ICMEM

The 10th International Conference on Management in Emerging Markets

IPO Return Prediction: A Systematic Review on Methods, and the Case for Auditable Scorecards

Erwin Atmadja* and Taufik Faturohman
School of Business and Management, Institut Teknologi Bandung, Bandung, Indonesia

Abstract*—*Predicting the direction of an initial public offering (IPO) first-day return is difficult because the firm has little history and disclosure sits in a single prospectus. This review synthesises recent predictive modelling of IPO outcomes by horizon, method family, and three evaluation dimensions: validation design, data leakage and class imbalance. Of the 42 studies read at full text from Scopus and IEEE Xplore, a two-condition rule on outcome and model retains 37; evaluation quality is deliberately not a condition, so studies never evaluated out-of-sample are reported rather than screened out. Ensembles and textual methods outperform linear baselines in-sample, yet only six of the 37 are validated in time order. The binding weakness is a missing comparator: no benchmarks against a majority-class rule (always predicting the more common outcome), so most models cannot be shown to beat a naive baseline; a minimum future work should report. For retail-dominated emerging markets, we identify the auditable credit-style scorecard as an open direction: a Weight-of-Evidence points table that is deterministic, so a subscriber or regulator can reproduce and verify it, a property interpretable but stochastic rule systems do not provide.

Keywords*—*initial public offering, machine learning, first-day return, systematic literature review, data leakage, class imbalance, credit scorecard, Weight of Evidence, auditability

I. INTRODUCTION

Initial public offerings (IPOs) allow private firms to raise capital; however, the first day of trading is persistently anomalous: closing prices often exceed offer prices by a wide margin, a phenomenon known as underpricing [1], [2]. In emerging markets, IPOs can be either sharply underpriced or, less frequently, overpriced [3]. Classical explanations attribute underpricing to asymmetric information and a winner's curse among uninformed investors [1] and ex-ante uncertainty and the certifying role of reputable intermediaries [2]. A later synthesis argues that the fixed offer-price constraint is the unifying core explanation binding these accounts together [4]. These theories explain the magnitude of underpricing, but they are less directly informative about the direction of an individual issue's first-day return, which is the information an investor wants to anticipate before subscribing. Forecasting IPO outcomes is intrinsically challenging. A newly listed firm has little market history; disclosure is concentrated in a single prospectus, and demand is only partially revealed before the pricing. Researchers have historically relied on ordinary least squares and logistic regression to relate pre-IPO variables to outcomes. These models are transparent but assume linearity and are sensitive to outliers that characterise skewed IPO returns [3], [8]. Over the last decade, the field has shifted decisively toward data-driven machine learning (ML), deep learning, and textual analysis; motivated by their ability to capture nonlinear structures and to absorb high-dimensional, unstructured inputs. Prior reviews map this space bibliometrically [5] or catalogue the determinants of underpricing [6]. What remains underserved is a synthesis organised around the predictive task itself: what is predicted and over what horizon, which model families are used, and whether the reported performance is obtained under evaluation procedures that would survive deployment. This review addresses this gap. It organises the predictive literature by horizon, compares method families on explicit strengths and weaknesses, codes every study at full text on three trust-determining dimensions: validation design, data leakage, and class imbalance, and makes the case for auditable scorecards, distinguished not by interpretability, which several prior models already offer, but by deterministic reproducibility, as a direction suited to retail-dominated emerging markets such as the Indonesia Stock Exchange (IDX).

This review makes three contributions. First, it organises the literature by forecasting tasks, that is, horizon and method family, rather than by determinants of underpricing. Second, it isolates a missing comparator; no screened study benchmarks against a majority-class rule, an absence that motivates a concrete reporting standard. Third, it identifies the auditable Weight-of-Evidence scorecard as an unexplored direction for direction prediction in retail-dominated emerging markets and operationalises it (Table II), linking the review to a companion empirical study on the Indonesia Stock Exchange (IDX). The novelty claim is scoped to the searched corpus and to the absence of a WoE/IV scorecard for IPO first-day-return direction. The practical implications are a reporting checklist, time-ordered validation, ex-ante feature cut-off, majority-class baseline, and a reproducible scorecard that a subscriber or regulator can audit.

**Corresponding Author: erwinatmadja1@gmail.com*

II. METHODOLOGY

This review followed the PRISMA 2020 statement. Records were retrieved from Scopus on July 1, 2026. Scopus was chosen for its breadth across finance, business, and computing, the three fields covered by this topic. The following query was applied to the titles, abstracts, and keywords:

TITLE-ABS-KEY (("initial public offering" OR IPO) AND (scor* OR logistic* OR predict* OR forecast* OR classif* OR "machine learning" OR "neural network" OR "random forest" OR "gradient boosting" OR "support vector" OR "deep learning" OR "data mining") AND (underpric* OR "first-day return" OR "initial return" OR mispric* OR perform*)) AND PUBYEAR > 2016 AND PUBYEAR < 2027 AND (LIMIT-TO (LANGUAGE , "English"))

A total of 398 records were retrieved. To reduce the single-database bias, the search was replicated in IEEE Xplore, which indexes much of the computing and engineering literature on this topic. Because IEEE Xplore provides no single field spanning title, abstract, and keywords, the equivalent search was executed across its All Metadata field using the same three concept blocks and the same 2017-2026 window; it returned 52 records.

A study was included if it (i) concerned an equity IPO, (ii) developed or evaluated a predictive or statistical model of an IPO outcome, namely, first-day or initial return, underpricing, post-listing or long-run performance, or survival, failure, and withdrawal, and (iii) was peer-reviewed. Records were excluded if they used IPO in a non-equity sense, fell outside the offering domain, contained no predictive modelling component, or were retracted items or front matter. Foundational theory [1], [2], [4], and prior reviews [5], [6] were retained for framing only.

Screening proceeded in two passes: a transparent rule-based prescreening of the title, abstract, and keywords, refined by manual reading of borderline cases. Because IPO is heavily overloaded in the indexed literature, the prescreen removed acronym collisions (optimisation algorithms such as the Improved Parrot Optimizer) and unrelated domains such as streamflow forecasting and medical imaging. Fig. 1 shows the combined flow across both databases, from 450 identified records to 42 core studies. Every study was read in full and assessed against a two-condition scope rule: the outcome must be a return referenced to the offer price of a conventional IPO, and the study must fit a predictive model of that outcome. Five fail it. Four model an outcome outside the rule: a de-SPAC merger [22], three-year abnormal returns against a market benchmark, a regulator's admission decision [29], and post-listing price levels [17]. The fifth estimates a difference-in-differences policy effect with no machine-learning component and no model fitted for prediction. The rule deliberately does not condition on evaluation quality. Five further studies use machine learning to construct a textual, vocal, or sentiment regressor, while OLS or logit models the outcome, reporting in-sample fit only and no out-of-sample evaluation [26], [30], [31], [51], [52]. Screening a corpus on the same evaluation practice that the review sets out to measure would select the dependent variable and bias the audit toward a flattering picture. Therefore, they are retained and reported as a distinct category in Table I, so that the absence of out-of-sample evaluation is recorded as a finding rather than removed from view. The remaining 37 form the evaluation corpus coded in Section IV; all 42 were retained for the descriptive synthesis in Section III and listed in the supplementary corpus workbook. The scope is deliberately bounded rather than narrow: Section III spans the full range of IPO horizons, from listing-day and medium-horizon to long-run and to failure, withdrawal, or delisting, across all 42 studies, and the corpus contracts to 37 only for the evaluation coding, where a common outcome definition is needed for the counts to be comparable. Broadening the databases and the outcome set is a natural extension. Notably, 22 of the 52 IEEE Xplore records were already indexed in Scopus, and the entire search contributed one further predictive study [43]; the marginal yield of roughly two per cent is evidence of coverage saturation rather than of narrow search. A final full-record review reclassified five studies as ineligible, including an auction-mechanism experiment and a bibliographic review.

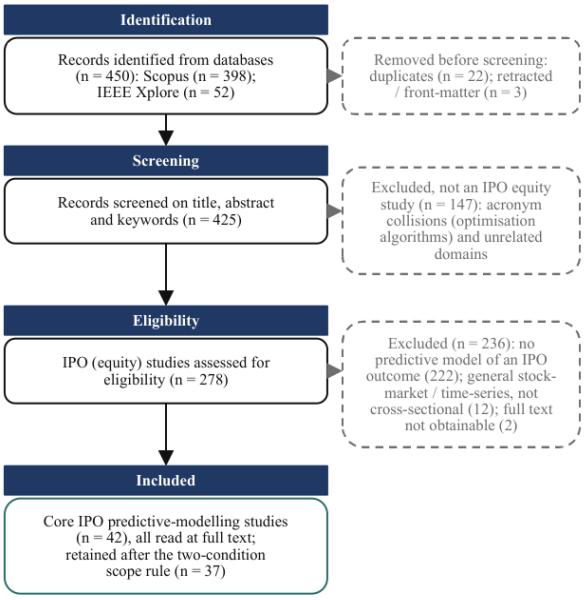


Fig. 1. PRISMA 2020 flow of identification, screening, and inclusion.

Each included study was then coded on three evaluation dimensions: the validation design reported, whether the predictors were framed as ex-ante, that is, available before the predicted event, and whether class imbalance was addressed. Each study was coded on three dimensions, namely validation design, ex-ante framing, and imbalance handling; the coding was done first from the bibliographic record and then re-derived from full text, with reported metrics recomputed from the published tables wherever a confusion matrix, class distribution, or split size allowed. This study had two limitations. First, although two databases were searched, other indexes such as Web of Science were not, and the field moves quickly enough that any corpus dates rapidly; the synthesis is therefore a current benchmark rather than a closed account. Second, because the included studies are heterogeneous in market, sample size and prediction target, no meta-analytic pooling is attempted. The first limitation bears on one gap we identify, the scarcity of Indonesian evidence: nationally

accredited or Bahasa Indonesia work is frequently absent from Scopus and IEEE Xplore, and so invisible to an English-language query. It surfaces instead through national indexes, chiefly Indonesia's SINTA (Science and Technology Index) portal and the Garuda repository it aggregates, alongside Google Scholar. A targeted post-hoc check, run outside the PRISMA flow and not counted in Table I, confirms the concern is material: a SINTA-3 accredited journal classifies 209 IDX issuers from the OJK e-IPO platform with seven algorithms [53]. Read at full text, it reproduces the failures documented here. SMOTE is applied before an 80:20 random split, so the 89.2% headline accuracy is measured on a synthetically rebalanced test set rather than at the sample's 77.0% underpricing prevalence; missing values and outliers are imputed from class-conditional means, routing the label into the predictors; and no majority-class comparator is reported. The local evidence is therefore not simply missing from the indexed corpus; it carries the same reporting gaps. Because SINTA aggregates rather than supporting reproducible Boolean querying, we did not treat it as a systematic source; a Garuda-based search is a scoped extension for the companion Indonesian study.

III. SYNTHESIS: PREDICTION TARGETS AND METHOD FAMILIES

A. Prediction targets: short-term and long-term

The single most important design choice in this literature is what is predicted. The target determines the relevant predictors, appropriate metric, and severity of the leakage and imbalance problems discussed later.

At the listing-day horizon, the target is the initial return, in two recurring formulations: regression on the magnitude of underpricing, and classification of its direction, that is, whether the issue closes above its offer price. The most comprehensive recent treatment assembles 863 pre-IPO variables and reports that gradient boosting (XGBoost) estimates both the direction and the magnitude of US first-day returns, with underwriter agency measures, price revision and free-float fraction most influential [10]. A Chinese binary-classification framework over nearly two thousand issues finds pre-listing fundamentals, especially earnings per share and net asset value, most informative, with random forest the strongest learner [11]. The 'breaking' variant, where a stock closes below offer, is studied through prospectus text [28], and a conference study classifies issues by both subscription status and positive versus negative listing gains [43].

At medium horizons, the target shifts to cumulative or abnormal returns over days to months: prospectus sentiment predicts the direction of price movement over the first 3 to 30 days [24], and a multilayer perceptron weighs macroeconomic, issue-specific, technical, and fundamental factors across 3-to-12-month windows [17]. At the long horizon, the target becomes long-run abnormal or buy-and-hold returns and post-IPO operating performance, where strong pre-IPO enthusiasm is associated with high initial returns but subsequent underperformance [30]. A distinct strand predicts outcomes rather than returns: failure or delisting [34], under-performance classes [35], and withdrawal [37]. One study examines underpricing, long-term underperformance, and delisting jointly [33], underlining that the horizons are related but not interchangeable.

B. Methodological families: strengths, weaknesses and comparison

Across the screened corpus, six method families recurred. The discussion below compares them and flags where common claims require qualification.

1) Linear and logistic regression

Regression remains the baseline against which other methods are measured and, for many Indonesian studies, is the primary tool [40], [42]. Its appeal is transparency; its weaknesses are linearity, limited interaction capacity, and sensitivity to outliers endemic to IPO returns [3], [8]. Interpretability keeps it central where decisions must be justified.

2) Tree-based ensembles

Random forests and gradient-boosting variants are the workhorses of recent literature, capturing non-linearities and interactions and yielding SHAP rankings. Random forest predicts US underpricing [7], outperforms robust regression on Borsa Istanbul [8], dominates fixed-price issues where demand proxies rank highest [9], and leads a binary comparison in China [11]. Gradient boosting is the strongest learner for first-day US return [10]. EDA pipelines report high gain-prediction accuracy [12], and a 2026 Malaysian analysis ranks random forest ahead of neural networks and linear regression [45]. The headline that ensembles beat linear models is robust in-sample but carries two qualifications taken up in Section IV: importance is associational, not causal, and small emerging market samples invite overfitting.

3) Neural networks and deep learning

Deep models target a hierarchical structure: a stochastic-frontier deep network estimates deliberate underpricing without distributional assumptions, isolating negative net income and EBITDA as drivers [18]. The cost is that deep models are data-hungry and unstable on small datasets typical of single-market IPO studies, and their opacity complicates the justification.

4) Neuro-fuzzy and fuzzy systems

A smaller strand pursues interpretability directly: a genetic rough-set neuro-fuzzy system exposes human-readable rules [15], and fuzzy techniques are applied to the same task [16], showing that nonlinear learning and transparency are not mutually exclusive, although the rules come from stochastic search, a limitation taken up in Section IV.D, and design complexity has kept adoption narrow.

5) Textual analysis and natural language processing

The fastest-growing strand mines the prospectus. Forward-looking statements inform price-revision analysis [23], roadshow tone is associated with first-day returns [25], and transformer-encoded news sentiment predicts US underpricing [26]. Textual features complement financial variables: adding S-1 text raises accuracy in a large US sample [27], and prospectus similarity, low readability, and negative sentiment raise the probability of ‘breaking’ in China [28]. The trade-offs are dimensionality, subjective tone labelling, and market-specificity; a subtler risk is leakage, since any disclosure that post-dates pricing must not predict listing-day outcomes.

6) Hybrid and interpretable approaches

The final family trades a little accuracy for transparency. A hybrid of logistic regression and shallow trees matches complex ML while surpassing it in interpretability [21]; ML outputs are fused with multi-

criteria ranking, so top-ranked issues outperform the lowest [13]; and multinomial logistic regression classifies Saudi IPOs into performance bands, evaluated with AUC and F1 [19]. Thai studies predict post-IPO performance and underpricing using class-based and hybrid schemes [20], [32], and the 2026 work continues in this vein [44], [46]. This family is the natural home of the scorecard idea developed in Section IV.D.

IV. FINDINGS AND DISCUSSION

Method choice receives the most attention in this literature, but whether a reported number is trustworthy depends on three evaluation decisions that are frequently under-reported. Table I reports how often each dimension is documented across the 37 studies retained by the scope rule of Section II, coded from the full text.

Table I.
Reporting of the Three Evaluation Dimensions across the Included Corpus

Evaluation dimension	Reporting category	Studies (n = 37); references
Out-of-sample evaluation and validation design	Explicitly described (of which time-ordered: 6)	22; [7], [10], [13], [15], [16], [18], [19], [21], [23], [24], [27], [28], [33], [34], [35], [37], [38], [39], [44], [45], [47], [50]
	Not stated or not determinable from the record	10; [8], [9], [11], [12], [14], [20], [32], [36], [43], [46]
	No out-of-sample evaluation performed (model fitted and reported in-sample only)	5; [26], [30], [31], [51], [52]
Ex-ante predictor framing	Explicitly stated as ex-ante	25; [8], [9], [10], [13], [18], [19], [20], [21], [23], [24], [26], [27], [28], [30], [32], [33], [34], [35], [37], [44], [45], [46], [50], [51], [52]
	Not stated	10; [7], [11], [12], [14], [15], [16], [31], [36], [38], [47]
	Implied by the predictors, not discussed	2; [39], [43]
Class imbalance	Explicit technique reported (SMOTE, SMOTE-Tomek, undersampling or cost-sensitive learning)	6; [13], [23], [28], [35], [36], [47]
	Not stated	19; [10], [11], [14], [15], [19], [20], [21], [24], [27], [32], [33], [34], [37], [43], [44], [46], [50], [51], [52]
	Not applicable (continuous target, not classification)	12; [7], [8], [9], [12], [16], [18], [26], [30], [31], [38], [39], [45]

Reference numbers correspond to the list in the References section. Not stated denotes absence of reporting, not absence of control.

A. Out-of-sample evaluation and validation design

Since IPO data are arranged chronologically, it is essential that validation adheres to this sequence. Random k-fold cross-validation shuffles issues across time, letting later issues inform predictions about earlier ones. The defensible alternative is temporal: expanding- or rolling-window validation that trains in the past and tests in the future. The most direct evidence for why this matters comes from a study of IPO withdrawal: random forest outperforms logit and lasso when training and test sets are drawn from the same period, yet when models are trained on past data and tested on future issues, all of them fail, a result attributed to concept drift in the predictor-outcome relationship [37]. Consistent with this, the strongest first-day-return study evaluated its model with a rolling-window procedure rather than a single random split [10]. Measured against that benchmark, the corpus divides sharply, and the division is only visible in the full text. Validation is systematically under-disclosed at the abstract level. What full-text reading does not rescue is temporal rigour. Only six of the 37 studies validate in a way that respects chronology, namely the rolling window of [10], the dual temporal-versus-cross-sectional comparison of [37], the in-fold feature selection with expanding origin of [34], the chronological hold-out of [28], and the blocked designs of [33] and [50]; the remaining 16 that describe a validation design use random splits or k-fold cross-validation that shuffle issues across time. Of these six, only [28] handles a mid-sample regulatory regime break explicitly, standardising its target on either side of a rule change that halved the mean initial returns. A further five studies never reached the question at all: they fit a predictive model of an offer-price-referenced return but report in-sample fit only, with no held-out evaluation of any kind [26], [30], [31], [51], [52]. In each, machine learning constructs a textual, vocal, or sentiment regressor and an OLS or logit specification then models the outcome, so the machine-learning component serves measurement rather than prediction. Counting these studies rather than excluding them is deliberate; they are the limiting case of the reporting problem that this section documents. The implication is that cross-sectional out-of-sample accuracy can substantially overstate live performance and that temporal validation should be the default.

B. Data leakage

Leakage occurs when information unavailable at the prediction time influences the model: temporal leakage from shuffling time-ordered data, feature leakage from variables realised at or after the moment of prediction,

and target leakage, where a feature is constructed using the outcome. Ex-ante framing is the one dimension that the corpus reports reasonably well: 25 of the 37 retained studies explicitly state that their predictors are drawn from pre-listing information, and 10 leave the question open. Full-text reading also shows the framing to be less reliable than the count suggests, because one study coded ex ante on its record is disqualified on inspection: it includes four predictors measured on the listing day itself, and carries both an at-issuance and a listing-day version of the same price-to-book ratio, the listing-day version scoring seven times higher in importance, which is a leak the abstract does not reveal. This distinction matters. A 2026 Malaysian study framed explicitly around ex-ante information nonetheless ranks investor demand as its most influential predictor [45]; oversubscription is known before listing, but only after the subscription window has closed, so it is available to a model forecasting the listing-day outcome and unavailable to an investor deciding whether to subscribe. Whether a variable is genuinely ex-ante, therefore, depends on whose decision the model is meant to support, and that reference point is rarely stated. The remedy is a strict ex-ante feature cutoff: only information verifiably available before the predicted event should enter the model, and any contemporaneous demand or disclosure variable should be justified as genuinely pre-listing. Textual pipelines deserve particular care because some disclosures and media postdate pricing [25], [26]. Read across the corpus, leakage falls into four tiers of increasing severity, sharing one failure of temporal and definitional discipline. In the mildest, a predictor is realised after the subscription decision but before listing, as with the final oversubscription in [45], [9], [39], and [46]: genuinely pre-listing, yet unavailable to the subscriber the model purports to serve. Second, a predictor is contemporaneous with the outcome, as in the listing-day valuation ratios noted above. In the third, the target reappears among the predictors measured at another date: one Indian study forecasts a stock's 3-, 6-, and 12-month price from that same stock's earlier prices, an almost perfectly autocorrelated series. Fourth and most directly, the target's own arithmetic components are entered as features: a conference study predicts whether the subscription ratio exceeds one from subscription rates by investor category, whose weighted sum is that ratio, and duly reports a confusion matrix with no errors of either kind. None require an argument about market microstructure; each is visible in the study's own variable list.

C. Class imbalance

When the target is the return direction, the classes are usually imbalanced because most issues close above offer. The overall accuracy is then misleading: a model can score well by predicting the majority class while failing on the loss-making issue that investors care about. Prevalence-robust metrics are therefore essential, chiefly AUC and precision-recall analysis defined on the minority class, with F1, precision, recall, and the confusion matrix [19]. On the modelling side, resampling (for example SMOTE), undersampling, class weighting, and threshold tuning are all used; one emerging-market framework combines SMOTE with hyperparameter optimisation and risk-adjusted metrics [36], and another is built explicitly for small, imbalanced datasets [35]. Two cautions apply. Resampling must be performed inside cross-validation folds, never before splitting, or it leaks information. Synthetic oversampling of a very small minority can fabricate structure that does not generalise. A 2026 study that classifies issues as underpriced versus overpriced treats class imbalance and high dimensionality as explicit design concerns, indicating that recent work increasingly foregrounds these controls [47]. An explicit technique is used in 6 of the 37, spanning SMOTE, SMOTE-Tomek, training-fold undersampling, and cost-sensitive learning; 12 target a continuous outcome for which the question does not arise, and 19 classify without stating whether or how skew was handled. The deeper problem is not the technique, but the comparison. No study in the corpus benchmarks its classifier against a majority class or constant-prediction rule, the natural floor for an imbalanced target. Three studies gesture at chance in weaker forms, a random-guess baseline tabulated at every horizon in [24], a proportional-random rule in [22], and the observation that an AUC of one-half is chance in [23]; however, the strongest of these still omits the class distribution needed to compute the majority rate, so none supplies the comparator that matters. Where the missing baseline conceals a failure, it is decisive; where it does not, as in [28], the omission still costs interpretation the authors could have supplied at no cost. Balanced accuracy, reported by [13] and referenced to half regardless of prevalence, is the cheapest reporting fix available.

D. Toward auditable IPO scorecards

Two properties are easily conflated but distinct in this argument. Interpretability refers to whether a model's reasoning can be understood by a human, that is, whether one can read off why a case received the score it received. Auditability is stronger and orthogonal: whether the model is deterministic and reproducible, yielding identical artefacts whenever it is re-fit to the same data, so that any third party can regenerate and verify it exactly. A model can be interpretable yet not auditable; a rule set produced by stochastic search is legible but changes from run to run, and this gap is the crux of our argument. The evidence frames a practical tension: black-box ensembles tend to win accuracy comparisons yet are hard to interpret. In retail-dominated markets, such as the IDX, a model that is both interpretable and auditable can be more useful than a marginally more accurate but unexplainable one.

Section III.B.6 identified several threads pointing toward such models [13], [15], [21].

This contribution rests on auditability rather than interpretability. At least four prior IPO models are already interpretable; they include the rule system of Quintana, Luque, and Isasi, the neuro-fuzzy system of [15], the fuzzy rule bases of [16], and the piecewise-linear tree of [21]. None provides a deterministic, reproducible artefact; their rule sets come from a stochastic search, and [15] presents its rules as one realisation of ten runs. The credit-scoring paradigm operationalises the alternative: predictors are transformed by weight-of-evidence (WoE) values, screened by information value (IV), and a logistic regression is converted into an integer-points scorecard in which each characteristic contributes a transparent, monotonic score, the same table on every re-fit, auditable by construction, and familiar to regulators. The method is documented end-to-end by Siddiqi [48] and underpinned by the classification review of Hand and Henley [49], both cited as methodological foundations outside the IPO corpus.

The search query explicitly included scoring terms (scor*), so a credit-style scorecard study would have been retrieved had one existed. None was: no screened study applies a WoE/IV scorecard to IPO first-day-return direction. That gap defines the companion empirical study, a directionally focused, auditable scorecard for a retail-dominated emerging market.

To make the transformation concrete, it helps to see how it carries over from consumer credit to the IPO direction. In a credit scorecard, the modelled event is default: each applicant is labelled good (repaid) or bad (defaulted), a characteristic such as income band is split into bins, and each bin is assigned a weight of evidence, WoE = ln(%good / %bad), the within-bin log-odds of good to bad. The IPO version keeps the machinery unchanged and only relabels the event; the outcome is the direction of the first-day return, good is a positive (above-offer) close and bad is a negative one, and a prospectus characteristic, for example, the underwriter reputation tier, is binned in exactly the same way. Table II shows one illustrative characteristic end-to-end; the figures are hypothetical and chosen only to expose the arithmetic.

Table II.
Illustrative WoE / IV Transformation for One IPO Characteristic (hypothetical figures)

Underwriter tier	n	Pos (good)	Neg (bad)	WoE	IV contrib.
High reputation	80	66	14	+0,452	0,072
Medium reputation	70	52	18	−0,038	0,001
Low reputation	50	32	18	−0,523	0,077
Total / IV	200	150	50	n/a	0,150

Each bin receives a signed, monotonic WoE: positive, where the above-offer closes are over-represented; negative, where they are scarce; and the information value, IV = Σ (%good − %bad) ·WoE = 0.15, which places this characteristic in Siddiqi's medium-strength band. The logistic regression is fitted on the WoE-coded columns, and its coefficients are rescaled into integer points so that the high tier adds points and the low tier subtracts them by a fixed, published rule. Because binning and scaling are deterministic, re-fitting the same data reproduces the identical table and the identical points; that reproducibility, not mere legibility, is what makes the artefact auditable.

E. Research gaps and future directions

Five gaps follow. First, validation should respect time order, and studies should report where concept drift degrades performance [37]. Second, leakage needs a documented ex-ante cut-off, so that only information available before the predicted event enters the model. Third, imbalance should be handled with prevalence-robust metrics, resampling should stay inside the training folds, and every predictive claim should be judged against a majority-class baseline as a minimum reporting condition. Fourth, deterministic and auditable models such as credit-style scorecards deserve attention for retail investors, rather than leaderboard accuracy. Fifth, direction prediction, which is what a subscriber most needs, is under-studied relative to magnitude. So is Indonesian evidence, which national indexes such as SINTA hold beyond the Scopus and IEEE Xplore corpus searched here [38], [39], [40], [41], [53].

V. CONCLUSION

This review synthesised the IPO prediction literature around horizon, method family, and evaluation rigour. Ensembles, textual analysis, and deep learning have displaced linear baselines, but headline accuracy is necessary, not sufficient: across the 37 retained studies a validation design appears in 22 (only 6 time-ordered), five never evaluate out of sample at all, and an imbalance technique appears in 6, while the binding weakness is a missing comparator; none benchmarks against a majority-class rule. The call to action follows reporting against a majority-class baseline should be a minimum condition of publication for any classifier on an imbalanced IPO target so that reviewers and practitioners can see whether a model adds value over doing nothing. For retail-dominated markets such as the IDX, auditable scorecards, valued less for interpretability, which prior models already offer, than for the deterministic reproducibility that lets a subscriber or regulator regenerate the exact artefact, are a credible, largely unexplored direction.

REFERENCES

[1] K. F. Rock, "Why new issues are underpriced," Journal of Financial Economics, vol. 15, no. 1-2, pp. 187-212, 1986, doi: 10.1016/0304-405X(86)90054-1.
[2] J. R. Ritter, I. Welch, "A review of IPO activity, pricing, and allocations," Journal of Finance, vol. 57, no. 4, pp. 1795-1828, 2002, doi: 10.1111/1540-6261.00478.
[3] D. N. Rathnayake, P. A. Louembe, D. F. Kassi, G. Sun, D. Ning, "Are IPOs underpriced or overpriced? Evidence from an emerging market Research in International Business and Finance, vol. 50, pp. 171-190, 2019, doi: 10.1016/j.ribaf.2019.04.013.
[4] S. L. Jones, J. C. Yeoman, "Survey and synthesis of the IPO underpricing literature: the fixed-offer price constraint as a unifying core explanation," The Oxford Handbook of IPOs, pp. 146-174, 2018, doi: 10.1093/oxfordhb/9780190614577.013.11.
[5] M. Joshipura, S. Mathur, H. Gwalani, "Decrypting IPO pricing: an integrated bibliometric and content analysis approach," Managerial Finance, vol. 49, no. 1, pp. 135-162, 2023, doi: 10.1108/MF-06-2022-0293.
[6] P. Padhi, L. Kosuri, "IPO underpricing: a bibliographic review on its determinants," Encyclopedia of Monetary Policy, Financial Markets and Banking, pp. 182-189, 2025, doi: 10.1016/B978-0-44-313776-1.00074-X.
[7] D. Quintana, Y. Saez, P. Isasi, "Random forest prediction of IPO underpricing," Applied Sciences, vol. 7, no. 6, art. 636, 2017, doi: 10.3390/app7060636.
[8] B. Baba, G. Sevil, "Predicting IPO initial returns using random forest," Borsa Istanbul Review, vol. 20, no. 1, pp. 13-23, 2020, doi: 10.1016/j.bir.2019.08.001.

[9] A. Albada, E. E. Abusham, C. Z. Ong, K. Al Qatiti, "Machine learning insights: probing the variable importance of ex-ante information," Managerial Finance, vol. 51, no. 1, pp. 12-31, 2025, doi: 10.1108/MF-12-2023-0765.
[10] G. Colak, M. Fu, I. Hasan, "Predicting IPO first-day returns: evidence from machine learning analyses," Journal of Banking and Finance, vol. 178, art. 107500, 2025, doi: 10.1016/j.jbankfin.2025.107500.
[11] B. Zhao, H. Long, "Machine learning in initial public offering underpricing: managerial implications for China," Polish Journal of Management Studies, vol. 32, no. 2, pp. 253-265, 2025, doi: 10.17512/pjms.2025.32.2.16.
[12] M. Munshi et al., "Artificial intelligence and exploratory-data-analysis-based initial public offering gain prediction for public investors," Sustainability, vol. 14, no. 20, art. 13406, 2022, doi: 10.3390/su142013406.
[13] S. Narang, R. P. Pradhan, "Determining IPO underpricing: an ML-MCDM approach," Managerial Finance, 2025, doi: 10.1108/MF-10-2024-0741.
[14] D. Vaidynathan, P. Kayal, M. Maiti, "A feature engineering based model architecture for modeling initial public offerings," Applied Soft Computing, vol. 174, art. 113035, 2025, doi: 10.1016/j.asoc.2025.113035.
[15] D. Wang et al., "An interpretable neural fuzzy inference system for predictions of underpricing in initial public offerings," Neurocomputing, vol. 319, pp. 102-117, 2018, doi: 10.1016/j.neucom.2018.07.036.
[16] D. Quintana, F. Chavez, R. M. L. Baena, F. Luna, "Fuzzy techniques for IPO underpricing prediction," Journal of Intelligent and Fuzzy Systems, vol. 35, no. 1, pp. 367-381, 2018, doi: 10.3233/JIFS-169595.
[17] A. K. Singh, M. K. Jain, S. Jain, B. Gupta, "A new modus operandi for determining post-IPO pricing: analysis of Indian IPOs using artificial neural networks," Indian Journal of Finance, vol. 15, no. 1, pp. 8-22, 2021, doi: 10.17010/ijf/2021/v15i1/157011.
[18] D. Pirayesh Neghab, R. Bradrania, R. Elliott, "Deliberate premarket underpricing: New evidence on IPO pricing using machine learning," International Review of Economics and Finance, vol. 88, pp. 902-927, 2023, doi: 10.1016/j.iref.2023.07.008.
[19] M. F. Alahmadi, M. T. Yilmaz, "Prediction of IPO performance from prospectus using multinomial logistic regression, a machine learning model," Data Science in Finance and Economics, vol. 5, no. 1, pp. 105-135, 2025, doi: 10.3934/DSFE.2025006.
[20] P. Supsermpol, V. N. Huynh, S. Thajchayapong, N. Chiadamrong, "Predicting financial performance for listed companies in Thailand during the transition period: A class-based approach using logistic regression and random forest algorithm," Journal of Open Innovation: Technology, Market, and Complexity, vol. 9, no. 3, art. 100130, 2023, doi: 10.1016/j.joitmc.2023.100130.
[21] P. Supsermpol, V. N. Huynh, S. Thajchayapong, N. Suppakitjarak, N. Chiadamrong, "Predicting post-IPO financial performance: a hybrid approach using logistic regression and decision trees," Journal of Asian Business and Economic Studies, vol. 32, no. 1, pp. 52-65, 2025, doi: 10.1108/JABES-06-2024-0292.
[22] C. J. Williams, "Picking winners: identifying features of high-performing special purpose acquisition companies (SPACs) with machine learning," Journal of Risk and Financial Management, vol. 16, no. 4, art. 236, 2023, doi: 10.3390/jrfm16040236.
[23] J. Tao, A. V. Deokar, A. Deshmukh, "Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach," Journal of Business Analytics, vol. 1, no. 1, pp. 54-70, 2018, doi: 10.1080/2573234X.2018.1507604.
[24] T. H. Ly, K. Nguyen, "Do words matter: predicting IPO performance from prospectus sentiment," Proc. 14th IEEE Int. Conf. Semantic Computing (ICSC), pp. 307-310, 2020, doi: 10.1109/ICSC.2020.00061.
[25] S. Bian et al., "Can management tone predict IPO performance? Evidence from mandatory online roadshows in China Pacific-Basin Finance Journal, vol. 68, art. 101588, 2021, doi: 10.1016/j.pacfin.2021.101588.
[26] E. Fedorova, S. Druchok, P. Drogovoz, "Impact of news sentiment and topics on IPO underpricing: US evidence," International Journal of Accounting and Information Management, vol. 30, no. 1, pp. 73-94, 2022, doi: 10.1108/IJAIM-06-2021-0117.
[27] A. G. Katsafados et al., "Textual information and IPO underpricing: a machine learning approach," Journal of Financial Data Science, vol. 5, no. 2, pp. 100-135, 2023, doi: 10.3905/jfds.2023.1.121.
[28] X. Diao, M. Cai, "IPO breaking anomaly in China: a prospectus-based textual analysis," Accounting and Finance, vol. 65, no. 3, pp. 2701-2723, 2025, doi: 10.1111/acfi.70014.
[29] M. Wei, M. Zhang, L. Wei, M. Chen, "IPOhelper: mining features in registration statements for listing prediction of technological innovation companies," Emerging Markets Review, vol. 68, art. 101343, 2025, doi: 10.1016/j.ememar.2025.101343.
[30] D. F. Vamossy, "Social media emotions and IPO returns," Journal of Money, Credit and Banking, vol. 57, no. 1, pp. 31-67, 2025, doi: 10.1111/jmcb.13152.
[31] H. Xia, J. Weng, S. Boubaker, Z. Zhang, S. M. Jasimuddin, "Cross-influence of information and risk effects on the IPO market: exploring risk disclosure with a machine learning approach," Annals of Operations Research, vol. 334, pp. 761-797, 2024, doi: 10.1007/s10479-022-05012-8.
[32] P. Supsermpol, V. N. Huynh, T. Karnjanapatchara, N. Chiadamrong, "Post-IPO underpricing prediction: a case study of Thai capital market," Journal of Open Innovation: Technology, Market, and Complexity, vol. 11, no. 4, art. 100671, 2025, doi: 10.1016/j.joitmc.2025.100671.
[33] G. Colak, M. Fu, I. Hasan, "Why are some Chinese firms failing in the US capital markets? A machine learning approach," Pacific-Basin Finance Journal, vol. 61, art. 101331, 2020, doi: 10.1016/j.pacfin.2020.101331.
[34] G. Colak, M. Fu, I. Hasan, "On modeling IPO failure risk," Economic Modelling, vol. 109, art. 105790, 2022, doi: 10.1016/j.econmod.2022.105790.
[35] M. Alahmadi, "A risk-optimized framework for data-driven IPO underperformance prediction in complex financial systems," Systems, vol. 13, no. 3, art. 179, 2025, doi: 10.3390/systems13030179.
[36] M. Alahmadi, "A deep learning-based ensemble framework to predict IPOs performance for sustainable economic development," Sustainability, vol. 17, no. 3, art. 827, 2025, doi: 10.3390/su17030827.
[37] A. Reiff, "Comparing the prediction performance of random forest, lasso, and logit in the context of IPO withdrawal," Intelligent Systems in Accounting, Finance and Management, vol. 32, no. 3, art. e70009, 2025, doi: 10.1002/isaf.70009.
[38] A. Dhini, L. Sondakh, "Predicting stock return of initial public offering in Indonesia stock exchange," AIP Conference Proceedings, vol. 2710, art. 050011, 2024, doi: 10.1063/5.0145600.
[39] N. W. D. A. Dhanendra, K. Halim, A. Pramudita, Y. Arifin, "Predicting potential returns of Indonesian IPOs on their listing day using machine learning techniques," Proc. Int. Conf. Informatics and Computational Sciences (ICICoS), pp. 35-40, 2025, doi: 10.1109/ICICoS68590.2025.11330081.
[40] L. Agustina, A. Clara, "An investigation of IPOs initial performance in Indonesian market," Global Business Review, vol. 26, no. 1, pp. 101-119, 2025, doi: 10.1177/0972150920976640.
[41] I. P. S. Hendrawan, C. A. Utama, "Do executive facial trustworthiness have impact on IPO underpricing in the Indonesia stock exchange?," Review of Behavioral Finance, vol. 16, no. 6, pp. 1059-1086, 2024, doi: 10.1108/RBF-12-2023-0327.

[42] L. Setia-Atmaja, Y. Chandera, "Impact of family ownership, management, and generations on IPO underpricing and long-run performance," Investment Management and Financial Innovations, Vol. 18, no. 4, pp. 266-279, 2021, doi: 10.21511/IMFI.18(4).2021.23.
[43] T. S. Pavan Reddy, S. Ajitha, R. Krishnamoorthy, R. Singh Bhandari, A. Haider Khan, "Leveraging artificial intelligence to forecast IPO outcomes," Proc. Int. Conf. Computational Engineering, Sensing Technology and Management (ICCETM), pp. 1-5, 2025, doi: 10.1109/ICCETM66557.2025.11557744.
[44] P. Supsermpol, V. N. Huynh, N. Chiadamrong, "Post-IPO performance prediction: a comparative study of logistic regression and machine learning techniques for Thai IPO firms," Studies in Big Data, vol. 181, pp. 99-131, 2026, doi: 10.1007/978-3-032-06179-9_6.
[45] M. S. Al-Daweri et al., "Applying machine learning to interpret the impact of ex-ante information on IPO underpricing," Journal of Systems Science and Information, vol. 14, no. 1, pp. 1-24, 2026, doi: 10.12012/JSSI-2025-0012.
[46] R. Janagi, A. Udhayagiri, S. S. Varsith Pranav, S. Saranya, "IPO performance predictor: using pre-listing sentiment and fundamentals," Proc. 2nd Int. Conf. Visual Analytics and Data Visualization (ICVADV), pp. 207-211, 2026, doi: 10.1109/ICVADV67766.2026.11470550.
[47] A. K. Singh, M. Nandy, S. Lodh, S. Kalra, "Sector signals: machine learning insights into IPO pricing dynamics," International Journal of System Assurance Engineering and Management, 2026, doi: 10.1007/s13198-026-03155-6.
[48] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ, USA: John Wiley & Sons, 2006.
[49] D. J. Hand, W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 160, no. 3, pp. 523-541, 1997, doi: 10.1111/j.1467-985X.1997.00078.x.
[50] B. Talukdar, R. P. Pradhan, "Predicting IPO performance using machine learning: evidence from a data-analytics approach," in Proc. 6th Int. Conf. Data Analytics for Business and Industry (ICDABI), IEEE, 2025, doi: 10.1109/ICDABI67967.2025.11547586.
[51] Y. Tsukioka, J. Yanagi, T. Takada, "Investor sentiment extracted from internet stock message boards and IPO puzzles," International Review of Economics and Finance, vol. 56, pp. 205-217, 2018, doi: 10.1016/j.iref.2017.10.025.
[52] S. Zhang, Y. Li, Y. He, R. Liang, "Do vocal cues matter in information disclosure? Evidence from IPO online roadshows in the SSE STAR market," International Review of Financial Analysis, vol. 93, art. 103229, 2024, doi: 10.1016/j.irfa.2024.103229.
[53] D. Hidayatullah, I. Jatnika, "Development of a classification model for underpriced issuers using machine learning algorithms," Journal of Applied Informatics and Computing, vol. 9, no. 4, pp. 1648-1654, 2025, doi: 10.30871/jaic.v9i4.9769.