SI670 Project Proposal

Chongdan Pan
pandapcd@umich.edu

Xin Ye
xinye@umich.edu

Fangzhe Li
fangzhel@umich.edu

November 1, 2022

Abstract

Bitcoin and cryptocurrency have been very hot topics these years due to the rapid growth of the market value. Due to the lack of regulation, the market is very volatile and risky for investors. This project tries to construct a systematic way to predict the volatility of the Bitcoin market through the microstructure of the market. The data source is a high-frequency orderbook, which is a snapshot of all the orders listed on the exchange, and they're collected from Tardis, a major cryptocurrency data collector. Various machine learning models and features will be investigated and their performance will be evaluated by different metrics as well.

1 Introduction

Cryptocurrency such as Bitcoin and Ethereum is a novel digital currencies that have increasingly become a popular topic in the field of economy. There are two reasons for the rapid growth of the cryptocurrency market. First, since cryptocurrency is running in a decentralized, there is no restrictive regulation stopping investors from entering and exiting the market. Such freedom greatly attracts retail investors as well as financial institutions. Second, the cryptocurrency is extremely volatile and it's open 24/7 every year. The volatility and long trading hour imply that there are countless opportunities for traders to make money. However, with high returns, there must be high risk, which should never be ignored by investors. This project, will not only focus on making a profit in the cryptocurrency market but also tries to dig out some meaningful information such as volatility from the market.

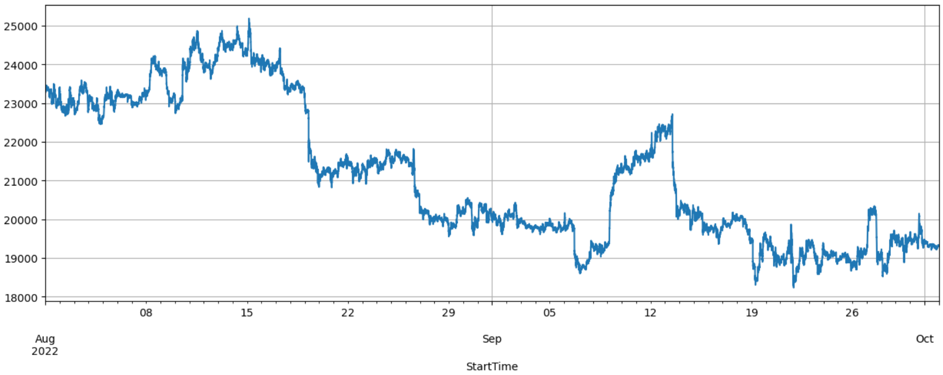
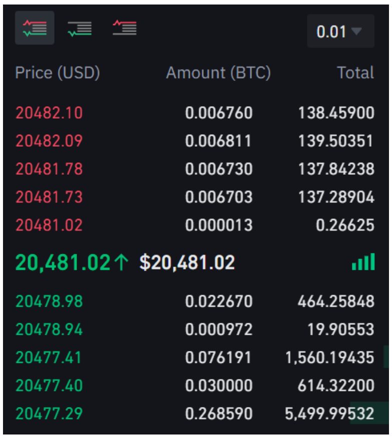


Figure 1: Bitcoin price

Orderbook will be the main source of the features, which is a snapshot of any market that stores all ”buy” or ”sell” orders information such as price and volume. Exchanges are usually made by matching orders from the price of the order book between buyer and seller. The order book data can provide insights into detailed trading information. Orderbook is generated extremely frequently because as long as anyone posts an order at the market, a new row of orderbook will be made. Therefore, the orderbook can be the detailed information source of the price movement. This project will not only use an orderbook but also construct predictors with feature engineering for further exploration of applied machine learning techniques.



Price (USD)	Amount (BTC)	Total
20482.10	0.006760	138.45900
20482.09	0.006811	139.50351
20481.78	0.006730	137.84238
20481.73	0.006703	137.28904
20481.02	0.000013	0.26625
20,481.02↑	$20,481.02	
20478.98	0.022670	464.25848
20478.94	0.000972	19.90553
20477.41	0.076191	1,560.19435
20477.40	0.030000	614.32200
20477.29	0.268590	5,499.99532

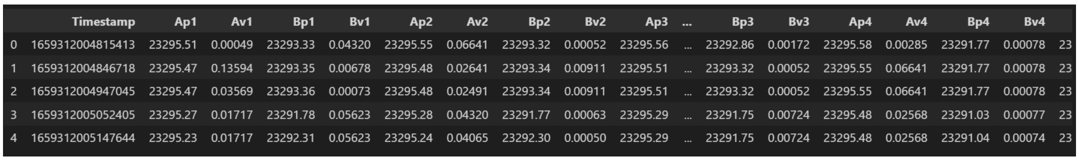
Figure 2: Sample live orderbook of Bitcoin

2 Related Work

Guo.T and Antulov-Fantulin [1] looked into the order book data of Bitcoins and used this to make a short-term prediction of the exchange price fluctuations towards the United States dollar. They derived a generative temporal mixture model and proved that the features of buy and sell orders significantly affect future high volatility. Guo et al. [2] also studied the problem of the Bitcoin short-term volatility forecasting based on the volatility history and order book data and proposed temporal mixture models which could successfully decipher the time-varying effect of order book features on volatility. Rathan et al. [3] studied the prediction of Bitcoin price by feature selection of different machine learning techniques. They transformed order book data into features over time and then used the decision tree and regression model to develop predictions of the Bitcoin price.

3 Datasets

The data source for the project is Tardis, a fine-granularity cryptocurrency data collector, which provides all historical trading information of cryptocurrencies from different exchanges. This project will focus on the data from Binance since it’s the largest cryptocurrency exchange in the world. Since the orderbook contains too many rows, the time range for the project will be passed 3 months, and only the top five ask and bid orders will be used. Bitcoin is chosen as the project’s subject because it’s the most popular cryptocurrency in the world with the largest market value. In summary, the dataset will have the following 21 fields as raw features:



	Timestamp	Ap1	Av1	Bp1	Bv1	Ap2	Av2	Bp2	Bv2	Ap3	...	Bp3	Bv3	Ap4	Av4	Bp4	Bv4	
0	1659312004815413	23295.51	0.00049	23293.33	0.04320	23295.55	0.06641	23293.32	0.00052	23295.56	...	23292.86	0.00172	23295.58	0.00285	23291.77	0.00078	23
1	1659312004846718	23295.47	0.13594	23293.35	0.00678	23295.48	0.02641	23293.34	0.00911	23295.51	...	23293.32	0.00052	23295.55	0.06641	23291.77	0.00078	23
2	1659312004947045	23295.47	0.03569	23293.36	0.00073	23295.48	0.02491	23293.34	0.00911	23295.51	...	23293.32	0.00052	23295.55	0.06641	23291.77	0.00078	23
3	1659312005052405	23295.27	0.01717	23291.78	0.05623	23295.28	0.04320	23291.77	0.00063	23295.29	...	23291.75	0.00724	23295.48	0.02568	23291.03	0.00077	23
4	1659312005147644	23295.23	0.01717	23292.31	0.05623	23295.24	0.04065	23292.30	0.00050	23295.29	...	23291.75	0.00724	23295.48	0.02568	23291.04	0.00074	23

Figure 3: Sample orderbook data of Bitcoin

- Timestamp: timestamp of the snapshot

- Bid Price[level]: The price of the top 5 buy orders with higher bid price.

- Bid Volume[level]: The volume of the top 5 buy orders with higher bid price.

- Ask Price[level]: The price of the top 5 sell orders with lower ask price.

- Ask Volume[level]: The volume of the top 5 sell orders with lower ask price.

4 Methods

First, since the orderbook is not aligned with a specific frequency, they need to be resampled by timestamp so that it can be processed by normal computers. At the same time, more features can be constructed, such as bid-ask spread, mid-price, weight bid and ask price, etc. With enough features and linear regression, decision tree can be models of interest.

5 Evaluation

Given a time range indexed by t, the price return of a given symbol can be defined as $R_t = \frac{P_t}{P_{t-1}} - 1$, then the volatility can be regarded as the standard deviation of the price return within a given time interval. In this project, the volatility of any thirty minutes will be the labels. Since this is regression task, MSE will be considered as a good evaluation metrics.

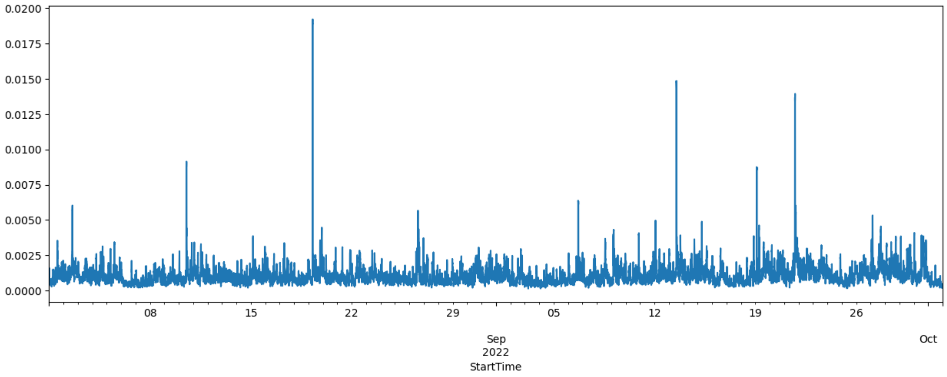


Figure 4: Bitcoin volatility

What's more, Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) can be an important baseline for comparison since it's a financial statistics model specially designed to model volatility. On the other hand, since the ultimate goal of any financial market research is to make a profit, the performance of the predictor can also be tested by constructing a trading strategy. If the trading strategy can achieve a higher return than the price of Bitcoin itself, then it implies that the models may successfully turn the information from the orderbook into profit.

6 Computing resources

Normal computers will be used as the computing resources.

7 Reference

[1] Guo, T., Antulov-Fantulin, N. (2018). Predicting short-term Bitcoin price fluctuations from buy and sell orders. arXiv preprint arXiv:1802.04065.

[2] Guo, T., Bifet, A., Antulov-Fantulin, N. (2018, November). Bitcoin volatility forecasting with a glimpse into buy and sell orders. In 2018 IEEE international conference on data mining (ICDM) (pp. 989-994). IEEE.

[3] Rathan, K., Sai, S. V., Manikanta, T. S. (2019, April). Crypto-currency price prediction using decision tree and regression techniques. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 190-194). IEEE.