Optiver Realized Volatility Prediction

Riccardo Croci, Alessandra Di Giacomo, Marco Giuliano, Luca Salvador

Machine Learning in Finance
29/05/2024

Objectives

- Utilize **limit order book data** to forecast high-frequency realized volatility. Used detailed order book data to predict **stock price fluctuations**.

- Implement and evaluate a variety of **predictive models**.

- Evaluate the performance of these models with **RMSPE and R-squared** metrics for volatility prediction.

- **Interpret** results to understand which methods and features are the most effective.

Objectives

Realized Volatility : Actual observed price fluctuation over a specific period.

$$r_{t,t+1} = \log\left(\frac{S_{t+1}}{S_t}\right)$$

$$\sigma = \sqrt{\sum_{t=1}^{T} r_{t-1}^2}$$

- **Optiver:** Market maker, provide liquidity in the markets and avoid exposure to market fluctuations. Volatility forecast for anticipating market movements.
- **Risk Management :** High-frequency traders need to manage risk precisely due to the rapid nature of their trades.
- **Understanding Market Microstructure :** High levels of volatility are generally associated with large bid ask spread in price and size indicating changes in market liquidity and dynamics.

📊 Data Structure

- 112 different stocks, 3,830 buckets for each stock.
- Each bucket is 10 minutes and they are not ordered.
- Order Book → Details of the most competitive buy and sell orders.
  - 1st and 2nd Order Size.
  - 1st and 2nd Bid and Ask Price.
- Trade Book → Data on trades that were actually executed.
  - Price, Order Size, Order Count.
- Target variable: Realized volatility computed over the 10-minute window immediately following the period covered by the feature data.

📊 Data Structure

Feature engineering

- Augmentation of the dataset.
- Forward filling the data in the seconds missing in the dataset.
- Split each 10 minutes bucket in 10 seconds windows; in each of these windows the new variables are computed.
- 10% of the data then is kept as a test set, where the remaining 90% is used to perform 5-fold cross validation.

📊 Data Structure

Variables

- Average Bid-Ask Spread within the interval.
- Ask Spread: Average difference between the two lowest ask prices.
- Bid Spread: Average difference between the two highest bid prices.
- Spread between logarithm of highest and lowest WAP.
- Trade: Indicates whether there is a trade in the interval.

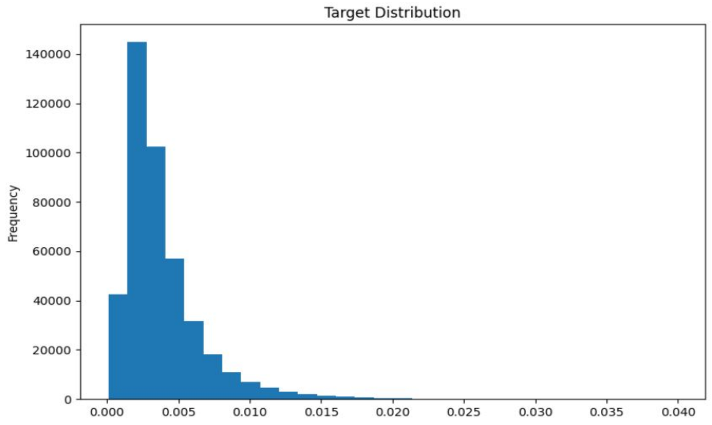
$$WAP = \frac{BidPrice \times AskSize + AskPrice \times BidSize}{BidSize + AskSize}$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

Data Structure

Summary Statistics

- Target Variable → Highly skewed
- Difference in term of volatility among stocks.
- Most volatile stocks presents:
  - Huge amount of trades
  - Large Bid-Ask spread
  - Imbalance between buy and sell orders
- If large trade does not widen the spread → market is highly liquid.



Statistic	Value
Mean	0.0039
Standard Deviation	0.0029
Minimum	0.0001
Maximum	0.07
Skewness	2.82
Kurtosis	14.96

📊 Data Structure

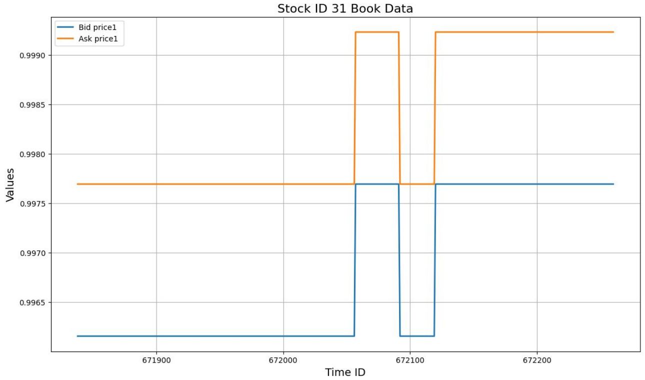
Summary Statistics

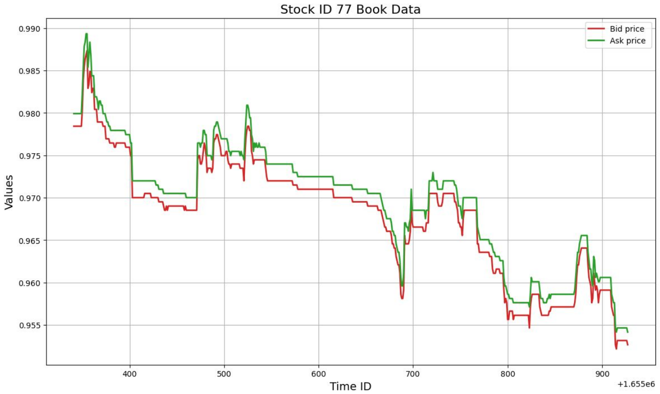
Table 1: Summary Statistics for Least Volatile Bucket

	bid_price1	ask_price1	ask_size1	bid_size1	price
mean	0.9968	0.9983	87543	141486.7	0.9977
std	0.0008	0.0008	47182.2	30536.5	0
min	0.9962	0.9977	13979	28170	0.9977
max	0.9977	0.9992	191692	186430	0.9977

Table 2: Summary Statistics for Most Volatile Bucket

	bid_price1	ask_price1	ask_size1	bid_size1	price
mean	0.9685	0.9691	10074.3	1539.8	0.9680
std	0.0071	0.0072	10935.8	1434.4	0.0080
min	0.9527	0.9537	16	30	0.9530
max	0.9879	0.9888	48222	8021	0.9880





Least and most volatile Bid-Ask Spread

Model definition

GARCH

Unsatisfactory results → RMSPE = 0.9

- Tried to predict next 600 values using 600 observations (or less)

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

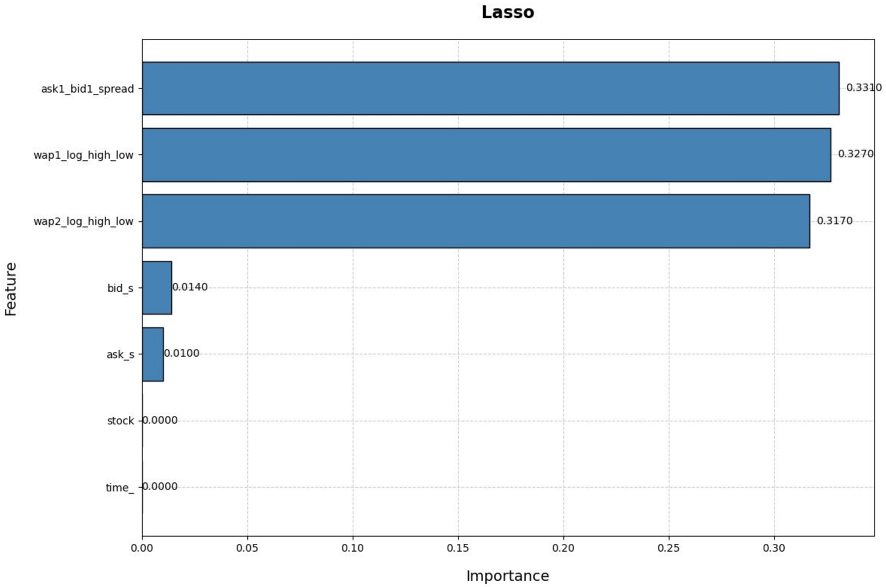
Model definition

Lasso

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Test Metrics:
  RMSPE = 0.308, R^2 =0.794

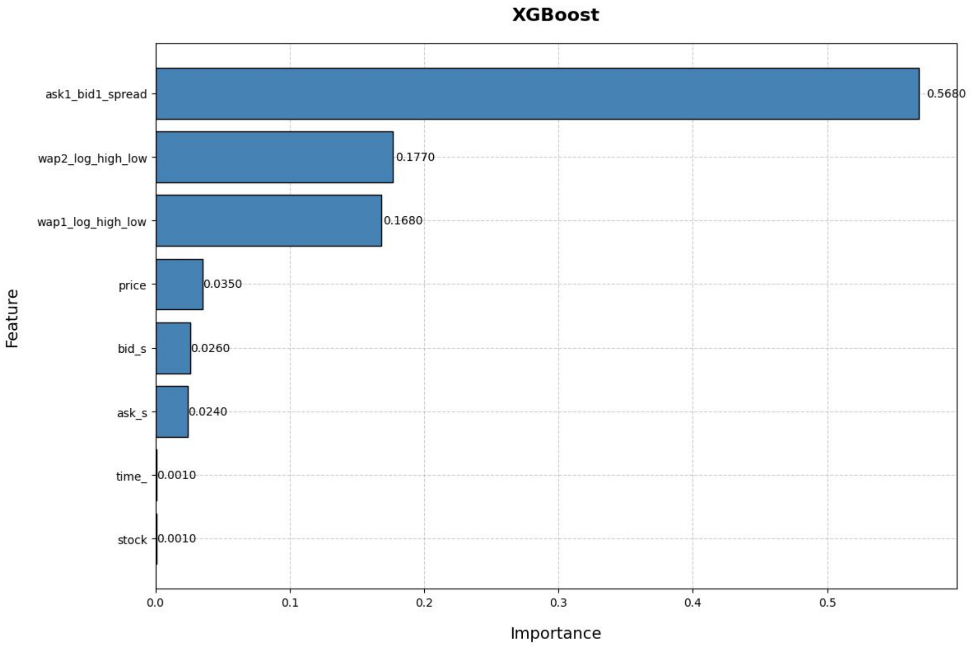
- The relevant features are:
  - Average Bid-Ask Spread
  - Spreads between highest and lowest log(WAP)



Model definition

XGBoost

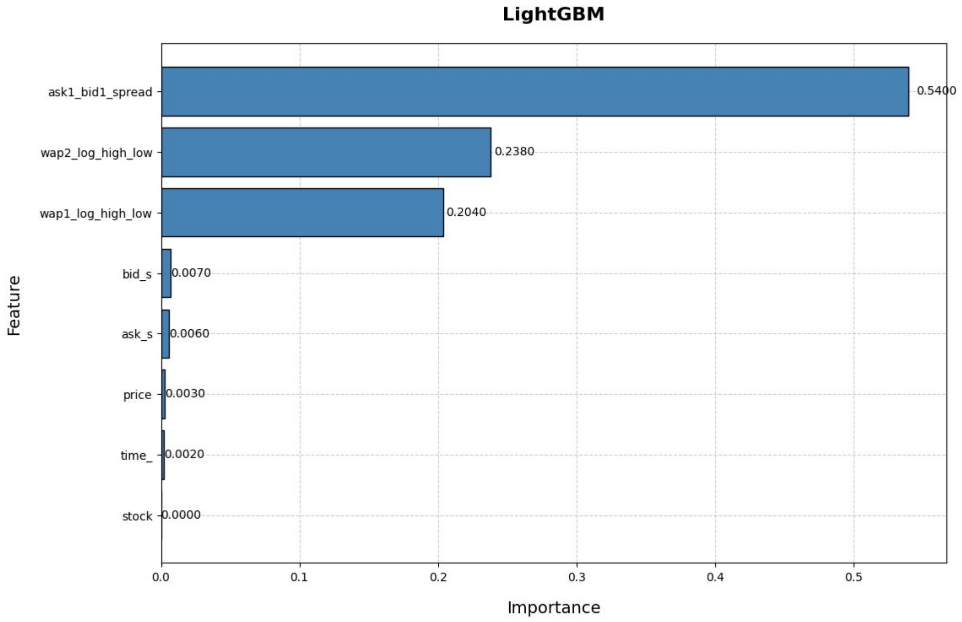
- Grid search with 5-fold cross-validation
- Optimal Hyperparameters:
  - Number of estimators: 600
  - Max depth: 7
  - Learning rate: 0.1
- RMSPE = 0.258, R^2 = 0.801
- Relevant features:
  - Bid ask spread
  - Spread between highest and lowest log(WAP)
  - Trade



Model definition

LightGBM

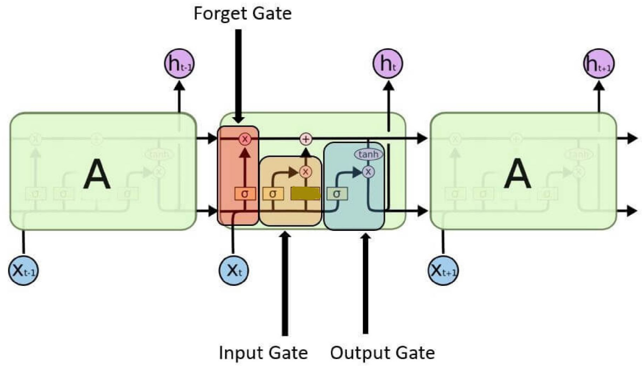
- Woks by constructing an ensemble of decision trees
- No significant improvement w.r.t XGBoost
- Slight reduction in complexity: Lower optimal number of estimator
- Feature importance : same as XGBoost. Trade not relevant any more.
- Performance metrics on test set :
  RMSPE = 0.276
  R^2 = 0.814



Model definition

LSTM

- Tanh activation function performs better than Linear and Sigmoid.
- 1 layer only : increase complexity does not improve results.
- Optimal Hyperparameters
    - Dimension of the hidden state: 96 neurons
    - Learning rate: 0.001



Prediction results similar to other models: $R^2 = 0.743$
RMSPE = 0.24
Unordered data reduce the power of LSTM network structure.

Model definition

MLP

- Multi-Layer Perceptron
  - Learning rate scheduler → prevent overfitting
  - 30 epochs
  - Drop out probability of 0.1
  - 3 hidden layer
  - RMSPE = 0.237 , R^2=0.74
- Feature importance
  - WAP
  - Bid Ask Spread

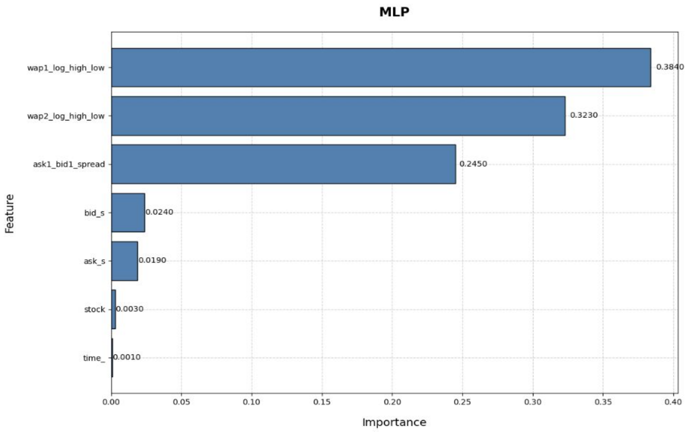


Figure 7: Feature importance with feature permutation for MLP

Results

	RMSPE	R^2
GARCH	0.9	x
LASSO	0.308	0.794
XGBoost	0.258	0.801
LightGBM	0.276	$\boxed{0.814}$
LSTM	0.240	0.743
MLP	$\boxed{0.237}$	0.742

📈 Results

- The most relevant features in predicting volatility are : **Bid-Ask Spread** and **Spread** of **log(WAP).**
- In our opinion, the bid-ask spread is highly relevant because it is correlated with volatility. The **bid-ask spread** tends to **widen** during periods of **high volatility** due to increased uncertainty among market participants.
- The range between the log highest and lowest WAP values captures the **price fluctuations** within a specific period. Large spreads indicate significant price movements, which are often associated with high volatility.
- The difference between the two most competitive ask and bid levels is not relevant since markets are highly **liquid.**

Conclusion

- **Standard econometrics** techniques don’t perform well on this task, especially because data is unordered, and only ordered in the 10 minute bucket. **Machine learning** techniques appear to perform better.
- The **presence of trades** in the interval seems **not** to be **relevant**. This can be related to the fact that number of trades is much lower than the total number of observations.
- The models highlight the significance or **recent market data**, particularly within the last few seconds of the buckets