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.

Commodity price forecasting and predicting is not a new area of interest. There has been some investigation in this area, some practical and academic headway is made, but mostly for precious metals and commodity metals such as copper or aluminum. Here are some academic journal references:

- Copper Pricing based on ML: https://studentnet.cs.manchester.ac.uk/resources/library/thesis_abstracts/MSc16/FullText/Olayiwola-Adegbenga-diss.pdf
- Gold pricing & Futures: http://univagora.ro/jour/index.php/ijccc/article/viewFile/2009/pdf
- The Prediction of Precious Metal Prices via Artificial Neural Network: http://alphanumericjournal.com/media/Issue/volume-5-issue-1-2017/the-prediction-of-precious-metal-prices-via-artificial-neura_SzcS45q.pdf

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:
  - Time range

There are two aspects of the problem I am exploring to solve. The first one is to pick one specific commodity, Steel Billets, and use historical data (over time) on one or more of the variables (features) listed on the "Factors and Features" sections to predict the "future" price indicator of the specific commodity.

Secondly, I will use existing commodity prices data as time-series forecasting. I will start with data over some long period, break it up into many time slices then try to define a linear regression model for each time slice to predict a future value (future relative to the time slice).

The output in the first case will be an indication on how the commodity price will change based on the feature importance. While in the second case the output will be a prediction of commodity based purely on the existing pricing data.

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: LIBOR rate, Iron pricing, Coal pricing, Petroleum prices, Construction, Interest rates, Currency fluctuations

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)
- Department of Commerce – Data catalog (data.gov)
- Historical Commodity Prices (datahub.io/dataset/commodity-prices)
- London Metal Exchange, (LME) for example: https://www.quandl.com/api/v3/datasets/LME/PR_CO?api_key=2rz93HrjCax7knFkwXAJ
- FRED Economic data: https://fred.stlouisfed.org/series/USDONTD156N

The data will be extracted from LME via Quandl APIs & download CSV from FRED.

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

For forecasting the pricing, I am considering implementing LSTM using keras library, however, another proposed solution could be a RandomForestRegressor or gradient-boosting regression.

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

For the benchmark model, I plan to use a simple linear regression of price vs time or a moving average of the price vs time.

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.

Since my target variable (price) is a continuous variable, I would use either of a linear regression problem and either MSE, RMSE or MAE (mean absolute error) as an evaluation metric.

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.

Development workflow will follow the sequence described below:

- Set Up
    - iPython Notebook and incorporate required Libraries (Keras, Tensor flow, Pandas, Matplotlib, Sklearn, Numpy)

- Prepare Dataset
  - Importing and preprocessing the data
  - Merge several dataset for a particular commodity from different sources
  - Process the requested data into Pandas Dataframe
  - Develop function for normalizing data
  - Scaling in Python using sklearn MinMaxScaler
  - Dataset will be used with a 80/20 split on training and test data across all models
- Feature importance techniques on 2-3 features impacting the commodity prices
- Develop Benchmark Model
  - Set up basic Linear Regression model with Scikit-Learn
  - Calibrate parameters
- Develop Basic LSTM Model
  - Set up basic LSTM model with Keras utilizing parameters from Benchmark Model
- Improve LSTM Model
  - Develop, document, and compare results using additional labels for the LSMT model
- Visualize Results
  - Plot Actual, Benchmark Predicted Values, and LSTM Predicted Values
  - Analyze and describe results for report.