

Peer-Review

Seung, Eunsu. 2026. "Evaluation of the Altman Z-Score Framework for Financial Distress Prediction in Korean Firms." *Journal of High School Science* 10 (3): 244–64.
<https://doi.org/10.64336/001c.165346>.

This is post hoc coefficient tuning to fit known outcomes. The model is validated on the same examples used to design it. This is circular validation.

In this manuscript, coefficients are chosen a priori based on narrative reasoning e.g., "Korean firms use more leverage → increase weight X

Adjusted until outputs look reasonable

Strong firms move into "safe zone"

Bankrupt firms move into "distress zone"

the model is then evaluated on the same cases used to motivate the change. But with such a small sample, many different coefficient combinations would achieve the same result. So the model is Underdetermined and non-unique

Further problems:

(a) No objective function. There is no formal criterion like:

maximize likelihood

minimize classification error

optimize AUC

(b) No statistical estimation

No fitting procedure (MLE, regression, etc.)

(c) No out-of-sample validation

Same data → design + evaluation

This is outcome-driven calibration (i.e., fitting to known answers).

1. Coefficients are manually adjusted:

Justification is qualitative (narrative), not statistically estimated. Validation is done using known historical cases. This is not model estimation—it is model reparameterization guided by desired outcomes.

2. Coefficients are chosen to:

Raise scores of "strong" Korean firms (Hyundai, POSCO)

Lower scores of bankrupt firms (STX, SsangYong)

This is effectively inverse fitting to known labels without formal optimization. The model is not identifiable as a statistical estimator—only as a heuristic transformation.

3. The strongest evidence presented is:

Distressed firms have lower K-Z scores

Healthy firms move into "safe zone" after adjustment

But:

These same firms were used to justify coefficient changes

No train/test split

No temporal holdout

No external dataset

This creates complete circularity.

4. The Korean analysis relies on ~6 companies (Hyundai, Kia, POSCO, STX, DSME, SsangYong)

This is critically inadequate:

No statistical power

No coverage of firm heterogeneity

Severe selection bias (handpicked "illustrative" firms)

This turns the study into a case study, not a predictive modeling paper.

5. The logistic regression parameters are derived by:

Forcing:

$P=50\%$ at $Z \approx 1.1$

$P \approx 92\%$ at $Z \approx 0$

Then solving for β_0 and β_1

Crucially, the paper states “The 92% value... is intentionally set by the researcher”. This is not regression. This is curve construction to match desired thresholds.

Reported probabilities (e.g., 94.7%) are not empirically derived. They are artifacts of imposed calibration constraints. So the model is a deterministic mapping designed to produce intuitive outputs—not a statistical estimator.

6. A predictive model paper should report:

AUC / ROC

Precision-recall

Confusion matrix

Calibration curves

Out-of-sample accuracy

None are provided.

Instead, evaluation is purely visual and anecdotal (figures showing trajectories over time).

7. The manuscript repeatedly claims:

“high discriminatory power”

“accurate assessment”

“superior detection”

These claims are unsupported because:

No benchmark comparison (e.g., original Z-score ROC)

No statistical testing

No uncertainty quantification

What should be done:

1. A scientifically valid approach also changes coefficients—but in a constrained, falsifiable framework: Step 1: Define a model. Same structure:

$Z = \beta_1 X(1) + \dots + \beta_5 X(5)$

2. Estimate coefficients from data

Using (1000+) or a large # of Korean firms:

Logistic regression or discriminant analysis

Coefficients obtained by:

maximizing likelihood

minimizing prediction error

You don't choose coefficients—you infer them from data

3. Separate training vs testing

Train on one dataset (1000+)

Evaluate on unseen firms or time periods (100+ that are NOT a part of the training dataset), OR (Training dataset from 2000-2020) and Testing dataset from (2021 - 2026). The same coefficients should still work on a different Korean dataset that the model is not trained on.

Prevents fitting noise or known outcomes and means that your model is generalizable across all Korean firms.

4. Quantify performance

ROC / AUC

Calibration

Confidence intervals

Now you can say whether the new coefficients actually improve prediction

5. Compare:

Original Altman Z against

Re-estimated Korean model, using Korean firms.

Otherwise “improvement” is unsubstantiated.

Summary:

The current manuscript relies on manually adjusted coefficients and post hoc calibration to known cases, which does not constitute a statistically valid modeling framework. To be suitable for publication as a predictive study, the model must be re-estimated using a sufficiently large dataset and evaluated using out-of-sample performance metrics.

Elevate novelty: Estimate the model coefficients separately for U.S. and Korean datasets using a consistent statistical method, then compare them and test whether cross-country differences can be explained by macroeconomic or structural variables. This would allow you to move from heuristic adjustment to empirically grounded cross-country calibration.

I cannot review this resubmitted manuscript unless - as communicated in the earlier review - the author uploads a word document listing my earlier review comments verbatim and describing where and how they have addressed each comment in this resubmitted manuscript.

Comment 1

“A scientifically valid approach also changes coefficients—but in a constrained, falsifiable framework. Step 1: Define a model. Same structure:
$$Z = \beta_1 X_1 + \dots + \beta_5 X_5$$
”

Response

Thank you for this important comment.

The revised manuscript substantially restructures the study from a manually adjusted coefficient framework into a formally estimated statistical prediction framework.

In the revised version, Section 2.1 (“Logistic Regression Re-estimation Based on Korean Firms”) (pp. 6–7) now introduces a logistic regression model that predicts financial distress directly from the five original Altman financial variables $X_1 = X_1$. Rather than manually modifying coefficients based on qualitative interpretation, the revised study statistically estimates the relationships between the five financial ratios and financial distress conditions using large-scale Korean firm-year observations.

In addition, Section 3.3 (“Logistic Regression Framework”) (p. 11) now formally defines the logistic regression structure and explains the statistical estimation procedure used in the study.

These revisions were implemented to replace the earlier heuristic coefficient-adjustment approach with a constrained and empirically estimated predictive framework.

Comment 2

“Estimate coefficients from data using (1000+) or a large number of Korean firms:

Logistic regression or discriminant analysis

Coefficients obtained by:

maximizing likelihood

minimizing prediction error

You don’t choose coefficients—you infer them from data.”

Response

Thank you for this valuable suggestion.

The revised manuscript now replaces the earlier manually adjusted coefficient framework with a large-scale statistical estimation procedure based on Korean firm-year observations.

First, Section 1.4 (“Empirical Evaluation of the Altman Z-score in Korean Firms”) (pp. 4–6) now introduces a Korean financial dataset consisting of approximately 12,651 Korean firm-year observations collected from the DART disclosure system between 2011 and 2024.

Using this large-scale dataset, the revised study statistically estimates the relationships between the five Altman financial variables $X_1 = X_5$ and financial distress conditions through logistic regression analysis.

Section 2.1 (“Logistic Regression Re-estimation Based on Korean Firms”) (pp. 6–7) now reports the estimated coefficients obtained directly from Korean firm-year observations rather than from manual adjustment.

In addition, Section 3.3 (“Logistic Regression Framework”) (p. 11) now formally explains the logistic regression estimation framework and maximum likelihood estimation procedure used in the study.

These revisions were implemented to replace the earlier heuristic coefficient-adjustment approach based on only seven illustrative Korean firms with a large-scale statistical estimation framework using more than 12,000 Korean firm-year observations. The revised study now estimates the logistic regression coefficients directly from data through maximum likelihood estimation rather than manually adjusting coefficients based on qualitative interpretation.

Comment 3

“Separate training vs testing.

Train on one dataset (1000+)

Evaluate on unseen firms or time periods (100+ that are NOT a part of the training dataset), OR (Training dataset from 2000–2020) and Testing dataset from (2021–2026).

Prevents fitting noise or known outcomes and means that your model is generalizable across all Korean firms.”

Response

The revised manuscript now introduces a strict temporal train/test separation framework in order to reduce circularity and evaluate genuine out-of-sample predictive performance.

Specifically:

- Korean firm-year observations from 2011–2020 were used for model training and coefficient estimation.
- Observations from 2021–2024 were reserved exclusively as an out-of-sample test dataset.

The revised manuscript explains that this temporal separation was adopted to evaluate whether the estimated relationships remain effective beyond the training period and to reduce overfitting and outcome-driven calibration concerns.

These revisions were incorporated into:

- Section 2.1 “Logistic Regression Re-estimation Based on Korean Firms” (pp. 6–7)
- Section 3.4 “Train/Test Split and Out-of-Sample Validation” (p. 11)

The earlier manuscript did not include a formal train/test validation procedure. The revised manuscript now introduces a chronological out-of-sample validation framework using 2011–2020 observations for training and 2021–2024 observations for testing.

Comment 4

The manuscript lacks standard predictive model evaluation measures such as ROC/AUC, confusion matrices, precision-recall analysis, and out-of-sample performance evaluation.

Response

This concern has been addressed by introducing a formal out-of-sample performance evaluation framework and additional classification metrics.

Specifically:

- Section 2.2 (pages 6–7) now provides a direct out-of-sample comparison between the original Altman Z-score model and the statistically re-estimated Korean logistic framework using the same 2021–2024 Korean test dataset. The revised manuscript now reports Accuracy, Precision, Recall, F1-score, confusion matrix results, ROC curves, and AUC values for both models.
- Section 3.5 (pages 11–12) additionally explains the statistical evaluation metrics and the rationale for using ROC/AUC and confusion matrix analysis in financial distress prediction.

The revised manuscript now includes quantitative predictive performance evaluation measures in addition to the previously presented qualitative and visual analyses.

Comment 5

“Compare:

Original Altman Z against

Re-estimated Korean model, using Korean firms.

Otherwise ‘improvement’ is unsubstantiated.”

Response

The revised manuscript now directly compares the predictive performance of:

1. the original Altman Z-score model, and
 2. the statistically re-estimated Korean logistic regression framework,
- using the same Korean out-of-sample test dataset.

The revised manuscript reports side-by-side comparisons of:

- Accuracy
- Precision
- Recall
- F1-score
- AUC

for both models.

In addition, ROC curve comparisons between the two frameworks are now included in Figure 5.

These revisions were incorporated into:

- Section 2.2 “Out-of-Sample Performance Evaluation” (pages 6–7)

The revised manuscript now provides direct comparative evaluation results between the original Altman model and the statistically re-estimated Korean logistic framework across multiple classification metrics.

The manuscript reports two substantially different counts of financially distressed observations without adequately explaining the distinction between them. In Section 1.4, the Korean dataset is described as containing only 117 distress-related observations, whereas Section 3.2 later states that approximately 1,755 observations were classified as financially distressed. It is unclear whether these figures refer to different subsets, different definitions of distress, or different stages of the analysis. Because the validity of the logistic regression framework depends critically on the definition and prevalence of the outcome variable, the authors should explicitly reconcile these numbers and clearly explain how each distress dataset was constructed and used throughout the study.

1. Although the revised manuscript appropriately avoids defining distress directly from the Altman Z-score itself, there remains a concern regarding overlap between the predictor variables and the criteria used to define financial distress. The logistic regression model uses financial ratios derived from working capital, profitability, leverage, and liquidity, while the distress label is constructed using conditions that include negative profitability measures, leverage thresholds, and liquidity-

related indicators. As a result, the model may partially be learning the same accounting characteristics used to generate the target labels. The authors should discuss this potential source of circularity and explain how much of the reported predictive performance may reflect prediction of an internally defined accounting-based distress condition rather than an independent financial outcome.

2. The manuscript is framed throughout as a bankruptcy prediction study, and the title explicitly references bankruptcy prediction. However, the outcome variable used for model development and evaluation is not actual bankruptcy, insolvency, restructuring, or legal failure events. Instead, the study predicts an accounting-based financial distress label constructed from financial statement indicators. While financial distress is an important and meaningful outcome, it is conceptually distinct from bankruptcy. The authors should clarify this distinction throughout the manuscript and either provide stronger justification for treating the distress label as a proxy for bankruptcy or revise the framing to reflect that the model predicts financial distress rather than actual bankruptcy events.

3. In addition, I have concerns regarding the statistical rigor of the reported performance improvements. The manuscript presents substantial gains in accuracy, recall, F1-score, and AUC for the proposed Korean logistic framework relative to the original Altman model. However, no confidence intervals, statistical significance tests, bootstrap analyses, or other measures of uncertainty are provided. Consequently, readers cannot determine whether the observed differences are statistically robust or potentially attributable to sampling variability. The authors should consider reporting confidence intervals for key performance metrics and, where appropriate, conducting formal statistical comparisons (e.g., DeLong testing for AUC differences) to support claims of superior predictive performance.

4. Finally, the reported predictive performance appears exceptionally strong and warrants additional scrutiny. The proposed model achieves an AUC of approximately 0.947 and identifies over 95% of distressed observations in the test dataset. Such performance is unusually high for real-world bankruptcy prediction problems, particularly when evaluated on future observations. Because the outcome variable is an internally constructed accounting-based distress label rather than actual bankruptcy events, the manuscript should explicitly discuss whether the reported performance may overstate the model's ability to predict genuine corporate failure. Additional validation using external datasets, alternative distress definitions, or actual bankruptcy outcomes would strengthen confidence in the model's practical applicability.

General Response to the Reviewers

We would like to express our deepest and most sincere gratitude to the reviewers for their invaluable guidance and academic generosity throughout this revision process.

As a developing researcher, the comprehensive and rigorous guidelines provided in the previous round of reviews were far more than a mere list of correction requirements—they served as a profound academic mentorship. Navigating and resolving each of the challenging tasks set forth by the reviewers became a deeply rewarding intellectual journey, one that far exceeded the boundaries of traditional classroom education. Thanks to your dedicated insights, we have truly learned the meaning of empirical rigor and data-driven inference.

Looking back, we feel a deep sense of humility regarding the foundational gaps in our initial draft, and we consider ourselves incredibly fortunate that the reviewers chose to invest such immense time and effort to guide us like a patient mentor rather than simply rejecting the manuscript.

In this second revision, we have strived to fully reflect your high standards of scientific integrity. We hope these thorough updates demonstrate our profound respect for your feedback and our earnest commitment to contributing a meaningful, high-quality study to the field. Thank you once again for your invaluable time, patience, and academic kindness.

Comment 1

The manuscript reports two substantially different counts of financially distressed observations without adequately explaining the distinction between them. In Section 1.4, the Korean dataset is described as containing only 117 distress-related observations, whereas Section 3.2 later states that approximately 1,755 observations were classified as financially distressed. It is unclear whether these figures refer to different subsets, different definitions of distress, or different stages of the analysis. Because the validity of the logistic regression framework depends critically on the definition and prevalence of the outcome variable, the authors should explicitly reconcile these numbers and clearly explain how each distress dataset was constructed and used throughout the study.

Response

We sincerely appreciate the reviewer's careful reading and for pointing out the inconsistency in the number of financially distressed observations between Section 1.4 and Section 3.2.

We would like to clarify that the figure "117" in Section 1.4 was a typographical error caused by an inadvertent copy-paste from Section 1.3 (which describes the U.S. dataset). The correct number of financially distressed observations in the Korean dataset is **1,755**, as correctly stated in Section 3.2. The Korean dataset consists of 12,651 firm-year observations, and based on our accounting-based distress criteria, 1,755 of these were classified as financially distressed. We have corrected the manuscript to reflect this consistent figure in Section 1.4.

Action Taken: We have revised the table and the accompanying text in Section 1.4 to ensure full consistency with the results presented in Section 3.2. The updated table now correctly displays the statistics for the Korean dataset as follows:

Category	Value
Firm-year observations	12,651
Observation period	2011–2024
Data source	DART (dart.or.kr)
Variables constructed	$X_1 \sim X_5$, Altman Z-score
Distress-related observations	1,755

Comment 2

Although the revised manuscript appropriately avoids defining distress directly from the Altman Z-score itself, there remains a concern regarding overlap between the predictor variables and the criteria used to define financial distress. The logistic regression model uses financial ratios derived from working capital, profitability, leverage, and liquidity, while the distress label is constructed using conditions that include negative profitability measures, leverage thresholds, and liquidity-related indicators. As a result, the model may partially be learning the same accounting characteristics used to generate the target labels. The authors should discuss this potential source of circularity and explain how much of the reported predictive performance may reflect prediction of an internally defined accounting-based distress condition rather than an independent financial outcome.

Response

We sincerely thank the reviewer for this insightful comment. We completely agree that there is an inherent potential for partial circularity and overlap between the explanatory accounting ratios (such as working capital, profitability, leverage, and liquidity) and the internally constructed financial distress label.

As the reviewer rightly pointed out, because these variables share common underlying accounting characteristics, the model's reported predictive performance may partially reflect its proficiency in

identifying these pre-embedded accounting conditions rather than predicting an entirely independent financial outcome.

We agree that this methodological boundary must be explicitly addressed to ensure scientific rigor. Accordingly, we have added a dedicated discussion acknowledging this potential source of circularity as the third limitation in Section 4 (Discussion). We also proposed incorporating external, non-accounting-based outcomes (e.g., formal bankruptcy filings, credit rating downgrades) for future research to mitigate this overlap.

The following text has been added to the **Discussion (Page 13)**:

"Third, there is an inherent potential for overlap between the predictor variables and the criteria used to define the financial distress label. The proposed logistic framework utilizes financial ratios that share common underlying accounting characteristics with the target distress definition. Although the distress definition was intentionally separated from the Altman Z-score itself to mitigate direct circularity, a portion of the exceptionally strong predictive performance—such as the AUC of 0.947—may reflect the model's proficiency in identifying internally defined accounting characteristics rather than predicting an entirely independent financial outcome. Future research could address this limitation by incorporating externally observed distress outcomes that are not directly derived from the same financial ratios, such as formal bankruptcy filings, debt restructuring events, credit rating downgrades, or involuntary delisting decisions."

Comment 3

The manuscript is framed throughout as a bankruptcy prediction study, and the title explicitly references bankruptcy prediction. However, the outcome variable used for model development and evaluation is not actual bankruptcy, insolvency, restructuring, or legal failure events. Instead, the study predicts an accounting-based financial distress label constructed from financial statement indicators. While financial distress is an important and meaningful outcome, it is conceptually distinct from bankruptcy. The authors should clarify this distinction throughout the manuscript and either provide stronger justification for treating the distress label as a proxy for bankruptcy or revise the framing to reflect that the model predicts financial distress rather than actual bankruptcy events.

Response

We accept the reviewer's excellent and constructive criticism. We completely agree that "financial distress" based on accounting indicators is conceptually distinct from legal "bankruptcy" or insolvency events. Blurring this distinction could mislead readers regarding the model's true target outcome.

Following the reviewer's recommendation, we have comprehensively revised the framing of the manuscript to clarify that our framework predicts **financial distress** rather than actual legal bankruptcy. Most notably, we have changed the title of the paper and revised the Abstract to precisely reflect this scope.

1. Title Revision:

- **Old Title:** *Evaluation of The Formal Bankruptcy Prediction Altmann Z-Score Model: Using the New K Score and Probability Model*
- **New Title:** *Evaluation of the Altman Z-Score Framework for Financial Distress Prediction in Korean Firms*

2. Abstract Revision: We have replaced the term "bankruptcy prediction" with "financial distress and credit risk" where applicable, and clarified that the logistic regression model derives "financial distress probabilities" to assess risk.

Comment 4

In addition, I have concerns regarding the statistical rigor of the reported performance improvements. The manuscript presents substantial gains in accuracy, recall, F1-score, and AUC for

the proposed Korean logistic framework relative to the original Altman model. However, no confidence intervals, statistical significance tests, bootstrap analyses, or other measures of uncertainty are provided. Consequently, readers cannot determine whether the observed differences are statistically robust or potentially attributable to sampling variability. The authors should consider reporting confidence intervals for key performance metrics and, where appropriate, conducting formal statistical comparisons (e.g., DeLong testing for AUC differences) to support claims of superior predictive performance.

Response

We are highly grateful to the reviewer for emphasizing the need for statistical rigor and measures of uncertainty. We acknowledge that reporting point estimates alone (such as AUC = 0.947) was insufficient to rule out the potential effects of sampling variability.

To address this crucial point, we performed a formal **bootstrap resampling analysis with 1,000 replications** to compute the 95% confidence intervals (CI) for the AUC values of both models.

The empirical results from the bootstrap analysis are as follows:

- **Proposed Korean Logistic Framework AUC:** 0.947 (95% CI: 0.935 – 0.960)
- **Original Altman Model AUC:** 0.859 (95% CI: 0.841 – 0.877)

The non-overlapping confidence intervals suggest that the observed difference in AUC is unlikely to be explained solely by sampling variability.

We have updated the manuscript in two places to reflect this rigorous validation:

1. **Section 3.5 (Performance Evaluation Metrics):** Added a description of the bootstrap resampling methodology (1,000 replications) used to compute the 95% confidence intervals.
2. **Section 2.2 (Out-of-Sample Performance Evaluation):** Reported the precise 95% CIs for both AUC metrics and added a discussion on the statistical significance of the non-overlapping intervals.

Comment 5

Finally, the reported predictive performance appears exceptionally strong and warrants additional scrutiny. The proposed model achieves an AUC of approximately 0.947 and identifies over 95% of distressed observations in the test dataset. Such performance is unusually high for real-world bankruptcy prediction problems, particularly when evaluated on future observations. Because the outcome variable is an internally constructed accounting-based distress label rather than actual bankruptcy events, the manuscript should explicitly discuss whether the reported performance may overstate the model's ability to predict genuine corporate failure. Additional validation using external datasets, alternative distress definitions, or actual bankruptcy outcomes would strengthen confidence in the model's practical applicability.

Response

We appreciate the reviewer's critical and valuable evaluation regarding the unusually high performance of our model (AUC of approximately 0.947 and over 95% recall). We agree that such high metrics warrant additional scrutiny and could overstate the model's practical applicability in real-world bankruptcy prediction.

As the reviewer noted, this exceptionally strong performance stems from the fact that our outcome variable represents an internally constructed, accounting-based financial distress label rather than actual corporate failure or legal bankruptcy events. While financial distress is conceptually associated with elevated default risk, they are not identical, and some distressed firms may recover.

To address the reviewer's concern and prevent any overstatement of our findings, we have added a thorough discussion as the fourth limitation in Section 4 (Discussion). We explicitly clarified that our metrics must be strictly interpreted as measures of financial distress identification rather than direct bankruptcy prediction, and emphasized the necessity of future validation using actual legal bankruptcy filings or external datasets.

The following text has been added to the **Discussion (Page 13)**:

"Fourth, because the outcome variable used throughout this study represents accounting-based financial distress rather than actual corporate failure, the reported classification metrics should be strictly interpreted as measures of financial distress identification rather than direct bankruptcy prediction. While financial distress is conceptually associated with elevated default risk, this relationship is not one-to-one; some distressed firms successfully recover without entering legal insolvency, whereas others fail due to qualitative or macroeconomic factors not fully captured by accounting indicators alone. Therefore, to evaluate the practical applicability and generalizability of the proposed framework, future validation against actual corporate bankruptcy filings or legal restructuring outcomes is warranted."

One of my biggest concerns remains unanswered. I would like to see you validate both models against an external failure dataset consisting of Korean firms that were subsequently delisted by KRX or entered formal bankruptcy/restructuring proceedings.

option 1: The Korea Exchange publicly publishes delisting decisions and lists of companies removed from KOSPI/KOSDAQ. Numerous studies use delisting as a proxy for corporate failure because the event is externally determined and independent of the accounting ratios being modeled. Your model and Altman's model should be validated against firms subsequently delisted from KOSDAQ or KOSPI. You already use DART and market data. Linking DART financial statements to KRX delisting outcomes is entirely feasible.

option 2: The Korean government's open data portal provides datasets and APIs on bankrupt financial companies, including bankruptcy registration dates and related information. These are not manufacturing firms, so they are not a perfect replacement, but they demonstrate that externally observed bankruptcy outcomes are publicly available.

option 3: The Korean bankruptcy literature includes datasets such as KRBDS (Korean Bankruptcy and Distress datasets), which have been used repeatedly in machine-learning bankruptcy prediction studies. Although some versions may require institutional access, they are established benchmark datasets rather than newly invented distress labels. It may be possible to request access for academic purposes.

This is an important experiment, because; without it; the current evidence does not establish superior prediction of real-world corporate failure.

Comment 1

One of my biggest concerns remains unanswered. I would like to see you validate both models against an external failure dataset consisting of Korean firms that were subsequently delisted by KRX or entered formal bankruptcy/restructuring proceedings.

option 1: The Korea Exchange publicly publishes delisting decisions and lists of companies removed from KOSPI/KOSDAQ. Numerous studies use delisting as a proxy for corporate failure because the event is externally determined and independent of the accounting ratios being modeled. Your model and Altman's model should be validated against firms subsequently delisted from KOSDAQ or KOSPI. You already use DART and market data. Linking DART financial statements to KRX delisting outcomes is entirely feasible.

option 2: The Korean government's open data portal provides datasets and APIs on bankrupt financial companies, including bankruptcy registration dates and related information. These are not

manufacturing firms, so they are not a perfect replacement, but they demonstrate that externally observed bankruptcy outcomes are publicly available.

option 3: The Korean bankruptcy literature includes datasets such as KRBDS (Korean Bankruptcy and Distress datasets), which have been used repeatedly in machine-learning bankruptcy prediction studies. Although some versions may require institutional access, they are established benchmark datasets rather than newly invented distress labels. It may be possible to request access for academic purposes.

This is an important experiment, because; without it; the current evidence does not establish superior prediction of real-world corporate failure.

Response

We thank the reviewer for raising this concern again and for outlining concrete paths forward. We agree that, without validation against an externally determined failure outcome, the evidence in the previous version of the manuscript did not establish that the re-estimated logistic framework's advantage over the Altman Z-score extends to real-world corporate failure, as opposed to reproducing an accounting-based distress label constructed from the same underlying financial ratios.

We adopted Option 1 — validating both models against Korean firms subsequently delisted from KOSPI or KOSDAQ — as this approach was directly compatible with our existing data infrastructure and, as the reviewer noted, straightforward to implement by linking DART financial statements to publicly available KRX delisting records.

Accordingly, we conducted the following external validation exercise using KRX delisting as the failure event:

1. New Section 2.4 — "External Validation on KRX Delisted Firms" (Results)

We identified 94 Korean manufacturing firms delisted from KOSPI or KOSDAQ between 2016 and 2024 using publicly available KRX disclosure records, and linked each firm to its pre-delisting financial statements obtained from DART, following the same extraction procedure used elsewhere in this study. We confirmed that none of these 94 firms overlapped with the 12,651 firms used to construct the original dataset, so that the validation sample is independent at the firm level. Because DART data are consistently available from 2015 onward, we collected financial data for each of the five fiscal years preceding delisting ($t-1$ through $t-5$), yielding approximately 400 firm-year observations for the delisted (failure) class.

As a non-distress comparison group, we used the held-out 2021–2024 test split from the original 12,651-firm sample ($n = 3,796$), from which 527 observations already classified as financially distressed under our accounting-based criteria were excluded, yielding a final comparison pool of 3,269 firm-year observations. This comparison pool was drawn exclusively from data not used to estimate the logistic regression coefficients, ensuring that both the failure and non-failure observations in this validation were genuinely out-of-sample with respect to model estimation.

At every pre-delisting horizon, the re-estimated logistic model achieved a higher AUC than the Altman Z-score ($t-1$: 0.854 vs. 0.781; $t-2$: 0.877 vs. 0.806; $t-3$: 0.864 vs. 0.783; $t-4$: 0.819 vs. 0.744; $t-5$: 0.816 vs. 0.743; pooled: 0.849 vs. 0.775), with 95% bootstrap confidence intervals (1,000 replications) for the AUC differences excluding zero at every horizon (pooled difference: -0.074 , 95% CI: -0.093 to -0.056). These results are presented in new Figure 8 and accompanying tables, together with confusion matrices and classification metrics (accuracy, precision, recall, F1)

computed at conventional, pre-specified thresholds ($Z < 1.81$ for Altman; predicted probability of 0.5 for the logistic model).

2. New Section 3.6 — "External Validation Using Delisted Firms" (Methodology)

We added a corresponding methodology section describing how delisted firms were identified from KRX records, how pre-delisting financial data were linked from DART, the rationale for examining five pre-delisting horizons, the construction of the non-distress comparison pool, and the bootstrap resampling procedure used to estimate confidence intervals.

3. Revised Discussion

We revised the limitation paragraphs that had previously identified the absence of external, independently determined failure outcomes as an open concern, to reflect that this concern has now been directly addressed through the KRX-based external validation. We also added a paragraph discussing why the AUC values obtained in this external validation (0.849 / 0.775) are lower than the corresponding in-sample test-set values (0.947 / 0.859), noting that this decline is expected given that delisting is a more demanding, independently determined outcome, and that the persistence of the logistic model's advantage under this stricter test strengthens rather than weakens our conclusions.

4. Revised Conclusion

We added a paragraph summarizing the external validation findings and their implication that the re-estimated framework's advantage generalizes beyond the accounting-based distress label used for estimation.

We believe this additional analysis directly responds to the reviewer's concern and provides evidence that the observed superiority of the re-estimated logistic framework extends to an independently determined, real-world failure outcome. We appreciate the reviewer's suggestion, which substantially improved the manuscript by motivating this external validation analysis. We acknowledge several limitations of this external validation, discussed explicitly in the manuscript (Section 2.4): the number of unique delisted firms (94) is modest, particularly at longer horizons; the non-distress comparison pool was reused across all five horizons, so the horizon-specific results are not fully independent and the pooled estimate should be regarded as the primary basis for inference; the cohort of firms represented at each horizon changes due to data availability constraints; bootstrap resampling was performed at the firm-year rather than firm level, which may understate true sampling variability; and the non-distress comparison pool's observation period (2021–2024) does not fully align with the fiscal years underlying the longer pre-delisting horizons ($t-4$, $t-5$).

Thank you for addressing my comments. Accepted.