A Novel Quantitative Behavioral Framework for Financial Markets Prediction

T. Ramesh Babu[1*] and M. Venkateshwarlu[2]
[1]Director, Centre for Quantitative Finance and Risk Analytics (CQFaRA), India
[2]Associate Professor, Department of Finance and Economics, NITIE, India
[2]Adjunct Professor, Indian Institute of Management Bangalore, India

Abstract

Effective prediction of financial asset prices has become a challenge in the present day volatile world. The use of mathematics have become very extensive in the financial world, most of the mathematical models concentrates on the market data rather than the behavior of the market from which the data has been generated. An attempt has been made for the first time to model the prediction of asset prices based on both the market data and the behavior of the market participants. The participants in the financial markets behave differently from each other, these behavioral differences can be attributed to the participants understating or/and his perception about the market. Each investor has his own perception about the market and he feel it is close to reality, but truly speaking it is not so. Each participant has his own impact on the market and the reality is the aggregation of each participant's perception. The impact of the investor's behavior has been modeled in the present quantitative behavioral approach by dividing the participants into broad categories based on their trading behavior. To model the participant's impact first one should predict the proportion of participants in each category. Most of the times, finding the exact number of participants in each category is not easily available from the market data, so an evolutionary based swarm intelligence model has been adopted in the present framework to find the proportion of the participants in each category. Finally the whole methodology has been applied to gold asset class (because gold is an international asset with increasing volatility these days) to validate the present method. The model is tested rigorously using different time varying samples to validate the present methodology; some interesting results have been obtained from the present study. The back testing results prove that the model presented in this paper is very effective in predicting the prices close to reality. The present frame work is very generic and can be applied to any financial asset class to estimate the returns close to reality.

1. Introduction

The prevalent theory of financial markets during the second half of the 20th century has been the efficient market hypothesis (EMH) which states that all public information is incorporated into asset

*corresponding author, ramesh.thimmaraya@gmail.com

prices, the market prices behave as though all traders were pursuing their self-interest with complete information and rationality. Toward the end of the 20th century, this theory was challenged in several ways. There were a number of large market events that cast doubt on the basic assumptions. On October 19, 1987 the Dow Jones average plunged over 20% in a single day, as many smaller stocks suffered deeper losses. To cater some of these deviations in the classical models a new discipline called Quantitative Behavioral Finance is emerging which uses mathematical and statistical methodology to understand behavioral biases in conjunction with valuation. The financial time series prediction in short term has gained much importance recently and most of the literature has been dominated by regression algorithms (Yang et. at. 2002). The use of mathematics have become very extensive in the financial world, most of the mathematical models concentrates on the market data rather than the behavior of the market from which the data has been generated. An attempt has been made in this paper to model the financial market based on both the market data and the behavior of the market. The market behavior is mainly influenced by the participants of the market (human factors or psyche). The market participants have been broadly divided into three categories; long term (investors), short term (speculators) and a small random component which may be attributed to the retail or noisy investors who move irrationally. The kernel of the present model is built on the heterogeneous expectations, which solves one of the major assumptions of the CAPM model of homogenous expectations if the present framework is applied to stock markets.

The schematic of the conventional methodology is shown in Chart-1, which tells us that each investor category forecasts the real market in his own way, so the result what he gets from the mathematics is his perception about the reality. The schematic of the present methodology is shown in Chart-2; this model is a blend of the quantitative techniques used for financial prediction and the behavior of the market participants who use these techniques to predict the financial asset prices. An important step of the present model is calculating the impact of each group of the participants on the market, to do this one should know the number of participants in each category, for most of the times the market data does not contain these facts. To resolve this issue a swarm intelligence algorithm called Particle Swarm Optimization has been used to predict the number of participants in each group.

The present model is very generic and can be applied to any asset class. In the present paper the framework has been applied to predict the gold prices, gold has been chosen because of its increasing volatility in the events of the present financial shocks. The remainder of this paper is organized as follows, in Section 2, we discuss about the present behavioral model. We have applied the whole methodology to predict gold prices. In section 3 we discuss about the back testing results from the gold data and finally we conclude in section 4.

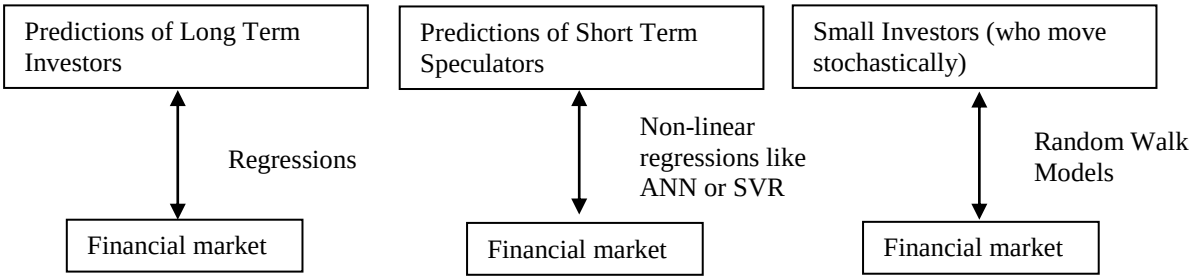

Chart 1: Conventional Prediction Models (Mathematical)

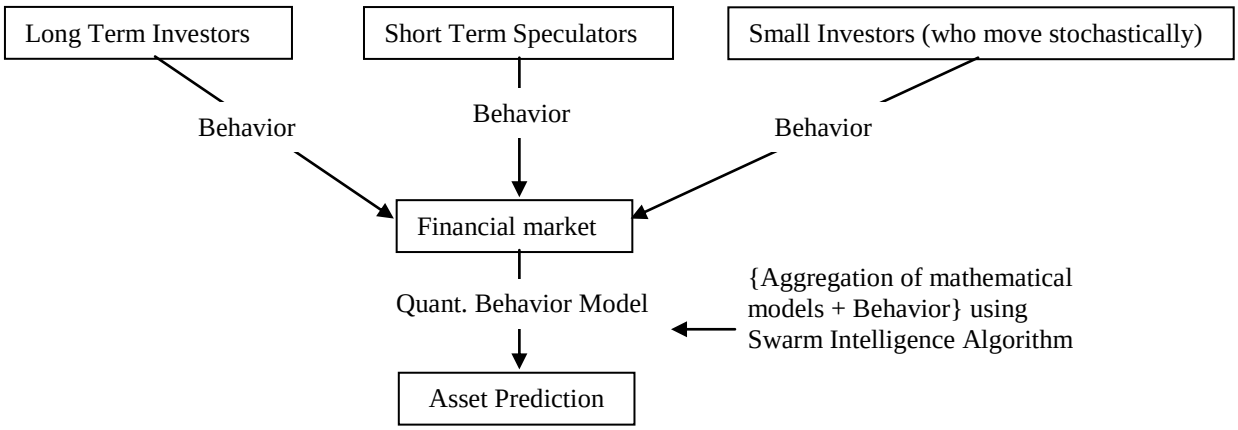

Chart 2: Present Model (Behavioral and Mathematical)

2. Principle of the Present Quantitative Behavioral Model

If we consider the present framework for the gold market, the participants are categorized into 3 groups (long term investors, short term speculators and retail or noisy investors). There are three major reasons why conventional methods or pure mathematics may fail to closely predict the actual market;

2.1 First Reason:

The first reason is that the long term investors predict the market from the long term trend of the market, most of the long term predictions are from simple linear OLS regression, which tells about the fundamental analysis and future projections of the growth. The results of the future projections (OLS regression) for the gold data are presented below. It is clearly seen from the Fig-1, that the OLS regression is a very good tool to estimate the long term trend of the asset prices time series.

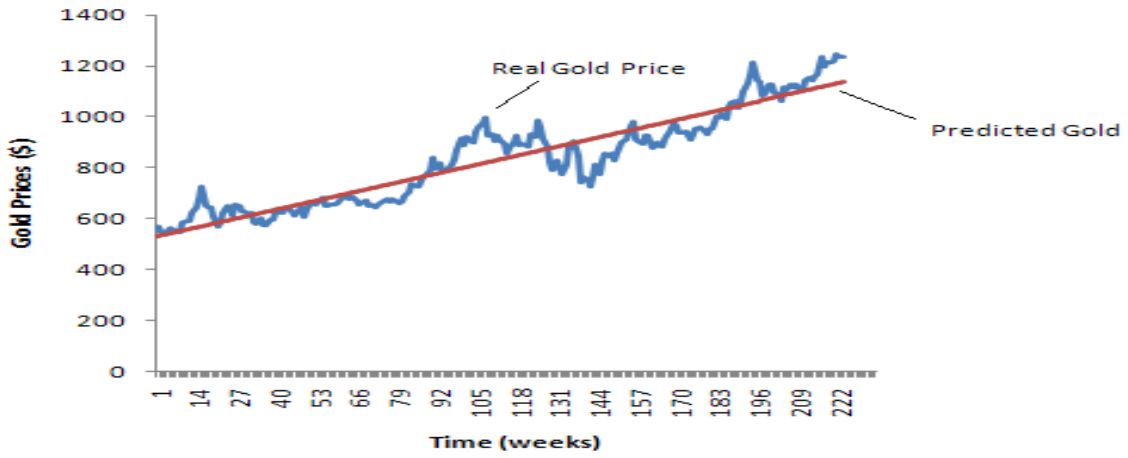

Figure 1: Long Term Prediction of Gold prices using regression analysis

The X-axis of Fig-1 has been normalized; where '0' indicates 1st week of Feb, 2006 and '235' indicate the last week of June, 2010. The Y-Axis is the gold prices in dollars.

2.2 Second Reason:

The second reason is that the speculators or short term investors predict the market based on the technical's or some complex patterns in the time series, these predictions are best estimated from very powerful non-linear mapping methods like Neural Networks or Support Vector Regression (SVR).

In the present paper we have used Particle Swam Optimization based Support Vector Regression to predict the short term gold prices; the methodology is as follows;

Suppose we are given training data $\{(x_1, y1),\ldots\ldots, (x_l, yl)\} \subset \chi \times \mathbb{R}$ where χ denotes the space of the input patterns (e.g. $\chi = \mathbb{R}^d$). The series y_i denote the gold prices measured at subsequent weeks and x_i denote the time in weeks. In ε-SV regression[vapnik,1995], our goal is to find a function f(*x*) that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this. This may be important if you want to be sure not to lose more than ε money when dealing with gold prices, for instance.

We begin by describing the case of linear functions *f*, taking the form

$$F(x) = < w, x > + b \text{ with } w \in \chi, \mathbf{b} \in \mathbb{R} \qquad (1)$$

Where $<.\ ,\ .>$ denotes the dot product in **χ.** Flatness in the case of eq. (1) means that one seeks a small *w*. one way to ensure this is to minimize the norm [3],i.e.,$\| w \|^2 = < w, w >$. We can write this problem as a convex optimization problem:

$$\text{Minimize} \quad \tfrac{1}{2}\| w \|^2$$

$$\text{Subject to} \begin{cases} y_i - < w, x_i > - b \leq \varepsilon \\ < w, x_i > + b - y_i \leq \varepsilon \end{cases} \qquad (2)$$

The tacit assumption in eq. (2) was that such a function f actually exists that approximates all pairs(x_i , y_i) with ε precision, or in the words, that the convex optimization problem is feasible. Sometimes, however this may not be the case, or we also may want allow for some errors analogously to the "soft margin" loss function in [Cortes and Vapnik [1995], one can introduce slack variables ξ_i, ξ^*_i to cope with otherwise infeasible constraints of the optimization problem eq. (2). Hence we arrive at the formulation stated in [Vapnik, 1995].

$$\text{Minimize} \quad \tfrac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi^*_i)$$

$$\text{Subject to} \quad \begin{cases} y_i - < w, x_i > - b \leq \varepsilon + \xi_i \\ < w, x_i > + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \qquad (3)$$

The constant C > 0 determines the trade-off between the flatness of *f and* the amount up to which deviations larger than ε are tolerated. This corresponds to dealing with a so called

$\varepsilon -$ Insensitive loss function $|\xi|_\varepsilon$ described by,

$$|\xi|_\varepsilon = \begin{cases} 0 & if |\xi| < \varepsilon \\ |\xi| - \varepsilon & \text{Otherwise} \end{cases} \qquad (4)$$

However for the sake of simplicity we will additionally assume 'c' to be symmetric and to have two (for symmetry) discontinuities at ±ε, $\varepsilon \geq 0$ in the first derivative and to be zero in the interval [−ε, ε]. Hence c will take on the following form.

$$C(x, y, f(x)) = \begin{cases} 0 & for \ |y - f(x)| \leq \varepsilon \\ \tilde{c}(|y - f(x)| - \varepsilon) & \text{otherwise} \end{cases} \qquad (5)$$

Note the similarity to Vapnik's ε- insensitive loss. It is rather straightforward to extend this special choice to more general convex cost functions. For nonzero cost functions in the interval [−ε, ε] use an additional pair slack variables. Moreover we might choose different cost functions $\tilde{c}_i$, $\tilde{c}_i^*$ and different values of ε_i, ε_i^* for each sample. At the expense of additional Lagrange multipliers in the dual formulation additional discontinuities also can be taken care of. Analogously to eq. (3) we arrive at a convex minimization problem [Smola et al., 1998]. We will stick however, to the notation of eq. (3) and will use C instead of normalizing by λ and *l*, as it contributes to the clarity of the exposition.

$$\text{Minimize} \quad \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\tilde{c}(\xi_i) + \tilde{c}(\xi_i^*)) \qquad (6)$$

$$\text{Subject to} \quad \begin{cases} y_i - < w, x_i > - b \leq \varepsilon + \xi_i \\ < w, x_i > + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Again by standard Lagrange multiplier techniques, exactly in the same manner as in the above case one can compute the dual optimization problem. We will omit the indices $_i$ and *, where applicable in order to avoid tedious notation.

This yield,

$$\text{Maximize} \begin{cases} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^* < x_i, x_j > \\ \quad -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) + C(T(\xi_i) + T(\xi_i^*)) \end{cases} \tag{7}$$

$$\text{Where} \quad \begin{cases} w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i \\ T(\xi) = \hat{c}(\xi) - \xi\, \partial_\xi \hat{c}(\xi) \end{cases}$$

$$\text{Subject to} \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha \leq C\partial_\xi \tilde{c}(\xi) \\ \xi = \inf\{\xi | C\,\partial_\xi \tilde{c} \geq \alpha \end{cases}$$

$$\alpha, \xi \qquad \geq 0$$

The crux of the SVR method is that the linear model in eq. 1 is made non-linear by introducing a kernel k (x,x') in place of vector x in eq.1. Most commonly used kernel is the Gaussian kernel which is $k(x, x') = e^{\|x-\bar{x}\|^2}/2\sigma^2$. The whole SVR models error minimization depends upon two parameters used in the algorithm they are C and σ.

However, most SVM practitioners select these parameters empirically by trying a finite number of values and keeping those that provide the least testing error. This procedure requires a grid search over the space of parameter values and needs to locate the interval of feasible solution and a suitable sampling step. Because of the computational complexity, grid search is only suitable for the adjustment of very few parameters. In farther researches, some intelligent algorithms such as evolution algorithms (EA) and genetic algorithms (GA) were employed to choose the parameters of a SVM model, and the improved model offers a superior performance to ordinary regression SVM model. The particle swarm optimization (PSO) algorithm, a relatively new evolutionary computation (EC) stochastic technique, can also be used as an excellent optimizer which originated as a simulation of the food-searching behavior of birds. Similar to EA and GA, PSO is a population based optimization tool, which search for optima by updating generations. However, unlike GA and EA, PSO has no evolution operators such as crossover and mutation. Compared to GA and EA, the advantages of PSO are that PSO is easy to implement and there are few parameters to adjust. Most versions of PSO have operated in continuous and real-number space.

PSO is a stochastic optimization technique introduced by [Kennedy and Eberhart, 1995], which is inspired by social behavior of bird flocking and fish schooling. The general principles for the PSO algorithm are stated as follows: Let us consider a swarm of size *n*. Each particle *Pi* (*i* =1, 2, . . . , *n*) from the swarm is characterized by: 1) its current position $X_i(k) \in R^d$ **,** which refers to a candidate solution of the optimization problem at iteration *k*; 2) its velocity $V_i(k) \in R^d$;and 3) the best position $P_{besti}(k) \in R^d$ that is identified during its past trajectory. Let $G_{besti}(k) \in R^d$ be the best global position found over all trajectories that are traveled by the articles of the swarm. Each of *n* particles fly

through the *d*-dimensional search space *Rd* with a velocity *V* (*k*) *i* , which is dynamically adjusted according to its personal previous best solution $P_{besti}(k)$ and the previous global solution $G_{besti}(k)$ of the entire swarm. The velocity updates are calculated as a linear combination of position and velocity vectors. The particles interact and move according to the following equations

$$V_i(k+1) = w(k).V_i(k) + C_1.R_1(k).(P_{besti}(k) - X_i(k)) + C_2.R_2(k).(G_{besti}(k) - X_i(k)) \quad (8)$$

$$X_i(k+1) = X_i(k) + V_i(k+1) \quad (9)$$

Where $V_i(k+1)$ is the velocity of $(k+1)^{th}$ iteration of i^{th} individual, $V_i(k)$ is the velocity of k^{th} iteration of i^{th} individual, w(k) is the inertial weight used as a tradeoff between global and local exploration capabilities of the swarm.

The results of the PSO-Support Vector Regression for the gold data are presented below. It is clearly seen from the Fig-2, that the PSO based SVR is a very powerful tool to estimate the short term pattern in the asset prices time series.

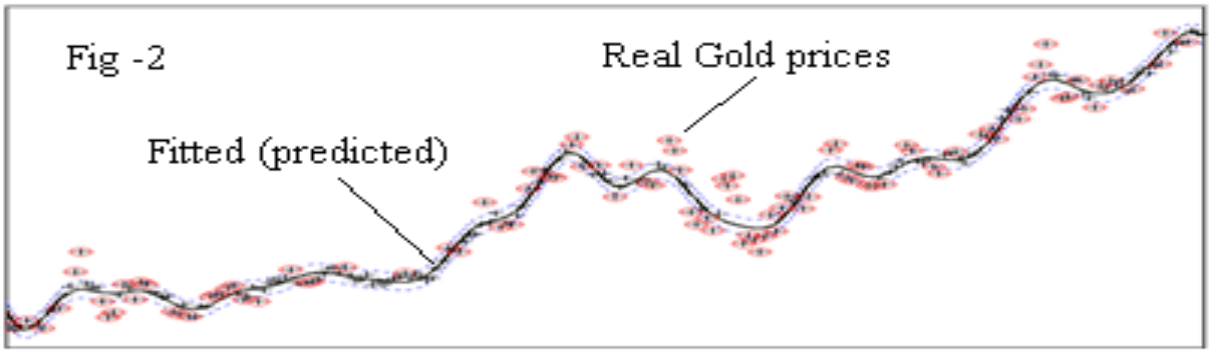

Figure 2: Short term prediction of Gold prices using PSO based Support Vector Regression, X-axis represents the time and Y-axis represents the gold prices.

2.3 Third Reason:

The third reason is that the retail investors who behave stochastically based the random information available publically. They move according to the market news created by some sources or news channels. Their behavior is generally irrational they don't use economic knowledge to judge the impact of the news. They don't use any prediction tools specifically, so their behavior can be closely traced from the random walk model and can be solved using Monte Carlo Simulations.

$$dy = \mu * y * dx + \sigma * y * \epsilon * (dx)\text{^}0.5 \quad (1)$$

Where, 'y' represents gold price, 'x' represents time, μ represents mean of the gold prices, σ represents the standard deviation of the gold prices, ϵ is the stochastic variable generated from the normal distribution.

2.4 Quantitative Behavioral Model (QBM):

Finally after observing the behavior of each of the participant, we know the actual market consists of many participants whose behavior is very different from each other. The actual market price is an aggregation of all the participant's expectations in the financial market. The impact of each participant depends upon the proportion of each participant; we call these as the weights which are dynamic in nature. The final model looks like;

$$E(Y_{QBM}) = w1 * E(Y_{Investors}) + w2*E(Y_{Speculators}) + w3*E(Y_{Retail}) \quad (2)$$

It is almost impossible to find the weights w1, w2 and w3 from the market data, so the present methodology has again adopted the evolutionary based optimization algorithm called the Particle Swarm Optimization to find these weights from the historical data.

The objective function for the PSO algorithm is the minimization of root mean square error obtained from eq.2. The parameters are w1, w2 and w3. For the gold data, the proportion of long term investors are about 52%, short term speculators are about 41% and stochastic retail investors are about 7 %, our weights predictions are consistent with the rarely available market data [6].

3. Back Testing

The data considered for the present analysis is the weekly gold prices from Feb, 2006 to August, 2010. The data for gold prices in $'s has been collected from the Bloomberg database and the data has been divided into two groups, one is the training set (Feb, 2006 to June, 2010) and the other is the validation set (July, 2010 to August, 2010).

3.1 Forecast Results

3.1.1 Long term investors

The long term investors are of the gold market always interested in the price appreciation or the trend, so most of them use OLS regression to predict the average return they can earn. The average return is nothing but the long term trend of the data; the predicted results for July, 2010 to August, 2010 using the regression analysis are shown in Fig-3.

3.1.2 Short term speculators

The short term speculators are more interested in the price movements in short periods of time. The price movements in the short term will be very complex and it depends on the pattern of the price. So these complex relationships can be explained using non-linear mapping methods like Artificial Neural Networks (ANN) or Support Vector Regression (SVR) for the prediction. As explained above that SVR has a better prediction power than the ANN because a major problem

with the ANN is over fitting of the data. The predicted results for July, 2010 to August, 2010 using PSO-SVR are shown in Fig-4.

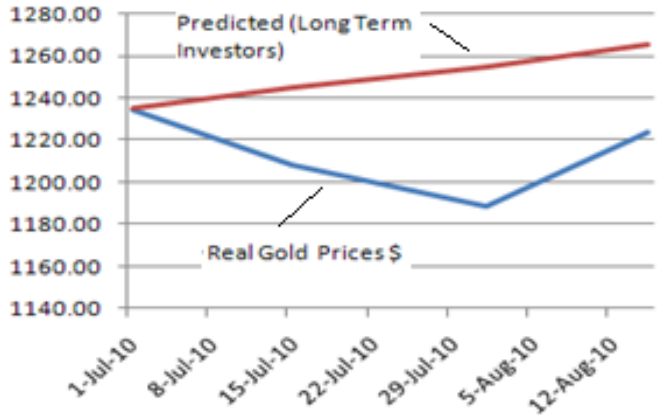

Figure 3: Forecasting of Gold prices using OLS Regression (Long Term Investors)

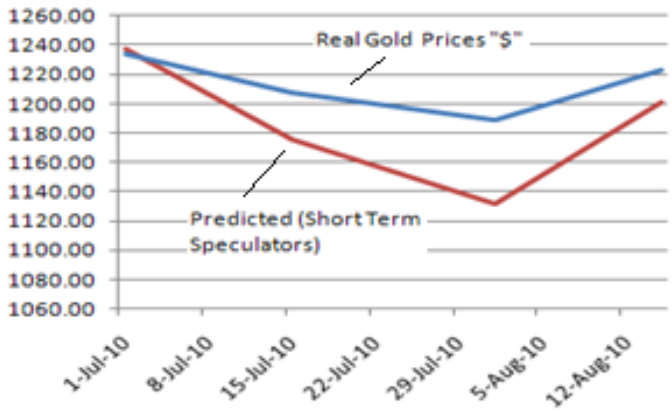

Figure 4: Forecasting of Gold prices using Support Vector Regression (Short Term Speculators)

3.1.3 Retail Investors (Random Component)

The retail investor's behavior can be modeled using random walk model described in the above section. The Monte Carlo simulation has been used to predict gold prices from July, 2010 to August, 2010; the results are shown in Fig-5.

3.1.4 Forecasting Gold Prices using the present Quantitative Behavioral Model

By applying the present model that we have discussed in section 2, the prediction of the gold prices from July, 2010 to August, 2010 are shown in Fig-6.

3.2 Comparative Performance of Conventional and the Quantitative Behavioral Model (QBM):

The performance is calculated from the root mean square error (RMSE) calculated from the difference between the real market prices and the predicted prices. The Table-1 shows the RMSE for all the models discussed in the present study. The error values in the table -1 clearly proves that the investors in each group think that they know the market better than the others, but each of them predict the market with huge errors. The QBM prediction has yielded a least error compared to the conventional methods; this proves that the QBM predictions are much closer to the reality.

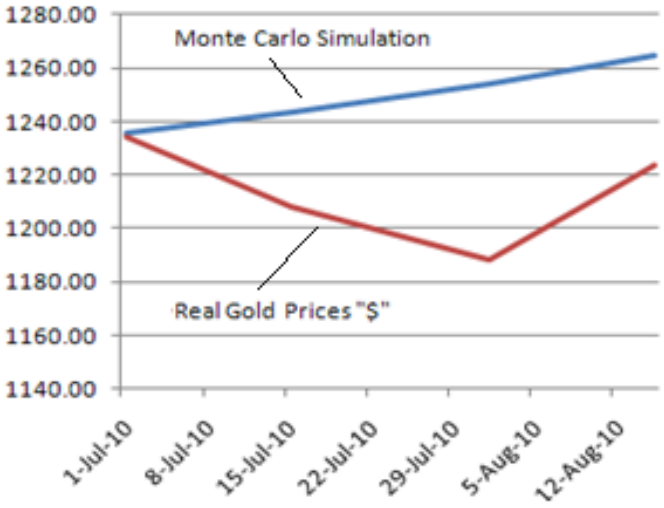

Figure 5: Forecasting of Gold prices using Monte Carlo Simulation (Retail Investors)

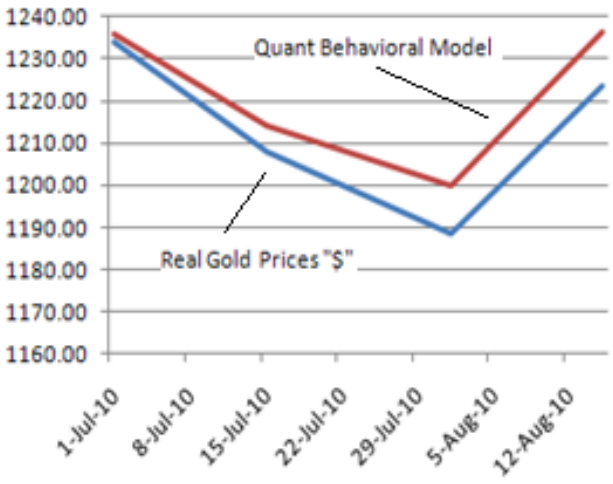

Figure 6: Forecasting of Gold prices using Quantitative Behavioral Model

Table 1: Performance of different models

Model	Long Term (Investors)	Short Term (Speculators)	Random Walk (Retail Investors)	QBM
RMSE ($)	21.73	17.19	21.26	4.53

5. Discussions and Conclusions

The participants in the financial markets behave differently, these behavioral biases can be attributed to the participants understating or/and his perception about the market. As observed from the results, each investor has his own perception about the market and he feel it is close to reality, but truly speaking it is not so. Each participant has his own impact on the market and the reality is the aggregation of each participant's perception. The present work is an attempt to model the aggregation of each participant's perception (at least in broad groups) to arrive close to the reality.

To validate the present novel quantitative behavioral model it has been applied to the gold asset prediction. The back testing results are as follows, it is observed from Table-1 that each individual has a root mean square error of around plus or minus 20 dollars in predicting the reality, but the present quantitative behavioral model has an error of around 5 dollars. The present approach has reduced the RMS error by around 75% which is very interesting. The model is tested rigorously using different time varying samples to validate the present methodology; the results indicate that in the best case the error was reduced by 75% and the worst case error reduction is around 50%. This indicates that the model presented in this paper is better in predicting the financial asset prices better than the conventional methods. The present frame work is very generic and can be applied to any

asset class to estimate the market returns close to reality. This research will have a great impact in predicting the implied CAPM return; as an extension to this study we are working on this frame work to estimate the stock market returns in a better way by developing an implied behavioral CAPM model.

References

[1] H. Yang, L. Chan, Laiwan and I. King. Support Vector Machine Regression for Volatile Stock Market Prediction. *IDEA. 2002 LNCS 2412*.2002: 391-396.

[2] Shi Y H, Eberhart R C. A Modified Particle Swarm Optimizer. *IEEE International Conference on Evolutionary Computation, Anchorage, Alaska* .1998: 69-73.

[3] A.Smola, B.Scholkopf, and K.R Muller. General cost function for support vector regression. *Proceedings. Of the Ninth Australian conference on neural networks.*1998: 79-83.

[4] C. Cortes and Vapnik. Support vector networks. *M. learning.* 1995, 20: 273 – 297.

[5] [4] Kennedy J, Eberhart RC. Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks, Perth, Australia.* 1995: 1942-1948.

[6] *http://www.technical indicators.com/gold.htm*