

RESEARCH ARTICLE

Explainable Artificial Intelligence for Retail Price Estimation and Pricing-Risk Detection Using Multivariate Product Attributes

Saumen Bandyopadhyay ^{a,*}^a*Baxter International Inc., Baxter Parkway, Deerfield, Illinois 60015, United States of America.***Abstract**

Retail price estimation for heterogeneous products requires models that represent nonlinear attribute interactions without concealing the basis of individual estimates. This study develops an Explainable Artificial Intelligence (XAI) system for reference-price estimation and pricing-risk screening using a real 53,940-record diamond transaction dataset containing carat, cut, colour, clarity, geometric measurements, and observed price. Twenty records with non-positive physical dimensions were removed, leaving 53,920 observations. To reduce analogue leakage, records were grouped by cut, colour, clarity, and 0.05-carat specification buckets before three-fold grouped cross-validation. Decision Tree, Random Forest, Extra Trees, histogram gradient boosting, and XGBoost regressors were evaluated on out-of-fold predictions after logarithmic target transformation. Extra Trees produced the lowest error, with MAE of USD 280.96, RMSE of USD 549.90, MAPE of 7.18%, and $R^2 = 0.9810$. XGBoost achieved RMSE of USD 596.87 and $R^2 = 0.9776$ and was retained as the explanation engine because exact TreeSHAP values are available for tree ensembles. SHAP analysis showed that geometric size variables and carat dominated the learned reference-price function, while clarity and colour provided secondary contributions. A residual-based audit layer translated reference-price deviations into controllable review workloads; a 20% deviation threshold flagged 2,825 records, or 5.24% of the cleaned dataset. The resulting system supports auditable price review while explicitly separating prediction, explanation, and managerial action.

Keywords: Retail price estimation, Explainable Artificial Intelligence (XAI), XGBoost, SHAP, Pricing risk.

Article information:

ISSN: 3107-9466 (Online)

Published by: **Krrish Scientific Publications Pvt. Ltd.**DOI: <https://doi.org/10.71426/jcdt.v1.si1.pp19-27>

Received: 13 Sep. 2025 | Revised: 12 Oct. 2025 | Accepted: 18 Oct. 2025 | Published: 22 Oct. 2025

Copyright ©2025 Author(s).

This is an open-access article distributed under the terms of the [Creative Commons Attribution-NonCommercial 4.0 International License \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/).**1. Introduction**

Retail pricing for differentiated goods is an inference problem over product characteristics, market context, and observed transactions rather than a simple markup calculation. Hedonic price theory formalized the idea that the price of a heterogeneous good can be represented as an equilibrium function of its measurable characteristics [1]. Subsequent work clarified that the estimated relationship is meaningful only when the chosen attributes pertain to the good and the functional form can represent the underlying characteristics frontier [2]. In operational retail systems, these principles are encountered in a more computational form: product-master data describe categorical quality grades, inspection or measurement systems provide continuous physical variables, and transaction systems

record realized prices. The analytical task is therefore naturally multisource even when the present benchmark is distributed as a single consolidated table.

Product-attribute valuation has long been used to support commercial decisions. Tomkovick and Dobie demonstrated that hedonic pricing and factorial surveys can inform product design and price sensitivity analysis [3], while Roheim, Gardiner, and Asche used retail scanner data to quantify the contribution of brand, species, package size, and processing attributes to frozen-fish prices [4]. Martínez-Garmendia further showed how scanner-data hedonic models can be adapted for consumer packaged goods, emphasizing that attribute valuation depends on the way observed transactions are represented [5]. These studies are valuable because they establish a disciplined connection between product characteristics and price; however, conventional linear or prespecified econometric forms can be restrictive when interactions, threshold effects, and nonlinear size relationships dominate.

*Corresponding author

Email address: saumen89b@gmail.com (Saumen Bandyopadhyay).

Diamonds provide an appropriate test domain because their prices depend on interacting physical and quality attributes. Carat, cut, colour, and clarity are central grading variables, but physical dimensions and proportions also encode aspects of size and geometry. Lee, Caudill, and Mixon showed that grading-related information can carry systematic premiums or discounts in retail diamond markets [6]. In a broader retail setting, Ferreira, Lee, and Simchi-Levi demonstrated that machine-learning forecasts can be integrated with operational pricing decisions, while also showing that the prediction layer and the optimization or policy layer are distinct components [7]. More recently, Kigo, Omondi, and Omolo compared supervised learning methods for diamond pricing and documented the strong predictive potential of nonlinear regressors on this domain [8]. These results motivate modern ensemble learning, but high predictive accuracy alone is insufficient for a price-review system used by analysts or managers.

The principal difficulty is governance. A price estimate can be numerically accurate yet operationally unhelpful if the analyst cannot determine which attributes drove the estimate or whether a large residual indicates a genuine pricing issue, an unusual but legitimate product, or an omitted market factor. XGBoost offers regularized tree boosting with scalable optimization [9], and TreeSHAP provides additive feature attributions for tree ensembles that connect local predictions with global feature effects [10]. Nevertheless, SHAP values are explanations of a fitted model, not causal estimates of what would happen if a merchant changed a product attribute or posted a different price. An auditable retail system must preserve this distinction.

A second difficulty is evaluation leakage. Product catalogues often contain many near-duplicate specifications. Randomly splitting individual rows can allow almost identical products into both training and test folds, producing an optimistic estimate of generalization. This problem is particularly relevant when product identity is summarized by a small set of highly structured attributes. Accordingly, the present study evaluates models with grouped cross-validation in which records sharing cut, colour, clarity, and a 0.05-carat bucket are assigned to the same fold. The test therefore asks whether a model can interpolate across related but non-identical specification groups rather than memorize close analogues.

This paper makes four contributions. First, it formulates a reproducible reference-price pipeline on a real 53,940-record retail-product dataset and applies explicit physical-validity cleaning without fabricating or augmenting observations. Second, it compares five nonlinear regression models under grouped out-of-fold evaluation and reports MAE, RMSE, MAPE, median absolute percentage error, and R^2 rather than relying on a single favorable statistic. Third, it separates the best predictive model from the explanation model: Extra Trees is selected by RMSE, whereas XGBoost is retained for exact tree-based SHAP analysis, avoiding the common but unjustified assumption that the explainable model must also rank first. Fourth, it defines a transparent residual-based pricing-risk screen that converts prediction deviations into an adjustable human-review workload. Recommendations are framed as audit prompts—for example, review possible overpricing or margin leakage—rather than

automated causal prescriptions. The resulting architecture links multisource-style product attributes, nonlinear price estimation, explanation, and controlled intervention while maintaining a clear boundary between statistical evidence and managerial action.

2. Literature Review

2.1. Attribute-based pricing and retail analytics

Hedonic analysis models differentiated products as bundles of characteristics. Rosen's equilibrium formulation [1] remains the conceptual basis for decomposing observed prices into implicit attribute valuations, while Edmonds [2] emphasized the consequences of functional-form choice and attribute selection. The retail literature subsequently extended these ideas from durable goods to consumer-product and scanner-data settings. The Parker Pen study [3] combined hedonic analysis with factorial surveys to connect observed market valuations with new-product design. Roheim *et al.* [4] demonstrated that brand, product form, and package size can be monetized from retail transactions, and Martínez-Garmendia [5] showed that quantity-aware scanner-data formulations can alter interpretation of hedonic coefficients. These studies establish two requirements that remain relevant to machine learning: the input representation must reflect defensible product attributes, and predictive associations should not automatically be interpreted as causal willingness-to-pay effects.

Diamond pricing makes this distinction concrete. Lee *et al.* [6] studied price premiums associated with grading laboratories in a large sample of online offerings, demonstrating that price formation includes both physical attributes and institutional signals. Kigo *et al.* [8] reported that modern supervised learners can model nonlinear diamond-price relationships effectively. The present study differs in three respects. It evaluates near-duplicate leakage through grouped splitting; it distinguishes the lowest-error ensemble from the model used for explanations; and it adds a residual-based review policy rather than ending with point prediction. The policy perspective is consistent with operational retail research such as Ferreira *et al.* [7], where predictive analytics supports but does not replace the subsequent pricing decision.

2.2. Tree ensembles for nonlinear price functions

Tree ensembles are well suited to product-pricing data because they represent threshold effects and interactions without imposing a global polynomial form. Breiman's Random Forest combines decorrelated randomized trees to reduce variance relative to an individual decision tree [11]. Extremely randomized trees increase split randomization and can provide additional variance reduction when enough observations are available [12]. Gradient boosting instead constructs an additive function sequentially by fitting learners to the current optimization direction [13]. XGBoost extends this principle through regularized objectives, efficient split search, subsampling, and systems-level optimizations [9]. Hyperparameters such as tree depth, number of estimators, feature subsampling, and minimum leaf size can materially affect ensemble behavior, as reviewed by Probst, Wright, and Boulesteix [14].

These model families expose a practical trade-off. Random Forest and Extra Trees often provide strong out-of-sample regression with relatively stable defaults, whereas boosting methods can produce compact high-performing functions but require more careful regularization. A price-estimation study should therefore compare model families under an identical evaluation protocol rather than select XGBoost solely because it is popular. In the present experiment, the difference is substantive: Extra Trees achieves the best RMSE, while XGBoost remains competitive but not first. This negative comparative finding is retained because the objective is a defensible system, not a predetermined algorithm ranking.

2.3. Explainability for pricing decisions

Post-hoc explanation methods attempt to expose how complex predictors use their inputs. Ribeiro, Singh, and Guestrin introduced LIME as a model-agnostic local surrogate approach [15]. Broader XAI taxonomies distinguish local from global explanations, intrinsic from post-hoc interpretability, and feature-attribution methods from rule or example-based approaches [16]. Guidotti *et al.* surveyed black-box explanation methods and emphasized that explanation quality is multidimensional rather than reducible to a single fidelity notion [17]. For regression surfaces, individual conditional expectation plots show heterogeneity around average partial effects [18], while accumulated local effects can reduce extrapolation problems that occur when predictors are correlated [19].

TreeSHAP is particularly attractive for a tree-based pricing model because it computes additive feature attributions efficiently and consistently for tree ensembles [10]. If $f(x)$ is a prediction, SHAP expresses it as a baseline plus feature contributions. This form is convenient for audit because an analyst can see which product attributes push a reference price upward or downward for a particular record. Yet attribution is not intervention. With strongly correlated size variables such as carat, x , y , and z , importance can be distributed among related measurements. A large SHAP value for a dimension therefore means that the fitted model relied on that variable in the observed data distribution; it does not mean that physically changing that dimension while holding everything else fixed would produce the same monetary effect. Benchmarking work on explanation methods reinforces the need to evaluate XAI in relation to the model, task, and user question rather than to treat explanations as universally interchangeable [20].

2.4. Evaluation, error metrics, and leakage control

Regression metrics describe different aspects of predictive error. MAE retains the units of the target and provides a direct measure of absolute discrepancy; Willmott and Matsuura argued for its interpretive advantages when an average error magnitude is required [21]. R^2 captures explained variation and complements absolute-error metrics; Chicco, Warrens, and Jurman discuss its value alongside MAE, MAPE, MSE, and RMSE [22]. RMSE remains useful in pricing because its quadratic penalty gives more weight to large dollar errors that may be operationally expensive,

but it should not be interpreted as an average absolute error.

Cross-validation introduces another source of uncertainty. Nadeau and Bengio showed that repeated use of training sets induces dependencies that complicate inference from cross-validation estimates [23]. Varma and Simon demonstrated that using the same cross-validation process for model selection and unbiased error estimation can produce optimistic results [24]. The present work does not conduct an extensive hyperparameter search; instead it fixes conservative model configurations before the reported grouped evaluation. More importantly, grouping is defined from product specifications before model fitting. This prevents a close analogue of the held-out product from appearing in the training fold solely because individual rows were randomly shuffled. The resulting protocol is still an interpolation benchmark, not a temporal or cross-retailer validation, and that limitation is reported explicitly.

The literature therefore leaves a specific gap. Hedonic pricing studies provide interpretable attribute valuation but may constrain nonlinear functional form; machine-learning studies improve prediction but can overlook near-duplicate leakage or treat accuracy as sufficient; and XAI work provides feature attributions without specifying how explanations should enter a governed pricing workflow. This study addresses that intersection by combining grouped predictive evaluation, separate best-model and explanation-model roles, SHAP diagnostics, and an adjustable pricing-risk audit layer on a fully reproducible real dataset.

3. Methodology

3.1. Dataset, feature domains, and cleaning

The experiment uses the public `diamonds` table distributed in the `ggplot2/plotnine` data ecosystem. The local snapshot contains 53,940 observations and ten columns: carat, cut, colour, clarity, depth, table, price, and physical dimensions x , y , and z . The data are treated as a real observational retail-product benchmark, not as synthetic transactions. Consistent with a deployable enterprise architecture, the columns are organized into three logical source domains: product-master quality attributes (cut, colour, clarity), inspection/measurement attributes (carat, depth, table, x , y , z), and transaction-ledger outcome (price). This logical decomposition describes how such fields commonly arrive in retail systems; it does not claim that the benchmark itself was physically assembled from three independent source systems.

Twenty rows contain a non-positive value in at least one physical dimension. These records are excluded by a deterministic validity rule before any split or model fitting, leaving $N = 53,920$ observations. No price outliers are deleted. Cut, colour, and clarity are mapped to ordinal quality ranks using their conventional ordered levels. The final predictor vector is

$$\mathbf{x}_i = [c_i, q_i^{cut}, q_i^{color}, q_i^{clarity}, d_i, t_i, x_i, y_i, z_i]^\top, \quad (1)$$

where c is carat, d is depth percentage, and t is table percentage. Price is transformed as $y_i = \log(1 + p_i)$ during fitting and inverse-transformed for all reported dollar metrics.

Table 1: Dataset and evaluation summary

Property	Value
Raw observations	53,940
Clean observations	53,920
Removed invalid-dimension rows	20 (0.037%)
Predictor variables	9
Observed price range	USD 326–18,823
Median observed price	USD 2,401
Specification groups	5,603
Cross-validation	3-fold GroupKFold
Random seed	42

3.2. Grouped evaluation and candidate models

A product-specification group is defined as

$$g_i = (\text{cut}_i, \text{color}_i, \text{clarity}_i, \text{round}(\text{carat}_i/0.05)). \quad (2)$$

All rows with the same g_i are assigned to a single cross-validation fold. The grouping yields 5,603 groups. This is more conservative than an unconstrained random split because near-identical combinations of the four principal commercial grades cannot be directly shared between train and test partitions. It is not equivalent to a future-period evaluation because the dataset does not provide a transaction timestamp.

Five models are evaluated: a single decision tree, Random Forest, Extra Trees, histogram gradient boosting (HistGB), and XGBoost. The ensembles use fixed moderate configurations in the supplied script rather than an extensive search on the same folds. Random Forest and Extra Trees use 60 trees; HistGB uses 140 boosting iterations; XGBoost uses 180 estimators with depth 4, learning rate 0.045, row subsampling 0.9, column subsampling 0.9, and ℓ_2 regularization 2.0. Every metric is calculated from out-of-fold predictions so each reported prediction comes from a model that did not fit that record.

For XGBoost, the fitted function is an additive ensemble

$$\hat{y}_i = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \quad (3)$$

and training minimizes the regularized objective is given by (4).

$$\mathcal{L} = \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|_2^2. \quad (4)$$

The current implementation uses squared error on log-price with $\lambda = 2$. Equations (1)–(4) correspond to the feature engineering and model-training stages in Figure 1.

3.3. SHAP explanation and pricing-risk screen

After cross-validation, XGBoost is refit on the cleaned dataset only for explanation. TreeSHAP decomposes a model output as (5):

$$f(\mathbf{x}_i) = \phi_0 + \sum_{j=1}^M \phi_{ij}, \quad (5)$$

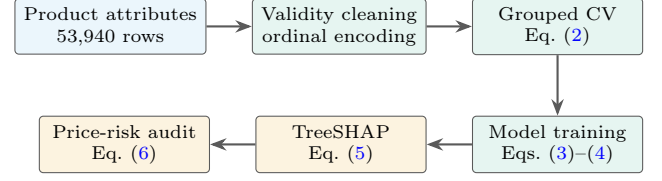


Figure 1: Leakage-controlled workflow. Prediction, explanation, and audit policy remain separate stages; arrows do not imply causal intervention.

Algorithm 1 Grouped price estimation, explanation, and audit

Raw product table D , threshold τ , seed s Remove rows with $x \leq 0$ or $y \leq 0$ or $z \leq 0$ Encode ordered cut, colour, and clarity grades Form specification groups using Eq. (2) Split groups with three-fold GroupKFold each candidate regressor m each fold k Fit m on training groups using $\log(1 + p)$ Store inverse-transformed predictions for held-out groups Compute MAE, RMSE, MAPE, median APE, and R^2 Select the lowest-RMSE model for reference-price accuracy Refit XGBoost; compute TreeSHAP attributions by Eq. (5) Compute out-of-fold XGBoost deviations r_i by Eq. (6) Route records with $r_i > \tau$ to human price review Metrics, explanations, flagged records, audit prompts

where ϕ_0 is the expected model output and ϕ_{ij} is the Shapley attribution for feature j . Global importance is summarized by $I_j = N^{-1} \sum_i |\phi_{ij}|$. The supplied code computes SHAP values on a fixed random sample of 5,000 cleaned records to control memory while retaining the full dataset for model training and evaluation. Because size variables are strongly related, SHAP values are interpreted as model reliance rather than isolated causal price effects.

The audit layer uses XGBoost out-of-fold predictions because the risk score must not be evaluated on in-sample fitted values. For observed price p_i and inverse-transformed reference prediction $\hat{p}_i$, relative deviation is (6):

$$r_i = \frac{|p_i - \hat{p}_i|}{\max(p_i, 1)}. \quad (6)$$

A record is routed for human review when $r_i > \tau$. Threshold τ is therefore a workload control, not a learned probability of mispricing. If $p_i > \hat{p}_i$, the interface recommends reviewing possible overpricing, grade/measurement anomalies, or omitted premium factors; if $p_i < \hat{p}_i$, it recommends reviewing possible margin leakage, promotion context, or data-entry errors. These prompts are intentionally non-causal. A merchandiser must consult commercial context before changing a price.

3.4. Review-capacity governance and decision traceability

A fixed error threshold is easy to communicate, but enterprise review teams are usually constrained by analyst capacity. The same residual distribution can therefore be parameterized by a review budget. Let B denote the maximum fraction of transactions that can be examined in a review cycle. A capacity-compatible threshold is written as (7), where Q_{1-B} is the empirical $(1 - B)$ quantile of out-of-fold relative errors.

$$\tau_B = Q_{1-B}(\{r_i\}_{i=1}^N), \quad (7)$$

Equation (7) does not optimize revenue or assume that large residuals are errors; it only maps a finite review

resource to a reproducible queue size. On the present benchmark, a 20% review capacity implies $\tau_B = 0.1263$, a 10% capacity implies 0.1647, a 5% capacity implies 0.2034, a 2% capacity implies 0.2653, and a 1% capacity implies 0.3149. The corresponding flagged counts are 10,784, 5,392, 2,696, 1,079, and 540 records. Mean relative deviation among the selected records increases from 18.62% at 20% capacity to 40.74% at 1% capacity. Thus, capacity reduction concentrates the queue on progressively more atypical transactions, but necessarily sacrifices coverage of moderate discrepancies.

This formulation provides a practical governance advantage over choosing a threshold after inspecting desirable business outcomes. A risk committee can state the available review capacity before scoring a batch, compute τ_B from the frozen out-of-fold or validation residual distribution, and then apply that threshold to a subsequent period. If the live flag rate shifts materially, the change becomes evidence of either distribution drift, altered market conditions, data-quality degradation, or model deterioration rather than a reason to silently retune the threshold. In a mature deployment, threshold changes should therefore be versioned alongside the model and approved separately from model retraining.

Decision traceability also requires that the system retain more than a single SHAP chart. For each reviewed item, the audit record should contain the input feature vector after validation, source-field lineage, model identifier, model-training cutoff, reference-price prediction, observed or proposed price, residual, threshold version, local SHAP vector, recommendation code, analyst disposition, and final action. This record enables three forms of retrospective analysis. First, false alarms can be stratified by omitted factors such as promotions or certification premiums. Second, frequently recurring data-quality causes can be routed upstream to master-data owners instead of repeatedly burdening merchandisers. Third, disagreement between analysts and the model can be monitored for systematic segments, revealing where a globally accurate model is locally weak.

The recommendation layer is deliberately rule-governed rather than learned from historical analyst actions. Learning directly from prior mitigation decisions would risk encoding legacy pricing habits, organizational incentives, or inconsistent reviewer behavior as if they were optimal labels. In the current design, SHAP identifies which recorded attributes influenced the reference model, the residual quantifies disagreement with the transaction price, and deterministic policy maps that evidence to a diagnostic checklist. For example, a large positive residual combined with strong size-related SHAP contributions triggers verification of measurement units and premium-context fields before any price reduction is considered. A large negative residual can trigger promotion and margin checks. Human approval remains mandatory because the observational dataset contains no counterfactual evidence about the financial effect of changing a price.

Finally, deployment should distinguish monitoring of predictive performance from monitoring of explanation stability. A model can preserve aggregate RMSE while shifting reliance among correlated variables. Consequently, the production monitor should track both error distributions and the distribution of grouped SHAP mass across size, quality,

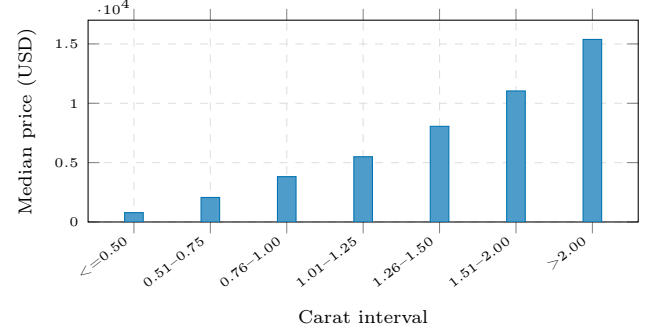


Figure 2: Observed median transaction price by carat interval after validity cleaning.

Table 2: Grouped out-of-fold model performance.

Model	MAE (USD)	RMSE (USD)	MAPE (%)	R^2
Extra Trees	280.96	549.90	7.18	0.9810
HistGB	285.57	565.49	6.96	0.9799
Random Forest	289.11	567.24	7.17	0.9798
XGBoost	307.61	596.87	8.00	0.9776
Decision Tree	356.59	678.16	9.20	0.9711

and proportion feature families. Large changes in either layer warrant investigation. This dual monitoring strategy is more informative than a single accuracy alarm because it can identify silent changes in the model’s internal dependence structure before they produce a large aggregate error increase.

4. Results and discussion

4.1. Dataset structure and price nonlinearity

The cleaned dataset spans a wide price range, from USD 326 to USD 18,823, with a median of USD 2,401. The relationship between size and price is strongly nonlinear. Figure 2 shows median observed price across carat intervals: the median rises from USD 788 for stones at or below 0.50 carat to USD 15,379 for stones above 2.00 carats. The acceleration across larger bins explains why a single global linear price-per-carat rule is inadequate. However, the bin plot is descriptive and does not control for quality grades, so it must not be read as a marginal causal carat effect.

4.2. Model comparison

Table 2 and Figure 3 report out-of-fold performance. Extra Trees ranks first by RMSE at USD 549.90 and by MAE at USD 280.96, with $R^2 = 0.9810$. HistGB is close, with RMSE USD 565.49 and the lowest MAPE, 6.96%. Random Forest produces RMSE USD 567.24. XGBoost reaches RMSE USD 596.87, MAE USD 307.61, MAPE 8.00%, and $R^2 = 0.9776$. The single decision tree is weakest, with RMSE USD 678.16. Thus, ensembling materially improves generalization, but the ranking does not support a claim that boosting universally dominates randomized tree ensembles on this dataset.

The ranking is stable across folds. Extra Trees fold RMSEs are USD 547.80, 553.86, and 548.01 (standard deviation 3.44), while XGBoost fold RMSEs are USD 593.95, 598.64,

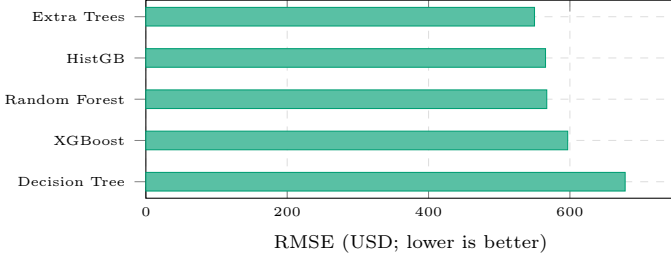


Figure 3: Model comparison under specification-grouped cross-validation.

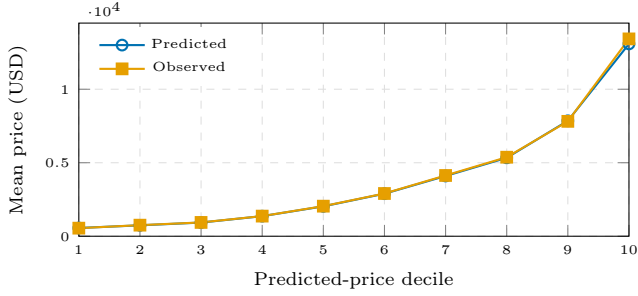


Figure 4: Out-of-fold XGBoost predicted and observed mean prices by prediction decile.

and 598.00 (standard deviation 2.54). These small across-fold variations indicate that no single partition drives the aggregate ranking. They do not establish external validity across retailers or time because neither retailer identity nor transaction dates are available.

4.3. Reference-price calibration by decile

Figure 4 compares mean XGBoost predictions with mean actual prices within deciles of the out-of-fold predicted price. The two curves remain close throughout most of the range. In the lowest decile, mean predicted and actual prices are USD 560.93 and USD 549.16; in the ninth decile they are USD 7,852.90 and USD 7,815.66. The largest visible discrepancy occurs in the top decile, where mean prediction is USD 13,083.18 and mean actual price is USD 13,424.57. This pattern is consistent with residual compression at the high-price tail and explains why RMSE remains materially larger than MAE.

4.4. SHAP-based model explanation

Figure 5 shows mean absolute TreeSHAP magnitude on the log-price scale. The y dimension is first (0.4647), followed by carat (0.2403), z (0.1506), clarity rank (0.1358), and x (0.1336). Colour rank contributes 0.0886, while cut, depth, and table have smaller global magnitudes in the fitted XGBoost model. Signed feature-SHAP correlations are positive for all nine variables under the chosen ordinal encodings; for example, carat has correlation 0.938 with its SHAP contribution, and clarity rank has 0.956. These correlations summarize direction within the fitted model and should not be treated as independent price elasticities.

The dominance of dimensions is plausible because price strongly scales with stone size, but it also exposes an interpretability limitation. Carat and dimensions are physically coupled. A model can substitute one correlated size proxy

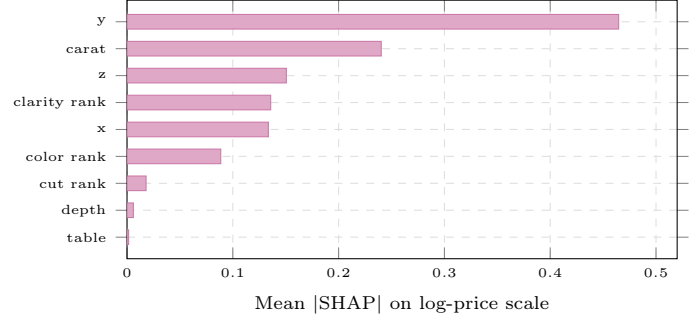


Figure 5: Global XGBoost TreeSHAP importance. Correlated size variables can share attribution.

Table 3: XGBoost out-of-fold pricing-risk audit thresholds.

τ	Flagged records	Flag rate (%)	Mean flagged error (%)
0.10	16,348	30.32	16.11
0.15	7,098	13.16	21.14
0.20	2,825	5.24	27.26
0.25	1,320	2.45	33.07
0.30	648	1.20	39.08

for another across nearby regions of feature space, so ranking y above carat does not establish that the y dimension is economically more fundamental. For operational use, the appropriate reading is that the model's reference price is highly dependent on the size-measurement block as a whole. This is precisely why the proposed interface retains both global importance and local attributions rather than translating a single SHAP ranking directly into pricing action.

4.5. Pricing-risk thresholds and mitigation prompts

The audit layer demonstrates how predictive residuals can be converted into manageable queues. Table 3 and Figure 6 show that a 10% threshold flags 16,348 records (30.32%), whereas a 20% threshold reduces the queue to 2,825 records (5.24%). At 30%, only 648 records (1.20%) are flagged. As expected, stricter thresholds reduce workload but concentrate more extreme errors: the mean relative error among flagged records rises from 16.11% at $\tau = 0.10$ to 39.08% at $\tau = 0.30$.

The threshold results should not be interpreted as mispricing prevalence because the reference model is imperfect and the dataset lacks promotion, retailer, certification-laboratory, inventory-age, and time variables. Instead, the screen is a prioritization mechanism. A flagged high-side transaction can trigger checks for a premium certification, unusual market scarcity, data-entry error, or deliberate premium positioning. A flagged low-side transaction can trigger checks for promotion status, stock-clearance policy, damaged-goods status, or margin leakage. If the explanation indicates that the reference price is driven primarily by size attributes, the reviewer can verify the measurement block; if quality grades dominate, the reviewer can verify cut/colour/clarity coding. This action mapping makes the explanation operational without claiming that changing the highest-SHAP feature will causally remove the residual.

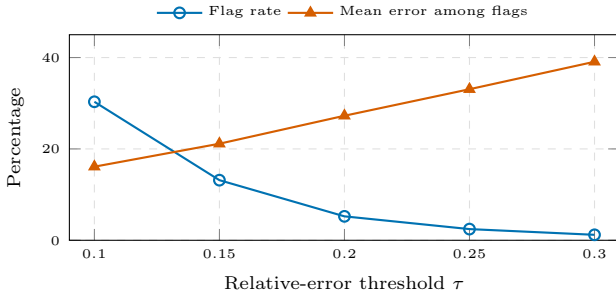


Figure 6: Review-workload trade-off for the residual-based pricing-risk screen.

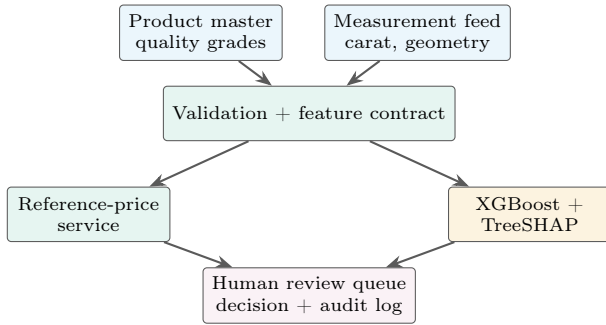


Figure 7: Deployment-oriented separation of validation, prediction, explanation, and human review.

4.6. Deployment architecture and limitations

A production implementation can separate four services: data-quality validation, reference-price inference, explanation generation, and review orchestration. Figure 7 illustrates this separation. Product-master and measurement feeds are validated before inference; the model registry exposes a versioned Extra Trees price estimator and a separately versioned XGBoost explanation model; the review service stores the model version, threshold, prediction, residual, and SHAP vector with the analyst disposition. This design allows a later audit to reconstruct why a record entered the queue and which evidence was shown to the reviewer.

Three limitations are important. First, the benchmark is cross-sectional and does not contain timestamps, retailer identity, inventory state, competitor prices, promotions, or macroeconomic context; consequently it cannot evaluate temporal drift or transfer across firms. Second, grouped cross-validation reduces specification analogue leakage but cannot eliminate all dependence among nearby products. Third, SHAP explanations inherit model and data dependencies, including correlated dimensions. The high R^2 therefore describes this benchmark under the stated grouping, not a universal retail pricing accuracy. External validation on time-stamped enterprise catalogues should be required before operational deployment.

5. Threats to validity

A) Construct validity:

The observed transaction price is used as the reference target, but it is not a direct measurement of intrinsic product value or optimal willingness-to-pay. Price can embed

promotion strategy, inventory pressure, retailer reputation, certification, and timing effects that are absent from the feature table. Consequently, a large residual measures disagreement with the fitted attribute-price function rather than verified commercial error. The term “pricing risk” is therefore restricted to audit prioritization throughout this study.

B) Internal validity:

All cleaning rules, ordinal mappings, groups, and model configurations are fixed in the supplied code, which limits researcher degrees of freedom after observing results. Nevertheless, the ordinal encodings impose monotone grade order and the candidate models are not exhaustively tuned. Alternative encodings or hyperparameters could change model ranking. The use of log-price reduces tail dominance during fitting, whereas final metrics are evaluated after inverse transformation; this design choice may favor models differently from direct dollar-scale training.

C) External validity:

The dataset contains many real product observations but represents one product class. It lacks timestamps and retailer identifiers, so neither temporal generalization nor cross-enterprise transportability can be tested. A production evaluation should use forward-chaining validation across transaction dates and, where multiple merchants are available, retailer-held-out tests. The present grouped protocol should be understood as a stronger analogue-control benchmark than random row splitting, not as a substitute for those deployment tests.

D) Conclusion validity:

The reported fold-level dispersion is descriptive because three folds do not support precise significance claims about small differences between ensembles. Extra Trees exceeds HistGB by only USD 15.59 in aggregate RMSE, and no claim of universal superiority is made. Similarly, SHAP rankings describe the fitted XGBoost explanation model, not the Extra Trees accuracy leader. Keeping these roles separate reduces the temptation to infer business causality from feature attribution or statistical certainty from a single benchmark.

6. Conclusion

This study developed an explainable AI architecture for retail reference-price estimation and pricing-risk screening using 53,920 valid observations from a real 53,940-record diamond dataset. Specification-grouped cross-validation was used to reduce near-duplicate leakage. Extra Trees achieved the best overall predictive result with MAE USD 280.96, RMSE USD 549.90, MAPE 7.18%, and $R^2 = 0.9810$; HistGB produced the lowest MAPE at 6.96%. XGBoost did not rank first, but its RMSE of USD 596.87 and $R^2 = 0.9776$ were sufficiently competitive to support a separate TreeSHAP explanation layer. Global SHAP magnitudes showed that the learned reference-price function relied most heavily on geometric size variables and carat, followed by clarity and colour, with the explicit caution that correlated size variables can share attribution.

The residual-based audit policy translated model disagreement into a transparent review workload. A 20%

relative-deviation threshold flagged 2,825 records, corresponding to 5.24% of the cleaned dataset, while a 30% threshold reduced the queue to 648 records. These flags are not labels of incorrect pricing; they identify transactions whose observed price is difficult for the reference model to reconcile with recorded attributes. The system therefore recommends verification actions rather than autonomous price changes. Its principal implication is architectural: prediction accuracy, explanation, and mitigation should be governed as separate layers with independent versioning and audit records.

Declarations and ethical statements

Conflict of interest: The author declares that there is no conflict of interest.

Funding statement: The author declares that no specific funding was received for this research.

Artificial Intelligence usage statement: During the preparation of this manuscript, the author utilized ChatGPT solely for language refinement and grammatical corrections. The author carefully reviewed and revised the generated content and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Availability of data and materials: The data and/or materials that support the findings of this study are available from the corresponding author upon reasonable request.

CRedit authorship contribution statement

Saumen Bandyopadhyay: Conceptualization, Investigation & Original draft writing.

Publisher's note

Krrish Scientific Publications Pvt. Ltd. and the *Journal of Computing and Data Technology* remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- [1] Rosen S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of political economy*. 1974 Jan 1;82(1):34-55. Available from: <https://doi.org/10.1086/260169>
- [2] Edmonds Jr RG. A Theoretical Basis for Hedonic Regression: A Research Primer. *Real Estate Economics*. 1984 Mar;12(1):72-85. Available from: <https://doi.org/10.1111/1540-6229.00311>
- [3] Tomkovick C, Dobie KE. Applying Hedonic Pricing Models and Factorial Surveys at Parker Pen to Enhance New Product Success. *Journal of Product Innovation Management: AN INTERNATIONAL PUBLICATION OF THE PRODUCT DEVELOPMENT & MANAGEMENT ASSOCIATION*. 1995 Sep;12(4):334-45. Available from: <https://doi.org/10.1111/1540-5885.1240334>
- [4] Roheim CA, Gardiner L, Asche F. Value of Brands and Other Attributes: Hedonic Analysis of Retail Frozen Fish in the UK. *Marine Resource Economics*. 2007 Jan 1;22(3):239-53. Available from: <https://doi.org/10.1086/mre.22.3.42629557>
- [5] Martínez-Garmendia J. Application of hedonic price modeling to consumer packaged goods using store scanner data. *Journal of Business Research*. 2010 Jul 1;63(7):690-6. Available from: <https://doi.org/10.1016/j.jbusres.2009.05.002>
- [6] Lee JA, Caudill SB, Mixon Jr FG. Shine bright like a diamond: a hedonic model of grading and pricing an experience good. *Applied Economics*. 2014 Jun 3;46(16):1829-38. Available from: <https://doi.org/10.1080/00036846.2014.884707>
- [7] Ferreira KJ, Lee BH, Simchi-Levi D. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & service operations management*. 2016 Feb;18(1):69-88. Available from: <https://doi.org/10.1287/msom.2015.0561>
- [8] Kigo SN, Omondi EO, Omolo BO. Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports*. 2023 Oct 12;13(1):17315. Available from: <https://doi.org/10.1038/s41598-023-44326-w>
- [9] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 2016 Aug 13 (pp. 785-794). Available from: <https://doi.org/10.1145/2939672.2939785>
- [10] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*. 2020 Jan;2(1):56-67. Available from: <https://doi.org/10.1038/s42256-019-0138-9>
- [11] Breiman L. Random Forests. *Machine learning*. 2001 Oct;45(1):5-32. Available from: <https://doi.org/10.1023/A:1010933404324>
- [12] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Machine learning*. 2006 Apr;63(1):3-42. Available from: <https://doi.org/10.1007/s10994-006-6226-1>
- [13] Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of statistics*. 2001 Oct 1:1189-232. Available from: <https://doi.org/10.1214/aos/1013203451>
- [14] Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*. 2019 May;9(3):e1301. Available from: <https://doi.org/10.1002/widm.1301>
- [15] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 2016 Aug 13 (pp. 1135-1144). Available from: <https://doi.org/10.1145/2939672.2939778>
- [16] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bénéto A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*. 2020 Jun 1;58:82-115. Available from: <https://doi.org/10.1016/j.inffus.2019.12.012>
- [17] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)*. 2018 Aug 22;51(5):1-42. Available from: <https://doi.org/10.1145/3236009>
- [18] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*. 2015 Jan 2;24(1):44-65. Available from: <https://doi.org/10.1080/10618600.2014.907095>
- [19] Apley DW, Zhu J. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2020 Sep;82(4):1059-86. Available from: <https://doi.org/10.1111/rssb.12377>
- [20] Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*. 2023 Sep;37(5):1719-78. Available from: <https://doi.org/10.1007/s10618-023-00933-9>
- [21] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing

- average model performance. *Climate research*. 2005 Dec 19;30:79-82. Available from: <https://doi.org/10.3354/cr030079>
- [22] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*. 2021 Jul 5;7:e623. Available from: <https://doi.org/10.7717/peerj-cs.623>
- [23] Nadeau C, Bengio Y. Inference for the Generalization Error. *Advances in neural information processing systems*. 1999;12. Available from: <https://doi.org/10.1023/A:1024068626366>
- [24] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*. 2006 Feb 23;7(1):91. Available from: <https://doi.org/10.1186/1471-2105-7-91>