Stakeholder report

Problem formulation

Identifying a way to predict the stock market return series has been on the agenda of academicians and practitioners for quite a while. Due to this we decided to make an attempt into forecasting the future value of stocks using Random Forest, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM).

Data source

The dataset consists of information obtained over the span of six years, 2010 to 2016, concerning the stock price of around 500 companies. The data was split into four csv files containing the stock prices, general description of each company and metrics extracted from annual SEC 10K fillings (2012-2016)

New York Stock Exchange - S&P 500 companies historical prices with fundamental data

https://www.kaggle.com/dgawlik/nyse

Stock market prediction

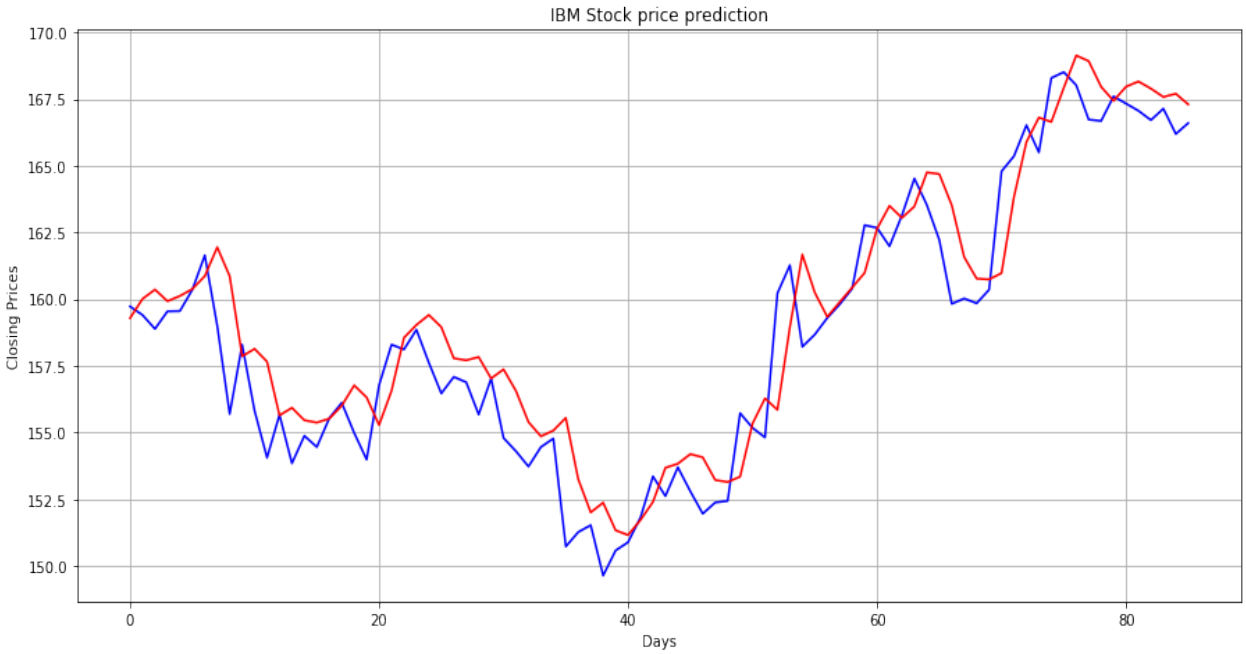
One of the main theories that needs to be taken into consideration when discussing this topic is The Efficient-Market Hypothesis (EMH). Based on this theory, efficient markets “rule out the possibility of trading systems based only on, the assumed information set, that have expected profits or returns in excess of equilibrium profits or returns” (Fama, 1970, p. 385) This means that the current price of the stock fully reflects the current available information about the company. (Oztekina, et al., 2016)

Another assumption for stock prediction is the random walk model which assumes that successive price changes are independent and identically distributed. This is the weakest form if EMH due to the fact that the information set is the past return history of the stock. (Oztekina, et al., 2016)

There are many factors that affect the stock prices such as: company news and performance, industry performance, investor sentiment, social media sentiment and economic factors. (Kusuma, et al., 2019) Due to this there are multiple types of analysis that can be conducted in order to be able to make a forecast. As mentioned before the dataset that we are using is split into four csv files. One of them is fundamentals.csv which concerns itself with analysing the company itself and it's credibility on the market. If we based our analysis on this, we would be looking into identifying whether or not the stocks of a company are over- or undervalued. The purpose of the analysis would be to develop a long-term strategy for the company which would require more time and it is not in our interests at the moment.

For this particular case we will be focusing solely on the prices of the stocks, meaning that we seek to determine the future prices of the stocks based on their previous values. Based on the information we gained from previous projects we have found that scholars mainly use Artificial Neural Networks and in particular time Recurrent Neural Networks in order to predict the stock fluctuations.

Because the dataset contained over 500 companies, we decided to limit ourselves to one, which is IBM. Our main focus will be on the closing price, which is the final price at which the stock traded during regular market hours. We want to see if the models used are capable of predicting closing price for a given time frame.



The red line represents the predicted closing price while the blue line is the actual price. The MSE is 0,00036 which means that on average the predicted line deviates over 0,00036% from the actual values. This would mean that the predictions are fairly accurate, but we are also dealing with a case of overfitting. Because it is based solely on the closing price and we are not feeding the model more information such as the fluctuations that occur during the regular market hour it is difficult to determine whether or not the predictions have a solid base or not. The accuracy of the model is over 80% but this does not mean that a person should base their decision for buying or selling stocks by using these results.

Future research

Because we are basing our analysis solely on the past return history of the stock, we are disregarding the economic environment as well as information on the company and its reputation which are factors that have great influence on the fluctuations of the stock. For future attempts we will try using more variables and perhaps combine it with a sentiment analysis of the company(perhaps Twitter dataset) in order to have a better image on how the stock values might change over time.