Copper Price Prediction Model- Q4 2023

Dataset:

https://www.investing.com/commodities/copper-historical-data

Data Preprocessing:

1) Data was described and divided into categorical and numerical types including the handling of date-time
2) Data was checked for duplicates and nulls
3) Data was scaled by a scaler according to unit variance
4) Outliers were dropped
5) Inter-quartile range, correlation, z scores were calculated

Feature Selection:

1) A pivot table summarized the highest correlated features to price
2) A scatter plot was drawn between each highly correlated variable and price
3) All variables have been changed to become of type int
4) A pair plot was drawn to summarize all features

Modelling and Evaluation:

1) LSTM, which is a time series model was used since prices are observed as time passes
2) Data was first scaled by a scaler that employed all columns. This is because a scaler that functions according to the unit variance yielded continuous results, so it didn't function well with the model
3) A random seed was set for reproducibility purposes
4) A look_back was set to allow the reference to a previous value in the dataset during each epoch
5) The training data size was set to 70% of the dataset
6) A custom dataset was created which returned a numpy array of all the dataset rows and look_back at previous row
7) The arrays were reshaped into the form [rows, time unit, columns], the rows represent the samples, the time unit how much time is set and the columns the features
8) Run the model's 100 epochs with a learning rate of 0.05, 1 input at a time and 4 hidden layers
9) Predict both the train and the test sets
10) Find the root mean square error in the train and the test

Conclusion:

1) Manual null handling doesn't work since it relies on the mode of the data field. An iterative imputer was better used since it relies on all the data fields to fill in the NaN values
2) The most frequent price is approximately at 3.8
3) A maximum of 2 outliers were found and they were dropped altogether, which means they lied in the same row or sample
4) Price, Open, High, Low are highly correlated with a range of 0.8-1.0
5) None of the features High, Low, Open shows a different overall trend than the other
6) The highly correlated features are positively and almost strongly correlated with Price
7) Without a random seed, a learning rate of 0.05 or a look_back of 1, the model may easily become overfit
8) Overall, the factors driving Price are High, Low and Open