Gold Market Analyzer using Selection based Algorithm

Sima P. Patil[1], Prof. V.M.Vasava[2], Prof. G.M.Poddar[3]

[1]Student Dept. of Comp. Engg., Gangamai COE, Dhule, Maharashtra, India
[2]Professor, Department Comp. Engg, Gangamai COE, Dhule, Maharashtra, India
[3]Head, Department Comp. Engg, Gangamai COE, Dhule, Maharashtra, India

Abstract— *Gold is the most important and valuable element right from its discovery. It is the most significant element at present and also the most valuable asset to. In the present market scenario the investors use gold as the security for their shares investment. During international trade all the countries use gold as their main mode of transaction. It is seen that all the currencies accepted by the international market are having the gold as the backup for their economy. The prices of gold are rising day by day continuously. As we see in the history of gold market the present prices of gold are much high as compare to the past values and that's why the gold market has attracted the most attention. The paper focus on the continuous changing in the gold rates, investment policies depend on the forecasting of trends in gold which will help the data mining companies to minimize the risk The description of the future situation on the basis of present trends is just not limited to the forecasting the prices. The knowledge discovers by the data mining techniques is gathered from the different gold related websites and also from the jewellers database. It is much more important for the ornaments making companies to know the demand and the requirements for the ornaments during the unstable (uncertain) market conditions. For the classification purpose the maid and sale database was gathered from the nearest jewellery shops of the past 5 years. The prediction is done after complete analysis of the gathered data set. With this the paper concentrate on making the information available about the government and private schemes related to gold market on one place. The paper proposes the system that gives the total access to the registered user and limited access to the unregistered user to get the required information of gold. The latest updates are provided to the registered user by sending mail or text message when the user was offline.*

Keywords—Historical gold market forecasting mineral prices, Trend Analysis, Correlation, Demand, Long-term trend reverting Jump and dip diffusion.

I. INTRODUCTION

Database Machine learning has often been applied to the prediction of financial variables, but usually with a focus on stock prediction rather than commodities. In my project, I chose to apply supervised learning to the prediction of gold prices in order to see what kind of success I could achieve. For the purposes of profitability, it is more important to predict the relative change in price tomorrow (i.e. whether the price will go up or down) than to predict the absolute price tomorrow, so I have formulated this as a classification problem: given historical price data up to a given day, my algorithm attempts to predict whether the gold price tomorrow will be higher or lower than it is today. The most reliable source of gold price data that I found was the London afternoon (PM) price fix, which I obtained from the USA Gold website. As gold is not traded in the same sense as stocks, there is no separate data available for, for instance, daily open vs. close prices or trading volume. I have gathered the daily price fix data spanning early 2007 to late 2013 - about 7 years, where 5 out of every 7 days are eligible data points, which yields about 1700 training examples. Another question that I sought to address in this project is whether or not inter market financial variables, like stock index prices, exchange rates, bond rates, and other commodity prices, can be effective inputs to the prediction of gold prices. Historically, these variables have exhibited a significant correlation; but I was curious as to the reliability of these relationships as well as their applicability in extremely short-term situations.

II. SURVEY REVIEW

1. Shelke focuses on mainly those techniques that are most commonly used for short term load forecasting. For classification purpose the month wise electric load consumption dataset is gathered from Symbiosis Institute of technology, Pune campus of the year 2012 and 2013.Results have been classified by measuring Holt-winters model parameter and results has been conducted that adding parameter like temperature and event summarized to highlight the accuracy.
2. Juberias developed a model which is based on a time series analysis methodology and includes the action

of meteorology as an explanatory variable. Information about the influence of meteorology on hourly electrical load is given to the model as an explanatory variable by using the daily electrical load forecast.

3. Bush currently using our Code Hawk abstract interpretation technology to produce such a formal metric for 6 C programs from NIST's SATE competition. A single mathematical definition of undefined memory access according to the C semantics covers 27 of the MITRE Common Weakness Enumeration categories, and allows us to identify a set of program locations where proof obligations guaranteeing memory-safety cannot be discharged.

III. FORECASTING TECHNIQUES

The paper mainly focuses on short term electric load forecasting techniques. Furthermore also discuss forecasting applications.

1. **Short term load forecasting using k-Nearest Neighbours (NN) algorithm**
   Short term load forecasting using k-NN algorithm by SARIMA and Holt winters. After load classification using 1-year training set and 1-year test set forecast was performed through two models. By measuring their mean absolute percentage error, Holt winters models was shown to have better performance in short term load forecasting model. Advantage of NN is algorithm estimate target concept locally and differently for each new instance to be classified. NN algorithm learns dataset quickly and it is robust to noisy training data. Disadvantage of NN algorithm is it requires large storage requirements because it has store all data. NN algorithm is slow during instance classification because all training instance have to be visited. Accuracy of NN algorithm degrades with increase of noise in training attributes.
2. **Short-term power load forecasting based on SVM**
   Short term electric load forecast is also carried out by machine learning method is support vector machine (SVM). From recent years support vector machine become commonly used in short term electric load forecasting. SVM perform binary classification and regression estimate task.
   They are increasingly popular as classification and learning because of two factors unlike other classification technique SVM minimize expected error rather than minimize classification errors. SVM shows that the forecasting result based on support vector machine is data also increase with irrelevant better than those other methods. Besides the advantages of SVM from a practical point of view perhaps the most serious problems predict events only few time periods such as days, weeks and months into future. Problem with SVMs is the high algorithmic complexity and extensive memory requirements in large scale tasks. Although SVMs have good generalization performance, they can be poorly slow in test phase.
3. **Short-term load forecasting with artificial neural network [ANN].**
   Short-term load forecasting has become an essential part of human's life. In the past few decades, many forecasting models have been presented. By using neural network model, which divides the electric load into two parts: the load scaled curve and the day maximal load and minimal load. The day maximal load and divides the electric load into two parts: The development of this model consist of three phases in which initially start from historical data, every day is classified according to its load profile by means of self-organizing feature maps the second consists of building and training the neural networks for each class and the third is an online operation phase in which prediction is carried out by previously trained recurrent neural network. In ANN structure number of input parameters, number of hidden neurons are system dependent mainly they determined by size of training set and number of input variables.

IV. ADDITIONAL FEARTURES

1. Inter market Variables

Gold prices and commodity prices in general, may also be related to other financial variables. For instance, gold prices are commonly thought to be related to stock prices; interest rates; the value of the dollar; and other factors. Therefore, I wanted to explore whether these other variables could be effective inputs to the prediction of gold prices. I collected the following data over the same time period as the gold price fix data, from early 2007 to late 2013:

- US Stock Indices: Dow Jones Industrial Average (DJI); S&P 500 (GSPC); NASDAQ Composite (IXIC)
- World Stock Indices: Ibo Vespa (BVSP); CAC 40 (FCHI); FTSE 100 (FTSE); DAX (GDAXI); S&P/TSX Composite (GSPTSE); Hang Seng Index (HSI); KOSPI Composite (KS11); Euronext 100 (N100); Nikkei 225 (N225); Shanghai Composite (SSEC); SMI (SSMI)
- COMEX Futures: Gold futures; Silver futures; Copper futures; Oil futures
- FOREX Rates: EUR-USD (Euro); GBP-USD (British Pound); USD-JPY (Japanese yen); USD-CNY (Chinese yuan)

- Bond Rates: US 5-year bond yield; US 10-year bond yield; Euro bund futures
- Dollar Index: Measures relative value of US dollar

2. Results with Additional Features

I ran logistic regression (the same format as above, with 10-fold CV) with the full set of features calculated for both gold prices and all intermarket variables listed above. The table below displays the results:

Table 1. Logistic Regression (Full Feature Set)

Precision	63.76%
Recall	63.89%
Accuracy	61.92

As shown above, adding extra features provided a major boost to the accuracy of the classifier: when I included the full feature set, accuracy jumped from 54.23% to 61.92%, and precision from 55.03% to 63.76%. These additional inter market variables are thus very useful for predicting the rate of change in gold price.

3. Computing Correlation

As adding a multitude of features that actually have little or no effect on gold prices would add noise and negatively affect my predictions, I first want to quantify the correlation that certain variables have with gold prices to determine whether they might be beneficial to include [3]. For all of the intermarket variables listed above, I calculated the Pearson correlation coefficient over the roughly 1700-day data period between (a) the London PM gold price fix on a given day, and (b) the close price of the variable. The table below displays all quantities by descending absolute-value correlation coefficient with the gold price fix:

The correlation values lead to a variety of very interesting observations, and here I discuss several salient ones. First, commodity futures are very highly correlated with the gold spot price, especially futures for gold itself. Also, when the dollar is valued high, it takes comparatively less dollars to buy the same amount of gold, so the gold price (measured in dollars) will be low. For this reason, when variables which increase with the value of the dollar (i.e. the dollar index or the USD-CNY exchange rate) are high, the dollar is high so the gold price is low. Conversely, when variables which decrease with the value of the dollar (i.e. the EUR-USD exchange rate) are high, the dollar is low so the gold price is high.

Table 2. Correlation Coefficients

1. Gold Futures	0.72
2. Silver Futures	0.50
3. Copper Futures	0.27
4. EUR-USD	0.24
5. Dollar Index	-0.21
6. Oil Futures	0.21
7. S&P/TSX Composite	0.18
8. USD-CNY	-0.17
9. GBP-USD	0.16
10. KOSPI Composite	0.13

4. Feature Selection Based on Correlation

Consider, for instance, the three US stock indices; these all have a very low correlation with gold prices, and thus are likely not helpful for gold price prediction. In fact, they may even add noise and cause the classifier to over fit the training data, leading to higher test error. Indeed, training and testing on the full data set with all of these features achieves a train accuracy of 81.69%, which is high compared to the test accuracy of 61.92%.

The outcome of this experiment was extremely positive. Ultimately, the optimal feature set included 10 of the top 15 most correlated variables (of 25 total). These are:

- 4 commodities futures: gold, silver, copper, oil
- Dollar index
- 3 exchange rates: EUR-USD, GBP-USD, USD-CNY Hang Seng Index and Nikkei 225

V. VARYING THE ALGORITHM

1 Revisiting SVM

Running SVM on the gold-price-only feature set was disappointing, as was running SVM on the feature set with all inter market variables. However, a linear-kernel SVM with l-1-regularization actually performed quite well on the final feature set (normalized by scaling, as in the SVM section above), achieving accuracy of 69.08% with 10-fold CV. Apparently these new features provide the SVM just enough information to make predictions with accuracy very close to, but still not quite matching, that of logistic regression.

2 Logistic Regressions

There are not too many parameters of the logistic regression model to experiment with, but I tested variations on (a) the norm used in penalization, and (b) the C parameter, which specifies inverse of regularization strength. I found that the optimal combination occurred with the l-1 penalty and the default C = 1.0, precisely the parameters of the model in Table 3. Thus this model attains the highest accuracy, precision, and recall of anything I tried, making my overall peak classification accuracy 69.30%.

VI. CONCLUSION

Finally, the simulation illustrates that this algorithm could actually have significant profitability. With a classification accuracy of nearly 70%, even the extremely simple trading agent set forth in the simulation was able to attain a considerable average daily profit. Overall, these experiments have further piqued my interest in the

applications of machine learning to financial prediction, as determining which features to add and which algorithms to use was quite interesting, and the end results were better than I expected. I would be interested in continuing to optimize this model and perhaps putting its predictions to practical use.

REFERENCES

[1] Megan Potoski, *"Predicting Gold Prices",* CS229, Autumn 2013

[2] G. Eason, B. Noble, and I. N. Sneddon, *"On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,"* Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[3] D.P Kothari& I.J Nagrath, *"Modelling and forecasting electricity loads"*, Tata McGraw Hill, 2003.

[4] H.Lee. Willis, *"Spatial Electric Load Forecasting*", Marcel Dekker, Inc, 2002.

[5] Juergen Schlabbach, Karl-Heinz Rofalski, *"Power System Engineering*", Wiley & Verley Co,2008.

[6] G. Juberias, R. Yunta, J. Garcia, Moreno C. *"Mendivil, A New ARIMA Model for Hourly Load Forecasting"*, Transmission and distribution Conference, 1999IEEE Vol.1, 11-16 April 1999, pp:314-319 vol.1

[7] Dae-Yong Kim, Chan-Joo Lee, Yun-Won Jeong, Jong- Bae Park, Joong-Rin Shin, *"Development of System Marginal Price Forecasting Method Using SARIMA Model",* Proceeding of conference KIEE Nov. 2005, pp.148-150.

[8] Hyeonjoong Yoo, Russel L. Pimmel, *"Short-Term Load Forecasting using a Self-Supervised Adaptive Neural Network"*, IEEE Transaction on Power System, Vol.14, No.2, May 1999.

[9] Jan Ivar Larsen. *"Predicting stock prices using technical analysis and machine learning"*. NTNU, 2010.

[10] Vatsal H. Shah., *"Machine learning techniques for stock prediction"*. NYU, 2007.

[11] Shunrong Shen, Haomiao Jiang, and Tongda Zhang., *"Stock market forecasting using machine learning algorithms"*. Stanford University, 2012.