Trigger System For Commodity Price Prediction Using Data Mining Techniques

Viraj Gada[1], Sumeet Deshpande[2], Apoorva Dhakras[3], Chetashri Bhadane[4]

Computer Engineering, Mumbai University
Vile-Parle

viraj gada1994 @g mail. co m

sumeet suhas deshpande @g mail. co m

apoorva. dhakras @g mail. co m

[4]*Assistant Professor*
Computer Engineering, Mumbai University, Vile-Parle

chetashri @g mail. co m

***Abstract*—In the modern era, markets have been becoming increasingly unstable. The prices of commodities ranging from valuable metals like gold, silver to utility metals like iron, copper and lifeline entities like gasoline, crude oil are becoming extremely volatile. Such a scenario acts as a deterrent for any investor with diminished possibilities of returns. In such times, any product that can bring stability and an increased sense of predictability to the markets possesses the ability to entirely alter the market dynamics and bring a much needed boost to the confidence of investors. While such experimentation and research on commodity prices has been conducted by the much-acclaimed quants in the Wall Street, there has been scarce research conducted on prediction of commodity prices in India. The perceived reasons for this is the lack of dataset availability, disorganized market and suspect accuracies of the algorithms being used. Also, atmost only 2 algorithms were used to test every commodity. In this project, we have made use of OHLCV (Open-High-Low-Close-Volume) charts which are used to generate the everyday movements of concerned commodities. We have developed an indigenous system to predict the open prices of many commodities using a trigger method that runs six algorithms to predict various parameters of accuracy. And after evaluating these parameters a decision is made and the best algorithm is selected for predicting the open prices of the commodities for that dataset. The results of our implementation and outputs of this project have been notably mentioned in this paper along with its further potential applications and effects on the Indian economy.**

***Keywords*— data mining, trigger system, svm, multilayer perceptron, rbf networks, m5 rules, linear regression, decision tree**

I. INTRODUCTION

The current market has become extremely volatile. There has been a loss of investor confidence due to unpredictability of the market prices. However, the changes in prices of commodities are caused due to a variety of factors. In an attempt to predict the prices of the commodity, the algorithms of machine learning have been employed to find patterns in data and predict the prices. This paper focusses on our technical analysis on the commodity price prediction done using data mining techniques and machine learning algorithms. Technical analysis refers to movement of price and their future price prediction. To conduct the technical analysis, we first used OHLCV (Open-High-Low-Close-Volume) charts. The trends of the commodities are spotted in these datasets and stock movements are viewed on a daily basis. The datasets are obtained from the website quandl which is a marketplace for financial and economic data delivered in modern formats for today's analysts. These datasets are divided into 70% training dataset and 30% testing set. Six algorithms are run on these datasets and these algorithms first train the datasets. After training, the 30% dataset is used to test the algorithms accuracy. The algorithm with the least root relative squared error and highest co-efficient correlation is used to predict the most accurate data. The other algorithms can also be used to predict the values as well. The implemented results have been mentioned in the outcome in the form of comparison table to enable easy interpretation.

II. PROBLEM DEFINITION

The technology has made increased forays into various aspects of modern life. In the financial sector, the concepts and algorithms of machine learning have been implemented to find patterns and trends in stock prices. As a result, the stock market prices have become more predictable along with more statistical analysis and increased research is underway to increase the accuracy of these algorithms further. However, the same cannot be said about the commodity prices. There has been a dearth of research on this prospect. The main hindrances were insufficient datasets and insufficient accuracies. Also, only one or two algorithms were used to test every commodity.The main motivation of this project is to create a system that implements the machine learning algorithms on datasets of commodities on the lines of stock price predictability. The concepts of machine learning have been used to conduct a technical analysis to find trends and patterns in datasets. A trigger system has been used to provide

increased accuracy for every commodity. Thus, this project involves a two-level problem addressing the issues of prediction and accuracy. Six algorithms have been used to predict the commodity prices and the best algorithm amongst them has been used to predict prices for that commodity. This not only increases the accuracy but helps us to know which algorithms works best for which dataset. Thus, this implementation with improved accuracy brings with it a renewed way of commodity price prediction. It also aims to increase investor confidence and this can help the market take appropriate measures to reduce the commodity prices.

III. Literature Review

Interest in forecasting of prices of various commodities has become a topic of interest in recent times. This section presents a very brief review of the related and recent studies. G.M.Nasira and H.Hemageetha [1] investigated the into the feasibility of performing for forecasting vegetable prices using back propagation neural networks. The author performed experimentation on matlab with 88% accuracy. Sarunas Raudy and Indre Zliobate [3] used neural networks and time series to reduce error rate for prediction. Kuncheva [4] states classification algorithms for possible improvements with gc relations to random noise, random trends and systematic trends. Moody states and shows in [6], that there exists an optimum for training window length at which test error is minimal. Regression analysis methods include a simple regression model, a multivariate regression model, etc. (Kleinbaum, Kupper, Nizam, et al., 2013). Kuo and Xu(1998) put forward a decision tree for sales prediction using fuzzy neural networks. This is used in our development of decision tree algorithms. Mark Orr[8] suggests RBF networks as a suitable alternative for predicting commodity prices as well. Usefulness of the above papers has been an invaluable part of our research as we have implemented this project.

IV. Algorithms

A. Support Vector Machine

Analysis of data for regression and classification analysis can be done efficiently by supervised learning models like Support Vector Machines. A model built by SVM allocates new instances to a category. Instances are represented in space as points and mapped so that instances of different category are divided by a proper gap. The same space is then used for mapping other instances and on basis of which category they fall, a prediction is made.

B. Linear Regression

The most common and basic analysis is done by linear regression. Relationship between one dependent variable and one or more independent variable can be properly estimated by regression. It is also used in the description of data. Regression involves fitting a single line through a scatterplot. It can be explained as y=m*x+c where, y is estimated variable, c is constant, b is regression co-efficient and x is independent variable. Issues involved in linear regression are concerned with model fitting. Increasing a variable leads to over fitting while underfitting happens when a cause-effect relation is tried to be proved by linear regression.

C. RBF Network

Set of radial basis function is implemented by hidden nodes in this 2-layer feed forward network. Similar to multi-layer perceptron, the output nodes employ linear functions in this algorithm. There are 2 stages to network training:

1. The values from input to hidden layers are made known.
2. Then, the training is quickly done and networks are implemented.

Classification is performed by RBFN by measuring the inputs similar to instance from training set. A sample prototype is stored by each neuron as an instance from training set. The Euclidian distance then measured by inputs and prototype is computed by each neuron to classify input. According to what the input looks similar to more closely, classification is done.

D. Decision Tables

The observed values about an item are mapped to conclusion about its target value by using a decision table learning. Data Mining, machine learning and statistics use decision tables as one of the predicitve models where finite values can be taken by a target variable. Branches represent feature conjunction while leaves represent class labels. Regression trees are those decision tables where target variables can take continous values.

The main aim of decision tables is to forecast the target variable value based on several input variables. Decision trees consisting of trees with nodes called root that has no incoming edges. An outgoing edge of a node is called test or internal node. Other nodes are decision nodes.

E. Multilayer Perceptron

A set of input data is mapped onto a set of appropriate outputs by multilayer perceptron. Multilayer Perceptron consists of an adirected graph with multiple node with each layer fully connected to other one. Every node has a activation which is non-linear and is a neuron. The standard linear perceptron is modified to form multilayer perceptron and the distinguished data is non-linearly separable.

In this algorithm, supervised learning is conducted by training on a dataset where number of dimensions for input is given by m and number of dimension outputs is given by O. By giving a target and feature set, we can learn approximattion for either regression or classification.

F. M5 Rules:

The M5 rules is at majority classified with a single conjunctive rule learner. Cross validation concept is used as well. The spliting is done by a categorical variable and the synthetic binary variable is chosen to split. The difference in data increases as categorical variables increases. Linearization to decision rukes is done for trees generated.

The method of building a tree in M5 is of 2 phases: growing phase and pruning phase. In growing phase, the tree starts from one node and recursively tries to split while in pruning phase it accomodates regression model to remove unimportant branches and prunes them. After that, smoothing is done. Smoothing increases accuracy but makes tree very difficult to interpret.

V. Outcomes and Discussion

TABLE I
Accuracy Parameters for Gold

Gold	L.R	D.T	MLP	RBF	SVM	M5
CC	0.998	0.995	0.999	0.911	0.998	0.997
RAE	1.29	8.65	1.53	34.87	1.31	1.61
RMS	0.538	3.04	0.597	13.08	0.539	0.653
RMSE	**1.691**	9.57	1.89	41.12	1.79	2.05

A sample output of one of the several commodities gold which we have implemented is given here. The acronyms are:
L.R- Linear Regression
D.T-Decision Tables
MLP-Multilayer Perceptron
RBF-Radial Basis Function
SVM-Support Vector Machine
CC-Correlation Co-efficient
RAE-Relative Absolute Error
RMS-Root Mean Square Error
RMSE-Relative Mean Squared Error.

As we see from table 1, the values for various accuracy parameters is given the most crucial being RMSE. The lowest RMSE value provided by Linear Regression is highlighted in the table. Similar outcomes have been conducted for commodities like Petroleum, Gasoline, Silver and many more.

VI. Accuracy Parameters

A variety of accuarcy parametrs have been considered and their values have been measured. Each of them has its own importance which has been mentioned below

A. *Correlation Co-efficient*

Correlation co-efficient is associated with the linearity of the variable under question with the other variables. The higher the correlation, the better the accuracy and results.

B. *Relative Absolute Error*

The absolute error is the magnitude of the difference between the exact value and the approximation. The relative error is theabsolute error divided by the magnitude of the exact value.

C. *Root Mean Squared Error*

The square root of the mean/average of the square of all of the error.The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions.

D. Relative Mean Squared Error

It gives us the average error of the dataset and is widely used as the parameter for calculating the accuracy. It helps us to know the error rate of every algorithm and whch is the best algorithm to be used.

VII. Modular description

The modular description of the entire working system is given in figure 1 on page 4. As shown, the data is first trained and during this time the algorithms starts learning. Then testing of algorithms is done. After that, the dataset is tested on all 6 algorithms and the best algorithm is triggered.

VIII. Conclusion

Thus we successfully implemented this system on various commodities the results of which can be seen in the outcome. We have also managed to keep the accuracy parameters high and near about the range of 90%. Thus, this indigenous trigger system developed by us holds the promise of providing a thorough statistical analysis and prediction for any commodity. This system acts as a basis for monte carlo simulation besides also providing an objective analysis into a highly uncertain domain. The prediction of the variable open is very close to that of the observed values while testing and gives us an insight into the accuracy of this system to predict the commodity prices for next day anytime we want. Thus, our aim to provide a system that could bring stability to the market commodities while also increasing accuracy has been fulfilled by the implementation of this project.

References

[1] G.M.Nasira and H.Hemageetha, *Forecasting of vegetable prices using backpropagation neural network*, International Journal of Computational Intelligence and Information, 2012.
[2] Wenjie Huang, Qing Zhang, Wei Xu, Hongjiao Fu, A Novel Trigger Model for Sales Prediction with Data Mining Techniques, China, Data Science Journal, May- 2015.
[3] Sarunas Raudy and Indre Zliobate, Time Series in Machine Learning, Oregon, Oregon State Journal, May-2015
[4] Kuncheva, Data Mining Classification Techniques, Russia, IEEE-Sochi, August 2012

[5] R. E. Erasmus, Machine Learning for Prediction, Boston, MITedu, March-2011
[6] L.Moody, Regression analysis for prediction, Stanford, May-2014
[7] M. Shell. (2002) Slideshare homepage on CTAN. [Online]. Available: http://www.slideshare.org/dmc-
[8] Mark Orr A Library for RBF Networks, Belgium,Journal of Machine Learning Research, Jaunuary-2014

FIGURE 1