Prediction on US Stock Market Performance: S&P 500

Danning (Danni) Lai, Sijia (Nancy) Li, Xu (Victoria) Tang, and Hanlin (Lily) Zhu

1 INTRODUCTION

In financial markets, traders and fund managers often aim to construct a simple and robust model to predict the US S&P 500 performance. There are two types of indicators: leading indicators and lagging indicators. Leading indicators are factors that change before the financial market changes (e.g., ISM manufacturing index, ISM non-manufacturing index, consumer sentiment), while lagging indicators respond some time after an exogenous shock (e.g., employment data, GDP). Due to time constraints, the scope of this project is narrowed down to focus on predicting the equity market from three other markets — energy, bullion, and bonds. The problem of interest is to predict returns on the S & P 500 index using indicators from these three markets. EDA results show evidence of cross-correlation and Granger causation between S&P 500 log return and other variables (crude oil futures log return, gold futures log return, change in 3-month treasury bill yield, and change in 10Y-3M treasury yield spread), making them promising predictors of the stocks.

For this project, the data frequency is weekly with data from November 10, 2002 to October 31, 2021 as the estimation sample, and data from November 7, 2021 to October 30, 2022 as the prediction period (test set). The forecasting scheme is fixed (rather than recursive or rolling) and the forecasting horizon is 3-step ahead. Three models are selected as the baseline: historical mean, last period naive, and autoregression. To improve on the baseline models, vector autoregression (VAR) and long short-term memory (LSTM) models are built and evaluated. The implementation of these models are described in the next section.

2 DATA DESCRIPTION AND PREPROCESSING

2.1 Data Collection and Frequency Adjustment

To predict the US stock market performance (S&P 500), five types of data are collected: S&P 500 historical prices, US 13-week treasury bill yield, US 10-year treasury yield, crude oil futures prices, and gold futures prices. The data collected are daily time series data from November 1, 2002 to October 31, 2022 (20 years). All data are obtained from the adjusted closing prices column on Yahoo Finance. When occasionally there is no data for a particular market (treasury, crude oil, or gold) on a specific trading day, forward fill is used to fill the missing value in the data. Resampling is used to convert the time series data into weekly frequency as daily time series data may be too noisy for analysis.

2.2 S&P 500 Index, Crude Oil Futures, and Gold Futures

To begin with, there are two ways to calculate financial returns: simple return and log return. Simple return is the percentage change in price after holding an underlying asset for a certain time period, while log return is the change in log of the price after holding an underlying asset for a certain time period. Log return ranges from negative infinity to positive infinity and is symmetrical around zero, but simple return is limited to negative 100%. Moreover, a negative movement of -20% ($100 to $80) cannot be reversed by the same amount of movement in the positive direction (going up by 20% from $80 to $96). Hence, the log return will be used for S&P 500 index prices and commodity prices in this project (i.e., crude oil and gold). One additional note is that the log return is multiplied by 100 during data preprocessing as this value is normally expressed as a percentage.

The weekly log return of S&P 500 index prices [link, ticker: ^GSPC] is included both as a predictor and a response, since the stock market performance in the future is related to the performance in the past. In addition, in finance, futures obligate buyers to purchase or sellers to sell an asset (e.g., crude oil, gold) at a predetermined price at a particular date in the future. In particular, the weekly log return of crude oil futures prices [link, ticker: CL=F] and the weekly log return of gold futures prices [link, ticker: GC=F] are included as predictors for stock market performance. Crude oil is important because it is the primary source of energy production and crude oil prices are indicative of the performance of the energy market. It is also important to consider Bullion (or gold of high purity) because gold is often seen as a great hedge against economic instability, which in turn, relates to the performance of the stock market.

2.3 Treasury Bills

The yield to maturity of US 13-week treasury bill [link, ticker: ^IRX] and US 10-year treasury note [link, ticker: ^TNX] are also obtained. The US Treasurys are fixed income securities backed by the US government. The 13-week treasury bills are short-term debt securities with a maturity of 13 weeks, while the 10-year treasury notes are intermediate-term debt securities with a maturity of 10 years. The yields of US Treasurys are often used as a benchmark for interest rates, an influencing factor of stock market performance. Two features, weekly changes in 13-week treasury bill yield and weekly changes in 10Y - 3M treasury yield spread, are created using these prices through first-order differencing. Specifically, 10 Year - 3 Month treasury yield spread is a popular metric as it reflects the slope of the yield curve and thus encapsulates market expectations.

2.4 Summary

In summary, five features are created using data obtained from Yahoo Finance: S&P 500 log return (for simplicity, we will denote it as S&P 500 in the rest of the report), crude oil futures log return (crude oil), gold futures log return (gold), change in 13-week treasury bill yield (T-Bill yield), and change in 10Y-3M treasury yield spread (yield spread). The final data range is from November 10, 2002 to October 31, 2022, with the first trading week in November 2002 removed after taking first-order differences during feature engineering. Finally, the data are then split into two sets: estimation sample (95% of data from November 10, 2002 to October 31, 2021) and prediction period (5% data from November 7, 2021 to October 30, 2022). Since we use a fixed scheme model, i.e., we do not expand the estimation set as the information set is updated, our model will not perform well too far into the future. For this reason, we limit our test set to the most recent 5% of the observations.

3 EXPLORATORY DATA ANALYSIS

3.1 Visualization of Original Time Series Data

We plotted the time series of the S&P 500 index, crude oil futures price, gold futures price, 3-month treasury bill yield, and the 10Y - 3M treasury yield spread. As shown in **Figure 1**, these time series are not covariance stationary, since they are not mean-reverting nor having constant variance. Therefore, to satisfy the assumptions of econometric models, we modified the time series through log transformation and first-order differencing, which are elaborated on in Section 2. It can be seen that, although we cannot make any conclusions yet, there are some relationships between S&P 500 with the other variables. For example, from 2004 to 2008, the S&P 500, crude oil futures price, gold futures price, and 3-month treasury bill yield all display a general upward trend, with the yield curve flattened. Moreover, when the pandemic first struck, S&P 500 took a steep dive, and so did the crude oil futures price and the T-Bill yield. This again shows the possible relationships between S&P 500 and the other markets.

Figure 1. Time Series Plot of Original Data

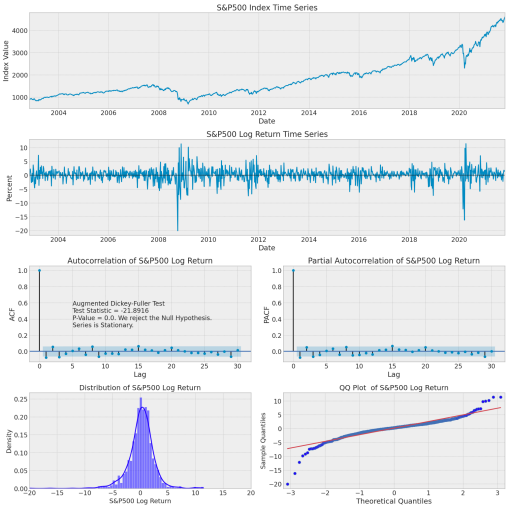

Figure 2. Time Series Analysis of S&P 500 Index

3.2 Serial Correlation and Stationarity

As mentioned above, transformations are applied to the data to make their probability distribution more time-invariant. Specifically, S&P 500 index price, crude oil futures price, and gold futures price are transformed into S&P 500 index log return, crude oil futures log return, and gold futures log return, respectively. For features related to US Treasurys, yields of 13-week (~3-month) treasury bills and 10-year treasury notes are converted into a change in 3-month treasury bill yield and a change in 10Y-3M treasury yield spread by first-order differencing. **Figure 2** shows the time series analysis for S&P 500 index. The time series analysis for other variables can be found in **Appendix A**. As shown in **Figure 2**, the log return of the S&P 500 index price appears stationary across the years. ACF (autocorrelation function) plot and PACF (partial autocorrelation function) plot are also provided. ACF is the correlation of the time series with its lagged values. PACF is the correlation between the residuals with the next lag value, where the residuals are what remain after removing the intermediate effects explained by previous lags. The blue-colored area in the plots represents the 95% confidence interval. As shown in the graphs, there is a sharp cutoff beyond the first displacement and no strong persistence across the lags, which indicates that the data is stationary. To confirm stationarity, the Augmented Dickey-Fuller (ADF) test is applied to the data. Given that the p-value is very small, the null hypothesis is rejected and it can be concluded that the S&P 500 index log return time series is stationary. In addition, the distribution of the S&P 500 index log return data is also examined. The distribution is centered around zero and forms a symmetrical bell shape. The data also roughly follow a straight line on the Q-Q plot. This shows evidence that the data is normally distributed.

3.3 Relationships among Variables

3.3.1 Correlation

In **Figure 3** and **Figure 4,** we find the correlation matrix

between features and provide visualizations below. Based on these plots, we found that no time series has high correlation ($> \pm 0.8$). The highest correlation among the variables is that between crude oil and S&P 500, which is 0.29. This indicates that crude oil may play an important role in our prediction model. Meanwhile, the small correlation of gold futures log return and S&P 500 may suggest that the gold market is a weak predictor of equities. In **Figure 5**, we also obtained the correlation of variables using rolling windows (up to 20 weeks), so as to provide a more valid result. We can see that the values in **Figure 5** correspond to those in **Figure 3** and **Figure 4**.

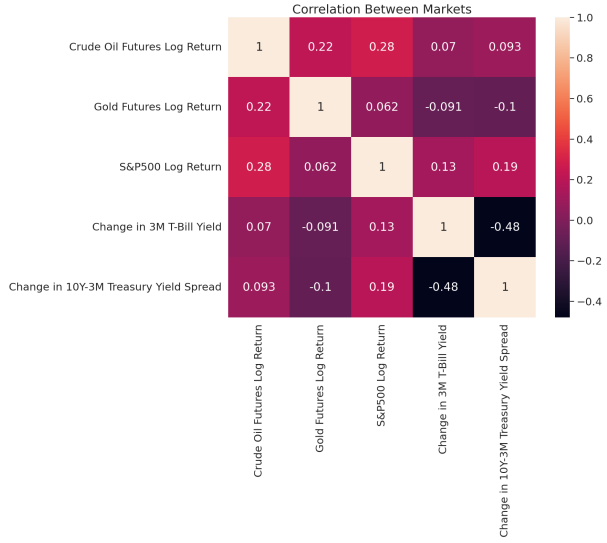

Figure 3. Correlation between Log Returns

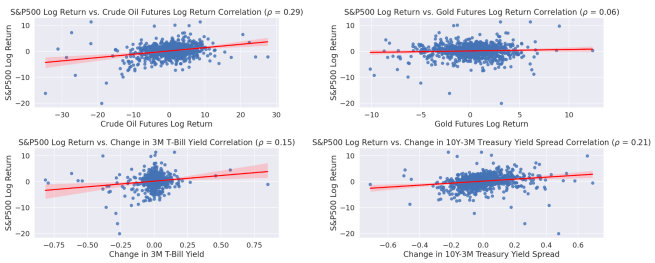

Figure 4. Correlation between Log Returns

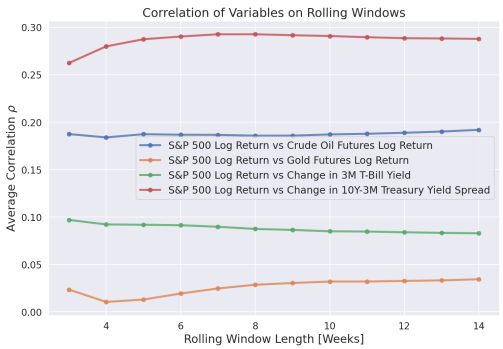

Figure 5. Correlation of Variables on Rolling Windows

3.3.2 Cross-correlation

The cross-correlation function is the correlation of two time series displaced relative to each other by k time steps (i.e, weeks, in our project). **Figure 6** shows the cross-correlation between the variables. We see that crude oil has some correlation with S&P 500 at lag 8 and lag 9, which means a potential lead-lag relationship. Similarly, T-Bill yield has a cross-correlation of 0.1 with S&P 500 at lag -4, which implies its moderate ability in predicting S&P 500. Notice that we use a maximum lag of 52 for the cross-correlation function and the Granger-causality test in the following, since we postulate that no significant lead-lag relationships spans over a year.

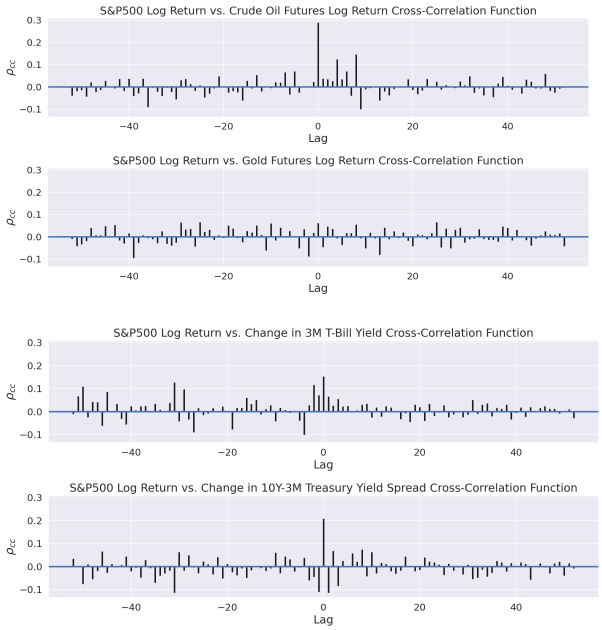

Figure 6. Cross-Correlation Function of Variables

3.3.3 Granger-Causality Test

The Granger-causality test is another way to determine whether one time series is useful serving as the predictor of another. The null hypothesis being set up is that time series **x** does not Granger-cause time series **y.** Therefore, if a test shows a small p-value (choosing the threshold to be 0.05), in which case we reject the null hypothesis, and we conclude x Granger-causes y. As a result, in order to find an ideal predictor of S&P 500 log return, we hope to find a time series x which Granger-causes the target. It would be most optimal if the target does not in turn Granger-cause x.

In our result, all of our features Granger-cause S&P 500. Remarkably, gold Granger-causes S&P 500 with a p-value of 0.0384, while S&P 500 does not Granger-cause gold. Therefore, we expect the gold market indicator to provide new information and improve the predictions of S&P 500. Similarly, change in 3M T-bill yield seems promising. **Table 1** shows the results of the Granger Causality test.

Table 1. Granger Causality Test Results

	Crude Oil	Gold	SP 500	T-Bill Yield	Yield Spread

Crude Oil	1.0000	0.0384	0.0000	0.0	0.0001
Gold	0.0803	1.0000	0.0527	0.0	0.0007
SP 500	0.0012	0.0008	1.0000	0.0	0.0001
T-Bill Yield	0.1046	0.2924	0.0844	1.0	0.0010
Yield Spread	0.0057	0.0026	0.0002	0.0	1.0000

4 BASELINE MODEL

There are various ways to approach this problem: average method, naive method, and autoregressive (AR) models. Average method takes the mean of all historical data as a forecast of future values. For the naive method, all forecasts are set to be the value of the last observation. A statistical model is autoregressive if it predicts future values on the basis of historical data.z

4.1 Historical Mean Model and Naive Model

In the historical mean model, future values are simply predicted as the average of all training data. **Figure 7** displays the performance of the historical mean model. From this plot, we see that the model's forecast is a horizontal line and does not depict the volatility of the stock return.

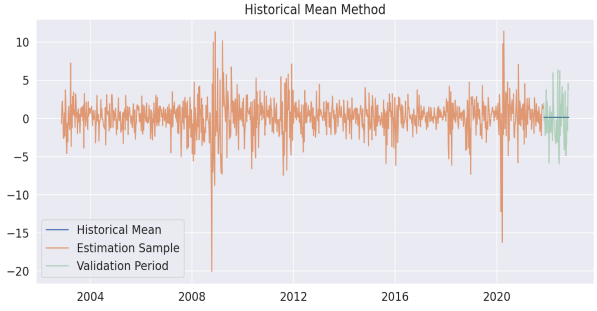

Figure 7. Results of Historical Mean Method

Naive model uses the latest period's value as the next period forecast of S&P 500 without training a model, the reason why it is called a "naive model". In this part, we take the S&P 500 index of the ending Friday in our estimation sample as the prediction. The result is illustrated in **Figure 8** below. Notice that for the above two baseline models, we use the fixed scheme.

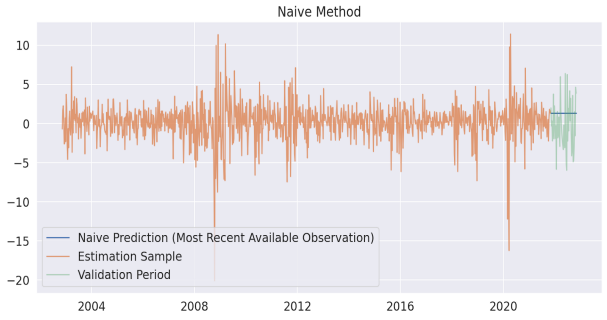

Figure 8. Results of Naive Method

4.2 Autoregression Model

AR is a common model for forecasting a time series based on its lagged values. In order to construct an AR(p) model, we first need to decide the order of the model, i.e., the value of p. In this case, we choose p among 0 to 52 lags (one year time frame), as we believe data within this range is more influential. Based on the result of Akaike information criterion (AIC), we set p = 4 lags. We use AIC to evaluate the model fit on the training data with an additional penalty term for model complexity. Comprehensive AIC results can be viewed in **Appendix B.1**. We then implement the AR model using the statsmodels package.

To evaluate the model, we first generate the in-sample static predictions, the fitted values, to see how well our model performs in our training data. The result can be seen on **Figure 9**. Out-of-sample forecasts of **Figure 10** displays the model performance as forecasts evolve into the validation period. Furthermore, we generate 3-step ahead forecasts on the prediction set as we aim to forecast three weekly stock returns. Each short colored line in **Figure 11** presents one such out-of-sample 3-step forecast. Since the forecast variance usually increases with the forecast horizon, we can't simply averaging out prediction errors at different forecast horizons. The performance for each of the three forecast horizons (one-week ahead, two-week ahead, and three-week ahead) can be viewed from **Figure 12**, and a more comprehensive graph can be found in **Appendix B.2**. To check that our model is a good fit for our data, through the Ljung-Box test, we check that the residuals are white noise, with diagnostic plots in **Appendix B.3**.

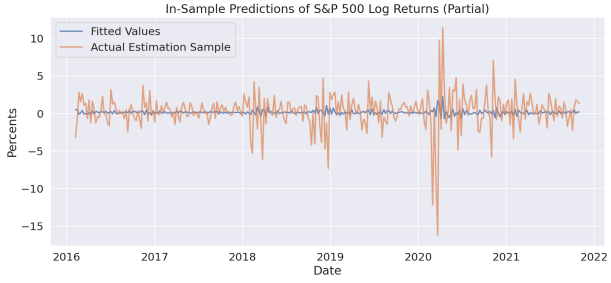

Figure 9. AR In-Sample Fitted Values

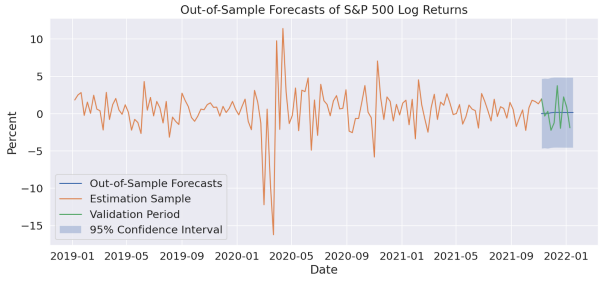

Figure 10. AR Out-of-Sample (OOS) Forecasts

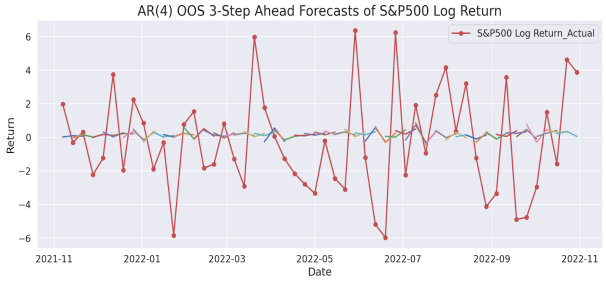

Figure 11. AR OOS 3-Step Forecasts

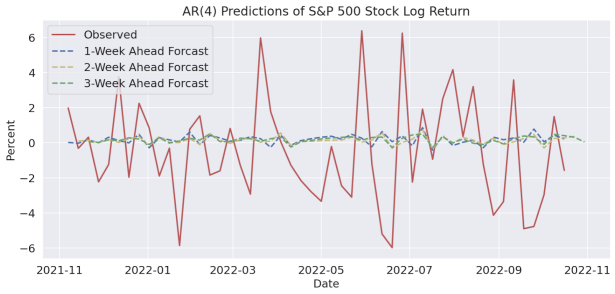

Figure 12. AR 1-, 2-, and 3-Week Ahead Forecasts

5 Advanced Models

5.1 Vector Autoregression Model

AR model imposes a unidirectional relationship of the time series. When it comes to multivariate time series and the variables influence each other, we implement a Vector Autoregressive (VAR) model. The VAR model is a generalization of the AR model, which specifies the bidirectional relationships between time series. The reason why we believe VAR is relevant in this case is because all four variables 'Granger-cause' the S&P 500 index (see **Table 1** above). Recall that using crude oil, gold, T-bill yield, and yield spread as x all gives p-values smaller than 0.05.

5.1.1 Model Selection - Optimal Lag

There are four available model selection criterions, AIC, HQ, SC and FPE, to determine the order p among 0 to 52. In this case, we select p = 10 by using AIC as it penalizes the complexity of the model. A general equation of VAR is shown as below:

$$y_{1,t} = c_1 + a^1_{1,1} y_{1,t-1} + a^1_{1,2} y_{2,t-1} + \cdots + a^1_{1,k} y_{k,t-1} + \cdots + a^p_{1,1} y_{1,t-p} + a^p_{1,2} y_{2,t-p} + \cdots + a^p_{1,k} y_{k,t-p} + e_{1,t}$$
$$y_{2,t} = c_2 + a^1_{2,1} y_{1,t-1} + a^1_{2,2} y_{2,t-1} + \cdots + a^1_{2,k} y_{k,t-1} + \cdots + a^p_{2,1} y_{1,t-p} + a^p_{2,2} y_{2,t-p} + \cdots + a^p_{2,k} y_{k,t-p} + e_{2,t}$$
$$\vdots$$
$$y_{k,t} = c_k + a^1_{k,1} y_{1,t-1} + a^1_{k,2} y_{2,t-1} + \cdots + a^1_{k,k} y_{k,t-1} + \cdots + a^p_{k,1} y_{1,t-p} + a^p_{k,2} y_{2,t-p} + \cdots + a^p_{k,k} y_{k,t-p} + e_{k,t}$$

In our 5-variable VAR model, we estimate five different equations. As our target variable of this project is S&P 500, we will focus on the equation with S&P 500 as the target, and itself, crude oil, gold, T-bill yield, and yield spread as the predictors. **Figure 13** shows the forecast of our VAR model. Then, we achieve multistep ahead forecasts from the VAR in a recursive manner. The out-of-sample forecast of S&P 500 can be viewed in **Figure 14**. The plots for the remaining four predictors can be seen in **Appendix C.2**.

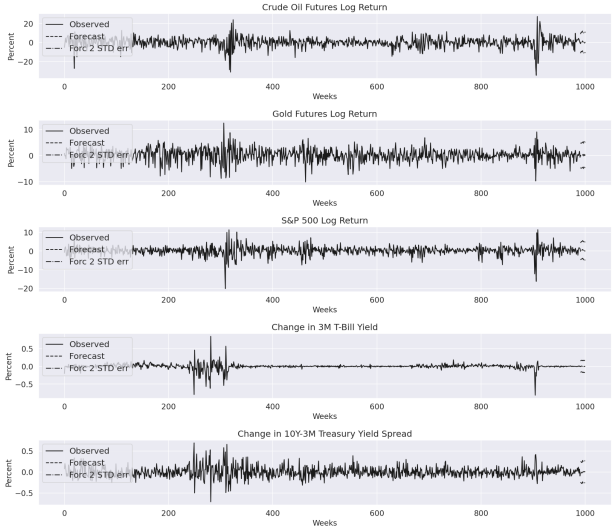

Figure 13. VAR OOS Forecasts

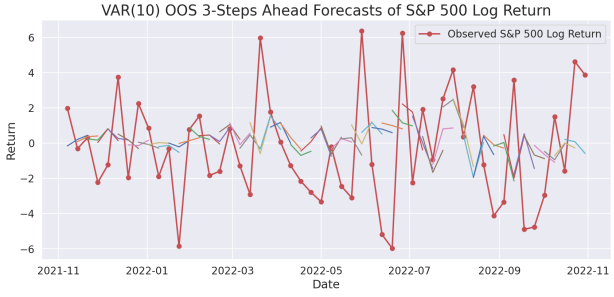

Figure 14. VAR OOS 3-Step forecasts for S&P 500 Returns

5.1.2 Impulse Response Functions (IRF)

As our main purpose is to describe the evolution of S&P 500 in reaction to a shock in other variables, we look at the impulse response functions. We examine both cross-variable impulse response (i.e, the effect of other variables' shock on S&P 500) and own-variable impulse response (i.e., the effect of S&P 500 on subsequent S&P 500).

Figure 15 shows the trend of S&P 500 log return after an external change (assume a positive unit shock). We find both the response at a time period and the cumulative response, but we will focus on the latter, as it tells more about the long-term predictability of the model and the effect over time. See **Appendix C** for the non-cumulative IRF. The region within the dash line displays the 95% confidence interval of the estimate.

In **Figure 15** below, we can see that S&P 500 substantially increases within the zero to four weeks after a unit increase in T-Bill yield. As the estimate with the confidence interval is almost completely above and away from the horizontal line at zero, we think this result is significant and Meanwhile, other results are not discernably significant. One unit increase of crude oil produces positive movement in S&P 500. Also, S&P 500 responds negatively to a unit shock of yield spread within six weeks, but goes up again gradually after six weeks. Similarly, a unit shock in gold somehow negatively, but weakly, affects S&P 500 in two to ten weeks. However, in these cases, the confidence interval still overlaps with (wraps around) the zero horizontal line, so, we conclude the result is not significant. On the other hand, S&P 500's unit shock does not have much impact on subsequent S&P 500 log returns. It slowly and slightly decreases after the shock and only shows small fluctuations within the 10 weeks.

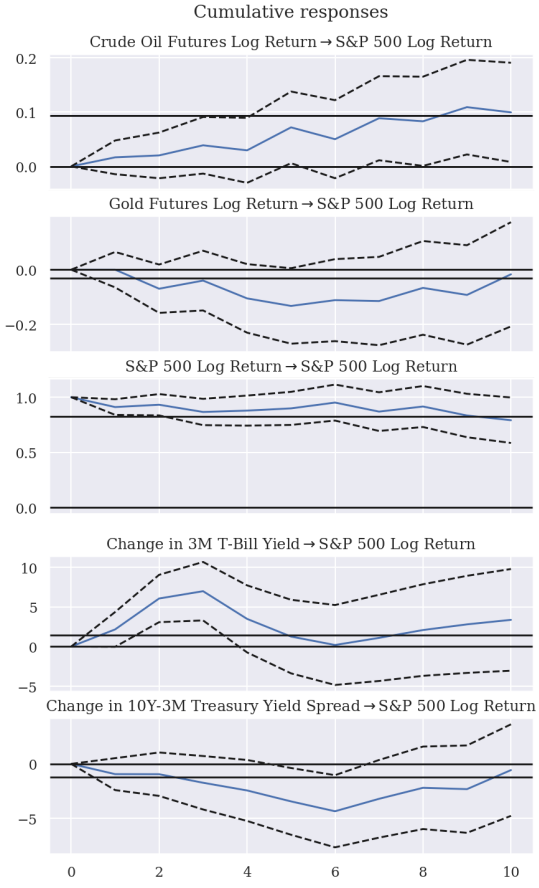

Figure 15. S&P 500's Cumulative Responses to Shocks

5.1.3 Residual Analysis – Durbin-Watson Test

Furthermore, similar to the Ljung-Box test we did for our autoregressive model, we check the autocorrelation of the VAR model residuals using the Durbin-Watson (DW) test. We expect a low degree of serial correlation left in the model residuals. The outcome of the DW test ranges from 0 to 4. A predictor is considered strongly positively autocorrelated if the outcome is close to 0, strongly negatively autocorrelated if the number is close to 4, and a result close to 2 refers to a low degree of autocorrelation. As observed in **Table 2**, we find that the DW test results for all five predictors are close to 2, indicating that there is much structure left in the residuals, which is in line with our expectation.

Table 2. Durbin Watson Test Results

	Crude Oil	Gold	S&P 500	T-Bill Yield	Yield Spread
Outcome	1.9987	2.0008	1.9881	1.9899	1.9986

5.2 Long Short-Term Memory Model

5.2.1 Introduction to Long Short-Term Memory Model

Long Short Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) with a chain of repeating modules of neural network that is able to handle long-term dependencies. Each repeating module in LSTM consists of four layers: forget gate, input gate, cell state, and output gate. The **forget gate layer** is given by the equation $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ whereby a sigmoid function is applied over the equation to decide whether to make the network keep or forget certain information from the previous timestamp (i.e., the network will forget everything when f_t is 0 or keep everything if f_t is 1). The **input gate layer** is used to determine how important the new information is from the input and this is given by the equation $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$, with sigmoid function applied on it. The new information passed to the cell state is then given by $N_t = tanh(W_N \cdot [h_{t-1}, x_t] + b_N)$ with tanh as the activation function. After deciding on the new information to be stored, the **cell state** is then updated with the equation $C_t = f_t \times C_{t-1} + i_t \times N_t$. Lastly, the **output gate layer** decides which parts of the cell state to output to the next module. This is given by the equations $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ and $h_t = o_t \times tanh(C_t)$.

5.2.2 Data Preparation

To begin with data preprocessing for the LSTM model, "StandardScaler" in scikit-learn is used to scale and transform the data to avoid any dominating features. In order to fit the data for LSTM model input and output, the data is reshaped into a 3 dimensional tensor. The reshaped data is then re-split into training and testing data. The final shape of x_train, y_train, x_test, and y_test are (939, 52, 5), (939, 3, 1), (50, 52, 5), and (50, 3 1). **Figure 16** shows the first data point in the test set. The input data is marked in green, with 52 weeks from 2020-11-08 to 2021-10-31 and 5 features (crude oil, gold, S&P 500, T-Bill yield, and yield spread), hence, has a shape of (1, 52, 5). The output data is marked in orange, with 3-step ahead S&P 500 forecasts from 2021-11-07 to 2021-11-21, hence has a shape of (1, 3, 1).

	Crude Oil Futures Log Return	old Futures Log Retur	S&P 500 Log Return	Change in 3M T-Bill Yield	Change in 10Y-3M Treasury Yield Spread
10/25/2020	-0.492453913	-0.045301142	-0.289099853	-0.066290980	0.802753621
11/1/2020	-2.006279859	-0.604372241	-2.483899323	0.016003296	0.153406254
11/8/2020	0.663113259	1.489167306	2.872486847	0.016003296	-0.297320004
11/15/2020	1.409607112	-1.450790356	0.819626459	0.016003296	0.565932989
11/22/2020	0.886385398	-0.356679046	-0.390196801	-0.277904535	-0.289680534
11/29/2020	1.404194239	-2.104623885	0.866113124	0.192348013	-0.007022613436
12/6/2020	0.272906189	1.151640731	0.620762750	0.039515972	0.963180689
12/13/2020	0.102424208	0.015766272	-0.471585719	-0.101559906	-0.495943971
12/20/2020	0.956444929	0.938149135	0.450162532	0.192348013	0.313832092
12/27/2020	-0.351282852	-0.197320235	-0.139796455	-0.042778217	-0.106335223
1/3/2021	0.089785560	0.215438473	0.522936767	-0.101559906	0.000619144101
1/10/2021	1.343883792	-1.36795847	0.684713979	0.192348013	1.32987222
1/17/2021	0.021416766	-0.178479318	-0.687010336	-0.042778217	-0.014662255
1/24/2021	-0.052759928	0.515721728	0.732267822	-0.042778304	0.000616402148
1/31/2021	-0.045734365	-0.256969045	-1.469996487	-0.24263561	0.191602378
2/7/2021	1.555640775	-0.886264590	1.821447399	-0.277904557	0.787474010
2/14/2021	0.811471228	0.170145398	0.442214803	0.157079066	0.145765656
2/21/2021	-0.092568797	-1.114114927	-0.366942316	-0.125072472	1.207641525
2/28/2021	0.670765247	-1.188379825	-1.098978835	0.157079066	0.795114212
3/7/2021	1.308923634	-0.790843123	0.266314401	-0.066290892	0.779835208
3/14/2021	-0.155650564	0.444205200	1.015481723	-0.019265651	0.649965355
3/21/2021	-1.240247863	0.447211019	-0.388831937	-0.219122989	0.902065754
3/28/2021	-0.156837863	-0.288110996	0.579752399	0.133566434	-0.618174780
4/4/2021	0.123910855	-0.206147905	0.403387573	-0.066290892	0.206880865
4/11/2021	-0.677255155	0.325475513	1.044047372	-0.042778274	-0.052859172
4/18/2021	1.129141731	0.759134511	0.498744274	0.074784807	-0.740405373
4/25/2021	-0.313008864	-0.117221357	-0.121232585	0.098297486	-0.091055484
5/2/2021	0.402290854	-0.295336451	-0.058521183	-0.125072467	0.588504663
5/9/2021	0.358679861	1.381385984	0.440202027	0.074784807	-0.442468583
5/16/2021	0.112439522	0.080672839	-0.650122721	-0.019265651	0.474259174
5/23/2021	-0.533948732	0.784518275	-0.248244258	-0.007509329	0.000616540886
5/30/2021	0.758569173	0.488090169	0.411816838	0.074784807	-0.419550212
6/6/2021	0.876221916	-0.345478179	0.185879134	0.098297486	-0.205647269
6/13/2021	0.318234922	-0.340784117	0.103780356	0.051272233	-0.763322615
6/20/2021	0.168256009	-2.532548067	-0.869596084	0.133566456	-0.159810409
6/27/2021	0.590337041	0.129914106	1.057155766	0.192348013	0.550653129
7/4/2021	0.253921166	0.066943027	0.622107864	-0.042778304	-0.755838773
7/11/2021	-0.169060813	0.553621786	0.095706807	0.074784883	-0.602896166
7/18/2021	-0.715306851	0.030539524	-0.474147713	-0.042778304	-0.38135316
7/25/2021	0.045799758	-0.367923751	0.737454239	0.098297301	-0.152170825
8/1/2021	0.454802900	0.18271388	-0.224706757	-0.042778304	-0.312599031
8/8/2021	-1.494838612	-1.277303994	0.320290915	0.051272266	0.379047204
8/15/2021	0.022269780	0.281060366	0.225707114	0.039515928	0.046453768
8/22/2021	-1.751705587	0.062453675	-0.314439839	-0.007509335	-0.259123304
8/29/2021	1.790566469	0.739478791	0.560942760	0.016003296	0.405505139
9/5/2021	0.126266202	0.250005799	0.171274653	-0.019265672	0.107582267
9/12/2021	0.093328588	-1.005632147	-0.779615212	-0.007509335	0.168828696
9/19/2021	0.565858086	-1.001687326	-0.380816270	-0.101559818	0.306192962
9/26/2021	0.487951756	-0.064127867	0.140633621	-0.019265651	0.718720422
10/3/2021	0.447541762	0.099379640	-0.998165890	0.051272244	0.130485970
10/10/2021	0.805180335	0.087469463	0.257782801	0.157079042	0.879147934
10/17/2021	0.648953814	0.182258555	0.683004690	-0.007509335	-0.198007921
10/24/2021	0.308404732	0.579560710	0.610199513	0.157079087	0.520095758
10/31/2021	-0.062933659	-0.357291097	0.481565759	-0.007509335	-0.725125623
11/7/2021	-0.536594508	0.689001149	0.755870941	-0.078047237	-0.725126847
11/14/2021	-0.130420540	1.073970591	-0.198741492	0.051272266	0.970820814
11/21/2021	-1.125924867	-0.438985074	0.065541335	0.039515928	-0.358435060
11/28/2021	-2.059550426	-1.555782603	-0.994022552	-0.007509335	-0.388992773
12/5/2021	-0.540604784	-0.146928782	-0.580521041	0.016003296	-1.053619828

Figure 16. One Sample in Training Data

5.2.2 Grid Search using Expanding Window Cross Validation

The overall LSTM model architecture used in this project is shown in **Figure 17** (with two LSTM layers). After each LSTM layer, a dropout layer is added to randomly set input units to zero in order to avoid overfitting. When working with time-related data, we cannot randomly shuffle the data, hence, time-based cross validation is used rather than the usual cross validation method. There are several approaches to conduct time-based cross validation. In particular, we are using the expanding window method whereby we successively consider each week as the testing data and then using all previous weeks as the training data. The number of splits is set to 3 for performing expanding window cross validation. **Figure 18** demonstrates an example of the splits in the expanding window cross validation method.

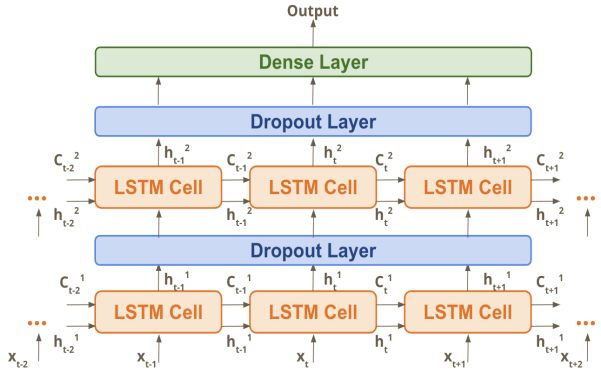

Figure 17. LSTM Model Architecture

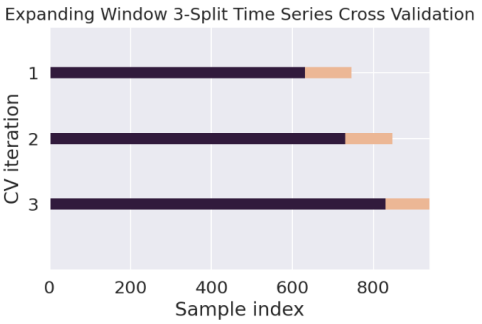

Figure 18. Illustration of Expanding Window CV

Hyperparameter tuning is performed to select the best LSTM model. This is conducted by performing rigorous grid search using the scikit-learn GridSearchCV based on expanding window cross-validation loss to find the optimal set of hyperparameters for the LSTM model. Notably, GridSearchCV only supports scikit-learn estimators. To overcome the limitation, we defined our own LSTM class wrapping around keras LSTM layers to match with the expected class interface. **Table 3** outlines the hyperparameters tuned. The best model has one layer with 64 nodes, a dropout rate of 0.2, a learning rate of 0.001, a batch size of 32, ReLU recurrent activation function, and 25 epochs. **Figure 19** shows the training loss over the iterations and a decreasing trend is seen.

Table 3. Hyperparameter Tuning For LSTM Model

Layer1 Nodes	Layer2 Nodes	Dropout Rate
[16, 32, 64, 128]	[0, 16, 32, 64, 128]	[0.1, 0.2, 0.3]
Batch Size	**Learning Rate**	**Epoch**
[16, 32, 64]	[0.01, 0.001, 0.0001]	Early Stopping

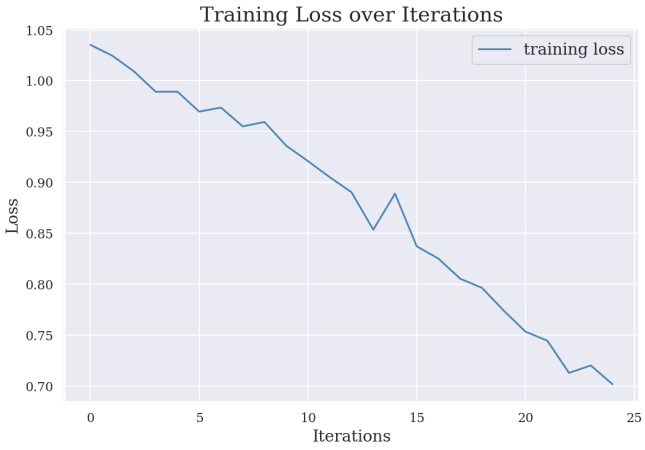

Figure 19. Training Loss Over Iterations (Best LSTM Model)

In addition, the out-of-sample three-week forecasts of S&P 500 log return is shown in **Figure 20** with each short segment representing one out-of-sample three-week forecast. The performance for each of the three forecast horizons (one-week ahead, two-week ahead, and three-week ahead) is illustrated in **Figure 21**. A more comprehensive graph that separates the three forecast horizons can be found in **Appendix D**.

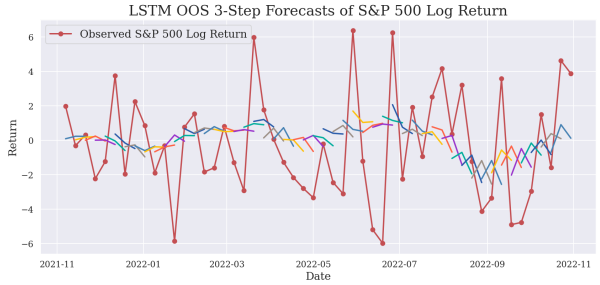

Figure 20. Best LSTM OOS 3-Step Forecasts

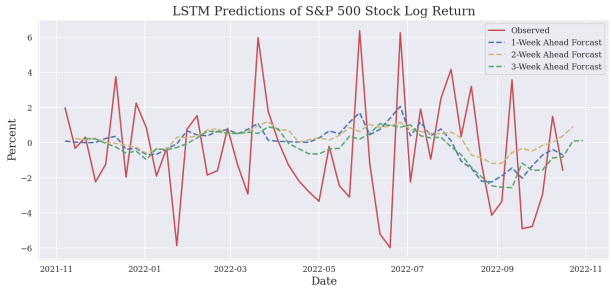

Figure 21. Best LSTM 1-, 2-, and 3-Week Ahead Forecasts

6 Results and Discussion

Model evaluation and comparison are conducted by both visual inspection and using quantitative metrics. By visual inspection (refer to **Figures 7, 8, 11, 14,** and **20**), both the historical mean model and the naive model predict a straight line, one uses the average of all training data and the other latest period's value. They do not capture any upward or downward trends in the true data. The AR model shows some small fluctuations in the OOS forecasts, capturing very little trend in the actual test data. The VAR model shows more fluctuations in the OOS forecasts than the AR model, showing that the forecasts match better with the actual test data in terms of following the ups and downs in the trend. Lastly, the final selected LSTM model has the best performance by visual inspection. For OOS forecasts, the LSTM model shows more fluctuations in terms of magnitude and these fluctuations also match well with the increases and decreases in the actual test data.

In addition, quantitative metrics are also used for model evaluation and comparison, namely Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-Squared (R2), as shown in **Equations (1), (2),** and **(3)**, respectively. In the equations, X_i denotes the actual values, $\widehat{X}_i$ denotes the predicted values, and N is the size of the testing data. The RMSE and MAPE values measure the discrepancies between the actual values and the predicted values. The smaller the RMSE and MAPE values are, the better the performance of the model. R2 measures how well the model explains the data. The closer the R2 is to 1, the better the model performance. In addition, the model is performing worse than solely predicting the mean if the R2 value is negative (i.e., R2 equals to zero when predicting the mean).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{X}_i - X_i)^2}{N}} \quad (1)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X_i - \widehat{X}_i}{\widehat{X}_i}\right| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\widehat{X}_i - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \quad (3)$$

Table 4 shows the results of the baseline models as well as the alternative models. For each of the three performance metrics (RMSE, MAPE, and R2), the best values for 1-step, 2-step, and 3-step ahead predictions are in bold fonts. Recall that "step" here can also be interpreted as "week". The results show that for 1-step ahead predictions, the final selected LSTM model is the best model in terms of RMSE, MAPE, and R2; and the historical mean model outperforms the naive, AR(4), and VAR(10) models based on RMSE, MAPE, and R2. For 2-step and 3-step ahead predictions, the historical mean model is the best model in terms of RMSE, MAPE, and R2; and the runner-up is the LSTM model (i.e., outperforming naive, AR(4), and VAR(10) models, with RMSE, MAPE, and R2 values only slightly worse than the historical mean model. However, as the advantage of LSTM in 1-step ahead forecasting is more outstanding (having a positive R2), we conclude that LSTM is overall the best model.

Table 4. RMSE, MAPE, and R2 Values Comparison

	Model	1-Step Ahead	2-Step Ahead	3-Step Ahead
RMSE	Historical Mean	3.0770	**3.0770**	**3.0770**
	Naive	3.4873	3.4873	3.4873
	AR(4)	3.0881	3.1324	3.1676
	VAR(10)	3.2412	3.2782	3.3015
	LSTM	**2.9108**	3.0824	3.1285
MAPE	Historical Mean	1.0689	**1.0689**	**1.0689**
	Naive	1.8922	1.8922	1.8922
	AR(4)	1.1201	1.1422	1.0768
	VAR(10)	1.3203	1.4188	1.2769
	LSTM	**1.0607**	1.2046	1.2122
R2	Historical Mean	0.0	**0.0**	**0.0**
	Naive	-0.2845	-0.2845	-0.2845
	AR(4)	-0.0629	-0.0470	-0.0303
	VAR(10)	-0.17095	-0.1467	-0.1193
	LSTM	**0.05565**	-0.005021	-0.005021

7 Conclusion

This project examines the predictability of S&P 500 returns within the time horizon of three weeks, using information taken from the crude oil, gold, and treasury markets. Five models are

constructed with 20 years of historical data and applied to generate the predictions of interest: the Historical Mean Model, Naive Model, AR, VAR, and LSTM. For statistical models such as AR and VAR, metrics such as the AIC are used for model selection. To determine the optimal design and hyperparameter values for our recurrent neural network, we rely on mean validation loss from expanding window time series cross-validation (k=3). With RMSE, MAPE, and R2 as the fundamental performance measures, multivariate LSTM demonstrated better predictive ability than linear, econometric models, and its 1-Step Ahead forecasts surpassed the historical mean method—the "horizontal line benchmark" widely embraced for financial and economic forecasting. Finally, the superior performance of multivariate LSTM also suggests implicit nonlinear dependencies within the financial data and constitutes as encouraging results for the effectiveness of gold, crude oil, and bond markets being predictors of the stocks. That is, these time series may indeed aggregate new information that facilitates stock market forecasting, a possibility which we will keep investigating as we continue the quest for the best input-output specification for equity forecasting.

8 FUTURE WORK

In order to further improve our forecasting, additional models and processing techniques should be taken into consideration. Our future works are explained below:

8.1 Conversion Back to Original Time Series

Currently, our objective is set to predict returns. A more conventional approach is to convert it back to the original S&P 500 index time series when evaluating the results.

8.2 A Comprehensive Forecasting Model with Trend, Seasonal, and Cyclic components

Because our objectives involve both forecasting and inference, we take the most prudent approach of making all the time series stationary. Instead of relying on first order differencing for detrending, we can attempt time series decomposition and fit models to all components, such as using Holt-Winters for handling trend and seasonality and autoregressions for modeling the cyclic dynamics. Alternatively, we can directly fit ARIMA and VECM models on the original data justified by the potential cointegration among the time series.

8.3 Other Time Series Forecasting Models

There are many other alternatives we plan to experiment with in the future, such as ConvLSTM, a method specialized in multiple variable, multi-step forecasting challenges, and state space models, such as Kalman Filter. In addition, we will try to combine forecasts based on different modeling approaches to minimize variance.

8.4 More rigorous residual diagnostics, stability assessment, and model selection

In our study, we conducted simple white noise tests on model residuals. For further analysis, we plan to try recursive parameter estimation and plot out the recursive residuals and the cumulative sum.

8.5 Application to Portfolio optimization

To see how our predicted returns work in practice, we plan to conduct asset allocation (among gold, energy, and stock markets) according to the Markowitz Portfolio Theory and find out the actual profitability of our strategy through backtesting.

9 REFERENCES

Two forecasting textbooks are found to be useful for guiding this project.

1. **González-Rivera, G. (n.d.). Forecasting for economics and business. Routledge.**
   This book discusses topics such as statistics and time series, and forecasting with autoregressive (AR) processes, which are relevant to the baseline methods that are implemented in this project.

2. **Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice. OTexts.**
   This book discusses time series regression models, time series decomposition, and advanced forecasting methods (e.g, vector autoregressions, neural network models), which can be relevant to developing improvement models (improvement on simple baseline models) for this project.

APPENDIX

A Time Series Analysis of Other Variables

The observations made for S&P 500 index log return in Section 2.2 are also observed for other variables. As shown in **Figures A.1 to A.4**, the crude oil futures log return, gold futures log return, change in 3-month treasury bill yield, and change in 10Y-3M treasury yield spread all appear stationary across the years. ACF (autocorrelation function) plots and PACF (partial autocorrelation function) plots are constructed. Since there are no extreme spikes in the ACF and PACF plots, there appears to be no autocorrelation in the time series, showing evidence that the data are stationary. To confirm stationarity, the Augmented Dickey-Fuller (ADF) test is applied to the data. Since the ADF test statistics are large negative values and the p-values are very small for all variables, the null hypothesis is rejected and it can be concluded that the crude oil futures log return, gold futures log return, change in 3-month treasury bill yield, and change in 10Y-3M treasury yield spread are stationary. Looking at their distributions, the distributions are centered around zero and form symmetrical bell shapes. The data also roughly follow a straight line on the Q-Q plots. This shows evidence that the data are normally distributed.

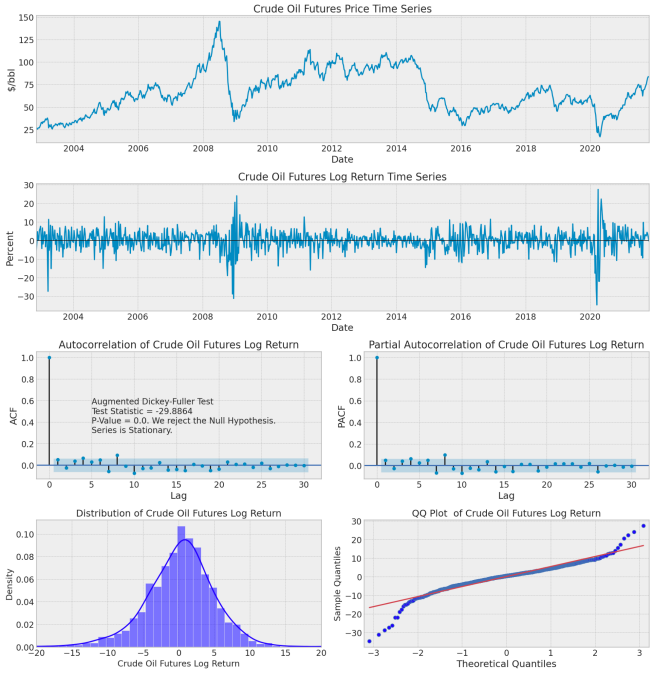

Figure A.1. Time Series Analysis of Crude Oil Futures Price

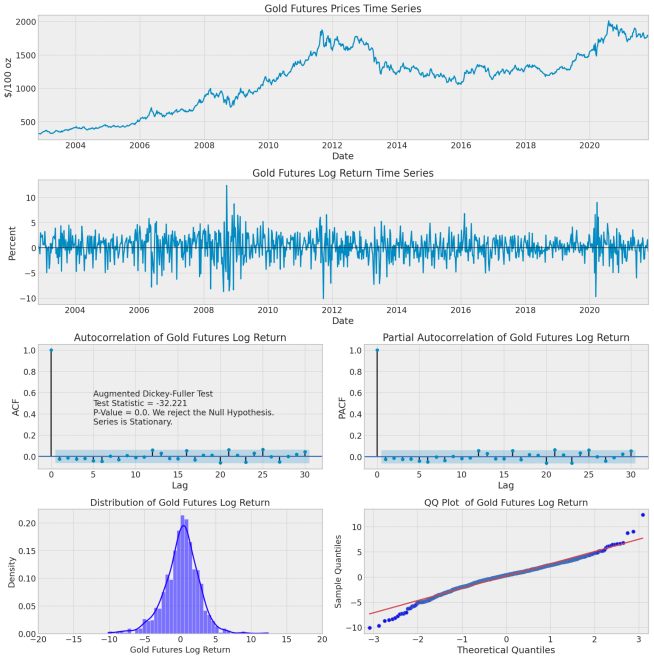

Figure A.2. Time Series Analysis of Gold Futures

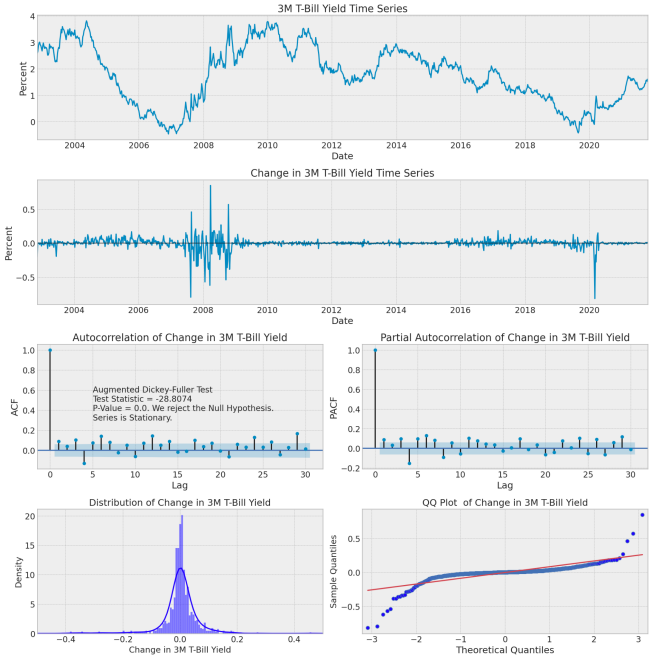

Figure A.3. Time Series Analysis of 3-Month Treasury Bill Yield

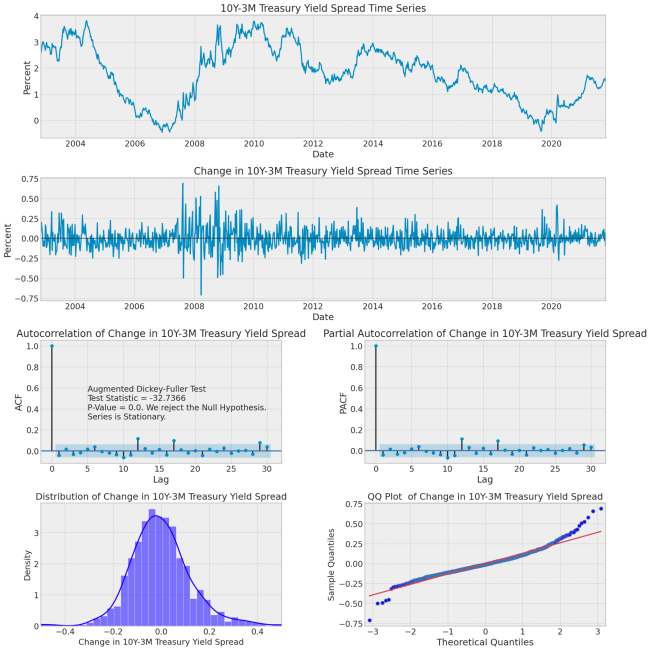

Figure A.4. Time Series Analysis of 10Y-3M Treasury Yield Spread

B Model Summary for AR(4)

The supplementary figures are plots resulting from the Autoregressive model with selected order of 4. **Figure B.1.** displays the model summary of AR(4) using AIC model selection criterion. Comprehensive model forecast can be viewed in **Figure B.2.** The additional residual check is referred to in **Figure B.3.**

AutoReg Model Results

Dep. Variable:	S&P 500 Log Return	No. Observations:	991
Model:	AutoReg(4)	Log Likelihood	-2260.410
Method:	Conditional MLE	S.D. of innovations	2.390
Date:	Tue, 06 Dec 2022	AIC	1.755
Time:	16:47:03	BIC	1.784
Sample:	12-08-2002	HQIC	1.766
	- 10-31-2021		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1825	0.077	2.376	0.018	0.032	0.333
S&P 500 Log Return.L1	-0.0707	0.032	-2.221	0.026	-0.133	-0.008
S&P 500 Log Return.L2	0.0470	0.032	1.478	0.139	-0.015	0.109
S&P 500 Log Return.L3	-0.0662	0.032	-2.083	0.037	-0.129	-0.004
S&P 500 Log Return.L4	-0.0417	0.032	-1.310	0.190	-0.104	0.021

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.3147	-1.5315j	2.0184	-0.1371
AR.2	1.3147	+1.5315j	2.0184	0.1371
AR.3	-2.1099	-1.2005j	2.4275	-0.4177
AR.4	-2.1099	+1.2005j	2.4275	0.4177

Figure B.1. Result for AIC

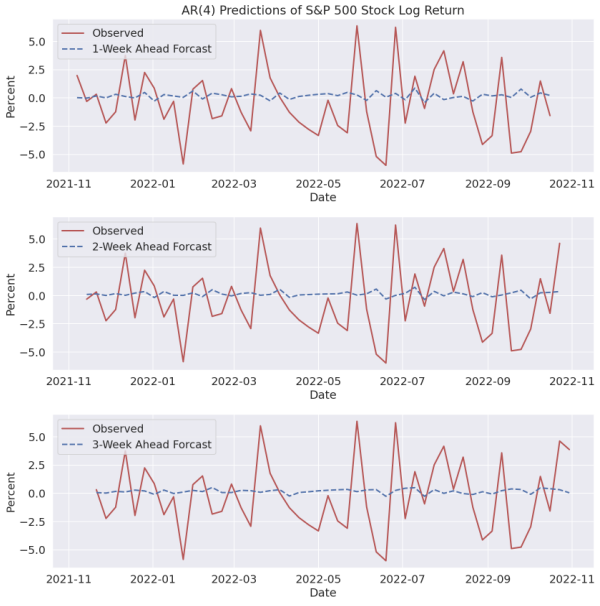

Figure B.2. AR OOS One-step, Two-step, and Three-step Ahead Model Performance

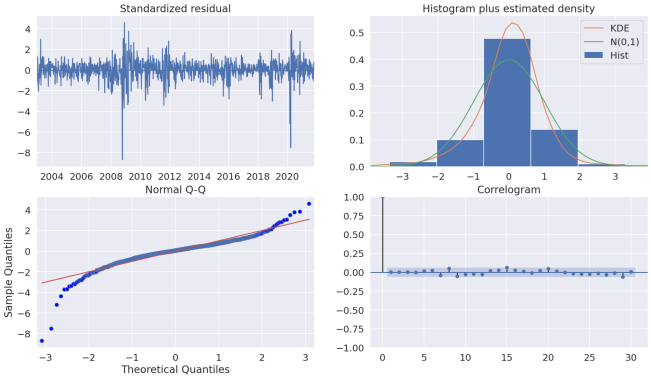

Figure B.3. Residuals Diagnostic Plots

C Forecast Performance for VAR(10)

The complementary plots in **Figure C.1.** are non-cumulative impulse responses. **Figure 3.2.** shows the forecasts for the other four predictors by the VAR (10) model. Lastly, the one-step, two-step, and three-step ahead forecast result of S&P 500 VAR(10) is shown in **Figure 3.3.**

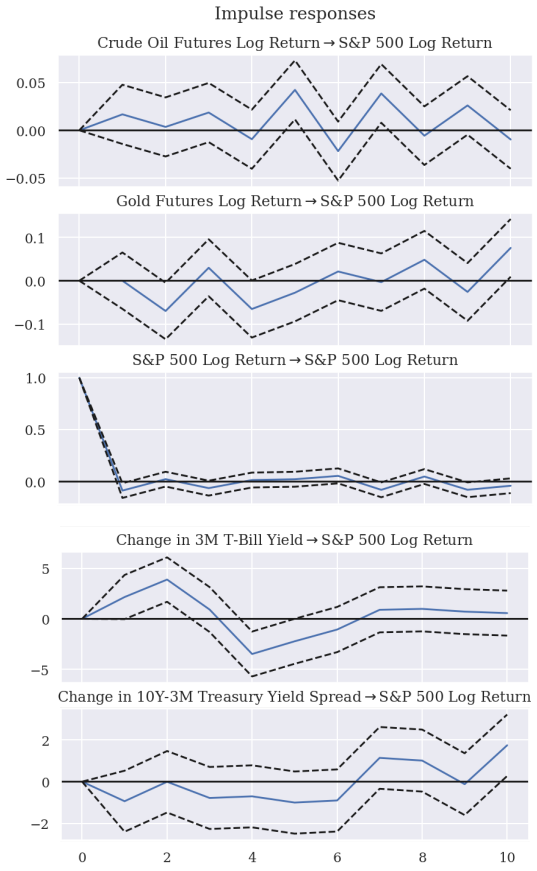

Figure C.1. Impulse response functions

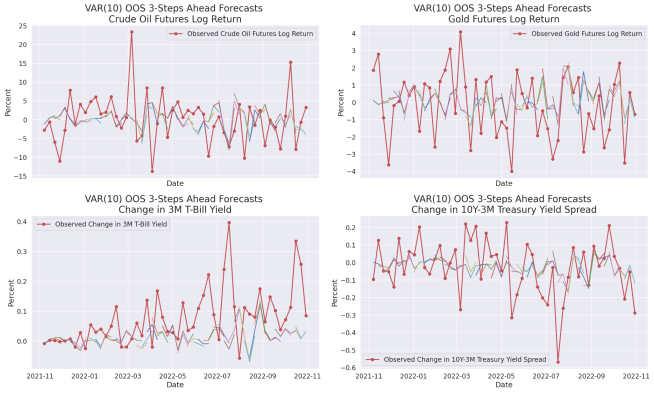

Figure C.2. VAR(10) OOS 3-Step Ahead Forecasts for Crude Oil, Gold, T-Bill Yield and Yield Spread

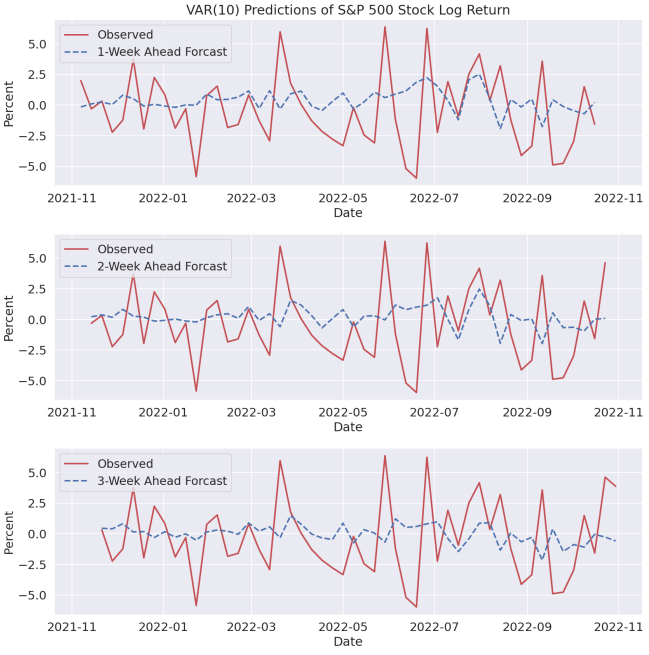

Figure C.3. VAR(10) OOS One-step, Two-step, and Three-step Ahead Model Performance

D Forecast Performance for Best LSTM Model

The one-step, two-step, and three-step ahead forecast result from the best LSTM model is shown below in **Figure D.1.**

Figure D.1. Best LSTM Model OOS One-step, Two-step, and Three-step Ahead Model Performance