Part 3 Final Project Rubric

Deliverable: 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. While reviewers will assess the Data Cleaning section onward, it is important to include all sections for the Project Topic and Data so that reviewers and graders can understand the project. 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					
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 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.	(5 pts) **Includes one of the following:** Includes clear explanations on how **or** why cleaning is performed.	(10 pts) **Includes both of the following:** Includes clear explanations on how **and** why cleaning is performed. E.g., for tabulated data, meeting the 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.			
Exploratory Data Analysis Does it include clear explanations on how and why an analysis (EDA) is performed?	(0 pts) EDA section not included or does not address any	(5 pts) EDA section has **proper visualizations**	(10 pts) EDA section has **proper visualizations**	(15 pts) EDA meets expectations. **Includes all of the**	●	

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	questions in the rubric.	(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.	(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, analysis of each visualization used, feature importance (if possible), etc.) **Missing or inadequate:** conclusions/discussions.	**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 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>Models</u> 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)	(0 pts) No models attempted	(5 pts) Model section does not choose any 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 of the following:** • addresses multilinear regression/collinearity	(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	(20 pts) Model section meets expectations. E.g. **proper single model** and **at least two** of the following: • addresses multilinear regression/collinearity • feature	

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?			● 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	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	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 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?	(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** **Does not include any of the following:** good amount of visualizations **or** tries different evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance	(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 evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance	(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 evaluation metrics **or** iterates training/evaluating and improving performance **or** shows/discusses model performance	
Discussion and Conclusion	(0 pts)	(5 pts)	(10 pts)			

	No discussion or conclusion attempted	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	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			
Write-up Is the write-up organized and clear?	(0 pts) No the write-up is not organized and clear	(2 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) 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 condensed version that focuses on the highlights.	(0 pts) Video presentation not included	(2 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		

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 (1 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 (1 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 (1 pts)	No (0 pts)