Group 20

STOCK MARKET ANALYSIS

Final Group Project

Prepared By:

+ Son Nguyen (snguye49)
+ Akshant Jain (ajain78)
+ Deep Patel (deeprp2)

Group 20

ABSTRACT OF THE PROJECT

Stock market is the aggregation of buyers and sellers of stocks, which represents ownership claims on business. Since the birth of the formal stock market, modern stock learning systems have been developed for people who are interested to learn how to trade and invest in the stock market. However, fews have a method of measuring stock performance and predicting if the stock will go up or down on a certain date. Therefore, we want to introduce you to our project of stock market analysis that will analyze the past five year records of particular stocks and do a predictive analysis on those dataset.

In the project, we used various training models, error computing methods and accuracy calculations to perform the prediction on the stock prices. All works are done on Jupyter Notebooks along with pre-existing libraries like sklearn, numpy and pandas. The models in Jupyter notebook do the price predictions of any blue chip stock by replacing stock variable value in the file with the stock symbol.

Our mission is to not only help people understand more about the stock market, but also to maximize their profit in a stock they choose to invest in. In the future, we will keep learning more about this subject and adding any necessary tools to improve our project.

Group 20

DATA SOURCE, FORMAT AND CLEANING

1. Source And Data Collection

Dataset is collected from Yahoo Finance (see Appendix) for all stocks in a span of 5 years starting from 8th February 2013 to 7th February 2018. Each stock will have a name, date, high, low, open, and close price.

Data Types

Columns	Type
Date	object
Open	float64
High	float64
Low	float64
Close	float64
Volume	int64
Name	object

Data Sample

Index	Date	Open	Close	Volumn	Name
0	2013-02-08	15.07	14.75	8407500	AAL
1	2013-02-11	14.89	14.46	8882000	AAL
...	...	...	...	...	...
619036	2018-02-02	77.53	76.78	2595187	ZTS
619037	2018-02-05	76.64	73.83	2962031	ZTS
619038	2018-02-06	72.74	73.27	4924323	ZTS

Group 20

2. Cleaning Data

+ Missing Data: Open, high, and low features have less than 10 missing data values. We decide to impute those with the average value from the same feature. Ten values is a small number that our data won't be affected at all if we impute it.

+ Outliers: There's only one outlier in the volume feature. We decide to keep this outlier because it's an important piece to why the volume of a stock spike in certain days, which can indicate good or bad news about that stock.

+ Noisy Data: For repetitive data, we check to find a column with more than 90% row value being the same but nothing got printed which means our data are mostly unique. For irrelevant data, we decide to drop high and low features because they're irrelevant to our project. We're more interested in the open and close value to try to predict if it goes up or down at the end of the day. For duplicated data, we check to see if there's any duplicate date and name together because those two should be unique as it describes a stock at a certain date. Lastly, for inconsistent data, we put all letters to uppercase for consistency and we also convert date column to date format for easier analysis later

* Additional Dataset Used:

After cleaning the first source of the dataset, we used another dataset of an Amazon stock to verify whether the model trained and used to predict if the values are working. We downloaded the dataset provided by Yahoo finance (see Appendix) which contains the value of amazon stock starting from 2nd December 2019 to 1st December 2020. This dataset was used to see whether our training dataset was producing the correct result or not according to the current values.

Group 20

METHODS USED FOR PREDICTION

For predicting the closing price of a stock based on the previous day's closing price, we used regression models because we have a continuous set of values. (The classification model predictions was beyond the scope of our project)

For predicting the values of the stocks, we used multiple regression methods:

- Linear regression
- SVM Regression (Linear, Radial Basis Function (RBF))
- KNNeighbour Regression

For the accuracy scores and errors, we used various methods/functions:

- Errors Methods:
    - Sklearn metrics absolute mean error
    - Sklearn metrics squared mean error
    - Sklearn metrics absolute median error
    - Sklearn metrics variance explained error
- Accuracy Scores Methods:
    - Sklearn metrics R2 accuracy score
    - Trained Model Accuracy Score
    - K-Fold cross validation Accuracy Score (Used different k-fold values but resulted in same accuracy score with a low differences between them)

Group 20

RESULTS

For the result demonstration, we explored Amazon and Amgen stock. However, the user can choose another stock to predict through the input box provided in the notebook.

For the two stocks, we did different types of graph analysis on the two stocks and

their prices such as predicting closing prices and predicting prices for next 20 days based on annual stock price data.

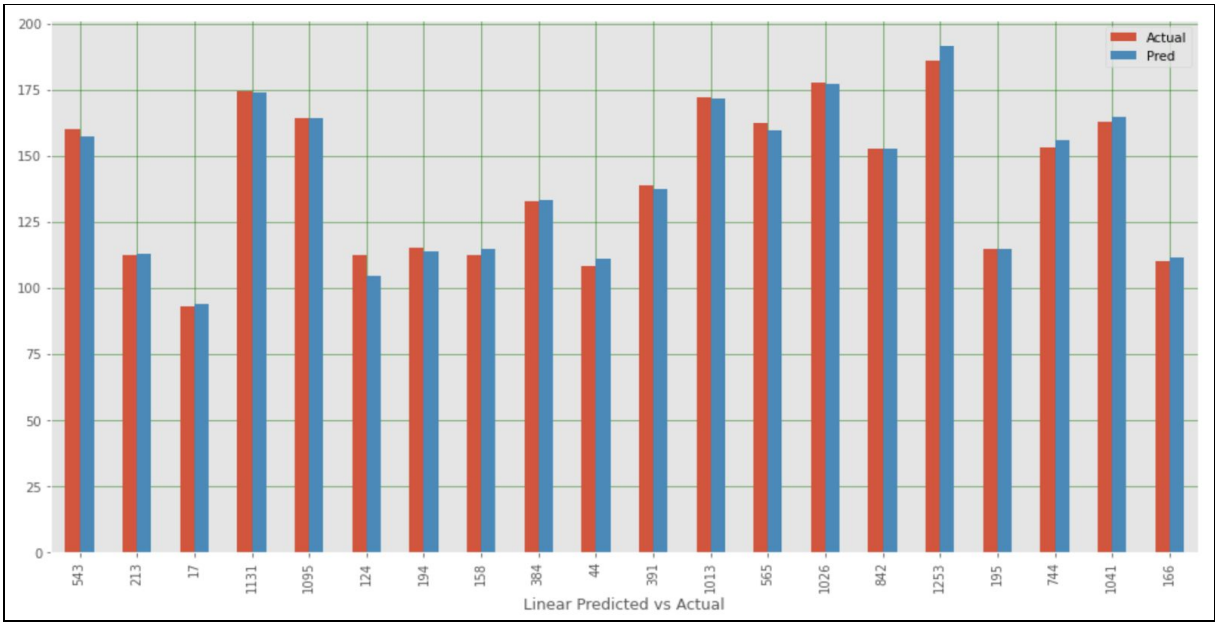
- Amgen Stock (AMGN)
- Amazon Stock (AMZN)

In the next page, we will be performing a regression analysis on the closing price of Amgen stock during the time of 2013-2018. We used linear regression, kNN, and SVM regression models to predict the closing price. For each model, we computed statistical data based on the k-fold cross validation since our models didn't have a large training dataset.

Group 20

Linear Regression Model

- Graph



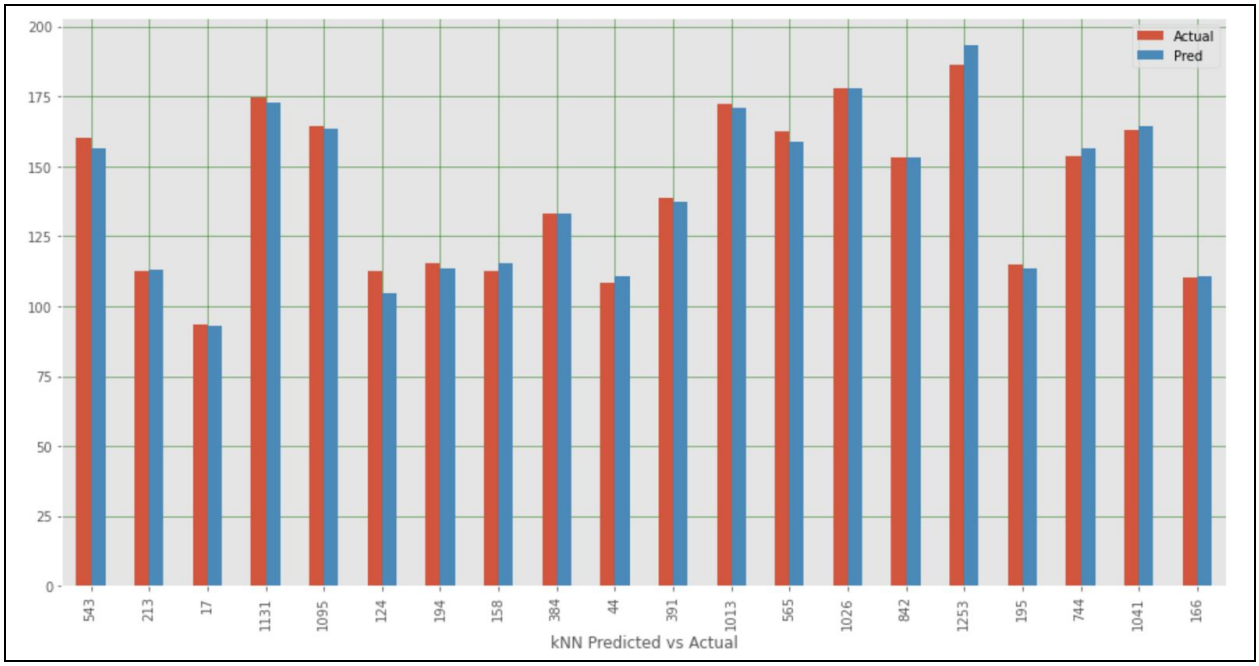
- Accuracy and Errors

```
K-Fold Accuracy:  99.457798652039
Score Accuracy:  0.994859101922841
Mean absolute error = 1.39
Mean squared error = 3.4
Median absolute error = 1.06
Explain variance score = 0.99
R2 score = 0.99
```

Group 20

KNNeighbour Regression Model

- Graph



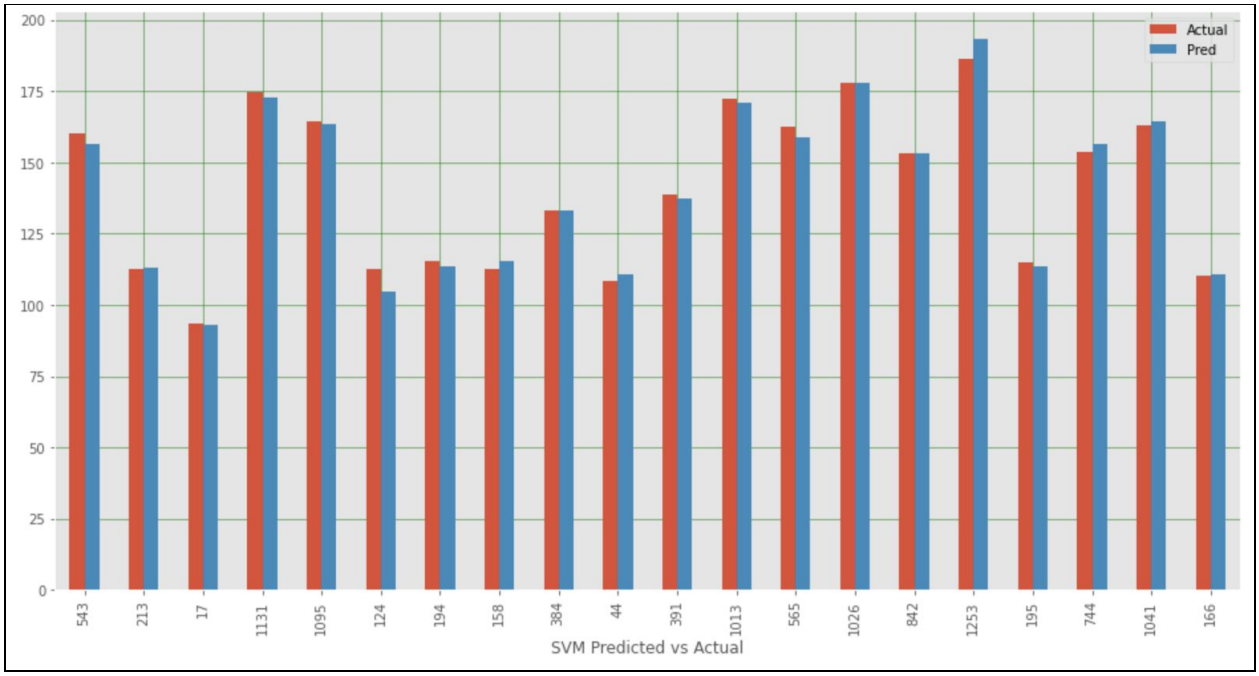
- Accuracy and Errors

```
K-Fold Accuracy:  98.95181580689155
Score Accuracy:  0.9930189068999671
Mean absolute error = 1.66
Mean squared error = 4.62
Median absolute error = 1.34
Explain variance score = 0.99
R2 score = 0.99
```

Group 20

SVM Linear Regression Model

- Graph



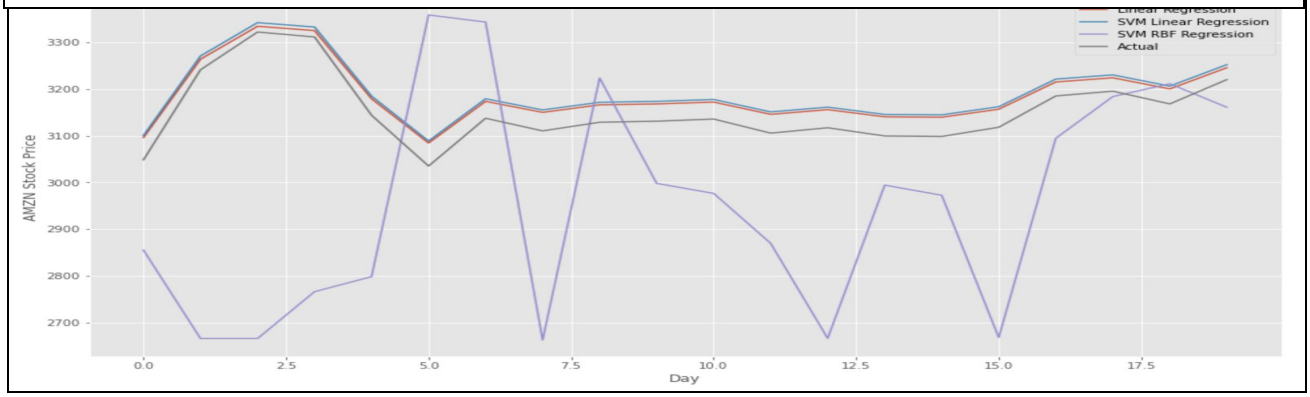
- Accuracy and Errors

```
Score Accuracy:  0.9945854997470462
Mean absolute error = 1.43
Mean squared error = 3.58
Median absolute error = 1.1
Explain variance score = 0.99
R2 score = 0.99
```

Group 20

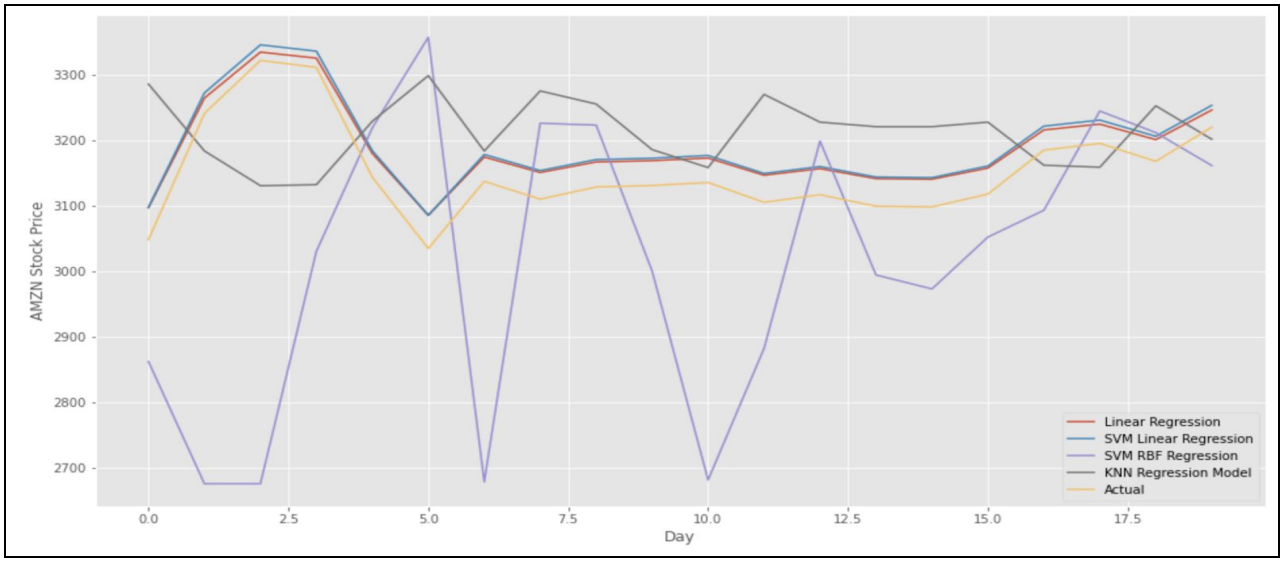
For amazon stock, we have used Linear regression, SVM Linear Regression, SVM RBF Regression models. We have used a multi-line chart for comparing the predictions by various models.

```
linear svm confidence:  0.8685427225846158
rbf svm confidence:  0.5946918726751149
lr confidence:  0.8674753550232022
```



We can see that the linear regression is better than the RBF regression for SVM regression models. Linear Regression. However, with the KNN model it looks like:

```
linear svm confidence:  0.793673175589624
Knn Confidence 0.9110828900264779
rbf svm confidence:  0.5067684817464231
lr confidence:  0.7902567749251435
```



FINAL CONCLUSION

Here are some conclusions that could be drawn from the above results:

- After analysing the regression models with the graphs, we can see that the linear regression models for both normal and SVM are better than any other models.

- We can see the high accuracy scores for the models. This is due to the continuous and small dataset which results in high accuracy scores and predicted values. We predicted stock prices solely based on the past historical data as our variable. We believe that the predictions would be skewed if we accounted for market sentiments and other factors.

- If a dataset with a sudden steep are analysed, then there could be accuracy errors because those are due to other factors such as natural calamity, board decisions, boycotts, pandemic etc. which may not be taken into account while using the ML at this stage.

- After the project, we've learned a lot about the stock market as well as how prediction can be made for a particular stock. We believe that there are many categories that make stock prices go up. Our project has addressed one of many factors to take into consideration to predict stock price in a short span of time. In the future, our team plans to keep working on our project to add more measuring factors to our list to help make prediction more accurate.

Group 20

APPENDIX

- Source for 5 year dataset of all the stocks by Kaggle:
  https://uofi.app.box.com/s/067ss2133hgyukhwva08oawq8iq6nsc7

- Source for the additional dataset for the amazon stock:
  https://finance.yahoo.com/quote/AMZN/history?p=AMZN