

Validating Selective Prediction Under Source Uncertainty in Small Enterprise Surveys

Suyono Thamrin^{1*}, M. Calvin Capnary², Iskandar Agung³

Faculty of Defense Management, Defence University, Bogor, Indonesia¹

Faculty of Management and Business, BINUS University, Jakarta, Indonesia²

Research Center for Education, National Research and Innovation Agency, Jakarta, Indonesia³

Abstract—Survey-based prediction is sensitive to unresolved source versions, repeated predictor profiles, and abstention rules. We evaluate these choices jointly in two Indonesian culinary-enterprise files, containing 206 aligned positions and 301 recorded rows, respectively. Each file has 177 predictor profiles, without verified respondent identities. Five fixed baselines are compared over 50 overlapping partitions, supplemented by training-only nested tuning, hierarchical conformal calibration, profile-stratified diagnostics, and explicit review-cost scenarios. In the second file, random-forest mean absolute error increases from 0.1966 under row splitting to 0.2814 under profile-disjoint splitting; ridge achieves 0.2369. A limited tuning sensitivity preserves this ordering. In the first file, simultaneous coverage of recorded profile outcomes and versions is 90.00 percent for profile-version calibration and 87.28 percent for row-midpoint calibration. Approximate dependence-adjusted uncertainty intervals include zero for the paired coverage gain, so superiority is not established. Aggregate coverage also conceals poor performance in some response and repetition strata. Exploratory abstention analyses reduce retained error but can lower coverage or require infinite intervals. Additional ordinal simulations expose failure under noise shift and unrecorded versions, while apparent coverage under misalignment does not authenticate the source. The contribution is an auditable evaluation workflow using established methods, with explicit coverage events, implementation choices, and decision-cost assumptions. Its transferable value is diagnostic; learner rankings, innovation effects, population validity, and operational benefits cannot be generalized from these files.

Keywords—*Selective prediction; conformal prediction; data provenance; grouped validation; survey data; small enterprises*

I. INTRODUCTION

Predictive systems for small enterprises often begin with survey records describing innovation, capabilities, and perceived business outcomes. These records offer accessible inputs when longitudinal financial data are unavailable. Their computational use is already visible in applied computer science. For example, Ducut evaluated decision trees, random forests, and logistic regression for business readiness using survey-based technology indicators [1]. The practical attraction is clear: a model can summarize a lengthy questionnaire and identify assessments that require further examination. However, this use also changes the evidential burden. An association estimated within a questionnaire does not establish prediction for an unseen enterprise, and a favorable held-out score does not establish that the underlying records are independent or that the response variable has a stable

definition. The target quantity and the population to which it applies must be stated before predictive performance can be interpreted [2].

Source uncertainty can enter this process in several forms. Two worksheets may contain different values for apparently corresponding records. A tabulation may repeat a long sequence of responses without identifying whether the repetitions represent separate respondents, copied entries, or an undocumented processing step. Aggregate demographic charts may exist even though corresponding record-level fields are absent. These conditions are consequential because preprocessing cannot resolve questions of provenance through model accuracy. A model trained on one worksheet can be evaluated against a different numerical target simply by changing the extraction rule. A random train-test split can place an identical predictor profile in both partitions. Work on leakage and reproducibility has established the need to examine these failures throughout the analysis pipeline [3, 4]. Dataset documentation and studies of downstream data problems further show why the source must be treated as part of the computational system [5, 6].

Uncertainty estimation provides only a partial response. Split conformal prediction calibrates prediction sets using held-out errors under an appropriate exchangeability assumption [7]. It does not determine which of two source versions is authoritative. Nor does ordinary row-level calibration automatically provide simultaneous coverage for all records belonging to a repeated profile. A further complication arises when a system withholds predictions having wide intervals. Coverage averaged over all test cases need not survive this selection. Recent work explicitly addresses ambiguous labels, hierarchical observations, and selection-conditional coverage [8, 9, 10]. An applied survey pipeline may encounter these conditions together, yet the decisions governing source extraction, validation units, and selective reporting are often assessed separately. Their interaction is especially relevant when a modest dataset is reused to justify decision support.

This study examines these conditions in two public survey deposits concerning Indonesian culinary enterprises [11, 12]. RQ1 asks how recorded-profile separation and model settings affect prediction and feature ablation. RQ2 asks how the calibration unit and recorded source versions affect simultaneous coverage, width, and the uncertainty of paired differences. RQ3 asks how selection changes the covered population, retained error, and calibration sample size. Reviewer-requested analyses further examine repetition and

*Corresponding author.

response strata, predictor redundancy, and hypothetical review costs. The contribution is a reproducible diagnostic workflow built from established methods. Its outputs inform the design and reporting of evaluations when similar file structures occur; the numerical model rankings are specific to these files.

The three choices interact through the score distribution and the population being evaluated. Profile separation first determines which recorded patterns are unseen at testing. Taking a maximum over members and versions then changes the calibration score and the event to be covered. A gate finally selects a subset of those profiles, changing both their residual distribution and the number available for calibration. Thus, a source-version interval can be appropriate before selection yet miscalibrated on retained profiles, and a low row-weighted error can coexist with poor coverage for frequently repeated profiles. Sections III–IV evaluate these links without treating them as separable causal effects.

II. RELATED WORK

A. Predictive Validity and Data Provenance

Leakage is broader than accidentally including a response column among the predictors. Evaluation can also be compromised when preprocessing uses the full dataset, when related observations cross partitions, or when a test sample does not represent the intended prediction setting [3]. Cawley and Talbot showed how model-selection procedures themselves can overfit a finite evaluation sample [13]. Reproducibility initiatives consequently emphasize executable code, documented data, and transparent reporting rather than a favorable score alone [14, 4]. Underspecification adds another concern: models with similar held-out performance can behave differently under deployment conditions that the evaluation did not identify [15]. The primary baselines use fixed settings and retain every scheduled split to concentrate on source and evaluation choices; a separate nested tuning sensitivity examines whether the model ordering persists.

Structured validation separates the unit recorded in a file from the unit about which generalization is claimed. Roberts et al. examine this distinction for data with hierarchical and other forms of dependence [16]. Here, reliable respondent identifiers are unavailable across the relevant records. Exact predictor signatures therefore provide a conservative, reproducible grouping rule. They are not an entity-resolution method. Different enterprises can give the same answers, especially on short ordinal scales. Group-disjoint evaluation in this study asks whether a predictor generalizes to an unseen recorded feature profile. It cannot establish performance on an independently sampled enterprise population. This narrower interpretation preserves the value of the computation without converting profile equality into an unsupported claim about respondent identity.

B. Conformal Inference with Uncertain Responses and Grouped Observations

Conformal prediction supplies a general calibration mechanism, but its guarantee is attached to a specified event and sampling structure [7]. Einbinder et al. investigate when noisy calibration labels permit coverage of an unobserved clean target [17]. Stutz et al. address ambiguity through

distributions over classification labels [8]. Clarkson et al. analyze contamination of calibration scores [18]. These settings differ from a workbook containing two observed numerical versions whose authority is unresolved. In the latter case, an analyst can target simultaneous inclusion of both recorded values without asserting that either value is the truth. That finite-source objective remains weaker than validation of an underlying construct.

Grouped observations create a separate issue. Lee et al. develop distribution-free inference for hierarchical data and distinguish the relevant exchangeability conditions [9]. Barber et al. extend conformal reasoning beyond ordinary exchangeability under specified departures [19]. The present construction takes the maximum residual across recorded members and available versions of a predictor profile. Applying split conformal calibration to these profile-level scores is a direct use of the standard rank argument. Its purpose is to align the reported event with the audit unit. The assumptions remain explicit, and empirical profile grouping is not presented as proof that those assumptions hold.

C. Selective Prediction and Conditional Coverage

Selective prediction allows a system to abstain, trading the fraction of retained cases against error among those cases. The reject-option literature formalizes this tradeoff and the role of an uncertainty ranking [20]. Risk-controlling prediction sets and conformal risk control address broader error objectives than ordinary marginal coverage [21, 22]. However, a low average error after abstention is not evidence that the retained intervals still satisfy their advertised coverage level. Jin and Ren develop selection-conditional conformal inference, and Bao et al. study selective and online procedures with false coverage-statement rate control [10, 23, 24]. This study uses a simpler independent-gate construction suitable for a transparent diagnostic. It does not claim the generality or efficiency of those methods.

D. Positioning of the Contribution

The closest prior methods address uncertain labels, hierarchical observations, and selection-conditional coverage [8–10, 17–19, 23, 24]. We do not propose a new conformal theorem: the two-version envelope is an algebraic identity, and calibration of profile maxima is standard split conformal prediction applied to a different unit. Table I distinguishes the prior targets from the present recorded-source event. We add a numerical hierarchical conformal prediction (HCP) comparator [9] and retain a Bonferroni version comparator. Ambiguous-label classification methods [8] require label probabilities rather than two fixed continuous responses; treating them as directly interchangeable would change the prediction task. The workflow contributes auditable comparisons and explicit failure reporting, rather than demonstrated superiority over those methods. Tabular prediction advances [25, 26] do not remove the need to define this evaluation target.

Three outputs make the contribution inspectable. The workflow specifies what is covered and which observations may cross partitions. The experiments distinguish accuracy, profile coverage, and interval width, including infinite intervals. The accompanying code and saved predictions

permit every reported metric to be recalculated. A positive result for one output is not treated as success for the others.

TABLE I. RELATION TO PRIOR METHODS

Strand	Prior focus	Present study
Business prediction [1]	Model comparison on survey outcomes	Source-sensitive evaluation before deployment claims
Label uncertainty [8, 17]	Ambiguous or noisy labels	Simultaneous coverage of observed worksheet versions
Hierarchical inference [9]	Grouped sampling and exchangeability	Predictor-profile stress test without inferred identities
Selective inference [10, 23]	Coverage after selection	Independent gate with small calibration-set accounting
This study	Joint source and evaluation audit	Executable ablations, retained coverage, and simulation

III. MATERIALS AND METHODS

A. Study Design and Source Boundaries

The study is a retrospective secondary computational analysis of two independent numerical case studies and controlled simulations. The original analysis was executed on 28 September 2026. The reviewer-requested extension was executed on 19 October 2026 after inspection of the submitted findings. Neither stage was prospectively registered. Nested tuning, HCP comparisons, uncertainty intervals, subgroup diagnostics, redundancy checks, cost scenarios, and ordinal stress tests are exploratory additions. Every scheduled partition and unfavorable result is retained. The datasets are never pooled, and neither serves as external validation of a predictor trained on the other.

Dataset A corresponds by filename and subject matter to the version-one Mendeley deposit by Yustian [11], describing 206 culinary-sector SME actors in West Java. The supplied workbook contains csv and datana worksheets with 206 relevant positions each. Dataset B corresponds to the version-one ICTIM deposit by Leksana et al. [12], describing 301 purposively sampled culinary MSME owners or managers. Its worksheet named TABULASI DATA, with a trailing space in the file, contains 301 rows of 25 numerical responses. These recruitment descriptions are attributed to the depositors. Both deposit pages specify CC BY 4.0. SHA-256 hashes identify the exact local workbooks and exports; binary equality with repository-hosted files has not been established. Every exported numerical cell was checked against the supplied workbook, without changing source values.

B. Extraction and Numerical Outcomes

In A, the csv worksheet contains four X columns, three XMed columns, Xmod, and five Y columns. The corresponding datana block supplies the same numerical roles but differs in some values. The four X columns are identical at all 206 positions. They provide the primary predictors, avoiding an arbitrary choice between conflicting covariates. Both worksheets retain their own five-column outcome mean. The midpoint of those two means is used only as a computational reference for interval comparisons. It is not a corrected observation. The observed component scores range from 8 to 15, and the underlying item aggregation is not documented sufficiently to interpret the numerical mean as a

validated scale. Positional alignment is assumed because csv lacks a shared record identifier. Agreement of the four X columns supports but does not prove that alignment.

In B, the first 20 columns contain five indicators each for Marketing Capability, Innovation Capability, Resource Capability, and Reputation. The final five columns are labeled Sustainable Competitive Advantage. The numerical response is their arithmetic mean, with an observed response scale from 1 to 5. Feature set F0 comprises the 15 Marketing Capability, Resource Capability, and Reputation columns. F1 adds the five Innovation Capability columns. These labels are retained from the source. Questionnaire wording, reverse-scoring rules, and individual recruitment records are not available in the supplied package, so the means are treated as numerical response summaries. No factor validity, financial performance, revenue growth, or causal effect is inferred. The raw workbook ranges were compared with the analysis exports, and all extracted numerical cells matched.

C. Source Audit and Validation Units

The audit compares literal values between A's worksheets, counts complete response patterns in B, and checks matching contiguous blocks. No value is imputed, reconciled, or removed. Demographic summaries are not expanded into individual records. Derived full-sample standardizations and earlier model outputs are excluded from the predictors. Table II summarizes the observable source structure and the associated inference boundaries.

TABLE II. SOURCE STRUCTURE AND ANALYSIS UNITS

Property	Dataset A	Dataset B
Recorded positions or rows	206	301
Core score fields	13 per version	25
Source versions analyzed	2 worksheets	1 tabulation
Predictor profiles	177 using 4 fields	177 using 20 fields
Complete response patterns	206 per version	185
Primary numerical response	Mean of 5 Y fields	Mean of 5 SCA fields
Main unresolved issue	Version authority and alignment	Origin of repeated blocks

Profile counts are not verified respondent counts. A and B are separate numerical case studies.

A predictor profile is the exact vector of the four invariant X values in A or the 20 predictor values in B. All records sharing a profile receive the same partition assignment in profile-disjoint experiments. The rule uses predictors only. It therefore avoids defining validation groups through the response being predicted. A has 177 such profiles and B independently has 177; this numerical coincidence does not make the samples interchangeable. The grouping is deliberately conservative because equal profiles can belong to different people. Its estimand is prediction for an unseen profile, with recorded multiplicities retained inside each profile.

D. Point Prediction and Repeated Partitions

Each primary profile experiment uses 50 partitions with seeds 20260928–20260977. Exact feature vectors are numbered lexicographically, and a NumPy default_rng

permutation assigns 106 of 177 profiles to training, 35 to calibration, and 36 to testing. All rows of a profile move together. B also uses row partitions of 180/60/61. Models and paired feature ablations share each partition. Standardization is fitted on training data only; calibration and test responses are excluded from model fitting. The row and profile designs differ in training diversity and prediction target. Their contrast is therefore split sensitivity, not an unbiased causal estimate of leakage inflation. The archive stores the assignment of every row.

Five fixed baselines are evaluated: the training median, ridge regression, random forest, histogram gradient boosting, and one-nearest-neighbor regression. Ridge uses training-only standardization and penalty 10 [27]. Random forest uses 100 trees, minimum leaf size three, and all predictors as split candidates [28]. Histogram boosting uses 100 iterations, learning rate 0.05, seven maximum leaves, minimum leaf size ten, L2 penalty one, and no early stopping. One-nearest-neighbor uses standardized Euclidean distance. Estimator random states equal the partition seed where applicable. No hyperparameter search is performed for these primary baselines; the separate nested tuning sensitivity is described below. Implementations use Python 3.12.14, NumPy 2.3.5, and scikit-learn 1.8.0 [29]. Fixed settings permit controlled comparisons but do not establish the best achievable learner.

In A, each model is fitted separately to csv, datana, and midpoint responses and evaluated against all three references. B compares F0 and F1. Row-weighted mean absolute error (MAE), root mean squared error (RMSE), and test R squared describe recorded-row prediction. Profile-balanced MAE averages absolute errors within each canonical profile before averaging profiles equally. Ablation contrasts are paired within partitions. Means and standard deviations (SDs) over 50 overlapping splits measure partition sensitivity [30, 31]; they are not 50 independent empirical replications. The original code and saved outputs are retained, and the reviewer extension supplies additional definitions and reproducibility checks.

A training-only tuning sensitivity uses the first 20 outer seeds under both row and profile splitting. Three inner folds use shuffled row folds for the row design and disjoint canonical profiles for the profile design. Ridge penalties are 0.1, 1, 10, and 100. Forest grids cross minimum leaf sizes 1, 3, and 5 with predictor fractions 1.0 and 0.5. Boosting grids cross 3, 7, and 15 leaves with L2 penalties 0 and 1. Other settings remain fixed. The smallest pooled inner-validation row MAE selects a configuration; exact ties follow the saved grid order. Neither outer calibration nor test outcomes enter selection. This limited grid checks sensitivity to settings, not global optimality.

E. A Finite-source Uncertainty Envelope

Let the two outcomes at aligned position i be y_A and y_D , with midpoint a and half-distance d . For a common prediction f , Eq. (1) is an identity, not a new uncertainty estimator. It makes the recorded-version coverage objective explicit.

$$\max_{v \in \{A,D\}} |y_i^{(v)} - f(x_i)| = |a_i - f(x_i)| + d_i \quad (1)$$

This identity separates prediction error relative to the midpoint from the observed disagreement between versions. Even an unrestricted position-specific prediction cannot reduce this worst-version error below d . The average half-distance is therefore a descriptive lower bound for this finite-source objective. It is not an estimate of measurement error in an unobserved true outcome. If both versions are incorrect, the construction does not protect the true value. If alignment is incorrect, the paired envelope loses its intended interpretation.

For a fitted predictor f and positive scale h , Eq. (2) defines the maximum standardized residual over every recorded member and available outcome version of profile g . A uses both worksheet outcomes. B has one version but may contain different outcomes within the same predictor profile. The maximum changes the event from covering a randomly selected member to covering the entire recorded profile.

$$S_g = \max_{i \in g, v} \{ |y_i^{(v)} - f(x_i)| / h(x_i) \} \quad (2)$$

The unnormalized construction sets h to one. The normalized random-forest variant sets h to the standard deviation of the 100 tree predictions plus 0.05. This forest dispersion is an uncertainty ranking, not an independently valid prediction standard deviation. Calibration determines its error scale. The added constant is fixed in response units and prevents division by zero.

F. Calibration and the Coverage Event

For m calibration profiles and miscoverage α equal to 0.10, the calibration radius q is the k -th ordered profile score, where k is the ceiling of $(m + 1)$ times $(1 - \alpha)$. If k exceeds m , q is infinite. Eq. (3) gives the interval centered on the fitted value, with half-width q times h . Closed endpoints include tied scores.

$$C_i = [f(x_i) - qh(x_i), f(x_i) + qh(x_i)] \quad (3)$$

Conditional on the fitted predictor and scale, assume that the calibration profiles and a new test profile are exchangeable as complete collections of recorded members and versions. The new maximum score then has a uniform rank, or a conservatively tied rank, among $m + 1$ scores. The probability of exceeding the selected order statistic is at most α , so all recorded outcomes and versions within one randomly drawn new profile are covered with probability at least $1 - \alpha$ [7, 9]. This is the standard split-conformal argument applied to a maximum score. It does not guarantee coverage for every covariate value, all future profiles simultaneously, unrecorded source versions, or an unobserved true business outcome. Random partitions support a finite-file diagnostic; the exchangeability needed to generalize to a survey population is not established.

A compares row-midpoint, profile-midpoint, and profile-version calibration around the same midpoint-trained ridge model. Bonferroni calibration uses each version's profile maximum at 95 per cent and takes the larger radius. The added HCP comparator [9] instead gives each calibration profile equal total weight and each member equal weight within that profile, using its maximum residual over the two versions. An additional mass of $1/(m + 1)$ is placed at infinity before taking the weighted 90th percentile. HCP targets a member's version

envelope under hierarchical exchangeability, whereas profile-maximum calibration targets all recorded members. Both are evaluated on the same simultaneous-profile event; a difference on this stricter event is not a violation of HCP's stated guarantee. B applies the corresponding single-version comparators. Widths use the original response units, without clipping to observed limits.

For each A calibration comparison, we compute the paired difference in simultaneous-profile coverage within each of the 50 splits. To expose dependence sensitivity, we report an approximate 95 per cent corrected repeated-holdout interval [32]: the paired mean plus or minus the 49-degree-of-freedom t critical value times the square root of the product of $(1/50 + 36/106)$ and the sample variance of paired differences. The test-to-proper-training ratio 36/106 is used explicitly; using 36/141 would give a smaller adjustment. This approximation extends a generalization-error variance correction to bounded coverage differences. It is not an exact confidence interval, a nested population bootstrap, or evidence that profiles are independent enterprises. We do not infer superiority when it includes zero; independent authenticated sampling remains necessary for population inference.

G. Selection and Calibration after Selection

The primary selection diagnostic uses B's F1 normalized random forest. Thresholds are calibration uncertainty quantiles for nominal retention 50%, 70%, and 90%; the 100% reference predicts every case. Comparators use nearest-training-profile distance after training-only standardization and seeded random retention. All selectors share point predictions and the normalized profile interval. Realized row retention, retained MAE, row coverage, and profile coverage are saved. Because the same calibration sample sets the primary threshold and interval, these curves are descriptive and have no selection-conditional guarantee [10, 23]. Different realized retention also precludes a claim of exact workload-matched dominance.

The independent-gate analysis reserves the final 22 profiles in the permuted training block for threshold setting and fits one forest on the remaining 84. Each gate profile contributes one dispersion or nearest-distance value. The gate uses no calibration or test response. Within each selector and retention setting, all-profile and accepted-profile calibrators share the same fitted model, threshold, and accepted test profiles. Constant residual scale and forest-dispersion normalization are both evaluated. Conditional on the fitted model, independently fixed gate, and exchangeability of accepted profiles, split calibration within the accepted population targets their simultaneous-profile coverage [10, 23, 24]. This does not guarantee an improvement over an already conservative comparator. With $\alpha = 0.10$, fewer than nine accepted calibration profiles necessarily produce an infinite radius; these runs remain in the results.

Algorithm 1. Evaluation with profile and version calibration

Input: fixed numerical exports, outcome versions, predictor-only grouping, learner, positive scale h , α , and retention level.

1. Audit source values and define the numerical target. Preserve all recorded rows and versions.
 2. Split profiles into proper training, calibration, and testing. When
-

selecting predictions, reserve independent training profiles to set the gate.

3. Fit preprocessing, predictor f , and scale h on proper training data. Fix the gate using one score per gate profile.

4. For every accepted calibration profile, compute the maximum residual over its members and versions, divided by h .

5. Select score order $\text{ceil}((m + 1)(1 - \alpha))$; use an infinite radius if this order exceeds m . Apply $f \pm qh$ to accepted test records.

6. Report retention, row error, simultaneous-profile coverage, interval width, and infinite-radius frequency. Save assignments and predictions.

H. Simulation and Additional Sensitivities

The original ten Gaussian-predictor scenarios, each with 200 independent trials, are retained in the supplement as calibration checks. Their known version displacement tests the intended coverage construction and does not establish realism for ordinal surveys. We add six ordinal stress scenarios with 200 trials each, 300 development clusters, and 1,000 separately generated external clusters. A common Gaussian factor and four correlated block factors generate 20 predictors rounded and clipped to integers from 1 to 5. Members of a synthetic cluster share predictors but have distinct outcomes, each averaging five rounded, bounded items. Synthetic cluster identifiers are known; equal predictor vectors across independently generated clusters need not denote the same cluster. This differs from the empirical predictor-only grouping assumption.

The latent outcome mean depends on the overall predictor mean and the first two block means; its noise increases with distance from the middle of the predictor range. The second version adds 0.8, bounded at 5, with probability 0.55 for high-mean profiles and 0.15 otherwise. Baseline cluster size is $1 + \text{Poisson}(2)$. The scenarios remove disagreement; retain the ordinal baseline; make repetition depend on predictor level; multiply external noise by 2.5; introduce an unrecorded third version shifted by 1.4; or permute second-version rows, breaking alignment. Exact equations, seeds, and clipping rules are supplied in code. Ridge is fitted to the mean of the first two versions on 180 clusters, with 60 for calibration; unused development clusters do not enter testing. All methods share the same fit and external sample. The target covers every recorded member and the evaluated versions; the unrecorded-version scenario deliberately expands that target at testing. Trial SD divided by the square root of 200 gives Monte Carlo standard error (MCSE). These are constructed failure scenarios, not fitted generative models of either survey.

Additional sensitivities examine grouping and source multiplicity. B's distinct 20-feature signatures are joined into connected components when their total absolute difference is at most one. A common grouping based on the 15 F0 inputs prevents identical reduced-feature inputs crossing partitions in either ablation. A separate training sensitivity assigns each row inverse profile-frequency weight, rescaled to mean one, for ridge and forest; both standardization and ridge fitting use these weights. Weighted forest fitting retains its usual row bootstrap, so this sensitivity does not eliminate every influence of multiplicity. All variants and the exploratory gate-scale comparison are reported, without selecting a preferred split.

I. Heterogeneity and Decision Costs

We stratify held-out predictions by profile repetition frequency (one, two, or at least three rows), by recorded response range, and by maximum source half-disagreement within an A profile (zero, up to 0.2, or above 0.2). Response bands use the observed outcome and are retrospective diagnostics, not deployable selectors. Row coverage is used for response bands because members of one profile may lie in different bands; repetition and disagreement strata use the complete-profile event. We report distinct profiles and repeated test appearances separately. Subgroup averages pool held-out appearances across the saved splits and do not create independent observations or subgroup coverage guarantees. A training-only ridge regression of each innovation item on F0, evaluated on held-out profiles, provides a predictive redundancy diagnostic without establishing psychometric validity.

For a fixed selector, the per-row decision loss is the mean of absolute prediction error for retained rows and a constant cost c for abstained rows. The scenarios $c = 0.10, 0.25$, and 0.50 are measured in response-score-error equivalents. They assume that manual review removes the prediction error and that its full burden is c . These are transparent utility assumptions, not observed prices or evidence of effective review. Costs are calculated per split before averaging, alongside the always-predict loss. If review leaves residual error, its expected magnitude must be added to c . The calculation cannot by itself select an operational policy.

J. Implementation Details

Feature columns retain their CSV order: A uses X1–X4; B F1 uses the first 20 columns, and F0 takes positions 1–5 and 11–20. Profiles use exact numerical equality and lexicographic numbering; rows within each partition retain ascending source position. StandardScaler uses training means and population variances. The forest uses squared-error splits, row bootstrap sampling, all predictors, no maximum depth, and one execution thread; boosting disables early stopping and categorical inference in the extension. Tree dispersion uses population SD plus 0.05. Gate thresholds use NumPy's linear quantile and accept equality. Conformal quantiles use the integer order statistic without interpolation; finite coverage comparisons use a $1e-12$ absolute tolerance. Saved decimal outputs are verified with $1e-9$ tolerance for boundary-rounding effects.

The supplement records complete estimator parameters, seeds, feature order, runtime versions, source hashes, and execution outputs. The original one-neighbor baseline retains scikit-learn 1.8.0's automatic search algorithm and row-order-dependent handling of equidistant neighbors. An additional deterministic sensitivity selects the lowest original training-row position among squared distances within $1e-12$ of the minimum. Thus, library-specific tie behavior is documented rather than assumed portable. The extension uses the pinned package versions in requirements.txt; its runtime manifest distinguishes the current execution environment from the original Python 3.12.14 analysis.

IV. RESULTS

A. Observable Source Uncertainty

The A worksheets differ in 58 of 2,678 core cells, affecting 46 of 206 aligned positions, or 22.33%. Thirty-eight differences occur in the five outcome columns and affect 32 outcome means. The mean absolute difference between the two outcome means is 0.0650 score units, with a maximum of 1.600. Their average half-distance is 0.0325. Thus, the aggregate finite-source disagreement is modest, although some individual discrepancies are appreciable. The csv response mean is 12.4126 and the dataNA response mean is 12.3709. Complete equality of X1–X4 permits a comparison with invariant inputs, but it does not establish which outcome version is authoritative.

B contains 185 unique complete 25-value patterns. The remaining 116 rows repeat a previously observed complete pattern, representing 38.54% of rows. Fifty-three repeated-pattern groups contain 169 rows when their first occurrences are included. Twenty-four rows are constant across all 25 columns. A sequence of 30 rows, at Excel rows 4–33, recurs identically at rows 211–240. Predictor-only grouping gives 177 profiles, four of which have multiple distinct numerical outcome means. These findings describe file structure. They do not establish duplicate respondents, and no row is removed on that basis.

B. Model Rankings under Fixed Settings and Tuning

Under ordinary row splitting in B, 49.64% of test rows have a 20-feature profile also present in training, averaged across splits. Training contains an average of 120.36 distinct profiles in this design, compared with 106 in the profile-disjoint design. The two designs have similar mean training-row counts, 180.00 and 179.76 respectively, but they expose the model to different feature-profile coverage. Table III reports performance with the full F1 feature set.

TABLE III. PREDICTION WITH ALL B PREDICTORS

Model	Row-split MAE (SD)	Profile-split MAE (SD)	Profile-split R ²
Median	0.519 (0.047)	0.511 (0.075)	-0.099
Ridge	0.214 (0.017)	0.237 (0.036)	0.745
Random forest	0.197 (0.020)	0.281 (0.046)	0.639
Histogram boosting	0.198 (0.028)	0.276 (0.050)	0.651
One neighbor	0.229 (0.038)	0.414 (0.094)	0.111

Row-weighted metrics by partition design; means and SDs over 50 dependent splits. MAE uses SCA score units. SD describes partition sensitivity, not a confidence interval. Full metrics are in [Supplementary Material](#).

Random forest has the lowest row-split MAE, 0.1966. Under profile-disjoint splitting, its MAE rises to 0.2814, a 43.1% relative increase, and mean test R squared falls from 0.8208 to 0.6389. Ridge becomes the strongest evaluated predictor, with MAE 0.2369 and R squared 0.7445. One-nearest-neighbor changes from MAE 0.2291 to 0.4135. These contrasts identify sensitivity to previously observed profiles, training diversity, and the test estimand together. Fig. 1 includes split SDs so the average ranking is not mistaken for uniform superiority on every partition.

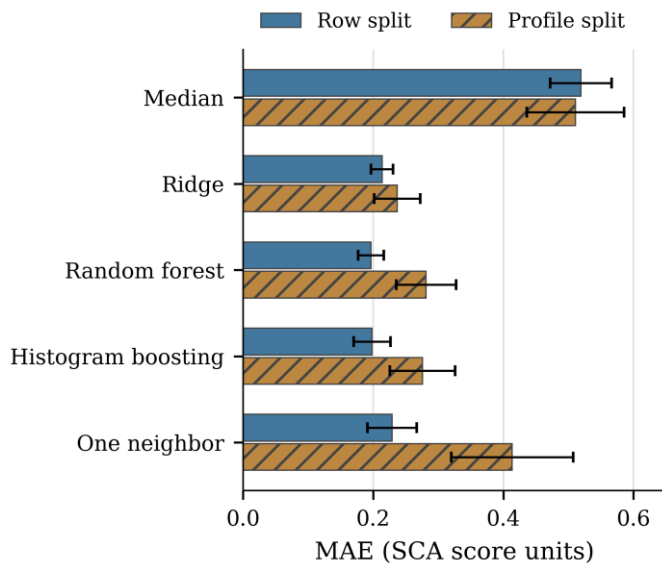


Fig. 1. B Prediction with all F1 inputs. Bars show mean row-weighted MAE and error bars show SD over 50 dependent splits, not confidence intervals. Source: saved point-prediction summaries.

Profile-balanced MAE supports the main ranking: ridge 0.2466 and forest 0.2630. The near-profile sensitivity reduces 177 profiles to 155 connected components and yields row MAEs of 0.2431 and 0.2820, respectively. Inverse-frequency training weights yield row MAEs of 0.2339 and 0.2723, with profile-balanced values 0.2415 and 0.2584. Ridge remains strongest among these two learners across the tested grouping and weighting choices. This robustness concerns the recorded numerical task and does not resolve respondent-level generalization.

The nested tuning sensitivity preserves the mean ordering for the tested grids. Across 20 outer splits, row MAEs are 0.2208 for ridge, 0.1754 for forest, and 0.1976 for boosting; profile-disjoint MAEs are 0.2495, 0.2896, and 0.2871, respectively. These figures use a different number of splits from Table III and are not paired against its 50-split averages. They reduce, but do not eliminate, the concern that the ranking reflects one arbitrary configuration. They do not establish a universal advantage of ridge or isolate the causal effect of the partition unit.

C. Incremental Prediction from Innovation Indicators

With the primary profile split, adding innovation indicators reduces ridge MAE from 0.2439 to 0.2369, a paired mean change of -0.0070 ; 72% of splits favor F1. Forest changes from 0.2828 to 0.2814 and histogram boosting from 0.2813 to 0.2757. One-nearest-neighbor worsens from 0.3980 to 0.4135. For ridge, the profile-balanced improvement is only 0.0003, from 0.2469 to 0.2466. These paired summaries show a small, learner-dependent incremental signal, with recorded multiplicity influencing its apparent size. The fraction of favorable splits is descriptive, not a hypothesis-test probability.

Primary 20-feature grouping permits identical F0 vectors to appear in training and testing for 12.59% of test rows on average. The common F0 grouping removes this overlap and yields 166 groups. Ridge row MAE then decreases from

0.2445 to 0.2401 with innovation, while canonical-profile-balanced MAE increases from 0.2431 to 0.2466. Forest row MAE changes from 0.2808 to 0.2813. The incremental information is therefore not stable across weighting choices. These numerical results cannot establish whether an innovation intervention improves enterprise outcomes.

Innovation redundancy is substantial for some items but not uniform. Predicting the five innovation indicators from F0 with training-only ridge gives mean held-out R squared values of 0.853, 0.895, 0.288, 0.730, and 0.208. This shows that F0 can reproduce much of the variation in some recorded items. It does not identify whether that overlap reflects related constructs, shared response tendencies, or duplicated wording. The incremental-outcome ablation and the item-prediction diagnostic answer different questions. Without instrument wording, scoring rules, and an independent measurement study, limited incremental prediction cannot be attributed to low substantive value or to a particular measurement defect.

D. Version Sensitivity and Simultaneous Coverage

A's invariant-input prediction task is substantially less predictable than B's numerical task, although their absolute errors cannot be compared because the score units differ. Ridge trained and evaluated on csv has MAE 0.8184 and mean R squared 0.1869. Holding that fitted model fixed while evaluating against datana increases MAE to 0.8791. Training on datana instead gives MAE 0.8778 against datana. The small improvement from changing the training version shows that response-reference choice matters more here than refitting alone. Midpoint-trained ridge has MAE 0.8460 and R squared 0.1780 against the midpoint. This computational average is not presented as the best representation of business performance.

Table IV compares intervals around the same midpoint-trained ridge predictor. Every coverage entry asks whether both source versions are included, even when the calibrator uses midpoint residuals only.

TABLE IV. SIMULTANEOUS VERSION COVERAGE IN A

Calibration	Profile coverage	Width
Row midpoint	87.28%	3.578
Profile midpoint	89.39%	3.862
Profile versions	90.00%	3.909
HCP envelope	89.00%	3.800
Bonferroni versions	96.44%	5.303

All methods share the midpoint-trained ridge model. Coverage requires both versions and every recorded member of a test profile. HCP's native guarantee concerns a member, so this column evaluates a stricter event. Means over 50 dependent splits; paired uncertainty is reported in the text.

Profile-version calibration has mean simultaneous-profile coverage of 90.00%, compared with 87.28% for row-midpoint calibration, 89.39% for profile-midpoint calibration, and 89.00% for HCP. Their widths are 3.909, 3.578, 3.862, and 3.800, respectively. Bonferroni calibration reaches 96.44% with a width of 5.303. The paired gain over row midpoint is 2.72 percentage points, with an approximate dependence-adjusted 95% interval of -1.89 to 7.33 points. The additional gain over profile midpoint is 0.61 points (-1.95 to 3.17), and the gain over HCP is 1.00 point (-1.85 to 3.85). These intervals include zero. The observed improvements therefore do not

establish an advantage beyond sampling and partition uncertainty. The defensible distinction is the target event and its assumptions, not empirical superiority.

In B, ridge with row-residual calibration has simultaneous-profile coverage of 88.11%. Profile-maximum calibration reaches 90.67%, with width increasing from 1.029 to 1.048. The normalized forest reaches 91.44% row coverage and 92.00% profile coverage, with width 1.369; its unnormalized profile calibrator reaches 90.11% profile coverage with width 1.191. This comparison makes the additional width explicit. A wider or more conservative interval is not automatically a better operating decision.

E. Abstention and the Price of Selected Calibration

For B's F1 forest, the primary width rule retains 47.74% of test rows at nominal 50% retention. MAE decreases from 0.2814 to 0.2453, but retained-row coverage falls from 91.44% to 87.46%. At nominal 70%, retention is 65.60%, MAE 0.2493, and coverage 89.19%. Lower retained error therefore does not certify retained coverage. Fig. 2 displays both quantities against actual retention.

The nearest-training-distance comparator retains 47.99% of rows at its nominal 50% threshold, with MAE 0.2309 and row coverage 93.71%. Random retention gives MAE 0.2843 at 49.18% retention. The interval-width rule is therefore not uniformly superior to a basic geometric selector. Their slightly different realized workloads and the unverified empirical sampling conditions remain relevant to interpretation.

Table V reports an exploratory independent-width-gate analysis with forest-dispersion normalization. At nominal 70% retention, 67.00% of test profiles and 66.71% of rows are retained. All-profile calibration yields 88.40% retained profile coverage and width 1.106; accepted-profile calibration yields 89.55% and width 1.333. The observed mean remains below 90%. These post hoc contrasts are descriptive; neither a marginal rank argument nor 50 dependent empirical splits requires every observed average to reach the target.

TABLE V. INDEPENDENT WIDTH GATE WITH NORMALIZED SCORES

Nominal retention	Calibration	Profile coverage	Width
50%	All profiles	85.25%	0.985
50%	Selected profiles	91.60%	Infinite*
70%	All profiles	88.40%	1.106
70%	Selected profiles	89.55%	1.333
90%	All profiles	90.72%	1.234
90%	Selected profiles	90.33%	1.260

Actual profile retention at the 50%, 70%, and 90% gates: 46.67%, 67.00%, and 84.83%. *Three of 50 splits have infinite radius. Comparators share the same predictor, gate, scale, and accepted test profiles.

At nominal 50% retention, normalized accepted-profile calibration raises coverage from 85.25% to 91.60%, but three splits accept only seven or eight calibration profiles and require infinite intervals. With constant-scale scores at nominal 70%, all-profile coverage is already 92.17%; selected calibration gives 89.62%, while width decreases from 1.238 to 1.205. For the independent distance gate with normalized scores at nominal 70%, coverage changes from 91.40% to 91.05% and

width increases from 1.322 to 1.578. These exploratory contrasts show why selected calibration must be evaluated jointly with the selector and score scale. The archive retains every setting; an increase in empirical coverage is not a universal property of the procedure.

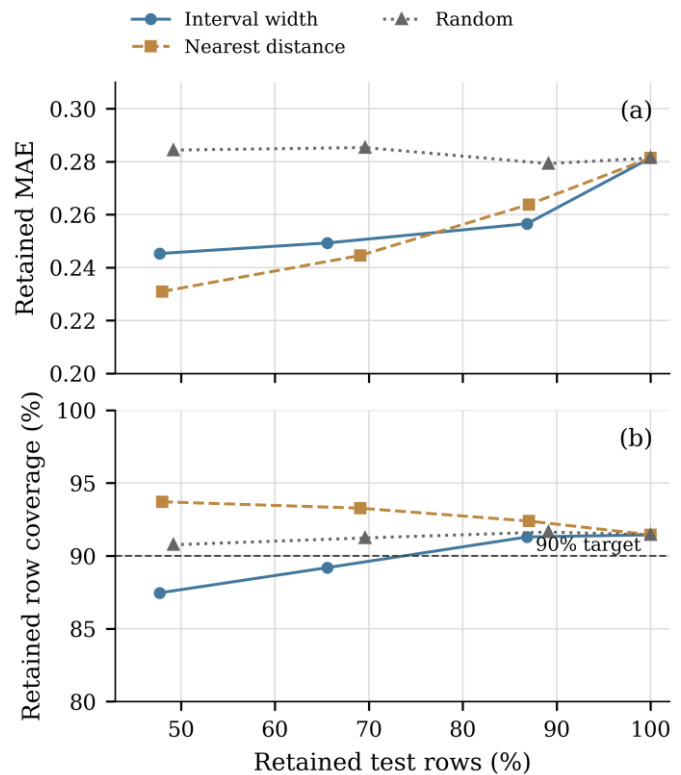


Fig. 2. Primary selection diagnostic in B. Each point averages 50 splits at a specified nominal retention: (a) retained MAE and (b) retained-row coverage. Lines connect operating points; axes are focused to show differences. The dashed line is the 90% target. These calibration-reusing curves do not have selection-conditional validity.

F. Ordinal Simulation Stress Tests

Table VI reports the added ordinal scenarios. Without source disagreement, profile-midpoint and profile-version intervals coincide, with 89.93% simultaneous-cluster coverage. In the ordinal baseline, profile-version coverage is 90.07% (MCSE 0.29 percentage points), while row-midpoint coverage is 62.47% and HCP coverage on the stricter whole-cluster event is 77.85%. The HCP figure must not be read as failure of its marginal member guarantee. Unlike the original identical-copy simulations, multiple members here have different responses, making simultaneous coverage more demanding.

Informative repetition gives 90.03% profile-version coverage, but the deliberately nonexchangeable noise-shift scenario reduces it to 52.92%. Adding an unrecorded third version at testing reduces it to 41.39%. Under row misalignment, the recorded-envelope coverage remains 89.58% despite the corrupted pairing, with width increasing to 3.730. This is a provenance failure that coverage cannot detect: an interval may cover the supplied numbers while the intended record correspondence is wrong. These stress tests demonstrate boundaries and failure modes; achieving near-nominal

coverage under a compatible generator is a construction check, not independent evidence of methodological novelty.

TABLE VI. ORDINAL STRESS TESTS OF SIMULTANEOUS COVERAGE

Scenario	Row midpoint	HCP	Profile versions	Width
No disagreement	75.83%	78.19%	89.93%	2.399
Ordinal baseline	62.47%	77.85%	90.07%	2.764
Informative repetition	66.55%	77.34%	90.03%	2.751
External noise shift	26.64%	37.43%	52.92%	2.737
Unrecorded version	8.98%	20.68%	41.39%	2.767
Row misalignment	24.48%	77.00%	89.58%	3.730

Percentages cover every member and evaluated version of an external synthetic cluster. Mean of 200 independent trials; 1,000 external clusters per trial. HCP targets a member and is assessed here on a stricter event. Width is for profile-version intervals. These are simulated scores, not additional survey observations.

G. Heterogeneity Hidden by Aggregate Summaries

Table VII separates canonical profiles by recorded repetition frequency. In B, ridge's pooled held-out MAE decreases from 0.252 for single-row profiles to 0.222 for profiles repeated at least three times, yet simultaneous-profile coverage drops from 91.77% to 81.36%. The strata contain 128 and 21 distinct profiles, respectively. A's four profiles in the highest repetition band are too few to support a stable subgroup conclusion. Frequency is a property of the file, not evidence that a respondent was duplicated.

TABLE VII. RIDGE RESULTS BY PROFILE REPETITION

File / rows per profile	N	Test appearances	MAE	Coverage
A / 1	153	1562	0.850	90.78%
A / 2	20	201	0.909	82.09%
A / 3+	4	37	0.533	100.00%
B / 1	128	1300	0.252	91.77%
B / 2	28	280	0.229	92.86%
B / 3+	21	220	0.222	81.36%

N denotes distinct profiles; appearances count repeated profile tests across 50 splits. MAE pools recorded-row errors, whereas coverage requires the whole profile. Coverage and error therefore use different explicit weighting rules. Widths and forest results are in the supplement.

TABLE VIII. DISAGREEMENT AND RESPONSE RANGE DIAGNOSTICS

File / stratum	N	MAE	Coverage	Width
A / 0	145	0.824	92.15%	3.912
A / (0,0.2]	25	0.767	88.03%	3.907
A / >0.2	7	1.505	59.55%	3.892
B / <3	10	0.630	37.69%	1.073
B / 3 to <4	98	0.192	96.00%	1.043
B / >=4	70	0.234	90.95%	1.053

A strata use profile-maximum half-disagreement and complete-profile coverage; B strata use row outcomes and row coverage. All results use ridge profile-version intervals. N is distinct profiles and may overlap across B response bands. Widths use each dataset's own score units; cross-dataset magnitudes are not comparable.

Source disagreement and response range expose further heterogeneity (Table VIII). The seven A profiles with maximum half-disagreement above 0.2 have lower simultaneous coverage than profiles with no disagreement

despite similar mean interval widths; this small stratum supports a diagnostic warning rather than a precise subgroup estimate. In B, outcomes below 3 have only 37.69% row coverage and MAE 0.630, compared with 96.00% and 0.192 in the 3-to-below-4 band. The low-response band contains 21 rows from 10 profiles and 199 repeated test-row appearances. These counts are not 199 independent observations. A's below-12 band similarly has 84.70% row coverage versus 99.81% in its 12-to-below-13 band. Constant-width ridge intervals conceal these local failures; no conditional validity is claimed.

H. Hypothetical Review-Cost Analysis

For B's primary nominal-50% width selector, average loss is 0.168, 0.247, and 0.377 at review costs 0.10, 0.25, and 0.50, respectively. Always predicting has loss 0.281. The distance selector gives 0.164, 0.242, and 0.372 at 47.99% retention, compared with 47.74% for the width selector. Thus, apparent value disappears at the highest assumed review cost even though retained MAE is lower. These retrospective cost scenarios use the calibration-reusing primary selectors and do not establish optimal thresholds or deployment savings. They quantify the missing decision assumption while preserving the coverage caveats.

V. DISCUSSION

For these files and evaluated settings, B's row split favors random forest and profile-disjoint evaluation favors ridge. Limited nested tuning preserves that mean ordering. This supports a sensitivity finding, not an algorithm recommendation for small enterprises. Equal responses do not establish equal respondents; the empirical profile split tests transfer to previously unseen recorded feature vectors. Repetition-stratified results show why row-weighted averages alone can miss weak coverage for multi-member profiles. The general lesson is to report the prediction unit and weighting rule, then inspect their consequences [3, 16].

The innovation block has a small, unstable incremental contribution to the recorded outcome. The additional item-prediction analysis identifies substantial overlap with F0 for some indicators, while leaving appreciable unexplained variation in others. It cannot distinguish construct redundancy from shared response tendencies or deficiencies in measurement. Consequently, neither the ablation nor its limited gain supports a claim about the real-world value of innovation. Better predictors [26] may change numerical performance but do not supply missing measurement documentation.

The interval comparisons apply established conformal methods to different coverage events. HCP addresses hierarchical sampling for a member; profile maxima target every recorded member and version. The empirical coverage gains have dependence-adjusted intervals spanning zero, so they do not demonstrate superiority. Width and subgroup diagnostics are essential even when aggregate coverage is near 90%. The ordinal simulations additionally show that exchangeability violations and unseen versions defeat the finite-source objective, while misalignment can remain invisible to coverage. Resolving version authority and recovering original scoring and linkage information remain necessary for stronger substantive claims.

Selection changes the evaluated population and consumes calibration information. An independently fixed gate permits calibration within the accepted population under the stated assumptions, but it need not improve a particular empirical result. Normalization, gate choice, and small calibration counts matter; the aggressive gate produces infinite intervals in three partitions. The cost scenarios demonstrate how the same retained-error reduction can be favorable or unfavorable depending on assumed review burden. Neither review effectiveness nor operational savings was measured. A deployment study would need externally justified costs, capacity constraints, authenticated inputs, and prospective evaluation of the full human–model process.

The two availability-selected datasets cannot estimate the prevalence of these problems in enterprise surveys. The file audits do not authenticate respondents, positional alignment, recruitment, or consent. The numerical outcomes have unverified measurement properties, and A has limited predictability from four invariant inputs. The 50 splits remain dependent; the variance correction is approximate and cannot replace independent data collection. Tuning covers only small grids, subgroup counts can be sparse, and all reviewer additions are exploratory. Ordinal simulation parameters are intentionally constructed rather than estimated from verified survey mechanisms. Beyond the two files, the work supplies an executable audit procedure, a comparison of coverage events, and examples of how reporting can fail. The numerical rankings and feature effects themselves do not generalize. Source checks and independent recomputation support computational correctness within these boundaries [4, 14].

VI. CONCLUSION

Source versions, validation units, and selection determine both the prediction target and the event an interval covers. In these two recorded-score case studies, profile separation changes the ranking of fixed and locally tuned baselines, while innovation ablations remain sensitive to weighting. Profile-version calibration targets simultaneous inclusion of recorded outcomes, but empirical gains over simpler or hierarchical comparators are uncertain. Subgroup analyses reveal failures concealed by aggregate means. Abstention trades retained error against review burden and can compromise coverage or yield infinite intervals. Ordinal simulations identify noise shift, unrecorded versions, and misalignment as distinct limitations. The contribution is a reproducible evaluation workflow using established methods. Authenticated sampling, documented instruments, and prospective cost evaluation are required before drawing enterprise-level or operational conclusions.

VII. DECLARATIONS

A. Data and Code Availability

The supplementary archive is publicly available at: <https://drive.google.com/file/d/1AZu5cYvHjIIFZjkP7qgK9Ym0N3KfRRRv/view?usp=sharing>. The original_study directory preserves the original numerical exports, code, notebook, 50-split results, 2,000 Gaussian simulation trials, and source attribution [11, 12]. The top-level code, data, and results directories contain the reviewer-requested extension, 1,200 ordinal simulation trials, tuning searches, paired coverage

intervals, subgroup diagnostics, redundancy checks, and cost calculations. PROTOCOL.txt records the retrospective specification and implementation clarifications. Running `code/run_revision.py` all followed by `code/summarize_validate.py` regenerates the extension. Original saved results are checked by `original_study/code/validate_results.py`. Full estimator parameters, software versions, hashes, and validation outputs are included. Numerical exports are byte-identical to the recovered original archive, and every exported value was rechecked against the supplied workbooks. Remote binary equality and survey authenticity remain unverified.

B. Declaration on Generative AI

ChatGPT (OpenAI) assisted with literature retrieval, code development, computational execution, analysis support, manuscript drafting, and revision, including the reviewer-requested analyses. Its contribution extended beyond language correction. The executable scripts, outputs, and validation checks are supplied for inspection. Generative AI is not an author or an evidentiary source. The named authors retain responsibility for scientific interpretation, originality, accuracy, critical review of the analyses, and the final manuscript; computational checks do not replace their scientific review.

FUNDING

This research received no grant funding.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ETHICS AND DATA USE

This secondary analysis used supplied numerical exports associated with public deposits whose repository pages specify CC BY 4.0 [11, 12]. No institutional ethics approval or formal exemption was obtained for this analysis. It involved no new recruitment, participant contact, intervention, or re-identification. The exports contain numerical response fields without direct personal identifiers. Repository availability does not establish the original recruitment or consent procedures, which this analysis does not authenticate.

REFERENCES

- [1] G. M. Ducut, “A comparative analysis of decision tree, random forest, and logistic regression models in predicting business readiness for digital technology integration,” *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 6, 2026, doi: 10.14569/IJACSA.2026.0170642.
- [2] I. Lundberg, R. Johnson, B. M. Stewart, “What is your estimand? defining the target quantity connects statistical evidence to theory,” *American Sociological Review*, vol. 86, no. 3, pp. 532–565, 2021, doi: 10.1177/00031224211004187.
- [3] S. Kapoor, A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, Art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
- [4] S. Kapoor et al., “REFORMS: consensus-based recommendations for machine-learning-based science,” *Science Advances*, vol. 10, no. 18, Art. eadk3452, 2024, doi: 10.1126/sciadv.adk3452.
- [5] T. Gebru et al., “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.
- [6] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, “Everyone wants to do the model work, not the data work:

- data cascades in high-stakes AI,” Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2021, doi: 10.1145/3411764.3445518.
- [7] A. N. Angelopoulos, S. Bates, “Conformal prediction: a gentle introduction,” *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.
- [8] D. Stutz, A. G. Roy, T. Matejovicova, P. Strachan, T. Cemgil, and A. Doucet, “Conformal prediction under ambiguous ground truth,” *Transactions on Machine Learning Research*, 2023. Available: <https://arxiv.org/abs/2307.09302>.
- [9] Y. Lee, R. F. Barber, and R. Willett, “Distribution-free inference with hierarchical data,” *ACM/IMS Journal of Data Science*, 2026, doi: 10.1145/3786352.
- [10] Y. Jin, Z. Ren, “Confidence on the focal: conformal prediction with selection-conditional coverage,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 87, no. 4, pp. 1239–1259, 2025, doi: 10.1093/jrsssb/qkaf016.
- [11] O. R. Yustian, “Tech-driven sustainability: the interplay of innovation, environmental orientation, and digital marketing in culinary SMEs,” *Mendeley Data*, version 1, 2025, doi: 10.17632/ngrp67bpt4.1.
- [12] J. Leksana, I. Gautama, S. B. Abidinagoro, L. Sanny, R. Kartono, and M. Hamsal, “ICTIM 2025 (2),” *Mendeley Data*, version 1, 2025, doi: 10.17632/nzv5rrk45f.1.
- [13] G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010. Available: <https://jmlr.org/papers/v11/cawley10a.html>.
- [14] J. Pineau et al., “Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program),” *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, 2021. Available: <https://jmlr.org/papers/v22/20-303.html>.
- [15] A. D’Amour et al., “Underspecification presents challenges for credibility in modern machine learning,” *Journal of Machine Learning Research*, vol. 23, no. 226, pp. 1–61, 2022. Available: <https://jmlr.org/papers/v23/20-1335.html>.
- [16] D. R. Roberts et al., “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
- [17] B.-S. Einbinder, S. Feldman, S. Bates, A. N. Angelopoulos, A. Gendler, and Y. Romano, “Label noise robustness of conformal prediction,” *Journal of Machine Learning Research*, vol. 25, no. 328, pp. 1–66, 2024. Available: <https://jmlr.org/papers/v25/23-1549.html>.
- [18] J. Clarkson, W. Xu, M. Cucuringu, and G. Reinert, “Split conformal prediction under data contamination,” *Proceedings of Machine Learning Research*, vol. 230, pp. 5–27, 2024. Available: <https://proceedings.mlr.press/v230/clarkson24a.html>.
- [19] R. F. Barber, E. J. Candès, A. Ramdas, R. J. Tibshirani, “Conformal prediction beyond exchangeability,” *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023, doi: 10.1214/23-AOS2276.
- [20] V. Franc, D. Prusa, and V. Voracek, “Optimal strategies for reject option classifiers,” *Journal of Machine Learning Research*, vol. 24, no. 11, pp. 1–49, 2023. Available: <https://jmlr.org/papers/v24/21-0048.html>.
- [21] S. Bates, A. Angelopoulos, L. Lei, J. Malik, M. Jordan, “Distribution-free, risk-controlling prediction sets,” *Journal of the ACM*, vol. 68, no. 6, pp. 1–34, 2021, doi: 10.1145/3478535.
- [22] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, “Conformal risk control,” *International Conference on Learning Representations*, 2024. Available: <https://arxiv.org/abs/2208.02814>.
- [23] Y. Bao, Y. Huo, H. Ren, C. Zou, “Selective conformal inference with false coverage-statement rate control,” *Biometrika*, vol. 111, no. 3, pp. 727–742, 2024, doi: 10.1093/biomet/asae010.
- [24] Y. Bao, Y. Huo, H. Ren, and C. Zou, “CAP: a general algorithm for online selective conformal prediction with FCR control,” *Journal of Machine Learning Research*, vol. 26, no. 287, pp. 1–74, 2025. Available: <https://jmlr.org/papers/v26/24-0452.html>.
- [25] L. Grinsztajn, E. Oyallon, G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?,” *Advances in Neural Information Processing Systems* 35, pp. 507–520, 2022, doi: 10.52202/068431-0037.
- [26] N. Hollmann et al., “Accurate predictions on small data with a tabular foundation model,” *Nature*, vol. 637, no. 8045, pp. 319–326, 2025, doi: 10.1038/s41586-024-08328-6.
- [27] A. E. Hoerl, R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.
- [28] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [29] F. Pedregosa et al., “Scikit-learn: machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>.
- [30] G. Varoquaux, “Cross-validation failure: small sample sizes lead to large error bars,” *NeuroImage*, vol. 180, pp. 68–77, 2018, doi: 10.1016/j.neuroimage.2017.06.061.
- [31] Y. Bengio and Y. Grandvalet, “No unbiased estimator of the variance of K-fold cross-validation,” *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004. Available: <https://jmlr.org/papers/v5/grandvalet04a.html>.
- [32] C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Machine Learning*, vol. 52, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.