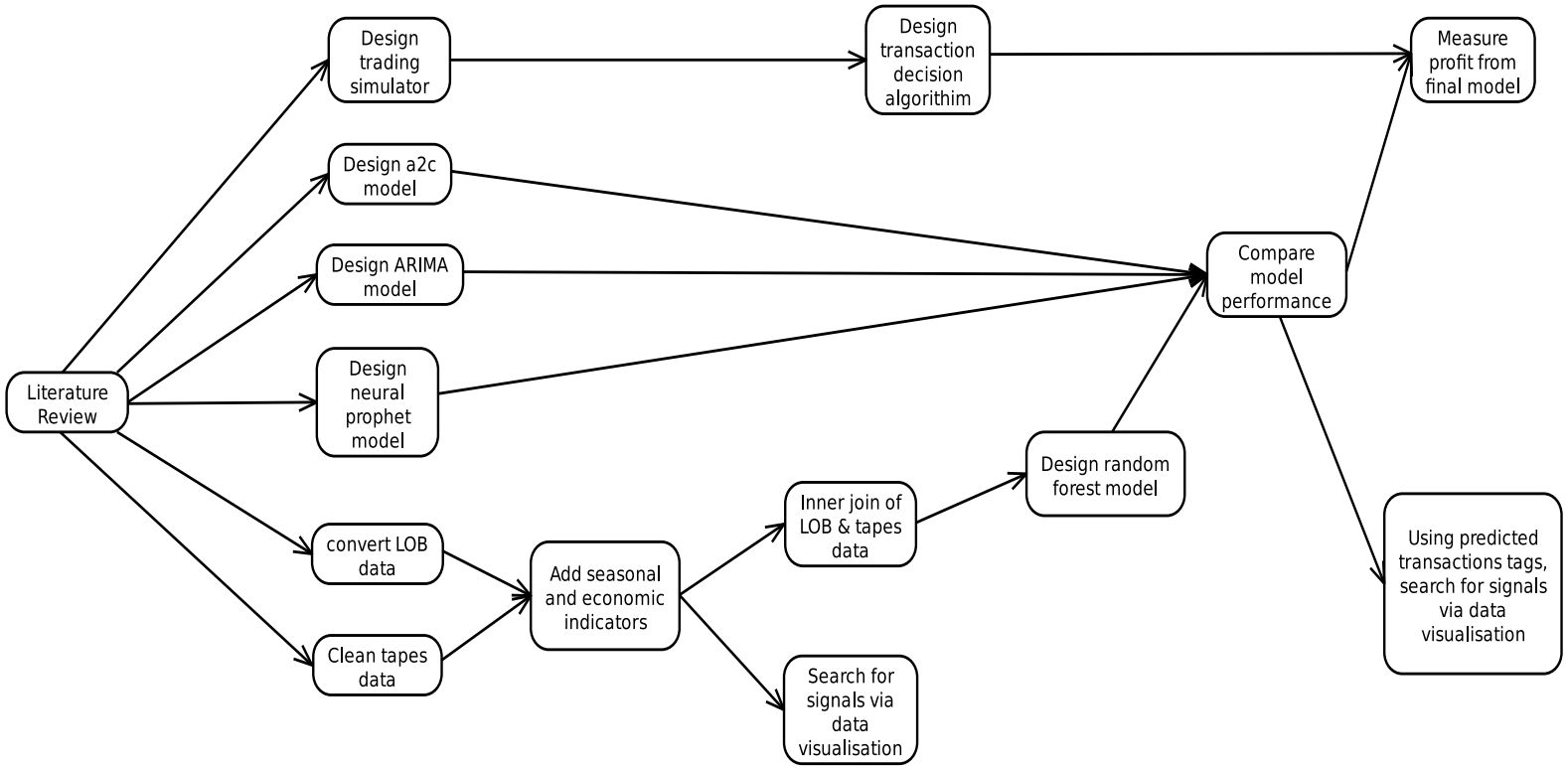
DATA SCIENCE MINI-PROJECT: HSBC GLOBAL MARKETS

Amy Gardiner, Brooke Grantham, Michael McCoubrey, Aditya Singh

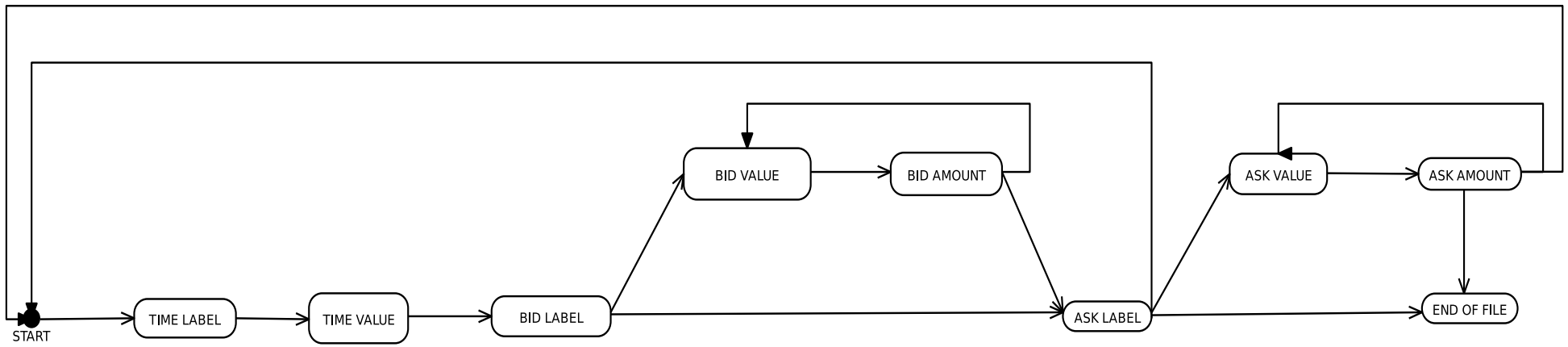
PRESENTATION STRUCTURE

- Data preparation
- Model descriptions and result discussions
- Further work considerations



DATA PREPARATION – TAPE DATA

- Followed a distributive approach using PySpark.
- Cleaned the data to retain only the required columns.
- Applied multiple transformations on cleaned data to derive important features required for training the model.




```
[
  "time",
  1.936,
  [
    "bid",
    [
      [
        183,
        4
      ],
      [
        27,
        3
      ]
    ]
  ],
  [
    "ask",
    [
      [
        739,
        5
      ]
    ]
  ]
]
```


DATA PREPARATION – LOB DATA

- The data needed to be converted into a tabular format.
- The original format didn't follow any standard format and needed a custom solution.
- The data files were extremely large and so the solution had to be memory efficient, otherwise a regular desktop would run out of memory.
- The additional dataset used a different format so couldn't use this.

DERIVING FEATURES – LOB DATA

Spread Features:

- Micro price

$$P_{micro} = \frac{(Q_a \times askBestPrice) + (Q_b \times bidBestPrice)}{Q_a + Q_b}$$

- Mid price

$$P_{mid} = \frac{askBestPrice + bidBestPrice}{2}$$

- Quoted Spread

$$QuotedSpread = bidBestPrice - askBestPrice,$$

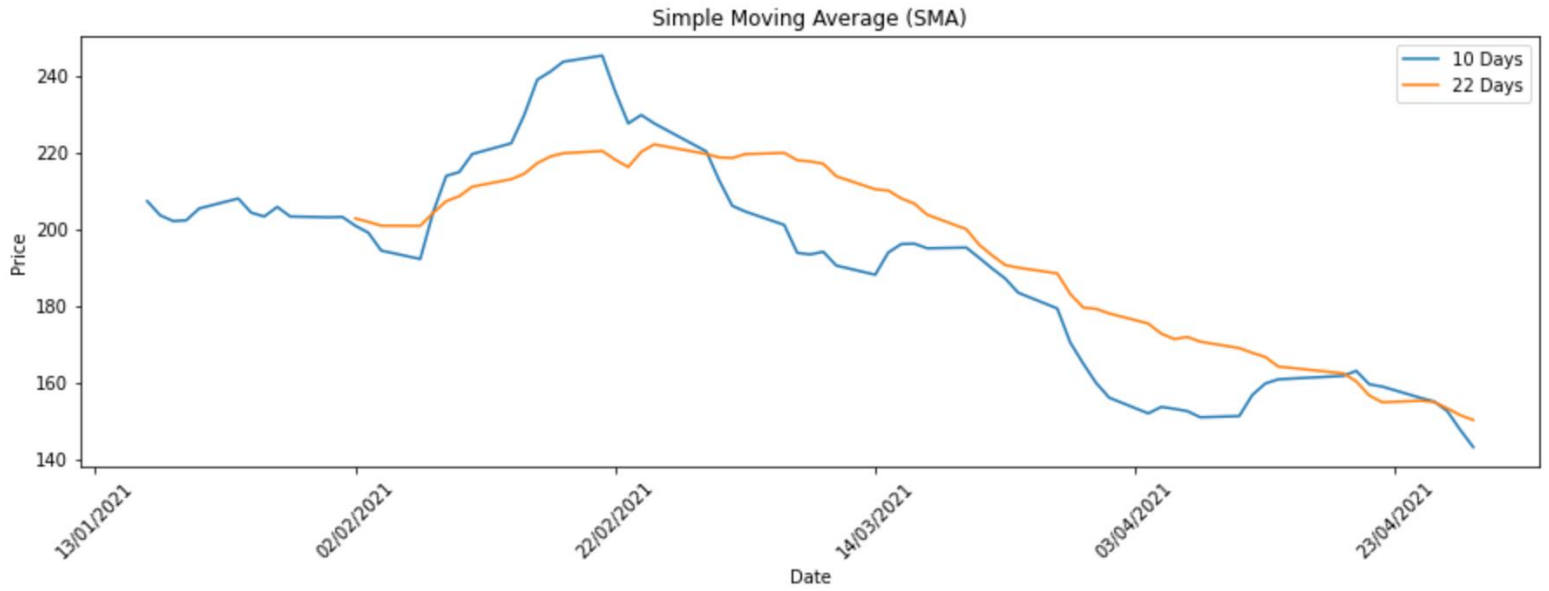
Liquidity Features:

- Median transaction price
- Interquartile range of transaction prices
- Best transaction price
  - Least expensive bid price
  - Most expensive ask price
- Transaction quantity at best price
- Median transaction quantity
- Interquartile range of transaction quantities
- Number of Transactions

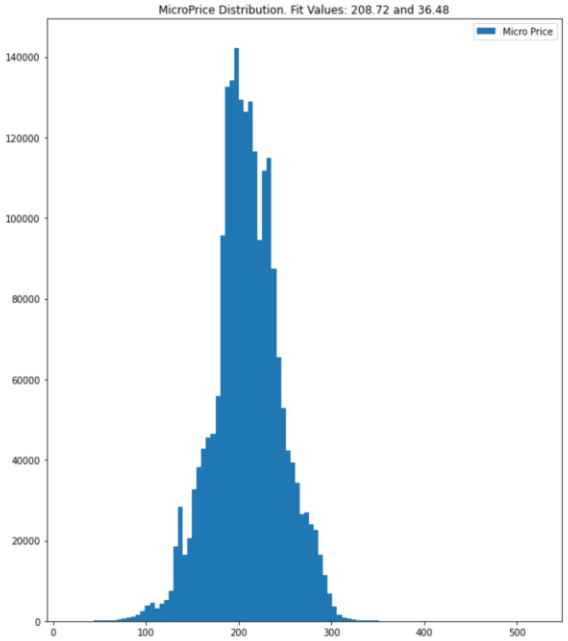

```
[
  "time",
  1.936,
  [
    "bid",
    [
      [
        183,
        4
      ],
      [
        27,
        3
      ]
    ]
  ],
  [
    "ask",
    [
      [
        739,
        5
      ]
    ]
  ]
]
```

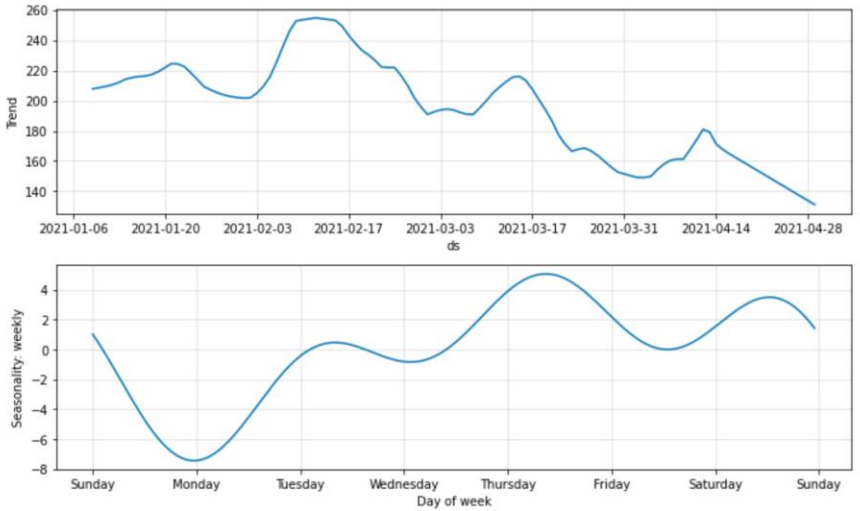
DATA EXPLORATION

- EMA and SMA moving averages
- Short term average dropping below long term average



DATA EXPLORATION

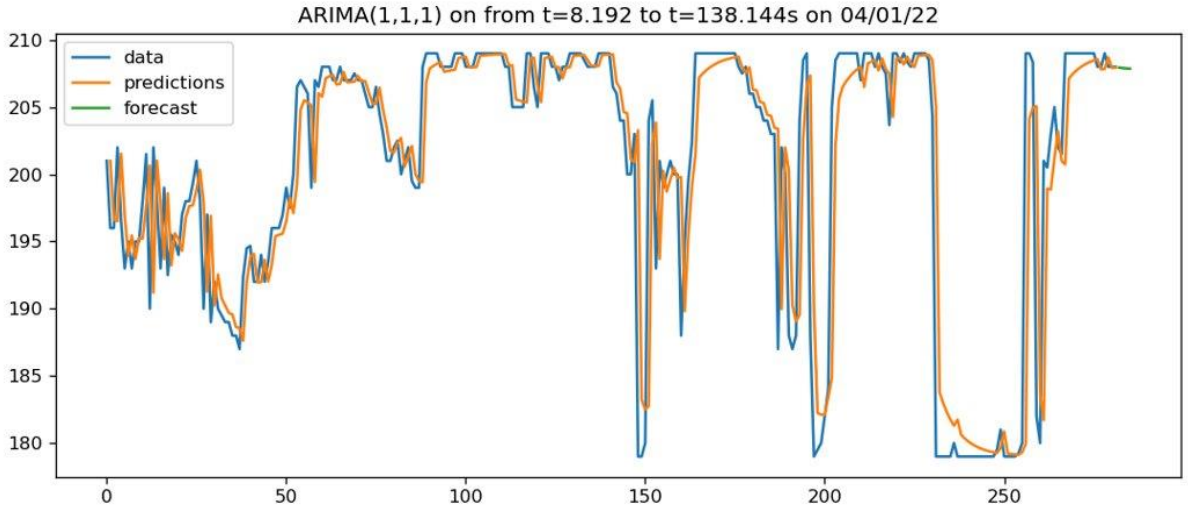


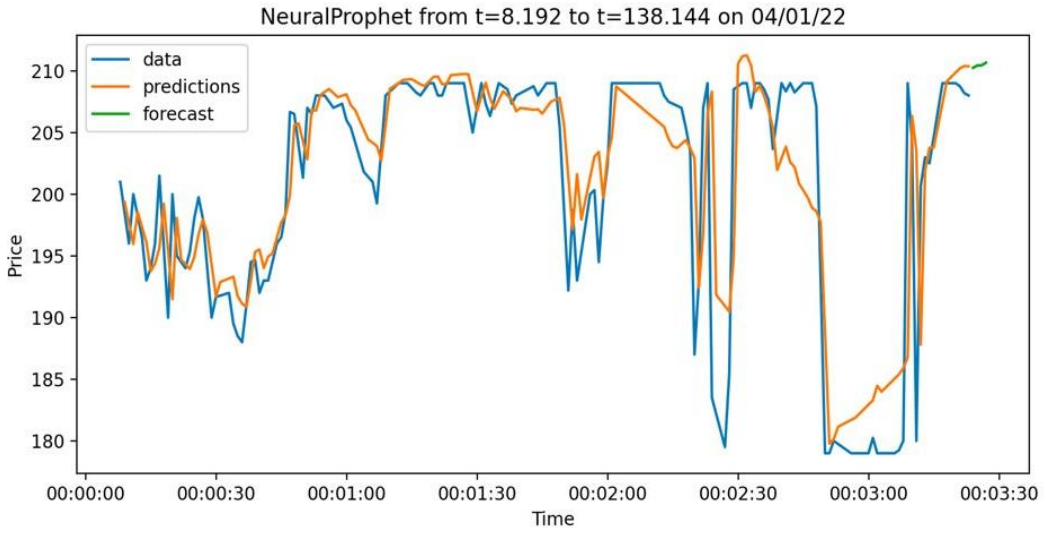


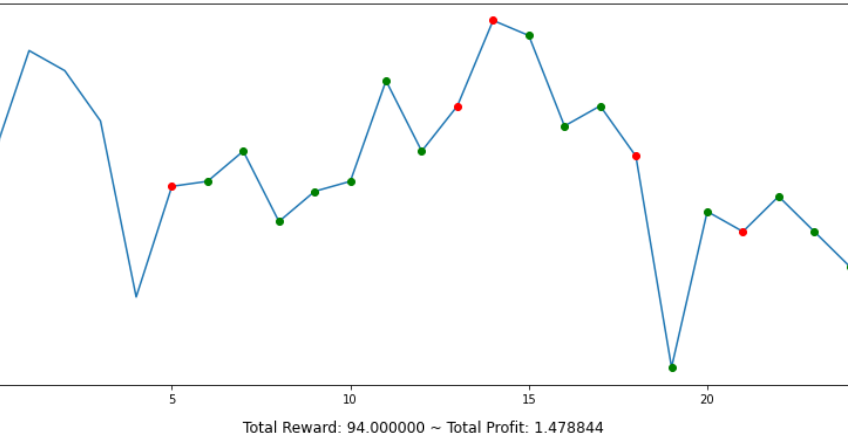
- Micro price showing multimodal Gaussian distribution
- Weekly seasonality trends derived from NeuralProphet

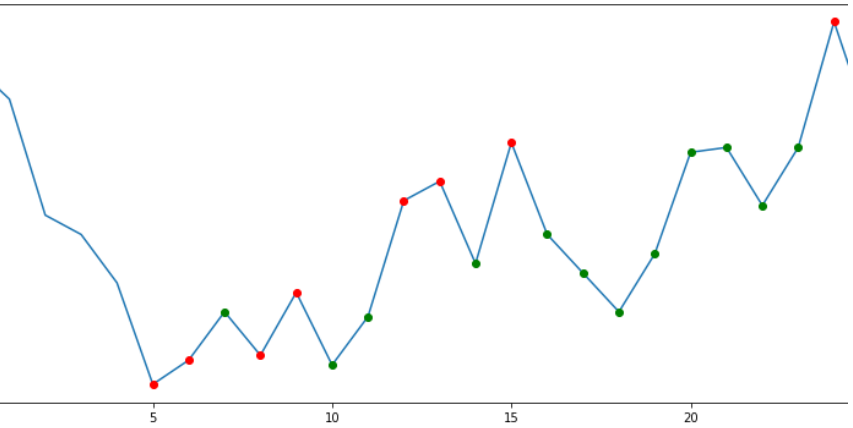
UNIVARIATE MODELS: ARIMA & NEURAL PROPHET

- Provided exploration into short-term price modelling
- Forecasting methods extremely limited within time-series algorithms
- ARIMA: ACF/PACF breaks down with N>250 datapoints









A2C

- Provides profitable strategies across the datasets
- Only considers tape data
- Trading Actions: Sell=0, Buy=1
- Trading Positions: Short=0, Long=1

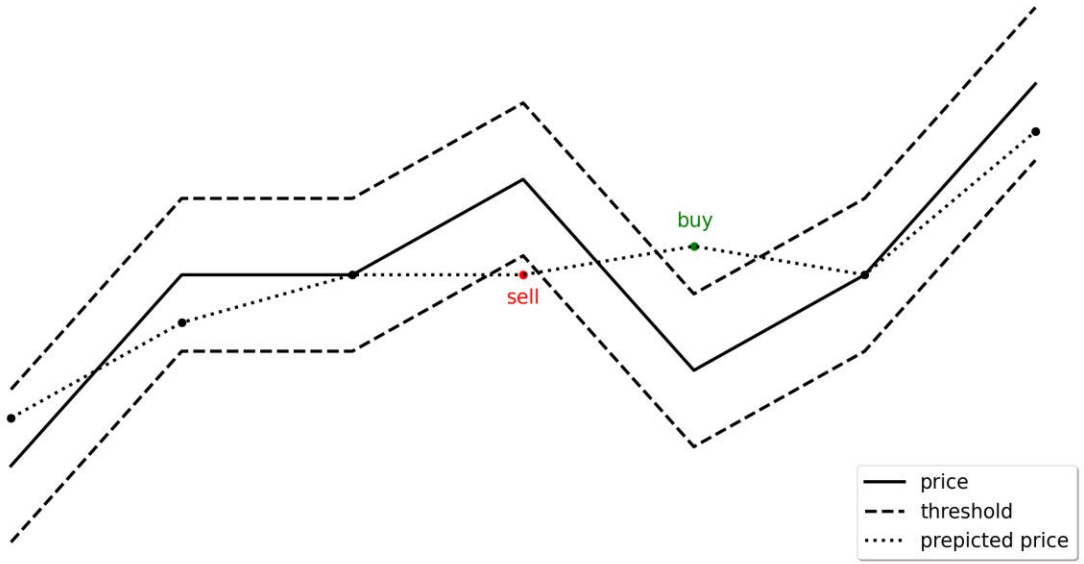
M. A. Haghpanah, "The most simple, flexible, and Comprehensive Openai Gym Trading Environment (approved by Openai Gym)," *GitHub*, 2022. [Online]. Available: https://github.com/AminHP/gym-anytrading.

TRANSACTION DECISION ALGORITHM

- Price predictions from random forest regressor:

MAE: 1.138 MSE: 8.337

- Our custom uses a random forest to predict price
- We mapped these predicted prices to transaction decisions
- The thresholds are based on the absolute error of the model

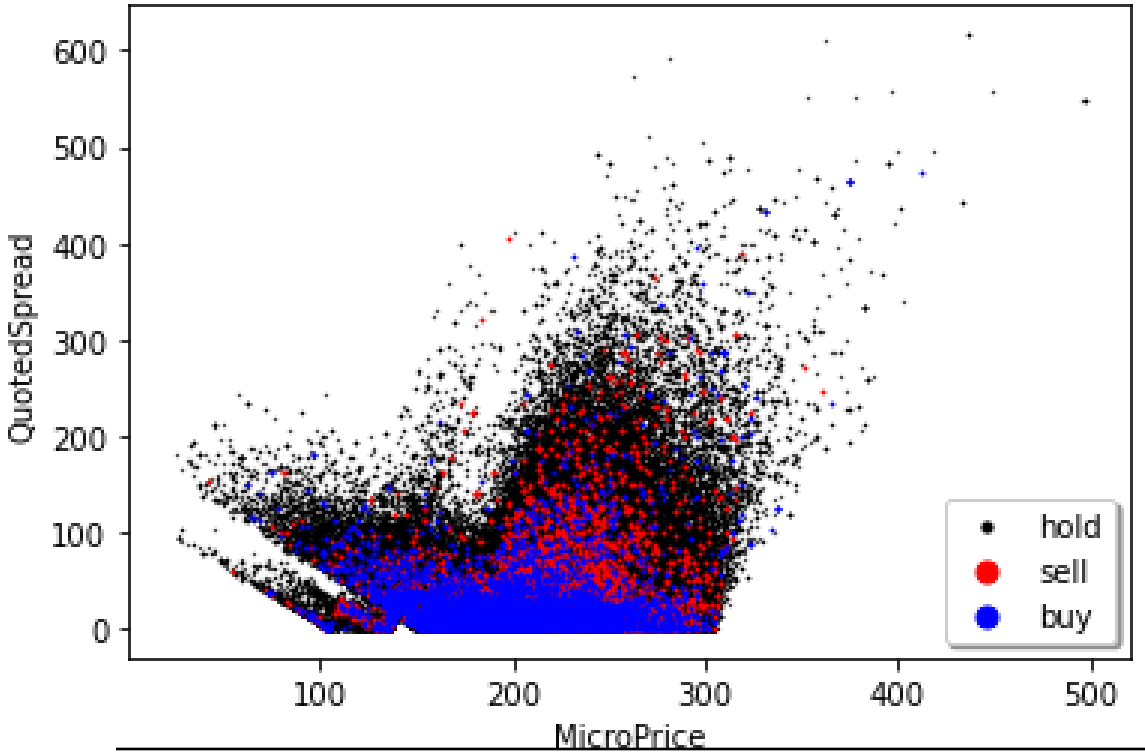


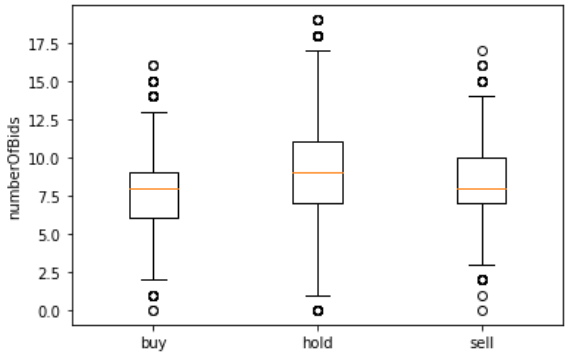
$$F(z) = \begin{cases} buy & \hat{p} > p + \mu + (z \times \sigma) \\ hold & p + \mu - (z \times \sigma) < \hat{p} < p + \mu + (z \times \sigma) \\ sell & \hat{p} < p + \mu - (z \times \sigma) \end{cases}$$

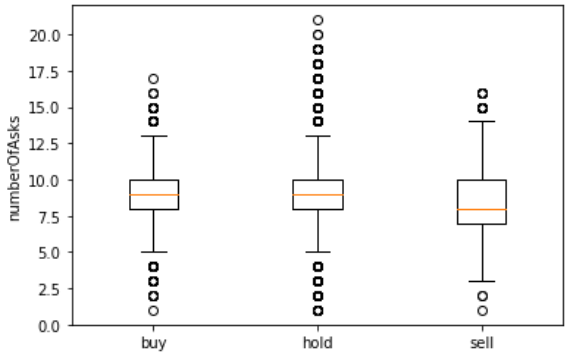
TRADING SIMULATOR

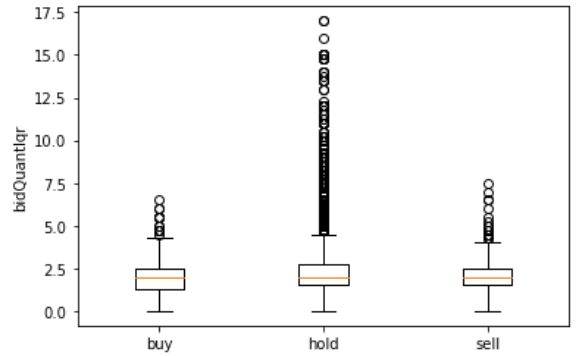
- Buy method computes the outstanding balance of money and number of shares after each buy transaction.
- Sell method computes the outstanding balance of money and number of shares after each buy transaction.
- Result is a significant profit ratio of 2.238 considering all transactions over 4 months.

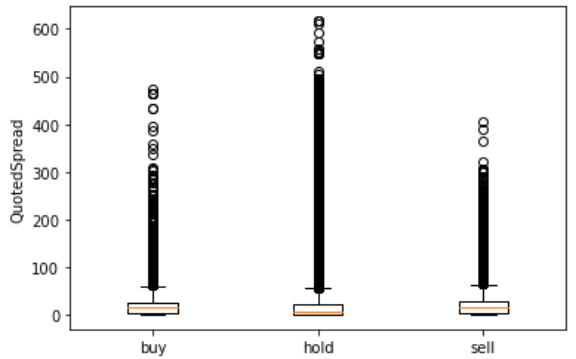
TRANSACTION TYPE PATTERNS











FURTHER WORK CONSIDERATIONS

- Adding more features to tape data for model training.
- Implement heterogeneous ensemble methods.
- Automating the end-to-end process for more than one asset.
- The transaction decision algorithm could be improved by purchasing a quantity of asset proportionally to numerous risk thresholds.