Executive Summary

Investors in the financial market have long been interested in predicting the future prices of equities or financial time series in general. This task is challenging as the data exhibits non-stationary and non-linear relationships between the future price and its predictors. Notably, the use of trading volume as a predictor has been extensively studied within the financial time-series forecasting field (Karpoff 1987).

The growth of computing power and available datasets has accelerated the use of machine learning in academia. Among the various supervised machine learning algorithms, gradient boosting has demonstrated its effectiveness across a diverse set of applications (T. Chen and Guestrin 2016). Consequently, the use of machine learning in combination with trading volume features to predict future prices has been extensively explored.

In the financial industry, the volume profile has historically been used by technical analysts to find support and resistance levels of stock prices. The volume profile of an asset over a period can be seen as a histogram-like shape, where the x-axis shows the prices at which the asset traded in that period, and the y-axis shows the total trade volume at that price in that period. Although trading volume has been extensively studied, the use of the volume profile in predicting future expected returns has not. In this paper, we fill a gap in the existing return predictability literature between high-frequency trading using order book data and low-frequency trading using open, high, low, close, and volume (OHLCV) data. This was achieved by utilizing the volume profile, which possesses more favorable properties than order book data. However, the volume profile is considered noisier than the OHLCV data.

The main contribution of this paper has been to build a methodology from the ground up to use the volume profile in a modern machine-learning context. The methodology consists of first defining the volume profile in a mathematical context, after which the volume profile is preprocessed to make it suitable for machine learning. Following the preprocessing steps, we treated the volume profile as a probability mass function of a discrete probability distribution. This allowed us to utilize the full range of probability theory and statistics to generate summary statistics of the volume profile, which were used as features. These features were used in an Extreme Gradient Boosting (XGBoost) model