UDACITY

< Back to Machine Learning Engineer Nanodegree

# Capstone Proposal

REVIEW

CODE REVIEW

HISTORY

## Requires Changes

2 SPECIFICATIONS REQUIRE CHANGES

This will be a very fun real world classification project. Just need to include a couple more things here in your proposal and you will be off and running. Some of these aspects are very important to be aware of before you begin your implementation. Check out some of the other ideas presented here and we look forward to seeing your next submission.

### 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.**

You have given good starting paragraphs to outline the project and have provided background information on the problem domain.

Therefore lastly for this section, make sure you also provide some research / links for other machine learning problems similar to yours(as you will need to do this in your final project anyway). It is always important to provide similiar research on such a topic to give some backing to your claims.

**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.**

> "for a given project, given a set of input features/project characteristics x, I aim to train a classifier f(x) such that it accurately predicts the target variable y, which indicates whether the project was successfully funded or not."

Love the ML problem statement here! Impressive! Problem statement is clearly defined here. And glad that you mention that this would be a classification problem in this section.

**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.**

Glad that you mention the size of the datasets, just need to document one more thing here

- Make sure you also give some ideas in terms of the distribution of the target class (successfully funded or not) in your training file. Do we have an unbalanced dataset? Balanced? Make sure you fully analyze the distribution of the target class, as this might determine what evaluation metric is appropriate in your problem and might be a big part of this particular dataset that you need to be aware of before diving in.

---

NOTE

> "Given the low dimensionality of the dataset, some new and hopefully illuminating features will also be constructed from the given features."

For sure! Especially with an 'interpretable' dataset like this one, understanding your dataset, creating features, and important pre-processing steps are always needed! Since this is a Kaggle competition, always check out the Kaggle forums and Kernels, you will always find really cool ideas there.

**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.**

I think you have a good Solution Statement here, as it is clear that you have thought a lot about what your approach is to this problem.

**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.**

> "The benchmark model for this project will be a simple logistic regression model (see here) which uses only the given features in the dataset (i.e. no feature engineering was conducted). This model achieved an accuracy score of 63.19%."

Good choice in a benchmark model here. Always good to get a baseline score for your dataset with a linear model.

Benchmarking is the process of comparing your result to existing method or running a very simple machine learning model, just to confirm that your problem is actually 'solvable'.

**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.**

If you do have an unbalanced dataset, just make sure you focus more on your F1 Score metric. As using accuracy is not the best for unbalanced datasets based on the accuracy paradox . As you do mention.

**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.**

You clearly have a good game plan in place here, very solid step by step process.

Nice ideas for potential algorithms. An Xgboost is a good choice. Could also look into using a LightGBM. Here might be a couple of examples of how to implement these powerful algorithms.

- Xgboost example
- LightGBM example

---

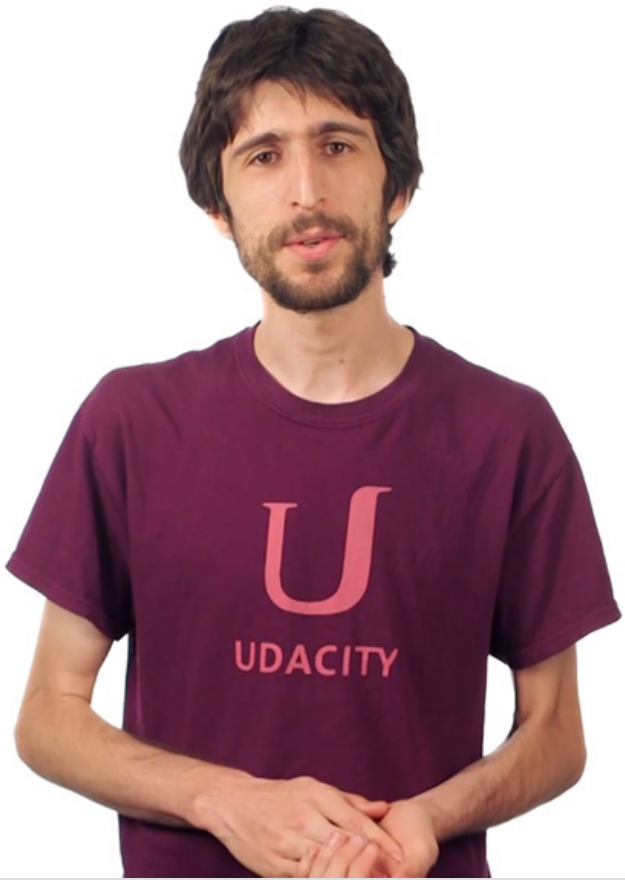
> " will explore the dataset for missing values and choose appropriate steps if they are encountered - either choosing a strategy for filling them in, or omitting rows containing missing value - depending on how frequently they occur and the circumstances in which they occur. "

Typically it is never recommended to drop any data regardless in the amount. Maybe these missing row had some great correlation with the target variable? Typically we add a new column indicating that the row was missing then input the missing value with the median or another idea. If you do want to get fancy, you could also run a supervised learning model to 'predict' the Nan value!

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

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)

RETURN TO PATH

Student FAQ