**Final Project Report: A S&P 500 Index Prediction Model**

**Group members:**

Yining Chen

Mohan Venkat Krishna Savata

Anil Kumar Anumola

Bharadwaja Vemulapally

1. **Abstract**

Stocks consist of all the shares by which ownership of a corporation or company is divided. A single share of the stock means fractional ownership of the corporation in proportion to the total number of shares.[1] The stock market has always been a popular field for technical analysis, which forecaste the trend of prices through the study of past dataset due to its digital nature.[2] In this project, the linear regression algorithonm is used on the “S&P 500 index 1791–2013 dataset” to train a proper model, which will be used to make predictions on new market data.

1. **Introduction**

The Standard and Poor's 500, or simply the S&P 500, is a stock market index tracking the stock performance of 500 of the largest companies listed on stock exchanges in the United States. It is one of the most commonly followed indices and includes approximately 80% of the total market capitalization of U.S. companies.[3] S&P 500 is often used as a vane to predict the economic direction of the United States and the world. In statistics, linear regression is a statistical model which estimates the linear relationship between dependent and independent variables. In this project, the linear regression algorithonm will be used to train a proper model to predict new market data.

1. **Dataset**

In this project, our team chose the ‘US Stock Price Index over 1791–2013’ dataset, which comes from the ‘Kaggle AI & ML community’. The contents of the dataset are as shown in the Figure 1.

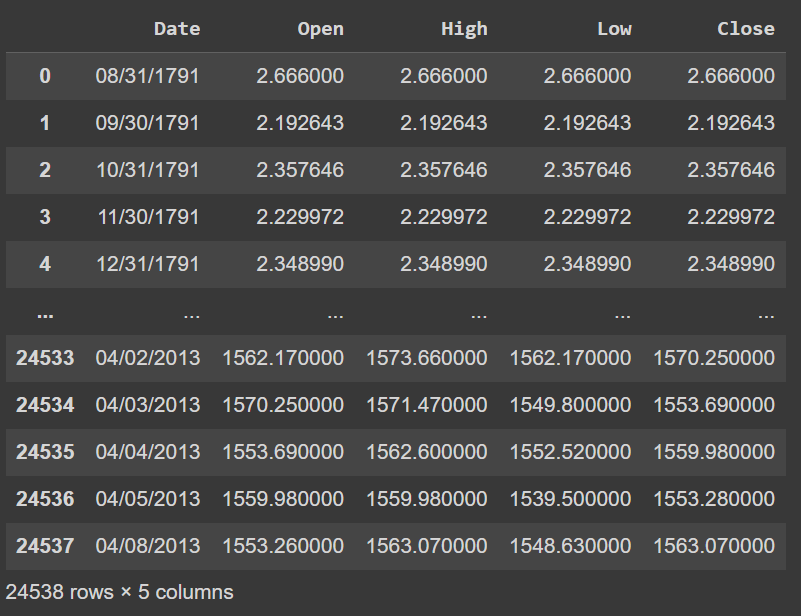


Figure 1, the ‘US Stock Price Index over 1791–2013’ dataset.

1. **Preprocessing**

First, add ten years, year and date as group index in order to facilitate access to data. The processed data set is shown in Figure 2.

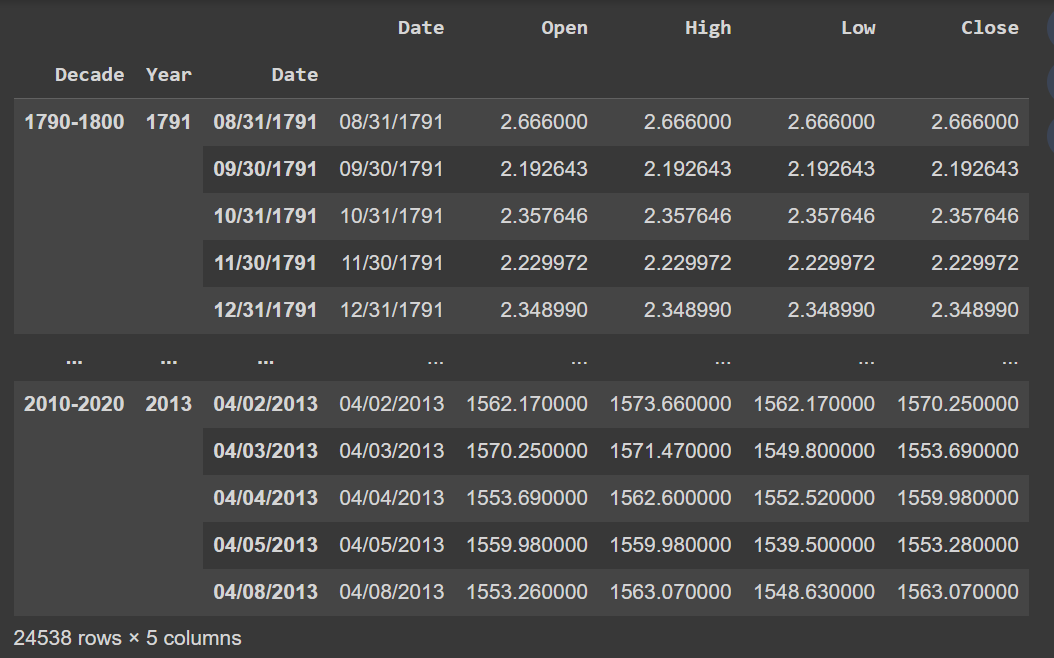


Figure 2, the dateset after adding group index.

Then convert “Date” column from string data type to numeric data type (julian date which is a continuous count of days in astronomy) to be calculated as a feature variable in linear regression. The processed data set is shown in Figure 3.

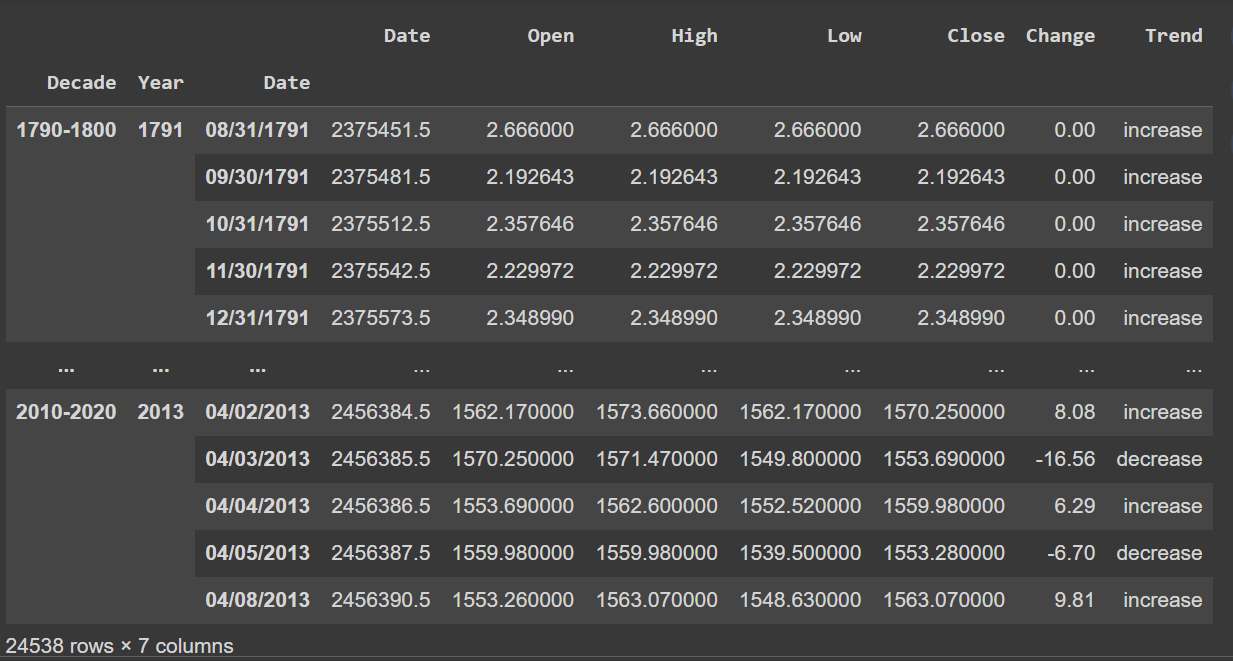


Figure 3, the dateset after converting data type.

1. **Feature Engineering**

We calculated the change between the ‘Open Price’ and ‘Close Price’, which is labeled by ‘Change’ and added to the dataset. We also converted the changes into ‘increase’ or ‘decrease’ to facilitate analysis, gave them a ‘Change’ label, and added them to the dataset. The processed data set is shown in Figure 4.

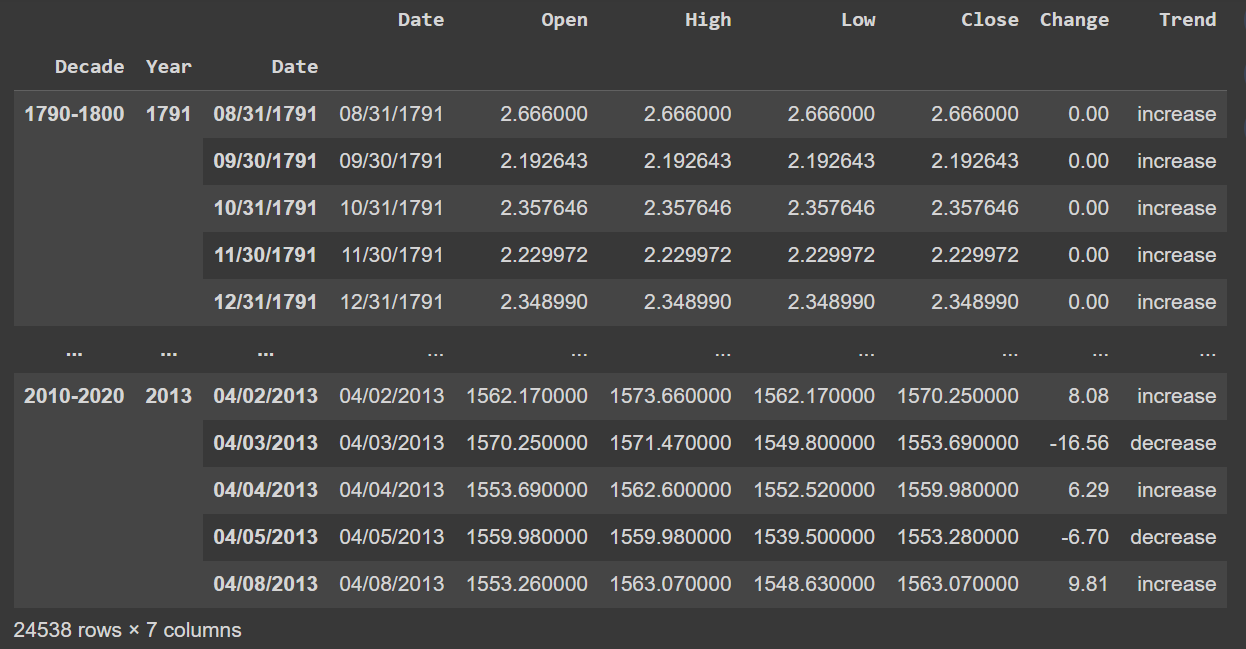


Figure 4, the dateset after feature engineering.

1. **Visualization**

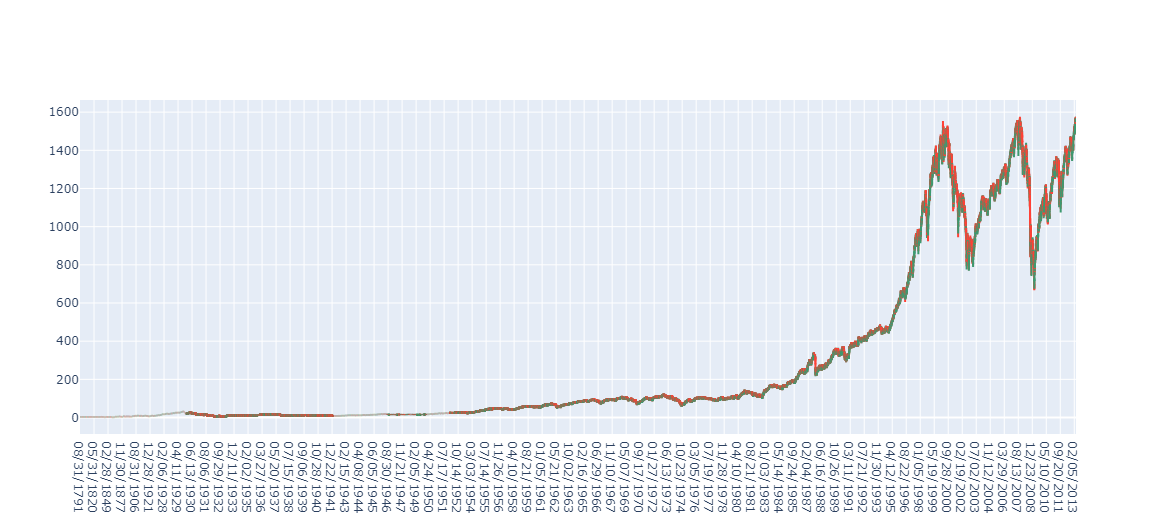
****

Figure 5, the general trends from 1791 to 2013.

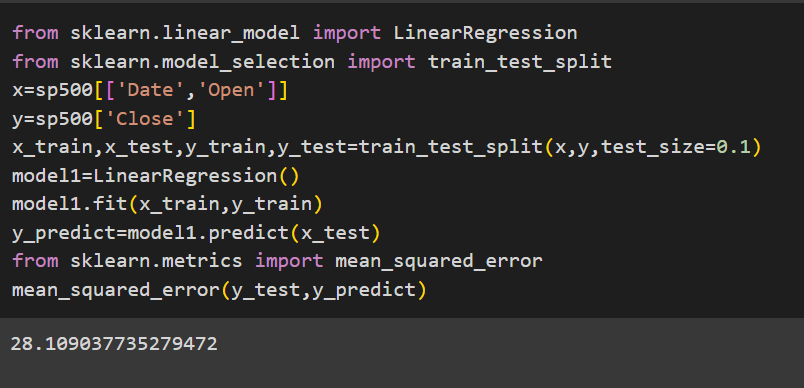


Figure 6, mean squared error of y\_test and y\_predict.

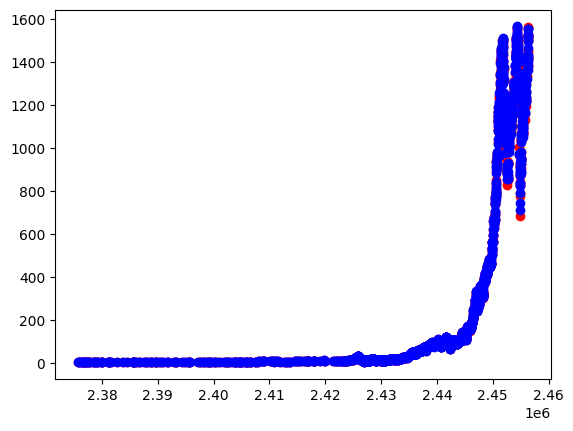


Figure 7, the scatter plot of y\_test (red) and y\_predict (blue).

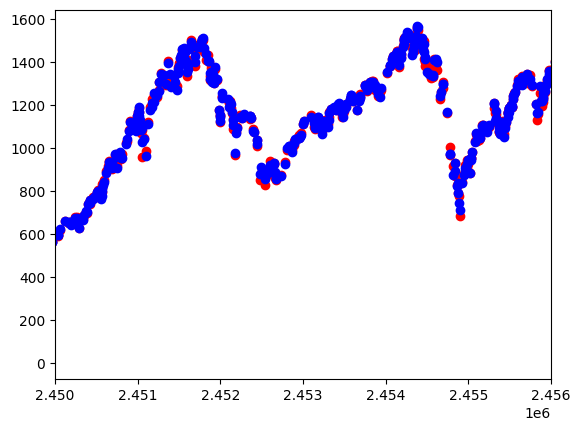


Figure 8, the scatter plot of y\_test (red) and y\_predict (blue) after zooming in.

1. **Results**

Figure 5 shows that the S&P 500 index has been relatively flat since 1791, with little fluctuation until it began to rise in 1980, and then began to experience dramatic changes in ups and downs in 2000. Therefore, analysis of the S&P 500 Index should focus on trends from 1980 to 2012.

Figure 6 shows steps to get mean squared error of y\_test and y\_predict. In statistics, the mean squared error (MSE) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. As it is derived from the square of Euclidean distance, it is always a positive value that decreases as the error approaches zero.

Figure 7 and 8 shows that the distribution of predicted values and true values. It can be seen from figures that this model has good prediction ability.

1. **Conclusion**

Since there are too many factors that affect the stock market, there is currently no perfect prediction model that can reduce MSE to a level close to 0. This model that only uses open price and date to predict close price can be used to assist in predicting the trend of the S&P 500 index.

1. **References**
2. Summers, Della (2007). Longman Business English Dictionary. Pearson Longman.
3. Kirkpatrick, Charles D.; Dahlquist, Julie R. (2006). Technical Analysis: The Complete Resource for Financial Market Technicians. Financial Times Press. Page 3
4. Valetkevitch, Caroline (May 6, 2013). "Key dates and milestones in the S&P 500's history".
5. **GitHub Repository for Code**

<https://github.com/YiningChen96/CS661-Final-Exam-Project>