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

6 SPECIFICATIONS REQUIRE CHANGES

Congratulations for a very strong first submission! In general, I think your report could benefit from a little less visualization and code and a little more justification and analysis. Since your notebook is very readable and well organized, it already takes care of the *how*, in your report, you should be more focused on the *why*. For instance, it's very important to justify your choice of algorithms to tackle this problem, as this is where you show that you know what you are doing. :)

I added several other comments and suggestions below, I hope they'll be of help. You are really close to meeting all specifications here - I wouldn't be at all surprised if your next submission is the one. Keep up the good work!

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.

Required

You did a good job here explaining where the project came from, but it's important to add a short paragraph or two to explain what you are trying to achieve here. What is the main goal of this Kaggle competition? The description section in the competition page can help you out here:

> Aspiring competitors will demonstrate insight into better ways to predict claims severity for the chance to be part of Allstate's efforts to ensure a worry-free customer experience.

Suggestion

Although it's an interesting information, I don't think you need to mention that a AWS instance was used to run the project.

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.

Awesome

Very good step-by-step description of your implementation. You also make sure to mention that this is a supervised learning problem and that the model should accurately predict the loss for new insurance claims. Great job!

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

Suggestion

To go above and beyond what is required here, you could discuss the differences between the mean absolute error and the root mean squared error.

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.

Required

As a general suggestion, for this and the following sections of your report as well, you could remove some (or even all) of the code and lean heavier on the explanations.

Here, specifically, there should be a more in-depth discussion about the variables. You show the distributions for all the variables, but what do these distributions *mean*? Is there any variable that strikes you as particularly important for the problem at hand? (Some 2-d analysis may be useful as well!)

Another important point to keep in mind, particularly for further data preprocessing: you mention that some of the categorical variables have many labels. Are these labels represented both in the train and test data sets? What would be the implication of a label that appears in the train set but not the test set, or vice-versa?

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.

Suggestion

I think you meet specifications a little too much here. :) Since all plots are in the notebook that you submitted along with your report, I think you can leave in your report only the visualizations that you deemed important. You could, for instance, add to your report only the violin plots for those numerical features which you deem worthy of further discussion.

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

Required

To meet specifications here, you should not only mention which regression models you'll be using, but also discuss how they work and justify your choice based on the problem at hand. Therefore, a short discussion of both Bayesian Ridge Regression and Gradient Tree Boosting is needed here.

Suggestion

As you probably know, many Kaggle competitors swear by Extreme Gradient Boosting, or XGB. Maybe you could give it a shot! :)

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

Suggestion

Great idea using the Kaggle benchmarks. If you are somewhat ambitious, you could try to beat a given proportion of other competitors - your goal would then be to score well enough to be among the top 25% competitors, for instance.

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.

Awesome

Solid and concise explanation for both steps taken in the preprocessing step. Don't forget to add to this section if more steps are required to deal with the "labels in categorical variables" issue I mentioned above!

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.

Required

One thing that was not clear for me here was why you used R-squared as a metric to pick the model, instead of going with the mean absolute error from the start. Could you expand on that? Also, make sure to mention whether you ran into any trouble implementing the initial instance of the model, as well as what the results were.

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.

Required

You should also justify your choice of parameters for tuning. What does `max_depth` do, for instance?

Suggestion

Even though you're not using XGBoost, its documentation provides notes on parameter tuning that can be of use. This post can also help!

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.

Required

You should also present information regarding the **robustness** of your model. You can use cross-validation results for this: do the results for each validation set vary a lot, or are they close to one another?

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.

Awesome

Congratulations for beating your benchmark! As suggested above, you may consider a more challenging benchmark for your next submission :)

Conclusion

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

Awesome

Good use of a violin plot here as a "sanity check" to make sure the distribution of the predicted loss function approaches that of the actual loss function.

Comment

> We can infer from the table that this change is a reflection of the limited maximum depth of the Gradient Boosting Regressor to avoid overfitting.

Can you expand on this idea? I'm not sure what inference you made here.

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

Awesome

Interesting, if short, discussion regarding both the running time of the algorithms and especially the comparison between the target variable and the numerical features.

Suggestion

> It may be because the features were given already preprocessed in a way that we don't know and/or because the categorical features have more influence in the 'loss' value than the numerical features.

You can use the `feature_importances_` attribute from a Gradient Boosting Regressor to check which features are more important for the learner.

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.

Awesome

Very good ideas here. I particularly like the mention of dummy variables and the removal of columns that represent values not present in the test set. Maybe in the next submission? :)

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.

Awesome

Your report is well written and easy to follow, congratulations!

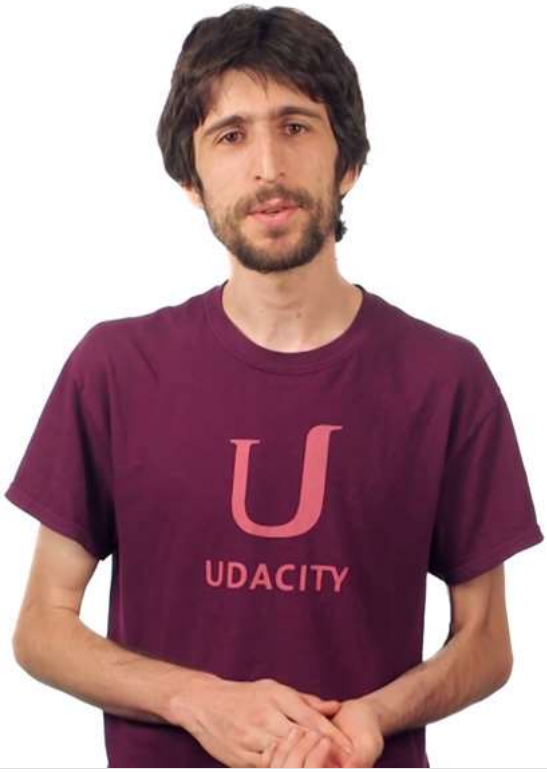
Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.

Suggestion

Your notebook is well organized and easy to follow as well. I only advise you to pay attention to the scikit-learn warnings and to try and fix your code so as to eliminate them. Most of them are deprecation warnings, meaning your code may not work with future versions of the package.

RESUBMIT

DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

Student FAQ