Final Project Rubrics & Tips

CSc 44700 S Spring 2025 Introduction to Machine Learning

Part 1: The Rubrics

The deliverables consist of a jupyter notebook, presentation and a report:

- Jupyter notebook: 100 points
- Presentation: 30 points
- Report: 30 points

The presentations are scheduled on 5/13/2025 (Tu) and 5/15/2025 (Th).

The notebook and report shall be submitted by 11:59 pm 5/20/2025 (Tu).

1.1 Notebook Rubrics

The length and depth of content in the Jupyter Notebooks will be considered as an indicator of the work you have put in.

I Machine Learning Question: 20 pts

A. Is the background context for the question stated clearly (with references)?
B. Is the hypothesis/problem stated clearly ("The What")
C. Is it clear why the problems are important? Is it clear why anyone would care? ("The Why")
D. Is it clear why the data chosen should be able to answer the question being asked?
E. How new, non-obvious, and significant are your problems? Do you go beyond checking the easy and obvious?

II Data Cleaning/Checking/Data Exploration: 20pts

A. Did you perform a thorough EDA (points below included)?
B. Did you check for outliers?
C. Did you check the units of all data points to make sure they are in the right range?
D. Did you identify the missing data code?
E. Did you reformat the data properly with each instance/observation in a row and each variable in a column?
F. Did you keep track of all parameters and units?

G. Do you have a specific code for reformating the data that does not require information not documented (eg. magic numbers)?
H. Did you plot univariate and multivariate summaries of the data including histograms, density plots, and boxplots?
I. Did you consider correlations between variables (scatterplots)?
J. Did you consider plot the data on the right scale? For example, on a log scale?
K. Did you make sure that your target variables were not contaminating your input variables?
L. If you had to make synthetic data was it a useful representation of the problem you were trying to solve?

III. Transformation, Feature Selection, and Modeling: 30pts

A. Did you transform, normalize, filter the data appropriately to solve your problem? Did you divide by max-min, or the sum, root-square-sum, or did you z-score the data? Did you justify what you did?
B. Did you justify normalization or lack of checking which works better as part of your hyper-parameters?
C. Did you explore univariate and multivariate feature selection? (if not why)
D. Did you try dimension reduction, and which methods did you try? (if not why)
E. Did you include 1-2 simple models, for example with classification LDA, Logistic Regression or KNN?
F. Did you pick an appropriate set of models to solve the problem? Did you justify why these models and not others?
G. Did you try at least 4 models including one Neural Network Model using Tensor-Flow or Pytorch?
H. Did you exercise the data science models/problems we described in the lectures showing what was presented?
I. Are you using appropriate hyper-parameters? For example, if you are using a KNN regression are you investigating the choice of K and whether you use uniform or distance weighting? If you are using K-means do you explain why K? If you are using PCA do you explore how many dimensions such as by looking at the eigenvalues?

IV. Metrics, Validation and Evaluation 20pts

A. Are you using an appropriate choice of metrics? Are they well justified? If you are doing classification, do you show a ROC curve? If you are doing regression, are you justifying the metric least squares vs. mean absolute error? Do you show both?
B. Do you validate your choices of hyperparameters? For example, if you use KNN or K-means do you use cross-validation to optimize your choice of parameters?
C. Did you make sure your training and validation process never used the training data?

D. Do you estimate the uncertainty in your estimates using cross-validation?
E. Can you say how much you are overfitting?

V. Visualization 10pts

A. Do you provide visualization summaries for all your data and features?
B. Do you use the correct visualization type, eg. bar graphs for categorical data, scatter plots for numerical data, etc?
C. Are your axes properly labeled?
D. Do you use color properly?
E. Do you use opacity and dot size so that scatterplots with lots of data points are not just a mass of interpretable dots?
F. Do you write captions explaining what a reader should conclude from each figure (not just saying what it is but what it tells you)?

VI. Code 20pts

A. Is the code provided can reproduce the entire work?
B. Is the data included or at least linked (externally) with instructions on how to download it?
C. Do you factor repeated operations into functions to avoid repetitively and error-prone copy-paste?
E. Do you use docstrings and numpy documentation style:
https://github.com/numpy/numpy/blob/master/doc/HOWTO_DOCUMENT.rst.txt
to make your code clear and readable?
F. Do you use markdown cells to explain every step of your code similar to Homeworks and some example notebooks?
G. Does the code demonstrate considerable work given the number of people on the project?

1.2 Presentation Rubrics

A. Do you tell a coherent story with a beginning, middle, and end?
B. Do you introduce why the problem is important?
C. Do you explain in the first couple of slides what you accomplished on solving the problem?
D. Are you careful not to have slides filled with text (keep in notes)?
E. Is data and evaluations presented as clear figures (mostly)?
F. Do you make sure to say what is "interesting" or should be learned from each figure?
G. Do you stay within your time limits 15 min?
H. Do you avoid useless padding slides of no relevance?

1.3 Report Rubrics

You may submit one report per group. The final report should be a polished and coherent narrative, not a casual or disorganized collection of calculations. Each section should be clearly labeled, and the following guidelines should be followed:

Remember, the final report should be a meticulously cleaned up and refined version of all the calculations and work done throughout the project. It should represent a collaborative effort, combining the contributions of each team member into a cohesive and self-contained narrative that thoroughly documents and explains your approach, findings, and conclusions.

- Explanations: Every calculation performed should be accompanied by an explanation of its purpose and rationale.
- Output and Visualizations: Whenever something is calculated, include the corresponding output, such as evaluations, checks, or, preferably, figures. Any figures or visualizations should be properly explained and contextualized.
- Cohesive Narrative: The report should stand on its own as a comprehensive document that explains the entire project. It should not resemble a scratchpad of calculations left for an archaeologist to decipher.
- Synthesis: The final report should synthesize the work of the entire team into a single, well-organized document, rather than presenting three separate reports.

Part 2: The Tips

2.1 Notebook

Ensure you meet the following requirements for grading:

- Apply Machine Learning Techniques: Use a dataset and apply one or more of the ML frameworks covered in class, such as multi-variable classification, non-linear regression, clustering, or dimension reduction.
- Project tackles a Real Problem: Attempt to solve a real problem. Merely showing relationships between variables is insufficient.
- Maintain Statistical Hygiene: Follow best practices for statistical analysis:
    - Your test data should only be used once for final evaluation, not iteratively during model selection or hyperparameter tuning.

  - Use a separate validation set (not the test set) to select the best model and tune hyperparameters.
  - Consider using cross-validation on the training set to create true train and validation subsets.
  - Ensure your input data never has access to the target variable during preprocessing, normalization, or cleaning.

Baselines

Comparing your model's performance against simple baselines is crucial and **REQUIRED**. If a coin flip, a constant classifier, or a straightforward threshold on a single variable outperforms your sophisticated SVM or Random Forest model, the fancy approach becomes worthless.

Balanced Classes

Highly imbalanced classes can become the most critical factor to address. Any further analysis without accounting for class imbalance is meaningless.

2.2 Presentation

The project presentation will be limited to 15 minutes. I will record the presentation in class using a video recorder, but this recording will not be shared publicly without your permission (it is solely for internal assessment purposes). During the presentation, you should:

- **Provide an Overview:** Briefly introduce the problem you are addressing and describe the dataset you are working with.
- **Outline Your Approach:** Explain the steps you followed in the machine learning pipeline, including data wrangling, cleaning, and any other relevant processes.
- **Present Your Conclusions:** Summarize the key takeaways and conclusions that someone reviewing your code, presentation, and report should draw from your analysis.
- Please keep the following in mind:
  - Do not discuss what you **would have done** if you had more time. Focus solely on the work you have accomplished.
  - Refrain from spending time explaining general machine learning concepts or techniques covered in the class.

The goal of the presentation is to concisely communicate the problem you tackled, the data you used, the methodological approach you followed, and the substantive insights and conclusions derived from your analysis. Remember, the presentation should be a focused and succinct summary of your project, allowing the audience to understand the essence of your work without delving into hypothetical scenarios or redundant explanations.

Note: Do not waste any time explaining what you personally learned from the process or the class during the presentation or report!

2.3 Report Tips

The report should roughly follow an academic paper format:

Title

Should express the problem in the least boring but factually accurate way. It might be interesting to ask a question.

Abstract

1-2 paragraphs of 200–250 words. Should concisely state the problem, why it is important, and give some indication of what you accomplished (2-3 discoveries)

Introduction

State your data and research question(s). Indicate **why** it is important. Describe your research plan so that readers can easily follow your thought process and the flow of the report. Please also include key **results** at the beginning so that readers know to look for. Here you can very briefly mention any important data cleaning or preparation. Do not talk about virtual results i.e. things you tried or wanted to do but didn't do. Virtual results are worse than worthless. They highlight failure.

Background

Discuss other relevant work on solving this problem. Most of your references are here. Cite all sources. There is no specific formatting requirement for citations but be consistent.

Data

Where you go the data. Describe the variables. You can begin discussing the data wrangling, and data cleaning. Some EDA may happen here. This includes your data source (including URL if applicable), any articles behind the data source.

Methods

How did you take your data and set up the problem? Describe things like normalization, feature selection, the models you chose. In this section, you may have EDA and graphs showing the exploration of hyper-parameters. Note: Use graphs to illustrate interesting relationships that are important to your final analyses. **DO NOT** just show a bunch of graphs because you can. You should label and discuss every graph you include. There is no required number to include. The graphs should help us understand your analysis process and illuminate key features of the data.

Evaluation

Here will to show your different models’ performance. It is particularly useful to show multiple metrics and things like ROC curves (for binary classifiers). Make sure it is clearly not just what the score is but for which instances in the data one has the largest errors (in a regression), or just sample examples miss-classified. Make an attempt to interpret the parameters of the model to understand what was useful about the input data. Method comparison and sensitivity analyses are absolutely CRUCIAL to good scientific work. To that end, you MUST compare at least 2 different methods from class in answering your scientific questions. It is important to report what you tried but do so SUCCINCTLY.

Conclusion

Summarize how well your solution works Characterize how robust you think the results are (did you have enough data?) Try for interpretation of what the model found (what variables were useful, what was not)? Try to avoid describing what you would do if you had more time. If you have to make a statement about “future work” limit it to one short statement.

Attribution

The grading for this project will explicitly consider each team member's individual contribution. To acknowledge significant effort, the highest contributor(s) may receive 1 to 2 bonus points. Conversely, if a team member's contribution is demonstrably and significantly lower than that of their peers, a deduction of 1 to 2 points may be applied.

A detailed breakdown of individual workload example is provided in the table below. We have strived for an honest and accurate representation of each member's responsibilities and effort.

Task breakdown	***Memeber1***	***Member2***	***Member3***
Literature review	40%	40%	20%
Data preprocessing	30%	20%	50%
Modeling & Eval	30%	40%	30%
Visualization	50%	30%	20%
Slides Prep & Present	20%	40%	40%
Report writing	40%	40%	20%
Sub-total	35%	35%	30%

Bibliography

References should appear at the end of the report/notebook. Again, no specific format is required but be consistent.

Appendix

If there are minor results and graphs that you think should be included, put them at the end. Do not include anything without an explanation. No random graphs just for padding!! However, let’s say you did a 50 state analysis of poverty and demographics, and your report focused on the 5 most interesting states, for completeness you could include all in an appendix. Be sure though to provide some (very short) discussion with each figure/code/result.