Project 3

math6010z_
Li_Liu_Nie_Wu

Kaggle: G-Research Crypto Forecasting

Li Chenghai 20828446
Liu Shifei 20734530
Nie Jialei 20747874
Wu Jiajun 20666111

CONTENTS

01 Introduction

Over $40 billion worth of cryptocurrencies are traded every day. They are among the most popular assets for speculation and investment, but they have proven wildly volatile. Fast-fluctuating prices carry huge risks, so we try to predict price movements in advance.

In this competition, we used machine learning to forecast short term returns in 14 popular cryptocurrencies, such as Bitcoin, Binance Coin, Monero and so on.

02 Data Description

There is a dataset of millions of rows of high-frequency market data dating back to 2018 which we used to build the model. In the whole calculation process, we used three data sets, namely train.csv, example_test.csv and asset_details.csv.

	timestamp	Asset_ID	Count	Open	High	Low	Close	Volume	VWAP	Target
0	1514764860	2	40.0	2376.580000	2399.500000	2357.140000	2374.590000	1.923301e+01	2373.116392	-0.004218
1	1514764860	0	5.0	8.530000	8.530000	8.530000	8.530000	7.838000e+01	8.530000	-0.014399
2	1514764860	1	229.0	13835.194000	14013.800000	13666.110000	13850.176000	3.155006e+01	13827.062093	-0.014643
3	1514764860	5	32.0	7.659600	7.659600	7.656700	7.657600	6.626713e+03	7.657713	-0.013922
4	1514764860	7	5.0	25.920000	25.920000	25.874000	25.877000	1.210873e+02	25.891363	-0.008264
...	...	...	...	...	...	...	...	...	...	...
24236801	1632182400	9	775.0	157.181571	157.250000	156.700000	156.943857	4.663725e+03	156.994319	NaN
24236802	1632182400	10	34.0	2437.065067	2438.000000	2430.226900	2432.907467	3.975460e+00	2434.818747	NaN
24236803	1632182400	13	380.0	0.091390	0.091527	0.091260	0.091349	2.193732e+06	0.091388	NaN
24236804	1632182400	12	177.0	0.282168	0.282438	0.281842	0.282051	1.828508e+05	0.282134	NaN
24236805	1632182400	11	48.0	232.695000	232.800000	232.240000	232.275000	1.035123e+02	232.569697	NaN

The training set contains information such as timestamp, open price, close price, high price, low price and Target.

02 Data Description

The example testing set is an example of the data that will be delivered by the time series API.

	timestamp	Asset_ID	Count	Open	High	Low	Close	Volume	VWAP	group_num	row_id
0	1623542400	3	1201.0	1.478556	1.486030	1.478000	1.483681	6.547996e+05	1.481439	0	0
1	1623542400	2	1020.0	580.306667	583.890000	579.910000	582.276667	1.227988e+03	581.697038	0	1
2	1623542400	0	626.0	343.789500	345.108000	343.640000	344.598000	1.718833e+03	344.441729	0	2
3	1623542400	1	2888.0	35554.289632	35652.464650	35502.670000	35602.004286	1.638115e+02	35583.469303	0	3
4	1623542400	4	433.0	0.312167	0.312600	0.311920	0.312208	5.855774e+05	0.312154	0	4
5	1623542400	5	359.0	4.832550	4.845900	4.822900	4.837583	4.714355e+04	4.836607	0	5
6	1623542400	7	541.0	55.223080	55.494000	55.182000	55.344680	6.625202e+03	55.298816	0	6
7	1623542400	6	2186.0	2371.194286	2379.200000	2369.670000	2374.380714	1.214129e+03	2374.335307	0	7
8	1623542400	8	35.0	1.003150	1.019800	0.987300	1.003300	7.061928e+03	1.002936	0	8
9	1623542400	9	560.0	161.933429	162.480000	161.730000	162.214714	1.485009e+03	162.231310	0	9
10	1623542400	10	61.0	2939.862750	2952.160000	2936.230000	2947.078025	9.584785e+00	2945.110614	0	10

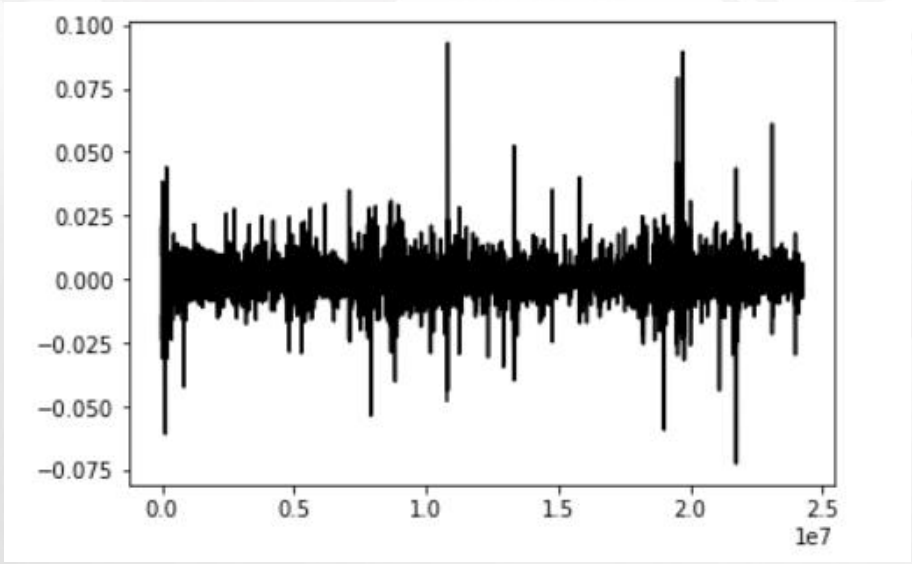
02 Data Description

The asset details set provides the real name of the cryptoasset for each Asset_ID and the weight each cryptoasset receives in the metric. This weight represents our investment ratio in each cryptocurrencies when we invest in cryptocurrencies.

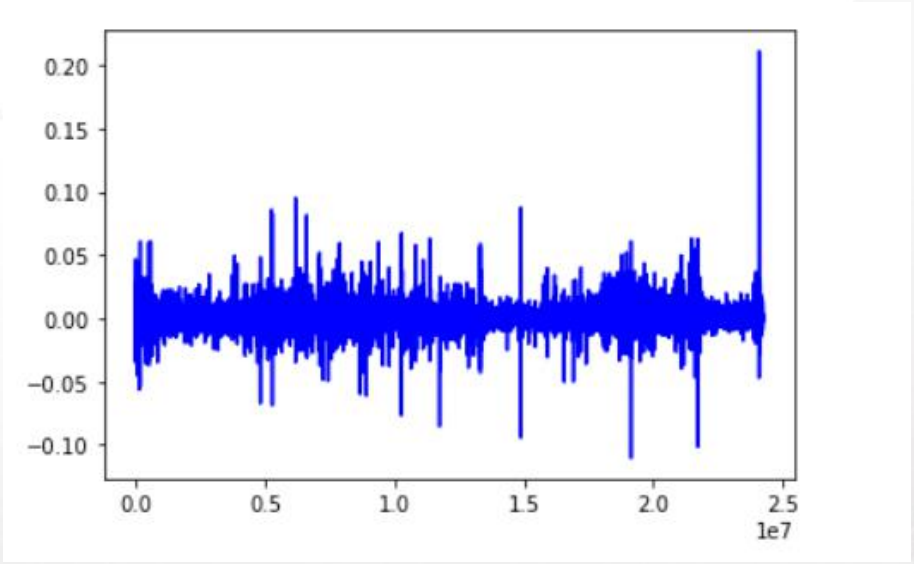
	Asset_ID	Weight	Asset_Name
1	0	4.304065	Binance Coin
2	1	6.779922	Bitcoin
0	2	2.397895	Bitcoin Cash
10	3	4.406719	Cardano
13	4	3.555348	Dogecoin
3	5	1.386294	EOS.IO
5	6	5.894403	Ethereum
4	7	2.079442	Ethereum Classic
11	8	1.098612	IOTA
6	9	2.397895	Litecoin
12	10	1.098612	Maker
7	11	1.609438	Monero
9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

02 Data Description

Bitcoin Return



Litecoin Return



Bitcoin's return is in a state of high volatility, with the highest being 0.09 and the lowest being less than -0.07. However, we can find that most of the returns are still concentrated between 0.025 and -0.025.

Litecoin's return is also in a state of high volatility, with the highest being higher than 0.2 and the lowest being less than -0.1. Similarly, we can find that most of the returns are still concentrated between 0.05 and -0.05.

03 Methodology

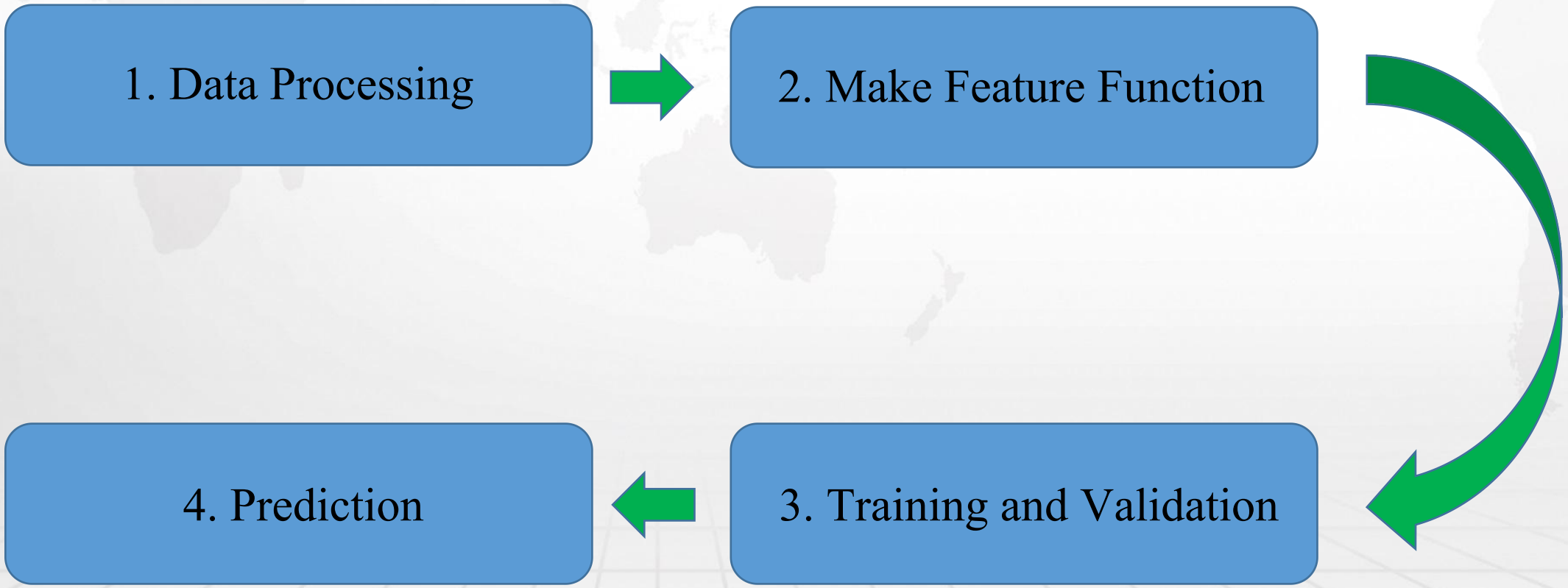
GBDT (Gradient Boosting Decision Tree)

Use weak classifiers (decision trees) to iteratively train to obtain the optimal model
Advantages:
good training effect & not easy to overfit

LightGBM (Light Gradient Boosting Machine)

Framework that implements the GBDT algorithm
Advantages:
Support high-efficiency parallel training
Faster training speed; Lower memory consumption; Better accuracy
Support for distributed
Quickly process massive amounts of data

03 Methodology

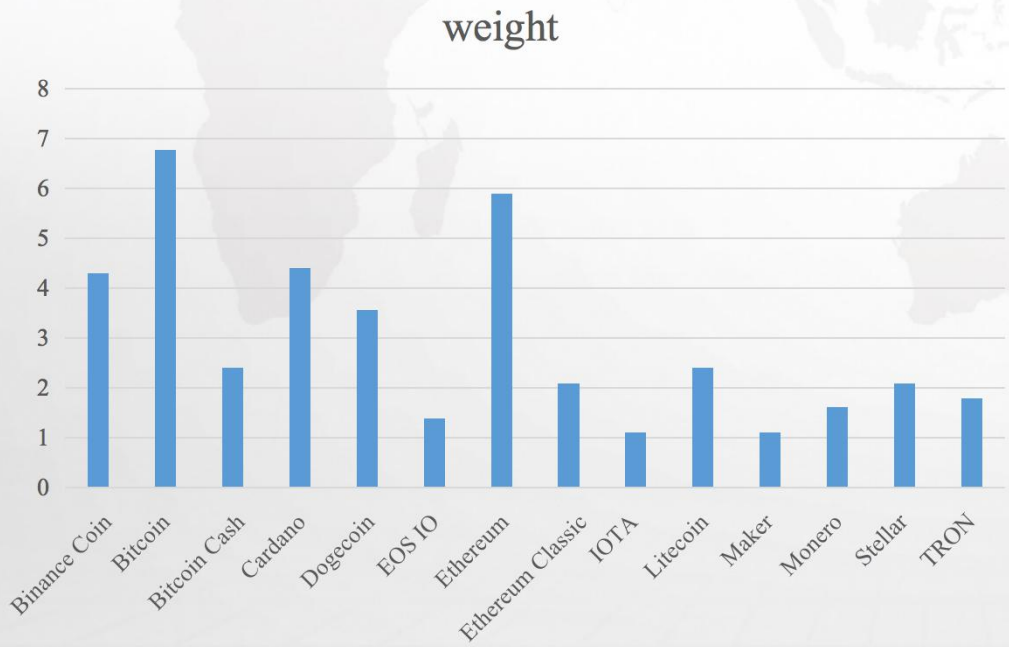


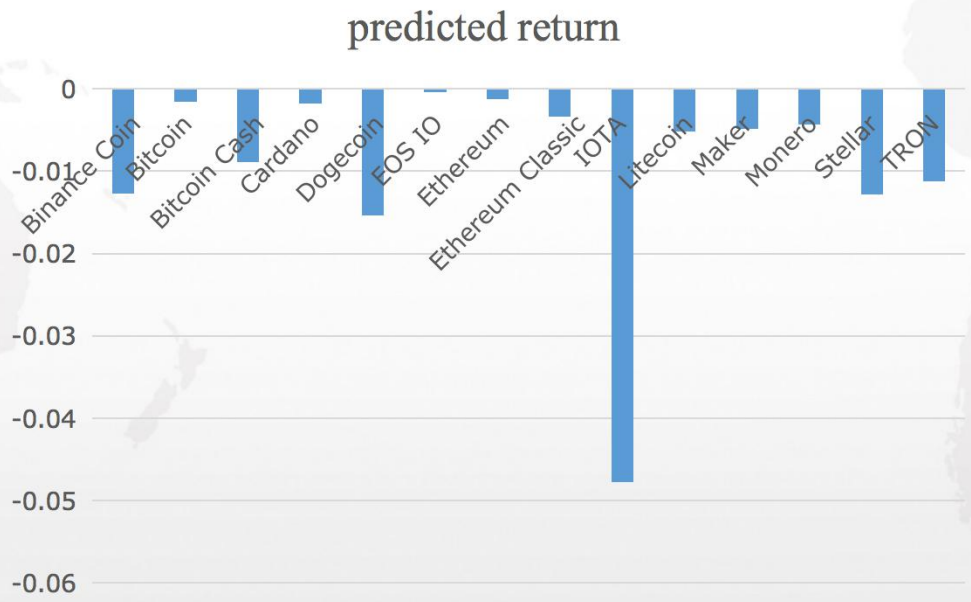
04 Prediction

	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

	row_id	Target
0	42	-0.012735
1	43	-0.001620
2	44	-0.008905
3	45	-0.001754
4	46	-0.015395
5	47	-0.000419
6	48	-0.001264
7	49	-0.003406
8	50	-0.047763
9	51	-0.005205
10	52	-0.004895
11	53	-0.004349
12	54	-0.012878
13	55	-0.011209

04 Prediction





Project 3

math6010z_
Li_Liu_Nie_Wu

Thank You!

Li Chenghai 20828446
Liu Shifei 20734530
Nie Jialei 20747874
Wu Jiajun 20666111

Project 3

math6010z_
Li_Liu_Nie_Wu

Video Link:

https://www.bilibili.com/video/BV1pL411E7rk/

Li Chenghai 20828446
Liu Shifei 20734530
Nie Jialei 20747874
Wu Jiajun 20666111