MSIS 680 Advanced Machine Learning

Final Project Guide

Purpose

This final project aims to put to work the tools and knowledge you learned throughout this course. This provides you with multiple benefits:

1. It will provide more experience using data analysis tools on real-life datasets.
2. It helps you become a self-directed learner. As a data analyst/data scientist, a large part of your job is to self-direct your learning and interests to find unique and creative ways to find insights into data.
3. It starts to build your data science portfolio. Establishing a data science portfolio is a great way to show potential employers your ability to work with data.

Project Goal

The principal goal of this project is for you to retrieve a real-life data set, work with the data, and perform exploratory analysis and predictive analysis/machine learning with the data. You can work individually or team up with your classmates (**maximum of three people** per group).

Group signup link: https://umassboston.instructure.com/courses/4516/groups#

Project Data

You are encouraged to seek your own real-life dataset and perform data analysis for this final project. This is because of the following reasons:

- Because you are working with a unique dataset, your analysis will be less likely to be similar to others, therefore, more attractive to your peers and me.
- Choose your own dataset that allows you to perform data analysis close to your field of interest. Project experience is one of the most important sections of your resume and is often what I will pay particular attention to when screening job applicants. Thus, you will want to work with a dataset close to your desired working industry. For example, if you are working/going to work in the marketing industry, performing a customer

segmentation analysis or review analysis will be an excellent project experience that light up your resume.

Your selected dataset should satisfy:

- Sufficiently large and complex, containing at least 2000 observations and more than 10 distinct features to ensure a comprehensive analysis.
- Complex dataset and data analysis (Social Networks, Time Series, and Natural Language Processing) will be granted 5pt bonus points.

Following are a list of useful resources & online data repository:

Websites You Can Look For New Datasets

- Kaggle Datasets
- UCL Machine Learning Repo
- Google Dataset Search

Other Interesting Datasets

- Coronavirus Data in the United States
- Petfinder.com Dog Data
- Hotel Booking
- NFL Stadium Attendance
- Spotify Songs
- Zillow Housing Data
- Yelp Review Dataset

Project Report

You will write a Jupyter Notebook (ipynb) file that provides the sections in the grading rubric below as your final project report. There is no need for an additional Word document. As a future data analyst or data scientist, it's essential that you develop the ability to organize your analysis and reports directly within a Jupyter Notebook. Familiarize yourself with Google Colab's capabilities in creating rich text and markdown documentation by following this link: Google Colab Tutorial | Markdown | Rich Text Documentation | Image Upload

You will need to import, assess, clean the data, and then come up with your research questions (3-5 questions) that you would like to answer from the data by performing exploratory data analysis and predictive analysis/machine learning. Some thoughts to help you:

- Your project should be a logical, cohesive story–not simply a bunch of graphs created for the sake of making them.
- Speaking of insights, keep in mind that your project should follow the chain of data -> insights -> actions. As a future data analyst (or data scientist), you work to create insights

that lead to actions, not to waste 40 hours on an awe-inspiring visualization that is ignored directly after a presentation and never used again.

- Simple descriptive statistics and exploratory data analysis can (and usually) yield more of an immediate impact than a complicated model. Brooke Watson (Links to an external site.) gave a compelling and enlightening presentation at the 2019 RStudio Conference on how the ACLU used various R packages to count and reunite families.
- Although each data set's data dictionary contains some additional questions worth pursuing, try to be creative in your analysis and investigate the data in a way that your classmates most likely will not. Creativity is an essential ingredient for a good data scientist!

Grading Rubric

Section	Standard	Possible Points
Introduction	**1.1** Provide an introduction that explains the problem statement you are addressing. Why should I be interested in this? **1.2** Provide a short explanation of how you plan to address this problem statement (the data used and the methodology employed) **1.3** Discuss your current proposed approach/analytic technique you think will address (fully or partially) this problem. **1.4** Explain how your analysis will help the consumer of your analysis.	20
Data Preparation	**2.1** Original source where the data was obtained is cited and, if possible, hyperlinked. **2.2** Source data is thoroughly explained (i.e., what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.). **2.3** Data importing and cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process. **2.4** Once your data is clean, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible. **2.5** Provide summary information about the variables of concern in your cleaned data set. Rather, provide me with a consolidated explanation, either with a table that provides summary info for each variable or a nicely written summary paragraph with inline code.	10
Exploratory Data Analysis	**3.1** Uncover new information in the data that is not self-evident (i.e., do not just plot the data as it is; rather, slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information). **3.2** Provide findings in the form of plots and tables. Show me you can display findings in different ways. **3.3** Graph(s) are carefully tuned for the desired purpose. One graph illustrates one primary point and is	20

	appropriately formatted (plot and axis titles, a legend if necessary, scales are appropriate, etc.). **3.4** Table(s) carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features. The size of the table is appropriate. **3.5** Insights obtained from the analysis are thoroughly, yet succinctly, explained. Easy to see and understand the interesting findings that you uncovered.	
Predictive Data Analysis	**4.1** What's your target (Y) variable and what are your features (X) variables; **4.2** What are your predictive models and why choose these models; **4.3**What are your evaluation metrics and why to choose these metrics; **4.4** What pre-processing (encoding, standardizing, etc.) methods you have applied; **4.5** How to fine-tune your models and how the predicting accuracy improves; **4.6** What's your final choice of model and what accuracy you have achieved.	20
Summary	**5.1** Summarize the problem statement you addressed. **5.2** Summarize how you addressed this problem statement (the data used and the methodology employed). **5.3** Summarize the interesting insights that your analysis provided. **5.4** Summarize the implications to the consumer of your analysis. **5.5** Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.	20
Formatting & Other Requirements	**6.1** Achievement, mastery, cleverness, creativity: Tools and techniques from the course are applied very competently and, perhaps, somewhat creatively. Perhaps the student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, and unusually sophisticated application of tools from the course.	10
Bonus Point	For doing analysis and modeling on more complex dataset (Social Networks, Time Series, and Natural Language Processing).	5

Total possible points: 105

I expect your report to tell a story with the data. I do not want you to just report some statistics that you find but, rather, to provide a coherent narrative of your findings.

Exception about additional group members:
While the course policy typically limits groups to three members to ensure fairness and workload balance, exceptions may be considered on a case-by-case basis for projects that demonstrate a strong justification for an additional group member. If you wish to request an exception for additional group members, please submit a comprehensive proposal detailing:

- The specific topic and dataset you plan to do.
- An explanation of the project's complexity or scope that necessitates additional members.
- A breakdown of the proposed roles and responsibilities for each team member.
- An outline of the project's objectives, methodology, and expected outcomes.

The proposal should clearly articulate the necessity of a larger group size due to the project's advanced topics, technologies, or significant workload that requires a broader range of skills and resources. Upon review of the proposal, a decision will be made by the instructor regarding the exception request. Please submit the proposal before project signup due date to ensure sufficient worktime.