Replication and Extension of Baseline Models for Limit Order Book Prediction

Project Report

Mustapha Aziz Laroussi

September 16, 2025

Abstract

This report documents my work on replicating and extending the benchmark study by Ntakarís et al. (2020) on short-term price forecasting using Limit Order Book (LOB) data. The original paper introduced Ridge Regression and Single-Layer Feedforward Neural Networks (SLFN) as transparent baselines. In this project, I faithfully replicated their methodology, including the dataset structure, normalization strategies, and fold-based evaluation. I then extended the benchmark by introducing **Extreme Gradient Boosting (XGBoost)**, a modern ensemble method. The results confirm the reproducibility of the original baselines while demonstrating that XGBoost consistently outperforms them across prediction horizons. I also discuss why this superiority can be misleading if interpreted without caution, as benchmark accuracy does not automatically translate into reliable trading performance.

1 Introduction

Financial markets are increasingly shaped by algorithmic trading systems, where decisions are made in fractions of a second and are often informed by market microstructure. The Limit Order Book (LOB), which records all buy and sell orders across multiple price levels, represents one of the richest datasets for studying price formation. Anticipating short-term price movements from LOB data can provide a measurable edge in execution strategies, risk management, and market making.

The study by Ntakarís et al. (2020) introduced a benchmark dataset for mid-price forecasting, designed to support reproducibility and comparability of machine learning methods in this domain. The authors evaluated two baselines: Ridge Regression, representing a linear model, and a Single-Layer Feedforward Neural Network (SLFN) with

radial basis function activations, representing a nonlinear approach. These baselines were never meant to be state-of-the-art, but to establish a foundation against which future work could be measured.

My objective in this project was twofold. First, to replicate the experimental pipeline of Ntakarís et al. (2020), ensuring that the Ridge and SLFN results could be reproduced. Second, to extend the benchmark by introducing Extreme Gradient Boosting (XGBoost), a modern ensemble method that has consistently shown competitive performance on structured tabular data. By comparing Ridge, SLFN, and XGBoost side by side, I evaluate both the reproducibility of the benchmark and the value of extending it with ensemble learning.

2 Literature Review

The task of predicting short-term price movements from LOB data lies at the intersection of market microstructure theory and machine learning. Market microstructure emphasizes how order flow, liquidity, and depth influence price discovery, while machine learning offers statistical tools for exploiting high-dimensional feature spaces.

Ntakarís et al. (2020) provided one of the first systematic baselines for this problem. The dataset consisted of ten trading days for five different stocks, processed into training and testing folds with 144 engineered features and label horizons of 1, 2, 3, 5, and 10 events. Two baselines were proposed:

- **Ridge Regression (RR)**: a linear classifier with ℓ_2 regularization. It is simple, interpretable, and robust to collinearity but cannot capture nonlinear interactions.
- **SLFN with RBF activations**: a shallow neural network with nonlinear hidden units constructed using k-means clustering. It provides some flexibility but is limited by its shallow architecture and sensitivity to hyperparameters.

While these baselines are transparent and computationally efficient, they are not powerful by modern standards. Recent literature has explored deep neural networks (CNNs, LSTMs, Transformers) to capture temporal dynamics in LOB data. Ensemble methods such as Gradient Boosting and XGBoost, though less studied in this context, are known for their strength on structured tabular data, which motivated my extension.

3 Dataset and Preprocessing

The dataset of Ntakarís et al. (2020) includes 144 engineered features per sample, derived from the top ten bid and ask levels of the LOB. Labels are defined at five horizons

$(h = 1, 2, 3, 5, 10)$, each discretized into three classes:

$$y \in \{1 = \text{up},\ 2 = \text{flat},\ 3 = \text{down}\}.$$

Nine folds were constructed using anchored forward splitting of trading days, allowing robust out-of-sample testing. Normalization strategies included z-score, min-max scaling, and decimal precision scaling. Both auction and no-auction variants were considered, though the results shown here focus on the z-score normalized auction data for fold 1.

4 Methodology

The models I evaluated follow the benchmark structure but with the addition of XGBoost.

4.1 Ridge Regression

Ridge regression solves a regularized least squares problem:

$$\hat{W} = \arg\min_{W} \ \|XW - T\|^2 + \alpha\|W\|^2,$$

where X is the feature matrix, T the one-vs-rest target encoding, and α the regularization parameter.

4.2 SLFN with RBF Kernels

The SLFN uses hidden nodes computed as:

$$h_j(x) = \exp\left(-\frac{\|x - c_j\|^2}{2\sigma^2}\right),$$

with centers c_j obtained via k-means clustering. Outputs are obtained via a linear layer:

$$\hat{y} = \arg\max_{k} \big(H(x)W_k\big).$$

4.3 XGBoost

XGBoost constructs an additive ensemble of decision trees:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), \quad f_m \in \mathcal{F},$$

with a regularized objective:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_m \Omega(f_m),$$

where $\Omega(f)$ penalizes complexity. Boosting allows subsequent trees to correct the errors of prior ones, resulting in strong predictive performance.

5 Experimental Results

I ran experiments on fold 1 (auction, z-score normalization), evaluating Ridge, SLFN, and XGBoost across horizons. Performance was measured using Accuracy, Precision, Recall, and F1-score (macro-averaged).

Table 1: Comparison of Ridge, SLFN, and XGBoost across horizons (Fold 1, z-score).

Model	**Acc**	**P**	**R**	**F1**
Horizon 1				
Ridge	0.424	0.420	0.443	0.390
SLFN	0.643	0.433	0.350	0.306
XGBoost	**0.663**	**0.530**	**0.448**	**0.461**
Horizon 2				
Ridge	0.467	0.446	0.453	0.442
SLFN	0.555	0.493	0.408	0.387
XGBoost	**0.714**	**0.689**	**0.630**	**0.647**
Horizon 3				
Ridge	0.442	0.440	0.441	0.436
SLFN	0.478	0.451	0.426	0.414
XGBoost	**0.609**	**0.593**	**0.585**	**0.587**
Horizon 5				
Ridge	0.438	0.439	0.439	0.436
SLFN	0.442	0.443	0.444	0.442
XGBoost	**0.523**	**0.535**	**0.520**	**0.516**
Horizon 10				
Ridge	0.458	0.442	0.444	0.436
SLFN	0.492	0.464	0.444	0.444
XGBoost	**0.537**	**0.533**	**0.454**	**0.430**

XGBoost consistently outperformed Ridge and SLFN across horizons, with the strongest gains at horizon 2.

Cautionary Note: While XGBoost appears dominant, this can be misleading. Its flexibility makes it powerful on benchmarks, but also prone to overfitting if hyperparameters are not tuned. Moreover, unlike sequential models, XGBoost does not explicitly capture temporal dependencies in LOB data. Thus, higher accuracy here does not guarantee profitability in live trading, where latency, execution costs, and regime changes matter.

6 Discussion

The experiments confirm the reproducibility of the baselines proposed by Ntakarís et al. (2020). Ridge regression served as a minimal linear benchmark, while the SLFN offered some nonlinear capacity but remained unstable. XGBoost, by contrast, delivered stronger predictive performance across all horizons.

However, there are important caveats. Ridge and SLFN were intentionally simple to ensure reproducibility, not to maximize accuracy. XGBoost is more expressive, but also less interpretable and not designed to exploit sequential structure. The results demonstrate that ensemble learning is a valuable addition to the benchmark, but they should not be mistaken for a final solution to the problem of LOB forecasting.

7 Conclusion

In this project I replicated the baseline models of Ntakarís et al. (2020), confirming their reproducibility. I then extended the benchmark with XGBoost, demonstrating consistent improvements across horizons. The findings highlight both the importance of reproducible baselines and the value of ensemble methods. Future work should include testing across all folds, exploring sequential deep learning models, and evaluating performance with real trading metrics.

References

Ntakarís, A., Magris, M., Kanniainen, J., Gabbouj, M., and Iosifidis, A. (2020). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*.