Problem Statement

Stock market prediction is defined as “the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange”. Predicting future values of stock prices can yield great profits for the company. However, the task is not particularly an easy one as it can involve numerous factors (both physical and psychological), making share prices particularly volatile. Investment banks are institutions that act in an advisory role to help companies, individuals or governments make financial transactions.

Major shareholders of the media conglomerate, Waystar Royco (WAYA US) have hired the investment bank you work at, as consultants to analyse their performance in the stock market and make future predictions. Your team, as business analysts in the investment bank, has to predict the closing stock price of the company from 30th July, 2021 to 10th September, 2021 as accurately as possible, given data from 14th August, 2015 to 29th July, 2021. The data you have been provided with, consists of the opening, high, low and closing price of Waystar Royco’s stock as well as the volume of stocks traded in a day, for 1500 days. Your team has been instructed to use ONLY this information and disregard any other factors while making predictions. In addition to this, the bank also wants to know what models are best suited for stock price predictions given similar data and wants your team to perform a comparison between regression and time series models so that they know what to adopt for future scenarios.

DISCLAIMER: Waystar Royco is a fictional company, so any information apart from what is given in the description above, is irrelevant to the problem statement at hand.

Dataset

The main dataset (train.csv) consists of stock prices of Waystar Royco from 14th August 2015 to 29th July 2021. These include the opening, low, high and closing prices on a particular day, as well as the volume of stocks traded in that day. The target attribute is the closing price. Please note that you will have to decide on train-test split yourselves and provide the rationale behind your decision.

Attributes

- Date: datetime, ID, consists only of weekdays
- Open: double, the price at the time trading begins in the stock market
- High: double, the highest price during trading hours of the stock market
- Low: double, the lowest price during trading hours of the stock market
- Volume: integer, the volume of stocks traded on that day
- Close: double, the price at the time trading closes in the stock market

Files

- train.csv: Main dataset
- test.csv: Data for which close price needs to be predicted
- sample_submission.csv: How your submissions must look

What you need to do

- Analyse the dataset provided
- Model using regression and time series techniques
- Provide comparative analysis based on performance metrics and technical domain knowledge (this needs to be done in your kernels itself, with the use of suitable documentation and/or graphs)
- Predict closing stock prices from 30/07/2021 to 10/09/2021 (Refer to sample_submission.csv)

The metric that will be used for evaluation is RMSE, or Root Mean Squared Error, given by the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

There are two leaderboards - a public and a private one. Scores for both are evaluated on different subsets of the solution dataset. The public leaderboard will be dynamic throughout the contest and is subject to change based on submissions. The private leaderboard is not open to all participants during the competition, but will display the final scores once the competition is complete.

The submission file must consist of the following:

- It must have Date and Close as the attributes (refer to sample_submission.csv)
- It needs to consist of 30 rows (excluding the header)

Example of a row: 14/08/2021, 124.67889