**Forecasting Gold Prices: Evidence from Support Vector Machines   
and Neural Networks**

HAYK MKRTCHYAN (ID: 300507458048) and NORAYR HAYRUNI (ID: 300007457843)

**ABSTRACT**

The price of gold has always played a key role in the financial market since it is considered to be a hedge by investors against many unprecedented events. This paper aims to predict the gold price return over the short- and long-term horizon. We have applied Catboost, Light Gradient Boosting Machine and SVR to forecast the gold price. In addition, we aim to understand the gold price movement relative to its base price over the 1- and 12-month horizon. For that classification problem, we applied SVM and ANN.

**Keywords:** Support Vector Machine, Catboost regression, Light Gradient Boosting Machine, Neural Networks, Gold Price

# **1. Introduction**

The gold price has followed an upward trend in the past decade (see **Figure 1** in Appendix). It has reached its historical high price of more than 2000 USD per ounce in August 2020 (Yahoo Finance). During the Corona pandemic, gold has gained high interest among potential investors by indicating that the metal is considered to be a hedge during uncertain times. Therefore, accurate prediction of the gold price can be used by private households and financial institutions to prevent or mitigate risks and reduce potential financial losses.

The traditional methods of predicting the price of gold include Autoregressive Integrated Moving Average (ARIMA) and Multiple Linear Regression (MLR) models. With the development of soft computing, Artificial Neural Networks (ANN) and Machine Learning Algorithms (ML) have also become popular among researchers to predict the gold price.

This paper aims to predict the price of gold by using ML and ANN algorithms. The paper is divided into five sections. Section 2 discusses the literature related to this. Section 3 describes the data and data preprocessing and presents the methodology applied in this paper. Section 4 provides the results, and section 5 concludes.

# **2. Literature Review**

There are many studies in the literature about the prediction of the gold price by applying various models of ML and ANN. Although ML and ANN are extensively used for predicting the gold price, other methods such as ARIMA and MLR are also being applied. Factors affecting the gold price are different among the studies. Also, the method of selecting those factors differ widely. However, there are a bunch of similar factors used by academicians to predict the price of gold. Some of the studies on predicting the price of gold are being discussed in the following sections.

Ismail, Yahya and Shabri (2009) applied the MLR method for forecasting the gold price. The price of gold used in their study was from London PM Fix. They identified different economic factors to construct their model. Features such as Commodity Research Bureau future index (CRB), Money Supply (M1), EUR/USD foreign exchange rate, Inflation Rate (IR), New York Stock Exchange Composite Index, Treasury Bill, US Dollar index, S&P 500 Index and others have been used as predictors for the gold price. The authors created two models. The first model included all predictors and found that the model explained 85.2% of sample variations in monthly gold prices. The second model performed better in terms of predictive accuracy. It explained 70% of sample variation and only included four variables; CRB, EUR/USD, IF and M1. Mean Square Error (MSE) was selected as an accuracy criterion.

Ahmed, Ping, Yaziz and Miswan (2015) investigated the hybrid model of ARIMA and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) to predict Malaysian gold price. Akaike Information Criterion (AIC) and Mean Absolute Percentage Error (MAPE) were used for the goodness of fit and forecasting performance. The results showed that a hybrid model of ARIMA-GJR (a variant of GARCH), designed to model asymmetric volatility, outperformed the ARIMA model. On the other hand, Sarangi and Dublish (2013) applied ANN and GARCH models and concluded that ANN performed with lower errors and provided better predicting power.

More recent research about predicting the gold price applies more complex models. Dubey (2016) used daily data ranging from 2007 to 2015 from the official bullion of Australia, Perth Mint, to predict the gold price. The author created models using Support Vector Machine (SVM) and Adaptive Neural Fuzzy Inference System (ANFIS). The three developed models concluded that the predictive accuracy of the Support Vector Regression (SVR) method with RBF kernel function is better than ANFIS. SVR performed better based on the evaluation criteria; Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and MAPE.

In addition, Ongsritrakul and Soonthornphisaj (2003) compared SVR with linear regression and neural networks based on the predictive accuracy of the gold price. They used daily data from 1998 to 2002, overall, 1000 observations. Factors such as exchange rates of South Africa, Canada, Australia, S&P 500 index, US Dollar index, Bond index, 30- and 10-year bond yields, and others were used to predict the gold price. Also, the authors applied decision trees to the feature selection task to increase the performance of the prediction. They found that neural networks perform weakly in comparison to the other models applied in their paper. Also, they concluded that the combination of SVR and decision tree leads to better predicting power.

Potoski (2013) applied Logistic Regression (LR) and SVM in an attempt to predict whether the price of gold will rise or fall compared with the current day's price. The author used daily data ranging from 2007 to 2013 from London PM Fix gold price. The first results of the models, where the author used only one factor, the previous price of gold, were disappointing. The accuracy for both models was below 60%. Later on, Potoski tried to include other factors in the model, such as stock indices, rate of change from open to close price, and others, to increase the models' performance. However, the results were still disappointing. After several tries, the author included variables that are most correlated with the gold price using Pearson correlation and increased the predictive ability of the models. The final model included variables such as Silver, Oil, Copper, Dollar index, Hang Seng Index and Nikkei 225, and 3 exchange rates; EUR-USD, GBP-USD, USD-CNY. Finally, the LR model showed 69.3% of accuracy and SVM showed 69.08%.

# **3. Methodology and Data Description**

## 3.1 Data description and data preprocessing

We examine the relationship between gold and different economic factors extensively used in the literature. Our study uses daily data for all the factors as well as for the per ounce of gold ranging from the 01st of January 2010 to the 01st of January 2020, overall, 2609 observations. Literature provides an extensive list of economic factors that affect the gold price. The factors that we use in our paper are the following; Silver price, Crude Oil price, Platinum price, Copper price, Soybean Futures, S&P 500 Index, Russel 2000 Index, 10 Year US T-Note futures, Nasdaq Composite Index, Euro-USD exchange rate, New York Stock Exchange Composite Index (NYSE), Euronext 100 Index and MSCI Emerging Markets ETF. All data is obtained from Yahoo Finance[[1]](#footnote-1). The missing values contained in the dataset were replaced by the most previous observations.

To develop a model, we first predict the gold return and then calculate its actual price. Thus, we first calculate and include short-term (7, 14 and 21 days) and long-term (60, 90, 180 and 250 days) returns of selected factors in the models. Also, simple and exponential moving averages of the gold price is included in the models.

The correlation matrix (see **Figure 2**) shows the relationship between gold and the other 13 factors. The highest positive correlation is observed between gold and silver, which accounts for 83%, and the highest negative correlation is between gold and Euronext 100 index, which is -49%. From figure 2, we can observe that gold has a negative relationship with all equity indices.

## 3.2 Methodology

This paper aims to apply SVM and ANN for creating robust models for predicting the gold price. In addition to that, Catboost regression and Light Gradient Boosting Machine analysis have been implemented as well. Since we aim to predict the continuous gold price and a discrete label output, both regression and classification have been used.

**Support Vector Machine**

SVMs have been extensively applied in the area of machine learning for the classification of supervised learning framework. SVM was first developed by Cortes and Vapnik (1995). In our paper, we apply SVM and SVR. SVR uses the same principle as SVM, but it deals with a regression problem – predicting a continuous variable rather than a discrete label.

SVM has a number of kernels – Linear, Nonlinear, Polynomial, Gaussian, Laplace, Sigmoid, RBF, Hyperbolic tangent and others. Each has its applications in numerous business problems. However, the most common kernels applied in the literature are – linear, RBF and Polynomial (see **Table 1**).

**Table 1.** Kernel Functions

|  |  |
| --- | --- |
| **Kernel** | **Function** |
| Linear |  |
| RBF |  |
| Polynomial |  |

The primary idea of SVM is to find a line or a hyperplane which would be able to separate the data into classes. Unlike the OLS simple regression, which tries to fit the best line to minimize the errors, SVR tries to establish a tube with a width of epsilon vertical to the axis. The tube is called ε-insensitive tube, and the points in our dataset that fall in this tube will not be included in the error computation. In other words, it is the margin of error that we are allowing our model to have. At the same time, the points that are outside the ε-insensitive tube should be taken into account. Therefore, the goal of SVM is to minimize the distance between the points outside the tube and the tube itself. In equation 1, epsilon and epsilon-star are the errors above and below the tube, respectively.

**Catboost Regression**

The Catboost ML algorithm was developed by the Russian Yandex Company in 2017. The primary advantage of Catboost over other ML algorithms is its capability to integrate a number of data types in a single framework.

Prokhorenkova, Gusev, Vorobev, Dorogush and Gulin (2018) have demonstrated that, unlike many ML algorithms that have difficulties handling non-numeric values, Catboost can handle them with a minimum of categorical feature transformation. A similar transformation from a non-numeric to a numeric state of our variable of interests can often be highly complicated with most ML algorithms. Catboost offers a gradient boosting framework and is built on the knowledge of decision trees. Another feature of the Catboost algorithm is its ability to create oblivious trees, unlike gradient boosting models. It also prevents overfitting due to its ordered boosting. For more technical explanation, see Prokhorenkova et al. (2018).

**Light Gradient Boosting Machine**

Light gradient boosting algorithm is originally developed by Microsoft. It also offers a gradient boosting framework that uses tree-based algorithms. LGBM is different to other gradient boosting algorithms in a sense that the trees are grown in a leaf-wise or vertically and not level-wise manner. This key feature enhances the speed of training of the dataset, and it requires lower memory usage. Ke et al. (2017) have shown that Light GBM is integrated with two main innovative techniques – One side sampling and exclusive feature bundling. The latter is of utmost significance when the dataset contains a large number of features. For more technical discussion, see Ke et al. (2017).

**Artificial Neural Network**

ANNs have been broadly used in many disciplines, such as system identification, pattern recognition, chemistry, face identification etc. An ANN is based on collections of numerous nodes which function like the neurons in the human brain. Each neuron has the capacity to transmit some signals to other neurons, and while transmitting, it assigns weights which are crucial in a sense that the neural network decides in every single case what single signal is important and what signal is not important to a certain neuron. These weights are being adjusted through the process of learning while minimizing the cost function. The ultimate goal of this algorithm is to find the output of the neuron, which is being done by taking the weighted sum of all inputs. The weighted sum then passes through the activation function to get the output. The backpropagation technique is used in order to find the optimal weights – that is, to minimize the error or the cost function in equation 2.

# **4. Results and Discussion**

**Regressions**

As for our regression analysis, our goal is to predict the return of gold, hence its price, over the 30-day horizon. For this purpose, we have implemented SVR with the radial kernel, Catboost and Light Gradient Boosting regressions to forecast the return of gold. MSE, RMSE and MAE (see equations 3-5) metrics have been used to evaluate the performance of our models. In addition, 10-fold cross-validation has been integrated into our regressions. One can notice that in terms of MSE, RMSE and MAE criteria, Catboost performed better than SVM and LGBM (see **Table 2**).

In **Table 4** (see Appendix), one can see the prediction results of our models over the 30-day horizon. For instance, on the 18th of January 2021, the gold price was 1829.3, and the SVM, Catboost and LGBM have forecasted it to be 1882.9, 1874.2 and 1894.2, respectively.

**Table 2.** Performance of the prediction models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Catboost | SVM | LGBM |
| MAE | 0.0151 | 0.0350 | 0.0167 |
| RMSE | 0.0210 | 0.0451 | 0.0232 |
| MSE | 0.0005 | 0.0020 | 0.0006 |

**SVM**

Furthermore, we apply SVM with the linear kernel to predict whether the current gold price is higher/lower than the 1-month and 12-month price in future[[2]](#footnote-2). A common practice for assessing the performance of a classification model is to calculate the accuracy and the precision of the model. The formulas are as follows. Accuracy = (TP + TN) / (TP + TN +FP + FN)[[3]](#footnote-3), Precision = TP/(TP + FP). Accuracy shows the correct predictions relative to all possible outcomes. Precision shows the rate of true positive predictions relative to the sum of true and false-positive predictions. We use 10-fold cross-validation (CV) for obtaining robust results. Also, following Viebig (2019), we also apply SVM with RBF kernel to ensure that the results are robust to different model specifications. The results are presented in **Table 3**. For the 1-month prediction, the best-obtained accuracy is 77.1% with the tuned RBF kernel. For the 12-month prediction, the best result is 91.95% with the linear kernel (all results are for KFOLD=10). It can be observed from **Table *3*** that prediction accuracy for the SVM with linear kernel increases from 76.9% to 91.95% (around 15% improvement) when the time horizon changes. This may suggest that SVM learn the pattern of gold price and better predict the long-run horizon. In addition, two different cross-validations have been applied to make sure that the SVM methodology is robust to changes in model specifications. There is a slight difference in the accuracies between the SVM with linear and RBF kernels. Overall, the accuracy and precision rates remain high for alternative resampling methods. So, it can be concluded that the SVM is a robust methodology for predicting the gold price.

**Table 3.** SVM results with different kernels and time horizons

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SVM | 1-month | 1-month | 12-month | 12-month |
| KFOLD 10 | **Accuracy** | **Precision** | **Accuracy** | **Precision** |
| Linear Kernel | 76.90% | 76.92% | 91.95% | 94.05% |
| RBF Kernel | 77.10% | 77.17% | 91.37% | 94.25% |
| KFOLD 5 | **Accuracy** | **Precision** | **Accuracy** | **Precision** |
| Linear Kernel | 76.47% | 77.50% | 92.54% | 94.59% |
| RBF Kernel | 76.32% | 75.37% | 90.49% | 93.20% |
|  |  |  |  |  |

**ANN**

In addition, we have carried out ANN to predict the gold price movement over a 1-month horizon. We divided the dataset into training and testing sets as before when implementing SVM algorithms. For the implementation of our ANN, we have chosen two hidden layers, each composed of six nodes. As for the activation functions for our hidden layers, we will stick with the most common rectifier function, which takes either 0 or 1. However, we decided to apply the sigmoid function (see equation 6) to produce the output. It is worth noting that the sigmoid function is between 0 and 1, so it will return us probabilities. Hence, we will consider all values that are greater than 0.5 to be 1, which means that the gold price will increase relative to its current price, and otherwise 0.

To evaluate how well our model forecasts the gold price, we will use the accuracy and precision metrics. The model has achieved around 90% accuracy over a 1-month horizon. The precision of the model is similar to the accuracy and compiles 89%. One can observe that ANN better predicts the gold price than SVM over a 1-month horizon in terms of accuracy performance metric.

# **5. Conclusion**

This paper has implemented numerous ML algorithms utilizing both regression and classification analysis to predict the gold price. Relatively new Catboost regression outperformed SVR and LGBM in terms of MSE evaluation metric. As for forecasting the relative gold price over a 1- and 12-month horizon, we applied SVM and ANN. It has been shown that using 10 cross-validations, the accuracy rate of SVM with linear kernel over a 1-month horizon is 76.9%. In addition, RBF SVM had a slightly improved performance in terms of accuracy metric (77.1%). SVM, with its linear and radial kernels, demonstrated the same performance for 12 months horizon as well. We have found out that for a 1-month prediction horizon, ANN performed better, and namely, the accuracy is 90% for the test set. Thus, ANN makes better prediction for the gold price over a 1-month horizon.

# **REFERENCES**

Ahmad, Maizah, Yean Ping Pung, Siti Roslindar Yaziz, and Nor Hamizah Miswan, 2015, Forecasting Malaysian gold using a hybrid of ARIMA and GJR-GARCH models, *Applied Mathematical Sciences 9*, 1491-1501.

Cortes, Corinna, and Vladimir Vapnik, 1995, Support - vector networks, *Machine Learning* *20*,273–297.

Dubey, Akash, 2016, Gold price prediction using support vector regression and ANFIS models, *International Conference on Computer Communication and Informatics (ICCCI),* 1-6.

GC=F historical prices, Gold the 21st of August. (n.d.). Retrieved the 05th of June, 2021 from https://finance.yahoo.com/quote/GC=F?p=GC=F&.tsrc=fin-srch

Ismail, Zuhaimy, A. Yahya, and A. Shabri, 2009, Forecasting Gold Prices Using Multiple Linear Regression Method, *American Journal of Applied Sciences 6*, 1509-1514.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, 2017, LightGBM: a highly efficient gradient boosting decision tree, *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

Ongsritrakul, Pedrudee, and Nuanwan Soonthornphisaj, 2003, Apply decision tree and support vector regression to predict the gold price, *Proceedings of the International Joint Conference on Neural Networks 4*, 2488-2492.

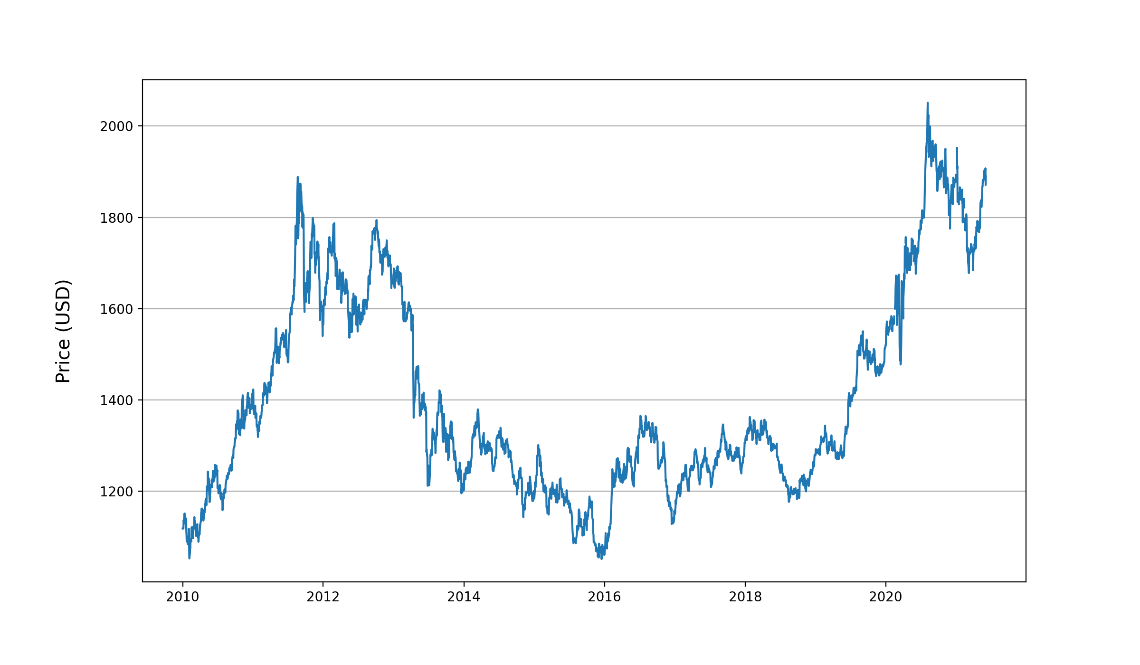
Potoski, Megan, 2013, Predicting Gold Prices, CS229.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, 2018, CatBoost: unbiased boosting with categorical features, *In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 6639–6649.

Sarangi, Prasant, and Shashank Dublish, 2013, Prediction of Gold Bullion Return Using Garch Family and Artificial Neural Network Models, *Asian Journal of Research in Business Economics and Management* *3*, 217-230.

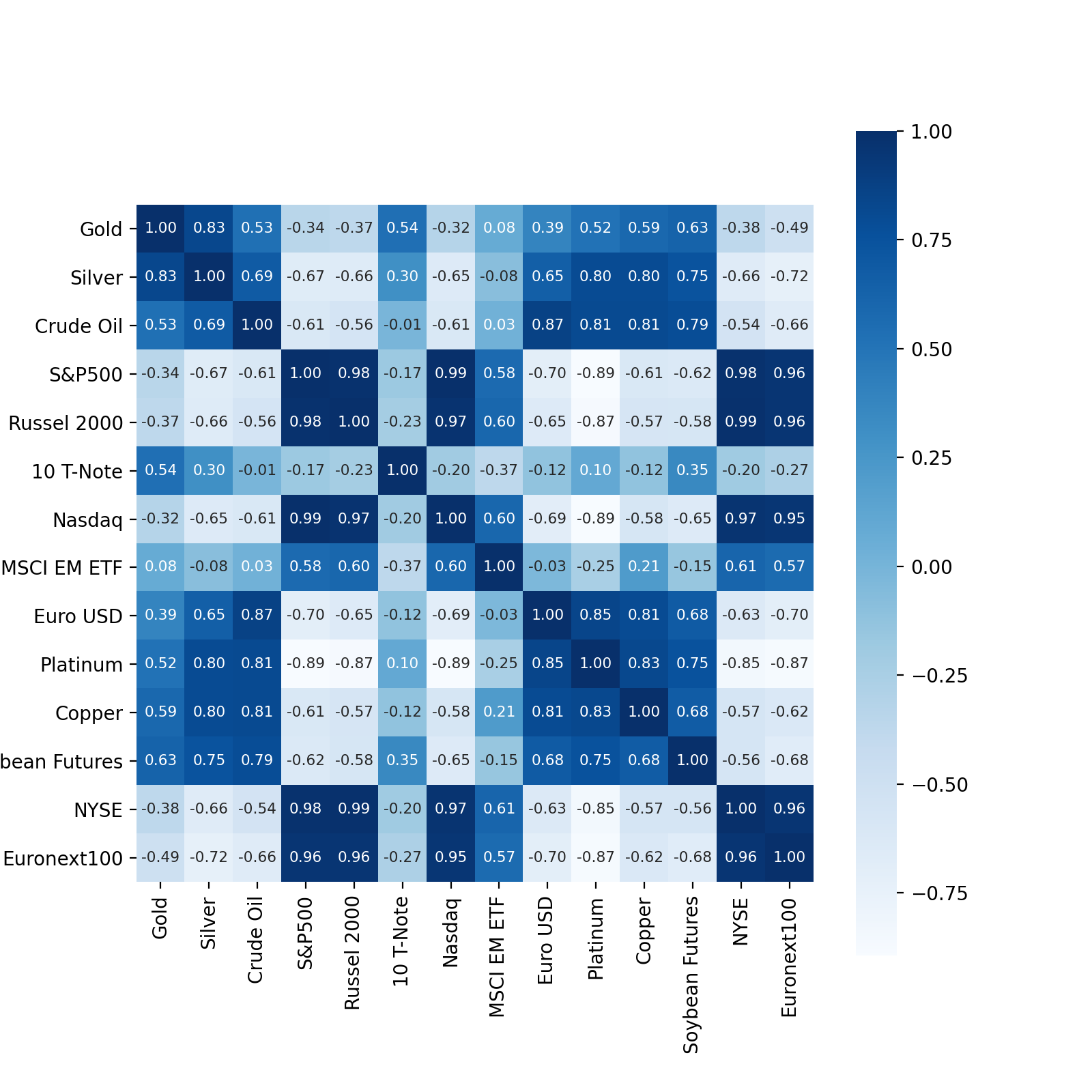
Viebig, Jan, 2019, Exuberance in Financial Markets: Evidence from Machine Learning Algorithms, *Journal of Behavioral Finance 21*, 1-8.

# **Appendix**



**Figure 1.** Price of Gold (per ounce) from 2010 to 2021.

**Figure 2.** Correlation Matrix.



**Table 4.** Actual and predicted Gold prices for January 2021

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Actual** | **SVM** | **Catboost** | **LGBM** |
| 2021-01-18 | 1829.3 | 1882.8 | 1895.4 | 1921.5 |
| 2021-01-19 | 1839.5 | 1881 | 1861.3 | 1911 |
| 2021-01-20 | 1865.9 | 1874.4 | 1863.7 | 1844.3 |
| 2021-01-21 | 1865.3 | 1862.3 | 1877.4 | 1852.6 |
| 2021-01-22 | 1855.7 | 1869.4 | 1881.7 | 1854.2 |
| 2021-01-25 | 1854.9 | 1869.4 | 1869.6 | 1854.1 |
| 2021-01-26 | 1850.7 | 1868.7 | 1861.7 | 1843 |
| 2021-01-27 | 1844.9 | 1870.5 | 1846.1 | 1838.9 |
| 2021-01-28 | 1837.9 | 1873 | 1842 | 1850.1 |
| 2021-01-29 | 1847.3 | 1884.3 | 1856.6 | 1826.7 |

1. For all variables Adj Close price was selected as a final price. [↑](#footnote-ref-1)
2. We classify the gold price as follows. If the price of gold at time is higher than the price of gold at time then {1} is assigned to the price of gold at time , and {0} is assigned on the other cases. [↑](#footnote-ref-2)
3. TP – true positives, TN – true negatives, FP – false positives, FN – false negatives [↑](#footnote-ref-3)