Vrije Universiteit Amsterdam

Research Proposal

Limit Order Execution Probability Modeling in Crypto-assets

Authors:

Mark de Kwaasteniet

Supervisors:

Norman J. Seeger

Amsterdam, April 7, 2022

Contents

1 Objective 1

2 Introduction 1

2.1 Research Question 2

3 Literature 2

4 Methodology 4

4.1 Data Structure 4

4.2 Variables: Time Series data 4

4.3 Variables: Cross-Sectional data 6

4.4 Data Scaling 7

4.5 Data Analysis 7

1 Objective

The primary focus of the paper is the evaluation of parametric and non-parametric limit order execution probability models in the Crypto-asset market, where non-parametric models are considered to have no distributional assumptions and a flexible number of parameters. The objective of this research paper is to optimize market making order placement strategies by incorporating execution probabilities when placing limit orders in the market.

Optional[1]: A benchmark bid/ask market making strategy is compared to the optimized execution probability market making strategy to test whether modeling execution probability significantly enhances the returns.

2 Introduction

In both the equity- and the Crypto-asset market there are considered directional and non-directional traders. Where directional traders (e.g., asset managers) are concerned with an up- or down movement of the price, non-directional traders (e.g., market makers) are less dependent on a directional price movement and more dependent on the trade activity or volatility in the market. Both types have multiple options for placing an order (i.e., a bid or ask) into the market. This paper focuses on market orders and limit orders. Where a market order is executed directly at the best bid- or ask price and a limit order is placed in the market for a certain price and executed only if a counterparty in the market is willing to buy/sell at that price. A market order is therefore guaranteed to be executed, providing a probability of 100% that the order is going to be executed within the next time period. For a limit order this execution probability might not be 100% as it is very dependent on multiple factors in the market. Factors like the spread between the limit order and the theoretical price, the current bid-ask spread, the order quantities, past trading volume, etc.

Although a market order will be executed right away, the price at which it is executed is the best bid/ask price in the market, providing a less favorable price for the trader. A limit order provides a more favorable price, but the trader is uncertain when or whether the order will be executed at all. Which might result in opportunity costs for the trader.

[1] Optional is used to indicate that this might not come back in the final version of the thesis as it is uncertain whether it is manageable to process this in the given thesis time period

This means that there is a crucial trade-off between the probability and the price at which an order is executed. By modeling the execution probability of a limit order, the trader will have the ability to calculate the expected value of the limit order placements making the financial trade-off between a market- and limit order more concrete.

2.1 Research Question

Therefore, this paper researches whether or not limit order execution probability can be accurately modeled given the order features available in the order book and transactions details. **(Optional)** In addition, the paper examines whether including an execution probability model in a margin market making strategy can outperform a simple bid/ask market making strategy.

> **RQ:** *Can market making strategies be optimized, in terms of P&L, by modeling limit order execution probabilities?*

The studies aims to answer this question by analyzing the importance of the order book features, valuating and comparing the accuracy of both parametric and non-parametric probability models, **(Optional)** assessing the expected order value when incorporating execution probabilities into a limit order placement strategy.

3 Literature

Concrete literature of market makers modeling execution probability for efficient order placement will be difficult to find as these profit organizations are motivated to keep their research inside the company, preventing competitors from copying their models and strategy. However, there has been extensive research on modeling limit order execution probability for directional traders. The referenced literature for this paper can be divided into the following subject matters:

- **Parametric models for limit order execution probability**
  This concerns papers that use models with certain assumptions regarding the data distribution (parametric models) in order to estimate execution probability. Yingsaeree

(2012) has done an extensive study about execution probability in the commodity and equity market, by reviewing empirical models in previous studies and evaluating proposed models in the current market. Concluding that the probabilities estimated with execution time models (i.e. a Survival Analysis), which indirectly compute the execution probability by studying the time it takes to execute an order, have the lowest standard errors in a short time horizon. The study also concludes that market factors, like bid-ask spread and past trading volume can significantly enhance the performance of the models, which is in line with the results found by Lo et al. (2002), also applying different types of parametric survival analyses to model limit order execution time. A main concern is that, although an execution time model is superior, it might be inappropriate as the distributional assumptions of a traditional proportional hazard model (type of survival analysis) do not hold when modeling variables like bid-ask spread, theoretical price, time to execute and order volume.

- **Non-parametric models for limit order execution probability**
  Various studies look at execution probability modeling using non-parametric models as these models don't have assumptions regarding the data distribution. The probability distribution can change continuously depending on a variety of market factors, making models with strict assumptions less appropriate and accurate for probability estimation. Maglaras et al. (2021) used a recurrent neural network (RNN) to minimize the implementation shortfall, which is the loss incurred by the trader due to the placement of a limit order instead of a market order accounted for the time horizon in which the trader wants the limit order to be executed. The study finds that the RNN significantly outperforms the Logistic Regression benchmark in terms of accuracy of the probability predictions (measured by the AUROC), the time to fill predictions (measured by the RMSE) and the overall implementation shortfall. Thereby not only showing the superiority of the non-parametric model when comparing the test results, but also while operating under minimal model assumption. These results are in line with the paper of Dixon (2018) that finds that supervised learning with a RNN outperforms a parametric proportional hazard model based on the accuracy of predictions.

4 Methodology

4.1 Data Structure

The Crypto-asset data used in this study is obtained through a partnership with Blocktraders. This study has been given access to high-frequency trading (HFT) data of cryptocurrencies supplied from Binance [2]. The data samples of this study consists of two data types:

- **Time Series data:** the limit order book data of various cryptocurrency exchange rates (i.e., BTC/USD) with, possibly, milliseconds time intervals (depending on the cross-sectional data).
- **Cross-Sectional data:** the details of the placed and executed orders in the Crypto-asset market.

The two data types will be concatenated into one data frame containing cross-sectional data of order execution observations.

4.2 Variables: Time Series data

Bid/Ask Spread

The bid/ask spread is simply calculated by subtracting the best bid price from the best ask price.

Order Book Imbalance

This variable calculates the difference between the quantities at the best bid and the best ask.

Order Book Inventory

This variable contains the unexecuted quantity in the complete order book of the cryptocurrency.

Theoretical Price

[2] https://www.binance.com

To calculate theoretical prices of the cryptocurrency, the bid and ask prices have to be combined to find an unbiased mid-point that does or does not take slippage into consideration. Various theoretical prices are calculated to use as explanatory variables for the probability model:

- **Unweighted Price**

$$P_t = \frac{P_t^{bid} + P_t^{ask}}{2} \tag{1}$$

- **Value Weighted Average Price**

$$P_t = \frac{(Q_t^{bid} P_t^{bid} + Q_t^{ask} P_t^{ask})}{(Q_t^{bid} + Q_t^{ask})} \tag{2}$$

- **Micro Price**

$$P_t = \frac{(Q_t^{bid} P_t^{ask} + Q_t^{ask} P_t^{bid})}{(Q_t^{bid} + Q_t^{ask})} \tag{3}$$

- **Deep Price**

  This is calculated using the bid and ask prices in combination with a quantity threshold. It ensures a certain quantity that can be bought before slippage will increase due to order book depth.

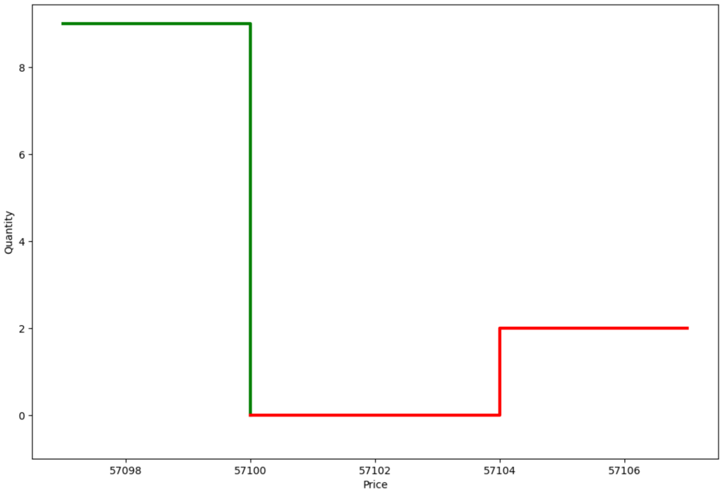
Table 1: Summary statistics

	XRP	LTC	ADA	SOL	BTC	ETH
count	1728000	1728000	1728000	1728000	1728000	1728000
mean	0.769	160.902	1.491	91.78	45580.23	2693.939
std	0.185	22.805	0.404	71.191	8525.218	891.892
min	0.449	104.882	1	13.641	28784.114	1543.538
25	0.575	149.737	1.216	19.259	38685.056	1850.555
50	0.804	157.106	1.301	75.853	46010.87	2463.141
75	0.946	176.075	1.594	158.578	51774.124	3458.045
max	1.07	206.642	2.345	197.469	59942.875	4151.402

This table contains the summary statistics for the calculated deep price of several randomly chosen cryptocurrencies exchange rates. Count considers the number of observations, the subsequent rows are based on the cryptocurrency perpetual prices and are expressed in US dollars.

An order book example for Bitcoin can be found in figure 1 below. For the buy orders, it shows that more than 8 bitcoins can be bought at a certain price. As for the sell orders, there is not enough liquidity to sell 1 bitcoin since the red line is near 0. The green line signifies bid prices and the red line signifies ask prices, with the overall line representing the accumulated quantity of the corresponding price.

Figure 1: Orderbook Bid- and Ask price



4.3 Variables: Cross-Sectional data

Order Type

This binary variable indicates whether the order is a bid (0) or an ask (1).

Order Volume

This variable constitutes the quantity of the bid or ask order that is placed in the market.

Order Price

This variable contains the price of the limit order when placed in the market (USD $)

Past Trade Volume

This variable is subdivided into multiple variables that differ in time windows. All of them contain the past trading volume at the moment the order is placed till $t - H$, where H depicts the time window used. This could result in four past trade volume variables using a 5, 30, 60 and 300 second time window.

Order Execution

This is the dependent Boolean variable which is True when the order being executed within time H and False if the order is not fully executed within time H.

4.4 Data Scaling

Scaling and transformation of variables is important for the accuracy of the model. Data will be scaled using a standardization approach when the distribution of the variable remotely resembles a normal distribution. Other variables will be transformed using a log transformation.

4.5 Data Analysis

In this section the methods of modeling execution probability are being discussed. The section will contain three different approaches for modeling the probability.

1. **Parametric Assumption Model:** Logistic Regression
2. **Non-Parametric Assumption Model:** Random Forest Regression
3. **Ensemble Model:** Soft Voting Regression

Where the Ensemble model consists of the Logistic Regression, Random Forest Regression and a Gaussian Process Regression.

4.5.1 Logistic Regression

Initially the logistic regression is modeled using all the proposed variables in the previous section, regardless of their significance on the dependent variable. This will result in the following model

$$\begin{aligned} Exec_i = c_0 + \beta_1 SPRD_i + \beta_2 IMB_i + \beta_3 INV_i + \Sigma_{j=1}^{4} \beta_4^j PRICE_i^j \\ + \beta_5 VOL_i^{order} + \beta_6 PRICE_i^{order} + \Sigma_{j=1}^{4} \beta_7^j PASTVOL_i^j + \epsilon_i \end{aligned} \tag{4}$$

Where the logistic regression is estimated separately for both bid and ask limit orders. After estimating the logistic regression, the study will inspect whether some variables are redundant in the model.

4.5.2 Random Forest Regression

Random Forest Regression/Classification was introduced by Breiman (2001) to reduce the forest error rate of single trees. Single regression trees are seen as decision trees where the algorithm puts an observation, with certain aspects, through the tree and makes a decision at every branch of the tree. After the path is completed, the algorithm reaches a conclusion based on all the past decisions that the observation has gone through. A random forest grows a significant amount of these single regression trees to reduce the forest error by combining the reached conclusion of every single tree.

Figure 2: Random Forest Model

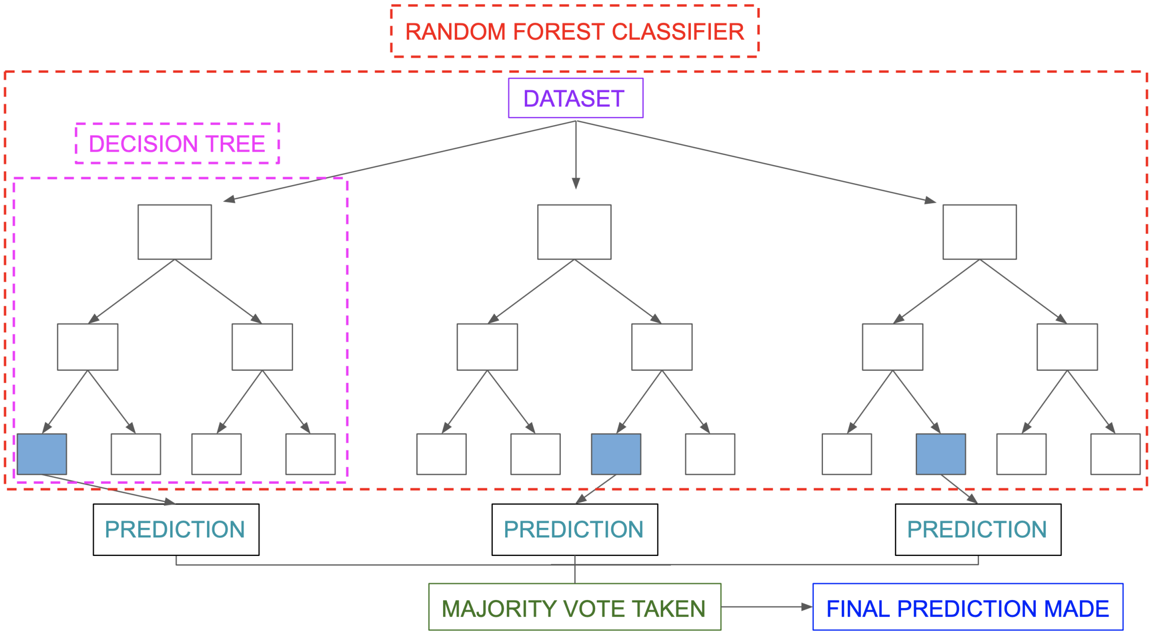


Figure Source

4.5.3 Soft Voting Regression

A soft voting regression model is not a model on itself, it is rather a combination of various trained models to produce the most accurate decisions. By combining the trained models and weight their importance in the soft voting regression model by evaluating the accuracy of every model at each observation. The underlying thought is that the errors of the individual models are somewhat independent of each other. That means that if we combine these individual models to predict together we will reduce the variance, because the aggregated model only makes a wrong prediction if more than half of the underlying models predict incorrectly.

4.5.4 Model Evaluation

All the models are being evaluated using the Area Under the Receiver Operating Characteristic (AUROC), which used to measure the accuracy of the model predictions on the test data set. The ROC curve is curve that plots the True Positive (sensitivity) predictions against the False Positive (specificity) predictions. Pure gambling on whether an order will be executed within H seconds or not, has a equal probability of predicting 'executed' (1) and predicting 'not executed' (0) (AUROC = 0,5). The AUROC value provides an interpretation of the accurateness of the predictions over the dependent variable.

References

Belgiu, M. and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114:24–31.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cao, C., Hansch, O., and Wang, X. (2009). The information content of an open limit-order book. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 29(1):16–41.

CoinMarketCap (2022). Binance trade volume and market listings.

Dixon, M. (2018). A high-frequency trade execution model for supervised learning. *High Frequency*, 1(1):32–52.

Lo, A. W., MacKinlay, A. C., and Zhang, J. (2002). Econometric models of limit-order executions. *Journal of Financial Economics*, 65(1):31–71.

Maglaras, C., Moallemi, C. C., and Wang, M. (2021). A deep learning approach to estimating fill probabilities in a limit order book. *Available at SSRN 3897438*.

Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680.

Omura, K., Tanigawa, Y., and Uno, J. (2000). Execution probability of limit orders on the tokyo stock exchange. *Available at SSRN 252588*.

Philip, R. (2020). Machine learning in a dynamic limit order market. *Available at SSRN 3630018*.

Stoikov, S. (2018). The micro-price: a high-frequency estimator of future prices. *Quantitative Finance*, 18(12):1959–1966.

Wang, J. and Zhang, C. (2006). Dynamic focus strategies for electronic trade execution in limit order markets. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*, pages 26–26. IEEE.

Yingsaeree, C. (2012). *Algorithmic trading: Model of execution probability and order placement strategy*. PhD thesis, UCL (University College London).