**Predicting Stock Exchange Price**

Jamelah Guimba C0853807

Ricardo Chacón C0854802

Sulabh Shrestha C0841540

Lambton College

CBD 2214 – Big Data Fundamentals – Data Storage Networking

Kritika Buttan

March 28, 2022

**Introduction**

Stocks has become a very important piece to many people not at all for professional investors just as well for common people, and this happened due covid time because many people research how to get incoming doing less using stocks, cryptocurrencies, etc.

Opportunities have been in any place and the main idea of this project is the creation of a model to train it through AI, using known frameworks as well NumPy, Matplotlib, Scikit-learn, etc.

**Motivation**

We are developing a training model to predict the behavior of stocks; this will be useful for ordinary people who want to get involved in the investment market and take advantage of the learning of AI.

As we mentioned early due to the covid pandemic many companies fired people and gave them compensation, others made their decisions to get early retirement; this impact economy because everyone want to save money for the hardest time, and the recession took more time than expected.

Thus, it was an opportunity to reinvent themself and research how to invest their money. So, doing this data analysis and selling our product to consumers would give us an immense market opportunity to start it.

**Preparation**

**Data Procuring**

We procure the data from Kaggle. The dataset contains the price data indexes if Stock Exchange Index all over the world such as United States, China, Canada, and more. Based on Kaggle, the data was collected from Yahoo Finance collected from for a long period of time (Kaggle, 2021). There are three set of data files available, and we use the indexData file (see Figure 1) as it contains the valuable data for predicting stock price using machine learning.

Table

Description automatically generated with medium confidence

**Figure 1.** Sample data from Kaggle for Stocks Exchange Price. (own photo)

**Data Identifying**

We have identified NASDAQ Stock Exchange Price with Index name 'IXIC' (see Figure 2) for this program. However, we can also be used for identifying other Stock Index such as NYA (NYSE) and SSMI (Swiss Market Index) that are in the Kaggle dataset.

Table

Description automatically generated

**Figure 2.** NASDAQ Stock Index (own photo)

There are 12,690 total rows available for NASDAQ Stock Index to be use for our training later. There are eight columns available, but we will be identifying the proper column to be use for training.

**Data Cleaning**

We make sure to process only valuable data in our training. Thus, we remove the row data with null values (see Figure 3).

Graphical user interface, text, application, email

Description automatically generated

**Figure 3.** Removing Null Data (own photo)

We found out that there are no null values for the NASDAQ Stock Index.

**Data Imputation**

Before we train our data, we are converting the data column from string to actual datetime. This is because the encoding takes datetime as a datatype, but our date column is a string datatype. To fit in our model, we first need to convert string datatype into date datatype (see Figure 4).

Table

Description automatically generated

**Figure 4.** Converting Date Column from String to Date Datatype (own photo)

**Identify Features**

To train our data label, the X\_value is raw date value that will not be accepted so it will be manipulated to fit for machine learning algorithm. So, it will be our X value to train which will predict our close which is Y (close) (see Figure 5).



**Figure 5.** Converting Date column to an ordinal value (own photo)

After converting the date column, we will get the proper columns for our training later. We want to determine the Closing Price for NASDAQ; thus, we choose the Date and Close columns (see Figure 6). The Close column is the closing value for the equivalent date.

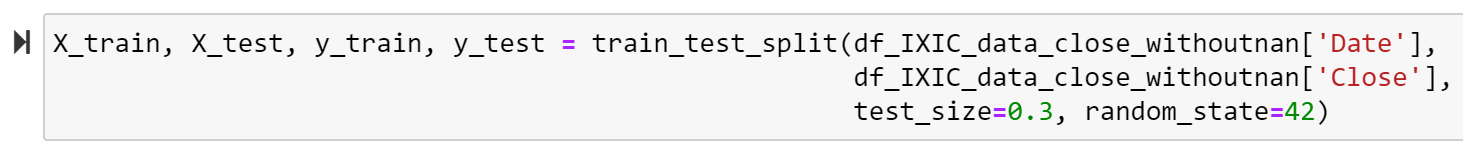
Table

Description automatically generated

**Figure 6.** Retrieve only Date and Close Column (own photo)

**Data Splitting**

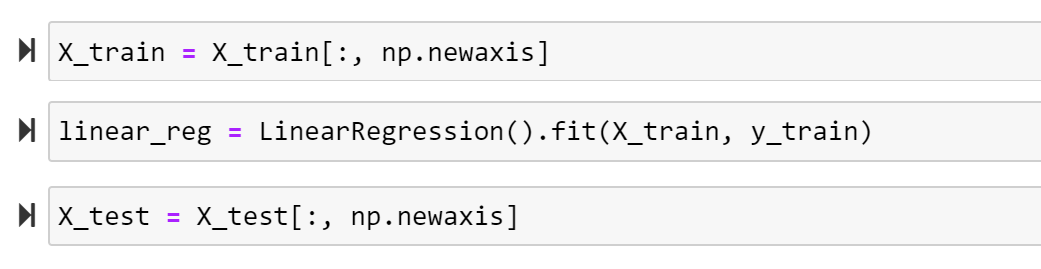
We are splitting the data into 70/30 ratio based on the size of our dataset (see Figure 7). Seventy percent of the data is for training set and thirty percent of the data is for testing set. We will first train our data using the seventy percent set of data and then we will use the thirty percent set of data for the testing to validate the accuracy of the resulting model. Based on some studies, best results are obtained from using twenty to thirty percent of data for testing.



**Figure 7.** Splitting the date in 70/30 ratio (own photo)

**Training**

We need to convert the Series into ND array for the algorithm to fit. The algorithm that we will using will be processing the data in array datatype (see Figure 8).



**Figure 8.** Converting the data from Series datatype to Array datatype (own photo)

**Linear Regression**

Linear regression is one of the prediction models which are simple but is proven to be dependable to predict the future. This model can be quickly trained and easily understood.

The advantages of Linear Regression are (Kumar, 2019)

* easy to implement and interpret
* less complex
* can be prone to over-fitting but can avoided reduction, regularization, and cross-validation techniques.

We follow the assumptions to consider for a successful linear-regression analysis:

* check for valid cases
* check for regression coefficient and mean absolute error
* plotting of data

Before using any ML algorithm, we will try to use our own algorithm here. So, this is how the data looks like as the Ground Truth (see Figure 9 and Figure 10).

A picture containing chart

Description automatically generated

**Figure 9.** Checking for Ground Truth (own photo)

Graphical user interface

Description automatically generated

**Figure 10.** Applying Linear Regression Algorithm (own photo)

We start creating the model for Linear Regression and determined the Coefficient of Determination (see Figure 11). The ideal value for Coefficient of Determination is one and the worst value is zero.

Graphical user interface

Description automatically generated with medium confidence

**Figure 11.** Coefficient of Determination for Linear Regression (own photo)

For the Mean Absolute Error, our prediction result is +/- 1,055.07 (see Figure 12). This is our margin of error. For example, if we have a 5,000 prediction the actual value will be from 3,944.93 to 5,055.07.

Text

Description automatically generated

**Figure 12.** Mean Absolute Error for Linear Regression (own photo)

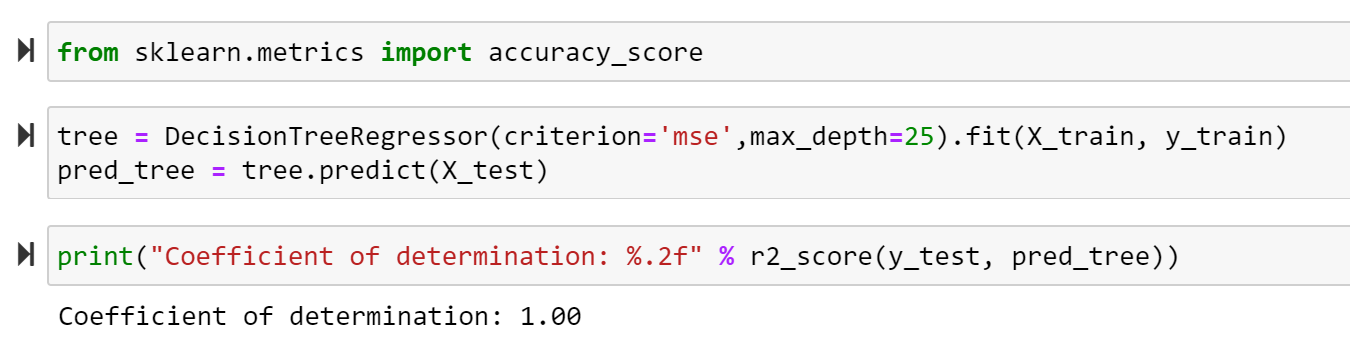
**Decision Tree**

Decision Tree is a regression model that transforms data into a tree representation. This is one of popular regression model as it requires less effort for data preparation and processing.

The advantages of Decision Tree are (Dhiraj, 2019):

* does not require normalization of data
* does not require scaling of data
* missing values in the data does not affect the process to a considerable extent
* intuitive and easy to explain

We start creating the model for Decision Tree Regression and determined the Coefficient of Determination (see Figure 13). The result of our Coefficient of Determination is one which is an exceptionally substantial value.



**Figure 13.** Coefficient of Determination for Decision Tree (own photo)

**Random Forest**

Random forest is a supervised learning algorithm that uses technique that can perform both regression and classification. This method uses multiple decision tree in parallel and randomly perform row and feature sampling forming sample dataset for every model. (Chakure, 2019).

Here are the advantages of Random Forest (Chakure, 2019):

* one of the most accurate machine learning algorithms
* can run efficiently on large database
* can manage thousands of input variables without variable deletion

We start creating the model for Decision Tree Regression and determined the Coefficient of Determination (see Figure 14). We got similar result as the Decision Tree where the Coefficient of Determination is one which is an exceptionally substantial value.

A picture containing text

Description automatically generated

**Figure 14.** Coefficient of Determination for Random Forest (own photo)

**Plotting the Predictions for Comparison**

We plot the original data (Y\_test which is the untouched data) first in blue dots.

Chart, scatter chart

Description automatically generated

**Figure 15.** Scatter Plot for the Y\_test (original data) (own photo)

The result of the Coefficient of Determination of our Decision Tree and Random Forest is one which is the ideal value. This can be seen from our scatter plot below (see Figure 16) where our result is almost identical to the original data.

Chart, scatter chart

Description automatically generated

**Figure 16.** Scatter Plot for the Original Data, Decision Tree, and Random Forest. (own photo)

**Plotting the Predictions for All Regression Algorithm**

We plot the original data of Y\_test which is the Closing Index Value and the X\_test which is the Date in blue line (see Figure 17).

Chart

Description automatically generated with low confidence

**Figure 17.** Scatter Plot for the Y\_test and X\_test. (own photo)

We can see from the graph (see Figure 18) that the Decision Tree and the Random Forest are almost overlapping which is the result of our training with Coefficient of Determination equal to 1.

Line chart

Description automatically generated with low confidence

**Figure 18.** Scatter Plot for the Three Regression Algorithm (own photo)

**Results**

We predict our own dataset by checking the Closing Index Value for the two random dates, 11/08/2011 and 04/23/2008 and compare the result with the actual Ground Truth.

Graphical user interface, text, application

Description automatically generated

**Figure 19.** Predicting for Specific Dates (own photo)

We can see from our result for 11/08/2011 (see Figure 19) that the value we predict which was 2,727.48999 is close to the actual value from our dataset below.

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 20.** Actual Data for November 2011 (own photo)

We can see from our result for 04/23/2008 (see Figure 19) that the value we predict which was 2,376.939941 is close to the actual value from our dataset below.

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 21.** Actual Data for April 2008 (own photo)

**Conclusion**

So, concluding our project, we brought the data into our system and then did all the data processing to make it ready for machine learning. Once the model is trained using machine learning models we compared to see which one is the best by using a random sample on one of the algorithms.

**References**

Kaggle (2021, June) *Stock Exchange Data.*

<https://www.kaggle.com/datasets/mattiuzc/stock-exchange-data>

Dhiraj K. (2019, May 26). *Top 5 advantages and disadvantages of Decision Tree Algorithm.*

<https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a#:~:text=Decision%20Tree%20solves%20the%20problem,both%20regression%20and%20classification%20problems>.

Naresh Kumar (2019, May 9). *Advantages and Disadvantages of Linear Regression in Machine Learning.*

<http://theprofessionalspoint.blogspot.com/2019/05/advantages-and-disadvantages-of-linear.html>

Afroz Chakure (2019, June 29). *Random Forest Regression.*

<https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>