Capstone Proposal

REVIEW | CODE REVIEW | HISTORY

Meets Specifications

Dear student

Great work on this proposal! I think that you're meeting the specifications and it's clear that you have a pretty good idea of what you want to do. Your suggestions are all feasible and I think you're on the right track.

About the capstone project:
While the code and implementation are both important, keep in mind that the capstone report is the most important element of your final project. This report simulates a formal submission to a journal for peer-review. Publishing your results is a key credential in machine learning and we want you to be ready for this!

You should have very little trouble quickly passing the project review if you initially follow the report template. Keep in mind that reviewers will be looking to see that you not only fully document how you implemented your project, but why you made the choices you made. This type of critical thinking is crucial to publishing in a peer-reviewed journal. Based on your proposal, I don't think you'll have much trouble with this, but I wanted to mention it up front.

I think you're definitely on solid ground and you've picked an interesting topic for your project. I'm looking forward to seeing the final result!

Cheers!

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.**

Great job giving the reader an introduction to the problem domain!

Suggested:

- If you include a link to your data source in this section, you can directly lift it into the 'Project Overview' section of your capstone report.

✓ **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.**

> As a work-around, we employ the following simplication: Clicks followed by app downloads are legitimate, whereas clicks that don't lead to downloads are fraudulent. With this simplification in place, we can now frame the problem as a supervised learning problem, and more specifically, we are to construct a binary classifier for predicting whether or not clicks are followd by app downloads.

Nice job! I think that this is very clear.

✓ **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.**

You've done several things that are quite excellent here! Aside from giving a thorough overview of the scope of the dataset, you've picked a nice sub-sample of a very large dataset to work with, and you've done a great job documenting how the dataset classes are balanced.

✓ **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.**

> Because the amount of no-download clicks is much higher than its counterpart, we have highly unbalanced classes at hand that will require special care during model selections and evluations.

Choosing AUC-ROC as your primary metric is definitely a good start to avoiding issues with a skewed dataset!

✓ **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.**

> Logistic regression is simple, fast, and easy to interprete. We will use it as the benchmark model both to evaluate the signal strengths of the input features and to compare its performance against those of more sophisticated models.

This should work well. Logistic regression is a common default implementation, so you'll be able to easily justify why this is the baseline for your project. However, it won't be overly hard to beat either.

✓ **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.**

✓ **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.**

I think that you're meeting the specifications here. Some suggestions:

- The XGBoost and LightGBM models are strong performers for this type of problem.
- Since you'll be creating multiple supervised learning models, you might consider combining them into a custom ensemble algorithm:

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

✓ **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.**

The template format is followed and the proposal is well written.

DOWNLOAD PROJECT

RETURN TO PATH