Predict ETH

Round 3

ocean

Desights

Data Collection

Binance & Freqtrade

- The ETH/USDT data from Binance (as the results were evaluated on it), along with the correlated pairs of BTC/USDT and ETH/BTC, was collected for the challenge using **freqtrade** as the **data source**, providing market information from as far back as 2017.

```
➜ freqtrade download-data --exchange binance --pairs-file user_data/data/binance/pairs.json -t 1h --timerange 20170817-
```

- The use of **correlated pairs**, such as BTC/USDT and ETH/BTC, provides additional market information and helps in creating a more robust model.

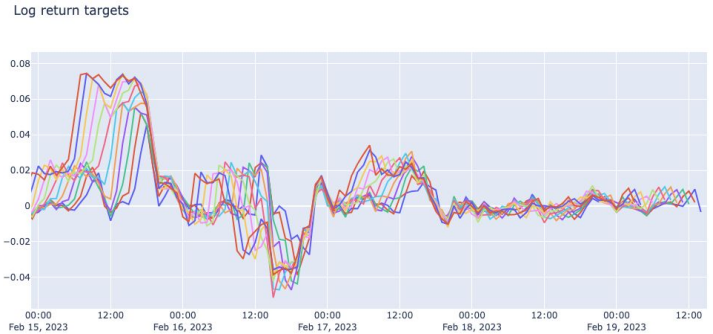
Target creation & selection

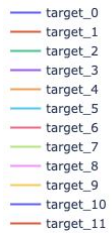
Log return

- The **log return** of ETH was chosen as the **target** for prediction instead of the *simple return* or *raw price*.

- Log returns are preferred in finance as they are **normalized** and provide a **more stable** target, allowing for better model performance compared to simple returns or raw prices. Additionally, log returns can capture both the **magnitude** and **direction** of price changes, which is important in financial forecasting.

- The last data point being on Sunday 19th February 2023 at 23:00:00 p.m, the first target (*target_2*) is here to predict the log return on Monday 20th February 2023 at 01:00:00 a.m (2h later), *target_3* at 02:00:00 a.m and so on (*target_4*, *target_5,*...) until traget_13 to predict the log return at 12:00:00 a.m. Then we convert all log returns to have the 12 prices of ETH.





ocean

Desights

Target creation & selection

Example

date	close	target_2	target_3	target_4	target_5	target_6	target_7	target_8	target_9	target_10	target_11	target_12	target_13
2023-01-31 06:00:00+00:00	1565.96	0.003652	0.003907	0.001002	0.003156	0.004905	0.014064	0.010880	0.014600	0.017365	0.017220	0.017967	0.019841
2023-01-31 07:00:00+00:00	1580.04	-0.005044	-0.007949	-0.005795	-0.004046	0.005113	0.001928	0.005648	0.008414	0.008269	0.009016	0.010890	NaN
2023-01-31 08:00:00+00:00	1571.69	-0.002650	-0.000496	0.001253	0.010412	0.007227	0.010947	0.013712	0.013568	0.014315	0.016188	NaN	NaN
2023-01-31 09:00:00+00:00	1572.09	-0.000751	0.000998	0.010158	0.006973	0.010693	0.013458	0.013314	0.014060	0.015934	NaN	NaN	NaN
2023-01-31 10:00:00+00:00	1567.53	0.003903	0.013062	0.009878	0.013597	0.016363	0.016218	0.016965	0.018839	NaN	NaN	NaN	NaN
2023-01-31 11:00:00+00:00	1570.91	0.010908	0.007724	0.011444	0.014209	0.014064	0.014811	0.016685	NaN	NaN	NaN	NaN	NaN
2023-01-31 12:00:00+00:00	1573.66	0.005975	0.009694	0.012460	0.012315	0.013062	0.014936	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 13:00:00+00:00	1588.14	0.000535	0.003300	0.003156	0.003903	0.005776	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 14:00:00+00:00	1583.09	0.006485	0.006341	0.007087	0.008961	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 15:00:00+00:00	1588.99	0.002621	0.003368	0.005241	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 16:00:00+00:00	1593.39	0.000602	0.002476	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 17:00:00+00:00	1593.16	0.002620	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 18:00:00+00:00	1594.35	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2023-01-31 19:00:00+00:00	1597.34	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

$$log\_return\_shift_i = \ln(\frac{price_t}{price_{t-i}}).shift(-i)$$

$$log\_return\_shift_{10} = \ln(\frac{1597.34}{1572.09}).shift(-10) = 0.015934$$

$$price_t = e^{(\log\_return\_shift_i + \ln(price_{t-i}))}$$

$$price_t = e^{(0.015934 + 1572.09))} = 1597.34$$

ocean

Desights

Model Selection

Gradient Boosting

- Gradient Boosting is a powerful machine learning algorithm for predicting **continuous target** variables.
- It is well-suited for time-series forecasting, capable of handling **complex non-linear relationships** and large datasets.
- Hyperparameter tuning and model customization are possible, providing **flexibility in fine-tuning** the model.
- A new approach was taken, different from traditional time-series models such as ARIMA and Prophet, in order to explore new solutions to the challenge.
- Gradient Boosting (LightGBM and XGBoost) may be a better choice in complex situations that require high accuracy and fine-tuning capabilities.

Features

TA-Lib

To improve the performance and robustness of our Gradient Boosting model, the following features were incorporated (from TA-Lib) at different time period *[2, 3, 5, 8, 13, 21, 23]*:

- Momentum Indicators *(measure the speed and rate of change in asset prices)*: Exponential Moving Average (EMA), Simple Moving Average (SMA), Relative Strength Index (RSI), and Average True Range (ATR).
- Time Indicators *(capture cyclical patterns)*: Time-related features, including hour, day of the week, day of the month, and week of the month, and encoded them using sin and cos functions.
- Volatility Indicators *(capture the risk and uncertainty associated with financial assets)*: Volatility-based features, such as annualized standard deviation of log returns and Williams %R indicator.
- Volume Indicators *(the level of buying or selling activity and the strength of trends)*: Features related to trading volume, such as On-Balance Volume (OBV), to capture.

The combination of these features helped to improve the accuracy and stability of our model.

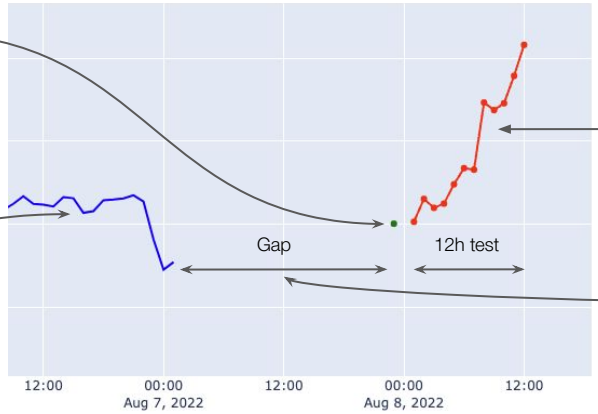
Cross Validation

Custom approach

7 folds that take up the **challenge's evaluation conditions**

Last data point (**Sunday 23:00:00 p.m**) on which we will make 12 predictions

Each training period is 3 months



To best simulate the conditions of the challenge, each test period **begins on a Monday at 01:00:00 a.m.**

The gap allows for temporal separation between the training data and the test data, which helps to reduce the risk of overfitting due to leaks

Here we put the price to better understand the process but the real target is log return

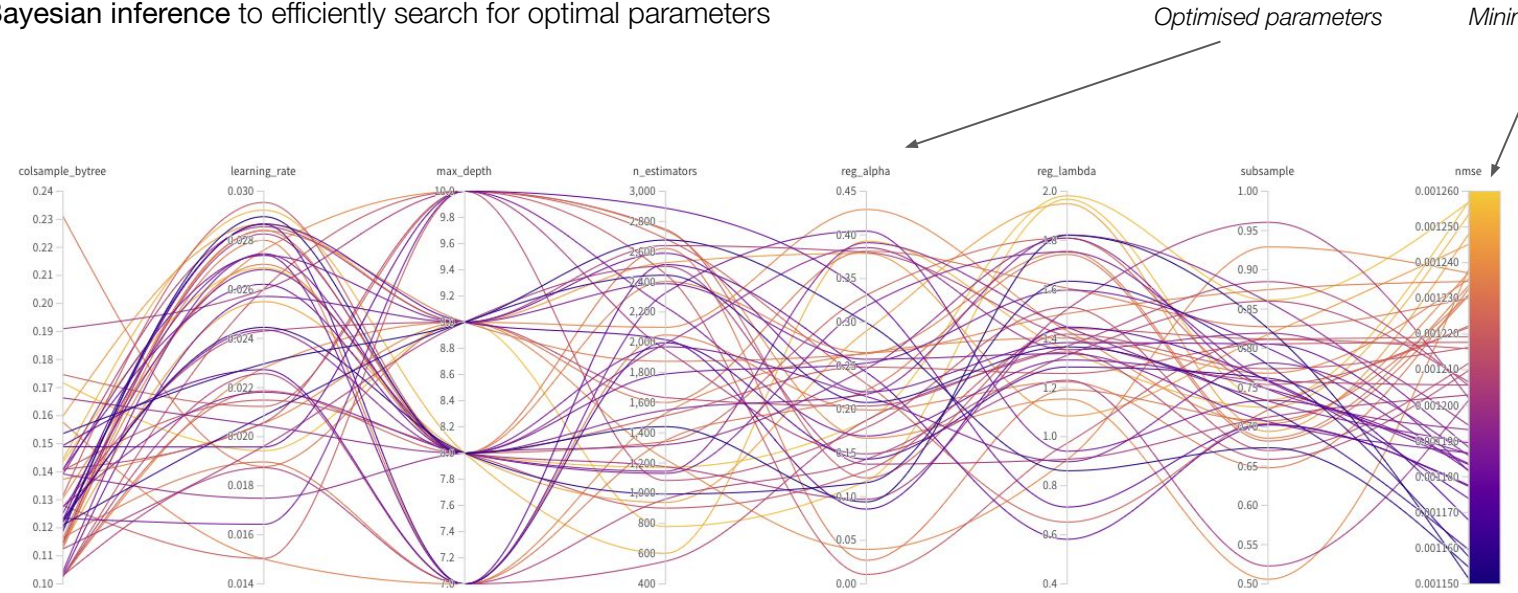
ocean

Desights

Hyperparameters Tuning

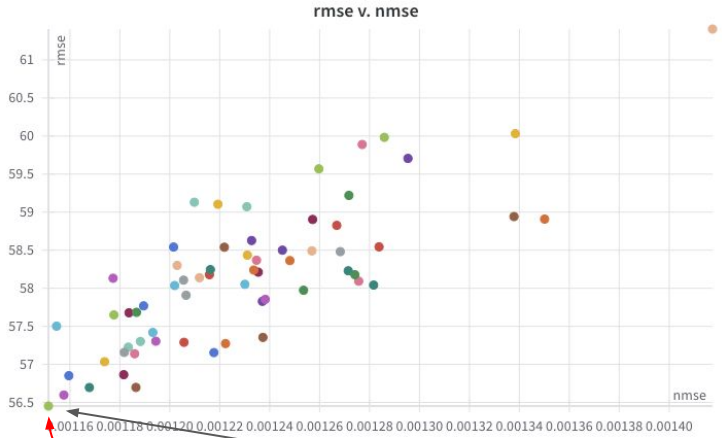
Sweep with Weight & Biases (Bayesian Optimization) on folds

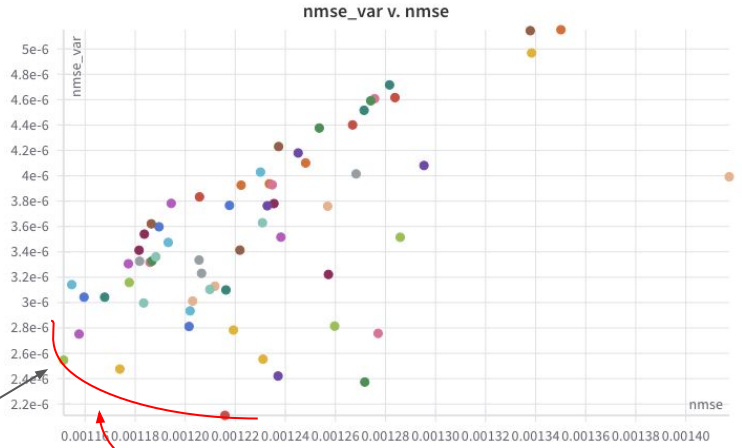
- Use **Weights & Biases Sweeps** for automated hyperparameter search and model exploration
- Sweeps offers visualization-rich, interactive experiment tracking
- **Bayesian inference** to efficiently search for optimal parameters



Hyperparameters Tuning

Hyperparameters selection





Best model

Best solution for the minimization of the two variables (RMSE + NMSE: Mean of all Folds)

Pareto Front minimization of the two variables (NMSE variance + NMSE : Mean of all Folds)

ocean

Desights

Model evaluation

Performance analysis with best params from Cross Validation (NMSE)

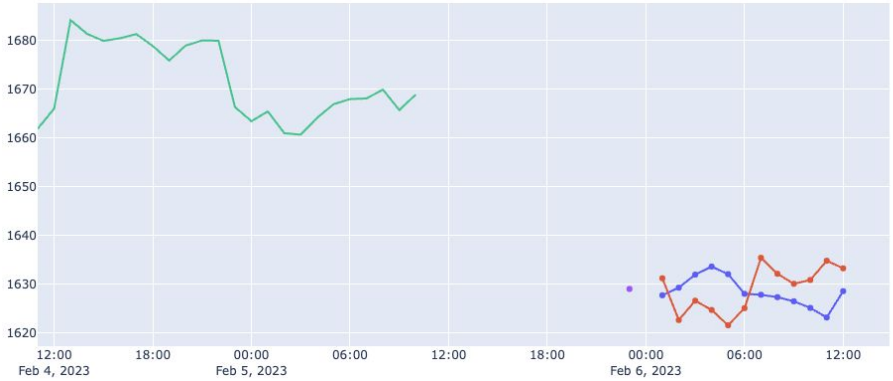
		LightGBM	XGBoost	Ensemble (XGBoost + LightGBM)
	Mean 7 folds	**0.001151**	0.001186	-
18 months of training	6 February 2023 *(01:00:00a.m to 12:00:00a.m)*	2.0683e-05	1.9053e-05	**1.7777e-05**
	13 February 2023 *(01:00:00a.m to 12:00:00a.m)*	0.0001902	0.0001877	**0.0001875**

ocean

Desights

Model evaluation

Performance analysis representation



ocean

Desights

Future Works

- Include other indicators with a **feature selection** approach (LASSO…)
- **Ensemble Methods**: Explore ensemble methods with a mix of Gradient Boosting, Prophet, and SARIMAX to improve the accuracy and robustness of the model.
- **Market Sentiment Analysis**: Utilize market sentiment indicators such as the Fear & Greed index to capture the psychological aspects of the market and improve prediction accuracy.

Submission

Ocean Library & Recommendations

1) Have the possibility to predict at **any time** for the next 12 hours with still a deadline per round. We can imagine having a limit of 3 predictions for the current round and we average these 3 predictions. I see 2 advantages to this improvement:

   a) Allow people from all over the world to predict and not wait until late at night to get the last data point (equitable)

   b) Avoid potential **network congestion** if everyone submits at the same time

2) Reduce the number of steps to share / publish a file on the Ocean Market or if you want to share it to an address. Especially when you want to share algorithmically. Maybe create a class with an **Sender** and **Receiver address** and a **file** ?

3) Allow directly on the Desights platform to publish a file, a report