Stock Market Prediction

INFO6105 Data Science Engineering Methods Project

Introduction :

The stock market is a place where investors purchase and sell company shares. It is a collection of markets where businesses list their shares and other securities for trade. In this project, we are using machine learning algorithms (regression algorithms) and testing which algorithm works best for the prediction.

Methods:

In this project, "Stock Market Prediction," we are using regression algorithms for the prediction process. The regression algorithms used are linear regression and support vector machine regression algorithms.

Linear regression is a technique used to predict the value of a variable based on the value of another variable. The variable that is used to predict is the independent variable,and the one we want to predict is the dependent variable. When we are using support vector machine regression, instead of fitting a line to the data points like linear regression, it determines which hyperplane in a continuous space best matches the data points.

We are determining regression metrics such as coefficient of determination r^2,mean squared error R_{sq},mean absolute error, maximum error, and average positive mean squared error. The mean absolute error is computed using the sum of all errors between predicted values and actual values and taking their average. The mean squared error is computed using the sum of the squared differences between each predicted value and actual value and then averaging the sum. The coefficient of determination measures how well the model predicts and explains future outcomes,but that alone won't be enough to tell which regression algorithm works well for prediction. We need to even check for errors and then conclude which regression algorithm is working well for the prediction process.

Results and Conclusions:

```
results of linear regression are:
Train R_squared:  0.9993430072802991

Test R_squared:  0.989485008050805

Train mean_squared_error:  1.8826931442342258

Test mean_squared_error:  47.78517167897713

Train mean_abs_errord:  0.8141109497354003

Test mean_abs_error:  4.397215030583551

Train_max_error:  15.327353030057623

Test_max_error:  44.567772540970225
```

Figure 1: Linear regression metrics obtained

```
results of support vector machine regression algorithm are:
Train R_squared:  0.45249422714674314

Test R_squared:  -7.24962151637302

Train mean_squared_error:  1568.9448818380747

Test mean_squared_error:  37490.24082482998

Train mean_abs_errord:  28.653536346171645

Test mean_abs_error:  179.5451267844229

Train_max_error:  152.1080332478199

Test_max_error:  392.8303774018592
```

Figure 2: SVM regression metrics obtained

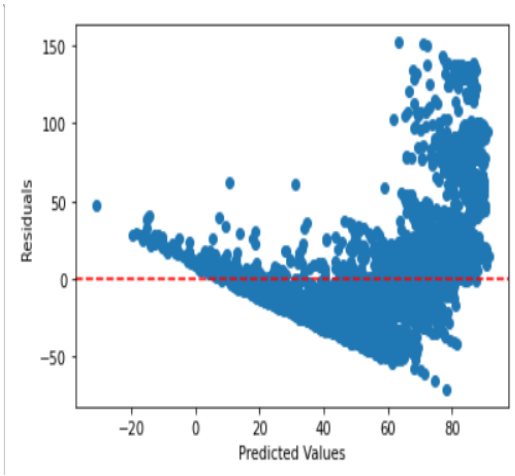

Figure 3: residuals vs predicted data for training data using svm regression

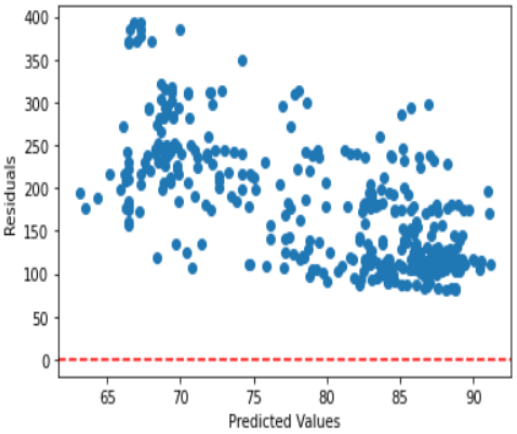

Figure 4: Residuals vs. data for testing data using svm

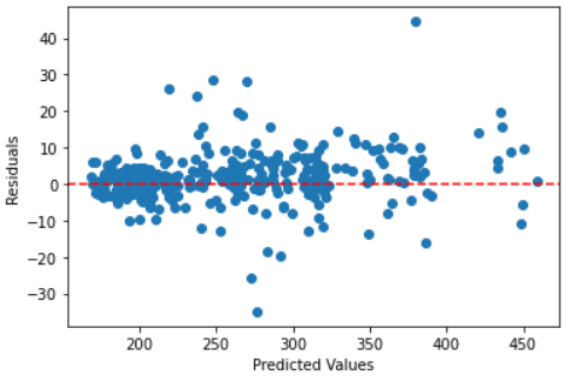

Figure 5: residuals vs predicted data for testing data using linear regression

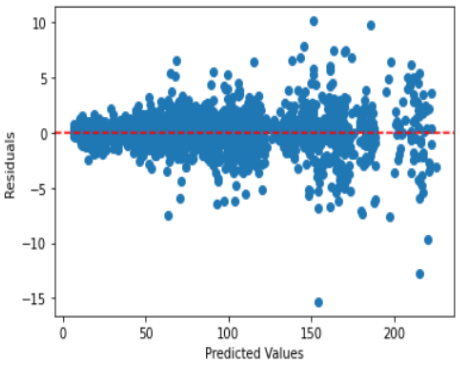

Figure 6: residuals vs predicted data for training data using linear regression

```
Average Positive MSE for training data using linear regression: 0    2.049695
dtype: float64

Average Positive MSE for testing data using linear regression: 0    55.204215
dtype: float64

Average Positive MSE for training data using svm regression: 0    1697.383085
dtype: float64

Average Positive MSE for testing data using svm regression: 0    3367.487251
dtype: float64
```

Figure 7: results obtained using K-cross validation

From the above figures, we can observe the metrics obtained using linear regression and the support vector regression algorithm for prediction. It can be observed that apart from the coefficient of determination, which is very high, we are also getting low mean squared error, mean absolute error, maximum error, and average positive mean squared error for linear regression compared to SVM regression.

From the residuals vs. predicted data plots, it can also be observed that for linear regression, the residuals are very low and are scattered around zero for both training and testing data. In contrast, for SVM regression, the residuals are very high and are scattered away from zero.

Hence, it can be concluded that the linear regression algorithm works well on our model compared to SVM regression for predicting the variable closing stock price.

References :

The dataset used for this project is taken from the Kaggle website,and the URL for the same is provided below:

https://www.kaggle.com/datasets/nikhilkohli/us-stock-market-data-60-extracted-features/data