Status Report: Team 8
Predicting Prices of Oil and Gold

Ayush Sengupta, Benjamin Lin, Komal Sanjeev, and Sreevathsan Ravichandran

Department of Computer Science, Stony Brook University,
Stony Brook, NY 11794-4400
{aysengupta,xianlin,ksanjeev,sravichandra}@cs.stonybrook.edu
http://www.cs.stonybrook.edu/~skiena/591/projects

1 Background Updates

1.1 Objective

Our objective is to predict the prices of Oil and Gold on January 1st 2015 as of December 1st in 2014 (a month in advance).

1.2 Baseline Models

In order to illustrate an improvement in the accuracy of our predictions, we compare our current autoregressive and multiple linear regressive models to the following baseline models:

- **Oil/Gold price is the same as the previous month's price:**
  $P_t = P_{t-1}$
- **Oil/Gold price is a weighted mean of price of previous k months:**
  $P_t = \frac{1}{\{k(k+1)\}^2} \sum_{i=1}^{k} (k-i+1)^3 P_{t-i}$

Error Metrics of Baseline Models The following are the error metrics for our baseline models:

Model	Relative Error	Mean Absolute Error	Root Mean Squared Error
$Model0.0$	6.73746	5.384215	7.101763
$Model0.1$	7.27174	5.761054	7.754635

Table 1. Error Metrics for Baseline Model - Oil

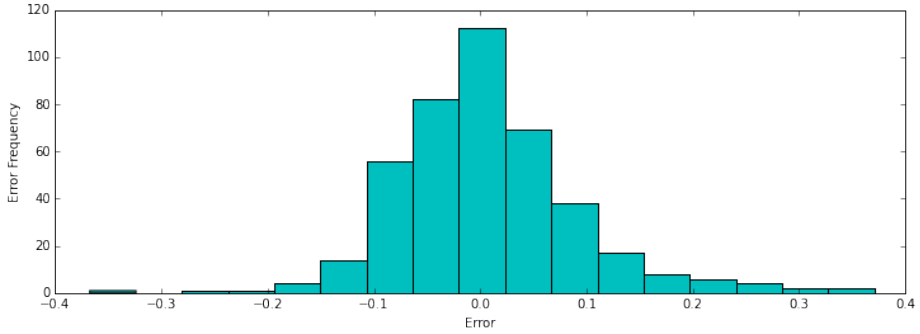

Fig. 1. Error plot for baseline model - oil/gold price is the same as the previous month's price

Model	Relative Error	Mean Absolute Error	Root Mean Squared Error
$Model0.0$	3.674560	29.142500	49.049942
$Model0.1$	3.606751	28.856644	47.804751

Table 2. Error Metrics for Baseline Model - Gold

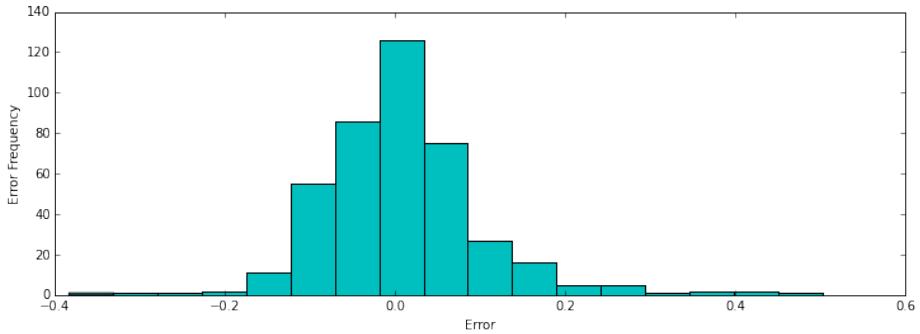

Fig. 2. Error plot for baseline model - oil/gold price is a weighted mean of price of previous k months

1.3 Autoregressive and Multiple Linear Regressive Models?

1.4 Spot price prediction using Futures prices

A futures contracts is an agreement between two parties to buy or sell a specific asset at a time in the future for a price decided today. One party takes a long position - it agrees to buy the asset or commodity, and the other party takes a short position - it agrees to sell the same asset or commodity. The purpose of these contracts is to hedge risks.

A futures market is a financial exchange where futures contracts are traded

2 Data Matrices

Our data consists of multiple time series of monthly Oil and Gold prices, and the following macroeconomic factors:

- S&P 500 Index
- New York Stock Exchange Index (NYSE)
- US Dollar Index
- Consumer Sentiment Index (CSI)
- EURO-USD Conversion Rate

Date	Oil Price	Gold Price	S&P 500	NYSE	USD Index	EUR-USD	CSI
10/31/2014	80.53	1164.3	2018.05	10845	80.8143	1.27103	86
9/30/2014	91.17	1216.5	1972.29	10702.93	81.0908	1.3401	86
8/31/2014	97.86	1285.8	2003.37	11046.29	77.9769	1.362	86
7/31/2014	98.23	1285.3	1930.67	10726.43	77.2128	1.3826	86
6/30/2014	106.07	1315	1960.23	10979.42	75.7271	1.3759	86

Fig. 3. Data frame for oil and gold price and its related macroeconomic factors

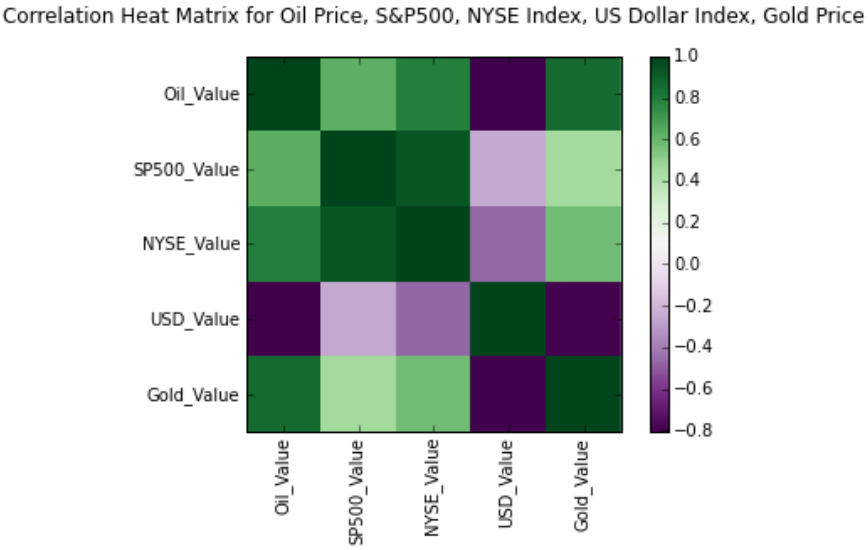

Fig. 4. Correlation Heat Map for Oil Price and related macroeconomic factors

	$S\&P\ 500$	$NYSE$	$USD\ Index$	$Gold\ Price$
$OilPrice$	0.718178	0.805763	−0.77365	0.872

Table 3. Correlation Matrix - Oil

The correlation heat map in Figure 4, and the corresponding table containing correlation coefficients show the correlation between the price of oil and related economic factors. They indicate a high correlation between oil price and certain macroeconomic factors - S&P 500, NYSE, US Dollar Index, Gold Price.

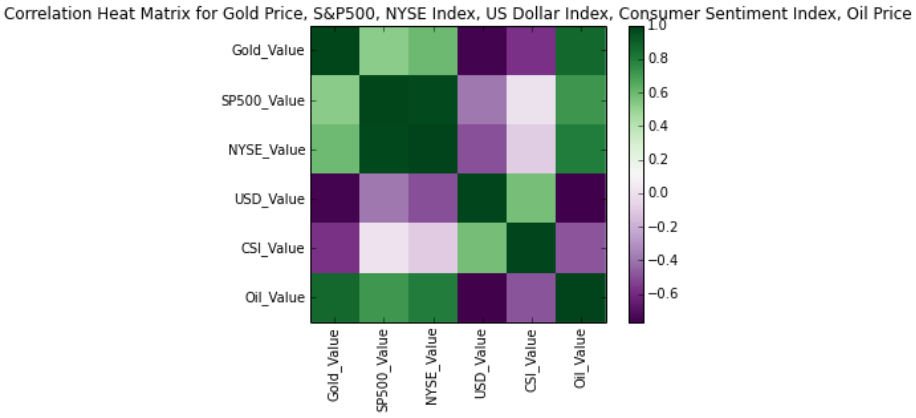

Fig. 5. Heat Map and Correlation Coefficient Matrix for Gold price and its related macroeconomic factors

	$S\&P500$	$NYSE$	USD	$EUR-USD$	CSI	$OilPrice$
$GoldPrice$	0.461485	0.578967	−0.792106	0.628074	−0.727766	0.860482

Table 4. Correlation Matrix - Gold

The correlation heat map in Figure 5, and the corresponding table containing correlation coefficients show the correlation between the price of oil and related economic factors. They indicate a high correlation between Gold Price and certain macroeconomic factors - US Dollar Index, EURO-USD Conversion Rate and Oil Price.

2.1 Data Sources

The data was obtained from the following sources:

Oil Price, Gold Price, S&P500, NYSE, USD, EURO-USD: `https://www.quandl.com`
Consumer Sentiment Index: `http://future.aae.wisc.edu/data/monthly_values/by_area/998?grid=true`

[ADD: HOW SATISFIED ARE YOU WITH THE DATA YOU COLLECTED?]

3 Development/Evaluation Environment

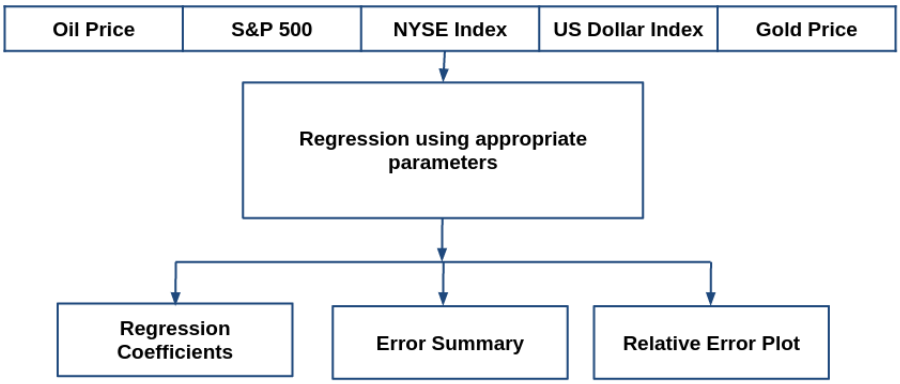

Fig. 6. The development and evaluation environment for the oil/gold price prediction with economic factors involved

Figure 6 shows the workflow of our prediction model. The model uses past prices of oil/gold, combined with certain macroeconomic factors, and performs regression to make predictions. Error metrics are used to determine the accuracy of our predictions.

The model takes as an input a data frame containing multiple time series of oil/gold prices and related macroeconomic factors. A parameter corresponding to each of these factors is also passed as a part of the input to the model. This parameter is a integer representing the number of lags in the autoregressive factor, and is used to predict the current value of the time series. We can also choose not to consider a particular factor by assigning 0 to the parameter.

The model is trained on the initial 60% of the time series, and tested on the remaining 40%. It tries to generate a linear function while assigning coefficients

to each of these economic factors. These regression coefficients are then used to make predictions for the next month.

In order to ensure that that our model does not over-fitting the curve, the data is cross validated using a size k slice of from training set(k is nearly equal to 80%). This slice is then slid across the training set, and finally an average of the different predictions of each slice is calculated. However, since there was no improvement in the results, we did not incorporate it as a part of our model.

Error Summary and Relative Error Plot. The following error metrics are calculated - Mean Relative Error, Mean Absolute Error, and Root Mean Square Error. A histogram of the relative error frequencies is also generated.

4 Current Model and Baseline

We have developed autoregressive and multiple linear regressive models to make oil and gold price predictions.

Autoregressive Model Autoregressive models are based purely on historical prices. They model the time series as a linear function of the values of the past 'p' months.

$$X_t = c + \sum_{i=1}^{p} \varphi_{t-i} X_{t-i}.$$

The Odinary Least Squares (OLS) method is used to estimate the parameters of the regression function. It tries to minimize the sum of squares of vertical distances between the predicted and the actual values.

Autoregressive and Multiple Linear Regressive Model Linear regression models the relationship between two variables - a dependent variable and an explanatory variable using a linear function. The process of modelling a variable based on more than one explanatory variables is called Multiple Linear Regression.

We initially develop a model which generates a purely autoregressive function. Then, the model is expanded to incorporate the factors which are highly correlated to the price of oil/gold we are trying to predict. As each factor is incorporated into the model, we perform a comparison of the error metrics between these models and try to estimate the model which makes predictions which best accuracy.

For predicting the price of oil, the following macroeconomic factors are taken into consideration:

- S&P 500 Index
- NYSE Index
- US Dollar Index
- Gold Price

For predicting the price of gold, the following factors macroeconomic are taken into consideration:

- S&P 500 Index
- NYSE Index
- US Dollar Index
- EURO-USD Exchange Rate
- Consumer Sentiment Index
- Oil Price

Autoregressive Moving Average Model (ARMA) ARMA models are used to understand and predict time series values as a function of two polynomials, an autoregressive function, and a moving average function.

$$X_t = c + \sum_{i=1}^{p} \varphi_{t-i} X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_{t-i} X_{t-i}.$$

In ARMA (p,q), p is referred to as the order of the autoregressive part and q is referred to as the order of the moving average part, i.e, the model is described using p autoregressive terms and q moving average terms.

Why other autoregressive models fail to perform much better than baseline? ACF: Autocorrelation Factor PACF: Partial autocorrelation Factor(autocorrelation with the linear dependence between variables removed)

[ADD]

4.1 OIL

Autoregressive and Multiple Linear Regression Models .

We initially try to predict the price of oil solely on the basis of the past values of oil prices. Figure 7 shows the performance of the autoregressive model as the number of months taken into consideration by the model is increased.

We then incorporate the macroeconomic factors into the model. We start by including one factor at a time to the autoregressive model and use the error metrics to compare the performance of these models. For each of these models, we plot a graph of the relative error against the number of months taken into consideration. Figure 8, Figure 9, and Figure 10 show the error plots for different models.

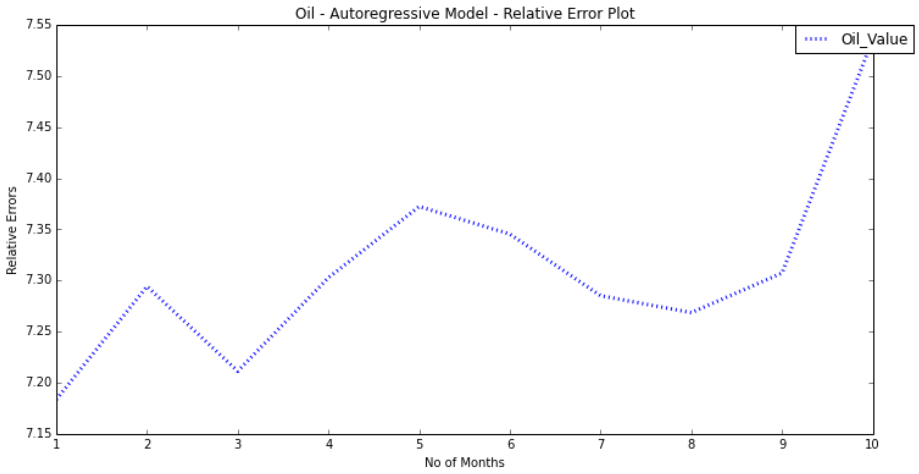

Fig. 7. Performance of the autoregressive model with increasing number of months

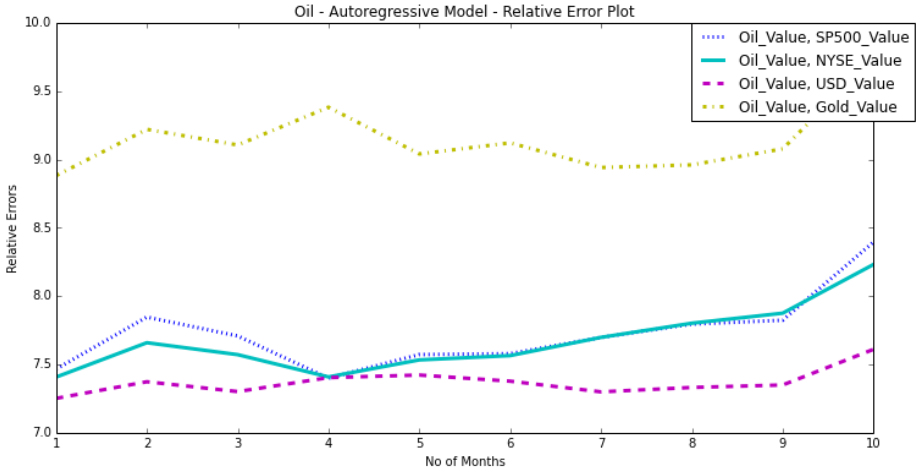

Fig. 8. Performance of the autoregressive model considering one factor with increasing number of months

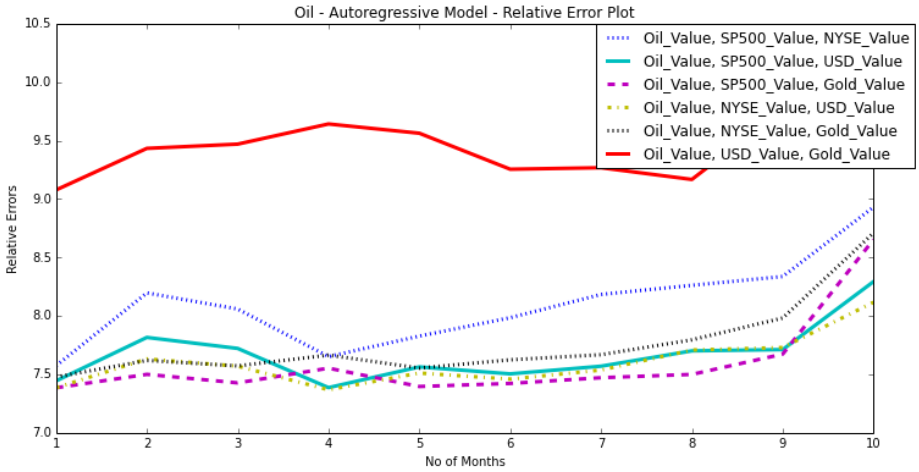

Fig. 9. Performance of the autoregressive model considering two factors with increasing number of months

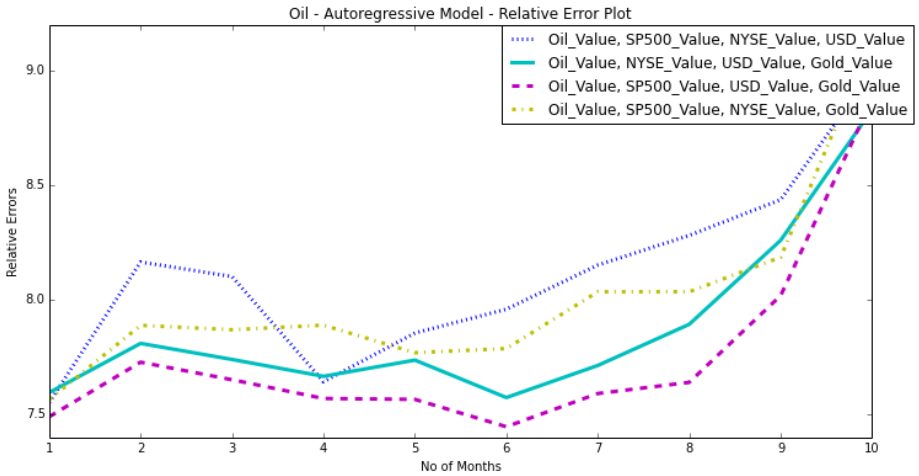

Fig. 10. Performance of the autoregressive model considering 3 factors with increasing number of months

ARMA Model

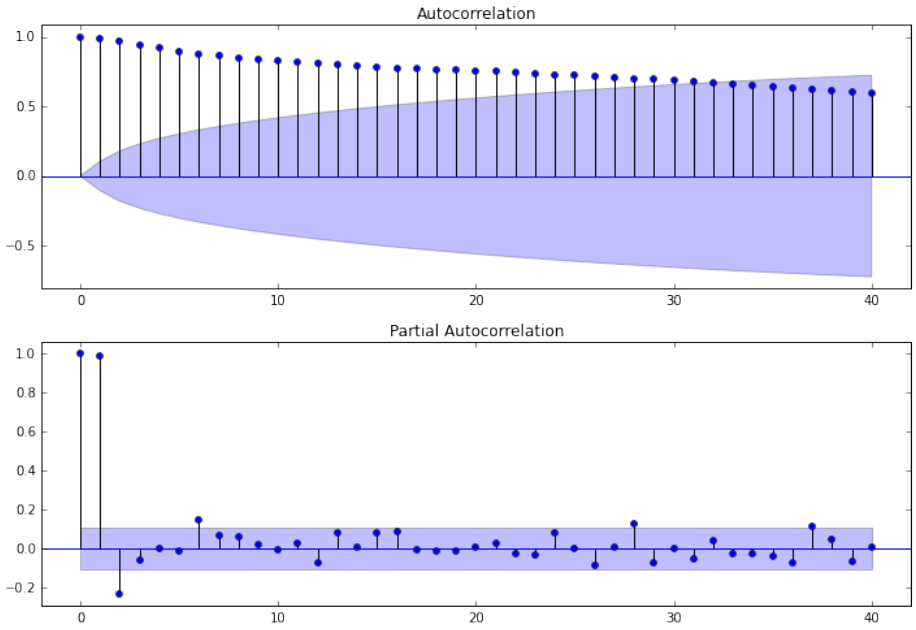

Fig. 11. ARMA Autocorrelation and Partial Autocorrelation factors

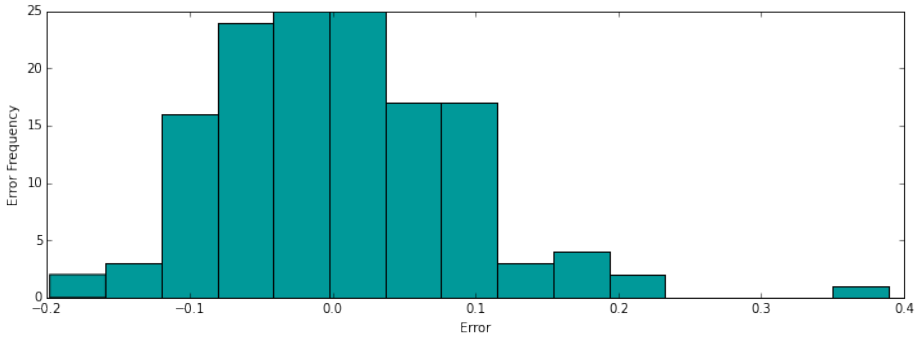

Fig. 12. ARMA Error Distribution

Comparison Summary - Oil The following is a summary of the comparison between autoregressive and multiple linear regressive models against the baseline models.

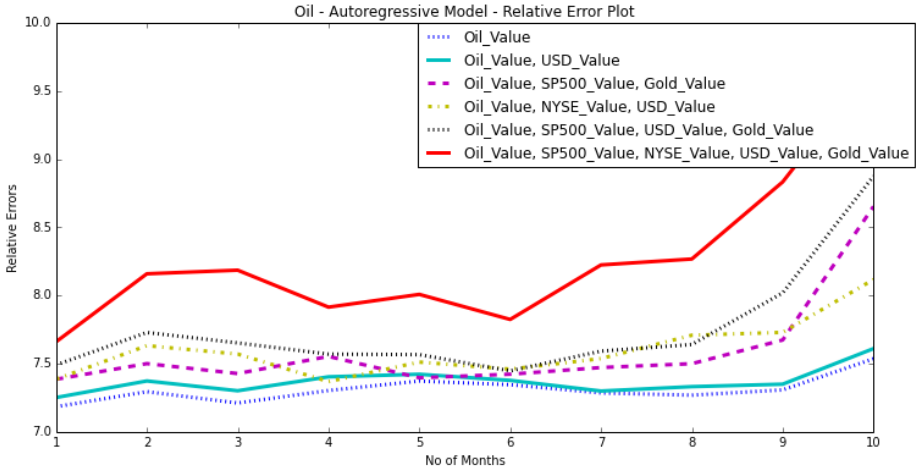

Fig. 13. Performance of the autoregressive model considering 3 factors with increasing number of months

Figure 13 shows a comparison of the best performing models considering various econometric factors.

The following table shows

Factors	Relative Error	Mean Abs Error	RMSE
$BaselineModel0.0$	6.73746	5.384215	7.101763
$BaselineModel0.1$	7.27174	5.761054	7.754635
$Autoregressive$	8.49	5.384215	6.672619
$USD\ Index$	8.69	5.097663	6.67896
$S\&P500,\ Gold\ Price$	9.04	5.160815	6.737639
$NYSE,\ USD$	9.05	5.241797	6.767898
$S\&P500,\ USD,\ Gold\ Price$	9.38	5.26046	6.79724
$S\&P500,\ NYSE,\ USD,\ Gold\ Price$	10.86	5.494503	7.000936
$ARMA\ (6,0)$	5.0322462866	6.42218763098	6.78951497738

Table 5. Summary of performance of various models for predicting oil prices

4.2 GOLD

Autoregressive and Multiple Linear Regression Models .

We follow a similar procedure for gold - we initially try to predict the price of gold solely on the basis of historical gold prices. We then add one economic factor at a time, compare the different models obtained to derive the best performing model. We only include factors with absolute correlation coefficients equal or greater than 0.55 with the gold price. Hence, we discard S&P500 for gold.
Figure 14 shows the performance of the autoregressive model as the number of months taken into consideration by the model is increased.

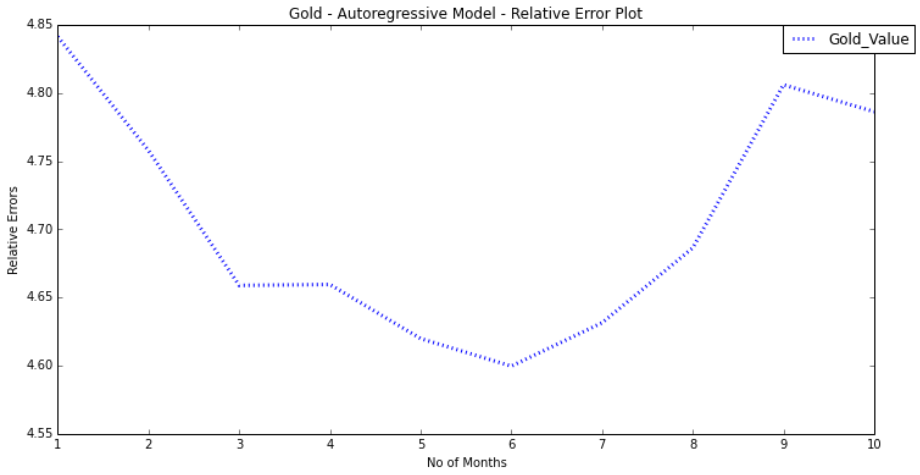

Fig. 14. Performance of the autoregressive model with increasing number of months

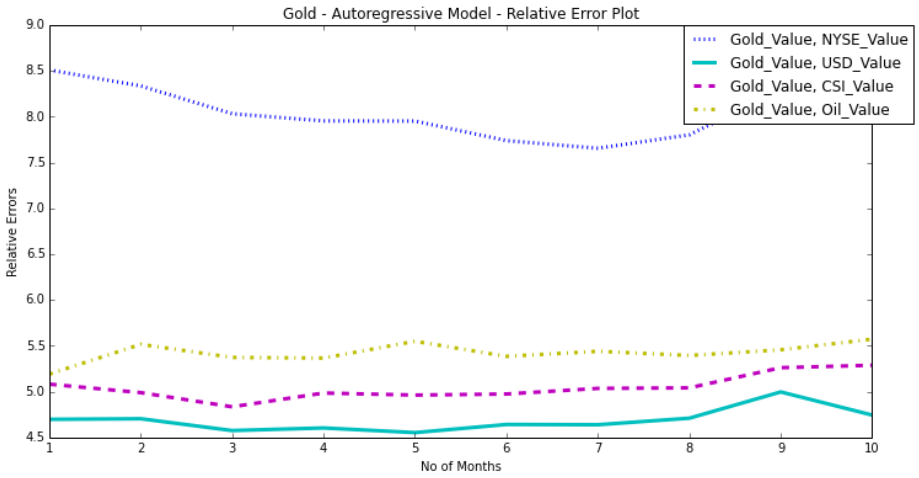

Fig. 15. Performance of the autoregressive model considering one factor with increasing number of months

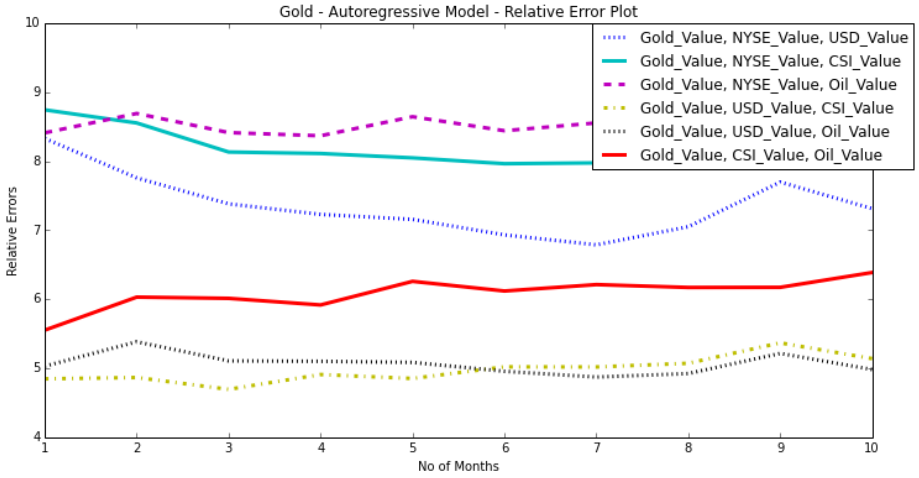

Fig. 16. Performance of the autoregressive model considering two factors with increasing number of months

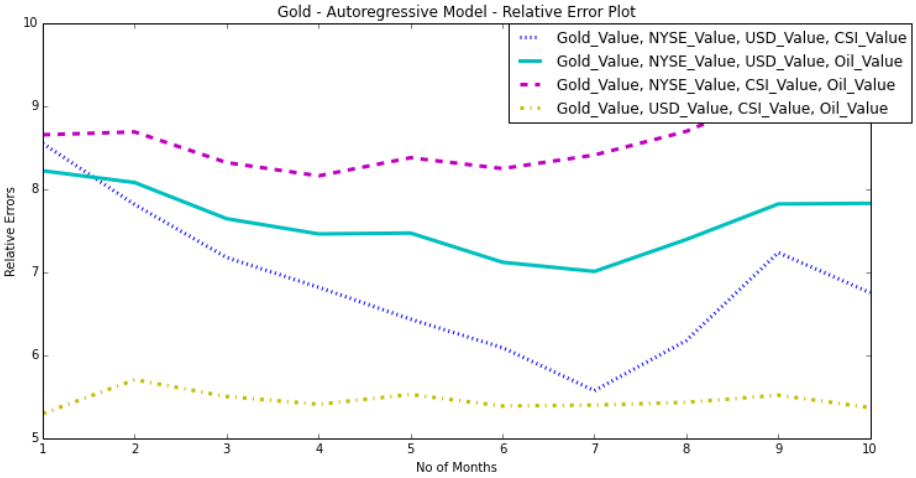

Fig. 17. Performance of the autoregressive model considering 3 factors with increasing number of months

Figure 15, Figure 16, and Figure 17 show the error plots for different models.

ARMA Model

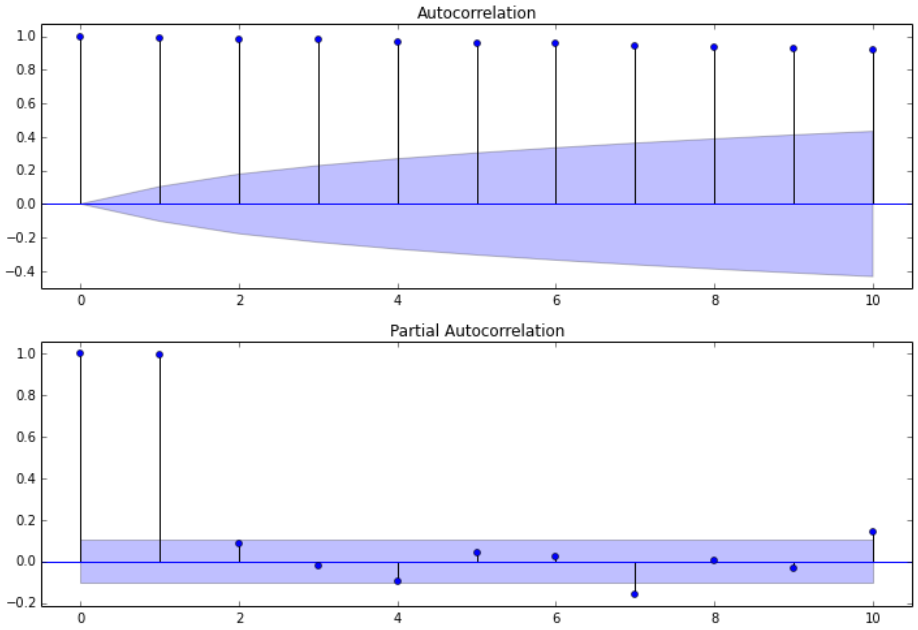

Fig. 18. ARMA Autocorrelation and Partial Autocorrelation factors

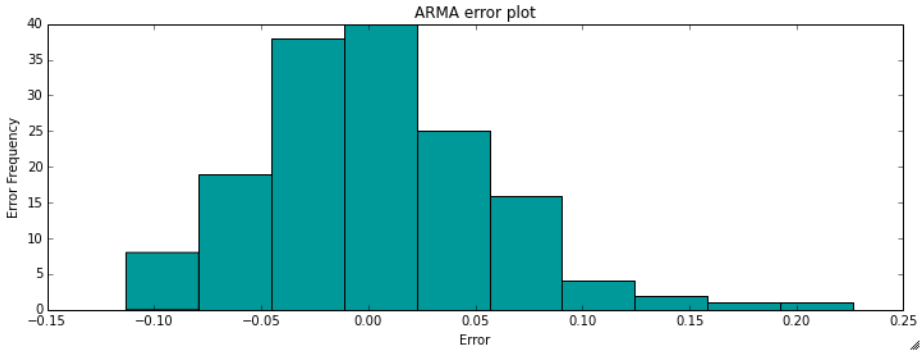

Fig. 19. ARMA Error Distribution

Comparison Summary - Gold The following is a summary of the comparison between autoregressive and multiple linear regressive models against the baseline models.

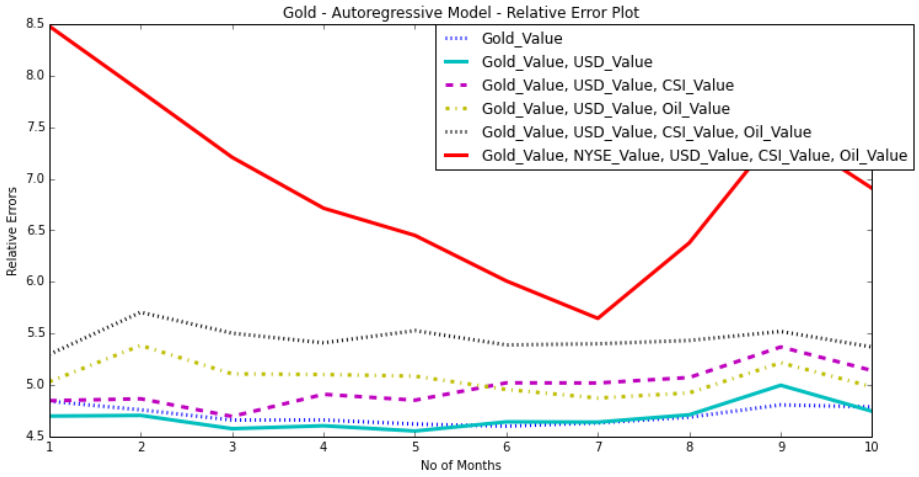

Fig. 20. Performance of the autoregressive model considering 3 factors with increasing number of days

Figure 20 shows a comparison of the best performing models.

Factors	Relative Error	Mean Abs Error	RMSE
$Model0.0$	3.674560	29.142500	49.049942
$Model0.1$	3.606751	28.856644	47.804751
$Autoregressive$	8.49	5.384215	6.672619
$USD\ Index$	8.69	5.097663	6.67896
$S\&P500$, $Gold\ Price$	9.04	5.160815	6.737639
$NYSE$, USD	9.05	5.241797	6.767898
$S\&P500$, USD, $Gold\ Price$	9.38	5.26046	6.79724
$S\&P500$, $NYSE$, USD, $Gold\ Price$	10.86	5.494503	7.000936
$ARMA\ (2,0)$	5.0322462866	6.42218763098	6.78951497738

Table 6. Summary of performance of various models for predicting gold prices

5 Current Prediction and Next Steps

5.1 Current Prediction

The predicted price of WTI Crude Oil on Dec 1 as of Nov 1 is
00.00 USD per Barrel

The predicted price of Gold on Dec 1 as of Nov 1
0000.00 USD per ounce

5.2 Next Steps

S&P 1200 Global: `http://en.wikipedia.org/wiki/S%26P_Global_1200`

[FUTURES]

[ADD "Present what you will do next to get a complete predictive model. Discuss any difficulties you will have to overcome in building a good model"]

Acknowledgments. Here acknowledge any other people who helped with this project.

6 Bibliography

For citations in the text please use square brackets and consecutive numbers: [**?**], [**?**], [**?**] – provided automatically by LaTeX's `\cite ... \bibitem` mechanism.

Please base your references on the examples below. a book [1]

References

1. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
2. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`