Executive Summary: Short-Term Volatility Prediction for Stocks

Li Zhu

May 31, 2024

1 Overview

This project aims to build an efficient model to predict short-term volatility for hundreds of stocks across different sectors. It is based on the Kaggle competition Optiver Realized Volatility Prediction. Volatility reflects the degree to which a stock price moves, usually defined by the standard deviation of the stock log returns over a one-year period. Volatility studies are extremely useful for short-term trading and intraday derivatives trading. Scalpers and day traders use volatility to trade in options, as both option buyers and writers expect high volatility for better returns over time.

2 Modeling Approach

The dataset contains six files with hundreds of millions of rows of highly granular financial data. To better understand the data, we first perform Exploratory Data Analysis (EDA) to investigate the datasets and summarize their main characteristics. We then calculate fundamental statistics to identify connections between features.

We begin by setting up a baseline model using the overall target mean value as the prediction for the next 10 minutes of volatility, achieving a Root Mean Square Percentage Error (RMSPE) of 1.110330. We gradually improve the RMSPE by using the target mean and median over `stock_id`, the realized volatility from trade book prices, and the realized volatility from WAP1/WAP2 from the order book.

During this process, we add more features and investigate their correlations with the target variable. We select the three most correlated features — wap1_realized_volatility, wap2_realized_volatility, and book_bid_ask_price_ratio_realized_volatility — to build linear regression and K Nearest Neighbors models. Ultimately, we use ensemble learning approaches, such as boosting, with a particular focus on the powerful gradient boosting package XGBoost.

3 Conclusion and Future Work

XGBoost demonstrated the best performance with an RMSPE value of 0.028044, significantly improving the prediction model. However, we need to verify the potential overfitting issue. Future work will involve investigating more efficient models, such as Graph Neural Networks, to further enhance volatility prediction.