Capstone Proposal. Predicting Future Stock Prices

Dias Irishev

March 2020

1 Background

Since the stock market was invented people have tried to come up with creative solutions to "beat it". With the invention of computers people have tried to use it's computational power to try and make a leverage on the stock market. Different areas of mathematics have given the rise to the quantitative analysis, where certain smart individuals have created tools and theories to predict the behaviour of the stock markets for their benefit. Today, according to some [1], up to 70% of trading is performed through the "trade-bots" and the algorithmic trading. Unlike basic algorithms (basic linear regressions and such) today algorithmic trading is way more sophisticated and wide range of machine learning algorithms can be used for different sets of data. For this project we will explore how deep neural networks can be applied to try and predict future behaviour of the stock market.

2 Problem

The goal of this project is to Predict stock prices using deep neural networks. Specifically, we will create a user-interface to predict a value of a user-specified stock from the list of NASDAQ indices. Machine Learning algorithms that we will use for this is LSTM Neural Networks and Linear Regression.

3 Dataset and Inputs

After some googling and considering different data sources (yahoo finance api, google finance api, and bloomberg api), we chose to use data provided by the service named Alpha Vantage [2]. This service provides us with the historical daily data on NASDAQ stocks for the past 20 years. The data contains information on open, high, close, low, volume values of different indices, and value of the stock.

Registration on the website will be required to obtain free API. With the API code, we can access 'json' files and download 'csv' files as well. These

datasets are already optimized to be used with python and pandas. However most probably some basic manipulations will have to be performed in order to input these datasets in the deep neural networks (such as cleaning, splitting data, scaling).

4 Solution

To predict the future prices, LSTM Neural Networks will be used. LSTM Neural Networks is a type of RNN, recurrent neural networks. Unlike regular neural networks, RNN neural networks have a capacity to store the previous information, and use it to make better decisions[3]. This seems especially useful when we deal with the time series data.

During the project different types of RNN architectures will be used, to decide which one has better predictive power.

Data will be split into the train and test data. Train data will be used to used to train data, and test data will be used to analyse the performance of the machine learning algorithm of our choice. After the machine leagning models will be trained, we will create a predicted values and be able to compare it with the real data.

5 Benchmark model

For this problem linear regression with OLS (ordinary least square) estimation will be used as a Benchmark model. Linear regressions is pretty simple and straight forward mathematical approximation of the graph, that gives reasonable approximations and predictions. Due to personal previous experience with time-series data and linear regressions, we would like to see how the LSTM Neural Network will perform against the Linear Regression Model.

6 Evaluation Metrics

To see how two models will perform, we will use mean squared error (MSE) to see how far the predicted data differs from the real data. MSE is a standard measure of a quality of an estimator. The smaller the MSE the more precise is the estimator [4].

In addition visualisation will be used to get a sense how well the models are predicting. Predicted data will be sketched against the real data.

7 Project Design

First we have to create a function that will use free-API of Alpha Vantage to extract the daily trade data of the NASDAQ stock indicies. After our data is acquired, a function that will split the data and clean it will be created. The

data will be split into what we want to predict (opening value of a stock on the next day) and the rest of the historical data about other stocks (the hgih, low, open, close, and volume traded). The data will be further split into test and train datasets. Train to feed into our machine learning algorithms, the test to see how the modes perform with predicting.

Next, we will create a user interface that will take as inputs the index of interest and the historical data window (the ammount of days of historical data on which we will train the model). This interface will make sure that users provide the correct inputs (the indices that exist within NASDAQ stocks, and the reasonable number of days to base the prediction on). The default index and number of days that we will use "GOOG" and '50' respectfully, to analyse the performance of our models.

After the user interface will be implemented, two machine learning models will be created. First one will be a linear regression and the second one will be LSTM Neural Network. Two functions will be created, inside whcih both models will be specified and trained. The output of these functions will be a graph of two datasets: the real vs the predicted datasets of price of the stock, along with the MSE's for both models.

During the process of creating LSTM Neural Network, we will try various architecture structures to see which one will perform better with the Google stock data and 50 days data window. For the final model the architecture that performed better (had lower MSE) will be used.

After the predicted data will be acquired, the performance of two models will be compered to each other, and further possible improvements will be stated.

References

[1] https://www.experfy.com/blog/the-future-of-algorithmic-trading.

[2] https://www.alphavantage.co/about.

[3] https://colah.github.io/posts/2015-08-understanding-lstms/.

[4] George Lehmann, E. L.; Casella. Theory of point estimation, 1998.