Stock Index Prediction using Historical Prices and News Headlines

Introduction

A market enables the exchange of buying and selling of a plethora of assets and the stock market is the place for buying and selling company stocks. Every stock exchange has its own stock index value which is the average value that is calculated by combining multiple stocks. It is a representation of the entire stock market and helps to portray the market's general movement over time.

Stock markets are an important part of a country's economy and are often revered as an indicator for economic health. Predicting the stock trends, on the other hand, seems to be an elusive concept which has been gaining popularity particularly with the application of quantitative analysis and machine learning techniques. Being able to predict stock prices efficiently can help minimize the risk of loss and maximize profits, so a lot of research and effort have been focused in this area.

Problem Statement and Context

Stock prediction has always been a controversial subject where even the most important experts differ in opinion whether stock market can actually be predicted or since the markets are efficient where all new information is already incorporated in the market behavior that there is hardly any room for prediction. Efficient Market Hypothesis claims that this is the case where all stocks are trading at their fair price, the market reflects all information already and all participants to the market interpret the information available to them in the same manner, so it would be impossible to predict the market.

In this project, rather than trying to prove any of these hypotheses is right or wrong, I attempted to do two things: firstly, I wanted to see if there is indeed some relationship between stock index prices and their past values and that we can use historical prices to predict future prices; secondly, I tried to see if I can use news headlines to predict the increase or decrease in stock index prices. The reason I chose two different approaches was to basically demonstrate the actual index price prediction using time series analysis which is a regression problem and predict the increase/decrease in index prices using text analysis techniques which is a classification problem.

My objective in this analysis is going to be two-fold: firstly, I will use common time-series forecasting techniques to predict the DJIA stock index, and secondly, I will implement various text analysis techniques to predict price increase/decrease using the daily news headlines. I will provide general guidelines on how to implement each time series regression methods and compare their performances by looking at root mean squared errors in predicting the index price. Then I will build classification models by implementing a few different NLP techniques to predict the stock index price increase/decrease.

Criteria for Success and Scope of Solution Space

There are a few starting points of intention for this project. I am aware the accuracy of the analysis models will be moderately questionable since the complexity of the stock market behavior is abundantly evident in real life. Due to the stochastic nature of the markets, achieving high levels of accuracy is not expected. However, for learning purposes and being able to apply both time series concepts and text analysis work to price prediction made this project desirable enough to pursue.

I will be comparing results from different time-series based machine learning models and compare their root mean squared errors to see which one performs better. I have chosen to apply a handful of common methods here, but the solution space can be extended to include more. For the classification problem, I will be applying again a few common NLP methods and comparing their accuracies.

Stakeholders

There may be few stakeholders who could benefit from this analysis especially in the financial sector. Stock prediction efforts have made quite a bit of headway in recent years where experts and quants in Wall Street are trying to maximize profits by exploiting knowledge that can be gained from prediction modelling.

Constraints

Price prediction by using only time series information may not be sufficient to forecast trends in the market. Same thing goes for just using news headlines as well. Due to time constraint, I have chosen not to do any extended feature engineering to build more proper regression and/or classification models i.e adding and/or creating various technical indicators, etc. I have done some basic feature addition in the linear regression model to showcase the possibilities. I have also kept the approaches separate at this time. If time permits, I may try to do modeling based on the combination of price and news headlines data.

Datasets

The dataset I chose for this project is from Kaggle's Daily News for Stock Market Prediction webpage and it includes both price data of Dow Jones Index Average (DJIA) and top 25 daily news headlines with price increase/decrease as the label. The original stock index data had been obtained from Yahoo Finance by the author and contains the Dow Jones Industrial Average (DJIA) that has a range from 2008-08-08 to 2016-07-01 (roughly eight years) with 1989 observations and 7 column headers.

The two main datasets are the stock index data: DJIA_table.csv and the news headlines data: Combined*News*DJIA.csv. I will combine these two datasets in the data wrangling part of the project.

Project Approach and Methodology

As I mentioned above, I will be tackling two sets of problems in the project: one is a regression and the other one a classification problem. I realize there may be quite a few methods to accomplish this, but I have chosen to predict the stock index price by using different time series approaches and will compare them to each other in terms of error performance. On the other hand, I will be building classification models by just using text analysis (NLP) methods on daily news headlines data to predict the increase/decrease in the stock index. Here is my methodology in a nutshell:

1) Data Collection, Wrangling and Exploratory Analysis (EDA)
2) Data Preprocessing and Modeling:
    a) Two main approaches
        i) Time Series Analysis of Stock Prediction
            (1) Moving Average
            (2) Linear Regression
            (3) KNN and SVR
            (4) ARIMA
            (5) LSTM
            (6) Prophet (Facebook)
        ii) Stock Prediction using NLP with Stock News
            (1) Baseline Model: TF-IDF Vectorizer
            (2) Word2vec
            (3) Doc2vec
            (4) LSTM

Deliverables

Deliverables for the project will be posted at the GitHub repository and will include:

a) Code (notebooks) for:
    i) Data Load and Wrangling and EDA
    ii) Data Preprocessing and Modeling
b) Project final report
c) Project slide deck