International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector Regression (SVR)

Navin[1], Dr. G. Vadivu[2]

[1]Department of Information Technology, Database Systems, SRM University, Katankulathur, Chennai, India

[2]Professor, Department of Information Technology, SRM University, Katankulathur, Chennai, India

Abstract: ***Develop a forecasting model for predicting and forecasting gold prices based on economic factors such as inflation, currency price movements and others. For investing the money, investors are putting their money into gold because gold plays an important role as a stabilizing influence for investment portfolios. Due to the increase in demand for gold in India, it is necessary to develop a model that reflects the structure and pattern of gold market and forecast movement of gold price. The most appropriate approach to the understanding of gold prices Support vector Regression and decision tree model. The experimental result will show the better performance from these two (Decision tree algorithm and support vector regression algorithm) algorithms.***

Keywords: R, RHadoop, SVR (Support Vector Regression), Decision tree, Gold price.

1. Introduction

Essentially there is two type of stock market present one is equity market and the commodity market. An equity market is aggregation of the producer and consumer of stocks and the trade in primary other than manufactured product is the commodity market. There are two type of commodity are present in the commodity market one is soft commodity in which wheat, coffee, cocoa and sugar are come and other is hard commodity in which gold, rubber and oil are come.

In Indian gold (hard commodity) play the crucial role in the market. Gold is the most popular as an investment the money of all the metals and investor buy the gold as per diversifying risk. Especially through the use of futures derivatives and contracts the gold market is subject to speculation as are other markets. Gold trades predominantly as a function of sentiment and its price is less affected by the laws of supply and demand. Gold is the storable and many people invest their money in the gold market.to invest the money, gold prediction is the very important way to predict and forecast the value (price) of the gold. There are so many method to predict and forecast the price such as linear regression method, logistic regression method, decision tree method, support vector regression method etc.in this paper we are describing the two forecasting method one is decision tree method and second is support vector regression method. All the method are predict the vale on the basis of factor of the gold. Forecasting is basically used for check the trend and to earn money. The price of the gold is depends upon the supply and demands just like other goods. Big data analytics is the process of gathering the data from the different sources, managing or organizing the data and apply analytics on large amount of the datasets to find the pattern and to check the tendency of the datasets. Big data can be any type of the data such as structured data, semi structured data and unstructured data or it can be mixed of these three datasets. Big data analytics is useful to find the correlation, customer preferences and various trends of the data. The main purpose of the big data analytics is to check the timing for enter into the market and exit to the market to invest the money. There are various tools are present to do the analytics such as BI tools, Statistical tools, data visualization tools. But most of the tools cannot support the large amount of data and if any tool supports the large data then it used to take so much time to process the data or to analyses of the data. Big data analytics is use to perform the data mining process, data forecasting, data prediction etc. For the forecasting of the gold price there are 4 or 5 factors are present such as Open Price, Close Price, Lowest Price, Highest Price and value of the Gold. From these factors they can find the percentage change of the price.

We know that one of the reasons of the gold price change is the external effects such as social problems, economic policies and environmental conditions political. One general assumption is made in such cases is that the historical data incorporate all those behavior. As a result, the historical data is the major input to the prediction process. In this hypothesis the external effects are modeled as noise, and the phenomena one considered as accidental.

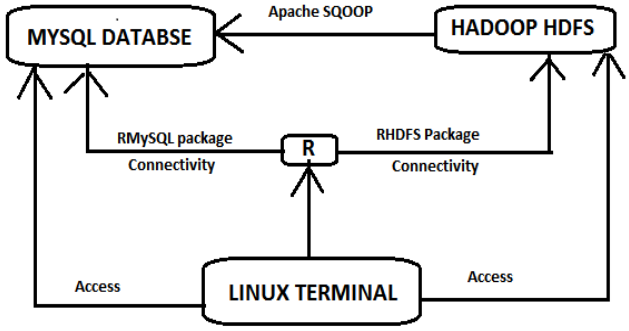

Fig.1. Connectivity Diagram

Before start the analysis, the analyst perform the data cleaning process or data pre-processing. analyst remove the

NULL values, data duplicity and data ambiguity from the datasets can connect to MySQL database to manipulate the data or clean the data with the package of RMySQL library and access all the database file through R. R can also connect to the hadoop to access the hdfs (hadoop distributed file system) file and to perform the analysis on that datasets and with the Linux terminal, analyst can access all the platform such as R, database and hadoop. Hadoop supports the additional software packages (eco-system tools) to analysis purpose, here the apache sqoop ecosystem tool is use for provide the connection between databases to hadoop. Big data can be analyzed with many software tools commonly used as part of advanced analytics disciplines such as data mining, statistical analysis, predictive analytics and text analytics. Many BI tools are support the analytics and visualization technique but the relational database cannot support the unstructured data or traditional data warehouse and data warehouses may not be able to handle sets of big data that need to be updated continually.

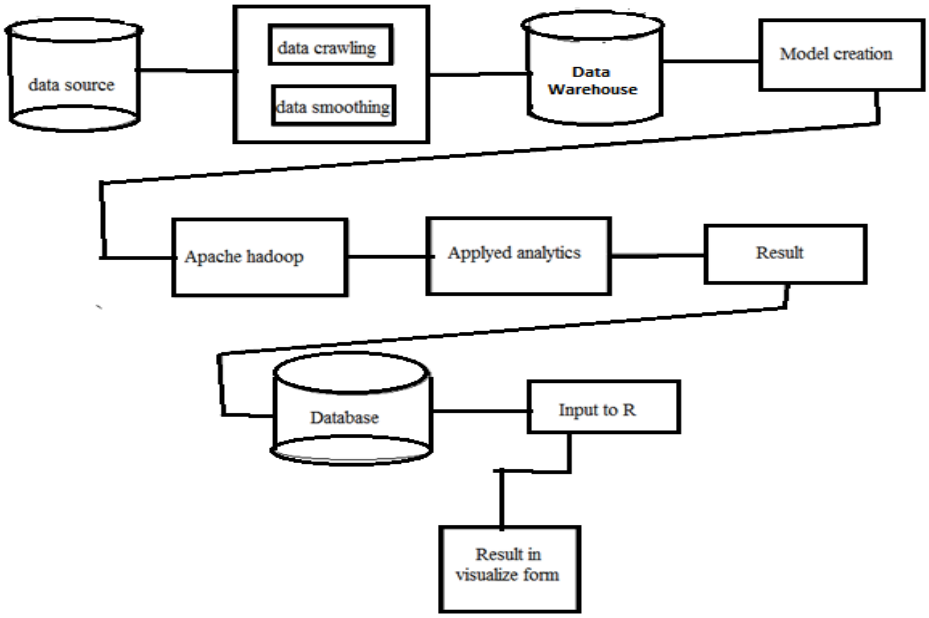

Fig.2. Solution Architecture

The one of the best framework is present which support the large datasets is hadoop. Apache hadoop is the open source framework written in java language to support the large amount of the data with map-reduce technique and hadoop-ecosystems tools (additional software packages) such as apache PIG, apache HIVE, apache SQOOP etc. the main part of the apache hadoop is HDHS (hadoop distributed file system). Real-time data on the performance of gold price. Many organizations used to gather, process and analyze big data have turned to a newer class of technologies that includes Hadoop and hadoop ecosystem tools such as YARN, MapReduce, Spark, Hive, MongoDB, Hbase and Pig as well as NoSQL databases. Those technologies form the hadoop framework that supports the processing of large data sets across clustered technology. This is the storage part of the hadoop and to perform the processing of the data hadoop uses the map-reduce technique. The main goal of hadoop is data locality which is to use a whole server in a large cluster, in which each server has internal disk drive.to provide the higher performance Map Reduce technique assign the total workload to these server and proceed to the for the data analysis.

2. R

R is a statistical software or data analytical software to analyse the data and to apply the predictive modeling with data visualization. R is used for many graphical and statistical methods, such as time series analysis, classification, cluster, classical statistical test, linear modeling and non-linear modeling and other with the use of its libraries. R support many languages such as C, C++, and PYTHON etc. to directly manipulate the R objects. For any specific function, specific area and specific language user upload the packages in R and because of this it used to become highly extensible. With the different packages R used to provide better connection, better analytics and better visualization. In the graphical representation (visualization), R uses lot of plots such as histogram, box plot, pie chart, line graph, bubble chart etc. to analyse the gold fluctuation the most valuable graph is line graph. R uses the visualization in the form of 2D and 3D. in R visualization can express the results, excavation process and it allowing user to find the exact problem after deeply understanding of the data and after analyse the data value it recognize which algorithm is best for the analysis.

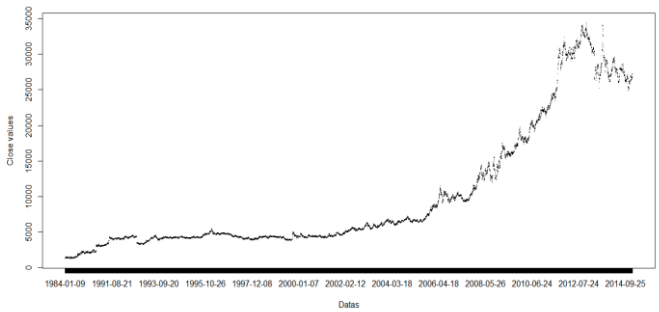

Fig.3. Simple plotting in R

R has some important features and it facilitate to the data

manipulation and calculation. It also include

- A facility of storage of data and data handling
- in the particular matrices, it suite or the calculation of array
- A large, integrated, coherent collection of intermediate tools for data analysis,
- It display the result either on softcopy or hardcopy and graphical facilities for data analysis
- It is simple and very effective language which includes many function such as (conditionals, loops, user-defined recursive functions) and input and output facilities.

3. RHadoop

R is the statistical software to analyse the data in the form of visualization and hadoop is the java open source framework which have two core component one is hdfs and other is map-reduce. To connect R with hadoop, basically three packages are required rhdfs, rmr2, rHbase in which rJava package comes under the rhdfs and for the storage purpose rhbase is required. rhdfs package provides basic connectivity to the (Hadoop Distributed File System) for access the data. R programmers can browse, read, write, and modify files stored in HDFS through R. rhdfs package install only on the node that will run the R client. rhbase package provides basic connectivity to HBASE which is used for the storage purpose, using the Thrift server. R programmers can access the tables stored in HBASE through R. rmr2 package support the hadoop map-reduce technique to perform the statistical analysis on the hadoop cluster and this package install in every node which are present in the hadoop cluster.

The connection of R and Hadoop using Streaming is an easy task but R should be installed in every datanode. RHadoop have many advantage such as scalability, data integration, flexibility, functionality, transparency.

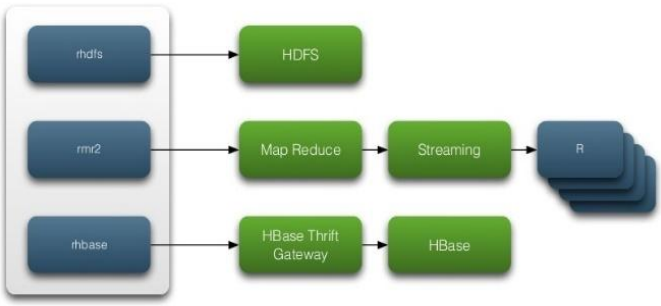

Fig.4. R with Hadoop Architecture

There are many advantage present in RHadoop which are given below:

- With hardware failure, it support the highly fault tolerance.
- Designed to be deployed on low-cost hardware
- RHadoop supports the SDA (Streaming Data Access)
- The combination of R and Hadoop support LDS (Large Data Sets) to perform the analysis.
- Portability Across Software Platforms and Heterogeneous Hardware

4. Support Vector Regression

Support vector Machine (SVM) is based on the statistical learning theory. For the regression and classification task, it has very powerful and useful tool. Basically support vector regression is used in the time series problem and regression problem. The best example of the time series data is Gold price data.

Now we are represent the basic concept of the support vector regression is a given dataset, $D = (X1, Y1) \dots \dots (Xp, Yp) \dots$

Where $i \in X$, $yi \in R$, P is the size of training data. X represents the space of the instance, R^P.The basic intension is to determine the function L based on the dataset D. there is the equation to find the value of the function L with respect to the function $\emptyset()$. The equation is

$$\mathbf{L}(x,c) = \sum_{i=1}^{D} c_i \phi_i(x) + b \quad \text{.....(i)}$$

Where the function $\{\phi_i(x)\}_{i=1}^{D}$ are called the features and $\{c_i\}_{i=1}^{\infty}$ and b are the coefficients which have to be determined by the dataset.

The featured dimensionality can be finite or infinite. The unknown coefficient R(c) determine by the given functional,

$$R(c) = \frac{1}{\mathbf{P}} \sum_{i=1}^{\mathbf{P}} \left| y_i - \mathbf{L}(x_i, c) \right|_{\varepsilon} + \lambda \|c\|^2 \quad \text{.....(ii)}$$

Where the symbol λ is the constant and the robust error function has defined, and in this function it finding the mod value of the function.

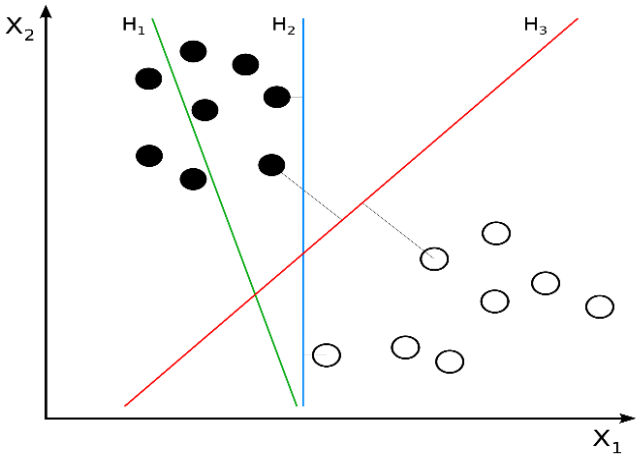

Fig.5. Basic concept of SVM

$$\left| y_i - \mathbf{L}(x_i, c) \right|_{\varepsilon} = \begin{cases} 0 & if \quad \left| y_i - \mathbf{L}(x_i, c) \right| < \varepsilon \\ \left| y_i - \mathbf{L}(x_i, c) \right| - \varepsilon & otherwise \end{cases} \quad \text{....(iii)}$$

In equation (ii) the function that minimizes the functional depends upon parameter which is finite and it has the given form,

$$\mathbf{L}(x, \alpha, \alpha^*) = \sum_{i=1}^{\mathbf{P}} (\alpha_i^* - \alpha_i) K(x, x_i) + b \quad \text{....(iv)}$$

Where $\alpha_i^* \alpha_i = 0$, $\alpha_i^*, \alpha_i \geq 0$, $i = 1, \dots, N$ and K(x,

y) is the kernel function and K(x, y) finding the inner product in the featured space,

$$K(x, y) = \sum_{i=1}^{D} \phi_i(x)\phi_i(y) \quad \text{.....(v)}$$

Which is satisfies the Mercer's condition and symmetric condition. K is the kernel function and there are many kernel are available such as Gaussian, trigonometric polynomial function and tensor product etc. the α and α^* are the coefficient and it determine the maximization and the equation is,

$$R(\alpha^*, \alpha) = -\varepsilon\sum_{i=1}^{p}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{p}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) \quad \text{...(vi)}$$

With respect to this equation the constraints $0 \le \alpha_i^*, \alpha_i \le C$ and $\sum_{i=1}^{N}(\alpha_i^* + \alpha_i) = 0$ and the number of the coefficient $\alpha_i^* - \alpha$ will be totally different from the Zero value.

The support vector regression (SVR) is used as a learning algorithm to understand the pattern from the input and to predict gold price as output based on that learning. This process is divided in two phases, training data phase and testing data phase and all the Stages from these two phases are given below:

1) Training phase
- Stage 1: Read the randomly selected training dataset from local repository.
- Stage 2: Apply windowing operator to transform the data into a generic dataset. This step will convert the last row of a window within the time series into a label or target variable. Last variable is treated as label.
- Stage 3: apply the cross validation process (CVP) of the produced label from that operator in order to feed them as inputs into support vector regression model.
- Stage 4: Select type of kernel and select special parameters of support vector regression
- Stage 5: apply that model into the dataset and observe the performance or accuracy.
- Stage 6: If accuracy is good than go to step 6, otherwise go to step 4.
- Stage 7: Exit from the training phase & apply trained model to the testing dataset.

2) Testing phase
- Stage 1: Read the randomly selected testing dataset from local repository.
- Stage 2: Apply the training model to test the out of sample dataset for gold price prediction.
- Stage 3: Produce the gold price predicted trends

5. Decision Tree

The decision tree is the visualization form that have a root node and the leaf node. The leaf node contain the results. There are two type of nodes are present in the decision tree one is inner node and other in terminal node. Basically there are two type of the decision tree can be drown in the gold price forecasting one is classification tree and other is regression tree. Classification tree analysis is when the predicted result is the class to which the data belongs that is called Classification tree analysis and when the predicted result can be considered a real number Regression tree analysis.

Decision tree is the method to find the target value and check the possibility of the trends with the different branches. In the decision tree all are instances are represented as the attribute values and it automatically perform the reduction of the complexity and selection of the features and regarding the predictive analysis its structure is vary understandable and interpretable.

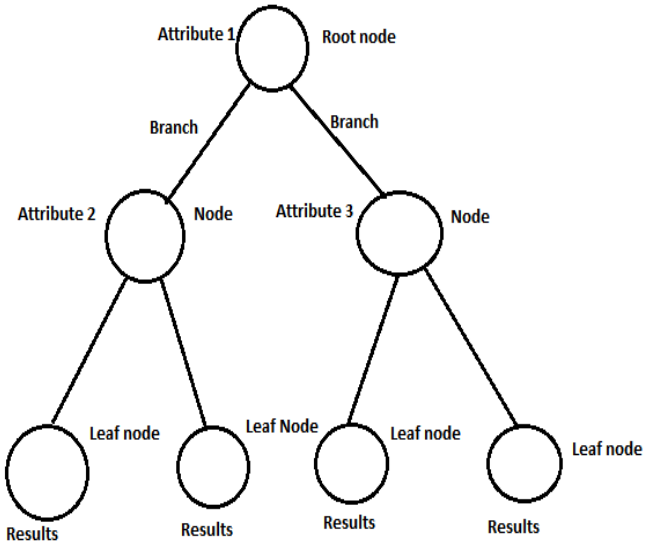

Fig.6. Decision tree architecture

It starts from the root node and step by step it goes down till terminal node to interpret the result. Decision tree is the best approach to predict the gold value and it give the best result at the time of prediction of the gold price. For each node, we calculate the EMV (expected monetary value), and place it in the node to indicate that it is the expected value calculated over all branches emanating from that node. There are four key advantage present in the decision tree
- It implicitly perform feature selection or variable screening
- It require relatively little effort from users for data prediction and data preparation
- Nonlinear relationships between parameters
- Easy to interpret and explain to executives

After performing all the experiments in the gold price data sets. We have to compute three error that are MSE, MAD, and MAPE.

$$\text{Mean Square Error (MSE)} = \frac{\sum E_i^2}{P} \quad \text{.....(vii)}$$

$$\text{Mean Absolute Deviation(MAD)} = \frac{\sum |E_i|}{P} \quad \text{....(viii)}$$

$$\text{Mean Absolute Percentage Error(MAPE)} = \frac{\sum \left[|E_i| \div X_i \right]}{P} \quad \text{.....(ix)}$$

Where the mean square error (MSE) is the average of the squares of the difference between the forecasted price and the actual price, the mean absolute deviation (MAD) is the average of the absolute values of the error and the mean absolute percentage error (MAPE) is the average of the absolute values of the percentage error of a forecast. Based on these three error we have checked the better algorithm for the gold forecasting. All the error are check the measurement and performance for both decision tree analysis and support vector regression.

6. Conclusion

There are five factor are present in the gold data which are open value, close value, low value, high value and volume. Gold provide an effective and useful means of diversifying a portfolio. The way to achieving success with the gold is to know your goals and risk profile before jumping in. The volatility of the gold can be harnessed to accumulate wealth, but left unchecked, it can also lead to ruin. Based on these attribute we have predicted the result from both method .decision tree are best for the feature selection and SVR are best for the large amount of the dataset. But there are some problem in the SVM. It takes long time to train the dataset. Decision tree takes less time to process the data. Decision tree have less mean square error then the SVM.

References

[1] K.Sahu, R.Panwar, “Exchange Forecasting Using Hadoop Map-Reduce Technique”, S.Tilekar, R.Satpute. April 2013
[2] Shahriar Shafiee “An overview of global gold market and gold price forecasting” , Erkan Topal 2010
[3] K.Sahu, R.Panwar, “Exchange Forecasting Using Hadoop Map-Reduce Technique”, S.Tilekar, R.Satpute. April 2013
[4] Daniel Keim “Big-Data Visualization” Huamin Qu, Kwan-Liu Ma 2013.
[5] Lucas, K. C. Lai, James, N. K. Liu, “Stock Forecasting Using Support Vector Machine,” In: Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, pp. 1607-1614, 2010.
[6] Tak-chung Fu, "Adaptive Data Delivery Framework for Financial Time Series Visualization ", Fu-lai Chung, Fu-lai Chung, Chun-fai Lam, Robert Luk 2005
[7] Z. Ismail “Forecasting Gold Prices Using Multiple Linear Regression Method”, A. Yahya, A. Shabri 2009
[8] Big data Decision tree analytics available online www.treeplan.com/chapters/introduction-to-decision-trees.pdf
[9] Ashesh Anand “Forecasting Gold Prices using Time Series Analysis”, Piyush Dharnidharka.
[10] Big data analytics available online “searchbusinessanalytics.techtarget.com/definition/big-data-analytics”.
[11] A. Smola and B. Scholkopf, “A Tutorial on Support Vector Regression,” Technical Report NeuroCOLT NC-TR-98-030, 1998.
[12] P.Chandarana,”Big Data analytics frameworks”, M.Vijayalakshmi, 2014.
[13] Xiaoyan Bai,"Context adaptive visualization for effective business intelligence”, White, D., Sundaram, D.
[14] Jinson Zhang,"5Ws Model for Big Data Analysis and Visualization", Mao Lin Huang.
[15] Hadavandi, E., "Developing a Time Series Model Based on Particle Swarm Optimization for Gold Price Forecasting", Ghanbari, A., Abbasian-Naghneh, S.
[16] Kim, Kyoung-jae, “Financial time series forecasting using support vector machines,” In: Neurocomputing 55, pp.307 – 319, 2003.