NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES

BIG DATA ANALYTIC PROJECT REPORT

ANALYSIS OF COMPANY SHARES SELLING IN PAKISTAN STOCK EXCHANGE

SAAD ATHER - 21L-7289
MOHSIN JUNAID - 21L-7293
ARUZA ASIF - 21L-7307

Supervisor
Dr. Zareen ALAMGIR

May 27, 2022

Abstract

Forecasting market performance and understanding the mechanism of price discovery is inherent to develop trading strategies. This project examines the predictive power of machine learning algorithms in forecasting stock index in distributed environment. The project applies various machine learning algorithms and suggests the best model for forecasting stock index. We ran our data set of Allied Bank Limited and BATA Pakistan on different ML models including linear regression, gradient boosted tree, generalized linear regression and random forest regression techniques. We found that the linear regression model provides an accurate forecast of the stock market index for the highest price of share on particular day compared to others tested models. The outcome of this project facilitate the investors to use the appropriate model for forecasting and take an informed decision by considering the nature of stock market volatility.

1 Introduction

Forecasting market return and predicting volatility play a significant role in the spot and derivatives market. Extant literature shows earlier studies have used statistical and econometric tools for forecasting. However, the models face constraints to issues such as parameter optimisation, feature selection, overfitting and underfitting [9]. Machine learning algorithms can overcome the limitations of the traditional model and improve forecasting accuracy. Further, the algorithms bring flexibility in dealing with big data and also improve forecasting accuracy. We test the forecasting accuracy of linear regression, gradient boosted tree, decision tree and random forest on predicting the stock market of Pakistan shares shuch as ABL (Allied Bank Limited), BATA (BATA Pakistan), ENFERT (Engro Fertilizer), ATLS (Atlas Honda Pakistan). Closing index (C) of the day based on open (O), high (H) and low price (L). These parameters jointly named as (OHLC) is a type of chart used to analyse the daily price movement of stock, currencies and commodities.

2 Related work

Adebiyi et al. [1] examined the forecasting performance of ARIMA and artificial neural networks model. The study found that neural networks are superior to ARIMA model in forecasting. Feuerriegel and Gordon applied text based model to examine the effect of new information on the short and long term stock index forecasting. The research found the text-based models succeed in reducing forecast errors. Fischer and Krauss [4] applied long- and short-term memory networks (LSTM) for predicting SP 500 and found LSTM outperform random forest, a deep neural net and logistic regression classifier. Göçken et al. [5] used hybrid artificial neural network (ANN) models which consist in exploiting capabilities of harmony search (HS) and genetic algorithm (GA) are found to be the dominant model for stock market forecasting. Cited books were a lot help in giving us the sight

on how to implement the recent approaches in spark distributed framework [2],[3],[6],[8].

3 Problem Statement

There are many factors involved in the prediction, such as physical and psychological factors, rational and irrational behavior, and so on. All these factors combine to make share prices dynamic and volatile. This makes it very difficult to predict stock prices with high accuracy. The aim of this project is to help discover the future value of company stock price on short term/time bases since market price is changing very rapidly and there are many factors like economical and social news that may inflict the price and other financial assets traded on Pakistan stocks exchange. So by looking at the highest price of a particular share we can give the future insight beforehand.

4 Methodology

Our entire idea is based to invest on a company which has good selling trends and to gain significant profits based on the share that has a highest price. Our approach is to predict the highest price of share by applying the machine learning algorithms on the huge dataset of 10 years using the spark data framework in distributed environment. So that an investor can buy that share before hand on a low price before the price of share reaches its peak on that day and can make profit of it.

We will use machine learning algorithms like Linear Regression, Generalized Linear regression, Random forest regression and gradient boosting regression on the dataset Fig.1, available at [7], and apply in distributed environment on Google cloud Platform (GCP) clusters where there is one master node and three worker nodes. Then we will run a pyspark job on cluster nodes. Each node having 2 cores of CPUs with memory size of 7.5GB and disk size of 500GB. And finally We will analyze the results of each model.

4.1 DATA SET

In this project we are going to take the original data set of different companies having shares in Pakistan Stock Exchange like **BATA Pakistan, Allied bank limited, Engro Fertilizers and Atlas Honda** from Pakistan stock exchange database [7]. And use this data set in spark for processing, then we will applying ML models, and will try to predict the Highest price of any share on that particular day.

The data set contains features, described as follow:

- OPEN: Opening price of a share when market opened.

Symbol	Date	Open	High	Low	Close	Volume
ABL	16-Jan-12	55.75	55.75	55.01	55.12	2153
ABL	17-Jan-12	56	56.5	55.5	56.31	12493
ABL	18-Jan-12	57	57	56	56.06	96350
ABL	19-Jan-12	56.75	57.21	56.4	57	34994
ABL	20-Jan-12	57.7	59.49	57.7	58.86	104409
ABL	23-Jan-12	60.99	61.5	59.99	61.03	166522
ABL	24-Jan-12	60.5	61.45	60.5	60.6	162023
ABL	25-Jan-12	60.5	61.8	60.5	61.14	47646
ABL	26-Jan-12	61.5	61.84	60.9	61.04	35355
ABL	27-Jan-12	61.02	61.5	60	61.03	129068
ABL	30-Jan-12	61	61.3	60	60.47	48006
ABL	31-Jan-12	60	61.28	60	60.51	12815
ABL	01-Feb-12	60.3	61	60.3	60.9	13378
ABL	02-Feb-12	60.51	61.5	60.5	61	39154
ABL	03-Feb-12	60.5	61.11	60	60.6	21971
ABL	06-Feb-12	61.45	62.2	61	61.37	102551
ABL	07-Feb-12	61.5	62.25	61.5	62.03	102793
ABL	08-Feb-12	62.01	62.9	62	62.51	59944
ABL	09-Feb-12	62.55	63	62	62.12	32101
ABL	10-Feb-12	62.02	62.75	62	62	62070
ABL	13-Feb-12	63.39	63.39	60.5	60.63	78752

Figure 1: ALLIED BANK LIMITED SHARE DATASET

- CLOSE: Closing price of a share when market closed.
- LOW: Lowest price at which it was sell on that day.
- VOLUME: Total amount of trading activity on that day.

Our model will predict the highest price, described as follow:

- HIGH: Highest price at which it was sell on that day.

4.2 Methodology Work Flow

The data set can be seen in a figure 1. Upon loading the data into our spark framework we first cleaned the data and converted the string type datatype to standard datatype format like 'Date', and plotted the Highest price against the each year as shown in figure 2. We first make a copy of spark Dataframe and converted our copied spark Dataframe to pandas Dataframe then we used the library called matplotlib for graphically visualization of the data. In machine learning models we can not send the each column to Model for predicting, therefore we have to convert the feature columns to feature vectors and have to convert the categorical features to vector indexes.

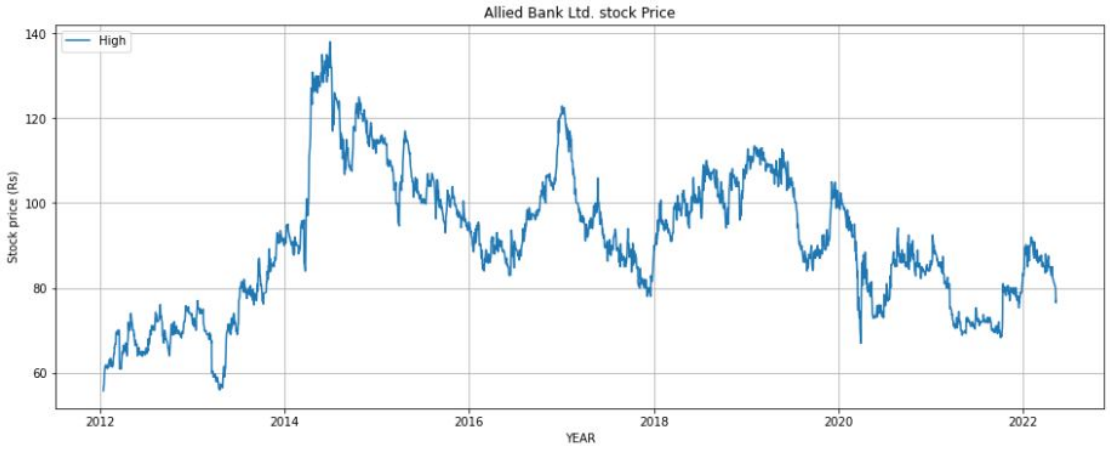


Figure 2: ALLIED BANK LIMITED SHARE HIGHEST PRICE PER YEAR

YARN helps to open up Hadoop by allowing to process and run data for batch processing, stream processing, interactive processing and graph processing which are stored in HDFS[6]. Here in our cluster of nodes, YARN had three node Mangers and one master node. YARN memory and its node managers are shown in figure 3.

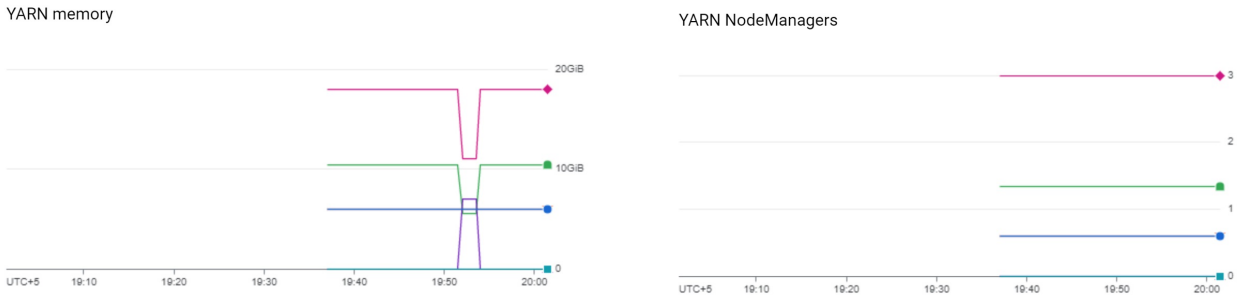
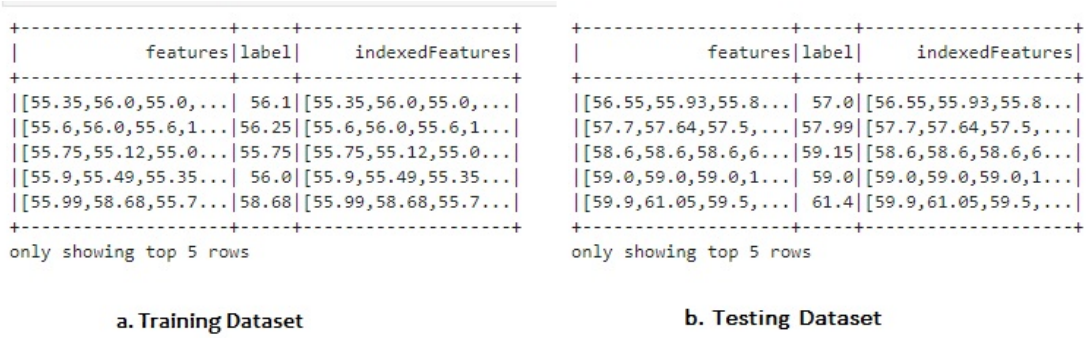


Figure 3: YARN Memory usage and nodeManagers in GCP

We split our dataset into training set and testing set with 80% and 20% of ratio. Then we created spark pipelines having feature-indexer, ML model and finally predicted the results. We use cross validation in case of Randomforest Regression for better predicting results. Figure 4. shows the training and testing dataset.

5 Results

Predicting the Highest stock price, considering we have a huge dataset of more than ten years it will take considerably a lot time to run on a local machine. So to overcome this issue Distributed file system comes in and a new framework was developed know as



```
+--------------------+-----+--------------------+
|            features|label|     indexedFeatures|
+--------------------+-----+--------------------+
|[55.35,56.0,55.0,...| 56.1|[55.35,56.0,55.0,...|
|[55.6,56.0,55.6,1...|56.25|[55.6,56.0,55.6,1...|
|[55.75,55.12,55.0...|55.75|[55.75,55.12,55.0...|
|[55.9,55.49,55.35...| 56.0|[55.9,55.49,55.35...|
|[55.99,58.68,55.7...|58.68|[55.99,58.68,55.7...|
+--------------------+-----+--------------------+
only showing top 5 rows
```

a. Training Dataset

```
+--------------------+-----+--------------------+
|            features|label|     indexedFeatures|
+--------------------+-----+--------------------+
|[56.55,55.93,55.8...| 57.0|[56.55,55.93,55.8...|
|[57.7,57.64,57.5,...|57.99|[57.7,57.64,57.5,...|
|[58.6,58.6,58.6,6...|59.15|[58.6,58.6,58.6,6...|
|[59.0,59.0,59.0,1...| 59.0|[59.0,59.0,59.0,1...|
|[59.9,61.05,59.5,...| 61.4|[59.9,61.05,59.5,...|
+--------------------+-----+--------------------+
only showing top 5 rows
```

b. Testing Dataset

Figure 4: a. shows Training Dataset b shows Testing dataset

Hadoop distributed file system. With further advancement in this field many frameworks were developed that were able to run the jobs or task on above of these HDFS. Spark is one of those recently developed framework that performs the tasks very efficiently.

So in this project we use spark framework and ran our models in distributed environment. We obtained the best results MSE of 0.69 when we used the Linear Regression for predicting. Further results are shown in Table 1 below.

Models	MSE Results
Linear Regression	0.69
Generalized Linear Regression	0.93
Gradient Boosting Regression	1.15
Random Forest Regression	0.99

Table 1: Models Result ran on GCP DataProc Jobs.

We ran all the above mentioned models separately on Google DataProc in separate Jobs. Figure 5 shows results of Linear regression and Gradient Boosting regression jobs ran on DataProc.

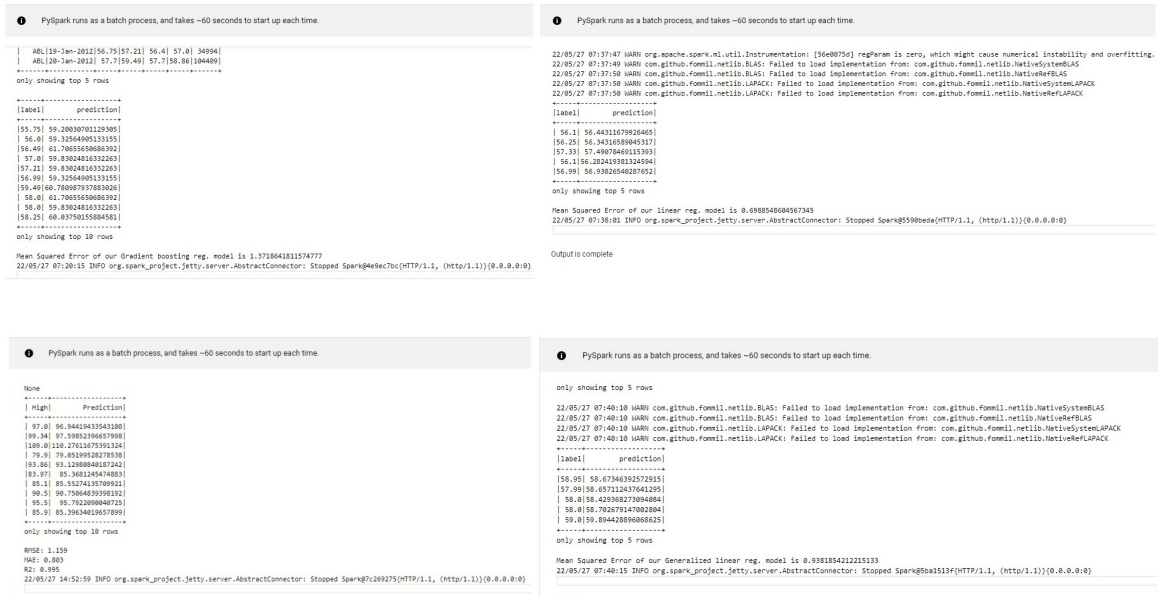


Figure 5: Top Left: Gradient Boosting, Top Right: Linear Regression Bottom left: RandomForest Regression, Bottom Right: Generalized Linear Regression Jobs on DataProc

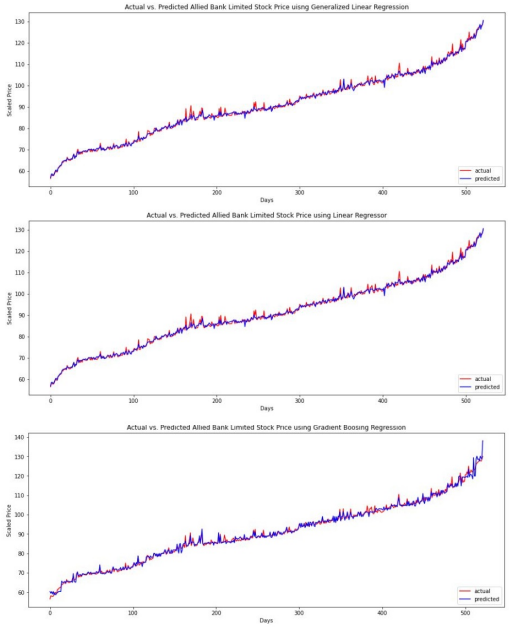


Figure 6: Predictions (Blue) vs Actual (Red)
Top: Generalized Linear Reg. Middle: Linear Reg. Bottom: Gradient boost.

6 Future Aspect and Discussion

For the model to be used efficiently in the financial market decision making, it should be tested across various stock exchanges data indexes. Key indicators would also be taken as a factor in the upcoming models, to improve the efficiency of the model. It is also expected to give more accurate results with least error by applying more advance Deep learning models. LSTM can give great results but to use it in distributed framework Elephas can be used which is a deep learning library using Kears in distributed environment. Also some modified ensemble technique can make the predictions more accurate.

Author's Contributions

Saad Ather contributions includes data extraction from Pakistan stocks database, data cleaning, applying linear, generalized linear and Gradient boosting regression for stock price prediction of Allied bank limited, plotting Graphs, running Jobs on GCP DataProc and report writing on Latex. *email l217289@lhr.nu.edu.pk*.

Mohsin Junaid contributions includes data loading in sparkframe work and data cleaning, features extraction and labeling applying Randomforest regression model on BATA Pakistan stock price prediction, evaluating results using Cross validation and finding Root mean square error of model, mean square error. *email l217293@lhr.nu.edu.pk*.

Aruza Asif contributions includes data loading in sparkframe work and data cleaning, features extraction and labeling, applying Randomforest regression model on Allied bank limited stock price prediction, evaluating results using Cross validation and finding Root mean square error of model, mean square error. *email l217307@lhr.nu.edu.pk*.

References

[1] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. "Comparison of ARIMA and artificial neural networks models for stock price prediction". In: *Journal of Applied Mathematics* 2014 (2014).

[2] Amrith Ravindra Ahmed Sherif. *Apache Spark Deep Learning Cookbook. Over 80 recipes of deep learning in distributed environment of Apache Spark*. Packt, 2018.

[3] Srinivas Duvvuri and Bikramaditya Singhal. *Spark for Data Science*. Packt Publishing Ltd, 2016.

[4] Thomas Fischer and Christopher Krauss. "Deep learning with long short-term memory networks for financial market predictions". In: *European Journal of Operational Research* 270.2 (2018), pp. 654–669.

[5] Mustafa Göçken et al. "Integrating metaheuristics and artificial neural networks for improved stock price prediction". In: *Expert Systems with Applications* 44 (2016), pp. 320–331.

[6] Guglielmo Iozzia. *Hands-on Deep Learning with Apache Spark: Build and Deploy Distributed Deep Learning Applications on Apache Spark*. Packt Publishing Ltd, 2019.

[7] Pakistan Stock Market. *Pakistan Stock Exchange Shares Database*. Ed. by KSE. URL: `http://www.ksestocks.com/QuotationsData`.

[8] Nick Pentreath. *Machine learning with spark*. Packt Publishing Birmingham, 2015.

[9] T Viswanathan and Manu Stephen. "Does Machine Learning Algorithms Improve Forecasting Accuracy? Predicting Stock Market Index Using Ensemble Model". In: *Advances in Distributed Computing and Machine Learning*. Springer, 2021, pp. 511–519.