Jupyter Report Rubric

Some guidelines for formatting an ML report as a Jupyter notebook. This is a draft of a living document so expect updates.

1. Title: something like "Predicting Voter Registration Data from Demographic Data." Tell us quickly what you're doing. It's OK to be clever if you can pull it off "To Eat or Not to Eat: Identifying Poisonous Mushrooms" (this is borderline.) "Report #2" (this is bad).
2. Intro: Say what you're solving, where the data comes from, and what techniques you're going to use. It's good to put the final result here too. "Ultimately our Random Forest model achieves a test accuracy of 95% and an MCC score of 0.92."
3. Outlining: Use sections and subsections as appropriate. The title is "#" in markdown. Sections are "##" and so on.
4. Sections. You should usually have (there can be variations, of course)
    1. Intro
    2. The Data (load the data and give us info – size, shape, features, target. Describe all the columns if possible or maybe point to a website that does if there's a lot of data. Deal with basic problems here – is it balanced? is there missing data?)
    3. Preprocessing (usually this is where you OneHotEncode, maybe drop obvious distractors, impute missing data, rebalance, pick a subset of the rows)
    4. Data Analysis – more detailed search for co-linearity, predictive features, relationship to the target field, scaling, normality and skew, histogram plots, heatmaps, etc.
    5. Feature Extraction – we haven't done this much yet, outside of TF-IDF, but this can be a very big part of your analysis. Especially with datasets like images, audio, video, time series, EEG, ECG, biology and genetics, etc.
    6. Modeling – Now you do some ML. Depending on the purpose of the report you will give your best model here, or several attempts. Give a detailed analysis; don't just print "accuracy = 0.92". Use relevant metrics: confusion matrix, ROC curves, precision, recall, F1, MCC, RMS, etc. You should use cross-validation and discuss the variance of the results.
    7. Refinement – Pick your best model (or a few) and make them better. There are many techniques here. GridSearch is an obvious one but massaging the data, combining models, better preprocessing to find salient features, dropping fields, etc. Lots of room here to be creative. This might be a place for ensemble methods.
    8. Conclusion – What's your best result. Make sure to address the pros and cons of close performing models (variance, precision, recall, in larger problems there are resource requirements, deployment.) It is also a good time to reflect on your data more generallly – how well

does the dataset represent the problem here. What would you like to look at next if you had more time, more resources, more data. What are the shortcomings in your own analysis. Do you trust your own results? How much does this generalize? It's good to give the reader something to think about beyond your paper, both to make your paper more interesting, and also to prove that you are capable of interesting thoughts.

5. General voice: The entire document should be narrative. You should rarely have a computation cell without some text before, setting up what you're doing "let's remove the columns for country and state because they are irrelevant" and tell us what you learned "it looks like pH and log[H+] are collinear so we will drop one of them" or "this dataset is highly imbalanced". Do *not* expect the reader to read or even look at your Jupyter code. It's just there to prove you're not cheating. It's the text in the document that carries the weight.
6. Format: Every graph has a title, every axis has a label. If your heatmap is too small to read, then fix it somehow (maybe do 2 heatmaps, maybe drop features from the heatmap that won't be helpful). It's fine and expected to *draft* your visuals before you *publish* the final versions. We don't need to see the drafts.
7. Citations: Tell us where to get your data. If you use any non-standard library (what's standard? that's a valid question) just point it out and say how to install it (probably 'pip'). If a brilliant paper or website gave you a brilliant idea, don't pretend like it's your idea. Eventually somebody will realize they have seen it before and you'll be cooked. If you actually publish something, we'll work harder to make citations, but for classwork just make sure you leave sufficient. breadcrumbs.
8. Final version: I think that a .pdf of your notebook might be best. That may take some fiddling. let's experiment but I hope we can get .pdf to work. Then submit your .ipynb and .pdf files (and any relevant datasets – either upload them or if they're huge link to them. Do *not* make the mistake of relying on data in the cloud to stay there. Make your own copy. Put it in your own cloud. Link to that!)