Boston University Virtual

Milestone One Guidelines and Rubric

Module C: Data Science Capstone

Overview

In Milestone One, you will use machine learning models to complete data analysis and evaluation of multiple datasets. Each week's content will cover a new topic or revisit topics from previous Mods. Using the project that you selected in Module B, you will apply each week's topic(s) and perform multiple analyses on your project's datasets. Your work should be reflected in your weekly Jupyter Notebook homework. Your project will be solved individually, and your solutions will be submitted as an individual summary document in Week 7.

In this course, there are no specific questions to answer in your Jupyter Notebook files; your general goal is to analyze your data using the methods you have learned in this course and in this program, and to draw interesting conclusions based on the project you have chosen. In some cases, you may already have a clear question you want to answer based solely on what you see in your data set. In other cases, you may need to explore multiple methods and analyses before determining which questions should or need to be answered.

This is an iterative process; one that is designed to be self-governed and open-ended. As you complete each milestone, you are developing the narrative behind the data set and, by the time you are ready to complete your Integrated Capstone Project, determining the key questions that a client related to your project would be interested in having answered.

Requirements

Criteria for Milestone One include:

1. A statement or description of the chosen project

2. An explanation of information learned from using the following modeling techniques. Furthermore, you must select one to two of these weeks (a–f) to discuss in a thorough and interesting way:
    a. Polynomial and interaction terms (Week 1)
    b. Lasso, ridge, elastic net regression (Week 2)
    c. Forward and backward selection, principal component regression (PCR), partial least squares regression (PLSR) (Week 3)
    d. Logistic regression and feature scaling (Week 4)
    e. Support vector machines, the kernel trick, and regularization for support vector machines (Week 5)
    f. Decision trees and random forests (Week 6)

3. For each topic above (2a–2f), an explanation of the following (as applicable to your dataset):
    a. How did you avoid overfitting?
        i. You should mention the techniques used to prevent overfitting, why these techniques were expected to be helpful, and the results.
    b. What metric(s) did you use, and what was the result? Did you use any hyperparameter tuning?
        i. You should describe the metrics and hyperparameter tuning performed to determine hyperparameters and include a discussion of the meaning of these parameters.
    c. What aspects of the results were expected or unexpected?
    d. How did your Exploratory Data Analysis help with the modeling?
    e. What sources did you rely on, apart from the course materials, to learn about this model.
        i. In this context, it is expected that you will find external sources, such as free educational URLs, and share and discuss them with other students through YellowDig.

4. Specific and supported conclusions drawn from the models.
    a. How has your work up to this point helped address any research questions or other questions of interest?
    b. Provide an explanation as to how you know the conclusions are true, potentially addressing any or all of the topics mentioned in Requirement 3 (i.e., overfitting, metrics, etc.)

   c. This explanation should show that:
      i. The conclusions are true from a quantitative perspective, and
      ii. Explain why the conclusions are relevant and/or important.

In each case, it is up to you to choose and/or emphasize the items that are most interesting to talk about, but you must strive for both *breadth* and *depth.* That is, your grade will be based on:

- Breadth: Briefly discussing all of the topics in Requirement 2, and
- Depth: Choosing and discussing 1–2 weeks' topics in a very thorough and interesting way. Make sure you explicitly state in your document which topics you are delving deeply into explaining.

Formatting

- Your summary document must be an 8- to 10-page Microsoft Word document with double spacing, 12-point Times New Roman font, and one-inch margins.
- Sources should be cited according to APA style.
- Your document must contain an appendix consisting of your Jupyter Notebook completions from each of the first six weeks of the course, showing how you applied the topics of each week to your project.
    - There is no page limit for the appendix containing your Jupyter Notebook and they are not included in the page requirement of this assignment.
    - It is okay, and expected, that some material in the summary document will be duplicated in the Jupyter Notebook files. For instance, you will likely have to show graphs from these files. Material taken from these files provides the evidence for your claims in the summary.

Milestone One Rubric

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Problem Statement/ Description	5 points Provides a clear and detailed description of the chosen project, including insight into the potential impact of the project.	4.25 points Provides a description of the chosen project but lacks details or insight into the potential impact of the project.	3.5 points Provides a brief or incomplete description of the project; provides minimal details into the potential impact of the project.	0 points No submission or provides little to no description of the chosen project.	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Conclusions: Depth	30 points Provides accurate and insightful descriptions of how the chosen week(s)' modeling techniques were used on the project datasets, including evidence-supported conclusions that were derived from them and how the issues of those models	25.5 points Provides accurate descriptions of how the chosen week(s)' modeling techniques were used on the project datasets but lacks some insight or depth of analysis in the conclusions.	21 points Provides descriptions of how modeling techniques were used on the project datasets, but is a general, superficial analysis or aspects of the analysis are inaccurate. It may be unclear which week's topics were chosen for the in-depth analysis.	0 points No submission or provides incorrect information or shows no evidence of in-depth analysis.	30

	(overfitting, etc.) were mitigated.				
Conclusions: Breadth	30 points Provides clear and detailed information about all modeling techniques covered in Weeks 1–6, including evidence-supported conclusions that were derived from them and accurately describing the issues of those models (overfitting, etc.)	25.5 points Provides information about a range of models and their respective issues but conclusions may be limited due to omitting some key models and/or issues.	21 points Provides information on a narrow range of cases.	0 points No submission or provides incorrect information or shows no evidence of breadth analysis	30

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Overfitting	10 points Provides a comprehensive and insightful explanation of overfitting including techniques to prevent overfitting, how these techniques were	8.5 points Provides a general explanation of overfitting that covers techniques to prevent overfitting, how these techniques were expected to help, and	7 points Provides minimal explanation of overfitting that somewhat covers including techniques to prevent overfitting, how these techniques were	0 points No submission or provides incorrect information or shows no discussion of overfitting.	10

	expected to help, and the results.	the results, but lacks detail and insight.	expected to help, and the results. Lacks significant detail and makes no insightful connections to the problem.		
Metrics and Hyperparameter Tuning	5 points Provides a comprehensive and insightful explanation of how metrics and hyperparameter tuning were performed and the meaning of the hyperparameters as it connects to the problem.	4.25 points Provides an explanation of how metrics and hyperparameter tuning were performed and the meaning of the hyperparameters. The connection to the problem is established but lacks some key details.	3.5 points Provides a minimal explanation of how metrics and hyperparameter tuning were performed and the meaning of the hyperparameters. Lacks significant detail and makes no insightful connections to the problem.	0 points No submission or provides incorrect information or shows no discussion of metrics and hyperparameter tuning.	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Expected/ Unexpected	5 points Provides detailed and insightful discussion of expected/unexpected	4.25 points Provides detailed discussion of expected/ unexpected aspects of data as evidenced by	3.5 points Provides some discussion of expected/	0 points No submission or does not discuss expected/	5

	aspects of data as evidenced by models.	models, but insight is lacking.	unexpected aspects of data as evidenced by models.	unexpected aspects of data.	
Exploratory Data Analysis (EDA)	5 points Provides detailed and insightful discussion of the usefulness/non-usefulness of EDA for model results.	4.25 points Provides detailed discussion of the usefulness/non-usefulness of EDA for model analysis, but insight is lacking.	3.5 points Provides some discussion of EDA's relevance or non-relevance.	0 points No submission or does not discuss the relevance or non-relevance of EDA to model analysis.	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Sources and Citations	5 points Provides and describes many sources outside of the course materials and cites them in proper APA format. The sources are used to support the understanding of models described and/or the conclusions being offered.	4.25 points Provides and describes some sources outside of the course materials and cites them in APA format. The sources are used to support the understanding of models described and/or the conclusions being offered.	3.5 points Provides limited sources outside of the course materials and/or cites them in a non-standard format, or the sources are minimally used to support the understanding of the models described and/or the conclusions being offered.	0 points No submission or does not provide sources outside of the course materials or the sources do not support the understanding of the models described and/or the conclusions being offered.	5

Communication	5 points Consistently and effectively communicates ideas clearly, concisely, and effectively for a specific audience.	4.25 points Generally communicates ideas for a specific audience clearly and effectively.	3.5 points Communicates ideas with some clarity but is inconsistent in organization or effectiveness.	0 points No submission or shows no evidence of consistent or effective communication.	5

Criteria	Exemplary	Proficient	Needs Improvement	Not Evident	Points
Total					**100**