Norwegian School of Economics
Bergen, Fall 2018

Properties of the Gold Price

An investigation using fractional Brownian motion and supervised machine learning techniqes

Theodor Bruu Røine & Edvard Kvalø Holter

Supervisor: Jan Ubøe

Master Thesis, MSc in Economics and Business Administration

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

This thesis investigates properties of the gold price. Two different aspects have been analyzed using two different methods. The concept of fractional Brownian motion has been utilized in the search for evidence of long memory in the gold price. Further, the machine learning techniques Gradient Boosting Machine and XGBoost are used in the investigation of the relative importance of financial and economic variables in a prediction of the gold price. Both aspects are studied across multiple time periods in order to examine the potential change.

We find evidence of long memory in the gold price, however not in all examined time periods from 1979 until now. The first ten years from 1979, as well as the ten years from the start of the 2008 financial crisis to today, appear to have the property of persistence, while the years in between display no evidence of long memory. The second part of the analysis, the relative importance of variables in a prediction, is more focused on the years before and after the 2008 financial crisis. We find that variables such as crude oil and durable goods orders are relatively important before the crisis, while crude oil and US 10-year treasury bonds are relatively important after. Both parts of the analysis indicate that the properties of the gold price change over time.

Preface

This thesis completes our Master of Science in Economics and Business Administration in Financial Economics at the Norwegian School of Economics (NHH). The idea behind this thesis was combining one of the most prominent entities of the financial markets with the application of various analysis tools and techniques inherited after five years of studies. The writing process has been challenging, but above all rewarding and educational.

We would like to express our deepest gratitude to our supervisor, Jan Ubøe, for advice and guidance throughout the semester. Without him there would be no master thesis. Further, we would like to thank Constantin Gurdgiev for topic suggestions and inspiration. We would also like to thank Walter Pohl for providing us with an understanding of machine learning techniques. Lastly, we would like to thank our family and friends for motivating and pushing us to complete this thesis.

Bergen, 19.12.2018

Theodor Bruu Røine

Edvard Kvalø Holter

Contents

Introduction

Gold is a commodity being known to mankind for ages. The scarcity, visual appearance and physical durability have made gold valuable. Its roles range from jewelry, money, industrial usage and being an investment asset. The fact that the gold price extends beyond indicating the value of a certain amount of a commodity, makes it an interesting entity to research.

This thesis aims to further look into some of the gold price' properties, predominantly in the context of gold as an investment asset. Even though gold has been functioning as an investment asset for centuries, we are mostly interested in the last decades, following the fall of the Bretton-Woods system in the 1970s. This was the first occasion the dollar was completely independent from either gold or silver since 1792, and marks a significant change in the market for gold (Hillier, Draper and Faff, 2006). This is when the gold's role as an investment asset truly emerges, and the monetary role diminishes in the world's largest economies, at least formally.

In the investigation of the properties of the gold price we will look at two aspects. Firstly, we will look for evidence of long memory. Secondly, we will look at gold's relationship to financial and economic factors through the creation of a predictive model.

In the search for evidence of long memory, we have applied the concept of fractional Brownian motion. This concept differs from the ordinary Brownian motion, among other things, by its characteristic Hurst exponent. By estimating this exponent, we will be able to get an indication of whether the process is persistent or anti-persistent. The phenomenon of long memory plays a significant role in many fields, with implications in regard to forecasting skills and trends (Graves et al. 2017), and in this thesis we are looking for evidence that the gold price is an entity influenced by this phenomenon.

With the possible presence and implications of long memory, the further analysis of the gold price properties will be conducted through a predictive model. The utilization of supervised machine learning methods, namely Gradient Boosting Machine (GBM) and XGBoost, enables the opportunity to display the relative importance of the input variables. The input variables are a selection of economic and financial factors and indicators, chosen on background of their approved relevance in literature, or simply because they appear interesting based on economic intuition. Our aim is not to examine the causal relationships, but rather the relative importance in a prediction.

The main motivation for investigating the properties of the gold price is to study if they change over time. Especially, we will emphasize the comparison of the periods before and after the 2008 financial crisis. The height of the 2008 financial crisis is now a considerable number of years ago, meaning that the amount of post-crisis data is now of significant volume, which we hope to exploit in this thesis.

The structure of the thesis is divided in two main parts. Part one will address the analysis of the possible presence of long memory in the gold price. Here the existing research on the topic will be presented, followed by the methodology chosen for estimating the Hurst exponent in section 1.1. The results of this analysis are shown and discussed shortly in section 1.2.

Part two is dedicated to the variable importance in a prediction of the gold price. Again, the existing research will be presented first, followed by a description of the supervised machine learning techniques applied in the analysis in section 2.1. Section 2.2 goes through the input data and the data preparations conducted before running the models. Section 2.3 contains the results. First, the variance importance plots are presented in section 2.3.1. In section 2.3.2 each variable will be looked at individually, before section 2.3.3 compares the time periods analyzed in a more general fashion, and investigates what differences or similarities that can be found across time. In order to assess the quality of the variable importance plots, section 2.3.4 will look at the predictive accuracy of the applied models. Section 2.3.5 discusses limitations and weaknesses of the analysis.

1. Fractional Brownian motion and the Hurst exponent

Fractional Brownian motion is a concept used to model and describe a wide range of features. Examples ranges from water levels and temperatures (Yerlikaya-Özkurt et al., 2013), applications in the field of biology (Menglong, 2013), and description of movements in the financial market (Mandelbrot and Hudson, 2004). In this text, the concept will be used in the examination of the properties of the gold price. More specifically, we will look further into the presence of long memory in the movements of the gold price. In general, the literature is split on evidence for long memory in the gold price or financial data in general, with for example Mandelbrot (1972), Greene and Fielitz (1977), Caporale et al. (2014) and Mynhardt et al. (2014) all providing evidence of the existence of long memory, while Lo (1991), Jacobsen (1995), Batten et al. (2005) to mention a few do the opposite.

Fractional Brownian motion is a non-stationary random process with stationary self-similar increments, characterized by the Hurst exponent, $0 < H < 1$ (Lacasa et al., 2009).
If $H = \frac{1}{2}$, the fractional Brownian motion corresponds to ordinary Brownian motion which has independent increments (Yerlikaya-Özkurt et al., 2013), meaning that the movement is independent of the past. By allowing this Hurst exponent to differ from ½, the fractional Brownian motion can be said to be a generalization of the Brownian motion (Shevchenko, 2014). A fractional Brownian motion with a Hurst exponent $H > \frac{1}{2}$ is called a persistent process, where the increments of the process are positively correlated. In this situation, the fractional Brownian motion is more likely to keep its trend. Oppositely will $H < \frac{1}{2}$ indicate negatively correlated increments, and the process being anti-persistent (Yerlikaya-Özkurt et al., 2013). This means that if the fractional Brownian motion decreased in the past, it is more likely to increase in the future.

An important note on the use of fractional Brownian motion when modelling financial assets, is the presence of arbitrage opportunities that comes with allowing the H to differ from $\frac{1}{2}$ (Rogers, 1997). In recent years though, the use of fractional Brownian motion has been accepted to model financial assets when imposing certain conditions. For example will the introduction of transaction costs eliminate arbitrage opportunities (Guasoni, 2006).

Through fractional Brownian motion we can also examine other properties, such as Fractal dimension, d. This is a term considering the roughness of a model and was introduced by Mandelbrot (Mandelbrot, 1983). The fractal dimension can be estimated by using the Hurst exponent. A larger fractal dimension means that the oscillations of the curve will increase at any scale. Conversely, smaller values of d imply a greater degree of smoothness to the graph of the curve (Granero, 2011). The Hurst method results in a fractal dimension of $d = 2 - H$ (Aref, 1998). Thus, a fractal dimension close to 1.5 implies a rough and volatile time series, as the Hurst exponent would be around 0.5 (Voss, 2013).

In our examination of the gold price and its possible long memory, we will estimate the Hurst exponent as well as take a look at the roughness of the model through fractal dimension. Both monthly and daily data will be examined. The results will also display how the Hurst exponent has changed over time. In section 1.1.1, multifractional Brownian motion will briefly be introduced, a potentially interesting extension of fractional Brownian motion that allows for a dynamic Hurst exponent.

1.1 R/S Analysis

Mandelbrot (1971) suggested that Rescaled Range (R/S) analysis could be used in studies of economic data. The rescaled range was developed by Hurst (1951), a hydrologist studying the floods of the Nile, and he discovered that the water movements were not independent. The R/S analysis is used to estimate the Hurst exponent, and thus finding potential evidence of long memory in a time series. There are alternative ways of estimating the Hurst exponent, namely detrended fluctuation analysis (DFA), a method that produces a fluctuation function $F(n)$ as a function of lag$n$ (Morariu et al., 2007). We have chosen the R/S analysis because it is the most appropriate in the case of a financial series, and it has been well developed and tested for more than 50 years (Kristoufek, 2010, Caporare et. al., 2017).

The following describes the method of estimating the R/S statistic and the Hurst exponent.

Select applicable time series.

Determine ranges to be analyzed.

Calculate the mean for each range means

$$mean_s = m_s = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where,

s = series (series 1 is all observations in the dataset, series 4 is four ranges of ¼ of whole dataset)

n = size of range

X_i = value of one element

Establish new time series adjusted with the mean for each range:

$$Y_t = X_t - m; for\ t = 1,2,\dots,n$$

Y = New time series

Create cumulative series with total running deviations from the mean:

$$y = \sum_{i=1}^{t} Y_t$$

y = running total of the deviations from the mean of each series:

Calculate the widest spread in the series of deviations:

$$R_t = \max(Y_1, Y_2, \dots, Y_t) - \min(Y_1, Y_2, \dots, Y_t)\ ;\ for\ t = 1,2,\dots,n$$

Calculate the standard deviation for each range:

$$\sigma = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(X_i - m)^2 \; ; for\ t = 1,2,\dots,n}$$

Calculate the rescaled range for each range, which is the widest spread in the time series measured by the standard deviation for each range:

$$rescaled\ range = \left(\frac{R}{S}\right)_t = \frac{R_t}{\sigma_t}; for\ t = 1,2,\dots,n$$

$\frac{R}{S}$ = Rescaled range for each range in the time series

Find the mean of the rescaled range values for each series to summarize each range

Now to calculate Hurst exponent:

Find the logarithmic values for the size of each series and for the rescaled range of each series.

Run a linear regression on the logarithm of the size of each series against the logarithm of the rescaled range.

Find the Hurst exponent by calculating the slope of the regression.

1.1.1 Multifractional Brownian Motion

The fractal Brownian motion is a continuous time Gaussian process $B_H(t)$ on $[0,T]$, starting at zero, has expected value of zero for all t, and has the following covariance function:

$$E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}$$

Here, H represents the Hurst index of the associated fractal Brownian motion. Lèvy used the Riemann-Liouville fractional integral to define a model where the next state of the model will depend on previous states of the model (Lizorkin, 2001). This process is called the Riemann-Liouville fractional Brownian motion (RLfBm) and is given by:

$$B_t^H = \frac{1}{\Gamma(\mathrm{H} + \frac{1}{2})}\int_0^t (t-s)^{H-\frac{1}{2}} dB_s$$

An extension of the Riemann-Liouville fractional Brownian motion was introduced by Vehèl and Peltier in 1995. They replaced the constant Hurst exponent with a Hölder function $h(t)$, with $0 < h(t) < 1$, thus allowing the long memory to vary in time. This allows the model to accept non stationary continuous processes. Vehèl and Peltier defined this as the multifractional Brownian motion which generalizes the fractal Brownian motion with $t \, \epsilon [0, \infty)$. This expansion of the RLfBm is defined by:

$$B_t^h := \frac{1}{\Gamma(h_t + \frac{1}{2})}\int_0^t (t-s)^{h_t - \frac{1}{2}} dB_s$$

This model is useful in various instances such as asset price, internet traffic and geological modelling. Estimating the Hurst index for such examples, one could often find that the Hurst is changing with time in the dataset. Thus, the extension of RLfBm above seems appropriate.

1.2 Results

The estimated Hurst exponent using the R/S analysis on the time series containing gold price returns in the period 1979 to 2018 is 0.83 for monthly returns, and 0.63 using daily returns (See table 1). With a $H > \frac{1}{2}$ in both instances, they have the property of persistence, where the past trend is more likely to sustain in the future. This is an evidence of long memory (Graves et al., 2017). Figure 1 shows the movement of the daily recorded gold price over the period and is consistent with a somewhat persistence time series. This result complies with the findings in other studies on the Hurst exponent of the gold price (Shaikh et al. 2014, Caporale et al. 2017, Uthayakumar and Jayalalitha 2018). We also observe that the Hurst exponent of the monthly data is higher than the one of the daily data. This finding corresponds with research conducted by Caporale et al. (2017), which shows that lower frequency data in general displays higher persistence.

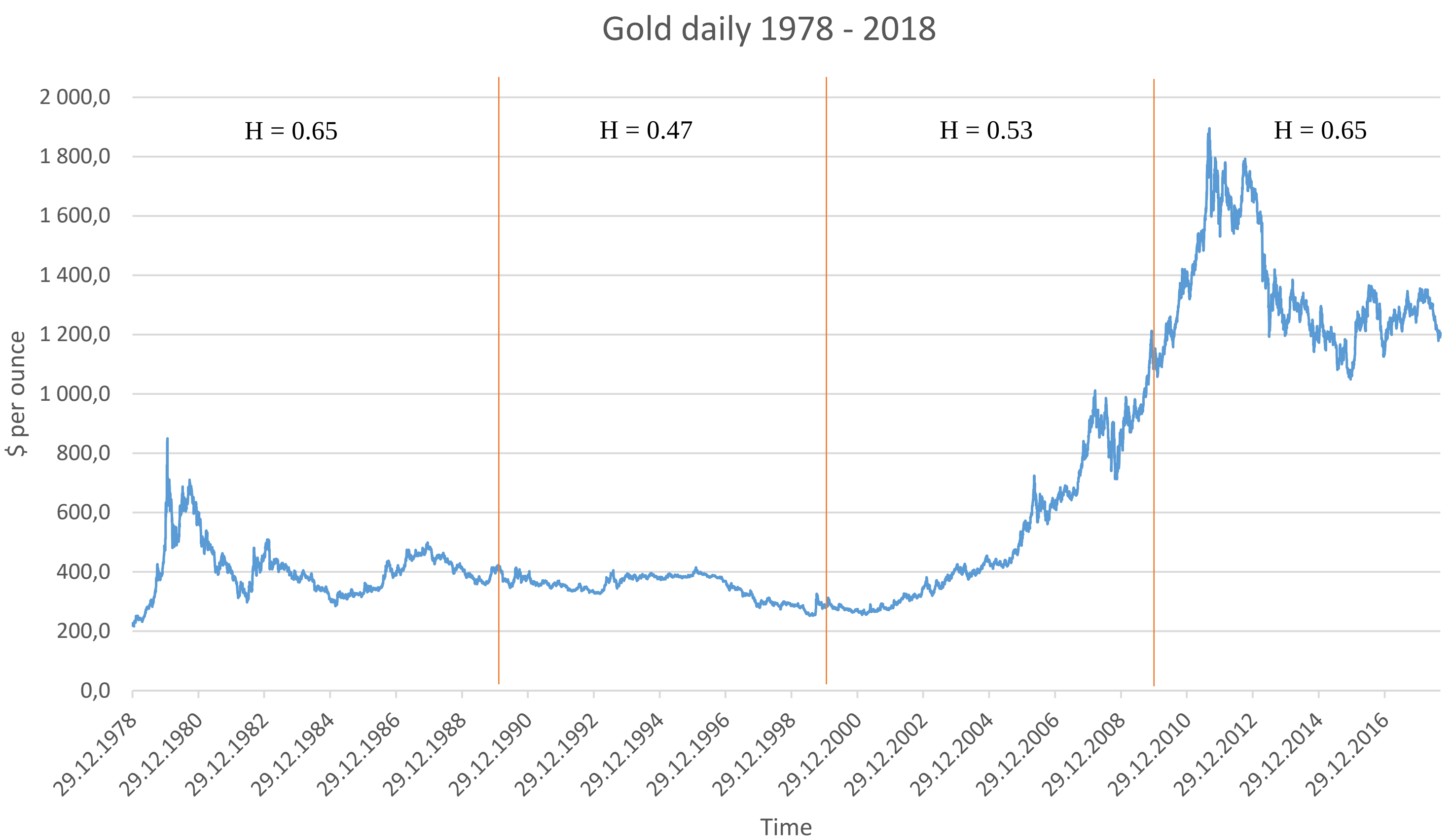
At the same time, it is worth mentioning that the Hurst exponents 0.63 and 0.83 are somewhat higher than what other studies have uncovered when examining other financial data such as stocks, currencies or indexes (Voss 2013 and Caporale et al. 2017). Considering the previously mentioned fact that many studies have doubted the presence of long memory properties in financial data, it is important to look at the results with a critical view.

The Hurst exponent calculation of the daily data implies a fractal dimension of 1.37, which corresponds to a fairly smooth graph. For the monthly data, the fractal dimension of 1.17 is closer to 1, implying an even smoother graph.

In our results, the Hurst exponent differs across the time series, as displayed in table 1. The Hurst exponent is close to $\frac{1}{2}$ in the two middle periods, from December 1988 to November 1988, and from November 1988 to October 2008. For the first period of our dataset, December 1978 to December 1988, as well as the last, October 2008 to September 2018, the Hurst exponent is slightly higher at 0.65 for both. The change of persistence over time is a phenomenon discussed by Corazza and Malliaris (2002) and Glenn (2007) among others. This change in persistence over time could be a source to the different findings and opinions regarding long memory in financial data in previous studies. For further research, the previously presented concept of multifractional Brownian motion could be used to examine how the Hurst vary across time through the dynamic Hurst.

Table 1: Results of the R/S Analysis

	Full series (monthly)	Full series (daily)	Oct 2008 – Sep 2018	Nov 1998 – Oct 2008	Dec 1988 – Nov 1998	Dec 1978 – Dec 1988
Intercept	-0.55	-0.17	-0.33	-0.06	0.15	-0.31
Hurst exponent	0.83	0.63	0.65	0.53	0.47	0.65
R-squared	0.99	1.00	0.96	0.99	0.98	0.99
Fractal dimension	1.17	1.37	1.35	1.47	1.53	1.35



Source: World Gold Council (2018)

Figure 1: The gold price with the estimated Hurst exponent

2. Variable importance in a prediction of the gold price

Numerous research have investigated the relationship between the gold price and different determinants; economic, financial or the price of other commodities. The relationship between inflation and gold has been widely discussed, where studies by Worthington and Pahlavani (2007), Wang et al. (2011) and Beckman and Czudaj (2013) all indicate that holding gold protects an investor against changes in the inflation rate. Furthermore, the gold's function as a "safe haven" against stocks in economic and financial crises has been examined in numerous researches, such as Hillier, Draper and Faff (2006), Baur and McDermott (2010) and Baur and Lucey (2010). Melvin & Sultan (1990), Christie-David et al. (2000) and Cai et al. (2001) among others have examined the link between movements in the gold price and factors such as interest rates, commodity prices and stance in the business cycle, while Tully & Lucey (2007), Sjaastad (2008) and Reboredo (2013) have investigated the relation to exchange rates. All the above mentioned factors are in the different studies considered to varying degrees to be determinants of the gold price. To further examine these relationships, this thesis will assess which variables that are important in a prediction of the gold price. Previous studies to a large extent use the above mentioned variables when forecasting the gold price (Ismail et al. 2009, Shafiee & Topal 2010). We will mostly focus on the demand side, as the supply of gold is relatively inelastic. Today, there are still new gold mines being discovered, but the supply will be adjusted by incline in older mines (Sykora, 2010).

The economy has largely been shaped by the financial crisis of 2008, and the variables influencing the gold price could have been affected by the aftermath of the crisis. The gold price rose above $1000 for the first time in 2009, and has stayed above since (Hale, 2017). After the financial crisis, the US economy has experienced significant growth, as corporate profits are at record high while the stock market has quadrupled since the recession in 2009 (DePillis, 2018). Relationships between the gold price and factors examined by Melvin & Sultan, Sjaastad and Tully & Lucey and others could be different today as an effect of the financial crisis. Therefore, it could be interesting to provide an updated view on the different variables, and to see if the results are in line with findings from previous research.

2.1 Methodology: Gradient Boosting Machine and XGBoost

For the examination of variable importance in a prediction of the gold price, various supervised machine learning techniques will be applied. Supervised machine learning is training a learning function on a data set where both the input and output values are known to the learning function. The learner will then use its generalized knowledge from the training data set to unseen and new datasets (Castle, 2017).

By utilizing previously unused methods in an otherwise well-researched field of gold and the gold price behavior, we hope to bring a new perspective. For this thesis we have chosen to use two models within the same branch of supervised learning, Gradient Boosting Machine (GBM) and XGBoost. They are both ensemble methods, which are techniques that combine several base models in order to produce one optimal predictive model (Lutis, 2017).

In the process of building a prediction, the learning function can display the most important variables for the predictive function. Consequently, the methods can display which variables that are the most important in a prediction of the gold price. One important reason for choosing GBM and XGboost is their ability to accept more variables as inputs than alternatives within ordinary econometrics would allow, as well as being robust to correlations between the independent variables (Chen et al, 2018). Especially this robustness is beneficial, as many of the relevant input variables potentially are correlated. In addition, as GBM and XGBoost in this case are based on decision trees, no variable scaling or normalization is necessary. By running two quite similar models we enable the opportunity to cross-check and verify the results achieved.

The following sections will take a further look at the models utilized in this thesis. Firstly, in section 2.1.1, GBM and XGBoost will be introduced in a general manner. Then, section 2.1.2 will break down the models and present the basis of the models, by describing decision trees and boosting. Section 2.1.3 and 2.1.4 present the algorithms of GBM and XGBoost, primarily with reference to the developers of the models, Friedman and Chen respectively, as well as

Nielsen (2016). Lastly, in section 2.1.5, we will introduce how these model display the relative importance of the variables used in the prediction, which is the overall aim of this thesis.

2.1.1 An introduction to the methods

Gradient Boosting Machine is based on an algorithm built by J. H. Friedman in 1999, while XGBoost is an extension of this algorithm developed by Tianqi Chen in 2014 (Chen, 2015). As other ensemble methods, GBM and XGBoost combines relatively weak and simple base learners to one stronger ensemble prediction. As boosting methods, GBM and XGBoost adds new learners to the ensemble sequentially (Natekin & Knoll, 2013). These are constructed to be maximally correlated with the negative gradient of a loss function, associated with the whole ensemble. The choice of loss function is up to the researcher, which makes GBM and XGBoost highly customizable to any particular data-driven task (Natekin & Knoll, 2013). The flexibility makes it attractive to use in the investigation of the properties of the gold price.

Especially XGBoost has been popular in recent years (Brownlee, 2016b). Some of the main strengths relative to other machine learning techniques is its ability to deal with missing values and cross-validation (Chen, 2016). Cross-validation is a resampling method typically used to evaluate a machine learning model on a limited data sample, which is the case in this thesis. Cross-validation is highly recognized for its ability to generate less biased and less optimistic results (Brownlee, 2018).

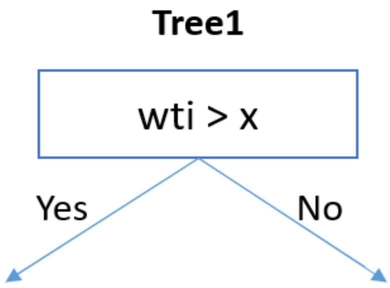
In the world of data science and machine learning, deep learning models are being aggrandized, but they are usually outperformed by Gradient Boosting methods in a majority of applications, especially in the general business domains (Pafka, 2016). This makes our models attractive as tools in the investigations of gold price properties, especially the XGBoost as it is with its accuracy and flexibility currently one of the most popular boosting methods.

2.1.2 Breakdown of GBM and XGBoost

Decision trees

There are different algorithms within the field of decision tree learning. Classification and Regression Tree (CART) analysis is a term introduced by Breiman et al. (1984), which includes procedures for building decision trees for classification and regression problems. These decision tree algorithms are preferred by Friedman (1999) in GBM, and are also used in XGBoost (Chen & Guestrin, 2016). In CART analysis the learning tree is built in a top-down approach using binary splits. For each tree node the model considers all splits parallel to the coordinate axes, and selects the one minimizing the objective. This procedure is repeated until a stop criterion is reached. As the goal of this thesis is to investigate the properties of a continuous variable such as the gold price, regression tree is the type of decision tree generated in our models.

The following three figures are a stepwise example of how a binary regression tree could be built in our case. All values shown in the figures are fictional, but the figures are still a good representation of how a regression tree is built up. In contrast to the sections describing boosting, GBM or XGBoost, there is in this scenario only one tree, called Tree 1.



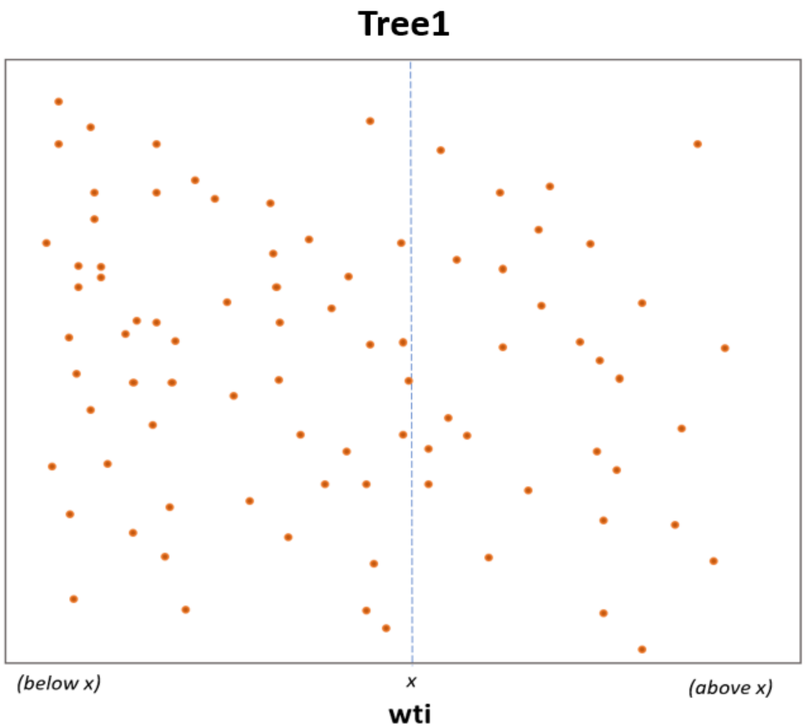


Figure 2: Step 1 of the regression tree example

The algorithm selects a feature and a splitting point on that feature based on what minimizes the expected value of a loss function. Here in step 1 displayed in figure 2, the algorithm has chosen *wti* (crude oil) as the most appropriate feature to partition the data with, and will further assess for which value of *wti* the observations should split in order to minimize loss. This partitioning is done in a greedy fashion, meaning that the model does not take the totality into consideration when conducting a split. On the left is tree 1 visualized as a binary regression tree, that currently only consists of a splitting point visualized with a square. This is called a node. To the right is tree 1 visualized differently, this time as a plot. This plot shows how the upcoming split will divide the observations.

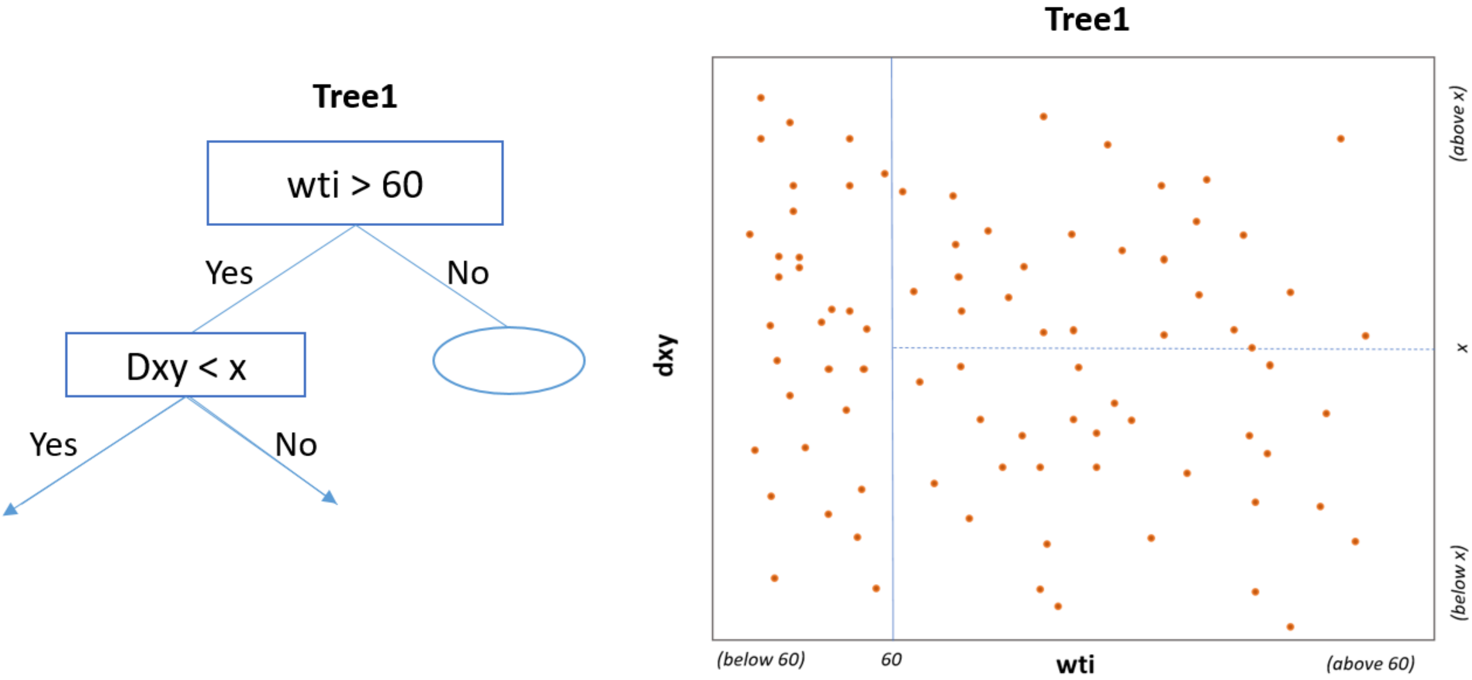


Figure 3: Step 2 of the regression tree example

From step 2 in figure 3 we see that the algorithm, based on the loss function, chose to divide the observations on whether the *wti* was above or below 60.

In the binary tree to the left in figure 3, the algorithm now is at the next, lower level in the tree, and again has to decide on the next split. For the observations with a *wti* lower than 60, the algorithm did not find it beneficial to do any further splitting. The circle thus represents a terminal node. On the other side does however the algorithm find a solution that further minimizes loss by doing another split. For this node, the process in step 1 repeats itself, but this time *dxy* (the dollar index) is found to be the best feature to partition the data. The plot to the right shows how the observations with a *wti* above 60 now will be further divided.

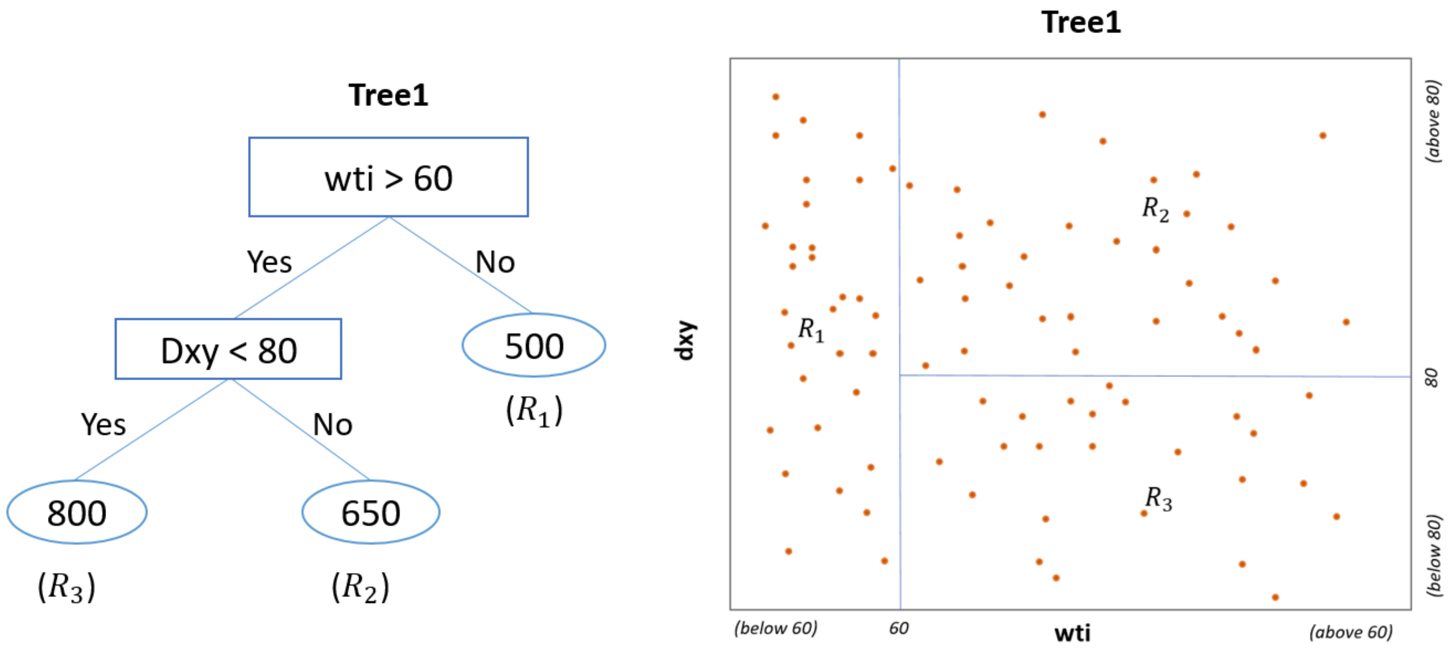


Figure 4: Step 3 of the regression tree example

Step 3 shown in figure 4 is the last step of this regression tree. Now the algorithm have concluded the splitting process, and does not find it beneficial to do any further splitting. The three resulting terminal points now formulate a prediction score for the observations within their region, which is the term used for the areas formed by the partitioning. The prediction score can be interpreted accordingly: If the value of *wti* is above 60, but the value of *dxy* is below 80, tree 1 imposes a gold price of 800 (dollars per ounce). Looking at the plot to the right, all observations with the corresponding characteristics would be placed in R_3.

The splitting decisions in figures 2-4 are fundamentally determined by a loss function, which is used by the algorithm to assess which split that achieves the largest gain. The loss function for a learning tree f with T terminal nodes gives:

$$L(f) = \sum_{j=1}^{T} \sum_{i \in I_j} L\big(y_i, w_j\big) = \sum_{j=1}^{T} L_j$$

Here, L_j represents the aggregated loss at node j. The gain of a split at node k, can be written as:

$$\begin{aligned} &Gain = L\big(f_{before}\big) - L\big(f_{after}\big) = \\ &\sum_{j \neq k} L_j + L_k \; - \; \sum_{j \neq k} L_j + L_{kL} + L_{kR} = \\ &L_k - (L_{kL} + L_{kR}) \end{aligned}$$

For an individual split, the split gain is calculated for every possible split for every possible feature and the model choses the one with the largest gain. Here, L_{kL} denotes the left node of the split and L_{kR} the right node. The weight for a region R_j is calculated by:

$$w_j = \arg min_w \sum_{i \in I_j} L(y_i, w)$$

The weight w_j indicates the relative importance of region R_j during a prediction. For example could the algorithm find region 1(R_1) in figure 4 relatively more influencing and emphasize this terminal node to a larger extent in the prediction.

A general tree model that splits observations into M regions $R_1, \dots, R_M$ can be written as:

$$f(x) = \sum_{m=1}^{M} w_j I(x \in R_m)$$

Again, w_j is the weight of its according region, and $I(x \in R_m)$ denotes the basis function for the decision tree. This indicator function indicates which region each specific observation is assigned to in figure 4.

Boosting

The idea of boosting is to combine many weak learners into a strong ensemble learner. To build on the previous regression tree example, boosting is the process of adding possibly countless of trees. In such a process, trees are introduced sequentially such that each subsequent tree has the objective to reduce the errors of the previous tree (Vidhya, 2018). Thus, a new tree in the sequence will learn from the residuals corresponding to the previous tree.

Turning back to our visual example, figure 5 is now expanded to include more trees than the regression tree example from figures 2-4, which represents the distinct characteristic of boosting methods described above: Growing multiple trees using information from previously grown trees. The trees are typically small, meaning that the model does not learn much from each tree. However, by growing many trees, the model will slowly improve, and by growing them sequentially, the model can focus on improving in areas where the performance previously was not adequate.

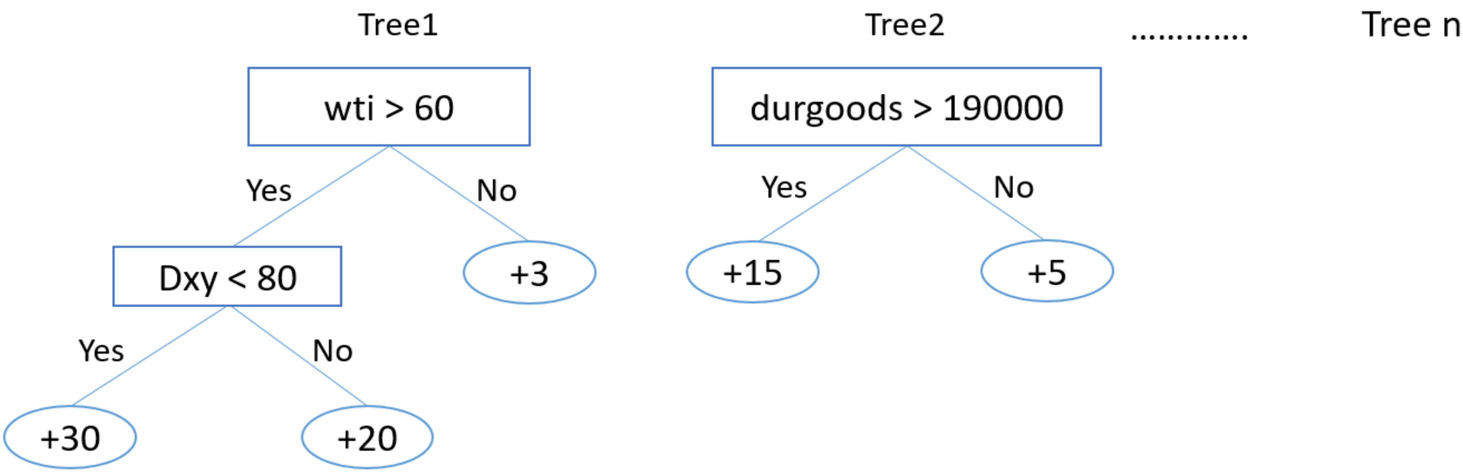


Figure 5: Boosting example

Tree 1 is followed by tree 2 sequentially, where tree 2 has the objective to reduce the errors of tree 1. This process continues all the way to tree n. Input factors such as how many trees the model builds, the number of splits per tree and learning rate, will be further described in section 2.1.3 and 2.1.4.

The interpretation of the terminal nodes in figure 5 is somewhat different from figures 2-4, as we now have grown many small trees. The prediction is now a sum of the scores from the whole ensemble of trees. Using *wti* as an example: Among all the trees, several will include *wti* as a decisive variable. The predicted gold price in situations where the *wti* is above 60, and the *dxy* is below 80, will be the sum of all prediction scores from the all generated trees. In this equation will tree 1 contribute with a positive change of 30.

The following paragraphs will describe boosting in a general manner, based on Friedman (2002) and Nielsen (2016). As previously mentioned, boosting is a type of ensemble model, and can be written as the following:

$$f(x) = \sum_{m=0}^{M} f_m(x)$$

This means that boosting uses multiple base learners in order to create improved predictive performance (which is trees in our case). The expression above can be expanded as adaptive basis function models:

$$f_m(x) = \theta_0 + \sum_{m}^{M} \theta_m \, \phi_m(x)$$

Here, $\theta_0 = f_0(x)$ is the initial starting point of the ensemble method, before further models are added. For each iteration $m = 1, \dots, M$ the ensemble function is given by $f_m(x) = \theta_m \phi_m(x)$, where θ_m is the weight given to a base learner $\phi_m(x)$. This base learner is a general notation of the basis function $I(x \in R_m)$ from the previous section. Boosting functions by using the base learner $\phi_m(x)$ to sequentially add further models that improve the prediction of the current model. The choice of base learner is flexible for the general boosting method, and in this thesis we opt to use previously elaborated regression tree as the base learner for both the GBM and XGBoost methods.

The boosting approach will solve the following general model:

$$(\theta_m, \phi_m) = \arg min_{\theta_m, \phi_m} \sum_{i=1}^{N} L(y_i, f^{(m-1)}(x_i) + \theta_m \phi_m(x_i))$$

In this process, the function $f^{(m-1)}$ is the current estimation, the product $\theta_m \phi_m(x_i)$ can be viewed as the optimal step toward the data-based estimate of the boosting function. One optimal step takes the current estimation, and use the base learner with according weight, to minimize the given loss function.

2.1.3 Gradient Boosting Machine (GBM)

Below is the algorithm used in the gradient boosting package in R. Some of the steps uses the same principles as the more general tree- and boosting functions previously displayed. A further description of the steps follows after the presentation of the algorithm.

Algorithm 1: Gradient boosting machine

Input: Data set D.

A loss function $L(y, f_m(x))$.

The number of iterations M.

The learning rate v.

The number of terminal nodes T.

1. Initialize model with a constant value:

$$f^{(0)}(x) = f_0(x) = \theta_0 = argmin \sum_{i=1}^{n} L(y_i, \theta)\,;$$

2. For $m = 1,2, \dots, M$:

3. $g_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f^{(m-1)}(x)}$;

4. Determine the structure $\{R_{jm}\}_{j=1}^{T}$ by selecting splits which maximize

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} + \frac{G_{jm}^2}{n_{jm}}\right];$$

5. Determine leaf weights $\{w_{jm}\}_{j=1}^{T}$ for the learnt structure:

$$w_{jm} = argmin_{wj} \sum_{i \in I_{jm}} L(y_i, f^{(m-1)}(x_i) + w_j);$$

6. $f_m(x) = v \sum_{j=1}^{T} w_{jm} I(x_i \epsilon R_{jm})\,;$

7. $f^{(m)}(x) = f^{(m-1)}(x) + f_m(x);$

Output: $f(x) = f^{(M)}(x) = \sum_{m=0}^{M} f_m(x)$

The goal of the gradient boosting machine is to find an approximation of $f(x)$ that minimizes the expected value of a chosen loss function $L(y, f_m(x))$. In this thesis, we opt to use a loss function of squared error:

$$\frac{1}{2}(y_i - f_m(x_i))^2$$

The data set D contains observations of all explanatory variables x and the dependent variable, y. Further, another input in the algorithm is the number of iterations M. This parameter represents the number of regression trees included in the model. The learning rate v is a shrinkage factor that modifies the learning of the model. Next, the number of terminal nodes T is the maximum interaction between the variables in the model. It represents the maximum depth of regression trees in the model.

Further, the model uses the following steps for each iteration M to find the most fitting output for the model.

Step 3 represents a forward stage-wise additive model by implementing gradient descent in a nonparametric function space. The best steepest-descent step direction is given by the data-based unconstrained negative gradient of the loss function. The step takes the derivative of the loss function with respect to $f(x_i)$. This ultimately means that at each iteration, a regression tree is fitted to predict the negative gradient, which provides information about the direction of the step:

$$-g_m(x_i) = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

The next step is to find the optimal structure for $\{R_{jm}\}_{j=1}^{T}$, which implies to splitting regions $R_1, \dots, R_M$. Here, I_{jm} represents the set of indices of $x_i$that falls in region R_{jm}. The optimal splitting of regions is found by selecting the splits that maximize gain, expressed by:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} + \frac{G_{jm}^2}{n_{jm}}\right];$$

Now that the structure of the trees are set, step 5 will learn the final leaf weights of the model. The weights $w_1, \dots, w_T$ are what makes it possible for the complexity of the tree model by adjusting the weight assigned to each individual region. The optimal weights are found by minimizing the loss function using the previously learned structure. This is given by:

$$w_{jm} = argmin_{wj} \sum_{i \in I_{jm}} L(y_i, f^{(m-1)}(x_i) + w_j)$$

Step 6 introduces the model to the learning rate, v, also called the shrinkage parameter. It was introduced by Friedman as a regularization of the model to enhance the model performance, and to deal with the typical problem of overfitting, which we will return to in section 2.2.2. The objective of this parameter is to scale each value computed by the model. The shrinkage factor varies between 0 and 1, and controls the degree of fit. In other words, it adjusts the learning from the trees using the optimal weights from the previous step. Now, each step taken at each iteration m can be written:

$$f_m(x) = v \sum_{j=1}^{T} w_{jm} I(x_i \epsilon R_{jm})$$

The function of the next step is to update the model by using the optimal learning tree structure and optimal weights.

$$f^{(m)}(x) = f^{(m-1)}(x) + v \sum_{j=1}^{T} w_{jm} I(x_i \epsilon R_{jm})$$

$$f^{(m)}(x) = f^{(m-1)}(x) + f_m(x)$$

2.1.4 XGBoost

Below is the algorithm used in the XGBoost in R.

Algorithm 2: XGBoost

Input: Data set D.

A loss function $L(y, f_m(x))$.

The number of iterations M.

The learning rate v.

The number of terminal nodes T.

1. Initialize model with a constant value:

$$f^{(0)}(x) = f_0(x) = \theta_0 = argmin \sum_{i=1}^{n} L(y_i, \theta)\,;$$

2. For $m = 1,2, \dots, M$:

3. $g_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f^{(m-1)}(x)}$;

4. $h_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=f^{(m-1)}(x)}$

5. Determine the structure $\{R_{jm}\}_{j=1}^{T}$ by selecting splits which maximize

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} + \frac{G_{jm}^2}{H_{jm}}\right];$$

6. Determine leaf weights $\{w_{jm}\}_{j=1}^{T}$ for the learnt structure:

$$w_{jm} = -\frac{G_{jm}^2}{H_{jm}};$$

7. $f_m(x) = v \sum_{j=1}^{T} w_{jm} I(x_i \epsilon R_{jm})\,;$

8. $f^{(m)}(x) = f^{(m-1)}(x) + f_m(x)$;

Output: $f(x) = f^{(M)}(x) = \sum_{m=0}^{M} f_m(x)$

The algorithm for XGBoost is very similar to the one from GBM shown in algorithm 1, but there are a few key differences. XGBoost is based on Newton Boosting, which introduces a new term, the Hessian. The inputs and initial point is identical for the XGBoost and the GBM model. One difference between the two methods is that where GBM divides the optimization problem into two parts by first determining the structure and then finding optimal final leaf weights, XGBoost find optimal structure and leaf weight simultaneously. This means that in XGBoost the final leaf weights are the identical weights learnt when finding the tree structure.

Because the two algorithms are similar to such an extent, we will not elaborate each step of the algorithm of XGBoost. However, the steps that differ from GBM will be further examined.

Step 4 in the XGBoost algorithm differs from GBM in the way that it finds the Hessian $h_m(x_i)$ at a current estimate. The Hessian is a term representing the second order derivative at the current estimate, and is given by:

$$h_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=f^{(m-1)}(x)}$$

When learning the optimal structure $\{R_{jm}\}_{j=1}^{T}$ the model seeks to find the optimal splitting of regions in addition to learn the optimal weights of those regions. This is achieved by expanding the loss function as:

$$L(f^{(m)}) = \sum_{j=1}^{T_m} \sum_{i \in R_{jm}} \left[g_m(x_i)w_{jm} + \frac{1}{2}h_m(x_i)w_{jm}^2\right]$$

Calling the sum of gradient in region j for G_{jm}, and the sum of hessian in region j for H_{jm}, the equation can be rewritten as:

$$L(f^{(m)}) = \sum_{j=1}^{T_m} \left[G_{jm}w_{jm} + \frac{1}{2}H_{jm}w_{jm}^2\right]$$

Now, optimal structure is found by maximizing the expression for gain in step 5. The expression is similar to the corresponding version in the GBM algorithm, but in this case the sum of Hessian is used in addition to the sum of gradient. The expression can be written as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} + \frac{G_{jm}^2}{H_{jm}}\right]$$

Further, step 6 differs from algorithm 1 as the optimal leaf weight is already found during the learning of optimal structure. The optimal weight is a fairly simple term using the sum of gradient and sum of Hessian:

$$w_{jm} = -\frac{G_{jm}^2}{H_{jm}}$$

2.1.5 Relative importance of input variables

The aim of this thesis is to investigate the relative importance of variables when predicting the gold price. This is possible to do with GBM and XGBoost due to their ability to interpret the derived approximation f(x). This includes receiving information of which of the input variables that are most influential in contributing to its variation.

The relative influence I_j, is a measure of the relative influence of the individual inputs x_j on the variation of f(x). It can be written as the following:

$$VI_j = \left(E_x\left[\frac{\partial \hat{f}(x)}{\partial x_j}\right]^2 var_x[x_j]\right)^{\frac{1}{2}}$$

This expression finds the relative importance by calculating the improvement or gain of the total model from the derivative of the function with respect to an input variable. The variable importance score is calculated by the total improvements of the total amount of splits associated with a given variable across the whole ensemble of trees in the model.

When requesting the information on variable importance from R, the output generated is feature importance plots. These plots present the variable importance as bar graphs. The graph presents each included feature as a horizon bar, in which the length of a bar indicates the relative importance of a feature. The features are presented ranked in order of decreasing relative importance.

Further does the XGBoost provide more information about the variables and their interactions with the model, in the form of a table. It displays descriptive attributes such as gain, cover and frequency, which give further insight in the variable importance. The gain is the fractional contribution of each variable based on the total model gain from the splits where this variable is represented. A higher value indicates a relatively more important predictive feature. Cover metric represents the relative number of times a feature is used to split the data across all trees weighted by the number of observations that is related to those splits. To illustrate, a model could include 1000 observations, five variables and three trees. If variable 1 is used to decide the node for 50, 25, and 10 observations in the trees, then the metric cover for this feature would be 85 observations. The frequency percentage is the relative number of times a particular feature or variable occurs in the trees of the model. In the example above, if variable 1 occurs in two splits in each of the three different trees, then the weightage for variable 1 would be six. Subsequently the frequency for this variable is the percentage weight over weights of all variables.

2.2 Data set and data set preparations

This section describes the inputs of the models. Firstly, will the content of data set D be described, and the necessary actions conducted on order to prepare this data set to be a sufficient input. Secondly will section 2.2.2 deal with the other inputs described in the methodology through the parameter tuning.

The relevant data containing the gold price and the relevant variables are collected from various sources, including Bloomberg database, the World Gold Council, and United States Census Bureau. They are presented in table 2.

Table 2: Input variables

Variable	R-abbreviation	Further information
The gold price	goldsupply	The price per ounce of gold in dollars.
US inflation	usinflation	US Inflation measured by the consumer price index (CPI), including food and energy.
FED Funds Rate, effective	FF_eff	The weighted average rate of which borrowing banks pays to the lending banks to borrow funds.
10-year Treasury Yield	TCMNOM_Y10	The constant maturity yield of the US 10-Year Treasury Note.
US Durable Goods	durgoods	A good that yields utility for more than three years, such as cars, home furniture and electrics, medical equipment and airplanes.
Unemployment rate	unemployment	Unemployment Rate Total in Labor Force Seasonally Adjusted (USURTOT)
S&P500	s.p500	Stock market index containing 500 large companies listed on the New York Stock Exchange (NYSE) and NASDAQ. The index is a capitalization-weighted index.
Dow Jones Industrial Average	dji	Stock market index containing 30 large, publicly owned companies on NYSE and NASDAQ. It is weighted differently than the S&P 500: It is an arithmetic mean of the sum of the price of stock for each component company.
Nikkei225	N225	Stock market index for the Tokyo Stock Exchange, Japan.

Hang Seng	hsi	Stock market index for the Hong Kong Stock Exchange, China.
Shanghai Composite	shcomp	Stock market index for the Shanghai Stock Exchange, China.
The US Dollar Index	dxy	The value of the US dollar relative to a basket of foreign currencies. It is a weighted geometric mean of the dollar's value relative to: Euro (EUR): 57.6% Japanese yen (JPY): 13.6% Pound sterling (GBP): 11.9% Canadian dollar (CAD): 9.1% Swedish krona (SEK): 4.2% Swiss franc (CHF): 3.6%
USD/CNY	usdcny	Conversion rate between the US dollar and the Chinese Yuan Renminbi.
MSCI World Index	mswi	Large and mid-cap representations from 23 developed markets countries.
The CBOE Volatility Index (VIX)	vix	Popularly referred to as the fear index. It formulates a theoretical expectation of the stock market volatility in the near future by quoting the expected annualized change in the S&P 500 index over the following 30 days.
Crude Oil	wti	The West Texas Intermediate (WTI), Texas light sweet.
Gold supply	goldsupply	World Gold Council.

The time period examined is 1999 to the end of 2007 (January 1st, 1999 to September 7th, 2018).

As mentioned earlier was the gold price linked to the US dollar through the Bretton-Woods system up to the 1970s. This, along with the availability of sufficient data, makes any earlier starting point before the 1980s inadequate. The gold price is quoted in numerous currencies, and is in this thesis quoted in US dollar, as the major price discovery hubs of London and New York use the US dollar as the trading currency (Manly, 2018). Another reason for listing the gold price in US dollars is that it will exclude exchange rate effects which would occur by using a different currency.

Most of the macroeconomic variables included in the data set are based in the US rather than globally. This is the case for variables such as US inflation, 10-year treasury yield, unemployment rate and durable goods. One reason for this selection is that the US dollar is usually used as a benchmark for the gold price, which is the case for our data set as well. Further, the US is a global powerhouse and economic tendencies in the country is a good benchmark to represent the world. In addition, data from the US is both easily available and well documented for the time period chosen in the data set.

One of the main interests of this thesis is to examine any change in the variable importance in the creation of a predictive model after the 2008 financial crisis. The period examined is for this reason divided in three sub periods: Before, during, and after the 2008 financial crisis. For the time definition of the duration of the crisis, the Business Cycle Dating Committee of the National Bureau of Economic Research' (NBER) definition is used. They have stated January 1. 2008 as the starting point, and June 31. 2009 as the end of the recession (NBER, 2010). The inter-crisis dataset is thus the period between these dates.

The after-crisis period analyzed for this question is from July 1. 2009 to September 7. 2018, which was the date all the data was downloaded. The before-crisis period selected with the intention of having equally sized data sets, thus this period starts January 1. 1999 and ends December 31. 2007.

2.2.1 Interpolation

Financial and economic data comes in different formats and frequencies. While for example stock indices are calculated or tracked in real time, macroeconomic variables such as the inflation rate are released monthly or quarterly. Such indicators also are to larger extent estimations that require time to calculate, thus resulting in a considerable publication lag.

GBM and XGBoost require equal length of columns in the data frame (Chen et al. 2018). In other words the variables should have the same frequency, where daily data is preferred in this case. The choice of daily data instead of monthly is a choice balanced between the desire of having a sufficient number of observations, as well as studies favoring daily data when assessing the econometric trade-off between the two frequencies (Morse, 1984).

Interpolation is the operation of estimating new data points within a collection of known data points (Friedman, 1962). Interpolation methods range from linear interpolation to more advanced methods such as polynomial and spline interpolation. The latter is more fitting in this case, more specifically cubic splines, which together with polynomial interpolation incurs less error than linear interpolation (Moosa & Burns, 2013). The disadvantage of using polynomial interpolation lies with its issues when the dataset is large enough to require a higher order polynomial for the interpolation. In this case, the interpolation may, especially at the edges of the interval, exhibit oscillatory artifacts (Kvarving, 2008). This problem is referred to as Runge's phenomenon (Fornberg & Zuev, 2007). Cubic spline interpolation avoids this phenomenon by interpolating piecewise with lower degree polynomials, instead of using one, global, higher order polynomial (Kvarving, 2008).

As the relevant variables to be interpolated in this case contains at least 450 data points, a piecewise approach seems rational. Furthermore, it is worth reflecting on the relatively stable nature of the macroeconomic variables as for example the inflation rate in a developed country such as the US in the late 20th century and early 21th century. The price level of a basket of goods and services will not fluctuate significantly between the monthly observations.

The interpolated variables used in this thesis are durable goods, inflation, unemployment and gold supply. A graphical output of the interpolation process can be found in the appendix (A1).

2.2.2 Model Implentation

In order to achieve the most accurate variable importance plots, the fundamental prediction model should be as accurate as possible.

Implementing the GBM and XGBoost models are in itself not too challenging, but applying them with precision is more difficult. As described in section 2.1.2, the models need inputs such as learning rate and number of iterations along with the input data. The process of adjusting these parameters are called parameter tuning. The tuning of model parameters is challenging, and is as many other concepts within machine learning about the bias- variance trade-off (Chen et al. 2015). A biased model is oversimplified, where the difference between the average prediction of our model and the actual value we are trying to predict is high. Variance is the variability of model prediction for a given value, which indicates the spread of our data (Singh, 2018). A model with high variance is performing well on training data, but has a high error rate on test data.

After subdividing the dataset into three different time periods, each subset were further split in two. One part will serve as a train set, and the other as a test set. The train set will be used to fit the models, and the test set is for assessing the final prediction. The split was set to 0.75 and 0.25 using random sampling. Random sampling was used foremost to prevent the possible biases that can arise when using ordered data (Mallick, 2017). Furthermore, our aim is to compare the periods as a whole, and pooling the observations does not obstruct this.

No preprocessing such as scaling or normalization is conducted on the data set. This is due to the fact that XGBoost and GBM with tree based learners as used in this cased not are sensitive to monotonic transformations. The models pick adequate splitting points in the variables to split a tree node, where defining a split on one scale has a corresponding split on a transformed scale (Chen, 2014).

The following paragraphs describe the process of tuning the model parameters to achieve the most precise prediction models possible, in order to obtain an accurate overview of variable importance.

XGboosting

The XGBoost package were used in creating the XGBoost model. This package requires a matrix instead of a data frame as input, thus the first step was to convert the training and testing sets into matrices. The process of parameter tuning has no standardized tutorial, and is not well described in literature. The following process is a result of general proposed procedures from various sources, as well as troubleshooting.

Tong (2016), Saraswat (2017) among others suggest default values as a starting point in the parameter tuning process. Further, they point to the number of iterations (nround) and the learning rate (eta) as a good next step. Chen et al. (2015) describes the important learning rate more thoroughly as the step sized shrinkage used in each update to prevent overfitting, shrinking the feature weights to make the process more conservative.

One of the mentioned strengths with the XGBoost model is the opportunity to customize evaluation criteria. This is the loss function described in section 2.1.2 and 2.1.4. The default metric is the root-mean-square-error (RMSE). RMSE is one of the most popular evaluation criterions for continuous variables (Swalin, 2018), which makes it unnecessary to change from this default configuration.

The model with default-only parameters achieved a RMSE that differed significantly between the training and test set. A considerable difference in accuracy between the training and test set suggests that the model has issues with overfitting (Brownlee, 2016a). The further tuning of the parameters is characterized by the effort of making the model more conservative and balance the algorithm's learning on the training set, to deal with the issue of overfitting. Furthermore, the first tuning of parameters did reveal the difference between the three time periods, which makes sense given the difference in number of observations, especially in the case of the shorter data set that represents the period of the 2008 financial crisis.

Some tools exist to support the tuning of parameters. One of them is the xgb.cv function, which performs cross-validation at each iteration, returning the optimum numbers of trees (nrounds) for a given learning rate (eta) (Aarshay, 2016). The optimum number of trees is partly determined by the parameter early_stopping_rounds, which will make the *xgb.cv* function stop running when the performance does not improves after the given number of iterations. This will help prevent overfitting. In this case, with early_stopping_rounds equal to eight, the *xgb.cv* function returns an optimum number of trees of 160 for the time period prior to the

2008 financial crisis, 180 for the period after and 120 for the period in between.

The learning rate turned out to be too high when running the model with default values, and a possible source to the overfitting. Zhang's (2010) suggestion on the relationship between the number of trees and the learning rate is to divide the number two by the number of trees. This will in our case result in a considerably lower eta, which will cause the model to be more conservative and less prone to overfitting. After dividing the number two with the optimal number of trees found with the *xgb.cv* function, the learning rate was given only limited adjustments. The eta was set to 0.015 for both the two periods before and after the 2008 financial crisis, and 0.025 for the period in between.

The remaining parameters are to a large extent left at the default setting, or very close to default, as the default values in many cases provide sufficient accuracy (Aarshay 2016, Saraswat 2017). This includes the last mentioned input in section 2.1.4, the number of terminal nodes. The variance importance plot is obtained after the tuning process is completed, using the *xgb.importance* function.

Gradient Boosting

The gbm package was used when building the GBM model. As with the XGBoost, the literature does not provide extensive support in the parameter tuning process, even though there exist more information about the technique in general than the more recent developed XGBoost. As with the XGBoost, the learning rate and number of trees are two central parameters in the tuning process, this time labeled shrinkage and iterations. This is the regularization parameters, mentioned in step 6 of the GBM walkthrough in section 2.1.3.

Friedman (1999) suggests to initially set a large number of trees, and then tune the learning rate, shrinkage, to achieve the best results. However, other sources propose starting with a fixed learning rate, and then calibrate the optimal number of trees (Hastie et al. 2001). With the *gbm.perf* function available, we have tuned the two parameters in the order suggested by Hastie et al. (2001).

The *gbm.perf* function estimates the optimal number of boosting iterations for a gbm object, with an already given learning rate. The learning rate was set quite low to begin with, considerably lower than the default of 0.1. A lower learning rate will in general give better prediction performance, at the cost of computational time (Ridgeway, 2007). The size of our dataset is not too comprehensive, meaning that computation time is not of concern. With *gbm.perf*, followed by some fine tuning adjustments, the shrinkage and iterations for the periods were 0.0020 and 500 for before the 2008 financial crisis, 0.0010 and 250 during the crisis, and 0.0015 and 500 after the crisis.

The GBM model contains a built-in cross-validation function. Again, as the computational time is of no issue, a more extensive 7-fold cross-validation was applied, as suggested by Ridgeway (2007). The interaction depth parameter is the parameter adjusting the number of terminal nodes. In the GBM model this parameter is set to three in order to capture some variable interaction.

2.3 Empirical results

This section will present the results. Firstly, the variable importance plots produced by the GBM and XGBoost models will be presented, with a short description to each of the plots. Secondly, all features will be thoroughly discussed individually to understand their importance for the gold price and why they are included in the models. Thirdly, the results will be used to analyze the different periods, where the overall similarities and differences will be assessed. Fourthly, the performance of the models will be evaluated in order to understand the prediction abilities of the basket of variables. Lastly, limitations of the models will be assessed. Appendix section A2 and A3 contain graphs and correlation coefficients relevant to the discussion.

2.3.1 Variable importance

Before the 2008 financial crisis

Figure 6 and 7 present the importance plots for the period before the financial crisis. The first plot represents the feature importance from the XGBoost model. It shows the 15 variables with an impact on the prediction, and indicates that *wti* and *durgoods* are the features with the highest degree of relative influence. The GBM model generates a similar importance plot to the XGBoost, illustrated by the fact that the four most important variables are in the identical order for both models. For the features with a low relative influence, the importance order vary more between the two models.

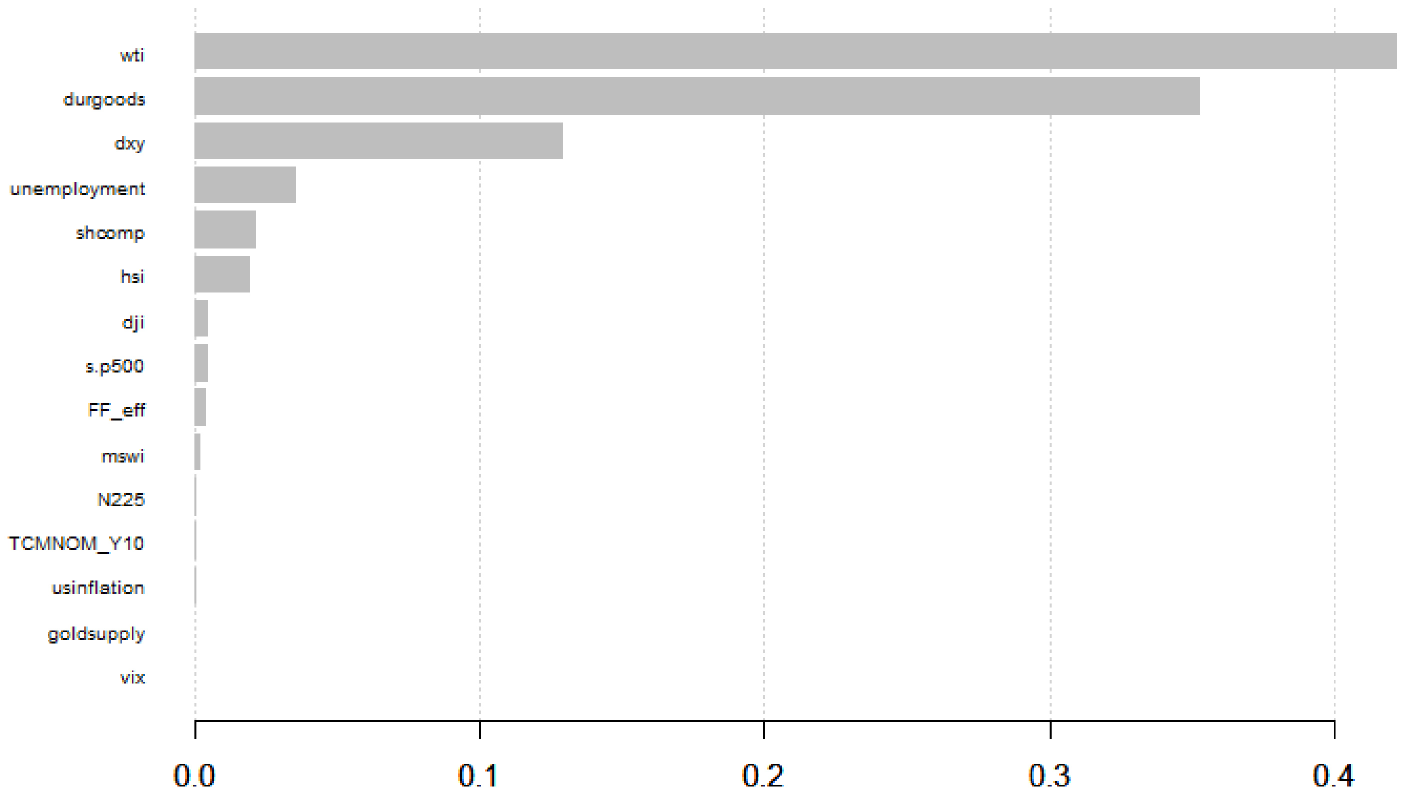


Figure 6: Variable importance plot. XGBoost, before the 2008 financial crisis.

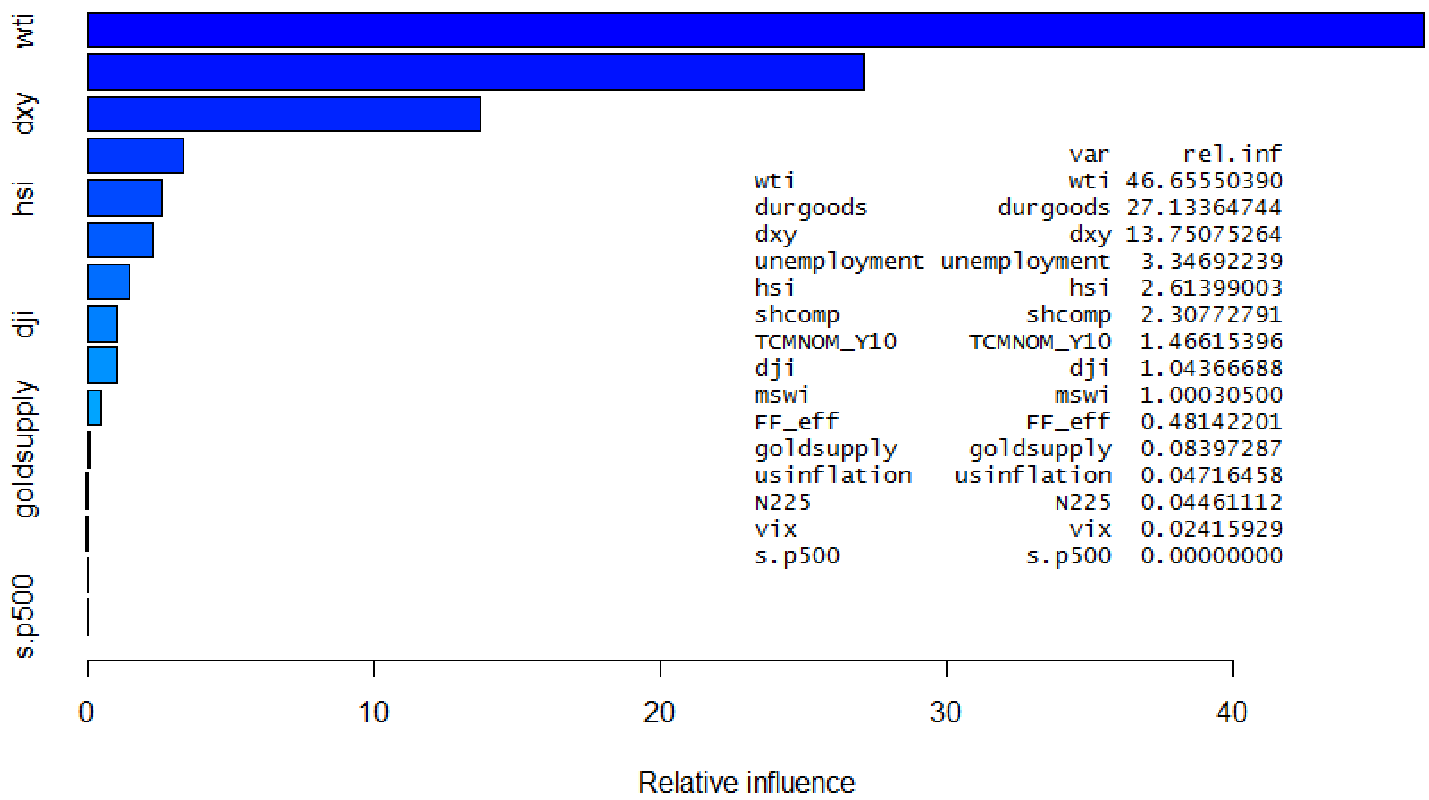


Figure 7: Variable importance plot. GBM, before the 2008 financial crisis.

Further, table 3 provides additional information about the variable importance in the XGB model for the period before the financial crisis. The gain column is visually presented in the feature plot, stating that the features towards the top of the table have the highest relative predictive abilities. For the cover and frequency, most of the features are approximately the same order as the gain. However, one interesting observation is that *dxy* has a substantially higher value of both frequency and cover compared to the other features. This means that the dollar index is the feature used to split the data across all trees the highest number of times.

Table 3: Variable importance table. R-output from XGBoost, before the 2008 financial crisis.

	Feature	Gain	Cover	Frequency
1:	wti	0.4187200254	0.14640723	0.139079852
2:	durgoods	0.3424739763	0.18160706	0.084611317
3:	dxy	0.1211911265	0.25498857	0.206240085
4:	unemployment	0.0389595254	0.08100792	0.113167636
5:	hsi	0.0220876031	0.04019658	0.052882073
6:	shcomp	0.0186370657	0.02551495	0.041248017
7:	vix	0.0083472596	0.01657915	0.008989952
8:	dji	0.0068668354	0.01323912	0.029085140
9:	mswi	0.0064832008	0.04436315	0.062929667
10:	s.p500	0.0056454154	0.04619896	0.042834479
11:	TCMNOM_Y10	0.0049975387	0.03511271	0.034902168
12:	FF_eff	0.0040203585	0.04147813	0.056054997
13:	N225	0.0009074212	0.02209868	0.049180328
14:	goldsupply	0.0003970325	0.02322776	0.030671602
15:	usinflation	0.0002656154	0.02798004	0.048122686

During the 2008 financial crisis

The feature importance plots for the XGBoost and GBM models using data from during the financial crisis are presented in figure 8 and 9. Unemployment is the relatively most important variable, but the Shanghai Composite, the VIX index and US inflation also have high influence in the XGBoost model. Further, the *dxy* is the only feature that has approximately the same level of importance as in the period before the financial crisis. The GBM model yields quite similar results. Although not in identical order, the four most important features are equal for both the models.

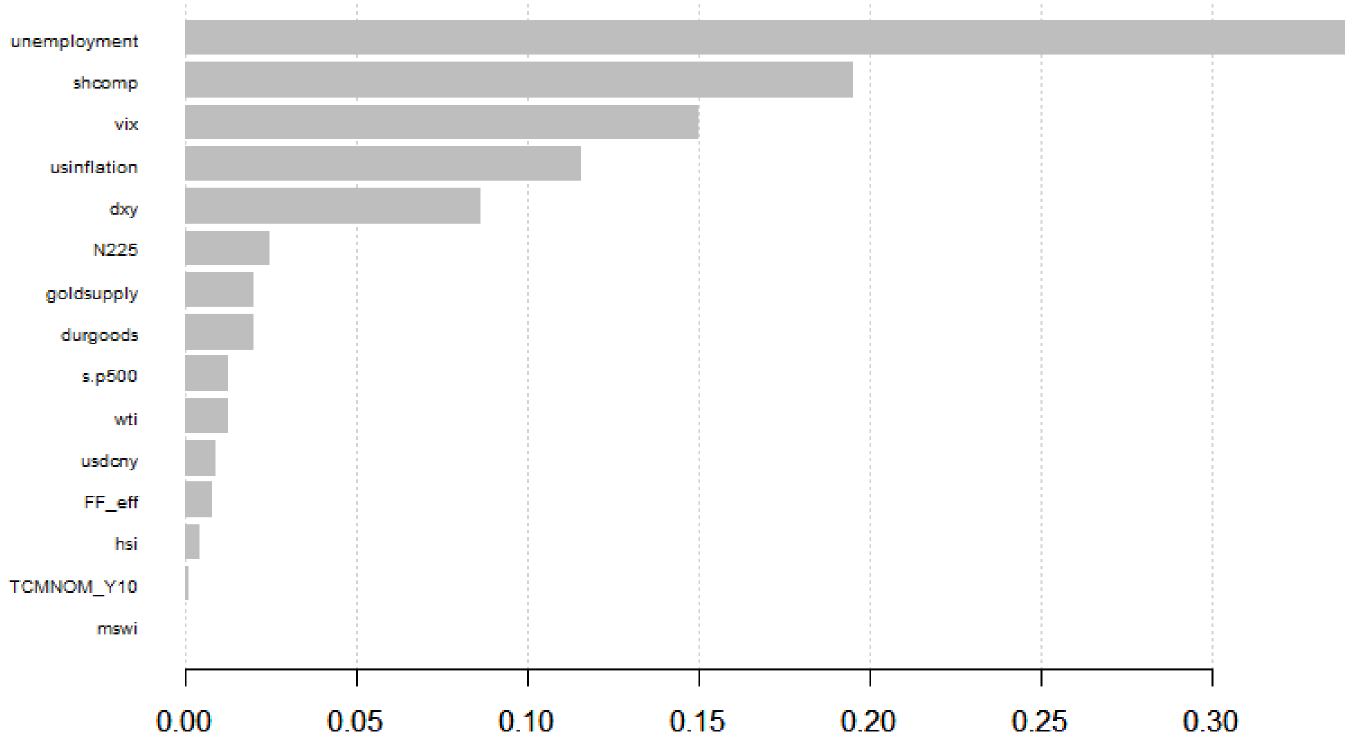


Figure 8: Variable importance plot. XGBoost, during the 2008 financial crisis.

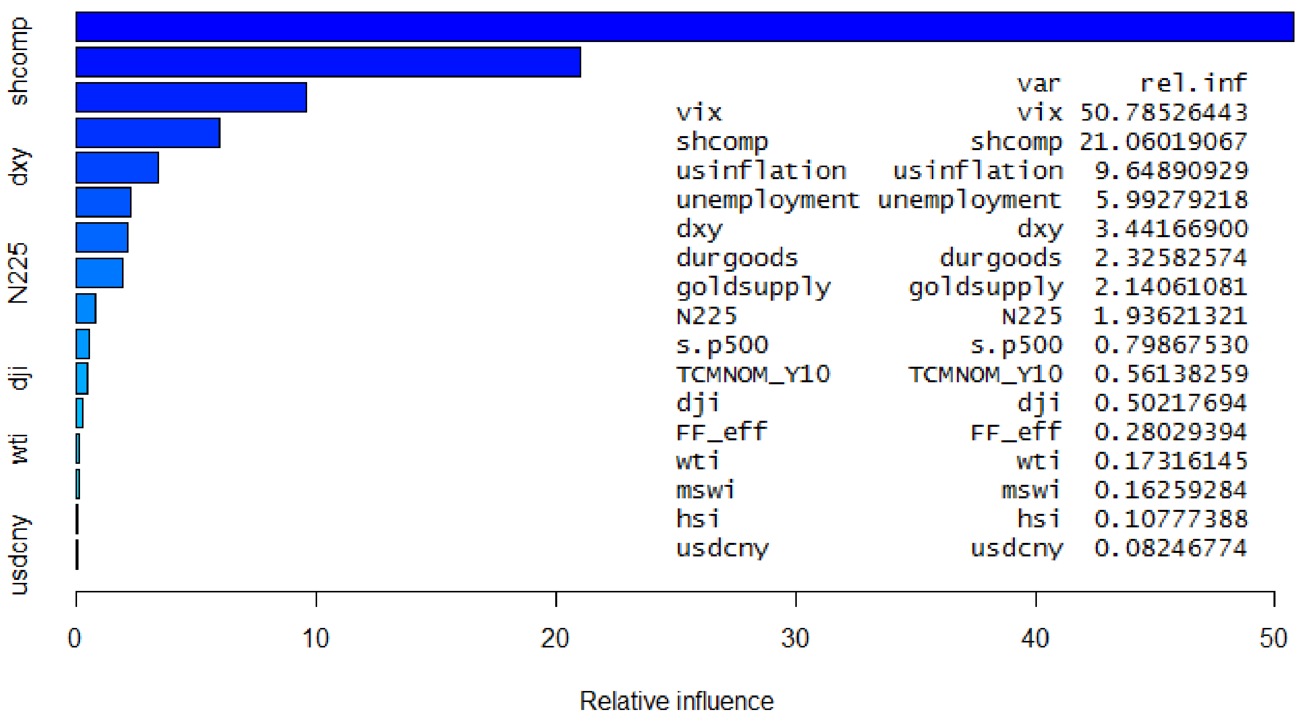


Figure 9: Variable importance plot. GBM, during the 2008 financial crisis.

For the period during the financial crisis, the XGB model yields the extended feature importance table presented below (table 4). The table indicates that there is a clear relationship between the relative influence, and the adjusted number of times a feature is used to split the data. One thing to note is that cover and frequency are more different from the previous period. This could be due to the relatively smaller observation size in this data set. US inflation has a relatively high frequency compared to the other two measurements, which indicates that this feature is used to decide the nodes for a high amount of observations across all trees. Further, the US dollar index has higher values of relative frequency and cover compared to the relative importance.

Table 4: Variable importance table. R-output from XGBoost, during the 2008 financial crisis

	Feature	Gain	Cover	Frequency
1:	unemployment	3.392760e-01	0.2518253912	0.168110919
2:	shcomp	1.955735e-01	0.1700507252	0.112651646
3:	vix	1.502286e-01	0.1507322998	0.098786828
4:	usinflation	1.160728e-01	0.1033507180	0.206239168
5:	dxy	8.648482e-02	0.1264699578	0.128249567
6:	N225	2.499666e-02	0.0351789669	0.041594454
7:	goldsupply	2.018593e-02	0.0249339144	0.019064125
8:	durgoods	1.983295e-02	0.0313067086	0.053726170
9:	s.p500	1.291513e-02	0.0351361006	0.029462738
10:	wti	1.259353e-02	0.0137315139	0.038128250
11:	usdcny	8.773585e-03	0.0370936629	0.045060659
12:	FF_eff	7.811147e-03	0.0117025077	0.025996534
13:	hsi	4.024675e-03	0.0038293920	0.013864818
14:	TCMNOM_Y10	1.201486e-03	0.0041866114	0.017331023
15:	mswi	2.918284e-05	0.0004715296	0.001733102

After the 2008 financial crisis

The feature importance plots in figures 10 and 11 indicate that features before and after the financial crisis are quite similar in regard to relative influence on the gold price. In this period, the WTI has the highest degree of relative influence. Further important variables are the Nikkei 225, the US inflation rate and durable goods. The dollar index is now less important than in the period before the financial crisis.

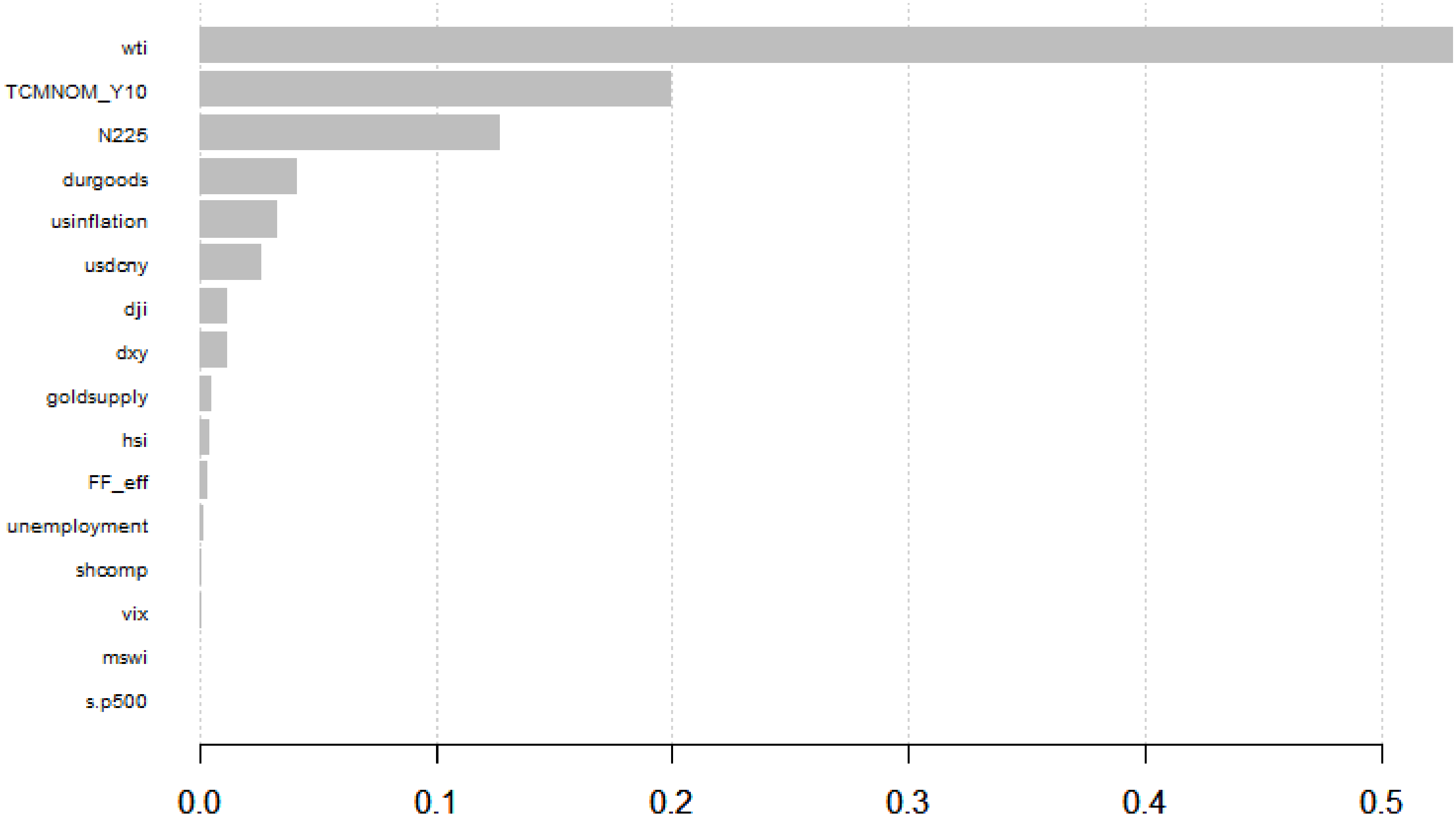


Figure 10: Variable importance plot. XGBoost, after the 2008 financial crisis.

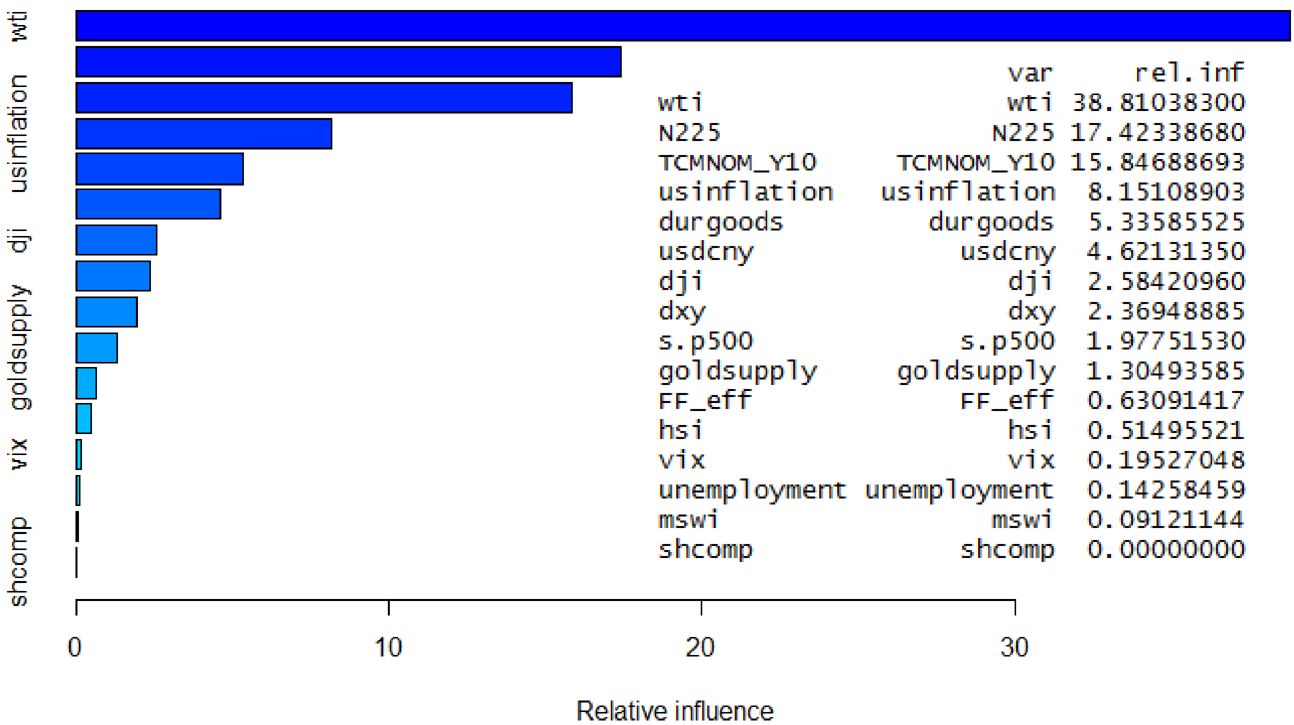


Figure 11: Variable importance plot. GBM, after the 2008 financial crisis.

Looking at table 5, the results are again quite similar to ones from the first period. The gain column is in this case more in line with cover, compared to frequency. The *wti* scores among the highest values of cover, in addition to the highest degree of relative influence. Regarding frequency, there seems to be less variation among the features now compared to table 3 containing data from before the financial crisis.

Table 5: Variable importance table. R-output from XGBoost, after the 2008 financial crisis

	Feature	Gain	Cover	Frequency
1:	wti	4.834593e-01	0.149154208	0.08803707
2:	TCMNOM_Y10	2.053405e-01	0.119186963	0.09814659
3:	N225	1.692717e-01	0.158971556	0.09098568
4:	durgoods	4.539189e-02	0.080819857	0.07455771
5:	usinflation	3.770322e-02	0.141139594	0.12342039
6:	usdcny	1.344213e-02	0.077566455	0.11162595
7:	dji	1.111034e-02	0.029452046	0.02906487
8:	dxy	9.488925e-03	0.074825487	0.09393429
9:	goldsupply	6.176199e-03	0.033969943	0.06318450
10:	FF_eff	5.920127e-03	0.034358877	0.02780118
11:	unemployment	5.546855e-03	0.035093427	0.06065712
12:	hsi	4.746161e-03	0.041794698	0.07118787
13:	s.p500	1.059373e-03	0.009309523	0.02948610
14:	shcomp	9.907289e-04	0.007933509	0.02021904
15:	mswi	2.713488e-04	0.004062605	0.01347936
16:	vix	8.126244e-05	0.002361251	0.00421230

2.3.2 Variables of predictive importance

Exchange rates: The Dollar index

The feature plots for both the GBM and XGB models in the pre-crisis period indicate that the US Dollar index is one of the relatively most important features when predicting the gold price. Joy (2011) discussed that on average, US Dollar returns have been correlated negatively with the gold price. This is in line with the role of gold as an investment hedge against the US dollar. Studies conducted by Capie et al. (2005) also concluded that gold serves as a hedge against fluctuation in the foreign exchange value of the US dollar, however to a lesser extent, as the effect was highly dependent on unpredictable political events and attitudes. One reason why gold serves as a hedge is due its nature as a homogeneous asset, thus easily traded in a continuously open market. Gold acquired the attributes of an asset mainly as it cannot be produced by authorities that produce currencies. Countries can indirectly decrease the value of their currency by increasing the supply of money, but cannot influence the value of gold (Capie et al., 2005).

Studies by Tully & Lucey (2007) and Shafiee & Topal (2010) indicate that oil price and inflation are the two macroeconomic variables influencing the gold price, both with more predictive power than the dollar. In the short term, however, the gold market's role as a safe haven and hedge against the dollar are both significant. During uncertain times when global financial markets crash and the global economy is in recession, investors seek to avoid financial markets as reliable investments. Consequently, investors typically switch to markets that does not include as much unpredictability, such as the gold market. This underlines the role of the gold price as a safe haven. Shafiee & Topal (2010) also argue that big companies often hedge gold against fluctuations in the US dollar. Subsequently, gold trading will influence the potential movement of real value in the short-term market toward US dollar oscillations.

In this research we have used gold price listed in US dollar, which makes it natural to assume that exchange rate variables are of high relative importance. Pukthuanthong & Roll (2011) suggest that the relationship between the gold price and the dollar creates an asymmetric implication for US and non-US investors. They argue how an investor not based in US dollar should take into account his implicit foreign exchange risk position when investing in gold or gold-linked instruments. Beckers and Soenen (1984) conclude that a non-US investor would want to invest in more gold because of its implicit foreign exchange hedging attributes. Studies by Sujit Ks and Kumar (2011) found that this exchange rate effect is not significant in Europe, as shocks in gold price listed in euro explains 1% of the variations in exchange rate.

Other commodities: WTI

The West Texas Intermediate, or WTI, is a grade of crude oil and works as a benchmark for oil traded in the US. The feature plots from GBM and XGB indicate that the WTI influence the predictions of the gold price. For the periods before and after the financial crisis, the models indicate that WTI has the highest relative influence in the prediction.

Research conducted Zhang and Wei (2010) suggest that the influence of crude oil on global economic development proves more extensive than the effect of the gold price. Le and Chang (2011) conclude that there is a long term relationship between the price of gold and oil through an inflation channel. Increasing oil price generates higher inflation, which in turn strengthens the demand for gold and thereby increasing the gold price. Furthermore, the non-linear effect between the variables indicate that the oil price can be used to predict the gold price.

Interest rates: The Effective Federal Fund Rate

The relationship between interest rates and the gold price has been widely discussed in literature. Christie-David et al. (2000), studying the effect of interest rate announcements on the gold price, finds evidence of a relationship, while Cai et al. (2001) does not find a significant relationship. The typical notion on the topic is that rising interest rates leads to a decrease in the gold price. An increase in interest rates will make fixed-income investments such as bonds more attractive, higher-yielding investments compared to gold.
The models in this analysis include the Effective Federal Fund Rate as a variable of importance for the prediction, but only to a limited extent, primarily in the period leading up to the 2008 financial crisis.

Unemployment

Cai et al. (2001) found, when examining 23 regularly scheduled US macroeconomic announcements, that employment reports were one of four announcements with a significant effects on the return volatility of the gold market, the others being gross domestic product (GDP), consumer price index (CPI) and personal income. A rapport conducted by the World Gold Council stated that the unemployment rate was the most important contributing factor on jewelry recycling (Christensen, 2015). As gold is one of the most inert metals and not easily oxidized, recycled gold is in a pure state and can easily be sold in the market (Ali, 2006). The World Gold Council found that in India, 1% rise in unemployment leads to a 1.3% increase in India’s share of global recycled gold. Recycled gold contributes to a large share of total gold supply, typically around 33%, reaching as high as 42 % of production in 2009 (Christensen, 2015 and Hewitt, 2015).Unemployment is a fairly important variable according to the feature importance plots. In the period before the financial crisis it has the fourth highest relative importance, but the gain is small compared to the values of crude oil and durable goods. During the financial crisis, both the XGBoost and the GBM model indicate that unemployment is the most important variable in the prediction.

US treasuries: The 10-year Treasury note

US Treasury bonds are debt obligations issued by the US government, where the 10-year note has been especially scrutinized and followed closely (Covarrubias et al. 2006). The role of the Treasury note goes beyond return on investment for the security itself, as they are fully backed by the US government (Krishnamurthy & Vissing-Jorgensen, 2012). This function enables the Treasury note’s function as a safe haven, and thus making it an attractive investment during times of uncertainty and market downturns (Flavin et al., 2014). A higher demand for Treasury notes leads to a fall in yields.

In the analysis conducted by Flavin et al. (2014), the 10-year Treasury note and gold both are found as a favorable safe haven asset during economic downturn from the perspective of an equity fund manager. The feature plots indicate that the 10-year Treasury yield is a relatively important variable in the prediction of the gold price, especially in the period after the 2008 financial crisis, which makes sense given the similar properties.

US inflation

The relation between inflation and the gold prices is frequently discussed in the literature, and is perhaps one of the more researched variables included in this thesis. The typical examined relationship is the hedging abilities of gold against inflation. The previous mentioned studies by Worthington and Pahlavani (2007), Wang et al. (2011) and Beckman and Czudaj (2013) all argue that there is evidence on a common long-term trend, and that gold and inflation are cointegrated, thus consistent with the notion that gold at least partially hedge inflation risk (Batten et al. 2014). The linkage between the two is on the other hand also questioned: Erb and Harvey (2013) points to 1982 as the only year with evidence of any connection, while Batten et al. (2014) rejects the hypothesis of cointegration for the whole period 1985 to 2012. Though, the latter finds that there is a significant increase in comovement between the inflation and gold the last years of their examined time period.

Our results show that US inflation is not of any significant importance for a prediction based on the time period before the 2008 financial crisis. The importance increases for the in-crisis time period. In the feature plot of the time period after the end of the 2008 financial crisis, the importance of inflation has decreased, but is still of more importance than in the years before the crisis. Thus our findings also display a change in the relationship over time. This is consistent with the standpoint of Batten et al. (2014), who proposes the choice of examined time period as one possible reason for the mixed opinions in literature.

Gold supply

For commodities in general, the supply of material to the market is a natural factor to examine when explaining or predicting the price. For gold, the supply side of the market is not of the same importance as it is for other commodities. Two important reasons for these characteristic are the inelasticity of supply and the renewability of gold.

Firstly, the time and cost of exploring, extracting and refining gold is extensive. The time from a potential discovery of ores to the first, finished product can be delivered takes on average as much as 10-20 years (World Gold Council, 2017). Thus will an overnight increase in gold prices not lead to an immediate increase of gold supply.

Secondly is the gold's mentioned property of being highly renewable. In contrast to other commodities and even to a large extent other precious metals like silver, gold is in almost every produced form recyclable, and not being consumed in a traditional fashion. This means that a vast majority of all gold ever extracted is still in circulation (Sverdrup et al., 2012). This accumulated sum of metal means that the supply of new material to the market has a diminishingly lower impact as a yearly production is proportionally small to the existing stock of gold in circulation.

Our results show that neither the GBM nor the XGboost model specify the gold supply as an important variable in the prediction of the gold price.

US durable goods orders

A durable good is a good that yields utility over time, in contrast to a consumable good that is completely consumed in one use (Engel & Wang, 2011). A durable good typically yields utility for more than three years, such as cars, home furniture and electronics, medical equipment and airplanes. In economics it is considered an important indicator because households or businesses to a larger extent invest in durable goods when they are confident the economy is improving (Engel & Wang, 2011). The data is presented through a report built on a survey passed out to numerous manufacturers in the US, where information about orders, inventories and shipments is provided. For this analysis, the orders are especially interesting since they provide a leading indicator of economic growth, as opposed to shipments which are more immediate measures of the economic situation (Hughes, 2018).

In the context of gold, this durable good is included as it somewhat represents the opposite in terms of hedging properties. As the traditional notion of gold is the precious metal being a safe haven when there is less confidence in the market, will an increase in orders of durable goods be an indicator of increased confidence.

In both models, especially before but also after the 2008 financial crisis, durable goods proves to be a variable of importance to the models in predicting the gold price.

Stock Indices

The relationships between the gold price and financial markets have been widely researched, and multiple papers seek to examine the role of gold as a hedge or as safe haven for stock markets (Baur & Lucey, 2010, Choudhry et al., 2015). In our model, we have opted to include variables representing the stock market in the US, Japan, China and an index which represents the world economy. Both the GBM and XGB model suggest that financial markets are important when predicting the gold price. Which stock markets have the highest degree of relative influence vary across time.

A study conducted by Baur and Lucey (2010) examined the relationship between the gold price and stock markets in the US, the UK and Germany. They argued gold is a safe haven for stocks in the US, the UK and Germany, but only during extremely negative market shocks. Further, the safe haven property is only for a short time period. Purchasing gold after an extreme shock to the economy yields a positive gold return, thereby implying the role of safe haven for the gold price for investors who buy gold the day after the occurrence of an extreme market shock. That gold is a safe haven only for a short period can be explained with the role of gold as a hedge for stocks. A hedge correlates negatively with another asset on average, indicating that if the price of one asset increases the price of the hedge asset falls. Because the stock market commonly increases some time following an extreme negative shock has occurred, the existence of a hedge works against a safe haven asset in the long run (Baur & Lucey, 2010).

Smith (2002) concluded that there is no cointegration involving a gold price and a stock price index. In other words, he found that there is no long run equilibrium and the series do not share a common stochastic trend. On the other hand, there are short run relationships between the gold price and multiple stock market indices. The results from the XGB model using the subseries before the financial crisis indicate that the Hang Seng is relatively important when predicting the gold price. Similarly, the Nikkei225 is important for the period after the financial crisis.

Fear Index: VIX

The Volatility Index (VIX) is an index of fear sentiment. It differentiates itself from other volatility estimators by being a market-determined forecast and do not employ certain degrees of smoothing from past volatility, and is generally considered the best tool for predicting the future volatility of the market (Qadan & Yagil, 2012). It is included as a variable in this analysis as another factor determining the participants’ confidence in the market, but with regard to the psychological sentiment and how this possibly can be an influential variable when predicting the gold price.

The relative importance of the VIX index is close to zero for both the XGboost and GBM model before and after the 2008 financial crisis. During the 2008 financial crisis, the VIX index appears to be of more relative importance. The literature examining the relationship between the VIX index and gold does in general provide evidence on linkage between the two, also for time periods that overlap with the ones examined in this period. Qadan & Yagil (2012) and Jubinski & Lipton (2013) point to a significant relationship for the whole period 1990/1995 to 2010, while Cohen & Qadan (2010) and Boscaljon & Clark (2013) find that the relationship was especially significant during the 2008 financial crisis.

US Dollar/ Chinese Yuan exchange rates

The movement of the exchange rate between the US dollar and the Chinese yuan (*usdcny*) has similarities to the development of the gold price for recent years. The correlation between gold and the yuan has become clear in 2018, and researchers and speculators have introduced the idea of gold being pegged in yuan terms (Brady, 2018, Baker, 2018). This is based on the increasingly narrow range in the gold and yuan exchange rate. China is now directly involved in gold pricing, and conversely, the role of the New York-based commodity exchange COMEX is diminishing. Brady (2018) uses an R-squared measure to evaluate the gold regression models based on the yuan, euro and the yen. Here, linear regression is used on the exchange rates, and COMEX gold futures. He concludes that for the year of 2018 until August, the yuan outperforms the other currencies when computing an expected value and range for gold prices.

Other studies indicate that the relationship between the yuan and gold prices is not as interesting as previously suggested. Sid Norris (2018) found that there is correlation between the yuan and the gold price in recent years, but argues how this correlation is not certain for the future. The argument is supported by the fact the occasional high correlation never holds, and even reverses to a negative correlation regularly. For our models, in the period before the financial crisis the *usdcny* is not included in the data set because the value of the yuan was pegged to the US dollar until 2005. Interestingly, after the crisis the relative influence has increased for this feature in both the XGBoost and GBM model.

2.3.3 Summary of the time periods: Differences and similarities

In the period before the financial crisis, *wti* was the variable with the highest relative influence in both the GBM and XGB model. An increase in the oil price affects the US in two ways: Job creation and growth in the large and important oil sector, but also higher transport and manufacturing costs for some consumers and businesses. One possible effect of this could be increased spending in the private sector, and individuals could be more likely to buy jewellery when the economy is expanding. As previously mentioned, Le and Chang (2011) found a positive relationship between the oil price and the price of gold through the inflation channel. Interestingly, however, the GBM and XGB models indicate that inflation is not of high relative importance before the financial crisis. Further, durable goods has the second highest degree of relative influence in this period. The US dollar index also has a high degree of relative importance, which corresponds to findings of Joy (2011) and Capie et al. (2005) about the role of gold as a hedge for the US dollar. Other significant variables in this time period are *unemployment*, *shcomp*, *hsi* and *dji*.

During the 2008 financial crisis, the models yield quite different results regarding the relative importance of variables. Now, unemployment is a very important variable in the predictions, having the highest relative influence in the XGB model, and is the fourth most important in the GBM model. Further, the results from the GBM model indicate that the VIX index is the most important variable in the prediction, in addition to being of high relevance in the XGB model. Another interesting aspect during this time period is that the Shanghai Composite is among the most significant variables in both of the models. Similarly, US inflation is a much more relevant variable during the financial crisis.

The time period after the financial crisis yields results that are more in line with the models from before the financial crisis. Crude oil is again the most important variable, with the highest degree of relative influence in both the GBM and the XGB model. Conversely, Nikkei 225 is among the variables with the highest degree of relative influence in this subsample.

When comparing the three time periods, it becomes clear that there are some similarities among all of them. First, the WTI includes some prediction power across all the different subsets, although to a lesser degree for the “Inter” models. Further, the dollar index appear among the most important variables in both of models for all time periods. This supports the theory that gold function as a hedge against the US dollar, due to the negative correlation between the two variables. Durable goods is also among the most relevant input variables across the different models. The positive correlation between the gold price and durable goods is interesting, and does not directly correspond to the theory of gold as a hedge against stock markets. Another important aspect is that the period before and the period after the 2008 financial crisis share a high amount of similarities compared to the period during the crisis.

The time period during the financial crisis clearly separates from the other two periods, but there are other differences to note as well. During the financial crisis, VIX, unemployment and inflation all have a significantly higher degree of relative influence. This could support the theory that gold function as a safe haven, but only during a time of financial stress in the economy. Another thing to note is that some variables that have a high relative influence before and after the financial crisis, are far less important in the during time period. This includes variables such as WTI and durable goods, which could suggest that the role of gold as a hedge is less prominent in times of financial distress. However, another possibility is the effect of the global financial crisis on the demand of jewellery. A major decrease in the global demand of gold jewellery could cause a fall in the price of gold. It is important to note that the models during this period could be heavily influenced by the small number of observations. The issues concerning this period will be further addressed in section 2.3.5. This could be the main driver behind the vastly different results from both models, and the issue will be questioned in the limitation section.

Comparing the periods before and after the financial crisis, there are some interesting differences to look further into. First, we see that the US 10-year Treasury note is of substantially higher relevance after the financial crisis. This might be the result of the negative correlation between the gold price and the bond yield being more decisive in the models in this period. Baur and Lucey (2010) concluded that gold generally does not function as safe haven for bonds, so it could be interesting to examine if this relationship has changed in recent years. Further, the Nikkei 225 is only among the most important variables in the period after 2008. Although the US dollar index is present among the relatively more important variables after the financial crisis, the relative importance has dropped substantially compared to the first period. This is the case for both the XGB and GBM model.

The WTI is suggested to influence the gold price through the inflation channel, but the feature representing inflation is not of high importance according to the models. This result is interesting, as the WTI is proposed to be the most influential feature in the models. One potential reason of how this is possible, could be due to the fact that the inflation index is originally not noted on a daily basis. Even though the observation is transformed to daily data using interpolation, the true effect could potentially not be captured by the model. This could also the case for the other economic variables included in this thesis, as they are usually not noted on a daily basis.

To summarize, the period during the 2008 financial crisis is clearly different from the other periods. This supports the theory that gold function as a safe haven during times of financial distress, although the models can be questioned due to the small observation count. Further, the periods before and after the crisis yield similar models, and the WTI is the relatively most important feature in both. On the other hand, the US 10-year treasury note and Nikkei 225 index are relatively more important in the predictions after the crisis. In addition, the relative influence of the US dollar index is less substantial in the last period.

2.3.4 Model accuracy

In order to assess the legitimacy of the variable importance plots, the predictive accuracy of the models is examined using three different performance measures. Two of them normalized to improve the comparison between the periods. We include multiple measures to achieve a comprehensive evaluation. Mean absolute error (MAE) is, as the previously mentioned RMSE, measuring of the difference between the actual and predicted observations. Swalin (2018) also quotes MAE together with RMSE as the most popular metrics for continuous variables in supervised machine learning models. One of their most useful properties is the fact that they are dimensional statistics, which means that they are in the same unit as the response variable.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

RMSE squares the errors before averaging them, which means that RMSE emphasizes the larger errors. This is the most prominent difference between RMSE and MAE, resulting in RMSE being larger than MAE in the cases where some of the individual prediction errors are more extreme.

R-squared, noted R^2, is also included as measure of model accuracy, as it is a common and familiar instrument. The R^2, is however interpreted with caution as it has weaknesses as a measure of performance (Shalizi, 2015). R-squared is the fraction of the total sum of squares explained by the respective models.

$$R^2 = \frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}_j\right)^2}$$

Where the sum of squared errors of our regression model is the numerator, and the sum of squared errors of our baseline model is in the denominator.

The two measures which provide absolute output values, MAE and RMSE, are normalized in order to create a more accurate comparison. Among the different normalization methods, normalization by the mean value of the measurement is used in case, as the goal is to adjust for the difference in gold price level before and after the 2008 financial crisis. Normalized MAE and RMSE, hereby NMAE and NRMSE, can be described accordingly:

$$NMAE = \frac{MAE}{\bar{y}}$$

$$NRMSE = \frac{RMSE}{\bar{y}}$$

Table 6 displays the achieved accuracy of the two models.

Table 6: Performance measures

XGBoost	**MAE**	**RMSE**	**R^2**	**NMAE**	**NRMSE**
Before the 2008 financial crisis	37.94	42.46	0.92	0.093	0.105
During the 2008 financial crisis	46.95	55.22	0.16	-	-
After the 2008 financial crisis	89.81	96.57	0.75	0.068	0.073
GBM					
Before the 2008 financial crisis	49.71	64.90	0.81	0.122	0.159
During the 2008 financial crisis	40.73	51.50	0.27	-	-
After the 2008 financial crisis	90.49	118.77	0.69	0.068	0.089

XGBoost yields the best predictions. The GBM has a larger spread between the MAE and RMSE, indicating that the GBM predictions have some larger individual errors in some of the predictions.

In table 6 containing the performance measures we see that both MAE and RMSE increase with time, being lower in the period before the 2008 financial crisis than after, for both models. This is not necessarily because of lower accuracy, as the R-squared might imply, but a result of the general higher gold price in later years. In the period after the 2008 financial crisis, the average gold price has been approximately $1328 per ounce, while the average price before were approximately $406 per ounce. This difference in scale is the reason for including the normalized versions of MAE and RMSE. Lower NMAE and NRMSE implies better accuracy, which means that the predictions from the period after the 2008 financial crisis has a somewhat better accuracy than the period before the crisis. NMAE and NRMSE are not calculated for the period during the crisis, as the normalization could make the measure even more sensitive to sample size. This is not regarded as a limitation as the comparison between the before and after periods are of more interest in this thesis.

In general, despite being somewhat different, we consider both the pre- and post-crisis accuracy to be sufficient to give the variable importance plots credibility.

We are in general skeptical to the prediction accuracy and thus also the variable importance plots of the period during the 2008 financial crisis. When working with the data, the results appeared fragile by changing considerably even with smaller adjustments in parameters. The main issue is few observations in the data frame, especially after the split for making training and test set. Further assessment of the weaknesses of this period are discussed in the section 2.3.5. The period itself was a time of uncertainty, with rapid and extraordinary movements in many economic and financial variables.

2.3.5 Limitations

This analysis of the properties of the gold price is based on simplifications, as any other attempt to model the complex and dynamic reality. The following section will discuss some of the limitations, weaknesses or potential issues concerning the analysis or model specifics.

When running the XGBoost and GBM models, the included bucket of variables is a product of relevant literature, data availability and economic intuition. The literature differs in its evidence on the relationship between the gold price and financial or macroeconomic variables. In this thesis we have included well documented variables, both when researchers agree about a relationship, but also when they do not. The limitations regarding this point is firstly the likely possibility that there are relevant variables that were not included simply because we did not discover them. Secondly, the availability of data differed. Some data were not available for downloading, while others were not available in the appropriate format or time. For example numerous macroeconomic variables are reported quarterly, where the potential use of interpolation for estimating daily values were considered too radical.

Additionally, one of the fundamental questions asked in this thesis has been whether or not the basket of the most important variables for a prediction of the gold price have changed. In the ever changing reality, we have tried to include new possible variables that perhaps have not been included in previous studies of the gold price. Some examples of such potential variables are macroeconomic and financial variables that in previous studies have been US specific, but now as nations like China are increasingly more important players in the world market, equivalent Chinese specific variables could have been included. Again, the availability of data became the decisive impediment.

When it comes to the results, the already mentioned issue concerning the time period between the start and end of the 2008 financial crisis has some issues. The number of observations are significantly lower for this period, only totaling 322 observations after cleaning. Although the results provide some indication, the variable importance plots are assessed with caution, and the overall focus should be directed to a larger degree on the time periods before and after the crisis.

The performance measures used to evaluate the accuracy of the models all have their strengths and weaknesses. Reporting three and not one performance measure is the main preventive action to attenuate each of the individual weaknesses of the different measures. The weaknesses of R-squared and the comparability issues concerning the MAE and RMSE are already mentioned, but the question regarding which of the performance measures is the most fitting in this case is still highly relevant. Furthermore, the relationship between better model accuracy and more precise variable importance plots is not unambiguous.

The parameter tuning of the models is a source of uncertainty. The process of tuning parameters is challenging and differ from case to case. Thus, literature or other sources will only provide a limited amount of support. Though, when fine tuning the model, with help from the limited framework literature provides, tuning the parameters did not drastically change from trial to trial, apart from the problematic inter-crisis period. We therefore consider the results somewhat robust, even when taking the uncertainty around parameter tuning into account. We also consider the similarities in the results between GBM and XGBoost to be evidence for the tuning to be of somewhat appropriate quality.

Conclusion

The objective of this thesis has been to investigate the properties of the gold price, and whether these have changed over time. The analysis of the properties has been twofold. Firstly, we have looked for evidence of long memory in the gold price. Secondly, we have analyzed the relative importance of different financial and macroeconomic variables in a prediction of the gold price. Both parts of the analysis have been conducted on different periods in order to be able to compare the results, and assess whether they have changed over time.

Using the concept of fractional Brownian motion we looked for evidence of long memory by estimating the Hurst exponent using R/S analysis. The whole period 1979 to 2018 proved to have a Hurst exponent of above $\frac{1}{2}$, which mean the process has the property of persistence, where the past trend is more likely to sustain in the future. This was true for both the monthly and the daily series, 0.83 and 0.63 respectively. We also find that the lower frequency data display higher persistence, which is consistent with literature on the topic. When the period was divided in four, the Hurst exponent differed across time.

In the second part, our thesis has utilized two supervised machine learning techniques to produce predictive models that were able to display the relative importance of the input variables. With more focus on the time before and after the 2008 financial crisis, the relative importance of variables have changed slightly from before to after the crisis. However, variables that proved to be of most relative importance in general maintained their position among the top. Crude oil and durable goods orders proved to be of high relative importance both before and after the 2008 financial crisis. Two of the most volatile variables were the US 10-year Treasury Yield that turned out to be of more importance after the crisis, and the dollar index that proved to do the opposite. All the above mentioned variables are discussed to be relevant in relation to the gold price in previous studies. Furthermore the fact that these types of relationships change over time is also well established.

To provide some kind of verification of the relative variable importance the model's predictive performance are also measured when producing the variable importance plots. We regard both models to be within reasonable accuracy using multiple different accuracy measures. The

GBM and XGBoost yield the same results and perform quite similarly, with the XGBoost providing somewhat higher accuracy. This is according to expectations.

The period during the crisis displays results very differently from the other two periods. This period is considerably shorter with fewer observations for the models to work with, but is also a period of highly unusual times in the world economy, providing extraordinary movements in the input variables. The results are included and interpreted, but will be regarded with skepticism.

This thesis uncovers some potential topics for further research of the properties of the gold price. For further investigations of long memory in the gold price, we would suggest looking at the concept of multifractional Brownian motion, introduced in section 1.1.1. It allows the long memory to change over time with its dynamic Hurst exponent, which appears suitable to the case of the gold price, where we found the Hurst to change across time.

In the second part of the thesis, data availability was a pressing issue, and constrained some possibilities. The world economy is ever changing, with for example China being of more importance the last years. We would suggest, if more data would become available, including more macroeconomic and financial variables from countries like China in a future investigation of the gold price.

References

Aarshay, J. (2016). Analytics Vidhya. *Complete Guide to Parameter Tuning in XGBoost (with codes in Python)*. Downloaded November 1st from https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning- xgboost-with-codes-python/

Ali, S. H. (2006). Gold mining and the golden rule: a challenge for producers and consumers in developing countries. Journal of Cleaner Production, 14(3-4), 455-462.

Aref, S. (1998). *Hurst Phenomenom and Fractal Dimension in Long-Term Yield Data*. Conference Statistics in Agriculture. Downloaded September 14th from https://doi.org/10.4148/2475-7772.1275

Baker, R. (2018). The Mystery of Gold and the Chinese Yuan. *Kitco.com*. Downloaded November 12th from https://www.kitco.com/commentaries/2018-07-31/The-Mystery-of-Gold-and-the-Chinese-Yuan.html

Batten et al. (2005). R*eturn Anomalies on the Nikkei: Are They Statistical Illusions?* Chaos Solitons and Fractals 23 (4). Downloaded September 17th from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=396680

Batten, J. A., Ciner, C., & Lucey, B. M. (2014). On the economic determinants of the gold–inflation relation. *Resources Policy*, 41, 101-108.

Baur, D. G.; Lucey, B. M. (2010). *Is Gold a Hedge or a Safe Haven? An Analysis of Stocks, Bonds and Gold*. Financial Review, Volume 45, Issue 2.

Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? International evidence. Journal of Banking & Finance, 34(8), 1886-1898.

Beckers, S & Soenen, L. (1984). Gold: more attractive to non-US than to US investors*? Journal of Business Finance and Accounting*, 11, pp. 107-112

Beckman, J.; Czudaj, R. (2013). *Gold as an inflation hedge in a time varying coefficient framework*. The North American Journal of Economics and Finance, Volume 24

Brady, D. (2018). China Takes Control of Gold. *Sprott Money Blog*. Downloaded November 12th from https://www.sprottmoney.com/Blog/china-takes-control-of-gold.html

Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brownlee, J. (2016a). Machine Learning Mastery. *Overfitting and Underfitting With Machine Learning Algorithms.* Downloaded November 1st from: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning- algorithms/

Brownlee, J. (2016b). *A Gentle Introduction to XGBoost for Applied Machine Learning*. Machine Learning Mastery. Downloaded October 4th from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Brownlee, J. (2018). *A Gentle Introduction to k-fold Cross-Validation*. Machine Learning Mastery. Downloaded November 16th from https://machinelearningmastery.com/k-fold-cross-validation/

Boscaljon, B., & Clark, J. (2013). Do large shocks in VIX signal a flight-to-safety in the gold market?. *Journal of Applied Finance, 23*(2), 120. https://search.proquest.com/docview/1503534056?accountid=37265

Business Cycle Dating Committee. (2010). Conference call statement September 20, 2010. Cambridge, Massachusetts. Downloaded October 30th from http://www.nber.org/cycles/sept2010.html

Capie, F., Mills, T. C., & Wood, G. (2005). Gold as a hedge against the dollar. *Journal of International Financial Markets, Institutions and Money, 15*(4), 343-352.

Caporale et al. (2014). *Fractional integration and cointegration in US.financial time series data*. Empirical Economics 47, 4. Downloaded September 17th from https://econpapers.repec.org/article/sprempeco/v_3a47_3ay_3a2014_3ai_3a4_3ap_3a1389-1410.htm

Caporale et al. (2017). *Long memory and frequency in financial markets*. DIW Discussion Papers, No. 1647. Downloaded September 14th from https://www.econstor.eu/bitstream/10419/156139/1/882064789.pdf

Cai, J., Cheung, Y. L., & Wong, M. C. (2001). What moves the gold market?. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 21(3), 257-278.

Castle, N. (2017). *Supervised Vs. Unsupervised Machine Learning*. Oracle + Datascience.com. Downloaded October 1st from https://www.datascience.com/blog/supervised-and-unsupervised-machine-learning-algorithms

Chen, T. (2014). Introduction to Boosted Trees. *University of Washington, presentation Oct. 22 2014*. Downloaded November 20th from https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf

Chen, T. (2015). *Story and lessons behind the evolution of XGBoost*. Paul G. Allen School of Computer Science & Engineering. Downloaded October 4th from https://homes.cs.washington.edu/~tqchen/2016/03/10/story-and-lessons-behind-the-evolution-of-xgboost.html

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

Chen, T.; He, T.; Benesty, T.; Tang, Y. (2018). *Understand Your Dataset with XGboost*. XGBoost documentation version 0.80. Downloaded October 15th from https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html

Christie–David, R., Chaudhry, M., & Koch, T. W. (2000). Do macroeconomics news releases affect gold and silver prices?. Journal of Economics and Business, 52(5), 405-421.

Choudhry, T., Hassan, S. S., & Shabi, S. (2015). Relationship between gold and stock markets during the global financial crisis: Evidence from nonlinear causality tests. International Review of Financial Analysis, 41, 247-256.

Cohen, G., & Qadan, M. (2010). Is gold still a shelter to fear. *American Journal of Social and Management Sciences*, *1*(1), 39-43. http://scihub.org/AJSMS/PDF/2010/1/AJSMS-1-1-39-43.pdf

Corazza, M.; Malliaris, A. G. (2002). *Multifractality in Foreign Currency Markets*, Multinational Finance Journal 6. Downloaded September 17th from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1084659

Covarrubias, G., Ewing, B. T., Hein, S. E., & Thompson, M. A. (2006). Modeling volatility changes in the 10-year Treasury. *Physica A: Statistical Mechanics and its Applications*, *369*(2), 737-744.

DePillis, L. (2018). *10 years after the financial crisis, have we learned anything*? Downloaded October 3rd from https://money.cnn.com/2018/03/16/news/economy/financial-crisis-10-years/index.html

Engel, C., & Wang, J. (2011). International trade in durable goods: Understanding volatility, cyclicality, and elasticities. *Journal of International Economics*, *83*(1), 37-52.

Flavin, T. J., Morley, C. E., & Panopoulou, E. (2014). Identifying safe haven assets for equity investors through an analysis of the stability of shock transmission. Journal of International Financial Markets, Institutions and Money, 33, 137-154.

Friedman, M. (1962). *The Interpolation of Time Series by Related Series*. National Bureau of Economic Research (NBER), Cambridge, Massachusetts. Chapter 1, Introduction to “The Interpolation of Time Series by Related Series”.

Friedman, J.H. (1999). *Greedy Function Approximation: A Gradient Boosting Machine*. The Annals of Statistics, Vol 29, No. 5. Downloaded October 4th from https://www.jstor.org/stable/2699986

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Fornberg, B., & Zuev, J. (2007). The Runge phenomenon and spatially variable shape parameters in RBF interpolation. Computers & Mathematics with Applications, 54(3), 379-398.

Glenn, L. A. (2007). *On Randomness and the NASDAQ Composite*. Independent working paper. Downloaded September 17th from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1124991

Granero et. al. (2011). *Introducing fractal dimension algorithms to calculate the Hurst exponent of financial time series*. The European Physical Journal B. Downloaded September 14th from https://link.springer.com/content/pdf/10.1140%2Fepjb%2Fe2012-20803-2.pdf

Graves et al., (2017). *A Brief History of Long Memory: Hurst, Mandelbrot and the Road to Arfima, 1951-1980*. Entropy 2017, 19, 437. Downloaded September 14th from http://www.mdpi.com/1099-4300/19/9/437

Greene, M. T., & Fielitz, B. D. (1977). Long-term dependence in common stock returns. Journal of Financial Economics, 4(3), 339-349.

Guasoni, P. (2006). No arbitrage under transaction costs, with fractional Brownian motion and beyond. Mathematical Finance, 16(3), 569-582.

Hale, T. (2017). *New era for post-crisis interest rates threatens gold*. Downloaded October 3rd from https://www.ft.com/content/0989e098-b4b0-11e7-a398-73d59db9e399

Hastie, T.; Friedman, J. H.; Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *2nd Edition. Springer Science and Business Media, New York*.

Hewitt, A. (2015). The Ups and Downs of Gold Recycling. *World Gold Council.* https://www.gold.org/research/gold-recycling

Hillier, D., Draper, P., & Faff, R. (2006). Do precious metals shine? An investment perspective. Financial Analysts Journal, 62(2), 98-106.

Hurst, H. E. (1951). *Long-term storage capacity of reservoirs*. Transactions of the American Society of Civil Engineers, Vol. 116, Issue 1, page 770-799.

Hughes, R. (2018). Durable-Goods Orders Jump In August. *Seeking Alpha, Economy market outlook September 28. 2018*. Downloaded November 13th from https://seekingalpha.com/article/4208827-durable-goods-orders-jump-august

Ismail, Z., Yahya, A., & Shabri, A. (2009). Forecasting gold prices using multiple linear regression method. *American Journal of Applied Sciences*, *6*(8), 1509.

Jacobsen, B. (1995). *Are Stock Returns Long Term Dependent? Some Empirical Evidence.* Journal of International Financial Markets, Institutions and Money 5 (2/3). Downloaded September 17th from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7459

Joy, M. (2011). Gold and the US dollar: Hedge or haven?. Finance Research Letters, 8(3), 120-131.

Jubinski, D., & Lipton, A. F. (2013). VIX, Gold, Silver, and Oil: How do Commodities React to Financial Market Volatility?. *Journal of Accounting and Finance,* 13(1), 70-88. http://www.na-businesspress.com/jaf/jubinskid_web13_1_.pdf

Kvarving, A. M. (2008). Natural Cubic Splines. *Department of Mathematical Sciences, Norwegian University of Science and Technology*. Presentation October 21. 2008. Downloaded October 20th from https://www.math.ntnu.no/emner/TMA4215/2008h/cubicsplines.pdf

Krishnamurthy, A., & Vissing-Jorgensen, A. (2012). The aggregate demand for treasury debt. Journal of Political Economy, 120(2), 233-267.

Lacasa et al. (2009). *The Visibility Graph: a new method of estimating the Hurst exponent of fractional Brownian motion*. Europhysics Letters 86(3):30001. Downloaded September 14th from https://arxiv.org/pdf/0901.0888.pdf

Le, T & Chang, Y. (2011). *Oil and gold: correlation or causation?* Downloaded October 27th, from https://mpra.ub.uni-muenchen.de/31795/1/MPRA_paper_31795.pdf

Lizorkin, P.I. (2001). Fractional integration and differentiation. *Michiel, Encyclopedia of Mathematics.*

Lo, A. W. (1991). *Long-term memory in stock market prices.* Econometrica 59, page 1279-1313.

Lutis, E. (2017). Ensemble Methods in Machine Learning: What are They and Why Use Them? Towards Data Science. Downloaded November 22nd from https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

Mallick, S. (2017). Bias-Variance Tradeoff in Machine Learning. *Learn OpenCV*. Downloaded November 15th from https://www.learnopencv.com/bias-variance-tradeoff-in-machine-learning/

Mandelbrot, B; Van Ness, J. W. (1968). *Fractional Brownian Motions, Fractional Noises and Applications*. SIAM Review. Downloaded September 13th from https://courses.physics.ucsd.edu/2016/Spring/physics235/Mandelbrot%20Van%20Ness.pdf

Mandelbrot (1971). When Can Arbitraged Efficiently? *A Limit to the Validity of the Random Walk and the Martingale Models*. Review of Economics and Statistics 53(3) Downloaded September 14th from https://ideas.repec.org/a/tpr/restat/v53y1971i3p225-36.html

Mandelbrot, B. (1972). *Statistical Methodology For Nonperiodic Cycles: From The Covariance To Rs Analysis*. Annals of Economic and Social Measurement 1, page 259-290.

Mandelbrot, B. (1983). *The Fractal Geometry of Nature*. San Francisco: W. H. Freeman.

Mandelbrot, B; Hudson, R. L. (2004). The misbehavior of markets: A fractal view of financial turbulence, New York: Basic Books.

Melvin, M., & Sultan, J. (1990). South African political unrest, oil prices, and the time varying risk premium in the gold futures market. *Journal of Futures Markets*, *10*(2), 103-111.

Menglong, F. (2013). *Hurst parameter estimation on fractional Brownian motion and its application to the development of the zebrafish*. University of Warwick. Downloaded September 13th from https://warwick.ac.uk/fac/cross_fac/complexity/study/emmcs/outcomes/studentprojects/fu_menglong.pdf

Moosa, I. A., & Burns, K. (2013). Interpolating flow and stock variables in a continuous-time dynamic framework. Applied Economics Letters, 20(7), 621-625.

Morariu et. al. (2007). *Detrended Fluctuation Analysis of Autoregressive Processes.* Cornell University Library. Downloaded September 17th from https://arxiv.org/pdf/0707.1437.pdf

Mynhardt et al. (2014). *Behavior of financial markets efficiency during the financial market crisis: 2007 – 2009.* Corporate Ownership and Control 11, 2. Downloaded September 17th from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2501414

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

Nielsen, D. (2016*). Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?* (Master's thesis, NTNU)

Norris, S. (2018). The Chinese Yuan vs Gold. How Correlated Are They? *Elliot Wave Plus. Downloaded November 12th from* https://elliottwaveplus.com/the-chinese-yuan-vs-gold-how-correlated-are-they/

Pafka, S. (2016). XGBoost Workshop and Meetup Talk with Tianqi Chen. DataScience.LA Downloaded October 3rd from http://datascience.la/xgboost-workshop-and-meetup-talk-with-tianqi-chen/

Pukthuanthong, K., & Roll, R. (2011). Gold and the Dollar (and the Euro, Pound, and Yen). Journal of Banking & Finance, 35(8), 2070-2083.

Qadan, M., & Yagil, J. (2012). Fear sentiments and gold price: testing causality in-mean and in-variance. *Applied Economics Letters*, *19*(4), 363-366.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007. Downloaded November 1st from http://www.saedsayad.com/docs/gbm2.pdf

Rogers, L. C. G. (1997). *Arbitrage With Fractional Brownian Motion*. Mathematical Finance, Vol 7, No.1.

Saraswat, M. (2017). Hackerearth. *Beginners Tutorial on XGBoost and Parameter Tuning in R*. Downloaded November 1st from: https://www.hackerearth.com/blog/machine-learning/beginners-tutorial-on-xgboost-parameter-tuning-r/

Shafiee, S., & Topal, E. (2010). An overview of global gold market and gold price forecasting. *Resources policy*, *35*(3), 178-189.

Shaikh et al. (2014). *Gold Price Data Analysis Using Rescaled Range (R/S) Analysis*. International Journal of Economics, Commerce and Research (IJECR). Downloaded September 17th from http://www.tjprc.org/publishpapers/--1390573068-2.%20Gold%20price.full.pdf

Shalizi, C. (2015). Lecture 10: F-Tests, R-squared, and Other Distractions. *Carnegie Mellon University, 36-401, Section B. Friday 16th October 2015*. Downloaded November 15th from: https://data.library.virginia.edu/is-r-squared-useless/

Shevchenko, G. (2014). *Fractional Brownian Motion In A Nutshell*. Cornell University Library. Downloaded September 14th from https://arxiv.org/abs/1406.1956

Singh, S. (2018). Understanding the Bias-Variance Tradeoff. Towards Data Science. Downloaded November 1st from https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Sjaastad, L. A. (2008). The price of gold and the exchange rates: Once again. Resources Policy, 33(2), 118-124.

Smith, G. (2002*). London Gold Prices and Stock Prices Indices in Europe and Japan.* Downloaded October 1st from http://www.spdrgoldshares.com/media/GLD/file/GOLD&EUJPStockIndicesFeb2002.pdf

Sujit, K. S., & Kumar, B. R. (2011). Study on dynamic relationship among gold price, oil price, exchange rate and stock market returns. *International journal of applied business and economic research*, *9*(2), 145-165.

Sverdrup, H., Koca, D., Granath, C., Schwanninger, M., Husemann, E., & Lane, D. (2012). Modeling the gold market, explaining the past and assessing the physical and economical sustainability of future scenarios. In *International Conference of the System Dynamics Society, St. Gallen, Switzerland*.

Swalin, A. (2018). Choosing the Right Metric for Evaluating Machine Learning Models - Part 1. *The Data Institute at University of San Francisco*. Downloaded November 20th from https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Tully, E.; Lucey, B. M. (2007). A power GARCH examination of the gold market. Research in International Business and Finance 21. Downloaded October 1st from https://doi:10.1016/j.ribaf.2006.07.001

Uthayakumar, P.; Jayalalitha, G. (2018). *A Comparison of Fractal Dimension Algorithms by Hurst Exponent using Gold Price Time Series*. International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 6, Issue 2. Downloaded September 17th from https://www.ijraset.com/fileserve.php?FID=13698

Vehèl, J & Peltier, R. (1995). *Multifractional Brownian Motion : Definition and Preliminary Results.* Downloaded October 21st, from https://hal.inria.fr/inria-00074045/document

Voss, J. (2013). *Is the S&P 500 Mean Reverting? Rescaled Range Analysis Provides the Answer*. CFA Institute. Downloaded september 14th from https://blogs.cfainstitute.org/investor/2013/02/18/is-the-sp-500-mean-reverting-rescaled-range-analysis-provides-answers/

Wang, K. M., Lee, Y. M., & Thi (2011). Time and place where gold acts as an inflation hedge: An application of long-run and short-run threshold model. Economic Modelling, 28(3), 806-819.

World Gold Council (2017). *How gold is mined: The lifecycle of a gold mine*. World Gold Council. Downloaded November 15th from https://www.gold.org/about-gold/gold-supply/how-gold-is-mined

World Gold Council (2018). Gold prices data. GoldHub. Downloaded September 10th from https://www.gold.org/goldhub/data/gold-prices

Worthington, A. C.; Pahlavani, M. (2007). *Gold investment as an inflationary hedge: Cointegration evidence with allowance for endogenous structural breaks.* University of Wollongong. Downloaded September 13th from https://ro.uow.edu.au/cgi/viewcontent.cgi?referer=https://www.google.no/&httpsredir=1&article=1211&context=commpapers

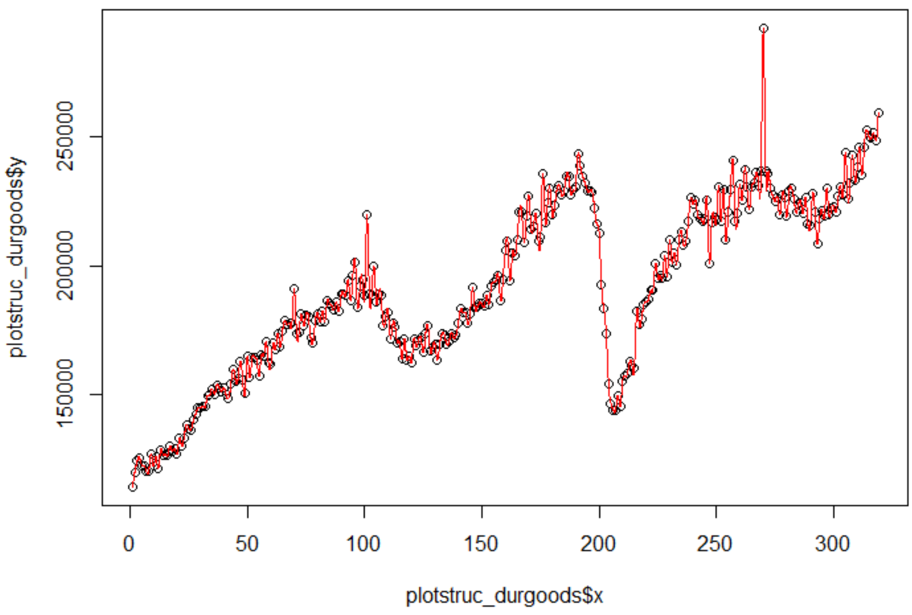
Yerlikaya-Özkurt, F., Vardar-Acar, C., Yolcu-Okur, Y., & Weber, G. W. (2014). Estimation of the Hurst parameter for fractional Brownian motion using the CMARS method. Journal of Computational and Applied Mathematics, 259, 843-850.

Zhang, Y. J., & Wei, Y. M. (2010). The crude oil market and the gold market: Evidence for cointegration, causality and price discovery. Resources Policy, 35(3), 168-177.

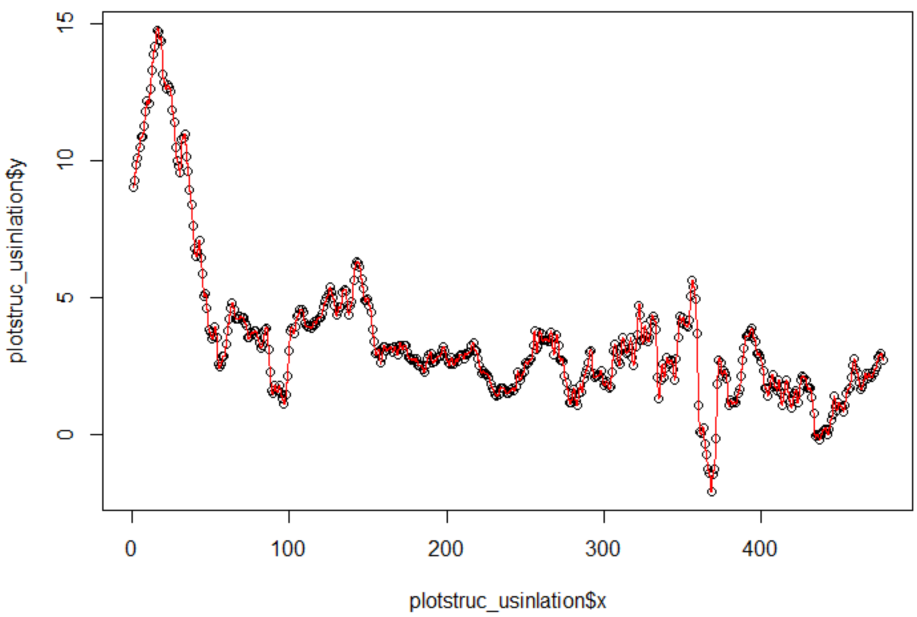
Appendix

A1: Interpolated variables: Original data points (dots) with interpolated estimations (red line).

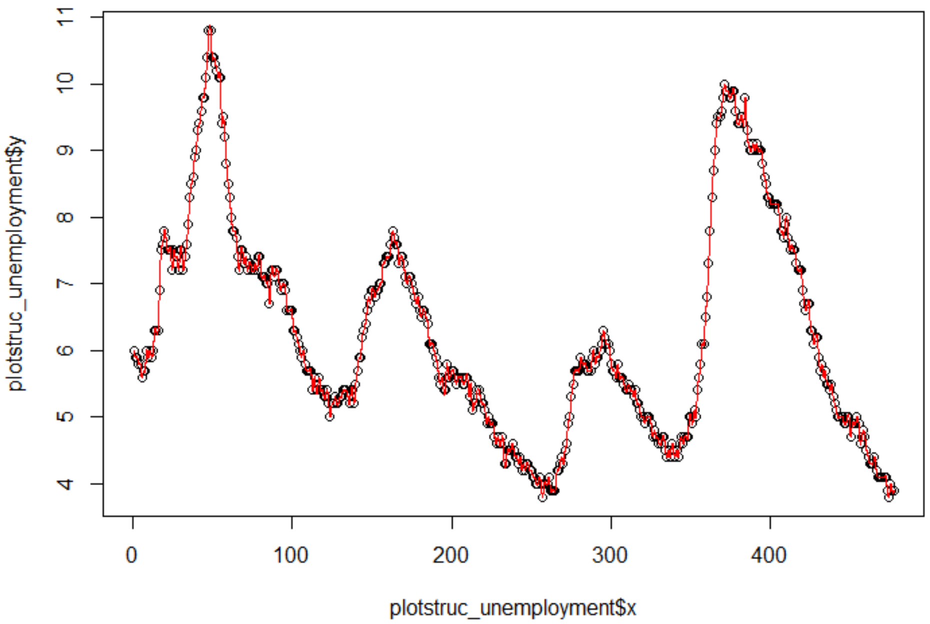
Durable goods



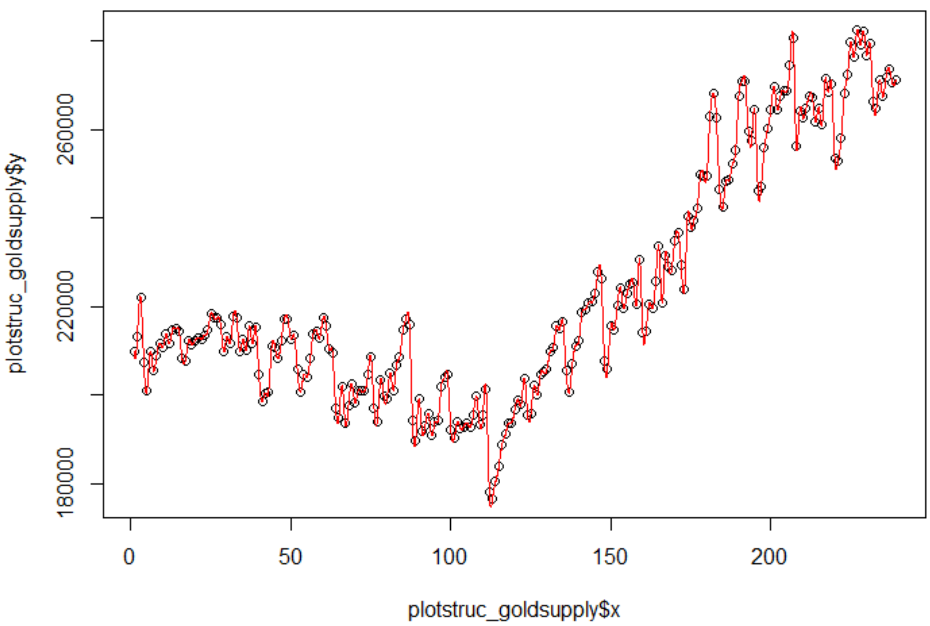
US inflation



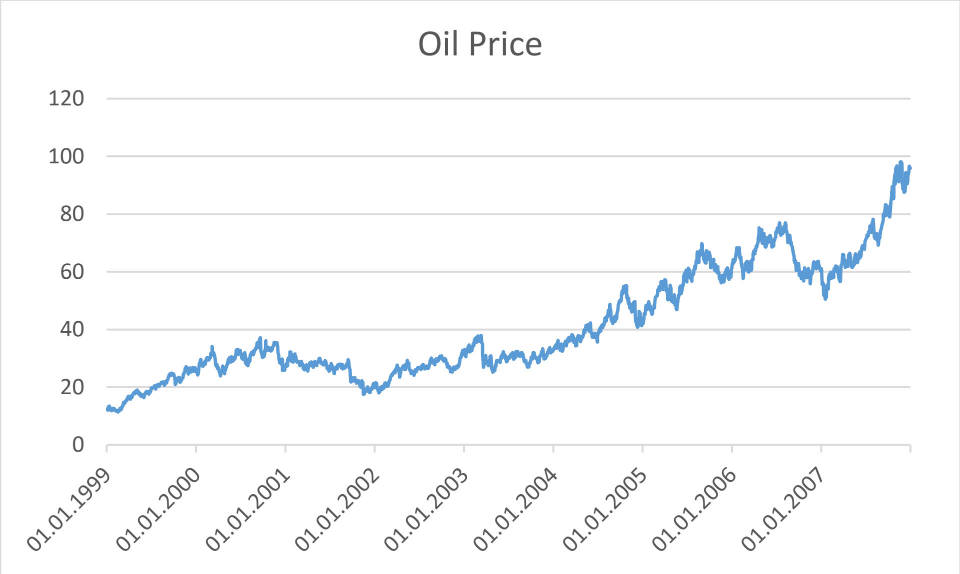
Unemployment

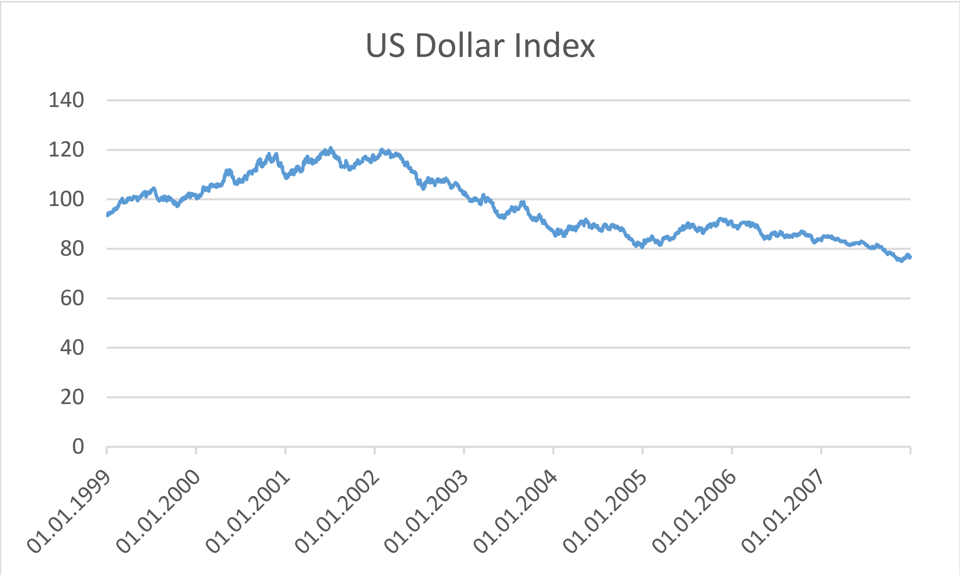


Gold supply



A2: Graphical comparisons: The gold price, crude oil and the Dollar index





A3: Correlation coefficient of the gold price and a selection of input variables

Variables	Before	During	After
dxy	-0.81	-0.37	-0.44
durgoods	0.84	-0.05	0.20
TCMNOM	-0.34	0.01	-0.45
wti	0.93	0.18	0.54

A4: Packages used in R

- `xgboost`
- `gbm`
- `stats`
- `caTools`
- `Matrix`
- `igraph`
- `perry`
- `hydroGOF`
- `tdr`

A5: R-script

The following R-script contain the majority of the code used to generate the results in this thesis. For space-saving purposes, most of the provided code is only for some example variables or subsets. This means identical code was written for the corresponding variables or subsets. This is specified in the beginning of each part of the code.

```
#Part 1 Data processing: goldprice and wti are used as examples
#Loading data
goldprice  <-  read.csv("M:/My  Documents/R/Masteroppgave/Input/Gold  price
daily 29.12.1978-08.09.2018_2.csv",
        header = TRUE, sep = ";", dec = ",")
wti <- read.csv("M:/My Documents/R/Masteroppgave/Input/wti.csv",
        header = TRUE, sep = ";", dec = ",")

#Replacing commas with dots
goldprice$Gold <- sapply(goldprice$Gold, gsub,
        pattern = ",", replacement= ".")
wti$wti <- sapply(wti$wti, gsub,
        pattern = ",", replacement= ".")

#Removing whitespace
goldprice$Gold <- gsub('\\s+', '', goldprice$Gold)
wti$wti <- gsub('\\s+', '', wti$wti)

#Converting all factor variables to numeric
goldprice$Gold <- sapply(goldprice$Gold, as.numeric)
wti$wti <- sapply(wti$wti, as.numeric)


#Part 2 Interpolation: usinflation are used as example
#Obtaining x and y coordinates for the plotting
xy.coords(usinflation$X.TIME., usinflation$X.Value.)
plotstruc_usinlation <- xy.coords(usinflation$X.TIME.,
        usinflation$X.Value.)

#Running spline with the xycoord
spline(plotstruc_usinlation, n = 10356, method = "fmm")
splineout_usinflation <- spline(plotstruc_usinlation, n = 10356,
        method = "fmm")

#Plotting new estimated observations (line) original observations (dots)
plot(plotstruc_usinlation)
points(splineout_usinflation$x, splineout_usinflation$y,
        type = "l", col = "red")


#Part 3 Data merging: goldprice and wti are used as examples
#Merge data frames by "Time"
goldprice_merge1 <- merge.data.frame(goldprice, wti, by.x = c("Time"), by.y
= c("Time"), all.x = TRUE, all.y = TRUE)
```

```
#Part 4 Data processing II: Pre-crisis subset is used as example
#Subdividing data frame on time
#Pre-crisis: 1999-01-01 -> 2007-12-31
pre_goldmerge <- goldprice_merge1[5221:7567,]
set.seed(200)
sample = sample.split(pre_goldmerge, SplitRatio = 0.75)
train_pre_goldmerge = subset(pre_goldmerge, sample == TRUE)
test_pre_goldmerge = subset(pre_goldmerge, sample == FALSE)

#Removing NAs
row.has.na <- apply(train_pre_goldmerge, 1, function(x){any(is.na(x))})
sum(row.has.na)
pre_train_goldmerge_filtered <- train_pre_goldmerge[!row.has.na,]
row.has.na <- apply(test_pre_goldmerge, 1, function(x){any(is.na(x))})
sum(row.has.na)
pre_test_goldmerge_filtered <- test_pre_goldmerge[!row.has.na,]

#Excluding specific columns
#"Time"
pre_train_goldmerge_filtered$Time <- NULL
pre_test_goldmerge_filtered$Time <- NULL

#"usdcny" (in pre_goldmerge_filtered as cny was pegged to usd until 2005)
pre_train_goldmerge_filtered$usdcny <- NULL
pre_test_goldmerge_filtered$usdcny <- NULL


#Part 5 XGBoost and GBM: Pre-crisis subset is used as example
#XGBoost
#Converting data frame into a xgb.Dmatrix as input for the xgboost model
pre_dtrain <- xgb.DMatrix(data = pre_train_goldmerge_filtered,
        label = pre_train_goldmerge_filtered$Gold)
pre_dtest <- xgb.DMatrix(data = pre_test_goldmerge_filtered,
        label = pre_test_goldmerge_filtered$Gold)

#Finding optimal number of nrounds, and running XGb-model
pre_params <- list(booster ="gbtree", eta = 0.015, gamma = 0,
        max_depth = 6, min_child_weight = 1,
        subsample = 1, colsample_bytree = 1)
pre_bstgold <- xgboost(params = pre_params, data = pre_dtrain,
        nround = 160)
pre_xgcv <- xgb.cv(data = pre_dtrain, nround = 500, nfold = 7,
        early_stopping_rounds = 8, maximize = FALSE)

#Feature importance plot
pre_impmat <- xgb.importance(model = pre_bstgold)
xgb.plot.importance(pre_impmat)
pre_impmat

#Prediction XGBoost
pred_pre_xgb <- predict(pre_bstgold, pre_dtest)
```

```
#GBM
pre_gbm <- gbm(formula = Gold ~ .,
        data = pre_train_goldmerge_filtered,
        distribution = "gaussian", cv.folds = 7,
        n.trees = 500, shrinkage = 0.0020,
        interaction.depth = 3)
#Optimising number of trees
pre_opt_trees <- gbm.perf(pre_gbm, method = "cv")
pre_opt_trees

#Feature importance plot
plot(pre_gbm, 'wti', return.grid = FALSE)
summary(pre_gbm)

#Prediction GBM
pred_pre_gbm <- predict(pre_gbm, newdata = pre_test_goldmerge_filtered,
        n.trees = 500)


#Part 6 Performance measures: Pre-crisis subset is used as example
#Mean Absolute Error (MAE)
#XGboosting (MAE)
MAE.XGb_pre <- function(actual, predicted) {mean(abs(actual- predicted))}
MAE.XGb_pre(pre_test_goldmerge_filtered$Gold, pred_pre_xgb)

#Gradient boosting (MAE)
MAE.gbm_pre <- function(actual, predicted) {mean(abs(actual- predicted))}
MAE.gbm_pre(pre_test_goldmerge_filtered$Gold, pred_pre_gbm)

#Root Mean Squared Error (RMSE)
#XGboosting (RMSE)
RMSE.XGb_pre <- rmspe(pre_test_goldmerge_filtered$Gold,
        pred_pre_xgb, includeSE = TRUE)
RMSE.XGb_pre

#Gradient boosting (RMSE)
RMSE.gbm_pre <- rmspe(pre_test_goldmerge_filtered$Gold,
        pred_pre_gbm, includeSE = TRUE)
RMSE.gbm_pre

#R-squared (R2)
#XGboosting (R2)
R2.XGb_pre <- 1 - (sum((pre_test_goldmerge_filtered$Gold - pred_pre_xgb)^2)
      /sum((pre_test_goldmerge_filtered$Gold-
      mean(pre_test_goldmerge_filtered$Gold))^2))
R2.XGb_pre

#Gradient boosting (R2)
R2.gbm_pre <- 1 - (sum((pre_test_goldmerge_filtered$Gold - pred_pre_gbm)^2)
      /sum((pre_test_goldmerge_filtered$Gold-
      mean(pre_test_goldmerge_filtered$Gold))^2))
R2.gbm_pre

#Normalized MAE and RMSE
#XGBoosting (NMAE and NRMSE)
NMAE.XGb_pre <- tdStats(pred_pre_xgb, pre_test_goldmerge_filtered$Gold)
NMAE.XGb_pre

#Gradient boosting (NMAE and NRMSE)
NMAE.gbm_pre <- tdStats(pred_pre_gbm, pre_test_goldmerge_filtered$Gold)
NMAE.gbm_pre
```