Machine Learning Engineer Nanodegree

Capstone Proposal

Predicting Short-term Single Stock End-of-day Close Prices using Machine Learning

Blake Stratton
August XX, 2018

Please note: I recognise the proposed problem is not particularly revolutionary however at this early point in my learning it will be feasible for me to complete and still test my ability to identify and solve a problem using machine learning (in an academic styled setting).

Proposal

Domain Background

We will investigate the prediction of single stock end-of-day close prices over 1 and 5-day trading windows using machine learning. Understanding the potential or likely path of stock prices is of interest to various participants in the equity financial markets. For short-term windows these may be short-term day traders taking advantage of short-term price movements or financial institutions making markets as agent (e.g. the sale of block stakes).

Stock prices reflect the value per ownership interest based on information available to investors and potential investors (the market). A company's ownership is split into a large number of equal ownership interests representing a small % of the company called single stock common equity. These stocks are traded during trading hours (generally several hours from morning to late afternoon) on markets across the globe generally for a number of days per week (Monday to Friday in most locations). An investor can generally buy and sell for any length of time they want. An investor gains if the stock they own increases in price or receives dividends, and, conversely, loses if the stock they own falls in price.

There has been significant research in this area including exploring deep learning techniques (see Selvin et al.), reinforcement learning (see Lee), over long-term (see see Milosevic) and short-term horizons (see Khare et al.). We aim to add to this body of research by exploring short-term forecasts using widely known machine learning approaches that range from the simple thru to complex versus a simple heuristic, to test if machine learning provides any value and if additional complexity is warranted.

Problem Statement

Predict 1- and 5-day (trading) price movements (hereafter referred to as "return") with a low margin of error using readily available historical price data only (constructed into features where there is predictive power) and machine learning techniques (linear regression and neural networks). The prediction error will be compared to the error from simply assuming prices stay flat (the benchmark model), which is a simple heuristic often used by financial institutions for estimating block trades (the sale of big parcels of equity for a client) less a small percentage for the risk of a detrimental price movement and to compensate the institution.

The intention is to see if machine learning offers any improvement to simple heuristics.

Datasets and Inputs

The primary dataset is a single stock daily price dataset downloaded from Kaggle for 501 New York Stock Exchange equities between 2010 and 2016. A supplementary dataset provided includes reference data including the company name, industry and date first listed. This Kaggle dataset has been used rather than a regular industry source of data (such as Bloomberg) because it is free, has been adjusted for splits, and has been judged of fair quality. Alternative data sources such as Google Finance and Yahoo Finance have become more difficult to use due to API changes in 2017.

The primary and supplementary datasets are available here: https://www.kaggle.com/dgawlik/nyse/data. The primary dataset is prices-split-adjusted.csv; the supplementary dataset is securities.csv.

The intention is to only utilise the primary dataset for the prediction and any features constructed from this dataset (such as prior 5/30/180 day returns). Note that it only adjusts for splits and not dividends. This a small issue which can not be adjusted at this point of time but noted as a future improvement.

The return for a time period is calculated as:

$$Return = \frac{P_{t+n} - P_t}{P_t}$$

Where:

'P' is the closing price on day t (from the primary dataset) ‘t’ is the time at which the return is being measured from

‘n’ is 1 trading day or 5 trading days depending on the return metric

Solution Statement

We propose utilising the primary dataset to train and test two models for predicting the 1 day and 5 day return using only the prices in the dataset (as a side aim) and engineered features from the stock prices that have predictive value. We will estimate a linear regression model, which can be seen as the simple solution to the problem, and a neural network model, which can be seen as the more complicated solution to the problem. We will leave the form of the neural network open until we explore the data but we aim to test a long short-term memory (LSTM) model as this type has received quite a bit of interest recently.

Benchmark Model

The benchmark model we have chosen is a simple heuristic often employed by financial traders when acting as agents for block trades which is to assume stock prices do not change for the 1 day and 5 day trading windows.

Financial traders often then reduce this price by a discount to account for potential downside moves and to compensate for the risk and work done for the client.

Evaluation Metrics

We will measure the performance of each model using the root mean squared percentage error between the stock's actual and predicted close price for the 1 day and 5 day windows separately. This will be averaged across all the stocks for each window, each model, and for the benchmark. A lower result is better.

Project Design

We intend to do the following:

Explore the data, perform basic statistics and look for any data issues.

Calculate the return metrics and engineer different features (expected to include 5/30/180 day prior returns etc). We may explore utilising the industry data point in the supplementary data to look for cross industry features. We will also normalise and or scale the features using methods such as MinMaxScalar.

Train and test each model on the return metrics and the engineered features, and assess which model is best using the RMSE evaluation metric described above. We may at this point elect to explore alternative models. To evaluate both models we will utilise the last 1 day trading period in the primary dataset for the 1 day return prediction, and the last 5 day trading period in the primary dataset for the 5 day return prediction. To train and test both models we will initially split the remaining data up to the test on the last 1 day or 5 day trading period before the respective evaluation trading period data separated out above. If we decide longer term feature variables are not of value then we will split the entire dataset into multiple non-overlapping windows of training and testing data (possible if longer term features are removed). Given this is times series data we will not be randomly assigning or shuffling data points for training and testing.

We aim to do this using minimal training cost or time (less than 4 hours, likely to be much less anyway).

Bibliography

K. Khare, O. Darekar, P. Gupta and V. Z. Attar, "Short term stock price prediction using deep learning," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 482-486. doi: 10.1109/RTEICT.2017.8256643

Jae Won Lee, "Stock price prediction using reinforcement learning," ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No.01TH8570), Pusan, South Korea, 2001, pp. 690-695 vol.1. doi: 10.1109/ISIE.2001.931880

Milosevic, Nikola. (2016). Equity forecast: Predicting long term stock price movement using machine learning.

Selvin, Sreelekshmy & R, Vinayakumar & Gopalakrishnan, E. A. & Menon, Vijay & Kp, Soman. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 1643-1647. 10.1109/ICACCI.2017.8126078.