# TXF LOB mid-price prediction based on tree-based model
# 基於樹模型預測台指期限價單數據簿的中間價格

組別: B460　　組員:110061273 吳昕庭 110033241 蔡宜蒨　　指導教授 :翁詠祿

## ABSTRACT

In this dynamic environment, many market participants leverage technology to streamline financial services, including lending, insurance, investments, trading, budgeting, and more. Notably, both individual retail investors and mutual fund agencies are actively engaged in predicting stock prices within this sector to maximize their trading gains. The use of Random Forest to forecast stock prices of fintech companies highlights the application of technological innovation in the financial industry. Additionally, other tree-based models have also been employed for similar purposes. The literature review section that follows outlines key research efforts related to stock price forecasting, high-frequency data, and the use of Random Forest and other tree-based models.

## ALGORITHMS

**1.Random Forest:** Many independent trees, average results.
- Create Trees: It generates multiple decision trees (hence "forest") using different subsets of the training data.
- Averaging for Regression: The final result is the average of the trees' predictions.

**2. Decision Tree:** Single tree, splits based on features.
- Splitting: Starting from the root, the tree selecting the best attribute to split the data based on metrics like Gini impurity or entropy.
- Leaf Nodes: Each branch continues to split until it reaches a "leaf", which represents the final prediction.

$$Gini\ Impurity: Gini(D) = 1 - \sum_{i=1}^{C} {p_i}^2$$

$$Entropy: Entropy(D) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

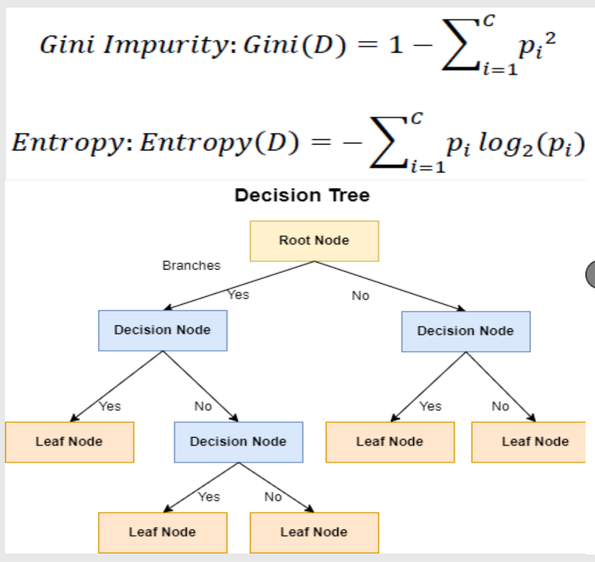


Fig.2 scheme of decision tree

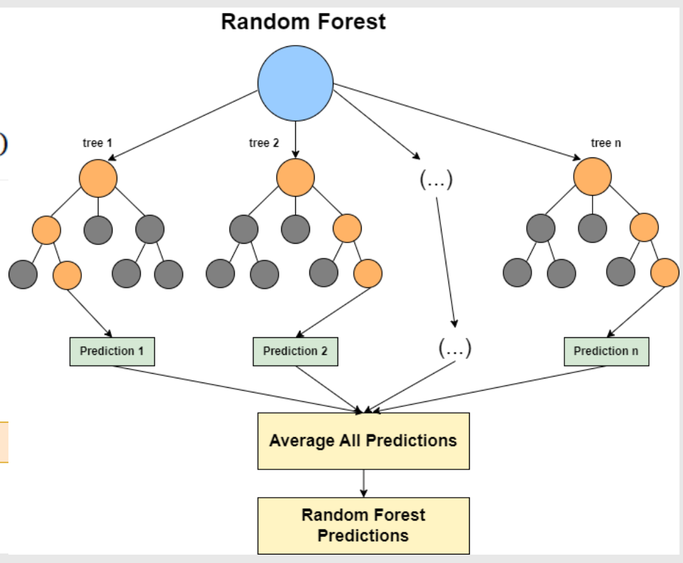


Fig.3 scheme of random forest

## PROPOSED SCHEME

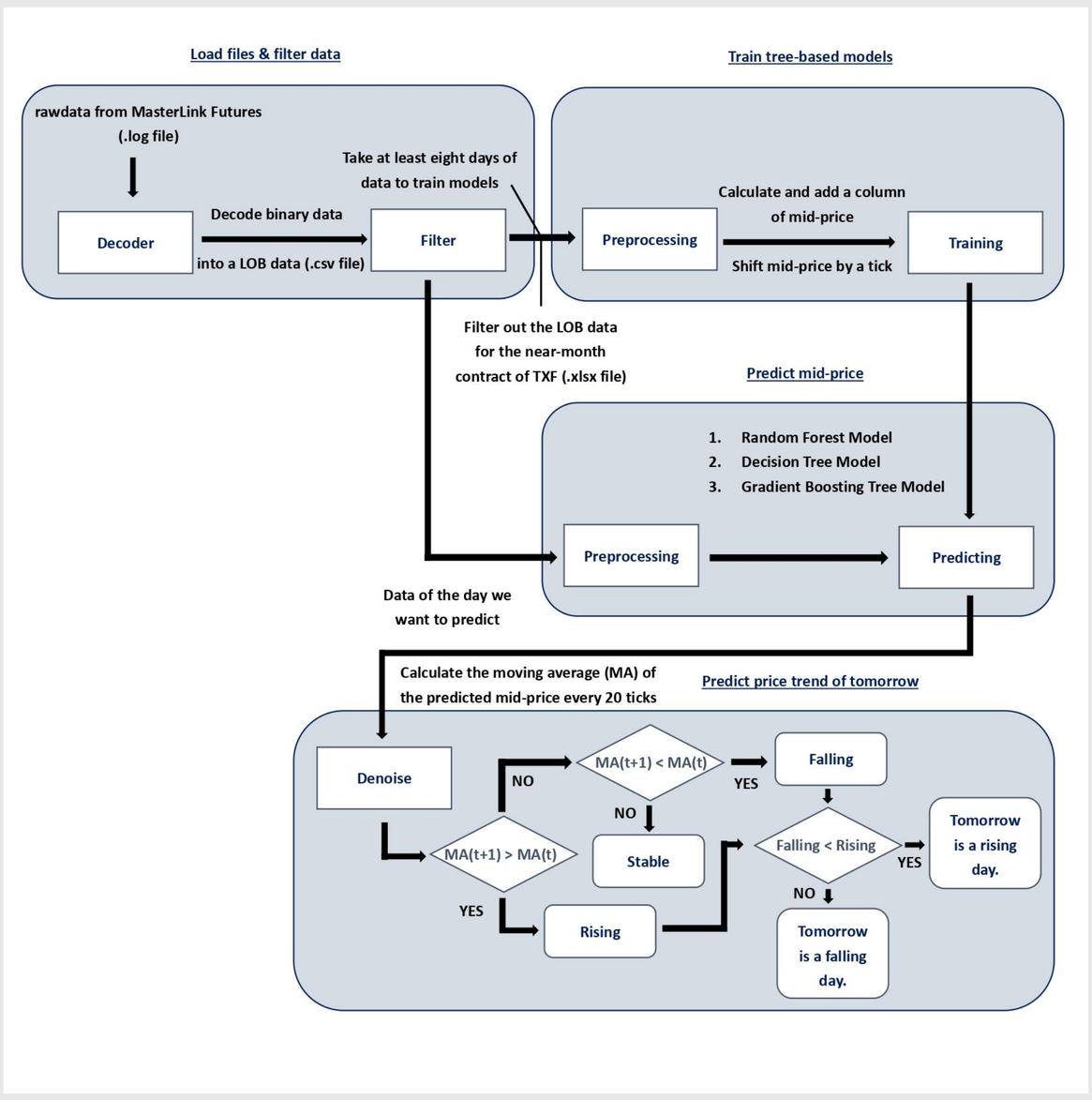


Fig.1 Flow chart

## EXPERIMENT & CONCLUSION

| Tree-based models | Random Forest | Decision Tree | Gradient Boosting Tree |
|---|---|---|---|
| RMSE | 30.79339 | 38.098049 | 64.3533157 |
| $R^2$ | 0.9999964 | 0.99999461 | 0.999968302 |
| performance | best | second | last |

Table.1 Comparison between different tree-based models

| Predicted date / Random Forest | 8/30 | 8/31 | 9/1 | 9/4 | 9/5 | 9/6 |
|---|---|---|---|---|---|---|
| RMSE | 12.002 | 4.375 | 7.635 | 4.441 | 9.756 | 4.6601 |
| $R^2$ | 0.9903 | 0.9985 | 0.9965 | 0.9984 | 0.9906 | 0.9978 |
| Compared to the conventional method, the sliding window method demonstrates a lower root mean square error. | | | | | | |

Table.2 Performance of sliding window method over one week

### Step1: Data Preprocessing

This study uses high-frequency limit order book data from Masterlink Futures for the Taiwan Futures Index (TX) contract expiring in 2023. The raw data is decoded and filtered to include only TX futures, with essential rows retained. This dataset provides features like the top five bid and ask prices and volumes for predicting mid-price changes.

### Step2: Model Training and Evaluation

Using previous tick data, we trained Random Forest, Decision Tree, and Gradient Boosting Tree models to predict the next tick's mid-price. Evaluation with RMSE and R-squared showed Random Forest as the best performer, followed by Decision Tree, with Gradient Boosting Tree performing worst.

### Step3: Enhance the Accuracy by Sliding Window Method

A sliding window approach was applied with Random Forest, using data from the past eight days to predict the next day's mid-price. Updating daily, this method outperformed static training by adapting to recent trends.

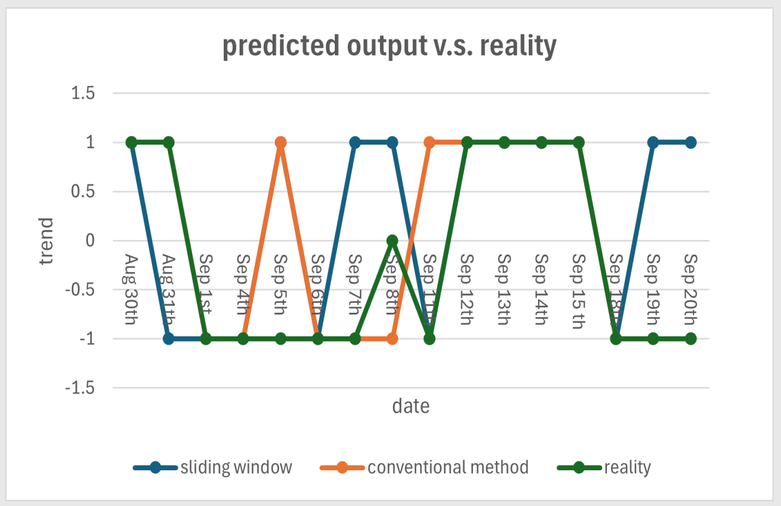


Fig.4 Trend of predicted output and reality

| | Conventional method | Sliding window method |
|---|---|---|
| precision | 62.5% | 62.5% |
| recall | 71.4% | 71.4% |
| F1-score | 66.6% | 66.6% |
| accuracy | 62.5% | 68.75% |

Table.3 Comparison methods

The conventional method achieved a precision and accuracy of 62.5%. In contrast, the sliding window method demonstrated superior performance with a precision of 62.5% and an accuracy of 68.75%, highlighting its effectiveness in enhancing predictive accuracy and reliability.