The Trade Information Matrix: Attributing the Performance of Strategies to Forecasting Models

The Trade Information Matrix: Attributing the Performance of Strategies to Forecasting Models

Matthew Dixon[1]

[1]Stuart School of Business, Illinois Institute of Technology, 10 West 35th Street, Chicago, IL 60616

November 17, 2016

Abstract

The confusion matrix is widely used for measuring the performance of discrete state predictive models ('classifiers'), however it fails to convey their economic utility for algorithmic trading. As a result, extensive backtesting must be performed and it can be difficult to attribute the P&L to the machine learning method. This paper introduces the concept of a 'trade information matrix' to attribute the profit and loss of classifiers under execution constraints, such as fill probabilities and position dependent trade rules, to correct and incorrect predictions. Such an approach is especially useful where execution constraints play a significant factor in alpha generation, such as high frequency trading. We further find through backtesting on Level II T-bond and E-mini S&P 500 futures history that machine learning methods have utility for market making but find no evidence to support the use of machine learning for market taking. Our conclusion is that while machine learning based price prediction may translate into economic utility through avoiding adverse selection in market making, it provides little if any advantage in gaining queue position, which is also a significant factor in strategy profitability.

1 Introduction

It is well understood that, over short time intervals, price changes are mainly driven by the order flow imbalance, defined as the imbalance between supply and demand at the best bid and ask prices (Cont et al., 2014; Cao et al., 2009; Kozhan and Salmon, 2012). Despite this imbalance being a strong indicator of price movement, prominent machine learning experts have concluded that any benefits from superior predict properties of machine learning are significantly offset by the costs of crossing the market (Kearns and Nevmyvaka, 2013) and do not justify 'market taking' - that is trading aggressively and paying the bid-ask spread.

A limit order book provides market participants with the ability to view different price levels away from the market price. Such prices and quantities are instantaneously executable modulo execution latency. Microstructure researchers seek to evaluate how the increased information stored in the limit order book informs price discovery and ultimately translates to consistent economic utility from trading.

There appears to be no consensus on the extent to which limit order books convey predictive information. Early seminal papers studying equities, including Glosten (1994); Seppi (1997), state that the limit orders beyond the best bid and offer contain little information. In contrast, several other studies state that such limit orders are informative(Parlour, 1998; Bloomfield et al., 2005; Cao et al., 2009; D. and Pollak, 2014; Zheng et al., 2013; Kearns and Nevmyvaka, 2013). In particular, Cao et al. (2009) study the information content of a limit-order book from the Australian Stock Exchange. They found that the book's contribution to price discovery is approximately 22% while the remaining comes from the first level data and transaction prices. They also demonstrate that order imbalances between the demand and supply exhibit a statistically significantly relationship to short-term future returns. Their conclusions have been supported by several more recent empirical papers (Kearns and Nevmyvaka, 2013; Zheng et al., 2013; Kercheval and Zhang, 2015).

Kempf and Korn (1999) empirically analyze the relationship between unexpected net order flow and price changes for German index futures with neural networks. Their study suggests order flow is informative but the impact on prices is non-linear: (i) the information content of order flow increases with its size; (ii) the information content from the buy side and sell side does not differ; and (iii) the best explanation for price changes is net order flow measured as contracts traded. Therefore Kempf and Korn (1999) argue that the study should be based on the whole price impact function instead of a single ratio, which could lead to different trading strategies for market participants.

Machine learning research puts the focus on exploring the study and construction of algorithms, fully using the data for prediction. It offers the key advantage of facilitating the discovery of non-linear relationships between input variables (”features”) and dependent variables, without restrictive requirements of independence of the input variables or apriori knowledge of recurrence as required in regression and time series models respectively. Using entropy to measure information content, machine learning can be applied to exploit the structure of the order book dynamics to efficiently extract information for prediction.

Anderson et al. (2008) apply an adaptive filtering algorithm in high frequency trading to investigate the information content of the order book. By reducing the impact of market shocks with support vector machines and independent component analysis, they conclude that the order book is an important source of information for predicting short-term fluctuations of returns.

Economic value The aforementioned theoretical and empirical research articles partially address the question of whether limit orders contain information beyond the best bid and ask prices. Through the proliferation of electronically traded exchanges, traders can use large numbers of variables, often available at every tick, when making trading decisions. Researchers are also able to use techniques that are more sophisticated than the standard time series analysis to forecast future price movements. However, until recently, there have been few studies focusing on whether this information can be efficiently and systematically translated to consistent economic profits. Despite finding statistically significant explanatory variables describing the structure of the limit order book, Kozhan and Salmon (2012) and Kearns and Nevmyvaka (2013) conclude, in their respective studies of the FX and Equity markets, that the information content of the limit order book does not seem to translate to greater economic profits through different high frequency trading rules. More precisely, these authors arrive at a similar conclusion that limit order book data alone is not robust enough to justify market taking.

2 Overview

While near-time price movements are relatively easy to predict, getting limit orders filled in order to capture such a movement without crossing the market is much more challenging. Consideration of such constraints should therefore be central to the assessment of the economic utility of machine learning from the limit order book. The main contribution of this paper is to introduce the concept of a trade information matrix - a weighted confusion matrix - which weights decisions by their P&L impact rather than the conventional approach of just simply comparing the model output with the actual directional price movement. We show why building very near-term predictive models from the limit order book does not justify market taking, despite some literature showing otherwise Kercheval and Zhang (2015). We further use the trade information matrix to understand how model parameters affect the P&L of a strategy and go on to demonstrate the existence of profitable market making strategies from a classifier trained on limit order book for various futures markets. Our conclusion is that while machine learning based price prediction does translate into economic utility through avoiding adverse selection, it provides little if any advantage in gaining queue position, which is also a significant factor in strategy profitability.

We begin in the next section by introducing the concept of a trade information matrix. Section 4 describes the preparation of the data used to train the classifier, referred to as the ’feature set’. Section 5 introduces our labeling methodology. Section 6 then presents results measuring the performance of the classifier. Finally in Section 7, we present the performance results of a simply market taking strategy, using the classifier to generate the trade entry and exit signals.

3 Trade Information Matrix

The concept of a confusion matrix is fundamental in the evaluation of state based predictive modeling and machine learning. In such an approach, the confusion matrix size is given by the number of states and the proportion of correct predictions form the diagonals and mis-predictions form the off-diagonal positions. It is instructive to characterize trading as an exercise in predictive modeling with additional constraints. These constraints may include path dependent decisions to enter into or exit a trade; The decision to buy or sell is rarely independent of the current position but rather constrained by it. For example, in a long only strategy, a sell signal is only actionable if a position is already held. Or buy signals may be ignored if a long position already exists. Furthermore, the weights could be set as the estimated fill probabilities. In general, the economic impact of a classification model is rarely independent of the true state or current position and we use this matrix to conveniently characterize this effect.

Note that this approach could also encompass 'trade expression', the skill of expressing an order to maximize alpha, although this is beyond the scope of the examples shown in this paper. We emphasize that we are not attempting to replace backtesting as a measure of expected P&L from a strategy but merely attribute the P&L to the correct and incorrect decisions of the classifier. Thus the 'trade information matrix', which we dub here, is simply a weighted confusion matrix.

3.1 Definitions

Let $\mathcal{D}$ denote the historical dataset of M features and N observations. When applying machine learning, or indeed fitting any predictive model, we draw a training subset $\mathcal{D}_{\text{train}} \subset \mathcal{D}$ of N_{train} observations and a test subset of $\mathcal{D}_{\text{test}} \subset \mathcal{D}$ of N_{test} observations where $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are mutually exclusive and $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}} \subseteq \mathcal{D}$. Common to all prediction models, we should avoid look ahead bias and thus the timestamp of observations $t(\mathcal{D}_{\text{train}}) < t(\mathcal{D}_{\text{test}})$.

Let $y(t) \in C := \{i \mid i : -n \to n\}$ denote a discrete response variable representing the range of discrete price movements over a prediction horizon starting at time t. Let $\hat{y}(t)$ denote the corresponding model prediction ("classification") for observation at time t in the test set $\mathcal{D}_{\text{test}}$.

Definition 3.1.1 (Trade Information Matrix) *The trade information matrix $\mathcal{TI} \in R_+^{M \times M}$, $M = 2n+1$ is defined as a weighted confusion matrix of the form*

$$\mathcal{TI}_{ij} := \sum_{k=1}^{N} \mathbb{1}_{y(t_k)=C_i} \cdot \mathbb{1}_{\hat{y}(t_k)=C_j} \cdot w_i(t), \tag{1}$$

where $w(t) \in R_+^M$ are weights corresponding to each true state i and observation at time t. When $w_i(t) = 1, \forall i, t$, we recover the classical confusion matrix.

Market Making Example To illustrate the utility of the trade information matrix for market making, suppose that the weights are defined by the product of the fill probabilities, the size of the true price movement and any additional path dependent trading constraints $\mathcal{R}(t)$ so that

$$w_i(t) := \frac{P(\mathcal{F}(t) \mid \hat{y}(t)) \cdot |\hat{y}(t)| \cdot \mathcal{R}(t)}{\sum_{k=1}^{N} P(\mathcal{F}(t) \mid \hat{y}(t)) \cdot |\hat{y}(t_k)| \cdot \mathcal{R}(t_k) \cdot \mathbb{1}_{y(t_k)=C_i}} \tag{2}$$

and the conditional probability of a fill given the prediction is

$$P(\mathcal{F}(t) \mid \hat{y}(t)) = \begin{cases} P(\mathcal{F}_{ask}(t)) & \hat{y}(t) = -1, \\ P(\mathcal{F}_{bid}(t)) & \hat{y}(t) = 1, \\ 0 & \textit{otherwise}. \end{cases} \tag{3}$$

Without loss of generality, we propose a simple estimation of the fill probability based on the order flow imbalance so that $P(\mathcal{F}_{bid}(t)) = 1 - \text{BAr}(t), P(\mathcal{F}_{ask}(t)) = \text{BAr}(t)$ and $\text{BAr}(t)$ is the ratio of best bid depth $V_{bid}^1(t)$ to the total inside market depth, defined as

$$\text{BAr}(t) = \frac{V_{bid}^1(t)}{V_{ask}^1(t) + V_{bid}^1(t)}. \tag{4}$$

Table 1 shows the confusion matrix for a classifier used to predict a trade entry signal. The row headers show the true state and the column headers show the predicted state. Each element of the matrix has been normalized to ensure that the rows sum to unity. The color coding indicates the relative contribution of the signal to the expected profit and loss of a market making strategy, ignoring transaction costs. Green indicates a profit due to correct predictions and red indicates a loss due to positioning on the wrong side of the market. Note that we ignore the effect of wrongly predicting stationary movements since this results in no trade entry signal, although this results in lost opportunities to capture market movements.

	-1	0	1
-1	0.86	0.06	0.08
0	0.15	0.70	0.15
1	0.14	0.12	0.74

Table 1: An example of an unweighted confusion matrix for a classifier used to predict a trade entry signal. The row headers show the true state and the column headers show the predicted state. Each element of the matrix has been normalized to ensure that the rows sum to unity. The color coding indicates the relative contribution of the signal to the expected profit and loss of a market making strategy, ignoring transaction costs. Green indicates a profit due to correct predictions and red indicates a loss due to positioning on the wrong side of the market.

The magnitude of the entries in Table 1 have limited interpretability as a measure of economic utility. First, the size of the market movement is not captured, although for near-term predictions this is very likely to be a price tick. Second, the signals have not been adjusted by the fill probability which is governed by a number of factors including the matching engine algorithm, order book imbalance and queue priority.

Table 2 shows the trade information matrix for the same classifier used to predict a trade entry signal. This matrix is strategy dependent and considers a simple market making strategy which quotes to buy at the best bid price if the classifier predicts an upward movement. Conversely, the strategy quotes to sell at the best ask price if the classifier predicts a downward movement. No trading constraints are placed in this simple example. Not only does the color coding indicate a profit or a loss from a correct or mis-prediction respectively, but the values now estimate the relative contribution to the expected P&L. Thus, on average, this market making strategy would be expected to make a profit because the expected profit contributions exceeds the expected loss contributions.

	-1	0	1
-1	0.66	0	0.34
0	0.45	0	0.55
1	0.33	0	0.67

Table 2: An example of the trade information matrix for a classifier used to predict a trade entry signal. The row headers show the true state and the column headers show the predicted state. Each entry of the matrix has been normalized to ensure that the rows sum to unity. The color coding indicates the relative contribution of the signal to the expected profit and loss of a market making strategy, ignoring transaction costs. Green indicates a profit due to correct predictions and red indicates a loss due to positioning on the wrong side of the market. The values represent the relative contribution to the expected P&L estimate.

In general, the concept that trading rules are path dependent and impose significant constraints is well understood and our proposed modification to the confusion matrix attempts to better characterize the likely performance of a strategy under such constraints. In particular, it helps a researcher to clearly identify if the information content of the order book or indeed any other source of data is sufficient to yield a profitable classifier under constraints such as fill probabilities. Moreover, as we shall demonstrate later in the paper, it can be used to assess the sensitivity of the strategy to execution latency and length of the prediction and training horizon.

The remainder of the paper, beginning with a description of the historical dataset, shall describe the application of the trade information matrix to evaluate a market making strategy.

4 The Data

Our dataset is a MDP3 encoded FIX format message feed from various Chicago Mercantile Exchange (CME) channels containing T-bond and ES futures from August 1, 2016 to August 31, 2016. This message feed provides information on each trading event, including time of occurrence, direction, price, volume, and transaction type etc. On each day, the quotes are captured from 6:00 C.T. to 16:00 C.T.

Packets are decoded into snapshots of 10-level deep limit order books and 'aggressors' - orders which cross the market, an extract of which is shown in Figure 4 to illustrate a typical sequence of order book updates. In the first row of this extract, the best bid is quoted at \$118.8984375 with a volume of 310 contracts provided by 62 orders. The second level bid is \$118.890625 with 960 contracts. The best ask is \$118.90625 with 228 contracts and the second level of ask is \$118.9140625 with 726 contracts. The second row represents a market crossing ('aggressor') limit order to buy 4 contracts at \$118.90625. The limit order book update is shown in the third row. We observe that the matching engine has filled the aggressor limit order reducing the quantity of best bid limit orders to 224.

```
ZNH6 (764660) SSN:12767358 ISN:2081542 Sent:1449044576396945631 Received:1449044576397017 (72)
     ( 62) 310 - 118.8984375|118.90625 - 228 (55 )
     ( 89) 960 - 118.890625|118.9140625 - 726 (81 )

ZNH6 (764660) SSN:12767360 ISN:2081543 Sent:1449044576428711249 Received:1449044576428779 Trade - 4 @ 118.90625 (Buy)

ZNH6 (764660) SSN:12767361 ISN:2081545 Sent:1449044576428728872 Received:1449044576428797 (69)
     ( 62) 310 - 118.8984375|118.90625 - 224 (53 )
     ( 89) 960 - 118.890625|118.9140625 - 726 (81 )
```

Figure 1: This figure shows a sequence of order book updates.

Note that for completeness, we include the distribution of inter-event times for ZNH6 in Figure 7 of the appendix and observe that the time between events can vary significantly. Thus when choosing the prediction horizon, we prefer a fixed length in time and not in event space.

4.1 Order Flow Imbalance

We briefly revisit the relationship between mid-price movements and order flow imbalance here. This relationship can be informally characterized by observing price movements conditioned on the BAr - the ratio of the best bid depth $V_{bid}^1(t)$ to the total inside market depth defined in Equation 4.

This ratio is chosen in preference to the bid/ask ratio as it conveniently scales from 0 to 1. Figure 2 considers a two-state representation of the ZNH6 tick-by-tick price history on 08-04-2016 in which all observations proceeding either an upward or downward mid-price movement define the universe. Conditional on the BAr either being greater than or equal to (left) or less than or equal to (right) a threshold $x \in [0, 1]$, we calculate the proportion of respective upward or downward movements as shown on the y-axis.

The left graph can be interpreted as follows: if, say, the BAr is at least 0.5 then there an approximate probability of 0.75 than the event shall be followed by an upward mid-price movement. For avoidance of doubt, it is helpful to emphasize that the plot should not be interpreted as a point-wise estimate of the conditional probability given an observed BAr, which would of course be more actionable for trading.

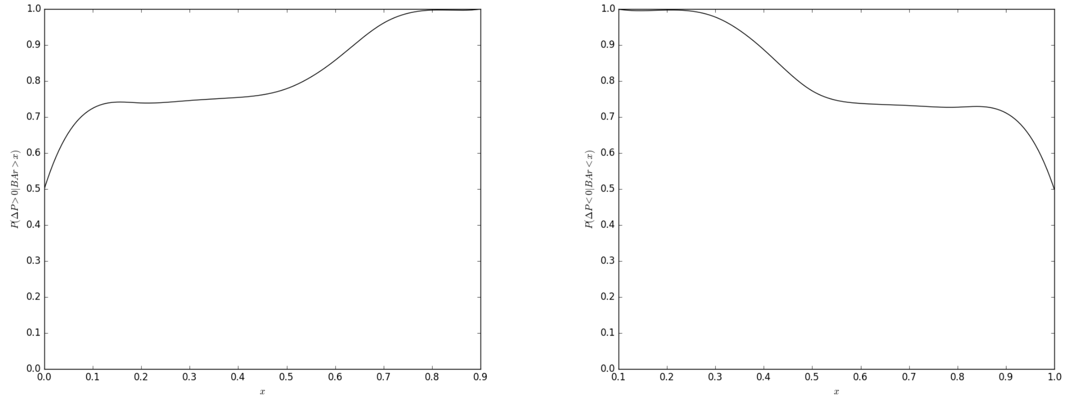


Figure 2: This figure shows the relationship between (left) upward and (right) downward mid-price movement of ZNH6 conditional on the BAr.

Such figures are well known by algorithmic traders and given such a strong indication of price movement, the reader may question why the BAr can not solely be used as a reliable price movement indicator. The answer lies with the fact that the universe defined above has been selected for pedagogical purposes. Once the stationary state is included, the strength of the BAr as a price indicator diminishes.

Figure 3 shows the distribution of BAr conditioned on a proceeding an downward (+1), stationary (0) or upward (+1) mid-price movement. The values in the parentheses denote the percentage of observations in each state. Even though the extreme values of BAr appear to be almost entirely uniquely associated with downward or upward movement, we point out that the number of stationary events significantly dominants the event space resulting in a large number, although small proportion, of extreme BAr values followed by stationary price movements.

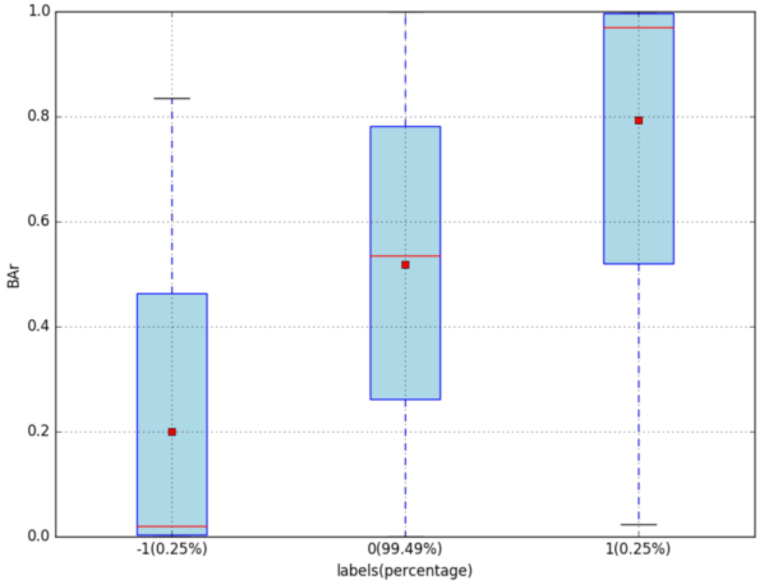


Figure 3: The figure shows the distribution of BAr conditioned on a following upward (+1), stationary (0) or downward (-1) mid-price movement.

It is standard practice by practitioners to not give up on such indicator but instead use it as one of many inputs in a classification model. The following section describes our approach to extracting 'features' from our data and using them to train a classification model.

4.2 Feature extraction

Following Kercheval and Zhang (2015), we compose our feature set of different levels of prices, volumes and number of limit orders on both the ask and bid side of the book. Each limit order book update is recorded as an observation. The average number of feature vector observations in the training set is 3,104,387 for ZN and 5,192,822 for ES.

Set	Content	Description
Limit order book	$\{P^i_{bid}, V^i_{bid}, O^i_{bid}, P^i_{ask}, V^i_{ask}, O^i_{ask}\}^n_{i=1}$	price, volume & number of orders at $n = 10$ levels

For completeness, Table 4.2 provides an example of a subset of the labelled ESU6 feature set on the 9th of August, capturing two levels of the limit order book.

Timestamp	Bid Pr. (1)	Bid Pr. (2)	Bid Vol. (1)	Bid Vol. (2)	Ask Pr. (1)	Ask Pr. (2)	Ask Vol. (1)	Ask Vol. (2)	Label
06:00:00.006	2175.75	2175.5	103	177	2176	2176.25	82	162	0
06:00:00.015	2175.75	2175.5	103	177	2176	2176.25	82	162	0
06:00:00.036	2175.75	2175.5	103	177	2176	2176.25	83	162	-1

Table 3: This table shows an extract of a subset of the features used in the labelled feature set for ES on the 9th of August. The timestamp corresponds to each book update or aggressor. This extract includes only price and volume for the first two levels of the limit order book. The full feature set compromises ten levels of price, quantity and volume on either side of the book. In addition to encoding the aggressors, all features are supplemented with their lagged counterparts at various lag intervals. Finally the feature set is labelled, as described in Section 5.

Table 4.2 summarizes the daily mid-price movements of ZNU6 and ESU6. Both contracts are sufficiently liquid to warrant a disciplined test case for machine learning. Illiquid contracts, by virtue of less price volatility, often exhibit higher predictive power but are less likely to yield profitable trading strategies. The range of ZNU6 and ESU6 mid-prices, over the one month period, is 32 and 46 ticks respectively, indicating that ESU6 is the more volatile contract.

	Price					Volume
Symbol	Mean	Max	Min	Std. Dev.	Range	Mean
ZNU6	132.2578	132.5085	132.0071	0.286401	32 Ticks	1,445,370
ESU6	2174.266	2180.049	2168.484	9.513321	46 Ticks	1,184,838

Table 4: This table provides a comparative summary of the daily mid-price movement of ZNU6 and ESU6 over the period Aug 01, 2016 to Aug 31, 2016.

5 Labeling Methodology

Application of machine learning classifiers to our feature data requires labeling each observation. The simplest approach is to delineate between a subsequent upward or downward mid-price movement, with a minimum size of a half a tick, or no movement. This approach has limited utility for use as a trade entry or exit signal since such a price movement may not be sufficient to offset trade execution costs. A pronounced undesired outcome results when a 'saw-tooth' sequence, that is a sequence of alternating up and down labels, arises from the model. This renders it impractical for trading from. A further impracticality of such a simplistic labeling approach is the absence of trade execution latency. After accounting for the computational overhead of prediction and message latency between the exchange and broker, such a signal is rendered useless.

Inspired by an approach introduced in Almgren (2013), we choose to instead label 'outlier' movements according to the labeling scheme defined here:

$$L = \begin{cases} 1 & \Delta P^h(t+l) \geq 1, \\ -1 & \Delta P^h(t+l) \leq -1, \\ 0 & otherwise \end{cases} \tag{5}$$

where L is the label and $\Delta P^h(t)$ is the net mid-price movement in ticks over a h duration bar ('observation period'), each movement being a minimum of half a tick. Accurate prediction of such an outlier as a trade entry or exit signal is likely sufficient to offset the costs of crossing the market and also enables a market maker to adjust quote volumes and levels accordingly.

Additionally, the labeling methodology builds in a tolerance for latency, represented by the 'lead time' l illustrated in Figure 5. Assuming a low-latency execution infrastructure, we suggest that the lead time should be a minimum of 1 millisecond (ms). The choice of h ms is entirely arbitrary and sensitivity of the classifier performance to h should be analyzed. Increasing the value of h results in a more balanced dataset.

It is well known that outlier detection results in unbalanced labels. In the terminology of machine learning, the outlier movements $L = \pm 1$ become the 'minority' class and no movements, the 'majority' class. Table 5 shows the percentage of labels of ZNU6 in our training set. The average percentage of label 0 is 99.9%. If left unresolved, such a degree of imbalance would result in a trained classifier always predicting 0 with a classification error of 0.01%.

date	% of label -1	% of label 0	% of label 1
8/3/2016	0.04%	99.92%	0.04%
8/4/2016	0.04%	99.91%	0.05%
8/5/2016	0.05%	99.91%	0.04%
8/8/2016	0.07%	99.88%	0.05%
8/9/2016	0.05%	99.89%	0.06%

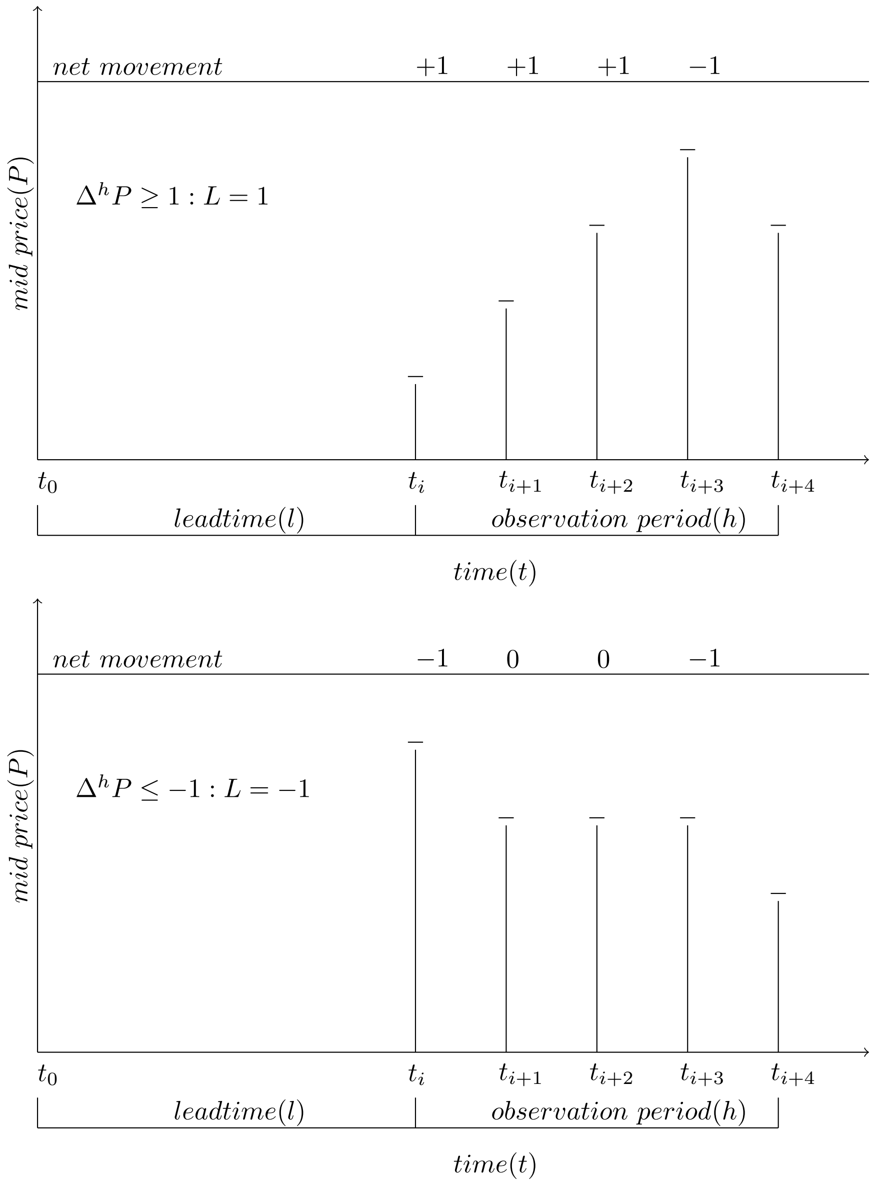
Table 5: This table shows the percentage of labels of ZNU6 on a selection of days in the training set.

To construct a balanced training set, the minority classes are oversampled with replacement and the majority class is undersampled without replacement. The resulting balanced training sets have, on average, 43630 training instances for ZNU6 and 298062 for ESU6.

6 Results

This section describes the performance of the classifier over ZNU6 and ESU6 futures dataset described in Section 4. We train a gradient boosting classifier on a historical period, test on the next day and then slide the training horizon forward by one day. This technique is referred to as 'walk forward optimization' (see, for example, Tomasini and Jaekle (2011)).

Table 6 provides two confusion matrices showing the effect of the length of the training horizon on the performance of the classifier applied to ZNU6 contracts over the month of August. All predictions are shown out-of-sample. As the number of training days increases, the classifier gets "smarter" and produces higher true positive and true negative rates. Additionally the variance in the accuracy also decreases, as shown by the standard deviations in parentheses). We use these results to determine a suitable training horizon length and trade-off computational complexity and the potential risk of using stale limit order book information with accuracy and variance.


net movement
+1
+1
+1
−1
$\Delta^h P \geq 1 : L = 1$
mid price(P)
t_0
t_i
t_{i+1}
t_{i+2}
t_{i+3}
t_{i+4}
leadtime(l)
observation period(h)
time(t)
net movement
−1
0
0
−1
$\Delta^h P \leq -1 : L = -1$
mid price(P)
t_0
t_i
t_{i+1}
t_{i+2}
t_{i+3}
t_{i+4}
leadtime(l)
observation period(h)
time(t)

	1ms latency		
	-1	0	1
-1	0.81 (0.08)	0.09 (0.05)	0.10 (0.04)
0	0.13 (0.04)	0.74 (0.07)	0.13 (0.03)
1	0.11 (0.03)	0.09 (0.04)	0.80 (0.06)

	2ms latency		
	-1	0	1
-1	0.71 (0.11)	0.12 (0.06)	0.17 (0.06)
0	0.14 (0.04)	0.70 (0.06)	0.15 (0.03)
1	0.17 (0.06)	0.11 (0.04)	0.72 (0.06)

Table 7: This table compares confusion matrices showing the effect of execution latency on the out-of-sample performance of the classifier applied to ZNU6 contracts over the month of August.

	1 Training Day		
	-1	0	1
-1	0.59 (0.19)	0.32 (0.21)	0.09 (0.04)
0	0.08 (0.04)	0.83 (0.08)	0.08 (0.05)
1	0.08 (0.04)	0.35 (0.23)	0.57 (0.20)

	3 Training Days		
	-1	0	1
-1	0.80 (0.09)	0.11 (0.07)	0.09 (0.04)
0	0.12 (0.03)	0.75 (0.07)	0.13 (0.04)
1	0.11 (0.03)	0.11 (0.07)	0.78 (0.07)

Table 6: This table show two confusion matrices comparing the effect of a one day or three day training horizon on the performance of the classifier applied to ZNU6 contracts over the month of August

6.1 Execution latency

Table 7 compares confusion matrices showing the effect of execution latency on the out-of-sample performance of the classifier applied to ZNU6 contracts over the month of August. We observe a sharp decay in the true positive and true negative rates and an increase in the variance of the true positive rate if the execution latency is doubled from 1ms to 2ms. However the impact of the increased rate of mis-classification, due to execution latency, on P&L is not clear here and in Section 7.2 we compare the trade information matrices.

6.2 Overfitting

Overfitting can be assessed by comparing the bias and variance of the in-sample and out-of-sample performance results. Overfitted models not only exhibit higher error on out-of-sample predictions than in-sample predictions, but the variance of the accuracy also increases. Tables 8 and 9 compare the in-sample and out-of-sample classifier performances for different prediction horizons $h = 0.5ms$ and $h = 5ms$, both assuming an execution latency $l = 1ms$. Each confusion matrix is formed by averaging the daily confusion matrices over the month of August. The parentheses show the standard deviations.

By comparing the in-sample and out-of-sample confusion matrices in Table 8, we observe only modest signs of over-fitting as evidenced by a marginal decrease in the true positive, neutral and negative rates, and a moderate increase in the standard deviations.

	In Sample		
	-1	0	1
-1	0.90 (0.02)	0.04 (0.01)	0.06 (0.02)
0	0.12 (0.02)	0.71 (0.04)	0.11 (0.02)
1	0.06 (0.02)	0.04 (0.01)	0.90 (0.02)

	Out-of-Sample		
	-1	0	1
-1	0.81 (0.07)	0.09 (0.05)	0.10 (0.04)
0	0.13 (0.04)	0.74 (0.06)	0.13 (0.03)
1	0.11 (0.03)	0.09 (0.04)	0.80 (0.05)

Table 8: These confusion matrices compare the (left) in-sample performance of the ZNU6 classifier with the (right) out-of-sample performance of the classifier using an execution latency $l = 1ms$ and prediction horizon $h = 0.5ms$. Each confusion matrix is formed by averaging the daily confusion matrices over the month of August. The parentheses show the standard deviation.

Increasing the prediction horizon to 5ms has undesirable consequences, despite the presence of a more balanced dataset. From Table 9 we observe not only a significant decrease of the in-sample accuracy, compared to the $h = 0.5ms$ model, but also a significant increase in the variance between the in-sample and out-of-sample prediction results. The increase in bias is comparable to the $h = 5ms$ model. These results therefore dictate the use of a shorter prediction horizon, which is deemed to yield a more robust classifier, and suggest there is still some room for improvement in feature selection to reduce bias and variance.

	In Sample		
	-1	0	1
-1	0.72 (0.04)	0.16 (0.01)	0.12 (0.03)
0	0.28 (0.02)	0.44 (0.04)	0.28 (0.02)
1	0.12 (0.03)	0.16 (0.01)	0.72 (0.04)

	Out-of-Sample		
	-1	0	1
-1	0.61 (0.18)	0.25 (0.19)	0.14 (0.09)
0	0.28 (0.11)	0.42 (0.19)	0.30 (0.1)
1	0.14 (0.07)	0.24 (0.18)	0.62 (0.14)

Table 9: These confusion matrices compare the (left) in-sample performance of the ZNU6 classifier with the (right) out-of-sample performance of the classifier using an execution latency $l = 1ms$ and prediction horizon $h = 5ms$. Each confusion matrix is formed by averaging the daily confusion matrices over the month of August. The parentheses show the standard deviation.

6.3 Performance profiling

Figure 4 illustrates daily volume and accuracy of the classifier applied to ZNU6 and ESU6 futures as a contract approaches the expiry date. The accuracy rates on average range from 60% to 80% for both upward and down predictions. The high accuracy confirms that the limit order book is informative and allow algorithms to predict very short term price movements.These results are consistent with (Kearns and Nevmyvaka, 2013; Zheng et al., 2013; Kercheval and Zhang, 2015). Moreover, we find that the accuracy decreases when there are important economic events, such as non farm payroll report release and FOMC Meeting etc. The similar finding has been described by Almgren (2013). He shows that an extraordinary event can drawdown the performance suddenly and positions should be withdrawn. Moreover, as a contract is about to expire, volume shrinks and the accuracy rate deteriorates sharply. Based on this finding, we suggest avoiding this approach to trading contracts during economic events and near expiry dates.

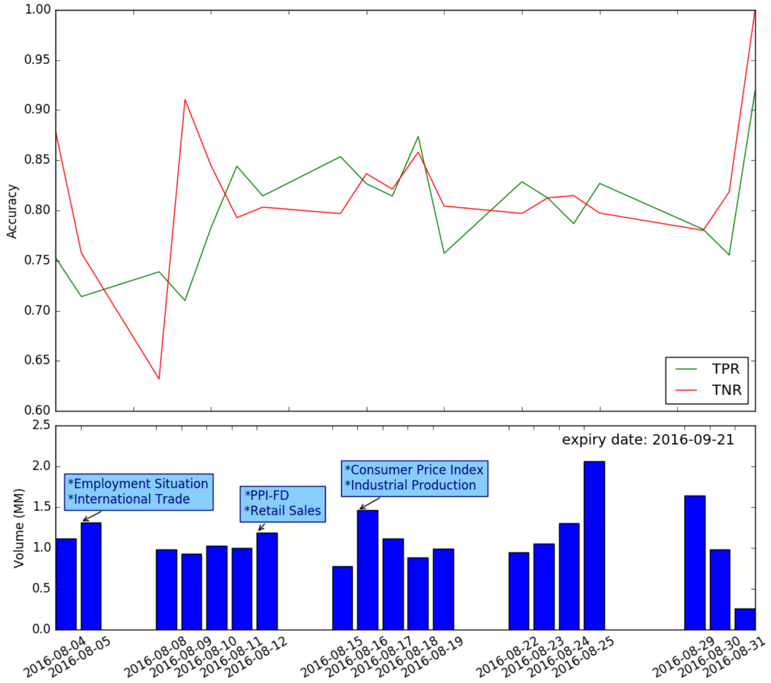


Figure 4: This figure shows daily volume and accuracy of the classifier as ZNU6 approaches the expiry date. The model assumes an execution latency of $1ms$ and predicts over a $h = 0.5ms$ horizon.

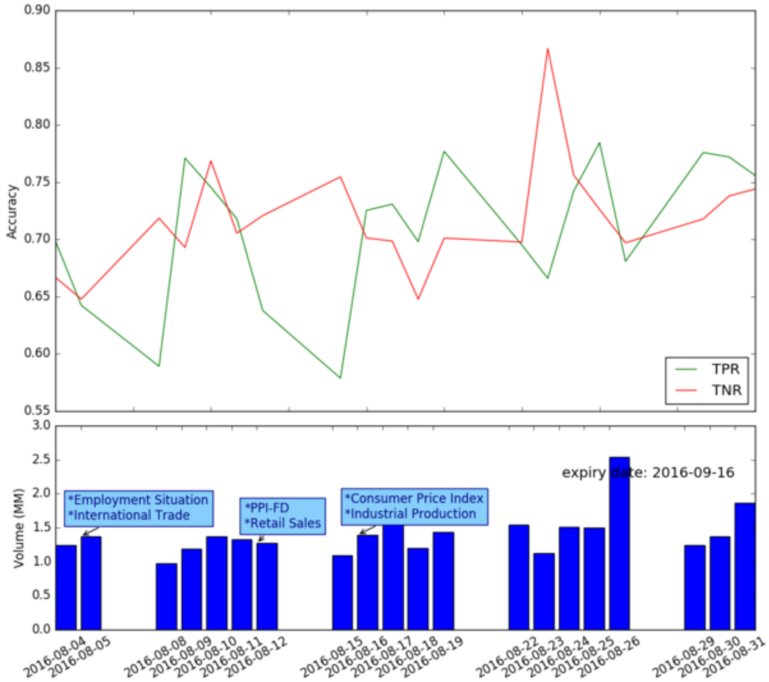


Figure 5: This figure shows daily volume and accuracy of the classifier as ESU6 approaches the expiry date. The model assumes an execution latency of $1ms$ and predicts over a $h = 0.5ms$ horizon.

7 Strategy Backtesting

7.1 Example trading strategy

In order to gain further insight into the utility of the trade information matrix, a simple market making strategy is chosen. For simplicity, the strategy only places one lot limit orders. The strategy quotes at the best bid level if there is anticipated upward movement and at the best ask level if there is an anticipated down movement. Once filled, the position is closed out when there is an predicted price movement. No quotes are placed if a corresponding position is already held.

- the account is opened with $100k of USD;
- there is sufficient surplus cash available in order to always maintain the brokerage account margin, through realization of the profit or otherwise;
- there are no limits on the minimum or maximum holding period;
- positions are closed out at market close;
- the margin account is assumed to accrue zero interest;
- execution latency is at most 1ms;
- transaction costs are assumed to be 15 cents per contract; and
- the strategy does not trade on the roll date.

The assumptions on execution latency and transaction costs are best suited to CME institutional members without low latency execution infrastructure.

7.2 Performance

Table 10 shows the trade information matrices for the market making strategy in ZNU6 over the month of August 2016, assuming a 1ms (left) and 2ms (right) execution latency respectively and $h = 0.5ms$. The fill probabilities are given by the model in Section 3.1.

The trade information matrix describes the importance of low execution latency over and above the corresponding confusion matrices given in Section 6.1. The relatively magnitude of the true positive (TP=0.73) to the false negative (FN=0.27) and the true negative (TN=0.61) to the false positive (FP=0.39) suggest that any losses from quoting on the wrong side of the market will be, on average, more than offset by trading gains from correct positioning.

When the execution latency is assumed to be 2ms, we observe that gains from correct short positions will be on average offset by bidding ahead of a downward price movement as evidenced by the respective TN=0.51 and FP=0.49 terms. There is still some comparative gain from correctly bidding versus incorrectly asking ahead of upward price movement, although it is less pronounced at the this higher execution latency.

	1ms latency		
	-1	0	1
-1	0.61 (0.06)	0	0.39 (0.12)
0	0.40 (0.11)	0.21 (0.02)	0.39 (0.08)
1	0.27 (0.07)	0	0.73 (0.05)

	2ms latency		
	-1	0	1
-1	0.51 (0.08)	0	0.49 (0.15)
0	0.45 (0.11)	0.14 (0.02)	0.41 (0.09)
1	0.36 (0.13)	0	0.64 (0.06)

Table 10: These tables show the trade information matrices for the market making strategy in ZNU6 over the month of August 2016, assuming a 1ms (left) and 2ms (right) execution latency respectively.

Figure 6 shows the cumulative P&L of the market making strategy in ZNU6 over August 2016. Note that the P&L contribution on the roll date has been dropped due to the severe drop in the performance of the classifier. Further details of the performance of the strategy are provided in the Table 11. Note that the TN, FP, TP and FN terms of the daily trade information matrix are shown in the four far right columns and attribute the daily P&L.

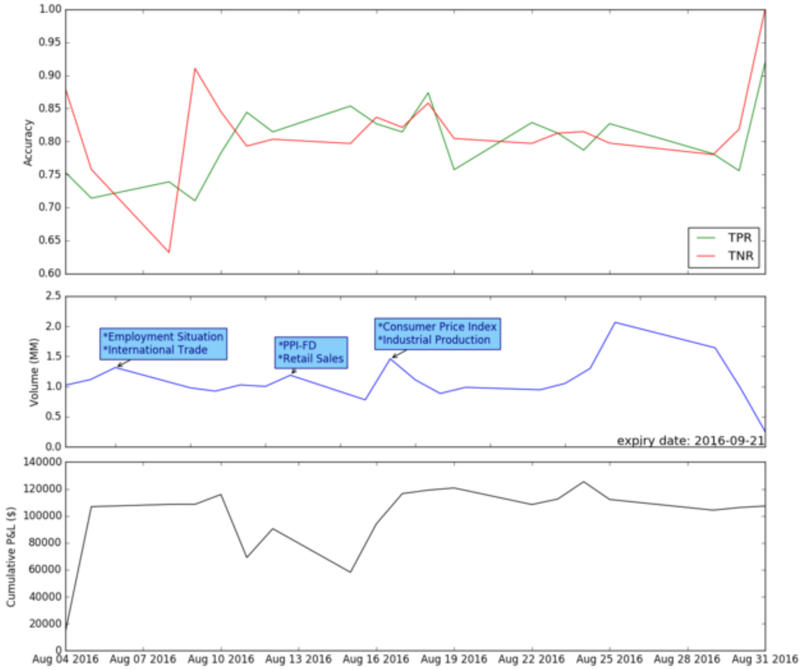


Figure 6: This figure shows the cumulative P&L of the market making strategy in ZNU6 over the month of August 2016.

Date	#trades	#long	#short	P&L (high)	P&L (low)	Profit	TN	FP	TP	FN
8/4/16	736	365	370	14086.77	-568.5	13945.75	0.57	0.43	0.53	0.47
8/5/16	1163	549	613	93880.6	-29719.67	92911.05	0.65	0.35	0.59	0.41
8/8/16	448	249	198	6501.32	-2893.78	1666.02	0.51	0.49	0.51	0.49
8/9/16	374	168	205	3560.98	-3991.48	74.25	0.50	0.50	0.51	0.49
8/10/16	559	290	268	7909.22	-606.1	7290.25	0.55	0.45	0.52	0.48
8/11/16	758	534	223	2164.3	-76061.78	-46822.98	0.44	0.56	0.36	0.64
8/12/16	932	427	504	23668.98	-17386.77	21439.03	0.60	0.40	0.57	0.43
8/15/16	542	384	157	9830.28	-36137.97	-32387.97	0.48	0.52	0.39	0.61
8/16/16	871	428	442	36065.52	-2725.88	36010.7	0.55	0.45	0.55	0.45
8/17/16	915	410	504	33852.05	-365.75	22392.15	0.54	0.46	0.52	0.48
8/18/16	571	264	306	3964.1	-732.23	2596.35	0.52	0.48	0.51	0.49
8/19/16	482	250	231	1869.15	-2710.23	1587.5	0.53	0.47	0.51	0.49
8/22/16	429	144	284	351.38	-20791.9	-12269.85	0.50	0.50	0.41	0.59
8/23/16	538	204	333	15118.33	-470.68	4054.3	0.53	0.47	0.42	0.58
8/24/16	475	190	284	12902.03	-4668.12	12902.03	0.57	0.43	0.52	0.48
8/25/16	624	410	213	5391.55	-13317.45	-13210.87	0.49	0.51	0.40	0.60
8/29/16	296	122	173	1383.85	-7944.2	-7944.2	0.49	0.51	0.45	0.55
8/30/16	282	152	129	4930.35	-830.1	1939.2	0.53	0.47	0.50	0.50
8/31/16	189	93	95	1165.12	-4977.13	1133.87	0.50	0.50	0.53	0.47
						107306.58				

Table 11: This table shows the performance of the ZNU6 classifier over the month of August 2016 assuming an execution latency $l = 1ms$ and a prediction horizon of $h = 0.5ms$. Note the TN, FP, TP and FN terms of the daily trade information matrix are shown in the four far right columns and attribute the daily P&L

8 Conclusion

Bloomfield et al. (2005) characterize 'market making' as providers of liquidity when the value of their information is low and is ultimately adopted by the traders who are least subject to adverse selection when placing limit orders. It is well understood that, over short time intervals, price changes are mainly driven by the order flow imbalance, defined as the imbalance between supply and demand at the best bid and ask prices (Cont et al., 2014; Cao et al., 2009; Kozhan and Salmon, 2012). Despite this imbalance being a strong indicator of price movement, prominent machine learning experts have concluded that any benefits from superior predictive properties of machine learning are rarely not offset by the costs of crossing the market (Kearns and Nevmyvaka, 2013).

This paper introduced the concept of a 'trade information matrix' to attribute the profit and loss of classifiers under execution constraints, such as fill probabilities and position dependent trade rules, to correct and incorrect predictions. Such an approach is especially useful where execution constraints play a significant factor in alpha generation, such as high frequency trading. We further find through backtesting on Level II T-bond and E-mini S&P 500 futures history that machine learning methods have utility for market making but find no evidence to support the use of machine learning for market taking. Our conclusion is that while machine learning based price prediction does translate into economic utility through avoiding adverse selection in market marking, it provides little if any advantage in gaining queue position, which is also a significant factor in strategy profitability.

A Additional Results

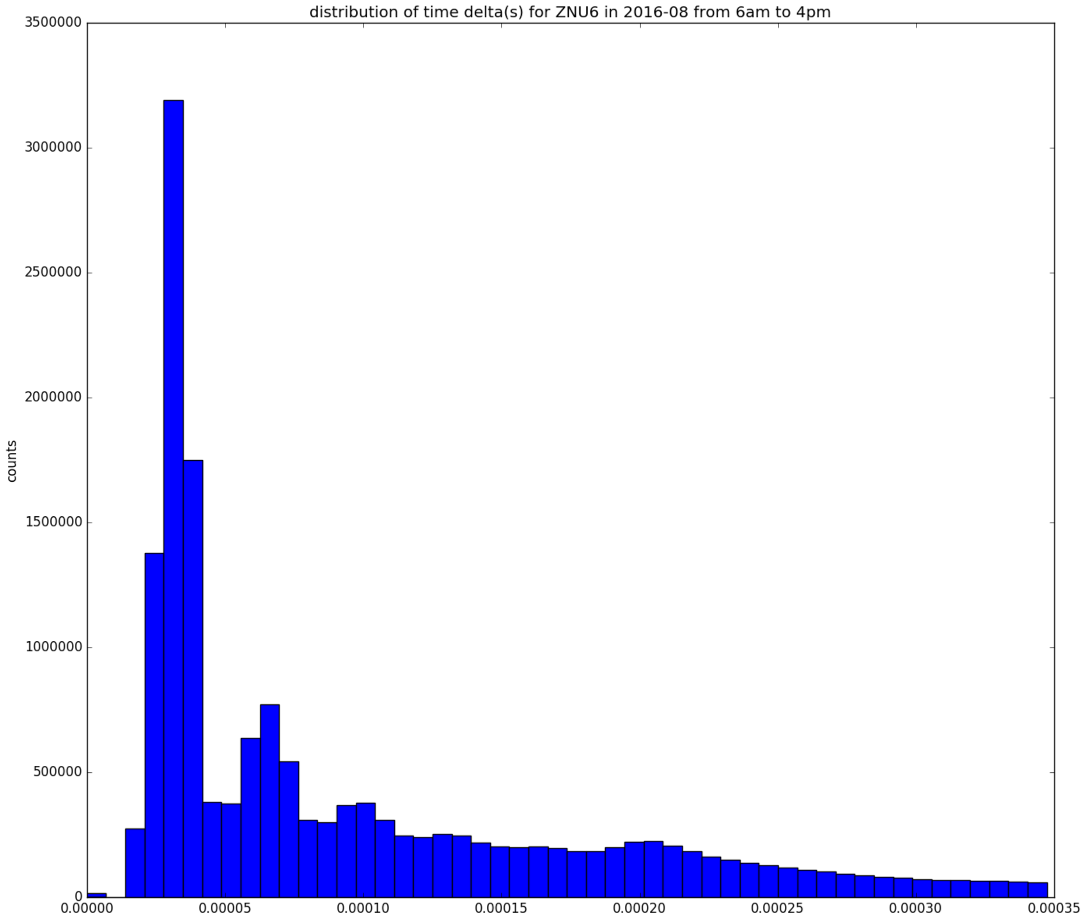


Figure 7: This figure shows the distribution of inter-event times for ZNH6.

References

R. Almgren. Execution strategies in fixed income markets. *High Frequency Trading; New Realities for Trades, Markets and Regulators*, 2013.

E. Anderson, P. Merolla, and P. Pribula. Adaptive strategies for high frequency trading, 2008.

A. Baasher and W. Fakhr. Forex trend classification using machine learning techniques. 2011.

R. Bloomfield, M. O'Hara, and G. Saar. The "make or take" decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics*, 75(1):165–199, 2005.

C. Cao, O. Hansch, and X. Wang. The information content of an open limit order book. *Journal of Futures Markets*, 29:16–41, 2009.

R. Cont, S. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12:47–88, 2014.

P. D. and I. Pollak. Mid-price prediction in a limit order book. *IEEE Journal of Selected Topics in Signal Processing*, 10, 2014.

M. Evans and R. Lyons. Meese and rogo redux: Micro-based exchange rate forecasting. *American Economic Review Papers and Proceedings*, 95:405–414, 2005.

M. Evans and R. Lyons. Understanding order flow. *International Journal of Finance and Economics*, 11: 3–23, 2006.

M. Evans, D. Martin, and R. Lyons. Order flow and exchange rate dynamics. *Journal of Political Economy*, 110:170–180, 2002.

T. Fletcher and J. Shawe-Taylor. Multiple kernel learning with fisher kernels for high-frequency currency prediction. *Computational Economics*, 42:217–240, 2013.

L. Glosten. Is the electronic open limit order book inevitable? *Journal of Finance*, pages 1127–1161, 1994.

M. Kearns and Y. Nevmyvaka. Machine learning for market microstructure and high frequency trading. *High Frequency Trading - New Realities for Traders*, 2013.

A. Kempf and O. Korn. Market depth and order size. *Journal of Financial Markets*, 2:29–48, 1999.

A. Kercheval and Y. Zhang. Modeling high-frequency limit order book dynamics with support vector machines. *Journal of Quantitative Finance*, 15(8):1315–1329, 2015.

M. King, L. Sarno, and E. Sojli. Timing exchange rates using order flow: the case of the loonie. *Journal of Banking and Finance*, 34:2917–2928, 2010.

R. Kozhan and M. Salmon. The information content of a limit order book:the case of an FX market. *Journal of Financial Markets*, 15, 2012.

C. Kuan and T. Liu. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics*, 10(4):347–64, 1995.

C. Parlour. Price dynamics in limit order markets,. *Review of Financial Studies*, 11:789–816, 1998.

V. Plakandaras, T. Papadimitriou, P. Gogas, and K. Diamantaras. Market sentiment and exchange rate directional forecasting. *Algorithmic Finance*, 4, 2015.

D. Rime, L. Sarno, and E. Sojli. Exchange rate forecasting, order flow and macroeconomic information. *Journal of International Economics*, 80:22–88, 2010.

M. Sager and M. Taylor. Commercially available order flow data and exchange rate movements: Caveat emptor. *Journal of Money, Credit and Banking*, 40:583–625, 2008.

D. Seppi. Liquidity provision with limit orders and a strategic specialist. 10:103–150, 1997.

E. Tomasini and U. Jaekle. *Trading Systems*. Harriman House Limited, 2011. ISBN 9780857191496. URL `https://books.google.com/books?id=xGIQSLujSmoC`.

B. Zheng, E. Moulines, and F. Abergel. Price jump prediction in a limit order book. *Journal of Mathematical Finance*, 3:242–255, 2013.