GROUP WORK PROJECT 01
GROUP NUMBER: 8458

MScFE 642: Deep Learning for Finance

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Sujan Antony S	India	sonnyantony05@gmail.com	
Regulavalasa Krishna Vamsi	India	krishnavamsi8262@gmail.com	
Phan Van Nguyen	Vietnam	nphanvan10@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	**Sujan Antony S**
Team member 2	**Regulavalasa Krishna Vamsi**
Team member 3	**Phan Van Nguyen**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.

Step 1

a.Data Collection

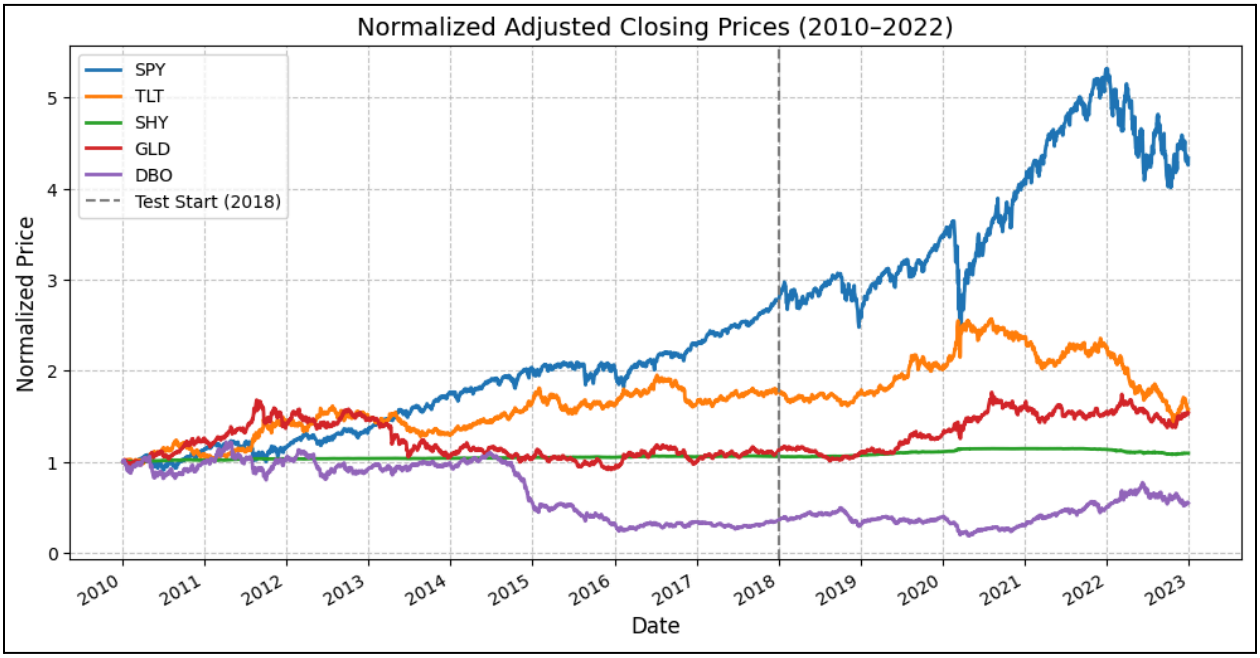
The daily price data for the five ETFs is downloaded using the yfinance library. The ETFs and their corresponding asset classes are:

- Equity (SPY): SPDR S&P500 ETF
- Fixed Income (TLT): iShares 20+ Year Treasury Bond ETF
- Cash-like (SHY): iShares 1-3 Year Treasury Bond ETF
- Precious Metals (GLD): SPDR Gold Shares
- Crude Oil (DBO): Invesco DB Oil Fund

The period is from January 1, 2010 to December 30, 2022. The Testing period is from January 1, 2018 to December 30, 2022.

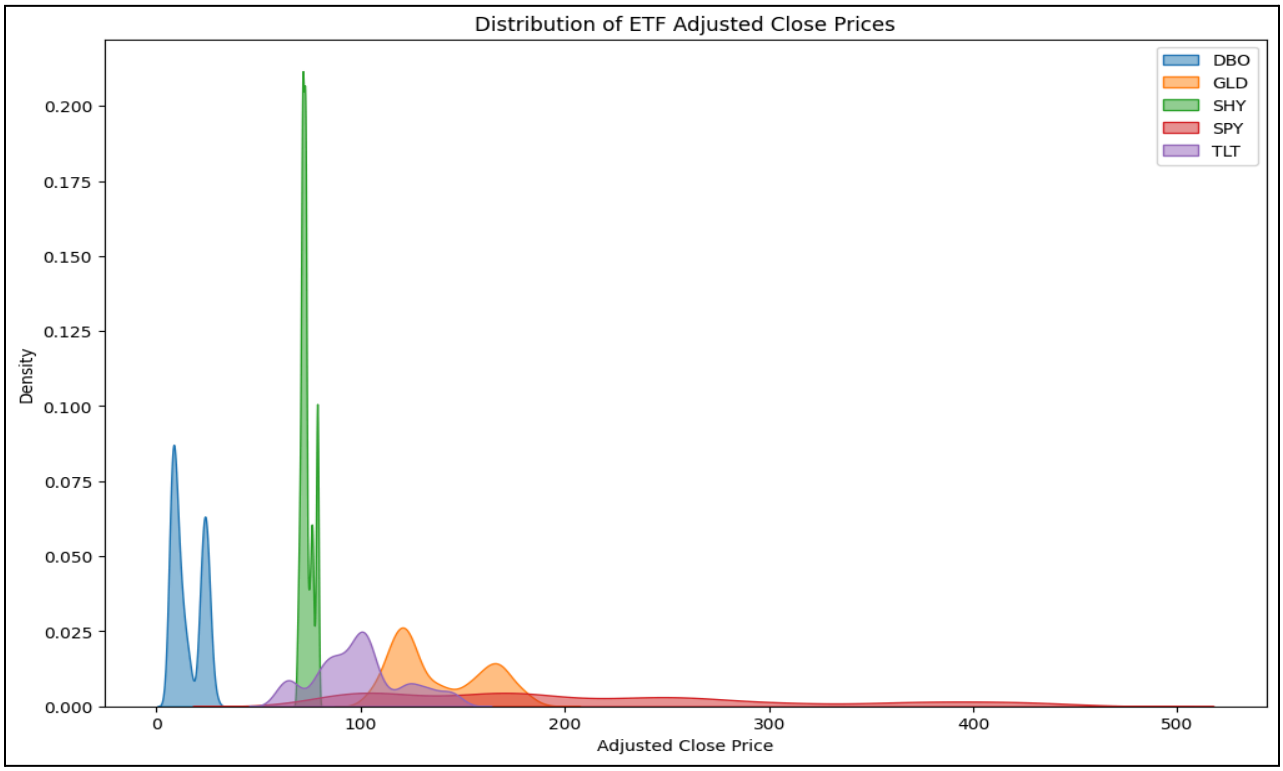
b. Exploratory Data Analysis of Five ETFs (2010–2022)

Normalized Adjusted Closing Prices (2010–2022):
The below graph shows the relative performance of SPY, TLT, SHY, GLD, and DBO over 12 years. Prices are normalized to start at 1 in 2010, enabling a comparison of growth trends regardless of initial price levels. A vertical dashed line at 2018 marks the start of a test period.



- **SPY**: It progressively climbs until 2016, then accelerates, peaking at more than 4.5 in 2020-2021, followed by modest dips. This illustrates the strong success of the US equities market, especially after 2018, which was fueled by economic recovery and positive market circumstances.
- **TLT**: Shows modest increase, peaking at 2.5-3 in 2020-2021 before declining somewhat. This implies a consistent rise in long-term bond prices, most likely due to low interest rates in recent years.
- **SHY**: Remains constant near 1, indicating minor price movement. This is consistent with SHY's role as a low-risk, short-term Treasury ETF geared toward stability.
- **GLD**: Has an inconsistent performance, reaching around 2 in 2011-2012 and 2020 but mainly stabilizing between 1 and 1.5. Gold's volatility indicates its vulnerability to economic instability and inflationary predictions.
- **DBO**: Starts with a peak near 2 in 2011–2012, then declines sharply to below 1 by 2016, remaining low through 2022. This indicates poor long-term performance in the oil market, possibly due to oversupply or shifting energy dynamics.

Distribution of ETFs:



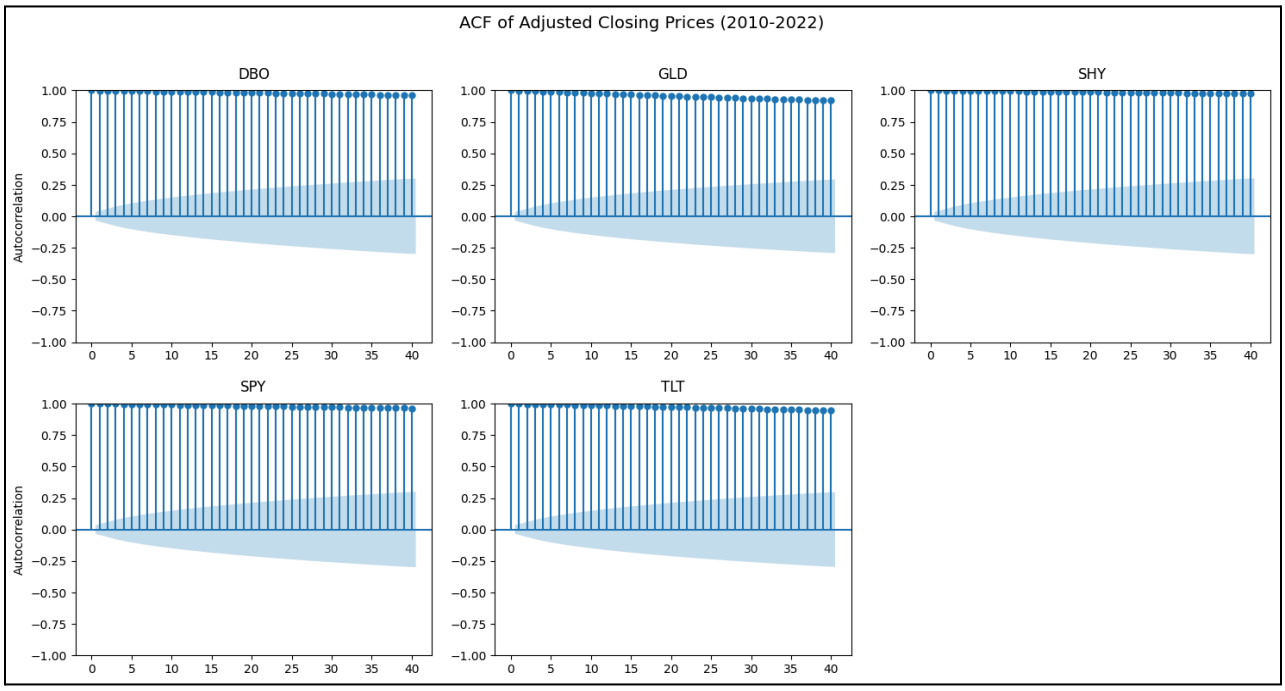
- The distribution demonstrates that SPY (red) has the broadest price range, exceeding $500.
- SHY (green) is firmly grouped around a low price range, which is expected given that it represents short-term Treasury bonds.
- GLD, TLT, and DBO have broader distributions, indicating more volatile price changes.

Summary Statistics on Returns:

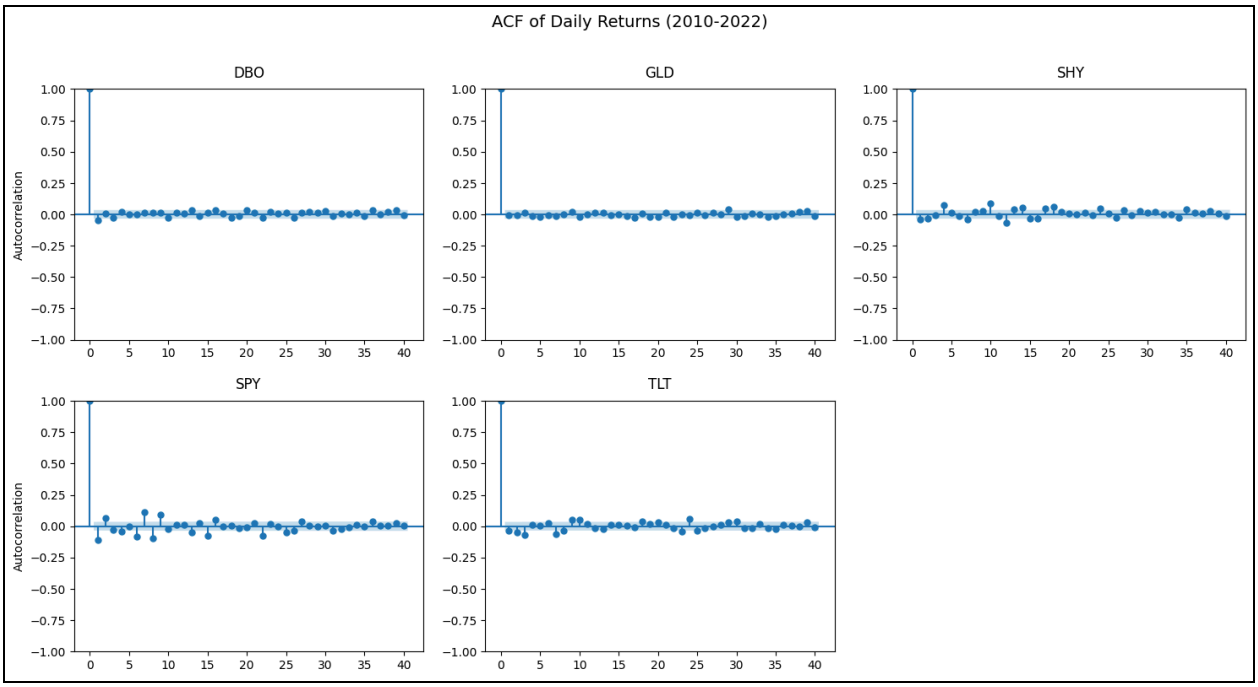
ETF	Mean	STD	Skewness	Kurtosis
SPY	0.051%	1.109%	-0.52	10.81
TLT	0.018%	0.951%	0.06	4.21
SHY	0.002%	0.069%	0.11	8.75
GLD	0.018%	0.988%	-0.44	4.73
DBO	0.0002%	1.914%	-0.52	5.07

From the above summary stats we can see that SPY offers high returns with significant risk (high volatility, negative skewness, fat tails). TLT balances moderate return and risk, SHY prioritizes stability, GLD shows moderate risk with diversification potential, and DBO carries high risk with little average gain.

ACF on Daily Prices and Returns:



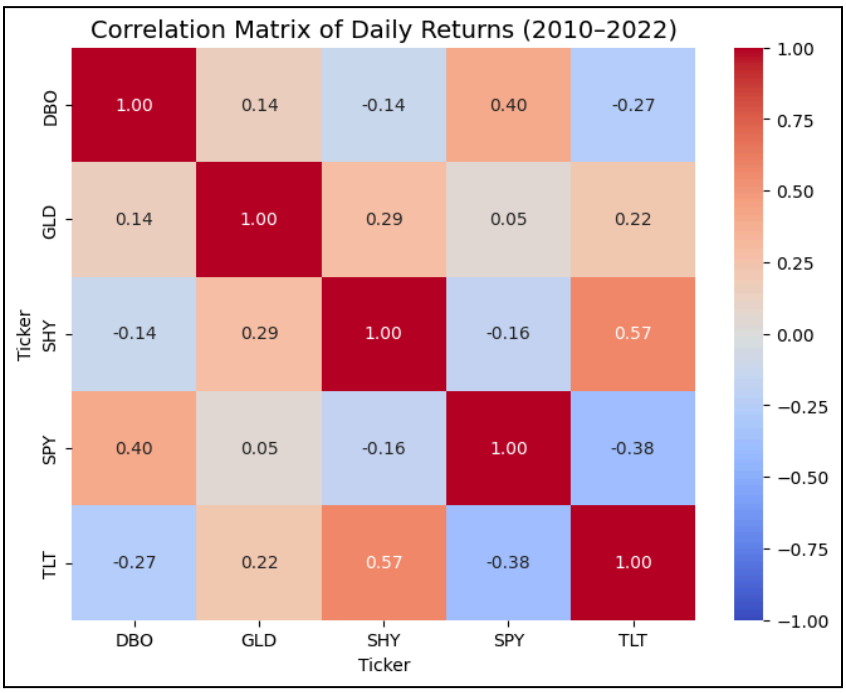
The ACFs for the price series (first set of plots) begin high and decline gradually. This is typical of non-stationary or trending series (such as most price series), in which today's price is highly associated with yesterday's price, and so on.



The ACFs for daily returns are typically near zero at various lags. This shows that returns have little to no autocorrelation, which is compatible with the popular "random walk" or "efficient market" assumption. Essentially, knowing yesterday's return has limited predictive value for today's return.

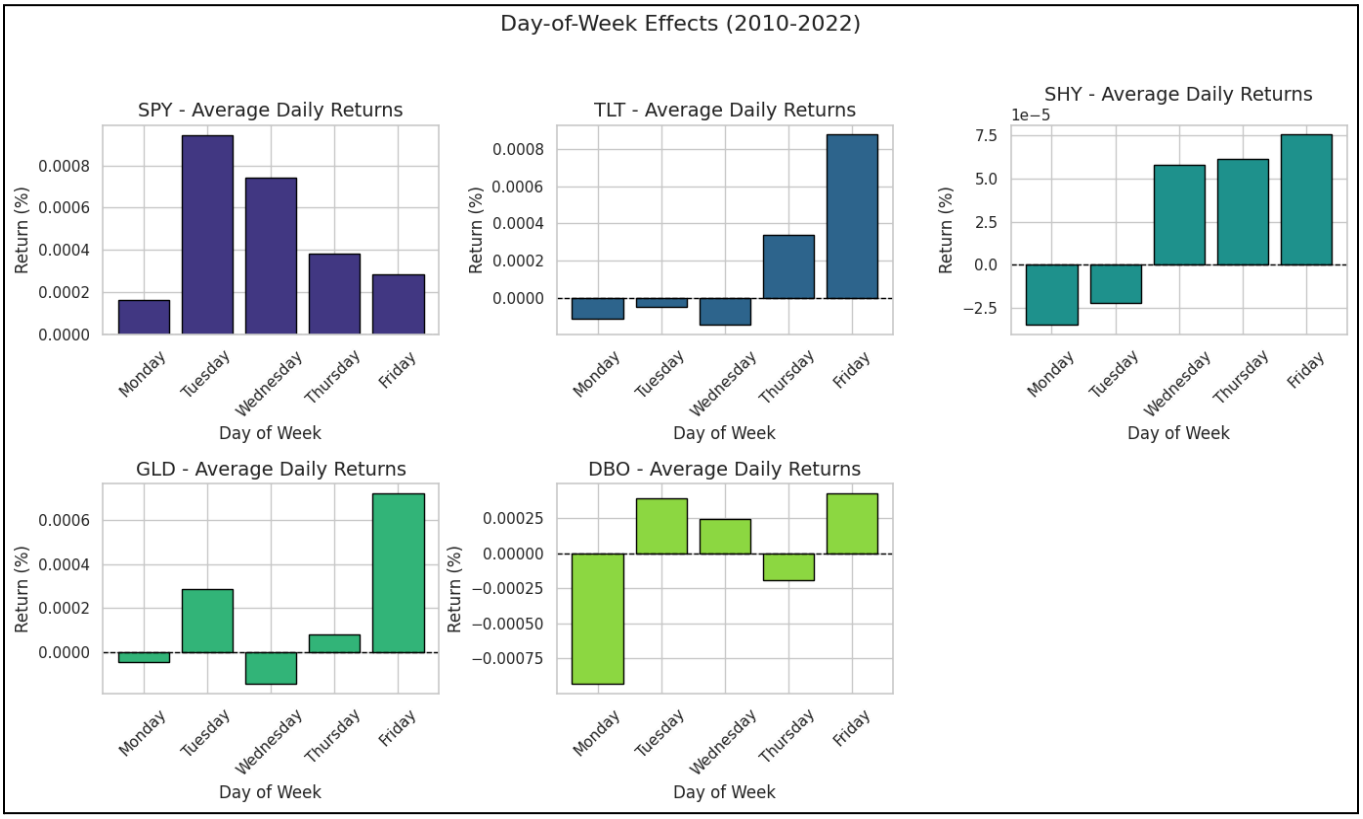
In summary, prices frequently exhibit substantial autocorrelation (they "remember" previous levels), whereas returns are considerably closer to white noise with minimal serial dependency.

Correlation Matrix Of ETFs Daily Returns:



The matrix identifies diversification opportunities: SPY and TLT's negative correlation suggests a potential hedge, but GLD's low correlations indicate that it can lower portfolio risk. SHY and TLT's positive correlation shows their similar bond qualities, while DBO's link to SPY relates to equities market patterns.

Day-of-Week Returns plot & Kruskal-Wallis Test:



Kruskal-Wallis Results:

- SPY: H-stat = 0.76, p-value = 0.944
- TLT: H-stat = 9.61, p-value = 0.048
- SHY: H-stat = 16.91, p-value = 0.002
- GLD: H-stat = 4.40, p-value = 0.355
- DBO: H-stat = 4.74, p-value = 0.315

The plot indicates that TLT and SHY have differing returns on different weekdays. The Kruskal-Wallis test reveals that these changes are unlikely to be attributable to chance, while SPY, GLD, and DBO show no meaningful pattern. As a result, Treasury bond ETFs (TLT and SHY) exhibit significant weekday performance disparities, implying a possible calendar effect specific to bond markets.

Calendar Effects – Dunn's Test Analysis:

To analyze the day-of-week effects revealed by the original Kruskal-Wallis tests, Dunn's post-hoc testing with Bonferroni correction are performed on TLT and SHY data

Dunn's test for TLT:

	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1	1	1	0.59	0.03
Tuesday	1	1	1	1	1
Wednesday	1	1	1	1	0.34
Thursday	0.59	1	1	1	1
Friday	0.03	1	0.34	1	1

A statistically significant difference can be observed between Monday and Friday ($p = 0.03$), showing that TLT's returns on Friday varied from those on Monday. This could indicate a calendar effect related to long-term Treasury bond returns.

Dunn's test for SHY:

	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1	1	0.06	0.05	0.01
Tuesday	1	1	0.32	0.22	0.07
Wednesday	0.07	0.32	1	1	1
Thursday	0.05	0.22	1	1	1
Friday	0.01	0.07	1	1	1

Significant differences are found, especially among Monday and Friday ($p \approx 0.01$) as well as Monday and Thursday ($p \approx 0.050$), suggesting day-of-week effects in short-term Treasury returns.

These results reveal that whereas stock (SPY), gold (GLD), and crude oil (DBO) do not display significant day-of-week anomalies, Treasury bond ETFs (TLT and SHY) may be influenced by calendar effects.

Step 2

a. Model architecture

The chosen model architecture is a Sequential neural network designed for time series prediction using Long Short-Term Memory (LSTM) layers. It has an Input layer of 30 (the number of past observations chosen) and 1 (the number of features, which is 1 because only data of 1 ETF is trained for each model)

The first LSTM layer consists of 50 units and is configured to return sequences, allowing it to output a sequence of hidden states for each input time step. This is followed by a Dropout layer with a rate of 0.2, which helps to prevent overfitting by randomly setting 20% of the input units to zero during training.

The second LSTM layer also contains 50 units but does not return sequences, instead outputting only the final hidden state, which summarizes the learned information from the entire input sequence. This second layer is also followed by another Dropout layer, with the same purpose of regularization.

The final layer is a Dense output layer, which produces the final prediction based on the LSTM's output. The model is compiled using the Adam optimizer and is set to minimize the mean squared error, making it suitable for regression tasks.

b. Training and in-sample predictive performance evaluations

A model is trained separately for each of the 5 ETFs, with the final output being 5 independently trained models.

During training, the model processes 30 previous observations for each 25-day ahead returns prediction. The training loop for each ticker involves 100 epochs and a batch size of 32, allowing for efficient gradient updates.

Below are the in-sample predictive performance evaluations using RMSE and MAE:

	SPY	TLT	SHY	GLD	DBO
RMSE	0.0372	0.0402	0.0023	0.0498	0.0786
MAE	0.0279	0.0310	0.0018	0.0391	0.0617

DBO has the highest RMSE (0.0786) and MAE (0.0617) out of the 5, drastically higher than the rest, suggesting it has the most significant prediction errors among the models. In contrast, **SHY** shows an astonishingly low RMSE and MAE of 0.0023 and 0.018, signaling that it has learned the dataset well. Meanwhile, **SPY, TLT, and GLD** show relatively low and similar RMSE and MAE (~0.03 - 0.04), indicating that they are relatively good at capturing the patterns within the data.

Overall, it seems that aside from the DBO model, the remaining 4 models can somewhat be reliable for developing a trading strategy. This result might be because of DBO's high variance as stated in the EDA.

c. Out-sample predictive performance evaluations

Below are the out-sample predictive performance evaluations using RMSE and MAE:

	SPY	TLT	SHY	GLD	DBO
RMSE	0.0591	0.0481	0.0052	0.0422	0.1112
MAE	0.0423	0.0369	0.0035	0.0328	0.0889

The trend shown in the in-sample evaluation remains, **DBO** still performs the worst, with an RMSE of 0.1112 and an MAE of 0.0889, significantly higher than the rest**. SHY** continues to outperform, with a low out-of-sample result of 0.0052 and 0.0035, demonstrating its strong predictive capability even outside the training dataset. **SPY, TLT, and GLD** all have relatively good out-of-sample results, with their MAE level still at around 0.03 - 0.04.

Overall, similar to the in-sample result, with the exception of DBO, the remaining 4 ETFs' models all seem to generalize well to unseen data, as indicated by their similar RMSE and MAE to the in-sample evaluations. This suggests that the output of these 4 models can be used for the trading strategy.

d. Trading Strategy

Since each asset class has its own returns characteristics, we have decided to base our trading strategy on each asset class' mean and standard deviations.

Whenever an asset's predicted 25-day returns would cross the upper (lower) limit, we would long (short) the asset within that 25-day period.

The upper (lower) limit is defined as the mean of the asset plus (minus) 1.5 standard deviations within the training period. This is to avoid noises and to attempt to capture the most meaningful signals.

The trading strategy is applied to 4 ETFs, except for DBO due to the high likelihood of false positives.

e. Backtest

We perform backtesting on the testing period (from 2018 to 2022).
The resulting total period returns for the asset classes are as follows:

- SPY: -5.2%
- TLT: 0.94%
- SHY: 3.56%
- GLD: 0% (did not have any buy or sell signal)

The result was consistent with the out-of-sample evaluation findings. The asset classes with lower RMSE and MAE performed the best, as the trading strategy heavily relied upon the prediction output of the models.

During the same period, the assets had the following returns:

- SPY: 56.50%
- TLT: -11.99%
- SHY: 3.26%
- GLD: 36.55%

Our strategy outperformed on TLT and SHY, and way underperformed on SPY and GLD. Similar to the above findings, it is best that we only make use of models with low errors.

Step 3

a. Model architecture

The chosen model's architecture is similar to Step 2, with the only difference being the number of features being 5 instead of 1.

As a result, both the initial Input layer and the Dense output layer have a parameter of 5 instead of 1.

b. Training and in-sample predictive performance evaluations

A model is trained on the data of all 5 ETFs, with the final output being only 1 trained model.

The training loop remains the same, with 100 epochs and a batch size of 32.

Below are the predictive performance evaluations using RMSE and MAE:
In-sample

	SPY	TLT	SHY	GLD	DBO
RMSE	0.0342	0.0361	0.0021	0.0470	0.0697
MAE	0.0261	0.0280	0.0016	0.0371	0.0562

Out-sample

	SPY	TLT	SHY	GLD	DBO
RMSE	0.0567	0.0468	0.0050	0.0400	0.1107
MAE	0.0403	0.0363	0.0033	0.0315	0.0897

Overall, it seems although the actual number for RMSE and MAE of the combined, multi-output model is slightly lower than that of the individual models, the general trend of the dataset remained the same, with SHY outperforming and DBO underperforming.

c. Trading Strategy

Since both the in- and out-of-sample performance evaluations have been relatively similar, we decided to apply the same trading strategy as Step 2.
However, as a test, we decided to keep DBO to check whether it would have any meaningful results.

d. Backtest

We perform backtesting on the testing period (from 2018 to 2022).
The resulting total period returns for the asset classes are as follows:

- SPY: 43.91%
- TLT: -4.45%
- SHY: 2.47%
- GLD: 7.96%
- DBO: -33.33%

During the same period, the assets had the following returns:

- SPY: 56.50%
- TLT: -11.99%
- SHY: 3.26%
- GLD: 36.55%
- DBO: 52.47%

The multi-output model was able to pick up the correct signals on SPY and GLD that the previous model missed, resulting in outperformance in both, relative to the strategy in Step 2.
However, TLT and SHY remain the individual models' strengths, where they continue to have superior returns. Nevertheless, the multi-output model's performance is better than that of the average market for TLT, where it had lower negative returns.
DBO is still a weak spot for both models, as both of them failed to capture the growth that DBO had in the testing period.