Proceedings of CONF-CDS 2026 Symposuim: Data-Centric AI Security: Securing Models, Learning Agents, and Autonomous Systems
DOI: 10.54254/2755-2721/2026.CH34669

Leakage-Safe Machine Learning for Short-Horizon Volatility Forecasting Using Disclosure Signals

Yunlin Shi

Department of Accountancy, Shanghai University of Finance and Economics, Shanghai, China
2024110273@stu.sufe.edu.cn

Abstract. Predicting short-horizon volatility is a difficult machine-learning problem in finance because the data are noisy, temporally dependent, and sensitive to leakage. This paper asks whether simple disclosure-based signals improve next-day volatility forecasting under a leakage-safe walk-forward design and against the strong benchmarks. Using a daily panel of the 800 A-share stocks from 2020 to 2025, next-day absolute returns are predicted from price-based, disclosure-based, and volatility-history features. The empirical framework combines the expanding-window walk-forward testing, train-only thresholding, and comparisons across naive different baselines including Ridge regression, and XGBoost. The results show that the volatility-history benchmarks explain most forecastable variation, so disclosure variables deliver only modest average gains once leakage is controlled for. Their predictive value is more visible in rare high-attention states with heterogeneous same-day disclosures, but the evidence remains statistically fragile because such states are sparse. Overall, disclosure signals appear more useful for conditional than for broad unconditional volatility forecasting.

Keywords: volatility forecasting, disclosure signals, machine learning, leakage

1. Introduction

Forecasting short-horizon volatility remains difficult, even though volatility is central to many problems in finance. A large literature shows that volatility is clustered, time-varying, and hard to model, motivating econometric approaches ranging from Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models to realized-volatility frameworks [1-3]. More recently, machine learning has renewed interest in volatility forecasting by providing flexible predictive tools that can capture nonlinearities and interactions in noisy financial environments. This motivation is especially relevant in short-horizon financial prediction, where the data are noisy, the signal-to-noise ratio is low, temporal dependence is strict, and predictive relationships may shift over time. In such settings, the value of machine learning lies not simply in predictive flexibility, but in whether nonlinear structure can be learned in a way that survives genuinely out-of-sample evaluation under a real-time information set.

A related literature studies how information arrival, firm disclosures, and investor attention affect trading activity and volatility. Classic evidence shows that major firm announcements are associated

DOI: 10.54254/2755-2721/2026.CH34669

with abnormal trading volume and heightened return volatility [4]. Later work further suggests that public information arrival and investor attention help explain variation in returns and volatility [5-8]. These findings imply that disclosure-based variables may contain information relevant for short-horizon volatility forecasting. However, most of this literature focuses on market reactions to information arrival rather than on whether simple, real-time disclosure measures improve stock-level forecasts in a true out-of-sample setting.

The central issue here is forecastability rather than association. In panel forecasting, measured performance can change materially with the design of the forecasting exercise. The forecasting literature has long treated out-of-sample testing as part of credible empirical evaluation rather than as a routine robustness check [9]. The data-mining literature likewise shows that leakage can substantially inflate predictive performance [10]. In financial panels, even small implementation choices, such as defining thresholds using the full sample rather than the training window, can introduce future information and overstate results. More importantly, simple linear specifications may be too restrictive for this problem. Disclosure signals may interact with recent price movement, generate threshold-like volatility responses, and matter only in particular information states. Such patterns are difficult to represent with linear additivity alone, but are natural targets for modern machine-learning models designed to capture nonlinear effects and feature interactions. XGBoost, in particular, is well suited to this setting because tree boosting can model nonlinear decision structure and interactions efficiently while remaining competitive on tabular prediction tasks.

The empirical analysis examines whether simple disclosure-based signals provide incremental predictive value for next-day volatility forecasts beyond strong price- and volatility-based baselines in a daily panel of 800 A-share stocks from 2020 to 2025. The A-share market is a useful setting because firm disclosures are frequent and often clustered within the same day, creating substantial variation in firm-level information environments. The predictors are observable by the end of day and include current-day absolute return, the number of firm disclosures, and a disclosure-conflict indicator that captures heterogeneity in same-day disclosure content. The core predictive analysis uses XGBoost to model disclosure–volatility interaction structure under an expanding-window walk-forward design, while Ridge regression and naive volatility proxies are retained as strong benchmark models. To avoid leakage, all threshold-based variables are defined using the training sample only and then carried forward to the corresponding test period. The results indicate that strong volatility-history baselines already account for most predictable variation in next-day volatility, leaving limited room for stable incremental predictive value from simple disclosure variables on average. Once leakage is ruled out, disclosure activity delivers only modest predictive gains in the full sample, and the remaining gains are concentrated in rare high-attention states with heterogeneous disclosures. Overall, the evidence suggests that disclosures may coincide with higher volatility without delivering robust average out-of-sample forecasting improvements.

2. Method

2.1. Task formulation

This study examines a supervised stock-level panel forecasting task: predicting next-day short-horizon volatility using information observable by the end of the current trading day. The empirical design emphasizes reproducible out-of-sample evaluation rather than in-sample fit, because predictive conclusions in financial panels are highly sensitive to time ordering and implementation details [9, 10]. Accordingly, the analysis combines association tests with a leakage-safe walk-forward forecasting protocol. From a machine-learning perspective, the objective is not merely to

DOI: 10.54254/2755-2721/2026.CH34669

test whether disclosure variables are statistically related to future volatility, but to determine whether they improve predictive performance under a deployment-realistic temporal validation setting.

2.2. Data and prediction target

The sample is constructed from daily A-share stock data and firm announcement records obtained through Tushare Pro, a financial data platform that provides market and corporate disclosure data through SDK and API interfaces [11]. The study uses a daily panel of 800 A-share stocks from 2020 to 2025. For each stock-day observation, the prediction target is next-day volatility, proxied by the absolute value of the next trading day's return, denoted as $y_{i,t+1} = |r_{i,t+1}|$. Using absolute return as a volatility proxy is standard in empirical forecasting settings when the object of interest is short-horizon variation rather than latent conditional variance itself [3].

2.3. Feature construction

The predictor set combines price-based and disclosure-based information. The baseline price-side predictor is abs_return, which captures recent realized movement and serves as the main price benchmark. Disclosure-side variables are constructed from firm announcement titles. First, ann_count is defined as the number of announcements released by a firm on a given day and is used as a simple attention or information-arrival measure, consistent with the literature linking public information arrival to volatility [4, 5]. Second, conflict_day is a binary indicator equal to one when positive-type and negative-type announcements appear on the same day for the same firm, capturing heterogeneity in same-day disclosure content. Third, sent_mean is the average title-level sentiment score computed from a rule-based dictionary over announcement titles. In addition, several historical-volatility features are constructed for stronger baselines: the 5-day rolling mean of absolute returns, the 10-day rolling mean of absolute returns, the 5-day rolling standard deviation of returns, and a 5-day exponentially weighted moving average (EWMA) of absolute returns. These variables provide simple volatility-history benchmarks alongside the disclosure features.

2.4. Temporal alignment and leakage control

To preserve the real-time information set, all announcement-based variables are lagged by one trading day within each stock before modeling. In the feature-construction pipeline, disclosure aggregates merged on calendar date t are shifted so that the predictors used on row t correspond to the previous trading day's disclosure aggregates rather than unlagged same-day announcements. Thus, the model combines contemporaneous price-based variables with lagged disclosure information to forecast $y_{i,t+1}$, avoiding look-ahead bias [10]. Missing announcement-side values after merging are filled with zeros, so non-announcement days enter the sample as zero-disclosure days. The final target variable is then created by shifting abs_return one day forward within each stock. No full-sample normalization is applied. This avoids introducing transformations that mix information across time and is also unnecessary for the tree-based models used later. From an ML perspective, this step is central: it ensures that all feature engineering remains consistent with the information that would be available at prediction time.

DOI: 10.54254/2755-2721/2026.CH34669

2.5. Walk-forward evaluation protocol

The main forecasting evaluation uses an expanding-window walk-forward design [9]. Let the available years be 2020 through 2025. The first test fold trains on 2020 and tests on 2021, after which the training window expands year by year. For example, a later fold trains on 2020–2022 and tests on 2023, and the next fold trains on 2020–2023 and tests on 2024. This procedure is repeated year by year, and performance is aggregated across all test folds. This design is preferred to a random 80/20 split because random splitting breaks the temporal structure of financial forecasting and can contaminate the test set with information patterns that would not be available in real time [9, 10]. In ML terms, the walk-forward design evaluates whether learned predictive structure remains stable under strict temporal dependence and evolving market conditions, rather than under an artificially easy random split.

The same principle is applied to subsample definitions. HighAttention is defined within each fold using the top 20% of ann_count among stock-days with at least one announcement, but the threshold is estimated on the training segment only and then applied to the corresponding test segment. HighUncertainty is defined analogously from the uncertainty variable; because conflict_day is binary in the main specification, the operational rule reduces to conflict_day = 1 in the test segment when appropriate. This training-only thresholding is a key anti-leakage step [10]. It also matters substantively, because state-dependent predictive patterns would otherwise be overstated by thresholds estimated with future information.

2.6. Models and baselines

Association tests are first conducted using ordinary least squares (OLS). The OLS specifications are intended to evaluate linear association and hypothesis consistency, rather than to establish the paper's main forecasting result. In particular, the regressions examine whether ann_count and conflict_day are positively associated with next-day volatility after controlling for abs_return.

The main predictive analysis then compares several baselines and machine-learning models. Two constant benchmarks, Mean and Median, predict all test observations using the training-sample mean or median of the target. Two naive volatility baselines, Naive Roll-5 and Naive EWMA-5, use same-day rolling or exponentially weighted volatility proxies as direct forecasts. Linear predictive models are implemented with Ridge regression, which applies L2 regularization to stabilize estimation under correlated predictors [12]. Three Ridge variants are used: Ridge (Price), Ridge (Vol), and Ridge (Full). Nonlinear models are implemented using XGBoost, a gradient-boosted tree method well suited to modeling nonlinear effects and feature interactions in tabular data [13]. The XGBoost specifications include XGB (Price), XGB (Vol), XGB (+Attention), XGB (+Text), and XGB (Full). The full model combines current-day price information, disclosure intensity, disclosure conflict, and title-based sentiment. In this setup, XGBoost is the core nonlinear learner used to test whether disclosure–volatility interaction structure, threshold effects, and state-dependent patterns can be captured more effectively than by linear baselines.

2.7. Evaluation metrics and statistical tests

Forecast accuracy is evaluated using mean absolute error (MAE), root mean squared error (RMSE), and Spearman information coefficient (IC). In the main walk-forward evaluation, IC is measured as the pooled Spearman rank correlation across all stock-day observations in the relevant out-of-sample sample. By contrast, the state-dependent analysis reports IC as the average of daily cross-sectional

DOI: 10.54254/2755-2721/2026.CH34669

Spearman rank correlations. MAE and RMSE measure absolute and squared prediction error, respectively, while IC captures rank-order predictive association between forecasts and realized next-day volatility. Because the IC definitions differ across these two settings, the corresponding values are not intended to be compared one-for-one across tables.

Full-sample results are reported after concatenating all out-of-sample test folds. Additional results are also reported for HighAttention and HighUncertainty subsamples. To assess incremental predictive value, the analysis compares richer models against the price-only benchmark at the fold level. In particular, the differences between XGB (Full) and XGB (Price), and between XGB (+Text) and XGB (Price), are tested using paired t-tests and Wilcoxon signed-rank tests across walk-forward folds. Finally, robustness is examined by varying the HighAttention cutoff across multiple top_pct values rather than relying on a single threshold choice. This combination of aggregate metrics, fold-level paired tests, and robustness analysis is intended to evaluate not only average predictive performance, but also the stability of incremental gains across temporally ordered test folds.

3. Results and discussion

3.1. Full-sample predictive performance

Table 1 reports full-sample walk-forward forecasting performance across all benchmark and machine-learning models. In this table, IC is computed as a pooled Spearman rank correlation after concatenating all stock-day observations in the out-of-sample test folds, so the reported IC values summarize overall rank association in the pooled test sample rather than date-by-date cross-sectional ranking performance. Several findings stand out. First, volatility-history baselines clearly dominate disclosure-based specifications in the full sample. XGB (Vol) achieves the best overall performance, with MAE = 0.0145, RMSE = 0.0227, and IC = 0.3300, while Ridge (Vol) performs similarly. By contrast, the price-only and disclosure-augmented models cluster much closer together: XGB (Price) records MAE = 0.0153, RMSE = 0.0235, and IC = 0.1877, whereas XGB (Full) yields MAE = 0.0153, RMSE = 0.0235, and IC = 0.1885. Thus, adding announcement intensity, conflict, and text tone produces only a very small improvement in pooled IC over the price-only benchmark. The broader implication is that short-horizon volatility is already strongly captured by recent realized-volatility information, leaving limited room for simple disclosure variables to generate large unconditional gains.

Table 1. Full-sample prediction performance (walk-forward evaluation)

Model	MAE	RMSE	IC	N
Mean (Train)	0.0161	0.0244	-0.0208	928,657
Median (Train)	0.0146	0.0252	-0.0603	928,657
Naive Roll-5	0.0152	0.0241	0.3091	928,657
Naive EWMA-5	0.0149	0.0237	0.3199	928,657
Ridge (Price)	0.0153	0.0235	0.1903	928,657
Ridge (Vol)	0.0145	0.0227	0.3264	928,657
Ridge (Full)	0.0153	0.0235	0.1906	928,657
XGB (Price)	0.0153	0.0235	0.1877	928,657
XGB (Vol)	0.0145	0.0227	0.3300	928,657
XGB (+Attention)	0.0153	0.0235	0.1886	928,657

DOI: 10.54254/2755-2721/2026.CH34669

Table 1. (continued)

XGB (+Text)	0.0153	0.0235	0.1883	928,657
XGB (Full)	0.0153	0.0235	0.1885	928,657

Notes: MAE = mean absolute error; RMSE = root mean squared error; IC = pooled Spearman correlation. Metrics are aggregated across yearly walk-forward test folds.

3.2. Fold-level stability and incremental value

Table 2 summarizes fold-level stability for selected models in the full sample. The pattern is stable across the five walk-forward test folds. XGB (Price) has an average IC of 0.1872 with a fold standard deviation of 0.0108, while XGB (Full) has an average IC of 0.1880 with a standard deviation of 0.0098. The difference is therefore positive but economically small. The same table also shows that volatility-history models remain consistently stronger across folds, with XGB (Vol) averaging IC = 0.3111 and Ridge (Vol) averaging IC = 0.3086. This result is important for interpretation: the paper is not comparing disclosure variables against weak benchmarks, but against strong price and volatility baselines under a leakage-safe out-of-sample design.

The fold-level paired tests further clarify the size of incremental gains. Relative to XGB (Price), XGB (Full) improves MAE enough to pass the paired t-test at the 5% level ($t = -3.4628$, $p = 0.0258$), but not the Wilcoxon test ($p = 0.0625$). For RMSE and pooled IC, the differences are not statistically significant; the fold-level test for IC yields $p = 0.1505$. A similar pattern holds for XGB (+Text): MAE improves under the paired t-test ($t = -2.8737$, $p = 0.0453$), whereas RMSE and IC remain insignificant. These results suggest that disclosure variables may slightly reduce absolute prediction error, but do not consistently improve rank-order forecasting performance in the full sample.

Table 2. Walk-forward performance across folds (selected models, full sample)

Model	MAE mean (std)	RMSE mean (std)	IC mean (std)	Avg. N	Folds
XGB (Vol)	0.0145 (0.0016)	0.0227 (0.0023)	0.3111 (0.0191)	185,731	5
Ridge (Vol)	0.0146 (0.0016)	0.0227 (0.0023)	0.3086 (0.0172)	185,731	5
XGB (Price)	0.0153 (0.0012)	0.0235 (0.0023)	0.1872 (0.0108)	185,731	5
XGB (+Text)	0.0153 (0.0012)	0.0235 (0.0023)	0.1879 (0.0099)	185,731	5
XGB (Full)	0.0153 (0.0012)	0.0235 (0.0023)	0.1880 (0.0098)	185,731	5

Notes: Each entry reports the mean and standard deviation across the five yearly walk-forward test folds.

3.3. Subsample performance

Table 3 reports the corresponding subsample results for HighAttention and HighUncertainty observations. The subsample analysis sharpens the full-sample conclusion rather than reversing it. In the HighAttention subsample, disclosure-augmented XGBoost models do not outperform the price-only benchmark: XGB (Price) achieves IC = 0.1825, whereas XGB (Full) and XGB (+Text) fall to 0.1720 and 0.1729, respectively. In the HighUncertainty subsample, the same pattern remains: XGB (Price) reaches IC = 0.2310, while XGB (Full) drops to 0.1827. More importantly, the volatility-history models again dominate both subsamples. This means that broad attention or uncertainty partitions alone are not sufficient to generate average gains from disclosure variables.

DOI: 10.54254/2755-2721/2026.CH34669

Table 3. Subsample prediction performance (highattention and highuncertainty)

Subsample	Model	MAE	RMSE	IC	N
HighAttention	XGB (Price)	0.0159	0.0230	0.1825	21,794
HighAttention	XGB (Vol)	0.0152	0.0220	0.3153	21,794
HighAttention	XGB (+Text)	0.0161	0.0229	0.1729	21,794
HighAttention	XGB (Full)	0.0161	0.0230	0.1720	21,794
HighUncertainty	XGB (Price)	0.0153	0.0215	0.2310	641
HighUncertainty	XGB (Vol)	0.0149	0.0207	0.3566	641
HighUncertainty	XGB (+Text)	0.0157	0.0215	0.1971	641
HighUncertainty	XGB (Full)	0.0159	0.0218	0.1827	641

Notes: Representative XGBoost specifications are reported to highlight the comparison between price-only, volatility-history, and disclosure-augmented models in each subsample.

3.4. State-dependent mechanism

Table 4 and Figure 1 reveal a more nuanced state-dependent mechanism. The state experiment partitions the test sample into mutually exclusive groups: routine days ($S0$), attention-only days ($S1$), uncertainty-only days ($S2$), and attention-plus-uncertainty days ($S3$). Unlike Table 3, the state analysis defines IC as the average of daily cross-sectional Spearman rank correlations rather than as pooled Spearman over the concatenated sample, so these IC values should be interpreted within the state analysis and not compared numerically one-for-one with the pooled IC values reported earlier. Table 4 shows that the incremental value of disclosure variables is negligible on routine days and negative on attention-only days, but much larger in the combined $S3$ state. Specifically, in $S3$ the IC of XGB (Price) is 0.1315, while the IC of XGB (Full) rises to 0.1835, implying $\Delta IC = 0.0520$. By contrast, S0 shows only $\Delta IC = 0.0006$, and $S1$ shows $\Delta IC = -0.0029$. Figure 1 visually reinforces this pattern: the largest positive improvement appears in $S3$, whereas $S0$ is close to zero and $S1$ is slightly negative.

At the same time, the statistical evidence for $S3$ remains limited because the state is rare. The fold-level state tests report mean $\Delta IC = 0.0561$ in $S3$, but with high dispersion across folds and only four valid folds, so the one-sample t-test is not significant (p = 0.5278). Likewise, the paired comparison of S3 versus S0 yields no significant difference in either ΔIC or ΔMAE. A more accurate conclusion is therefore that the incremental value of disclosure information is economically concentrated in the narrow Attention + Uncertainty state, but the evidence remains imprecise because such observations are sparse.

Table 4. Marginal value of disclosure signals by information state

State	N	MAE (Price)	MAE (Full)	ΔMAE	Mean Δ|e|	IC (Price)	IC (Full)	ΔIC
S0: Routine	906,863	0.0153	0.0153	0	0	0.1746	0.1752	0.0006
S1: Attention only	21,477	0.0159	0.0161	0.0002	0.0002	0.1551	0.1522	-0.0029
S2: Uncertainty only	--	--	--	--	--	--	--	--
S3: Attention + Uncertainty	317	0.0161	0.0166	0.0006	0.0006	0.1315	0.1835	0.052

Notes: $\Delta MAE = MAE(Full) - MAE(Price)$; $\Delta IC = IC(Full) - IC(Price)$. In this table, IC is defined as the average of daily cross-sectional Spearman correlations.

DOI: 10.54254/2755-2721/2026.CH34669

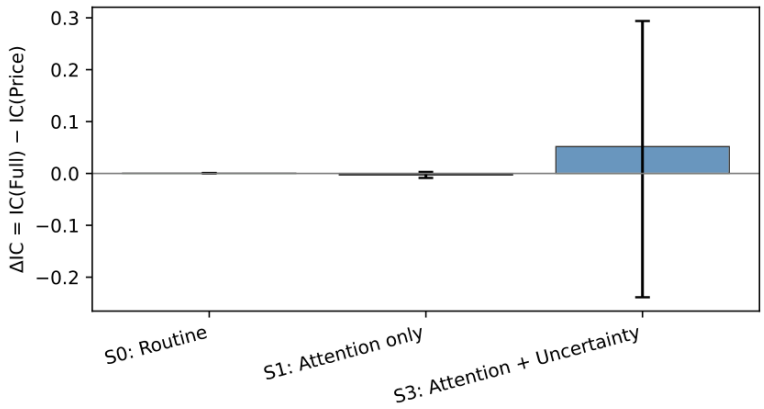

Figure 1. State-dependent marginal value of disclosure signals (picture credit: original)

3.5. Robustness to the highattention threshold

Finally, the robustness exercise confirms that the weak average performance of disclosure variables in the HighAttention subsample is not driven by a single arbitrary threshold choice. When the HighAttention cutoff is varied across top 10%, 20%, and 30%, XGB (Price) remains stronger than XGB (Full) and XGB (+Text) throughout within that subsample. Thus, the main message is stable: disclosure variables do not generate broad average improvements in the HighAttention subsample, and their value appears to be concentrated in a much narrower subset of days characterized by both elevated attention and heterogeneous disclosures.

3.6. Discussion

Overall, the evidence supports a restrained interpretation. Strong volatility-history baselines explain most of the forecastable variation in next-day absolute returns, and simple disclosure variables do not add large unconditional gains on average. However, disclosure signals are not irrelevant. Their predictive value appears conditional rather than universal, becoming most visible in rare states where attention is high and same-day disclosures are heterogeneous. This result fits the central argument of the paper: disclosure variables may be associated with volatility, but their true forecasting value depends critically on evaluation design, baseline strength, and information state. From a machine-learning perspective, the main lesson is not that nonlinear models automatically unlock large gains, but that evaluation discipline matters: once walk-forward validation, strong baselines, and train-only thresholding are enforced, the remaining predictive contribution of disclosure signals is much narrower than association-based evidence alone would suggest.

4. Conclusion

This paper examines whether simple disclosure-based signals improve next-day volatility forecasting in a stock-level A-share panel under a leakage-safe out-of-sample design. Using daily data for 800 stocks from 2020 to 2025, the analysis combines lagged disclosure features with price-based predictors in an expanding-window walk-forward framework. The results show that strong volatility-history baselines explain most forecastable variation in next-day absolute returns, leaving limited room for broad average gains from disclosure variables. Once leakage is controlled for, announcement intensity, disclosure conflict, and title sentiment provide only modest incremental improvements on average. Their predictive value appears more concentrated in rare high-attention

DOI: 10.54254/2755-2721/2026.CH34669

states with heterogeneous disclosures, although the evidence remains statistically fragile because such states are sparse. Overall, disclosure signals are more useful for conditional than for broad unconditional volatility forecasting.

References

[1] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1007.
[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.
[3] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579-625.
[4] Beaver, W. H. (1968). The information content of annual earnings announcements. Journal of Accounting Research, 6, 67-92.
[5] Kalev, P. S., Liu, W.-M., Pham, P. K., & Jarnecic, E. (2004). Public information arrival and volatility of intraday stock returns. Journal of Banking & Finance, 28(6), 1441-1467.
[6] Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461-1499.
[7] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168.
[8] Andrei, D., & Hasler, M. (2015). Investor attention and stock market volatility. The Review of Financial Studies, 28(1), 33-72.
[9] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437-450.
[10] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), Article 15.
[11] Tushare Pro. (2026). Official platform and documentation. Available at: https: //tushare.pro/
[12] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.
[13] 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).