UDACITY

Logout

PROJECT

Capstone Proposal

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

Requires Changes

4 SPECIFICATIONS REQUIRE CHANGES

SHARE YOUR ACCOMPLISHMENT

Hi,
Numerai is certainly an interesting data source to explore and you present a solid explanation about why it is relevant to analyze it. Even though you already discuss some interesting ideas to solve the problem, there are still some aspects of your proposal to be expanded, just make sure that you clearly present more details about your plan to solve the problem (e.g. which algorithms you are going to explore first). You will find more details about these in the rest of this review.

Best Regards.

Project Proposal

Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.

Awesome

- You discuss how hedge funds take advantage of machine learning to obtain better predictions of financial products.
- Some examples of hedge funds were included

Suggestion

- I was going to ask for more details about the relevance of the specific area for which your project is related. However, you do that in the Problem statement. My suggestion then is that you move part of your discussion from the Problem Statement to this section. Specifically, from "The difficulty to compete [...]" up to "[...] designed algorithms of the big quantitative funds." could be moved to your Domain Background section. The idea behind this suggestion is that you introduce numerai while discussion the background and focus in the specific problem definition in the Problem Statement section.

Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.

Awesome

- You correctly mention that the inputs are encrypted.
- The seasonality for which the data is presented is mentioned.

Required

- You mention the inputs of the problem, but you do not mention how the expected outputs look like. Even if you are not aware of the semantics, it is important that you specify that the machine learning problem is a binary classification problem.

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

Awesome

- You mention the source of the data (numerai)
- The volume of training and testing data was properly defined.
- The problem was identified as a binary classification problem.

Required

- I know it is difficult to discuss the input features as the data is anonymized, but it would be useful to mention here what is the domain of the features (e.g. numeric features only?) and how many features there are.

Suggestion

- If possible, also discuss whether or not the class label is evenly distributed, i.e., verify if there is an imbalance between class 0 and 1. Maybe this changes from week to week.

Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

Awesome

- This looks like a really interesting idea:

  > "I ANTICIPATE TO USE CLUSTERING METHODS FOR DIVIDING THE DATASET"

Required

- It is understandable that at this point you do not know exactly which algorithm (or set of algorithms) will generate your final model. However, you have to start somewhere, right? Define a list of algorithms to try and verify their performance. That is what is expected in the proposal: at least mention which specific algorithms you planning to try. See my suggestion.

Suggestion

- To cover different biases, try at least decision trees, k-nearest neighbors, naive bayes, SVM and multi-layer perceptrons. It is highly recommended that you also use ensembles of decision trees, such as random forest and extra trees, and other ensemble methods such as XGBoost and Gradient Boosting might be good to explore as well.

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

Awesome

- Indeed 50% performance is a reasonable baseline for this project.

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

Awesome

- Accuracy is a reasonable metric for this problem.

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

Required

- It is not clear what you mean by this:

  > "I PLAN TO SLICE THE TRAINING SET INTO 4 PARTS, WITH THE FIRST PART FOR TRAINING AND PROTOTYPING A MODEL, AND THE REMAINING 3 PARTS FOR IMPROVING THE MODEL, WITHOUT CHANGING THE STRUCTURE OF THE MODEL."

  What kind of model do you have in mind? I.e. what learning algorithm are you focusing on while discussing this? Because, one can interpret that you are going to try different neural network structures using the first 1/4 and then use the other 3/4 to train it, without changing the structure. In the other hand, it is not clearly specified that you are going for a neural network, so perhaps when you mention "structure" you are referring to something else?
- In your solution statement you mentioned using unsupervised learning to group data before training. This should be represented in your workflow as well.

Suggestion

- If you apply transformations to the input data, for example, grouping using unsupervised learning, then test the learning algorithm on both datasets (grouped and not grouped) to assess how much the transformation improved the results.

Proposal 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 and referenced are properly cited.

☑ RESUBMIT PROJECT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Student FAQ