

Analyzing and forecasting of stock index price applying machine learning techniques

Kausalya Kola¹, Kesavarao Seerapu²

¹M.Tech Scholar(CSE) in Department of Computer Science and Engineering,

²Assistant Professor, Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vizianagaram (Dist), Andhra Pradesh, India.

Abstract

Machine Learnings mining is a significant subject in the investigation of information mining Data mining is the way toward finding substantial, valuable and reasonable example in information. Because of the huge size of data sets, significance of data put away, and important data acquired, finding concealed examples in information has gotten progressively huge. A period arrangement informational collection comprises of groupings of qualities or occasions that change with time. Time arrangement information is mainstream in numerous applications, for example, the every day shutting costs of an offer in a securities exchange. Stock information mining assumes a significant part to imagine the conduct of monetary market. AI calculations can be utilized to find all thing affiliations (or rules) in a dataset that fulfill client indicated requirements, for example least help and least certainty. Since just a single least help is utilized for the entire information base, it is certainly accepted that all things are of a similar sort as well as have comparative frequencies in the information. Examples are assessed by methods for creating itemsets utilizing a predefined backing and Machine Learnings with a higher certainty level. The example created by the continuous itemset of size three is discovered to be same as being reflected by methods for acquired Machine Learnings. The example so created causes speculators to assemble their portfolio and utilize these examples to study venture arranging and monetary market.

Key words- Machine learning, data mining, explicit model.

I. Introduction

In stock trading markets, pay improvement rate is a fundamental pointer for financial specialists to foresee the creating firms later on. The creating firms represent explicit firm that can advancing make and more get builds benefit per offer [1]. Financial exchange recommender system considering Machine

Learning (ARM) that endorses a course of action of stocks. The objective of this recommender structure is to help securities exchange dealers, particular monetary trained professionals and resource bosses in their decisions by proposing interest in a social affair of significant worth stocks when strong affirmation of possible profit by these trades is open [2]. Stock rundown measuring is one of the huge activities of cash related firms and private theorists in making adventure decisions. Since securities exchanges is an amazing, extraordinary, and nonlinear component structure its gauge is considered as a testing task [3]. Stock is the most popular of cash related business area instruments. It tends to be portrayed as a sign of capital interest of a man or an endeavor in an association or a limited danger association [4]. Stock expenses depend on upon various segments, the basic ones being the market assumption, execution of the business, procuring results and extended profit, takeover or consolidation, presentation of another item or presentation of a current item into new business sectors, share repurchase, declarations of profits or rewards, expansion or expulsion from the file [5]. Financial exchange expectation has been a district of remarkable eagerness due to the ability of getting an uncommon yield on the put money in a short period of time [6]. The straightforward accessibility to information bases with recorded costs makes the weak structure market adequacy test a less requesting undertaking to realize differentiated and the profitability preliminary of the other two structures [7]. The divulgence of explicit models, which are brand name for time course of action and therefore have an exceptional hugeness, is a basic task in the locale of time plan assessment, and similarly, the disclosure of certain model, to be explicit of Machine Learnings, is an indispensable task in the examination of significant worth based data [8]. To envision risk and return factors decisively, the fruitful parts should be recognized. In all honesty, one of the main points of interest of stock desire layout lies on the most ideal approach to pick specialist features for gauge [9]. Developmental calculation generally

creates individuals with a particular ultimate objective to assemble their health, and it procures the ideal or close ideal individuals which address movement rules when they are used as a piece of expert control structures [10]

II. Related work

Mining Machine Learnings is one of the main field of use of Data Mining. A bunch of client exchanges on things is given and the primary intention is to decide the relationships inside the deals of things. Mining Machine Learnings, is known as Market Basket Analysis, which is additionally an application field of Data Mining. It is basic to look at the client' buy conduct and help with expanding the deals and save stock by zeroing in on the retail location exchange information. This functions as a wide region for the analysts to build a superior information mining calculation. This section examines a review about the current Machine Learning.

Zhixin et al., [1] suggested an improved arrangement method dependent on Predictive Machine Learnings. Order Dependent Predictive Machine Learnings (CPAR) is one of the kinds of affiliation arrangement technique which coordinates the advantages of acquainted grouping and regular standard based characterization. For age of the standard, CPAR is more productive than the regular principle based grouping, since a large portion of the duplicate count is overlooked and numerous literals can be chosen to make various guidelines simultaneously. Despite the fact that the advantage referenced above stays away from the recreate computation in standard age, the expectation measures have the detriment in class rule dissemination irregularity and interference of wrong class rules. Further, it is ineffectual in cases that fulfill no principles. To evade these challenges, the creator suggests Class Weighting Adjustment, Center Vector-based Pre-arrangement and Post-handling with Support Vector Machine (SVM).

Bartik [2] introduced affiliation based characterization for social information and its utilization in web mining. Arrangement as per the mining Machine Learnings is a superior and human justifiable order plot. The goal of the creator is to constrain an adjustment of the crucial affiliation based characterization strategy that can be utilized in social affair information from the Web pages. The change of the strategy and fundamental discretization of numeric qualities are given.

Mining Interesting Rules by Association and Classification Algorithm is advanced by Yanthy et al., [4]. The primary reason in information mining is to uncover concealed information from the information and various procedures have been recommended up until now. Nonetheless, the drawback is that a couple of segments of the made standards would bear some significance with some gave client. Along these lines, a few estimates like certainty, uphold, lift, data and gain, have been proposed to find the best or exceptionally fascinating standards. Be that as it may, a few procedures are incredible at making rules high in one intriguing quality measure however not extraordinary in other intriguing quality measures. The association between the methods and intriguing quality proportions of the made guidelines isn't evident at this moment. The creator considered the association between the methods and the intriguing measures. The creator utilized manufactured information with the goal that the outcome isn't restricted to specific circumstances.

Wang et al., [5] plays out an examination on Machine Learnings Mining dependent on metaphysics in internet business. Business activities did with the utilization of Internet are extremely famous. A lot of exchange logs are delivered, which assists with gathering valuable data through information mining. Subsequently, Machine Learning is significant in web based business. Notwithstanding, there are a few issues that emerge in the current Machine Learnings mining frameworks. The current ordinary procedures can't tackle these issues. The objective of noting these challenges better, it recommends Machine Learnings mining relying upon cosmology. There are three sections during information mining: (1) strategies for metaphysics creation and standards of product classification; (2) streamlining R-fascinating dependent on real circumstances; (3) actualizing Machine Learnings mining relying upon philosophy by improved Apriori. Moreover, it tests the improved calculation utilizing FoodMart2000, Java as the advancement language and Jena as the cosmology motor, finishes the entire cycle of mining, and checks the legitimacy of the calculation by the case of the information base.

III. AI on Stock Market:

Agrawal et al. (1993) finding Machine Learnings is a significant information mining issue, and there has been impressive examination on utilizing Machine

Learnings in the field of information mining issues. The associations' rules calculation is utilized basically to decide the connections between things or highlights that [9] happen simultaneously in the information base. For example, if individuals who purchase thing X likewise purchase thing Y, there is a connection between thing X and thing Y, and this data is helpful for chiefs. Consequently, the principle motivation behind actualizing the Machine Learnings calculation is to discover coordinated connections by examining the irregular information and to utilize these connections as a kind of perspective during dynamic (Agrawal et al., 1993). One of the main issues in current money is finding proficient approaches to sum up and picture the securities exchange information to give people or foundations valuable Hajizadeh et al. 115 data about the market conduct for venture choices. The tremendous measure of important information produced by the securities exchange has pulled in scientists to investigate this difficult space utilizing various philosophies. ShuHsien et al. (2008) examined financial exchange venture issues on Taiwan securities exchange utilizing a two phase information mining approach. The first stage apriori algorithm is a methodology of Machine Learnings, which is implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose stock category association and possible stock category investment collections. Then the K-means algorithm is a methodology of cluster analysis implemented to explore the stock cluster in order to mine stock category clusters for investment information. By doing so, they propose several possible Taiwan stock market portfolio alternatives under different circumstances (Shu-Hsien et al., 2008).

IV. The Issues in Machine Learning

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below.

1. Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision

making. In addition, when data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amount of sensitive and private information about individual and companies is gathered and stored. Moreover, Data mining could disclose new implicit knowledge about individual or groups that could be against privacy policies, especially if there is potential dissemination of discovered information.

2. User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps user better understand their needs. There are many visualization ideas and proposals for effective data graphical presentation.

3. Mining methodology issues: These issues pertain to the data mining approaches, applied and their limitation. It is often desirable to have different data mining methods available since different approaches may perform differently depending upon data at hand. Moreover, different approaches may suit and solve user's needs differently. Most algorithms assume the data to be noise free.

4. Performance issue: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issue if scalability and efficiency of the data mining methods when processing considerably large data.

5. Data Source issues: There are many issues related to the data sources. Some are practical such as diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data that we can handle and that we are still collecting data at an ever higher rate.

V. Requirements and Challenges of Machine Learning

Arranged by lead compelling Data mining, one necessities to initially inspect what sort of highlights an applied information disclosure framework is relied upon to have and what sort of difficulties one may look at the improvement of Data Mining procedure.

1) Handling of various sorts of information

Since there are numerous sorts of information and information bases utilized in various applications, one may expect that information disclosure framework should have the option to perform compelling information mining on various types of information.

Explicit information mining framework should be developed for information mining on explicit sort of information, for example, frameworks devoted to information mining in social information bases, exchange information bases, spatial information bases, media data sets, and so forth

2) Efficiency and unwavering quality of Data Mining calculations

To proficiently separate data from a tremendous measure of information in data sets, the information revelation calculations should be productive and versatile to enormous data sets. That is, the running season of an information mining calculation should be unsurprising and adequate in enormous data sets. Calculations with dramatic or even medium request polynomial intricacy won't be pragmatic use.

3) Usefulness, conviction and expressiveness of information mining results.

The disclosure information ought to precisely depict the substance of the data set and helpful for specific applications. The imperfectness should be communicated by proportions of vulnerability, as inexact standards or quantitative guidelines. Commotion and remarkable information should be dealt with richly in information mining frameworks. This likewise propels a deliberate investigation of estimating the nature of the found information, including intriguing quality and dependability, by compression of measurable, logical, and simulative models and instruments.

4) Expression of inquisitive sorts of information mining solicitations and results

Various types of information can be found a colossal measure of information. Additionally, one may lime to look at found information from various perspectives and protect them in various structures. This expects us to communicate all the information mining demands and found information in elevated level dialects or graphical UIs with the goal that the information mining undertaking can be indicated by

no specialists and the find information can be reasonable and straightforwardly useable by clients. This additionally requires the disclosure framework to receive expressive information portrayal strategies.

5) Interactive mining information at different reflection levels.

Since it is hard to foresee what precisely could be found from information base, an advanced information mining inquiry should be treated as a test which may reveal some fascinating follows for additional investigation. Intelligent disclosure should be energized, which permits a client to intuitively refine an information demand, powerfully change information centering, dynamically extend an information mining measure deftly see the information mining results at numerous reflection levels and from various stages.

6) Mining Information From various wellsprings of information

The generally accessible neighborhood and wide zone PC organization, including web, interface numerous wellsprings of information and structure tremendous circulated, heterogeneous information bases. Mining information from various wellsprings of arranged or unformatted information with different information semantics presents new difficulties to information mining. Then again, information mining may help reveal the advanced information normalities in heterogeneous data sets which can barely be found by straightforward question frameworks.

7) Protection of protection and information security

At the point when information can be seen from a wide range of points and at various reflections levels, it is the objective of ensuring information security and guarding against the attack of protection. It is imperative to examine when information revelation may prompt an intrusion of protection, and what security majors can be produced for forestalling the divulgence of delicate data.

Drugs Sector:

From BSE SENSEX analyst has chosen

Cipla Ltd, Dr.Reddy's Lab, SunPharma India Ltd, Glenmark Ltd, Orchid Chemicals Ltd to compute relationship between these equivalent area contents.

Here least help is the normal of all the determined help.

Furthermore, the MINIMUM SUPPORT: amount of help/absolute number of contents

=56/10

=5.6 %

In this way, least help is 5.6%

Least CONFIDENCE: amount of certainty/all out number of contents

= 350/10

=35%

So least certainty is 35%.

Specialist applied the above standard to ascertain min.Support and min.Confidence to acquire result for other area contents.

Script1	Script2	Support	Confidence	Is in rule
Cipla	Dr.Reddy's Lab	10%	67%	TRUE
Cipla	Sun Pharma India ltd	9%	60%	TRUE
Cipla	Glenmark Lab	1%	7%	FALSE
Cipla	Orchid Chem	4%	27%	FALSE
Dr.Reddy's Lab	Sun Pharma India ltd	12%	67%	TRUE
Dr.Reddy's Lab	Glenmark Lab	1%	6%	FALSE
Dr.Reddy's Lab	Orchid Chem	5%	28%	FALSE
Sun Pharma India ltd	Glenmark Lab	4%	18%	FALSE
Sun Pharma India ltd	Orchid Chem	7%	32%	FALSE
Glenmark Lab	Orchid Chem	3%	38%	FALSE

So from the above information examination analyst can reason that Cipla ltd and Dr.Reddy's Ltd go connected at the hip and furthermore Dr.reddy's lab And Sun Pharma India Ltd goes inseparably. So analyst can say these contents are emphatically connected with one another.

V. Proposed Method

In this our goal is to gauge the future securities exchange lists utilizing time-arrangement models, for example, Linear relapse model, ARIMA model and Vector Auto relapse model using the connection between opening cost and shutting cost.

DATA DESCRIPTION

The S&P 500 record gives the market estimation of 500 stocks. In spite of the fact that it incorporates stocks from various enterprises S&P 500 has some separated qualities, for example, affectability, consistency, versatility, to give some examples. We have utilized all the files for future computations. An example of information acquired from [www.yahoofinance.com] is appeared in table. Close, Open, High, Low, Volume, Change Percentage are considered as characteristics for this examination. Here Date is the date when the information is recorded, Price is the end value, Open is the Opening value, High and Low is the Highest and Lowest cost of the stock lists.

Fig : Stock indices prices of NSE500

DATA PREPARATION

There are a couple significant subtleties to take note of the manner in which the information can be pre-prepared to be found a way into the relapse models. The stock information are not accessible for quite a long time and occasions. We can disregard the missing information or we can supplant them with normal information. Alternate path if there should be an occurrence of ARIMA model in Python language is to set the recurrence as "B", which implies the model will set the week end days of a month as "0". In the event of VAR model and Simple Linear Regression model we have overlooked the missing qualities.

FEATURE SELECTION

Securities exchange close cost and open cost are significant snippet of data that is helpful for each momentary broker. The nearby costs are significant, particularly for swing brokers and position merchants. It additionally has suggestions for viable day exchanging numerous day exchanging frameworks. The financial exchange close value level gives significant data about the overall state of mind of speculators. It enlightens a ton regarding the considering enormous financial specialists that apportion huge measure of cash into the securities exchange for their resource the executives purposes.

The open cost is firmly connected with close cost and bad habit section.

V. Simulation Results

Opening Stock indices price				Closing stock indices price			
Test data	Simple Linear prediction	ARIMA prediction	VAR prediction	Test data	Simple Linear prediction	ARIMA prediction	VAR prediction
9968.4	9636.724864	9938.95822	9921.39	9861.45	9615.2424	9954.16898	9922.2985
10025.35	9642.377452	9955.11732	9919.4552	9919.2	9620.715	9976.25977	9920.2767
10006.95	9648.168032	9948.76347	9917.5204	10009.3	9626.3212	9985.39887	9918.2549
10023.5	9654.099973	9995.30044	9915.5856	9956.95	9632.0643	10006.1409	9916.2331
10046.45	9660.176727	9995.72582	9913.6509	9999.6	9637.9475	10035.2062	9914.2113
10019.95	9666.401829	9977.416	9911.7161	10083.7	9643.9744	10028.0984	9912.1895
9961.8	9672.778899	9944.34168	9909.7813	10021.4	9650.1484	10020.9446	9910.1677
10029.55	9679.311648	9920.7206	9907.8465	9947	9656.4731	9975.5091	9908.1459
10001.85	9686.003877	9929.91994	9905.9117	9987.55	9662.9523	9966.03998	9906.1241
10174.95	9692.859478	9945.37368	9903.977	10025.7	9669.5896	9984.30311	9904.1022
10096.4	9699.88244	9937.39736	9902.0422	10118.7	9676.3889	9994.38682	9902.0804
10099.95	9707.076849	9869.70538	9900.1074	10110.05	9683.3542	9966.18383	9900.0586
10070	9714.44689	9904.9757	9898.1726	10087.05	9690.4896	9910.64757	9898.0368
10051.2	9721.99685	9921.52922	9896.2379	10075.85	9697.7991	9931.96027	9896.015
10012.8	9729.731122	9915.2871	9894.3031	10042.5	9705.2871	9941.51467	9893.9932
9985	9737.654205	9960.10077	9892.3683	9978	9712.9579	9961.34893	9891.9714
9894.05	9745.770709	9960.28994	9890.4335	9943.9	9720.8159	9989.84689	9889.9496
9708.7	9754.085356	9941.32949	9888.4987	9792.2	9728.8658	9982.13926	9887.9278
9814.45	9762.602982	9909.61136	9886.564	9805.4	9737.1122	9974.7454	9885.906
9896.45	9771.328543	9886.43505	9884.6292	9747.4	9745.5599	9930.52164	9883.8842
9966.8	9780.267115	9895.21035	9882.6944	9941.65	9754.2139	9921.12491	9881.8624
9903.25	9789.423897	9910.47656	9880.7596	9980	9763.0791	9938.45426	9879.8406
9897.95	9798.804218	9902.47035	9878.8249	9888.55	9772.1607	9948.8123	9877.8188

9921.5	9808.413534	9837.15758	9876.8904	9872.55	9781.464	9921.49204	9875.7971
9931.6	9818.257435	9871.38289	9874.954	9924.7	9790.9945	9868.16226	9873.7752
9852.15	9828.341649	9888.33083	9873.0174	9908.7	9800.7576	9888.69686	9871.7512
9871.45	9838.672041	9882.20044	9871.1092	9817.2	9810.759	9898.66656	9869.7335
9905.35	9849.254622	9925.24376	9869.1745	9868.4	9821.0046	9945.52364	9867.741
9882.95	9860.095548	9905.63269	9866.8313	9897.65	9831.5003	9937.21622	9865.6261
9899.5	9871.201126	9875.27074	9865.3191	9905.35	9842.2523	9929.58229	9863.2516
9866.15	9882.577816	9847.69693	9869.2107	9892.3	9853.2667	9884.10184	9863.0226

Table 1: Predicted opening and closing prices

PERFORMANCE METRICS

There are certain performance metrics that could measure the accuracy of model. The less the difference between original value and predicted value, the better the model will be.

Table 2: Forecast of 10 days

Performance metrics	Simple Regression		ARIMA		VAR	
	Open	Close	Open	Close	Open	Close
MSE	5480.23	11059.50661	1407.486	2615.431	925.2118	2053.133
RMSE	74.02858	105.1641888	37.51648	51.14129	30.41729	45.31151
MAE	61.96637	88.81612822	28.84965	45.78778	25.3615	35.86036
MAPE	0.006256	0.008955566	0.002913	0.004632	0.00256	0.003617

Table 3: Forecast of 20 days

Performance metrics	Simple Regression		ARIMA		VAR	
	Open	Close	Open	Close	Open	Close
MSE	33121.84	43519.17672	9702.449	10498.49	8505.191	10145.38
RMSE	181.9941	208.6125037	98.50101	102.4622	92.22359	100.7243
MAE	139.2437	161.7852841	70.57418	82.42754	67.14157	79.92172
MAPE	0.013943	0.016182255	0.0071	0.008309	0.006742	0.008018

Table 4: Forecast of 30 days

Performance metrics	Simple Regression		ARIMA		VAR	
	Open	Close	Open	Close	Open	Close
MSE	68136.18	72863.41693	10095.14	7967.305	12274.47	11059.52
RMSE	261.0291	269.9322451	100.4746	89.25976	110.7902	105.1642
MAE	215.6521	227.6077012	74.93858	69.5365	87.40315	85.2634
MAPE	0.021522	0.022746265	0.007501	0.006994	0.008732	0.00853

VI. Conclusion

This research entailed the development of various methods to find the most accurate model for prediction of prices of the NSE500 stock. We have analysed from the prediction result that for short term time period, VAR model could predict the Opening price and closing price slightly closer to original price as compared to ARIMA and Simple Linear Regression. But as we go for further long term prediction, the ARIMA model gives better prediction than other two models. Although ARIMA and Vector Auto Regression model gives approximately same result for these two variables, ARIMA model is better as compared to VAR model for stock price prediction because in ARIMA model we have options to choose both Auto Regression and Moving Average as and when required. In contrast, if the other relevant external variables, on which the price of particular stock or indices varies, could be identified and included in the VAR model, it could predict more accurately. The trader can consult the VAR model to get the trend and ARIMA model to get the closest value.

References

- [1] Alor-Hernandez, G., Gomez-Berbis, J. M., Jimenez-Domingo, E., Rodríguez-González, A., Torres-Niño, J., "AKNOBAS: A Knowledge-based Segmentation Recommender System based on Intelligent Data Mining Techniques", Computer Science and Information Systems, Vol. 9, No. 2, 2012, pp. 713-740.
- [2] Han, J., Kamber, M., "Data mining concepts and techniques 2nd edition", Morgan Kaufman, 2006, pp. 227-378.
- [3] Agrawal R., Imielinski T., "Mining associations between sets of items in large databases", Proceedings

of the ACM SIGMOD International Conference on Management of Data. pp. 207-216.

- [4] Konda, S., "Web Data Mining Based Business Intelligence and Its Applications", IJCST, Vol. 4, No. 4, 2013, pp. 112- 116.

- [5] Aher, B., "Machine Learning in Data Mining". IJCST, Vol. 4, No. 3, 2013.

- [6] Nagabhushanam, D., Naresh., N., "Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient's Data", IJCST, Vol. 4, No. 4, 2013, pp. 262-265.

- [7] Surya, K., Priya, K., "Exploring Frequent Patterns of Human Interaction Using Tree Based Mining", IJCST, Vol. 4, No. 4, 2013.

- [8] Rajnikanth, J., "Database Primitives, Algorithms and Implementation Procedure: A Study on Spatial Data Mining", IJCST, Vol. 4, No. 2, 2013.

- [9] Quinlan, J. R., "C4.5: Program for machine learning. CA", Morgan Kaufmann, San Francisco, 1992

- [10] Clark, P., Niblett, T., "The CN2 induction algorithm. Machine Learning", Vol. 3, No. 4, 1989, pp. 261-283.

Authors



KAUSALYA KOLA Holds a B.Tech Degree in Computer Science & Engineering from MVGR College, chinthalavalasa, vizianagara m. She is presently Pursuing M.Tech (CSE) in Department of Computer

Science and Engineering from Avanthi Institute of Engineering and Technology Cherukupalli(V), Near Tagarapuvalsa Bridge,Vizianagaram (Dist) - 531162,Andhra Pradesh. Area of interest include Web Technologies, web programming and C programming, Machine learning and Data science.



Mr.Kesavarao Seerapu
Working as Asst. Professor,
Department Of Computer
Science and Engineering in
Avanthi Institute of
Engineering and Technology
Cherukupalli(V), Near
Tagarapuvalsa
Bridge,Vizianagaram (Dist)

- 531162,Andhra Pradesh. His Area of interests include Machine Learning,Classification Data Mining and Knowledge Discovery,Sentiment Analysis.