Feature Engineering and Selection of Multi-step LSTM Models

1. Introduction

Our goal is to accurately forecast various target variables from our dataset using a variety of Long Short-Term Memory (LSTM) models. To achieve this, we have undertaken extensive feature engineering. The primary objective is to enhance the predictive power of our models by incorporating a diverse set of features that capture various aspects of the market dynamics.

The next section of the report provides a detailed explanation of the features we have created and the rationale behind using them or dropping them.

2. Feature Engineering

Feature engineering is a critical step in the modeling process, especially when working with time series data. The features we have created can be broadly categorized into four groups:

1. Rolling and Exponential Moving Averages
2. Lagged Features
3. Momentum
4. Volatility Features
5. Additional Technical Indicators.

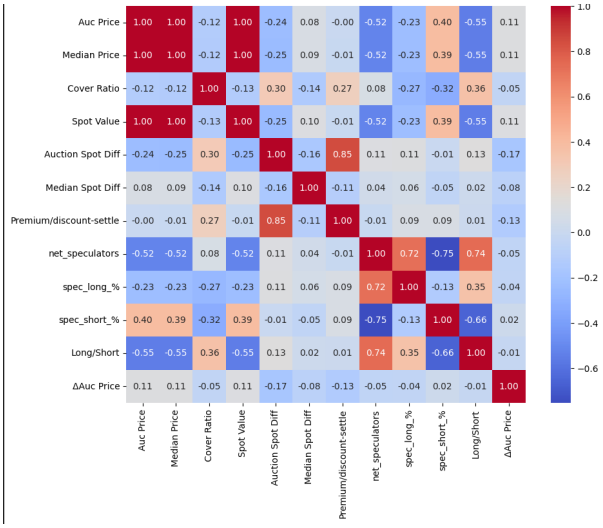
2.1. Original Features from COT and Auction Data

The initial features that we have after combining both the COT and Auction Data are as follows:

AUCTION Columns:
'Date', 'Auc Price', 'Median Price', 'Cover Ratio', 'Spot Value', 'Auction Spot Diff', 'Median Spot Diff', 'Premium/discount-settle'

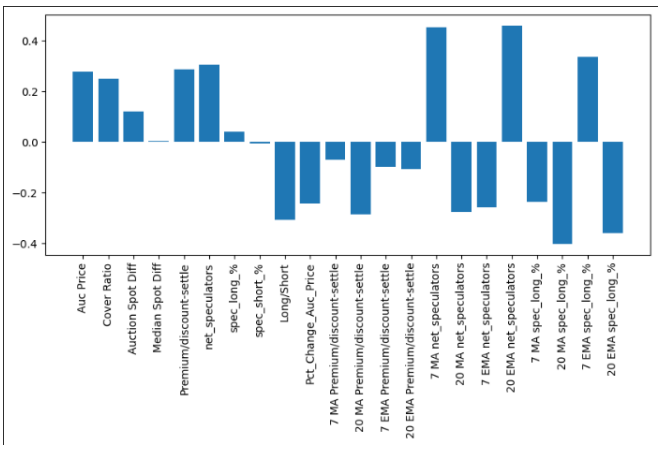
COT Columns:
'Date', 'net_speculators', 'spec_long_%', 'spec_short_%', 'Long/Short'

We used feature correlation to identify if there are features that are highly correlated. Highly correlated features can lead to multicollinearity, which occurs when one feature can be linearly predicted from another with a high degree of accuracy. This can make it difficult to determine the individual effect of each feature on the target variable, leading to unreliable or unstable coefficient estimates in models.



The correlation heatmap essentially tells us that there is a high correlation between ***Auc Price, Median Price, Spot Value***. Therefore, we will choose to keep 1 feature out of these 3 so that the model generalizes better on the rest of the features.

Additionally, we utilize a linear model (as opposed to a deep neural network) to understand what features have a considerable impact considering “Auc Price” as the target output. These features can differ depending on the target output.



The merged data of Auction and Cot looks as follows before adding more features:

	Date	Auc Price	Cover Ratio	Auction Spot Diff	Median Spot Diff	Premium/discount-settle	net_speculators	spec_long_%	spec_short_%	Long/Short	Pct_Change_Auc_Price
1	2018-06-17	15.210	1.9025	-0.0575	-0.0900	-0.003876	52749.70	5.46	0.36	15.166667	-4.037855
2	2018-06-24	14.456	2.6080	0.0280	-0.0760	0.001912	49905.83	5.37	0.52	10.326923	-4.957265
3	2018-07-01	14.990	2.5825	-0.0325	-0.0750	-0.002153	49170.57	5.38	0.65	8.276923	3.693968
4	2018-07-08	15.280	3.5700	0.0580	-0.0560	0.003868	49675.31	5.37	0.62	8.661290	1.934623
5	2018-07-15	16.005	3.0400	-0.0750	-0.1025	-0.004669	51059.21	5.50	0.69	7.971014	4.744764

The final shape of the Data is as follow:

- # of Rows: 2243
- # of Columns (Features): 10

2.2. Rolling and Exponential Moving Averages

Objective: Capture trends in key market variables

Moving averages are widely used in trading to identify the direction of the trend. By incorporating both short-term (7-period) and long-term (20-period) moving averages, our model can differentiate between short-term market noise and longer-term trends.

- **Premium/discount-settle Rolling Mean and EMA**:
  The Premium/discount-settle reflects market sentiment through the difference between the auction settlement price and the spot price. Rolling mean and EMA help smooth short-term noise and reveal underlying trends, particularly in how the market perceives future prices.

- **net_speculators and spec_long_% Rolling Mean and EMA**:
  These indicators track speculative activity in the market. Applying moving averages allows us to capture persistent trends in speculative sentiment, which can significantly impact the Auc Price.

2.3. Lagged Features

Objective: Incorporate historical context into the model, enabling it to learn from past patterns

In time series analysis, past values are often strong predictors of future values. By incorporating lagged features, we allow the model to account for temporal dependencies, which are essential for accurate multi-step forecasting. The lagged differences also help in capturing the momentum, providing additional context on whether the market is accelerating or decelerating.

- **Lagged Differences:**
  We have created features that represent the difference between the current value and its value 2 and 4 periods ago for variables such as ***Auc Price, Long/Short, net_speculators, spec_long_%, and spec_short_%***. These lagged differences capture the momentum or speed of change in these variables.

- **Lagged Values:**
  Similarly, we have introduced lagged values for the same variables, where the value at time t-2 or t-4 is used as a feature at time t. This helps the model to incorporate the temporal dependencies in the data.

2.4. Momentum and Volatility Features

Objective: Capture the strength and volatility of price movements, which are key indicators of market behavior.

Momentum and volatility are two of the most critical aspects of market behavior that traders monitor closely. Momentum indicators like RSI help in identifying potential reversal points, while volatility measures provide insights into the risk and stability of the market. By incorporating these features, our model becomes more robust in capturing the dynamic nature of the market, which is essential for accurate multi-step forecasting

- **Relative Strength Index (RSI):**
  RSI is a momentum oscillator that measures the speed and change of price movements. It helps identify overbought or oversold conditions in the market, which can be precursors to reversals or corrections.

- **Momentum:**
  We have calculated the momentum of the Auc Price, which is simply the difference between the current price and the price 10 periods ago. This feature highlights the rate of price change over time.

- **Volatility:**
  The rolling standard deviation of the Auc Price has been calculated to capture the volatility in the market. High volatility often precedes significant market moves, making it a critical feature for our model.

After incorporating all the new features in the data, the final shape of the updated data is as follow:

- # of Rows: 2243
- # of Columns (Features): 45

Below is an image of the top 5 rows of the data.

Date	Auc Price	Cover Ratio	Auction Spot Diff	Median Spot Diff	Premium/discount-settle	net_speculators	spec_long_%	spec_short_%	Long/Short	Pct_Change_Auc_Price	...	delta_spec_long_4	spec_long_T-2	spec_long_T-4	delta_spec_short_2	delta_spec_short_4	spec_short_T-2	spec_short_T-4	RSI_Auc_Price	Momentum_Auc_Price
2018-06-24	14.78	3.45	0.08	-0.05	0.005442	49905.83	5.37	0.52	10.326923	0.000000	...	-0.09	5.46	5.46	0.16	0.16	0.36	0.36	34.433498	-0.66
2018-06-25	15.03	3.06	-0.03	-0.06	-0.001992	49905.83	5.37	0.52	10.326923	1.691475	...	-0.09	5.46	5.46	0.16	0.16	0.36	0.36	39.479580	0.41
2018-06-26	14.86	2.36	0.04	-0.06	0.002699	49905.83	5.37	0.52	10.326923	-1.131071	...	-0.09	5.37	5.46	0.00	0.16	0.52	0.36	37.373248	0.24
2018-06-27	14.86	2.36	0.04	-0.06	0.002699	49905.83	5.37	0.52	10.326923	0.000000	...	-0.09	5.37	5.46	0.00	0.16	0.52	0.36	37.373248	0.24
2018-06-28	15.22	2.24	-0.03	-0.08	-0.001967	49905.83	5.37	0.52	10.326923	2.422611	...	0.00	5.37	5.37	0.00	0.00	0.52	0.52	44.628662	0.70

**All columns are not visible*

This data is now used for modeling trainnig and evaluation. The data is split into 3 categories: *training data, validation data, testing data.* The training and validation data is used during the model training process.

The segregation of data is based on time. We have picked data for training till Jan-2024. The validation data is Feb-2024 – May-2024. The tesitng data is Jun-2024 onwards.