Final 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 Challenge

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

Oil and Gold are extensively traded commodities, and there are several macroeconomic factors which affect the prices, thus causing the prices of these commodities to be extremely volatile. Certain factors are also interdependent, which makes it extremely difficult to estimate the extent to which an individual factor could affect the price of a commodity. Moreover, the relationship between the factors and the prices could also vary over time. This makes the prediction of oil and gold prices a very complex and challenging problem.

1.1 Oil

Oil is a non-renewable resource which occurs in the earth. It is extracted and sent to refineries where different petroleum products such as gasoline, petrol, heating oil, etc are separated. Almost two-thirds of our energy demands are met by oil. Oil is the most heavily and actively traded commodity, which accounts to almost 10% of the world's trade.

The price of oil is determined by supply and demand. An increase in demand results in an increase in the price of oil, and vice versa. The supply and demand themselves are determined by various factors such as economy, weather and geopolitics.

1.2 Gold

Gold has been viewed as a symbol of wealth since the ancient times of the human history. About 50% of the gold produced is used in jewellery, 40% in investments, and the rest in the electronics industry, medicine, etc [1].

For a long period of time, the price of gold was fixed. After 1968, with the breakdown of the Bretton Woods Currency Arrangement, the price of gold began to be determined by the market. Gold is also used as an investment to hedge against market fluctuations.

2 History/Background

2.1 Factors Affecting the Price of Oil

Crude oil prices are determined by the balance between supply and demand. An increase (or decrease) in demand causes the price of oil to rise (or fall). Consequently, a cutback in the supply of oil results in an increase in oil prices. There are several factors which disturb the supply-demand balance, thus resulting in oil price fluctuations.

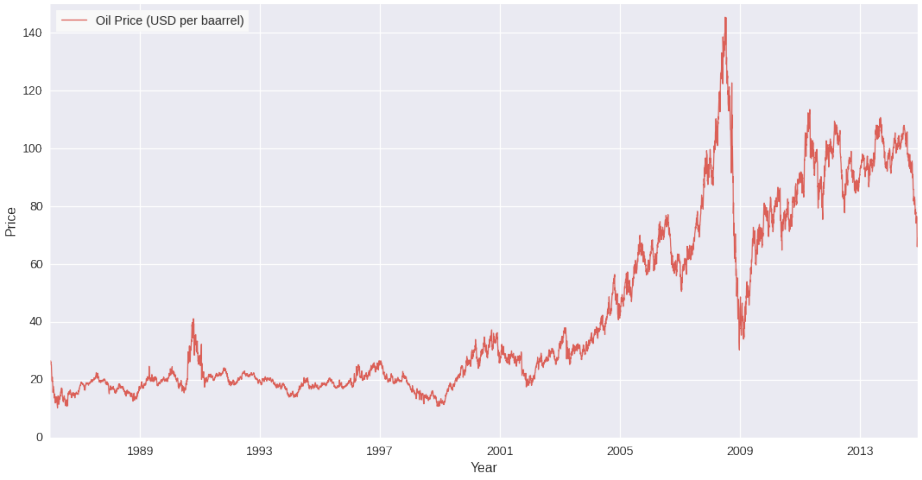

Fig. 1. Historical Oil Prices - US Dollar per Barrel - 1986 to 2014

Organization of the Petroleum Exporting Countries (OPEC) Supply OPEC is an organization of 12 oil exporting nations, namely Algeria, Angola, Ecuador, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela. It aims at coordinating and unifying petroleum prices of its member countries [2]. Oil supply from the OPEC member countries represents about 40% of the worlds crude oil, and their actions can affect the prices of oil to a significant extent. For example, limiting the oil production from OPECs major oil producers such as Saudi Arabia can influence crude oil supply and

affect the prices [3].
Oil prices not only depend on the current demand and supply, but also on the projected future supply and demand. OPEC adjusts the oil productions of its member countries based on current and future demand.

Non-OPEC Supply The non-OPEC countries produce about 60% of the world's crude oil. A reduction of supply from the non-OPEC countries creates additional pressure on the OPEC countries, which can also contribute to a rise in oil prices [3].

Stock Market The stock market can be used as an indicator of the economy. As economic conditions improve, there is an increase in demand for several commodities including oil, which results in an increase in oil prices. The Standard and Poor's 500 (S&P 500) index is a common benchmark for the stock market of USA. It is a weighted index of the market capitalization of 500 companies. It is the most commonly used indicator of the US economy.

Seasonal Effects Certain crude oil products such as heating oil and gasoline tend to have a seasonal variance. For example, there is an increase in demand for oil in the fourth quarter due to the cold weather, and a subsequent reduction in demand during late winter as the weather gets warmer. Gasoline prices also tend to rise in the summer due to an increased consumption.

US Dollar Oil is traded in dollars, and hence, any change in the strength of the dollar relative to other currencies can cause oil prices to shift. The US Dollar Index is a measure of the dollar relative to a basket of foreign currencies. Several studies support the negative correlation between the US Dollar Index and the price of oil [3] [4]. One reason is that the depreciation in the dollar exchange rate causes oil to be cheaper in countries outside the US, thus leading to an increase in demand, which in turn causes oil prices to rise. However, it has been observed that the relationship between oil prices and the dollar exchange rate has not been stable over the years.

2.2 Factors Affecting the Price of Gold

Gold is a precious metal which was used as a currency in several major civilizations in the past centuries. In the Unites States of America, gold had been at a fixed price of $20.67 per ounce, from the early 19th century until 1934 when President Franklin Roosevelt raised gold price to $35 per ounce. In 1968, with the breakdown of Bretton Woods Currency Arrangements, the gold price became market determined. From Figure 2 below, we observe that the two peaks of the gold price coincide with two significant economic recessions in our history - one in 1980, and the other 2008.

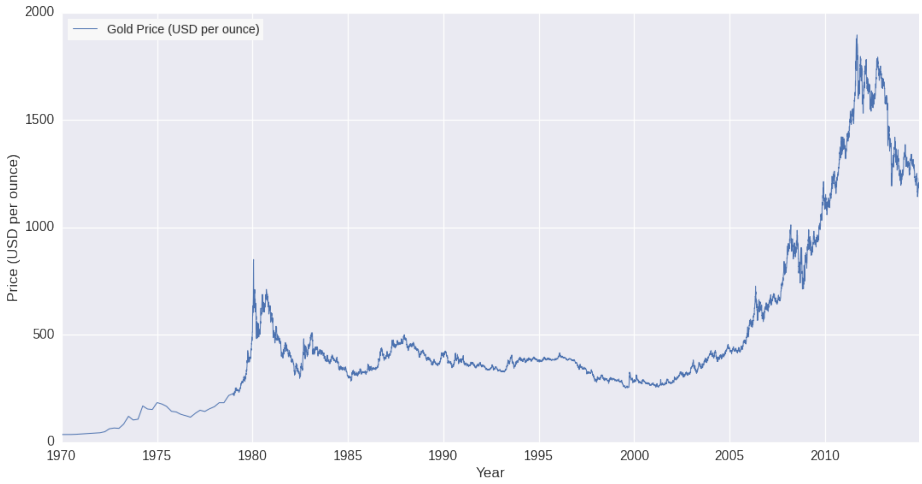

Fig. 2. Historical Gold Prices - US Dollars per Ounce - 1970 to 2014

The price of gold is determined by several factors, the most significant being, Crude Oil Price, Commodity Research Bureau Future Index (CRB), EURO/USD Exchange Rate (EURO/USD), Inflation Rate (INF), Money Supply (M1), and the US Dollar Index (USDX) [5][6][8].

Consumer Sentiment Index (CSI) The Consumer Sentiment Index or Consumer Confidence Index, is an index measuring the consumers' confidence over the market [11][12]. Similar to the effects of USDX over the gold price, a high CSI indicates consumers generally feel optimistic about the overall economy and their ability of obtaining and keeping their jobs, so they would be less likely to keep precious metals like gold.

Crude Oil Price Crude oil prices are highly correlated to gold prices. Rising oil prices may lead to an increase in the gold price, but the converse may not be true [5][6]. The gold-oil relation suggests that the crude oil price could partly account for inflation. An increase in the oil price results in increased prices of gasoline. Gasoline being more expensive results in an increase in the cost to transport goods, thus causing a possible hike in prices of goods. The final result is an increased price level, in other words, inflation. Gold tends to appreciate with inflation. Therefore, elevated oil prices can eventually lead to higher gold prices [9].

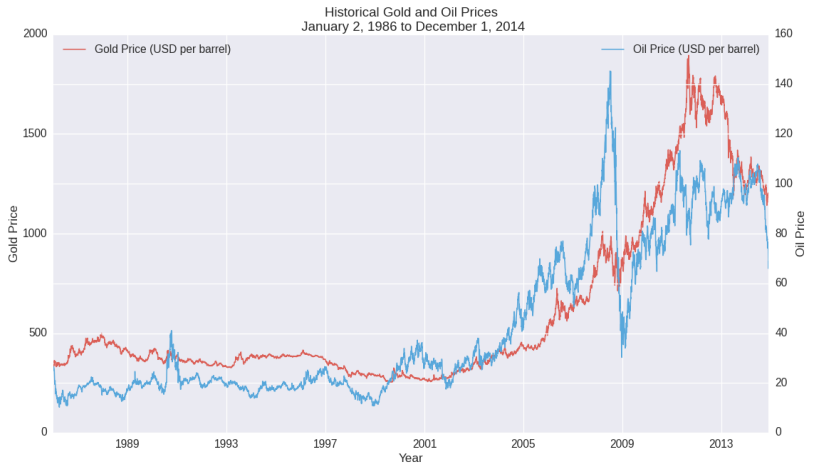

Fig. 3. Historical oil and gold prices showing correlation

2.3 Futures Contracts

A forward or a futures contract is an agreement to buy or sell an asset at a certain time in the future for a certain price. Two parties are involved in a futures contract, where one assumes a long position and agrees to buy the asset for a specific price at a particular time, and the other assumes a short position, and agrees to sell the same asset at the same price and time [25].

Futures contracts can be used to hedge market risks. Consider the current price of crude oil to be \$70 a barrel, and an individual speculates that the price is going to increase in the future, he can enter a long contract to purchase 100 barrels of crude oil at \$70 a barrel on a date 3 months from now. If the price of oil does increase to say \$75 a barrel, the individual benefits because he pays only \$70.

Futures contracts are also used by speculators to make profits. Suppose an individual enters a long contract to buy 100 barrels of oil at \$70 a barrel on April 1 2015. He purchases the asses worth $\$70,000$, but pays only a certain sum, called the *margin*, which is usually worth about $5\% - 15\%$ of the total value of the asset (say 10% here). In this situation, the individual pays \$700, hoping that he would be able to sell the contract later at a higher price. If the price of crude oil rises to \$75 a barrel on March 1 2015, he can sell the contract early, thus making a profit of $(75 - 70)100 = \$500$.

The prices of futures contracts are determined by the supply and demand [25]. If there are more sellers than buyers, the prices go down. If more traders wish to buy the commodity rather than sell, the prices goes up. As the delivery period for a futures contract is approached, the futures price converges to the spot price. For example, consider that the current price of a futures contract is higher than the spot price. In such a scenario, more individuals will be willing to go short and sell contracts, thus increasing the supply and causing the price to decrease and eventually converge to the spot price. The converse is also true when the price of the futures contract is lesser than the spot price. More individuals will be interested in buying contracts, this increasing the demand, and consequently the price of the contracts.

The prices of these contracts constantly change in response to newly obtained information in the market. Information can be obtained from various sources, for example, the US Energy Information Administration (EIA) provides weekly data of crude oil supply. This newly obtained information affects the prices of the underlying crude oil futures contracts.

These futures contracts encapsulate information about future trends, because they contains collective information of what several individuals think about the future market trends. Several experts claim that the spot prices hold as much information as the futures prices, and are equally good estimators of the future trends.

3 Literature Review

A significant amount of research has been done to understand and predict oil and gold prices. Various studies have developed different predictive models based on different techniques and factors. Some studies try to make predictions based on historical oil and gold prices. Others focus on the economic aspects and try to explain the correlations between the prices of oil and gold with respect to a variety of economic factors. Therefore, we evaluate and summarize some widely used models into the following two categories: Standard Time Series Models and Structural Models Considering Economic Factors.

3.1 Standard Time Series Models(Technical Models)

Standard time series models attempt to predict the oil price using the current and historical oil prices. The same strategy applies to predicting the gold price as well. These models are useful in the following scenarios:

- The prices show autocorrelation and autoregressive behavior, i.e., there is a pattern or a significant correlation between current and the previous prices.
- There are a large number of explanatory variables and it is difficult to understand them well because they interact with each other in a very complicated manner.
- Forecasting the dependent variable may require predicting the explanatory variables. And prediction of the explanatory variable might in turn be a harder problem.
- Not all explanatory factors and variables are known.

The most basic time series models that have been applied to model oil prices are the autoregressive models. In general, an autoregressive model **AR(p)** tries to model the current value of a time series based on the value of the last p instances in the time series.

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

The coefficients are regressed to predict the current price. Here the term ϵ_t is known as the *white noise*. It is a random variable with zero mean, constant variance. Also, $corr(\epsilon_t, \epsilon_{t-1})$ is 0 $\forall i > 1$ and it is 1 for $i = 0$. In other words, ϵ_t is a random variable with no auto-correlation.

Similarly, Autoregressive Moving Average (ARMA) models take in to account the moving average factor. They try to predict the randomness based on the

historical prices. Hence **ARMA(p,q)** can be written as:

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

Further, it is known that oil price changes(volatility) follow GARCH/ARCH properties. Least square models generally assume that the expected value of all error terms, when squared, is constant. This assumption is termed as homoskedasticity. Data(time series) in which this conditions fail to hold, are heteroskedastic. ARCH and GARCH models treat heteroskedasticity as a variance which is then modelled autoregressively.[13]

In short, GARCH models split the error-terms ε_t into a stochastic component z_t and a time dependent variance σ_t^2. Thus, $\varepsilon_t = z_t + \sigma_t$. The series σ_t^2 in **ARCH(q)** is modelled by:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$$

Similarly **GARCH(p,q)** is modelled by:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

GARCH and ARCH models have consistently been used in the literature to predict oil prices with varying degrees of accuracy.

3.2 Structural Models Considering Economic Factors

For the price of oil, structural models consider the oil price to be modelled as a function of certain explanatory variables such as oil consumption and production, OPEC behaviour, interest rates, exchange rates, and other commodity prices. The major drawback of using structural models to predict oil prices is that the models are extremely complex, and there is a strong inter correlation between factors themselves. Hence, there have not been many studies that focus on structural analysis to forecast oil prices.

According to Huntington[17], structural demand and supply models are generally not successful in predicting oil prices due to inaccurate forecasts of GDP and the oil supply from different countries. Another reason was not taking into account the market participation expectation of OPEC countries.

However, some interesting work has been done based solely on structural models. Most of these studies use models and results that instead of trying to predict the price, try to understand the nature of the oil market. Also, these models predict short-term oil prices, and it is unclear if they could be used for long term forecasting. One such interesting study by Pindyck[18] shows that long term oil prices are mean reverting around shifting trend lines.

In another direction Yang et al. [22] introduces a model to determine the factors affecting US oil prices. First, they highlight the unstable demand structure of the oil market. Then, they use a GARCH model (general autoregressive conditional heteroskedasticity) to investigate the volatility of oil prices. Using the co-efficients they generated, they estimate that the future oil price will be 0.987 times the current oil price if the US GDP decreases by 5%.

Similarly, structural models considering a variety of economic factors also apply to the gold price prediction. According to Ismail et al.[8], they design models with the gold price as the only dependent variable, alongside different numbers of independent variables. Initially, they propose that the gold price is dependent on the following 8 factors: Commodity Research Bureau future index (CRB), USD/Euro Foreign Exchange Rate (EUROUSD); Inflation rate (INF); Money Supply (M1); New York Stock Exchange (NYSE); Standard and Poor 500 (SPX); Treasury Bill (T-BILL) and US Dollar index (USDX). Their first-order regression model, which they call a naive model, is represented by:

$$\hat{Y} = -560.618 + 0.712X_1 + 161.740X_2 - 7.836X_3 + 0.424X_4 - 0.010X_5 + 0.010X_6 + 3.198X_7 + 0.580X_8$$

where $\hat{Y}$ is the predicted gold price; X_1 is CRB; X_2 is EUROUSD; X_3 is INF; X_4 is M1; X_5 is NYSE; X_6 is SPX; X_7 is T-Bill; X_8 is USDX.

Then, they show that using stepwise regression, the number of independent variables can be reduced from 8 to 4, thus resulting in a enhanced model which can be represented by:

$$\hat{Y} = -311.939 + 0.474X_1 + 113.258X_2 + 0.3279X_4$$

where $\hat{Y}$ is the predicted gold price, X_1 is CRB; X_2 is EUROUSD; X_3 is INF; X_4 is M1.

Performance Comparison: The paper reports that their models have a Root Mean Squared Error (RMSE) of 76.40 and 75.60, whereas our advanced models models considering macroeconomic factors (Advanced Models, Section 7) perform better than these models with a RMSE of 65.67 (Advanced Model 3).

Ismail's paper provides us an intuition of the factors we could use to build advanced models (autoregressive and multiple linear regression models) to predict the price of gold. In our advanced models, we include S&P 500 Index, NYSE Index and the US Dollar Index mentioned in Ismail's paper, as well as Consumer Sentiment Index and the Oil Price that are not. The details will be discussed in Section 7.

3.3 Non Standard Models

Non standard methods have gained popularity because most of the linear time series model fail to take into account the non-linearity of the data. Oil prices and gold prices have a strong non-linear and chaotic time series and some experts think that non-linear models might fit quite well in this context.

Several non standard and non-linear methods have been applied to the time series in recent times. One such interesting method is the Emperical Mode(EMD) Decomposition method[7]. This paper assumes that the data depending on its complexity, may have several different co-existing modes of oscillations. The authors try to extract these modes of oscillation and then they add them to forecast oil prices. EMD is a relatively innovative approach for modelling time series data and they seem to give relatively good results for long term predictions.

Support Vector Machines have also gained popularity for oil price forecasting. These methods have been used for forecasting oil consumption as well. For example, Dong et al [19] have used it to forecast oil consumption in tropical regions. Lin and Pai [20] have tried to use a hybrid model of Support Vector Machines and ARIMA model to model oil prices. Here the ARIMA models the linear aspects of the time series, whereas the SVM tries to model the non-linear aspects. They evaluated their results based on real data and seem to get very positive results.

Another interesting tool being used in recent times for oil price forecasting(and stock price predictions in general) is Artificial Neural Networks. ANNs are computational models inspired by the central nervous system and are used to estimate functions that can depend on large number of unknown inputs. Multiple ANN approaches have been used to predict oil prices and they all achieve varying degrees of accuracy. One interesting hybrid ANN as well as regression approach is the NARX(non linear autoregressive model with eXogenous input) model [21]. They claim that the NARX model is more accurate than time series and static ANN models in predicting oil prices in general as well as in predicting the occurence of oil price shocks.

4 Data Sets

Our data consists of multiple time series of monthly Oil and Gold prices [23], and related macroeconomic factors.

Crude oil benchmarks are reference prices for buyers and sellers. There are several benchmarks for the prices of oil, out of which the most popular ones are West Texas Intermediate (WTI), Brent, Dubai. Spot price is the current price at which a particular security can be bought or sold at a specified time and place.
Historical WTI Crude Oil prices and spot prices of gold were used for making predictions.

Date	Oil Price	Gold Price	S&P 500	NYSE	USD Index	CSI
11/30/2014	75.74	1182.8	2067.56	10955.41	80.81	86
10/31/2014	80.53	1164.3	2018.05	10845	80.8143	86
9/30/2014	91.17	1216.5	1972.29	10702.93	81.0908	86
8/31/2014	97.86	1285.8	2003.37	11046.29	77.9769	93.4
7/31/2014	98.23	1285.3	1930.67	10726.43	77.2128	90.3
6/30/2014	106.07	1315	1960.23	10979.42	75.7271	86.4

Fig. 4. Data matrix for oil and gold price and its related macroeconomic factors

The following macroeconomic factors were used:

- S&P 500 Index [23]
- New York Stock Exchange Index (NYSE) [23]
- US Dollar Index [23]
- Consumer Sentiment Index (CSI) [24]

The aforementioned data sets are monetary time series, and it is important to adjust for inflation in order to eliminate the unnecessary correlation caused due to inflation. Inflation adjustment is done by dividing the monetary time series by a price index, which is a number representing the price level relative to a base year. The Consumer Price Index (CPI) was used to adjust for inflation.

5 Observations

The correlation matrices in Figure 5 and Figure 6, show the correlation between the values of oil/gold price and related economic factors adjusted for inflation. The coefficients in the matrices indicate the level of correlation between the parameters. A coefficient close to 1 indicates a high positive correlation, and close to -1 indicates a high negative correlation. Coefficients closer to 0 indicate low correlation.

These matrices indicate a high positive correlation between the oil price and NYSE Index, gold price, and a high negative correlation between the oil price and the USD Index. Similarly, they indicate high correlation between the gold price and USD Index, CSI, Oil Price, and a very low correlation between gold price and S&P 500, NYSE Index.

Prior to adjusting for inflation, the correlation between these factors was much higher. However, despite using inflation adjusted data and factors with low correlation, the performance of our predictive models (discussed in Section 6 and 7) improved. This indicates two important things. First, it is important to adjust for inflation to discard meaningless correlation caused purely due to inflation. Second, highly correlated factors do not always guarantee better results, and factors with low correlation should not be discarded because they could help building a better predictive model.

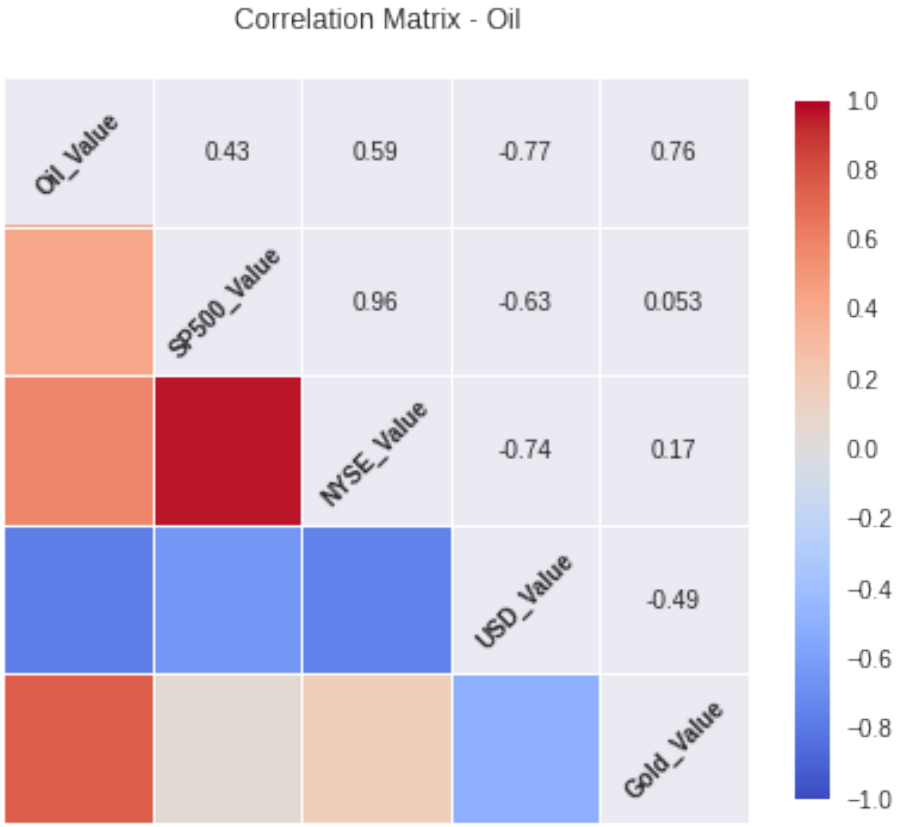

Fig. 5. Correlation matrix for oil price and related macroeconomic factors

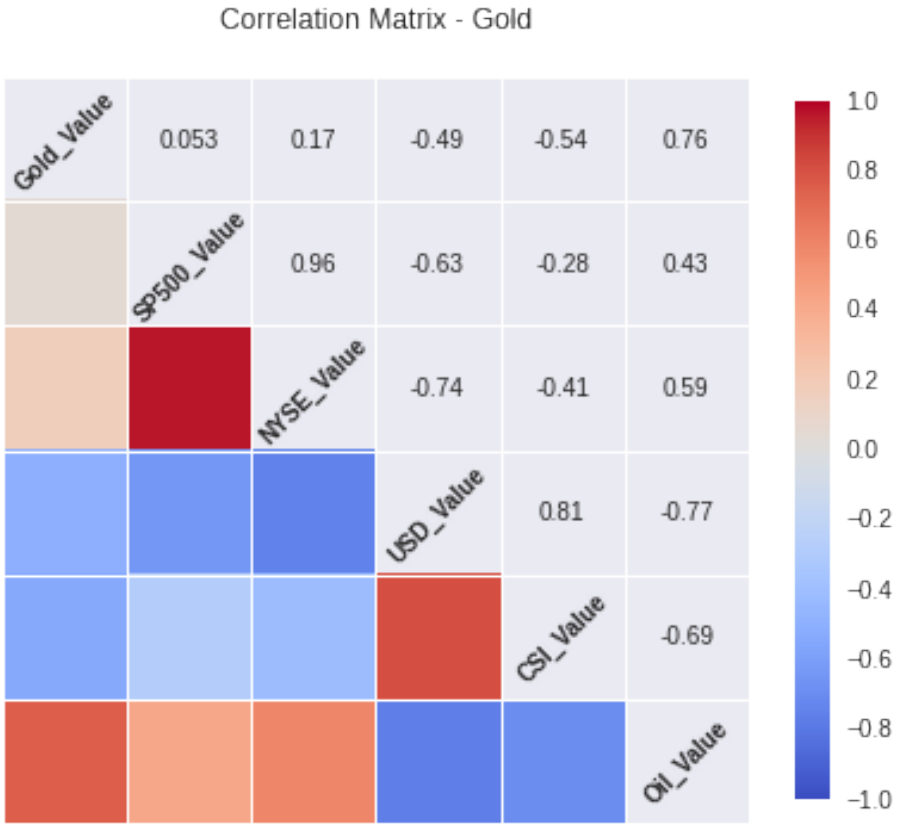

Fig. 6. Correlation matrix for gold price and related macroeconomic factors

6 Baseline Models

We initially developed simple baseline models which predict the price solely as a function of the past few months. In order to illustrate an improvement in the accuracy of our predictions, we compare our advanced models to the following baseline models:

- **Baseline Model 1** The price stays the same as the previous month's price: $P_t = P_{t-1}$
- **Baseline Model 2** The price is a cubic weighted average of the last 3 months: $P_t = \frac{1}{\{k(k+1)\}^2} \sum_{i=1}^{k} (k-i+1)^3 P_{t-i}$

While developing the baseline models, it was observed that the oil/gold price tends to be closer to the previous month's price, as compared to earlier months. Hence, assigning more weight to the previous month's price (in this case, cubic weights) resulted in an improved model.

7 Advanced Models

We developed autoregressive models operating only on price data, autoregressive models with multiple linear regression to incorporate macroeconomic factors, as well as regression models using futures data to make our predictions.

7.1 Autoregressive Models

Autoregressive Model 1: Autoregressive Model These models are purely based on the historical prices. It models 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 Model 2: 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.

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.
The following equation is a representation of the ARMA model:

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

Here, ϵ_t is the random white noise, a random variable with zero mean, constant variance and zero auto-correlation.

7.2 Autoregressive and Multiple Linear Regressive Models Considering Macroeconomic Factors

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

The pure autoregressive model was expanded to incorporate the macroeconomic factors which are highly correlated to the price of commodity being predicted. As the various factors are incorporated into the model, a comparison of the error metrics is performed to estimate the set of macroeconomic factors which make predictions with the 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
- Consumer Sentiment Index
- Oil Price

Several combinations of economic factors were incorporated into the models, and the best preforming models were selected. The following table enlists the various multiple linear regressive models used and the factors they incorporate.

Model Name	Macroeconomic Factors (Oil)	Macroeconomic Factors (Gold)
Advanced Model 1	*Gold Price*	*CSI*
Advanced Model 2	*S&P* 500, *Gold Price*	*CSI*, *Oil Price*
Advanced Model 3	*NYSE*, *Gold Price*	*S&P* 500, *NYSE*, *Oil Price*
Advanced Model 4	*NYSE*, *USDX*, *Gold Price*	*S&P* 500, *NYSE*, *CSI*, *Oil Price*
Advanced Model 5	*S&P* 500, *USDX*, *Gold Price*	*S&P* 500, *NYSE*, *USDX*, *CSI*, *Oil Price*

Table 1. Various multiple linear regressive models and the respective macroeconomic factors taken into consideration

7.3 Regression using Futures Data Models

Futures Model 1 Past futures prices are used to predict the current spot price. A simple regression is used to make predictions. This model can be represented as follows:

$$X_t = \sum_{i=1}^{k} \alpha_i F_{t-i}^t + c.$$

Futures Model 2 The Futures Model 1 is improved, and instead of trying to predict the spot price, we try to predict the difference in the current spot price and the last months spot price, using the difference in the futures price at lag i and the spot price at lag i. This model can be represented as follows:

$$X_t - X_{t-1} = \sum_{i=1}^{k} \beta_i (F_{t-i}^t - X_{t-i}) + c.$$

7.4 Evaluation Metrics

The models are trained on the initial 70% of the time series, and tested on the remaining 30%. They try to generate a linear function while assigning coefficients to each of the parameters. These regression coefficients are then used to make predictions.

The performance of our models is evaluated based on the following error mertics:

- **Mean Relative Error (MRE)**
  $MRE = \frac{1}{n} \sum_{i=1}^{n} \frac{|P_t^* - P_t|}{P_t}$
- **Mean absolute Error (MAE)**
  $MAE = \frac{1}{n} \sum_{i=1}^{n} |P_t^* - P_t|$

– **Root Mean Square Error (RMSE)**

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(P_t^* - P_t)^2}$$

A histogram of the distribution of relative error is also generated.

7.5 Performance of Models Predicting the Price of Oil

The following is a comparison of the performance all the models used for predicting the price of oil.

Model	Mean Rel Error	Mean Abs Error	RMSE
Baseline Model 1	**6.58%**	**5.93**	**7.85**
Baseline Model 2	7.17%	6.37	8.58
Autoregressive Model 1	7.36%	5.73	7.29
Autoregressive Model 2	**6.78%**	**5.03**	**5.42**
Advanced Model 1	7.27%	5.67	7.21
Advanced Model 2	**6.99%**	**5.52**	7.16
Advanced Model 3	7.06%	5.56	7.16
Advanced Model 4	7.17%	5.61	7.14
Advanced Model 5	7.04%	5.54	**7.12**
Futures Model 1	3.75%	3.80	4.41
Futures Model 2	**3.03%**	**3.63**	**4.32**

Table 2. Error metrics of various models for predicting the price of oil

Figure 7 shows the comparison of the relative errors of different predictive models. Figure 9 shows the error histograms generated by different models.

Summary

– Among the autoregressive models, Autoregressive Model 2 (ARMA) performed the best with a mean relative error of 6.78%.

– Among the regressive models considering economic factors, the Advanced Model 2, considering S&P 500 and the gold price performed best with a mean relative error of 6.99%. This performance of this model was better as compared to the Baseline Model 2.

– The Autoregressive Model 2 (ARMA) also performs better in comparison to any of the models considering the economic factors.

– Futures Model 2 (with a mean relative error of 3.03%), which considers the difference between spot and futures prices, performs slightly better as compared to Futures Model 1 (with a mean relative error of 3.75%).

– Overall, the futures models perform significantly better (almost twice as good) than any of the autoregressive or multiple linear regressive models because the futures prices tend to encapsulate more information than just the few economic factors considered in the advanced models, thus giving better results.

7.6 Performance of Models Predicting the Price of Gold

The following is a summary of all the models used for predicting the price of gold.

Model	Mean Rel Error	Mean Abs Error	RMSE
Baseline Model 1	4.61%	56.11	75.36
Baseline Model 2	**4.56%**	**55.78**	**73.27**
Autoregressive Model 1	**4.24%**	46.61	65.91
Autoregressive Model 2	4.32%	**45.73**	**63.71**
Advanced Model 1	**4.24**%	**46.62**	65.87
Advanced Model 2	4.3%	47.24	65.86
Advanced Model 3	4.68%	49.17	**65.67**
Advanced Model 4	5.61%	55.02	69.51
Advanced Model 5	5.81%	56.09	70.06
Futures Prices Model 1	**2.34**%	**26.86**	**35.42**
Futures Prices Model 2	4.03%	42.71	79.49

Table 3. Error metrics of various models predicting the price of gold

Figure 8 shows the comparison of the relative errors of different predictive models. Figure 10 shows the error histograms generated by different models.

Summary

– Among the autoregressive models, the Autoregressive Model 1 considering only historical gold prices performed the best with a mean relative error of 4.24%. It performs better than both the baseline models.

– Among the multiple linear regressive models considering economic factors, Advanced Model 1 performed best with a mean relative error of 4.24%. This model also performed better compared to the Baseline Model 1 and 2. It

performs as good as the pure autoregressive model.

– Models considering other combinations of economic factors (Advanced Models) fail to perform better than the Autoregressive Model 1.

– Overall, the Futures Models 1, with a mean relative error of 2.34% performs significantly better (almost twice as good) as compared to any of the autoregressive or multiple linear regressive models because the futures prices tend to encapsulate more information than just the few economic factors considered in the advanced models, thus giving better results. Futures Model 2, which considers the difference between spot and futures prices, does not perform better than the Futures Model 1.

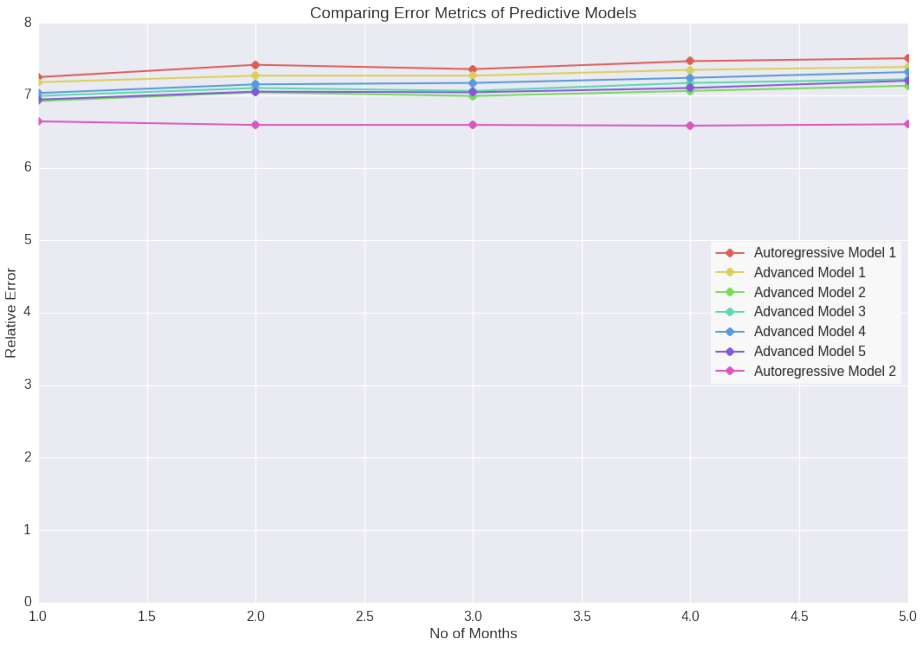

Fig. 7. This plot shows the relative performance of the autoregressive models, and the advanced models incorporating macroeconomic factors predicting the price of oil. It shows the relative error percentage of various models with an increase in the number of past values (months) of the parameters taken into consideration by the model. The advanced models 1-5 considering certain combinations of economic factors perform better than the pure Autoregressive Model 1. The Autoregressive Model 2 (ARMA) performs better than the Autoregressive Model 1 and the Advanced Models.

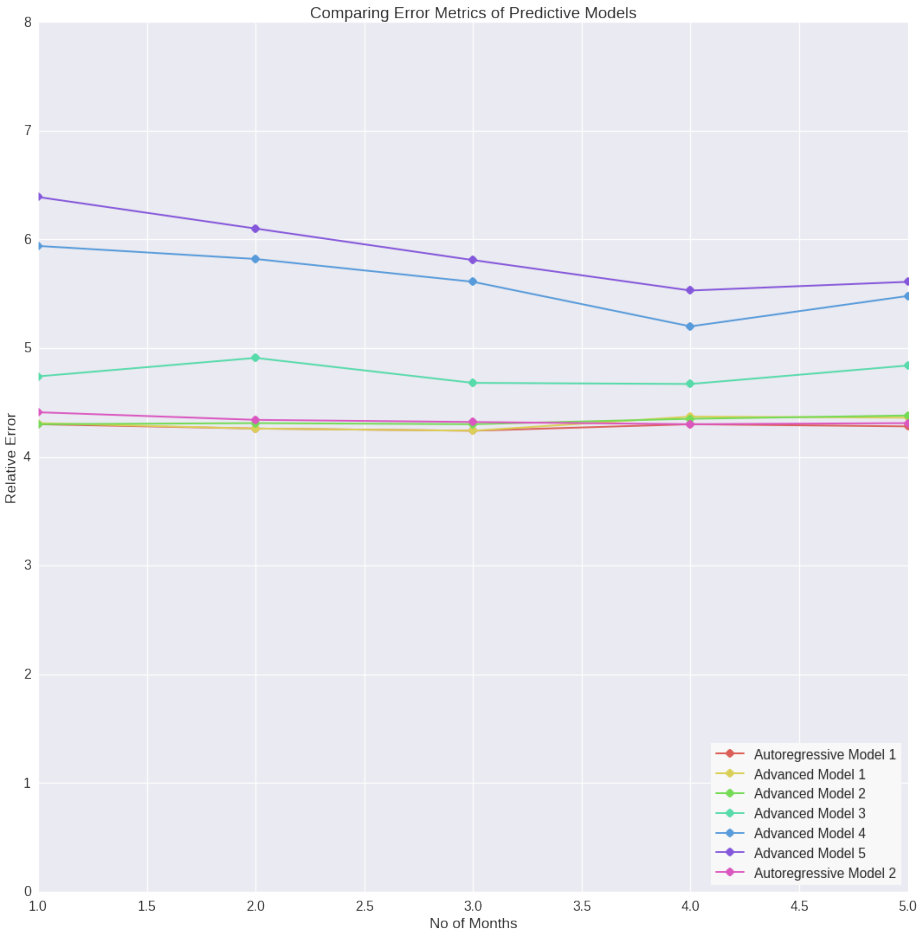

Fig. 8. This plot shows the relative performance of the autoregressive models, and the advanced models incorporating macroeconomic factors predicting the price of gold. It shows the relative error percentage of various models with an increase in the number of past values (months) of the parameters taken into consideration by the model. The advanced model 1, and the autoregressive models 1,2 perform almost equally well, and better than the other advanced models.

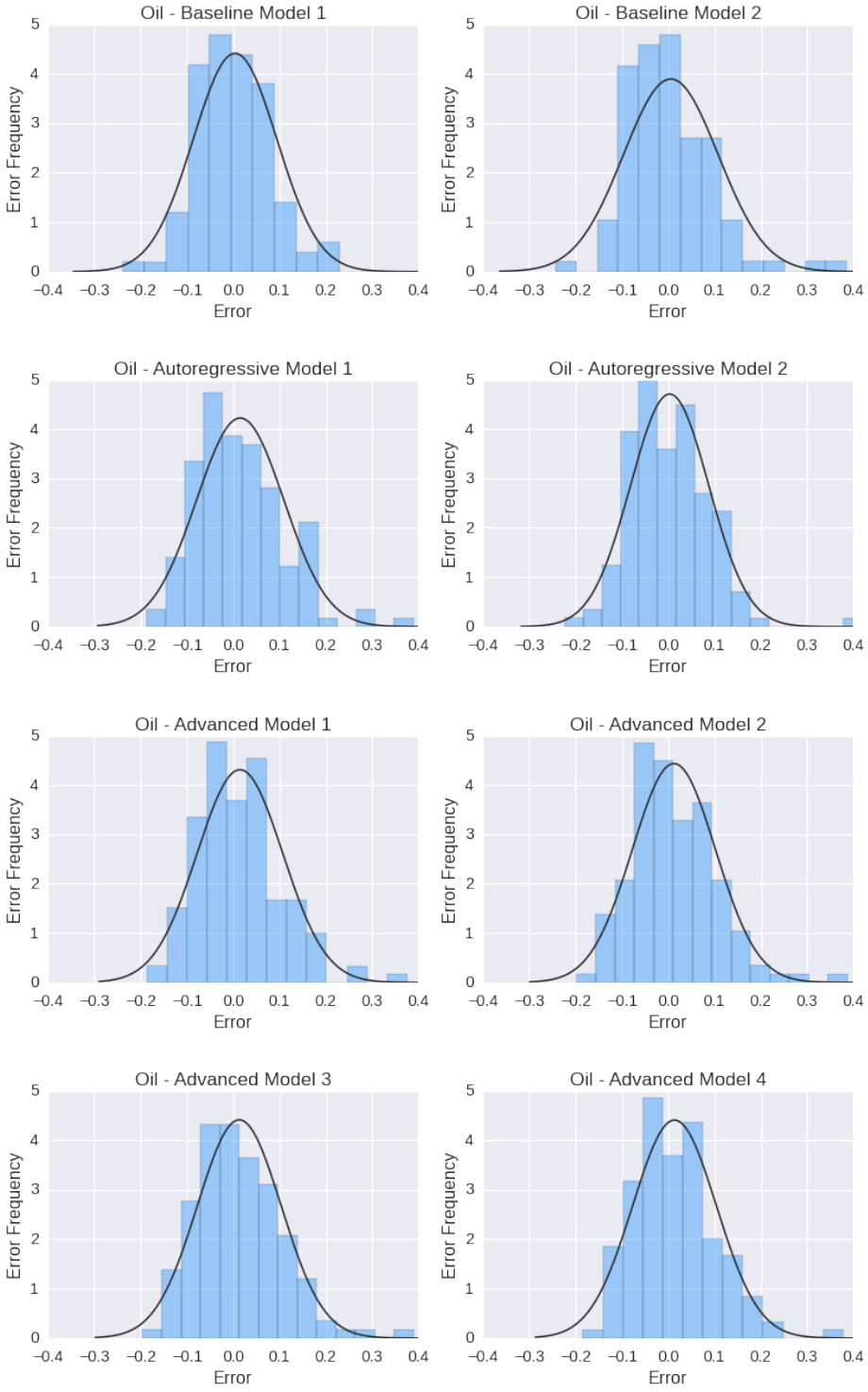

Fig. 9. Histograms of percentage error in models for predicting the price of oil

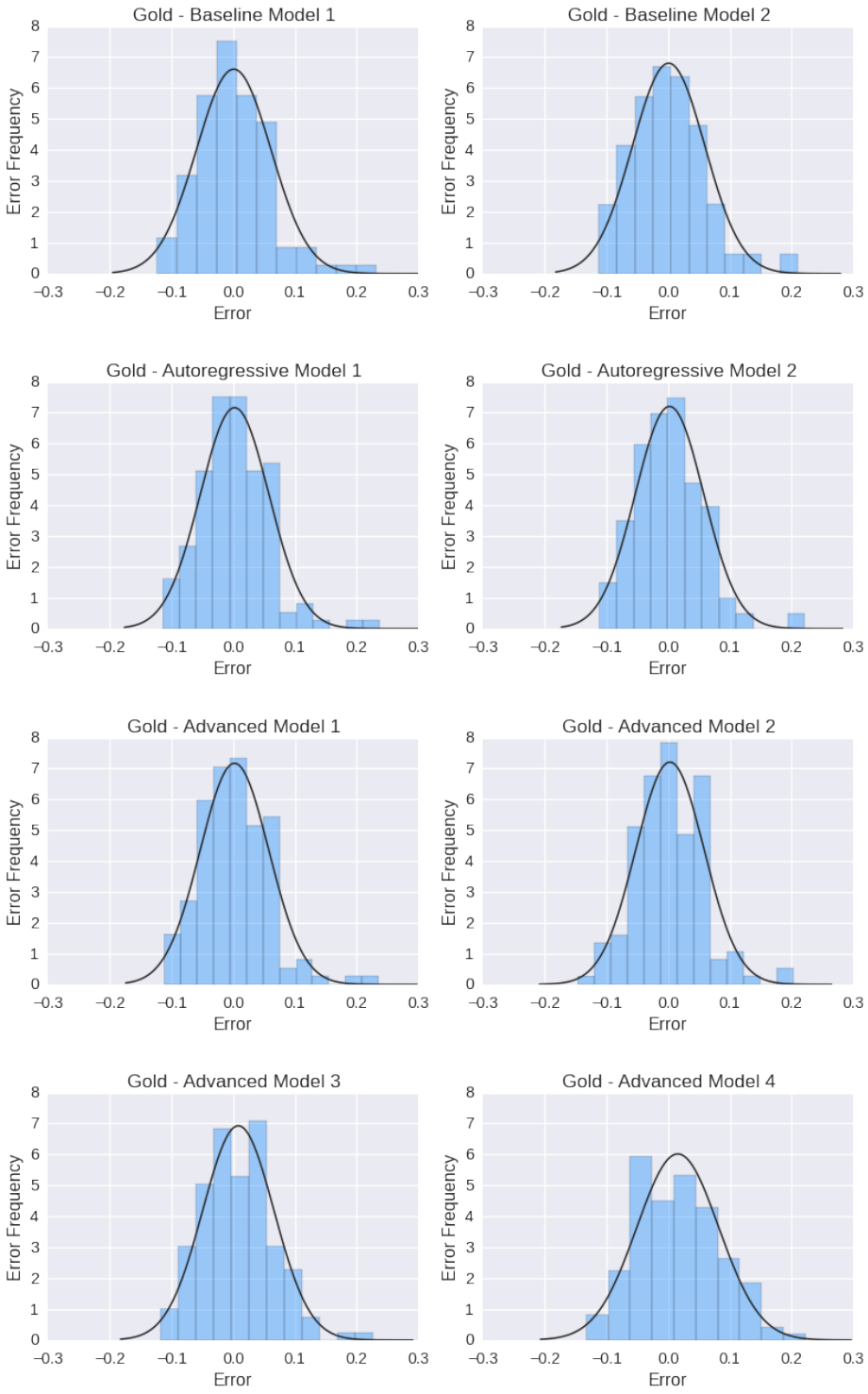

Fig. 10. Histograms of percentage error in models for predicting the price of gold

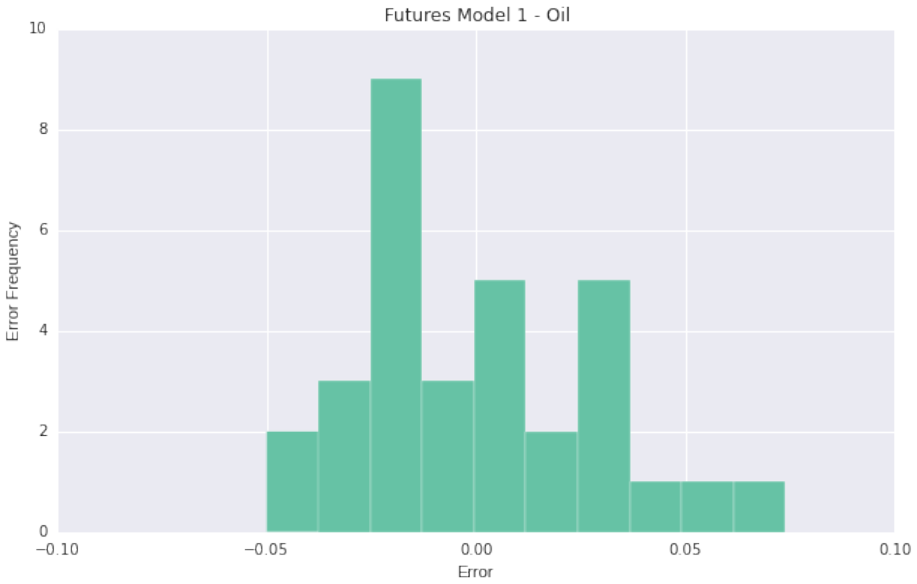

Fig. 11. Histogram of percentage error in Futures Model 1 for predicting the price of oil

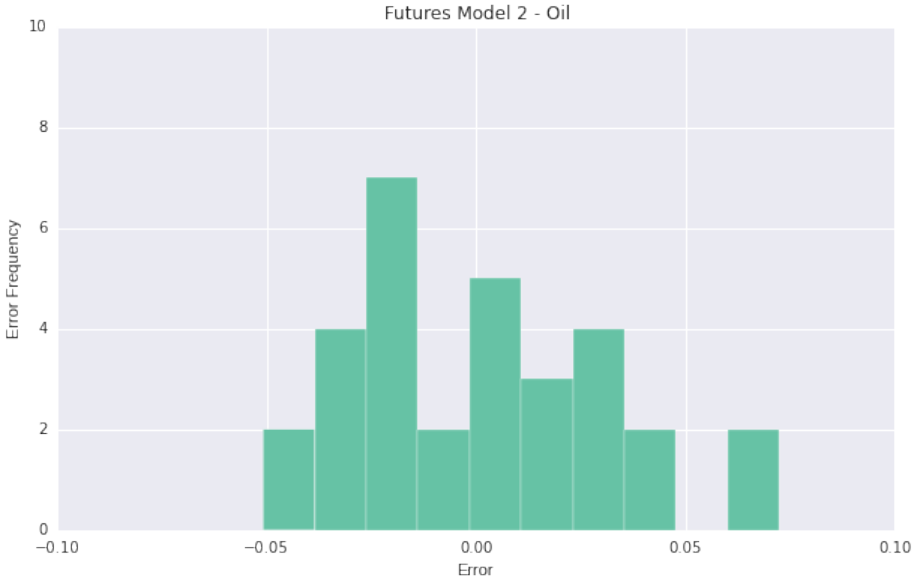

Fig. 12. Histogram of percentage error in Futures Model 2 for predicting the price of oil

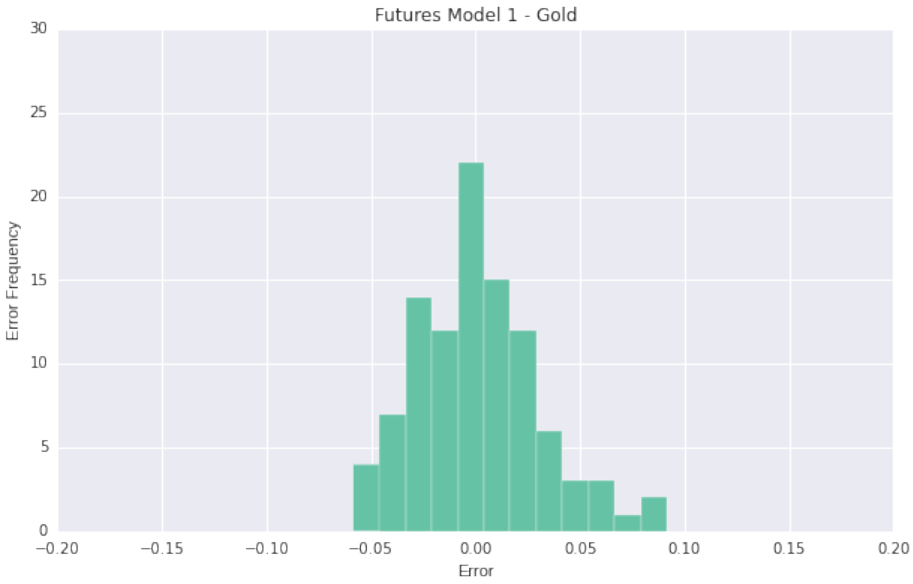

Fig. 13. Histogram of percentage error in Futures Model 1 for predicting the price of gold

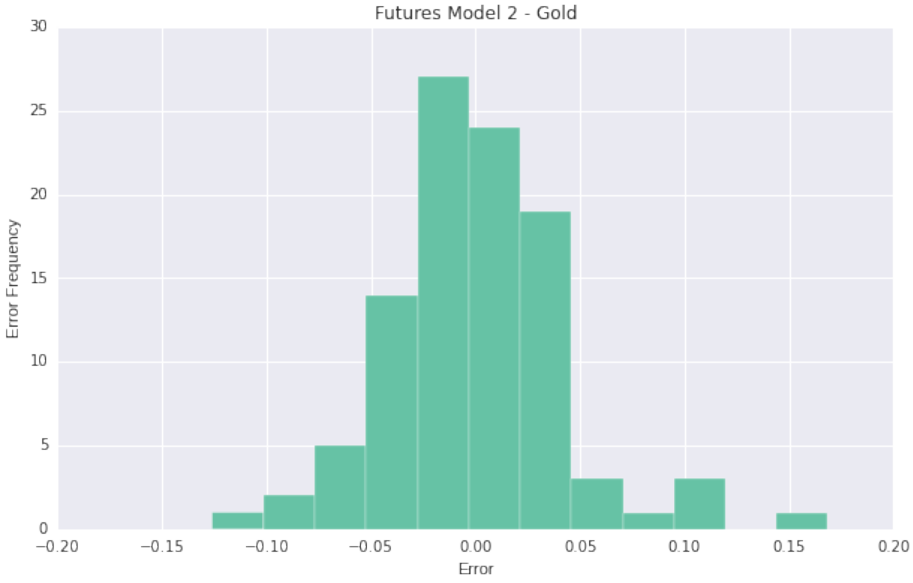

Fig. 14. Histogram of percentage error in Futures Model 2 for predicting the price of gold

8 Final Prediction and Conclusions

8.1 Final Prediction

Our prices are predicted as a probability distribution. The distribution gives the probability that the relative error percentage would lie within a certain value.

$E = [e1, e2, e3... \ en]$
$P = [p1, p2, p3... \ pn]$

E is an array consisting of the relative error percentages and P is the corresponding cumulative probability distribution. The i-th value of the probability distribution array P gives the probability that our relative error lies within $E[i]$.

Oil Price Prediction

Our predicted price of **WTI Crude Oil** on **Jan 1, 2015** as of Dec 1, 2014 is:
57.33 USD per Barrel

The price of oil is predicted with the following probability distribution:

$E = [0.01, 0.02, 0.03, 0.040.050.060.070.08]$

$P = [0.12, 0.56, 0.66, 0.88, 0.94, 0.97, 0.98, 1.0]$

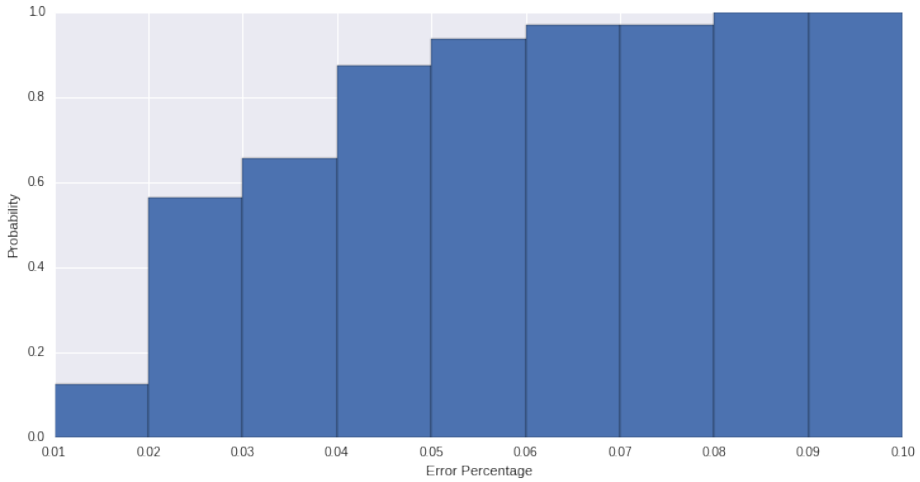

Fig. 15. Cumulative probability distribution of percentage error for predicted oil price

Gold Price Prediction

Our predicted spot price of **Gold** on **Jan 1, 2015** as of Dec 1, 2014 is
1175.29 USD per Ounce

The price of gold is predicted with the following probability distribution:

$E = [0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16]$

$P = [0.15, 0.37, 0.63, 0.76, 0.84, 0.90, 0.92, 0.93, 0.95, 0.95, 0.97, 0.98, 0.99, 0.99, 0.99, 0.99, 1]$

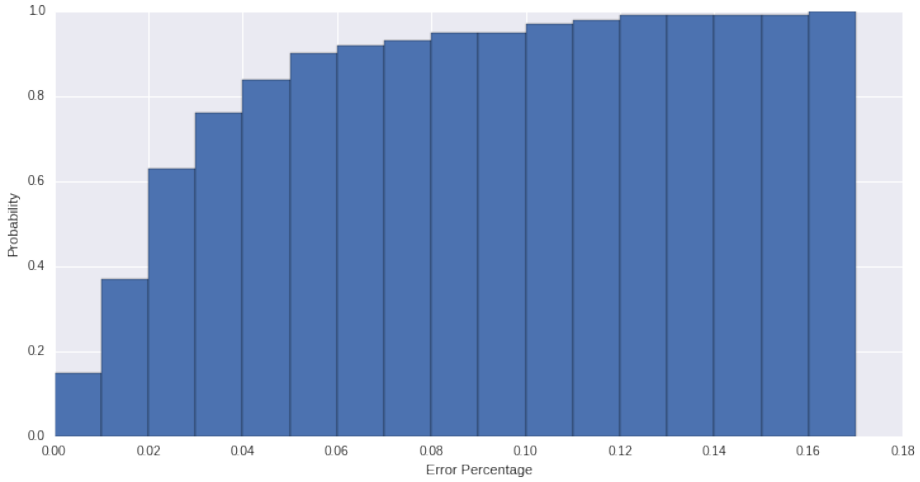

Fig. 16. Cumulative probability distribution of percentage error for predicted gold price

8.2 Difficulties We Had to Overcome in Building Good Models and Investigations for Subsequent Groups

- **Missing Data**
  We initially faced a problem of missing data rows in the data sets of certain economic parameters. This resulted in a dimension mismatch error while merging multiple time series into a single data frame. Using an inner join to combine data frames resolved our issue.

- **Inflation Adjustment**
  The data we used for our predictions were monetary time series, and it is important to inflation adjust the data before performing a correlation. We initially had not adjusted for inflation, which resulted in very high correlation between the oil/gold price and related macroeconomic factors. Most of this correlation was meaningless and could be attributed to inflation.

  Adjusting the data for inflation reduced the correlation to a large extent, and also resulted in an improvement in our models. Earlier, the performance of models incorporating economic factors was worse compared to all other models. Adjusting for inflation resulted these models performing better than the baseline and the pure autoregressive models. This is definitely an indicator that highly correlated factors are not always the best predictors.

- **Data Unavailability**
  We intended to include more macroeconomic factors into our model, but certain factors could not be included due to the unavailability of historical data. For instance, Euro-USD exchange rate was a factor we failed to incorporate because the Euro was introduced only in 1992, and we were using past 30 years data (since 1986) for the other factors.

- **Downloading Futures Data**
  The futures data is available as individual files for every month of the year. Downloading approximately 360 files, and reading them into a single data frame was a challenge.

- **Outliers in Futures Data**
  The futures model was giving us an RMSE ¿13, which was a very high number compared to the other models. However, the mean error and variance were not as bad. We had almost disregarded this model assuming that it wasn't good enough.

  But when we plotted the error histogram, we noticed an outlier. Most of the errors were lying between -0.1 and +0.1, but this outlier had an error of -0.9 which was causing the large RMSE. When we investigated the data set, we found that the outlier was caused due to a data entry being 0. We then removed the outlier from our data sets, we realized that the futures model

was by far the best model we had developed.

The important lesson we learnt was that we should always plot graphs to detect outliers, and it is important to know the reason behind the outlier. Also, since a RMS amplifies the error, we shouldn't simply disregard the results. Instead, we should look deeper into the data and investigate the cause.

Acknowledgments. Dr. Steven Skiena. Professor Keli (Andrew) Xiao.

References

1. Soos, Andy (6 January 2011). "Gold Mining Boom Increasing Mercury Pollution Risk". Advanced Media Solutions, Inc. (Oilprice.com). Retrieved 26 March 2011.
2. "OPEC : Home.". http://www.opec.org/
3. "Energy and Financial Markets: What Drives Crude Oil Prices". http://www.eia.gov/finance/markets
4. Zhang, Yue-Jun, et al. "Spillover effect of US dollar exchange rate on oil prices." Journal of Policy Modeling 30.6 (2008): 973-991.
5. Shahriar Shafiee, Erkan Topal.:An overview of global gold market and gold price forecasting. ScienceDirect. Volume 35, Issue 3, September 2010, 178–189 (2010)
6. Zhang, Yue-Jun, and Yi-Ming Wei. "The crude oil market and the gold market: Evidence for cointegration, causality and price discovery." Resources Policy 35.3 (2010): 168-177.
7. Zhang, Xun, Kin Keung Lai, and Shou-Yang Wang. "A new approach for crude oil price analysis based on empirical mode decomposition." Energy Economics 30.3 (2008): 905-918.
8. Z. Ismail, A. Yahya and A. Shabri. "Forecasting Gold Prices Using Multiple Linear Regression Method." American Journal of Applied Sciences 6 (8) (2009): 1509-1514.
9. The Relationship between Gold and Crude Oil Price, http://www.marketoracle.co.uk/Article38141.html
10. What's the difference between consumer confidence and consumer sentiment?, http://www.investopedia.com/ask/answers/09/consumer-confidence-sentiment-difference.asp
11. What's the difference between consumer confidence and consumer sentiment?, http://www.investopedia.com/ask/answers/09/consumer-confidence-sentiment-difference.asp
12. Consumer confidence index, http://en.wikipedia.org/wiki/Consumer_confidence_index
13. Engle, Robert. "GARCH 101: The use of ARCH/GARCH models in applied econometrics." Journal of economic perspectives (2001): 157-168.
14. Hadavandi, Esmaeil, Arash Ghanbari, and Salman Abbasian-Naghneh. "Developing a Time Series Model Based on Particle Swarm Optimization for Gold Price Forecasting." Business Intelligence and Financial Engineering (BIFE), 2010 Third International Conference on. IEEE, 2010.
15. Khashei, Mehdi, Mehdi Bijari, and Gholam Ali Raissi Ardali. "Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs)." Neurocomputing 72.4 (2009): 956-967.

16. Khashei, Mehdi, Seyed Reza Hejazi, and Mehdi Bijari. "A new hybrid artificial neural networks and fuzzy regression model for time series forecasting." Fuzzy Sets and Systems 159.7 (2008): 769-786.
17. Huntington, Hillard G. "Oil price forecasting in the 1980s: what went wrong?." The Energy Journal (1994): 1-22.
18. Pindyck, Robert S. "The long-run evolution of energy prices." The Energy Journal (1999): 1-27.
19. Dong, Bing, Cheng Cao, and Siew Eang Lee. "Applying support vector machines to predict building energy consumption in tropical region." Energy and Buildings 37.5 (2005): 545-553.
20. Pai, Ping-Feng, and Chih-Sheng Lin. "A hybrid ARIMA and support vector machines model in stock price forecasting." Omega 33.6 (2005): 497-505.
21. Godarzi, Ali Abbasi, et al. "Predicting oil price movements: A dynamic Artificial Neural Network approach." Energy Policy 68 (2014): 371-382.
22. Yang, C. W., Ming-Jeng Hwang, and Bwo-Nung Huang. "An analysis of factors affecting price volatility of the US oil market." Energy Economics 24.2 (2002): 107-119.
23. Quandl - Find, Use and Share Numerical Data. `https://www.quandl.com`
24. `http://future.aae.wisc.edu/data/monthly_values/by_area/998?grid=true`
25. Hull, John. Options, futures and other derivatives. Pearson education, 2009.