Machine Learning Engineer Nanodegree

Capstone Proposal – The Winton Stock Market Challenge

John Yiu 7 August 2018

Domain Background

In the financial market, price movement of stock prices are never entirely predictable. Having a reliable way to forecast stock prices is crucial for most trading strategies. Many techniques have been developed to make predictions, ranging from fundamental techniques that involve analysing company financial statements to technical analysis which utilises indicators derived from the prices.

In particular, price prediction using machine learning algorithms have become very popular in recent years, due to the success in leveraging data mining techniques in the other fields in science and technology, as well as the increased availability of data. The topic is chosen based on my personal interest in the field. In addition, it was a competition hosted on Kaggle which should also allow me to compare against the others, and prepare for data science competitions in the future.

Problem Statement

In the Equities market, stocks of listed companies are traded on exchanges. Different kind of market participants perform trades on the stock which causes the prices to vary intraday. These price movements form a non-stationary time series data. Although the series formed are often very noisy, one can extract signals from historical price trends in some occasions and forecast future prices.

Based on a given dataset containing historical stock performance (close price of the last few days) and various masked features, this project aims to predict the intraday and end of day stock returns using deep learning techniques. The problem was a live competition hosted jointly by Kaggle and Winton Capital.

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).

Datasets and Inputs

The dataset was provided by Winton as part of the challenge, its link can be found at the end of this section. It consists of a 5-day time frame, days D-2, D-1, D, D+1 and D+2. The returns in days D-2, D-1 and part of Day D have been provided and we have to predict the returns of the remaining period of days D, D+1 and D+2.

Specifically, for day D, intraday returns containing 180 minutes of data, from t=1 to t=180 are provided. The size of test set has been predefined, and in this challenge only the first 120 minutes of intraday return are provided. In addition, for each 5-day window, 25 anonymous features (Feature_1 to Feature_25) are given, these could be useful for the prediction.

The data is given in a tabular form where each row is an arbitrary stock at an arbitrary 5-day time window. There are 40k rows in total.

It is also worth noting that these returns are calculated by Winton, where the methodology is not revealed. However, they are designed to be representation of real world data.

Link to the dataset - https://www.kaggle.com/c/the-winton-stock-market-challenge/data

Solution Statement

If there exist patterns or trends in the stock price time series dataset provided, the problem could be approached by using a neural network to exploit such possible patterns in the training set, and to make forecast. Specifically, the input variables would be the returns of D-2, D-1, 120 minutes of intraday return of D, plus 25 additional features, and the output would be the remaining 60 minutes of returns from t=121 to t=180 of D, and end of day returns for D+1 and D+2. The model should be optimised in a way such that the weighted mean absolute error between the actual and predicted returns are minimised.

Benchmark Model

The model can be benchmarked against a dummy model which predicts returns using a uniform random distribution. Most modern financial models assume stock returns to be normally distributed. Thus, for our model to be sensible, its weighted mean absolute error should be lower than a naive model that uses a uniform distribution ranging from the minimum and maximum returns between return D-2 to D.

Evaluation Metrics

Specifically, the prediction model will be evaluated using the weight mean absolute error (WMAE), where the actual return will be compared against the predicted return:

$$WMAE = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot |y_i - \hat{y}_i|$$

where w_i is the weight associated with the return i, y_i is the predicted return, $\hat{y}_i$ is the actual return, and n is the number of predictions.

A total of 62 returns have to be generated by our model for each 5-day window (row in the input dataset).

Project Design

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.

The proposed theoretical workflow includes data pre-processing, model design and training, validation with testing set, and post processing to adhere to submission format required by the competition.

To solve the stock market challenge, the training and testing sets would be acquired from Kaggle. Preliminary check should be performed on the dataset to filter out any extreme or erroneous data points. Then the additional anonymous features should be examined, and one hot encoding can be applied if there exist non-numerical data types.

We will then attempt to train a neural network to predict the required data output. Due to the nature of stock price movements, long-short term memory network should be suitable for this problem. The trained model will be evaluated using the weight mean absolute error metric defined above, and we would adjust the hyperparameters in our algorithm, such as the number of layers to be used in the neural network, the size of each layer, the activation functions, and whether to include a dropout layer to reduce overfitting, etc.

The output from our model would be transformed into a comma separated file consisting of two columns. The solution will be written in Python, leveraging deep learning libraries such as Keras and Tensorflow.

References

https://www.kaggle.com/c/the-winton-stock-market-challenge

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180944

https://www.sciencedirect.com/science/article/pii/S2405918818300060