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 the 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

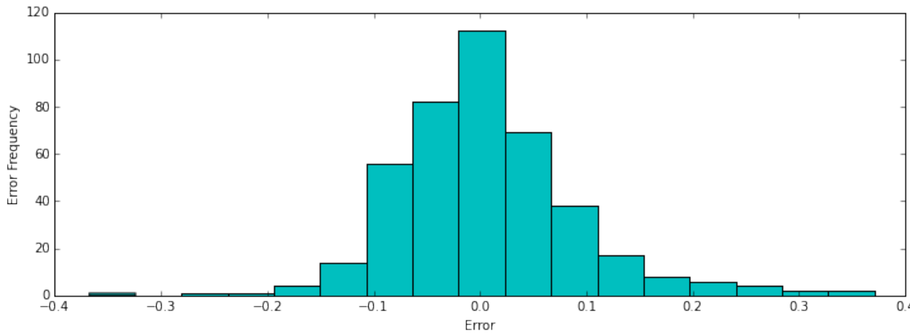

Fig. 1. Error plot for baseline model - oil/gold price is the same as the previous month's price

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

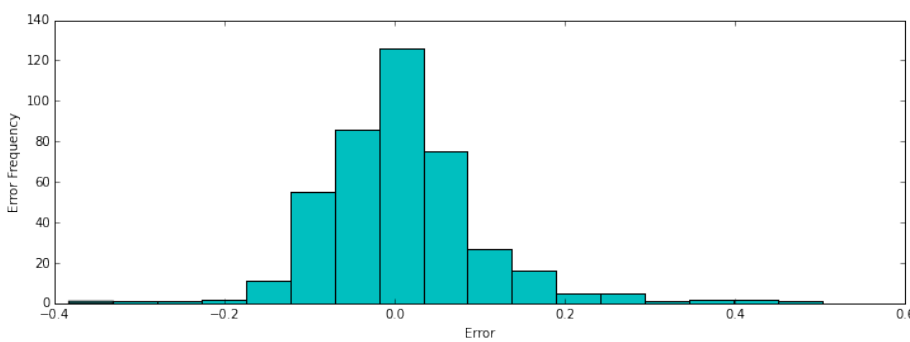

Fig. 2. Error plot for baseline model - oil/gold price is a weighted mean of price of previous k months

1.3 Spot price prediction using futures prices

A futures contract is an agreement between two parties to buy or sell a specific asset at a time in the future for a price decided today [1]. 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. If there are more sellers than buyers, which is indicative of a reduction in demand of a particular asset, the futures prices of that asset go down. However, more number of buyers than sellers indicates that the supply is unable to meed the demands, which results in an increase in the price of that asset.

Most futures contracts do not lead to delivery because traders choose to close out their positions prior to delivery, however it is this factor that ties the futures price to the spot prices.

Commodities such as oil and gold are also traded on the futures market. The are are two types of assets - investment assets and consumption assets. Gold is considered as an investment asset and has a net income associated with it. Oil is a consumption asset and has storage costs associated.

The value of the futures contracts and the futures prices of oil and gold hold some information in them which can help us determine the price of these commodities.

2 Data Matrices

Our data consists of multiple time series of monthly Oil and Gold prices[2], and the following macroeconomic factors:

- S&P 500 Index [2]
- New York Stock Exchange Index (NYSE) [2]
- US Dollar Index [2]
- Consumer Sentiment Index (CSI) [3]

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

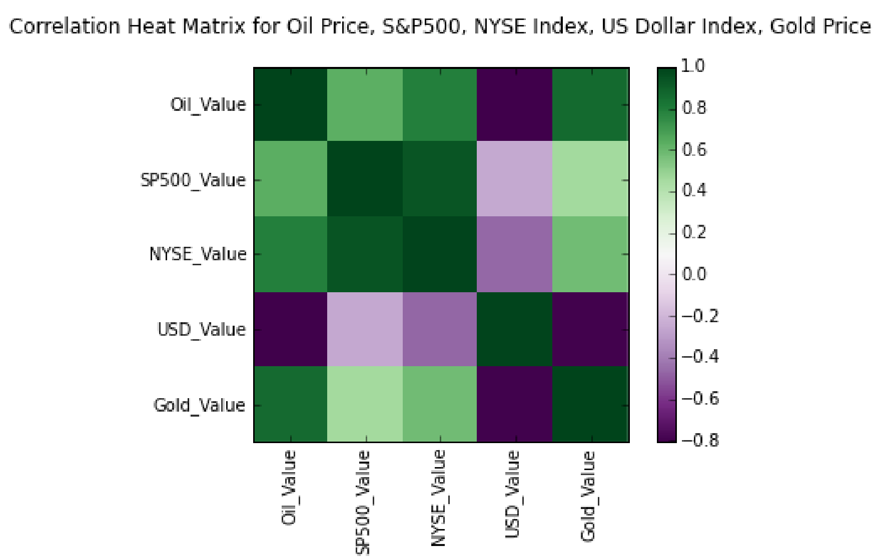

Fig. 4. Correlation Heat Map for Oil Price and related macroeconomic factors

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

Table 3. Correlation Matrix - Oil

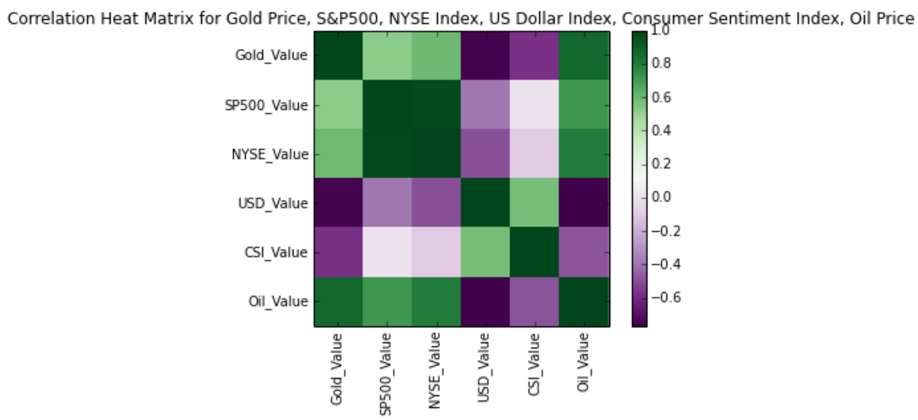

Fig. 5. Correlation Heat Map for Oil Price and related macroeconomic factors

	$S\&P500$	$NYSE$	USD	CSI	$OilPrice$
$GoldPrice$	0.526883	0.594711	−0.751085	−0.574481	0.87159

Table 4. Correlation Matrix - Gold

The correlation heat map in Figure 4 and Figure 5, and the corresponding tables containing correlation coefficients show the correlation between the price of oil/gold 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, and a high correlation between Gold Price and certain macroeconomic factors - US Dollar Index, NYSE and Oil Price.

Our goal is to predict the prices of oil and gold on January 1st, 2015 as of December 1st, 2014, which is a month in advance. Therefore, we have used monthly data, i.e, oil/gold price and values of other economic factors on the last day of every month. But this severely limits the amount of data we can obtain. Although we collected data for the past 30 years, we have just 360 entries in our time series, which isn't a good enough number to generate good enough error metrics. This also limits our prediction power.

3 Development/Evaluation Environment

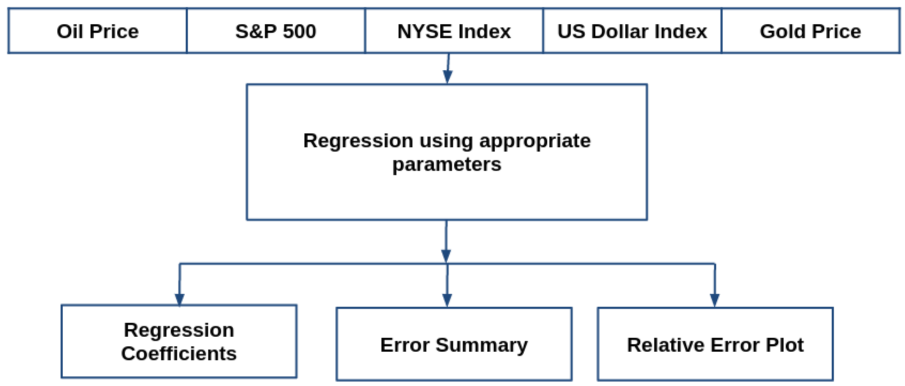

Fig. 6. The development and evaluation environment for the oil/gold price prediction with economic factors involved

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 a data frame containing multiple time series of oil/gold prices and related macroeconomic factors as an input. A parameter corresponding to each of these factors is also passed as a part of the input to the model. This parameter is an 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, our time series 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.

The Mean Relative Error, Mean Absolute Error, and Root Mean Square Error are calculated, and a histogram of the relative error frequencies is 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 Ordinary 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 with a better 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
- 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.

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.

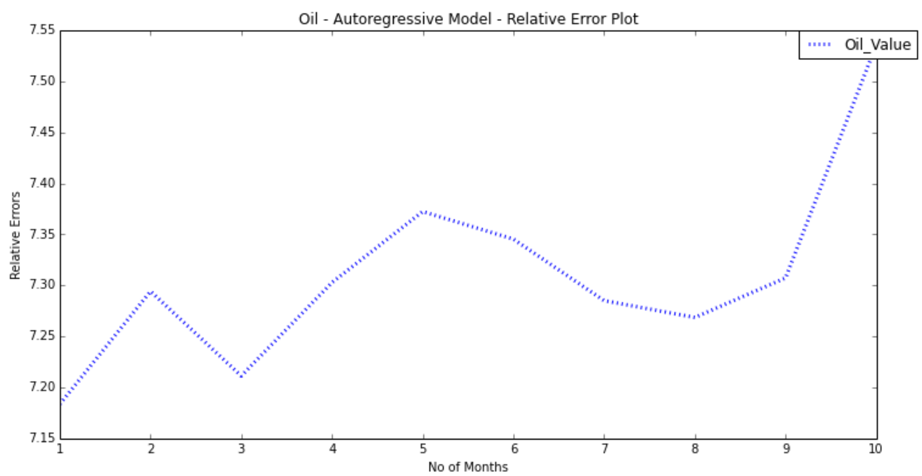

Fig. 7. Performance of the autoregressive model with increasing number of months

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 predicting oil prices.

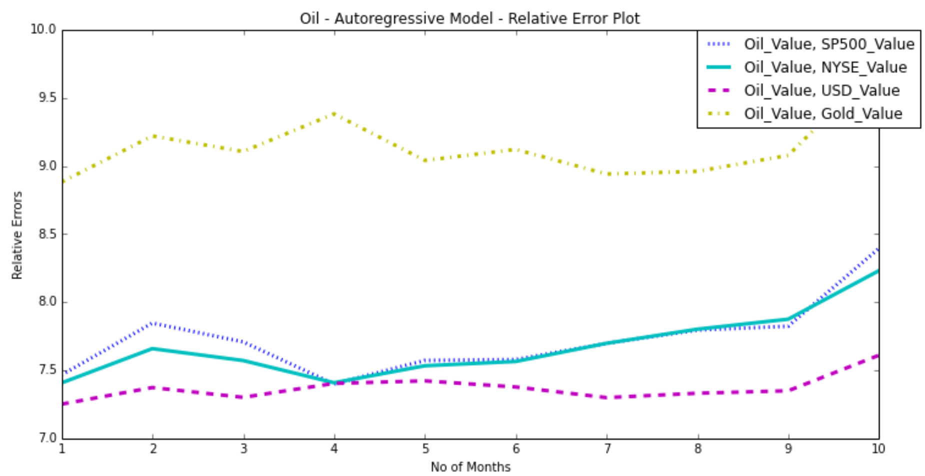

Fig. 8. Performance of the autoregressive model considering one factor with increasing number of months

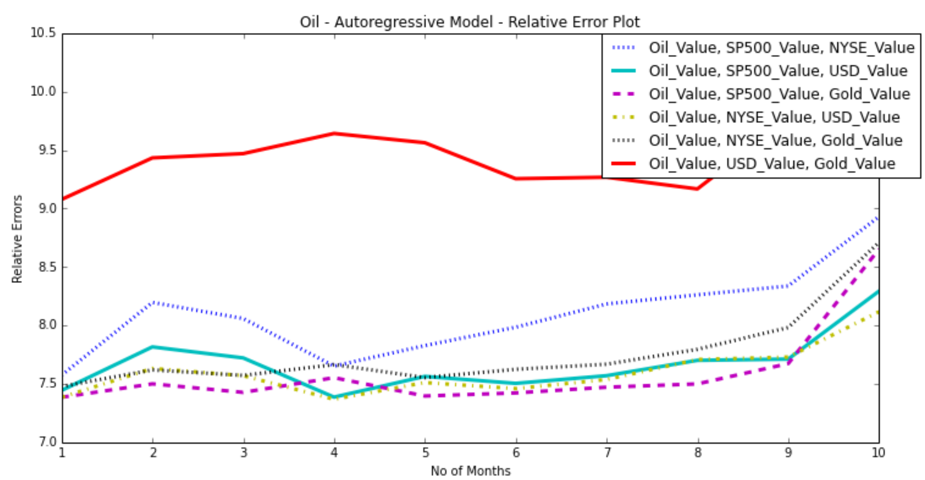

Fig. 9. Performance of the autoregressive model considering two factors with increasing number of months

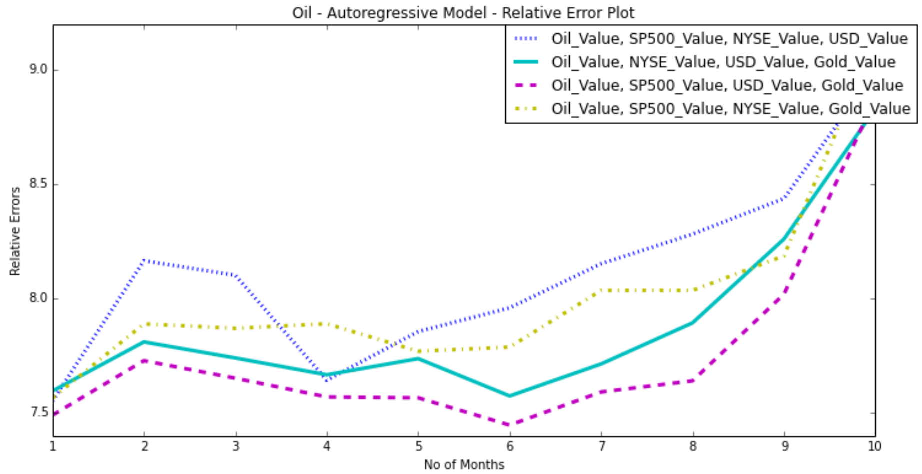

Fig. 10. Performance of the autoregressive model considering 3 factors with increasing number of months

ARMA Model .

We also developed an ARMA(p,q) model to predict the price of oil. After testing the ARMA model with various values of p and q, we observed that the ARMA (6,0) model was giving us the best results for predicting the oil price.

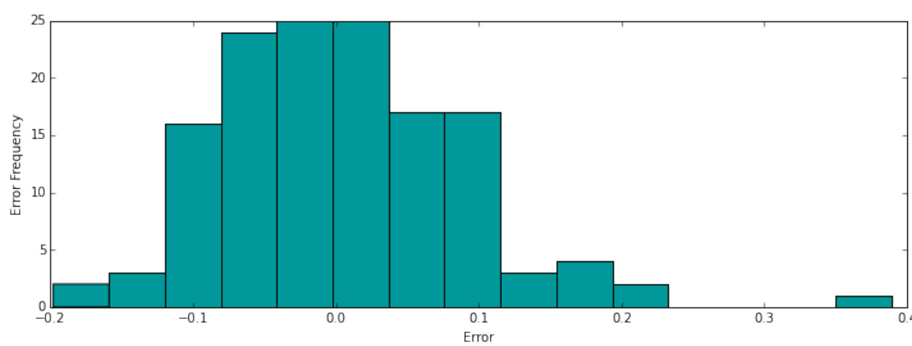

Fig. 11. ARMA Error Distribution

Figure 11 shows a typical relative error distribution of the ARMA model.

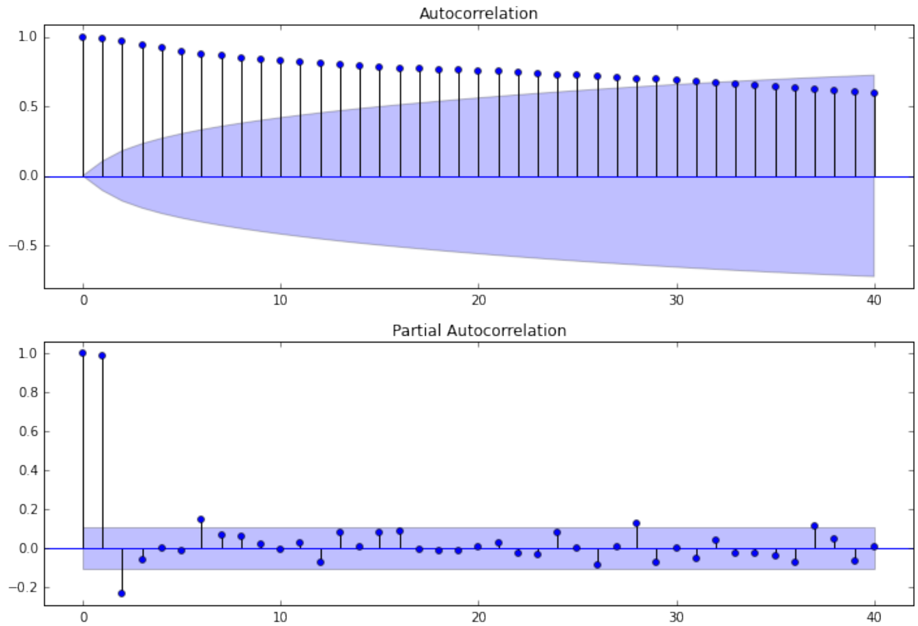

Fig. 12. ARMA Autocorrelation and Partial Autocorrelation factors

Figure 12 shows a plot of the autocorrelation and the partial autocorrelation factors between the prices of oil over the past 40 days.

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

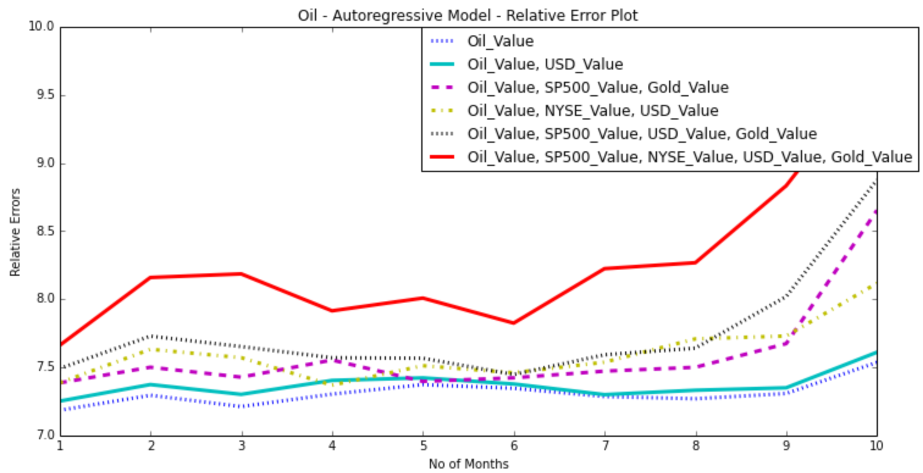

Fig. 13. Best performing models for predicting oil prices

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

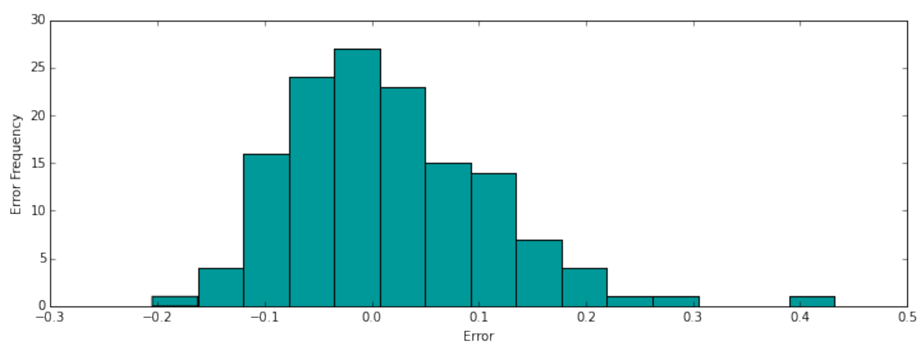

Fig. 14. Typical error distribution of an autoregressive and multiple linear regressive model

Figure 14 shows a typical error distribution of an autoregressive and multiple linear regressive model.

Factors	Relative Error	Mean Abs Error	RMSE
$BaselineModel0.0$	6.73746	5.384215	7.101763
$BaselineModel0.1$	7.27174	5.761054	7.754635
AR	7.211242952	5.097662905	6.672619397
$USD\ Index$	7.30095138	5.137570958	6.67896
$S\&P$ 500, $Gold\ Price$	7.427236625	5.160815	6.737639
$NYSE$, USD	7.570348752	5.241797	6.767898
$S\&P$ 500, USD, $Gold\ Price$	7.651851728	5.26046	6.79724
$S\&P$ 500, $NYSE$, USD, $Gold\ Price$	8.185639121	5.494503	7.000936
$ARMA$ $(6,0)$	5.0322462866	6.42218763098	6.78951497738

Table 5. Summary of error metrics of various models for predicting oil prices

Table 5 summarizes the performance of various predictive models. It includes - the baseline models (Baseline Model 0.0,0.1), pure autoregressive models (AR and ARMA), and multiple linear regressive models taking into consideration various macroeconomic factors.

- The Autoregressive model, with a relative error of 7.21%, performs slightly better as compared to the baseline model considering oil prices of the past three days.

- ARMA (6,0) is the best performing model among the autoregressive models, with a relative error of about 5.03%.

- The model considering only one factor - US Dollar Index performs the best amongst multiple linear regressive models, with a relative error of about 7.30%. This model performs slightly worse compared to the pure autoregressive model and the baseline model.

- ARMA(6, 0) performs better than multiple linear regressive model, and also performs significantly better than the baseline model 0.0 having a relative error of about 7.27%.

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 15 shows the performance of the autoregressive model as the number of months taken into consideration by the model is increased.

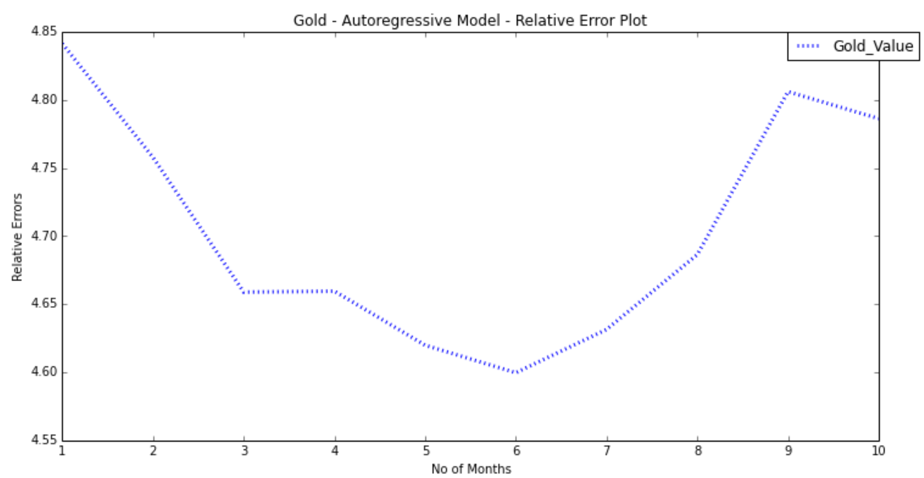

Fig. 15. Performance of the autoregressive model with increasing number of months

Figure 16, Figure 17, and Figure 18 show the error plots for different models predicting gold prices.

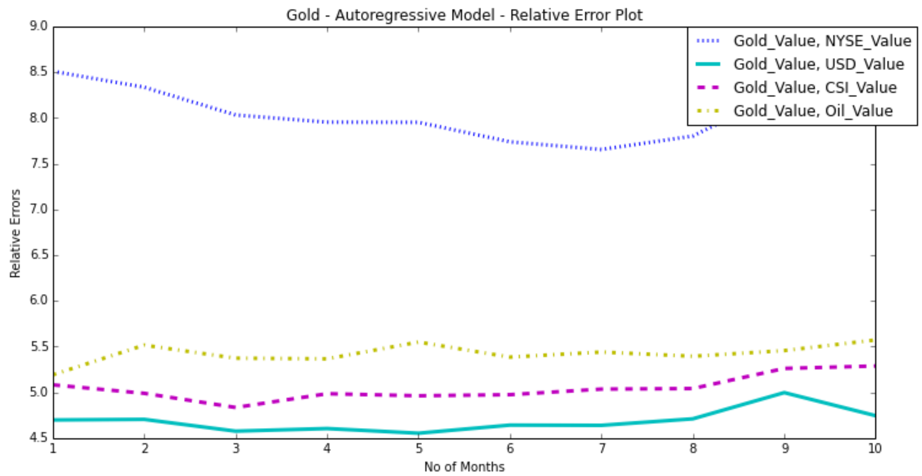

Fig. 16. Performance of the autoregressive model considering one factor with increasing number of months

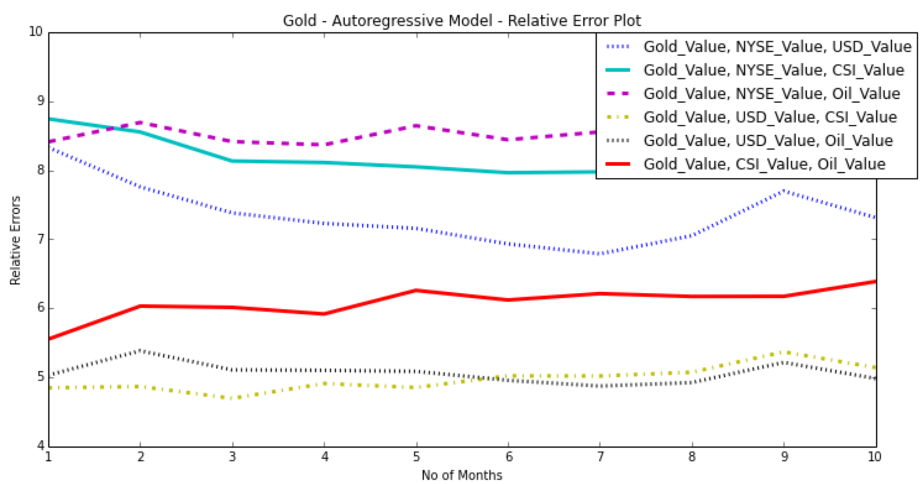

Fig. 17. Performance of the autoregressive model considering two factors with increasing number of months

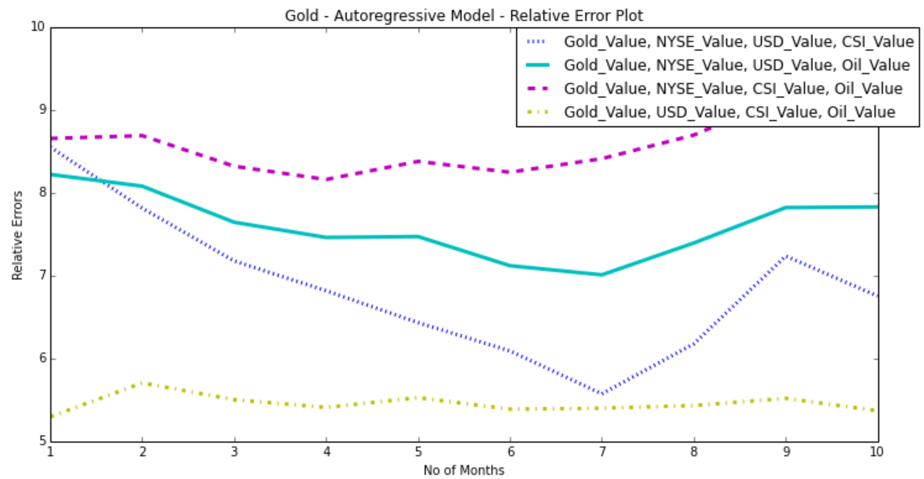

Fig. 18. Performance of the autoregressive model considering three factors with increasing number of months

ARMA Model .

Similar to oil, the ARMA model was also used to predict gold prices. Here, we observed that ARMA (2,0) was giving us the best results.

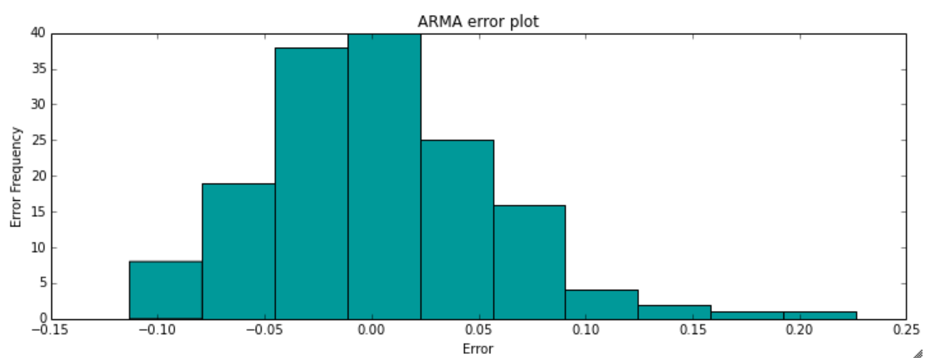

Fig. 19. ARMA Error Distribution

Figure 19 shows a typical relative error distribution of the ARMA model.

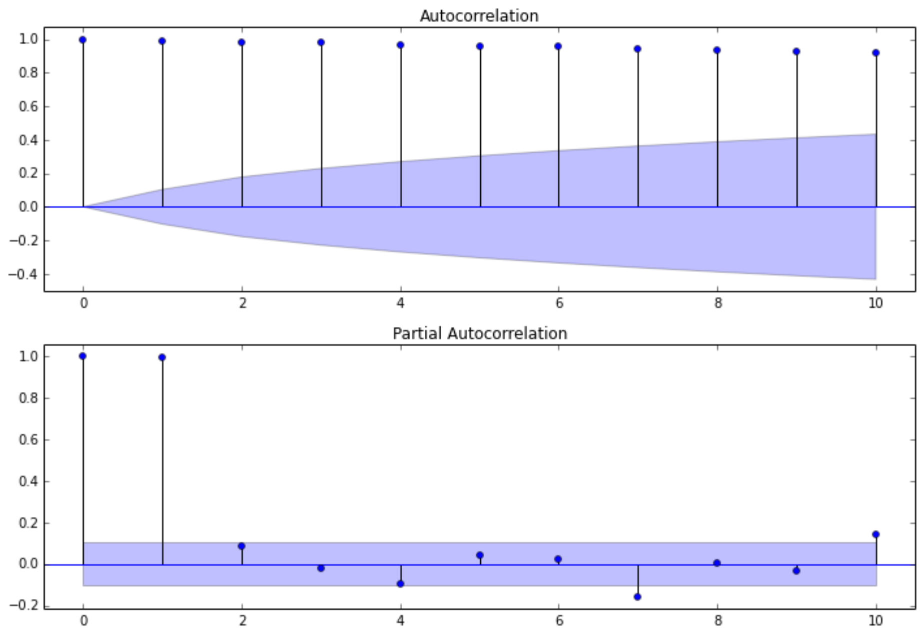

Fig. 20. ARMA Autocorrelation and Partial Autocorrelation factors

Figure 20 shows a plot of the autocorrelation and the partial autocorrelation factors between the prices of oil over the past 10 days.

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

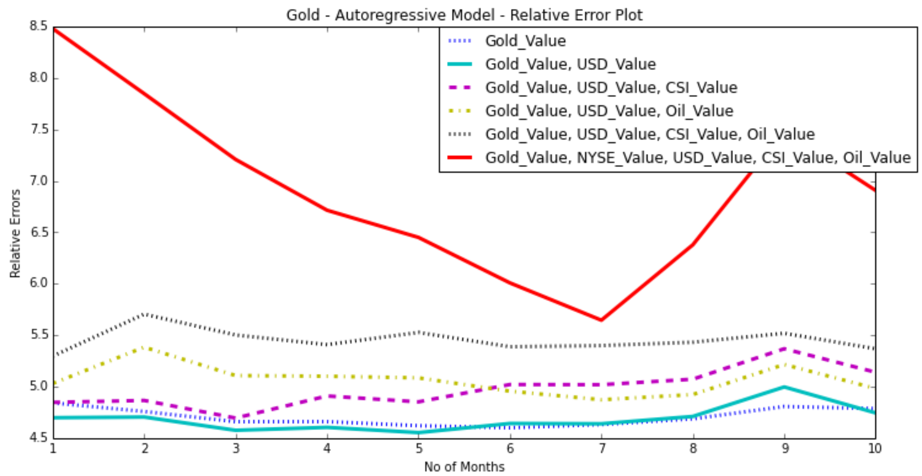

Fig. 21. Best performing models for predicting oil prices

Figure 21 shows a comparison of the best performing models.

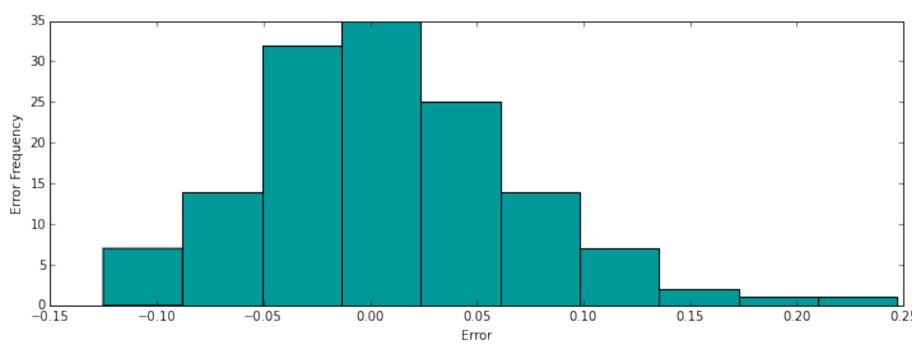

Fig. 22. Typical error distribution of an autoregressive and multiple linear regressive model

Figure 22 shows a typical error distribution of an autoregressive and multiple linear regressive model.

Factors	Relative Error	Mean Abs Error	RMSE
Baseline Model 0.0	4.43662163812	42.94499881	62.29470885
Baseline Model 0.1	4.317908747	42.46524296	60.68314337
Autoregressive	4.65886115765	44.86082675	62.74391971
USD Index	4.57491310135	44.3165759	62.01865518
USD Index, CSI	4.69383444887	44.65872598	62.12632004
USD Index, Oil Price	5.1068833463	46.8515039	62.64738884
USD Index, CSI, Oil Price	5.50132391575	63.70140951	64.68856
NYSE, USD Index, CSI, Oil Price	7.20780269939	71.68320365	67.96092
ARMA (2,0)	4.24125955354	40.5505169891	58.9195927119

Table 6. Summary of error metrics of various models predicting gold prices

Table 6 summarizes the performance of various predictive models. It includes - the baseline models (Baseline Model 0.0,0.1), pure autoregressive models (AR and ARMA), and multiple linear regressive models taking into consideration various macroeconomic factors.

- ARMA (2,0) is the best performing model among the autoregressive models, with a relative error of about 4.24%.

- The model considering only one factor - US Dollar Index performs the best amongst multiple linear regressive models, with a relative error of about 4.58%. This model performs almost slightly better than the pure autoregressive model, but slightly worse than the baseline model.

- ARMA(2, 0), with a relative error of 4.24% performs better than multiple linear regressive model, and also better than the baseline model 0.1, with a relative error of about 4.32%.

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
74.17 USD per Barrel

The predicted price of Gold on Dec 1 as of Nov 1
1165.16 USD per ounce

5.2 Next Steps

- **Better Models** Since the ARMA model is the best performing model, we need to try an incorporate the various macroeconomic factors as a part of the ARMA model.
- **Futures** We need to understand how the futures prices can give us more information about the price of a commodity, and use them to make our predictions.
- **Recent Trends** We need to expand our model to take into account the recent trends, and evaluate whether it can give us better results.

Acknowledgments. Dr. Steven Skiena.

6 Bibliography

References

1. Hull, John. Options, futures and other derivatives. Pearson education, 2009.
2. Quandl - Find, Use and Share Numerical Data. https://www.quandl.com
3. http://future.aae.wisc.edu/data/monthly_values/by_area/998?grid=true