Prediction of Stock Price for Large-cap companies

Introduction:

This project aims on predicting the future price changes of a stock of a large cap company. This uses the previous prices and financial news related to that particular company. The data required is taken from Nasdaq.com. 5 years worth stock price data has been gathered for about 20 large cap companies.The datasets used contain data related to stock price, opening, closing values and the highest, lowest price that particular stock reached. Firstly, Exploratory Data Analysis is performed on the numerical dataset inorder clean up the data, and change all values to desirable datatype and to derive insights from the data. The insights from this numerical data has been plotted as graphs. Then the newsheadlines data is preprocessed, trained and labelled to apply Naive Bayes on the data. The insights derived from these two datasets help predict future stock price changes of that particular stock.

Exploratory Data Analysis on Stock price dataset

Any dataset that is to be used to perform analysis on, needs to be prepped before we conduct analysis on that.The data available might contain some discrepancies, and might not be consistent, to avoid all these affecting our analysis , Exploratory Data Analysis was performed on the numerical datasets (in csv format). In our EDA we have,

- Imported the data
- Imported all the required libraries
- Got to know about the data, Know the head, tail, get a sample from the data. Find the sum of all missing values. As there were no missing values in our data set, we moved on to change data into their desired data types.
- Plotted insights from the EDA.
- we have Open ,Close\Last,high ,low and volumn columns

Time Series Data:

The stock market data is time series data as in it changes by the time . A time series is a sequence of numerical data points taken at successive equally spaced points in time.In investing, a time series tracks the movement of stock price, over a specified period of time.

Univariate and Bivariate Analysis

- Univariate analysis has been performed to draw graphs on History of opening,closing ,high and low prices and on History of volume.
- Bivariate analysis has been done on Closing - Open, Closing - Low price, Closing - High prices.

Inference from univariate and bivariate analysis:

- Closing Price is never smaller than Low Price.
- Closing Price is always smaller than High Price.
- Closing Price is sometimes larger or sometimes smaller than the Opening Price.

We have performed Resampling(It involves changing the frequency of your time series observations.) and Zooming-in to the data set and Inference: We can say that profit is not predictable for a short amount of time (week,month). There is a chance of gain and loss at the same time if we buy stock for a short period of time.But for long term there is very less chance of getting loss and profit is not that much good. So we can say that it is neutral.

PreProcessing the data

Basically it means manipulating the data according to our use so for that we have use some technique:-

- Functions to calculate moving averages
- Used shift function

Creating the Model and validating it

1)As it is time series data we can't just split the data into train and test it is not applied for time series data insted of we have to take the data for a specific time using date so for that we have used iloc() function and specifies the range of the data into the arguments paranthesis train_x,train_y these two will have all the data basically it is training data and test_x,test_y these two are testing data . 2)We had created logistic regression model

- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- Logistic Regression is used when the dependent variable(target) is categorical.
- Logistic regression is easier to implement, interpret, and very efficient to train.
- It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).
- It is very fast at classifying unknown records.

To import the module:-

- from sklearn.linear_model import LogisticRegression
- After importing the file we have created the model object and then using fit function we are training the data to the model

3)Using text_x we had made a prediction and than compared with the actual value which is test_y for this the accuracy score is 98%

4)Using train_x we had made prediction and than comapred with the actual value which is train_y for this the accuracy score is 99%

5)As per the classification report the F1 for both 0 and 1 is 98% (Basically its harmonic mean It conveys the balance between the precision and the recall.

2*((precision*recall)/(precision+recall))

Deployment of the model:

Fronted

Technology used Flask:-

- Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website
- HTML/CSS

Hosting

Technology used Heroku:-

- Heroku is a container-based cloud Platform as a Service (PaaS). it is used to deploy, manage, and scale modern apps. Our platform is elegant, flexible, and easy to use, offering developers the simplest path to getting their apps to market.