

Stock Market Price Prediction

Project-Based Internship 2020 Report

Submitted

To

DataRitz Technologies

Duration: Six Weeks

By

Srajit Srivastava
1803210151

Vishal Maurya
1803210182

ABES Engineering College

Dr. A.P.J. Abdul Kalam Technical University

Under the guidance of

Mr. Gopal Gupta

Mr. Shashank Shekhar

TABLE OF CONTENTS

	Page No.
DECLARATION	iii
CERTIFICATE	iv
ACKNOWLEDGEMENT	v
ABSTRACT	vi
LIST OF FIGURES	ix
CHAPTER 1 INTRODUCTION	10
1.1 Problem Definition	10
1.2 Motivation	10
1.3 Objective of the Project	10
1.4 Scope of the Project	10
1.5 Need of Work	11
CHAPTER 2 RELATED WORK	12
CHAPTER 3 PROPOSED METHODOLOGY	13
3.1 Dataset Description	13
3.2 Methods	16
3.3 Hardware / Software Requirements	17
3.3 Our Methodology	18
CHAPTER 4 EXPERIMENT AND RESULT ANALYSIS	26
CHAPTER 5 CONCLUSION	28
5.1 Discussion	28
5.2 Future Work	28
REFERENCES	29

DECLARATION

We hereby declare that the work being presented in this report entitled “**STOCK MARKET PRICE PREDICTION**” is an authentic record of our own work carried out under the supervision of “**Mr. Gopal Gupta**” and “**Mr. Shashank Shekhar**”. The matter embodied in this report has not been submitted by us for the award of any other degree.

Dated: **07/07/2020**

Signature of Student:

SRAJIT SRIVASTAVA

VISHAL MAURYA

Department:

**COMPUTER SCIENCE AND
ENGINEERING**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Signature of Supervisor

CERTIFICATE

This is to certify that Project Report entitled “STOCK MARKET PRICE PREDICTION” which is submitted by Srajit Srivastava and Vishal Maurya in partial fulfillment of the requirement for the summer internship of Data Analysis and Machine Learning Using Python in Department of Computer Science and Engineering of ABES Engineering College, is a record of the candidates’ own work carried out by them under our supervision.

Supervisor 1: Mr. Gopal Gupta

Supervisor 2: Mr. Shashank Shekhar

Date: 07/07/2020

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the Project Based Internship 2020 undertaken during “Data Analysis and Machine Learning Using Python - 2020”. We owe special debt of gratitude to Mr. Gopal Gupta, Lead Technical Architect – Data Science & AI and Mr. Shashank Shekhar, Project Consultant – Data Science & AI, DataRitz Technologies for his constant support and guidance throughout the course of our work. His constant motivation have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of team members of DataRitz Technologies for their full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the motivation of Computer Science and Engineering Department of ABES Engineering College to provide us the opportunity to undergo training at DataRitz Technologies.

*Signature: SRAJIT SRIVASTAVA
VISHAL MAURYA*

*Name: Srajit Srivastava
Vishal Maurya*

*Roll No.: 1803210151
1803210182*

Date: 07/07/2020

ABSTRACT

The Stock Market refers to the public markets where exchanges and activities of Buying, Selling and Issuance of shares of companies takes place on regular basis. During the processing of these activities there is an exchange in the capital and therefore consists of risk due the fluctuation of the market. These fluctuations in the stock market in turn gives the investors a second thought whether to invest in a particular company or not. So to gain the confidence of the investors to buy and invest in the market and to maintain a regularity in the market there should be a mechanism to predict the market trends of the future. In order to gain this confidence of the investors and maintain regularity of the market we have implemented a model using machine learning which could predict the trends of the market and give the approximate value of the stock of a company on a particular date. There are some models which are doing this task but our model is made using data analysis and some machine learning algorithms which best fits on the dataset taken to build the model.

Keywords: Stock Market, Machine Learning Algorithm

Project Summary

Region/Unit	DataRitz Technologies
Location	Ghaziabad
Program	DataRitz Technologies.<<programcode>>.<<version>>
Project Number	DataRitz Technologies. <<projectcode>>.<<version>>
Project Description	Stock Market Price Prediction

Document Control

Prepared by:	Srajit Srivastava Vishal Maurya
Title:	Stock Market Price Prediction
College:	ABES Engineering College
Department:	Computer Science and Engineering
Location:	Ghaziabad
Version date:	07/07/2020
Status:	Initial Draft

Version history

Version no.	Date	Changed by	Nature of amendment
1.0	07/07/2020	Srajit Srivastava Vishal Maurya	Initial Draft

Endorsement and Approval

Project Customer

I approve the business requirements specifications in this document.

Name	<<customer name>>		
Position	<<customer position>>		
Signature		Date	

The following officers have **endorsed** this document

Project Sponsor

Name	<<sponsor name>>		
Position	<<sponsor position>>		
Signature		Date	

Project Manager (= Component Project Customer)

Name	<<project manager name>>		
Position	Lead Technical Architect		
Signature		Date	

Component Project Sponsor

I accept the business requirements specifications in this document.

Name	Dr B P Sharma		
Position	Country Head – Delivery		
Signature		Date	
Comments			

The following officers have **endorsed** this document

Component Program Manager

Name	Mr. Gaurav Kansal		
Position	Chief Operating Officer		
Signature		Date	

LIST OF FIGURES

Figure No.	Figure Description
3.1	A view of top rows of the dataset
3.2	A view of bottom rows of the dataset
3.3	Data types of the attributes of the dataset
3.4	Steps involved in a ML Process
3.5	Distribution of Train, Cross-validation and Test data
3.6	Predictions done on cross-validation data
3.7	Graph between RMSE and no. of samples on cross-validation data
3.8	Graph between r2_score and no. of samples on cross-validation data
3.9	Graph between actual and predicted value of 'Adj Close' using no. of sample = 5 on cross validation data
3.10	Predictions done on Test data
3.11	Graph between actual value and predicted value on test data
3.12	RMSE and r2_score for test data
3.13	GUI to take input from the user
3.14	Sample input from the user
3.15	Output for the sample input
4.1	Sample Input
4.2	Sample Output

CHAPTER 1

INTRODUCTION

1.1 Problem Definition:

Stock Market refers to public markets where stocks are traded i.e. investors buy and sell stock. There is very high risk of capital invested on any stock. In order to decrease the risk factor, we have to predict the trend of the stock market in an efficient way. To predict the trend, we have used previous data of the company's stock values.

1.2 Motivation:

The motivation of the project is basically to overcome the risk investors face on investing in any company's stock and give an overview whether it would be beneficial to invest in that company's stock or not by predicting the approximate value of the stock on a particular date using some of the previous data of company's stock values. And to implement some of the machine learning algorithms and analyze data we have learnt so far, we made this project.

1.3 Objective of the Project:

The objective of the project is to predict the approximate value of the company's stock on a particular date using the data of stock value of the previous days of the company. We have made an interface where we take input of the date and number of days to be used for the predictions and give the predicted value of stock on that date.

1.4 Scope of the Project:

It is very important to define the scope this project as there are already some projects having same purpose of predicting stock value. In this project we are using Linear Regression as after analyzing the data we have seen that the attributes of the data are linearly varying with respect to each other except the volume attribute so at last we have concluded that the predicted price can be calculated with the date and adj. close only using

Linear Regression between the date and adj. close. We are using the data of some previous days to plot a linear regression and from that plotting we fetched the trend of the graph with the value on the given particular date.

1.5 Need of Work:

The need of work is that we are living in an era full of technologies and to utilize that in stock market we are using machine learning algorithms to predict trend of the market and reduce the risk of capital of investors investing in the stock market. There had been a huge growth in the technologies day-by-day and to give it a chance to benefit investors we are doing this project.

CHAPTER 2

RELATED WORK

There are many related works being done on stock market price predictions previously and going on too. Some of them are mentioned below:

1. "Stock price prediction using the ARIMA model"

ARIMA - Autoregressive Integrated Moving Average Model. This model is used for short-term predictions.

2. "Stock price prediction using LSTM, RNN and CNN-sliding window model."

This model works on deep learning concepts and uses both linear and non-linear algorithms.

3. "Stock price prediction using neural network with hybridized market indicators."

This model is based on ANN (Artificial Neural Network) which is used as mining tool and financial forecasting.

4. "Integrating metaheuristics and artificial neural networks for improved stock price prediction."

This model uses Integrating metaheuristics and ANN for improved stock price prediction and both topology of ANN and the number of inputs are optimized.

5. "Stock price prediction using reinforcement learning."

This model proposes a method using reinforcement learning to predict stock market prices for long-term.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Dataset Description:

The dataset we chose for our project is downloaded from Yahoo Finance.

The dataset consists of stocks of Tata Motors of approx. 10 years from 09/04/2010 to 08/06/2020 .

The attributes of dataset are as following:

- ❖ Date - Corresponding Date for stock values.
- ❖ Open - Opening price of a stock on a particular day. It is the price at which the financial security opens in the market when trading begins.
- ❖ High - Highest selling stock value for a day. It is the highest price at which a stock traded during the course of the trading day.
- ❖ Low - The lowest value of the selling price of a stock on a given day. It is the lowest price at which a stock trades over the course of a trading day.
- ❖ Close - Contains closing value of a stock on a given day. It is the last price at which stock is closed for a trading day.
- ❖ Volume - The number of shares traded or brought on a given day.
- ❖ Adj Close - The closing price of a stock after paying dividends to the investors. Adjusted close price amends a stock's closing price accurately reflects that stock's value after accounting for any corporate actions.

A view of how dataset looks is given below:

```
#View of dataset
data.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2010-04-09	155.334000	161.141998	154.343994	160.212006	128.496658	29347238.0
1	2010-04-12	160.270996	160.270996	153.968994	154.552002	123.957092	29781931.0
2	2010-04-13	154.938004	155.669998	150.604996	153.296005	122.949738	49489395.0
3	2010-04-15	154.343994	157.787994	152.761002	153.423996	123.052391	30353385.0
4	2010-04-16	153.354996	156.214005	153.216995	155.283997	124.544197	17014036.0

Fig.3.1 View of data using data.head()

```
data.tail()
```

	Date	Open	High	Low	Close	Adj Close	Volume
2505	2020-06-02	90.000000	97.300003	89.750000	96.500000	96.500000	125407771.0
2506	2020-06-03	100.000000	101.449997	97.400002	98.750000	98.750000	90598067.0
2507	2020-06-04	99.000000	101.400002	96.800003	98.500000	98.500000	72078707.0
2508	2020-06-05	100.449997	112.449997	99.050003	110.750000	110.750000	187209208.0
2509	2020-06-08	114.000000	119.150002	113.349998	115.449997	115.449997	136943010.0

Fig.3.2 View of data using data.tail()

3.1.1 Datatype of Dataset:

The datatype of the attributes of dataset can be seen in the figure given below:

```
#Check datatype
data.dtypes
```

Date	object
Open	float64
High	float64
Low	float64
Close	float64
Adj Close	float64
Volume	float64

Fig.3.3 Checking datatypes of the attributes

From figure, we can conclude that the datatype of the attributes of the dataset is:

- Date is in the form of Object. For further calculations we need to convert it to in the pandas 'DateTime' format.
- Open, High, Low, Close, Adj Close, Volume are in numerical from and of Float 64-Bit type.

3.2 Methods

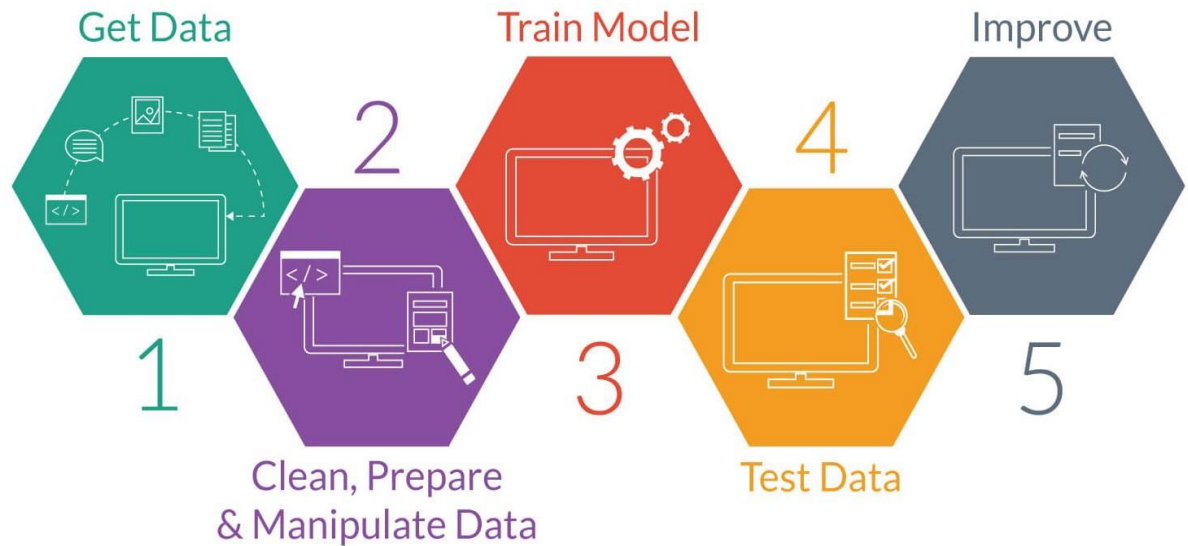


Fig.3.4 Steps in machine learning

- **Get Data:** For a machine learning project a good dataset is necessary. For our project we have chosen the dataset of Tata Motors Ltd. We have downloaded the dataset from the yahoo finance. The dataset is of 10 years from date 09/04/2010 to date 08/06/2020.
- **Clean, Prepare and Manipulate Data:** In this step we removed the NaN values from our project and then checked for outliers. After cleaning the data we visualized the data to find correlations between different attributes. After above processes we divided the data into train, cross-validation and test data.
- **Train Model:** After splitting the dataset fit the train data into the model to get the model.
- **Test Data:** After fitting the train data into the model we got the model for our project. Then we tested the model using the test dataset.
- **Improve:** Check the accuracy of the data and improve the model to increase efficiency.

3.3 Hardware / Software Requirements:

Hardware Requirement:

Minimum hardware requirement to download Python on your Windows operating System as follows:

- Processor: AMD R5 2.00 GHZ
- RAM: 4 GB
- Hard Disk: 1 TB

Software Requirement

Nowadays, Anaconda Navigator is supported by almost every operating system, whether it is a Windows, Macintosh and Unix all supports the Python application development. So, you can download any of the operating system on your personal computer. Here, is the minimum requirement.

- Operating System: Windows 7/8/10
- Jupyter Notebook

IDE

To develop this project Jupyter Notebook IDE is used. Jupyter Notebook is a well-known IDE for Python development.

3.4 Our Methodology:

- ❖ In our project we are trying to predict the value of 'Adj Close' for a trading day. The reason to choose this attribute of dataset is because 'Adjusted close price' amends a stock's closing price accurately reflect that stock's value after accounting for any corporate actions.
- ❖ After choosing the desired attribute we need to find a suitable model for our project and fit the data.
- ❖ We have splitted our dataset into three parts:
 - **Train Data:** Train data is used to fit the model into the linear regression model. Train data is used to make sure the machine recognizes the patterns in data.
 - **Cross Validation Data:** Cross validation data is used to ensure better accuracy and efficiency of the algorithm used to train the model.
 - **Test Data:** Test Data is used to see how well our machine can predict the new values based on its training on train data.

Below figure depicts the distribution of train, cross-validation and test data by Date

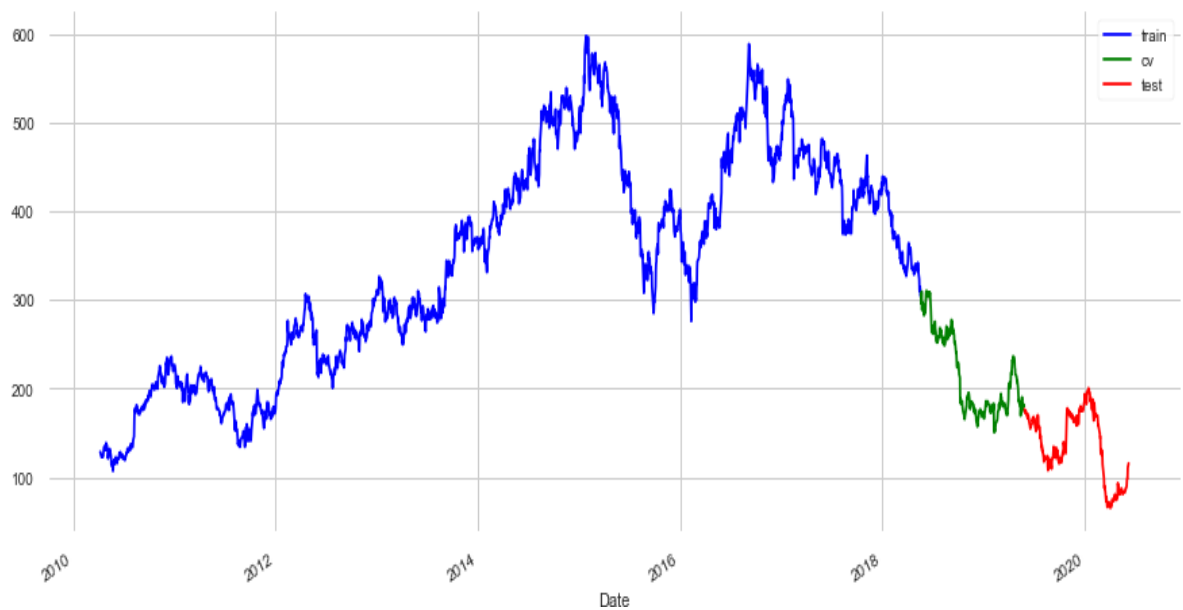


Figure.3.5 Distribution of Train, Cross-validation and Test data

- ❖ After splitting the data task was to fit the data into the linear regression model.

We used a number of past days stock prices to predict the stock prices for upcoming days. Now we needed to know how many days should be used to predict the price for upcoming days, we came out with a solution. We call the number of days as ‘number of samples’ in our project.

We chose the sample of maximum of 30 days to be used in our model. It means we will be using the prices of past 30 days to predict the price for upcoming days. We first fitted our model using day 1 and predicted the price, then we used 2 days to fit our model and predict the price and then 3 days and so on till 30 days and predicted the prices. We used Cross-Validation dataset to predict the prices using 1 to 30 days. Below is the result of the above process

	Adj Close	Volume	pred_for_N_1	pred_for_N_2	pred_for_N_3	...	pred_for_N_21	pred_for_N_22	pred_for_N_23	pred_for_N_24	pred_for_N_25	pre
34	295.899994	11549490.0	304.600006	294.300018	304.266673	...	316.780002	317.398704	317.837550	318.047284	316.862502	315
12	307.700012	17199292.0	295.899994	287.199982	286.133331	...	310.992382	311.855196	312.612057	313.188406	313.536000	312
12	309.450012	17698475.0	307.700012	319.500030	305.833343	...	308.191909	308.905198	309.756525	310.511960	311.098002	311
34	288.649994	46104103.0	309.450012	311.200012	317.900024	...	305.915482	306.800006	307.472338	308.284244	309.011505	305
34	294.149994	19240172.0	288.649994	267.849976	282.883321	...	299.754765	301.055199	302.037752	302.813591	303.718504	304

Figure.3.6 Predictions on Cross-Validation Data

We saw a variation in the values predicted using different number of samples. Now the task was to find what number of samples gave us the best result. To find it we decided to find the error in the prediction against actual value.

To check the error in each prediction done by our model, we used the mathematical technique ‘root-mean-squared-error’ (RMSE). Lesser the value of RMSE, better the predictions.

We plotted a graph between the number of days used in our project and the corresponding RMSE which is shown on the next page

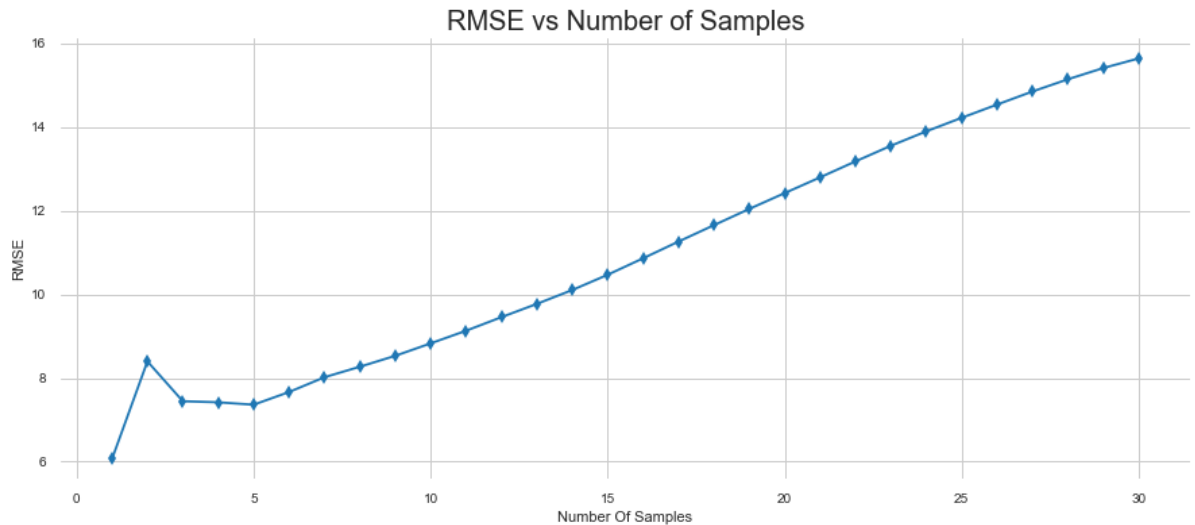


Figure.3.7 Graph between RMSE and number of samples

It is obvious from the graph that the lowest RMSE is when we take the number of samples 1, and after that lowest RMSE is achieved on the number of samples = 5.

Since we are using linear regression model for our project, so taking number of samples = 1 is not suitable. Hence we will choose the number of samples = 5 which gave us the next lowest RMSE on the data. Hence we will use number of samples = 5 throughout our project.

Another way to know the accuracy of the predictions is 'r2_score'. **R-squared (R^2)** is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Higher the r2_score, better the data fitted into model.

Below is the graph representing the variation of r2_score with respect to the number of samples

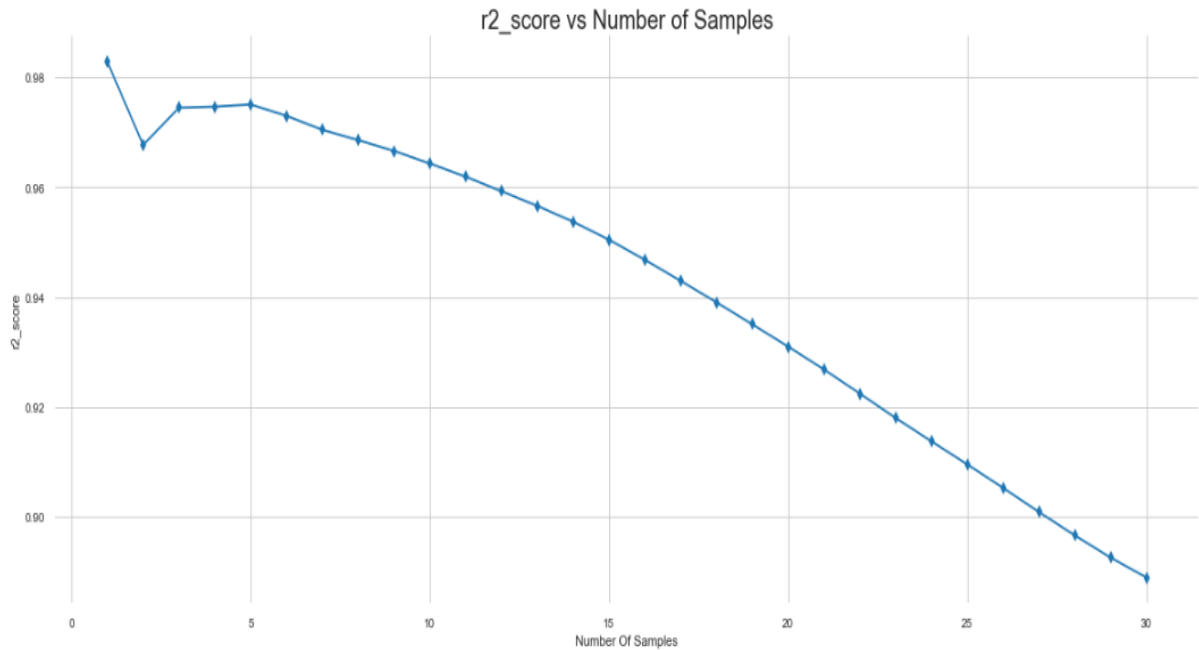


Figure.3.8 Graph between r2_score and number of samples

It is obvious from the graph that highest r2_score is achieved by number of samples equal to 1, but we can't choose number of sample = 1, so next highest value of r2_score is achieved on the number of samples = 5. Also our RMSE was lowest on the number of samples = 5.

Hence by the above observation we decided to choose the number of samples = 5.

On the next page there is the distribution of actual value of 'Adj Close' and the value predicted by our project using number of samples = 5.

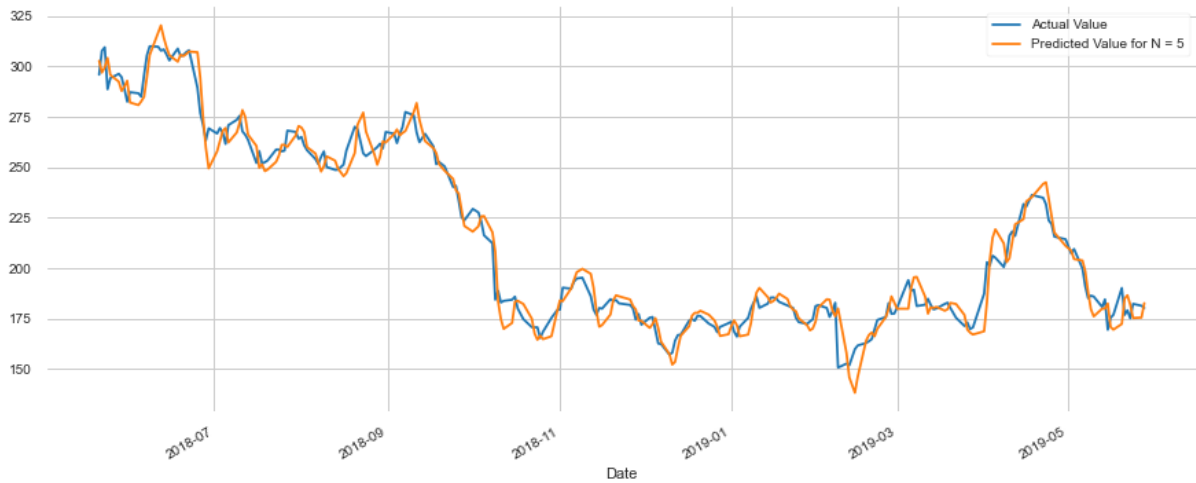


Figure.3.9 Graph between Actual value and predicted value for number of samples = 5

For choosing the best number of sample, we have used our cross validation dataset.

Now after finding the best number of sample and training our linear regression model, the next task was to check the predictions on the test data.

For the test data we used number of sample = 5 and done the prediction using the model. We obtained following result

	Date	Open	High	Low	Close	Adj Close	Volume	Predicted Adj Close
2259	2019-05-29	179.850006	180.000000	175.350006	176.350006	176.350006	15222008.0	182.180002
2260	2019-05-30	177.500000	178.500000	173.100006	175.149994	175.149994	18078726.0	179.185008
2261	2019-05-31	177.800003	177.850006	170.000000	172.600006	172.600006	21488115.0	173.289996
2262	2019-06-03	170.500000	175.000000	168.399994	174.500000	174.500000	21288078.0	170.360001
2263	2019-06-04	174.250000	177.199997	172.399994	173.250000	173.250000	17428731.0	171.195001

Figure.3.10 Predictions done on Test data

For better understanding we plotted a graph between actual value and predicted value of 'Adj Close' for our test data

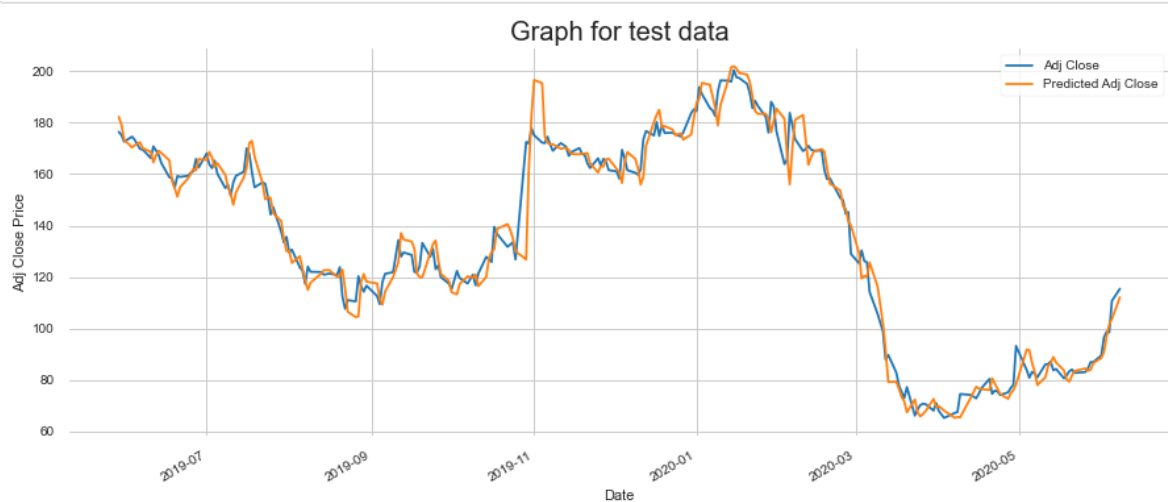


Figure.3.11 Graph between Actual value and predicted value for test data

Now its time to check the error in our prediction.

```
1 #root_mean_squared_error
2 rmse = mse(test_data['Adj Close'],pred,squared = False)
3 r2 = r2_score(test_data['Adj Close'],pred)
4 print(rmse)
5 print(r2)
```

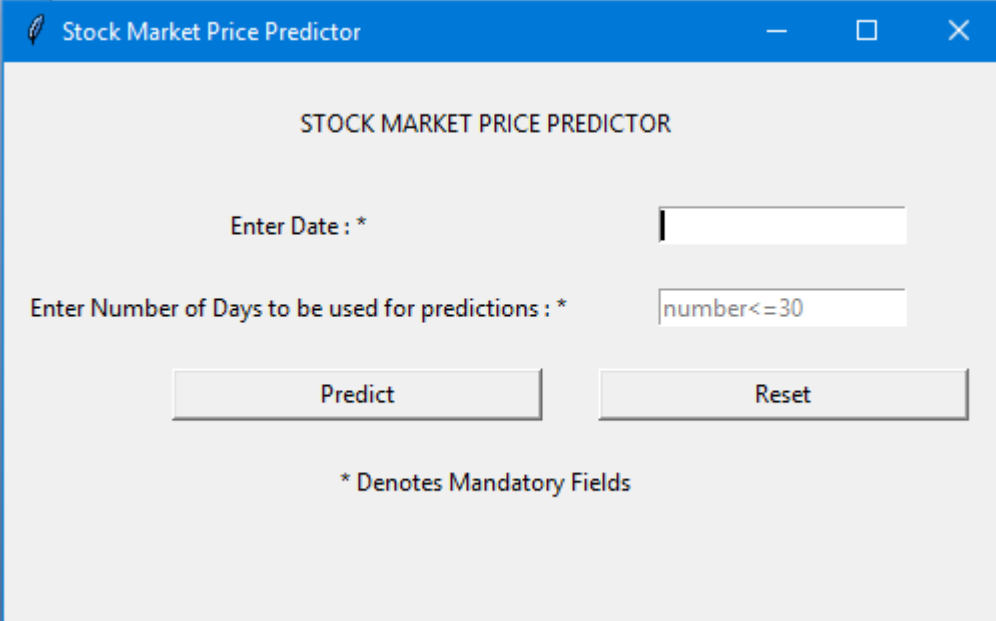
```
9.411367153360919
0.9462723644643782
```

Figure.3.12 RMSE and r2_score for test data

The **RMSE** in our prediction done by our model on test data is **9.411367153360919**.

The accuracy of the prediction is decided by **r2_score** which is **0.9462723644643782** which is very close to 1.

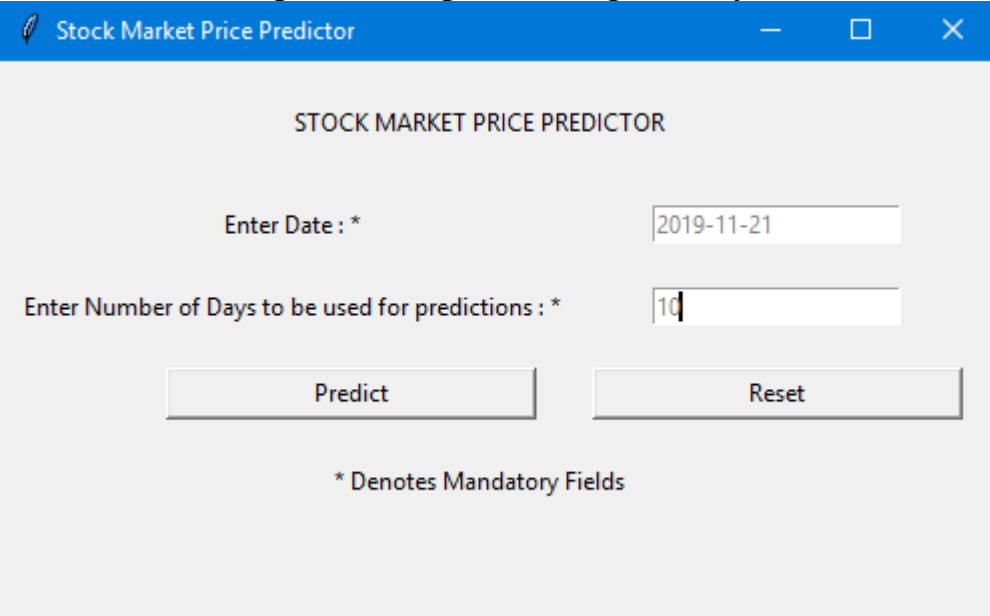
- ❖ After finding the best fitted model on the dataset, we added a simple GUI in our project to provide easiness to the user. We need to take 2 values input by the user 'Date' and the 'max number of samples' to be used in the predictions.



The screenshot shows a window titled "Stock Market Price Predictor". Inside, the title "STOCK MARKET PRICE PREDICTOR" is centered. Below it, there are two input fields. The first is labeled "Enter Date : *" and is empty. The second is labeled "Enter Number of Days to be used for predictions : *" and contains the text "number<=30". Below these fields are two buttons: "Predict" and "Reset". At the bottom, there is a note: "* Denotes Mandatory Fields".

Figure.3.13 GUI to take input from the user

- ❖ Now the task is to predict the price for a given day.



The screenshot shows the same window as Figure 3.13, but with sample input. The "Enter Date : *" field now contains "2019-11-21". The "Enter Number of Days to be used for predictions : *" field now contains "10". The "Predict" and "Reset" buttons remain at the bottom, along with the note "* Denotes Mandatory Fields".

Figure.3.14 Sample Input from the user

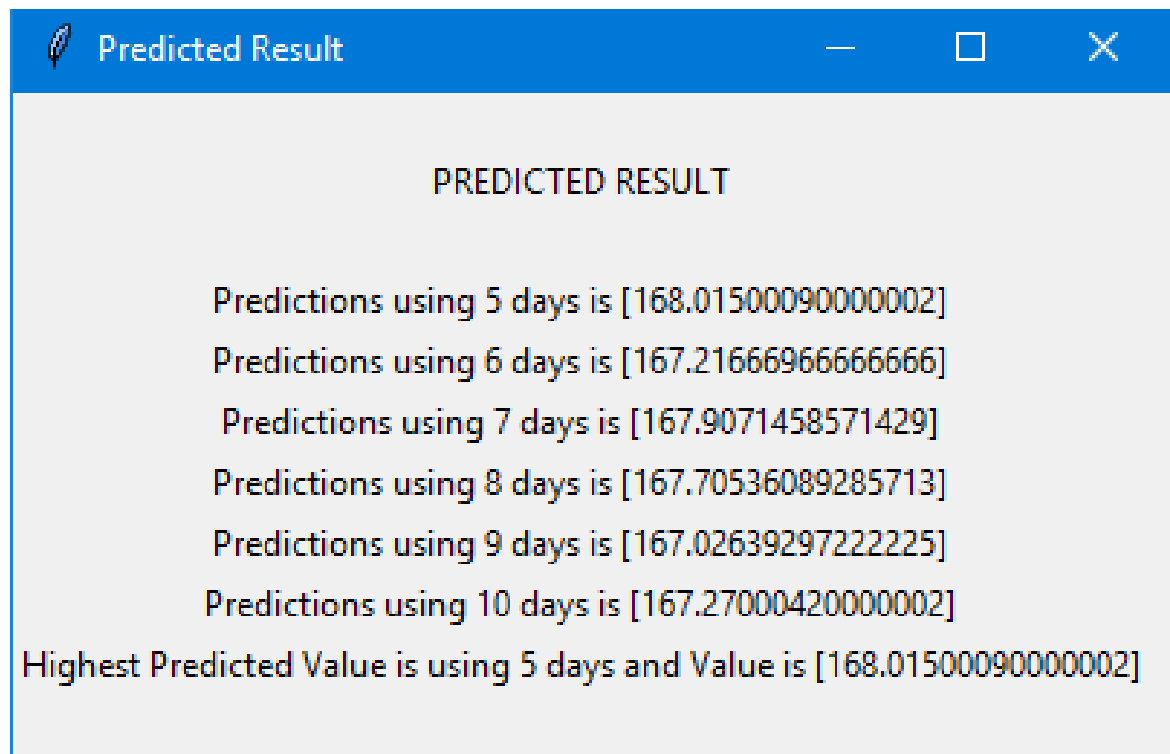
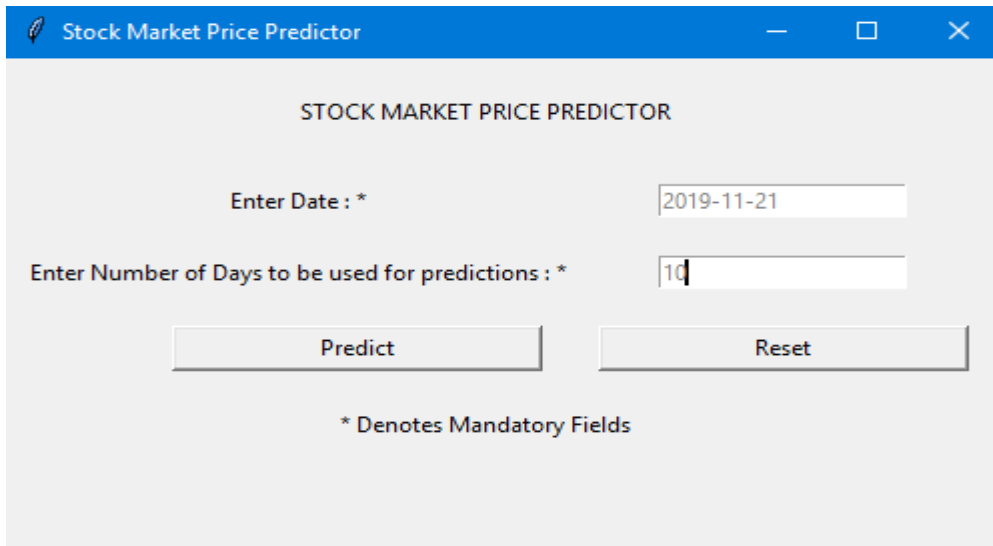


Figure.3.15 Output of the input values

CHAPTER 4

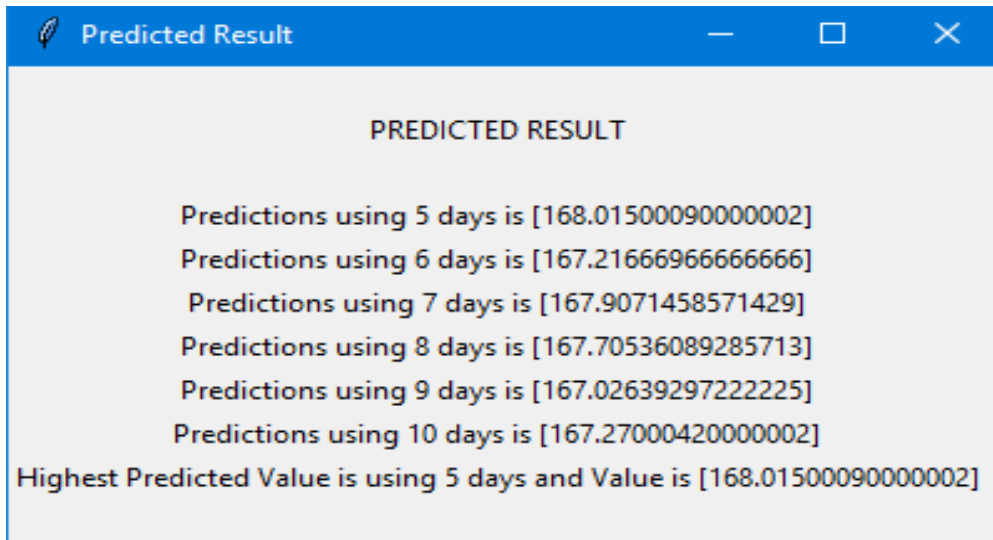
EXPERIMENT AND RESULT ANALYSIS

To predict the price for a day the user needs to enter the date of the desired day and the number of samples to be used to do the predictions. In the below figure, a sample input and output is shown.



The screenshot shows a window titled "Stock Market Price Predictor". Inside, the title "STOCK MARKET PRICE PREDICTOR" is centered. There are two input fields: "Enter Date : *" with the value "2019-11-21" and "Enter Number of Days to be used for predictions : *" with the value "10". Below these fields are two buttons: "Predict" and "Reset". At the bottom, a note states "* Denotes Mandatory Fields".

Figure.4.1 Sample Input



The screenshot shows a window titled "Predicted Result". Inside, the title "PREDICTED RESULT" is centered. The output text is as follows:

```
Predictions using 5 days is [168.01500090000002]
Predictions using 6 days is [167.21666966666666]
Predictions using 7 days is [167.9071458571429]
Predictions using 8 days is [167.70536089285713]
Predictions using 9 days is [167.02639297222225]
Predictions using 10 days is [167.27000420000002]
Highest Predicted Value is using 5 days and Value is [168.01500090000002]
```

Figure.4.2 Sample Output

In the output for a given input we have used the max number of samples given by the user and predicted the price using the given number of samples. We have provided the facility to user to decide how much number of days the user uses to know the price. On the basis of predictions using several days the user can decide whether to buy or sell the stock.

Also we show the result using the best number of samples which we found earlier and tell the user about same.

CHAPTER 5

CONCLUSION

5.1 Discussion

We have used linear regression technique to predict the trend of the stock price, which is based on the previous values provided to the model. As it is known that stock prices are very much dependant on the other factors also like economy of the country where the company is, total capital of the company, NPA of the company etc.

So it is not a good practice to be completely dependent upon the predictions done by the machine. The user also needs to look other factors as well. The user may use the machine for his help.

5.1.1 Limitations: In our project a limited dataset is provided to the model, so the predictions done by the model for a date that is after one or two months far from the last date in the provided dataset, the predictions will not be accurate because all the predictions will be done using the already provided dataset. So the model can't predict the values for the day which is after a month.

Also our model doesn't consider other factors that affect the stock market.

5.2 Future Work:

This project mainly focuses on analyzing data and predicting the price of the stock market on a particular date and using a given number of days. In this project there is an interface which provides convenience to the user to input date and number of days to be used for the predictions and a new window pops which gives the predicted price of the stock market.

We can automate the process of choosing the best number of samples using some techniques.

The project can be extended in terms of predicting the price of stock market using any provided dataset. This could be done by providing a pre-processed and cleaned dataset to the model in the form of '.csv' file.

REFERENCES

1. Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. "Stock price prediction using the ARIMA model." In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pp. 106-112. IEEE, 2014.
2. Selvin, Sreelekshmy, R. Vinayakumar, E. A. Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. "Stock price prediction using LSTM, RNN and CNN-sliding window model." In *2017 international conference on advances in computing, communications and informatics (icacci)*, pp. 1643-1647. IEEE, 2017.
3. Adebiyi, Ayodele A., C. K. Ayo, Marion O. Adebiyi, and S. O. Otokiti. "Stock price prediction using neural network with hybridized market indicators." *Journal of Emerging Trends in Computing and Information Sciences* 3, no. 1 (2012): 1-9.
4. Göçken, Mustafa, Mehmet Özçalıcı, Aslı Boru, and Ayşe Tuğba Dosdoğru. "Integrating metaheuristics and artificial neural networks for improved stock price prediction." *Expert Systems with Applications* 44 (2016): 320-331.
5. Lee, Jae Won. "Stock price prediction using reinforcement learning." In *ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570)*, vol. 1, pp. 690-695. IEEE, 2001.