Machine Learning Engineer Nanodegree

Capstone Proposal

Ömer Faruk BÜLBÜL
March 2nd, 2019

Investment and Trading Capstone Project

Stock Price Indicator

Domain Background

Investment firms, hedge funds and even individuals have been using financial models to better understand market behaviour and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

Since finance area dynamics are very complex and stock prices depend on so many factors, machine learning can be a good solution if a suitable subset of these factors is selected and a reasonable performance is targeted. In this project I will try to investigate prediction success of close price of stocks given 5 or 10 days previous stock market data with a trained machine learning algorithm which is trained with a dataset set described below.

There are many successful implementations to make an estimator for stock prices one of which creates a framework with neural networks and decision forests.[1] In their "A machine learning based stock trading framework using technical and economic analysis" work they have managed to beat S&P500 Index by far according to charts provided.

Problem Statement

The problem is predicting the actual value of adjusted close price of a requested stock for a requested day with 5-10 days of previous data.

For this project, I will try to build a stock price predictor that takes daily trading data over 5 or 10 days as input of requested stock, and outputs projected estimates for given query dates. The predictor given open price(open), highest price(highest), volume, adjusted close price(adjusted close) will only predict Adjusted Close price for a given stock. The predicted value can easily be compared and a benchmark can easily be set up with actual adjusted price of the requested date for various dates upon request.

Datasets and Inputs

I will use 13 different stocks' historical stock prices saved to data folder of this project. Each file is in the csv format and containing approximately 9 years of historical stock market data. Data saved are from Yahoo Finance website and open to public use. The symobols of the stocks in alphabetical order is: AAPL, AMZN, AVGO, CSCO, MA, MSFT, NVDA, NVS, PFE, QCOM, TXN, V, WNT. Historical Data of stock prices can be easily gathered using https://pypi.org/project/yahoo-finance/ project.

In the csv files cloumns are: Date, Open, High, Low, Close, Adjusted Close, Volume. The date column will be necessary for retrieving 5 or 10 previous days of requested date adjusted close price prediction.

If close, adjusted close price of a stock is greater than open price it clearly shows a bullish market rather than a bearish one.

So relations between open price and close price will be very important for predicting the requested days adjusted close price. Also volume is very important since volume reflects the intensity (strength) of a stock. Volume also provides an indication of the quality of a price trend and the liquidity of a stock.

Highest price and lowest price values when compared to open price and close price are used to interpret trend of the stock by financers.

A trend analysis can be done with volume and price like in the below table [2].

General Rules in Volume Analysis:

Volume	Price	Interpretation
Increasing	Rising	bullish
Decreasing	Falling	bullish
Increasing	Falling	bearish
Decreasing	Rising	bearish

Here is an example of data structure:

Date	Open	High	Low	Close	Adj Close	Volume
2019-01-28	155.789993	156.330002	153.660004	156.300003	155.632523	26192100
2019-01-29	156.250000	158.130005	154.110001	154.679993	154.019440	41587200
2019-01-30	163.250000	166.149994	160.229996	165.250000	164.544296	61109800

Solution Statement

Since this prediction is a regression problem we can define supervised machine learning algorithms to predict adjusted close price at the end of the day. We can use supervised machine learning techniques to gather trend information of previous 5-10 days and use that trend to predict next day's adjusted close price. Once the predictor is trained it can be used to predict for various requested dates.

I am planning to implement a bunch of supervised machine learning algorithms like SVM and LSTM and compare success of the models with the evaluation metrics RMSE as specified below.

Benchmark Model

We can use Linear Regression algorithm as a benchmark for our solutions to compare with. This simple machine learning model could be used to determine whether we achieved a better estimation or not.

Evaluation Metrics

I would like to use RMSE(root mean squared error) for evaluation metrics.

$$\sqrt{\frac{\sum_{i=1}^{n}(Predicted - Actual)^2}{N}}$$

In RMSE since the errors are squared before averaging, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable. That is a very suitable property for a stock price indicator when real data is used as a benchmark.

Project Design

I am planning to follow the below worksteps in order to achive a good stock price indicator:

1- Domain Research:

I am planning to search for successful financial models which are frequently used by investors. For example moving average, bollinger bands, rsi etc. Also these models have some parameters to work with. For example moving average can be calculated for different timesacles where investors look for short, mid or long term insvestments. Since we are dealing with only one days' adjusted close, 5-10 days of moving average would be more suitable. Both of them can be used alternatively to investigate which one is better. The moving average of 5-10 days will be added to data as an additional column. Here we can use volume for weight parameter for a weighted moving average where higher volume will have a higher weight in moving average.

$$\mathrm{WMA}_M = \frac{np_M + (n-1)p_{M-1} + \cdots + 2p_{(M-n+2)} + p_{(M-n+1)}}{n + (n-1) + \cdots + 2 + 1}$$

Also bollinger bands adds standard deviation to the moving average information for which can contribute to a potential predictor, estimator. The purpose of Bollinger Bands is to provide a relative definition of high and low prices of a market. By

definition, prices are high at the upper band and low at the lower band. This definition can aid in rigorous pattern recognition and is useful in comparing price action to the action of indicators to arrive at systematic trading decisions. I am planning to add this information as an additional column to historical data.

BB = (MA(n) +- Kσ) where first term is Moving Average of n days, k is a weight constant for σ standard deviation.

 [3]

RSI(Relative Strength Index) is also another technical indicator which can contribute pattern recognition. The relative strength index was developed by J. Welles Wilder and published in a 1978 book, New Concepts in Technical Trading Systems, and in Commodities magazine (now Futures magazine) in the June 1978 issue.[4] It has become one of the most popular oscillator indices.[5]

2- Data Exploration and Normalization

Historical data retrieved from Yahoo Finance can have some problems, discontinueties etc. So I am planning to examine if there are null or enexpected values inside csv files. We should either compensate these problems by providing alternatives like manually fixing the data or I should eliminate lines having problematic data.

After fixing the abnormal values within the data, we can normalize it in order not to have weighted solution in terms volume and price of different stocks. Our aim is to find a predictor working equally for all of the stocks.

3- Constructing the Model

I would like to try SVM and LSTM to data alternatives constructed with the work flow step 1 and normalized at step 2. The principle I must follow is not to test the trained data. I will try to evaluate the performance by RMSE and choose best performing method to serve ui. I am planning to apply %80-%20 train-test ratio.

4- Building UI for user experience

I am planning to provide a minimal user interface for selecting stock and date or several dates for user to request easily. Maybe I could provide some successful results I came accross during the model construction phase.

5- Readme

Finally, I am planning to finish readm.md file for the github repository.

References

[1] A machine learning based stock trading framework using technical and economic analysis http://cs229.stanford.edu/proj2017/final-reports/5234854.pdf

[2] Trading volume: What it reveals about the market from https://www.rediff.com/money/special/trading-volume-what-it-reveals-about-the-market/20090703.htm

[3] Bollinger Bands: http://www.wikizeroo.net/index.php?q=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvQm9sbGluZ2VyX0JhbmRz

[4] RSI: J. Welles Wilder, New Concepts in Technical Trading Systems, ISBN 0-89459-027-8

[5] John J. Murphy (2009). The Visual Investor: How to Spot Market Trends (2nd ed.). John Wiley and Sons. p. 100. ISBN 9780470382059.