

DOI: https://doi.org/10.48009/2_iis_2026_132

Predicting m&a announcement returns with machine learning: firm size, model complexity, and the role of time-consistent validation

Jianyu Ma, *Robert Morris University*, ma@rmu.edu

Yun Chu, *Robert Morris University*, chu@rmu.edu

John Stewart, *Robert Morris University*, stewartj@rmu.edu

Abstract

This study examines whether machine learning methods improve the prediction of abnormal returns around mergers and acquisitions (M&A) announcements. While event-study methods are widely used to estimate average market reactions, they are not designed for out-of-sample prediction at the transaction level. Six machine learning models—Support Vector Machines, Random Forests, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, and Multi-Layer Perceptrons—are evaluated under both time-ordered and random sampling frameworks. The results show that more complex models do not consistently exhibit noticeably higher predictive metrics than simpler alternatives in predicting abnormal returns. Predictive performance varies noticeably with firm size, with smaller firms exhibiting substantially higher predictability. In addition, random sampling yields noticeably inflated performance estimates relative to time-consistent validation, underscoring the importance of proper evaluation design. Overall, the findings suggest that predictive gains in M&A settings are driven more by data characteristics and firm heterogeneity than by model complexity, and that evaluation methodology plays a critical role in assessing true out-of-sample performance.

Keywords: Mergers and Acquisitions, Machine Learning, Return Predictability, Model Complexity, Time-Consistent Validation, Look-Ahead Bias

Introduction

Mergers and acquisitions (M&A) are among the most significant forms of corporate restructuring, and therefore, have been widely analyzed within the corporate finance literature. Prior studies have investigated whether M&A transactions create value for acquiring firms and shareholders (Ma et al., 2009; Moeller et al., 2003; Ross et al., 2022; Savor & Lu, 2009). Although existing research provides important insights into average market reactions surrounding acquisition announcements, less research has been conducted on whether abnormal returns can be systematically predicted using observable firm and deal characteristics.

Traditional empirical finance research has relied primarily on linear econometric models and hypothesis-driven specifications. While these approaches have provided important insights, they are often limited in capturing nonlinear relationships and complex interactions among variables. More recently, advances in machine learning have shifted attention toward out-of-sample forecasting performance, rather than in-sample inference (Gu et al., 2020; Mullainathan & Spiess, 2017). However, evidence regarding the effectiveness of machine learning methods in predicting corporate event outcomes remains limited and mixed.

Ma et al. (In press) indicate that KNN can be used to predict abnormal returns for bidding firms and yield a meaningful result only for small companies. However, their study does not examine whether the prediction results can be improved by using different machine learning algorithms. Building on prior research (Ma et al., In press), we examine whether more complex models yield superior predictive performance relative to simpler benchmarks and whether predictive ability varies systematically across firm characteristics. We focus on firm size as a key source of heterogeneity, given its well-documented role in information asymmetry and market efficiency.

To address these questions, we implement six widely used machine learning models—Support Vector Machines, Random Forests, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, and Multi-Layer Perceptrons—and evaluate their performance under a consistent out-of-sample framework. Importantly, we compare both time-ordered and random sampling schemes to assess the role of evaluation design in predictive accuracy. This setting allows us to address three related hypotheses. First, whether more complex machine learning models exhibit numerically higher predictive metrics than simpler alternatives in predicting abnormal returns. Second, whether predictive performance varies systematically with firm size. Third, whether time-consistent validation produces materially different performance estimates relative to conventional random sampling approaches. The results yield three main findings. First, we do not find strong evidence that more complex models consistently exhibit numerically higher predictive metrics than simpler benchmarks. Second, predictive performance is noticeably higher for smaller firms, consistent with greater informational friction in this segment. Third, random sampling substantially overstates predictive accuracy, highlighting the importance of time-consistent validation in financial prediction tasks. The remainder of the paper is organized as follows. Section 2 reviews the relevant literature and develops the hypotheses. Section 3 describes the data, explains how the variables are constructed, and outlines the empirical methodology and machine-learning framework. Section 4 presents empirical results and their interpretation. Section 5 summarizes key findings and discusses the implications. Section 6 addresses the limitations and Future Research.

Review of Literature and Research Hypotheses

The empirical literature on mergers and acquisitions (M&A) has primarily relied on event-study methodologies to estimate average abnormal returns around announcement dates (Ma et al., 2009; MacKinlay, 1997; Moeller et al., 2004). While these methods are well-suited for identifying average treatment effects and testing cross-sectional determinants of announcement returns, they are not designed for individual-level prediction. As a result, they provide limited guidance for forecasting the direction of abnormal returns in specific M&A transactions. More recently, advances in machine learning have shifted the focus of empirical finance from inference toward prediction. Machine learning methods provide flexible modeling frameworks that accommodate nonlinearities and high-dimensional interactions, relaxing the parametric assumptions typically imposed in traditional econometric models (Hastie et al., 2009; James et al., 2021). In financial applications, recent studies document that machine learning models can improve predictive performance in asset pricing and return forecasting tasks relative to traditional linear models, although gains are often context-dependent and sensitive to model specification and validation design (Feng et al., 2020; Gu et al., 2020). At the same time, concerns remain regarding overfitting, instability across samples, and limited robustness in noisy financial environments (Varian, 2014).

Despite this progress, two key gaps remain. First, there is limited evidence on the relative predictive performance of different machine learning algorithms for M&A announcement returns. Existing studies

typically focus on a single class of models or a narrow subset of methods, making it difficult to assess whether gains in predictive accuracy are driven by model complexity or data structure. Second, literature provides limited insight into cross-sectional variation in predictability, particularly with respect to firm size. Prior work documents that firm size is a fundamental determinant of M&A announcement effects (Ma et al., 2020; Moeller et al., 2004), reflecting differences in information asymmetry, analyst coverage, and market efficiency. These characteristics suggest that predictability may also vary systematically across firm size, although direct evidence remains scarce. To address these gaps, this study develops and tests the following hypotheses:

H1: More complex machine learning models do not necessarily exhibit noticeably higher predictive metrics than simpler models in predicting abnormal returns.

H2: Predictive performance varies systematically with firm size, with smaller firms exhibiting higher predictability.

H3: Time-consistent validation produces lower but more reliable estimates of predictive performance than random sampling approaches.

Methodology

This study employs a structured empirical framework that integrates standard event-study methods with machine-learning techniques to assess the predictability of abnormal returns in bidding firms' stock prices following M&A announcements. Specifically, abnormal returns are first estimated using conventional market-model procedures, and machine-learning models are then used to predict cross-sectional variation in these returns. The research design emphasizes consistency, comparability, and rigorous out-of-sample evaluation, consistent with recent advances in the financial machine-learning literature.

Data

The sample data consist of 1,253 completed merger and acquisition (M&A) transactions involving Chinese firms between 2010 and 2018. To examine heterogeneity in predictive performance, firms were partitioned into percentile-based subsamples according to total assets, as presented in Table 1 below. This approach enables systematic comparison of model performance across firm-size segments.

The sample construction procedures follow established conventions within the M&A literature and are consistent with prior studies (Ma et al., 2026). Abnormal returns were estimated using a standard event-study framework that was designed to isolate firm-specific excess returns surrounding acquisition announcements while controlling for broader market movements and systematic risk factors.

The dependent variable was defined as a binary indicator equal to 1 when the bidding firm experienced a positive abnormal return, and equal to 0 otherwise. Consistent with prior research in financial forecasting and machine learning applications (James et al., 2021), this specification frames the analysis as a directional prediction problem. Three event windows were examined: Day (+1), Day (+2), and the cumulative abnormal return over Days (+1) to (+2), denoted CAR (1,2).

Table 1. Description of Sample Size Used for Testing

% of Smallest Firms	Number of Firms	Training Data	Testing Data
15	188	132	56

20	250	175	75
25	313	219	94
30	376	263	113
40	501	351	150
50	626	438	188
100	1253	877	376

Explanatory Variables

The empirical analysis includes four explanatory variables that capture key characteristics of transactions. The *form of acquisition* (merger, stock acquisition, or asset acquisition) reflects differences in deal structure, integration complexity, and expected synergies. The *method of payment* (cash, stock, or hybrid) proxies for financing choices and signaling effects related to managerial confidence and firm valuation. The *industry relatedness*, defined as whether the acquirer and target operate within the same industry, captures the trade-off between operational synergies and diversification. The *firm size*, measured by total assets, reflects differences in information environment, market efficiency, and managerial behavior.

The four selected variables are the most fundamental, theoretically validated determinants of M&A announcement abnormal returns established in mainstream Chinese and international M&A literature:

- Form of acquisition: Captures integration complexity, synergy potential, and market valuation differences across merger, stock acquisition, and asset acquisition deals.
- Method of payment: Acts as a classic proxy for managerial overconfidence, acquirer valuation, and financing risk, a foundational predictor of bidder stock returns in event-study research.
- Industry relatedness: Distinguishes between synergy-driven horizontal/vertical acquisitions and diversification-motivated cross-industry transactions, a core cross-sectional driver of market reaction.
- Firm size (total assets): The central moderating variable of our paper, capturing information asymmetry, analyst coverage, and market efficiency gaps between small and large acquirers, serves as the basis for our heterogeneous subsample analysis.

Model Specification

The study evaluates six classification models to compare predictive performance across varying levels of model complexity: Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting (GB), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Multi-Layer Perceptrons (MLP). These models span a spectrum from parametric and relatively interpretable approaches to flexible, nonlinear algorithms, as commonly employed in the machine learning literature (Breiman, 2001; Hastie et al., 2009). The analysis focuses on comparative predictive performance within a unified empirical framework. All models are estimated to use the same dataset and feature set to ensure comparability across specifications. Hyperparameters are selected via five-fold cross-validation on the training sample to maximize out-of-sample performance. This procedure mitigates overfitting and enables consistent comparisons across models.

Data Processing and Validation

The dataset is preprocessed to ensure compatibility across model specifications. Categorical variables are encoded numerically, and firm size is log-transformed to reduce skewness and improve comparability across features. The empirical design employs a time-ordered sample split, with 70% of earlier observations used for model estimation and the remaining 30% reserved for out-of-sample evaluation. This approach preserves the temporal structure of the data, mimics realistic forecasting conditions, and avoids look-ahead bias (Varian, 2014).

Evaluation Metrics

Model performance is evaluated by using multiple classification metrics to provide a comprehensive assessment. Precision measures the accuracy of positive predictions, recall captures the ability to identify realized positive outcomes, and the F1-score balances these two dimensions. In addition, the area under the ROC curve (ROC-AUC) summarizes overall classification performance across thresholds.

Evaluating multiple metrics is important in classification settings where class distributions may be imbalanced and different types of prediction errors carry distinct economic implications (Saito & Rehmsmeier, 2015). To assess heterogeneity, all metrics are reported separately across firm-size subsamples, including the smallest 15%, 20%, 25%, 30%, 40%, 50%, and the full sample. This design allows for a detailed examination of how predictive performance varies across firm segments.

This study does not implement formal statistical significance tests to evaluate whether observed numerical gaps in Precision, Recall, F1-Score, and ROC-AUC between machine learning models reflect meaningful, statistically distinguishable performance differences. Two practical constraints prevent such testing within the current empirical framework. First, the out-of-sample test partitions are extremely small—resulting in insufficient statistical power for bootstrap, permutation, or paired classification difference tests, which would yield unreliable and uninformative p-values. Second, our stored empirical outputs only retain aggregated summary performance metrics for each model and subsample, rather than the complete raw predicted probabilities and individual holdout-sample classification outcomes required to execute distribution-based statistical comparison procedures.

Small-Sample Constraints and Mitigation Strategies

This study faces two interrelated empirical constraints inherent to our Chinese M&A dataset: a modest overall sample size of 1,253 transactions and drastically reduced observations within stratified firm-size subsamples, as well as a deliberate parsimonious feature space of only four explanatory predictors. Access to comprehensive, fully standardized M&A transaction databases covering Chinese listed acquirers is limited, restricting the total pool of usable complete deal records spanning 2010–2018. When partitioned by firm asset percentiles, the smallest subsample contains just 188 total observations. Such limited sample volume creates well-documented risks for nonlinear machine learning algorithms, including Random Forest, Gradient Boosting, and Multi-Layer Perceptrons: heightened susceptibility to the curse of dimensionality, unstable hyperparameter optimization, elevated overfitting risk, and unreliable out-of-sample performance estimates when feature counts expand. Following standard small-sample machine learning guidance from foundational statistical learning textbooks, we restrict the model input space to four theoretically essential, fully non-missing covariates—form of acquisition, payment method, industry relatedness, and firm size—rather than incorporating additional financial and deal characteristics proposed in broader M&A literature. Expanding the feature set would introduce dimensionality imbalance that our tiny training partitions cannot accommodate, breaking fair cross-model comparisons and rendering our tests of model complexity untrustworthy.

To offset biases and instability stemming from small sample sizes, we implement a series of rigorous mitigation procedures throughout the empirical pipeline. First, all hyperparameter tuning for every machine

learning model relies exclusively on five-fold cross-validation applied only to training data, avoiding leakage of test-set information and stabilizing parameter selection despite limited training observations. Second, we adopt temporally consistent train-test partitioning (70% historical training data, 30% future holdout sample) instead of random sampling; this eliminates upward-biased predictive performance estimates driven by look-ahead bias, a critical correction that yields conservative, more generalizable forecasting metrics even with limited data. Third, uniform preprocessing rules are enforced across all models: categorical deal attributes are numerically encoded, and firm size is logarithmically transformed to reduce skewness, compress extreme-value ranges, and improve convergence for nonlinear learners. Fourth, we report multiple complementary evaluation metrics (Precision, Recall, F1-Score, ROC-AUC) rather than relying on a single accuracy measure, enabling robust cross-verification of consistent performance patterns across all subsamples and model specifications.

It is important to clarify that the primary research objective of this analysis is not to maximize absolute predictive accuracy via large-scale high-dimensional machine learning, but to systematically compare directional performance differences across model complexity tiers, firm-size segments, and validation frameworks within a controlled experimental environment. While our small-sample setup limits the scope of feature engineering, the uniform, consistent empirical design yields coherent, replicable directional results aligned with our three core hypotheses. We acknowledge that future research using larger, more comprehensive proprietary Chinese M&A databases would be well positioned to expand the predictor set with additional financial and deal-level variables and to replicate our findings, further validating their external generalizability.

Empirical Results and Discussion

Prediction Performance Across Models

Table 2 reports out-of-sample predictive performance for six machine learning models using a time-ordered validation framework. The results directly support H1, which predicts that more complex machine learning models exhibit noticeably higher predictive performance than simpler alternatives for predicting abnormal returns. The evidence provides only partial support for H1. While flexible models such as SVM and MLP rank among the best-performing methods in smaller-firm subsamples, their performance advantage is neither uniform nor monotonic across firm sizes. In several subsamples, simpler models such as logistic regression and KNN achieve comparable predictive accuracy.

More importantly, Gradient Boosting does not consistently exhibit noticeably higher predictive metrics than simpler benchmarks, despite its higher model complexity. This suggests that greater algorithmic sophistication does not necessarily translate into superior predictive power in the presence of noisy financial data. Overall, the results indicate that model performance does not increase monotonically with complexity, and that predictive gains from advanced machine learning methods are conditional on firm characteristics rather than on the structural superiority of the algorithms. Accordingly, H1 is not supported in its strong form.

Time-Ordered vs. Random Sampling

Table 3 compares predictive performance across time-ordered and random sampling designs. These results directly test H3, which posits that time-consistent validation yields lower but more reliable performance estimates than conventional random sampling. The evidence partially supports H3. Across all models and subsamples, random sampling noticeably inflates predictive performance relative to the time-ordered framework.

This discrepancy is consistent with look-ahead bias, in which information from future observations contaminates the training process under random partitioning. Consequently, random sampling yields upward-biased estimates of true out-of-sample performance. In contrast, the time-ordered design yields more conservative but economically plausible performance estimates, reinforcing its suitability for financial forecasting applications. These findings confirm that the evaluation design materially affects empirical conclusions in machine-learning-based prediction settings.

Firm Size Effect

Tables 2–7 examine performance across firm-size subsamples and different event windows, providing evidence relevant to H2, which predicts that predictive performance varies systematically with firm size. The results provide strong support for H2. Across all models and specifications, predictive accuracy is noticeably higher among smaller firms and declines monotonically with firm size. In the smallest subsamples (15%–30%), models frequently achieve high F1-scores, whereas performance in the full sample approaches near-random classification levels.

This pattern aligns with the view that smaller firms face greater informational frictions, lower analyst coverage, and less efficient price discovery, which together produce more predictable return dynamics. In addition, predictability is time-sensitive, with stronger performance on Day (+1) than on Day (+2). This suggests that most of the exploitable information in M&A announcements is quickly incorporated into prices, limiting the persistence of predictable patterns. Importantly, the firm-size effect persists under both time-ordered and random sampling, confirming that it is not an artifact of the validation design but a structural feature of the data.

Table 2. Prediction Results Using Past to Predict Future with Different Classification Models for CAR (1,2)

Model	Precision	Recall	F1-Score	ROC-AUC
SVM	0.67/0.69/0.71/0.63/0.59/0.60/0.56	0.67/0.69/0.71/0.63/0.59/0.60/0.56	0.68/0.69/0.71/0.63/0.59/0.61/0.58	0.67/0.69/0.71/0.63/0.59/0.60/0.56
RF	0.58/0.68/0.63/0.67/0.63/0.62/0.55	0.58/0.68/0.63/0.67/0.63/0.62/0.55	0.57/0.68/0.63/0.67/0.63/0.64/0.55	0.58/0.68/0.63/0.67/0.63/0.62/0.55
GB	0.51/0.61/0.59/0.63/0.57/0.58/0.54	0.51/0.61/0.59/0.63/0.57/0.58/0.54	0.52/0.61/0.59/0.63/0.57/0.58/0.55	0.51/0.61/0.59/0.63/0.57/0.58/0.54
LR	0.63/0.69/0.70/0.60/0.59/0.60/0.56	0.62/0.69/0.70/0.60/0.59/0.60/0.56	0.61/0.69/0.70/0.59/0.59/0.60/0.58	0.62/0.69/0.70/0.60/0.59/0.60/0.56
KNN	0.67/0.71/0.70/0.67/0.57/0.53/0.54	0.67/0.71/0.71/0.67/0.57/0.53/0.54	0.66/0.71/0.71/0.67/0.57/0.53/0.55	0.67/0.71/0.71/0.67/0.57/0.53/0.54
MLP	0.64/0.73/0.71/0.69/0.65/0.61/0.59	0.62/0.73/0.70/0.66/0.61/0.59/0.55	0.58/0.72/0.69/0.64/0.63/0.63/0.63	0.62/0.73/0.70/0.66/0.61/0.61/0.59

Table 3. Testing Results Using Randomly Pooling Data with Different Classification Models for CAR (1,2)

Model	Precision	Recall	F1-Score	ROC-AUC
SVM	0.75/0.62/0.70/0.68/0.67/0.62/0.60	0.75/0.62/0.70/0.68/0.66/0.62/0.60	0.75/0.63/0.70/0.68/0.67/0.62/0.60	0.75/0.62/0.70/0.68/0.66/0.62/0.60
RF	0.62/0.57/0.61/0.59/0.64/0.63/0.58	0.61/0.57/0.60/0.58/0.64/0.63/0.58	0.63/0.59/0.61/0.60/0.64/0.63/0.58	0.61/0.57/0.60/0.58/0.64/0.63/0.58
GB	0.62/0.63/0.67/0.68/0.60/0.64/0.59	0.62/0.63/0.67/0.67/0.60/0.63/0.59	0.63/0.64/0.67/0.68/0.60/0.63/0.59	0.62/0.63/0.67/0.67/0.60/0.63/0.59
LR	0.72/0.61/0.69/0.67/0.70/0.62/0.61	0.70/0.60/0.69/0.67/0.70/0.62/0.61	0.70/0.61/0.69/0.67/0.70/0.62/0.61	0.70/0.60/0.69/0.66/0.70/0.62/0.61
KNN	0.64/0.63/0.65/0.63/0.64/0.59/0.59	0.63/0.63/0.64/0.63/0.63/0.59/0.59	0.65/0.64/0.64/0.64/0.64/0.59/0.59	0.63/0.63/0.64/0.63/0.63/0.59/0.59
MLP	0.66/0.64/0.73/0.73/0.71/0.67/0.60	0.66/0.64/0.73/0.73/0.71/0.67/0.60	0.67/0.64/0.72/0.73/0.70/0.67/0.60	0.66/0.64/0.73/0.73/0.70/0.67/0.60

Table 4. Time-Ordered Testing (F1) for Event Window Day (+1)

Model	15%	20%	25%	30%	40%	50%	100%
SVM	0.70	0.64	0.65	0.59	0.63	0.52	0.52
RF	0.62	0.56	0.56	0.54	0.67	0.52	0.47
GB	0.64	0.56	0.61	0.59	0.63	0.49	0.48
LR	0.68	0.61	0.65	0.59	0.59	0.51	0.52
KNN	0.64	0.57	0.61	0.59	0.67	0.49	0.44
MLP	0.70	0.61	0.63	0.60	0.64	0.52	0.52

Table 5 reports the F1-score results for the Day (+1) event window under the random sampling validation framework. Relative to the corresponding time-ordered results presented in Table 4, predictive performance is systematically higher across nearly all machine learning models and firm-size subsamples. In particular, Support Vector Machines (SVM) and Logistic Regression (LR) achieve F1-scores exceeding 0.70 within several smaller-firm categories, while K-Nearest Neighbors (KNN) and Multi-Layer Perceptrons (MLP) also demonstrate notable improvements.

Table 5. Random Sampling (F1) for Event Window Day (+1)

Model	15%	20%	25%	30%	40%	50%	100%
SVM	0.67	0.65	0.66	0.71	0.69	0.61	0.52
RF	0.63	0.55	0.64	0.60	0.60	0.50	0.47
GB	0.67	0.57	0.61	0.65	0.61	0.57	0.48
LR	0.70	0.65	0.63	0.72	0.69	0.63	0.52
KNN	0.72	0.63	0.60	0.66	0.62	0.62	0.44
MLP	0.72	0.65	0.64	0.72	0.66	0.62	0.52

Table 6 presents time-ordered test results for the Day (+2) event window. Relative to Day (+1) in Table 4, predictive performance is generally weaker and more variable across models. While some models—such as SVM and MLP—retain moderate performance on smaller subsamples (F1-scores around 0.60–0.68), the overall pattern indicates that predictability declines as the event window moves further from the announcement date. The firm-size effect persists but is less pronounced than on Day (+1). In several cases, performance differences across subsamples narrow, particularly for larger firm groups. This finding suggests that market adjustments for overreaction or slow reaction on Day (+1) may become more efficient over time, reducing the persistence of predictable patterns. The results indicate that short-term predictability is strongest immediately after announcements and diminishes quickly thereafter.

Table 7 presents the F1-scores for Day (+2) under random sampling. As in the comparison between Tables 4 and 5, the results again show higher performance than in the time-ordered framework in Table 6. Models such as SVM, LR, and MLP achieve F1-scores above 0.70 in smaller subsamples, indicating strong apparent predictive performance. However, this improvement is largely attributable to random sampling, which introduces information leakage across time periods. As in earlier comparisons, the discrepancy between time-ordered and random-sampling results underscores the risk of overestimating model effectiveness when temporal dependencies are ignored. Despite the higher reported metrics, the underlying predictive signal

for Day (+2) remains weaker than for Day (+1), reinforcing the conclusion that predictability diminishes over time.

Table 6. Time-Ordered Testing (F1) for Event Window Day (+2)

Model	15%	20%	25%	30%	40%	50%	100%
SVM	0.68	0.60	0.61	0.56	0.54	0.60	0.62
RF	0.57	0.55	0.52	0.52	0.47	0.55	0.58
GB	0.52	0.56	0.57	0.54	0.52	0.54	0.57
LR	0.61	0.55	0.56	0.52	0.53	0.51	0.61
KNN	0.66	0.55	0.61	0.55	0.50	0.54	0.57
MLP	0.59	0.61	0.60	0.56	0.52	0.59	0.63

Table 7. Random Sampling (F1) for Event Window Day (+2)

Model	15%	20%	25%	30%	40%	50%	100%
SVM	0.75	0.65	0.61	0.63	0.67	0.60	0.61
RF	0.59	0.60	0.54	0.58	0.61	0.56	0.55
GB	0.63	0.65	0.60	0.63	0.64	0.53	0.56
LR	0.70	0.72	0.61	0.63	0.67	0.60	0.60
KNN	0.63	0.60	0.62	0.61	0.60	0.58	0.57
MLP	0.74	0.67	0.62	0.64	0.66	0.59	0.60

Overall, this performance inflation is particularly evident in mid-sized subsamples (e.g., 30%–40%), where several models achieve their peak F1 scores. However, this improvement should be interpreted cautiously because random sampling disrupts the data's temporal structure and introduces potential look-ahead bias. As a result, the reported metrics likely overstate true predictive ability. The contrast between Tables 4 and 5 underscores the importance of time-ordered validation for obtaining realistic performance estimates in financial prediction tasks. However, performance deteriorates noticeably with larger subsamples, and most models converge to F1-scores near 0.50 in the full sample, indicating limited predictive power beyond random classification. Among the models, SVM and MLP show relatively stable performance across subsamples, whereas KNN and RF show greater variability. The results reinforce the presence of a firm-size effect and highlight the challenges of predicting short-term market reactions for larger firms under realistic, time-consistent validation.

As a result, performance estimates from random sampling are likely to be upward-biased and do not reflect the true out-of-sample predictive performance. In contrast, the time-ordered validation framework provides a more realistic assessment of model performance in practical forecasting settings, particularly in financial applications where temporal dependence is inherent (Guo et al., 2020). Taken together, these findings underscore the importance of temporal validation in financial machine learning applications. They also reinforce the broader conclusion that while machine learning models can provide meaningful predictive insights—especially for smaller firms—their effectiveness depends critically on both model selection and appropriate evaluation design.

Nuanced Performance Differences Across Validation Schemes: Random Variance vs. Systematic Bias

This study observes that some performance metrics from random sampling (RS) validation are lower than those from time-consistent (TC) validation, a pattern that requires explicit clarification to avoid misinterpretation. The two validation schemes differ fundamentally in their stability and reliability: TC validation uses a deterministic, fixed split of transactions by announcement date (70% earliest deals as training, 30% latest as test), with no randomness involved, producing a single, stable set of performance metrics. RS validation, by contrast, uses a stochastic random shuffle of all transactions to create the 70/30 train/test split, with no regard to deal timing, making it subject to random sampling variance.

The pattern of some RS metrics being lower than TC metrics is driven by three interconnected factors tied to our dataset: (1) extremely small firm-size subsamples (the smallest containing only 188 total observations) amplify random split variance, creating "unlucky" splits where the training set is not representative of the test set; (2) inherent class imbalance in the M&A abnormal return target variable creates sensitivity to mismatched class distributions between train and test sets, which random sampling can produce; and (3) nonlinear machine learning models are highly sensitive to distribution mismatches between training and test data, leading to poorer generalization in some random splits.

Crucially, this pattern does not contradict the core principle that RS upward-biases performance estimates. The systematically expected value of RS performance is reliably higher than TC performance, due to look-ahead bias and data leakage. Individual random splits, however, can produce metrics higher or lower than the expected value due to random variances, especially in tiny subsamples. This pattern illustrates why random sampling is unreliable for time-series data: it yields erratic, unstable performance estimates, whereas TC validation produces stable, consistent estimates that replicate the real-world prediction problem faced by market participants.

Key Findings and Implications

This study evaluates the predictive performance of six machine learning models for forecasting abnormal returns around M&A announcements, with explicit attention to model complexity, firm heterogeneity, and evaluation design. The empirical evidence yields three main findings that align with the hypotheses. First, H1 is not strongly supported. While flexible machine learning models such as SVM and MLP perform well in specific subsamples, particularly among smaller firms, the results do not indicate a systematic or monotonic advantage of more complex models over simpler alternatives. Logistic regression and KNN perform comparably in many settings, and Gradient Boosting does not consistently exhibit noticeably higher predictive metrics than benchmark methods. This suggests that predictive performance in financial markets is driven more by the structure and informational content of the data than by algorithmic complexity. Second, H2 is strongly supported. Predictive accuracy varies noticeably with firm size, with substantially higher performance observed among smaller firms. Across all models and event windows, smaller firms exhibit more predictable return dynamics, whereas performance deteriorates markedly in larger-firm subsamples. This pattern is consistent with the view that smaller firms are associated with greater information asymmetry, lower analyst coverage, and less efficient price discovery, which together enhance predictability. Third, H3 is partially supported. The comparison between time-ordered and random sampling demonstrates that random partitioning noticeably inflates predictive performance in some cases. This confirms that ignoring temporal dependence introduces look-ahead bias and leads to overly optimistic assessments of model effectiveness. Time-consistent validation, therefore, provides a more reliable basis for evaluating predictive models in financial settings.

The findings relate to the financial machine learning and market efficiency literature. Weak support for H1 questions the assumption that complex models always predict better, especially in markets with weak signals and noise, where simple models often perform well, supporting the value of parsimony in finance. Strong support for H2 indicates that market efficiency varies across firms, with firm size influencing forecast ability, aligning with classical asset pricing theory and extending it to machine learning. A partial support for H3 emphasizes evaluation design as an econometric issue; the disparity between random and time-ordered validation shows that ignoring temporal structure can overstate predictive ability, thereby affecting the interpretation of prior studies that rely on random cross-validation.

From a practical perspective, the results impact investors and decision-makers. The lack of an advantage for complex models (H1) suggests practitioners don't need sophisticated architecture for good performance. Simpler models like logistic regression can be just as effective, especially with noisy data. The strong firm-size effect (H2) shows that predictive strategies work better in less efficient market segments, notably smaller firms, informing targeted investments. Validation findings (H3) highlight the need for a larger sample and for avoiding random sampling in time-dependent data to prevent overly optimistic performance estimates.

Overall, the evidence suggests that predicting M&A-related abnormal returns depends less on model sophistication and more on data structure, firm characteristics, and evaluation methodology. Machine learning models can provide meaningful predictive insights, but their effectiveness is conditional and concentrated in specific market segments, particularly among smaller firms and in short post-announcement windows. These findings reinforce the view that in financial markets, predictive performance is ultimately constrained by market efficiency and data-generating processes rather than by algorithmic complexity alone.

From an economic perspective, the observed predictive gains—particularly F1-score improvements of 0.10-0.15 among smaller firms—appear practically meaningful. To illustrate, an investor employing an SVM-based strategy for the smallest 15% of firms would correctly identify approximately 67-70% of positive abnormal return events. While these estimates are promising, we caution that transaction costs, market impact, and short-sale constraints may erode realized returns, particularly for less liquid smaller-firm stocks.

A particularly important practical consideration concerns model interpretability. While complex models such as SVM and MLP achieved comparable or slightly superior performance in some subsamples, simpler approaches—particularly Logistic Regression and KNN—offer substantially greater interpretability... For practitioners in regulated financial institutions, the ability to explain model decisions to regulators, clients, and internal stakeholders may outweigh marginal improvements in predictive accuracy. Simpler models require less computational infrastructure, are easier to maintain and monitor, and are less prone to concept drift over time. We note that for many practitioners, these practical advantages may outweigh marginal improvements in predictive accuracy from more complex algorithms.

We further discuss the generalizability of our empirical results, given that our sample is restricted to M&A transactions involving Chinese listed firms completed between 2010 and 2018. Our three core documented patterns—the stronger return predictability for smaller acquirers, the systematic upward bias of random sampling relative to time-consistent validation, and the absence of consistent predictive advantages for complex machine learning models over simpler alternatives—derive from fundamental economic and statistical mechanisms that transcend China's specific market setting, implying the directional trends should hold across other capital markets, regulatory frameworks, and time horizons. The firm-size predictability

effect is rooted in universal information frictions: smaller firms everywhere suffer from thinner analyst coverage, weaker institutional monitoring, slower price incorporation of public news, and higher information asymmetry, all of which create short-run predictable return patterns around corporate events. The methodological advantage of time-ordered validation is a general principle applicable to all time-series financial datasets, as random data partitioning inevitably introduces look-ahead bias and temporal information leakage that inflate out-of-sample forecasting accuracy estimates, regardless of geographic jurisdiction or sample era.

Implications for AI and Financial Forecasting: Our findings have broader implications for the application of contemporary AI approaches in financial forecasting. The findings on time-consistent validation have direct implications for evaluating large language models (LLMs) and generative AI in financial applications. Just as random sampling introduces look-ahead bias in traditional machine learning, improper evaluation of LLMs (e.g., using overlapping time periods for training and testing when evaluating financial text analysis) can lead to inflated performance estimates. Model complexity and sample size constraints have implications for the use of pre-trained foundation models in finance. While foundation models may offer advantages when fine-tuned on large datasets, their benefits may be limited in small-sample financial prediction problems—consistent with our core finding that data characteristics often matter more than algorithmic sophistication. As the financial industry continues to adopt increasingly sophisticated AI technologies, careful attention to evaluation design, sample size constraints, and model interpretability will remain essential for generating reliable and actionable predictions.

Limitations and Future Research

An important caveat concerns the limited sample size, particularly within the smaller-firm subsamples, where training sets comprised as few as 132 observations. Under these conditions, the relatively modest performance of complex models—such as Random Forest, Gradient Boosting, and MLP—may reflect overfitting rather than a genuine inability of these methods to capture predictive signals in the data. This suggests that the performance of complex models relative to simpler alternatives may depend on sample size, and we encourage future research to test these relationships on larger datasets.

A richer feature set would provide additional insight into the determinants of predictive performance and potentially strengthen the machine learning models. These variables include, but are not limited to, relative deal size (Deal value / Bidder total assets/market capitalization), deal completion probability and competing bids, public vs. private target status, pre-acquisition financial performance: high profitability (ROE, ROA, operating margin), abundant free cash flow, growth opportunities (Tobin's Q, P/E ratio), corporate governance and ownership, and industry, market, and macroeconomic variables. We encourage future research to explore even richer features of engineering, including textual analysis of merger announcements and the use of alternative data sources.

While the time-consistent validation framework provides evidence of temporal generalizability within the sample period, the external validity of our findings beyond the Chinese context warrants discussion. China's unique institutional environment—including the prevalence of state-owned enterprises, regulatory approval requirements for M&A transactions, and evolving disclosure standards—may shape the magnitude and nature of abnormal returns in ways that differ from those in developed markets.

China's regulatory environment, state-owned enterprise involvement, and unique information disclosure requirements may limit the generalizability of our findings to developed markets (e.g., the U.S., Europe) or other emerging markets (e.g., India, Brazil). Our sample period (2010-2018) predates significant recent

regulatory changes in China and excludes the COVID-19 pandemic period. We encourage future research to test whether our findings are held in more recent years and during periods of market stress.

References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3), 1327–1370. <https://doi.org/10.1111/jofi.12883>
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
- Guo, J. M., He, Q., Xin, J., & Liu, J. (2020). Managerial overconfidence and M&A performance: Evidence from China. *International Journal of Banking, Accounting and Finance*, 11(3), 342–360.
<https://doi.org/10.1504/IJBAAF.2020.107946>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
- Ma, J., Chu, Y., & Beaves, R. (2020). Firm size effects on stock returns in mergers and acquisitions in Asian emerging markets. *International Journal of Business and Systems Research*, 14(4), 398–410.
- Ma, J., Chu, Y., & Stewart, J. (2026). Prediction of abnormal returns on bidding firms in mergers and acquisitions [Unpublished manuscript]. *International Journal of Revenue Management*.
- Ma, J., Pagán, J. A., & Chu, Y. (2009). Abnormal returns to mergers and acquisitions in ten Asian stock markets. *International Journal of Business*, 14(3), 235–250.
- Ma, J., Pagán, J. A., Chu, Y., & Soydemir, G. (2012). Revenue enhancement through mergers and acquisitions: Wealth effects of method of payment. *International Journal of Revenue Management*, 6(3), 274–290. <https://doi.org/10.1504/IJRM.2012.050402>
- MacKinlay, A. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1), 13–39.
- Moeller, S., Schlingemann, F., & Stulz, R. (2003). Do shareholders of acquiring firms gain from acquisitions?
- Moeller, S., Schlingemann, F., & Stulz, R. (2004). Firm size and the gains from acquisitions. *Journal of Financial Economics*, 73(2), 201–228. <https://doi.org/10.1016/j.jfineco.2003.07.002>
- Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *The Journal of Economic Perspectives*, 31(2), 87–106.

Ross, S., Westerfield, R., & Jaffe, J. (2022). *Corporate finance* (13th ed.). McGraw-Hill Education.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>

Savor, P. G., & Lu, Q. (2009). Do stock mergers create value for acquirers? *The Journal of Finance*, 64(3), 1061–1097.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. <https://doi.org/10.1257/jep.28.2.3B>