Prediction of Closing Prices of Various Stocks

Name:	**Aditya Sneh**
Registration No./Roll No.:	19017
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	January 19, 2022
Date of Submission:	April 24.2022

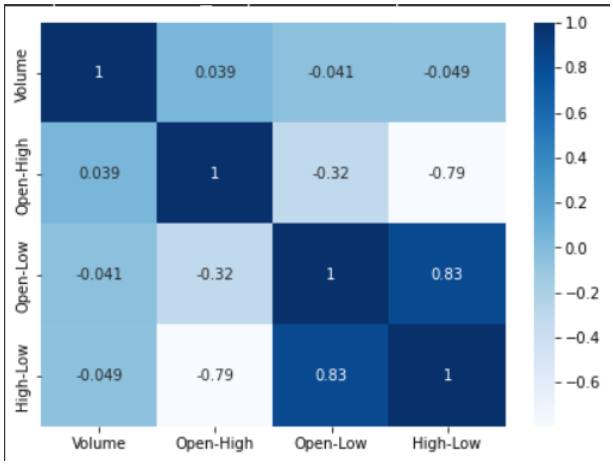
1 Introduction

The stock price prediction dataset is continuous data comprised of 4 Features [Open, High, Low, Volume] with 97732 rows. In this problem, we need to predict the closing price with the help of training data set. To predict stock closing prices we use classical techniques like Linear Regression, Random Forest Regressor KNeighborsRegressor. To select the best features for processing I have used the mutual_info_regression technique to analyze whether dropping the features helps or not or Which features to drop. Hyperparameter tuning (Grid Search) is done to find optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any dataset. To measure the efficiency of the model, three Standard Strategic Indicators were used which are RMSE (Root-mean-square deviation)and MAE (Mean Absolute Error). At last, we compare each RMSE value of the algorithm to get an idea of which one is best for stock market prediction.

2 Methods

2.1 Data Analysis :

Each feature (Open, High, Low, Volume) has an equal number of data points of 97729.

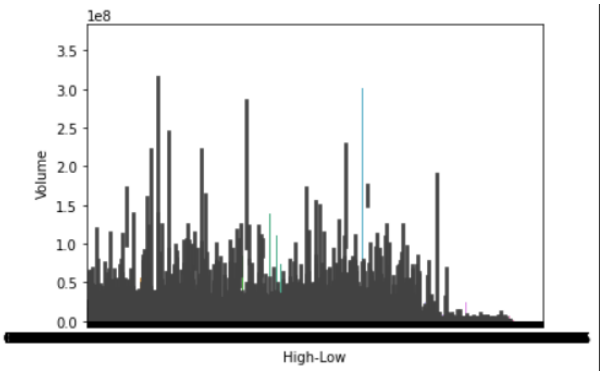


This plot shows the correlation between two variables or features. There are lots of 1's which shows it has high positive correlations and are interrelated. This might be possible because of the comparatively very small difference between those values.

- Variables having a large correlation value with volume represent that those numbers might have intrigued a large number of buyers and sellers.

- Correlation between those features and the volume feature will tell us how a change in that feature impacts the number of stocks traded that day.
- High difference in opening and highest value of the stock might attract more buyers. Whereas fewer differences may attract more sellers.

Now to analyze in-depth, let's study the seaborn barplot of Volume vs High-Low. Volume is high for smaller values of High-Low and lowers for high values of High-Low.



2.2 Data Cleaning :

As dataset contains 12 Null Values with 3 null values in each feature. To treat null values we used Mean Imputation Method. Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable. We calculate the mean of each feature column the fill the value with respective null places.

The mean of the observed data is preserved when the mean is computed. As a result, even if all of the data is missing at random, the mean estimate remains impartial. You can also keep your sample size up to the full sample size by imputing the mean.

2.3 Splitting Dataset :

Splitting the training dataset in an 80:20 ratio for further processing. 80% of the dataset goes into the training set and 20% of the dataset goes into the testing set. To avoid overfitting, it is important to divide the data into the training set and the testing set. We first train our model on the training set, and then we use the data from the testing set to gauge the accuracy of the resulting model.

2.4 Regression Model Used:

We Fit all the regression models on (X_train,y_train) and then predict values on X_test for getting RMSE and MAE Score. Then use the trained model on the test CSV file for Stock Closing Price Prediction.

Linear Regression: Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Random Forest Regressor: Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. In the case of a regression problem, the final output is the mean of all the outputs.

KNN Regressor: KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

2.5 Feature Selection:

As the dataset consists of 4 features and it's not appropriate to remove any features. It might lose some information. To counter whether we can drop features or not. I used a feature selection method called mutual_info_regression which gives a rank score to each feature. Open: 4.335536, High: 4.993344, Low: 4.929872 ,Volume: 0.353862. The result after testing the top 3 features is the same as using all 4 features. MSE : 5164.148 with sel_k: 3 MSE 5164.148 with sel_k: 4 which is same.

2.6 Hyperparameter Tuning:

A GridSearchCv method is used on random forest, KNN regressor to find the best parameter on which the model performs very well. As well from observation data, the accuracy after and before tuning is almost similar.

Experimental Analysis:

hyperref

Observation Excel Sheets : Click on link to view observation EXCEL SHEETS LINK

Google Collab : Click on link to view Code On Google Collab CODE LINK

Github : Click on link to view Github Github LINK

Model	MAE	RMSE	Observation
Linear Regression (Best)	6.124797	69.488751	It performs well on variables that you are specifically testing along with other variables that affect the response in order to avoid biased results.
Random Forest	8.82843567265821	106.198140510133	A trained forest may require significant memory for storage, due to the need for retaining the information from several hundred individual trees.
KNN Regressor	49.9976466727374	164.26406666616	KNN model is that it lacks interpretability. It takes longer to fit.

2.7 Discussions

Describe the merits, limitations, and future scopes of the proposed method or framework developed. Explain your opinions or thoughts about this work e.g., if the work can be extended or it can be used for other applications. You must report the significant findings of this work and scopes of future works if any.

2.8 Merits

The model works very well on a continuous dataset with higher accuracy. Predicting stock Closing price is now more accurate and easy with Linear Regression. Almost all of the algorithms provided

similar accuracy with very close MSE values.

2.9 Limitations

Since the dataset is a simple regression problem Predicting Stock Price is more desirable when the Date Company Stock's Name were given. On Weekends and holidays the market is closed, so it's very difficult to predict the market on that day. Treating missing values in volatile and non-linear datasets is still a debating task. As algorithm is test on small dataset, we need huge pile of date to evaluate.

2.10 Future Scopes

Time series forecasting (predicting future values based on historical values) applies well to stock forecasting which can be improved we use a lag period for fulfilling the missing days and using LSTM (NEURAL NETWORK).

This work can be extended by focusing on the Missing Values On weekends when the market is closed.