Conformal Prediction for Reliable Stock Selections

Ipek Kaya IPEK.KAYA.2025@LIVE.RHUL.AC.UK
Royal Holloway University of London, Surrey, United Kingom
Khuong An Nguyen KHUONG.NGUYEN@RHUL.AC.UK
Royal Holloway University of London, Surrey, United Kingdom

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

A major challenge in quantitative finance is not just predicting which stocks will outperform but quantifying the uncertainty and reliability of those predictions. This is critical because financial markets are inherently noisy, volatile, and affected by countless unpredictable factors, meaning that even the best models can be dramatically wrong (Virgilio and Paz López (2024)). Reliable measures of uncertainty are essential for risk-aware investment decisions: they help portfolio managers judge when to trust a prediction, size positions appropriately, and avoid overconfidence that can lead to costly losses.

Currently, most machine learning approaches for stock selection produce only point predictions (Gu et al. (2020)), offering no meaningful measure of confidence, which limits their practical value for investors who need to manage risk. Thus, in this paper, we benchmark classical and deep learning models for US stock selection (Fu et al. (2018)), and apply conformal prediction (CP) to generate well-calibrated prediction sets. Across all models, CP achieves empirical coverage closely matching the nominal confidence level, with most prediction sets being singletons.

1. Methodology

While many practitioners approached stock selection by predicting raw returns, applying arbitrary thresholds, or optimising factor signals such as momentum or value, these methods often overlook the importance of risk-adjusted metrics (Gu et al. (2020)). For this reason, we rank all stocks by their return-to-volatility ratio (Shin and Kim (2025)), which gives us a robust measure widely used to identify likely outperformers on a risk-normalised basis.

At each time step, we rank all stocks by their return-to-volatility ratio over a forward window. Without loss of generality, the return-to-volatility ratio is computed as:

$$S_i(t;h) = \frac{P_i(t+h) - P_i(t)}{P_i(t)} \Bigg/ \sqrt{\frac{1}{h-1}\sum_{k=1}^{h}\left(\ln\frac{P_i(t+k)}{P_i(t+k-1)} - \frac{1}{h}\sum_{\ell=1}^{h}\ln\frac{P_i(t+\ell)}{P_i(t+\ell-1)}\right)^2}$$

where $P_i(t)$ is the price of stock i at time t, and h is the forward window length.

After calculating $S_i(t;h)$ for all stocks at each time t, we sort the stocks by this ratio. We assign a label $y_i(t) = +1$ (outperformer) to stocks in the top q quantile, and $y_i(t) = -1$ (underperformer) to stocks in the bottom q quantile; stocks in the middle are discarded. This 'tail and head' labeling approach focuses on clear outliers and reduces ambiguity (Fu et al. (2018)). Formally, for quantile threshold $q \in (0, 0.5)$:

$$y_i(t) = \begin{cases} +1, & \text{if } S_i(t;h) \text{ in top } q\text{-quantile} \\ -1, & \text{if } S_i(t;h) \text{ in bottom } q\text{-quantile} \end{cases}$$

Each data sample thus consists of a feature vector $x_i(t)$, representing 17 technical and fundamental indicators (calculated only from data available up to t), and the binary label $y_i(t) \in \{+1, -1\}$ derived as above. To avoid overfitting, we use a Genetic Algorithm (GA) for feature selection, which searches for optimal subsets of features by maximising model performance on a validation set (Fu et al. (2018)).

We then train four classification models: Logistic Regression (LR), Random Forest (RF), Deep Neural Network (DNN), and a Stacking ensemble, on these labeled samples, i.e., using the dataset $\{(x_i, y_i) : y_i \in \{+1, -1\}\}$. These models represent a range of capacities from linear to highly non-linear, and the inclusion of a stacking ensemble allows us to assess the robustness of conformal prediction across diverse model types.

We then apply Conformal Prediction (CP) to each model's probabilistic outputs to address a key limitation of traditional point predictions in finance (Shafer and Vovk (2008)). After fitting a probabilistic classifier $\hat{p}(y \mid x)$, we define for each labeled calibration sample (x_i, y_i) a nonconformity score (the probability-inverse nonconformity score): $\alpha_i = 1 - \hat{p}\big(y_i \mid x_i\big)$. For a new sample x_{n+1} and each candidate label $y \in \{+1, -1\}$, we compute $\alpha_{n+1}(y) = 1 - \hat{p}\big(y \mid x_{n+1}\big)$ and the CP p-value: $p(y) = \frac{\#\{\, i=1,\ldots,n : \alpha_i \geq \alpha_{n+1}(y)\,\}+1}{n+1}$. The conformal prediction set at significance level ε is $\Gamma_{1-\varepsilon}\big(x_{n+1}\big) = \{\, y : p(y) > \varepsilon \,\}$.

2. Empirical Results

We curated a dataset comprising historical data from 764 US stocks, each described by 17 technical and fundamental features (such as price-to-earnings ratio, momentum indicators, and trading volume), covering the period from 03-01-2022 to 01-01-2024. The sample includes both large-cap and mid-cap US companies drawn from the S&P 500 and S&P 400 indices. This diversity ensures the findings are representative and generalisable across a broad range of real-world investment scenarios. The historical stock data was randomly split into training and test sets, with 30% reserved for testing.

Figure 1 demonstrates both the empirical error rate and the average prediction set size for all models. As ε increases, the error rates for all models rise accordingly, confirming the validity of CP. At the same time, the average prediction set size drops rapidly and approaches singleton for moderate values of ε.

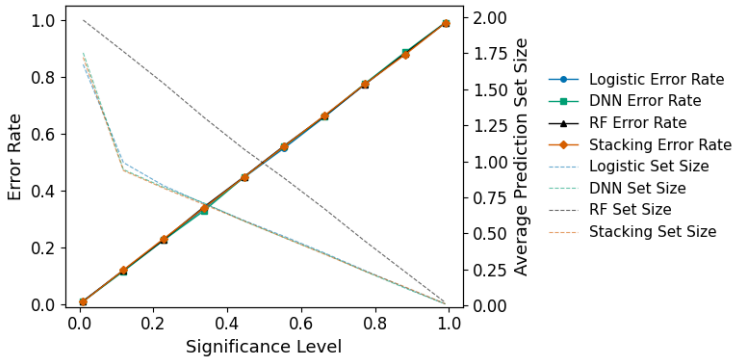

Figure 1: Empirical error rate and average prediction set size for all 4 models.

References

XingYu Fu, JinHong Du, YiFeng Guo, MingWen Liu, Tao Dong, and XiuWen Duan. A machine learning framework for stock selection. *arXiv preprint arXiv:1806.01743*, 2018.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Jungcheol Shin and Daehwan Kim. Active style drift and mutual fund performance. *Finance Research Letters*, 81:107498, 2025. doi: 10.1016/j.frl.2025.107498.

Gianluca P. M. Virgilio and Manuel Ernesto Paz López. Revisiting noise—fischer black's noise at the time of high-frequency trading. *Risk Management*, 26, 2024. doi: 10.1057/s41283-024-00151-7.