Stock Price Prediction with Linear Regression

V. Sellam
Assistant Professor,SRM Institute of Science and Technology, Chennai, India

Ayush Sharma
Student, SRM Institute of Science and Technology, Chennai, India

Kishan Agarwalla
Student,SRM Institute of Science and Technology, Chennai, India

Vikram Sinha
Student,SRM Institute of Science and Technology, Chennai, India

Abstract – This paper looks at the hypothesis and routine with regards to relapse procedures for expectation of stock value slant by utilizing a changed informational collection in ordinal information arrange. The first pretransformed information source contains information of heterogeneous information types utilized for treatment of cash esteems and monetary proportions. The information organizes in cash esteems and budgetary proportions give a procedure at calculation of stock costs. The changed information set contains just an institutionalized ordinal information compose which gives a procedure to gauge rankings of stock value patterns. The results of the two procedures are inspected and evaluated. The essential plan depends on relapse examination from WEKA machine learning programming. The stock value development in Bursa Malaysia is utilized as our exploration setting. The information sources are corporate yearly reports which included accounting report, salary articulation and income proclamation. The factors incorporated into the informational index were framed in light of securities exchange exchanging key examination approach. Classifiers in WEKA were utilized as calculations to create the results. This investigation demonstrated that the results of relapse strategies can be enhanced for the expectation of stock value incline by utilizing a dataset in institutionalized ordinal information organize.

Keywords- Techniques of regression; ordinal data type; machine learning ; linear regression; classifiers.

1. INTRODUCTION

Securities exchanges furnish openings alongside related dangers for financial specialists to make benefits. Researchers and expert speculators have done many research and created hypotheses and speculation on securities exchange timing methods. This paper investigate the use of relapse methods as prescient logical on stock value patterns. The detailing of dataset is in light of principal examination approach which is a prevailing school of thought in contributing. The highlights utilized in the dataset are factors comprised of factual proportions. In this investigation, all information in numerical qualities are changed into ordinal or identified qualities to frame the dataset. Relapse based classifiers from WEKA are then utilized as prescient examination to test the ordinal information. The results were analyzed and assessed.

2. LITERATURE REVIEW

Since the presence of securities exchanges, a great deal of research had been done in creating models to make forecasts on stock value developments. Two models are outstanding and they are the proficient market speculation and the arbitrary walk hypothesis.In the nineteen sixties, Eugene Fama composed a PhD exposition on the productive market theory. Fama's contention demonstrated that stock cost will be estimated properly also, mirror all accessible data in a functioning business sector. Since advertise is productive, there is no quality investigation can be conveyed out to result in predominant execution of a suitable benchmark. Louis Bachelier's PhD paper titled "The Theory of Theory" was an examination on the irregular walk speculation's idea. The speculation contends that securities exchange cost develops as per arbitrary walk and that the market stock cost can't be anticipated. The theory is in accordance with the efficientmarket theory. Proficient financial specialists support two predominant schools of thought on contributing which are central investigation and specialized investigation. Principal examination approach recognizes forthcoming stocks by investigating their basic properties. Insights from budgetary reports, for example, monetary record, money book and benefit furthermore, misfortune articulation are utilized to consider the characteristic estimations of organizations. Money related proportion insights that incorporate working execution, corporate valuation, development balance, budgetary use and corporate liquidity frame the premise of essential traits. A related procedure of key investigation is known as the contrarian procedure. This procedure joins factor of human enthusiastic inclinations with basic examination. A contrarian trusts that over response of ravenousness and dread in swarm conduct prompts exploitable mispricing in stock costs. A contrarian speculator takes an opposite position in purchasing offers of stocks that are performing inadequately and after that offering them when they perform well. Specialized examination approach distinguishes outline designs based on an organization's chronicled share cost. This methodology does not gain knowledge into the business side of an organization; it expect the accessible open data does not offer an aggressive exchanging advantage. This procedure predicts slants ahead of time through diagram designs. The examination on securities exchange expectation methods has in the end moved into the mechanical domain. Machine learning approach is one of the regular strategies. The approach of machine learning is by looking at a possibly straight or non-direct relationship exists with the accessibility of enough markers. Machine learning is a part of fake insight. This methodology discover designs in preparing datasets what's more, shape their own guidelines which are then utilized for making gauges in testing datasets. Relapse procedures are a piece of the machine learning approach. In 1805, Legendre distributed the technique for slightest squares, which was the most punctual type of relapse. In 1821, Gauss distributed a further improvement of the hypothesis of slightest squares which incorporate the Gauss-Markov hypothesis. In the nineteenth century, Francis Galton utilized the expression "relapse" to depict an organic marvel. Galton's work was later formed into the factual setting by Udny Yule and Karl Pearson. Regular relapse examination includes contributions of numerical information which may comprise of vast or an extensive variety of qualities. In this exploration, we begin by social event numerical information in realvalued arrange utilizing the principal examination approach.

3. DATA AND METHODOLOGY

A. Data

Data of companies

Table 1-Data collections

No.	Stock code	Duration
1	3689	2003-2010
2	3255	2004-2011
3	3921	2003-2010
4	4707	2003-2010
5	4065	2003-2010
6	7084	2004-2011
7	4588	2003-2010
8	5584	2003-2010
9	3107	2004-2011
10	9466	2005-2010
11	6033	2004-2011
12	6599	2004-2010
13	5032	2003-2010

Table 2-Dataset 1

Feature	Data Type
NTA	Real-valued
LA	Real-valued
DE	Real-valued
ZS	Real-valued
AT	Real-valued
Price	Real-valued

Table 3- Dataset 2

Feature	Data type
NTA	Ordinal
LA	Ordinal
DE	Ordinal
ZS	Ordinal
AT	Ordinal
PriceRank	Ordinal

The kinds of factors under examinations were recognized in view of crucial investigation approach. The chose autonomous factors are Net Tangible Asset (NTA), Liquid Resource (LA), Debt to Equity (DE), Altman Z-Score (ZS) and Resource Turnover (AT). Net Tangible Asset is a proportion of the unmistakable worth of an organization, short any immaterial resources. This is one conceivable proportion of an organization's offer worth. Fluid Resource is a benefit that can be immediately changed over into money. It is a great pointer on the money related quality of an organization. Obligation to Value proportion estimates the level of obligation with respect to the value of an organization. This estimation demonstrates whether an organization has nearly nothing or over introduction to obligation. Altman Z-Score joins five money related proportions to decide the likelihood of chapter 11 for an organization. Resource Turnover estimates the capacity of a organization in turning out deals in view of its accessible resource. A organization which can create deals productively shows higher introduction to benefit . The reliant variable is named Price for Dataset 1. The Value variable contains the genuine esteemed information of the anticipated cost. The reliant variable is named PriceRank for Dataset 2. The PriceRank variable contains the positioning of the anticipated value incline in clear cut ordinal esteem. Every one of the factors contain either a positive or a negative an incentive for Dataset 2. A positive esteem suggests positive relationship on value slant (cost positioning) and the other way around for a negative esteem. The results of both Price and PriceRank factors are subject to their associations with alternate factors.

B. Limitations of Data

The information for every one of the factors in Dataset 2 comprise of just the values in the scope of {-2, - 1, 1, 2}. The scope of qualities for look into results could increment if the information go comprises of a Zero esteem. The same is valid if more qualities are incorporated into the information go. The start of the money related revealing period contrasts among the organizations.

This is basic since the bookkeeping rehearse enables organizations to pick their own bookkeeping periods as long as every one of the bookkeeping time frames comprises of a year. Most gathered information were dated from year 2003 to year 2011. A few information were dated from 2004 or 2005 on the grounds that the information were not accessible preceding those periods.

C. Methodology

Insights on companies and dataset highlights are produced through essential examination. Information was screened and preprocessed to evacuate out-of-bound qualities. This procedure can anticipate issues of delivering deluding results .

The target of the change procedure is to make the information more organized. Pre-prepared information contains general realvalued information which incorporate the cash sum and rate groups. In the information change process, the pre-handled information is institutionalized into rate situated information. A rate to ordinal transformation table contains the scopes of rate values related with their ordinal specified qualities. Each counted esteem is relegated to the dataset in light of the change table. This methodology gives indistinguishable classifications of counted values for various factors despite the fact that the range of qualities for one factor varies from another variable. This approach likewise elucidates the classifications inside a variable where its numerical qualities swing generally starting with one territory then onto the next go. Amid the information preparing stage, in a steady progression each relapse classifier was utilized as prescient systematic on the dataset. A rate split determines a relapse classifier to split the dataset into preparing information and testing information relatively. Preparing information gives learning procedure to each classifier to define its very own relapse rules. The relapse lead was utilized on the testing information for forecasts of future stock value patterns. The test outcome was then assessed.

In a relapse show, an anticipated esteem Y is identified with a

capacity of x and b.

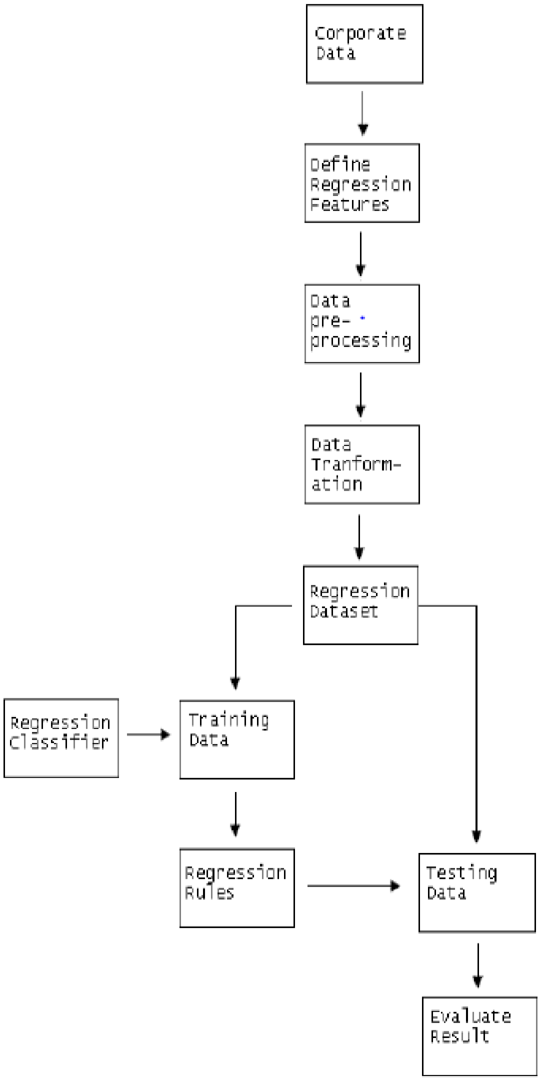
$Y = f(x, b)$ (1)

Y is the needy variable, x is the autonomous variable furthermore, b is the obscure parameter. A straight relapse demonstrate takes the shape

$Y = b0+b1 x1+ ... +bnxn+ e$ (2)

where x1 to xn are autonomous factors, and e is known as the blunder term. A direct relapse condition written in vector frame is

$Y= a+bx+e$ (3)



4. FINDINGS AND DISCUSSION

A. Findings

The accompanying demonstrates the test outcomes and assessment measures for relapse classifier exhibitions in light of two datasets with various information composes:

TABLE3 -Dataset 1 having Real-valued Data Type

Regression Algorithm	**Correlation Coefficient**	**Mean Absolute Error**	**Root Mean Square Error**
Additive Regression	0.3747	15.2275	19.0615
Linear Regression	0.3755	14.4419	18.5975
Regression by Discretization	0.0397	18.8868	24.8674
Simple Linear Regression	0.3049	15.3287	19.3164
SMO Regression	0.4173	14.2697	19.2356

TABLE 4-Dataset 2 with Ordinal Data Type

Regression Technique	**Correlation Coefficient**	**Mean Absolute Error**	**Root Mean Square Error**
Additive Regression	0.5336	0.7317	0.8641
Linear Regression	0.5742	0.7141	0.8360
Regression by Discretization	-0.0378	0.9491	1.3084
Simple Linear Regression	0.5742	0.7141	0.8360
SMO Regression	0.6079	0.5462	0.8164

Dataset 1 contains the first wellspring of information in genuine numbers. Dataset 2 contains the changed qualities in ordinal shape from dataset 1. Test on Dataset 2 shows that outcome has enhanced when changed ordinal information is utilized. Analyze to Dataset 1, everything except one relapse systems utilized on Dataset 2 indicate higher connection coefficient rates and lower blunder rates. Among the relapse strategies utilized on Dataset 2, the SMO Regression strategy showed a sensible outcome.

The system acquired a relationship coefficient of 0.6079. Likewise, the system has the most minimal mistake rates among the relapse systems with a mean outright blunder of 0.5462 and a root mean square mistake of 0.8164. The additional common occurrence of the 2008 worldwide budgetary emergency had added to nature swing of stock costs however out the world amid the period. Therefore, the connections among stock costs had been profoundly influenced and contorted.

A. Discussion

Each regression classifier is learned from the training data. The classifier’s rule is a regression function of the independent variables which produces the estimation target. SMO Regression model equation-

Forecast =

- 0.042 * (standardized) NTA

- 0.0184 * (standardized) LA

+ 0.0208 * (standardized) DE

+ 0.2803 * (standardized) ZS

- 0.0409 * (standardized) AT

+ 0.7783

The capacity of "(4)" is a likelihood appropriation that portrays the likelihood of irregular factors taking certain values. In this SMO Regression demonstrate, at first the PriceTrend takes the straight relapse shape from "(2)" as demonstrated as follows:

Y = b0+b1 x1+ ... +bnxn+ e

The estimation of b0 is zero and the estimation of e is 0.7783. To finish everything of the direct relapse condition, the SMO Regression standardized the factors where the yields depend on the institutionalized information [14]. Subsequently, the capacity of "(4)" is shaped. The "+" signs indicate positive relationships for the parameters, while "- " signs mean negative connections for the parameters in "(4)". The understandings of the connections change starting with one contributing methodology then onto the next. Contrarian approach has inclination to go specifically against the exploration discoveries by the ordinary stock experts who communicate their sees . Relapse procedures are helpful apparatuses for foreseeing the estimation of a needy variable in light of the estimations of other free factors. The procedure includes a straight change of the chose parameters into the anticipated variable. The choice of parameters depends on the slightest squares basis to accomplish an ideal choice run.

5. CONCLUSION

By this research we found that regression techniques outcomes can be improved by using the input data as standardized into a common data type.The use of an ordinal data type for prediction based on ranking system provides a different dimension for predicting outcomes regression techniques in the experiment.The outcomes of the regression techniques can be improved by using different data types.The outcomes are favorable when less structured data are transformed into more structured data in ordinal form. Because of availability of different data types,more research can be conducted for prediction of stock price trend.

REFERENCES

[1] Ashish Sharma ; Dinesh Bhuriya ; Upendra Singh (2017), “Survey of stock market prediction using machine learning approach”, International conference of Electronics, Communication and Aerospace Technology (ICECA)

[2] B.V. Phani ; B. Chandra ; Vijay Raghav (2011), “Quest for Efficient Option Pricing Prediction model using Machine Learning Techniques”, The International Joint Conference on Neural Networks

[3] Osman Hegazy ; Omar S. Soliman ;Mustafa Abdul Salam (2013), “A Machine Learning Model for Stock Market Prediction”, International Journal of Computer Science and Telecommunications