



SCIENTIFIC OASIS

Journal of Intelligent Decision Making and Granular Computing

Journal homepage: www.jidmag.org
ISSN: 3042-3759

JIDMGC

Journal of
Intelligent
Decision
Making and
Granular
Computing

Scientific Oasis

SCIO <https://doi.org/10.31181/jidmgc31202754>

Predicting First Observed Appearance in a Corporate Governance Rating List and Subsequent Upgrades with Interpretable Machine Learning

Abdullah Kürşat Merter^{1,*}

¹ Department of Business Administration, Gebze Technical University, Gebze, Kocaeli, 41400, Türkiye

ARTICLE INFO

Article history:

Received 7 Jun 2026
Received in revised form 22 August 2026
Accepted 9 October 2026
Available online 12 October 2026

Keywords:

Corporate governance; Machine learning;
Explainable artificial intelligence; Index
certification; Rare events

ABSTRACT

Corporate governance certification provides a public assessment of organizational practices. Previous research has primarily examined the relationship between certification and firm outcomes, while the timing of certification adoption represents a distinct prediction problem. This study examines whether the first observed appearance on a governance rating list and subsequent rating upgrades can be predicted using statutory accounting disclosures. Rating history is also incorporated into the upgrade model. The institutional setting is the Borsa Istanbul Corporate Governance Index, where licensed agencies evaluate four governance sections and eligibility is determined by published thresholds. The dataset combines rating records for 49 non-financial firms from 2013 to 2024 with financial statements from 438 listed non-financial firms. Two interpretable modeling modules employ an out-of-time design, with predictor variables measured before the corresponding outcomes. The entry module compares L2-regularized logistic regression, Firth-penalized logistic regression, and gradient-boosted trees within a discrete-time hazard specification. The trajectory module estimates next-year rating upgrade scores for rated firms using accounting information and rating history. Both linear entry models achieve areas under the receiver operating characteristic (ROC) curve above 0.50, while the regularized logistic model shows a small positive difference in average precision relative to the training-prevalence benchmark. A single deflated firm-size measure provides discriminatory performance comparable to that of the multivariate entry models. The combined upgrade models achieve ROC areas close to 0.60, while models incorporating rating history achieve higher ROC areas than accounting-only models. Feature attributions also differ between the two tasks, with firm size playing a prominent role in entry predictions and rating history contributing more strongly to upgrade predictions. These findings identify the information sets associated with observed rating adoption and progression. The proposed framework provides rankings of candidate firms and feature-level explanations to support governance monitoring and outreach planning.

* Corresponding author.

E-mail address: author.mail@gmail.com

<https://doi.org/10.31181/jidmgc31202754>

© The Author(s) 2026 | [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Institutions that make corporate practices observable are central to corporate accountability and sustainable development. Third-party certification of corporate governance compliance provides a dated, public assessment of organizational practices [1, 2]. Signaling theory relates the informational value of certification to its observability, verifiability, and the costs of obtaining it [2, 3]. These characteristics are relevant in emerging markets, where disclosure, enforcement, and access to external assessment differ across firms [4]. Corporate governance certification consequently links firm-level disclosure practices with wider questions of transparency and accountable economic development. Recent machine learning research examines governance, policy evaluation, and financial decisions through prediction and decision-support applications [5, 6]. This literature provides a methodological basis for studying the timing and subsequent development of externally assessed governance outcomes.

Research on corporate governance has examined its relationship with firm value, financing conditions, and creditor assessments. Agency theory identifies conflicts among shareholders, managers, and creditors as relevant to corporate contracting [7]. Empirical studies associate disclosure quality and governance attributes with debt costs, credit ratings, and contract terms [8–13]. Board structure and internal control quality are also related to creditor assessments [14–17], while corporate political transparency is associated with debt pricing [18]. International evidence documents differences in these relationships across ownership structures, creditor-rights regimes, and financing arrangements [19–22]. Studies of governance reforms report heterogeneous borrowing-cost responses across institutional settings and changes in shareholder and creditor protection [23, 24]. Index membership and certification have also been examined in relation to information availability, market visibility, and financing conditions [25–30]. Evidence from Türkiye links governance quality with firm value, compliance ratings, and the cost of debt [31–34]. A related study of the Borsa Istanbul Corporate Governance Index reports lower post-inclusion debt costs relative to synthetic counterfactuals [35]. These contributions establish the relevance of external governance assessment to financial relationships and provide institutional context for examining which firms subsequently obtain ratings.

The timing of rating adoption presents a different empirical question from the financial outcomes associated with certification. Research on voluntary governance adoption examines ownership concentration, listing-segment requirements, and financing needs [36, 37]. Machine learning studies predict governance and ESG scores from accounting and disclosure information [38–42]. The present design focuses on first observed appearance in a governance rating list under annual out-of-time evaluation and compares this outcome with subsequent rating upgrades. These targets differ from score-level prediction and forecasts of future ESG performance based on contemporaneous or historical scores [41, 42]. The comparison is relevant to exchanges and regulators considering outreach to potential rating candidates, as well as investors monitoring changes in rated-firm coverage. It also distinguishes the accounting information associated with obtaining an external assessment from the information associated with later changes in that assessment. Balance-sheet scale and financing structure provide observable indicators of organizational capacity before rating adoption. Subsequent grade changes may be associated with information in rating histories and with assessments collected during agency review. Delegated information production offers a theoretical account of the role of external assessors in these information relationships [43, 44].

Regularized linear models and tree ensembles are widely used in financial prediction. Credit-scoring benchmarks document their performance across sample sizes and validation settings [45–47]. Related applications estimate firm performance and firm size from accounting variables [48, 49], while research on artificial intelligence examines its use in corporate governance decisions [50].

Explainable methods provide feature-level decompositions of model scores and are increasingly used in finance and sustainability applications [42, 51–53]. For governance-rating prediction, the evaluation design combines annual ordering of predictors and outcomes with probability calibration under class imbalance. Temporal ordering and the distinction between class-weighted scores and calibrated probabilities are incorporated in the estimation design. These elements inform the estimators and evaluation measures used in the present analysis.

The Borsa Istanbul Corporate Governance Index provides an institutional setting with published rating requirements and annual assessments. The index has operated since 2007 and currently requires an overall grade of at least 8 on a ten-point scale, with section minima of 7 [54–56]. Licensed agencies renew ratings annually. The sample period includes substantial inflation and exchange-rate volatility, accompanied by pronounced changes in nominal financial statements [57, 58]. Research on policy uncertainty and corporate relationships associates these conditions with variation in financing costs and the value of observable corporate characteristics [59, 60]. The main specification adjusts assets and sales using the annual price-level index. A within-year percentile-rank specification provides a separate comparison based on firms' relative positions in the annual cross-section.

The study addresses three research questions. The first concerns the out-of-time predictability of first observed rating adoption from statutory accounts. H1 specifies that the two linear entry estimators achieve ROC areas above chance. Average-precision differences relative to the training-prevalence benchmark are evaluated separately for each estimator, with uncertainty summarized by firm-clustered intervals. The second question concerns the distribution of predictive information across accounting features. H2 specifies that scale has the largest feature attribution and that a small set of capacity-related features retains substantial discrimination. The third question concerns differences between adoption and subsequent rating progression. H3 specifies that the predictive information associated with statutory accounts differs across the two tasks, as reflected in their feature-attribution rankings. The analysis evaluates common-calendar performance and feature importance separately to distinguish model discrimination from the composition of the information used by each model.

Both linear entry models achieve pooled ROC areas above chance. The regularized logistic model has a positive average-precision difference relative to the training-prevalence benchmark, with its lower interval endpoint close to zero. Its annual ROC area is at least 0.70 in four of the five test years. A single deflated firm size measure has discrimination comparable to the multivariate models. Feature-level decompositions associate entry scores with scale, profitability, financial leverage, and other accounting characteristics. For subsequent upgrades, rating-history specifications have higher ROC areas than accounting-only and combined specifications. The feature-attribution rankings also differ across the two tasks. Two departures from the rating list are examined through the firms' recorded grade histories.

The study contributes an annual prediction design for first observed governance-rating adoption, with explicit comparisons among regularized, bias-reduced, and nonlinear estimators. The firm-size-only benchmark identifies the discrimination associated with a single observable firm characteristic, while reduced specifications examine the concentration of predictive information. Comparing entry with subsequent upgrades provides evidence on differences between the accounting and rating-history information sets. The feature decompositions relate these differences to specific financial characteristics and prior assessment records. The framework also supplies operating characteristics for candidate prioritization, including event capture at a fixed screening capacity and probability-based net benefit. These outputs extend decision-support applications of machine learning to governance monitoring.

2. Materials and Methods

2.1. Institutional Setting and Data

The Borsa Istanbul Corporate Governance Index has operated since 2007 and tracks firms assessed by rating agencies licensed by the Capital Markets Board of Türkiye. The rating methodology covers shareholders, public disclosure and transparency, stakeholders, and the board of directors. The four section scores enter the overall grade with weights of 25, 25, 15, and 35 percent, respectively. Current index requirements specify an overall grade of at least 8 and a minimum grade of 7 in each section. Earlier eligibility regimes used different thresholds [54, 55]. Agencies assess public filings and information obtained through firm contact and document review [61]. Ratings are disclosed through the Public Disclosure Platform and renewed annually. These records provide consecutive observations of overall and section-level governance assessments. The same institutional setting is examined in related research on governance certification and financing [35].

The study combines annual governance-rating records with financial statement data. The rating list contains overall grades and four section scores for 49 non-financial firms between 2013 and 2024, comprising 367 firm-year observations. The financial source contains annual statements from 2002 to 2025 for 588 firms, compiled from exchange and Public Disclosure Platform releases. Ticker codes provide the matching key, with 48 rated firms represented in the statement source. Firms classified under the BIST financials index are excluded because their accounts and regulatory arrangements differ from those of industrial, services, and technology firms. The working non-financial panel covers 2011 to 2024 and contains 438 firms with 4,399 firm-year observations. The entry module uses accounting features and sector indicators. The trajectory module combines these features with lagged overall ratings, lagged rating changes, and time in the rating list.

2.2. Sample Construction

The entry sample is constructed as a discrete-time hazard panel for first observed governance-rating adoption. A firm contributes an observation for each feature year in which it has not yet appeared in the rating list and leaves the risk set after its first appearance. Feature years cover 2013 to 2023, with outcomes recorded in the following calendar year. Of the 438 firms in the non-financial panel, 68 contribute no eligible observations. These comprise 23 firms already rated in 2013, 16 firms first observed in the financial panel in 2024, and 29 firms without usable financial statements throughout their observed histories. Observations with positive total assets are retained. Before applying listing eligibility, the candidate sample contains 2,829 firm-year observations from 370 firms.

Financial records are identified by firm and year, with the first record in source order retained when multiple records share the same key. Listing eligibility is determined from annual listing archives and issuer announcements, including admissions to the exchange's pre-market trading platform. Firms listed before 2013 are eligible throughout the observation window. For subsequent listings, a financial observation is retained when its outcome year is equal to or later than the firm's first trading year. This rule excludes 147 observations from 66 firms. Two of these firms contribute no eligible observations because their first trading years fall after the outcome window. The final entry sample contains 2,682 firm-year observations from 368 firms. All retained rating-adoption events occur after listing eligibility. The replication package provides the firm-level listing records and duplicate-record audit.

The rating list contains 26 first observed appearances between 2014 and 2024. Two lack usable pre-entry financial information. One firm has no matching financial statements, while another has an empty statement for the preceding feature year. The remaining 24 appearances form the analytic event sample, corresponding to a prevalence of 0.89 percent. Across the complete rating list, 48 firms match the financial panel, comprising 30 industrial, 15 services, and 3 technology firms. Missing

feature values are concentrated in real sales growth, which is unavailable for 4.7 percent of observations, and operating margin, for which missingness is below one percent. Missing model inputs are replaced by medians estimated within the corresponding training window. Table 1 summarizes the sample construction.

Table 1

Construction of the entry and rating-transition samples

Sample stage	Firms	Observations
Panel A. Entry sample		
Non-financial listed-firm panel	438	4,399
Firms rated at the start of the observation window	23	
Firms first observed in the financial panel in 2024	16	
Firms without usable financial statements	29	
Candidate sample before listing eligibility	370	2,829
Observations excluded by listing eligibility	66	147
Final entry sample	368	2,682
First observed rating appearances	24	24
Panel B. Rating-transition sample		
Firms with consecutive annual ratings	47	318
Upgrades		230
Downgrades		23
Unchanged ratings		65

Note. Entry observations are firm-years; rating-transition observations are consecutive annual rating pairs. Blank cells indicate counts that are not reported separately.

2.3. Outcome Definitions

For the entry module, the outcome equals one when a firm appears in the rating list for the first observed time in the year following the feature year. Other eligible observations receive a zero label. The first appearance is established from the annual rating records for 2013 to 2024. Firms rated in 2013 constitute the initial cohort and contribute no entry-risk observations. Subsequent appearances identify the adoption events used by the hazard specification. Entry-risk observations end after the firm's first appearance. The outcome therefore describes the observed timing of obtaining a rating within the study period. Annual rating-list records provide the event definition, while the institutional index requirements describe the setting in which these assessments are issued. With T denoting a firm's first observed rating year, the entry indicator is defined in Eq. (1) for eligible firm-years. T is assigned infinity when the firm has no observed rating appearance. Subscripts i and t identify firms and feature years, respectively, and the indicator function takes the value one when its condition holds.

$$y_{i,t+1}^{\text{entry}} = \mathbb{1}\{T_i = t + 1\}, \quad T_i > t \quad (1)$$

Accounting features are calculated from statements for the feature year, and rating outcomes are taken from the following calendar year. This ordering provides a common annual prediction horizon across firms. The design follows the typical disclosure calendar, under which annual statements are released during the early months of the following year. In the entry module, financial records for year t are associated with the first-appearance indicator for year $t + 1$. In the trajectory module, financial records and rating information for year t are associated with the upgrade indicator ending in year $t + 1$.

The trajectory outcome equals one when the overall grade increases by more than 0.005 points from the previous year. Observations with smaller changes or declines receive a zero label. This specification distinguishes upgrades from the remainder of the annual transitions. The 0.005-point

band separates small numerical changes from increases in the recorded overall grade. The same consecutive-year pairing structure is used for all information-set comparisons. A separate threshold specification uses an increase exceeding 0.05 points. Section-level models apply annual upgrade labels to each of the four governance scores. Eq. (2) expresses the overall upgrade label, where G denotes the recorded grade and δ denotes the increase threshold. The main specification uses δ equal to 0.005, and the alternative specification uses 0.05.

$$y_{i,t+1}^{\text{upgrade}} = \mathbb{1}\{G_{i,t+1} - G_{i,t} > \delta\} \quad (2)$$

2.4. Feature Set

The accounting feature set contains eleven continuous measures and two sector indicators, defined in Table 2. The measures comprise firm size, financial leverage, return on assets, operating margin, real sales growth, the current ratio, asset tangibility, the financial expense ratio, the equity ratio, net foreign currency position, and financial-panel tenure. Ratios are calculated from the reported financial items. Continuous inputs are winsorized at the 1st and 99th percentiles using bounds estimated within each training window. Missing values are replaced by training-window medians. The same bounds and medians are applied to the corresponding test observations. The two sector indicators represent services and technology firms, with industrial firms as the reference category. The maximum variance inflation factor in the estimation sample is 2.74.

The monetary features use a compound GDP deflator index for Türkiye constructed from the World Development Indicators annual growth series [62]. The index for each year equals the preceding level multiplied by one plus the annual growth rate, with the 2010 level normalized to one. Firm size is the natural logarithm of total assets divided by this index. Real sales growth is the annual proportional change in deflated sales. Eq. (3) defines firm size and real sales growth, where d is the annual deflator growth rate in percent, A is total assets, and S is reported sales. Ratios of items reported for the same year retain their original values. A percentile-rank specification replaces model inputs with their within-year cross-sectional ranks. This specification evaluates predictions based on relative financial positions within each annual sample. The nominal specification retains unadjusted assets and nominal sales growth for comparison with the main deflated measures.

$$D_{2010} = 1, \quad D_t = D_{t-1} \left(1 + \frac{d_t}{100}\right) \quad (3)$$

$$\text{Firm size}_{it} = \ln\left(\frac{A_{it}}{D_t}\right), \quad \text{Real sales growth}_{it} = \frac{S_{it}/D_t}{S_{i,t-1}/D_{t-1}} - 1$$

Table 2

Definitions of accounting features and sector classification

Feature	Definition.
Firm size	Natural logarithm of the ratio of total assets to the deflator level index
Financial leverage	Short-term and long-term financial debt divided by total assets
Return on assets	Net income over total assets
Operating margin	Operating profit over sales
Real sales growth	Annual growth rate of deflated sales
Current ratio	Current assets over current liabilities
Asset tangibility	Property, , plant and equipment divided by total assets
Financial expense ratio	Financial expenses over total assets
Equity ratio	Equity attributable to the parent over total assets
Net foreign currency position	Foreign currency assets minus foreign currency liabilities divided by total assets
Financial-panel tenure	Years since the firm's first observed financial-panel record
Sector classification	Classification as an industrial, services or technology firm

Note. Ratio definitions use annual consolidated financial statement items.

2.5. Models and Estimation

Three estimators are compared. The first is L2-regularized logistic regression with balanced class weights and inverse regularization parameter C set to 0.5. Standardized features enter an additive linear index. The second is a gradient-boosted tree ensemble implemented in LightGBM 4.6.0 [63]. Its split structure accommodates nonlinear relationships and interactions among the accounting features. Financial prediction benchmarks examine these models across credit-scoring and default settings [45, 46]. The third estimator is Firth penalized logistic regression, which applies the Jeffreys-prior adjustment to the likelihood score equations [64, 65]. The implementation uses iterative coefficient updates, a maximum of 300 iterations, and a coefficient-step tolerance of 1e-9. Firth estimation uses standardized inputs and no class weights. Comparison with the weighted logistic model examines the predictive results under two different approaches to estimating a linear index. Eq. (4) defines the logistic event-class score q for the standardized feature vector. Eq. (5) gives the weighted penalized criterion over a training window W . The observation weight w is n divided by twice the number of training observations in its outcome class. Here n is the number of observations in W . The intercept is excluded from the L2 penalty. Eq. (6) gives the Firth criterion, where the unweighted Bernoulli log likelihood is augmented by one half of the log determinant of its Fisher information matrix. The coefficient vector θ includes the intercept.

$$q_{it} = \sigma(\beta_0 + \tilde{x}_{it}^T \beta), \quad \sigma(v) = \frac{1}{1 + \exp(-v)} \quad (4)$$

$$\ell_W^{L2}(\beta_0, \beta) = \sum_{(i,t) \in W} w_{it} [y_{it} \ln q_{it} + (1 - y_{it}) \ln(1 - q_{it})] - \frac{\|\beta\|_2^2}{2C} \quad (5)$$

$$\ell_W^F(\theta) = \ell_W(\theta) + \frac{1}{2} \ln \det I_W(\theta) \quad (6)$$

The entry boosting model uses 400 trees, 15 leaves per tree, and at least 40 observations per leaf. The trajectory model uses 300 trees, 7 leaves per tree, and at least 15 observations per leaf. Both models use a learning rate of 0.03, a column-sampling fraction of 0.8, and an L2 penalty of 5. Row bagging is disabled through a subsampling frequency of zero. Positive-class weights equal the ratio of non-events to events within the training window. The estimator settings remain constant across test years. Eq. (7) represents the boosted score as the logistic transformation of an initial score and the learning-rate-scaled sum of M fitted trees. The tree functions operate on the preprocessed feature vector. A sensitivity specification replaces class weighting with SMOTE-based synthetic minority oversampling [66]. The reduced entry specification contains firm size, financial leverage, return on assets, the current ratio, and financial-panel tenure. Its discrimination and average precision are compared with those of the full specification on the same test observations.

$$q_{it} = \sigma(F_0 + \eta \sum_{m=1}^M f_m(x_{it})), \quad \eta = 0.03 \quad (7)$$

Estimation uses Python 3.12, scikit-learn 1.7.2, and LightGBM 4.6.0. The supplementary materials contain the core analysis code, event list, feature matrices, calibration diagnostics, and out-of-sample predictions.

2.6. Temporal Validation and Calibration

Out-of-time evaluation uses expanding training windows. For each test year, the models are fitted to observations from all earlier feature years and applied to that year's held-out observations. The entry module evaluates feature years 2019 to 2023, with 1,404 test observations and 17 first appearances in the following calendar years. The trajectory module evaluates transitions ending in 2021 to 2024, with 140 test observations and 98 upgrades. Table 3 gives the training and test counts for each entry window. Training-window winsorizing bounds, medians, and standardization

parameters are applied to the corresponding held-out year. The first entry training window contains seven events. Model estimation and probability calibration both use the observations available within the outer training window. The ten test events in feature year 2022 correspond to rating-list appearances in 2023. In the first calibration window, the contiguous blocks cover 2013 to 2014, 2015 to 2016, and 2017 to 2018. The first block supplies training history with four events and produces no predictions. The second block has one event and is predicted from the first; the third has two events and is predicted from the preceding blocks.

Table 3

Training and test samples for the entry models

Test feature year	Training observations	Training firms	Training events	Test observations	Test events
2019	1,278	228	7	228	1
2020	1,506	235	8	239	1
2021	1,745	247	9	285	4
2022	2,030	294	13	308	10
2023	2,338	321	23	344	1

Note. Event counts refer to rating-list appearances in the calendar year following the feature year.

Logistic, boosting, and Firth model outputs are mapped to probabilities through sigmoid calibration [67]. Event-class scores are clipped to the interval from $1e-6$ to $1 - 1e-6$ and transformed to log odds. A logistic regression is fitted to these transformed scores with C set to $1e10$, the lbfgs solver, a tolerance of $1e-4$, and a maximum of 1,000 iterations. The very large C approximates an unpenalized mapping, and coefficients are stored as double-precision values. Calibration inputs are generated within each outer training window using three contiguous year blocks. The first block supplies training observations, the second is predicted from the first, and the third is predicted from the first two. These internal predictions are generated from the matrix already winsorized and imputed within the outer training window. Each internal fit then estimates an additional set of clipping bounds and scaling parameters from its preceding blocks. For Firth regression, internal predictions are generated by fitting the unweighted Firth estimator within the same forward blocks. Each model-specific calibrator pools its available internal score-label pairs. If there are ten or fewer usable pairs or they contain only one outcome class, a constant-score fit uses the outer-window event prevalence. A non-positive fitted calibration slope assigns the training prevalence to that window's test observations. All entry-window slopes are positive. The trajectory boosting model uses the prevalence assignment in the 2022 and 2023 windows. Equation (8) gives the score transformation and calibration rule. Parameters a and b are fitted for each test window τ , ϵ equals $1e-6$, and π denotes the event prevalence in the corresponding outer training window. The firm-size-only benchmark uses training-window winsorized log assets as its direct calibration input, with no upstream classifier and no log-odds transformation. A separate firm-size-only logistic model with C equal to 0.5 and balanced class weights provides raw probabilities for the Brier comparison. The direct firm-size mapping is fitted on the complete outer training window. For this benchmark, z in Eq. (8) is replaced by the winsorized log-asset score. Calibration accuracy is evaluated with the Brier score and a window-specific training-prevalence predictor [68]. The supplementary code exports fitted calibration parameters and internal score-label pairs.

$$\tilde{q}_{it} = \min(1 - \epsilon, \max(\epsilon, q_{it})), \quad z_{it} = \ln\left(\frac{\tilde{q}_{it}}{1 - \tilde{q}_{it}}\right)$$

$$p_{it} = \begin{cases} \sigma(a_\tau + b_\tau z_{it}), & b_\tau > 0 \\ \pi_\tau, & b_\tau \leq 0 \end{cases} \quad (8)$$

2.7. Evaluation Protocol and Explainability

Average precision is the primary performance measure and is calculated with scikit-learn's stepwise precision-recall summary. Precision-recall evaluation is suited to outcomes with low prevalence [69]. ROC area and Brier score provide secondary summaries of discrimination and probability accuracy. Average precision and ROC area are calculated from raw model scores pooled across test years, with annual estimates also reported. Calibrated probabilities are used for Brier scores and decision curves. Within-year top-decile rankings use the raw scores. Eq. (9) defines average precision as the sum of precision values weighted by successive recall increments. Index k runs over distinct score thresholds in descending order, with initial recall equal to zero. Equal scores enter the same threshold group. The Brier score is the mean squared difference between the binary outcome and its calibrated probability across N test observations. Confidence intervals are obtained from 1,000 block-bootstrap samples of firms. Each sample retains the selected firm's complete set of test-year observations. Model differences are calculated on identical bootstrap samples to preserve the paired comparison. This procedure is applied to the firm-size-only and training-prevalence benchmarks and to the trajectory information sets. The reported intervals are percentiles of the resampled fixed-prediction statistics. Class imbalance is addressed through estimator-level weighting in the main specifications.

$$\begin{aligned} AP &= \sum_{k=1}^K (R_k - R_{k-1}) P_k, \quad R_0 = 0 \\ \text{Brier} &= \frac{1}{N} \sum_{j=1}^N (y_j - p_j)^2 \end{aligned} \quad (9)$$

Two screening summaries accompany the statistical metrics. Top-decile capture is the proportion of observed events among the highest-scoring ten percent of eligible firm-years within each test year. The number selected in each year is the integer part of one tenth of its candidate pool. The total number flagged is reported alongside the events captured. Decision curve analysis evaluates net benefit from calibrated probabilities at specified thresholds, using the threshold odds to weight false positives relative to true positives [70]. Eq. (10) expresses net benefit for threshold c , where TP and FP count true-positive and false-positive selections among N test observations. Flag-all and flag-none rules provide the decision-curve benchmarks. The thresholds specify statistical trade-offs within the observed candidate pools.

$$NB(c) = \frac{TP(c)}{N} - \frac{FP(c)}{N} \frac{c}{1-c}, \quad 0 < c < 1 \quad (10)$$

Feature attributions are calculated for the final fitted models. The logistic decomposition multiplies each standardized feature by its estimated coefficient. TreeSHAP provides the decomposition for the uncalibrated boosting model fitted to the full analytic sample [71]. Mean absolute attributions summarize feature importance. Spearman correlations between feature values and their signed attributions summarize direction, and quartile-level mean attributions describe non-monotone relationships. The two model classes are evaluated separately because their functional forms allocate contributions differently across correlated features. Fold-wise importance rankings are also compared across expanding windows.

3. Results

3.1. Rated Firms, Entrants, and Transitions

The number of rated firms increased from 23 in 2013 to 47 in 2024. The largest additions occurred in 2022 and 2023. Table 4 gives the annual grade distribution. The mean overall grade increased from 9.02 to 9.41, while annual standard deviations ranged from 0.26 to 0.43. Across the complete rating panel, stakeholders and public disclosure had the highest section means, at 9.56 and 9.51. Shareholders and the board of directors both averaged 9.15. The mean overall grade was 9.32 for

subsequent entrants and 9.29 for the 2013 cohort. The annual series therefore combines an expansion in coverage with relatively concentrated grade distributions. New rating-list appearances contributed to this expansion alongside the continuing assessments of the initial cohort. The two lowest overall ratings recorded in 2014 are 7.62 and 7.79. Annual rating-list coverage and index eligibility follow separate institutional criteria.

Table 4

Annual coverage and overall governance ratings

Year	Rated firms	Mean overall rating	Standard deviation	Minimum	Maximum
2013	23	9.02	0.34	8.05	9.34
2014	25	9.01	0.43	7.62	9.42
2015	27	9.13	0.37	8.00	9.55
2016	25	9.24	0.30	8.08	9.60
2017	26	9.27	0.28	8.20	9.63
2018	26	9.31	0.26	8.38	9.70
2019	28	9.36	0.27	8.51	9.72
2020	29	9.40	0.26	8.54	9.72
2021	31	9.39	0.35	8.01	9.75
2022	35	9.42	0.28	8.52	9.75
2023	45	9.37	0.36	8.15	9.76
2024	47	9.41	0.34	8.16	9.79

Note. Overall ratings are measured on a ten-point scale.

Future entrants have larger balance sheets than never-rated firms in the feature-year comparison in Table 5. Mean log deflated assets are 20.78 and 18.99, respectively, with corresponding medians of 20.50 and 18.99. The firm size difference is significant at the one percent level in both the pooled and firm-level comparisons. Entrants also have greater financial leverage, with means of 0.30 and 0.23 and medians of 0.34 and 0.19. The financial leverage difference has a pooled p-value of 0.014 and a firm-level p-value of 0.005. Profitability, real sales growth, and financial-panel tenure have p-values below 0.05 in the pooled comparison and above 0.05 in the firm-level comparison. Of the 24 modeled appearances, 13 are industrial firms and 10 are services firms. These financial and sector characteristics describe the composition of the observed entrant group.

Table 5

Financial characteristics of future entrants and never-rated firms

Feature	Entrant mean	Never-rated mean	Entrant median	Never-rated median	Pooled p-value	Firm-level p-value
Firm size	20.78	18.99	20.50	18.99	<0.001	<0.001
Financial leverage	0.30	0.23	0.34	0.19	0.014	0.005
Return on assets	0.11	0.06	0.10	0.05	0.029	0.153
Operating margin	0.18	0.39	0.15	0.10	0.020	0.091
Real sales growth	0.31	0.87	0.22	0.07	0.034	0.326
Current ratio	1.61	2.90	1.35	1.42	0.756	0.161
Asset tangibility	0.34	0.30	0.23	0.28	0.654	0.857
Financial expense ratio	0.07	0.06	0.06	0.04	0.130	0.357
Equity ratio	0.44	0.46	0.42	0.47	0.365	0.091
Net foreign currency position	-0.07	-0.05	-0.04	0.00	0.738	0.442
Financial-panel tenure	6.29	7.54	5.00	7.00	0.043	0.092

Note: The comparison includes 24 pre-entry observations and 2,584 never-rated firm-years. Firm-level p-values use one financial summary per never-rated firm across 344 firms. Both comparisons use the same rank-sum test.

Rating histories are persistent once a firm is rated. No rated firm leaves the sample and returns within the window. Upgrades outnumber downgrades by a wide margin, at 230 against 23 of 318 transitions, and 65 transitions stay flat. The board section is the most dynamic, with 204 upgrades and 36 downgrades. The stakeholders section moves least, with 111 upgrades and 22 downgrades.

3.2. Predicting First-Time Entry

Table 6 reports pooled out-of-time performance for the entry models. Regularized logistic regression has average precision of 0.089, with a firm-clustered 95 percent interval of 0.014 to 0.259, compared with test-event prevalence of 0.012. Its ROC area is 0.713, with an interval of 0.583 to 0.814. Firth logistic regression has average precision of 0.030 and a ROC area of 0.691, with respective intervals of 0.014 to 0.104 and 0.582 to 0.798. The ROC intervals for both linear estimators exceed 0.500, consistent with H1. Their average-precision comparisons with the training-prevalence predictor differ. The mean paired bootstrap difference is 0.077 for regularized logistic regression, with an interval of 0.001 to 0.228, and 0.024 for Firth regression, with an interval of -0.001 to 0.088. The regularized model's lower endpoint is positive and close to zero, while the Firth interval spans zero. Boosted trees have average precision of 0.022 and a ROC area of 0.598, with intervals of 0.011 to 0.049 and 0.432 to 0.739. The linear estimators thus have higher pooled discrimination point estimates than the tree ensemble. Their relative performance is consistent with financial prediction benchmarks in which regularized linear models remain competitive with more flexible estimators [45, 46]. Figure 1 summarizes the placement of entry events across the pooled model rankings.

The firm-size-only benchmark has average precision of 0.046, with a 95 percent interval of 0.017 to 0.197, and a ROC area of 0.719, with an interval of 0.600 to 0.829. Its discrimination is close to that of regularized logistic regression. Mean paired bootstrap ROC differences relative to firm size are -0.006 for regularized logistic regression, -0.030 for Firth regression, and -0.122 for boosting, with all three intervals spanning zero. The observed average-precision difference between regularized logistic regression and firm size is 0.043. The mean difference across paired bootstrap samples is 0.028, with an interval of -0.096 to 0.147. The observed difference is calculated from the original pooled test sample, whereas the bootstrap mean averages differences across resampled firm histories. Firm size therefore accounts for substantial pooled entry discrimination within the selected financial information set. The multivariate models provide additional descriptions of how profitability, financial leverage, liquidity, and other accounting characteristics contribute to individual scores. Their higher logistic average-precision point estimate reflects different event positions near the top of the pooled ranking. The pooled comparison and the year-specific screening results capture different aspects of candidate prioritization because each annual screen selects firms within that year's eligible pool.

Deleting one entry event at a time yields logistic average precision between 0.025 and 0.094 and ROC areas between 0.695 and 0.752. The event with the largest influence on average precision gives deletion estimates of 0.025 and 0.695, respectively. This contrast shows that average precision responds more strongly than ROC area to the position of an individual event near the top of the ranking. Table 6 also identifies differences between discrimination and probability accuracy. Regularized logistic regression has a raw Brier score of 0.1286 and a calibrated score of 0.0119. The corresponding boosting scores are 0.0141 and 0.0120. Firth regression has a raw score of 0.0121 and a calibrated score of 0.0120. Its smaller change is consistent with the unweighted estimator producing probabilities closer to the observed event rate before recalibration. In contrast, the balanced logistic models assign greater weight to the rare event class during fitting, so their uncalibrated outputs are not directly aligned with its frequency in the test population. The firm-size-only weighted logistic specification has a raw Brier score of 0.1825, whereas the direct firm-size-based probability mapping has a score of 0.0119. This comparison concerns 2 probability specifications based on the same financial characteristic, rather than the addition of predictors. The

firm size comparison reflects differences in probability estimation and weighting, while its discrimination metrics continue to use the deflated log-asset ranking. The calibrated scores are close to the prevalence benchmarks at the reported precision. Relative to the training-prevalence predictor's Brier score of 0.0120, calibrated logistic regression has a paired difference of -0.00009 , with an interval of -0.00040 to 0.00011 . The calibrated firm size benchmark has a paired difference of -0.00013 , with an interval of -0.00052 to 0.00019 . Both intervals span zero despite the lower point estimates. Firth regression also has calibrated squared-error performance close to the prevalence benchmark at the reported precision. Figure 2 compares mean predicted probabilities with observed entry frequencies across quintile bins. The probability comparisons describe the accuracy of estimated event frequencies, while average precision and ROC area summarize the ordering of firms within the candidate pool.

Table 6
Out-of-time performance of the entry models

Model	Average precision	95% CI	ROC area	95% CI	Raw Brier	Calibrated Brier
Logistic	0.089	[0.014, 0.259]	0.713	[0.583, 0.814]	0.1286	0.0119
Firth logistic	0.030	[0.014, 0.104]	0.691	[0.582, 0.798]	0.0121	0.0120
Boosted trees	0.022	[0.011, 0.049]	0.598	[0.432, 0.739]	0.0141	0.0120
Firm-size benchmark	0.046	[0.017, 0.197]	0.719	[0.600, 0.829]	0.1825	0.0119
Constant-prevalence benchmark	0.012		0.500		0.0120	0.0120
Training-prevalence predictor	0.014		0.499		0.0120	0.0120

Note: The test pool contains 1,404 observations and 17 entry events. Confidence intervals use 1,000 firm-clustered bootstrap samples. The firm-size row uses different probability specifications for raw and calibrated Brier scores. Prevalence benchmarks receive no additional calibration.

Annual ROC areas for regularized logistic regression are 1.000 in 2019, 0.899 in 2020, 0.747 in 2021, 0.730 in 2022, and 0.455 in 2023. Four of the five estimates exceed 0.70. The 2019 and 2020 test years each contain one entry event. The unweighted mean of annual ROC areas is 0.766 for regularized logistic regression, 0.777 for Firth regression, 0.704 for boosting, and 0.763 for firm size. These annual averages give each test year equal weight, whereas the pooled estimates compare scores across the complete test sample. At a fixed annual top-decile capacity, regularized logistic regression captures five of the 17 events, reduced logistic regression captures seven, boosting captures six, and firm size captures five. Each rule selects a total of 137 firm-years across the five annual pools, with approximately 1.7 events expected under random selection at the same capacity. The reduced logistic screen captures two more events than the firm size rule. Its pooled average precision is 0.082 and its ROC area is 0.783, compared with 0.089 and 0.713 for the full logistic model. Reduced Firth regression has a pooled ROC area of 0.784. The reduced specifications retain substantial discrimination, while the full logistic model has the higher pooled average-precision estimate. The operating comparisons therefore depend on both model specification and the part of the ranking selected for outreach. Decision-curve analysis based on calibrated full-model logistic probabilities gives positive net benefit at the 1, 2, 6, and 7 percent thresholds and negative net benefit at the 3, 4, and 5 percent thresholds. Positive values exceed the flag-none benchmark of zero. The flag-all benchmark is positive at 1 percent and negative above the test-event prevalence of 1.21

percent. These threshold-specific results complement the fixed-capacity screens by incorporating the relative weight assigned to false-positive selections.

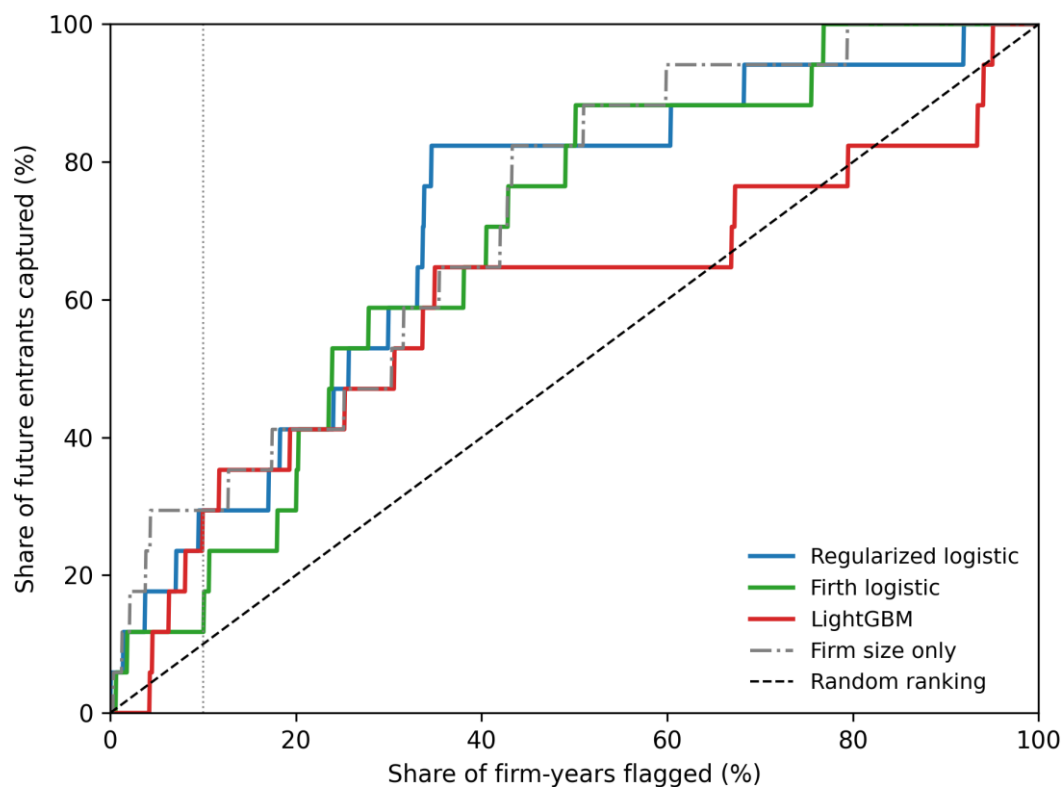


Fig. 1. Cumulative gains for the entry models

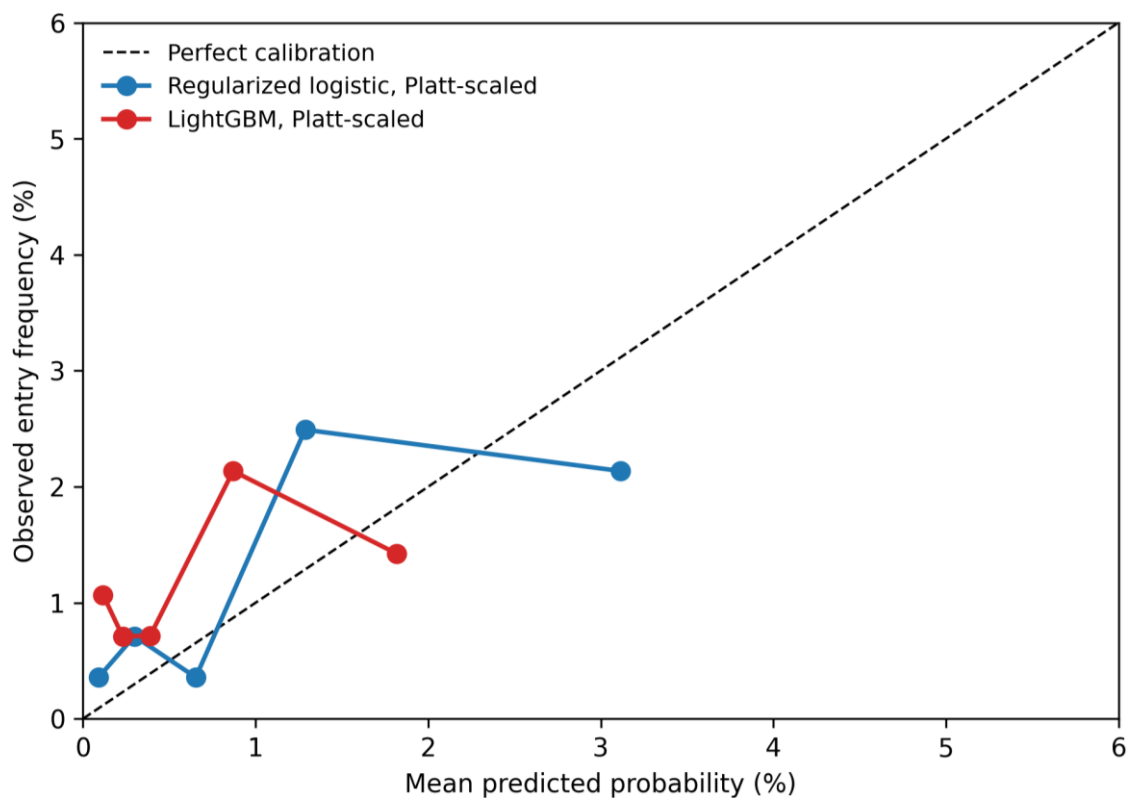


Fig. 2. Calibration of entry probabilities across quintile bins

3.3. Accounting Features Associated with Entry Scores

Figure 3 and Table 7 summarize feature attributions from the regularized logistic and boosting models. Table 7 reports mean absolute magnitudes for the eleven continuous accounting features. Logistic attributions are products of standardized feature values and their estimated coefficients. TreeSHAP values decompose the uncalibrated boosting score. Both summaries use the final models fitted to the complete analytic sample. Figure 3 also includes the sector indicators and normalizes each model's attributions by its largest value. Firm size has the largest attribution in both models, consistent with the feature-concentration component of H2. The remaining importance rankings differ, with a Spearman correlation of -0.14 across the eleven accounting features. Return on assets and financial-panel tenure follow firm size in the logistic decomposition. Financial leverage, asset tangibility, operating margin, and net foreign currency position have relatively large attributions in the tree decomposition. Attribution magnitudes describe the allocation of importance within each fitted model. Attribution direction is summarized separately by the Spearman association between feature values and signed attributions. Eight features have the same directional association across models. Firm size, financial leverage, return on assets, real sales growth, the equity ratio, and net foreign currency position are positively associated with entry scores. The current ratio and financial-panel tenure have negative associations in both decompositions. Operating margin, asset tangibility, and the financial expense ratio have different directional associations across the two models. Mean fold-wise correlations of importance ranks are 0.65 for logistic regression and 0.71 for boosting. The joint comparison identifies a common concentration in firm size alongside model-specific allocations of the remaining financial information.

Table 7

Mean absolute feature attributions for the entry models

Feature	Logistic mean absolute attribution	Mean absolute TreeSHAP value
Firm size	1.225	0.958
Return on assets	0.887	0.444
Financial-panel tenure	0.621	0.363
Current ratio	0.493	0.213
Equity ratio	0.397	0.226
Financial expense ratio	0.319	0.272
Financial leverage	0.299	0.651
Net foreign currency position	0.235	0.490
Real sales growth	0.221	0.277
Operating margin	0.120	0.510
Asset tangibility	0.104	0.511

Note: Features are ordered by logistic attribution magnitude. Values summarize importance within each fitted model.

Real sales growth has a positive attribution direction in both models under the deflated specification. Under the nominal specification, its direction in the tree decomposition is negative. The specification comparison therefore associates the price-level adjustment with a change in the sales-growth attribution. Entrants have median real sales growth of 22 percent, compared with 7 percent among never-rated firms. This difference is consistent with the positive direction in the deflated entry models. The pooled and firm-level comparisons in Table 5 provide the corresponding distributional summaries, while the model decomposition describes real sales growth conditional on the other included features.

The entry models associate rating adoption with scale and a set of financing and operating characteristics. Firm size represents the largest contribution in both decompositions, while financial leverage and profitability provide additional financial information. Financial-panel tenure contributes negatively, associating higher scores with firms earlier in their observed financial-panel histories.

These features describe the resources and financial relationships of prospective rated firms. Related evidence from the same institutional setting documents financial differences between rated firms and the wider listed population before rating adoption [35]. The present feature profile adds an annual predictive comparison to those pre-adoption differences.

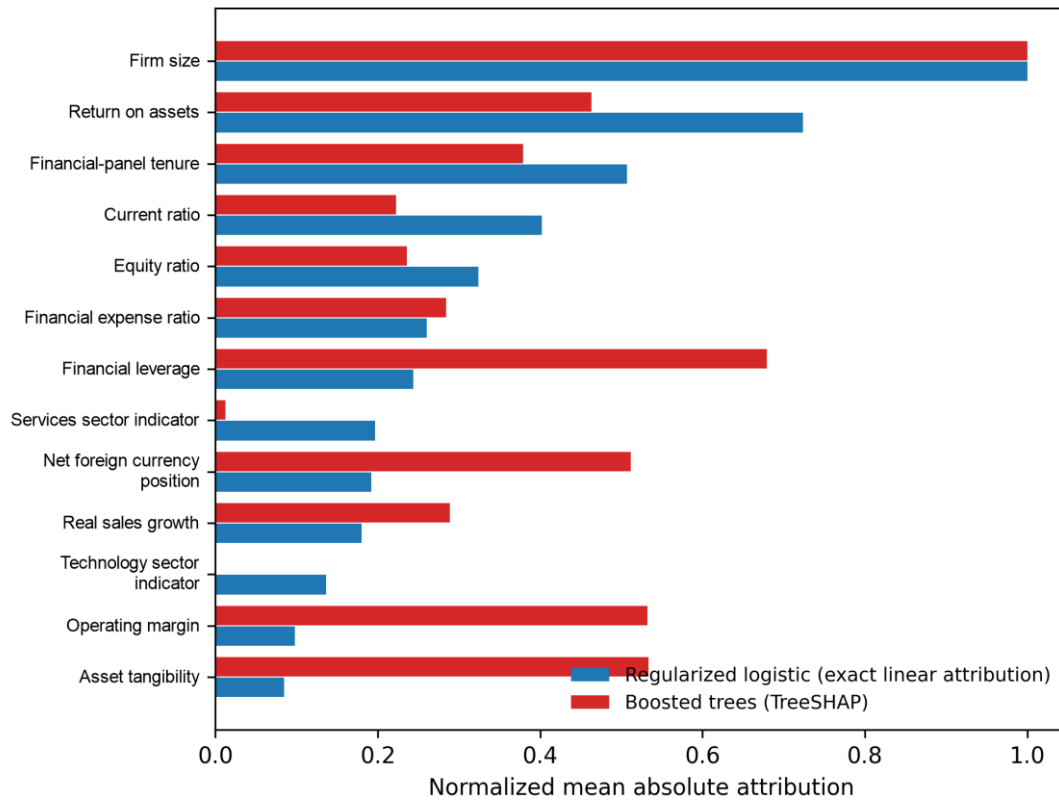


Fig. 3. Normalized feature attributions for the entry models

3.4. Predicting Post-Adoption Trajectories

The combined trajectory models have pooled ROC areas of 0.598 for logistic regression, 0.593 for boosting, and 0.592 for Firth regression. Their 95 percent intervals are 0.493 to 0.690, 0.491 to 0.686, and 0.500 to 0.685, respectively. Average precision ranges from 0.771 to 0.778, compared with upgrade prevalence of 0.700. Discrimination point estimates are similar across the 3 model families. Calibrated Brier scores are 0.211 for logistic regression, 0.213 for boosting, and 0.209 for Firth regression. The constant-prevalence benchmark has a Brier score of 0.210, while the training-prevalence predictor has a score of 0.213. Firth therefore has the lowest calibrated Brier point estimate at the reported precision, with a small difference relative to both prevalence benchmarks. These values place the estimated probabilities close to the squared-error performance of prevalence-based predictions. The small differences in calibrated squared error contrast with the similar ROC estimates and characterize the probability outputs at the reported precision. Table 8 summarizes discrimination and probability accuracy on the same test transitions. The high frequency of upgrades contributes to the level of average precision, while the Brier comparisons assess the distance between estimated probabilities and realized outcomes. The information-set analysis examines whether these predictions vary when accounting characteristics and prior ratings are used separately. All four Firth calibration slopes are positive, so its test probabilities use the fitted sigmoid mappings throughout. The constant benchmark assigns the pooled upgrade rate of 0.700, for which the Brier score equals 0.700 multiplied by 0.300. Annual boosting ROC areas are 0.643, 0.579, 0.600, and 0.741 on 29, 31, 35, and 45 transitions, respectively.

Table 8

Out-of-time performance of the trajectory models

Model	Average precision	95% CI	ROC area	95% CI	Calibrated Brier
Boosted trees	0.771	[0.666, 0.873]	0.593	[0.491, 0.686]	0.213
Logistic	0.778	[0.692, 0.872]	0.598	[0.493, 0.690]	0.211
Firth logistic	0.771	[0.684, 0.871]	0.592	[0.500, 0.685]	0.209
Constant-prevalence benchmark	0.700		0.500		0.210
Training-prevalence predictor	0.685		0.426		0.213

Note: The test pool contains 140 transitions and 98 upgrades. Confidence intervals use firm-clustered bootstrap samples. Prevalence benchmarks receive no additional calibration.

Table 9 compares accounting information, rating history, and their combined feature set on the same test transitions. Rating history consists of the lagged overall rating, the lagged rating change, and rating-list tenure. The history-only ROC areas range from 0.642 to 0.651 and exceed the accounting-only and combined point estimates in each model family. Paired combined-minus-history differences range from -0.039 to -0.057, with intervals spanning zero. The average-precision ranking differs across model families, whereas the history-only specification has the highest ROC area in all three. Accounting-only ROC areas range from 0.540 to 0.570, and combined estimates range from 0.592 to 0.598. Prior assessment records therefore contain more ROC discrimination than the selected accounting features. Combining the information sets yields scores with ROC areas between the accounting-only and history-only estimates. Mean paired combined-minus-history ROC differences are -0.039 for logistic regression, -0.057 for boosting, and -0.048 for Firth regression. Combined-minus-accounting differences are 0.027, 0.053, and 0.024, respectively, with intervals including zero at their boundaries.

Table 9

Trajectory discrimination across accounting and rating-history information sets

Information set	Logistic ROC	Boosted ROC	Firth ROC
Accounting features only	0.570	0.540	0.568
Rating history only	0.642	0.651	0.643
Combined	0.598	0.593	0.592

Note: All specifications use the same 140 test transitions. Accounting features are defined in Table 2; rating-history measures are defined in the text.

Figure 4 presents mean absolute TreeSHAP values for the trajectory boosting model. Rating-list tenure and lagged rating change have the largest attributions. The current ratio lagged overall rating, and financial-panel tenure followed. Both tenure variables have negative directional associations with their signed attributions, while the current ratio has a positive association. Red and blue bars denote positive and negative directional associations, respectively. Firm size has a smaller relative magnitude than in the entry model. The lagged overall rating has a non-monotone attribution pattern across quartiles. Mean attributions are -0.083, -0.271, and -0.041 in the first three quartiles and 0.386 in the highest quartile. The unconditional transition distribution has its highest upgrade rate, 79 percent, in the lowest lagged overall rating quartile. Mean signed rating change declines from 0.103 in the lowest quartile to 0.022 in the highest. The conditional attribution and unconditional transition summaries describe different relationships with the prior assessment. The latter identifies a concentration of larger recorded increases among firms with lower preceding ratings.

The common-calendar comparison uses outcomes ending in 2021 to 2023. Entry features therefore come from 2020 to 2022. Entry ROC areas are 0.749, 0.710, and 0.601 for logistic regression, Firth, and boosting. The corresponding trajectory values are 0.622, 0.627, and 0.535, with overlapping intervals. Positive-class prevalence is 0.018 for entry and 0.632 for upgrades. Base-rate-

normalized average precision gives the trajectory task higher relative values, whereas ROC point estimates are higher for entry. The two metrics provide different cross-task comparisons. H3 is reflected in the feature decompositions, where firm size is prominent for first appearances and rating history is prominent for upgrades. The common-calendar estimates place this difference in information composition alongside the task-specific discrimination measures. The comparison uses the same outcome years while retaining each task's outcome definition, candidate population, and associated information set.

Under the 0.05-point upgrade threshold, the positive-class prevalence is 0.279. The boosting model has a ROC area of 0.724 and an average precision of 0.503. These estimates describe the alternative upgrade target on the same 140 test transitions. Increasing the threshold changes the positive labels and the allocation of events across the score ranking. The comparison therefore provides results for two distinct definitions of a recorded upgrade. The 0.005-point specification captures smaller annual improvements, while the 0.05-point specification concentrates on larger recorded changes. Their respective class frequencies and model metrics characterize the two monitoring targets.

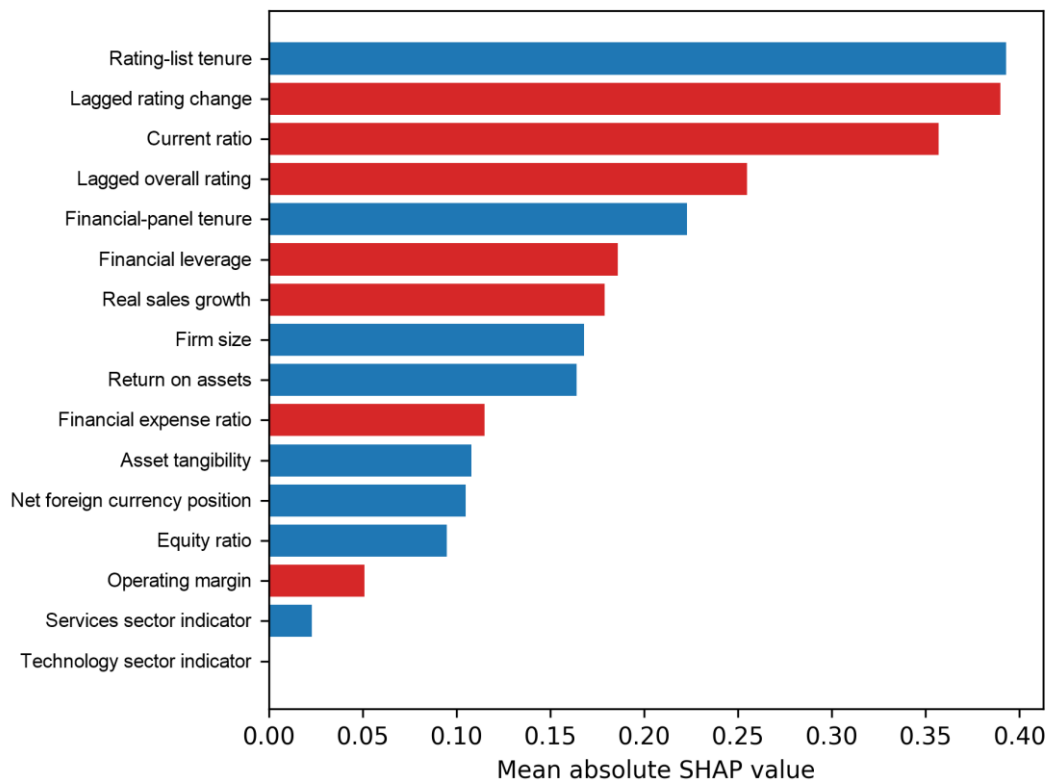


Fig. 4. Feature attributions for the trajectory boosting model

3.5. Pillar-Level Predictability

Table 10 compares the four section-level upgrade models. The stakeholders model has a ROC area of 0.676, with an interval of 0.568 to 0.774. The shareholder, public-disclosure, and board models have ROC areas of 0.558, 0.510, and 0.492, respectively. Their intervals include 0.500. Stakeholders therefore has the highest section-level ROC estimate and an interval entirely above chance. Its average precision is 0.455, compared with section-upgrade prevalence of 0.286. Section-level event frequencies differ across the test transitions. The board has the highest upgrade prevalence, at 0.629, and the stakeholders section has the lowest, at 0.286. The estimates identify

variation in the association between the selected financial features and annual improvements across governance sections. Upgrade counts are 46 for shareholders, 53 for public disclosure, 40 for stakeholders, and 88 for the board of directors.

Table 10

Out-of-time prediction of governance-section upgrades

Section	Upgrade prevalence	Average precision	ROC area	95% ROC interval
Shareholders	0.329	0.478	0.558	[0.430, 0.670]
Public disclosure	0.379	0.427	0.510	[0.409, 0.595]
Stakeholders	0.286	0.455	0.676	[0.568, 0.774]
Board of directors	0.629	0.653	0.492	[0.403, 0.581]

Note: Each governance-section model uses 140 test transitions. Confidence intervals are firm-clustered bootstrap intervals.

3.6. Robustness

Table 11 compares alternative input specifications and estimation choices. With 2nd and 98th percentile winsorization, entry ROC areas are 0.709 for logistic regression and 0.702 for Firth regression. Removing sector indicators preserves the ordering of the 3 model families. Excluding the 2022 feature year removes the 10-event appearance wave in 2023 and leaves 7 test events. Logistic average precision is 0.174 in this test subset, and Firth has a ROC area of 0.774. SMOTE-based training gives average precision of 0.088 for logistic regression, 0.086 for Firth regression, and 0.020 for boosting. The response to this training choice differs across estimators. The reduced 5-feature specification gives ROC areas of 0.783 for logistic regression and 0.784 for Firth regression, compared with 0.713 and 0.691 in their full specifications. Reduced logistic average precision is 0.082, slightly below the full-model estimate of 0.089, while reduced Firth average precision is 0.055 compared with 0.030. Under the same feature reduction, boosting has average precision of 0.032 and a ROC area of 0.603, compared with 0.022 and 0.598 in the full model. Feature reduction is therefore associated with a larger change in ROC discrimination for the linear estimators than for the tree ensemble. The contrasting changes in average precision and ROC area also indicate that a reduced input set can alter event placement differently across parts of the ranking. Concentrated accounting information retains substantial entry discrimination in the linear specifications, while the boosting comparison shows a smaller difference in its pooled ROC estimate.

The nominal-feature specification has a logistic ROC area of 0.765, compared with 0.713 under the main deflated specification. This comparison concerns the treatment of assets and real sales growth in the predictive inputs. Within-year percentile ranks give ROC areas of 0.724 for logistic regression and 0.747 for Firth regression. Ranking represents each financial observation relative to the other firms in the same year. Both linear estimators have ROC point estimates above 0.70 under the nominal and rank-based specifications. For the trajectory module, 2nd and 98th percentile winsorization gives average precision of 0.771 for logistic regression, 0.766 for Firth regression, and 0.773 for boosting. Their ROC areas are 0.594, 0.586, and 0.596, respectively. These ROC estimates remain close to the main trajectory values of 0.598, 0.592, and 0.593. The corresponding changes in average precision are also small, from main estimates of 0.778, 0.771, and 0.771. Altering the clipping bounds therefore leaves the broad level of trajectory discrimination similar in this sample, although the ordering of the model point estimates changes. Table 11 separates this trajectory comparison from the entry specifications and reports both performance measures for every estimator. The results describe sensitivity to the stated input and estimation choices, with different responses across model families and prediction tasks.

Table 11

Sensitivity of entry and trajectory predictions to alternative specifications

Specification	Logistic AP	Logistic ROC	Firth AP	Firth ROC	Boosted AP	Boosted ROC
Main specification	0.089	0.713	0.030	0.691	0.022	0.598
Winsorization at the 2nd and 98th percentiles	0.088	0.709	0.027	0.702	0.019	0.580
No sector indicators	0.111	0.725	0.033	0.706	0.021	0.624
Exclusion of the 2022 feature year	0.174	0.764	0.033	0.774	0.013	0.573
Nominal features	0.115	0.765	0.031	0.722	0.023	0.662
Percentile ranks	0.084	0.724	0.059	0.747	0.047	0.613
SMOTE training	0.088	0.705	0.086	0.699	0.020	0.660
Reduced five-feature specification	0.082	0.783	0.055	0.784	0.032	0.603
Trajectory winsorization at the 2nd and 98th percentiles	0.771	0.594	0.766	0.586	0.773	0.596

Note: AP denotes average precision and ROC denotes area under the receiver operating characteristic curve. All rows except the final row refer to entry models. The final row uses the trajectory test sample in Table 8.

3.7. Exit Events as Case Evidence

Two firms left the rating list during the sample window, and both sat near the bottom of the cross-sectional grade distribution before leaving. CEMAS held the lowest grade among rated firms in both of its rated years. Its grade was 7.62 in 2014 and 8.00 in 2015, and its board section score was 7.12 in the first year. PETKM entered with 8.91 in 2013 and held grades near 9.0. These absolute levels still placed the firm at or near the bottom quarter of the rated cross-section. Its ranks from the bottom were five of 23 in 2013, seven of 25 in 2014, and six of 27 in 2015.

For the two firms' final observed rating transitions, the trajectory model assigned upgrade scores between 0.60 and 0.93. Both firms recorded improvements in their last rated year and subsequently left the list. Their histories therefore combine relatively low grades, final-year increases, and discontinuation of the annual rating record. The score estimates relate to the observed upgrade transitions preceding departure. CEMAS and PETKM provide two firm-level sequences in which an increase in the overall grade was followed by the end of the recorded assessment series.

4. Discussion

4.1. Entry Predictability in the Adoption and Certification Literatures

First observed governance-rating adoption is associated with financial characteristics recorded in statutory accounts. Both linear estimators achieve ROC intervals above chance, while the regularized logistic model has a positive paired average-precision difference relative to training prevalence. This annual comparison adds a predictive dimension to studies of voluntary governance adoption. Research on certification has primarily examined financing conditions and firm outcomes [28, 29, 35]. Here, the accounting information precedes the observed assessment. The concentration of discrimination in firm size is consistent with a relationship between organizational resources and participation in externally assessed governance arrangements. Voluntary adoption studies similarly associate governance choices and listing-segment participation with ownership characteristics, free-float requirements, and financing needs [36, 37]. Signaling theory relates the credibility of certification to the resources involved in establishing and communicating organizational practices [1–3]. Large firms may have more extensive reporting systems and greater demand for external assessment from creditors and investors. Financial leverage represents the scale of financial obligations relative to assets, while profitability describes operating resources available alongside those obligations. Their joint presence in the entrant profile connects the predictions with the financing relationships examined in certification research [72, 73]. The theoretical relevance lies in

the association between observable financial capacity and participation in a governance-assessment institution.

Rating-list coverage expanded markedly in 2022 and 2023, while the distribution of overall grades remained concentrated. Later entrants have an average grade close to that of the initial cohort and are concentrated toward the upper part of the asset distribution. These observations distinguish the growth in the population assessed from changes in assessment levels. That distinction matters when index coverage is used to characterize governance development in a market. Research on index membership examines corporate visibility and information availability around changes in the population represented by an index [25]. An annual entry ranking addresses a related question before the observed appearance. It identifies firms whose financial profiles resemble those of subsequent entrants. The coverage series and the predictive results together describe the financial composition of the institution's expansion.

4.2. Parsimony as a Finding

A single asset measure accounts for substantial entry discrimination. Its ROC area is close to the regularized logistic estimate, and paired intervals include zero for every multivariate comparison. The reduced linear specifications retain similar discrimination with five financial features. Reduced logistic regression has a higher ROC point estimate than the full model and slightly lower average precision. With the same 5 features, boosting has average precision of 0.032 and a ROC area of 0.603, compared with 0.022 and 0.598 in its full specification. This smaller change in ROC area indicates that the discrimination associated with feature reduction varies across model families. Financial prediction benchmarks document comparable cases in which regularized models remain competitive with flexible estimators [45–47]. In the present setting, this comparison has a substantive implication beyond estimator choice. Much of the accounting information associated with adoption is concentrated in scale and a limited financial profile. Adding further ratios changes the allocation of individual scores more than it changes aggregate discrimination. The multivariate decompositions show how profitability, liquidity, financial leverage, and financing structure contribute alongside assets. A firm size ranking remains a transparent alternative, while the decompositions provide greater financial detail. The fixed-capacity comparison gives seven captured events for reduced logistic regression and five for firm size. This difference describes the candidate lists selected by the two rules and complements their pooled performance estimates.

For upgrades, information selection is more prominent than the difference between linear and tree-based estimators. Combined-set ROC estimates are similar across model families, but rating history has higher ROC point estimates in every family. The comparison places the observed assessment record at the center of progression prediction. Financial scale is associated with entry into the rating system, whereas the preceding grades describe the assessment position from which a rated firm progresses. These distinctions are relevant to explainable financial applications, where model performance and feature contributions serve different analytical purposes. Brier scores add a further dimension by comparing calibrated probabilities with window-specific prevalence. Calibrated Firth regression has the lowest trajectory Brier point estimate, at 0.209, close to the constant-prevalence value of 0.210 and the training-prevalence value of 0.213. The small differences describe the probability outputs alongside the similar combined-set ROC estimates. Joint examination of discrimination, calibration, and attribution links the present design with research on explainable finance and probability calibration [51, 68].

4.3. Accounting Information and Rating History

The entry and trajectory decompositions address the information-composition difference specified in H3. Firm size has the largest attribution in both entry models, while rating-list tenure and lagged rating change are prominent in the trajectory model. Initial adoption and subsequent upgrades thus involve different associations with the observed firm characteristics. The adoption model ranks firms before their first recorded assessment and uses a financial profile associated with participation. The trajectory model starts from an existing grade and uses the history of that assessment alongside the accounts. Its history variables summarize prior levels, recent changes, and time under rating coverage. The common-calendar comparison gives higher ROC point estimates for entry, whereas base-rate-normalized average precision gives a different ordering. Performance comparisons therefore depend on the evaluation measure as well as the event distribution. The feature findings contribute a distinct result about which information is associated with each stage. This distinction connects the annual predictions to the institutional sequence of obtaining an assessment and subsequently updating it.

Delegated-monitoring and disclosure frameworks describe how external assessors organize information relevant to corporate relationships [43, 44]. Governance-rating agencies examine public disclosures and information collected through firm contact and document review [61]. A rating history records the accumulated outcomes of this assessment process. Its contribution to subsequent upgrade scores has a parallel in ESG research, where historical scores are informative for forecasts of future performance [41]. Accounting ratios describe resources, operating outcomes, and financing positions. Governance-specific records describe board arrangements, ownership, and the procedures evaluated by the agency. The difference between these information categories helps explain the relevance of comparing accounting-only and history-only models. Prior grades also locate a firm within the assessment scale and summarize earlier changes. Their predictive contribution is therefore related to the sequence and level of the observed assessments. The history-only results connect progression prediction with that institutional record, while the accounting-only results describe the contribution of the selected financial characteristics.

4.4. Section-Level Evidence and Exits

Stakeholder assessments have the highest section-level ROC estimate. This section covers relationships with employees, creditors, and other corporate counterparties [55, 56]. Liquidity, operating performance, and financing obligations are financial aspects of these relationships that appear in statutory accounts. The stakeholders result is consistent with predictive information in those financial measures for its recorded upgrades. Shareholder rights, public disclosure, and board arrangements have lower ROC estimates under the same inputs. Their assessments concern voting arrangements, reporting procedures, and governance documentation more directly. A uniform accounting feature set may consequently relate differently to the four sections. The disaggregated analysis identifies where financial information is most strongly associated with observed score changes. It also provides a basis for considering section-specific inputs in governance monitoring.

The CEMAS and PETKM records distinguish rating changes from continuity of coverage. Both firms were near the lower part of the contemporaneous grade distribution and recorded increases in their final assessed years. A model of upgrades follows changes among firms with consecutive ratings. A model of departures would instead concern the end of the observed series. These outcomes are institutionally related but involve different labels and candidate populations. Monitoring systems that track both governance grades and participation would therefore require separate records of assessment changes and coverage status. The two histories illustrate their distinction within the observed rating panel.

4.5. Implications for Sustainable Governance Monitoring

The candidate rankings have practical relevance for voluntary governance outreach. Selecting the highest-scoring decile in each test year flags 137 firm-years. Full logistic regression captures five appearances, and the reduced specification captures seven. Exchanges and regulators could compare these candidate groups when planning information sessions, certification guidance, or engagement with firms considering external assessment. The feature decompositions describe their scale, financial leverage, profitability, and liquidity. This financial detail can support differentiated discussions about reporting preparation and assessment requirements. A firm size ranking offers a simple alternative with comparable discrimination. The choice of rule consequently depends on the candidate composition and the detail needed for the outreach activity. Investors monitoring the growth of rated-firm coverage may also use the rankings to organize subsequent research on prospective participants. Adoption propensity concerns participation in external assessment, while the grades themselves concern assessed governance practices. An inclusive outreach objective also extends to smaller firms outside the highest-scoring groups. These applications place predictive information within a defined governance-monitoring activity and extend decision-support research in sustainable development [5, 6]. They complement the financing literature on the same certification setting [35] and the financial-reporting context of the Turkish market [57, 58].

4.6. Limitations

The outcome definition and observation period impose several limitations. First appearance is measured in the 2013 to 2024 rating list. This event differs from a dated index admission, because rating publication and index membership follow separate records and eligibility requirements. The index predates the observation window, having operated since 2007. Pre-2013 rating histories are unavailable, so the first observed appearance may differ from a firm's first historical assessment. Although no firm leaves and returns within the observed window, earlier participation is outside the recorded history. Listing eligibility is also measured at annual resolution. The outcome-year rule approximates exposure during the first trading year and does not use dated rating cycles. Financial-panel tenure starts with observed records from 2011 and is left-censored for firms with longer reporting histories.

The timing of predictor availability is based on annual reporting calendars. Publication dates differ across firms and years, and they have not been matched to individual rating-announcement dates. The archived statement source may also contain later restatements. Retaining the first record in source order produces unique keys but does not establish the earliest publicly available reporting version. These features make the analysis a retrospective annual prediction design. The financial data do not provide a complete point-in-time information set for every rating event. The 2023 transition to TMS 29 introduces a related reporting-basis issue. Statements prepared under this standard are expressed in end-of-period purchasing power, while the annual GDP deflator applies a common price-level adjustment. Their combination involves different conventions. Within-year ranks remove the common annual price level from the cross-sectional comparison but retain firm-specific differences arising from asset composition, acquisition dates, and restatement practices.

Probability calibration is conditional on preprocessing learned from the complete outer training window. Internal forward predictions use earlier blocks for model fitting, but the matrix supplied to these fits already contains outer-window clipping and median replacements. Subsequent internal blocks therefore contribute to the feature-distribution estimates used in that preprocessing. Reapplying clipping within an earlier block does not recover the original missing-value mask or restore previously truncated observations. This internal dependence differs from use of the held-out outer test year, which remains excluded from fitting and preprocessing estimation. It limits the

interpretation of the internal calibration scores as fully fold-independent forward predictions. The prevalence assignment for non-positive calibration slopes is an additional algorithmic convention. It responds to inversions in the fitted mappings and is evaluated here within the same retrospective analysis. The resulting probabilities describe the implemented calibration procedure rather than prospective individual-firm risk estimates.

The entry sample contains 24 events, of which 17 occur in the test pool. Annual results in the first two test years are based on one event each. Average precision is sensitive to individual event positions, as the single-event deletion estimates show. Firm-level bootstrap intervals account for repeated test observations but remain conditional on the estimated models and supplied scores. They exclude uncertainty from refitting, preprocessing estimation, and model selection. The paired intervals do not establish multivariate superiority over the firm size benchmark. Similarly, the top-decile capture differences and decision-curve points provide observed operating summaries without interval-based evidence of improved decision value. Their threshold trade-offs are not linked to monetary costs. The high share of flagged non-entrants and sensitivity to the choice of screening objective are relevant to any prospective use.

The reduced entry specification, alternative upgrade threshold, section-level comparisons, and calibration guard are exploratory components of the analysis. Their selection and evaluation use the available sample, and the comparisons are not based on a pre-registered protocol. Section-level ROC intervals are marginal comparisons without a joint multiple-testing adjustment. Attributions summarize final fitted models rather than population effects. Fold-wise rankings share observations through expanding windows and do not constitute independent stability tests. The two exit sequences provide case descriptions but insufficient observations for estimating a separate departure model. These features constrain inference about model superiority, section-specific mechanisms, and the causes of changes in rating-list participation.

The information set contains accounting items, sector indicators, and the specified rating-history measures. Ownership structure, free float, auditor identity, board composition, market prices, and textual disclosures are not represented. Firm size may be associated with disclosure capacity, market visibility, investor demand, and agency-client selection. The financial variables do not distinguish among these channels. Likewise, the contribution of rating history may reflect cumulative assessment, omitted public governance information, measurement conventions, or differences among agencies. Concentration near the upper end of the ten-point scale may affect the distribution of recorded upgrades. The analysis estimates predictive relationships and does not identify causal effects of financial characteristics on adoption or of certification on financing. The setting comprises one emerging market and a rating list ending in 2024. Results may differ under other institutional arrangements or coverage regimes. A 2025 outcome extension requires a later rating-list vintage, even though the financial source contains statements through that year.

4.7. Future Research

Further research could combine the accounting features with ownership, free float, board composition, and disclosure characteristics. These inputs would provide a direct comparison between financial capacity and governance arrangements in the adoption models. Firm-level publication timestamps from the Public Disclosure Platform would support prediction horizons tied to the actual availability of each financial statement. Archived statement versions would also allow the observation date and reporting basis of the predictors to be specified jointly. Textual disclosures could provide measures of governance information beyond the ratios used here. Longer rating histories and multi-market datasets would expand the analysis of coverage continuity and provide observations for separate exit models. Prospective evaluations could compare the composition and

event capture of candidate lists generated by firm size and multivariate specifications. Linking annual candidate scores with subsequent bond-market outcomes would connect this predictive comparison with the certification and index-financing literature [26, 27, 35, 73].

5. Conclusions

This study examines first observed governance-rating adoption and subsequent upgrades in the Borsa Istanbul setting. It combines non-financial rating records for 2013 to 2024 with statutory statements from the listed-firm panel. Expanding-window evaluation compares regularized logistic regression, Firth penalized logistic regression, and gradient-boosted trees. Both linear entry models have pooled ROC intervals above chance. The regularized logistic model has a positive paired average-precision difference relative to training prevalence, with a lower interval endpoint close to zero. Deflated firm size provides discrimination comparable to that of the multivariate entry models, and reduced financial specifications retain substantial ROC performance. For upgrades, rating-history models have higher ROC point estimates than accounting-only and combined models. Feature decompositions distinguish the prominence of scale in adoption from that of assessment history in progression. These results characterize the information associated with entering an external rating system and updating an established assessment. Candidate rankings, feature contributions, and fixed-capacity event capture provide complementary outputs for governance monitoring. The study links observable financial profiles with the annual coverage and development of externally assessed corporate governance.

Acknowledgment

This research was not funded by any grant.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability

Governance-rating records derive from public disclosures of Borsa Istanbul and the Public Disclosure Platform. Financial statements derive from public annual filings. The GDP deflator growth series is available through the World Development Indicators [62]. Supplementary materials contain the event and listing records, feature matrices, rating pairs, section scores, calibration diagnostics, internal predictions, and out-of-sample scores. The accompanying code estimates the core entry specifications, the combined trajectory model, and their calibration mappings. The supplied predictions also support recalculation of performance summaries for the reported information sets.

References

- [1] Terlaak, A., & King, A. A. (2006). The effect of certification with the ISO 9000 Quality Management Standard: A signaling approach. *Journal of Economic Behavior & Organization*, 60(4), 579-602. <https://doi.org/10.1016/j.jebo.2004.09.012>
- [2] Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39-67. <https://doi.org/10.1177/0149206310388419>
- [3] Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355-374. <https://doi.org/10.2307/1882010>
- [4] Claessens, S., & Yurtoglu, B. B. (2013). Corporate governance in emerging markets: A survey. *Emerging Markets Review*, 15, 1-33. <https://doi.org/10.1016/j.ememar.2012.03.002>
- [5] Canitez, F. (2026). An Explainable AI Framework for Identifying Transferable Sustainable Urban Mobility Policies: Evidence from Istanbul. *Journal of Expert Systems and Sustainable Development*, 2(1), 152-168. <https://doi.org/10.65069/jessd21202619>

- [6] Yüksel, S., Olaru, G. O., Eti, S., & Dinçer, H. (2026). Psychological Determinants of Investment Decisions: An Integrated Sierpinski Triangle Fuzzy-based Decision-Making Model for Enhancing Financial Well-Being. *Journal of Expert Systems and Sustainable Development*, 2(1), 169-190. <https://doi.org/10.65069/jessd21202620>
- [7] Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305-360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
- [8] Sengupta, P. (1998). Corporate disclosure quality and the cost of debt. *The Accounting Review*, 73(4), 459-474. <https://doi.org/10.2308/tar-1325290>
- [9] Bhojraj, S., & Sengupta, P. (2003). Effect of corporate governance on bond ratings and yields: The role of institutional investors and outside directors. *The Journal of Business*, 76(3), 455-475. <https://doi.org/10.1086/344114>
- [10] Ashbaugh-Skaife, H., Collins, D. W., & LaFond, R. (2006). The effects of corporate governance on firms' credit ratings. *Journal of Accounting and Economics*, 42(1-2), 203-243. <https://doi.org/10.1016/j.jacceco.2006.02.003>
- [11] Anderson, R. C., Mansi, S. A., & Reeb, D. M. (2004). Board characteristics, accounting report integrity, and the cost of debt. *Journal of Accounting and Economics*, 37(3), 315-342. <https://doi.org/10.1016/j.jacceco.2004.01.004>
- [12] Klock, M. S., Mansi, S. A., & Maxwell, W. F. (2005). Does corporate governance matter to bondholders? *Journal of Financial and Quantitative Analysis*, 40(4), 693-719. <https://doi.org/10.1017/S0022109000001940>
- [13] Armstrong, C. S., Guay, W. R., & Weber, J. P. (2010). The role of information and financial reporting in corporate governance and debt contracting. *Journal of Accounting and Economics*, 50(2-3), 179-234. <https://doi.org/10.1016/j.jacceco.2010.10.001>
- [14] Bradley, M., & Chen, D. (2015). Does board independence reduce the cost of debt? *Financial Management*, 44(1), 15-47. <https://doi.org/10.1111/fima.12068>
- [15] Palmieri, E., Geretto, E. F., Polato, M., & Miani, S. (2025). Alternative finance in bank-firm relationship: how does board structure affect the cost of debt? *Journal of Management and Governance*, 29(1), 111-143. <https://doi.org/10.1007/s10997-024-09700-4>
- [16] Dhaliwal, D., Hogan, C., Trezevant, R., & Wilkins, M. (2011). Internal control disclosures, monitoring, and the cost of debt. *The Accounting Review*, 86(4), 1131-1156. <https://doi.org/10.2308/accr-10043>
- [17] Wang, K. T., Wu, Y., & Ho, K. Y. (2021). Internal control reporting and cost of bond financing: Evidence from China. *International Review of Economics & Finance*, 76, 1323-1346. <https://doi.org/10.1016/j.iref.2019.11.016>
- [18] DeBoskey, D. G., Li, Y., Lobo, G. J., & Luo, Y. (2021). Corporate political transparency and the cost of debt. *Review of Quantitative Finance and Accounting*, 57(1), 111-145. <https://doi.org/10.1007/s11156-020-00941-6>
- [19] Boubakri, N., & Ghouma, H. (2010). Control/ownership structure, creditor rights protection, and the cost of debt financing: International evidence. *Journal of Banking & Finance*, 34(10), 2481-2499. <https://doi.org/10.1016/j.jbankfin.2010.04.006>
- [20] Fields, L. P., Fraser, D. R., & Subrahmanyam, A. (2012). Board quality and the cost of debt capital: The case of bank loans. *Journal of Banking & Finance*, 36(5), 1536-1547. <https://doi.org/10.1016/j.jbankfin.2011.12.016>
- [21] Mande, V., Park, Y. K., & Son, M. (2012). Equity or debt financing: does good corporate governance matter?. *Corporate Governance: An International Review*, 20(2), 195-211. <https://doi.org/10.1111/j.1467-8683.2011.00897.x>
- [22] Ghouma, H., Ben-Nasr, H., & Yan, R. (2018). Corporate governance and cost of debt financing: Empirical evidence from Canada. *The Quarterly Review of Economics and Finance*, 67, 138-148. <https://doi.org/10.1016/j.qref.2017.06.004>
- [23] Liu, Y., & Wu, X. (2023). How does shareholder governance affect the cost of borrowing? Evidence from the passage of anti-takeover provisions. *Journal of Accounting and Economics*, 75(2-3), Article 101569. <https://doi.org/10.1016/j.jacceco.2022.101569>
- [24] Chiu, H. Z., Lin, C., & Wei, L. (2023). How do board reforms affect debt financing costs around the world?. *Journal of Financial and Quantitative Analysis*, 58(1), 217-249. <https://doi.org/10.1017/S0022109022000801>
- [25] Goyal, V. K., Urban, D., & Zhao, W. (2025). Index Creation, Information Changes, and Financing. *Journal of Accounting Research*, 63(5), 2181-2228. <https://doi.org/10.1111/1475-679X.12626>
- [26] Faulkner, M., Frost, T., & Ivanov, S. I. (2025). To be or not to be included in the S&P 500: cost of debt implications. *Managerial Finance*, 51(8), 1239-1256. <https://doi.org/10.1108/MF-04-2024-0270>
- [27] Jing, J., Rao, S., & Song, Y. (2025). Emerging stock market integration and corporate bond credit spreads. *Emerging Markets Finance and Trade*, 61(3), 684-699. <https://doi.org/10.1080/1540496X.2024.2399539>
- [28] Ullah, B. (2020). Signaling value of quality certification: Financing under asymmetric information. *Journal of Multinational Financial Management*, 55, Article 100629. <https://doi.org/10.1016/j.mulfin.2020.100629>
- [29] Sobhan, A., Bose, S., Miah, M. S., & Razzaque, R. M. R. (2024). Does certification of corporate governance compliance pay off? Evidence from a unique regulatory setting. *Corporate Governance: An International Review*, 32(4), 670-702. <https://doi.org/10.1111/corg.12563>

- [30] Darsono, D., Ratmono, D., Tujori, A., & Clarisa, T. Y. (2025). The relationship between ESG, financial performance, and cost of debt: the role of independent assurance. *Cogent Business & Management*, 12(1), Article 2437137. <https://doi.org/10.1080/23311975.2024.2437137>
- [31] Ararat, M., Black, B. S., & Yurtoglu, B. B. (2017). The effect of corporate governance on firm value and profitability: Time-series evidence from Turkey. *Emerging Markets Review*, 30, 113-132. <https://doi.org/10.1016/j.ememar.2016.10.001>
- [32] Özer, G., Çam, İ., & Çerez, S. (2023). Kurumsal yönetimin borç maliyetine etkisi: Borsa İstanbul'da bir uygulama. *Business and Economics Research Journal*, 14(2) 217-226. <https://doi.org/10.20409/berj.2023.412>
- [33] Biçer, M., & Şit, A. (2023). The impact of corporate governance quality on firm value: A case study on corporate governance index of Borsa İstanbul. *Central European Business Review*, 12(3), 1-19. <https://doi.org/10.18267/j.cebr.324>
- [34] Güney, C. (2020). Kurumsal Yönetim İlkelerine Uyum Derecelendirmesi: Borsa İstanbul Kurumsal Yönetim Endeksine (BIST-XKURY) Tabi Şirketler Üzerine Bir Analiz. *İşletme Araştırmaları Dergisi*, 12(2), 1297-1307. <https://doi.org/10.20491/isarder.2020.912>
- [35] Merter, A. K. (2026). Does Corporate Governance Certification Reduce the Cost of Debt? Evidence from the Borsa İstanbul Corporate Governance Index. *Borsa İstanbul Review*. Article 100885 <https://doi.org/10.1016/j.bir.2026.100885>
- [36] Anand, A. I., Milne, F., & Purda, L. D. (2006). Voluntary adoption of corporate governance mechanisms (Queen's Economics Department Working Paper No. 1112). Queen's University, Department of Economics. <https://doi.org/10.2139/ssrn.921450>
- [37] Sella, L. C., & Bortolon, P. M. (2022). Factors associated with the non-migration to the New Market of companies of the one share, one vote type. *Revista Contabilidade & Finanças*, 33(90), e1492 <https://doi.org/10.1590/1808-057x20221492.en>
- [38] Svanberg, J., Ardeschiri, T., Samsten, I., Öhman, P., Neidermeyer, P. E., Rana, T., Semenova, N., & Danielson, M. (2022). Corporate governance performance ratings with machine learning. *Intelligent Systems in Accounting, Finance and Management*, 29(1), 50-68. <https://doi.org/10.1002/isaf.1505>
- [39] D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2021). Fundamental ratios as predictors of ESG scores: A machine learning approach. *Decisions in Economics and Finance*, 44, 1087-1110. <https://doi.org/10.1007/s10203-021-00364-5>
- [40] Kartal, M. T., Kılıç Depren, S., Pata, U. K., Taşkın, D., & Şavlı, T. (2024). Modeling the link between environmental, social, and governance disclosures and scores: the case of publicly traded companies in the Borsa İstanbul Sustainability Index. *Financial Innovation*, 10, 80 <https://doi.org/10.1186/s40854-024-00619-1>
- [41] Taskin, D., Sariyer, G., Acar, E., & Cagli, E. C. (2025). Do past ESG scores efficiently predict future ESG performance?. *Research in International Business and Finance*, 74, 102706. <https://doi.org/10.1016/j.ribaf.2024.102706>
- [42] Zhang, J., & Zhao, Z. (2026). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems with Applications*, 295, Article 128809. <https://doi.org/10.1016/j.eswa.2025.128809>
- [43] Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *The Review of Economic Studies*, 51(3), 393-414. <https://doi.org/10.2307/2297430>
- [44] Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1-3), 405-440. [https://doi.org/10.1016/S0165-4101\(01\)00018-0](https://doi.org/10.1016/S0165-4101(01)00018-0)
- [45] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- [46] Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161, Article 113567. <https://doi.org/10.1016/j.eswa.2020.113567>
- [47] Dasilas, A., & Rigani, A. (2024). Machine learning techniques in bankruptcy prediction: A systematic literature review. *Expert Systems with Applications*, 255, Article 124761. <https://doi.org/10.1016/j.eswa.2024.124761>
- [48] Saha, D., Young, T. M., & Thacker, J. (2023). Predicting firm performance and size using machine learning with a Bayesian perspective. *Machine Learning with Applications*, 11, 100453. <https://doi.org/10.1016/j.mlwa.2023.100453>
- [49] Hezam, Y., Luong, H., & Anthonysamy, L. (2025). Machine learning in predicting firm performance: a systematic review. *China Accounting and Finance Review*, 27(3), 309-339. <https://doi.org/10.1108/CAFR-03-2024-0036>
- [50] Ustahaliloğlu, M. K. (2025). Artificial intelligence in corporate governance. *Corporate Law and Governance Review*, 7(1), 123-134. <https://doi.org/10.22495/clgrv7i1p11>
- [51] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8). 216. <https://doi.org/10.1007/s10462-024-10854-8>

- [52] Bueff, A. C., Cytryński, M., Calabrese, R., Jones, M., Roberts, J., Moore, J., & Brown, I. (2022). Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. *Expert Systems with Applications*, 202, Article 117271. <https://doi.org/10.1016/j.eswa.2022.117271>
- [53] Xu, Y., Wei, P., & Ji, Y. (2025). Machine learning for predicting corporate carbon emissions: The role of corporate governance. *Journal of Digital Economy*, 5, 1-16. <https://doi.org/10.1016/j.jdec.2025.11.008>
- [54] Borsa İstanbul. (2025). BIST Corporate Governance Index ground rules. Borsa İstanbul A.Ş. <https://www.borsaistanbul.com>
- [55] Capital Markets Board of Türkiye. (2005). Corporate governance rating methodology and index eligibility framework. Capital Markets Board
- [56] OECD. (2023). OECD Corporate Governance Factbook 2023, OECD Publishing, Paris, <https://doi.org/10.1787/6d912314-en>
- [57] Gürkaynak, R. S., Kisacıkoğlu, B., & Lee, S. S. (2023). Exchange rate and inflation under weak monetary policy: Turkey verifies theory. *Economic Policy*, 38(115), 519-560. <https://doi.org/10.1093/epolic/eiad020>
- [58] Ozkaya, A., & Altun, O. (2024). Domestic and global causes for exchange rate volatility: Evidence from Turkey. *Sage Open*, 14(2), Article 21582440241243200. <https://doi.org/10.1177/21582440241243200>
- [59] Kaviani, M. S., Kryzanowski, L., Maleki, H., & Savor, P. (2020). Policy uncertainty and corporate credit spreads. *Journal of Financial Economics*, 138(3), 838-865. <https://doi.org/10.1016/j.jfineco.2020.07.001>
- [60] Lins, K. V., Servaes, H., & Tamayo, A. (2017). Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *The Journal of Finance*, 72(4), 1785-1824. <https://doi.org/10.1111/jofi.12505>
- [61] JCR Eurasia Rating. (n.d.). Corporate governance rating methodology. <https://www.jcrer.com.tr/en/methodology/corporate-governance-rating-methodology>
- [62] [dataset] World Bank. (2025). World Development Indicators, GDP deflator, Türkiye, series NY.GDP.DEFL.KD.ZG [Data set]. World Bank. <https://databank.worldbank.org/source/world-development-indicators>
- [63] 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. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- [64] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38. <https://doi.org/10.1093/biomet/80.1.27>
- [65] Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409-2419. <https://doi.org/10.1002/sim.1047>
- [66] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- [67] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625-632. <https://doi.org/10.1145/1102351.1102430>
- [68] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The achilles heel of predictive analytics. *BMC Medicine*, 17(1), Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
- [69] 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>
- [70] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565-574. <https://doi.org/10.1177/0272989X06295361>
- [71] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- [72] Bellucci, A., Borisov, A., Giombini, G., & Zazzaro, A. (2023). Information asymmetry, external certification, and the cost of bank debt. *Journal of Corporate Finance*, 78, 102336. <https://doi.org/10.1016/j.jcorpfin.2022.102336>
- [73] Sufi, A. (2009). The real effects of debt certification: Evidence from the introduction of bank loan ratings. *The Review of Financial Studies*, 22(4), 1659-1691. <https://doi.org/10.1093/rfs/hhm061>