Part 2 Final Project Rubric

Deliverable: Jupyter Notebook or PDF Report

The Jupyter notebook should have Project Topic, Data Cleaning/EDA, and Plans for Model Approach sections. If your work doesn't fit into one notebook (or you think it will be less readable by having one large notebook), make several notebooks or scripts in the GitHub repository and submit a report-style notebook or pdf instead.

Part of the project is to create a GitHub repository with your work. This repository needs to be specifically for this project. While Part 2 does not require you to submit the link, it would be a good idea to set up the repository and commit to it throughout the project.

Prompt	Points					
Project Topic Is there a clear explanation of what this project is about? Does it state clearly which type of problem? E.g. classification or regression	(0 pts) Not included in the project	(1 pts) States the type of problem (e.g. classification or regression). **Missing** explanation of what project is about.	(2 pts) Gives a clear explanation of what the project is about. **Missing** the type of problem (e.g. classification or regression).	(3 pts) Gives a clear explanation of what the project is about **and** clearly states the type of problem (e.g. classification or regression).		
Project Topic Is the goal of the project clearly stated? E.g. why it's important, what goal the author wants to achieve, or wants to learn.	(0 pts) Not included in the project	(1 pts) Needs improvement — attempts but doesn't get across the motivation or goal for the project	(2 pts) Very Good — clearly states the motivation or the goal for the project			
Bonus (Optional Extra Credit) ***Is the project topic creative?***-- For the final project, it is a valid strategy to solve an existing online problem (e.g., go on the UCI Machine Learning Repository and work on the default task.) ***A bonus is available if you want to stretch and define your own project problem.*** For the bonus, you can use your own data or existing data (e.g., available online). If you are trying for the bonus, make sure to add a **Bonus section** in your Jupyter notebook/report and provide a brief explanation of why your topic meets the bonus criteria.	(0 pts) No extra credit attempted/did not define a creative project problem.	(2 pts) Includes a **Bonus section** in the Jupyter notebook/report with a **brief description** of how the learner defined their own creative project topic.				

Data Is the data source properly cited and described? (including links, brief explanations)	(0 pts) **Missing both of the following:** Does not include a brief explanation of where the data is from/how it was gathered **and** does not include a citation (using the format of a style manual like APA) for either a public or unpublished dataset.	(1 pts) **Includes both of the following:** a brief explanation of where the data is from/how it was gathered **and** cites the dataset (using the format of a style manual like APA) for either a public or unpublished dataset.				
Data Is the data description explained properly? The data description should include the **data size.** ● E.g. for **tabulated data**: number of samples/rows, number of features/columns, bytesize if a huge file, data type of each feature (or just a summary if too many features- e.g. 10 categorical, 20 numeric features), description of features (at least some key features if too many), whether the data is multi-table form or gathered from multiple data source. ● E.g. for **images**: you can include how many samples, number of channels (color or gray or more?) or modalities, image file format, whether images have the same dimension or not etc. ● E.g. **sequential data**: texts, sound file; please describe appropriate properties such as how many documents or words, how many sound files with typical length (are they the same or variable), etc.	(0 pts) Does not include any description of the data or the data size	(2 pts) Partially describes the data but ***does not*** refer to the data size **or *does not*** describe the data size appropriately for the type of data.	(4 pts) Describes the data including the data size appropriately for the type of data.			

<u>Data Cleaning</u> 1. Does it include clear explanations on how and why cleaning is performed? a. E.g. the author decided to drop a feature because it had too many NaN values and the data cannot be imputed. b. E.g. the author decided to impute certain values in a feature because the number of missing values were small and he/she was able to find similar samples OR, he/she used an average value or interpolated value, etc. c. E.g. the author removed some features because there are too many of them and they are not relevant to the problem, or he/she knows only a few certain features are important based on their domain knowledge judgment. d. E.g. the author removed a certain sample (row) or a value because it is an outlier.	(0 pts) **No attempt or missing both of the following:** Includes clear explanations on how **and** why cleaning is performed.	(3 pts) **Includes one of the following:** Includes clear explanations on how **or** why cleaning is performed.	(5 pts) **Includes both of the following:** Includes clear explanations on how **and** why cleaning is performed.			
<u>Exploratory Data Analysis</u> Does it include clear explanations on how and why an analysis (EDA) is performed? 1. Does it have proper visualizations? (E.g. histogram, correlation matrix, etc. Project should include a mix of simple plots like histograms and box plots with at least one more complicated plot like a correlation matrix.) 2. Does it have adequate analysis? (E.g. clear explanations of how and why specific steps were performed, analysis of each visualization used, feature importance (if possible), etc.) 3. Does it have conclusions or discussions? (E.g. summary of data cleaning and findings, discussing	(0 pts) EDA section not included or **does not address any questions in the rubric**.	(4 pts) EDA section has **proper visualizations** (including a mix of simple plots like histograms and box plots with at least one more complicated plot like a correlation matrix). **Missing or inadequate:** analysis and conclusions/discussions.	(7 pts) EDA section has **proper visualizations** (including a mix of simple plots like histograms and box plots with at least one more complicated plot like a correlation matrix). **and** **adequate analysis (**including clear explanations of how and why specific steps were performed,	(10 pts) EDA meets expectations. **Includes all of the following: proper visualizations** (including a mix of simple plots like histograms and box plots with at least one more complicated plot like a correlation matrix). **and** **adequate analysis (**including clear explanations of	

foreseen difficulties and/or analysis strategy.)			analysis of each visualization used, feature importance (if possible), etc.) **Missing or inadequate:** conclusions/discussions.	how and why specific steps were performed, analysis of each visualization used, feature importance (if possible), etc.) **and conclusions/discussions** (E.g. summary of data cleaning and findings, discussing foreseen difficulties, and/or analysis strategy)		
<u>Model Approach</u> This section is for planning purposes. It's good enough if it contains one or more of the following: 1. Does it mention which models the author will try and why? 2. Does it describe how certain models might have some difficulties (and some ways to overcome them)? It is excellent to think about the following: • If the data has highly correlated features, how does it affect the proposed models? • If the data has too many features, how does it affect the model? Can some methods be used to overcome this? • If the data's class (in a classification task) is very imbalanced, what happens to the proposed model? • Based on the findings from EDA, is there anything to be cautious about selecting types of models or features?	(0 pts) Model approach section not included or **does not address either of the questions in the rubric.**	(5 pts) **Answers either of the following:** mentions which models the author will try and why **or** describes how certain models might have some difficulties and ways to overcome them.				