Stock Predictions

Dinesh Daultani

Illinois State University

Author Note

Include any grant/funding information and a complete correspondence address.

Abstract

Stock prices fluctuate rapidly with the change in world market economy. There are many techniques to predict the stock price variations, but in this project, New York Times’ news articles is used to predict the change in stock prices. We are using NY Times Archive API to gather the news website articles data over the span of 10 years. Sentiment analysis of the headlines are going to be performed and then the output of the sentiment analysis is going to be fed into machine learning models to predict the price of DJIA stock indices.

Keywords: DJIA – Dow Jones Industrial Average, NLTK – Natural Language Toolkit, MLP – Multi Layer Perceptron

# Introduction

According to EMH (Efficient Market Hypothesis) the stock prices are based on new information such as news rather than present and past prices of stocks. Hence, we will be using news articles to predict the change in stock indices rather than predicting the prices by historical stock prices.

# Data Gathering

Two types of data gathered for this project:

1. **Stock indices:** As in general, researchers predict stock prices of composite index instead of predicting individual company’s stock prices. Because those composite index prices reflect the overall change in the stock market. We have used DJIA stock indices to predict the overall change in US top companies.

2. **NY Times API's:** News article data is collected from NY Times Archive API for last 10 years from 2007 to 2016.

# Data Processing

## Articles Filtering:

Articles collected from NY Times archive contains some other categories of articles, which are not related to stocks at all such as Biography, Obituary, and Schedule etc. Therefore, we have removed those kind of articles from the lists. In addition, news, which doesn’t affect stocks, are also filtered out from the most relevant articles. Articles sections that are kept are as follows: 'Business', 'National', 'World', 'U.S.' , 'Politics', 'Opinion', 'Tech', 'Science', 'Health' and 'Foreign'. Approximately 400k articles are filtered out after applying above filters out of 1M articles.

## Merge stock indices with articles:

After filtering out the relevant articles, single string was concatenated from all the articles headlines for each day. After getting single string for a day, it was merged with appropriate date and its DJIA stock indices.

## Pickling the processed data:

After preprocessing of the data, data frame is stored (pickled) to be read for further model training.

# Sentiment Analysis

NLTK package in python is most widely used for sentiment analysis for classifying emotions or behavior through natural language processing. Vader Sentiment Analyzer, which comes with NLTK package, is used to score single merged string for articles and get a positive, negative and neutral score for that string.

# Training models

Output of sentiment analysis is been fed to machine learning models to predict the stock prices of DJIA indices. We have trained various machine learning models based on Random Forest, Linear Regression and MLP Classifiers.

# Challenges

1. **Missing stock indices:** As the stock market is closed on weekends and US holidays, there is no open/close prices for any of the stocks on those days. Which affects our DJIA values as well. To fill the missing data for indices, we have used interpolation of the prices to fill in the missing values.
2. **Filtering of the news articles:** Some of the articles contained ‘none’ as articles section, for those articles all the articles are included for that day.
3. **High fluctuations in prices:** As the prices of the stocks fluctuate a lot, we have used technique called smoothing which is used in financial markets to take moving average of the values, which results in comparatively smooth curves.
4. **Price change during testing and training:** As the starting price predicted for testing does not match with the ending price and it rather matches with starting value of the training DJIA values. We have took average for 10 starting days for predicted and actual values for testing and then added the difference in predicted values.

# Conclusion

MLP and Random forest shows better results than logistic regression trained models. Although graphs generated does not shows satisfactory results, further optimization in the process could lead to better accuracy in the results.