UDACITY

PROJECT

Machine Learning Capstone Project

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

8 SPECIFICATIONS REQUIRE CHANGES

This is a solid solution to an interesting problem, and your work is well presented in many areas. The report is well written with some good details; there are certainly some places however that would benefit from more discussion, as I've outlined. Fundamentally, this is a great start, and I look forward to your next submission.

Definition

Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.

- Solid introduction to the problem you're solving, with a good focus on the information available to solve it
- It's clear how machine learning is a viable solution in this situation, based on your discussion here

The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

- The input and output are well defined, which makes for a solid problem statement
- Your approach to the problem is clear, and it's good that you have multiple goals to attain to, as this will add depth to your analysis

Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

- These metrics are well defined here
- This section should also focus on justifying your choices. What makes these the optimal metrics for this situation?

Analysis

If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.

- The source of the data, its size, the features and their meaning are all clearly described, which makes for a solid overview of the key characteristics
- A data sample is provided as required
- You've done excellent work in digging into the univariate distributions and bivariate relationships in your data. This is valuable exploratory analysis that gives your reader (and I'm sure gave you) important insight into the data and problem

A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

- These are important data qualities and certainly things well worth visualizing; good use of this section here
- The visualizations themselves are clean and well presented, with appropriate labels and identifiers, and the right visual encoding for the data type
- Feature importance in particular is a great secondary focus of any supervised learning work, as it gives us greater understanding of the problem on top of helping us model better
- You've done an excellent job addressing critical characteristics like outliers and missing data. Again, this is valuable EDA work that cannot be overlooked

Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

- Your explanations of how each of these algorithms work is quite thin, and would benefit from more specific details
- It's okay to stay away from mathematically rigorous details, but for example, one way you could expand on your discussion here is by breaking down some of the terms used. You've called random forests an "ensemble" without really defining the term ensemble; to someone unfamiliar this may not mean much, but even to someone more familiar with the term, they may not know whether it's a bagging technique or a boosting technique. To ensure that your writing reaches as wide of an audience as possible, it's best if you can provide details on how the learned decision trees are combined
- Another example is the term "weak learner". What is a weak learner?
- In addition, I would recommend diversifying your algorithms more. These are all decision tree ensembles, and because of that they'll likely perform comparably. There's not much need for comparing GradientBoostingRegressor to XGBoost, since they're the same algorithm. Why not consider a neural net, or logistic regression, naive bayes or an SVM? There are many other choices that aren't tree based. By considering many *different* algorithms, we ensure that we're finding the best approach for our dataset and problem

Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.

- Comparing to well defined, objective, concrete models / results is always the best approach for obtaining a baseline for our own work, and that's what you have here. Good choice

Methodology

All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

- This is a solid step by step overview of the work required to prepare your data for proper training, and again it's written in a way that's clear and detailed
- It's easy to see what the structure of the data would be before and after each of these transformations
- This is the benefit of good exploratory work; you've already identified all that needs to be done, and now you have a clean dataset

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

- The goal of this section is to make our work as reproducible as possible; for any future researchers that read your work and wish to expand on it, they'll have to start by re-implementing what you have done, and they can only do that if your explanation of your work through this report is detailed and accurate. So beyond listing the libraries and filenames, your focus here should be on describing your process in this way
- One of the main ways we help with reproduciblity is by clearly documenting the challenges we faced, and how we overcame them. This helps out those that are following our work not to get stuck on the things that we got stuck on. So including this is also a good idea

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

- You've stated that you *will* do hyperparameter tuning, but at this point, it should be done, and you should now be describing it
- The hyperparameters tuned, the values tried, and the results obtained should all be clearly and cleanly recorded, fully characterizing the refinement process

Results

The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

- Your final results are presented in a way that's easy to analyze and compare
- What's missing is a discussion and interpretation of these results. Primarily, the focus here is on robustness. Can we trust your model based on these results? Why or why not? Try to be as objective as possible in your analysis here; if you can perform a well defined statistical test to ensure robustness, that would be optimal

The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

- Your linear regression benchmark doesn't appear to have been included in your discussion of results here. How did it compare? Did you beat your benchmark?

Conclusion

A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.

- Visualization is used to represent your results, but that's the results section; this section (Free Form Visualization) should be separate, and contain some discussion of a quality about your data, model, or problem that you found interesting
- Feature importance was already addressed, so that's out, but what about further examining other residual plots? One of the nice things about regression is that we have quite a few diagnostic plots to play with, and digging further into this would certainly help your discussion of robustness along. This is, of course, just one potential idea

Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

- More of a recap of the overall process is needed; what were the steps you took to produce this final model?
- It's great to see how much you learned from this work and how much you benefitted from it. Ultimately, the capstone is a learning experience, so your personal takeaways are the most important lasting effect

Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.

- You've got good ideas here touching on improving both the data and the algorithms, which is great
- It's true that neural networks require lots of data when approximating complex functions. Consider for example, facial recognition. When every pixel is a feature, there's a *lot* of features, and a lot of interactions between features, for a model to learn about. This is the primary reason why lots of data is needed: it's a high-dimensional problem. However, when you're applying NNs to simpler problems, this becomes less of a restriction. So I would definitely recommend trying this. It also helps expand your skillset, as deep learning is an incredibly valuable tool to have right now. In a lot of ways, it's the face of machine learning at this point

Quality

Project report follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used to complete the project are cited and referenced.

Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH