BVM Engineering College

Project Report in Advanced Programming Lab

on

Share Price Prediction

Using Regression Method

Author:

18CP047 Pavan Gabani

18CP042 Mautik Donda

Supervisor:

Mr. Udesangsir Jaliya

Submission Date: Nov 5,2020

Abstract

The prediction of a share market direction may serve as an early recommendation system for short-term investors and as an early financial distress warning system for long-term shareholders. Forecasting accuracy is the most important factor in selecting any forecasting methods. Research efforts in improving the accuracy of forecasting models are increasing since the last decade. The appropriate stock selections that are suitable for investment is a very difficult task.The key factor for each investor is to earn maximum profits on their investments. We use Multiple Linear Regression methods to find predicted shares.This is a simple and Easy approach to predict shares next week. The results will be used to analyze the stock prices and their prediction in depth in future research efforts.

TABLE OF CONTENTS

1 INTRODUCTION

1.1. OBJECTIVE

In the past decades,there is an increasing interest in predicting markets among economists, policymakers,academics and market makers. The objective of the proposed work is to study And improve the supervised learning algorithms to predict the stock price.

Technical Objective

The technical objectives will be implemented in R. The system must be able to access a list of historical prices. It must calculate the estimated price of stock basedon the historical data. It must also provide an instantaneous visualization of the market index.

Experimental Objective

Two versions of prediction system will be implemented; one using Decision trees and others using Support Vector Machines. The experimental objective will be to compare the forecasting ability of SVM with Decision Trees. We will test and evaluate both the systems with same test data to find their prediction accuracy.

1.2. WHAT IS THE PROBLEM?

Investors are familiar with the saying, “buy low, sell high” but this does not provide enough context to make proper investment decisions. Before an investor invests in any stock, he needs to be aware how the stock market behaves. Investing in a good stock but at a bad time can have disastrous results, while investment in a mediocre stock at the right time can bear profits. Financial investors of today are facing this problem of trading as they do not properly understand as to which stocks to buy or which stocks to sell in order to get optimum profits.

Predicting long term value of the stock is relatively easier than predicting on a day-to-day basis as the stocks fluctuate rapidly every hour based on world events.

1.3 WHY THIS IS A PROJECT RELATED TO THIS CLASS?

The solution to this problem demands the use of tools and technologies related to the field of data mining, pattern recognition, machine learning and data prediction. The application will predict the stock prices for the next trading day. The requirements and the functionality of this application correlates it to the class.

1.4 WHY OTHER APPROACH IS NO GOOD?

The other approach makes use of Neural Networks. Neural networks have the following drawbacks:

1. Slow Convergence Rate The neural network takes a lot of time to train.

2. Local Minima and Maxima Neural Networks are based on gradient descent method to find the local extreme value and they have a tendency to get stuck on the local minima and maxima and therefore it is difficult to find global minima and maxima. In the approach previously discussed, the author has used pattern matching to overcome this problem.

1.5 WHY YOU THINK YOUR APPROACH IS BETTER?

The proposed approach makes use of Support Vector Machines (SVM) and Decision Trees. The benefit of using Decision trees over Neural Network are: 1. They are easy to program. 2. The top nodes in the tree will give the information about what data affects the prediction. 3. Trees are interpretable and provide visual representation of data. 4. Performs faster than Neural Networks after training. The benefits of using SVM over neural networks are [2]: 1. SVM has strong founding theory. 2. Global optimum guaranteed. 3. Requires less memory to store the predictive model. 4. Yield more readable results and a geometrical interpretation.

1.6 STATEMENT OF THE PROBLEM

Financial analysts investing in stock market usually are not aware of the stock market behavior. They are facing the problem of trading as they do not properly understand which stocks to buy or which stocks to sell in order to get more profits. In today's world, all the information pertaining to stock market is available. Analyzing all this information individually or manually is tremendously difficult. As such, automation of the process is required. This is where Data mining techniques help. Understanding that analysis of numerical time series gives close results, intelligent investors use machine learning techniques in predicting the stock market behavior. This will allow financial analysts to foresee the behavior of the stock that they are interested in and thus act accordingly. The input to our system will be historical data from Yahoo Finance. Appropriate data would be applied to find the stock price trends. Hence the prediction model will notify the up or down of the stock price movement for the next trading day and investors can act upon it so as to maximize their chances of gaining a profit. The entire system would be implemented in

1. Slow Convergence Rate The neural network takes a lot of time to train. 2. Local Minima and Maxima Neural Networks are based on gradient descent method to find the local extreme value and they have a tendency to get stuck on the local minima and maxima and therefore it is difficult to find global minima and maxima. In the approach previously discussed, the author has used pattern matching to overcome this problem. 1.5 WHY YOU THINK YOUR APPROACH IS BETTER? The proposed approach makes use of Support Vector Machines (SVM) and Decision Trees. The benefit of using Decision trees over Neural Network are: 1. They are easy to program. 2. The top nodes in the tree will give the information about what data affects the prediction. 3. Trees are interpretable and provide visual representation of data. 4. Performs faster than Neural Networks after training. The benefits of using SVM over neural networks are [2]: 1. SVM has strong founding theory. 2. Global optimum guaranteed. 3. Requires less memory to store the predictive model. 4. Yield more readable results and a geometrical interpretation. 1.6 S TATEMENT OF THE PROBLEM Financial analysts investing in stock market usually are not aware of the stock market behavior. They are facing the problem of trading as they do not properly understand which

stocks to buy or which stocks to sell in order to get more profits. In today's world, all the information pertaining to stock market is available. Analyzing all this information individually or manually is tremendously difficult. As such, automation of the process is required. This is where Data mining techniques help. Understanding that analysis of numerical time series gives close results, intelligent investors use machine learning techniques in predicting the stock market behavior. This will allow financial analysts to foresee the behavior of the stock that they are interested in and thus act accordingly. The input to our system will be historical data from Yahoo Finance. Appropriate data would be applied to find the stock price trends. Hence the prediction model will notify the up or down of the stock price movement for the next trading day and investors can act upon it so as to maximize their chances of gaining a profit. The entire system would be implemented in Python/Java and R language using open source libraries. Hence it will effectively be a zero cost system.

1.7 AREA OR SCOPE OF INVESTIGATION

This project requires investigation in the following areas: Stock Market Investigating trends in stock market and factors affecting the stock prices. Data mining techniques Investigating the available tools and techniques for data mining and then selecting those that are best fit to solve the problem.

2 METHODS AND METHODOLOGIES

2.1. WHAT IS MULTIPLE LINEAR REGRESSION (MLR)?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

2.2. Formula and Calculation

$$yi = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + ... + \beta p xip + \epsilon$$

where, for i=n observations:

yi =dependent variable

xi=explanatory variables

β_0 = Y-intercept (Constant)

β_p = slope coefficient for each explanatory variable

2.3. WHAT MLR CAN TELL YOU

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- y_i observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R^2 by itself can't this be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.[1]

When interpreting the results of multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

2.4. HOW WE USE MLR IN R PROJECT

In our project, to find share of our company for next week, we find relation between our past values of different factors. So, our project is based on MLR.

3 IMPLEMENTATION RESULTS & DISCUSSION

2.1. CODE

```
library("ggthemes") #theme for graph

library(tidyverse) #graph

mydata <- read.csv('E:\\Lab\\R\\project.csv', header = TRUE) # reading file

print(head(mydata)) #default 10


#training ####################

model <- lm( EstimatedSharesOutstanding ~ CashRatio + CurrentRatio + FixedAssets +
ForYear + EarningsPerShare +

        TotalAssets + TotalCurrentAssets + TotalCurrentLiabilities + TotalEquity +
Investments + NetCashFlow +

        ProfitMargin + TotalRevenue , data = mydata )

print(summary(model)$coef)


#PRACTICAL with ORIGINAL Data

newdata <-
data.frame(CashRatio=58,CurrentRatio=115,FixedAssets=1156000000,ForYear=2015,Earnin
gsPerShare=4,  TotalAssets=8869000000,TotalCurrentAssets=1817000000,
```

```
TotalCurrentLiabilities=1583000000,TotalEquity=2183000000,Investments=-
35000000,NetCashFlow=683000000,ProfitMargin=12, TotalRevenue=6282000000)

predicted_Estimated_Shares_Outstanding<-predict(model,newdata)  # prediction

print(predicted_Estimated_Shares_Outstanding)


#ACCURACY

print("accuracy :")

accuracy<-predicted_Estimated_Shares_Outstanding/183255152.7 #this is ans of ORIGINAL
DATA which is known

print(accuracy)

#GRAPH PLOTTING


print(ggplot(mydata, aes(x=EstimatedSharesOutstanding)) + geom_point(aes(y = CashRatio),
color = "steelblue")
                                        + geom_point(aes(y = CurrentRatio), color = "darkred")
                                        + geom_point(aes(y = ProfitMargin), color = "black")
                                        + ggtitle("Share market")
                                        + theme_wsj()+ scale_colour_wsj("colors6"))
```

2.2. RESULT

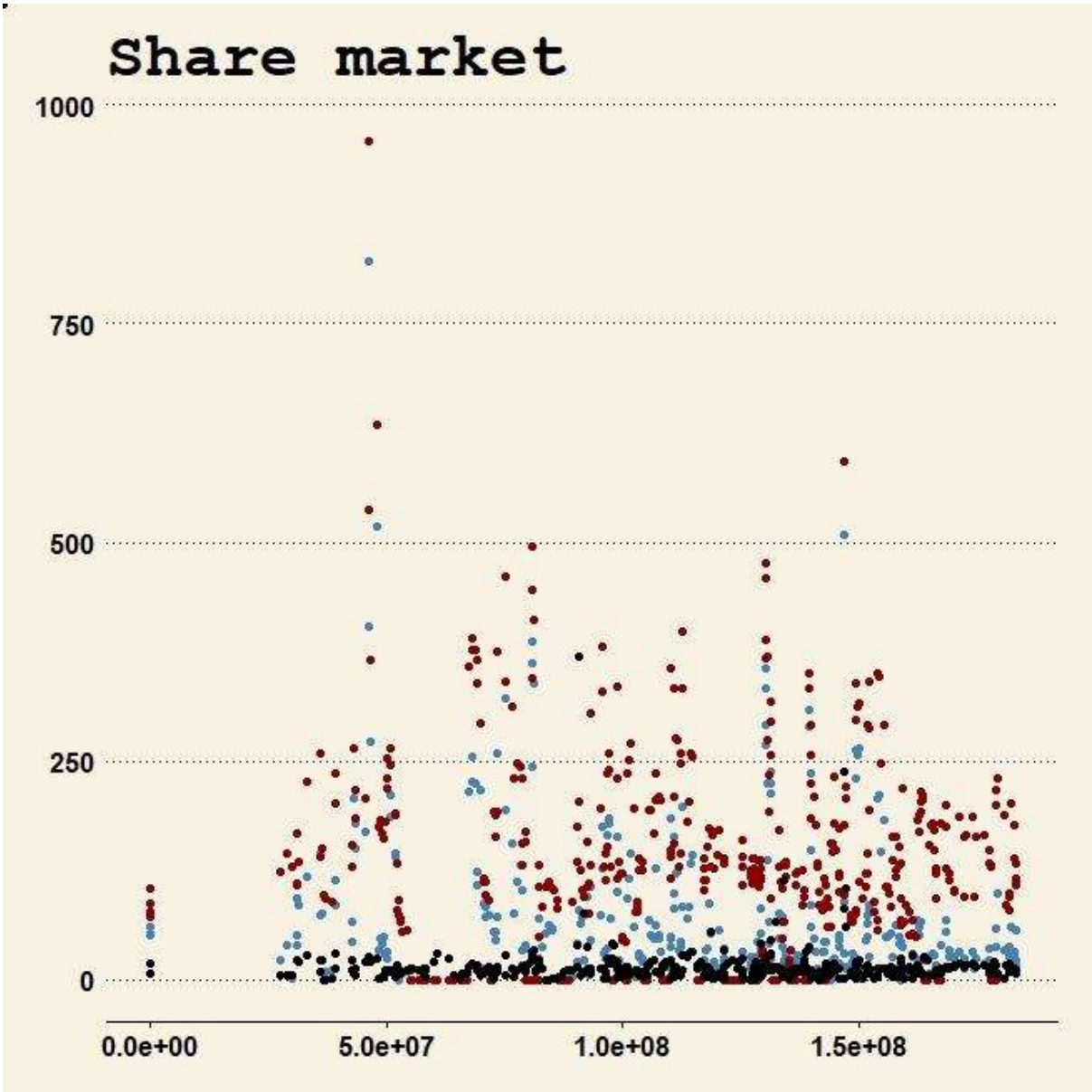
	index	CashRatio	CurrentRatio	FixedAssets	ForYear	EarningsPerShare	TotalAssets
1	0	53	78	13402000000	2010	-5.60	23510000000
2	1	75	104	19259000000	2010	-11.25	42278000000
3	2	60	88	23084000000	2010	4.02	43225000000
4	3	51	73	27510000000	2010	11.39	48415000000
5	4	23	124	1292547000	2010	5.29	4613814000
6	5	40	144	1286034000	2010	5.36	5564774000

	TotalCurrentAssets	TotalCurrentLiabilities	TotalEquity	Investments
1	7072000000	9011000000	-7987000000	3.060e+08
2	14323000000	13806000000	-2731000000	-1.181e+09
3	11750000000	13404000000	2021000000	1.799e+09
4	9985000000	13605000000	5635000000	4.430e+08
5	3184200000	2559638000	1210694000	0.000e+00
6	3989384000	2764785000	1516205000	0.000e+00

	NetCashFlow	ProfitMargin	TotalRevenue	EstimatedSharesOutstanding
1	197000000	8	24855000000	0
2	660000000	7	26743000000	0
3	-146000000	7	42650000000	0
4	-604000000	19	40990000000	0
5	540210000	6	6205003000	27672157
6	514360000	6	6493814000	28884799

Coefficient-

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.947531e+10	6.099791e+08	-81.10985243	2.983079e-284
CashRatio	-2.469591e+04	1.198370e+04	-2.06079176	3.985556e-02
CurrentRatio	2.756669e+04	8.714587e+03	3.16328088	1.657886e-03
FixedAssets	4.000318e-05	3.938906e-05	1.01559110	3.103305e-01
ForYear	2.462914e+07	3.029586e+05	81.29537116	1.062544e-284
EarningsPerShare	7.346099e+04	8.892187e+04	0.82612968	4.091367e-01
TotalAssets	-1.926068e-06	6.439785e-06	-0.29908887	7.650004e-01
TotalCurrentAssets	4.803084e-05	1.030152e-04	0.46624988	6.412460e-01
TotalCurrentLiabilities	-3.601650e-06	1.676652e-04	-0.02148120	9.828706e-01
TotalEquity	9.156195e-06	5.655914e-05	0.16188710	8.714622e-01
Investments	-1.944374e-04	1.040290e-04	-1.86906963	6.221560e-02
NetCashFlow	1.228919e-05	2.433024e-04	0.05050996	9.597368e-01
ProfitMargin	-3.224006e+04	2.357830e+04	-1.36736160	1.721453e-01
TotalRevenue	-4.434746e-05	2.754965e-05	-1.60972863	1.081078e-01



X = EstimatedSharesOutstanding

y = CashRatio , color = "blue"

y = CurrentRatio , color = "red"

y = ProfitMargin, color = "black"

Predicted_value – 153907699

4 CONCLUSION

4.1 CONCLUSION

By using Multiple Linear Regression mode, we have found the accuracy

"accuracy" : 0.8398547

We try one data and measure accuracy of output with original output, which is than 83%.

REFERENCES

In this project, to create efficient code and retrieve result, we use many References,defines below:

https://www.kaggle.com/lesibius/selecting-stocks-from-predicted-p-e-ratio

https://datatofish.com/multiple-linear-regression-in-r/

https://www.investopedia.com/terms/m/mlr.asp

********** END OF REPORT **********