UDACITY

< Return to Classroom

# Data Scientist Capstone

REVIEW

CODE REVIEW

HISTORY

## Requires Changes

### 3 specifications require changes

Dear student

Nice start on your final project! I've noted some things that you should be sure to add/change in your blog post, but I think you'll see that these are mostly minor issues and shouldn't take you long at all.

- Please be sure to explain or justify why MAE/RMSE/MSE/R^2 were chosen as the regression metrics for this particular dataset.
- Please be sure to note or document the process of tuning or refining one or more of the models.
- Please add a `Model Evaluation and Validation` section to the blog post (see review for more details).

I've left the rubric information for the sections that need to be added or updated, but I'll also leave the project rubric link here too in case this would be helpful:

https://review.udacity.com/#!/rubrics/2345/view

Cheers!

Suggested reading for dealing with datasets with sparse features:

https://towardsdatascience.com/working-with-sparse-data-sets-in-pandas-and-sklearn-d26c1cfbe067

## Project Definition

✓

**Student provides a high-level overview of the project. Background information such as the problem domain, the project origin, and related data sets or input data is provided.**

> Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offers during certain weeks. Every offer has a validity period before the offer expires.
> Based on the information given above, we can infer that any transaction made by a customer is something really desired. No matter what offer you'd provide to the customers, still, you will deserve a transaction. Therefore, this project will try to predict how much money a customer will spend.

Nice overview of the problem domain! I love the focus on the real-world impact of the application.

Suggested:

- It's a good idea to cite some of the studies where the machine learning techniques that you're using were pioneered. This shows that you really know the field and it gives credit to the inventors.

✓

**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.**

> To solve the problem, we will use all datasets provided.
> We will employ several machine learning algorithms. As this will be a regression problem, we will evaluate the algorithms based on several regression metrics, i.e. mean absolute error, mean squared error, root mean squared error, and R2 score.

You've done a great job restating the problem clearly!

Suggested:

- This is a good point to begin to justify why your solution is a good 'fit' for the problem. If you were submitting this to a journal for peer review, you'd want to keep the readers focused on what you want them to think about. If they get distracted, they can ask for random things in subsequent revisions (which can significantly drag out the process and lead to arguments).

🔄

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

**For example, explain why you want to use the accuracy score and/or F-score to measure your model performance in a classification problem,**

> As explained earlier, this problem is a regression problem. Therefore, the metrics we used to evaluate all models are those for regression problems. So, we used:
> mean absolute error
> mean squared error
>
> root mean square error
> R2 score.

Nice work here! Please be sure to add a little bit more explanation for why you chose these specific regression metrics for this dataset or problem. For example, what properties of the dataset make these metrics optimal? Why not use other regression metrics like RMLSE, Mallow's Cp etc.?

From the rubric for this section:

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

## Analysis

✓

**Features and calculated statistics relevant to the problem have been reported and discussed related to the dataset, and 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.**

Well done! I think you've done a fantastic job explaining the scope and nature of the dataset for the reader. A sampling of the raw data is shown and you've noted some interesting properties of the data. Perfect!

✓

**Build data visualizations to further convey the information associated with your data exploration journey. Ensure that visualizations are appropriate for the data values you are plotting.**

> There's something fishy here. The number of agaes equal to 118 is 2175. Earlier, we found that the number of NaNs for each gender and income column is also 2175. Maybe, they have relationship each other. Let's check it.

Nice job visualizing a number of important properties of the dataset. In this case, 118 is a missing value code for this feature so you can definitely drop these data points or impute them.

> As explained earlier, we assume that the age equal to 118 means default value for unavailable data. Therefore, we're going to remove those ages.

This is definitely a 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.

There was a fair bit of preprocessing to document, but I think that you've done a great job explaining everything. Nice work!

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.

One of the goals for this project is to fully document everything in enough detail so that someone could (more or less) reproduce your results using only the blog post description. I think that you've definitely achieved that here. One other thing that I'd recommend adding would be to specifically note if there were any complications or difficulties that you encountered during the coding process.

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.

Please be sure to document the process of tuning or refining one of the models. For example, you could tune one of the models and document which parameters were tuned, what values you searched over for each parameter, and what the final parameter settings were.

From the rubric for this section:

> 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.

## Results

If a model is used, the following should hold: 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. For example, you can use cross-validation to find the best parameters.

Show and compare the results using different models, parameters, or techniques in tabular forms or

charts.

Alternatively, a student may choose to answer questions with data visualizations or other means that don't involve machine learning if a different approach best helps them address their question(s) of interest.

Please be sure to add a `Model Evaluation and Validation` section to the blog post. The idea with this discussion is to talk about the specific properties/parameter settings of your best model and discuss how they match up well with the problem.

- Please be sure to provide some discussion about the final parameters or characteristics of your best model. How do these align with the characteristics of the dataset? Why would these be a robust solution to the problem?
- Another approach that you could take would be to demonstrate that your optimized model is robust would be to perform a k-fold cross validation. In this case, you'd document how the model performs across each individual validation fold. If the validation performance is stable and doesn't fluctuate much, then you can argue that the model is robust against small perturbations in the training data.

From the project rubric for this section:

> If a model is used, the following should hold: 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 final results are discussed in detail. Explain the exploration as to why some techniques worked better than others, or how improvements were made are documented.

> From all four models, we can see that each model performs standard results. Random Forest provides the best performance in this case. Even, the NN model which should give better performance is unable to deliver the best performance.

I think you've really done a good job justifying that the solution to the problem is adequate. I'd also recommend adding a discussion about how you've managed to answer the big picture questions that you set out to address.

## Conclusion

✓

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

> I found this project challenging, mainly due to the structure of the data in the transcript dataset. My goal is to predict the amount will be spent by a customer. So, the decision-makers would know how customers behave in purchasing Starbucks products. Therefore, I need all rows with amount

existing, i.e. the rows with event col equal to transaction. However, using only this subset of the dataset, in fact, the models I've implemented can't give good performance. Even, by checking the correlations among all columns, we are unable to capture any strong relationship between the amount and other columns.
Perhaps, in order to get broader factors, we should consider the other events, i.e. offer received,

offer viewed, and offer completed. But, if we include the three, we will have many missing values. This is contradictory.

Great job summing up each of the main steps you took in the project. Keep in mind that this is also a great opportunity to brag a bit about anything exciting or innovative in your implementation.

✓

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

Due to time reasons, I don't have a chance to perform enhancement on data or model tuning. For example, I could do more experiments on feature engineering to see if adding/substracting features can improve the model; or I could try other combinations of model hyperparameters to see if this will affect the model performance.
In addition, the analysis and modelling should also consider the condition of customers who don't make any transaction as such a condition also provide more insight.

Great ideas!

Suggested:

- The XGBoost and CatBoost models could be good supervised learning approaches to try here.
- Since you're creating multiple supervised learning models, you could try combining them all together into a custom ensemble model:
  http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/
  https://www.kaggle.com/arthurtok/introduction-to-ensembling-stacking-in-python

## Deliverables

✓

**If the student chooses to provide a blog post the following must hold: Project report follows a well-organized structure and would be readily understood by a technical 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.**

**If the student chooses to submit a web-application, the following holds: There is a web application that utilizes data to inform how the web application works. The application does not need to be hosted, but directions for how to run the application on a local machine should be documented.**

The blog post follows the project rubric and is easy to read (thanks...this makes the reviewing process much

faster!).

Student must have a Github repository of their project. The repository must have a README.md file that communicates the libraries used, the motivation for the project, the files in the repository with a small description of each, a summary of the results of the analysis, and necessary acknowledgements. If the student submits a web app rather than a blog post, then the Project Definition, Analysis, and Conclusion should be included in the README file, or in their Jupyter Notebook. Students should not use another student's code to complete the project, but they may use other references on the web including StackOverflow and Kaggle to complete the project.

The Github repository meets all of the specifications and the README file is complete. Nice job!

Code is formatted neatly with comments and uses DRY principles. A README file is provided that provides. PEP8 is used as a guideline for best coding practices.

Best practices from software engineering and communication lessons are used to create a phenomenal end product that students can be proud to showcase!

The code is well commented and cleanly written. I don't think that a skilled programmer would have any difficulty getting oriented quickly when they look at the project code for the first time. Additionally, the code appears to produce the documented output. Perfect!

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH