CAPSTONE PROJECT

Machine Learning Nanodegree

Edgar Velázquez

October 7th, 2017

Definition

Project Overview

In the 1600s, one of the first publicly recognized traded company was the East India Company or East India Trading Company. Investors financed or bought shares of ships for the exploration of new corners of the world searching for gold and treasures in exchange for a percentage of the values obtained. This represented high risk investments for investors because there was a 33% chance of the ships being seized by pirates and sometimes some of them never even returned or got lost at sea. This then resulted in the loss of shares and thus a loss of money for the investors.

Nowadays, although technologies and interests have changed from buying shares of ships to buying stocks and company values or debts, the risks are still similar. The difference from then to now is that back then there was not the same access to information nor the same amount of data about a certain topic or endeavor that we have today. We live in the age of information. Data is everywhere and valuable information can be gathered about companies that can give some insights about the future of such companies and thus the values of their shares.

Hedge funds, investment firms, and data analysts have been using the access to data to their advantage for years by reading the news, studying company history and trends in order to predict the value of stock shares of a company. The first attempt to use a financial model to predict stock prices was made by two statisticians named Box and Jenkins in 1976 using quantitative and statistical analysis.

I was inspired by a video on youtube by Siraj Ravel where he talks about using news headlines from the New York Times to predict the direction of the stock market (http://bit.ly/2yRNj01). I was inspired to collect my own version of the headlines dataset from the New York Times Developer API to work on this project.

Problem Statement

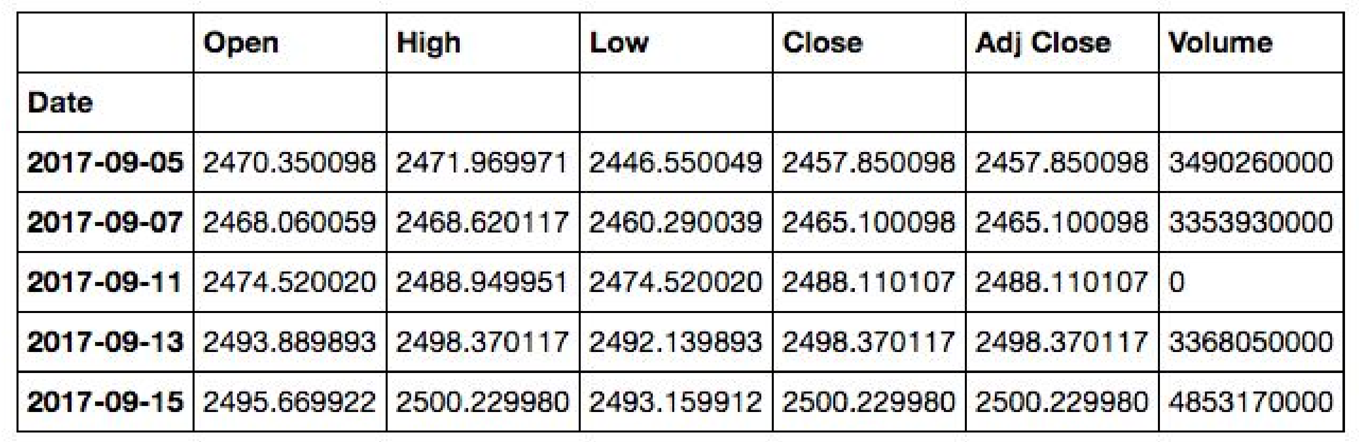
The challenge is to accurately forecast the random future price of a stock in a period of days using current time series, and the sentiment analysis of current news headlines. For instance, the latest time series displaying the volatility percent, volume, adjusted close, and percent change will be used to forecast the next 1 to 10 days of future prices of the stock. As part of the features of the data frame, the result of the sentiment analysis such as compound, negative percentage, neutral and positive percentage of all the headlines of each date in the data time series will be included. The source of the news will be New York Times and the most relevant headlines in related categories such as politics, business, economy, tech, etc.

Datasets and inputs

In this example the chosen stock will be S&P 500 index but the models are going to be tested against other stocks such as Google and Apple stocks as well. The S&P 500 is composed of 500 large-cap U.S based companies stocks which market cap is $5.3 billion or more. More than 10 years of data is collected for this process starting from January, 01 2007 to September 15 2017.

The following dataset is the one that is going to be used during this process.

- Open: The initial price of the stock at the beginning of the day.
- High: The peak or highest value reached by the stock in the day.
- Low: The lowest price reached by the stock during the day.
- Close: The final price of the stock at the end of the trading day.
- Adjusted Close: The final price with the adjusted stock splits.
- Volume: The amount of stock traded in the day.

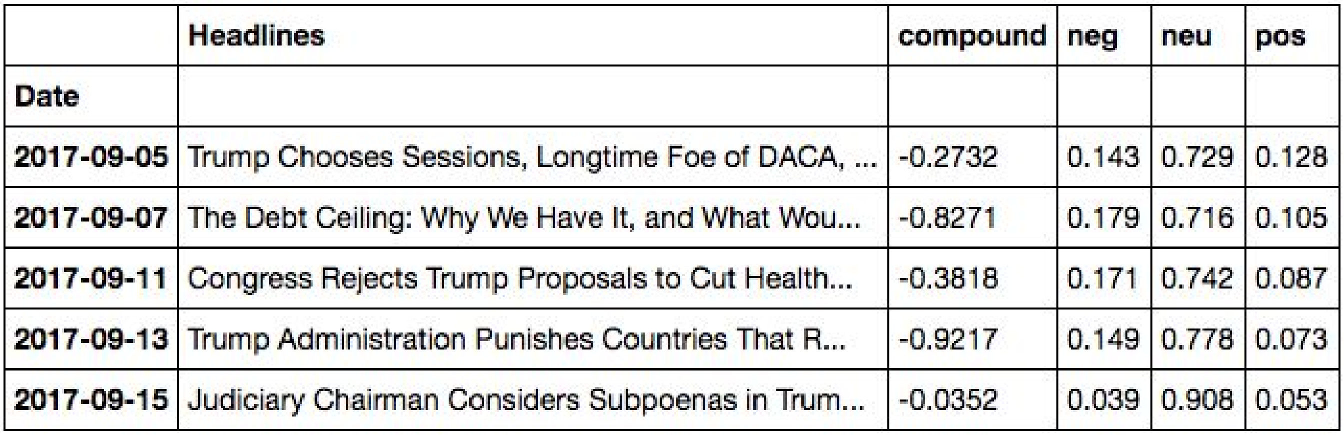


	Open	High	Low	Close	Adj Close	Volume
Date						
2017-09-05	2470.350098	2471.969971	2446.550049	2457.850098	2457.850098	3490260000
2017-09-07	2468.060059	2468.620117	2460.290039	2465.100098	2465.100098	3353930000
2017-09-11	2474.520020	2488.949951	2474.520020	2488.110107	2488.110107	0
2017-09-13	2493.889893	2498.370117	2492.139893	2498.370117	2498.370117	3368050000
2017-09-15	2495.669922	2500.229980	2493.159912	2500.229980	2500.229980	4853170000

This time series dataset was obtained through the Yahoo Finance API (https://finance.yahoo.com/).

The following data frame was my own personal collection and the current result was acquired using the New York Times Developers API (https://developer.nytimes.com/). The compound, neg, neu, and pos was obtained using the SentimentIntensityAnalyzer tool from the Natural Language Toolkit (http://www.nltk.org).

According to the Natural Language Toolkit documentation, this tool returns the number of positive, negative and neutral words and a percentage of the overall, neutrality, positivity, and negativity of the whole paragraph.



	Headlines	compound	neg	neu	pos
Date					
2017-09-05	Trump Chooses Sessions, Longtime Foe of DACA, ...	-0.2732	0.143	0.729	0.128
2017-09-07	The Debt Ceiling: Why We Have It, and What Wou...	-0.8271	0.179	0.716	0.105
2017-09-11	Congress Rejects Trump Proposals to Cut Health...	-0.3818	0.171	0.742	0.087
2017-09-13	Trump Administration Punishes Countries That R...	-0.9217	0.149	0.778	0.073
2017-09-15	Judiciary Chairman Considers Subpoenas in Trum...	-0.0352	0.039	0.908	0.053

Solution Statement

The seemingly random nature of the stock market constitutes the usual problem at the time of predicting a stock price, but experts in the area use their statistical knowledge and the news comprehension to understand future stock trends. The idea is to mimic what the experts are doing using machine learning trying to best predict and beat the experts prediction by a smaller margin of error.

In this project we will build machine learning models using quantitative statistical analysis, data time series and we will include major headlines of the New York Times for each date of the times series to test if we can predict the price of a specific stock more accurately. As part of the statistical technical analysis we will resort to feature engineering and get the Adj Close rolling mean and Bollinger bands in an N days window where N is the number of days in the future we are trying to predict.

The goal is to get the lowest margin of error, the difference, and the percentage difference between the predicted price and the actual future price from different machine learning models predictions. Next, find the mean of all predictions to see if the average result is closer to the actual result than each model's individual result. Also, to repeat this process with more than just one stock to further strengthen the quality of the results.

Benchmark model

As for the chosen benchmark model, the Dr. Eric Brown (http://ericbrown.com/) created a very compelling analysis using the S&P 500 index trying to predict the future price of the stock as well. In his analysis, he used both R^2 and the MAE as evaluation metrics getting 0.91 with the R^2 metric and 36.16 % in the MAE metric. The analysis can be found in http://pythondata.com/stock-market-forecasting-with-prophet.

Evaluation Metrics

To test the results of the model predictions the difference between the actual future price and the model's predicted price are going to be used to get the percentage of difference. The percentage of difference need to be lower than 20% in order to assume that the stock price can be safely predicted with this technique.

Actual Price: $100.00

Model's Predicted Price: $90.00

Difference: 100/90 = $10.00

Difference %: Actual Price / Difference = 10%

The percentage of difference will be applied to every prediction of every single day in the future and then get the average of all the percentage of differences. To further confirm the accuracy of the results, the Mean Absolute error regression loss will be included as well to measure the quality of the prediction.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|prediciton_i - real_i|$$

The Mean Absolute error is the outcome of the sum of all predicted prices minus the real prices for each day divided by the total number of stock prices. This method is commonly used in regression models to tell the margin of error between the real price and the predicted price.

The last Evaluation metric is going to be R^2 (R squared) coefficient of determination to give a percentage of accuracy to our results.

$$R^2 = \frac{SSR}{SST} = \frac{\sum \hat{y}_i^{\,2}}{\sum y_i^{\,2}}$$

The use of all of these evaluation methods will result in very solid and accurate measurable outcome of the predicted values.

Project Design

To start the process, the S&P 500 time series dataset of over 10 years of length from January 2007 to September 2017 will be used. Using feature engineering, from the Adj Close and the High columns the percentage of volatility will be calculated to create a new feature for the dataset.

*Percentage of Volatility = (High - Adj Close) / Adj Close * 100*

Also, another feature is going to be added using a similar formula. The percentage of change, which is calculated by using the Adj Close and Open columns.

Percentage of Change = (High - Open) / Open 100*

These newly engineered features and the columns with the most relevant related features (Adj Close, Volume, Percentage of Volatility, Percentage of Change), will be used to start the process.

To further complete the dataset and generate a label, the Adj Close will be shifted N numbers of days to the past so that this column becomes a new feature called Future Price. To this, as we mentioned before, we are going to add the Bollinger Bands and the Rolling mean as new features calculated from the current Adj Close in an N days window. The new columns after the mentioned calculations are Rolling mean, Upper Band, and Lower Band that are added to the existing columns (Adj Close, Volume, Percentage of Volatility, Percentage of Change, Rolling mean, Upper Band, Lower Band). The resulting dataset is what we are going to feed to our models first.

Rolling Mean = N days moving average

Upper Band = N days moving average + (N Days standard deviation of price x 2)

Lower Band = N days moving average - (N Days standard deviation of price x 2)

The process of generating trained machine learning models will be divided in three steps. First, before the sentiment analysis of the headlines is used, the dataset without the headlines analysis will to serve to pre train and test the chosen machine learning algorithms. The result of this first try will then work as a guideline to then attempt the second try which consists of using just the Adj Close and the sentiment analysis of the headlines as features and the Future Price as label. The columns of the dataset then will be (Adj Close, compound, neg, neu, pos, Future Price). The third step will be the combination of the first two steps and all mentioned columns will be implemented together as such (Adj Close, compound, neg, neu, pos, Volume, Percentage of Volatility, Rolling mean, Upper Band, Lower Band, Percentage of Change, Future Price). The result of the third step will then be compared with the results of the previous steps to observe which of the steps is the most accurate and thus selecting the best of the three steps to predict future stock prices.

The algorithms considered for the purpose of this projects are the Linear Regression algorithm, the K Nearest Neighbor algorithm, the Random Forest algorithm, the Support Vector Machine algorithm, and a Deep Neural Networks algorithm. After all algorithms have been trained and all results of all the predictions have been compared, the mean of the sum of all predictions are going to be considered as the final result.

References

- https://bebusinessed.com/history/history-of-the-stock-market/
- https://link.springer.com/article/10.1007/BF00167127
- https://www.fool.com/knowledge-center/what-is-the-sp-500.aspx
- http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- https://www.kaggle.com/wiki/MeanAbsoluteError
- https://developer.nytimes.com/
- https://finance.yahoo.com/
- http://www.nltk.org/
- https://www.youtube.com/watch?v=JuLCL3wCEAk&t=66s
- https://www.udacity.com/course/machine-learning-for-trading--ud501
- https://pythonprogramming.net/regression-introduction-machine-learning-tutorial/
- https://www.linkedin.com/pulse/python-tutorial-bollinger-bands-andrew-hamlet/
- http://pythondata.com/stock-market-forecasting-with-prophet/