Capstone Proposal

Machine Learning Engineer Nanodegree

# Domain Background

\_(approx. 1-2 paragraphs)\_

In this section, provide brief details on the 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 in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required.

* Financial predications used to understand market and make profitable trades.
* Can a better prediction be made if other data sources are used as well as the market data.
* Lots of info out there to help predict.
* Twitter has lots of info
* Any academic research into this topic
* Want to do this as want to play around with time series data and twitter api
* Need to mention the type of learning that I will be using
* Also mention sentiment analysis background

# Problem Statement

\_(approx. 1 paragraph)\_

In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms) , measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).

* Can stock market predictions be made better by using twitter data.
* How to quantify this
* How to measure this
* Over multiple stocks
* Using different twitter data (the company, mentions of the company etc.)

# Datasets and Inputs

\_(approx. 2-3 paragraphs)\_

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary 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.

* Where to get historical stock data
  + need training data, use this to quantify the performance
  + describe dataset
* How to get historical twitter data
  + twitter api
  + sentiment analysis module?
  + describe dataset
* Use pandas-datareader to query yahoo finance for historical stock data.
  + Will read from web then save to csv to speed up future requests.
  + returns open, high, low, close, adjusted close and volume.

# Solution Statement

\_(approx. 1 paragraph)\_

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

* What type of model for timeseries data?
* compare one with twitter info to one without
* maybe just use the output of the basic model then adjusted by the sentiment analysis
* any other ways other than sentiment analysis

# Benchmark Model

\_(approximately 1-2 paragraphs)\_

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail.

* What are some benchmark models for finance predictions
* buy and hold SPY
* Just a linear fit?
* Same as before

# Evaluation Metrics

\_(approx. 1-2 paragraphs)\_

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

* Need to look up some standard metrics that are used for this kind of thing

# Project Design

\_(approx. 1 page)\_

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

* Import stock data
* Basic model with no twitter data
  + Compare to benchmark
* advanced model with twitter data
  + sentiment analysis of tweets
    - Companies own twitter
    - tweets that mention the company
    - Related companies
  + stats about tweets
    - number of tweets by company
    - number of tweets that mention the company (compared to average?)
  + How to combine this info, do I take the predication from the basic model and pass this into a different model with the twitter info to give a final prediction. Do I just try and predict the delta between my model and the basic model.
* Pseudocode?
* Visualistaion?

-----------

\*\*Before submitting your proposal, ask yourself. . .\*\*

- Does the proposal you have written follow a well-organized structure similar to that of the project template?

- Is each section (particularly \*\*Solution Statement\*\* and \*\*Project Design\*\*) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

- Would the intended audience of your project be able to understand your proposal?

- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?