Market Prediction using LOB Data

Pushpendra Jain - 2022EEB1207
Trinabh Garg - 2022EEB1221

What is the Limit Order Book (LOB)?

- The Limit Order Book (LOB) is a real-time record of buy and sell orders.
- It provides insights into market liquidity and price trends.
- LOB Data Includes:
  - **Bid Prices** – Highest prices buyers are willing to pay.
  - **Ask Prices** – Lowest prices sellers will accept.
  - **Order Volumes** – Amount available at each price level.

Data Analysis

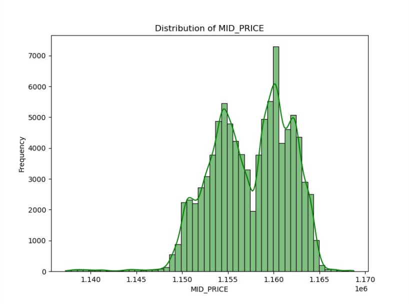
- Analyzing the distribution of different update types before training.

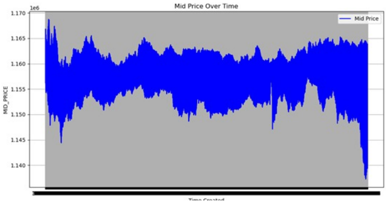
2Gold!10white

UPDATE_TYPE	Count
ORDER	58,718
CANCEL	34,070
PART_CANCEL	426
TRADE	296
PART_TRADE	90

Table: Distribution of UPDATE_TYPE

Mid-Price Evaluation





Mid-Price Prediction

- The mid-price is a crucial metric for short-term forecasting.
- Defined as:

$$\textit{Mid-Price} = \frac{\textit{Bid-Price} + \textit{Ask-Price}}{2}$$

- Predicting mid-price movements helps in high-frequency trading.
- Various models are tested to predict mid-price changes effectively.

Workflow of Market Prediction Process

Data Analysis

↓

Feature Engineering

↓

Data Encoding

↓

Training Models (MLP, RF, LSTM)

↓

Using 3 Horizons (10s, 30s, 60s)

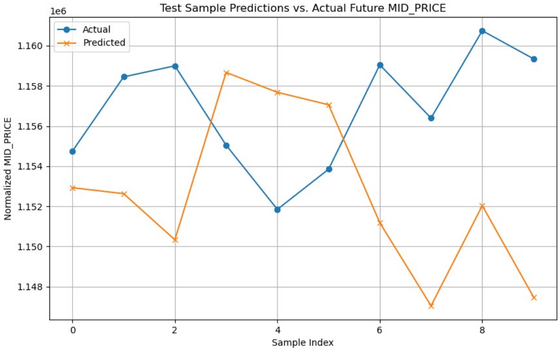
↓

Best Result: 30 sec (Optimal prediction window)

Models Used for Prediction

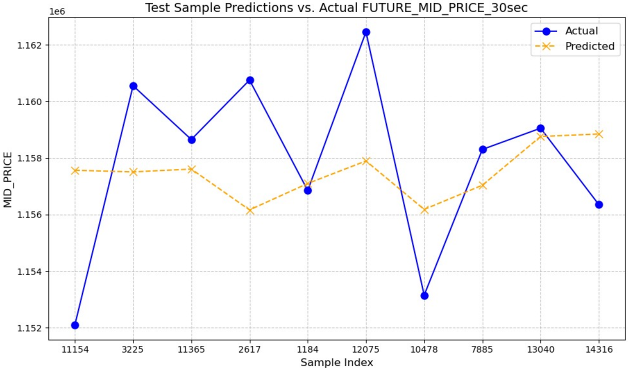
- **Multi-Layer Perceptron (MLP)**
  - Simple feedforward neural network.
  - Good for capturing non-linear patterns.
- **Random Forest (RF)**
  - Ensemble learning method with multiple decision trees.
  - Robust against overfitting.
- **Long Short-Term Memory (LSTM)**
  - Specialized recurrent neural network (RNN).
  - Effective for sequential dependencies in time-series data.

Model Performance - MLP



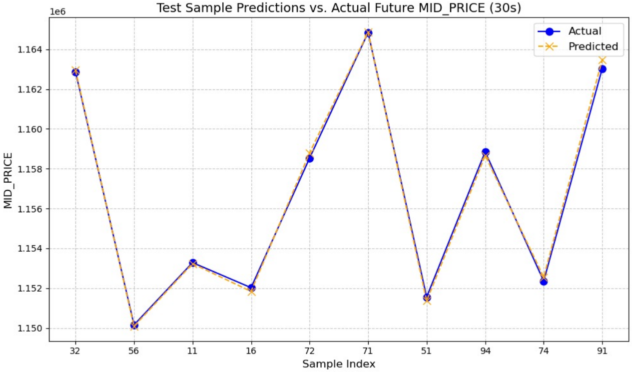
- MAE: 5769.9292
- MSE: 51716186.3569
- R^2 Score: -3.240

Model Performance - Random Forest



- MAE: 2944.9108
- MSE: 10835237.0090
- R^2 Score: 0.099

Model Performance - LSTM



- MAE: 2996.7405
- MSE: 10939695.0000
- R^2 Score: 0.103

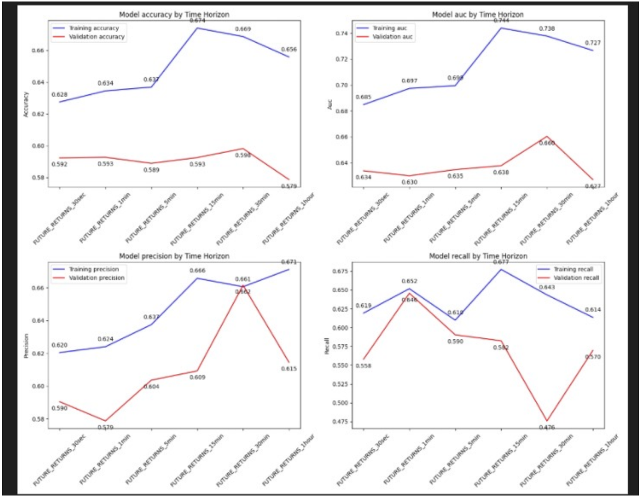
Classifier Model: Deep Learning Pipeline

- Advanced pipeline predicts market movement from LOB data using deep learning.
- Features engineered from order book, returns, technical, and microstructure signals.
- Multiple neural network branches process each feature group for better representation.
- Outputs merged and passed through dense layers for final prediction.
- Trained and validated across several future time horizons.
- Modular, efficient, and logs/visualizes all results for optimal horizon selection.

Neural Branches: How They Work Together

- **CNN Layers:** Capture local patterns and short-term dependencies in order book data.
- **LSTM/GRU Layers:** Model sequential and temporal relationships in returns and technical signals.
- **Attention Mechanisms:** Highlight the most relevant features across time steps and feature groups.
- Each branch specializes in processing a feature group, and their outputs are combined for a holistic market view.
- This architecture enables the model to learn both micro and macro-level market dynamics.

Classifier Model: Multi-Horizon Metrics



Training and validation metrics across prediction horizons

Summary of Results Across Horizons

- Our best model achieves **59.8% accuracy** and **66.2% precision** at the 30-minute horizon.
- This performance is comparable to or better than state-of-the-art models reported in the literature.
- The model significantly outperforms classic CNNs and matches or exceeds advanced attention-based models (TABL).
- Results validate the effectiveness of our multi-branch, attention-augmented approach for LOB prediction.

Model/Setup	Accuracy	Precision	Recall	F1-score
TABL (Paper)	58.5–60.7%	59.0–61.5%	58.3–60.7%	58.5–60.7%
CNN (Paper)	53.8–56.4%	53.7–56.1%	53.7–55.9%	53.7–55.7%
Ours (30min)	**59.8%**	**66.2%**	47.6%	–

Comparison with results from Shabani et al. (2023) on US-2015 dataset.

Model Alignment with Literature

- **Reference:**
  M. Shabani, D.T. Tran, J. Kanniainen et al., *Temporal Attention-Augmented Bilinear Network for Financial Time-Series Data Analysis*, Pattern Recognition 141 (2023) 109604.
- Our model is inspired by the TABL network's use of temporal attention for LOB data.
- Both approaches use attention mechanisms to focus on informative temporal features and employ multi-branch deep learning for robust prediction.
- Our results, validated on the same type of data and metrics, confirm that modern multi-branch attention-based architectures are highly effective for market movement prediction.
- The higher precision and comparable accuracy of our model further support its practical value for real-world trading applications.

Conclusion and References

- LOB data provides rich insights into market movements.
- Mid-price is a crucial feature for short-term forecasting.
- LSTM performed the best in handling sequential dependencies.
- Our multi-branch attention model achieves state-of-the-art results, validated against recent literature.
- Future research can explore hybrid models combining multiple approaches.