate model for your dataset (e.g., regression for numerical data, clustering for unsupervised learning).
2. Train and validate your model, documenting key performance metrics.
3. Interpret the results in the context of your dataset and project goals.

Milestone: A Jupyter Notebook or equivalent file containing your modeling process and results, due by the end of Module 11.

Phase 3: Creating a Data Story

Storytelling and Visualization

In this phase, you will focus on creating a compelling narrative around your data set, integrating the principles of effective storytelling and data visualization.

Tasks:

1. Develop a narrative that contextualizes your findings and answers the questions asked during the exploratory phase.

2. Create visuals that effectively communicate your insights, adhering to the design principles discussed in class.
3. Use tools such as Tableau, PowerPoint, or Python to design your visuals.

Milestone: A storyboard outlining your narrative and visualizations, due by the end of Module 13.

Final Presentation

The culmination of your project will be a live presentation where you share your data story with the class. Choose a medium that best fits your project (e.g. slide deck, interactive dashboard).

Tasks:

1. Present your findings and narrative in a 15-20 minute live session.
2. Ensure that all group members actively participate in the presentation.
3. Prepare to answer questions from the audience.

Milestone: Live presentation, scheduled during the final 2 weeks of the course.

Conclusion

This project is designed to challenge you to apply everything you have learned in the course, from data exploration and modeling to storytelling and visualization. Use this opportunity to showcase your skills, creativity, and expertise as a data professional. Good luck!