Final Project: Working Towards Success

This document describes both requirements and advice in working towards success in your final project. First, please re-read the original project description, as I do not try to repeat that information here.

Requirements: What to Do Next

This project is designed to be open ended, so that you can proceed in the directions you find most interesting or promising. Some options include:

- Work with the attributes. Review the “Pre-process the Data” discussion in the original project description. Run experiments using different approaches. Study the domain more deeply and see how that affects your use or processing of certain attributes.
- Work on tuning GradientBoostingRegressor. It has many parameters, and the original project description discussed some basic tuning strategies.
- Apply alternative algorithms and tune them as well.

Each of the above could be done in a simplistic way, or in a very deep and complex way. You don’t have to go deep in every category above, but do please go deep somewhere. Carefully log everything you do and keep writing things up as you go! Check out the original project description for more advice on this.

Requirements: The Final Paper

In your final paper, please include the following sections and subsections:

- Abstract
- Introduction
- Data Description
- Experiment
    - Pre-Processing
    - Algorithms and Parameterization
    - Results
    - Analysis
- Conclusion

The Abstract is a quick overview of what the reader should expect from this paper. Keep it short (1 paragraph) and treat it like a section to “hook” the reader, making the reader interested in reading the entire paper.

For Data Description, explain the key points of the dataset that any reader would need to know. In particular, be certain you explain facts about the data that you’ll need to refer to later.

The Experiment section should have subsections as listed above. If all of your work can fall under one broad “experiment”, then put it all in one section here. If one experiment leads to another and another, and it wouldn’t work to lump them into a single “experiment”, then feel free to repeat this section and subsections multiple times, with each “Experiment” section labeled a little differently based on what you’re describing.

For Pre-Processing, don’t feel like you need an absolutely exhaustive list, but explain the most important/interesting steps, and summarize the rest. If pre-processing steps will be tested in the experiment, explain precisely what will be tested and how.

For Algorithms and Parameterization, describe in English what the algorithm does, if it’s something we’ve discussed in class. If we haven’t discussed the algorithm in class, feel free to keep this brief and general. For parameterization, again, an exhaustive list is not necessary, but any intentional decisions and points of interest should be described. If different parameterizations are tested in the experiment, explain precisely what will be tested and how.

For Results, state the results of your experiment. It is important that analysis *not* be done here. The results section is strictly showing and describing the facts of what happened. There should be no interpretation of what those results mean here.

For Analysis, explain what you think the results mean. Why do you think things worked out that way? Why do you think one pre-processing approach / parameterization / algorithm worked better than another? What does this mean about the problem and/or the dataset? What additional questions does this lead to?

For Conclusion, provide a summary of the most important points arising from your experiments: the key conclusions for someone that might only read the Abstract and Conclusion sections.

If you feel that you need to adjust the organization above, that may be fine, but please talk with me about it first.

LaTeX Requirements

If you’re going for the optional extra credit of doing the paper in LaTeX, then you must also include the following:

- A figure
- A table
- Non-trivial mathematical notation. At the very least, each of you should be describing GradientBoostingRegressor as one of your algorithms, so you could use such notation there.

You must also cite everything correctly, including websites, as demonstrated in the sample LaTeX materials.

It would probably improve your grade to make effective use of the above in a Word paper too.

Advice: Interpreting Experimental Results

The default scorer in cross_val_score is whatever is specified by the predictor. For GradientBoosting-Regressor, as for many predictors, the scorer used is R2 (“R squared”), also known as the “coefficient of determination.” It’s a measure of “goodness of fit” of the model to the dataset, with 0 meaning no fit at all, and 1 meaning perfect fit. To put it another way, 0 means “when I know the inputs, they’re not helpful at all in knowing the output”. 1 means, “when I know the inputs, I can compute precisely the correct output every time”.

Note that “accuracy” isn’t clearly defined for regression models. Accuracy is for classification (categorical output variable) – either the model’s hypothesis was correct, or it wasn’t. For regression (quantitative output variable), though, what would it mean for the hypothesis to be “correct”? Exactly right? Within some value? Within some percentage? It’s not clearly defined.

So in particular, note that R2 is not formally a “percentage accuracy”. Rather, R2 is a measure of how much of the total variance in the dependent variable is explained or captured by the model.

There is much you could study about R2. For our purposes for now, we can just see it as a way to compare different models. The magnitude of R2 for a particular model depends not just on the model output, but on the inputs; after all, it is a measure of the relationship between the two! So use R2 to compare models and pre-processing strategies.

You may find as you work that you try something you think will be significant, only to find a pretty small improvement to R2 (or even a reduction!). Even small improvements may nevertheless be important. For example, with millions of homes sold in a year, and with them being the largest purchase most people will ever make, every little change has a cumulative effect. Small differences matter to those trying to set a price for a home to sell, and to those bidding on a home. If you can get a small improvement after a matter of hours of work, that contains real-world value.

Also keep in mind that GradientBoostingRegressor is a powerful algorithm, well-suited to this dataset, so even a basic application of it has some good success.

So don’t worry if you have only small improvements to R2 given your changes. The most important thing is that you are engaging deeply with this project:

- Making a variety of educated hypotheses
- Implementing your ideas correctly
- Organizing, analyzing, and clearly reporting your results
- Using this in a feedback loop to lead to more hypotheses

Advice: The Number of Folds

Recall that cross_val_score has a cv parameter for setting the number of folds. One thing we saw in our series of homework assignments is that a higher number of folds means a larger training set, and a larger training set typically means better model performance. But it’s not the model (or the pre-processing) that has improved; a higher number of folds just means that the test has changed. So don’t get caught into doing lots of tests with different numbers of folds. Pick a number of folds that is computationally reasonable for your work, and stick with it consistently for tests you’d like to compare directly.

You may wish to pick a number of folds that is higher than what we’ve been doing, though. When a model is ready to be deployed to the real world, we’d first fit it with the entire training set. Thus, a higher number of folds creates a test that is a little closer to this eventual reality. So testing with a higher number of folds gives us a somewhat more accurate estimate of real-world performance. But it’s also slower. So strike that balance.