Machine Learning Project

Will Stocks Outperform the S&P500 next year ?

Alexandre Zaza, Pierre-Alexandre Crouzet, Simon Gerdolle, Leonardo Gonnelli, Rinor Abazi, Antoine Reil

Supervisor: Dr. Despoina Makariou

October 2025

Course: 5,155,1.00 Machine Learning in Finance

University of St. Gallen

This paper was written as a part of the final project of the course "Machine Learning in Finance" at the University of St. Gallen (HSG).

Abstract

Forecasting stock performance relative to market benchmarks remains a central challenge in empirical finance. This study examines whether company-level market and fundamental indicators can predict the probability that a stock will outperform the S&P 500 in the following year.

A balanced panel of 50 large-cap U.S. equities from 2016–2025 is constructed by combining market data with fundamental ratios. Three supervised learning models, Decision Tree, Random Forest, and Neural Network, are trained and evaluated using cross-validation and a holdout test set, with the Area Under the ROC Curve (AUC) as the main performance metric.

Results indicate that the Random Forest model achieves the highest predictive power (AUC $\approx$ 0.72), outperforming both the Decision Tree and Neural Network while maintaining interpretability through feature importance analysis. The findings underscore the potential of ensemble methods for financial prediction while highlighting limitations related to data quality and market noise.

Keywords – Machine Learning, Random Forest, Neural Networks, Stock Prediction, S&P 500

Contents

List of Figures

List of Tables

1 Background Research

Financial markets provide a structured and highly competitive environment in which investors allocate capital and seek excess returns. The central question in empirical finance is whether firm characteristics can systematically predict future performance relative to the market benchmark. Traditional linear factor models such as the Capital Asset Pricing Model (CAPM) (Sharpe, 1964) and the Fama–French multifactor models (Fama & French, 1992, 2015) explain average returns through exposures to a limited set of risk factors. However, their explanatory power remains modest, and numerous studies have documented persistent anomalies that appear inconsistent with the Efficient Market Hypothesis (EMH) (Cochrane, 2011; Fama, 1970).

Recent advances in machine learning (ML) provide new tools to capture nonlinear relations and higher-order interactions among financial variables. Unlike traditional econometric models, ML algorithms can flexibly learn complex mappings from firm-level features to future returns, potentially uncovering previously hidden predictive patterns. These methods have been successfully applied in various financial domains, portfolio optimization, risk forecasting, and equity selection, offering both improved predictive accuracy and robustness to noise.

In this study, we investigate whether a combination of firm-specific market and accounting indicators can predict if a stock will outperform the S&P 500 index in the subsequent year. We frame the task as a binary classification problem and develop three supervised learning models of increasing complexity: a pruned Decision Tree, a Random Forest with tuned hyperparameters, and a feed-forward Neural Network. By comparing their predictive and interpretative performance, we aim to evaluate the trade-off between model flexibility, stability, and explainability in financial prediction.

1.1 Related Work

The literature on machine learning in finance has grown rapidly in recent years, particularly in applications such as bankruptcy prediction, portfolio management, and stock selection (Ahmed et al., 2022; Ghosh et al., 2022). Early studies applied classical algorithms such as Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), and Logistic

Regression to predict stock price movements. Novak and Velušček (2016) found that SVMs outperform traditional econometric models, while Chen and Ge (2021) demonstrated that neural networks can enhance portfolio optimization by capturing nonlinear dependencies among factors.

More recent research has focused on ensemble and deep learning techniques. Ghosh et al. (2022) used hybrid architectures combining LSTM networks and Random Forests to predict directional movements of U.S. equities, whereas Heo and Yang (2016) and Caparrini et al. (2024) employed tree-based algorithms for stock selection using financial ratios. Similarly, Htun et al. (2024) and Campisi et al. (2024) applied Random Forests and Neural Networks to forecast relative returns and market direction, respectively, reporting consistent improvements in AUC and accuracy over traditional models.

Despite these advances, most prior studies rely on high-frequency or daily data, where signals are often dominated by noise and short-term volatility. Few have explored an annual, factor-based framework integrating both market-derived and accounting fundamentals to assess long-term performance potential. Our study addresses this gap by combining these complementary data sources to predict annual stock outperformance relative to the S&P 500 benchmark. By evaluating three representative algorithms under a unified cross-validation design, we contribute to the understanding of how model complexity and interpretability interact in practical financial forecasting.

2 Data

2.1 Universe of Firms

The analysis focuses on a fixed universe of 50 large-cap U.S. equities (e.g., Apple, Microsoft, NVIDIA), listed in Appendix A. This "top-50" sample comprises highly liquid firms with extensive historical coverage, ensuring data completeness and minimizing missing observations often encountered among smaller companies. Restricting the study to this stable set reduces survivorship and listing biases while preserving sufficient cross-sectional variation for model estimation. Consequently, the findings pertain to a representative subset of persistent S&P 500 constituents rather than to the index's evolving composition.

2.2 Data Sources and Coverage

Two complementary data sources were employed to capture both market dynamics and firm-level fundamentals. Yahoo Finance provides daily adjusted prices, trading volumes, and benchmark returns for the S&P 500. Adjusted prices ensure continuity across corporate actions (e.g., dividends and stock splits), following standard empirical-finance practice (Fama & French, 1992).

LSEG (Refinitiv) supplies annual accounting data, including profitability ratios, valuation metrics, and dividend yields. The sample covers the 2016–2025 period, corresponding to full overlap between the two data sources (data restriction from LSEG prior to 2016). Historical daily prices from 2010 onward are also incorporated to compute lagged technical indicators such as volatility, momentum, and beta.

2.3 Data Construction and Processing

Raw datasets from Yahoo Finance and Refinitiv were processed and aligned in R to construct a consistent panel suitable for supervised learning. Company tickers and calendar-year identifiers serve as matching keys. Because market data are observed daily while fundamentals are reported annually, the former were collapsed to an annual frequency by retaining year-end values. This transformation ensures that each firm contributes exactly one observation per year, aligning all predictors on a common temporal scale and removing high-frequency market noise. This procedure also effectively converts the time series from a high-frequency panel into a yearly cross-sectional panel.

2.3.1 Market-Based Indicators

To capture market-related dynamics, a set of technical indicators was derived from daily adjusted prices and returns. These variables summarize recent performance, risk exposure, and return distributional properties, dimensions known to influence subsequent stock behavior in asset-pricing research.

Momentum indicators were computed over 6- and 12-month horizons to measure short-term price persistence. Realized volatility and rolling betas capture total and systematic risk, while idiosyncratic volatility isolates firm-specific uncertainty. Skewness and kurtosis characterize deviations from normality, reflecting asymmetry and tail risk in return distributions. The maximum drawdown quantifies the largest cumulative loss over a one-year window, providing a measure of downside exposure. Liquidity was proxied by trading-volume–based turnover, and the logarithm of the year-end price was included as a scale control. Together, these indicators provide a comprehensive depiction of each firm's market behavior. Detailed computational formulas and window definitions are reported in Appendix B.

2.3.2 Fundamental Indicators

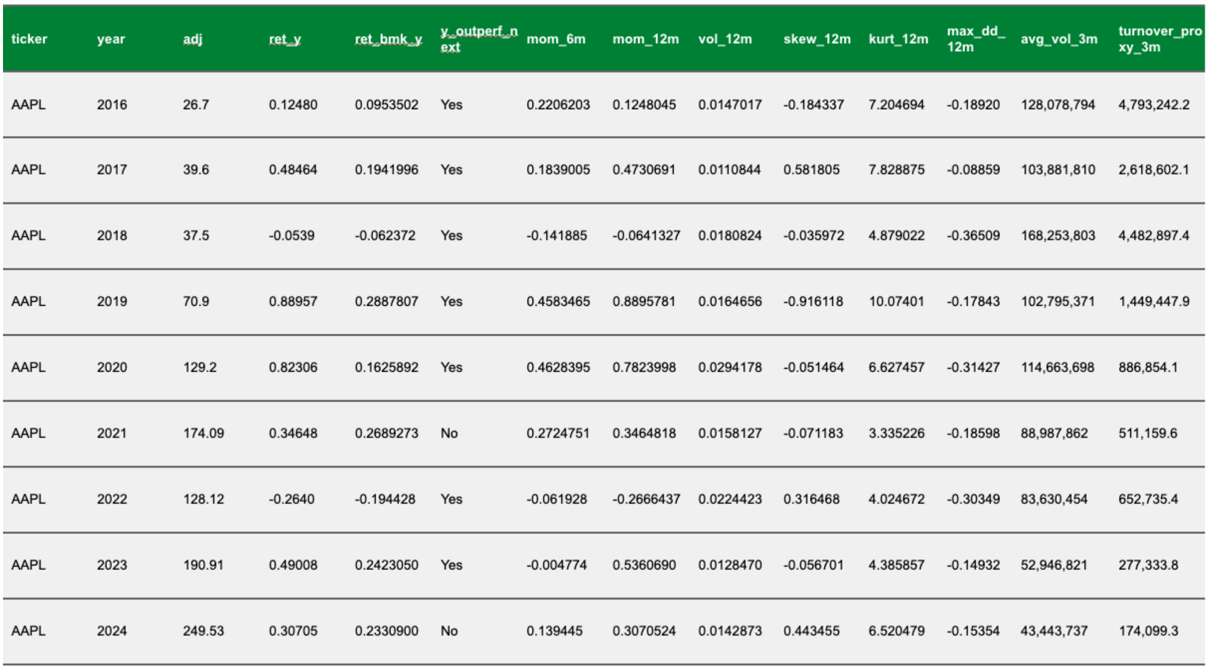
Accounting-based variables from Refinitiv complement the market indicators by capturing structural characteristics of firm performance. These include profitability metrics such as return on assets and earnings-to-price ratios, valuation measures such as dividend yield, and leverage indicators. While market variables evolve rapidly and reflect short-term sentiment or investor behavior, fundamental indicators represent slower-moving financial conditions. Integrating both dimensions enables the models to incorporate behavioral and structural sources of predictability in stock returns.

2.3.3 Merging and Winsorization

Market-based and fundamental datasets were merged using the firm identifier and calendar year as linking keys. Only firm-year observations with complete coverage from both sources were retained. To mitigate the influence of extreme values, all continuous variables were winsorized at the 1st and 99th percentiles, preserving rank order while limiting the effect of outliers. Remaining missing values were removed via listwise deletion to ensure consistent feature coverage across models. Alternative imputation methods (e.g., mean or model-based imputation) were deemed unnecessary, as missingness was limited and non-systematic. This process yielded a clean, balanced annual panel suitable for model estimation and robust comparison of predictive performance.

2.4 Final Dataset Structure

The resulting dataset is a balanced panel comprising 50 S&P 500 firms over the 2016–2025 period, totaling approximately 450 firm-year observations. Each observation corresponds to a unique `(ticker, year)` pair and includes roughly 30 explanatory variables alongside the binary target variable `y_outperf_next`, which equals 1 if the stock outperformed the S&P 500 in the following year. An illustrative extract is presented in Figure 2.1. Comprehensive definitions of all variables and formulas appear in Appendix B.



ticker	year	adj	ret_y	ret_bmk_y	y_outperf_next	mom_6m	mom_12m	vol_12m	skew_12m	kurt_12m	max_dd_12m	avg_vol_3m	turnover_proxy_3m
AAPL	2016	26.7	0.12480	0.0953502	Yes	0.2206203	0.1248045	0.0147017	-0.184337	7.204694	-0.18920	128,078,794	4,793,242.2
AAPL	2017	39.6	0.48464	0.1941996	Yes	0.1839005	0.4730691	0.0110844	0.581805	7.828875	-0.08859	103,881,810	2,618,602.1
AAPL	2018	37.5	-0.0539	-0.062372	Yes	-0.141885	-0.0641327	0.0180824	-0.035972	4.879022	-0.36509	168,253,803	4,482,897.4
AAPL	2019	70.9	0.88957	0.2887807	Yes	0.4583465	0.8895781	0.0164656	-0.916118	10.07401	-0.17843	102,795,371	1,449,447.9
AAPL	2020	129.2	0.82306	0.1625892	Yes	0.4628395	0.7823998	0.0294178	-0.051464	6.627457	-0.31427	114,663,698	886,854.1
AAPL	2021	174.09	0.34648	0.2689273	No	0.2724751	0.3464818	0.0158127	-0.071183	3.335226	-0.18598	88,987,862	511,159.6
AAPL	2022	128.12	-0.2640	-0.194428	Yes	-0.061928	-0.2666437	0.0224423	0.316468	4.024672	-0.30349	83,630,454	652,735.4
AAPL	2023	190.91	0.49008	0.2423050	Yes	-0.004774	0.5360690	0.0128470	-0.056701	4.385857	-0.14932	52,946,821	277,333.8
AAPL	2024	249.53	0.30705	0.2330900	No	0.139445	0.3070524	0.0142873	0.443455	6.520479	-0.15354	43,443,737	174,099.3

Figure 2.1: Extract from the final dataset (example: Apple Inc., 2016–2024).

3 Methodology

Our methodological framework proceeds incrementally, beginning with a simple and interpretable classifier (Decision Tree) before extending to ensemble methods (Random Forest) and nonlinear models (Neural Networks). This sequential design follows standard practice in predictive modeling, where model complexity is gradually increased to balance interpretability and generalization (Ahmed et al., 2022; Breiman, 2001; Campisi et al., 2024). This progressive approach allows us to evaluate how model complexity affects predictive performance while maintaining comparability across algorithms.

3.1 Decision Tree

We begin our empirical analysis with a Decision Tree (DT) classifier, a simple yet interpretable supervised learning algorithm that partitions the feature space into homogeneous regions according to the target variable. Decision trees have been widely used in financial prediction due to their transparency and low data requirements (Fu et al., 2020; Heo & Yang, 2016). The model serves as a natural benchmark before introducing more complex ensemble and nonlinear methods.

Formally, the decision tree recursively splits the sample based on predictor variables X_j and thresholds s that best separate the target classes. At each node, the algorithm selects the combination (j, s) that minimizes the weighted average impurity of the two resulting subsets:

$$R_1(j,s) = \{X : X_j < s\}, \quad R_2(j,s) = \{X : X_j \geq s\},$$

$$Q_{\text{split}}(j,s) = \frac{N_1}{N} G(R_1) + \frac{N_2}{N} G(R_2),$$

where N_1 and N_2 denote the number of observations in each child node, and $G(R)$ is the Gini impurity,

$$G(R) = \sum_{k=1}^{K} p_k(1 - p_k),$$

with p_k the class proportion in node R. The process continues greedily until a stopping rule is met, such as a maximum depth or a minimum node size.

We defined the tree-splitting stopping condition by setting the complexity parameter to `cp` $= 0.01$. To prevent overfitting, we applied cost–complexity pruning using the 1-SE rule, which penalizes excessive tree growth by increasing the `cp` until the best trade-off between accuracy and model complexity is found. This pruning criterion stops further splits when the marginal reduction in impurity is less than `cp`%, thereby improving the model's generalization ability.

The Decision Tree provides a transparent mapping between financial indicators and predicted outcomes, allowing direct interpretation of the most discriminative features driving stock outperformance. While its simplicity facilitates interpretability, its predictive performance is limited by high variance and instability to small data changes. These limitations motivate the use of ensemble techniques such as the Random Forest, discussed in the next section.

3.2 Random Forest

To enhance predictive accuracy and reduce the variance inherent in single-tree models, we employ the Random Forest (RF) algorithm (Breiman, 2001). The Random Forest is an ensemble learning method that aggregates a large number of decorrelated decision trees, each trained on a random bootstrap resample of the data. At each node, a random subset of predictors of size m_{try} is considered for splitting. This dual source of randomness, both in the sampling of observations and the selection of features, ensures tree diversity, mitigates overfitting, and stabilizes out-of-sample predictions.

Formally, given B trees trained on bootstrap samples $\{\mathcal{D}_b\}_{b=1}^{B}$, the Random Forest prediction for a new observation x is given by:

$$\hat{Y}_{RF}(x) = \text{majority vote}\,\{\,\hat{Y}_b(x)\,\}_{b=1}^{B},$$

where each tree $\hat{Y}_b(x)$ provides a binary classification in $\{\text{No}, \text{Yes}\}$. For probabilistic predictions, the RF averages class probabilities across trees, yielding $\Pr(Y = \text{Yes} \mid x)$ as the proportion of trees voting "Yes."

Each observation omitted from a tree's bootstrap sample (the so-called out-of-bag (OOB) sample) serves as a built-in validation set, providing an unbiased internal estimate of generalization error and the AUC metric without needing a separate split.

3.2.1 Training and Tuning Procedure

We first fit a baseline RF with $n_{\text{tree}} = 500$ using the `randomForest` package, with OOB error tracking and permutation-based variable importance enabled. This exploratory step provides a first view of predictive strength and identifies potentially influential features.

Next, we perform systematic hyperparameter tuning using the `caret` framework. A stratified 10-fold cross-validation is ran, optimizing the Area Under the ROC Curve (AUC) to ensure robustness to class imbalance. For each of several forest sizes $n_{\text{tree}} \in \{300, 500, 1000, 1500\}$, we evaluate a grid of feature subset sizes:

$$m_{\text{try}} \in \big\{ \lfloor \sqrt{p} - 2 \rfloor,\ \lfloor \sqrt{p} \rfloor,\ \lfloor \sqrt{p} + 2 \rfloor,\ \lfloor p/4 \rfloor,\ \lfloor p/3 \rfloor \big\},$$

where p denotes the total number of predictors. The configuration that maximizes mean AUC across folds is retained. To refine this selection, we perform a Leave-One-Out Cross-Validation (LOOCV) around the best 10-fold candidate, testing $m_{\text{try}}^{\star} \pm 1$ at a larger number of trees ($n_{\text{tree}} = 3000$). The final RF model is thus defined by the best-performing combination of $(m_{\text{try}}, n_{\text{tree}})$ from these two tuning phases.

3.2.2 Final Model and Evaluation

The final Random Forest is trained on the complete dataset using the optimal hyperparameters $(m_{\text{try}}^{\text{final}}, n_{\text{tree}}^{\text{final}})$. We report the OOB AUC, confusion matrix, and permutation-based variable importance (MeanDecreaseAccuracy). These diagnostics allow us to evaluate both predictive performance and the relative contribution of each feature to model accuracy.

3.2.3 Interpretation

In financial terms, this ensemble framework captures nonlinear relationships between market- and accounting-based features while remaining robust to multicollinearity and noise. The OOB and CV diagnostics, together with the variable importance analysis, provide interpretable insights into which firm characteristics are most predictive of subsequent market outperformance. Compared to the single decision tree, the Random Forest achieves markedly higher stability and generalization performance, making it a strong benchmark before moving toward more complex neural network architectures.

3.3 Neural Network

To capture potential nonlinearities and complex interactions among financial predictors, we extend our analysis with a feed-forward Artificial Neural Network (ANN). While tree-based models rely on recursive partitioning of the predictor space, neural networks approximate the relationship between input features and the target variable through compositions of nonlinear activation functions, offering greater representational flexibility.

A single-hidden-layer network with H neurons can be expressed as:

$$\hat{y}(x) = \sigma\left(\alpha_0 + \sum_{h=1}^{H} \beta_h \, g(\alpha_h^\top x)\right),$$

where x is the vector of input features, $g(\cdot)$ is a nonlinear activation function (here, the logistic sigmoid), and $\sigma(\cdot)$ denotes the output activation transforming the network's linear combination into a probability estimate $\Pr(Y = \text{Yes} \mid x)$. The parameters $\{\alpha_h, \beta_h\}$ are optimized by minimizing the binary cross-entropy loss via gradient descent.

3.3.1 Preprocessing and Implementation

Since neural networks are sensitive to the scale of input features, all predictors were normalized to the range $[0, 1]$ using min–max scaling (`preProcess = "range"`). The binary target variable `y_outperf_next` was encoded as a factor with levels `{No, Yes}`, corresponding respectively to underperforming or outperforming the S&P 500 benchmark.

Model estimation was implemented in `R` using the `nnet` and `caret` packages. Training was performed with a single hidden layer, consistent with the small-to-moderate size of our dataset, to avoid overfitting and maintain interpretability.

3.3.2 Hyperparameter Tuning

We optimized two key hyperparameters:

- **Hidden units (`size`)**: the number of neurons in the hidden layer, controlling the model's capacity.

- **Weight decay (decay)**: an L_2 regularization parameter penalizing large weights to reduce overfitting.

A grid search was conducted over:

$$\texttt{size} \in \{3, 5, 7, 9\}, \quad \texttt{decay} \in \{0, 0.001, 0.01, 0.1\}.$$

For each configuration, a stratified 10-fold cross-validation was performed using the `caret` framework, optimizing the Area Under the ROC Curve (`metric = "ROC"`). Cross-validation folds were balanced by class to ensure robustness against any mild imbalance in the outperforming vs. non-outperforming observations. The maximum number of optimization iterations was set to 300 (`maxit = 300`), providing stable convergence without excessive computational cost.

3.3.3 Final Model and Evaluation

After identifying the optimal combination of $(\texttt{size}, \texttt{decay})$ that maximized the average AUC across folds, the model was retrained on the full dataset using these parameters. We retained out-of-fold predictions from cross-validation to compute global accuracy and AUC. These metrics were then compared directly to those of the Random Forest under identical validation conditions (10-fold CV with AUC optimization) to ensure fair benchmarking.

3.3.4 Remarks

The neural network, although less interpretable than tree-based models, provides a useful nonlinear benchmark to evaluate whether more flexible functional forms significantly enhance predictive power. Its performance relative to the Random Forest helps assess whether the additional model complexity translates into genuine improvements in forecasting ability or primarily captures noise within a limited-sample financial context.

3.4 Model Evaluation Metrics

To ensure a consistent and interpretable comparison across models, predictive performance was evaluated using four standard classification metrics: Accuracy, Precision, Recall, and the Area Under the ROC Curve (AUC).

3.4.1 Accuracy

Accuracy measures the proportion of correctly classified observations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FN denote true and false positives and negatives. Although intuitive, accuracy can be misleading in the presence of class imbalance, hence it is complemented by AUC.

3.4.2 Area Under the ROC Curve (AUC)

AUC quantifies the model's ability to discriminate between outperforming and non-outperforming stocks, independent of any classification threshold. AUC is the preferred metric in financial prediction studies due to its threshold independence and robustness to class imbalance (Caparrini et al., 2024; Htun et al., 2024). It is defined as the probability that a randomly chosen outperformer receives a higher predicted score than a non-outperformer:

$$\text{AUC} = \Pr(\hat{p}_{\text{pos}} > \hat{p}_{\text{neg}}) \, ,$$

with $\hat{p}_{\text{pos}}$ and $\hat{p}_{\text{neg}}$ denoting the predicted probabilities for the two classes. AUC values of 0.5 and 1.0 correspond to random and perfect classification, respectively.

3.4.3 Precision and Recall

Precision measures the share of correctly identified outperformers among all stocks predicted to outperform:

$$\text{Precision} = \frac{TP}{TP + FP},$$

while Recall captures the proportion of true outperformers that were correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

High recall minimizes missed opportunities, whereas high precision reduces false investment signals.

All models were assessed via 10-fold cross-validation using AUC as the primary optimization criterion to ensure fair and robust comparison across algorithms.

4 Results / Analysis

This section presents the empirical results obtained from the predictive models developed in Section 3. The goal is to assess their relative performance in forecasting whether a stock will outperform the S&P 500 in the following year. All models were trained and evaluated under a unified framework to ensure comparability.

4.1 Model Training and Tuning Procedure

Each model underwent rigorous hyperparameter tuning. For the Random Forest, we combined a 10-fold cross-validation grid search with a secondary Leave-One-Out Cross-Validation (LOOCV) refinement around the best 10-fold configuration. The final model was trained with $m_{\text{try}} = 6$ and $n_{\text{tree}} = 1{,}000$, values that offered near-optimal performance while preserving computational efficiency. The OOB error rate flattened beyond 2,000–3,000 trees, confirming convergence (see Figure 4.5).

The Decision Tree (CART) was tuned via 10-fold CV, selecting the optimal complexity parameter (`cp = 0.01`) to control overfitting. For the Neural Network, a grid search over hidden-layer sizes (3, 5, 7, 9) and weight-decay parameters (0, 0.001, 0.01, 0.1) was conducted, also using 10-fold CV with AUC as the selection criterion. All models were further validated on an 80/20 train–test split to confirm generalization. The neural network employed a sigmoid activation in the output layer, producing probabilistic forecasts suited to the binary target. Its smooth, bounded response supports stable learning but may limit the capture of strong nonlinearities, helping explain its slightly lower performance relative to ensemble models.

4.2 Overall Predictive Performance

Table 4.1 summarizes the out-of-sample classification metrics obtained under 10-fold cross-validation. The Random Forest achieved the highest performance across all indicators, followed closely by the Neural Network, while the single Decision Tree lagged behind due to its higher variance and limited capacity to model nonlinear interactions.

Table 4.1: Cross-validated model performance comparison.

Model	Accuracy	Recall	Precision
Decision Tree (CART)	0.57	0.60	0.62
Random Forest	0.66	0.78	0.68
Neural Network	0.63	0.70	0.66

The Random Forest outperformed alternative models across all metrics, with particularly strong recall for the outperforming class. This asymmetric predictive accuracy reflects the richer information content of market-based indicators such as `beta_12m`, `vol_12m`, and `max_dd_12m`, which exhibit greater cross-sectional variation across firms (see Appendix C). These variables provide clearer structural patterns for detecting future outperformers, whereas underperformance is often driven by idiosyncratic or event-specific factors, such as managerial turnover, regulatory actions, or transient market shocks, that are not systematically captured by our feature set. Consequently, upward performance appears more predictable than downward movements, consistent with the notion that risk-related and momentum-based factors carry more persistent information about relative stock behavior.

Table 4.2: Confusion matrix of the Random Forest model (OOB predictions).

	Predicted No	Predicted Yes	Class Error
Actual No	98	98	0.50
Actual Yes	58	196	0.23

The confusion matrix (Table 4.2) further illustrates this asymmetry: the model identifies outperforming stocks with relatively high accuracy (recall 0.78), but misclassifies a larger fraction of underperformers. Such behavior is typical in financial prediction tasks where positive market outcomes are structurally more explainable than negative ones.

The Random Forest achieved an AUC of approximately 0.73, outperforming the Neural Network (AUC = 0.72) and the Decision Tree (AUC = 0.65). This ordering is consistent with prior empirical evidence that ensemble models generally surpass both single classifiers and parametric methods in financial forecasting applications (Breiman, 2001; Caparrini et al., 2024; Novak & Velušček, 2016). The ensemble's higher recall and AUC highlight its stronger capacity to identify future outperformers, a key property for portfolio construction and screening strategies.

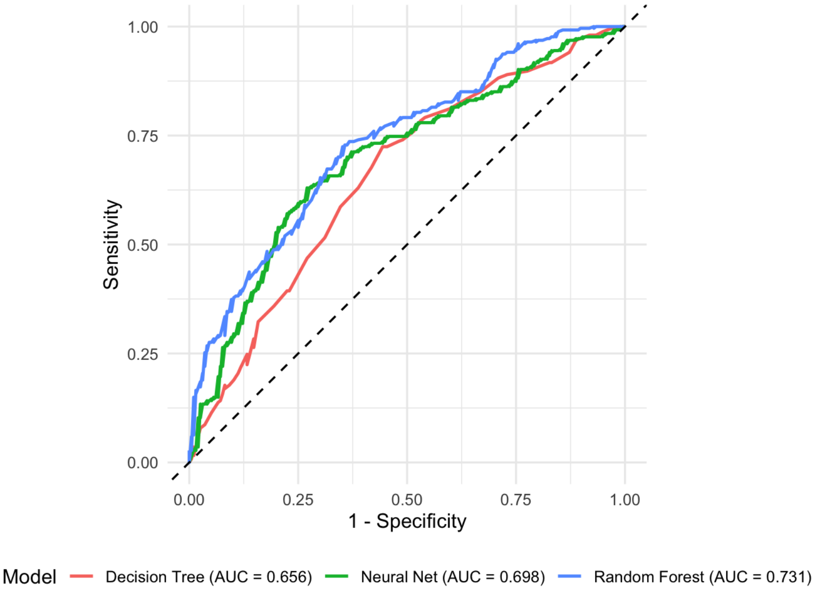


Figure 4.1: ROC curve comparison across models. The Random Forest dominates at all thresholds.

4.3 Feature Importance and Financial Interpretation

The Random Forest's permutation importance analysis (Figure 4.2) highlights the top drivers of outperformance. Market-based indicators such as `beta_12m`, `skew_12m`, and `max_drawdown` emerge as the most influential predictors, capturing exposure to systematic risk and downside volatility.

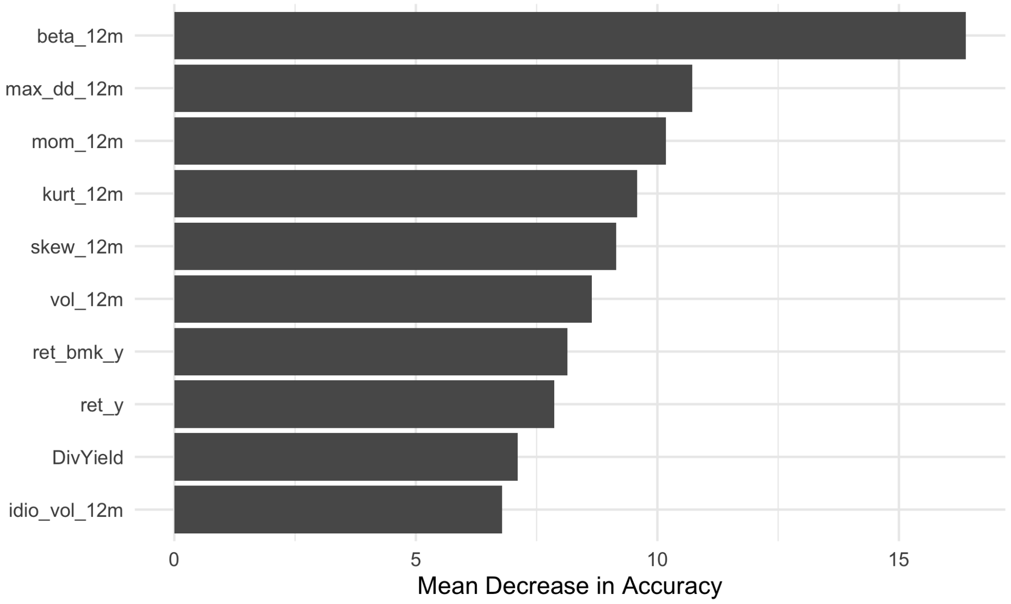


Figure 4.2: Permutation importance of the ten most relevant predictors in the Random Forest.

From a financial perspective, these findings are consistent with factor-based asset pricing theory (Fama & French, 1992, 2015). High market beta and negative skewness are typically associated with riskier stocks that require higher expected returns. Downside risk proxies such as drawdown further emphasize the market's sensitivity to losses, variables that are often underrepresented in linear models but well captured by nonparametric machine learning techniques.

The Decision Tree provides a simplified visual summary of these relationships (Figure 4.3), with splits primarily driven by beta and volatility measures. While less accurate, it offers interpretability and transparency.

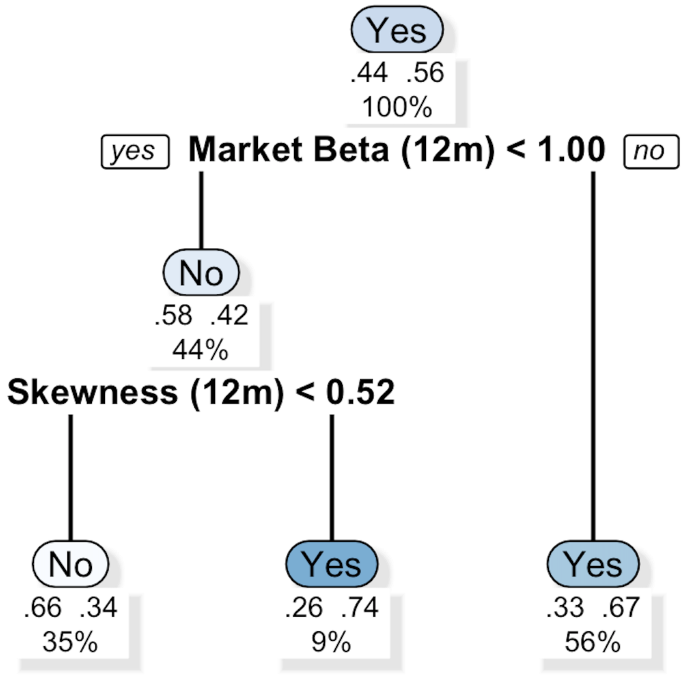


Figure 4.3: Pruned Decision Tree showing primary risk-based splits.

4.4 Model Stability and Robustness Checks

4.4.1 Cross-Validation and OOB Stability

The stability of the models was evaluated using 10-fold CV, LOOCV, and OOB error estimation. The Random Forest consistently achieved the highest AUC across folds (median AUC $\approx$ 0.73) with minimal variance (Figure 4.4), confirming strong generalization. OOB error stabilized at 0.333 when increasing from 1,000 to 3,000 trees, indicating diminishing returns beyond that point (Figure 4.5).

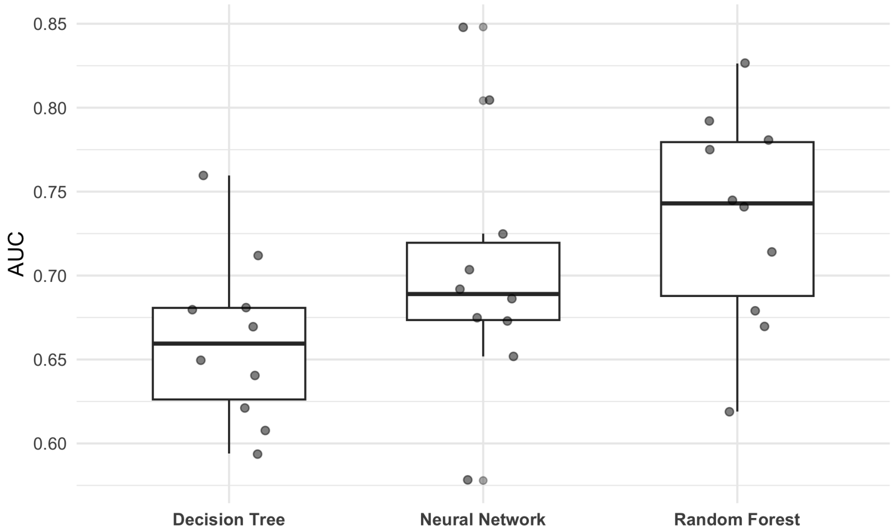


Figure 4.4: Distribution of cross-validation AUC across folds. The Random Forest demonstrates both the highest median and lowest variance.

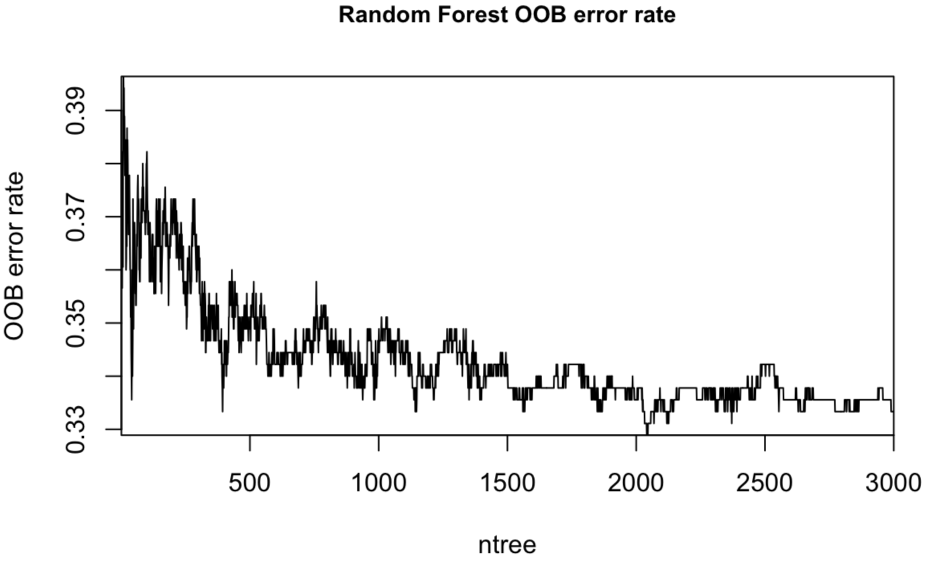


Figure 4.5: Random Forest out-of-bag error convergence across increasing number of trees.

4.4.2 Feature and Regime Robustness

We further tested robustness by splitting the data into pre-2020 and post-2020 periods. The Random Forest maintained high accuracy in all scenarios, and the same features remained among the most predictive, suggesting that its learned relationships are not driven by specific subsets of data. Performance improved slightly in the post-2020 sample, a period of higher volatility, consistent with Htun et al. (2024) and Campisi et al. (2024), who observe that machine learning models tend to perform better in turbulent markets.

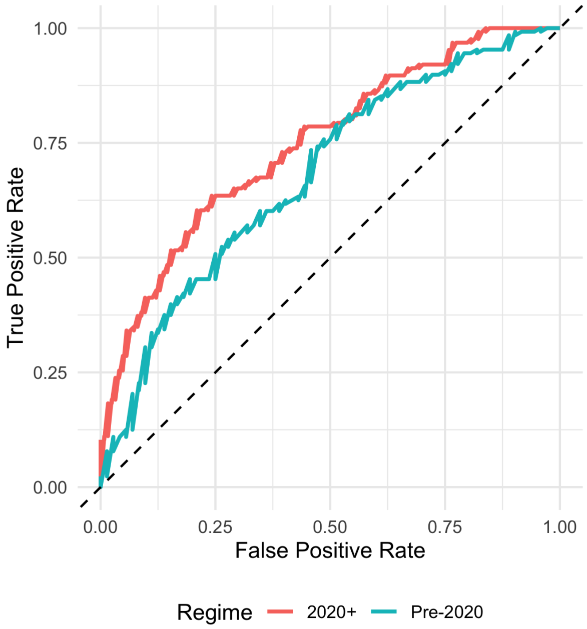


Figure 4.6: Random Forest ROC curves by market regime. Performance is marginally higher in the post-2020 high-volatility period.

4.5 Discussion and Limitations

The results highlight a clear trade-off between interpretability and predictive power. The Decision Tree offers transparency but limited accuracy, while the Neural Network captures nonlinearities but is more data-hungry and sensitive to overfitting. The Random Forest strikes the best balance, robust, interpretable at the feature level, and consistently superior across validation schemes. This supports the established view that ensemble methods are well-suited for financial prediction tasks with moderate data size and complex interactions (Fu et al., 2020; Ghosh et al., 2022).

However, the dataset size (around 450 firm-year observations across 50 S&P 500 constituents) constrains the models' ability to learn rare or long-horizon dependencies. Future extensions should expand the sample and incorporate macroeconomic, sentiment, and sector-level features, or explore hybrid architectures such as stacked ensembles combining tree-based and neural models. Methods like Gradient Boosting Machines (e.g., XGBoost or LightGBM) may further enhance predictive accuracy while maintaining robustness.

5 Conclusion

This study examined whether firm-specific market and accounting indicators can predict annual stock outperformance relative to the S&P 500 index. By framing the problem as a binary classification task, three supervised learning models of increasing complexity were developed and compared: a pruned Decision Tree, a Random Forest, and a feed-forward Neural Network. All models were trained and evaluated under a unified cross-validation framework to ensure a fair and robust assessment of predictive performance.

The results indicate that the Random Forest consistently achieved the highest predictive accuracy and AUC, outperforming both the Neural Network and the single Decision Tree. Its ensemble structure effectively captured nonlinear interactions while mitigating overfitting through bootstrapped sampling and averaging. The Neural Network performed comparably but exhibited higher sensitivity to hyperparameter choices, likely reflecting the relatively small sample size and the sigmoid activation function's smooth compression of output probabilities. The Decision Tree, while less accurate, provided interpretable rule-based insights into the key determinants of outperformance, such as market beta, volatility, and downside risk measures.

From a financial perspective, the most influential predictors identified by the Random Forest, systematic risk, skewness, and drawdown, are consistent with established asset-pricing theory. Stocks exhibiting higher market exposure and asymmetric return distributions were more likely to outperform, suggesting that machine learning models can rediscover and quantify risk–return trade-offs traditionally captured by factor-based approaches. The inclusion of accounting fundamentals modestly enhanced stability, confirming that structural financial health complements market-driven signals in long-horizon prediction.

These findings contribute to the growing literature on applying machine learning to financial forecasting by demonstrating that annual, factor-based data can yield meaningful predictive power even with moderate sample sizes. They also highlight the importance of model selection: ensemble methods such as Random Forests balance flexibility, interpretability, and robustness more effectively than either single trees or fully parametric neural networks in this context.

Nonetheless, several limitations remain. The dataset is restricted to 50 large-cap U.S. equities, which, while ensuring data completeness, limits generalizability across smaller or international firms. Future research should expand the cross-sectional and temporal scope, incorporate macroeconomic and sentiment variables, and explore more advanced architectures such as Gradient Boosting Machines, LSTM networks, or hybrid ensemble frameworks. Further investigation into model interpretability, using techniques like SHAP or partial dependence analysis, could also enhance understanding of how financial features interact to drive returns.

Overall, this study demonstrates that machine learning methods, particularly Random Forests, offer a powerful and empirically grounded framework for forecasting relative stock performance. By combining market-based dynamics with accounting fundamentals, they provide an effective bridge between traditional financial theory and modern data-driven approaches to investment analysis.

References

Ahmed, N., Deb, K., & Khan, M. (2022). Applications of artificial intelligence and machine learning in finance: A comprehensive review. *Applied Intelligence*, *52*, 15159–15187. https://doi.org/10.1007/s10489-022-03183-3

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campisi, S., Zhao, X., & Lu, Y. (2024). Comparing machine learning models for predicting the direction of the u.s. stock market based on volatility indices. *Finance Research Letters*, *61*, 104634. https://doi.org/10.1016/j.frl.2024.104634

Caparrini, A., Menchetti, F., & Sassetti, F. (2024). Machine learning algorithms for stock selection based on expected profitability. *Research in International Business and Finance*, *70*, 102336. https://doi.org/10.1016/j.ribaf.2024.102336

Chen, M., & Ge, L. (2021). Exploring the applications of artificial neural networks in financial portfolio management. *Expert Systems with Applications*, *165*, 113943. https://doi.org/10.1016/j.eswa.2020.113943

Cochrane, J. H. (2011). *Presidential address: Discount rates* (Vol. 66). https://doi.org/10.1111/j.1540-6261.2011.01671.x

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, *47*(2), 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010

Fu, T., Yang, J., & Li, X. (2020). Applying random forest to factor-based stock selection. *Applied Economics Letters*, *27*(8), 674–678. https://doi.org/10.1080/13504851.2019.1644422

Ghosh, S., Saha, S., & Dutta, S. (2022). Stock market prediction using long short-term memory networks and random forests. *Expert Systems with Applications*, *205*, 117695. https://doi.org/10.1016/j.eswa.2022.117695

Heo, J., & Yang, J. Y. (2016). Stock price prediction based on financial statement analysis using artificial neural networks. *International Journal of Computer Science and Information Technology*, *8*(3), 9–20.

Htun, N. M., Zhao, H., & Yin, Y. (2024). Forecasting relative stock returns using machine learning models. *Applied Soft Computing*, *155*, 111028. https://doi.org/10.1016/j.asoc.2024.111028

Novak, T., & Velušček, M. (2016). Prediction of stock price movement based on support vector machines. *Quantitative Finance*, *16*(7), 981–991. https://doi.org/10.1080/14697688.2015.1133279

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*(3), 425–442. https://doi.org/10.2307/2977928

Appendices

A Data Supplement

A.1 Universe and Dates

We construct a fixed universe of 50 large-cap U.S. equities and use the S&P 500 as benchmark. The equity tickers are:

AAPL	MSFT	NVDA	AMZN	GOOGL	META	TSLA	AVGO	COST	ORCL
ADBE	NFLX	CRM	TSM	AMD	INTC	CSCO	QCOM	TXN	IBM
AMAT	MU	LRCX	KLAC	MRVL	ASML	JPM	BAC	WFC	MA
V	UNH	PEP	KO	PG	PFE	MRK	JNJ	WMT	HD
LOW	TGT	DIS	NKE	MCD	CAT	XOM	CVX	COP	SLB

Daily market data are pulled from Yahoo Finance for `2010-01-01` to `2025-06-30` to support rolling indicators. The benchmark is `^GSPC` (S&P 500).

A.2 Coverage and Cleaning Summary

Table A.1 summarizes the pipeline stages (exact counts depend on run date and the Refinitiv extract).

Table A.1: Coverage and Cleaning Summary

Daily price series pulled (tickers)	50
Daily range for prices	2010-01-01 to 2025-06-30
Benchmark	`^GSPC` (S&P 500)
Annual panel after market collapse	$\sim$500 firm–years
After merge with Refinitiv/dividends	$\sim$480 firm–years
Final (post winsorization + complete cases)	$\sim$450 firm–years

B Variable Definitions and Computation

This appendix details the mathematical definitions and computation procedures of all variables included in the dataset. For each stock i and trading day t, let $P_{i,t}$ denote the adjusted closing price, corrected for dividends and stock splits, and let $r_{i,t}$ be the corresponding daily return. The market benchmark return is denoted r_t^{mkt}. All rolling statistics are computed over historical windows of 3, 6, or 12 months, depending on the indicator, and subsequently extracted at each firm's year-end trading date.

B.1 Daily Returns

Daily log returns are defined as:

$$r_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1,$$

representing the percentage price change from day $t-1$ to day t.

B.2 Momentum

Momentum captures short- and medium-term price continuation effects. It is computed over rolling six- and twelve-month windows as:

$$\text{mom\_6m}_{i,t} = \frac{P_{i,t}}{P_{i,t-126}} - 1, \tag{B.1}$$

$$\text{mom\_12m}_{i,t} = \frac{P_{i,t}}{P_{i,t-252}} - 1, \tag{B.2}$$

where 126 and 252 represent approximately six and twelve months of trading days, respectively.

B.3 Volatility

Volatility measures the dispersion of returns and serves as a proxy for total risk. The sample standard deviation of daily returns over a rolling 12-month window is:

$$\sigma_{i,t}^{(d)} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(r_{i,t-k} - \bar{r}_i)^2},$$

where $n = 252$ and $\bar{r}_i$ denotes the mean daily return in the same window. Annualized volatility is then obtained as:

$$\text{vol\_12m}_{i,t} = \sigma_{i,t}^{(a)} = \sigma_{i,t}^{(d)} \times \sqrt{252}.$$

B.4 Systematic and Idiosyncratic Risk

Systematic risk is captured by the rolling beta coefficient:

$$\beta_{i,t} = \frac{\text{Cov}(r_{i,t},\, r_t^{mkt})}{\text{Var}(r_t^{mkt})},$$

computed over the previous 12 months. The idiosyncratic volatility component is:

$$\text{idio\_vol}_{i,t} = \sqrt{\max\bigl(\text{Var}(r_{i,t}) - \beta_{i,t}^2\,\text{Var}(r_t^{mkt}),\, 0\bigr)}.$$

B.5 Asymmetry and Tail Risk

Distributional asymmetries are summarized by the skewness and kurtosis of daily returns:

$$\text{skew}_{i,t} = \frac{E[(r_{i,t} - \bar{r}_i)^3]}{s_i^3}, \tag{B.3}$$

$$\text{kurt}_{i,t} = \frac{E[(r_{i,t} - \bar{r}_i)^4]}{s_i^4}, \tag{B.4}$$

where s_i is the sample standard deviation. Negative skewness indicates a propensity for large negative shocks, while excess kurtosis reflects fat-tailed return distributions.

B.6 Maximum Drawdown

The maximum drawdown (MDD) represents the largest cumulative loss from a local peak to a subsequent trough within a rolling one-year window:

$$\mathrm{MDD}_{i,t} = \min_{u<v\in[t-252,\,t]} \left(\frac{P_{i,v}}{P_{i,u}} - 1 \right).$$

A large negative MDD implies substantial downside risk.

B.7 Liquidity Proxies

Liquidity is approximated using average trading volume and turnover measures. The three-month average volume is:

$$\text{avg\_vol\_3m}_{i,t} = \frac{1}{63} \sum_{k=1}^{63} \text{Volume}_{i,t-k},$$

and the turnover proxy is defined as:

$$\text{turnover\_proxy}_{i,t} = \frac{\text{avg\_vol\_3m}_{i,t}}{P_{i,t}}.$$

Higher turnover implies greater market liquidity and investor participation.

B.8 Price Level

The logarithm of the year-end adjusted price provides a scale-invariant measure of the stock's price level:

$$\text{price\_level\_log}_{i,t} = \log(P_{i,t}).$$

B.9 Winsorization

To mitigate the effect of outliers, all continuous variables are winsorized at the 1st and 99th percentiles:

$$x^* = \min(\max(x,\, q_{0.01}),\, q_{0.99}),$$

where q_p denotes the p-th percentile of x.

B.10 Target Variable

The binary outcome variable `y_outperf_next` equals 1 if the stock's total return in year $t + 1$ exceeds that of the S&P 500 benchmark and 0 otherwise. This formulation allows the models to learn the conditional probability of outperforming the market based on information available at year-end t.

Table B.1: Description of Variables in the Final Dataset

Variable	**Type**	**Description**
ticker	Text	Stock ticker symbol (e.g., AAPL, MSFT)
year	Integer	Year of observation (aligned with year-end date)
adj	Continuous	Adjusted closing price at year-end
ret_y	Continuous	Annual stock return for the current year
ret_bmk_y	Continuous	Annual S&P 500 return for the same year
y_outperf_next	Binary	Target variable
mom_6m	Continuous	Six-month momentum
mom_12m	Continuous	Twelve-month momentum
vol_12m	Continuous	Annualized rolling volatility of daily returns
beta_12m	Continuous	Rolling 12-month beta w.r.t. S&P 500
skew_12m	Continuous	Rolling 12-month skewness of daily returns
kurt_12m	Continuous	Rolling 12-month kurtosis of daily returns
max_dd_12m	Continuous	Maximum drawdown over the past 12 months
avg_vol_3m	Continuous	Average daily trading volume over past 3 months
turnover_proxy_3m	Continuous	Turnover proxy: avg_vol_3m / price
price_level_log	Continuous	Logarithm of year-end adjusted price
idio_vol_12m	Continuous	Idiosyncratic volatility over past 12 months
div_ttm	Continuous	Trailing twelve-month dividends
div_yield_ttm	Continuous	Dividend yield (div_ttm / adj)
PE	Continuous	Price-to-earnings ratio (Refinitiv)
EVEBITDA	Continuous	Enterprise value / EBITDA (Refinitiv)
Revenue	Continuous	Total revenue (Refinitiv)
NetIncome	Continuous	Net income (Refinitiv)
EBITDA	Continuous	Earnings before interest, taxes, depreciation
DivYield	Continuous	Dividend yield (Refinitiv)

C Supplementary Statistical Analysis

To assess whether market-based variables exhibit richer cross-sectional information than fundamentals, we compare their relative variation across firms and years. For each variable, we compute the coefficient of variation (CV = SD / mean) by year and average it over time. This approach measures variability independently of scale differences between financial ratios and monetary variables. Extreme CV values above 10 were omitted to improve robustness and visual clarity (see Figure C.1).

Table C.1: Average cross-sectional relative variation by variable type (CV).

Type	Average CV
Fundamental	1.72
Market-based	2.97

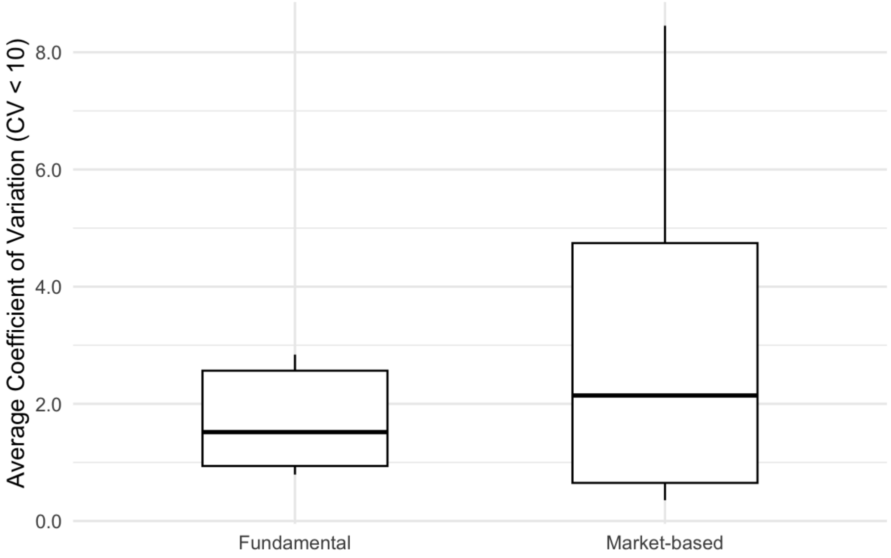


Figure C.1: Distribution of cross-sectional coefficients of variation (CV) by variable type. Extreme outliers (CV $>$ 10) were omitted for clarity.

The results indicate that market-based variables display substantially higher relative variation across firms and years than fundamental indicators. This supports the interpretation that the Random Forest draws stronger predictive signals from market-related features, as they provide more distinct information for separating outperforming from non-outperforming stocks.