Novel Modeling of High-Frequency Stock Trading Data

Presentation

Yuying Huang[1], Ke Xu[2], Li Xing[3], Xuekui Zhang[1]

[1]Dept. of Math & Stat, University of Victoria, [2]Dept. of Economics, University of Victoria, [3]Dept. of Math & Stat, University of Saskatchewan

Master's Project Defence 2021

Background Introduction

Research Background

High-frequency trading (HFT) rises from increased full electronic automation in stock exchanges, which features the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders. Investment banks, hedge funds and institutional investors design and implement the algorithmic trading strategy to identify the emerging surge in stock price and then speculate from the stock market.

What to Address

- It's impossible to analyze a whole dataset
- Information redundancy issue
- Most machine learning models assume independence among the data, while close observations in high-frequency stock price data are highly correlated

Innovation & Contribution

- Improve the stock price prediction accuracy by pre-processing raw data and constructing proper input variables (predictors) from high-frequency data
  - Propose novel strategy to decrease data redundancy while avoiding information loss in high-frequency data during the data thinning process
  - Propose novel strategy to break correlation between highly correlated data points while maintain robust model performance
  - Model long-term stock price input and evaluate its impact

Definition and Notification

1. Limit Order Book (LOB) : timestamp, best ask price, best bid price
2. Stock mid-price, as our dependent variable, is defined as:

$$MP = \frac{Ask_{best} + Bid_{best}}{2}$$

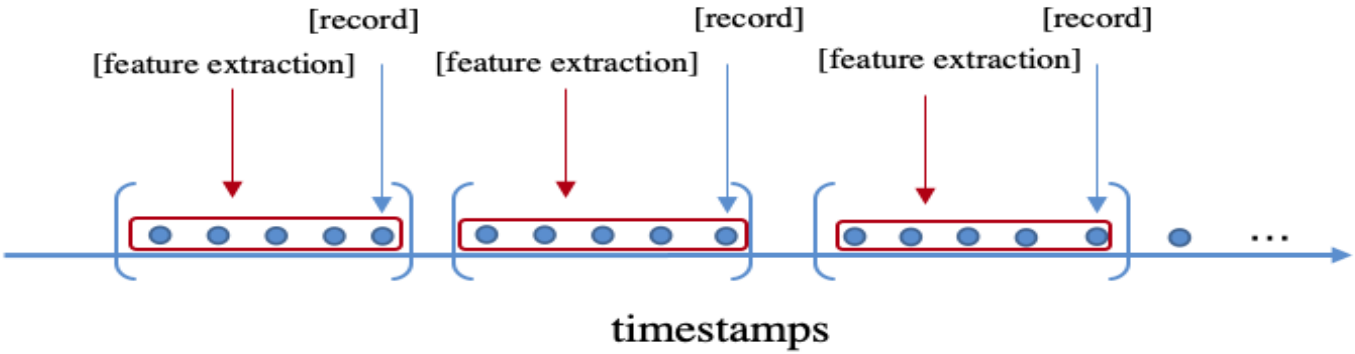
3. three-class classification problem: mid-price movement Y = {Upward, Downward, Stationary}
4. 

Figure 1: Illustration of event-based inflow framework

Definition and Notification

5. We regard the last observation in every Δt window as the "record" events, based on which, the stock mid-price movement can be labelled and features can be extracted accordingly. In our case, Δ t=5 events, blocks are independent with each other

7.

$$\begin{cases} \textit{Upwards} & \textit{if } \dfrac{MP_{t+1}}{MP_t} > 1 + 10^{-5} \\ \textit{Downwards} & \textit{if } \dfrac{MP_{t+1}}{MP_t} < 1 - 10^{-5} \\ \textit{Stationary} & \textit{otherwise} \end{cases} \tag{1}$$

Novel Strategies

Strategy I: data thinning with summary features–proposed feature set

To make amends for the lost information, during the data pre-processing, we involve new variables to extract and generalize summary features within each Δt window

Table 1: novel data thinning strategy with lately proposed "within-window" feature set

Attributes (k=5)	Category	Description
$V_4 = \{(P^{ask}_{t-i} - P^{bid}_{t-i})/MP_{t-i}\}$	within window	bid-ask spread return
$V_5 = (P^{bid}_{t-1} - P^{bid}_{t-k})/P^{bid}_{t-k}$	within window	best bid price difference return
$V_6 = (P^{ask}_{t-1} - P^{ask}_{t-k})/P^{ask}_{t-k}$	within window	best ask price difference return
$V_7 = (P^{bid}_{t-1} - P^{ask}_{t-k})/P^{ask}_{t-k}$	within window	bid-ask spread crossing return
$V_8 = (P^{ask}_{t-1} + \cdots + P^{ask}_{t-k})/k$	within window	mean best ask price in nanosecond
$V_9 = (P^{bid}_{t-1} + \cdots + P^{bid}_{t-k})/k$	within window	mean best bid price in nanosecond
$V_{10} = (MP_{t-1} + \cdots + MP_{t-k})/k$	within window	mean mid-price in nanosecond
$V_{11} = \sum_{t-k}^{t-1} V^{ask}$	within window	best ask price market depth
$V_{12} = \sum_{t-k}^{t-1} V^{bid}$	within window	best bid price market depth
$V_{13} = \{SD^{sec}_{t-i}\}$	within window	volatility in a second
$V_{14} = 1/\{time\ difference\ within\ the\ window\}$	within window	window slope

Novel Strategies

Strategy II: “Sampling + Ensemble" Model

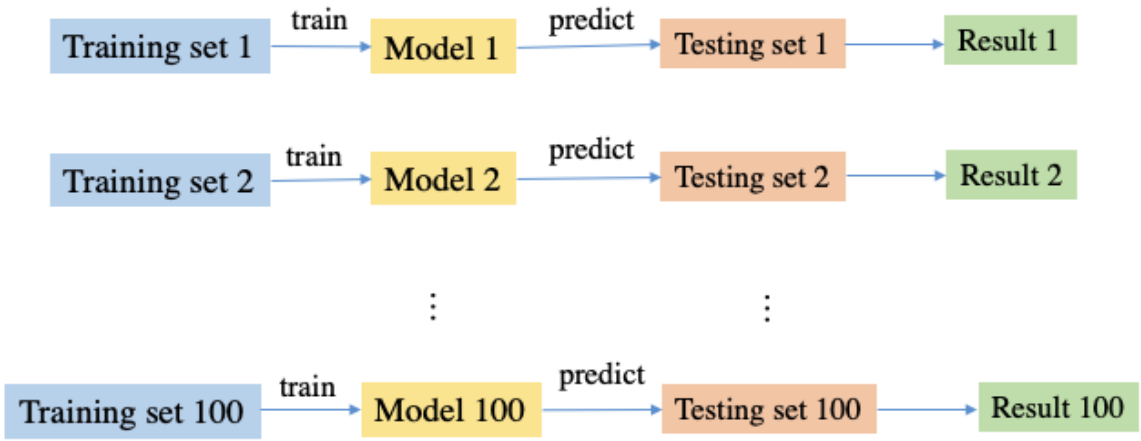


Figure 2: In an original framework, data sets are randomly split, trained and performed separately

Novel Strategies

Strategy II: "Sampling + Ensemble" Model

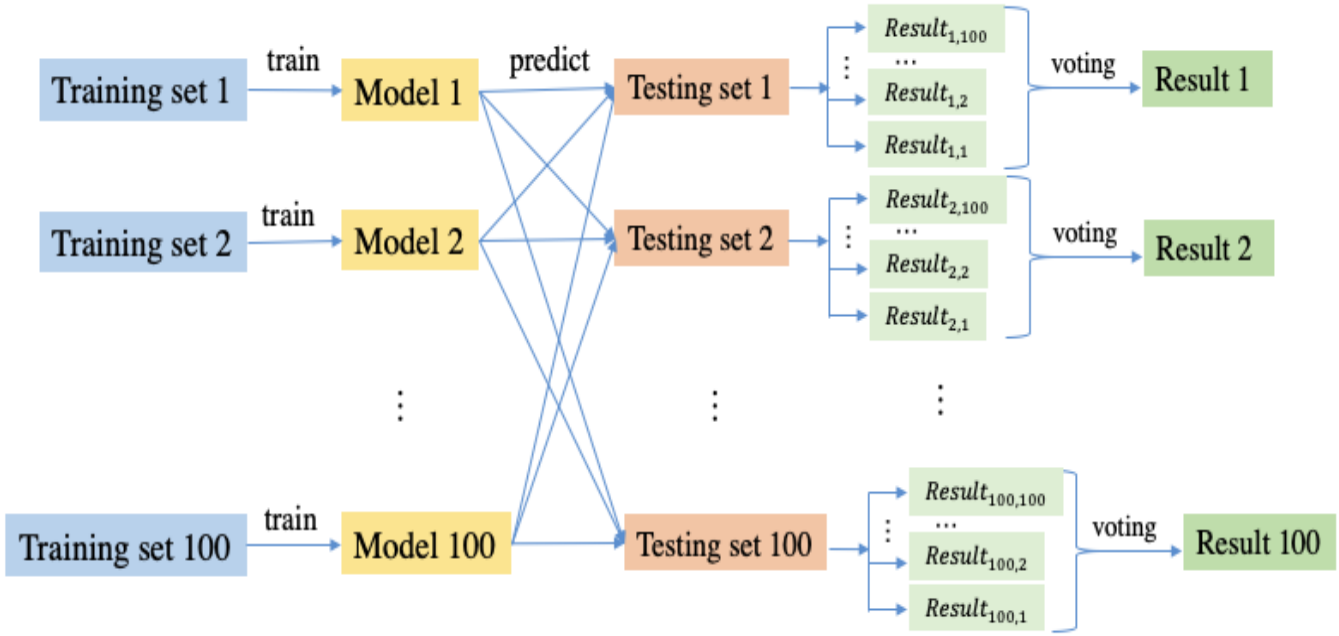


Figure 3: In a "Sampling + Ensemble" framework, data sets are randomly split. The same testing data set will be predicted through different trained models and a voting scheme will be applied to select the final prediction results

Novel Strategies

Strategy III: Combination of Long-term and Short-term Resolution

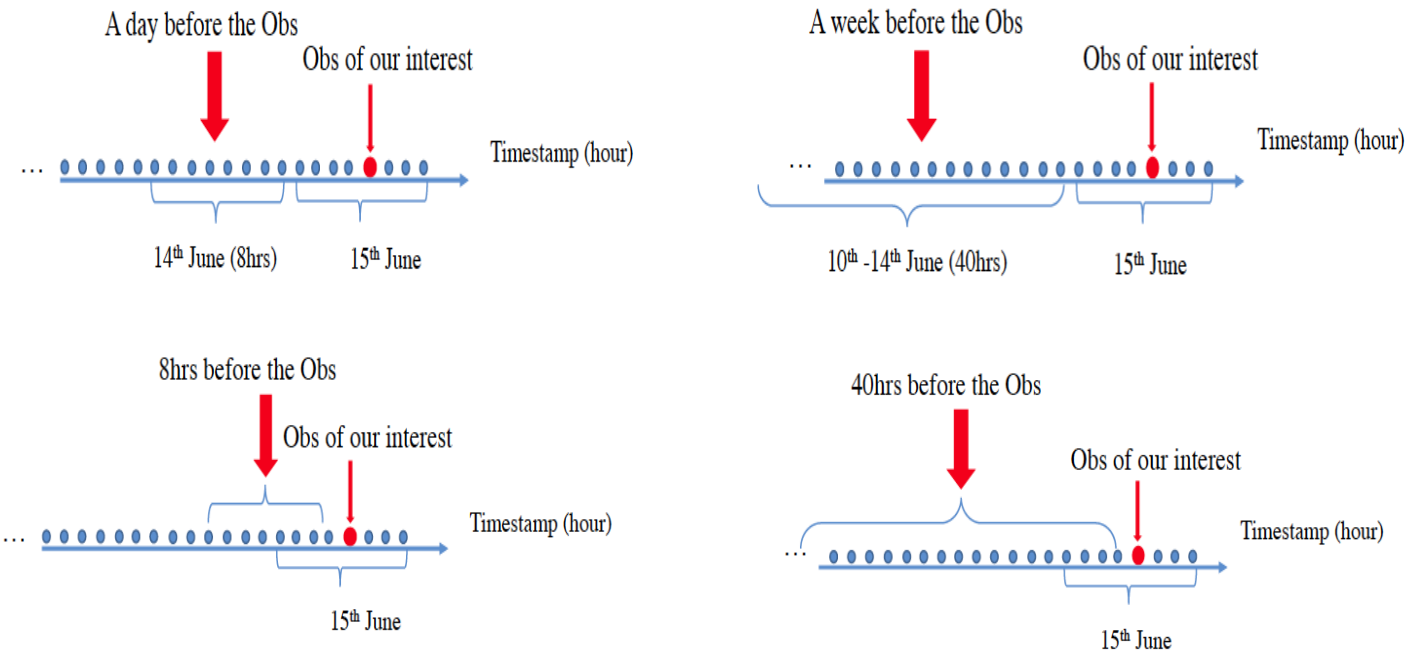


Figure 4: Illustration of two definitions of long-term curves. The upper two graphs represent the recurrence of daily pattern (definition 1) while the bottom two graphs represent the recurrence of weekly pattern (definition 2)

Novel Strategies

Strategy III: Combination of Long-term and Short-term Resolution

Table 2: Demonstration of 4 kinds of proposed FPCA scores including their signs, the number of variables used to represent them and the input data's time scope

Name	Definition	Prediction input
FPCA1	definition 1	8hrs (the day before today)
FPCA2	definition 1	40hrs (5 days before today)
FPCA3	definition 2	8hrs before this hour
FPCA4	definition 2	40hrs before this hour

Data Source

A glimpse at the data

- Stock high frequency intra day order-level data with nanosecond timestamps(e.g. HHMMSSxxxxxxxxx)
- NYSE database
  - Provides best bid- and ask prices and their volume of the limit order book
  - Transaction prices and quantities for each trade
- Sample period: Jun 1, 2017 to Aug 31, 2017 (NYSE launched on May 31, 2019)
- Massive records from 30 most popular high-tech companies' stocks by market capitalization, such as APPLE, Microsoft

Data Source

Data Manipulation

- Filter

  Initially remove all transactions recorded outside official trading time and clearly misrecorded transactions

- Winsorization & normalization

  We first compute the 1st and 3rd quantiles Q1 and Q3 of our train sample, get the difference length IQR=Q3-Q1. And then, we replace the observations falling outside [Q1-1.5IQR, Q3+1.5IQR] with the lower bound Q1- 1.5IQR and upper bound Q3+1.5IQR, respectively.

- Experiment and examine strategies

  Repeated each of the experiment 100 times with different random seeds to subsample our data set and calculate the Recall, Precision, and F1 score. Compare the model performance between model with and without our novel strategies

Evaluation Indicator

Evaluate single strategy prediction accuracy

- Classification Indicator: Precision, Recall and F_β.
  - Precision P = (True Positive)/(Predicted Positive)
  - Recall R = (True Positive)/(Positive)
  - $F_\beta = (1+\beta^2)PR/(\beta^2 P + R)$,the weighted harmonic mean of P and R. When $\beta = 1$, meaning that both Precision and Recall are equally important, we obtain the $F_1 = 2PR/(P+R)$

Evaluate the difference between novel strategies

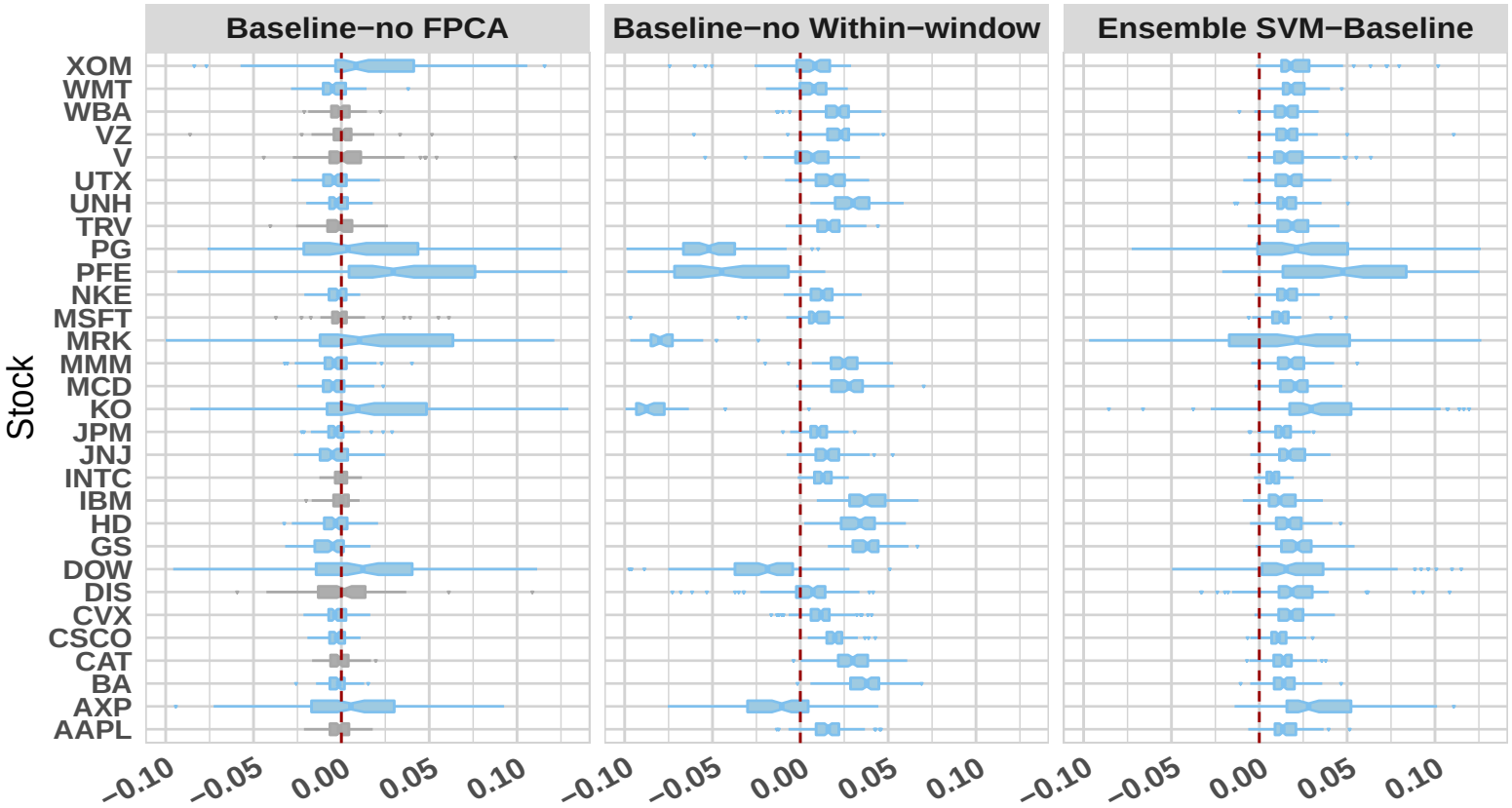
Two-sided Wilcoxon sign rank test (non-parametric test)

Result

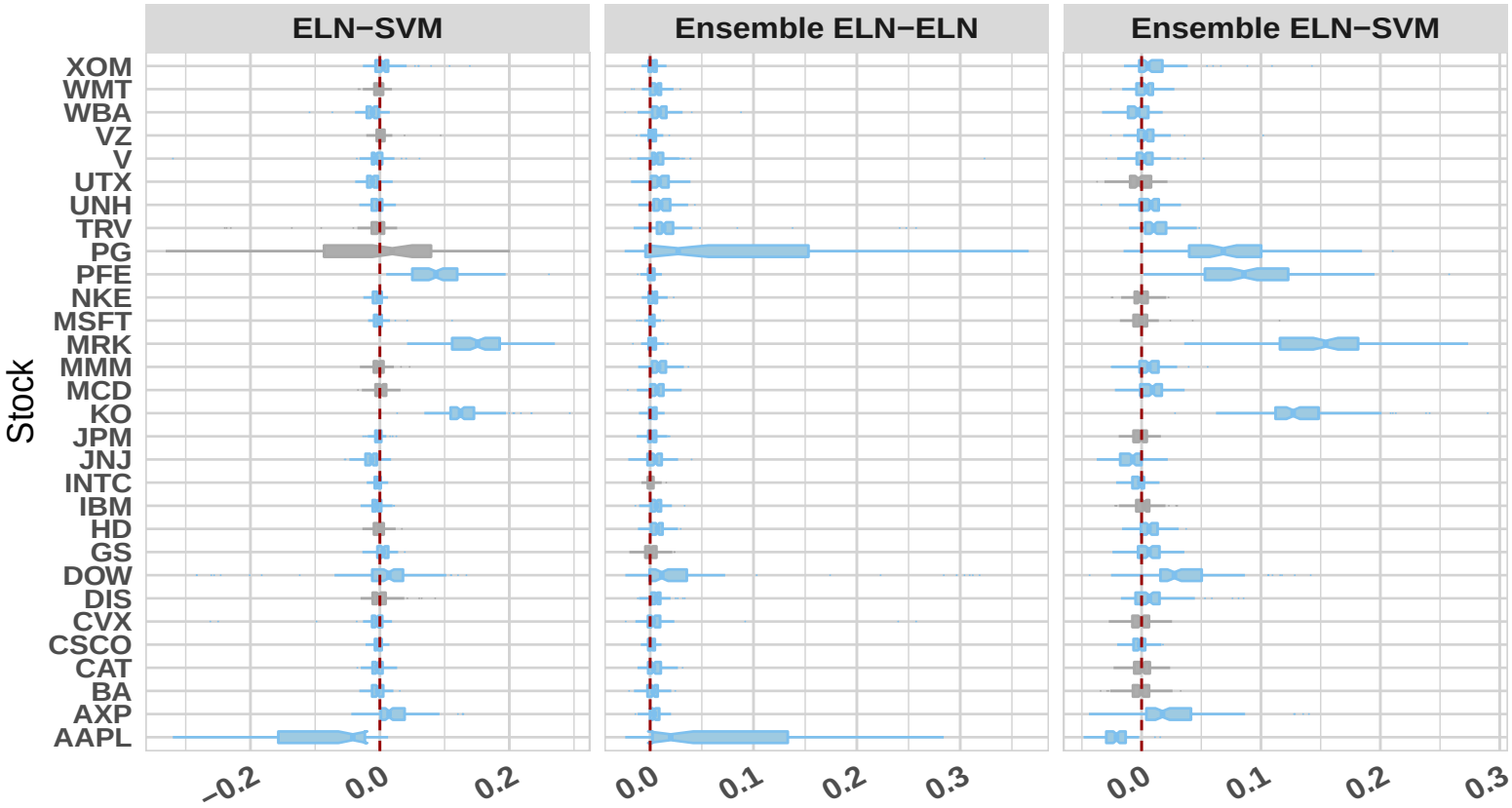
Comparisons

- baseline model: Strategy I and Strategy III (SVM and ELN)
- comparison between baseline model vs ensemble baseline model (SVM and ELN)
- comparison between baseline model vs baseline model without long-term variables (SVM)
- comparison between baseline model vs baseline model without within-window variables (SVM)
- variables selection (ELN)

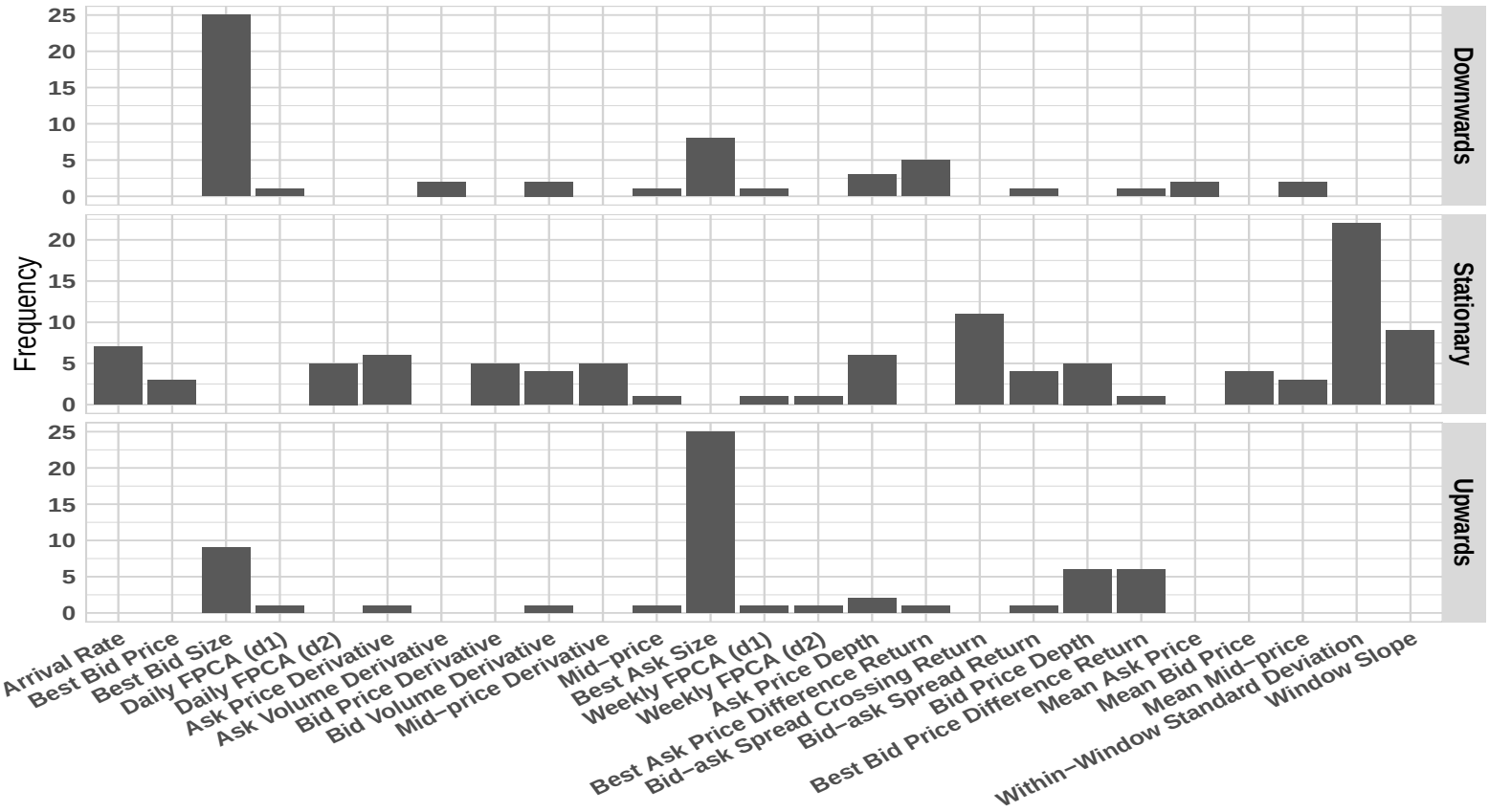
Result


Baseline−no FPCA
Baseline−no Within−window
Ensemble SVM−Baseline
Stock
XOM
WMT
WBA
VZ
V
UTX
UNH
TRV
PG
PFE
NKE
MSFT
MRK
MMM
MCD
KO
JPM
JNJ
INTC
IBM
HD
GS
DOW
DIS
CVX
CSCO
CAT
BA
AXP
AAPL
−0.10
−0.05
0.00
0.05
0.10

Result


ELN−SVM
Ensemble ELN−ELN
Ensemble ELN−SVM
Stock
XOM
WMT
WBA
VZ
V
UTX
UNH
TRV
PG
PFE
NKE
MSFT
MRK
MMM
MCD
KO
JPM
JNJ
INTC
IBM
HD
GS
DOW
DIS
CVX
CSCO
CAT
BA
AXP
AAPL
−0.2
0.0
0.2
0.0
0.1
0.2
0.3
0.0
0.1
0.2
0.3

Result


Frequency
Downwards
Stationary
Upwards
0
5
10
15
20
25
Arrival Rate
Best Bid Price
Best Bid Size
Daily FPCA (d1)
Daily FPCA (d2)
Ask Price Derivative
Ask Volume Derivative
Bid Price Derivative
Bid Volume Derivative
Mid−price Derivative
Mid−price
Best Ask Size
Weekly FPCA (d1)
Weekly FPCA (d2)
Ask Price Depth
Best Ask Price Difference Return
Bid−ask Spread Crossing Return
Bid−ask Spread Return
Bid Price Depth
Best Bid Price Difference Return
Mean Ask Price
Mean Bid Price
Mean Mid−price
Within−Window Standard Deviation
Window Slope

Thanks!