UDACITY

Logout

PROJECT

Machine Learning Capstone Project

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW | CODE REVIEW | NOTES

Requires Changes

SHARE YOUR ACCOMPLISHMENT

3 SPECIFICATIONS REQUIRE CHANGES

This is a truly exceptional project here, as I am very impressed with your understanding and techniques. You just have a few sections to perfect here and you will be good to go. Look forward in seeing 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.**

I am glad that you have built a solid grasp on supervised machine learning here. I would recommend going into more detail in the specific problem you are planning on solving here, the Allstate Claims Severity Challenge. Any background information, problem domain, the project origin, why this is important? etc...

✓ **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.**

> "we can predict the claim cost by using Supervised Learning algorithms, specifically Regression algorithms to predict the continue value of a Claim Loss."

Problem statement is clearly defined here. And nice work with your simple strategy for solving the problem, would be easy to replicate.

I would recommend mentioning what type of machine learning "models" you plan on using here, to give the reader some ideas in what is to come in your report and how you plan on solving this important task.

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

> "I preferred MSE over R2_score because it is relatively easy to understand since reported values are in the same units as the target variable, so we could see how far the predicted values were from the actual ones."

This comment isn't quite right here, as the reported values are actually NOT in the same units as the target variable. Remember what MSE stands for(mean **squared** error). Therefore they are not in the same units. As the metric of Root Mean Squared Error(RMSE) is in the same units, since we take the square root at the end to negate this squared term.

Note(Optional): With your comment

> "While MSE is greater it is a worst result"

Would recommend switching this around, and mention that we are trying to minimize MSE. A bit more intuitive.

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.**

Very nice job describing your dataset and very nice with the plots and descriptive stats. As the reader can clearly get an understanding of the structure of the data you are working with.

Rate this review

☆☆☆☆☆

> "The target value by the other hand was highly skewed and had pretty obvious outliers."

Might be a good idea to look into applying a log or square root transformation to the target variable to see if you get any more desirable results. As this is quite a large skew(which I see that you do later, nice job. Could also mention this here).

✓ **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.**

Nice visuals here, as pandas is great! These are definitely needed and you have provided thorough discussion. Does this give any ideas of maybe trying out something like PCA on the features to reduce the dimensionality?

Could also check out the library seaborn

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

Nice job addressing the previous reviewers concerns. I would recommend also giving some ideas in terms of *why* each of these would be good algorithms for this particular problem and dataset. Do these deal well with high dimensional datasets? etc...

Note: You really should describe PCA here as well, since this another big algorithm that you use.

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

I would say that a SVM model might be a bit too advanced for a simple benchmark model. You can definitely still use this, but I would also recommend using like a very simple decision tree, linear regression or Naive Bayes to get a very basic baseline to get an idea of how a simple model could perform.

Since this is a Kaggle competition, could also use the leaderboard as a benchmark. Maybe strive for the top 5% or top 200, etc...

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.**

Good work documenting your pre-process steps here. Nice idea to use PCA and transform the target variable(both are definitely needed)! Couple other ideas

- Another idea to check out would be the feature_selection module in sklearn. As SelectKBest and SelectPercentile could also give you important features.
- Could also look into running a tree based model and get the feature_importances (http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

⟳ **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.**

As the previous reviewer had also mentioned, just make sure you also give some ideas if any complications occurred during the coding process. You might have briefly touched on this, but please expand a bit more. Anything go wrong here? Was there any part of this process that was more difficult than the others? If not, mention this.

Notes:

- What are these training/prediction times in? Seconds?
- Why did you choose 20% of the data? Any reasoning here?

✓ **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.**

Great job documenting your refinement here. Could look into creating a table to clearly outline the changes and the results.

> "And the result was 4082608.64216, so it seems that my hypothesis was wrong about transforming the target data. It had a better result when we use the raw target and without reduced dimensionality."

Glad that you played around with the results for using the raw target with reduced dimensionality and results using the transformed target without reduced dimensionality.

Could also check out combining these all in one step with the notion of Pipelining, check out this blog post brought to you by Katie from lectures, as this would be a good example.

- (https://civisanalytics.com/blog/data-science/2016/01/06/workflows-python-using-pipeline-gridsearchcv-for-compact-code/)

Rate this review

☆☆☆☆☆

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.**

> "The model is robust as when I made small perturbations to the data the model didn't get greatly affected by it"

This is a great idea, but can you show a bit of concrete justification for this? What was the MSE you obtained after you made small perturbations?

Couple other ideas

- Could look into using KFold CV
- Or since this a Kaggle competition, what was the result on the leaderboard?

✓ **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.**

Conclusion

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

Nice visual and discussion here, as it seems that your model can't pick up on the larger values.

You could also check out using a residual plot. A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

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

Solid end-to-end summary of your report here. This is one of the best I have seen!

✓ **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.**

> "I think that maybe using an ensemble model to average the responses from the different models could get better results."

Great idea! As these often win Kaggle competitions. However this approach is not really applicable for the real world.

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.**

Your writing is very clean and it is very easy to understand what you are saying. I personally thank you as this report is very easy to read :)

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

Code looks great and ipython notebooks are very cool. Nice job with the comments.

I would recommend setting `random_state` in your models for reproducible results.

☑ RESUBMIT PROJECT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

Rate this review
☆☆☆☆☆

RETURN TO PATH

Rate this review

☆☆☆☆☆