PROJECT

Capstone Proposal

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

Meets Specifications

SHARE YOUR ACCOMPLISHMENT

This should be a very interesting project and really suitable for the real world. Check out some of the other ideas presented here, but you are clearly ready to get started on this final report. Wish you luck 😄

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.**

I would recommend describing the problem a bit more(in this section) and why this is a real world problem and why machine learning can solve this. But nice personal touch here.

✓ **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.**

> "we can predict the claim cost by using Supervised Learning algorithms, specifically Regression algorithms to predict the continue value of a Claim Loss."

Your problem statement is clearly defined and glad that you have mentioned that this is a regression problem. I would also recommend mentioning some of the inputs(features) that will be in this analysis.

✓ **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.**

Nice simple description of your dataset and you have done a good job analyzing the feature presented to you from Kaggle.

> "a total of 116 categorical features and 14 continues features,"

Seems that you have a lot of features here! Definitely after you one-hot-encode the categorical ones. Therefore feature selection might be an important step. Therefore one idea to check out would be the feature_selection module in sklearn. As SelectKBest and SelectPercentile could also give you important features.

One more idea(and you should definitely do this in your final report) would be to analyze the distribution of the target variable. Is it normally distributed? Do we need any data transformations?

✓ **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 strategy for solving the problem and really like the details in your different algorithms you plan on using! Maybe even try combining these models and use something like a VotingClassifier

✓ **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.**

> "As benchmark, I will be using a Support Vector Machine model to compare the MSE to the MSE obtained for the Capstone Project"

Rate this review

☆☆☆☆☆

You can definitely use an SVM for a benchmark model. I would recommend using a more simple machine learning model instead. Maybe instead look into using a simple linear regression model or Naive Bayes to get a very simple baseline to get an idea of how a simple model could perform.

✓ **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.**

Could also check out RMSE, as the benefit this provides over MSE is that the error is in the same *units* are the prediction. So you can quickly see how far off your prediction is.
(https://www.vernier.com/til/1014/)

✓ **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.**

> "Reduce the dimensionality: As we have several features and it will grow a lot when we encode the categorical features, we need to find a way to reduce the dimensionality of the features. We will be using Sklearn’s PCA and SelectKBest to achieve this."

This will be a key step here to get a good model. As this is typically where most of the time and effort is needed. Maybe another idea would be to run like a Random Forest / Decision Tree and check out the forest_importances.

Maybe also look into combing some steps 2 and 5 with the notion of Pipelining, check out this blog post brought to you by Katie from lectures

- (https://civisanalytics.com/blog/data-science/2016/01/06/workflows-python-using-pipeline-gridsearchcv-for-compact-code/)

✓ **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.**

DOWNLOAD PROJECT

RETURN TO PATH

Student FAQ

Rate this review

☆☆☆☆☆