ML & Fintech
Fall, 2023

Cryptocurrency Trend Prediction

Future Trends Based on Data Analysis and Machine Learning

312554011 謝翊庭

Report Date: 12/22

Motivation

Upon discovering relevant competitions on Kaggle related to this topic, I was motivated to undertake this project as a practical application.

Problem Description

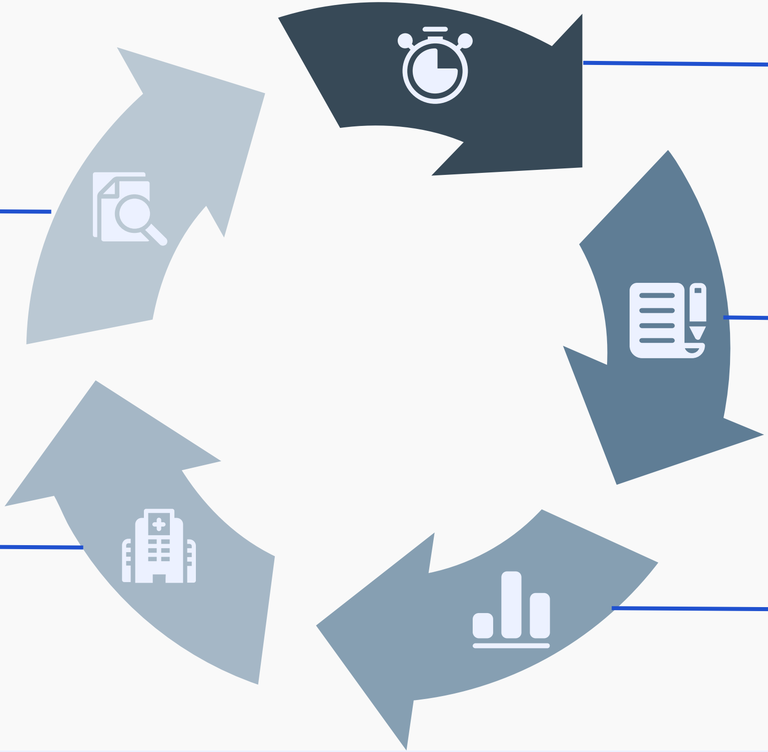
- Input:

  The dataset includes time, virtual currency ID, financial features such as opening price, closing price, highest price, lowest price, and sales volume in the financial context.

- Output:

  Calculate the target values for each virtual currency over a specific future period.

Data Description



Train:Test

7:3

Download source

Kaggle

Number of fields

10

Original source

G-research

Number of data

530k

Data Description (Cont.)

Name	Description
Timestamp	A timestamp for the minute covered by the row
Asset_ID	An ID code for the cryptoasset
Count	The number of trades that took place this minute
Open	The USD price at the beginning of the minute
Close	The USD price at he end of the minute

Data Description (Cont.)

Name	Description
Low	The lowest USD price during the minute
High	The highest USD price during the minute
Volume	The number of cryptoasset units traded during the minute
VWAP	The volume weighted average price for the minute
Target	15 minutes residualized returns

Analysis Workflow

Data Analysis & Processing

The raw data has been processed relatively cleanly. Although there are some missing values, we can start by imputing the data and then proceed with attribute selection and transformation.

Feature Enigeering

1. Analyze the relationships between different currencies through correlation
2. Add various features based on the findings.

Modelling

Attempt to use the LGBM model for prediction.

Evaluation

Due to uncertainty about the correct scoring method on Kaggle, we adopt the rMSE approach to assess the magnitude of errors.

Knowledge

Through the results from the model, the goal is to identify meaningful patterns.

Data Analysis & Processing | Feature Enigeering | Modelling | Evaluation | Knowledge

Data description

Display the basic indicators for each feature.

	timestamp	Asset_ID	Count	Open	High	Low	Close	Volume	VWAP	Target
count	5300895.000	5300895.000	5300895.000	5300895.000	5300895.000	5300895.000	5300895.000	5300895.000	5300895.000	5299020.000
mean	1620791296.830	6.500	788.235	3649.352	3654.925	3643.968	3649.356	786662.153	3649.326	0.000
std	6559688.080	4.031	1619.291	11652.151	11668.836	11635.865	11652.164	4905453.219	11652.060	0.006
min	1609430400.000	0.000	1.000	0.005	0.005	0.005	0.005	0.000	0.005	-0.253
25%	1615110660.000	3.000	104.000	0.881	0.893	0.867	0.881	240.668	0.881	-0.002
50%	1620790380.000	6.000	284.000	66.781	66.921	66.647	66.777	2714.060	66.777	-0.000
75%	1626472320.000	10.000	849.000	596.934	597.879	596.020	596.954	165715.583	596.952	0.002
max	1632153600.000	13.000	165016.000	64805.944	64900.000	64670.530	64808.537	759755403.142	64799.822	0.305

	Asset_ID	Weight	Asset_Name
1	0	4.304065	Binance Coin
2	1	6.779922	Bitcoin
0	2	2.397895	Bitcoin Cash
10	3	4.406719	Cardano
13	4	3.555348	Dogecoin
3	5	1.386294	EOS.IO
5	6	5.894403	Ethereum
4	7	2.079442	Ethereum Classic
11	8	1.098612	IOTA
6	9	2.397895	Litecoin
12	10	1.098612	Maker
7	11	1.609438	Monero
9	12	2.079442	Stellar
8	13	1.791759	TRON

Data Analysis & Processing | Feature Enigeering | Modelling | Evaluation | Knowledge

Figure Display

Use candlestick graphs to represent the time series of closing prices for different currencies.

Data Analysis & Processing

Feature Enigeering

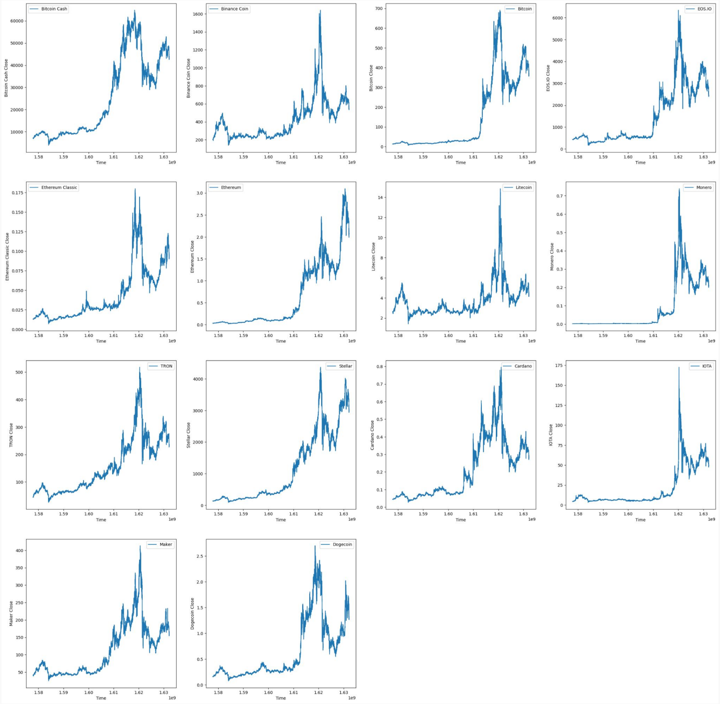
Modelling

Evaluation

Knowledge

Figure Display

Use line charts to represent the time series of closing prices for different currencies.

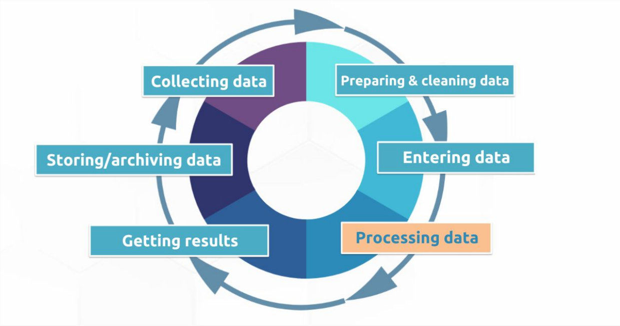


Data Analysis & Processing | Feature Enigeering | Modelling | Evaluation | Knowledge

Data processing

Process the data as listed on the right

- Seperate instances based on asset_ID
- Perform data type transformation
- Fill in missing time intervals
- Impute missing values



Data Analysis & Processing

Feature Enigeering

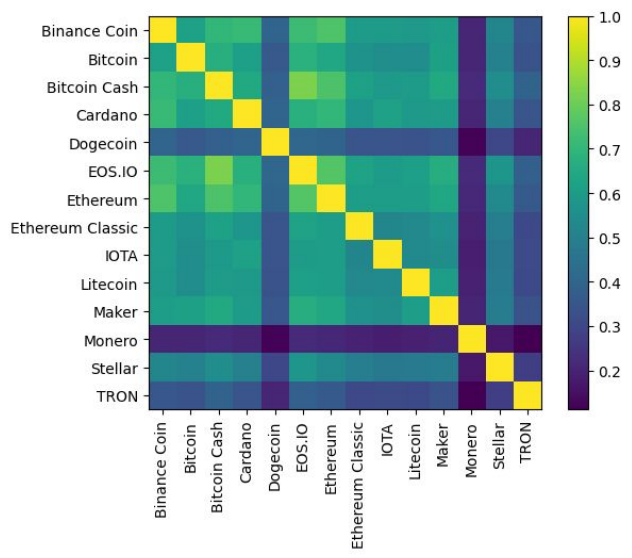
Modelling

Evaluation

Knowledge

Correlation

Using the Pearson correlation coefficient, identify the relationships between different currencies

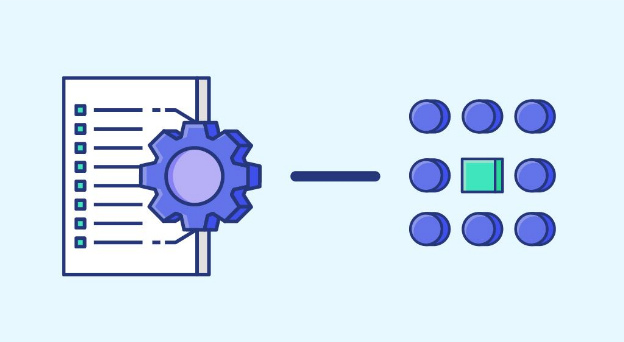


Data Analysis & Processing | **Feature Enigeering** | Modelling | Evaluation | Knowledge

Add some new feature

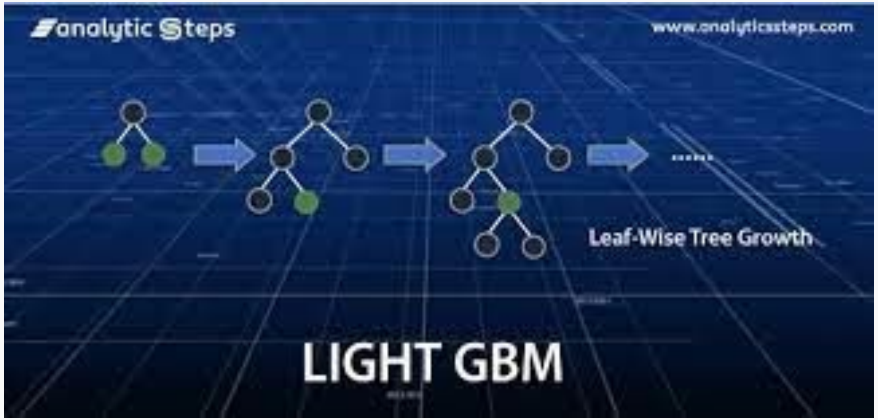
Add some additional features to enhance the model's utility.

- Price change or price decrease
- Rate of price change
- Difference between the highest and lowest prices
- Ratio of highest price and the mean and so on



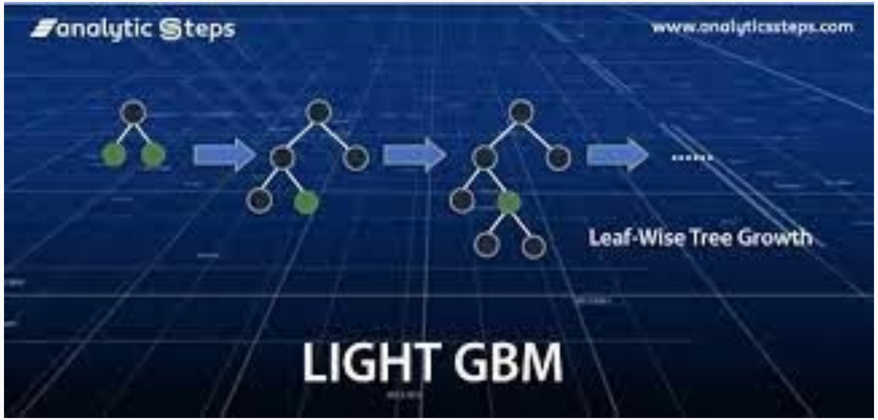
Data Analysis & Processing | Feature Enigeering | Modelling | Evaluation | Knowledge

- LGBM (lightgbm)
  - Efficient training speed
  - Low Memory Consumption
  - Powerful Fitting Ability
  - Convenient Parameter Tuning



Data Analysis & Processing → Feature Enigeering → Modelling → Evaluation → Knowledge

- initial parameters
  - n_estimators=1500
  - num_leaves=500
  - objective="regression"
  - metric="rmse"
  - boosting_type="gbdt"
  - learning_rate=0.05
  - random_state=1221
  - force_col_wise=True

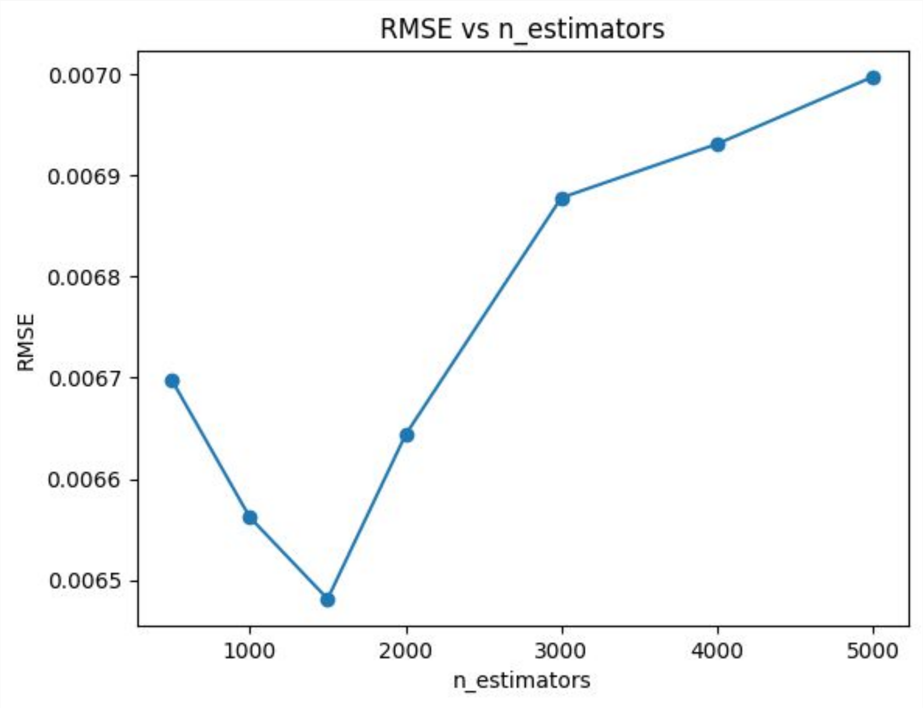


Data Analysis & Processing → Feature Enigeering → Modelling → Evaluation → **Knowledge**

Hand over the model results to professionals for further analysis, aiming to identify specialized or deeper patterns in the data.

Result Analysis

- Using grid search on different n_estimators



Conclusion

- A certain degree of correlation between certain cryptocurrencies
- Through grid search, it was found that the optimal number of n_estimators for the model is around 1500.

Improvement

- Incorporate the potential correlations between pairs of currencies as features into the training considerations
- Adjust other parameters
- Try using other models, such as LSTM

ML & Fintech
Fall, 2023

Thank you for your attention!

312554011 謝翊庭