Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression

Author: Ashray Anand

November 22, 2020

1 Abstract

The key purpose behind this study is to use logistic regression (LR) and various financial ratios and accounting ratios as independent variables to investigate indicators that significantly affect the performance of stocks as dependent variable, actively traded on the Indian stock market. The study sample consists of the ratios of 50 large market capitalization companies (part of NIFTY index) over a five-year period for annual data from 2015-2020. The study identifies and examines eight financial ratios that can classify the companies into two categories good or poor based on their rate of return. The model developed could enhance an investor's ability of stock price forecasting and enable them to select out-performing stocks. Macroeconomic variable is not considered to forecast stock return performance in this paper.

2 Introduction

Recently predicting stock market trends is gaining more consideration, perhaps because of the fact that if the trend of the market is successfully forecasted the traders may be well directed. The profitability of trading in the stock market to a large extent rest on the predictability. More over the forecast trends of the market will support the regulators of the market in taking corrective measures. Forecasting in stock market have been built on traditional statistical prediction methods. Ordinary least square method (OLS) have been the midpoint of all traditional method.But, these methods have rarely proved successful owing to the presence of non-linearity and noise in the time series. Nonlinear methods suggests an innovative way of observing stock prices, and it proposes new methods for empirically measuring their nature. This new technique proposes that past prices help determine future prices, but not in a straightforward way.

The objective of this paper is to apply statistical methods to survey and analyze financial data in order to develop a simplified model for interpretation. This study aims to develop a model for classifying stocks into two categories (good or poor), based on their rate of return. A company's stock is classified as good if its share returns perform above the market returns provided by the National Stock Exchange composite index of India; i.e., the NIFTY. In this study, the logistic regression (LR) method has been used to classify selected companies, based on their performance and stock market trends. Logistic model is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable(i.e., a class label) based on one or more predictor variables (features). The LR method is used to predict the probability of good stock performance by fitting the variables to a logistic curve. Thus, LR is used to classify a set of independent variables into two or more mutually exclusive categories. It involves finding a linear combination of independent variables that reflect large differences in group means. The model will use the preprocessed data set of NIFTY Index. The data set encompassed the trading days from 31st March, 2015 to 31st March, 2020 from CMIE database. Various new ratios, such as sales growth, Cash Earnings per Share, Book Value, Price/Cash Earnings Per Share, Price/Earning, Profit Before Interest Depreciation and Tax/Sales, Sales/Net Assets and Price/Book value have been included in this discipline for share valuation. This study, based on financial ratios of companies, tries to develop a logistic regression model which may helpful for potential investors to look after stock performance of companies before investing. This study aims, therefore, to answer two questions.

1. Can the returns of stocks be explained with the help of different financial ratio?
2. Can we analyze stock returns using a logistic regression model?

2.1 Problem Statement

Analysis of Model-Logistic Regression

Logistic regression is used in our study because we assume that the relation between variables is non-linear. Also this type of regression is preferred when the response variable is binary which means that can take only two values 1 or 0. Logistic regression could forecast the likelihood, or the odds ratio, of the outcome based on the predictor variables, or covariates. The significance of logistic regression can be evaluated by the log likelihood test, given as the model chi-square test, evaluated at the $p < 0.05$ level, or the Wald statistic. Logistic regression has the advantage of being less affected than discriminant analysis when the normality of the variable cannot be assumed.

It has the capacity to analyze a mix of all types of predictors. Logistic regression, which assumes the errors are drawn from a binomial distribution, is formulated to predict and explain a binary categorical variable instead of a metric measure. In logistic regression, the dependent variable is a log odd or logit, which is the natural log of the odds. In the logistic regression model, the relationship between Z and the probability of the event of interest is described by this link function.

$$p_i = \frac{e^{zi}}{1+e^{zi}} = \frac{1}{1+e^{-z}}$$

$$z_i = \log\left(p_i/1 - p_i\right)$$

Where p_i is the probability the i^{th} case experiences the event of interest z_i is the value of the unobserved continuous variable for the i^{th} case

The z value is the odds ratio. It is expressed by

$$z_i = c + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Where x_{ij} is the j^{th} predictor for the i^{th} case β_j is the j^{th} coefficient.P is the number of predictors.βs are the regression coefficients that are estimated through an iterative maximum likelihood method. However, because of the subjectivity of the choice of these misclassification costs in practice, most researchers minimize the total error rate and, hence, implicitly assume equal costs of type I and type II errors.

In order to carry out logistic regression analysis, first a method is needed for classifying returns as a "1" or "0" for a given day. In this study we use a method that is simple and objective, if the value of a return in day j is above the return in day j-1, it is noted as a "1"; otherwise, it is classified as a "0".

The study uses the LR for prediction of stock performance because of the following reasons:

- In LR, linearity assumption between dependent and independent variables does not exist. Rather, it is S-shaped curve.
- No need of normality assumption for independent variables is required. Also variables can be of interval scale or ratio scale or nominal scale.
- The groups (dichotomous or multinomial) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one the groups.

Application of Logistic Regression

Dataset description: The sample in this study is based on Top 50 companies of NSE, popularly known as NIFTY50, on the basis of Market Capitalization for every year from 2015 to 2020. The relevant data taken for this analysis are from CMIE database. The study consists of a sample size of 66026 distinct companies' daily observations(data points).

Sample Data Set (2 Observations)

co-stkdate	company-name	nse-closing-price	nse-returns	nse-pe	nse-pb	ceps-on-stkdate	equity-bv-on-stkdate	ca-net-sales	ca-sales-net-fixed-assets	ca-pbdita
31-03-2000	ASIAN PAINTS LTD.	423.14	0.97	17.34	4.75	31.34	3574.00	14699.60	376.72	2435.00
31-03-2000	BHARAT PETROLEUM CORPN. LTD.	35.0500	1.01	3.13	0.38	14.72	1484.40	393423.00	531.43	31115.80

Link: Referance to Datasource

The main objective of this study is to test the model efficacy on prediction based on the financial ratios for examining the out-performing firms. Specific objectives are:

1. To study the effect of each financial ratio in determining the performance of the Company's Stock.
2. To examine the efficiency of financial ratios in the suggested Logistic regression model.

Table 2, shows the dichotomous classification of stock performance.Table 1, shows the 8 Independent Variables (Financial and Accounting Ratios). As a dichotomous dependent variable, stock performance was classified as POOR=0 and GOOD=1, to signify the investment choice Top 50 firms listed in NSE.

Table 1: The eleven independent (predictors) variables are

Name of the variable	Description of the variable
CEPS	Cash Earnings Per share
NS	Percentage Increase in Net Sales
BV	Book Value
PECEPS	Price/Cash Earnings Per Share
PE	Price/Earning
PBIDTS	Profit Before Interest Depreciation and Tax/Sales
SNA	Sales/Net Assets
PEBV	Price/Book value

Table 2: The dependent variable is

Type of Company (based on stock market return)	Description of the variable
GOOD	Return above Market return
POOR	Return below Market return

Table 3: Dependent Variable Encoding

Original Value	Internal Value
GOOD	1
POOR	0

The final logistic regression equation is estimated by using the maximum likelihood estimation for classifying the stock performance:

$$Z_{it} = c + \beta_1 CEPS_{it} + \beta_2 NS_{it} + \beta_3 BV_{it} + \beta_4 PECEPS_{it} + \beta_5 PE_{it} + \beta_6 PBIDTS_{it} + \beta_7 SNA_{it} + \beta_8 PEBV_{it} + v_{it}$$

where: z$^{=}$ $\log\left(\frac{p}{1-p}\right)$ and p ' is the probability that the outcome is GOOD.

Exploratory Data Analysis on Stocks data

	nse_closing_price	nse_returns	nse_pe	nse_pb	ceps_on_stkdate	equity_bv_on_stkdate
count	1346.000000	1346.000000	1346.000000	1346.000000	1346.000000	1346.000000
mean	1233.257318	1.000908	61.309557	15.407737	23.824559	77048.814413
std	320.216482	0.016593	7.043813	1.561011	5.159708	18375.976863
min	698.550000	0.859700	45.813600	11.800400	15.386000	42302.500000
25%	969.900000	0.990000	56.418225	14.259950	20.719700	64269.000000
50%	1165.525000	1.000000	61.486400	15.458650	21.747500	77981.500000
75%	1423.012500	1.010000	65.117675	16.465125	27.695200	94522.000000
max	2031.200000	1.088500	85.394700	20.075100	35.155300	110448.100000

Figure 1: Some Description of Data

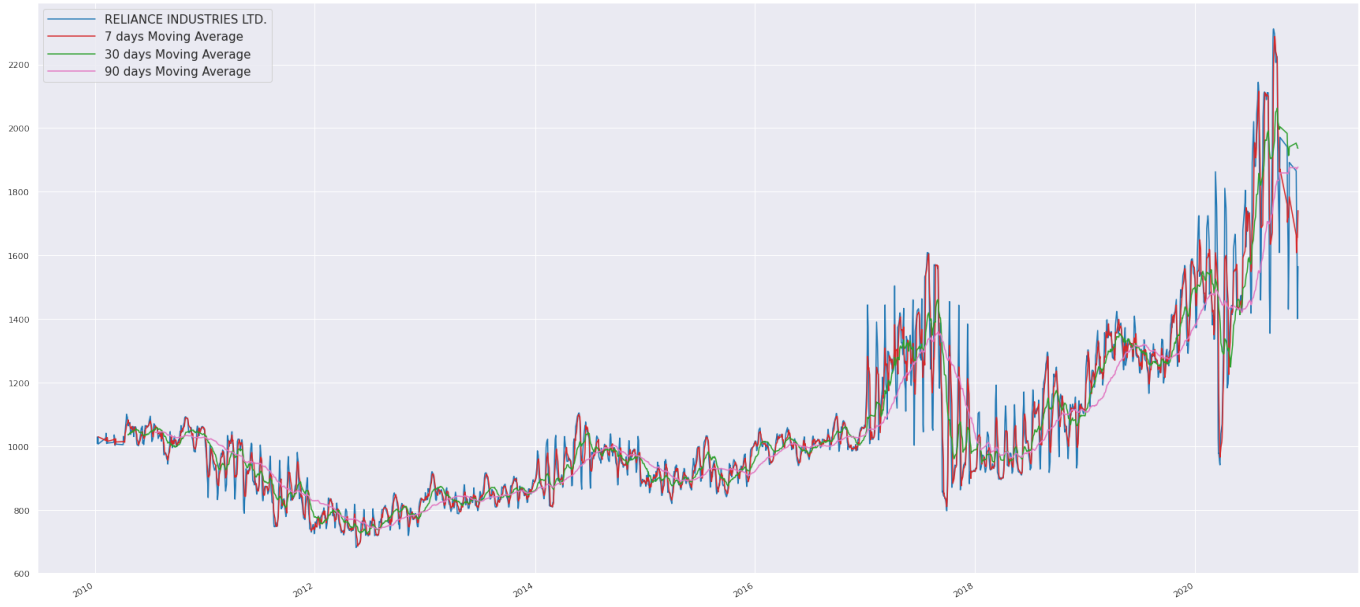


Figure 2: Closing Price of RELIANCE INDUSTRIES LTD. from 2015-20

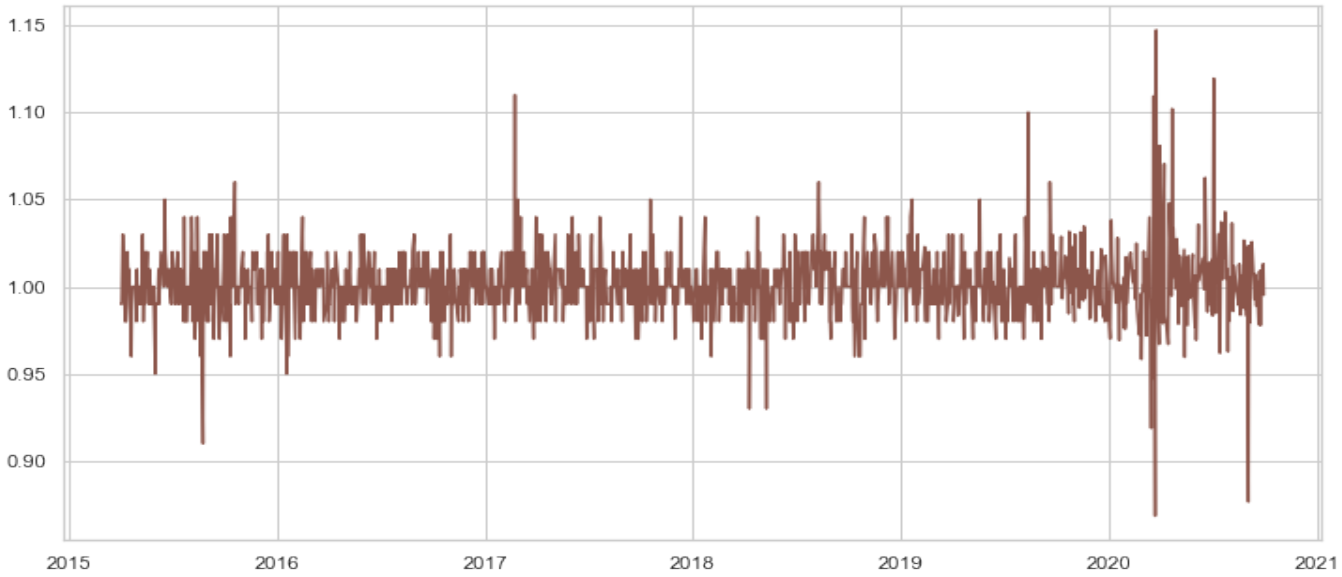


Figure 3: Returns of RELIANCE INDUSTRIES LTD. from 2015-20

Above charts shows the adj. closing price and returns of all companies from 1st Jan 2020 to 30th Oct. 2020. It can be observed that, all companies fell drastically from start of march 2020 (Due to Covid-19). It can be seen that, RELIANCE INDUSTRIES LTD. fell the lowest in mid- April.

Kurtosis tells the 'fatness' of the tail and it is important because it tells how 'extreme' can the values get. In our case, the value is positive, so this indicates that the chance of 'extreme' values are rare.
Positive values for the skewness indicate data that are skewed right.

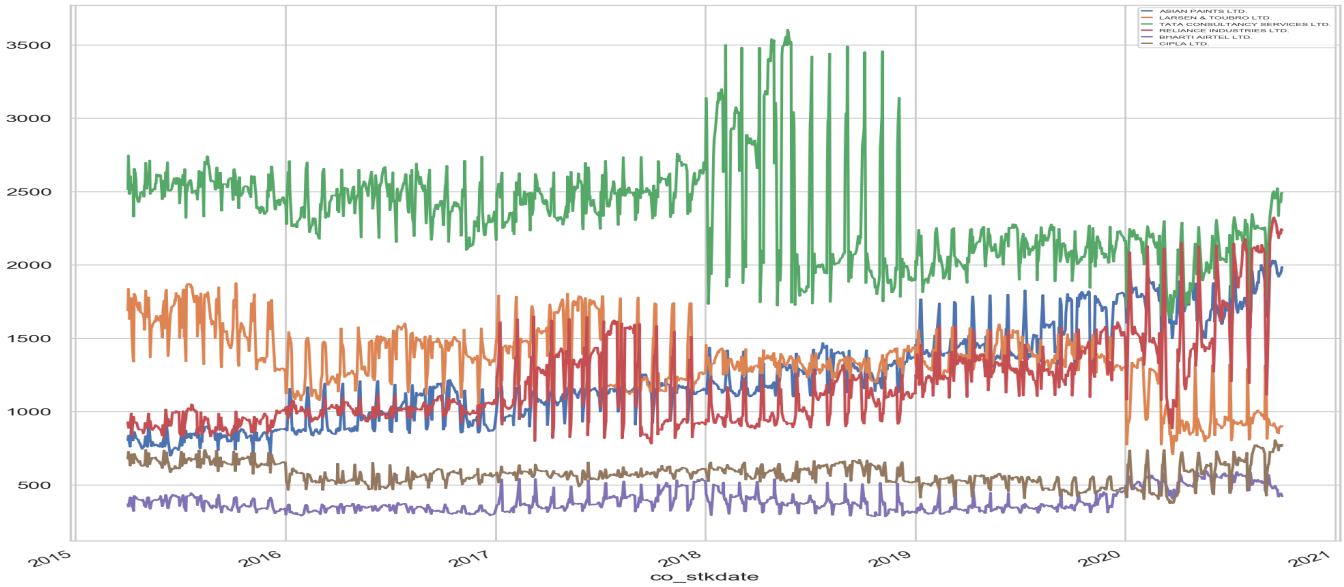


Figure 4: Closing price of Top 5 companies for 2015-20

The behaviour of one of the major Indian stock indices, Nifty, over the last five years is shown in Figure 6. It can be seen from Figure 6 that the stock index gradually rises to higher levels (based on countryâs growth story as well as impact of global stock market movement) but also many bumps with sudden large variations have been noticed on the

growth road of the market.

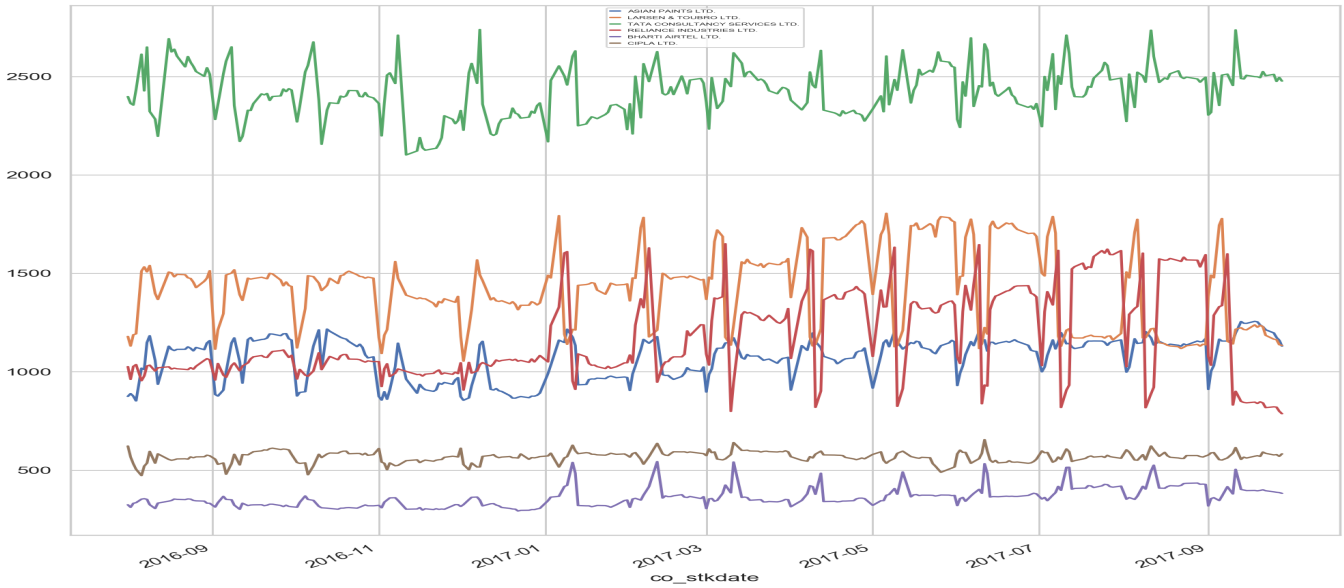


Figure 5: Closing price of Top 5 companies for 2016-17

The Government of India declared demonetization on 8th November 2016. NIFTY dropped 541.30 points. But slowly the market recovered. This can be verified from Figure 7, as closing price fell sharply around that period for 5 companies that is being studied. But later on they recovered. In the following year i.e. in 2017, another major economic event took place, the so called Goods Services Tax (GST) was launched on July 2017. As widely anticipated stock market did a good job and no major (sudden) change in Nifty due to implementation of GST has been noticed.

3 Methodology

The paper contains the practical implications of using the binary logistic regression method to predict the probability of good stock performance by analyzing relation between financial ratios and stock performance of the firms. The research examines sales growth, Cash Earnings per Share, Book Value, Price/Cash Earnings Per Share, Price/Earning, Profit Before Interest Depreciation and Tax/Sales, Sales/Net Assets and Price/Book value for the prediction of stock performance. We will use CMIE database for this study which include data of 50 large market capitalization companies (listed in NIFTY) over a period for daily data from 2015-2020.

For the purpose of carrying out logistic regression analysis, first a method is required for classifying a company as a good or poor investment choice for a given year. Although there is no definitive method for defining a market investment as good or poor, in this study we will use a method that is simple and objective namely, if the value of a companyâs stock over a given year rose above market return, it is classified as a good investment option; otherwise, it is classified as a poor investment option. Here, the NIFTY return has been taken as proxy for market return. To obtain the return, the closing price for each day is used. The return is calculated using the following formula:

$$\text{Return of stock } = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100$$

where, $t =$ Price in the t^{th} day

$$\mathrm{P}_{t-1} = \text{ Price in the } (t-1)^{\text{th}} \text{ day}$$

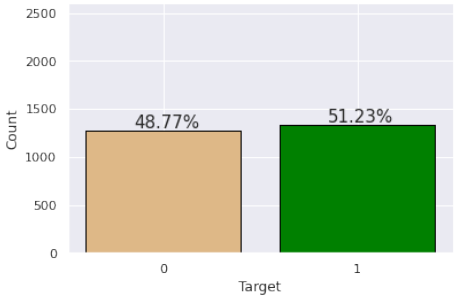
4 Simulation Study

4.1 Feature engineering

Features Rescaling:

nse_traded_qty= nse_traded_qty/10000
equity_bv_on_stkdate = equity_bv_on_stkdate/1000000
equity_shares_on_stkdate=equity_shares_on_stkdate/1000000000
nse_no_of_trans=nse_no_of_trans/10000

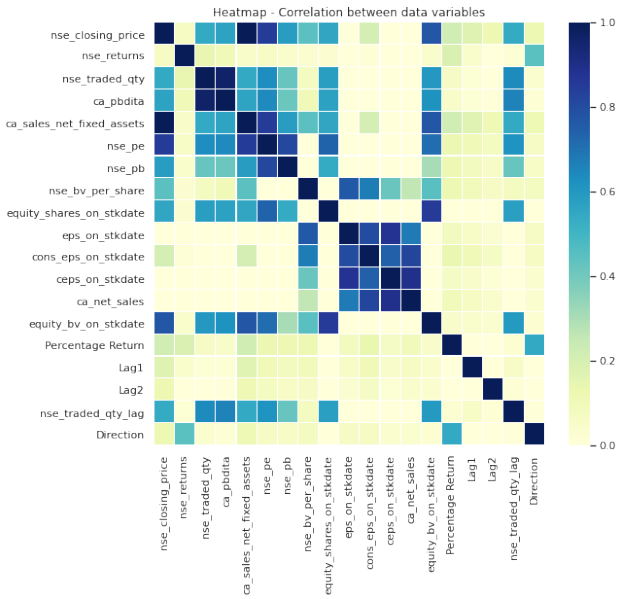
Class Imbalance



There is a balanced distribution between the number of times the price went up (class 1) or down (class 0), as expected for this kind of financial data.

Figure 6: Visualisation of Class Imbalance

Features Selection: Next, the features correlation is analysed:



From the correlation analysis is interesting to see how the additional features (Lag1, Lag2 and nse_traded_qty_lag) are linearly correlated with the closing price and net_sales, pbdita and sales_net_fixed_assets have low correlation, so we removed these features from our subsequent models. Preliminary tests shown that these additional features significantly helped the model to predict the stock behaviour. This analysis can help to select other technical indicators or market indexes in the future.

Figure 7: Correlation coefficient heat map

Regression Models:

We used Logistic regression and **Artificial Neural Network to** capture and represent complex input/output relations.In ANN, Logistic function is chosen as transfer function for both hidden layers to the output layer because of the dependent variable is bounded [0,1].

1. **Logistic Regression**

Results:

Features	Without Lag	With Lag
nse_closing_price	0.002214	0.005295
nse_returns	0.193416	2.618142
nse_traded_qty	0.000525	0.000467
nse_pe	0.229114	0.075120
nse_pb	0.025442	-0.020840
nse_bv_per_share	0.003565	0.002421
equity_shares_on_stkdate	-0.675631	-0.845687
eps_on_stkdate	0.060058	0.023996
cons_eps_on_stkdate	-0.097556	-0.092790
ceps_on_stkdate	-0.021964	-0.022647
equity_bv_on_stkdate	-0.309857	0.009751
Lag1	0	-4.69492012
Lag2	0	-2.12193313
nse_traded_qty_lag	0	0.000094

Table 4 : Regression Coefficients

	precision	recall	f1-score	support
0	0.61	0.26	0.37	308
1	0.54	0.84	0.66	315
accuracy			0.55	623
Macro avg	0.58	0.55	0.51	623
weighted avg	0.58	0.55	0.51	623

Table 5 : Performance Metrices for Logistic Regression Model

2. **Artificial Neural Network (ANN) model:**

To further improve the performance matrices, neural network is used. The number of input and output layers were set. The network offers only limited possibilities of adaptation, with one hidden layer is capable, with a sufficient number of neurons to approximate all continuous function. A second hidden layer takes into account the discontinuities or detect relationships and interactions between variables. In our study the use of hidden layers is very useful to detect all non-linear relationships between variables in the model. After several experimentation, 4 layers were used. Next step is to determine the number of neurons in layers. Large number of nodes, allows better match the data presented but decreases capacity generalization of the network. The size of the hidden layer must be either equal to that of the input layer or is equal to 75% of the latter or is equal to the square root of the number of neurons in the input and output layer. In the study we chose the approach of setting a maximum number of nodes in each hidden layer. Then, we eliminated those have no utility for the learning procedure. So, the number of input nodes is equal to number of independent variables which is 11. And the number of output nodes is equal to number of dependent variables which is 1. The mean absolute error (MAE) was used to study the performance of the trained forecasting models for the testing years and Mean square error (MSE) is used for evaluating the prediction accuracy of the model.

The network is trained by repeatedly presenting it with both the training and test data sets. To identify when to stop training, two parameters, namely the MAE and MSE of both the training and test sets were used. After 70 iterations, the MAE and MSE of the training set was 0.2352 and 0.1183 respectively, while those of test set was 0.2660 and 0.1368.

The training was stopped after 70 iterations as there was no significant decrease in both parameters. Thus, any further training was not going to be productive or cause any significant changes. The prediction accuracy of test data is 71% .

	precision	recall	f1-score	support
0	0.66	0.82	0.73	308
1	0.77	0.58	0.66	315
accuracy			0.70	623
Macro avg	0.71	0.70	0.70	623
weighted avg	0.71	0.70	0.70	623

Table 6 : Performance Metrices for ANN Model

3. **Logistic regression with AR Variables:**

   At last we created AR models by creating Lag1 and Lag2 variable on 'percentage return' variable and nse_traded_qty_lag for nse_traded_qty variable.

   Result of Logistic regression with AR variables:

	precision	recall	f1-score	support
0	0.67	0.26	0.37	308
1	0.55	0.88	0.67	315
accuracy			0.57	623
Macro avg	0.61	0.57	0.52	623
weighted avg	0.61	0.57	0.53	623

Table 7 : Performance Metrices for Logistic regression Model with AR Variables

We can see that there is some definite improvement with respect to Logistic regression model created without Auto-Regressive Lag variables.

4. **ANN model with AR variables:**

   Now to further improve the above result, we will simulate ANN, with dataset with Lag variables with same parameters that we used earlier in ANN.

	precision	recall	f1-score	support
0	0.87	0.91	0.88	308
1	0.75	0.73	0.73	315
accuracy			0.82	623
Macro avg	0.81	0.82	0.80	623
weighted avg	0.82	0.82	0.80	623

Table 8 : Performance Metrices for ANN with AR Variables

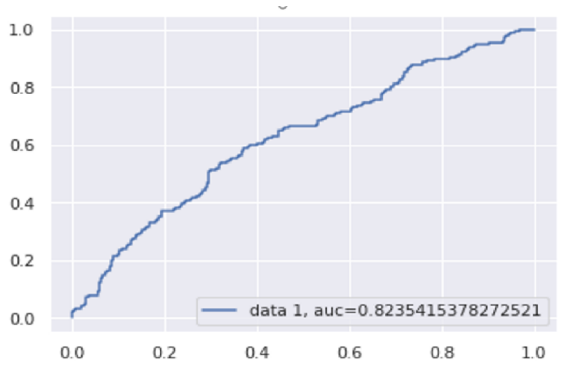


Figure 8: (ROC) curve

ROC Curve Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the trade-off between sensitivity and specificity. AUC score for the case is 0.82. AUC score 1 represents perfect classifier, and less than 0.5 represents a worthless classifier.

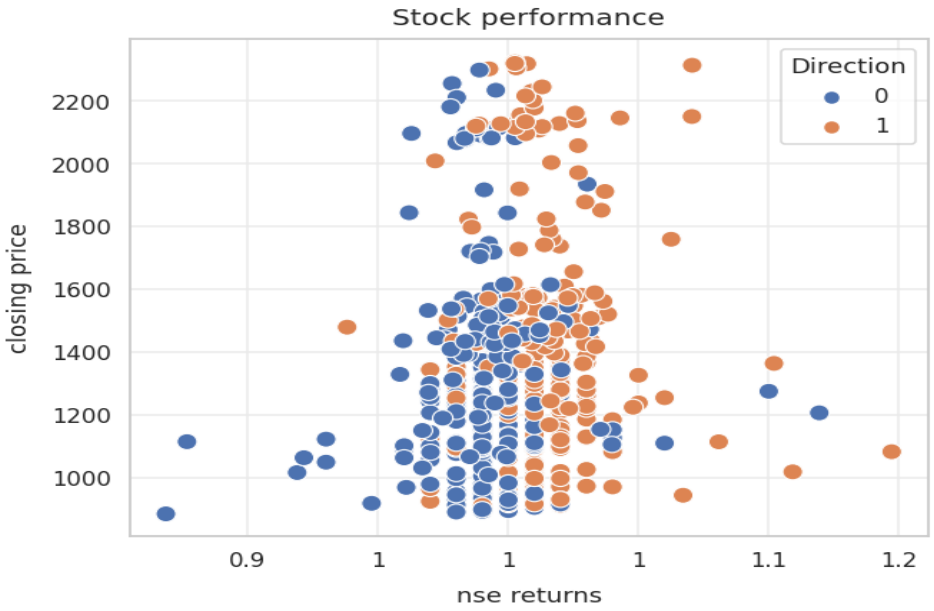


Figure 9: Visualization of two features (nse returns and closing price) and two class labels (price went up as 1 or down as 0)

5 Conclusion

In this paper, an attempt is made to explore the use of logistic regression to determine the factors which significantly affect the evolution of the stock index. Logistic regression method helps the investor to form an opinion about the time to invest. It may be observed that 10 variables and lag features can classify up to 71 % into two categories up and down. This prediction rate is very good, so it can be used for prediction with higher accuracy. This study has also employed neural network to predict the direction of stock index return. Multi- layer perceptron network is trained using Tan-sigmoidal algorithm. The prediction accuracy of the model is high for test data (82 %). We can deduce from this study that the use of logistic regression and neural network give us very close result. But, both methods give us a good results of prediction with very high accuracy using a mixture of fundamental and technical variables. So this study must be extended to use these methods in the prediction and classification of stock returns.

5.1 Scope for Further Study

The present study is based on financial ratios as the solitary factor influencing stock prices of banks, but there may be several other fiscal and organizational factors that may also influence stock prices. Therefore, the scope for further research lies in focusing on usage of quarterly or monthly data and diverse type of qualitative data for evaluating stock performance. The contemporary study deliberated logistic regression to construct a model, but for advance studies, a plethora of other methodologies can be considered to increase prediction ratio.

References

Hands-On Machine Learning with Scikit-Learn, Keras Tenserflow (Aurelien Geron 2nd Edition) Chapter 4,5,6 .

Dutta, A., Bandopadhyay, G., Sengupta, S. (2008). Classification and Prediction of Stock Performance using Logistic Regression. An Empirical Examination from Indian Stock Market: Redefining Business Horizons: McMillan Advanced Research Series, 46-62. view at Google scholar

Guresen, Erkam, et al. 2011. Using artificial neural network models in stock market index prediction, Expert Systems with Applications 38 (8),10389-10397.