Prediction of Closing Prices of Various Stocks

Name:	**Dishant Digdarshi**
Registration No./Roll No.:	19108
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	January 21, 2022
Date of Submission:	April 24, 2022

Introduction

Prediction of stock prices is a demanding task since the financial sector is highly volatile.For our project the data set utilized for analysis was selected from Yahoo Finance.It has over 97000 rows of the required Stock price and other pertinent information.The data set consisted of 4 attributes namely ”Open”,”High”,”Low”,”Volume”.Data prepossessing,analysis was done with the help of multiple python libraries.The test set was limited to 20 % of the total dataset. This project focuses on regression based model architecture. The Regression-based Model is used to predict continuous values from a set of autonomous values. Regression makes predictions by using a given linear function to predict continuous values of the attribute ”Close”.The stock prices vary from a minimum of 2.13 dollar of a stock to 272800 dollar.

Methods

The proposed workflow for my project is :

1. Importing the data set and exploratory data analysis
2. Data Prepossessing
3. Data Splitting
4. Build and train the Models
5. Model Evaluation

Importing the data set and exploratory data analysis :

After reading the dataset, I found out that the train data set had 97732 rows and 4 columns. the stock prices vary from a minimum of 2.13 dollar of a stock to 272800 dollar.The correlation between the open,high,low is 1. 1 indicates a perfectly positive linear correlation between two variables.

Data Prepossessing:

After Dataset reading, I have performed preprocessing operation on the dataset for implimentation of some models. Here I apply Standard Scaler to preprocess the dataset. In preprocessing operation StandardScaler performs the task of Standardization.Standardizing a dataset involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1.

Data Splitting:

After preprocessing, I separated the dataset into two groups: training and testing. The training portion of the data is utilised for 80% of the data, while the testing portion is used for 20% of the data .train_test_split method was used to slit the dataset.This method is used to split arrays or matrices into random train and test subsets.

Build and train the Models:

1.Multiple Linear Regression:

Multiple linear Regression allows the analyser to consider multiple variables which affect the quantity to be predicted.The regression coefficient of an independent variable is the measure by which the dependent variable varies as a result of a unit change in the independent variable.In the implementation of our model we have used the "Open","Low","High","Volume" as independent variables to predict the "Close" values of the stock.I have also performed feature selection for identifying and selecting a subset of input variables that are most relevant to the target variable."Open","Low","High" came to be the best variables to predict "Close"

2.Lasso Regression:

In this method I used a Least Absolute Shrinkage and Selection Operator (LASSO) method based on a linear regression model to predict the closing prices of the stocks.Lasso regression uses a linear regression model but also does L1 regularisation, which is a technique of adding more information to avoid overfitting.StandardScaler was used to rescaling the distribution of prices so that the mean of observed values is 0 and the standard deviation is 1.Lasso performs best when all numerical features are centered around 0 and have variance in the same order.To find the optimal value of alpha, we use scikit learns lasso linear model with iterative fitting along a regularization path. The best model is selected by cross-validation.

3.Random Forest Regressor:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression.Random Forests are formed by multiple Decision Trees. The machine will learn which trees are high influence and set weights on each feature point.The dataset must be split into a training and testing dataset before modeling.The hyperparameters in the random forest model are either used to increase the predictive power of the model or to make the model faster.I have performed hyperparameter tuning to optimize the model performence.

4.KNN Regressor:

The kNN algorithm is a non-parametric algorithm that can be used for either classification or regression.The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.I have used the grid search method to find the best value of k for the model for our problem.

5.Model Evaluation:

Statistical metrics are error metrics for regression, and I used them for calculating risks. Model evaluation is critical, and it needs to be evaluated to reduce risks and increase model performance.I have used the following three metrics to evaluate our models :

1. MAE
2. MSE
3. RMSE

GITHUB LINK : https://github.com/dishant2009/Close-Price.git

Experimental Analysis

1.MAE:

It measures the average magnitude of the errors in a set of predictions without considering their direction. It is the average absolute difference between the prediction and the actual observation where all individual differences have equal weight.

2.MSE:

It takes the sum of the absolute value of error. The mean squared error determines the model performance too.

3.RMSE:

It is the standard deviation of the residuals (prediction errors). Residuals measure how far data points are from the regression line.

Table 1: Performance Of Different Models Using RMSE and MAE

Model	RMSE	MAE	MSE
LR	65.07607345054765	5.8846671624926055	4234.895335741074
KNN	194.24779212790006	60.26588930909649	37732.204746563875
RF	106.13789801810465	9.00976676192824	11265.253395701584
L1R	169.8183235849896	15.032213051734601	28838.263025216238

From our project we got to know that the linear regression models performed best for our data set with the least RMSE value and the highest accuracy.

Conclusion

Modelling of capital markets has authentically been done in partial equilibrium.Such machine learning algorithms do, in fact, provide us a better understanding of the stock market. As a result, we've taken advantage of this opportunity to learn more about the market. My best model was the multiple linear regression technique in this project, which is one of the most commonly used techniques in the capital market for making forecasts.