Ocean Protocol

ETH Price Prediction #3

Predict the price of ETH over the course of the next 12 hours from Monday Feb 20^{th}, 2023

Problem Statement

Data Source

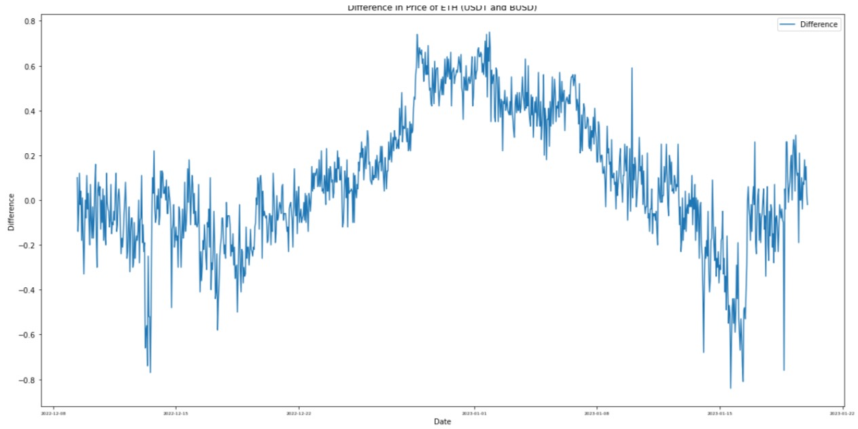
Link : https://cexa.oceanprotocol.io/ohlc?exchange=binance&pair=ETH/BUSD&period=1h

* Output : OHLC data

- Extracts the most recent 1000 rows of hourly data on Binance

Visualisation

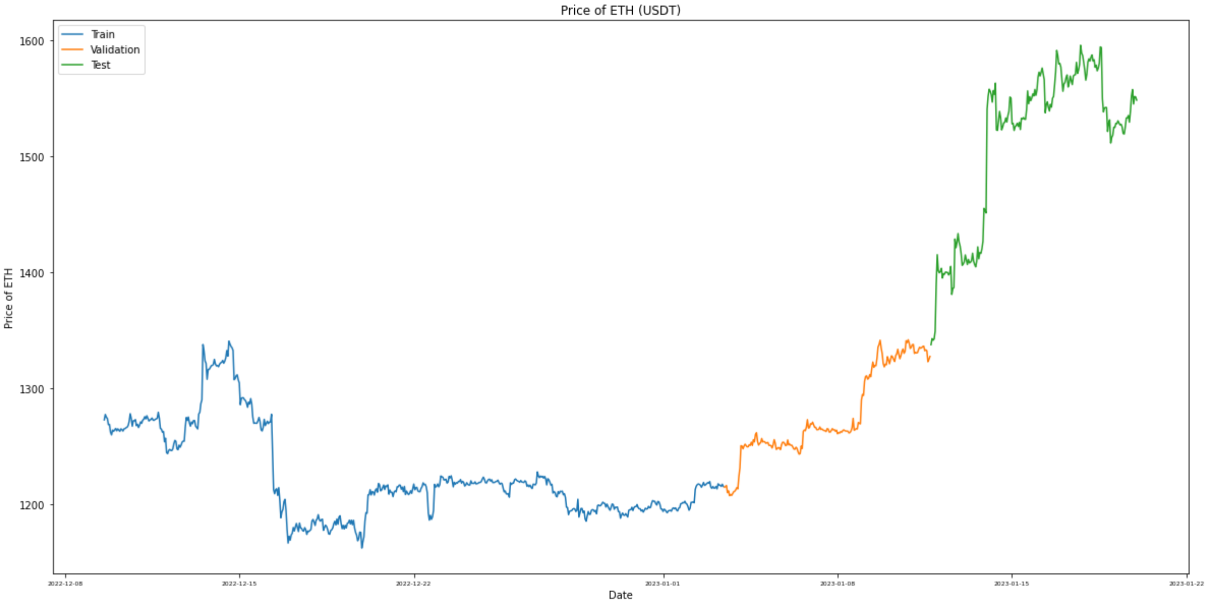
To ensure minimal discrepancy in the asset pair, base assets USDT and BUSD were used as a comparison



Minimal differences were observed and it can be assumed that USDT and BUSD essentially trade at parity of 1 : 1.

USDT as the base asset was chosen given the higher daily volumes recorded

Train-Test Split



Dataset Split

- 60% train set
- 20% validation set
- 20% test set

EDA & Data Preprocessing

Selected Features	•Close Price feature has a lag time of n days, where n is [1,5], to address the look ahead bias totalling up to 5 selected features E.g. “Date_lag_1”, “Date_lag_2”
Moving Average	Simple Moving Average implemented with initial value of window = 5 to get the mean and standard deviation of ETH
Missing Data	KNN Imputation implemented with initial value of n_neighbours = 5 given the volatility of ETH
Scaling	Scaling using a MinMaxScaler() was done to fit the data for machine learning

EDA & Data Preprocessing

	lag_1	lag_2	lag_3	lag_4	lag_5	Close_Price_mean	Close_Price_std
0	0.631121	0.692938	0.702923	0.686487	0.687104	0.662711	0.130041
1	0.619510	0.631121	0.692938	0.702923	0.686487	0.616320	0.149889
2	0.645538	0.619510	0.631121	0.692938	0.702923	0.630300	0.088751
3	0.632692	0.645538	0.619510	0.631121	0.692938	0.630360	0.058034
4	0.625961	0.632692	0.645538	0.619510	0.631121	0.628582	0.047294
...	...	...	...	...	...	...	...
592	0.310428	0.286195	0.297695	0.290010	0.289561	0.267493	0.039198
593	0.304482	0.310428	0.286195	0.297695	0.290010	0.270698	0.040820
594	0.303416	0.304482	0.310428	0.286195	0.297695	0.273579	0.036053
595	0.297919	0.303416	0.304482	0.310428	0.286195	0.273627	0.035960
596	0.308296	0.297919	0.303416	0.304482	0.310428	0.278375	0.011378

Example

X_train_scaled

XGBoost Model

- Decision tree based ensemble model
- Adopts gradient boosting framework

** Feasible model since it adapts quickly to evolving conditions, especially since ETH price is volatile.

Regularisation

Penalise more complex models via L1 & L2

Shrinkage Estimates

Scales newly added weights after each step through column sub-sampling

XGBoost Model – Hyperparameter Tuning

Iteration 1 : Default Parameters

Default values of each variables and the objective function as ‘reg:squared error’

Iteration 2 : n_estimators, max_depth

•n_estimators : Range of 10 to 100, with a step of 5
•max_depth : Range of 1 to 10, with a step of 1

Iteration 3 : learning_rate, min_child_weight

•learning_rate : Range of 0.0001 to 1 with a step of 0.0005
•min_child_weight : Range of 1 to 21 with a step of 1

Iteration 4 : subsample, gamma

•subsample : range of 0.1 to 1 with a step of 0.1
•gamma : range of 0.1 to 1 with a step of 0.1

Iteration 5 : colsample_bytree, colsample_bylevel

•colsample_bytree : range of 0.5 to 1 with a step of 0.1
•colsample_bylevel : range of 0.5 to 1 with a step of 0.1

Evaluation Metrics

1.Mean Absolute Percentage Error (MAPE)

- Quantifies error in terms of percentage
- Easier to interpret and understand
- High percentage → high presence of error
- Cannot detect error of zeros and extreme values

2.Root Mean Square Error (RMSE)

- Measures standard deviation between predicted and actual values
- High value → poor performance

3.Normalized Mean Square Error (NMSE)

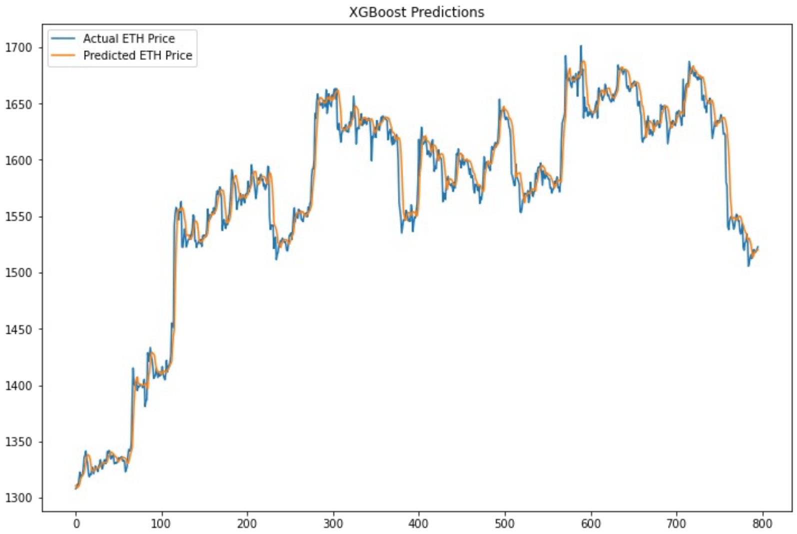
- Measures conformance or nonconformance in proficiency testing where the uncertainty in the measurement result is include
- High value → poor performance

Summary Statistics

	param	original	after_tuning
0	n_estimators	100.000000	10.000000
1	max_depth	3.000000	1.000000
2	learning_rate	0.100000	0.010000
3	min_child_weight	1.000000	1.000000
4	subsample	1.000000	0.500000
5	colsample_bytree	1.000000	0.500000
6	colsample_bylevel	1.000000	0.500000
7	gamma	0.000000	0.010000
8	rmse	14.931330	13.956443
9	mape	0.005856	0.005427

For each parameter, the optimal number would correspond to the lowest RMSE recorded

In Sample Analysis

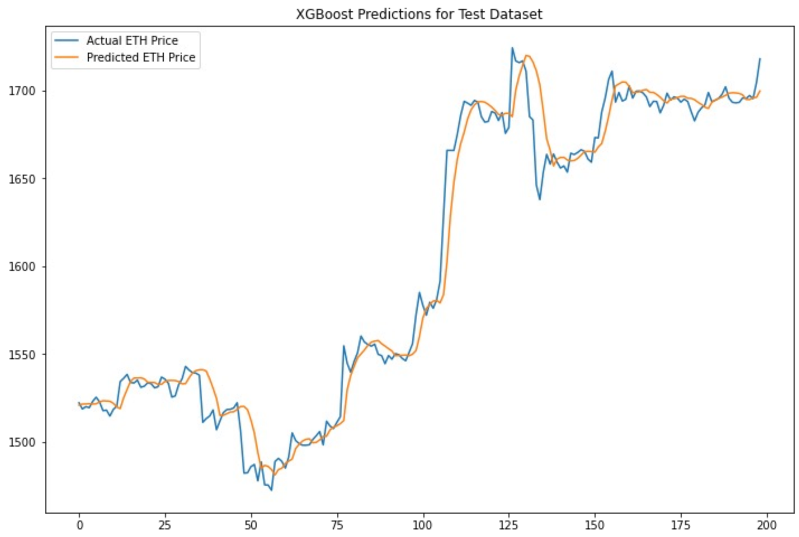


In sample analysis was performed where the model fitted the scaled versions of the training and validation datasets.

This was used to predict the same dataset, yielding the following results:

- MAPE = 0.591
- RMSE = 14.55
- NMSE = 8.59e^-05

Out Sample Analysis

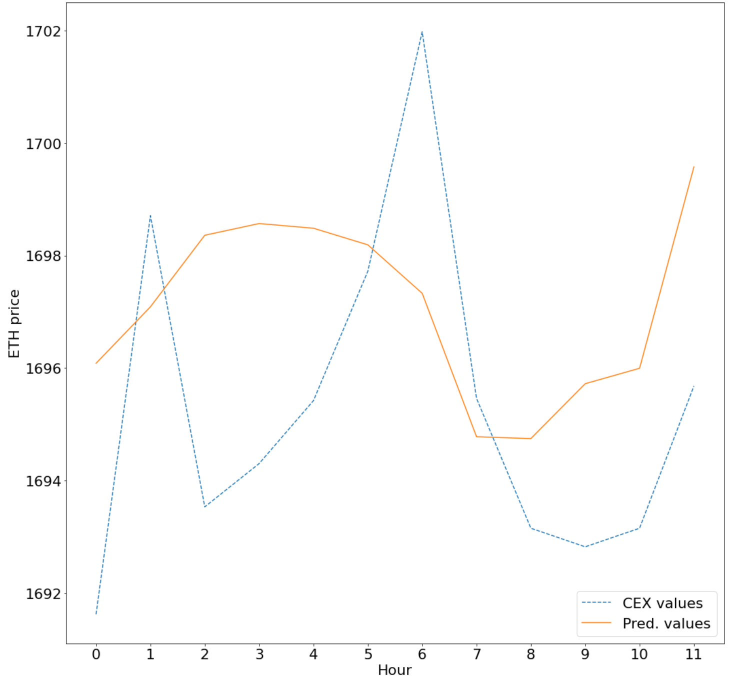


Out sample analysis was performed where the model fitted the scaled versions of the training and validation datasets.

This was used to predict the scaled version of the test dataset, yielding the following results:

- MAPE = 0.523
- RMSE = 13.97
- NMSE = 7.59e^-05

Prediction Method



Using the second last 12 rows of the scaled test set as the feature for prediction, alongside the rolling averages of the mean and standard deviation,

This was used to predict the next 12 hours. In this case, this was compared against the actual values once again

Limitations & Looking Ahead

- Limited dataset of only 1000 datapoints. More data and features such as longer time horizon, daily transaction activity, news sentiment etc can be added to improve the learning and account for potential volatilities. This is because XGBoost is sensitive to outliers and the classifiers are forced to fix previous errors. As a result, this may lead to huge deviations from the final model, incurring bigger estimation losses.

- Incorporate other machine learnings to layer and improve the learning (eg. LSTM)

Nonetheless, model could predict the short term movement of the ETH market with slightly limited predictability since MAPE approaches 1% for the prediction analysis.