DATA3888 - Optiver

Optiver-07

March 26, 2025

1 Define Business Objectives

The goal is to develop a robust volatility prediction model that directly supports trading and risk management decisions. By forecasting volatility accurately, the model helps traders adjust pricing, manage risk, and allocate capital more effectively.

2 Determine Use Case

The solution will be integrated into a trading analytics platform, delivered via an interactive dashboard or web app. End-users (traders and risk managers) will receive real-time alerts and visualizations, enabling timely and informed trading decisions.

3 Evaluate Existing Solutions

Currently, models like the GARCH family are common in the industry. However, they assume linearity and may not capture the complex, nonlinear behavior of high-frequency trading data. Their limitations include:

- Sensitivity to parameter settings.
- Oversimplification of market dynamics.
- Reduced performance during turbulent periods.

Our supervised autoencoder-MLP approach aims to address these issues by learning nonlinear representations and reducing label leakage through joint training.

4 Analyze Similar Problems

Similar challenges exist in both academic research and industry practice:

- Deep learning models have been applied to capture complex market behavior.
- Hybrid approaches combining traditional econometric models with machine learning have shown improved accuracy.

These insights support our use of a hybrid supervised autoencoder-MLP framework for volatility prediction.

5 Manual Problem-Solving Approach

Manually tackling volatility prediction would involve:

1. Collecting and cleaning high-frequency trading data.
2. Computing basic metrics (e.g., bid-ask spread, weighted average price).
3. Analyzing time-series trends to spot volatility periods.
4. Calculating historical volatility using statistical formulas.

5. Comparing these measures to known market events.

This process, while informative, is impractical for real-time trading due to the volume and complexity of the data.

6 Data Exploration

Data Source: The provided dataset consists of high-frequency stock data with 31 columns (bid/ask prices, sizes, spreads, mid-prices, and rolling statistics).

Assumptions:

- The data is complete and free of missing values.
- Timestamps and bucket identifiers accurately reflect market events.
- Historical data is representative of typical market behavior.

7 Understand Data Owner Priorities

Data owners focus on:

- **Data Quality:** Accuracy and consistency.
- **Timeliness:** Ability to process data in real time.
- **Interpretability:** Clear insights from model predictions.
- **Scalability:** Efficient handling of large volumes of data.
- **Compliance:** Transparency and regulatory adherence.

8 Model Selection

Potential models for volatility prediction include:

- **GARCH Models:** Standard but limited by linearity.
- **LSTM/GRU Networks:** Good for temporal patterns.
- **Supervised Autoencoder-MLP:** Combines dimensionality reduction with classification to capture nonlinear behavior.
- **Transformer-Based Models:** Emerging option for long-term dependencies.

Our chosen approach strikes a balance between complexity and interpretability.

9 Define Performance Metrics

We propose evaluating model performance using:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Classification Accuracy and F1 Score (for multi-label tasks)
- Precision and Recall for high-volatility events
- Backtesting metrics (e.g., profitability, risk-adjusted returns)