CSE523: Machine Learning

Weekly Report-3

Submitted to: Prof Mehul Raval

Date of submission: 22/02/2022
Group No.4
Optimizers

Name	Enrolment Number
Digvijaysinh Gohil	AU1940199
Satya Shah	AU1940288
Smit Shah	AU1940291
Tirth Kanani	AU1920144

Insights from the EDA:

1. **book_[train/test].parquet:**

A parquet file partitioned by stock_id. Provides order book data on the most competitive buy and sell orders entered into the market. The top two levels of the book are shared. The first level of the book will be more competitive in price terms, it will then receive execution priority over the second level.

- **stock_id -** ID code for the stock. Not all stock IDs exist in every time bucket.
- **time_id -** ID code for the time bucket. Time IDs are not necessarily sequential but are consistent across all stocks.
- **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.

2. **trade_[train/test].parquet:**

A parquet file partitioned by stock_id. Contains data on trades that actually executed. Usually, in the market, there are more passive buy/sell intention updates (book updates) than actual trades, therefore one may expect this file to be more sparse than the order book.

- **stock_id -** Same as above.
- **time_id -** Same as above.
- **seconds_in_bucket -** Same as above.
- **price -** The average price of executed transactions happening in one second. Prices have been normalized and the average has been weighted by the number of shares traded in each transaction. The formula for which was mentioned in the last report.
- **size -** The sum number of shares traded.
- **order_count -** The number of unique trade orders taking place.

3. **train.csv:**

- **stock_id -** Same as above, but since this is a csv the column will load as an integer instead of categorical.
- **time_id -** Same as above.
- **target -** The realized volatility computed over the 10 minute window following the feature data under the same stock/time_id.

4. **test.csv**

- **stock_id -** Same as above.
- **time_id -** Same as above.
- **row_id -** Didn't understand much have to research more about it.

Work for next week: We will continue with the data exploration as the data is quite large and we are trying to understand it better in order tune the features in the best possible way. We will try to fit some models if possible but due midterms it may get pushed back a week. We will also read some of the research articles mentioned in the previous reports in order to get more clarity on the data and the model building part.

References:

https://www.kaggle.com/kentata/time-series-data-exploration

https://towardsdatascience.com/common-time-series-data-analysis-methods-and-forecasting-models-in-python-f0565b68a3d8

https://ieeexplore.ieee.org/document/8626097