PERFORMANCE EVALUATION OF MACHINE LEARNING TECHNIQUES USING BIG DATA IN PREDICTIVE ANALYTICS

A.Kalpana
Research Scholar
School of computing sciences,
VISTAS
Chennai ,India
kalpana_ambi12@yahoo.co.in

Dr.K.Rohini
Associate Professor
School of computing sciences
VISTAS
Chennai ,India.
rrohini16@gmail.com

Abstract :

Big Data has arisen as a significant area of interest of study and exploration among specialists and academics. Big data is a great source of information from the frameworks to opposite end-clients. In fact, with the big data spread and constant increase logical systems assume more significant role and inevitability in organizations. So, Predictive analytics is used to find the relations and forms in the data so as to predict future by observing the past and making good decisions. In statistical and analytical techniques the term substantially used is predictive analytics. This term is drawn from Optimization techniques, database techniques, statistics and machine learning. It has been derived from classical statistics. Using the models of predictive analytics, the future events and behaviour of variables can be predicted. The predictive analytics have many advantges. A scoring technique is provided for predictive analytics models. A higher score shows the higher probability of occurrence of an event and a lower score demonstrates the lower probability of occurrence of an event. To find solution for various commercial and technical problems, the past and transactional data patterns are broken by these models. The predictive analytics models have dominated due to the growth of attention in the decision support solutions. This paper, presents applications and techniques of predictive analytics is reviewed.
Application of Machine learning Algorithms such as Regression Modelling and ARIMA model.ARIMA (Autoregressive Integrated Moving Average) model and Regression model are applied for Gold price forecasting.

Keywords— Big data technology, predictive analytics, Machine Learning, Techniques of predictive analytics
DOI Number:10.14704/nq.2022.20.8.NQ44431 NeuroQuantology 2022; 20(8): 4001-4011

INTRODUCTION

Predictive analytics, a branch from advanced analytics is used to predict the future . To predict future it analyzes current and historical data. It is obtained from machine learning, database techniques, statistics and optimization techniques[1]. To Predict the future it combines Information technology, management and business modeling process. To gain profit in business, predictive analytics in combination with big data can be effectively applied. Centering the data, it assist the organization to be dynamic, modern and to predict its progress. Substantial growth in big data have made it to grow expressively.

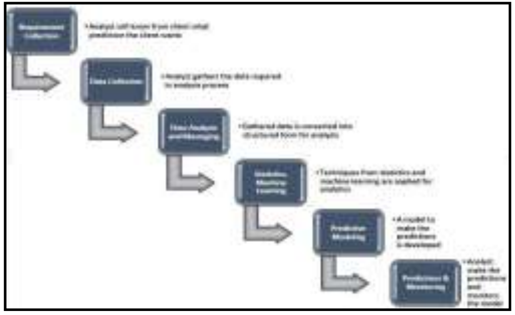
Predictive Analytics lets in us to achieve understandings to predict the future movements and unknown activities. In the area of Data Mining, Predictive Analytics joins fingers with statistical evaluation to supply a most stimulating combination of strategies that results in knowledge discovery. The word Analytics is deduced from technology of data analysis this is generally associated with another term Business Intelligence to explain the provisioning of selection.

Support in companies. Big wide variety of applications of Predictive Analytics tiers from each academia and industries. It is a massive area having big scope in present and future.

Predictive analytics isn't always best restricted its software with e- retailing. The domain names of software of it's miles big which include insurance, Banking sectors and so forth, institutions concerned in fiscal investments identify the stocks that offers good returns on their investment and that they indeed forecast the future overall performance of shares grounded at the beyond records and cutting-edge performance. Diverse corporations practice predictive models to predict sale for his or her commodities when they make investment for manufacturing. Scientific groups identifies drugs that has lower sales in a specific location and come alert of those drugs expiry[4].

2. PREDICTIVE ANALYTICS- PROCESS

Predictive analytics contains of diverse steps over which a records analyst forecast the future based on past and cutting-edge data. The following diagram depicts the process of predictive analytics.



Predictive Analytics Process

2.1 COLLECTION OF REQUIREMENT

To enhance to a forecasting model, first it must be clean at the purpose of forecasting. From forecasting the end result that must be done must be described. The data analyst consults with client to get the prerequisite to expand a predictive version and the way client gets profited from these predictions. While growing the version, which sort of data the client requires will be identified.

2.2 DATA COLLECTION

To develop the predictive version, the analyst first is aware of the constraint of the client and begins to acquire the specified datasets from various sources. It includes an entire list of customers individuals who take a look at or use the made from the business enterprise. The information can be in unstructured or dependent sample. Analyst assessments the information accumulated from the customers at their very own website.

2.3 DATA ANALYSING AND MESSAGING

The analyst examines the data that has been accumulated and create it in an arrangement that may beanalysed and which may be used in the model. Here, during this process the unstructured data is transferred to structured data. When data is made completely available its quality is tested. Numerousopportunitiesincluding presence of faultyrecords in most importantrecords set or attributes values is probably missing. Effectiveness of predictive model solelyrely upon the quantity of data. Analysis phase is noted be data munging or messaging the datathis iswhileraw data is transformed to a format which may be used for analytics.

2.4 MACHINE LEARNING, STATISTICS

In the process of predictive analytics numerous statistical and machine learning techniques were employed. Here, for analyzing the most important methods such as regression analysis and theory of probability were often used. In various predictive analysis task tools of machine learning such as support vector machines, artificial neural networks, decision tree are widely used. Statistical technique or machine learning acts as a base for predictive analytic models. While comparing Machine learning with statistical techniques machine learning has greater advantage. The techniques from statistics are incorporated to develop any predictive model.

2.5 PREDICTIVE MODELLING

In modelling stage, a model is made by using statistical techniques with sample dataset and machine learning. Once, after the development the data undergoes testing phase to find the validity of the model by using test data which is a part of original data set collected. When the phase executes effectively, the model is said to be fit. Once the model is fitted it can accurately predict new data. To solve this problem numerous applications uses the multi-model solution.

2.6 MONITORING AND PREDICTION

The model is deployed at client site to make day to day prediction and decision-making process, later it produces successful test in predictions. The results and reports are generated by the model nor managerial process. Consistently the model is monitored and ensured that it provides exact results and forecast accurately. We have seen here, that predictive analysis is not a single step process to make predictions on future. It is a step-by-step process. It has many processes starting from requirement collection to deployment, and monitors for effective usage of system that makes it as a system for decision making process.

3. OPPORTUNITIES IN PREDICTIVE ANALYTICS

A huge history to work with predictive analytics is there. It has been usually applied in several domains for decades, Now-a-days predictive analytics is widely used due to the progression of technologies and on data dependency [5]. To increase the bottom line and profit various organizations are adopting towards predictive analytics.

4.CATEGORIES OF PREDICTIVE ANALYTIC MODELS

a) Predictive Model: Models of this type analyse the performance of past to foretell about future.

b) Descriptive Model: Models of this type measures the relationship in data. To classify datasets into groups descriptive models are used.

c) Decision Model: Models of this type defines relationship with all variables to make a decision which helps in order to predict the results [6]

The predictive analytic model can be defined precisely as a model that predicts a detailed level of granularity. For each individual a predictive score is generated. It is similar to a technology that learns from practice so that it can make prediction about future performance of an individual. It is helpful to make better decisions. Accuracy of results of the model is based on data analysis level.

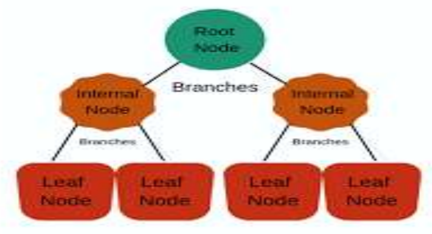
5. TECHNIQUES OF PREDICTIVE ANALYTICS:

The models for predictive analytics can be obtained from classification models and regression models. The relationship of values for specific classes are anticipated using Classification model whereas number is anticipated in regression model. Here we currently list the important algorithms that are popularly used to develop predictive models.

5.1 ARIMA: It is a measurable examination model that uses time series information to either better understanding of the data set or to anticipate future patterns. A statistical model is autoregressive on the off chance that it predicts future qualities in light of past qualities. To assess ARIMA Model, the dataset is spitted into training and test sets. Go through time steps in the test dataset. Train the model.

It gives one-step forecasting. Save the forecast to find and save actual perception. Ascertain blunder score for forecasts contrasted with anticipated values.

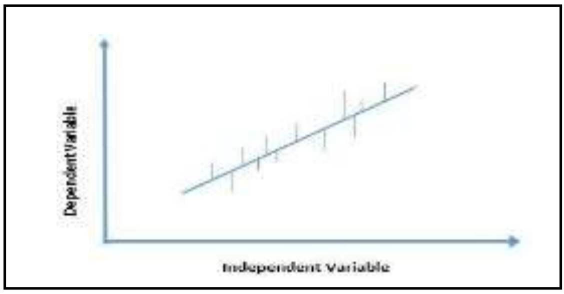
5.2 Decision tree: In regression we use a decision tree which is a classification model or binary tree representation model. Every node represents a single information variable (x) and a split point on that variable (assume the variable to be numeric).The leaf nodes of the tree hold an output variable (y) using which the forecast can be made. By walking the splits of tree until a leaf mode is arrived predictions can be made. Trees make predictions rapidly[15]. The reason to make the decision trees popular to use is that it can be easily understood and interpreted. A classic model of decision tree is shown in the below diagram



Decision tree

By using this model, primer factors can be selected and it has the property by which the missed data can be handled. Commonly they are denoted as generative models of induction rules which works on t observed data. Most of data from data sets are used and questionnaire level is minimized.

5.3 Regression Model: To estimate the relationship between variables, regression is used which is one among the best statistical techniques. The relationship between independent and dependent variables are modeled. It analyzes how value of dependent variable varies on capricious the values for independent variables in modeled relation[9]. Modeled relation of dependent and independent relation is shown in the below diagram.



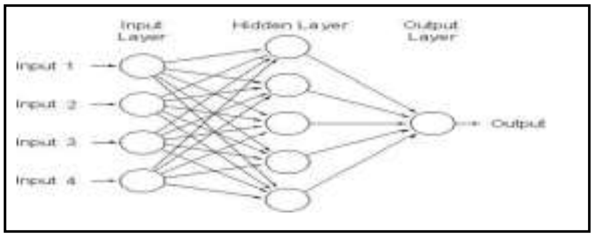
Regression Model

The connection among dependent and independent variables are modeled. It analyzes how the value of dependent variable fluctuates on impulses for the qualities for independent variables in modeled relation. In the framework of continuous data, where it is expected that it has ordinary distribution, the regression model finds the key pattern in huge datasets. It is valuable in observing the impact of specific elements which impact the development of variables. In regression, on the basis of the predictor variable the value of the response variable is anticipated. Here, map independent variables with dependent variables regression function is used. In this strategy, the difference of dependent variables is categorized utilizing prediction of a regression function that utilizes a probability distribution.

Linear regression model, and the logistic regression models are the two sorts of regression models that are utilized in predictive analytics for prediction. To show the direct connection among linear relations between dependent and independent variables linear regression model is applied. In this model a linear function is utilized as a regression function. Coordinated factors relapse is utilized when there are classes of ward factors. . Here, unknown values of discrete variables are predicted on the source of known values of independent variables. Only few number of values are assumed in prediction

5.4 Artificial Neural Network

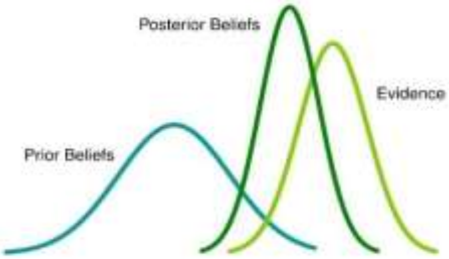
A network of artificial neurons based on biological neurons is Artificial neural network. It is used to simulate the human nervous system. Artificial neural networksprocess the input signals and produce the outputs. It can handle extremely complex relations. The architecture of artificial neural network is represented in diagram



Artificial neural networks are largely used in predictive learning. It is used to predict prices on a weekly basis. Predictive analytics works on the existing data and makes predictions on new data. Data present in the input layer is processed and passed to the hidden layer. Based on the requirement of the output various functions are performed on the input neurons. The output of one neuron is passed to the next layer and the output layer gives the prediction of the new data. Artificial neural networks have various models and each uses different algorithms. Artificial neural networks and clustering are unsupervised learning methods. Both can handle non-linear data more effectively. Both methods evaluate regression models and decision trees. Both models are effective for pattern recognition and widely used in image data.

5.5 Bayesian Statistics

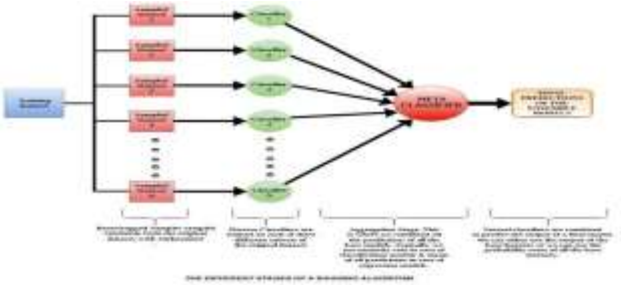
This method belongs to the statistics that take parameters as random variables and the term that is used as “degree of belief” to outline the likelihood of incidence of an occurrence [11]. The Bayesian statistics is predicated on Bayes’ theorem that terms the events priori and posteriori. In probability, the method is to search out the likelihood of a posteriori event on condition that priori has occurred. On the opposite hand, Bayes' theorem finds the likelihood of priori event provided that posteriori has already occurred. The BayesiaStatistics is represented in diagram



It uses a probabilistic graphical model that is termed the Bayesian network that represents the conditional dependencies among the random variables.

5.6 Ensemble Learning

In the branch of machine learning, Ensemble learning belongs to the class of supervised learning. These models are developed by working with several similar types of models and eventually combining their results on prediction. During this method, the accurateness of the model is upgraded. Development during this method reduces the bias and reduces the variance of the model. It helps in finding the most effective model to be used with new data [8].

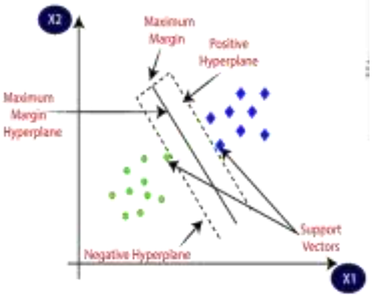


5.7 Gradient Boost Model

This is a predictive analytic machine language which is mainly used in classification. And regression based applications. It is similar to esemble model. It is a boosting approach, which is applicable for any type of data sets. It resamples the datasets frequently and generates weighted average the data sets.Various advantages such as, less prone to overfitting and improves fitting of data.

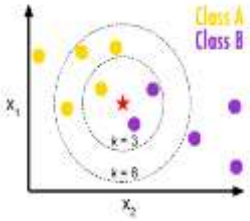
5.8 Support Vector Machine

Support vector machine learning technique is used in predictive analysis. It analyzes data for classification and regression which is mostly used in classification applications. Support vector machine is a discrimination classifier used to classify examples into categories with a clear gap. The new examples are predicted on the gap.



5.9 k-nearest neighbors (k-NN)

k-NN is a nonparametric method for classification and regression. The input is the k-closest training examples and the output is the membership of a class in the regression problems. It is the simplest machine learning algorithm.



5.10 MLP Regression

MLP Regressor additionally upholds multi-yield relapse, in which an example can have more than one objective. Multi-facet Perceptrons (MLPs) can be utilized effectively for order Class MLPRegressor executes a multi-facet perceptron (MLP) that trains involving backpropagation with no actuation work in the result layer, which can likewise be viewed as involving the character work as initiation work.

6. Proposed Methodologies

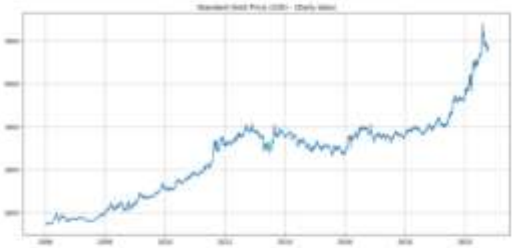
An empirical collaborative model is obtained by combining the regression model with Gradient Boost Model(XGB),Linear model, MLP Regressor and Support vector Regressor.

Sample Data

Date	Country	State	Location	Pure Gold (24 k)	Standard Gold (22 K)
01-02-2006	India	Tamilnadu	Chennai	768	711
01-03-2006	India	Tamilnadu	Chennai	770.5	713
01-04-2006	India	Tamilnadu	Chennai	784.5	726
01-05-2006	India	Tamilnadu	Chennai	782.5	725
01-06-2006	India	Tamilnadu	Chennai	776	719
01-07-2006	India	Tamilnadu	Chennai	787.5	729
01-09-2006	India	Tamilnadu	Chennai	790	732
01-10-2006	India	Tamilnadu	Chennai	791	732
01-11-2006	India	Tamilnadu	Chennai	788	730
01-12-2006	India	Tamilnadu	Chennai	789	731
01-13-2006	India	Tamilnadu	Chennai	790	732

Datashape
[2]: (4971,6)
Columns
Index([‘Date’,’Country’,’state’,’Location’,’Pure Gold(24 k)’,’Standard Gold(22 K)’], dtype=’object’)
The price indicated for Pure Gold(24 k) and Standard Gold(22 K) is 1 Gram weight.

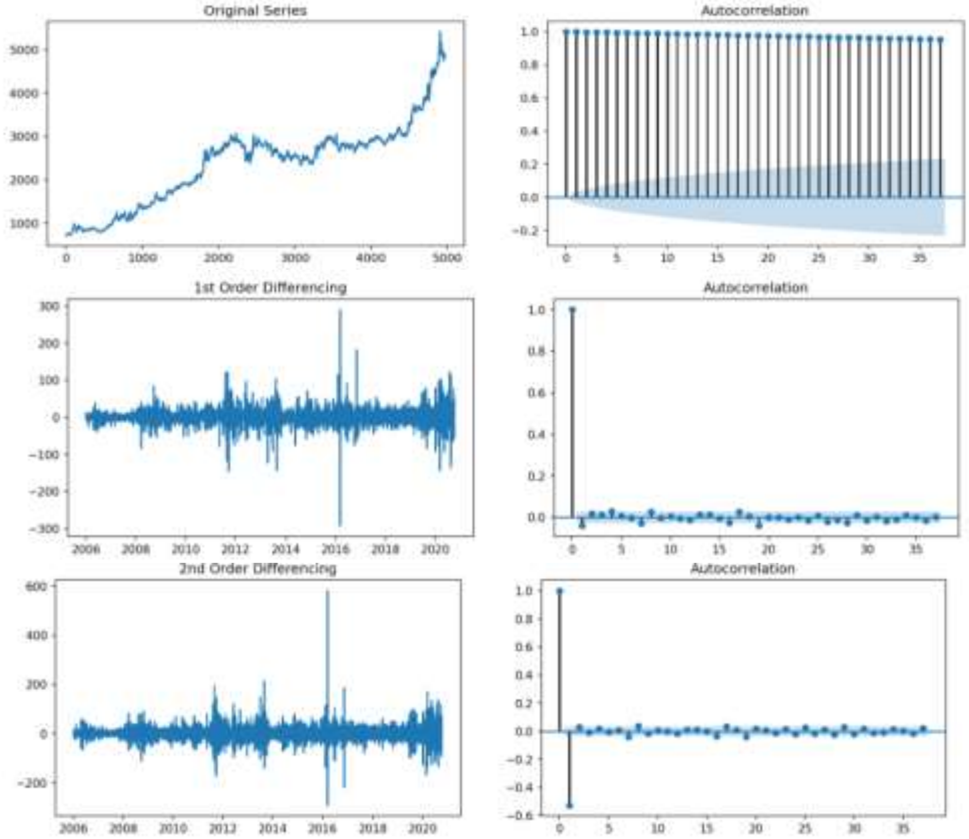
The date ranging from 2006-01-02 00:00:00 to 2020-10-10 00:00:00



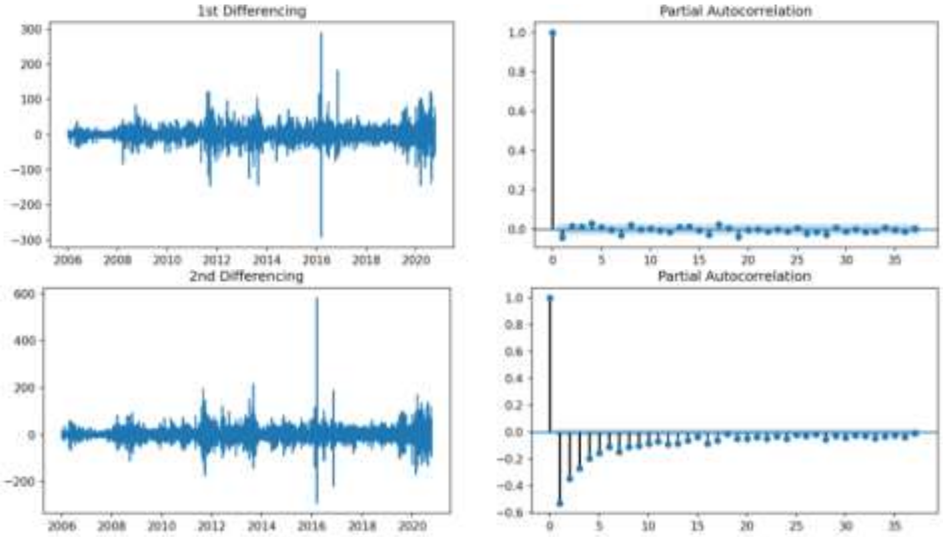
ARIMA model

ADF Statistic: 0.542589

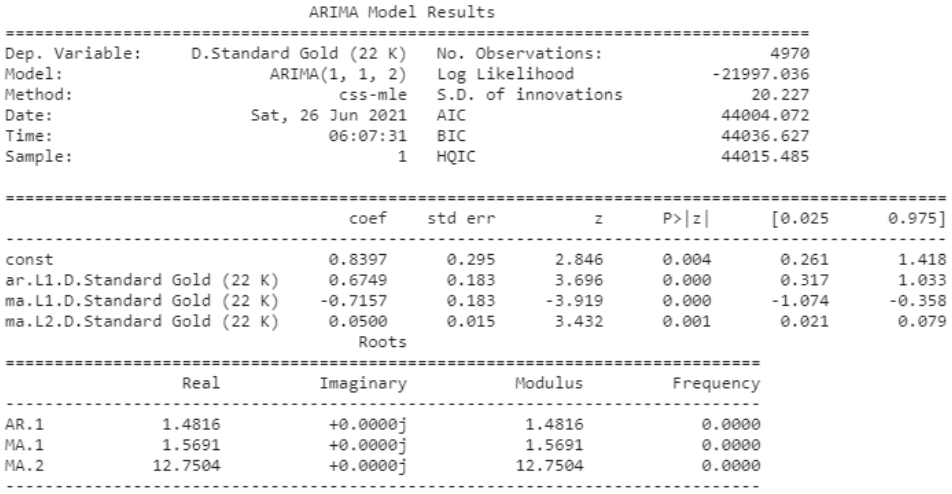
p-value: 0.986112



Partial correlation



P value is tentatively fix 1

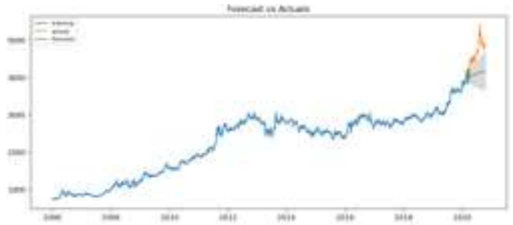


```
                             ARIMA Model Results
==============================================================================
Dep. Variable:     D.Standard Gold (22 K)   No. Observations:                 4970
Model:                 ARIMA(1, 1, 2)   Log Likelihood              -21997.036
Method:                       css-mle   S.D. of innovations             20.227
Date:                Sat, 26 Jun 2021   AIC                          44004.072
Time:                        06:07:31   BIC                          44036.627
Sample:                             1   HQIC                         44015.485

=================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          0.8397      0.295      2.846      0.004       0.261       1.418
ar.L1.D.Standard Gold (22 K)   0.6749      0.183      3.696      0.000       0.317       1.033
ma.L1.D.Standard Gold (22 K)  -0.7157      0.183     -3.919      0.000      -1.074      -0.358
ma.L2.D.Standard Gold (22 K)   0.0500      0.015      3.432      0.001       0.021       0.079
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.4816           +0.0000j            1.4816            0.0000
MA.1            1.5691           +0.0000j            1.5691            0.0000
MA.2           12.7504           +0.0000j           12.7504            0.0000
-----------------------------------------------------------------------------
```

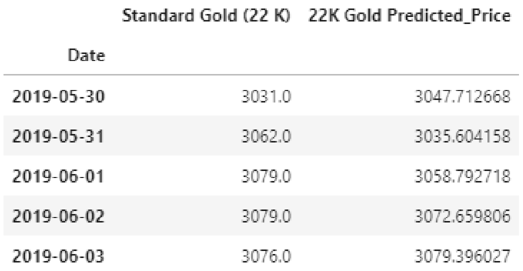
Training and Testing Shape

(4771,) (200,)



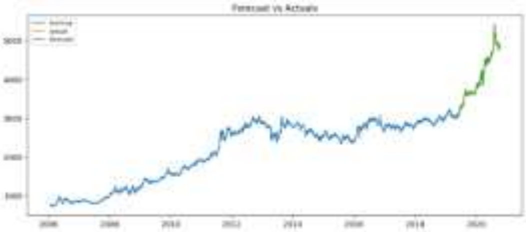
Here results are not satisfied.

The Empirical Collabrative



	Standard Gold (22 K)	22K Gold Predicted_Price
Date		
2019-05-30	3031.0	3047.712668
2019-05-31	3062.0	3035.604158
2019-06-01	3079.0	3058.792718
2019-06-02	3079.0	3072.659806
2019-06-03	3076.0	3079.396027

Test data Results Predicted Price

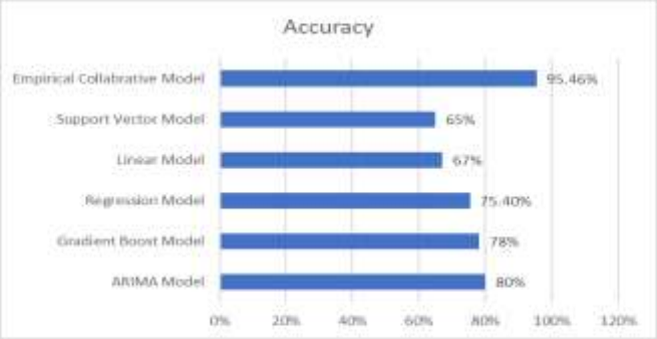


RMSE 34.94649666644517

Comparison of Machine Learning Algorithm

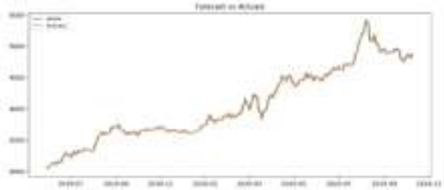
Model	Accuracy
ARIMA Model	80%

Gradient Boost Model	78%
Regression Model	75.40%
Linear Model	67%
Support Vector Model	65%
Empirical Collabrative Model	95.46%



The empirical collaborative model that is obtained by combining Gradient Boost Model,Regression Model, Linear Model & Support Vector Model yields higher accuracy with huge amount of data. In contrast, the ARIMA model fails to produce high accuracy with large data set. Hence when Arima model is compared with empirical collaborative model yields better results.

closer to Predicted Vs Actual



Comparing to ARIMA, Empirical collabrative model gives better results

8. CONCLUSION AND FUTURE SCOPE

A huge history is there to use predictive model for the task of predictions. In past, Statistical models were used for predictive modelling which were based on sample data from huge data set. With the improvements in the field of computer science and the drastic growth of computer technology, new techniques have grown up and various new algorithms were introduced over time period. The machine learning model is been used by predictive models. Based on parameters that is given as input, the future or output of any value can be predicted. This paper opens a scope of development of latest models for the task of predictive analytics. There's additionally a chance to feature further options to the present models to enhance their performance within the task.

9. REFERENCES

[1]. Charles Elkan, 2013, “Predictive analytics and data mining”, University of California, San Diego.

[2]. Eric Siegel, 2016, “Predictive Analytics”, John Willey and Sons Ltd.

[3]. Charles Nyce, 2007, “Predictive Analytics White Paper”, American Institute of CPCU/IIA.

[4]. W Eckerson, 2007, “Extending the Value of Your Data Warehousing Investment”, The Data Warehouse Institute.

[5]. Sue Korn, 2011, “The Opportunity of Predictive Analytics in Finance”, HPC Wire.

[6]. Mohsen Ahauran, SharminAhauran 2018 “Opportunities and Challenges of ImplementingAnalysis for Competitive Advantage”, International Journal of Business Intelligence Research Vol-9,Issue-2.

[7]. V Dhar, 2001, “Predictions in Financial

Markets: The Case of Small Disjuncts", ACM Transaction on Intelligent Systems and Technology, Vol-2, Issue-3.

[8]. J Osheroff, J Teich, B Middleton, E Steen, A Wright, D Detmer, 2007, "A Roadmap for National Action on Clinical Decision Support", JAMIA: A Scholarly Journal of Informatics in Health and Biomedicine, Vol-14, Issue- 2, Pages-141-145.

[9]. B Kaminski, M Jakubczyk, P Szufel, 2018, "A framework for sensitivity analysis of decision trees", Central European Journal of Operations Research, Vol- 26, Issue-1, Pages-135-159.

[10]. J S Armstrong, 2012, "Illusions in regression analysis", International Journal of Forecasting, Vol-28, Issue-3, Pages-689-694.

[11]. W S McCulloch, Walter Pitts, 1943, "A logical calculus of the ideas immanent in nervous activities", The bulletin of mathematical biophysics, Vol-5, Issue-4, Pages-115- 133.

[12]. Peter M Lee, 2012, "Bayesian Statistics: An Introduction, 4th Edition", John Willey and Sons Ltd.

[13]. R Polikar, 2006, "Ensemble based Systems in decision making", IEEE Circuits and Systems Magazine, Vol-6, Issue-3, Pages-21-45.

[14]. J H Friedman, 1999, "Greedy Function Approximation: A Gradient Boosting Machine", Lecture notes.

[15]. C Cortes, 1995, "Support-vector networks", Machine Learning, Vol-20, Issue-3, Pages- 273-297.

82.

[16]. Ben Hur et al, 2001, "Support Vector Clustering", Journal of Machine Learning Research, Vol-2, Pages- 125-137.

[17]. J Lin, E Keogh, S Lonardi, C Chiu, 2003, "A symbolic representation of time series, with implications for streaming algorithms", Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, Pages-2-11.

[18]. N S Altman, 1992, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", The American Statistician, Vol-46, Issue-3, Pages- 175-185.

[19]. H Abdi, L J Williams, 2010, "Principal component analysis", WIREs: Computational Statistics, Vol-2, Issue-4, Pages-433-459.

[20]. K Das, GS Vidyashankar, 2006, "Competitive Advantage in Retail Through Analytics: Developing Insights, Creating Values", Information Management.

[21]. N Conz, 2008, "Insurers Shift to Customer-Focused Predictive Analytics Technologies", Insurance & Technology.

[22]. J Feblowitz, 2013, "Analytics in Oil and Gas: The Big Deal About Big Data", Proceeding of SPE Digital Energy Conference, Texas, USA.

[23]. G H Kim, S Trimi, J-H Chung, 2014, "Big-data applications in the government sector", Communications of the ACM, Vol-57, Issue-3, Pages-78-85.

[24]. Vaibhavkumar, M.L.Garg, 2018,"Predictive analytics: A Review of trends and Techniques",International Journal of Computer Applications, vol-1