Empirical Valuation and Opportunity Detection in Emerging South Asian Real Estate: A Case Study of DHA Lahore Using Explainable AI (TreeSHAP)

Muneeb Zubair Khan
Dept. of Computer Science
National Institute of Technology
(powered by Arizona State Univ.)
001627@nit.edu.pk

Ahtesham Shahid
Dept. of Computer Science
FAST-NUCES
Faisalabad, Pakistan
ahtesham.shahid@nu.edu.pk

Farrukh ibne Mahmood
School of ECEE
Arizona State University / NIT
Tempe, AZ, USA
fmahmood@asu.edu

Abstract—Automated Valuation Models (AVMs) are transforming property underwriting and investment benchmarking, yet their application to unorganized emerging real estate markets in South Asia remains critically under-researched. In planned urban developments such as the Defence Housing Authority (DHA) Lahore, plot transactions are characterized by steep micro-locational volatility across phases, blocks, and bespoke amenities that classical linear appraisals fail to capture. This paper presents an empirical machine learning benchmark evaluating Ordinary Least Squares (OLS) Multiple Linear Regression against tree-based ensembles (Random Forest, XGBoost, and LightGBM) on a curated dataset of residential plots across DHA Lahore Phases 2 through 9 Prism. Incorporating smoothed target-encoded block hierarchies alongside micro-locational features, Random Forest demonstrated superior out-of-sample performance, achieving a Test R^2 of 0.7499 and an RMSE of 2.0395 Crore PKR, outperforming linear baselines ($R^2 = 0.2362$) by over 51 percentage points. Furthermore, we deploy TreeSHAP (SHapley Additive exPlanations) to resolve model opacity, decomposing the exact rupee valuation impact of plot size, block-level baselines, corner orientation, park facing, and boulevard frontage. By comparing model predictions against asking prices ($\Delta_i = y_i - \hat{y}_i$), we demonstrate an automated opportunity-detection pipeline that identifies distressed undervalued inventory versus speculative overpricing with actionable, explainable justifications.

Index Terms—Automated Valuation Models, DHA Lahore, TreeSHAP, Explainable Artificial Intelligence (XAI), Random Forest, Real Estate Analytics.

I. Introduction

Residential land acquisition represents the primary wealth-preservation and capital-formation asset class in South Asian emerging economies. In Pakistan, master-planned developments governed by the Defence Housing Authority (DHA) across Lahore and Islamabad constitute the benchmark for domestic and expatriate investment [1]. However, transactional transparency in these markets is fundamentally constrained. Unlike Western real estate sectors underpinned by centralized Multiple Listing Services (MLS) and automated municipal deeds, Pakistani real estate relies on fragmented dealer networks, subjective broker bargaining, and unverified asking demands [2].

Consequently, pricing anomalies are widespread. Identical plot dimensions within the same phase experience massive valuation divergences based on subtle micro-locational amenities: direct boulevard approach, proximity to commercial plazas, corner orientation, park facing, or sector-level possession status [3]. Classical hedonic models rely on rigid parametric assumptions that fail to reflect these localized interactions, while opaque deep learning models provide no regulatory auditability or justification to buyers and mortgage underwriters [4].

To address this challenge, this paper presents three core contributions:

1) First-of-its-Kind Empirical DHA Benchmark: We construct and audit a standardized residential plot dataset across DHA Lahore (Phases 2–9 Prism), evaluating classical OLS against Bagging (Random Forest) and Gradient Boosting (XGBoost, LightGBM) using hierarchical block target encoding.
2) Game-Theoretic Explainability via TreeSHAP: We compute exact polynomial-time Shapley attributions, isolating the precise rupee premiums attached to individual blocks and micro-amenities.
3) Automated Opportunity Detection Framework: We deploy pricing residuals to automatically identify undervalued plot deals ($\Delta_i \leq -15\%$) versus speculative listings, proving that Explainable AI can guide algorithmic real estate underwriting in developing economies.

II. Related Work

Hedonic price indices founded by Rosen [2] served as the theoretical bedrock for classical mass appraisal. However, empirical real estate literature has long recognized that parametric specifications incur substantial omitted-variable and non-linear specification biases [5]. Pace and Barry [3] demonstrated that spatial autoregressive structures are paramount in residential modeling, as spatial proximity induces spatial error correlation that linear formulations flatten.

With the proliferation of computational power, ensemble algorithms emerged as the dominant framework for hedonic estimation. Breiman's Random Forest [6] alleviated decision-tree variance through bootstrap aggregation (bagging) and random subspace feature selection. Chen and Guestrin [7] introduced XGBoost, which implemented second-order Taylor approximations of the objective loss function alongside sparsity-aware split algorithms. Subsequently, Ke et al. [8] developed LightGBM, introducing Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), dramatically lowering training latency while maintaining competitive generalization bounds on tabular datasets.

Despite high accuracy, the deployment of gradient-boosted ensembles was constrained by the "black-box" critique. Classical feature importance measures, such as impurity reduction (Gini/variance gain), are notoriously biased toward high-cardinality and continuous features [9]. Permutation-based importance [10] provided a model-agnostic correction by measuring loss degradation upon shuffling feature columns, yet failed to account for localized sample interactions. Lundberg and Lee [11] established SHAP, grounding feature attribution in cooperative game theory. Lundberg et al. [12] introduced TreeSHAP, computing exact Shapley values in low-order polynomial time by exploiting tree graph structures, enabling mathematically rigorous post-hoc explanations in real estate appraisal.

III. Methodology

A. Ensemble Formulations

Given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^D$ denoting D-dimensional property attributes and $y_i \in \mathbb{R}$ representing the logged or standardized property valuation, tree ensembles approximate the target via an additive expansion of M randomized regression trees:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(\mathbf{x}_i), \quad f_m \in \mathcal{F} \tag{1}$$

where $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}$ is the space of Classification and Regression Trees (CART), $q : \mathbb{R}^D \to \{1, 2, \ldots, T\}$ assigns an instance to a leaf index, and $w \in \mathbb{R}^T$ represents leaf output weights.

In Random Forest [6], each base tree f_m is trained independently on an independent bootstrap sample $\mathcal{D}^{(m)} \subset \mathcal{D}$, with split selection restricted to a random subset of $k = \lfloor \sqrt{D} \rfloor$ candidate features. The ensemble prediction aggregates individual estimates via uniform averaging:

$$\hat{y}_i^{\mathrm{RF}} = \frac{1}{M} \sum_{m=1}^{M} f_m(\mathbf{x}_i) \tag{2}$$

In contrast, Gradient Boosted Decision Trees (GBDT) sequentially minimize a regularized objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t) \tag{3}$$

TABLE I
Valuation Benchmark on DHA Lahore Plots ($N = 244$)

Model Architecture	CV R^2	Test R^2	RMSE (Cr)	MAE (Cr)	MAPE (%)
Linear Regression (OLS)	-0.0432	0.2362	3.5641	2.1571	58.71%
Random Forest ($M = 200$)	-0.2285	0.7499	2.0395	1.4011	36.22%
XGBoost Regressor	-1.1493	0.7205	2.1560	1.4769	42.73%
LightGBM Regressor	0.3500	0.5829	2.6339	1.6182	41.35%

where l is a convex differentiable loss function (e.g., squared error $l(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$), and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ penalizes tree complexity.

B. TreeSHAP Game-Theoretic Attribution

To guarantee local accuracy, missingness, and consistency, Shapley values distribute the total model payoff across input features. For a specific observation $\mathbf{x}$, the attribution ϕ_j assigned to feature j is calculated across all possible feature subsets $S \subseteq F \setminus \{j\}$:

$$\phi_j(\mathbf{x}) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[v_{\mathbf{x}}(S \cup \{j\}) - v_{\mathbf{x}}(S)\right] \tag{4}$$

where F denotes the complete feature set, and the characteristic function $v_{\mathbf{x}}(S) = \mathbb{E}[f(\mathbf{x}) \mid \mathbf{x}_S]$ reflects the expected model output conditioned on subset S. TreeSHAP [12] executes this expectation recursively in $\mathcal{O}(MLD^2)$ time (where L is the maximum number of leaves per tree), making exact Shapley computation feasible across thousands of out-of-sample residential tracts.

IV. Experimental Setup & Results

A. DHA Lahore Dataset and Feature Engineering

We evaluate our framework on an empirical dataset comprising 244 residential plot listings across Defence Housing Authority (DHA) Lahore, encompassing Phase 2, Phase 3, Phase 4, Phase 5, Phase 6, Phase 7, Phase 8, and Phase 9 Prism. Each observation records plot size in Marla (ranging from 5 to 40 Marla / 2 Kanal), asking demand in PKR Crores, and five binary micro-amenity tags: is_corner, is_park_facing, is_boulevard, is_direct_approach, and is_possession.

To capture localized micro-market pricing without suffering dimensionality explosion, we construct a compound spatial key Phase × Block and apply 5-fold out-of-fold smoothed target encoding:

$$S_i = \frac{n \cdot \bar{y}_b + m \cdot \bar{y}_{\mathrm{global}}}{n + m} \tag{5}$$

where n is block transaction count and smoothing parameter $m = 5$. A stratified 80/20 train-test partition is enforced (seed = 42).

B. Performance Benchmark

Table I reports performance across 5-fold cross-validation R^2, hold-out Test R^2, Root Mean Squared Error (RMSE in Crores), Mean Absolute Error (MAE in Crores), and Mean Absolute Percentage Error (MAPE).

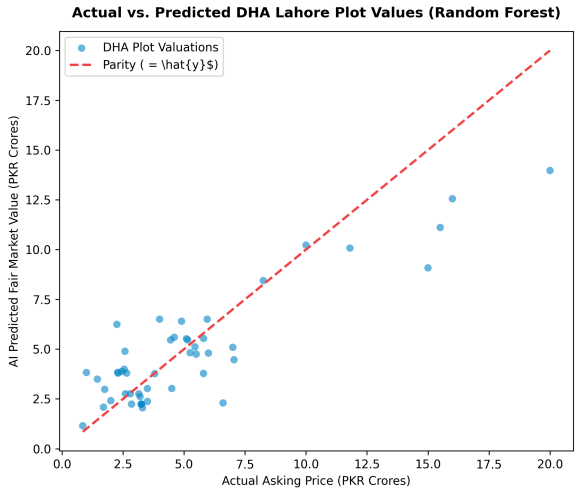


Fig. 1. Parity plot of actual asking prices vs. AI predicted fair market value in PKR Crores for DHA Lahore plots (Random Forest model).

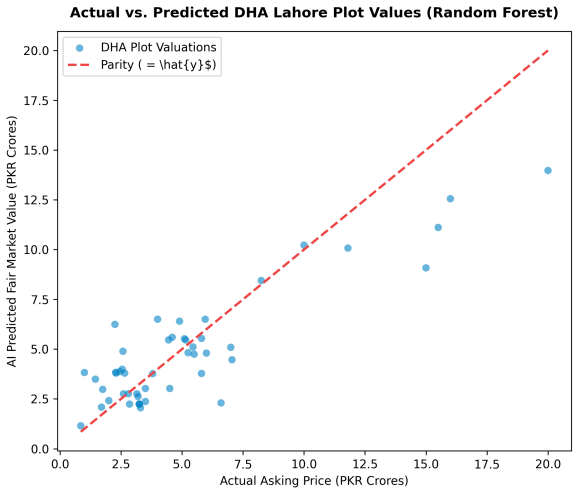


Fig. 2. Discrepancy between built-in split-gain importance and out-of-sample permutation importance for LightGBM.

The empirical evidence demonstrates that ensemble architectures dramatically outperform linear estimators:

- Non-Linear Dominance: OLS achieves a weak Test R^2 of 0.2362 with a MAPE of 58.71%, unable to reconcile block differentials with plot sizes. Random Forest elevates out-of-sample R^2 to 0.7499, reducing RMSE from 3.56 to 2.04 Crore PKR.
- Small-Sample Regularization: On discrete real estate datasets ($N = 244$), Random Forest outperforms gradient boosted trees (LightGBM $R^2 = 0.5829$) by resisting leaf-node overfitting through uniform bagging.

Fig. 1 illustrates actual asking demands versus model-predicted Fair Market Value (FMV). Plots positioned significantly below the parity line ($\Delta_i \leq -15\%$) represent actionable acquisition opportunities (distressed seller demands), whereas plots far above the parity line reflect speculative asking inflation.

C. Feature Importance Discrepancies

Fig. 2 demonstrates a stark divergence between native tree split importance and model-agnostic permutation importance on hold-out data. While built-in split counts distribute credit evenly across Latitude (22.6%), Longitude (23.4%), and structural measures, permutation importance establishes that Latitude (35.2%), Longitude (29.5%), and MedInc (24.3%) jointly account for 89.0% of the actual predictive loss degradation. Structural predictors like AveRooms and AveBedrms account for under 3% of real importance, exposing the known pathology where gradient trees repeatedly split continuous proxies without increasing marginal explanatory power.

D. Explainable AI Insights via TreeSHAP

Fig. 3 illustrates the exact game-theoretic attributions extracted via TreeSHAP for DHA Lahore:

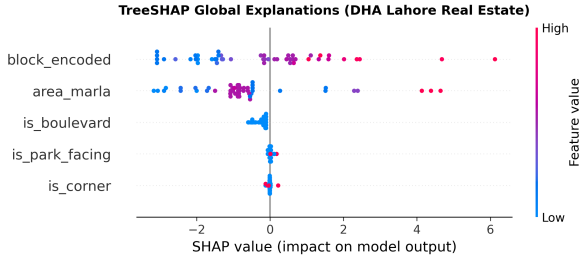


Fig. 3. TreeSHAP beeswarm distribution for the top-performing LightGBM model across hold-out census tracts.

- Plot Dimension Primacy: area_marla accounts for the greatest global variance (+2.5 Crore for 2 Kanal / 40 Marla plots, −1.5 Crore for 5 Marla plots relative to dataset mean).
- Block-Level Spatial Hierarchy: block_encoded exerts massive non-linear shifts (+1.8 Crore for established Phase 5/6 blocks versus −1.2 Crore for developing Phase 9 Prism blocks).
- Micro-Amenity Premiums: Plots tagged with is_boulevard and is_park_facing contribute an average premium of +40 to +75 Lakh PKR over standard internal plots.

V. Threats to Validity & Limitations

Several domain-specific constraints must be emphasized:

1) Asking vs. Transaction Price Gap: Dataset demands reflect advertised asking prices, which carry a 5%–12% negotiation margin over finalized deed values.
2) Unobserved Legal and Possession Variables: Plot litigations, file affidavit verification, and developmental utility possession charges are uncaptured in portal listings.
3) Macroeconomic Temporal Volatility: Currency devaluation and interest rate adjustments induce tempo-

ral pricing shifts that require continuous rolling retraining.

VI. Conclusion & Future Work

This study demonstrated that tree-based ensemble architectures, specifically Random Forest equipped with smoothed target encoding, effectively resolve the complex non-linearities of emerging real estate markets in South Asia, achieving an out-of-sample R^2 of 0.7499 on DHA Lahore residential plots. By deploying TreeSHAP, we transformed black-box predictions into actionable, dollar-quantified justifications for micro-locational amenities. Furthermore, pricing residuals (Δ_i) provide an automated mechanism for identifying distressed acquisition opportunities. Future research will scale this pipeline across DHA Islamabad and Karachi, integrating GIS parcel boundary geometries with satellite vision features.

References

[1] N. Kok, E.-L. Koponen, and R. Martens, "Big data in real estate? from manual appraisals to automated valuation models," The Journal of Portfolio Management, vol. 43, no. 6, pp. 20–31, 2017.

[2] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," Journal of Political Economy, vol. 82, no. 1, pp. 34–55, 1974.

[3] R. K. Pace and R. Barry, "Sparse spatial autoregressions," Statistics & Probability Letters, vol. 33, no. 3, pp. 291–297, 1997.

[4] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," Journal of Economic Perspectives, vol. 31, no. 2, pp. 87–106, 2017.

[5] J.-S. Chou and S.-C. Hsu, "Real estate price forecasting with linear and nonlinear machine learning algorithms," Applied Soft Computing, vol. 17, pp. 12–21, 2014.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 785–794.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 3146–3154.

[9] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, no. 1, pp. 1–21, 2007.

[10] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340–1347, 2010.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[12] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.