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 prediction, we compare our current autoregressive and multiple linear regressive models to the following baseline models:

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

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

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

[ADD: PLOT]

1.3 Autoregressive and Multiple Linear Regressive Models?

1.4 Spot price prediction using Futures prices

[Futures???]

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

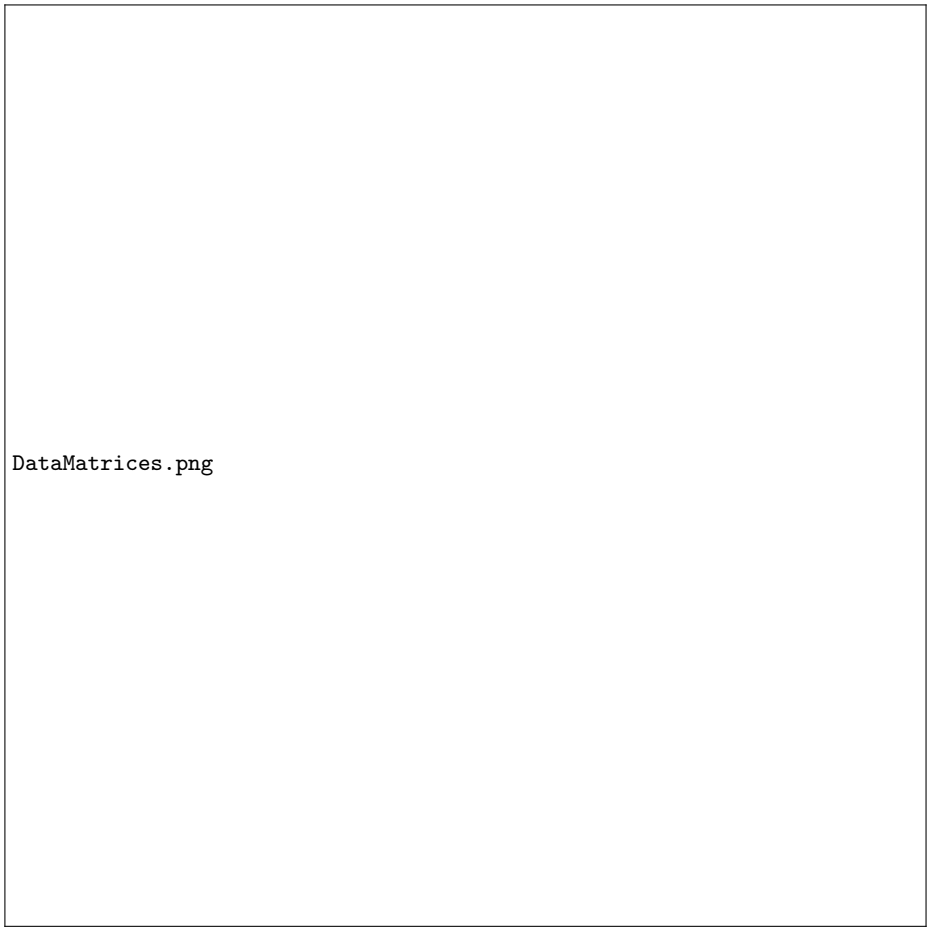

Fig. 1. Data frame for oil and gold price and its related economic factors

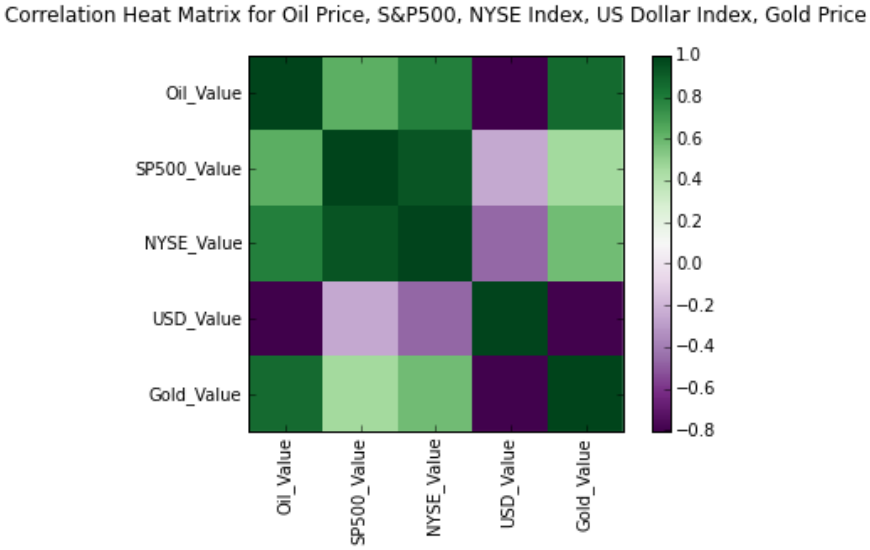

Fig. 2. Correlation Heat Map for Oil Price and related macroeconomic factors

Correlation Matrix					
	$OilPrice$	$S\&P500$	$NYSE$	$USDIndex$	$GoldPrice$
$OilPrice$	1	0.718178	0.805763	−0.77365	0.872

The correlation heat map in Figure 2, 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.

Correlation Matrix							
	$GoldPrice$	$S\&P500$	$NYSE$	USD	$EUR-USD$	CSI	$OilPrice$
$GoldPrice$	1	0.461485	0.578967	−0.792106	0.628074	−0.727766	0.860482

The correlation heat map in Figure 3, 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

For both the daily and monthly data, they are obtained from
Oil Price: `https://www.quandl.com`
Gold Price: `https://www.quandl.com`

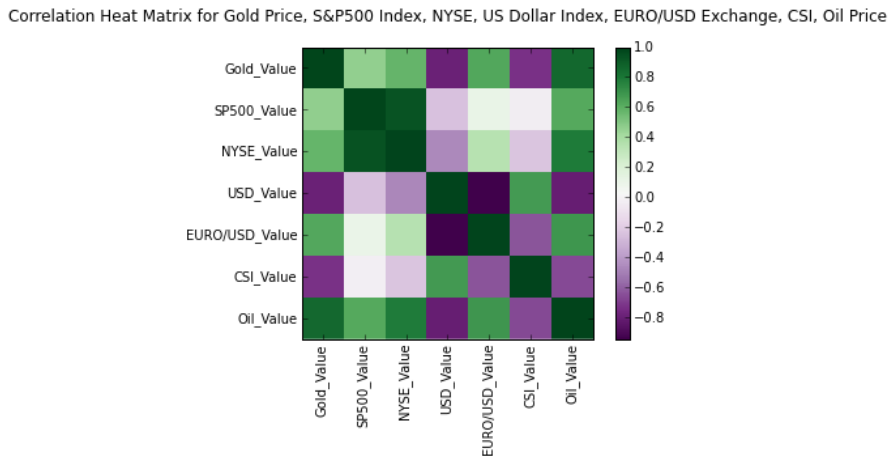

Fig. 3. Heat Map and Correlation Coefficient Matrix for Gold price and its related macroeconomic factors

SP500: `https://www.quandl.com`
NYSE: `https://www.quandl.com`
USD: `https://www.quandl.com`
EURO/USD: `https://www.quandl.com`
CSI: `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

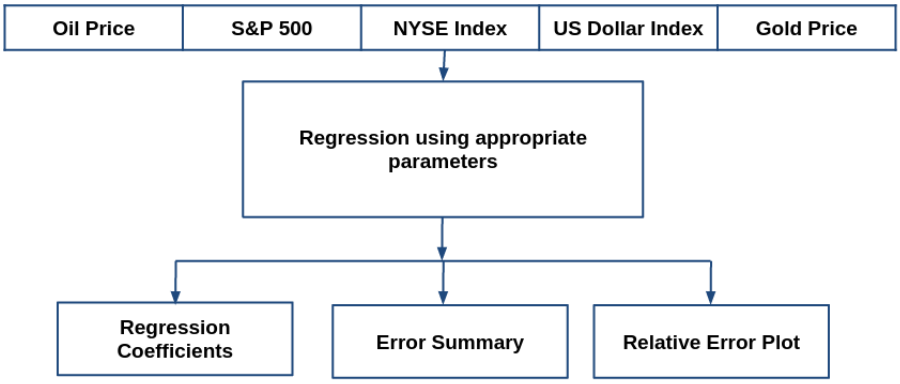

Fig. 4. The development and evaluation environment for the oil price prediction with economic factors involved

Figure 4 shows the work flow of our prediction model. The model uses past prices of oil/gold, combined with certain macroeconomic factors, and performs regression to make predictions. The accuracy of the model is then determined on the basis of error metrics.

A data frame containing multiple time series of oil/gold prices and related macroeconomic factors is generated. A parameter corresponding to each of these factors is also passed as a part of the input to the model. This parameter is a number which represents the number of lags in the autoregressive factor, that 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%. The model tries generate a linear function while assigning coefficients for each of these economic factors. These regression coefficients are then used to make predictions for the next month. To make sure that we are not over-fitting the curve, we also cross validated our data using a size k slice of from training set(k is nearly equal to 80%). We then slided this slice accross the training set and finally took the average of the different predictions of each slice.

Error Summary. 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.

[ADD]

4 Current Model and Baseline

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

[ADD]

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' days.

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

The Odinary Least Squares (OLS) method is used to estimate the parameters in the regression model. 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 of two variables - a dependent variable and an explanatory variable using a linear function. The process of modeling a variable based on more than one explanatory variables is called Multiple Linear Regression.

An initial model is developed which is 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.

[ADD]

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

Performance of Autoregressive Models against Baseline Models A comparison.

[CHANGE]

Module Name	Mean Relative Error	Mean Absolute Error	RMSE	
Baseline Model				
Oil price same as last day price	6.527614	2.718182	4.288619	
Weighted mean of the last 3 days Oil price	6.985914	2.91937	4.826848	
Autoregressive Models				
Oil price of last 1 day	1.622386201	1.308279448	1.850814385	
Oil prices of past 3 days	1.624936069	1.309915	1.848537056	
Autoregressive Moving Average (ARMA) Models				
ARMA (2,0)	**1.701360053**	**0.8342170483**	**1.304953411**	PERFORMS SLIGHTLY BETTER
ARMA (3,1)	**1.703171384**	**0.8355328195**	**1.303745335**	

Fig. 5. Current Autoregressive Models against base

Comparison Summary
Autoregressive Models "ARMA(2,0)" and "ARMA(3,1)" have lower errors than the baseline models. ARMA(2,0) has slightly lower Mean Relative Error and Mean Absolute Error than ARMA(3,1), but a slightly higher RMSE than ARMA(3,1). Multiple Linear Regression Model "Oil Price, S&P 500, NYSE and USD Index of past 3 days" has the lowest RMSE. It has lower errors than the baseline models.

[EDIT]

Module Name	Mean Relative Error	Mean Absolute Error	RMSE
Baseline Model			
Oil price same as last day price	6.527614	2.718182	4.288619
Weighted mean of the last 3 days Oil price	6.985914	2.91937	4.826848
Factors being considered: S&P 500, NYSE, US Dollar Index, Gold Price			
Multiple Linear Regression Model with 1 factor			
Oil Price and S&P 500 Index of past 3 days	1.624450329	1.309749966	1.845501306
Oil Price and NYSE Index of past 3 days	1.624267213	1.309533504	1.843488943
Oil Price and USD Index of past 3 days	1.624688158	1.309750162	1.848171195
Oil Price and Gold Price of past 3 days	1.631910352	1.315978511	1.8558882
Multiple Linear Regression Model with 2 factors			
Oil Price, S&P500, NYSE of past 3 days	1.625388806	1.310179341	1.842075871
Oil Price, S&P 500 and USD Index of past 3 days	1.623829806	1.30932709	1.844316414
Oil, S&P 500 and Gold Price of past 3 days	1.639090652	1.322648417	1.861609629
Oil Price, NYSE and USD Index of past 3 days	1.62386608	1.309186423	1.841737277
Multiple Linear Regression Model with 3 factors			
Oil Price, S&P 500, NYSE and USD Index of past 3 days	**1.625237208**	**1.310016037**	**1.84083677**
Multiple Linear Model with all 4 other factors			
Oil Price, USD, S&P 500, NYSE and Gold Price of past 3 days	1.63182826	1.315517678	1.847603297

Fig. 6. Current Multiple Linear Regression Models against base

4.2 GOLD

Performance of Autoregressive Models against Baseline Models A comparison.

Current Multiple Linear Regression Models Note: After generating the correlation coefficient heat map [Fig 4.], we only concerning the combination of Gold Price, USD, EURO/USD, CSI, Oil Price for predicting gold price.

Performance against Baseline Models

[CHANGE]

Module Name	Mean Relative Error	Mean Absolute Error	RMSE	Best one in All
Baseline Model				
Gold price same as last day price	4.154336 %	22.50364	39.667392	
Weighted mean of the last 3 days gold price	4.216399 %	22.719552	38.998358	
Auto Regressive Models				
Gold price of last 1 day	0.909893675973 %	10.62531691	15.6143533	
Gold prices of past 3 days	0.911377728085 %	10.63567494	15.61209638	
Autoregressive Moving Average(ARMA) Models				
ARMA(2,0)	0.896845911091 %	9.940274096	14.82209422	
ARMA(3,1)	0.696338469176 %	5.371189697	9.988917016	SIGNIFICANTLY BETTER
ARMA(2,2)	0.698170840577 %	5.377845349	9.989045436	SIGNIFICANTLY BETTER

Fig. 7. Current Autoregressive Models against base

Module Name	Mean Relative Error	Mean Absolute Error	RMSE	Best one in All
Baseline Model				
Gold price same as last day price	4.154336 %	22.50364	39.667392	
Weighted mean of the last 3 days gold price	4.216399 %	22.719552	38.998358	
Multiple Linear Model with 1 other factor				
Gold Price and S&P 500 Index of past 3 days	0.910989470538 %	10.63411189	15.61509994	
Gold Price and NYSE Index of past 3 days	0.908913469643 %	10.61488888	15.59285781	
Gold Price and USD Index of past 3 days	0.910740053257 %	10.62421809	15.60111514	
Gold Price and Euro/USD Exchange Rate of past 3 days	0.913922103832 %	10.65708401	15.62858645	
Gold Price and CSI of past 3 days	0.911209607511 %	10.63316055	15.61160527	
Gold Price and Oil Price of past 3 days	**0.906189820768 %**	**10.58470852**	**15.51158146**	**Lowest Mean Relative Err and Mean Absolute Err**
After generating the correlation coefficient heat map, we only concerning combination of USD, EURO/USD, CSI, Oil				
Multiple Linear Model with 2 other factors				
Gold Price, USD and EURO/USD of past 3 days	0.909034610537 %	10.6077233	15.58611912	
Gold Price, USD and CSI of past 3 days	0.910204441426 %	10.6186964	15.59982143	
Gold Price, USD and Oil Price of past 3 days	0.907577054439 %	10.59470598	15.50859888	
Gold Price, EURO/USD and CSI of past 3 days	0.913284526674 %	10.64898094	15.62476428	
Gold Price, EURO/USD and Oil Price of past 3 days	0.908706241253 %	10.60923026	15.52201089	
Gold Price, CSI and Oil Price of past 3 days	0.906784413338 %	10.59035719	15.5167956	
Multiple Linear Model with 3 other factors				
Gold Price, USD, EURO/USD and CSI of past 3 days	0.908442364248 %	10.60134702	15.58003276	
Gold Price, USD, EURO/USD and Oil Price of past 3 days	**0.907727253341 %**	**10.59562283**	**15.50604703**	**Lowest RMSE**
Gold Price, USD, CSI and Oil Price of past 3 days	0.908974041264 %	10.60838368	15.51978355	
Gold Price, EURO/USD, CSI and Oil Price of past 3 days	0.909469026042 %	10.61644669	15.5296531	
Multiple Linear Model with all 4 other factors				
Gold Price, USD, EURO/USD, CSI and Oil Price of past 3 days	0.907849241727 %	10.59608486	15.50836909	

Fig. 8. Current Autoregressive Models against base

[EDIT]

Comparison Summary

Autoregressive Models "ARMA(3,1)" and "ARMA(2,2)" have significantly lower errors than the baseline models. ARMA(3,1) has slightly lower errors than ARMA(2,2).

Multiple Linear Regression Model "Gold Price and Oil Price of past 3 days" has the lowest mean relative error and the lowest mean absolute error; "Gold Price, USD, EURO/USD and Oil Price of past 3 days" has the lowest RMSE. Both have significant lower errors than the baseline models.

4.3 Summary

[CHANGE FIG DESCRIPTION]

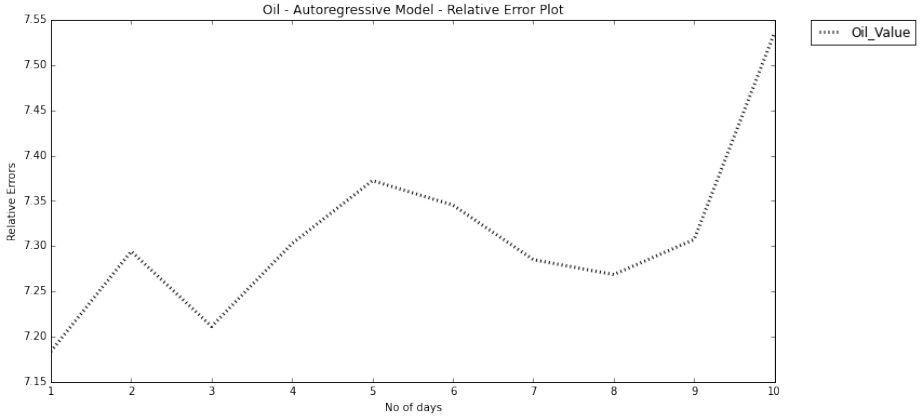

Fig. 9. Performance of the autoregressive model with increasing number of days

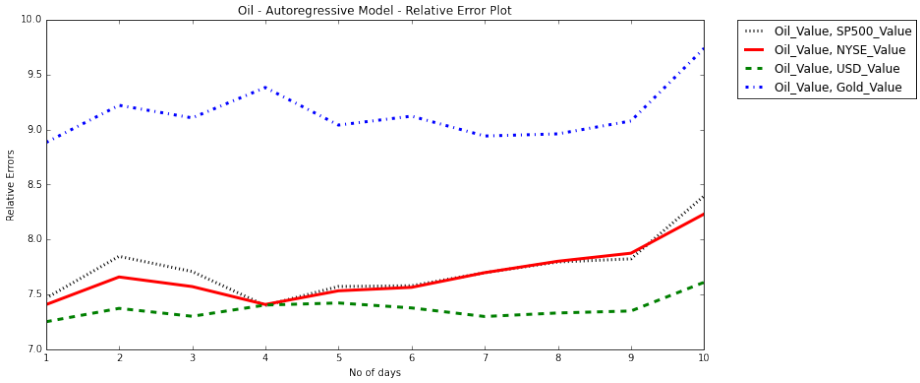

Fig. 10. Performance of the autoregressive model considering one factor with increasing number of days

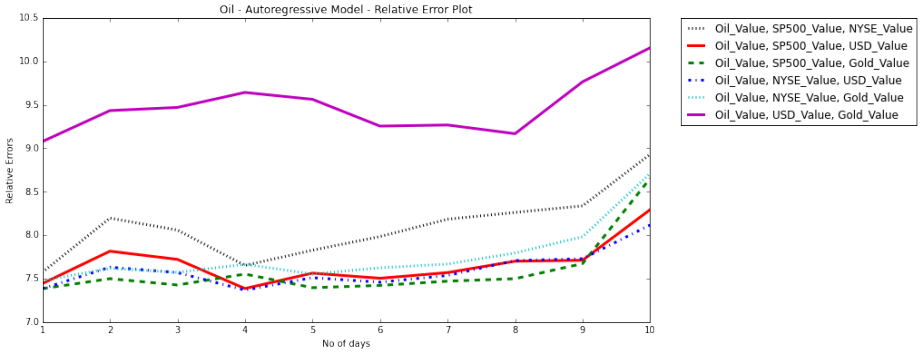

Fig. 11. Performance of the autoregressive model considering two factors with increasing number of days

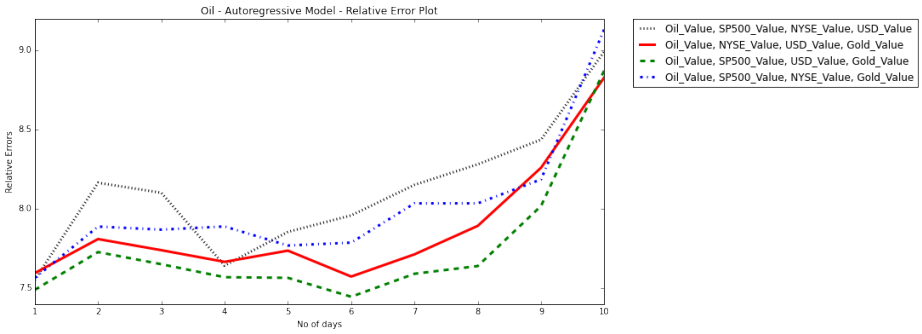

Fig. 12. Performance of the autoregressive model considering 3 factors with increasing number of days

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`

We will further investigate how we may include futures price as one of the economic factors to help predict oil and gold price.

[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

The correct BibTeX entries for the Lecture Notes in Computer Science volumes can be found at the following Website shortly after the publication of the book:
`http://www.informatik.uni-trier.de/~ley/db/journals/lncs.html`

For citations in the text please use square brackets and consecutive numbers: [1], [2], [4] – provided automatically by LaTeX's `\cite ... \bibitem` mechanism.

Please base your references on the examples below. The following section shows a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4] and [5], as well as a URL [6]. Please note that proceedings published in LNCS are not cited with their full titles, but with their acronyms!

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`