Supervised Learning Project Rubric

Prompt 1 — Submit Deliverable One: Jupyter Notebook or PDF Report

The Jupyter notebook should show a brief problem description, EDA procedure, analysis (model building and training), results, and discussion/conclusion. 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 (as deliverable 3) and submit a report-style notebook or pdf instead.

If your project doesn't fit into jupyter notebook format (E.g. you built an app that uses ML), write your approach as a report and submit it in a pdf form.

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. type of learning and type of the task.	(0 pts) Not included in the project	(1 pts) Provides **one** of the following**:** explanation of what the project is about **or** the type of learning/algorithms **or** the type of task	(2 pts) Provides **two** of the following: explanation of what the project is about **or** states the type of learning/algorithms **or** states the type of task	(3 pts) Gives a clear explanation of what the project is about **and** clearly states both the type of learning/algorithms **and** type of task.		
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			
Data Is the data source properly cited and described? (including links, brief explanations)	(0 pts) Does not include a brief explanation of where the data is from/how it was gathered **or** does not include a citation (using the format of a style manual like APA) for a public dataset	(1 pts) Includes a brief explanation of where the data is from/how it was gathered **and** if the data is from a public source, cites the dataset using the format of a style manual like APA.				

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.			
Data Cleaning To receive full points for this section, the learner must address the three questions below: 1. Does it include clear explanations on how and why a 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	(0 pts) Uses a dataset that hasn't been cleaned without attempting any cleaning.	(5 pts) **One of the following situations**: Uses a clean dataset and indicates that the dataset didn't require further cleaning **or** attempted a data cleaning but was missing at least one of the following: clear explanations of how and why cleaning steps were performed or conclusions/discussions (E.g., the	(10 pts) Includes **all three of the following**: clear explanations of how and why cleaning steps were performed **and** conclusions or discussions (E.g. the data cleaning summary, findings, discussing foreseen difficulties and/or analysis strategy.) **and** proper visualizations. E.g., for tabulated data, meeting the			

certain features are important based on their domain knowledge judgement. d. E.g. the author removed a certain sample (row) or a value because it is an outlier. 2. Does it have conclusions or discussions? a. E.g. the data cleaning summary, findings, discussing foreseen difficulties and/or analysis strategy. 3. Does it have proper visualizations? a. For example, for tabulated data, meeting the benchmark for moderate data cleaning could include: data type munging, drop NA, impute missing values, check for imbalance, look for any data-specific potential problems, and address issues found. b. If the data is not in tabulated form (E.g., image, sound, text, etc.), focus on answering the three questions above. Note: if you are using a dataset that is already clean, you can receive five points if you note that the dataset is already clean and do not do anything further for this section.		data cleaning summary, findings, discussing foreseen difficulties and/or analysis strategy.) or proper visualizations.	benchmark for data cleaning could include: data type munging, drop NA, impute missing values, check for imbalance, utilize visualizations to look for any data-specific potential problems, and address issues found. If the data is not in tabulated form (e.g. image, sound, text, etc.), focus on including all three of the components above.			
Exploratory Data Analysis Does it include clear explanations on how and why an analysis (EDA) is performed? 1. *Does* it have proper visualizations? 2. Does it have proper analysis? E.g., histogram, correlation matrix, feature importance (if possible) etc. 3. Does it have conclusions or discussions? E.g., the EDA summary, findings, discussing foreseen difficulties and/or analysis strategy.	(0 pts) EDA section not included.	(5 pts) EDA does not have proper visualization, analysis, or conclusions and discussions.	(10 pts) EDA does not address all of the questions in the rubric. E.g. **simple plots** like histograms and box plots without analysis or conclusions.	(15 pts) EDA meets expectations. E.g. in addition to **simple plots** with an explanation of how and why, the author included **at least one of the following**: a correlation matrix with analysis, **or** extra EDA (e.g. statistical tests), **or** good analysis and conclusions/discussions.	(20 pts) EDA above and beyond expectations. E.g. in addition to simple plots, the author included **at least two** of the following (or similar): ● good analysis and conclusions / discussions ● correlation matrix with analysis ● extra EDA (E.g. statistical tests	

Models Some questions to consider: ● Is the choice of model(s) appropriate for the problem? ● Is the author aware of whether interaction/collinearity between features can be a problem for the choice of the model? Does the author properly treat if there is interaction or collinearity (e.g., linear regression)? Or does the author confirm that there is no such effect with the choice of the model? ● Did the author use multiple (appropriate) models? ● Did the author investigate which features are important by looking at feature rankings or importance from the model? (Not by judgment- which we already covered in the EDA category) ● Did the author use techniques to reduce overfitting or data imbalance? ● Did the author use new techniques/models we didn't cover in the class?	(0 pts) No models attempted	(5 pts) Model section does not choose an appropriate single model	(10 pts) Model section needs improvement and does not address most of the rubric. E.g. **One proper single model** without any other project component	(15 pts) Model section does not meet expectations. E.g. **proper single model** and **at least one** of the following: ● addresses multilinear regression/collinearity ● feature engineering ● multiple ML models ● hyperparameter tuning ● regularization or other training techniques such as cross validation, oversampling/undersampling/SMOTE or similar for managing data imbalance ● uses models not covered in class	(20 pts) Model section meets expectations. E.g. **proper single model** and **at least two** of the following: ● addresses multilinear regression/collinearity ● feature engineering ● multiple ML models ● hyperparameter tuning ● regularization or other training techniques such as cross validation, oversampling/undersampling/SMOTE or similar for managing data imbalance ● uses models not covered in class	(25 pts) Model section above and beyond expectations. E.g. **proper single model** and **at least three** of the following: ● addresses multilinear regression/collinearity ● feature engineering ● multiple ML models ● hyperparameter tuning ● regularization or other training techniques such as cross validation, oversampling/undersampling/SMOTE or similar for managing data imbalance ● uses models not covered in class
Results and Analysis Some questions to consider: ● Does it have a summary of results and analysis? ● Does it have a proper visualization? (E.g., tables, graphs/plots, heat maps, statistics summary with interpretation, etc.) ● Does it use different kinds of evaluation metrics properly? (E.g., if your data is imbalanced, there are other metrics (F1, ROC, or AUC) that are better than mere accuracy). Also, does it explain why they	(0 pts) No results or analysis attempted	(5 pts) Results and analysis section does not meet expectations. Attempt does not have basic results and analysis	(10 pts) Results and analysis section needs improvement. E.g. **includes** a summary with basic results and analysis	(15 pts) Results and analysis section does not meet expectations. E.g. **includes** a summary with basic results and analysis and **one of the following:** good amount of visualizations **or** tries different	(20 pts) Results and analysis section meets expectations. E.g. **includes** a summary with basic results and analysis and **two of the following:** good amount of visualizations **or** tries different	(25 pts) Results and analysis section goes above expectations. E.g. **includes** a summary with basic results and analysis and **three of the following:** good amount of visualizations **or** tries different

chose the metric? ● Does it iterate the training and evaluation process and improve the performance? Does it address selecting features through the iteration process? ● Did the author compare the results from the multiple models and make appropriate comparisons?				evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance	evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance	evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance
Discussion and Conclusion	(0 pts) No discussion or conclusion attempted	(5 pts) Discussion and conclusion section needs improvement. E.g. **includes one of the following:** discussion of learning and takeaways **or** discussion of why something didn't work **or** suggests ways to improve.	(10 pts) Discussion and conclusion section meets expectations. E.g. **includes two of the following:** discussion of learning and takeaways **or** discussion of why something didn't work **or** suggests ways to improve.	(15 pts) Discussion and conclusion section goes above expectations. E.g. **includes three of the following:** discussion of learning and takeaways **or** discussion of why something didn't work **or** suggests ways to improve.		
Write-up Is the write-up organized and clear?	(0 pts) No the write-up is not organized and clear	(5 pts) Yes the write-up is organized and clear				

Prompt 2 — Submit Deliverable Two: Video Presentation

Record a video of a presentation or demo of your work. The presentation should be a condensed version as if you're doing a short pitch to advertise your work; so please focus on the highlights:

1. What problem do you solve?
2. What ML approach do you use, or what methods does your app use?
3. Show the result or run an app demo.

The minimum video length is 5 min, the maximum length is 15 min. The recommended length is about 10 min. Submit the video in the .mp4 format.

Prompt	Points			
Does the video explain the following?: 1. What problem do you solve? 2. What ML approach do you use, or what methods does your app use? 3. Show the results or run an app demo.	(0 pts) Video presentation not included	(3 pts) Presentation needs improvement. E.g., **includes one of the following:** problem the project solves **or** the ML approach and methods used **or** shows the results or runs an app demo	(7 pts) Average presentation. E.g., **includes two of the following:** problem the project solves **or** the ML approach and methods used **or** shows the results or runs an app demo	(10 pts) Excellent presentation. E.g., **includes all of the following:** problem the project solves **and** the ML approach and methods used **and** shows the results or runs an app demo
Is the video clear and organized? Consider the following: • The presentation follows a logical sequence. • The structure gives appropriate time to each section, so the video is about 10 minutes in length (between 5 and 15 minutes). • The presentation is a well-rehearsed, condensed version that focuses on the highlights.	(0 pts) Video presentation not included	(1 pts) Video presentation is not clear or organized, and does not seem to have been rehearsed. Presentation doesn't follow a logical sequence or give appropriate time to each section. Video does not meet time length requirements.	(3 pts) Average quality video presentation. E.g., Presentation **includes two of the following:** follows a logical sequence **or** gives appropriate time to each section, focusing on the highlights **or** is well-rehearsed.	(5 pts) Presentation has very good clarity and organization. E.g., presentation **has all of the following:** follows a logical sequence **and** gives appropriate time to each section, focusing on the highlights **and** is well-rehearsed

Prompt 3 — Submit Deliverable Three: GitHub Repository Link

Create a public project GitHub repository with your work (please include the git repo URL in your notebook/report and slides). It is essential that it is public so your peers will be able to access it. This repository needs to be specifically for this project.

Data by-product: If your project creates data and you want to share, please do not upload the data in git. An excellent way to share would be through a Kaggle dataset or similar. Similarly, please do not upload videos to git- if you want, you can upload to youtube and post those link(s) to your git.

Prompt	Points	
Does the project have a public GitHub repository with code specifically for this project?	Yes (5 pts)	No (0 pts)
Does the code include comments to help you understand the code? E.g., the comments indicate **why** the code is there **or** the **what/how** for tricky code.	Yes (5 pts)	No (0 pts)
Is the code organized? E.g., The file repository structure makes sense and the code is generally easy to read and follow. git	Yes (5 pts)	No (0 pts)