Short-Term Volatility Prediction for Stocks

Data Science Boot Camp (May-Summer 2024)

THE ERDŐS INSTITUTE

Li(Julie) Zhu

May 31, 2024

Problem Description

1 Introduction

- Volatility: a reflection of the degree to which stock price moves, usually defined by the standard deviation of the stock log returns in 1-year period
- Design a model to forecast volatility for hundreds of stocks across different sectors over 10-minute periods
- Source: Kaggle Competition — Optiver Realized Volatility Prediction
- Stakeholder: Traders for options, ETFs, cash equities, bonds and foreign currencies on numerous exchanges around the world.

Data Files

1 Introduction

- **book_[train/test].parquet** [1]
  - Provides order book [2] data on the most competitive buy and sell orders entered into the market.
  - `stock_id`: ID code for the stock, `time_id`: ID code for the time bucket
  - `seconds_in_bucket`: Number of seconds from the start of the bucket, always starting from 0.
  - `bid_price[1/2]`: Normalized prices of the most/second most competitive buy level
    `ask_price[1/2]`: Normalized prices of the most/second most competitive sell level
  - `bid_size[1/2]`: The number of shares on the most/second most competitive buy level
    `ask_size[1/2]`: The number of shares on the most/second most competitive sell level.
- **trade_[train/test].parquet**
  - `stock_id`, `time_id`, `seconds_in_bucket`
  - `price`: The average price of executed transactions happening in one second.
  - `size`: The sum number of shares traded.
  - `order_count`: The number of unique trade orders taking place.
- **train.csv**
  - `stock_id`, `time_id`
  - `target`: The realized volatility computed over the 10 minute window following the feature data under the same `stock_id`/`time_id`.
- **test.csv**
  - `stock_id`, `time_id`

[1] Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk.

[2] The term order book refers to an electronic list of buy and sell orders for a specific security or financial instrument organized by price level.

Fundamental Statistics

1 Introduction

- Weighted averaged price (stock valuation):

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

- Log returns:

$$r_{t_1,t_2} = \log\left(\frac{S_{t_2}}{S_{t_1}}\right),$$

  where S_t is the price (approximated by WAP) of the stock S at time t

- Realized volatility:

$$\sigma = \sqrt{\sum_t r^2_{t-1,t}}$$

- Root Mean Square Percentage Error:

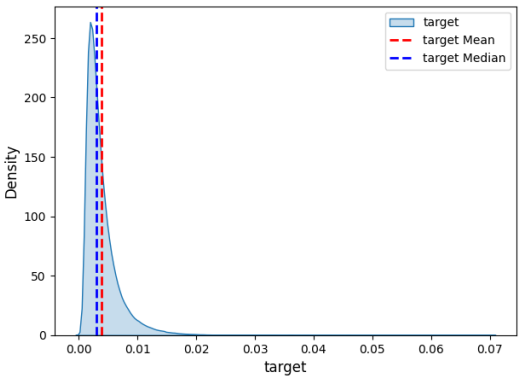
$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((y_i - \hat{y}_i)/y_i)^2}$$

Train.csv

2 Exploratory Data Analysis (EDA)

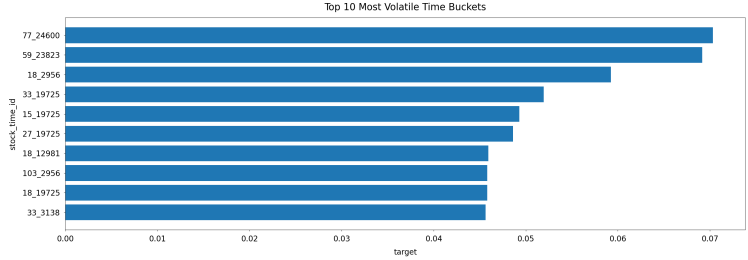
Observations

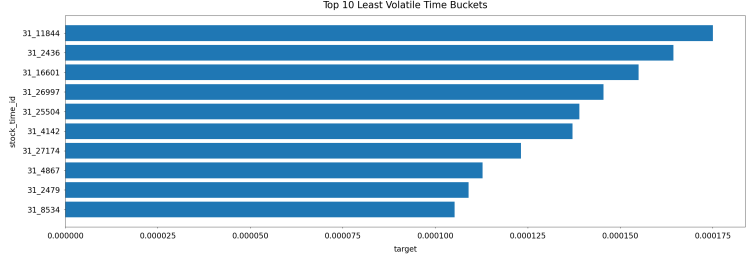
- There are 428932 rows
- The total number of different stocks is 112
- The total number of different time_id is 3830
- Mean: 0.0039 - Median: 0.0030 - Std: 0.0029
  Min: 0.0001 - 25%: 0.0020 - 50%: 0.0030 - 75%: 0.0047 - Max: 0.0703
  Skew: 2.8226 - Kurtosis: 14.9611



Train.csv

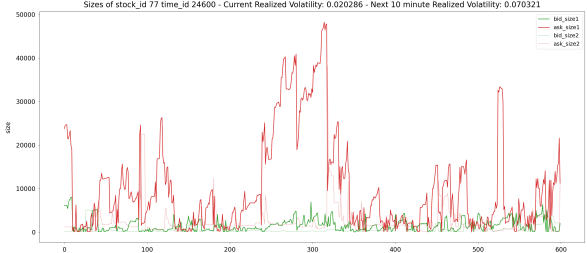
2 Exploratory Data Analysis (EDA)





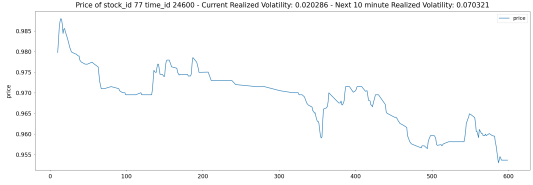
Order Book

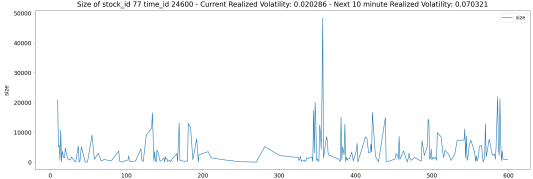
2 Exploratory Data Analysis (EDA)

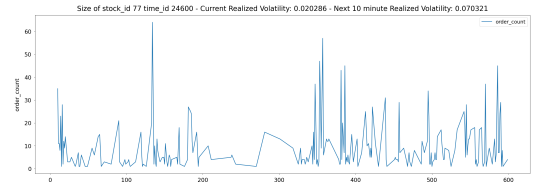


Trade Book

2 Exploratory Data Analysis (EDA)

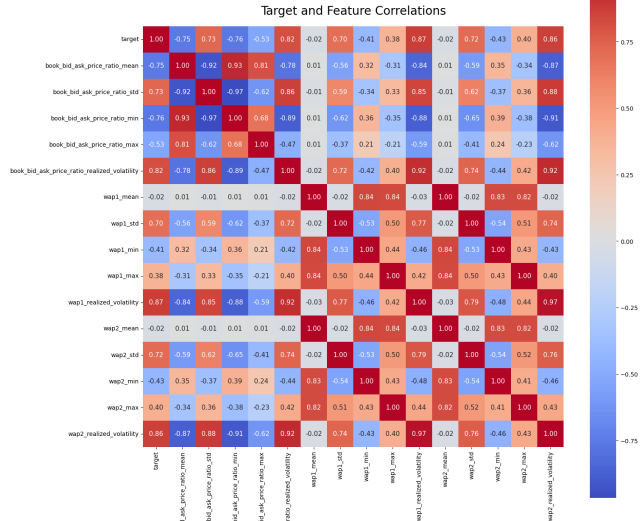






Feature Correlations

3 Feature Engineering



Models

4 RMSPE Comparison

- Baseline Model
  - target Mean: 1.110330
  - stock_id target Mean: 0.789618
  - stock_id target Median: 0.589135
  - Realized Volatility from WAP1: 0.341354
  - Realized Volatility from WAP2: 0.705453
  - Realized Volatility from price: 0.380267
- Linear Regression: 0.352226
- K-Nearest-Neighbors (KNN): 0.333281
- XGBoost: 0.028044

Conclusion and Future Work

5 Conclusion

- XGBoost works the best but maybe have overfitting issue
- Build GNN model (GNN for Realized Volatility Prediction)