Capstone Project 2 - Milestone Report 2

1. Introduction

The purpose of this Capstone Project is to test and evaluate machine learning and deep learning methods in performing prediction for gold prices in USD (also known as XAU/USD).

Background

Finance is at the forefront of technology, as banks and investment firms have invested large amounts of money in the latest technologies, from High-Frequency Trading which allows them to perform trades in milliseconds, to the use of Machine Learning and Deep Learning algorithms, both to predict price trends and the impact of news events on the market, among other applications.

In trading, there are two major types of analysis performed by traders and investors:

- Fundamental Analysis - This type of analysis focuses on analysing general economic indicators (economic growth, inflation, interest rates, etc), as well as, news events which can affect a particular trading instrument (company stock, futures, currency pair, etc).
- Technical Analysis - Technical Analysis, on the other hand, focuses more on trying to identify possible patterns or signals in the data of a particular instrument, in order to predict whether it will go up or down in price.

Although fundamental analysis is more widely used, particularly among institutional investors, there are traders who also rely on Technical Analysis for their investment strategies.

In academic circles, Technical Analysis is mostly dismissed as being equal to making future predictions based on chance, however, Andrew Lo, an MIT professor, who initially dismissed Technical Analysis, later published a paper titled "Foundations of Technical Analysis:Computational Algorithms, Statistical Inference, and Empirical Implementation"[1], in which he and his co-authors concluded that "several technical indicators do provide incremental information and may have some practical value."

Project Objectives

The project will focus on the price of gold and its purpose will be to perform two major tasks.

[1] Paper: https://www.cis.upenn.edu/~mkearns/teaching/cis700/lo.pdf

First, the project will analyze gold price data in order to determine if there are any trading patterns, for example:

1. Does the price trend in particular time intervals, such as particular hours, days or months?
2. Does the price exhibit any particular patterns repeatedly?

The second task of this project will be to utilize different machine learning and deep learning algorithms in order to predict future returns. The purpose in doing so will be to not only learn different machine learning approaches, but also to be able to compare them to deep learning approaches and then draw conclusions on their performance.

The project will be focusing only on historical gold price daily data. However, if necessary, the inclusion of other data sources which are relevant, such as stock index data, currency data and others, will be taken into consideration.

Even though trading a highly-competitive field, it is one which I have a particular interest in, as both a small investor and as a former MBA student who took several finance courses. Additionally, the practice that pursuing a project perform analysis on time series, will provide me with skills which can be applicable in other areas where time series and trend analysis is necessary.

2. Data Wrangling

Overview of Main Datasets

The data sets that were used in the project were the following:

- XAUUSD Price Data - This is the main dataset with historical Gold Price data from May 5th, 2003 up to April 30th, 2019
- XAGUSD Price Data - This is an additional dataset with historical Silver Price data from August 8th, 2003 up to April 30th, 2019
- EURUSD Price Data - This is an additional dataset with historical Euro/Dollar exchange rate data from May 5th, 2003 up to April 30th, 2019

Although there is an abundance of financial data available, the existence of free quality data is limited, in particular, if one is looking for data at either the tick or minute level.

After extensive research, the best means of obtaining the relevant data for the project was through QuantDataManager[2], a software tool which provides direct access to data from Dukascopy, a Swiss broker.

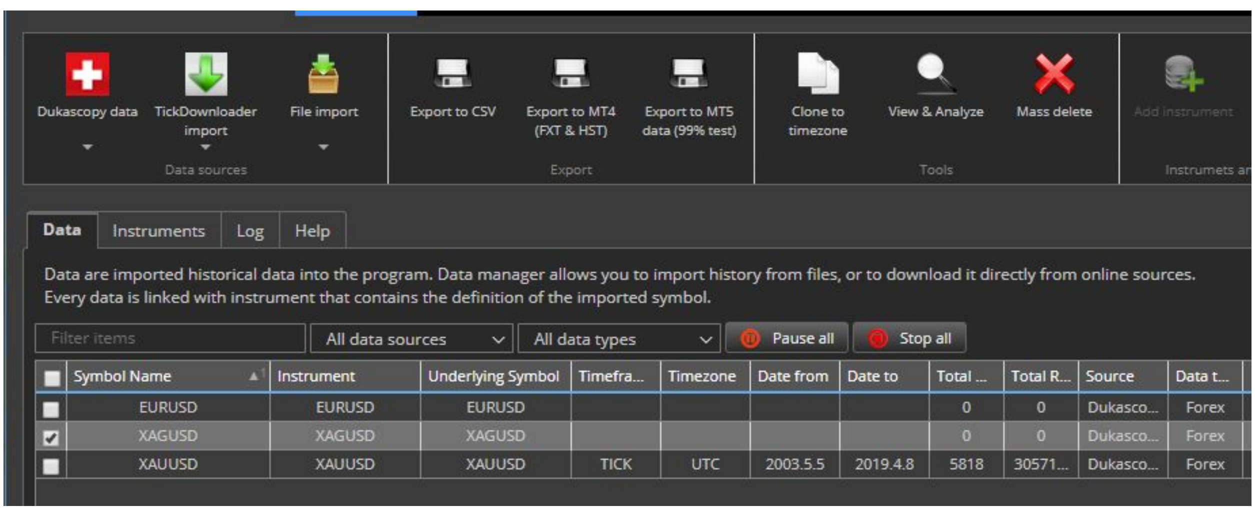

Figure 1 - Screenshot from QuantDataManager

The software allows for the download of historical tick data for a variety of instruments. If the data is downloaded at the tick level, the software then allows it to be saved in various timeframes, from minute data up to daily data.

[2] QuantDataManager - https://strategyquant.com/quantdatamanager/

The datasets were of excellent quality and did not require cleaning up. However, some of the historical data was incomplete in smaller time frames (1 Min). This may have to do with the fact that either the data did not exist (as data could have been refreshed only when it changed price) or the data provider did not have it available.

Data Import Process

The Data Import process was very straightforward as the software provided the option to download the data into various timeframes and save it as a CSV file.

Upon downloading the datasets at the tick level, they were saved in different timeframes, ranging from 1-Minute up to Daily data. Although only daily data was used in this project, this was done in order to have the data available for future work.

Data Cleaning Process

The data cleaning process involved just reviewing it using the appropriate pandas commands (info and describe) and performing a quick visual inspection.

The output of the info command on the daily dataset for XAUUSD was as follows:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 4961 entries, 2003-05-05 to 2019-04-08
Data columns (total 5 columns):
Open      4961 non-null float64
High      4961 non-null float64
Low       4961 non-null float64
Close     4961 non-null float64
Volume    4961 non-null int64
dtypes: float64(4), int64(1)
memory usage: 232.5 KB
```

The output of the ‘describe’ method was as follows:

	Open	High	Low	Close	Volume
count	4961.000000	4961.000000	4961.000000	4961.000000	4961.000000
mean	1047.044227	1053.937454	1039.867448	1047.199843	61623.476718
std	400.377123	402.381432	397.928167	400.255074	58346.230620
min	340.350000	342.590000	339.000000	341.230000	35.000000
25%	663.470000	668.340000	659.730000	664.150000	13977.000000
50%	1185.510000	1192.260000	1178.610000	1185.520000	46770.000000
75%	1309.640000	1315.530000	1302.540000	1309.600000	92550.000000
max	1909.110000	1920.660000	1879.570000	1909.080000	398293.000000

Table 1 - Output of the describe method for the daily XAUUSD price dataset

Data Wrangling Conclusions

Overall, all datasets were of very good quality, with no need to perform cleanup.

Data in the smaller time frames would require to be filled in in some spots, however, this was left to do, if and when necessary, according to the needs of the project.

3. Data Story

Initial Analysis

The initial analysis of the dataset focused on understanding the trends and patterns of the time series. Although the dataset provides Open, High, Low and Close prices, we chose to focus first on the Close prices for each day. A plot for the daily Close price is shown below:

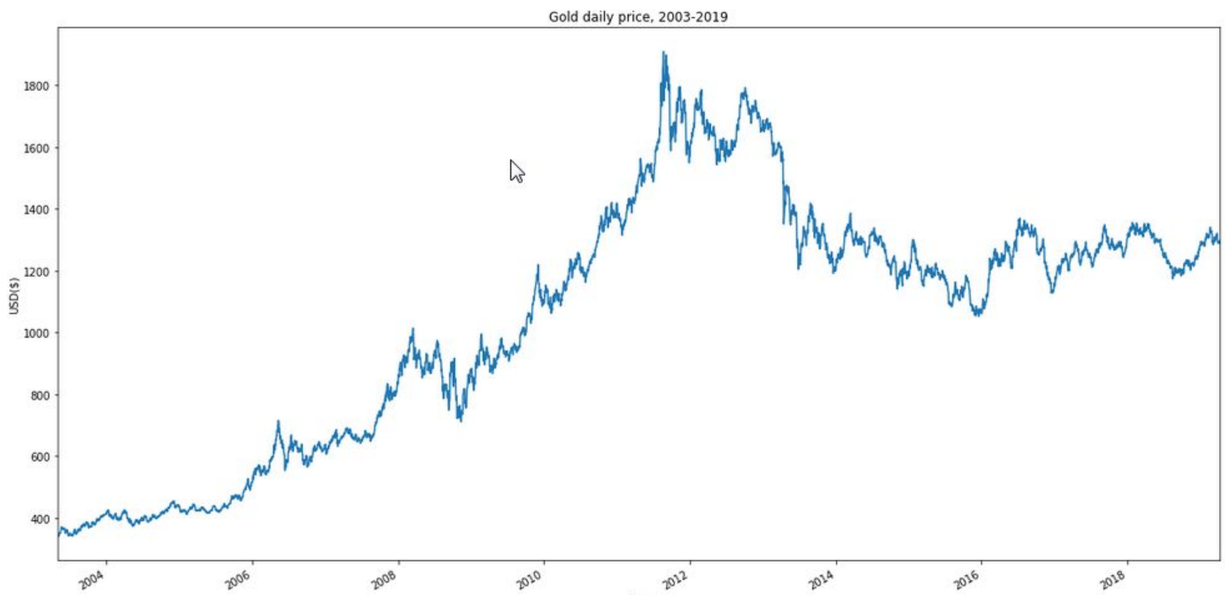

Figure 2 - Daily Close prices for Gold (May 2003 - April 2019)

As it can be seen in Figure 2, shown above, Gold prices were on an upward trend from at least 2003 to 2011, where they reached their peak at around 1900 $/oz. After hitting that peak value, the prices have slid down and have hovered between 1200-1400 $/oz.

Figure 3 shows the daily returns for the same time period:

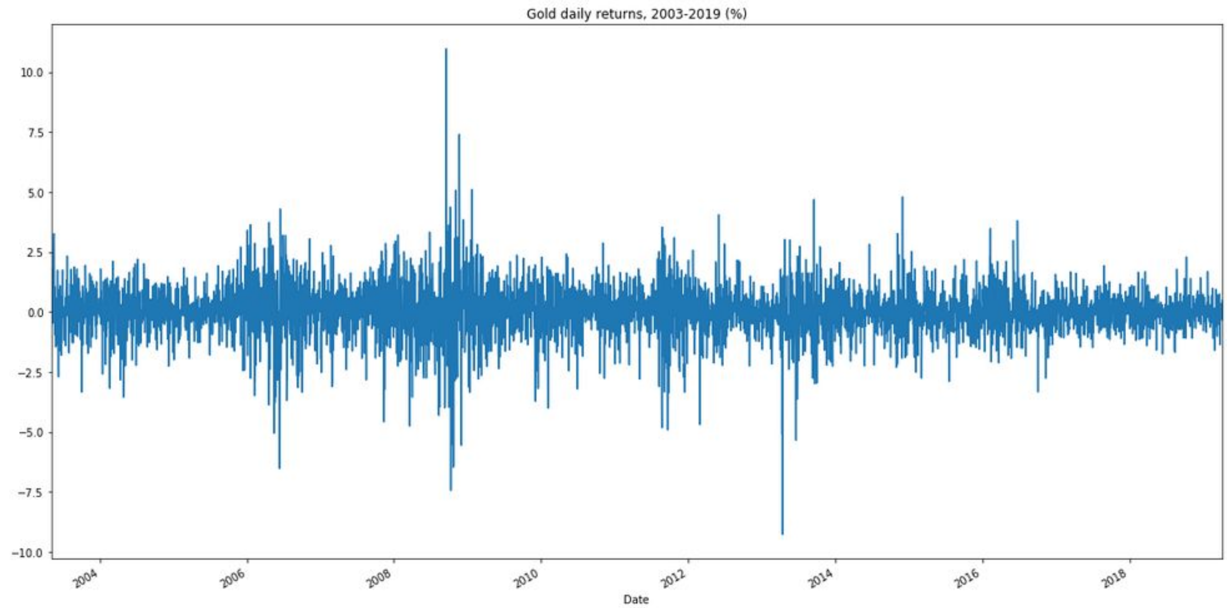

Figure 3 - Daily returns for Gold (May 2003 - April 2019)

The daily returns plot shows a different perspective. As we can see, the highest daily returns occurred in 2008, most likely, around the time when the 2008 economic crisis began, as investors moved their funds from the stock market into Gold.

The above plots were also repeated for Silver and Euro data. However, these instruments show different trends than Gold, as illustrated below:

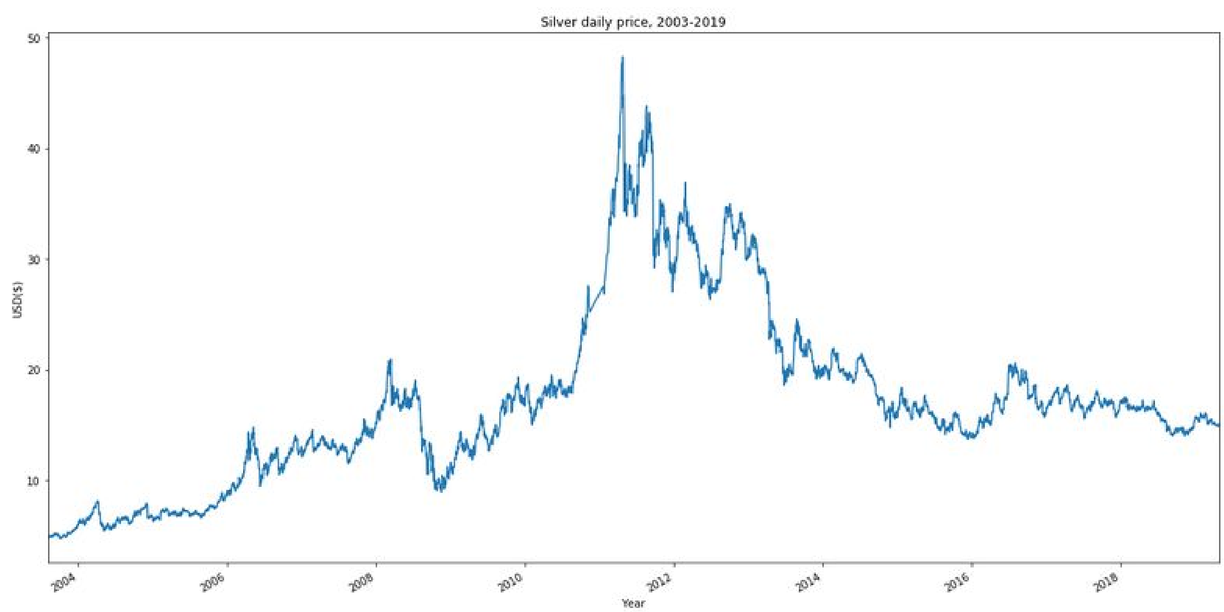

Figure 4 - Daily Close prices for Silver (August 2003 - April 2019)

As we can see in Figure 4, the price of Silver behaved somewhat differently than Gold. It also had a relative peak in 2008, followed by a steep decline in 2009, and then it peaked around 50 $/oz in 2011, followed by a decline in the years after.

Although Silver is a precious metal just like Gold, it is also widely used in a variety of industrial applications, which explains its different behavior. The peak in 2011 occurred due to a fear that there would be a shortage of silver, due to the increased industrial demand[3], as the world economy got back on the recovery path from the 2008 economic recession. However, these fears did not materialize and the price corrected and the price of Silver has since come down to its current level around 15 $/oz.

The daily returns for Silver are shown below:

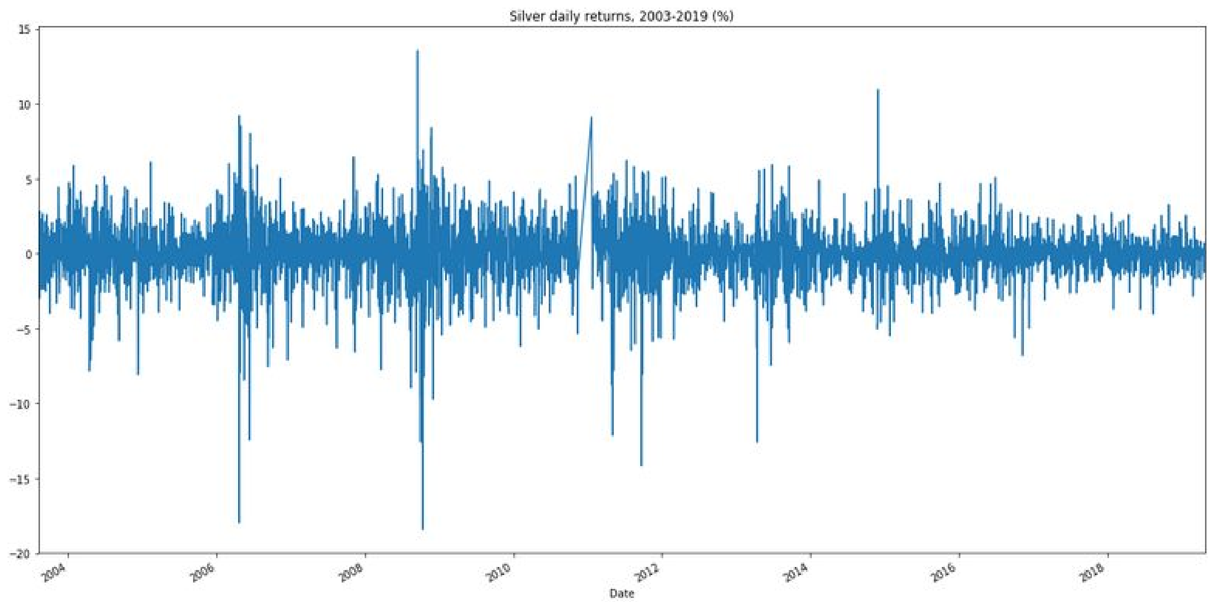

Figure 5 - Daily returns for Silver (August 2003 - April 2019)

The daily Silver returns plot shows greater volatility in Silver prices, however, some of the larger price movements seem to coincide with those of Gold, namely, during the 2008 crisis. The plot also shows a gap in the dataset in 2011, which will be dealt with when the data is integrated for further analysis.

The next plot shows the price trend for the Euro/USD exchange rate:

[3] “Silver $50: Three Years After the "Shortage" - https://www.bullionvault.com/gold-news/silver-2011-042320145

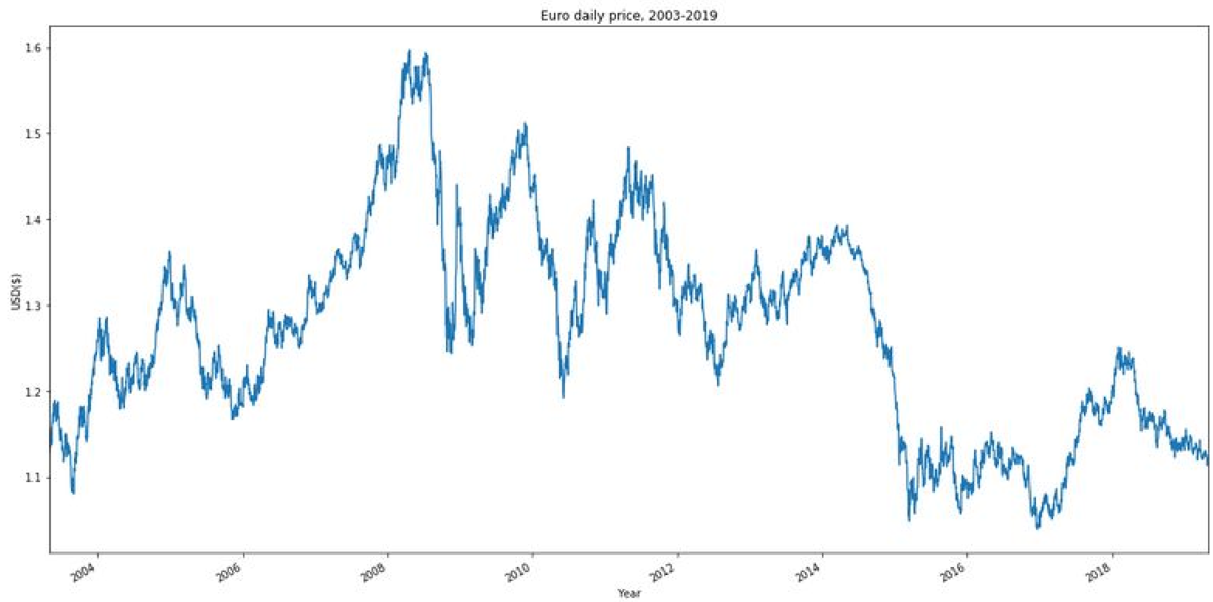

Figure 6 - Daily Close prices for EUR/USD (May 2003 - April 2019)

The Euro price trend displays some similarities with the trends of precious metals. Again, we have an all-time high of 1.60 USD/EUR in 2008, with a couple of other peaks occurring in 2009 and in 2011, followed by a steep decline around 2015.

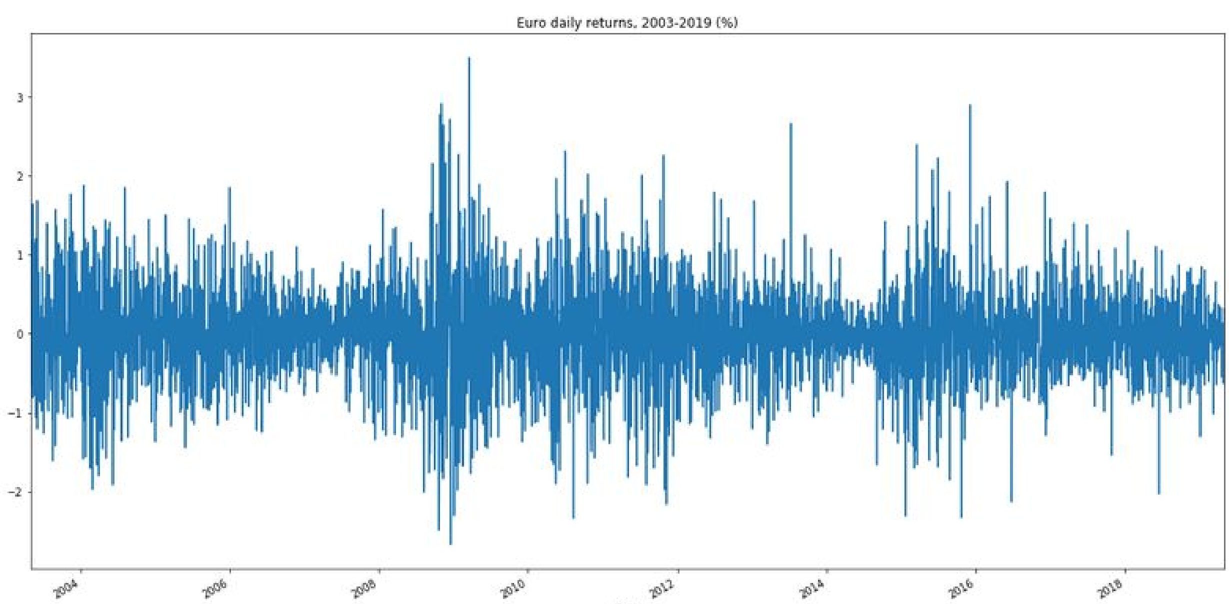

Figure 7 - Daily returns for EUR/USD (May 2003 - April 2019)

The daily return chart shows again some volatility taking place in 2008, however the range of daily returns is much narrower, not exceeding 4% in both directions.

First Analysis Conclusions

Based on this first analysis, we saw that there is some correlation between Gold prices and Silver and Euro prices. Major economic events such as the beginning of the 2008 economic recession affected these 3 instruments, as they reached peak levels in both price and volatility of their daily returns.

However, it will be necessary to further infer whether both Silver and the EUR/USD exchange rate can provide beneficial input into determining the return of Gold. This analysis will continue in the next sections of the report.

4. Inferential Statistics

Analysis of Gold returns

After performing the initial exploratory data analysis of Gold, Silver and the Euro, the next part of the project involved taking a closer look at the daily returns of Gold.

First, some summary statistics of the daily return values are presented below:

	Daily_return							
	count	mean	std	min	25%	50%	75%	max
Weekday								
Sunday	818.0	0.056797	0.319068	-1.358777	-0.096334	0.025627	0.176309	1.925301
Monday	834.0	-0.042003	1.017094	-9.275493	-0.448975	0.008109	0.436255	4.796461
Tuesday	835.0	-0.007508	1.102825	-6.524084	-0.584146	0.024008	0.572896	5.069485
Wednesday	833.0	0.047511	1.119512	-6.466444	-0.484326	0.071388	0.640283	10.972228
Thursday	834.0	0.002799	1.125786	-5.510074	-0.568201	-0.017806	0.581797	4.288629
Friday	824.0	0.134767	1.129819	-7.436854	-0.409948	0.141585	0.747231	7.401234

Table 2 - Output of the describe method for XAUUSD daily returns (%)

As we can see in Table 2, the maximum daily return values occurred on Wednesday and Friday, while the minimum daily return values took place on Monday and Friday. Although both maximum and minimum may have been events that only happened once, the daily return values for the 75% percentile, are also the highest on both Friday and Wednesday. On the other hand, the 25% percentile values are lowest on Tuesday and Thursday.

The following plots show a distribution of the daily returns for each day of the week. Sunday was excluded as it is not considered a major trading day.

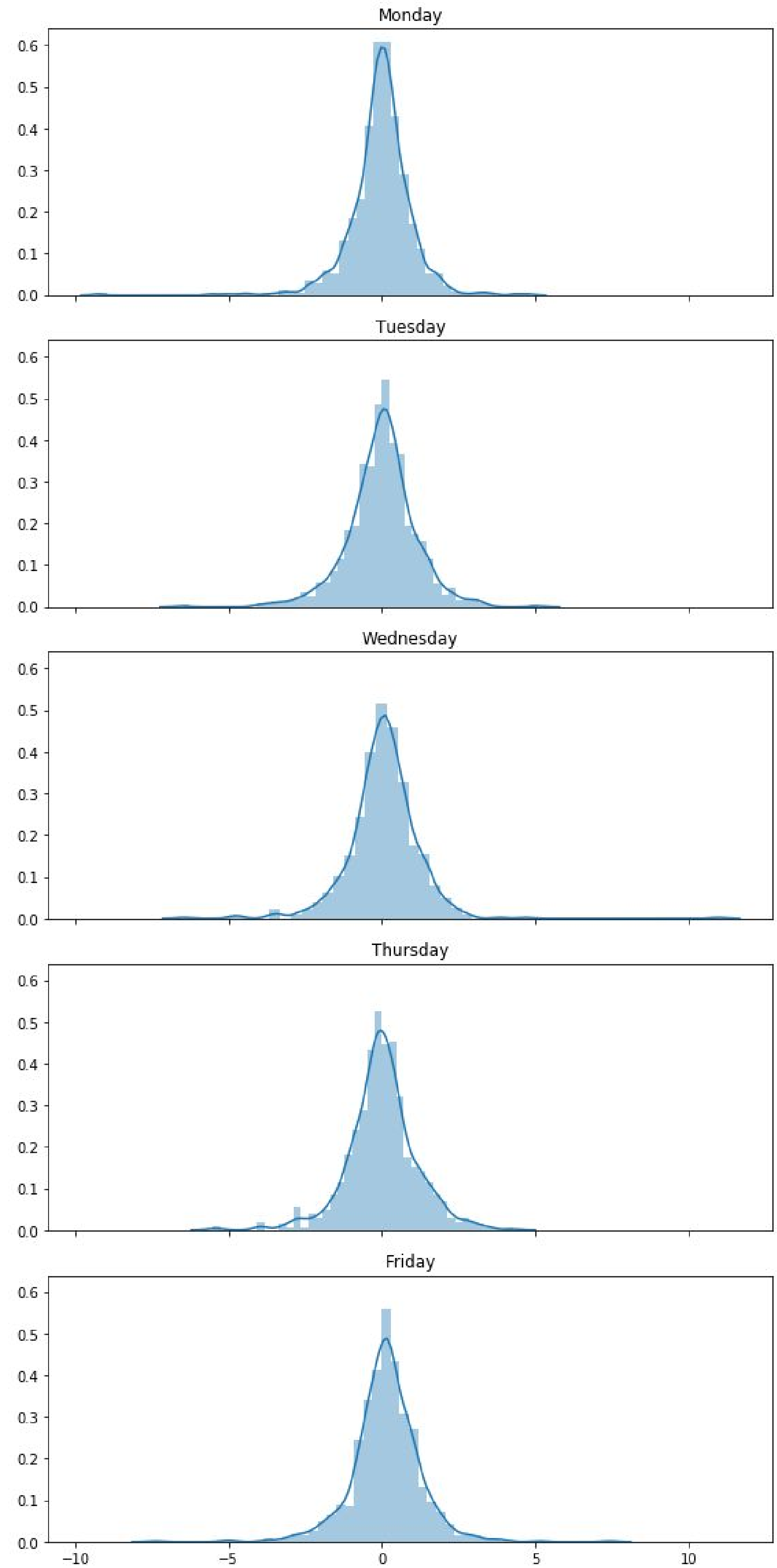

Figure 8 - Distribution of Gold daily returns per day of the week

The distribution plots show that the daily returns follow a normal distribution for all the days of the week, as it is also confirmed by the violin plot show below:

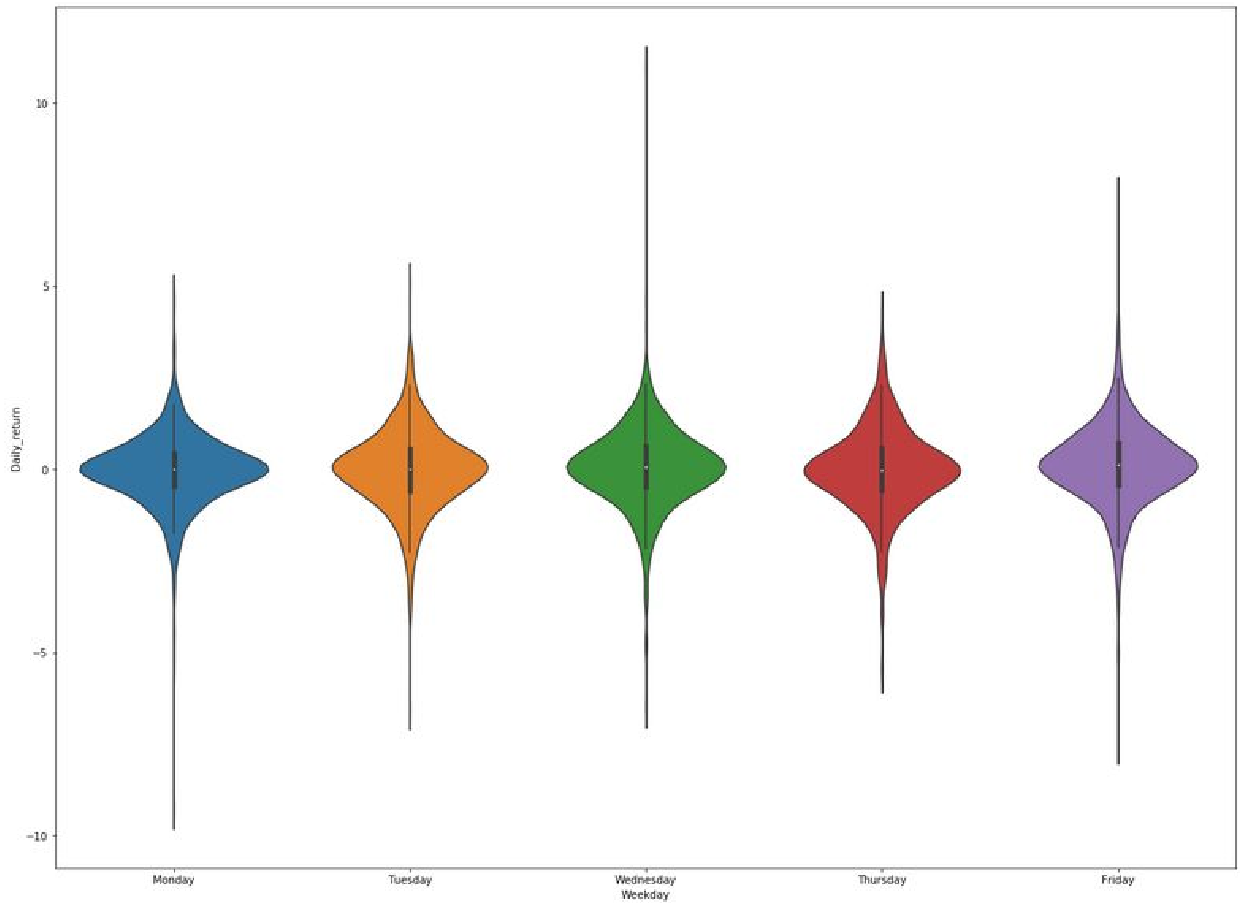

Figure 9 - Violin plot of Gold daily returns per day of the week

The violin plot, however, does seem to indicate that both Tuesday and Thursdays have had less volatility in terms of daily returns, when compared with the other days of the week. The greater volatility in both Fridays and Mondays can be related to both economic news (i.e US. employment statistics are usually announced on Fridays) and the fact that traders maybe reacting to news which took place over the weekend, or seeking to close their positions before the weekend if they are uncertain about their positions or expecting news which could affect their positions over the weekend.

The following violin plot shows the distribution of daily returns based on week of the year:

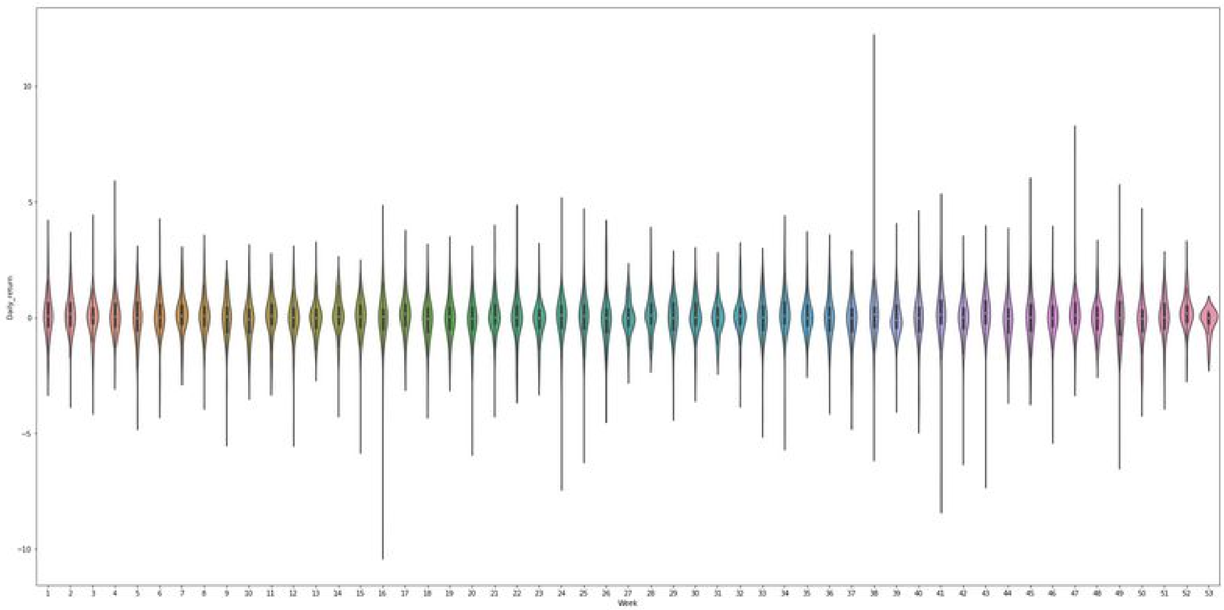

Figure 10 - Violin plot of Gold daily returns per week of the year

When looking at the daily returns from a weekly perspective, we can see that there a weeks with higher volatility followed by weeks of with lower volatility, in a somewhat cyclical fashion.

Correlation analysis

The last step in this inferential statistical analysis section was to perform an analysis of the correlation between Gold and both Silver and Euro closing prices and daily returns. The results obtained were as follows:

Correlation XAU - XAG closing prices: 0.8808017606279614
Correlation XAU - XAG daily returns: 0.7680463150559792

Correlation XAU - EUR closing prices: 0.018244845978475904
Correlation XAU - EUR daily returns: 0.40066063713051353

We can see that there is a strong positive correlation between Gold and Silver prices and daily returns. However, there is basically no correlation between Gold and Euro closing prices, while there is only a moderate positive correlation between Gold and Euro daily returns.

The following correlation matrix, shows the correlations between the Gold, Silver and Euro volumes, prices and daily returns:

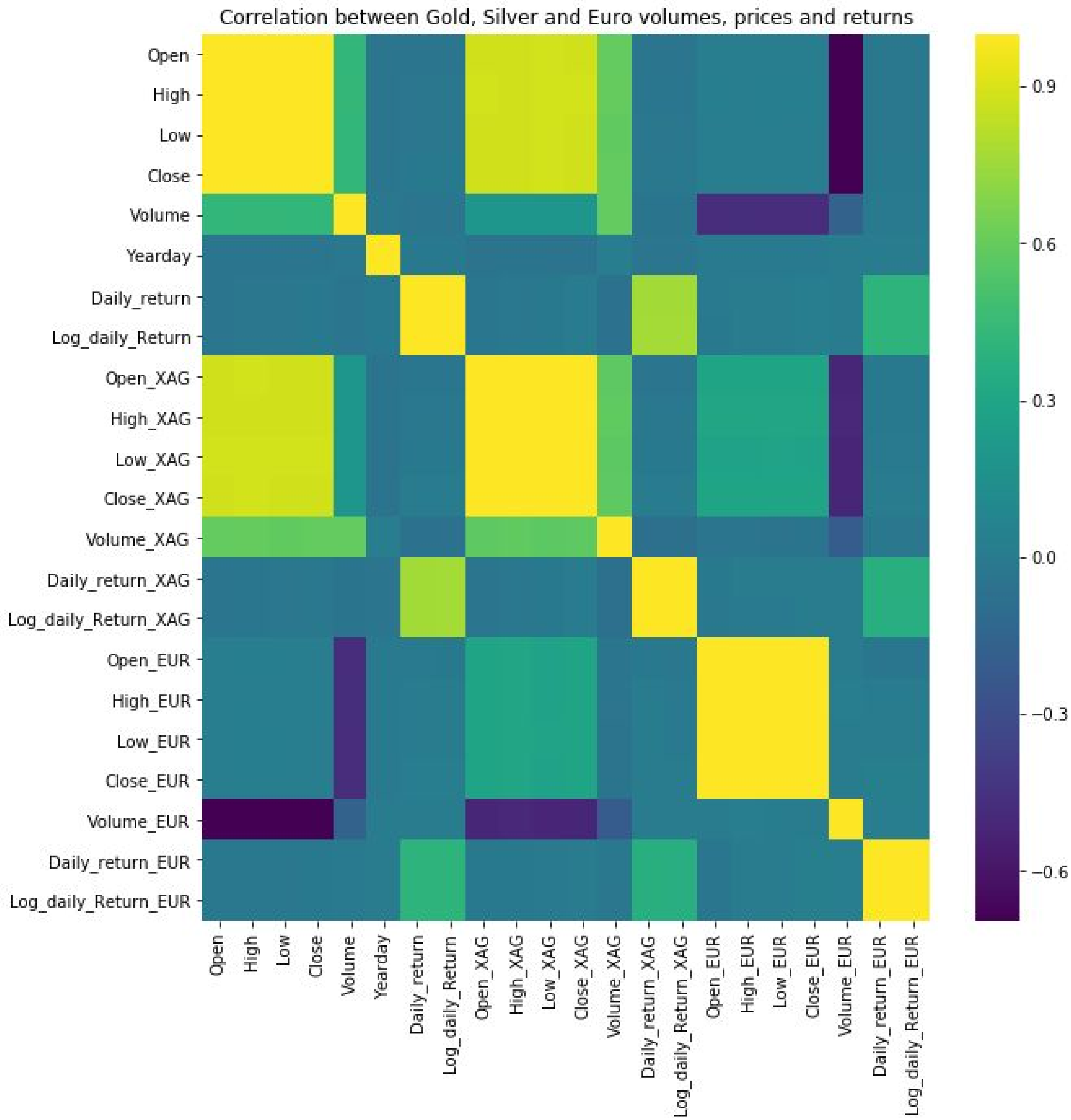

Figure 11 - Correlations matrix for Gold, Silver and Euro volume, prices and daily returns

The correlation matrix confirms our earlier conclusions regarding the correlation between Gold and Silver prices. Other than that, there seems to be not much relevant correlation between other variables.

Technical indicators

In order to further enrich the dataset, some technical indicators were also used for the three instruments, based on the existing price and volume data. The indicators were calculated using the Python Ta-lib library and they were the following: Simple Moving Average, Exponential Moving Average, Moving Average Convergence Divergence and Relative Strength Index.

The Simple Moving Average (SMA) is calculated by averaging the closing prices over a pre-determined period of days. In this project, we used a 50-day and a 200-day period SMA's. The formula for SMA[4] is as follows:

$$SMA_i = \frac{\sum_{k=i-n}^{i} x_k}{n}$$

The Exponential Moving Average (EMA) is somewhat similar to a Simple Moving Average, however, it makes use of weighted factors (alpha). We used both a 9-day and 21-day period EMA's. The formula for the EMA[5] is shown below:

$$EMA_i = EMA_{i-1} + \alpha * \left(x_i - EMA_{i-1}\right) \quad \alpha = \frac{2}{n+1}$$

The Moving Average Convergence Divergence (MACD) indicator combines both the difference between two EMA's (fast and slow periods) and the EMA of a signal period, together with a histogram series. The MACD formula utilized a fast period of 12 days, a slow period of 200 days and signal period of 9 days. The MACD[6] utilizes several formulas, as shown below:

$$MACD_i = EMA[fast\ period]_i - EMA[slow\ period]_i$$

$$signal_i = EMA[signal\ period]_i MACD$$

$$histogram_i = MACD_i - signal_i$$

---

[4] SMA Formula - https://docs.anychart.com/Stock_Charts/Technical_Indicators/Mathematical_Description#simple_moving_average

[5] EMA Formula - https://docs.anychart.com/Stock_Charts/Technical_Indicators/Mathematical_Description#exponential_moving_average

[6] MACD Formula - https://docs.anychart.com/Stock_Charts/Technical_Indicators/Mathematical_Description#moving_average_convergence_divergence

The Relative Strength Index (RSI) is a "momentum oscillator measuring the velocity and magnitude of directional price movement by comparing upward and downward close-to-close movements."[7] The RSI[8] formula uses upward and downward prices changes and moving averages, as shown in the below:

$$U_i = \begin{cases} close_i - close_{i-1}, \text{if } close_i > close_{i-1} \\ 0, \text{if } close_i \leq close_{i-1} \end{cases}$$

$$D_i = \begin{cases} close_{i-1} - close_i, \text{if } close_{i-1} > close_i \\ 0, \text{if } close_{i-1} \leq close_i \end{cases}$$

Upward and Downard Change formulas

$$MA_{U_i} = U_i + MA_{U_{i-1}} * \frac{period - 1}{period}$$

$$MA_{D_i} = D_i + MA_{D_{i-1}} * \frac{period - 1}{period}$$

Moving Average formulas

$$RSI_i = 100 - 100 * \frac{1}{1 + \frac{MA_{U_i}}{MA_{D_i}}}$$

Final RSI formula

In addition to the above standard technical indicators, a few more "custom" ones were added, namely:

- SMA Delta - A relative indicator which calculates the ratio of the difference between short (50-day period) and long SMA (200-day), over the closing price.
- EMA Delta - A relative indicator which calculates the ratio of the difference between short (9-day period) and long EMA (21-day), over the closing price.
- Gold/Silver ratio - Ratio of Gold closing price over Silver closing price

The following correlation matrix shows the correlations between these various technical indicators for the three instruments:

---

[7] Relative Strength Index - https://docs.anychart.com/Stock_Charts/Technical_Indicators/Relative_Strength_Index_(RSI)
[8] RSI Formula - https://docs.anychart.com/Stock_Charts/Technical_Indicators/Mathematical_Description#relative_strength_index

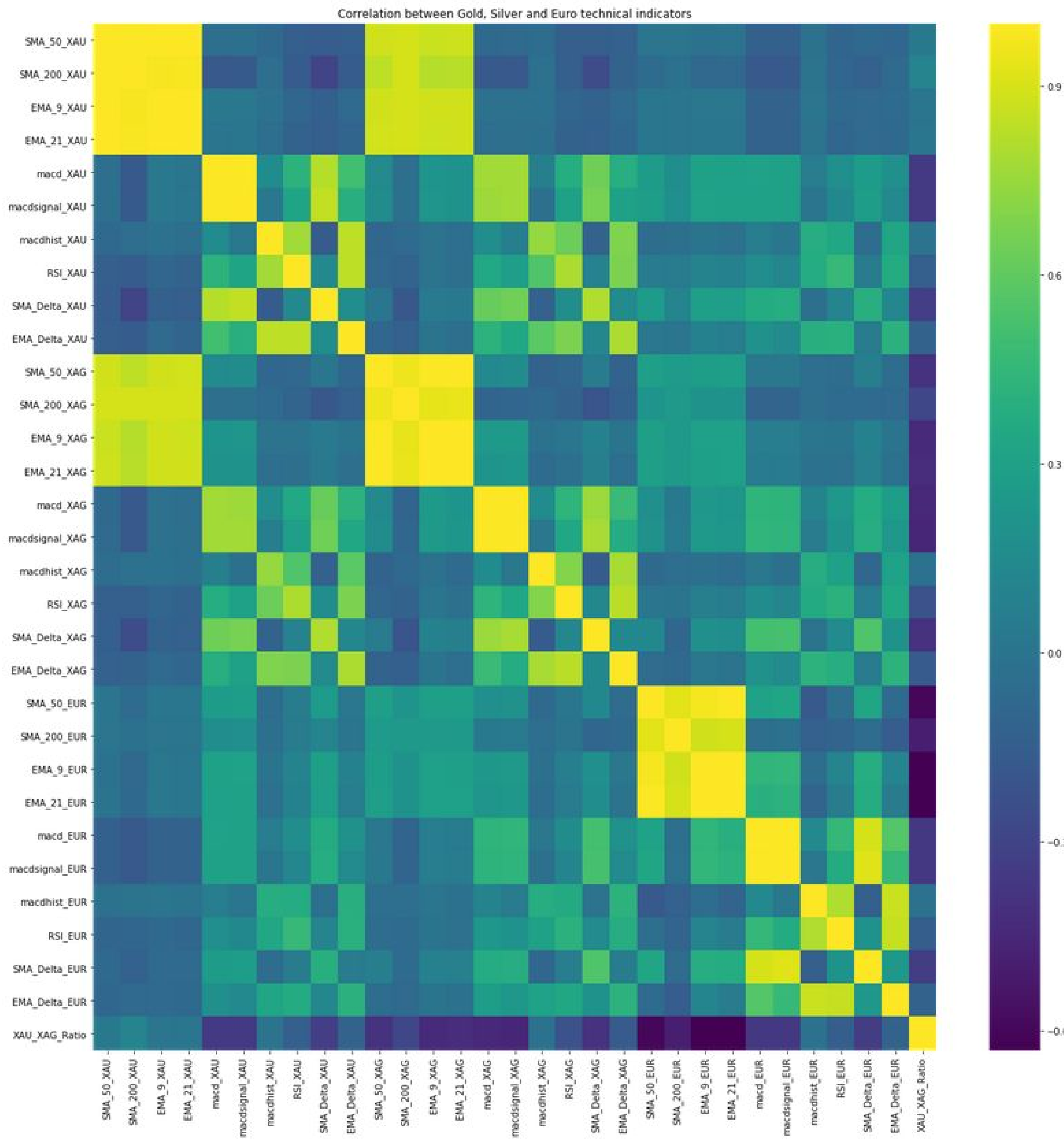

Figure 12 - Correlations matrix for Gold, Silver and Euro technical indicators

As it can be seen in the matrix above, we again see a correlation between Gold and Silver SMA’s and EMA’s, which would be expected, as there is a strong correlation between their prices. Also of interest, is the strong negative correlation between the Gold/Silver ratio and the Euro SMA’s and EMA’s, as well.

5. In-depth analysis using Time Series Methods

Traditional Time Series methods

The in-depth analysis of the dataset first began by using more traditional approaches to Time Series prediction, such as Autoregressive (AR), Moving Average (MA), ARMA and ARIMA models.

Before doing so, it was first necessary to determine if the Gold price time series data was stationary. In order to determine that, I plotted the both the autocorrelation and partial autocorrelation of the time series, as shown below:

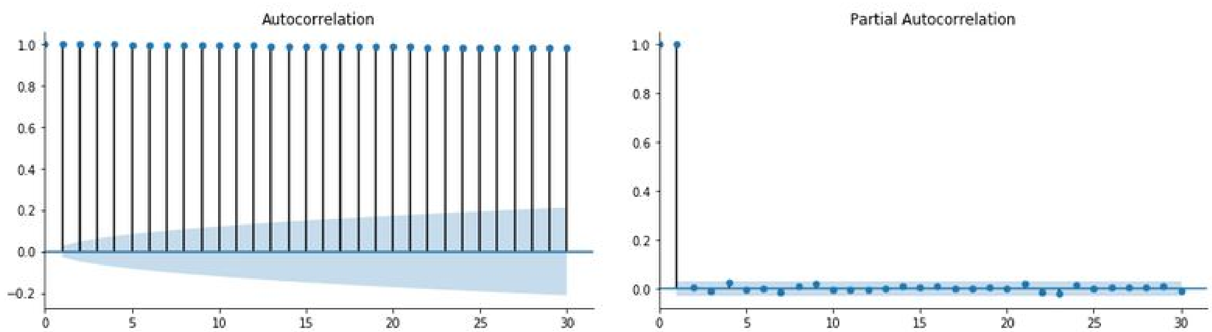

Figure 13 - Autocorrelation and Partial Autocorrelation plots for Gold price time series

As shown in the autocorrelation plot, the series is non-stationary, as the price data is highly autocorrelated.

After performing this analysis, some different time series predictive models were developed. The results were based solely on historical closing price data and the goal of the models was to predict between 8 and 30 days of future price data.

Unfortunately the results were not very good, as it can be seen by the graphs shown below:

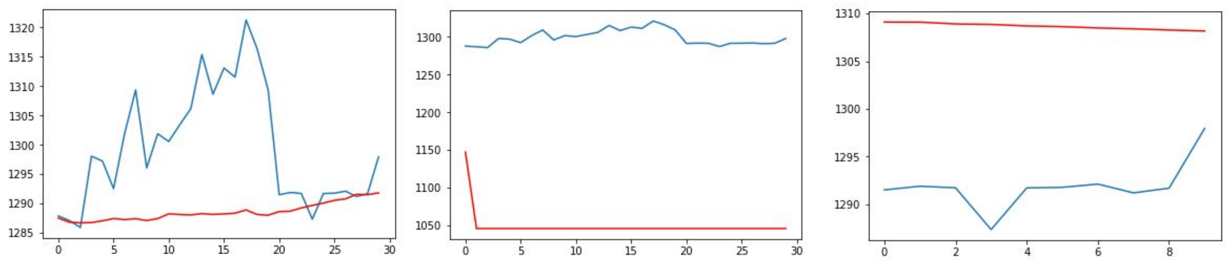

Figure 14 - AR, MA and ARMA plots

Although the results obtained were not that convincing, this exercise provided the opportunity to obtain a fundamental understanding of the more common time series predictive models.

6. In-depth analysis using Machine Learning

Feature Engineering

The next focus of the project them moved on towards using Machine Learning algorithms in order to predict Gold price returns.

In order to do so, the first step involved selecting which relevant features to include in the model. As one of the objectives of the project was to determine whether there were temporal patterns in data, many of the features used in the machine learning models revolved around temporal elements, such as day of the week, day of the year and week of the year.

In order to properly account for the cyclical nature of time data, the days of the week, days of the year and week of the year information, was converted as integers which were then multiplied by a sin and cosine function in order to better represent their cyclical nature.

In addition to the time data, some relevant technical indicators were also calculated through the use of the TA-lib python library, and were added as features in the dataset. As previously mentioned, the technical indicators utilized included:

- 50 and 200 day Simple Moving Averages (SMA-50 and SMA-200)
- 9 and 21 day Exponential Moving Averages (EMA-9 and EMA 21)
- Moving Average Convergence Divergence (MACD)
- Relative Strength Index (RSI)

In addition to the above, two more data columns were added with values derived from the above technical indicators:

- SMA Delta - This relative indicator was defined as the difference between the 50 and 200 day SMA, divided by the close price
- EMA Delta - Another relative indicator which is the difference between the 9 and 21 day EMA, divided by the close price

With regards to the target variable, given the difficulty during the first attempts in obtaining convincing results using regression, the problem was changed towards using classification algorithms, where the objective was to determine positive results above a certain threshold.

The threshold first began at 0.5% daily return, however, given the low number of such returns, the threshold was then reduced to 0.1%.

Logistic Regression

The first models which were implemented used the Logistic Regression Algorithm to predict returns greater than 0.1%.

As mentioned earlier, the difficulty in obtaining reasonable prediction results with higher return values at first, led us to lower the value of the return threshold. However, even after lowering the target to include all returns above 0.1%, the results obtained were not very good, as shown in the confusion matrix below:

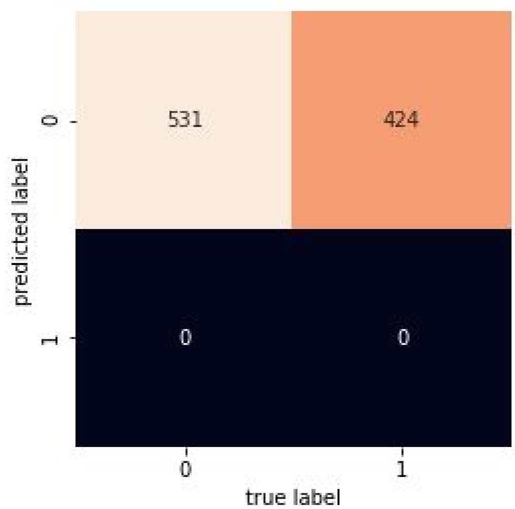

Figure 15 - Confusion matrix results for Logistic Regression

The fact that the data contains mostly 0 values (returns below 0.1%) made the dataset somewhat unbalanced. Therefore, another variation of this algorithm was attempted, this time using a SMOTE treatment to the dataset, in order to compensate for this fact. However, the results did not fare that much better:

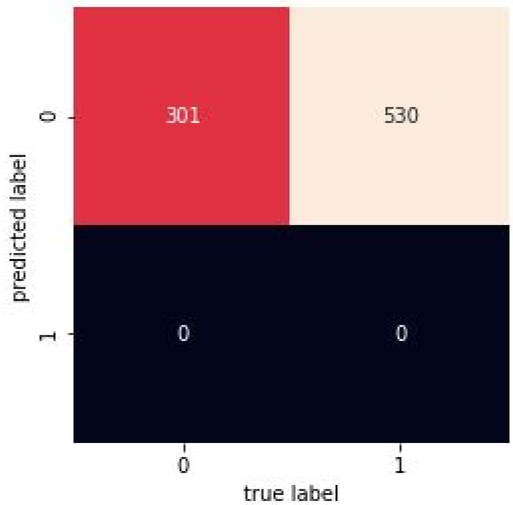

Figure 16 - Confusion matrix results for Logistic Regression with SMOTE

The results obtained by both Logistic Regression variations are summarized in the table below:

	Logistic regression without SMOTE				Logistic regression with SMOTE			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
0	0.56	1	0.71	531	0.36	1	0.53	301
1	0	0	0	424	0	0	0	530
micro avg	0.56	0.56	0.56	955	0.36	0.36	0.36	831
macro avg	0.28	0.5	0.36	955	0.18	0.5	0.27	831
weighted avg	0.31	0.56	0.4	955	0.13	0.36	0.19	831

First, neither variation is able to detect a single positive target value, which makes this algorithm unusable. Also, the logistic regression without SMOTE performed better than the one using the SMOTE treatment.

Random Forest Classifier

It was then time to move on to the next algorithm, Random Forest Classifier.

This algorithm made use of the exact same feature and target data, however, this time, a GridSearch function was utilized in order to determine the best combination of parameters.

The results obtained were as follows:

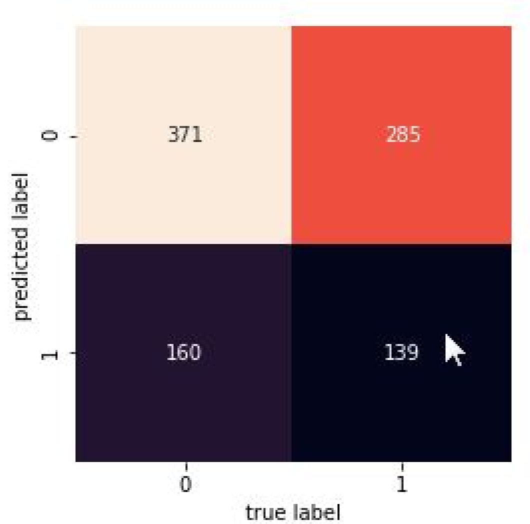

Figure 17 - Confusion matrix results for Random Forest Classifier

The results obtained with this algorithm were significantly better, however it still was not able to correctly identify almost 70% of the positive target returns.

This algorithm was re-run a second time, however, it now included additional daily price data (Open, Close, High, Low). The results were very similar, as shown by the confusion matrix below:

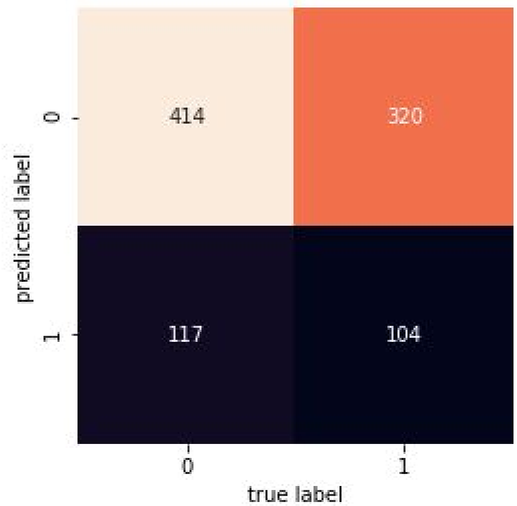

Figure 18 - Confusion matrix results for Random Forest Classifier with price data

The results of both algorithm variations are summarized in the table below:

	Random Forest Classifier				Random Forest Classifier (w/ Price Data)			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
0	0.57	0.7	0.63	531	0.56	0.78	0.65	531
1	0.46	0.33	0.38	424	0.47	0.25	0.32	424
micro avg	0.53	0.53	0.53	955	0.56	0.56	0.56	955
macro avg	0.52	0.51	0.5	955	0.53	0.55	0.51	955
weighted avg	0.52	0.53	0.52	955	0.68	0.56	0.6	955

Gradient Boost Classifier

After the results obtained with the Random Forest Classifier algorithm, it was then time to try out one last algorithm, Gradient Boost Classifier. Again, we used GridParameter search in order to find the optimal parameter combination. The confusion matrix is shown below:

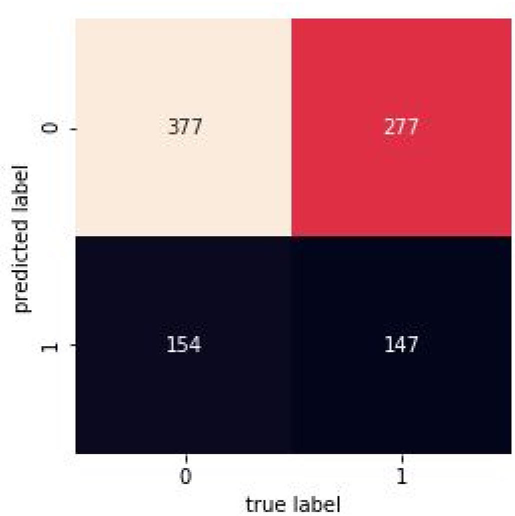

Figure 19 - Confusion matrix results for Gradient Boost Classifier

The same algorithm was again repeated with price date, as done previously with the Random Forest Classifier. The confusion matrix was as shown below:

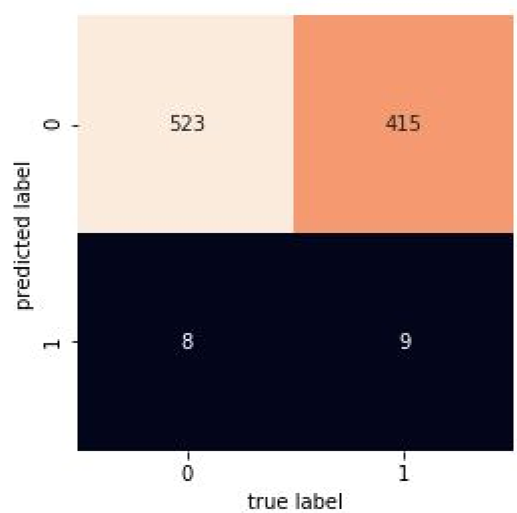

Figure 20 - Confusion matrix results for Gradient Boost Classifier with price data

The results with price date change dramatically when including price data, as the algorithm reduce significantly the number of positive returns that it can identify.

The complete results of both variations are presented in the table below:

	Gradient Boost Classifier				Gradient Boost Classifier (w/ Price Data)			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
0	0.58	0.71	0.64	531	0.56	0.98	0.71	531
1	0.49	0.35	0.41	424	0.53	0.02	0.04	424
micro avg	0.55	0.55	0.55	955	0.56	0.56	0.56	955
macro avg	0.53	0.53	0.52	955	0.5	0.54	0.38	955
weighted avg	0.54	0.55	0.53	955	0.97	0.56	0.7	955

Based on the results obtained from these three machine learning algorithms,we inferred that either the data that we have as features are not very helpful in terms of their predictive abilities, or, given the stochastic behaviour of markets, predicting positive returns is quite challenging and any modelling attempts may not produce better results than random selecting a positive or negative direction for tomorrow's results, based on historical data.

Integrating Euro and Silver Data

In order to better understand whether additional data from other instruments would be helpful in improving model results, the above three machine learning algorithms were again repeated with both daily returns and the same technical indicators mentioned earlier for both Silver and the Euro. Additionally, the Gold/Silver ratio, based on the Close price, was also added as a feature in the model.

The same methodologies for each algorithm were followed as described earlier. The confusion matrices for the three algorithms are shown below:

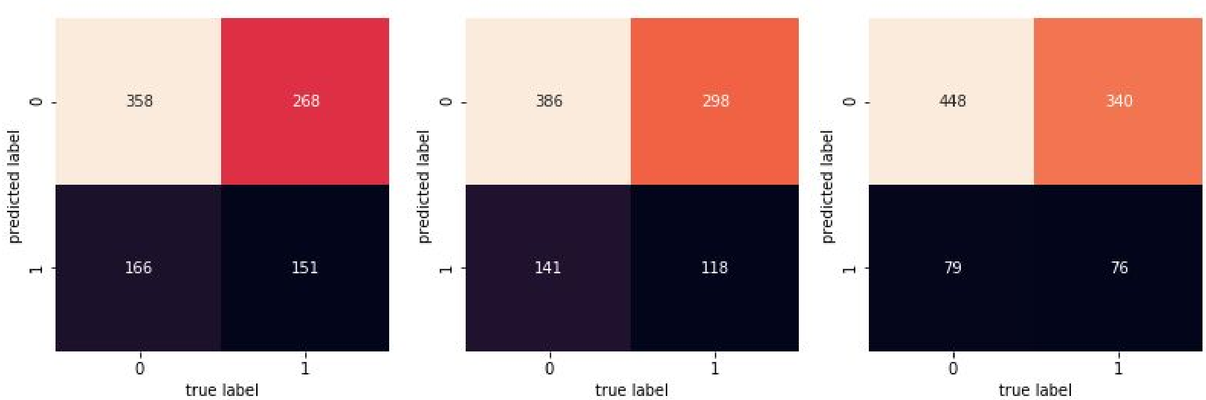

Figure 21 - Confusion matrices results for Logistic Regression (Left), Random Forest (Center) and Gradient Boost Classifier (Right) algorithms using Gold, Silver and Euro data and indicators

The results obtained are summarized in the table below:

	Logistic regression (XAG/XAU/EUR)				Random Forest Classifier (XAG/XAU/EUR)				Gradient Boost Classifier (XAG/XAU/EUR)			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
0	0.57	0.68	0.62	524	0.56	0.73	0.64	527	0.57	0.85	0.68	527
1	0.48	0.36	0.41	419	0.46	0.28	0.35	416	0.49	0.18	0.27	416
micro avg	0.54	0.54	0.54	943	0.53	0.53	0.53	943	0.56	0.56	0.56	943
macro avg	0.52	0.52	0.52	943	0.51	0.51	0.49	943	0.53	0.52	0.47	943
weighted avg	0.53	0.54	0.53	943	0.52	0.53	0.51	943	0.53	0.56	0.5	943

Even though Logistic Regression had the worst results with just the Gold data, the combined dataset helped it have the best results out of the three algorithms.

Unfortunately, the results even with additional data features, are not sufficient to be useful for any predictive purposes, as they are even below 50%.