SPE 486 Final Project Specifications

SPRING 2022

OBJECTIVES

Your goal is to employ best practices of data science, analytics and visualization to a domain challenge problem or opportunity of your choosing, to explore its veracity, performance & real-world relevance.

DELIVERABLES

- 15 minute, detailed professional conference style presentation satisfying the requirements outlined below
- 25 Page Paper
- Any and all code associated with your final product
- Late work will be accepted according to the course syllabus guidelines

REQUIREMENTS

1. Research Question

a. What is your research question(s)?
b. Why are you trying to do this?
c. Are you trying to explain or predict phenomena? Detail how and why choosing one or the other could change how you assess and perform best practice data analytics very differently.

2. Data

a. What data do you use?
b. What is the unit of analysis?
c. What type of data is it, cross section, time series, panel, etc.?
d. What is the DV, what are the IVs?
e. How do you operationalize variables and features?
f. Given only the data used, anything interesting or troubling that you might see about their operationalization for phenomenological inferences or validation later?
g. Given your type of data and your research question, which data science techniques are appropriate?

3. ETL Assessment

a. What, if any, ETL procedures do you perform? Why?

4. EDA + Visualization Assessment

a. What type(s) of EDA and visualization do you perform and provide?
b. What does the data tell us? What inferences can we make?
c. How does the EDA inform us on how best practices data science should be conducted on this particular topic?

5. EDA + Visualization Recommendation
   a. What other types of EDA and visualization could you be performing? Please detail.
   b. What types of techniques/graphs could you use?
   c. What data would you use for each?
   d. What would be the specific inferences you could make from each of your suggested EDA and visualization types?
   e. Why are you suggesting these other types of EDA and visualization?

6. Empirical Specifications
   a. Are you doing supervised or unsupervised learning?
   b. If supervised, what is the empirical f(x) model specification(s) you are testing?
   b. Write it out in both mathematical notation of the features/variables as well as natural language for clear communications (please remember, there might be multiple empirical model specifications you are testing across multiple predictive modeling techniques).

7. Predictive Modeling Techniques and Algorithms
   a. List all the predictive modeling techniques you use (please remember that multiple modeling techniques can be used with the same or different empirical modeling specifications)
   b. Are they supervised or unsupervised? Why?
   c. Why are these techniques appropriate to be used given your research question and available data?
   d. What are the pros and cons of each predictive technique you are using?
   e. What, if any, other types of predictive modeling techniques would you recommend that you did not perform? Why?

8. Bias-Variance Tradeoff
   a. How do you control for the bias-variance tradeoff? Please explain.
   b. What sort of re-sampling methods do you perform? Please detail.
      i. Why do you think they are appropriate and/or complete?
      ii. What, if any, other types of these techniques would you recommend given your data that you did not perform? Why?
   c. What sort of feature selection do you perform? Please detail.
      i. Why do you think they are appropriate and/or complete?
      ii. What, if any, other types of these techniques would you recommend given your data that you did not perform? Why?
   d. What sort of regularization do you perform? Please detail.
      i. Why do you think they are appropriate and/or complete?
      ii. What, if any, other types of these techniques would you recommend given your data that you did not perform? Why?
   e. What sort of out of sample predictive validation or ensemble methods do you perform? Please detail.
      i. Why do you think they are appropriate and/or complete?
      ii. What, if any, other types of these techniques would you recommend given your data that you did not perform? Why?

9. Predictive Performance

a. How do you predictive modeling technique(s) perform?
b. Specifically assess each of your empirical and predictive model(s) accuracy and fit, both individually and relatively if multiple. What insights can you make?
c. What criteria should be used to assess the accuracy and fit for each different type of predictive and/or empirical model? Why?
d. Are there other criteria you would suggest given your data and models? Why?
e. What are your recommendations to increase your predictive model performance? Why?

10. Empirical Insights

a. What are your specific empirical findings given the various data science methods you use?
b. Given your model(s) performance, what is the substantive interpretation of results for specific variables and features? Please write out and explain which features matter and why, including significance, directionality and magnitude of substantive effects.
c. What are your recommendations to increase the empirical findings? Why?

11. Your overall execution assessment

a. Assess how well did you think you executed your data science? Why?
b. Any shortcomings given the best practices on process we have developed? Why?
c. Any other predictive methods or techniques you could have used? Why?
d. What would be your next steps on this topic?