Project Report

Ralph R. Christian

Usnik Biswas

Ajay Sharma

Lambton College in Toronto

CBD 2214 Big Data Fundamentals- Data Storage Networking

Kritika Buttan

Abstract

For the linear regressor assignment for the Big Data Fundamentals, we have selected Netflix Stock Price as our topic. A Linear Regressor is a supervised machine learning program that is used to find the best fit linear line between independent variables. In simple words, it finds linear relationship between dependent and independent variable.

# Dataset

Graphical user interface, application, Word

Description automatically generated

Any type of prediction or analysis depends on the dataset that is given to the model, the data needs to be selected and collected from proper sources, and the data needs to be of high veracity to have better value.

The data that we have selected is of the Netflix Stock prices starting from the date 05-02-2018. Each line of data has 6 columns, namely, Date, Open, High, Low, Close, Adjacent Close and Volume.

* + Open: Shows what value the stock opened on that particular day. The open gives more detail about the price of the stock at the time trading is open for general population to invest. This is the time when there is the most activity in any stock price.
  + High and Low shows the highest and the lowest value of the stock on that particular day.
  + Adjacent Close: The adjacent close price is the value of the stock that is affected after the trading stops for the day, this price change is without the interference of buyers or sellers.
  + Close: What value did the stock close on that day.
  + Volume: What is the amount of stock available in the market to purchase.

.

# Preprocessing Steps

Graphical user interface, application, Word

Description automatically generated

Before the regression or any other big data analytics process is done, there are a few steps which are necessary to ensure that the whole cycle completes successfully and yields good results that are high in value. These steps are called preprocessing steps and they include tasks like Data Procuring, Data Cleaning, and Imputation.

Few of the key features are:

* Data Procuring: Data procuring is the process of data collection or mining, during this process the relevant data which was selected for the analysis during the data selection process is collected through various sources like external and internal. This data can be purchased from third party vendors as well and can be collected from company’s own database as well.
* The results in the form of information and knowledge depends heavily on the type of data that is fed to the regressor and other analytics processes, hence cleaning of the data is also important. Data cleaning is the process of cleansing data and removing or solving any anomalies the are present in the data that can potentially skew the results.

# Column Selection

Graphical user interface, chart, line chart

Description automatically generated

The whole data set is not pushed into the regressor, that means, not all columns are essential to the prediction, which is why selecting a column is of utmost importance. Here we have selected the column “Close” to train the regressor and make predictions.

Close is defined as the final price of the day when trading stops for public, this column is selected because the price variation after the close is not because of buyers and sellers, hence making predictions based on that data will not be possible. Selecting the close column will result in proper results that are based on the buyers and sellers dealing with the stock. Here we are splitting the data 80% to 20% using the following line:



# Linear Regressor

Chart, bar chart, histogram

Description automatically generated

Here we have used a linear regression model to represent the relationship between input values that are plotted against two axes, those two axes are X and Y. The regression model that we are using is a simple model that presumes the output values of Y given that we have the linear combination of values of Y. The form of the equation used is that of a slope which is MX + C, where M and C are the coefficients. We are using the Linear Regress() module of Python.

After the regression is done and results are obtained, those results are passed through some parameters to determine their accuracy and value. Those parameters include:

Mean Absolute Error: It is the error observed when comparing the predicted value to the actual value which is observed. It is more advantageous to compare this way instead of using mean square error which makes things more complicated.

Mean Squared Error: It is the difference between square of predicted value and the square of observed value, this is usually and appropriation for more complex cases.

Root Mean Square Error: This provides the standard deviation of the regression, and the data line points; this is basically the quadratic mean of predicted value and observed value.

Our results for these parameters are:

Graphical user interface, application, Word

Description automatically generated

References

Linear regression | Introduction to linear regression for data science. (2021, May 25). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/>

Netflix stock price prediction. (n.d.). Kaggle: Your Machine Learning and Data Science Community. <https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction>

Scikit-learn. (n.d.). scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation. Retrieved March 29, 2022, from <https://scikit-learn.org/stable/>

What is a closing price? (n.d.). Investopedia. <https://www.investopedia.com/terms/c/closingprice.asp>