Machine Learning Engineer Nanodegree

Capstone Proposal

Yee Jie Thay

19 May 2018

Domain Background

The main aim of the project is to create a tool in facilitating indivudals to get better entry points for trades and know when will be a good time to realise profits/ losses of trades so as to maximise returns on investments.

Utilising daily trading information avaliable, we have a rich set of time series data which can be used to understand behavioural patterns of investors. Such behavioural patterns such as bouts of euphoria or signs of panics will allow investors to better understand the environment they are deploying their capital in.

You are free to choose what form your project takes (a simple script, a web app/service, Android/iOS app, etc.), and any additions/modifications you want to make to the project (e.g. suggesting what trades to make). Make sure you document your intended features in your report.

Problem Statement

For this project, I will look to build a stock price predictor that takes daily trading data over a certain date range as input, and outputs buy/sell signals for given query dates.

The problem that investors often face with a stock is that thay know which companies are the industry leaders and who the winners are. However, investors do not know the opportune time to invest in it such as to maximise their profits. This results a significant reduction returns. This is especially so for stock whose prices that exhibits high volatility or for investors with a shorter time horizon.

One has to be mindful that different shares exhibit different behaviour. To reduce the universe of shares to investigate in this project, we will aim to have a model that specifically looks at the behaviour on the components of Dow Jones Technology (DJUSTC) Index. https://www.investing.com/indices/dj-technology (https://www.investing.com/indices/dj-technology)

Datasets and Inputs

I plan to use a magnitude of data coming from several sources.

1) Bloomberg API

a) I plan to get several different types of price information such as historical opening and closing pricing, volumes, intraday high and intraday low.

b) I also plan to get data such as earnings date, Central Bank rates decision dates, major economic data releases etc so as to be able to get align them with the share price moves and enrich our price data.

c) I will look to classify all the shares into one of the following 9 categories. However, given the magnitude of the data, it might be wise to focus on a single sector. The sector chosen will be the Technology sector and the index that we will be comparing to will be Dow Jones Technology (DJUSTC) which has 131 components and will form the basis of our analysis. This means that the current scope of data will focus solely on Technology stocks with an eye to the future of expanding it out to other sectors such as

1) Basic Materials
2) Consumer Goods
3) Consumer Services
4) Healthcare
5) Financials
6) Industrials
7) Oil and Gas
8) Utilities

c) Lastly, I will be looking to extract various information of the company at each point in time such as the earnings, sales and dividends.

2) Twitter

a) This mainly involves getting tweets which has the ticket name in it so as use it as proxy of the sentiments associated with the stock. This is done by running it through a word processor so as to get the words with the highest frequency associated with the word as a proxy of the sentiment of the stock.

Solution Statement

The solution to the problem is such that we perform better than just random guessing. If a buy signal is generated, we will look at the 30 days return of the share against our benchmark index. If the return is positive, we will classify it as a true positive. If the return is negative, we will classify it as a false positive. Similarly, if a sell signal is generated, we will use a similar method to classify true negatives adn false negatives. The solution will be defined as the area under the receiver operating characteristic curve (ROC curve) between the predicted probability and observed target.

Benchmark Model

Most benchmark models that we can use for direct comparison are proprietary and hence will not be easy for us to use as reference for performance benchmark. However, the benchmark for the financial industry has always been indices. Thus, a good reference benchmark model will be the returns achieved by various indices over the course of 30 days.

We will eventually classify the shares into 9 different sectors which has 9 different benchmark indices from the Dow Jones Index family. But currently, we will specifically look at the Technology index so as to make sure that we are comparing equivalents. Dow Jones Technology (DJUSTC) Index has 131 components to it and it will be the sector we will be applying our model to.

Evaluation Metrics

The model will be evaluated on area under the Receiver Operating Characteristic curve (ROC curve) between the predicted probability and the observed target.

We plot the curve by using the model to calculate True Positive Rate (True Positive divided by the sum of False Negative and True Positive)and False Positive Rate (False Positive divided by the sum of False Positive and True Negative) with different thresholds for the logistic regression.

An area under the curve of more than 0.5 would mean that we have performed better than randomly guessing.

Project Design

I envisage the workflow to be broken up to 5 main steps.

1) First and foremost, I look to splice the data set into a consistent set of data for data exploration. This will mainly be carried out in Python and will involve ensuring consistency in the data for interpretation at a later stage.

2) This will be the stage of data exploration where I will be spending time to explore the structure of the data. This involves plotting the distributions of various variables that are of intuitive importance as well as looking at the various correlations of the variables in the data.

3) Next I will embark on processing the data. This means to refining the data and creating new variables that are normalised or creating data that describe the data better such as ratios. It is also important in cleaning out dirty data as well as removal of outliers that might lead to unintended behaviour in our model.

4) I will look to built the neural network and different layers of it using Python Keras module and train the model. I will specifically set a subset of results aside for testing the model and will not use it for any training.

5) Once the results are in place, I will have to structure a report so as to correctly reflect the analysis that has gone through the whole project and ensure that it is properly explained and sufficiently rigorous in methodology.