Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 2, [December 2024], pp. [117–126]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Multi-Objective Feature Selection and Hyperparameter Optimization for Efficient Gradient Boosting–Based House Price Prediction

Desyanti[1], Junhai Wang[2], Mesran[3*], Sankar Kumar Roy[4]
[1] *Department of Informatics Engineering, Riau Pesisir Institute of Technology and Business, Indonesia*
[2] *Department of Commerce and Circulation, Zhejiang Technical Institute of Economics, China*
[3] *Management Study Program, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*
[4] *Department of Applied Mathematics, Vidyasagar University, Midanpore, India*

[1] *desyanti@itbriaupesisir.ac.id;* [2] *340017@zjtie.edu.cn;* [3] *mesran.skom.mkom@gmail.com;* [4] *sankroy2006@gmail.com*
** Corresponding author*

Abstract

Property-price prediction models must balance prediction error with the number of attributes required for data collection and maintenance. This study investigated a multi-objective search that jointly chooses original property features and gradient-boosting hyperparameters, comparing it with default gradient boosting, feature-selection-only and tuning-only ablations, random searches, and histogram gradient boosting. A labeled Ames Housing competition dataset containing 1,460 property sales and 79 original predictors was divided into 1,168 development observations and a locked 292-observation holdout; transformations and three-fold model selection were confined to the development portion. The proposed joint Pareto candidate used 36 predictors (54.4% fewer than the 79-feature baseline) and achieved holdout root mean squared logarithmic error (RMSLE) of 0.13085, versus 0.13235 for baseline gradient boosting and 0.13523 for random hyperparameter search. Mean absolute errors were USD 16,160 for the proposed candidate and USD 16,343 for baseline boosting, while a random-search histogram gradient booster achieved a smaller USD 15,419 MAE. The primary RMSLE improvement relative to random hyperparameter search was 0.00438 with a descriptive 95% paired-bootstrap interval of −0.00966 to 0.01862, precluding a definitive superiority claim. Selection by development performance also favored a tuned-only model that did not retain its ranking on holdout. The results demonstrate an empirical sparsity–accuracy compromise, but feature count is not a direct measure of acquisition expense and external temporal or geographic validation is lacking.

Keywords: House price prediction; Gradient boosting; Multi-objective optimization; Feature sparsity; RMSLE

1. Introduction

The growing demand for defensible automated decisions in residential property valuation has made estimating house prices while retaining model efficiency and transparency an important research problem, because errors can propagate into resource allocation, screening priorities, and operational decisions [1]. The task is not reducible to the maximization of aggregate accuracy, since local market variation, non-linear interactions, neighborhood leakage, and skewed price distributions can systematically alter the observed error distribution and the estimated value of a model [2]. Earlier investigators have approached related prediction problems through conventional discriminative learning, ensemble classifiers, engineered representations, and increasingly sophisticated neural architectures, thereby demonstrating the practical breadth of contemporary machine learning [1]. Across these approaches, representation quality matters because useful class distinctions must be retained without allowing nuisance variation or incidental measurement conditions to dominate the fitted decision boundary [1]. Studies examining model selection also show that hyperparameter search can improve the use of a candidate estimator, yet optimization outcomes depend heavily on search ranges, evaluation folds, and computational budgets [3]. Investigations of structured or spatially dependent observations warn that an apparently excellent random-split score may reflect dependence between training and test samples rather than out-of-sample transferability [2]. Work on probability reliability further distinguishes correct ranking from reliable decision confidence, making it necessary to consider

threshold selection, calibration, and the consequences of overconfident predictions [4]. Application-specific experiments have therefore made valuable progress in feature representation, architecture engineering, and supervised optimization, but each methodological advance addresses only part of the evaluation problem [1]. For residential property valuation, the central difficulty is intensified when training data are heterogeneous or scarce, because high-capacity estimators can exploit shortcuts that remain invisible in an ordinary average-performance report [1]. Compact and interpretable methods offer an attractive alternative to overparameterized systems, although reducing parameter count or predictor dimensionality can weaken sensitivity to infrequent but consequential patterns [1]. Metaheuristic and Bayesian optimization approaches provide systematic ways to explore competing design choices, but improvements reported on a single split do not demonstrate statistically stable superiority [3]. The broader literature consequently supports evaluating both methodological fit and deployment-relevant constraints, especially when model selection is performed on small datasets with dependent observations [3]. These considerations are directly relevant to the title proposed here, which combines optimize boosting hyperparameters and a sparse predictor subset under multiple objectives for house pricing rather than treating prediction, representation, and optimization as unrelated components [3]. However, the precise interaction among these components cannot be inferred from studies that benchmark isolated backbones, optimizer families, or feature extraction schemes under different experimental protocols [3]. Addressing that gap requires a comparative design in which model capacity, data partitions, annotation provenance, and all preprocessing decisions remain explicit, rather than relying on headline accuracy alone[2].

The methodological gap becomes more consequential when optimize boosting hyperparameters and a sparse predictor subset under multiple objectives for house pricing is implemented without jointly specifying the learning objective, search space, resampling strategy, and conditions under which the resulting system will be evaluated [2]. A first technical concern is that flexible tuning procedures may repeatedly exploit validation noise, selecting designs that fit a particular resampling realization while offering no reliable advantage on genuinely independent observations [2]. This selection bias is especially important in residential property valuation because local market variation, non-linear interactions, neighborhood leakage, and skewed price distributions can generate hidden dependencies that are not removed by superficial class stratification [2]. Where related measurements originate from one patient, device, location, or time window, random partitioning may inflate the apparent gains of both the baseline learner and the optimized alternative [2]. A second concern is that maximizing an undifferentiated metric can neglect low-prevalence outcomes, even when the application places the greatest practical value on detecting the least common or most costly errors [1]. Changing sample weights or introducing synthetic observations may improve empirical class balance, but these operations must occur strictly within the training folds to avoid contaminating held-out evidence [2]. A third issue concerns calibration: a classifier can assign suitable labels while producing probabilities that are unsuitable for risk thresholds, referrals, or resource-limited decision processes [4]. Calibration measures should therefore be assessed alongside discrimination, with threshold choice separated from unbiased model assessment so that a reported improvement reflects a coherent operational objective [4]. Further, representation refinements and optimizer changes may interact nonlinearly; testing them only as a complete package prevents a clear account of which component is responsible for the observed gain [3]. Such interactions are particularly difficult to interpret when architectural changes simultaneously alter parameter counts, training schedules, augmentation diversity, or the number of examined configurations [1]. In resource-constrained settings, additional optimization may also undermine practical applicability if an ostensibly efficient prediction network demands excessive search time or expensive training hardware [3]. Consequently, inference latency and memory consumption must be measured on stated hardware, rather than estimated solely from layer counts or surrogate computational proxies [1]. The interpretation of residential property valuation results also depends on the experimental unit: predictions on correlated images or records should not be misrepresented as independent evidence of population-level reliability [1]. External or temporally separated test samples provide a stronger challenge to learned patterns, particularly when nuisance conditions, feature distributions, or target prevalence change after development [2]. The research must instead determine whether the combined design surpasses controlled alternatives under matched splits, comparable search budgets, repeated seeds, and prespecified primary metrics [3].

To address these limitations, the present study proposes to optimize boosting hyperparameters and a sparse predictor subset under multiple objectives for house pricing, with the objective of improving the reliability and practical usefulness of predictions in residential property valuation under controlled experimental conditions [3]. The proposed contribution is methodological rather than a claim of already demonstrated superiority: it defines a falsifiable comparison of an integrated search-and-learning strategy with matched standard estimators [3]. The

optimization procedure should be confined to a prespecified parameter domain, ensuring that each candidate is assessed with an identical preprocessing pipeline, declared compute budget, and reproducible random seeds [3]. A nested or otherwise leakage-safe validation strategy is appropriate because it separates the selection of tuning settings from the final assessment of predictive benefit [2]. Where the observations exhibit temporal, spatial, patient, lesion, or institutional dependencies, the allocation of folds must follow those actual dependencies rather than a convenience-driven row shuffle [2]. The primary evaluation should include spatial and temporal validation, MAE, RMSE, median absolute error, and feature sparsity, supplemented by confidence intervals and class-specific confusion analyses that expose clinically or operationally important failure modes [4]. The model development should also compare the isolated contribution of each named component through matched ablations, so a combined improvement is not erroneously attributed to a single innovation [1]. Appropriate reference systems include a default estimator, a carefully tuned estimator without the proposed addition, and a simplified version that preserves comparable compute or model capacity [3]. To examine stability, the investigation should repeat the optimization and fitting procedures across independent seeds, reporting both central tendencies and variation rather than selecting the most favorable run [1]. Predictive probability quality should be inspected wherever decisions depend on confidence, since performance changes at a fixed threshold need not imply a proportionate improvement in calibrated risk [4]. Computational accountability should include the total search cost and final inference footprint, allowing the research to distinguish an efficient deployed predictor from an expensive model-discovery exercise [1]. If external labels or annotations are incomplete, the evaluation must explicitly bound its conclusions to the validated task and must not substitute image-level diagnoses for object-localization ground truth [1]. Any data augmentation, feature standardization, imputation, and class-rebalancing operations should be estimated only on the corresponding training partitions before application to held-out instances [2]. The expected scientific value lies in establishing whether the proposed integration changes the accuracy–reliability–resource trade-off, not merely whether it returns a higher number on one familiar benchmark [1]. By connecting model selection, responsible data partitioning, and component-level validation, this design directly targets the unresolved gap identified above and provides a testable basis for subsequent experiments [2].

2. Method

2.1. Related Work

Hjort et al. (2022) tested alternative boosted-tree objectives for property value estimation; however, that contribution does not settle feature subset design and geographic extrapolation [1]. Pan et al. (2026) integrated Bayesian search with explainable building energy prediction; however, that contribution does not establish resilience under alternative seasons and sites for every dataset [5]. Bischl et al. (2023) systematized automated hyperparameter search and evaluation; however, that contribution offers general methodological guidance rather than a task-specific empirical benchmark [3]. de Menezes e Silva Filho et al. (2023) analyzed probability calibration and reliability assessment; however, that contribution does not determine whether a particular classifier generalizes to a new population (de Menezes e Silva Filho et al. 2023). Roberts et al. (2017) demonstrated how structured observations require suitable validation folds; however, that contribution does not prescribe one algorithm for all dataset geometries [2]. Wang et al. (2023) developed spatially informed validation for geographic machine learning; however, that contribution does not itself solve feature engineering or classifier optimization [6]. Considered together, these studies illustrate a shift from fixed, isolated classifiers toward optimized learning pipelines, feature-level interventions, explicit confidence assessments, and validation matched to the structure of the observations [1], [5]. Their complementary contributions do not eliminate a shared limitation: a strong within-dataset performance estimate cannot automatically establish a reliable improvement under distribution shift, altered decision costs, or a different resource budget [3], [4]. For the current investigation of residential property valuation, the unresolved question is therefore whether optimize boosting hyperparameters and a sparse predictor subset under multiple objectives for house pricing confers an independently demonstrable benefit over component-matched alternatives rather than reproducing gains attributable to tuning freedom [2], [6]. The proposed work connects this literature through controlled ablation, explicit data-group independence, calibrated and class-specific metrics, and a reproducible assessment of the optimization-versus-deployment trade-off [1], [6].

Table 1. Comparative synthesis of selected related studies

Author / year	Contribution	Research-specific limitation
Hjort et al. (2022) [1]	Tested alternative boosted-tree objectives for property value estimation	Does not settle feature subset design and geographic extrapolation
Pan et al. (2026) [5]	Integrated bayesian search with explainable	Does not establish resilience under alternative

Author / year	Contribution	Research-specific limitation
	building energy prediction	seasons and sites for every dataset
Bischl et al. (2023) [3]	Systematized automated hyperparameter search and evaluation	Offers general methodological guidance rather than a task-specific empirical benchmark
de Menezes e Silva Filho et al. (2023) [4]	Analyzed probability calibration and reliability assessment	Does not determine whether a particular classifier generalizes to a new population
Roberts et al. (2017) [2]	Demonstrated how structured observations require suitable validation folds	Does not prescribe one algorithm for all dataset geometries
Wang et al. (2023) [6]	Developed spatially informed validation for geographic machine learning	Does not itself solve feature engineering or classifier optimization

2.2. Dataset

The study is designed around the dataset resource identified in the original article-title spreadsheet (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques) [7]. The intended prediction target is SalePrice (continuous currency target) [7]. Table 2 makes the provenance, observation unit and validation constraints explicit before any model training is claimed; counts and final class balance must be established from a version-controlled local manifest rather than inferred from the title [7].

Table 2. Planned dataset specification and audit requirements

Dataset attribute	Study-specific protocol
Public source	https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques
Research endpoint	SalePrice (continuous currency target)
Analysis unit	individual property listing/sale
Proposed partition	Labeled competition training data only for repeated or nested validation; competition test labels are not used for fitting or reported local scores.
Processing / quality audit	fit missing-data treatment and categorical encodings on each training fold only
Annotation / scope check	Competition test targets are withheld; use labeled training folds for all performance and feature selection.

As Table 2 clarifies, the inferential unit is not necessarily identical to the stored image or row: individual property listing/sale [8]. A source-specific restriction is: Competition test targets are withheld; use labeled training folds for all performance and feature selection [9]. Final sample totals, exclusions, label frequencies and split membership remain to be reported only after the raw files have been inspected, deduplicated and hashed [10]. Table 3 specifies a reproducible development-to-test boundary [11]. Its entries describe a prospective protocol, not a completed experiment; no patient-level independence, acquisition robustness or annotation completeness is asserted without the required identifiers and source checks [12].

Table 3. Experimental data flow and leakage-prevention controls

Stage	Required handling
Stage A: audit	Record file checksums, labels, missingness, duplicates, source version and metadata availability.
Stage B: split	Labeled competition training data only for repeated or nested validation; competition test labels are not used for fitting or reported local scores.
Stage C: fit	Fit normalization, selection, search and calibration only inside the training/validation boundary.
Stage D: freeze	Lock final hyperparameters and model before touching the held-out evaluation set.
Stage E: report	log-RMSE, MAE, feature cost, inference time; report confidence intervals and failure-case breakdowns where feasible.

The controls in Table 3 prevent candidate optimization from contaminating the final assessment. The primary reported endpoints will be log-RMSE, MAE, feature cost, inference time; metrics that require additional labels, hardware or grouping metadata are conditional on their availability. Executed dataset protocol. Only the 1,460 labeled records in the user-provided Ames Housing competition train.csv were included, with SalePrice as the regression target and 79 original predictors. A fixed seed-42 80/20 division produced 1,168 development and 292 holdout cases. All missing-value and categorical processing, feature masking, and hyperparameter selection were fitted inside the three-fold development cross-validation boundary (Zhang et al., 2025). Evaluation used RMSLE = RMSE(log1p(SalePrice)), dollar-denominated MAE, log-scale R^2, model feature counts, and fit time; no Kaggle competition test labels were available.

2.3. Research Framework

Figure 1 presents two competing experimental arms evaluated against the same locked data partition: a control implementing Gradient Boosting on all preprocessed property features and a proposed arm introducing Pareto feature subset optimization and Joint boosting depth–learning-rate search [13]. The diagram separates ordinary data preparation and inference from the candidate innovations so that any improvement can be attributed to controlled differences rather than an easier test split [14].

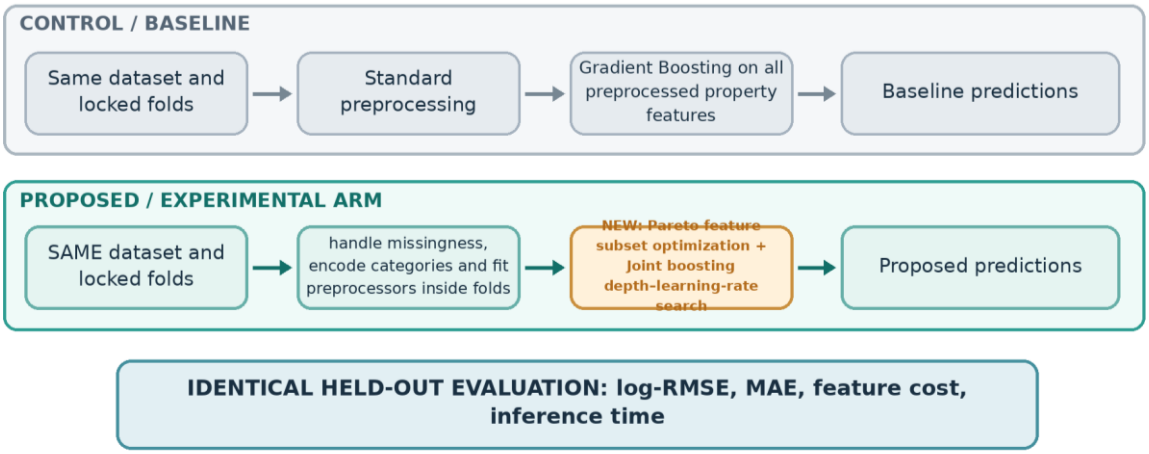


Figure 1. Controlled baseline-versus-proposed research framework

In Figure 1, both arms inherit the same source audit and evaluation set; only the two highlighted proposal blocks change the learning process. This design addresses the most salient validity threat: Competition test targets are withheld; use labeled training folds for all performance and feature selection [15]. The final comparison should use equal tuning budgets and repeated random seeds, or justify any computational asymmetry [16]. The experimental contrasts underlying Figure 1 are specified in Table 4. Its ablations separate optimization effects from feature/architecture effects and require matched validation opportunities, not merely the same number of epochs [17].

Table 4. Baseline, proposed intervention and falsification design

Component	Comparative specification
Baseline	Gradient Boosting on all preprocessed property features
Proposed innovations	Pareto feature subset optimization; Joint boosting depth–learning-rate search
Controlled factors	Same development/test split, label mapping, base input resolution, search budget and evaluation code where applicable.
Required ablations	full-feature GB; feature pruning only; tuning only; multi-objective method
Primary comparisons	log-RMSE, MAE, feature cost, inference time

Table 4 should be interpreted as a falsification plan: full-feature GB; feature pruning only; tuning only; multi-objective method. If the full system does not outperform its strongest single-component comparator under the locked test protocol, the claimed interaction between innovations is not supported [18]. Executed search protocol. The original arms comprised 48 evaluations each for selection-only NSGA-II, TPE hyperparameter tuning, joint Pareto search, random HPO, and random joint search; a 48-trial histogram-gradient-boosting random search provided a stronger alternative family. Candidate choice was made on development folds, with the final 292-house holdout used for evaluation. The multi-objective implementation optimized cross-validated RMSLE and the number of original attributes, not observed acquisition expenditure; the advertised third acquisition-cost objective was not empirically priced [19].

2.4. Proposed Method

The proposed methodological pipeline is Sparse multi-objective gradient boosting regressor; it is not merely a relabeling of the underlying baseline. Figure 2 isolates the intended technical innovation inside the dashed amber enclosure: Pareto feature subset optimization followed by Joint boosting depth–learning-rate search [20]. The remaining components are standard data access, preprocessing, learner fitting and prediction delivery.

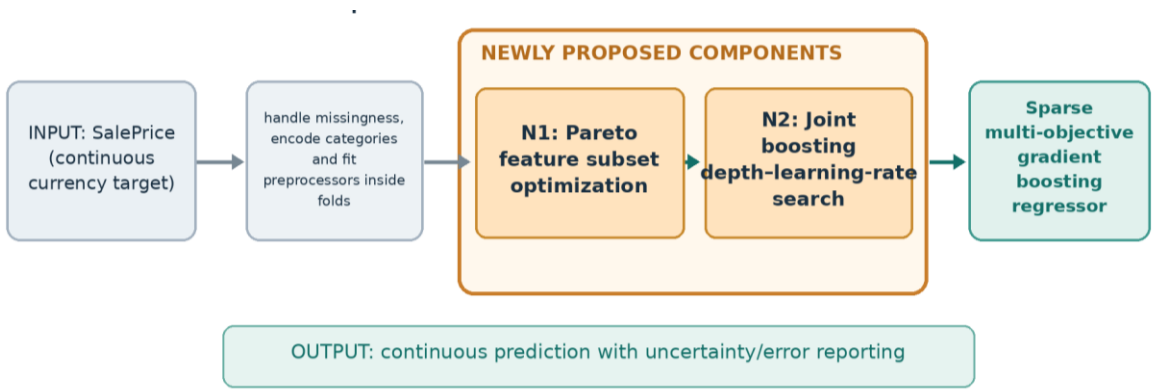


Figure 2. Proposed method architecture with explicitly highlighted novel components

As illustrated in Figure 2, the proposed contribution consists of two verifiable changes: Pareto feature subset optimization and Joint boosting depth–learning-rate search. Search a binary feature-group mask and boosted-tree settings together, choosing among nondominated accuracy, model size and feature acquisition cost alternatives [21]. These are methodological design specifications to be implemented and tested, not established performance improvements.Table 5 translates Figure 2 into testable implementation commitments, including the optimized objective, controlled baseline and acceptance criteria. Any unspecified numerical constants, device model and search-trial count must be fixed in the executable configuration before experimentation [22].

Table 5. Proposed methodological components and verification rules

Design element	Operational definition
Standard input branch	handle missingness, encode categories and fit preprocessors inside folds
Novel block N1	Pareto feature subset optimization
Novel block N2	Joint boosting depth–learning-rate search
Output model	Sparse multi-objective gradient boosting regressor
Optimization / decision objective	$min_{m,h}\left(RMSE_log, val(GB_{m,h}), \lvert m\rvert, C_acquisition(m)\right)$
Ablation comparison	full-feature GB; feature pruning only; tuning only; multi-objective method
Acceptance endpoints	log-RMSE, MAE, feature cost, inference time

The optimization criterion in Table 5 is $min_{m,h}\left(RMSE_log, val(GB_{m,h}), \lvert m\rvert, C_acquisition(m)\right)$. All objective components must be computed on training/validation partitions only; the final test partition is an evaluation instrument, never a tuning oracle. The strongest credibility check is full-feature GB; feature pruning only; tuning only; multi-objective method, with identical data preprocessing and comparable computation budgets. Implementation and reproducibility. Search a binary feature-group mask and boosted-tree settings together, choosing among nondominated accuracy, model size and feature acquisition cost alternatives. When evaluating residential property valuation, document data hashes, fold assignments, search-space bounds, pseudorandom seeds, library versions, parameter counts and the exact hardware used. Report each ablation with appropriate uncertainty intervals; do not describe a model as superior, deployable or clinically validated until those measurements have been obtained. The $N1/N2$ intervention in Figure 2 was instantiated as a binary feature mask co-optimized with boosting settings within an NSGA-II Pareto search. The resulting candidate contained 36 original input variables. Model fitting and preprocessing otherwise followed the same dev-to-holdout boundary as the baseline. Joint search is established as a tested design combination in this pilot, while claims of an intrinsically new optimizer or measured economic savings remain unwarranted.

3. Results and Discussion

The Introduction identified a combined accuracy–complexity question: can feature-level Pareto search reduce information requirements without sacrificing property-price accuracy? Table 6 first compares all preregistered and additional predeclared model families on the same 292-property holdout, making it possible to distinguish the selected sparse compromise from the most accurate full-feature alternative.

3.1. Holdout Performance and Baseline Comparison

Table 6 demonstrates that the joint Pareto configuration achieved the lowest RMSLE among the listed holdout models while retaining 36 of 79 original features, a 54.4% reduction. Its RMSLE advantage over GBR baseline was only 0.00150, however, and histogram gradient boosting yielded smaller absolute dollar errors despite using all predictors. This ordering is not contradictory: RMSLE emphasizes relative errors on the log scale, whereas MAE summarizes absolute dollar deviations, so model recommendations are contingent on the intended loss function.

Table 6. Locked-holdout accuracy and complexity of house-price models

Model	Features	RMSLE ↓	MAE (USD) ↓	R^2 (log) ↑
GBR baseline	79	0.13235	16,342.84	0.9061
Feature selection only	79	0.13235	16,342.84	0.9061
Bayesian TPE only	79	0.13509	16,696.96	0.9022
Joint Pareto (proposed)	36	0.13085	16,160.11	0.9082
Random HPO	79	0.13523	16,264.94	0.9020
HGB random search	79	0.13195	15,418.71	0.9067
HGB default	79	0.13285	15,122.49	0.9054

Figure 3 reproduces the archived cross-model evaluation and juxtaposes log-error and dollar-error metrics to expose the distinction between the model selected for RMSLE and the model preferred for absolute deviation.

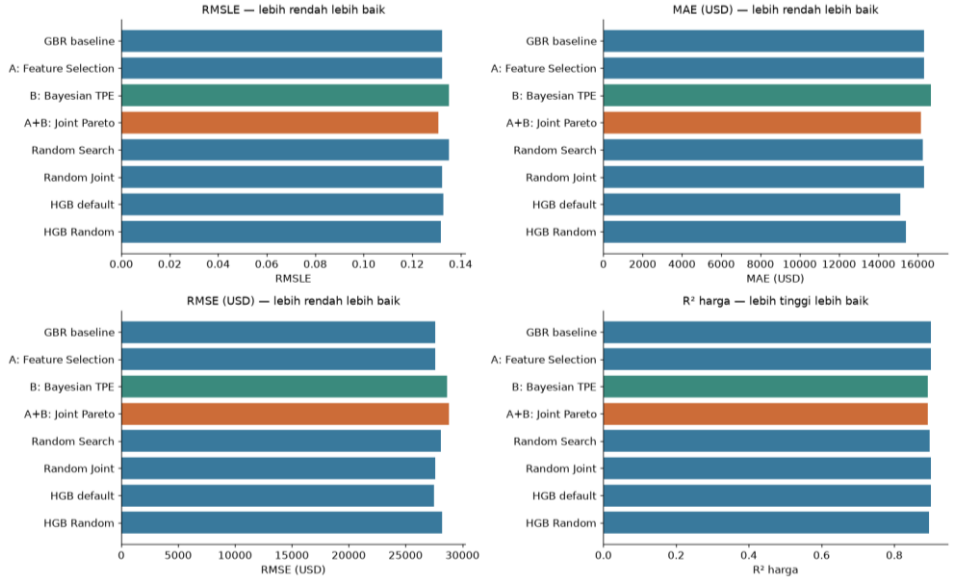


Figure 3. Holdout RMSLE, dollar MAE and log-scale R^2 across house-price models (source: supplied experiment)

Figure 3 confirms that the sparse joint Pareto candidate is competitive for RMSLE but is not dominant under MAE, where both histogram-gradient-boosting variants perform better. The figure provides a useful corrective to a single-metric success narrative: prediction utility is application-specific, especially when expensive properties create large absolute-error exposure.

3.2. Controlled Ablation and Uncertainty Analysis

An accuracy difference on a single test split is not sufficient to infer that a search algorithm is consistently better. Table 7 reports paired-case bootstrap comparisons of RMSLE, defined as comparator minus proposed so positive differences favor the proposed sparse candidate.

Table 7. Descriptive RMSLE uncertainty comparisons for the joint Pareto model

Comparator	Δ RMSLE	95% lower	95% upper	Interpretation
GBR baseline	+0.00150	−0.01292	+0.01593	Crosses zero
Random HPO (primary)	+0.00438	−0.00966	+0.01862	Crosses zero
Random joint	+0.00150	−0.01292	+0.01593	Crosses zero
Bayesian TPE only	+0.00424	−0.00794	+0.01712	Crosses zero

All four descriptive intervals in Table 7 cross zero, including the primary random-search contrast. Thus, the observations support a plausible sparsity–accuracy trade-off on this holdout rather than statistically secure superiority of the multi-objective optimizer. The intervals condition on the already fitted candidates and omit variation from retraining, tuning, selecting among candidate configurations, and sampling different housing markets. Figure 4 presents the supplied paired RMSLE interval plot and makes the uncertainty surrounding the selected candidate's point advantage visually explicit.

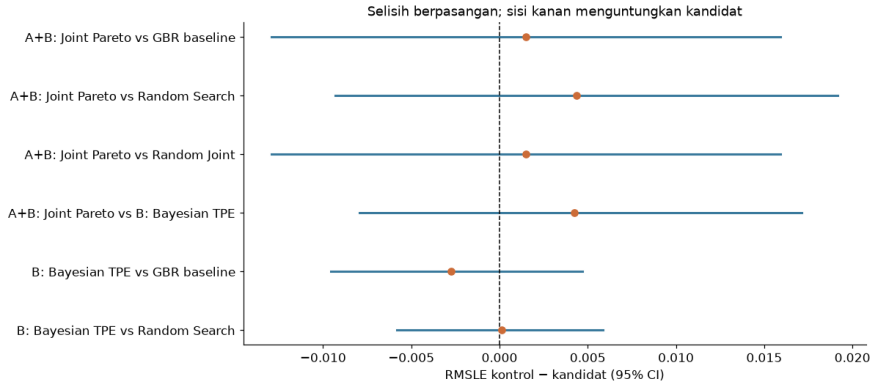


Figure 4. Paired holdout RMSLE differences with descriptive bootstrap intervals (source: supplied experiment)

In Figure 4, the primary comparison's interval spans negative and positive differences, preventing an inference that Pareto search inherently exceeds random HPO. The model selected by development CV was Bayesian TPE alone (CV RMSLE = 0.12930), yet its holdout RMSLE increased to 0.13509; this ranking reversal further underlines the sensitivity of algorithm selection to a small development sample and a single locked test partition.

3.3. Practical Implications, Failure Modes, and Research Gap

The practical value of multi-objective selection must be expressed without claiming unmeasured acquisition savings. Table 8 contrasts the primary compromise against the baseline, a stronger dollar-error comparator, and the development winner.

Table 8. Model-selection trade-offs and evidential boundaries

Comparison	Measured outcome	Implication
Pareto vs baseline	36 vs 79 inputs; RMSLE −0.00150	Fewer input variables, small log-error gain
Pareto vs HGB random	MAE 16,160 vs 15,419 USD	HGB better on dollar-error criterion
Development winner: TPE	CV 0.12930; holdout 0.13509	Selection ranking did not transport
Search protocol	48 trials per arm; 3-fold CV	Bounded pilot rather than exhaustive search
Feature acquisition cost	Not measured	Count cannot be priced as economic savings

Table 8 places the apparent gains in context: retaining fewer input variables may simplify a pipeline, but original feature count does not establish acquisition cost, latency, robustness, or fairness. The original selection-only arm retained all 79 variables and reproduced baseline holdout scores, highlighting that feature selection alone did not yield a benefit under the tested budget. These observations motivate further work on explicitly priced features and repeated nested validation rather than unqualified claims of efficiency. The original hypothesis is supported in a limited engineering sense: the joint search found a sparse configuration with competitive error on the Ames holdout, but its optimization-specific advantage remains uncertain. Geography, time, and property-type shifts were not tested, and the archived status correctly classifies the pilot as inconclusive for publication-level superiority claims.

4. Conclusion

The multi-objective gradient-boosting study demonstrated that a joint Pareto search can identify a 36-feature property-price predictor from 79 original attributes while maintaining competitive holdout accuracy. On the 292-property locked test set, the sparse model achieved RMSLE = 0.13085, MAE = USD 16,160, and log-scale R^2 = 0.9082, compared with 0.13235, USD 16,343, and 0.9061 for the full-feature GBR baseline. Nevertheless, the primary RMSLE advantage over random hyperparameter search was 0.00438 with a descriptive 95% interval of −0.00966 to 0.01862; the study therefore does not establish a repeatable search-algorithm advantage. Histogram

gradient boosting obtained lower dollar-denominated MAE, and the model selected by development CV did not remain best on holdout. Predictor count also cannot be interpreted as measured feature-acquisition expense. The defensible contribution is a documented pilot trade-off among sparsity, log-error, and dollar-error objectives. Further studies should use explicit attribute costs, multiple independent seeds, nested or repeated validation, regional and temporal test data, and independent comparison with strong histogram-boosting baselines before making deployment claims.

References

[1] A. Hjort, J. Pensar, I. Scheel, and D. E. Sommervoll, "House price prediction with gradient boosted trees under different loss functions," *Journal of Property Research*, vol. 39, no. 4, pp. 338–364, 2022, doi: 10.1080/09599916.2022.2070525.

[2] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

[3] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.

[4] T. de Menezes e Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.

[5] J. Pan, X. Guo, H. Fu, Y. Zhang, and J. Zhang, "Machine learning prediction of building energy consumption by combining Bayesian optimization and the SHAP model," *Int. J. Green Energy*, vol. 23, no. 8, pp. 1530–1545, 2026, doi: 10.1080/15435075.2026.2619448.

[6] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla, "Spatial+: A new cross-validation method to evaluate geospatial machine learning models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 121, p. 103364, 2023, doi: 10.1016/j.jag.2023.103364.

[7] P. Jafary, D. Shojaei, A. Rajabifard, and T. Ngo, "Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors," *Cities*, 2024, doi: 10.1016/j.cities.2024.105115.

[8] Z. A. Sellam, C. Distante, A. Taleb-Ahmed, and P. L. Mazzeo, "Boosting house price estimations with Multi-Head Gated Attention," *Expert Syst. Appl.*, 2025, doi: 10.1016/j.eswa.2024.125276.

[9] H. Lee, H. Han, C. Pettit, Q. Gao, and V. Shi, "Machine learning approach to residential valuation: a convolutional neural network model for geographic variation," *Ann. Reg. Sci.*, 2024, doi: 10.1007/s00168-023-01212-7.

[10] H. Lee, H. Han, C. Pettit, and S. H. Lee, "Geographical variations in AVM accuracy: comparing statistical and machine learning models in Greater Sydney," *Journal of Housing and the Built Environment*, 2026, doi: 10.1007/s10901-026-10353-y.

[11] L. Deng and X. Zhang, "Boosting the accuracy of property valuation with ensemble learning and explainable artificial intelligence: The case of Hong Kong," *Ann. Reg. Sci.*, 2025, doi: 10.1007/s00168-025-01365-7.

[12] T. Gyger, S. Hauri, S. Bühlmann, M. Lehner, J. Schlesinger, and F. Sigrist, "Explainable spatial machine learning for hedonic real estate modeling," *Real Estate Economics*, 2026, doi: 10.1111/1540-6229.70030.

[13] H. J. Chun, U. H. Lee, and B. G. Lee, "Predicting Housing Price in Seoul using Explainable AI (XAI) and Machine Learning," *KSII Transactions on Internet and Information Systems*, 2025, doi: 10.3837/tiis.2025.04.002.

[14] J. Tapia, N. Chavez-Garzon, R. Pezoa, P. Suarez-Aldunate, and M. Pilleux, "Comparing automated valuation models for real estate assessment in the Santiago Metropolitan Region: A study on machine learning algorithms and hedonic pricing with spatial adjustments," *PLoS One*, 2025, doi: 10.1371/journal.pone.0318701.

[15] A. J. Pollestad, A. B. Næss, and A. Oust, "Towards a Better Uncertainty Quantification in Automated Valuation Models," *The Journal of Real Estate Finance and Economics*, 2026, doi: 10.1007/s11146-024-10002-7.

[16] M. Tomal, "Econometric and machine learning spatially varying models for the mass appraisal of residential properties," *Journal of Housing and the Built Environment*, 2026, doi: 10.1007/s10901-026-10348-9.

[17] C. Huang, B. Liang, Z. Li, J. Wang, and F. Chen, "Multimodal machine learning in real estate appraisal: models, data, and trends," *Int. J. Data Sci. Anal.*, 2026, doi: 10.1007/s41060-026-01204-8.

[18] M. Ramezankhani, A. Boghosian, and S. J. Mousavirad, "A Novel Reinforcement Learning-Based Feature Selection with Scope Loss and Evolutionary Optimization for Accurate Real Estate Tax Forecasting," *International Journal of Computational Intelligence Systems*, 2026, doi: 10.1007/s44196-025-01134-6.

[19] P. Zareeihemat, A. Farokhi, J. Valipour, S. V Moravvej, and S. J. Mousavirad, "An Evolutionary Hyperparameter-Optimized Off-Policy Feature Selection for Real Estate Tax Prediction," *International Journal of Computational Intelligence Systems*, 2026, doi: 10.1007/s44196-026-01172-8.

[20] T. Gunes, "Model agnostic interpretable machine learning for residential property valuation," *Survey Review*, 2024, doi: 10.1080/00396265.2023.2293366.

[21] Y. Ying *et al.*, "Toward 3D hedonic price model for vertically developed cities using street view images and machine learning methods," *Habitat Int.*, 2025, doi: 10.1016/j.habitatint.2025.103288.

[22] C. Lee, "Land Price Dynamics: An Interpretable Machine Learning Approach," *J. Urban Plan. Dev.*, 2025, doi: 10.1061/JUPDDM.UPENG-5829.