Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD38878

Explainable Machine Learning for One-Year Stock Return Prediction Using Financial Indicators

Binhong Wu

School of Mathematical Sciences, Tongji University, Shanghai, China
2452997@tongji.edu.cn

Abstract. This study examines whether annual financial indicators can rank U.S. stocks by their subsequent one-year price changes. The dataset contains 22,077 firm-year observations and 221 numeric indicators from 2014–2018. Ridge, Lasso, Random Forest, XGBoost, and LightGBM are fitted using the 2014–2016 observations and evaluated on the 2018 financial-data cohort, whose target covers 2019. The 2017 cohort is excluded from model fitting. LightGBM records the highest test Spearman rank information coefficient, 0.0805, and a top-minus-bottom decile difference of 26.69 percentage points. Lasso has the lowest root mean squared error and mean absolute error, so the preferred model depends on the evaluation criterion. SHAP analysis of LightGBM identifies earnings yield, weighted-average shares outstanding, price-to-book ratio, the Energy sector indicator, and share issuance or buybacks as its five leading inputs. These attributions describe the fitted model rather than causal return determinants. The reported experiment uses full-sample winsorization and lacks filing timestamps; consequently, its chronological split does not eliminate information leakage. The findings are descriptive comparisons within one annual test cohort, and the decile difference should not be interpreted as an implementable investment return.

Keywords: stock return prediction, financial indicators, machine learning, gradient boosting, SHAP

1. Introduction

Annual accounts supply many possible predictors of stock returns, but selecting and combining them is difficult. The Fama–French three-factor and five-factor models organize return variation around market, size, value, profitability, and investment factors [1, 2]. These factor models provide economic context for a different task considered here: predicting individual firms' subsequent price changes from accounting variables. The large number of proposed return predictors also raises a multiple-testing problem, documented by Harvey, Liu, and Zhu [3]. A positive association in one sample therefore needs to be distinguished from evidence that will persist in another period.

Machine-learning studies offer useful comparisons for this task. Gu, Kelly, and Xiu evaluate a broad set of forecasting methods in empirical asset pricing [4], while Krauss, Do, and Huck study nonlinear learners in an S&P 500 statistical-arbitrage setting [5]. Their results motivate comparisons between regularized linear models and tree ensembles; they do not establish which model should

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD38878

lead in this shorter annual dataset. Interpretation is a separate concern. LIME explains predictions through local surrogate models [6], whereas SHAP assigns additive feature contributions [7]. Tree-specific SHAP methods make these contributions practical to compute for ensembles [8].

The present comparison uses 221 numeric financial indicators for U.S. listed firms. The 2014–2016 cohorts supply training observations, 2017 is held aside, and 2018 supplies the test predictors. Because the target field describes the following calendar year, the final test concerns 2019 price changes. Hyperparameters for four models are chosen by three-fold cross-validation within the training observations; LightGBM uses a fixed specification. Financial model validation requires attention to information timing and data reuse [9]. In this experiment, full-sample clipping and missing disclosure dates limit what can be inferred from the chronological split.

Three questions guide the analysis. Does the model with the smallest prediction error also rank stocks most accurately? Which inputs account for LightGBM's fitted predictions? How do the reported comparisons change after extreme-target exclusion and separate sector fits? LightGBM leads on rank correlation, while Lasso leads on prediction error. This distinction is more informative than a single claim of overall model superiority. The portfolio sorts and SHAP plots provide additional descriptions of the same fitted models.

Section 2 describes the observations and preprocessing. Section 3 specifies the learners and evaluation criteria. Sections 4–6 report the model comparison, SHAP analysis, and sensitivity checks. Section 7 discusses the scope of the findings and the requirements for a stronger temporal evaluation.

2. Dataset and preprocessing

2.1. Dataset description

The public Kaggle dataset "200+ Financial Indicators of US stocks (2014–2018)" supplies 22,077 firm-year observations [10]. Each record contains 221 numeric financial indicators, a sector classification, and a subsequent-year price-change label. The predictors cover income statements, balance sheets, cash flows, valuation ratios, and growth measures. Eleven sectors are represented, including Financial Services, Healthcare, and Technology. Table 1 gives the annual sample counts. File years refer to financial-data cohorts: the label in the 2018 file is 2019 PRICE VAR [%]. The target measures price variation and should not be assumed to include dividends or delisting returns.

Table 1. Dataset overview by financial-data year

Year	Split	Samples	Features
2014	Train	3,808	221
2015	Train	4,120	221
2016	Train	4,797	221
2017	Held aside	4,960	221
2018	Test	4,392	221
Total	-	22,077	221

DOI: 10.54254/2754-1169/2026.LD38878

2.2. Preprocessing pipeline

The Class column is removed because it encodes the sign of the future price change. Ticker and year identify observations and are excluded from the predictor matrix. The supplied implementation clips the target and eligible continuous predictors at their 1st and 99th percentiles before splitting the combined dataset by year. Consequently, the clipping thresholds use information from the held-aside and test cohorts. The errors and decile means reported below refer to the processed target, and the experiment cannot be described as free of test-set leakage.

The 2014–2016 cohorts contain 12,725 training observations. The 4,960 observations from 2017 are held aside as a buffer and are not used to choose the reported models; the 2018 test cohort contains 4,392 observations. Missing numeric values are replaced by training-set medians. Sector indicators are aligned to the training categories, with missing sector labels assigned to an Unknown category. Ridge and Lasso use features standardized with training-set means and scales; the tree models use imputed, unstandardized inputs. Imputation and scaling are fitted before the three-fold parameter search, rather than refitted inside each fold. The inner validation therefore also has preprocessing leakage.

A financial-data year is not a disclosure timestamp. The source does not specify when every accounting value became available, so matching a fiscal-year record directly to the next calendar year can include information published after that return window begins. Excluding the 2017 cohort does not resolve this timing issue or the full-sample clipping issue. These features of the original experiment are retained in the description so that the reported numbers are not confused with results from a fully chronological, point-in-time backtest.

3. Methodology

3.1. Models

Ridge and Lasso are linear regressions with different coefficient penalties. Ridge uses an L2 penalty to shrink coefficients [11]; Lasso uses an L1 penalty that can set coefficients to zero [12]. Their penalty grids are {0.1, 1, 10, 100} and {0.01, 0.1, 1}, respectively. Random Forest averages 200 randomized regression trees, with maximum depth selected from {7, unrestricted} [13]. Gradient boosting builds trees sequentially to reduce a loss function [14]. XGBoost uses 200 trees and searches maximum depth {4, 6} and learning rate {0.05, 0.1} [15]. LightGBM uses 200 trees, maximum depth 5, and learning rate 0.05 without a parameter search [16].

The four parameter searches use scikit-learn GridSearchCV with three folds and negative mean squared error as the scoring rule [17]. The selected estimators are refitted on the full 2014–2016 training set. The folds are not rolling-origin splits, and LightGBM does not receive the same search budget as the other learners. A random seed of 42 is used in the supplied implementation. These settings define the comparison; the results should not be generalized to all possible parameter choices for each model family.

3.2. Evaluation metrics

Root mean squared error (RMSE) and mean absolute error (MAE) compare predicted and processed observed price changes. RMSE gives greater weight to large errors; MAE measures average absolute error. Both are expressed in percentage points because the target is stored in percent. Spearman rank information coefficient (rank IC) is the correlation between predicted and observed

DOI: 10.54254/2754-1169/2026.LD38878

ranks, with tied values assigned average ranks. Unlike RMSE and MAE, it measures ordering rather than agreement in return levels.

For the portfolio comparison, test observations are sorted by predicted price change and divided into ten approximately equal-sized groups. D1 contains the lowest predictions and D10 the highest. The reported T–B statistic is the mean processed target in D10 minus that in D1, with equal weights within each group. Its unit is percentage points. This is a descriptive spread for one test cohort: the calculation includes neither trading costs nor dividends, financing costs, short-sale constraints, or risk adjustment. A positive spread alone does not establish statistical significance or achievable profitability.

4. Experimental results

Table 2 compares the five models on the 2018 financial-data cohort, using its subsequent 2019 target. LightGBM has the highest rank IC, 0.0805, followed by XGBoost at 0.0669. Its D10–D1 difference is 26.69 percentage points, compared with 24.01 for XGBoost. These are spreads in the processed target, rather than documented returns from an executed long–short strategy.

Table 2. Test performance on the 2018 financial-data cohort, ordered by rank IC

Model	RMSE	MAE	Rank IC	T–B (pp)
LightGBM	49.2106	35.1081	0.0805	26.69
XGBoost	48.9499	34.9690	0.0669	24.01
Lasso	48.5887	34.3607	0.0587	8.70
RandomForest	48.9946	34.6812	0.0459	18.67
Ridge	49.1807	35.0018	0.0449	9.24

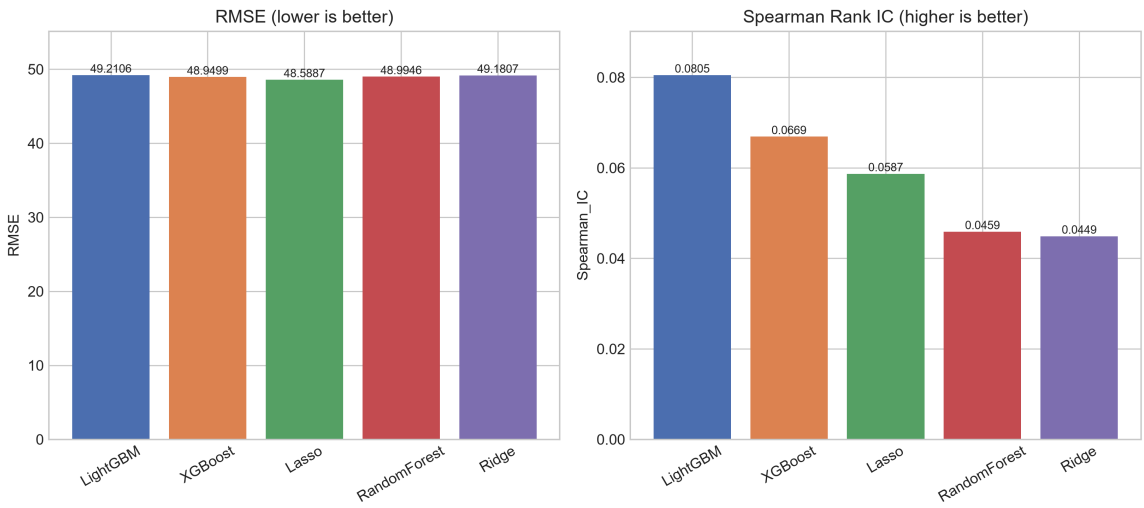


Figure 1. RMSE and Spearman rank IC by model

The error metrics give a different ordering. Lasso has the lowest RMSE (48.5887) and MAE (34.3607), whereas LightGBM's values are 49.2106 and 35.1081. Figure 1 displays the contrast between level error and rank correlation. Figure 2 shows the decile means used to calculate the spreads. All five rank IC estimates are positive, but the estimates are small and no confidence

DOI: 10.54254/2754-1169/2026.LD38878

intervals or repeated annual holdouts accompany this experiment. The results support a descriptive ranking comparison, rather than a statistically established advantage of nonlinear models.

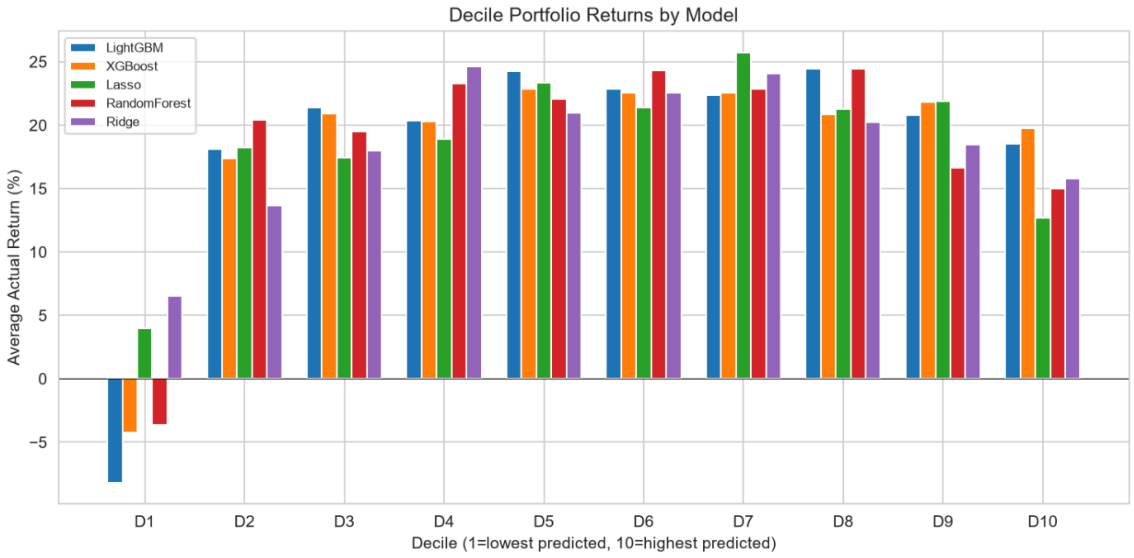


Figure 2. Decile means of the processed subsequent-year price-change target

5. SHAP-based model interpretation

5.1. Motivation and methodology

LightGBM is examined because it has the highest reported test rank IC. SHAP represents a prediction as a reference output plus feature contributions [7]. TreeExplainer provides the tree-model implementation [8]. In the supplied code, the global analysis uses a random sample of 2,000 test observations with seed 42, rather than every test observation. Separate explanations are computed for selected individual stocks. Since model choice for interpretation follows the test ranking, this analysis is exploratory.

5.2. Global feature importance

Figure 3 combines three views of the fitted LightGBM. Panel (a) plots the distribution of SHAP values for each displayed input; color denotes the input value, and horizontal position denotes its contribution to the prediction. Panel (b) orders features by their mean absolute SHAP contribution in the sampled test observations. This aggregation measures attribution magnitude without retaining its sign. A variable can therefore rank highly even when its contributions differ in direction across stocks.

The five leading inputs in Figure 3(b) are Earnings Yield, Weighted Average Shs Out, priceBookValueRatio, Sector_Energy, and Issuance (buybacks) of shares. These correspond to earnings yield, weighted-average shares outstanding, price-to-book ratio, Energy sector membership, and share issuance or buybacks. EPS ranks next. The chart therefore emphasizes valuation, share structure, and sector information; it does not place revenue growth or R&D expense growth among the top five. In panel (a), low earnings-yield values include large positive attributions, while high values cluster closer to zero or on the negative side. This pattern does not support a simple claim that the model reproduces a conventional value premium. Correlation among

DOI: 10.54254/2754-1169/2026.LD38878

accounting variables also means that the allocation of attribution depends on the fitted model and the SHAP setup.

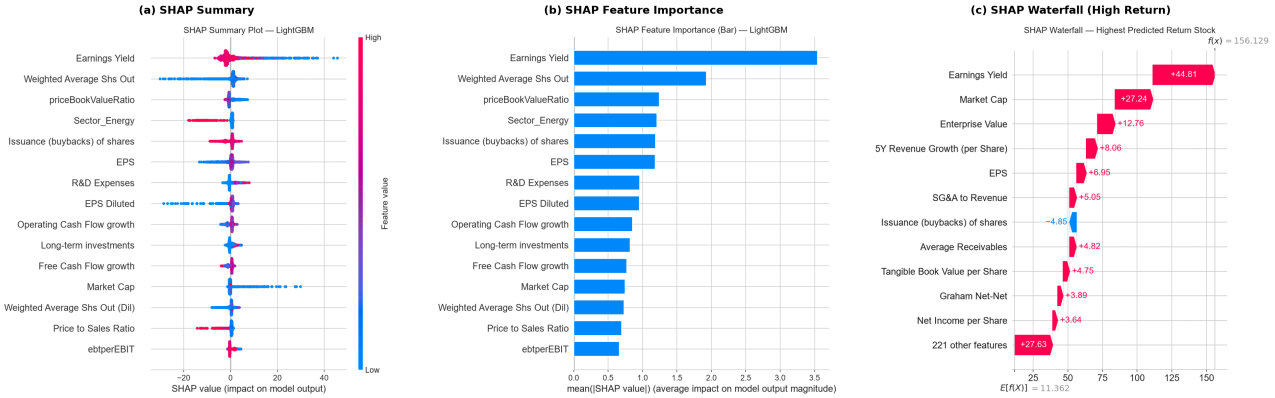


Figure 3. LightGBM SHAP results: (a) summary, (b) mean absolute attribution, and (c) highest-prediction observation

5.3. Local interpretability

Figure 3(c) explains the observation with the highest predicted price change. The reference output is 11.362 and the displayed prediction is 156.129. Earnings Yield contributes +44.81 percentage points, Market Cap +27.24, and Enterprise Value +12.76; Issuance (buybacks) of shares contributes −4.85. Other inputs account for the remaining difference, subject to rounding in the plot. These quantities explain how the fitted function produces this prediction. They do not show that the stock subsequently earned 156.129%, or that changing one accounting measure would cause a corresponding return change. A single high-prediction observation also cannot establish a general economic mechanism.

6. Sensitivity checks

6.1. Trimming extreme returns

Table 3 reproduces the reported sensitivity results after applying an absolute-target threshold of 200% to test observations. The code applies this rule to the already winsorized target, so it is not an independent filter on raw returns. LightGBM remains first by rank IC at 0.0704, followed by XGBoost at 0.0570 and Lasso at 0.0547. Lasso again has the lowest RMSE and MAE. The ordering by rank IC is unchanged, although each IC falls relative to Table 2. Lower error magnitudes after extreme-target exclusion are expected and, by themselves, do not establish robustness in a new population.

Table 3. Sensitivity results after excluding processed targets with absolute value above 200%

Model	RMSE	MAE	Rank IC
Ridge	44.8960	33.1517	0.0396
Lasso	44.0748	32.4605	0.0547
RandomForest	44.9069	32.8853	0.0358
XGBoost	45.0405	33.2257	0.0570

DOI: 10.54254/2754-1169/2026.LD38878

Table 3. (continued)

LightGBM	45.3615	33.3767	0.0704

6.2. Per-sector analysis

Separate LightGBM models are fitted for Financial Services, Healthcare, and Technology using each sector's training observations, numeric predictors, and training medians. Each fit uses 200 trees, maximum depth 5, and learning rate 0.05. Table 4 reports test rank IC values of 0.1059, 0.1008, and 0.1263, respectively. Healthcare has a substantially larger RMSE than Financial Services (74.8399 versus 28.7962). These differences describe sector-specific samples. Since the table does not compare each sector fit with the pooled model evaluated on exactly the same sector observations, it does not establish that sector-specific retraining improves accuracy.

Table 4. Separate LightGBM fits for the three largest test sectors

Sector	N_test	RMSE	Rank IC
Financial Services	824	28.7962	0.1059
Healthcare	691	74.8399	0.1008
Technology	636	54.4513	0.1263

7. Conclusion and future work

The five-model comparison gives different answers for ranking and level prediction. LightGBM records the highest rank IC, 0.0805, on the 2018 financial-data cohort with its 2019 target; Lasso records the lowest RMSE and MAE. LightGBM's top-minus-bottom decile difference is 26.69 percentage points on the processed target. Without uncertainty estimates or repeated test years, this result remains a sample-specific observation.

The SHAP plots identify the inputs used by the fitted LightGBM and allow individual predictions to be checked against their component attributions. Earnings yield and weighted-average shares outstanding lead the global ranking, followed by price-to-book ratio, the Energy indicator, and share issuance or buybacks. The observed directions are not uniformly consistent with a standard value or growth interpretation. Model attribution improves the description of the prediction function, but supplies neither causal identification nor independent confirmation of an asset-pricing anomaly.

The main constraints arise from the data and evaluation protocol. Full-sample winsorization uses held-out information, preprocessing is not refitted within each cross-validation fold, and accounting publication dates are unavailable. Only one subsequent return year is tested. In addition, the price-change target and portfolio calculations omit several components of investable total returns and trading costs. A stronger follow-up should estimate every preprocessing step inside a temporal training window, align predictors to documented disclosure dates, and evaluate raw adjusted returns across repeated rolling holdouts. Sector comparisons should use matched test observations, and explanation stability should be assessed after model refitting. These changes require new estimation; the existing tables cannot establish their likely outcome.

DOI: 10.54254/2754-1169/2026.LD38878

References

[1] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3–56.

[2] Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1–22.

[3] Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies, 29*(1), 5–68.

[4] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies, 33*(5), 2223–2273.

[5] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259*(2), 689–702.

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

[7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

[9] Lopez de Prado, M. (2018). *Advances in financial machine learning*. John Wiley and Sons.

[10] cnic92. (n.d.). 200+ financial indicators of US stocks (2014–2018). *Kaggle*. https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018

[11] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

[12] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B, 58*(1), 267–288.

[13] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[14] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.