Machine Learning Engineer Nanodegree

Capstone Proposal

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.

- Commodity is one of the largest traded (by volume) commodity Internationally as well as domestic.
- Commodity markets are highly fragmented pricing.
- No standard price, which is why there is no exchange for this product.
- No set price for any product, production cost differs from producer to producer.
- Multitude of categories of products/items/sizes/grades within the categories with substantial price differences.
- The scope is to pick one commodity item

My motivation to pursue this as a project because I am somewhat related to this industry where we face the pricing fluctuations on a daily basis. In addition, I am trying to develop a concept prototype which can be utilized by small businesses.

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).

- Use Machine Learning concepts to predict commodity pricing fluctuations.
- Determine significant correlations between features effecting the commodity pricing.
- Implement an interactive commodity price forecasting mechanism classifying them under the following categories:
  - Specific product
  - Specific region
  - Time range

Challenges with Commodities

- Very fragmented buying-selling process – in the form of brokers, agents, traders, consumers, stock-holders, producers.
- Unorganized and unregulated – special license or membership not required.
- Buying methods are still very old-fashioned.
- Highly relationship based.
- Price intelligence – mostly through verbal communication.
- Published journals give historical prices and informal projections but no specific price predictions.

Features Considered

Factors influencing Commodity prices:

- Production/availability/demand
- Imports
- Exports
- Iron pricing
- Coal pricing
- Petroleum prices
- Construction
- Infrastructure projects
- Bank financing
- Interest rates
- Currency fluctuations
- Government policies
- Weather conditions
- Labor market and policies
- Ocean freight
- Logistics
- Seasonality

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.

Due to the type of industry most of the time series datasets and inputs are scattered in industry journals and metal exchanges. Below are some of the source of data:

- Industry journals: Industry journals – AMM (USA), Metal Bulletin (UK)
- Specialized web-sites such as Platts, SBB, etc.
- IMF Primary Commodity Prices
- Department of Commerce – Data catalog (data.gov)
- Historical Commodity Prices (datahub.io/dataset/commodity-prices)

- London Metal Exchange, for example
  - https://www.quandl.com/api/v3/datasets/LME/PR_CO?api_key=2rz93HrjCax7knFkwXAJ

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).

The problem I am trying to solve here is to develop some sort of price guidance mechanism for commodity item, in particular a metal commodity product. I will be able to solve the problem is but determining correlation between various variables in order to predict a pricing trend of the commodity.

Furthermore, I will use the correlation matrix and Time Series Forecasting to achieve a high accuracy on the commodity pricing in the following categories:

- Time range
- Region specific
- Product category

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.

- Correlations generated results will be compared against historical data
- Achieve high accuracy of the predicted prices compared of historical data
- Attempt to predict pricing slightly better than London Metal Exchange

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).

Due to the nature of the datasets the evaluation metrics would be based on RMS (Root Mean Square) or similar average-based mechanism. However, I will compare the accuracy and the feature

importance with historical data & other forecasting algorithms results mentioned in the project design section.

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.

- Importing and preprocessing the data
  - Merge several dataset for a particular commodity from different sources
  - Commodity monthly dataset is available from 1980-2017
- Preparing training and test data
  - Training & testing data split chronologically 80:20
- Data scaling
  - Scaling in Python using sklearn MinMaxScaler
  - Activation functions to scale the data available
- Techniques
  - Use Supervised Learning & Apply Naive Bayes & Bayesian Models
  - Apply Classifiers & Logical regression to find correlations between several impacting factors
  - Convolution neural networks to apply differing weights for each of the features
  - Explore Forecasting Industry Standards
    - Autoregressive (AR)
    - Moving Average (MA)
    - Autoregression Moving Average