UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER NO. 3

Hedonic Housing Price Models for the United States: A Multi-Method Comparison of Parametric, Quantile, and Machine Learning Approaches

Simon-Pierre Boucher[1]

Département des sciences administratives
Université du Québec en Outaouais

First draft: May 2026

This version: August 5, 2026

Version 1.1

[1]Département des sciences administratives, Université du Québec en Outaouais (UQO), 283 boulevard Alexandre-Taché, Gatineau, QC J9A 1L8, Canada. Email: simon-pierre.boucher@uqo.ca. All errors are my own.

Abstract

This paper compares econometric and machine-learning approaches to hedonic housing valuation using 788,842 active Zillow listings across all 50 U.S. states and the District of Columbia. A semi-log OLS model with 62 regressors ($R^2 = 0.634$) provides interpretable listing-price gradients; adding ZIP3 fixed effects raises R^2 to 0.725, and Moran's $I = 0.27$ confirms strong residual spatial autocorrelation. Quantile regression reveals distributional heterogeneity, with inter-quantile Wald tests rejecting coefficient equality between $\tau = 0.10$ and $\tau = 0.90$ for 11 of 13 key variables. XGBoost achieves $R^2 = 0.833$ under random validation but only 0.425 under state-level geographic holdout; ablation analysis traces the predictive gain primarily to neighborhood-quality features (+17.6 pp) and shows that removing geographic features *improves* geographic holdout R^2 to 0.519, revealing spatial overfitting. SHAP importance rankings are stable across models (Spearman $\rho > 0.89$). Robustness checks confirm that the lot-size gradient triples when imputed observations are dropped, while other coefficients remain stable under winsorization and subsampling. Throughout, estimates are interpreted as conditional associations in listing prices, not causal willingness-to-pay parameters.

Keywords: hedonic pricing, housing markets, quantile regression, XGBoost, SHAP, spatial autocorrelation, geographic validation

JEL Classification: R31, C21, C45, C52

1. Introduction

Housing is the dominant asset in the typical American household's portfolio, the collateral underpinning the largest class of household debt, and a central object of study in urban economics, household finance, and public policy. How housing attributes map into prices matters far beyond academia: property-tax assessment, mortgage underwriting, and the automated valuation models (AVMs) that increasingly mediate transactions all rest on some version of that mapping. The hedonic pricing framework—formalized by Rosen (1974) on the consumer-theoretic foundations of Lancaster (1966), with empirical roots stretching back to Waugh (1928) and Court (1939)—provides the canonical approach: decompose observed prices into the implicit valuations of constituent characteristics. Yet after five decades of applied hedonic research, three methodological tensions remain unresolved, and the arrival of machine-learning valuation has sharpened rather than settled them.

First, the standard OLS hedonic model imposes a (log-)linear relationship between attributes and prices, an assumption chosen largely for robustness under attribute omission (Cropper et al., 1988) but one that may poorly approximate the non-linear, interactive structure of housing markets. Second, by estimating conditional mean effects, OLS constrains implicit prices to be constant across the price distribution—a restriction the quantile-regression literature has shown to be empirically untenable (Zietz et al., 2008, McMillen, 2008, Liao and Wang, 2012). Third, while gradient-boosted models deliver large predictive gains in property valuation (Bourassa et al., 2010, Kok et al., 2017), their black-box character limits adoption where economic interpretation matters (Mullainathan and Spiess, 2017, Athey and Imbens, 2019)—and, as this paper documents, their headline accuracy can be a partial illusion created by the way they are evaluated.

This paper confronts the three tensions on a single dataset of 788,842 active Zillow listings covering all 50 U.S. states and the District of Columbia, with 62 regressors spanning structural attributes, lot characteristics, amenities, neighborhood quality, market status, and six interaction terms. We estimate three families of models, each answering a distinct question:

1. **Semi-log OLS** with HC3 robust standard errors—and progressively finer geographic fixed effects—for interpretable conditional mean associations;
2. **Quantile regression** at five quantiles, with formal inter-quantile Wald tests, for distribution-specific listing-price gradients;
3. **Gradient-boosted ensembles** (XGBoost, LightGBM) with SHAP-based interpretation (Lundberg and Lee, 2017) for non-linear prediction.

The core methodological contribution is a systematic examination of **spatial leakage** in hedonic model evaluation. Housing data are spatially dependent: nearby properties share unobserved local price determinants (Dubin, 1998, Basu and Thibodeau, 1998). Standard random train-test splits therefore allow information to leak from training to test sets through spatial proximity, inflating out-of-sample performance—a phenomenon documented in ecology and geostatistics (Roberts et al., 2017, Ploton et al., 2020, Meyer and Pebesma, 2021) but rarely confronted in housing economics. We quantify its consequences through three complementary designs: (i) a *geographic holdout* in which ten entire states (438,315 listings) are withheld from training; (ii) a six-stage *ablation cascade* that traces the XGBoost predictive gain to specific feature groups under both validation schemes and reveals that region dummies actively *harm* geographic generalization; and (iii) a *coordinate experiment* in which adding raw latitude and longitude boosts random-split R^2 by 3.7 percentage points while *reducing* geographic holdout R^2 by 5.5 percentage points. Together these designs show that the gap between random and geographic validation is not a matter of degree: the two protocols measure qualitatively different model capabilities—interpolation within observed markets versus extrapolation to new ones—and they rank feature sets differently.

Our contribution is methodological and empirical, not causal. Following the identification literature (Bartik, 1987, Ekeland et al., 2004, Kuminoff et al., 2013, Bishop et al., 2020), we interpret all estimates as conditional associations in *listing* prices—capitalization gradients in asking prices—rather than structural willingness-to-pay parameters, and we engage the listing-price microstructure literature (Horowitz, 1992, Genesove and Mayer, 2001, Han and Strange, 2016) when drawing that distinction. Within that discipline, the paper makes four contributions:

1. **Geographic granularity in OLS.** Progressively finer spatial fixed effects (region → state → ZIP3) raise R^2 from 0.634 to 0.725, providing national-scale, prediction-oriented evidence for the specification advice of Kuminoff et al. (2010); residual Moran's I of 0.27 shows that even ZIP3 controls leave substantial spatial structure unabsorbed.
2. **Formal distributional heterogeneity.** Inter-quantile Wald tests reject coefficient equality between the 10th and 90th conditional percentiles for 11 of 13 key attributes; the pattern (garage and living-area gradients concentrated at the bottom of the distribution, pool and lot-size gradients at the top) is stable across ten independent subsamples.
3. **Anatomy of the ML advantage.** XGBoost's $R^2 = 0.833$ under random validation falls to 0.425–0.547 under state-level holdout; removing all geographic features *improves* geographic holdout R^2 to 0.519. The ablation traces the random-validation gain primarily to neighborhood-quality variables (+17.6 pp) and shows their contribution largely fails to transfer across states.
4. **Cross-model stability of SHAP.** Importance rankings correlate at $\rho = 0.89$–0.99 across XGBoost, LightGBM, and random forests, supporting SHAP as a description of predictive structure while our framing—informed by the explainability debate (Rudin, 2019)—keeps it distinct from Rosen's implicit prices.

For practitioners, the message is direct: in AVM deployment contexts that require generalization to unfamiliar markets, validation design is not a technicality but the difference between a model that appears excellent and one that actually transfers; and geographic features, however helpful in-sample, can be counterproductive out-of-market (Steurer et al., 2021). For researchers, the results argue that random-split accuracy comparisons between ML and hedonic models—now common in the literature—systematically overstate the ML advantage whenever the deployment question involves new locations.

The remainder of the paper is organized as follows. Section 2 situates the study in the hedonic, quantile, machine-learning, and spatial-validation literatures. Section 3 describes the data, sample construction, variable coding, and the listing-price caveat. Section 4 presents the econometric and machine-learning methodology, the three validation designs, and the interpretation framework. Section 5 reports estimation results across the three model families. Section 6 presents imputation, winsorization, and subsample-stability checks. Section 7 interprets the findings against the literature, Section 8 enumerates limitations, and Section 9 concludes.

2. Literature Review

This paper sits at the intersection of four literatures: the hedonic pricing tradition in housing economics, the econometrics of distributional heterogeneity, the rapidly growing machine-learning strand of property valuation, and the methodological literature on validating predictive models under spatial dependence. We review each in turn, emphasizing in every case how it bears on the design choices of this study and where the gap addressed by this paper lies.

2.1. Hedonic Pricing: Origins, Theory, and the Interpretation of Implicit Prices

The practice of regressing prices on product characteristics predates its theoretical justification by several decades. Waugh (1928) priced quality attributes of vegetables, Court (1939) constructed hedonic price indexes for automobiles, and Griliches (1961) revived the method for quality-adjusted price measurement. In housing, Ridker and Henning (1967) provided the first regression-based estimates of how a local disamenity—air pollution—is capitalized into residential property values, inaugurating the property-value approach to valuing local public goods. The theoretical foundation arrived with Lancaster (1966), who recast goods as bundles of characteristics, and Rosen (1974), who showed that in a competitive market the equilibrium price schedule $P(\mathbf{z})$ traces out a double envelope of bid and offer functions, so that the gradient $\partial P/\partial z_k$ equals the marginal implicit price of attribute z_k. Sirmans et al. (2005) synthesize 125 empirical applications, documenting robust positive gradients for living area, bathrooms, and garages and negative gradients for property age—the same qualitative pattern our conditional-mean estimates reproduce at national scale.

A central and often underappreciated distinction is that first-stage hedonic gradients are equilibrium price phenomena, not preference parameters. Recovering willingness-to-pay functions requires a second stage whose identification demands exclusion restrictions rarely available in practice (Bartik, 1987, Ekeland et al., 2004). Modern syntheses formalize when hedonic estimates can be trusted: Kuminoff et al. (2013) survey the equilibrium-sorting framework that now underpins structural interpretation of housing-market regressions, and Bishop et al. (2020) distill best practices for credible hedonic estimation, emphasizing fine spatial controls, robustness to specification, and transparent treatment of data limitations. Particularly relevant to our fixed-effects progression, Kuminoff et al. (2010) show in large-scale simulations that specifications with rich spatial fixed effects dramatically outperform sparse cross-sectional specifications in recovering marginal willingness to pay when omitted spatial variables correlate with amenities. Our region → state → ZIP3 exercise provides complementary, purely predictive evidence on the same point. Throughout, we follow this literature's counsel and interpret coefficients as conditional associations—listing-price capitalization gradients—rather than structural demand parameters.

2.2. Functional Form and Specification

The semi-log specification became the workhorse of applied hedonics after Cropper et al. (1988) showed in Monte Carlo experiments that simple forms outperform flexible ones when some attributes are omitted or measured with error—precisely the situation of scraped listing data; Hill (2013) surveys the subsequent evolution of hedonic specification practice in residential applications. Earlier work had already cautioned against atheoretic flexibility: Halvorsen and Pollakowski (1981) demonstrated the sensitivity of Box-Cox hedonic estimates, and Anglin and Gençay (1996) documented gains from semiparametric estimation of the price function. This literature frames one of our central questions: how much of the predictive advantage of gradient-boosted trees reflects genuine non-linearity that the semi-log form misses, and how much reflects something else—in our case, spatial memorization.

2.3. Spatial Dependence in Housing Markets

Housing data are spatial data. Can (1992) showed that ignoring spatial autocorrelation biases hedonic coefficient estimates; Dubin (1998) provides an accessible treatment of why residential prices are spatially correlated (shared unobserved neighborhood attributes, market-mediated spillovers); and Basu and Thibodeau (1998) document strong spatial autocorrelation in Dallas transaction prices even after conditioning on rich attribute sets. The formal apparatus of spatial econometrics—spatial lag, spatial error, and spatial Durbin models—is developed in Anselin (1988) and LeSage and Pace (2009), with Anselin (2010) providing a retrospective. We deliberately stop short of estimating spatial econometric models: our contribution is diagnostic. Using Moran's I (Moran, 1950) on OLS residuals, we quantify the spatial dependence that remains after 62 attributes and regional controls, and we show that it is a stable, structural feature of the data rather than a sampling artifact.

Pace and Hayunga (2020) pursue a complementary strategy, using trees and forests to examine what information hedonic and spatial model residuals still contain; their finding—that machine-learning methods extract signal from residual spatial structure—anticipates our result that tree-based models exploit location rather than only attribute non-linearity.

2.4. Quantile Regression and Distributional Heterogeneity

Quantile regression (Koenker and Bassett, 1978, Koenker and Hallock, 2001) replaces the conditional mean with a family of conditional quantiles, allowing implicit prices to vary across the price distribution. Zietz et al. (2008) introduced the approach to housing, finding that square footage and other attributes are priced differently at different quantiles; Mak et al. (2010) document quantile-varying gradients in Hong Kong; and Liao and Wang (2012) combine quantile regression with spatial methods. McMillen (2008) shows that changes in the *distribution* of house prices over time are driven largely by changes in coefficients rather than in characteristics—evidence that quantile-specific pricing is economically meaningful, not a statistical curiosity. More recently, Waltl (2019) provides a comprehensive quantile analysis of the Sydney market, documenting systematic variation across both price segments and locations. Our contribution to this strand is scale and formality: we estimate five quantiles on a national sample, verify coefficient stability across ten independent subsamples, and subject the tail differences to formal inter-quantile Wald tests, rejecting coefficient equality for 11 of 13 key attributes.

2.5. Listing Prices, Search, and Seller Behavior

Because our dependent variable is an asking price, the microstructure of listing behavior matters for interpretation. Theory and evidence establish that list prices are strategic objects: Horowitz (1992) models the list price as a commitment device in seller search; Knight (2002) shows that overpricing lengthens time-on-market and reduces eventual sale prices; Genesove and Mayer (2001) demonstrate that loss aversion leads sellers—especially those facing nominal losses—to set systematically higher asking prices; and Han and Strange (2016) show that the asking price plays a directing role in buyer search, so that its information content varies across market segments. Two implications follow for our estimates. First, listing-price gradients need not equal transaction-price gradients, and the wedge is likely correlated with attributes (luxury segments and distressed properties exhibit different list-to-sale gaps). Second, this wedge affects all three modeling frameworks identically, so *comparisons across frameworks*—our primary object of interest—are unaffected even where levels must be interpreted cautiously.

2.6. Capitalization of Local Public Goods and Amenities

Several of our regressors proxy local public goods, for which a mature capitalization literature exists. Oates (1969) initiated the study of property-tax and public-spending capitalization; Black (1999) used school-attendance boundaries to isolate the value of school quality, finding that parents pay 2.5% more for a 5% increase in test scores—about half the naive hedonic estimate; and Bayer et al. (2007) embed boundary discontinuities in an equilibrium sorting model, showing that naive cross-sectional estimates confound school quality with neighbor characteristics. This literature disciplines our reading of the counterintuitive negative school-rating coefficient in the OLS results: without boundary-style identification, school ratings in a national cross-section absorb correlated neighborhood and fiscal variation (the property-tax rate enters separately), and the coefficient should not be read as the value of school quality. Similarly, Pivo and Fisher (2011) document a walkability premium using Walk Score—the same measure we employ—while our specification, which conditions simultaneously on bike and transit accessibility, illustrates how collinear accessibility measures split the premium in ways that resist attribute-by-attribute interpretation. Finally, the growing climate-capitalization literature (Baldauf et al., 2020, Bernstein et al., 2019, Murfin and Spiegel, 2020) identifies a class of price-relevant risk variables that are entirely missing from our data—a limitation we return to in Section 8.

2.7. Machine Learning in Property Valuation

The econometrics profession has converged on a division of labor in which machine learning excels at prediction ($\hat{y}$) problems while classical methods target parameter ($\hat{\beta}$) problems (Mullainathan and Spiess, 2017, Varian, 2014, Athey and Imbens, 2019). In real estate, this maps onto automated valuation models (AVMs): Bourassa et al. (2010) compare methods for exploiting spatial dependence in prediction; Park and Bae (2015) document early machine-learning gains in county-level housing data; Kok et al. (2017) describe the shift from manual appraisal to big-data valuation; and Steurer et al. (2021) catalogue the metrics appropriate for evaluating AVM performance, emphasizing that headline R^2 figures conceal economically relevant tail behavior. The tree-ensemble methods we deploy—random forests (Breiman, 2001), gradient boosting (Friedman, 2001), and their modern implementations XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017)—dominate tabular prediction tasks of this kind. Against this backdrop, our contribution is to ask not *whether* boosting beats OLS (it does, by 20 percentage points of R^2 under random validation) but *what that gap is made of*: the ablation design decomposes it into feature-group contributions, and the geographic holdout reveals that a substantial share reflects spatial memorization rather than transferable attribute-price structure.

2.8. Explainability: SHAP and Its Limits

SHAP values (Lundberg and Lee, 2017), rooted in the cooperative-game solution concept of Shapley (1953) and computable efficiently for trees via TreeSHAP (Lundberg et al., 2020), have become the de facto standard for interpreting ensemble predictions. Yet the explainability literature itself urges caution: local surrogate explanations can be unstable (Ribeiro et al., 2016), and Rudin (2019) argues that post-hoc explanations of black-box models should not be conflated with intrinsically interpretable modeling, particularly in high-stakes settings. In the hedonic context the danger is specific: SHAP values are prediction decompositions, sensitive to feature correlation, and do not satisfy the equilibrium conditions under which Rosen's gradient equals a marginal implicit price (Rosen, 1974, Bishop et al., 2020). We therefore use SHAP for what it can do—describe the predictive structure of the fitted model—and we probe the robustness of that description by comparing importance rankings across three different tree ensembles, finding rank correlations of 0.89–0.99.

2.9. Validation Under Spatial Dependence

A methodological literature largely developed in ecology and geostatistics warns that random cross-validation overstates predictive skill whenever observations are spatially dependent, because information leaks from training to test folds through spatial proximity. Roberts et al. (2017) systematize blocking strategies for structured data; Valavi et al. (2019) provide the standard software implementation of spatial blocking; Ploton et al. (2020) demonstrate, strikingly, that large-scale ecological mapping models with excellent random-CV scores lose most of their skill under spatial validation; and Meyer and Pebesma (2021) formalize the "area of applicability" of spatial prediction models. The lesson is not uncontested: Wadoux et al. (2021) show that when the goal is map accuracy over a sampled region—an interpolation problem—spatial cross-validation can be *pessimistically* biased, and design-based random validation is appropriate. This debate sharpens rather than undermines our design: predicting prices in entirely unobserved states is an *extrapolation* task, for which held-out-region validation is the relevant benchmark, while our random split answers the interpolation question. Reporting both, and showing that feature sets rank differently under each, is precisely what the debate prescribes. To our knowledge, this framing has not previously been brought to bear on national-scale hedonic housing models.

2.10. Research Gap

Three gaps emerge from this review. First, the hedonic and AVM literatures rarely meet: machine-learning papers report accuracy without engaging identification and interpretation, while econometric papers report coefficients without assessing predictive generalization. Comparative studies on a single large dataset that treat OLS, quantile regression, and boosting as complementary lenses—each answering a different question—remain scarce. Second, the spatial-validation insights of ecology have barely penetrated housing economics, despite housing being a canonically spatial asset; the interaction between geographic *features* and validation *design* (our ablation and coordinate experiments) is essentially unexplored at national scale. Third, the literature offers little guidance on how explanation tools like SHAP behave across model families in housing applications. This paper addresses all three, at the scale of 788,842 listings spanning every U.S. state, while maintaining the interpretive discipline the identification literature demands.

3. Data and Variable Construction

3.1. Data Source

The dataset is sourced from Zillow, the largest online real estate marketplace in the United States. The raw data contain 839,313 residential properties with 116 variables, representing active for-sale listings as of 2025–2026. The data include listing prices, structural characteristics, geocoordinates, neighborhood quality scores, and tax/transaction histories.

Important limitations of the data source. The sample comprises active Zillow listings, not the universe of U.S. residential properties. Several sources of non-representativeness should be noted: (i) Zillow coverage varies by region, with some MLS-linked markets better represented than others; (ii) active listings capture supply at a point in time, not the full housing stock; (iii) the sample excludes off-market properties, FSBOs not listed on Zillow, and recently transacted properties that are no longer listed; (iv) Southern states are overrepresented (56.6% of the sample vs. approximately 38% of U.S. housing units), likely reflecting higher listing volumes in fast-growing Sun Belt markets. We do not claim national representativeness and caution against interpreting our estimates as population parameters for the entire U.S. housing market.

3.2. Listing Prices versus Transaction Prices

A critical limitation is that our dependent variable is the *listing* (asking) price, not the realized transaction price. Listing prices are strategic objects: theory models them as commitment devices in seller search (Horowitz, 1992) and as instruments that direct buyer attention (Han and Strange, 2016), and empirically they embed seller behavior—overpricing lengthens time-on-market and lowers eventual sale prices (Knight, 2002), while loss-averse sellers systematically set higher asking prices (Genesove and Mayer, 2001). List-to-sale price ratios therefore vary by market condition, property type, and price tier: luxury properties are more frequently overpriced, distressed properties may be strategically underpriced, and regional norms for overbidding versus negotiation differ substantially. Throughout this paper, we interpret the estimated coefficients as **listing-price capitalization gradients**—conditional associations between attributes and asking prices—rather than transaction-price implicit prices. If listing premiums correlate systematically with property attributes (e.g., if waterfront properties are more frequently overpriced), the estimated gradients will reflect this pricing behavior in addition to underlying valuation differences.

3.3. Sample Construction

Table 1 documents the sample construction.

Table 1. Sample Construction

Step	N remaining	Removed	Criterion
Raw Zillow records	839,313	—	—
Valid price	817,473	21,840	\$10,000 < P < \$10,000,000
Valid living area	797,380	20,093	200 < sqft < 20,000
Valid bedrooms/bathrooms	790,799	6,581	1 ≤ bed ≤ 10, 1 ≤ bath ≤ 10
Valid geocoordinates	789,199	1,600	Non-missing lat/lon
U.S. states + DC only	788,842	357	Exclude PR, VI
Final analytical sample	**788,842**	**50,471**	**(6.0% removed)**

Notes: Filters applied sequentially. The 6.0% removal rate suggests the raw data are reasonably clean, though the retained sample may still contain measurement error in secondary variables.

3.4. Variable Description

Our specification includes 62 regressors organized into six categories.

3.4.1. Structural Attributes (8 Variables)

The natural logarithm of living area (ln sqft), number of bedrooms, number of bathrooms, property age (2026 – year_built) and its square, number of stories, the bathroom-to-bedroom ratio, and square feet per bedroom. Living area enters in log form to accommodate the well-documented concavity of the size-price relationship.

3.4.2. Lot Characteristics (1 Variable)

The natural logarithm of lot size (ln lot). Missing lot sizes (16.3% of raw data) are imputed using state-level medians.

3.4.3. Amenity and Quality Indicators (11 Variables)

Binary indicators for swimming pool (39.8% prevalence), spa (6.1%), basement (24.3%), fireplace (39.6%), garage (67.2%), waterfront location (8.9%), central air conditioning (68.1%), forced air heating (25.8%), and hardwood flooring (26.7%). The number of parking spaces and a composite luxury score (0–7) supplement these indicators.

3.4.4. Neighborhood and Location Variables (7 Variables)

Walk Score (0–100), Bike Score, Transit Score, average GreatSchools rating (1–10), school count, distance to nearest school, and local property tax rate. Walk Score is the accessibility measure for which a capitalization premium has been documented in the literature (Pivo and Fisher, 2011); school ratings and tax rates proxy the local public goods whose capitalization into house prices is the subject of a large identification literature (Oates, 1969, Black, 1999, Bayer et al., 2007).

Data quality note. Bike Score has a maximum value of 248 in our data, exceeding the expected 0–100 range. Only 38 observations (0.005%) exceed 100, and these are retained without truncation. Results are robust to capping Bike Score at 100.

3.4.5. Market Status Variables (5 Variables)

Condominium indicator (14.6%), HOA membership (37.1%), log annualized HOA fees ($\ln(1 + \text{HOA}_{\text{annual}})$), new construction (5.9%), and foreclosure status (0.5%).

3.4.6. Interaction Terms (6 Variables)

- $\ln(\text{sqft}) \times$ age: depreciation-size interaction;
- pool $\times$ South: climate-dependent pool gradient;
- waterfront $\times \ln(\text{sqft})$: size-dependent waterfront gradient;
- basement $\times$ North: region-dependent basement gradient;
- condo $\times$ walk_score: walkability gradient for condominiums;
- age $\times$ luxury_score: luxury mitigation of depreciation.

3.4.7. Categorical Controls (24 Dummy Variables)

Simplified indicators for roof type (6 dummies), construction material (8 dummies), foundation type (7 dummies), and Census region (3 dummies; Midwest reference).

3.5. Missing Data and Imputation

Table 2 reports missingness rates for key variables.

Table 2. Missing Data Rates and Imputation Methods

Variable	Missing (%)	Method	Notes
Year built	19.4	State median	Affects age, age^2, interactions
Lot size	16.3	State median	Large variation
Walk Score	3.1	State median	
Bike Score	6.1	State median	Max = 248 (38 obs.)
Transit Score	70.1	Set to 0	Suburban/rural properties
Garage	35.1	Set to False	Conservative
Climate risk factors	100.0	Excluded	Not available

Notes: State-level median imputation is used to preserve geographic variation. The high Transit Score missingness (70.1%) likely reflects properties in areas without public transit infrastructure, for which zero is a reasonable value. Climate risk factors (flood, fire, heat, wind, air) are entirely missing and excluded from the analysis.

3.6. Data Quality Audit

We conduct three data quality checks to assess the sensitivity of key results to imputation and outliers.

Imputation prevalence. Two variables—year built (19.4% missing) and lot size (16.3% missing)—account for the bulk of imputed values. To assess the consequences, we re-estimate the baseline OLS model on subsamples that exclude imputed observations (Section 6.1). The lot-size coefficient is the most sensitive: $\ln(\text{lot})$ triples from 0.018 to 0.059 when imputed lot sizes are dropped, suggesting that state-median imputation attenuates the lot-size gradient substantially.

Outlier prevalence. We winsorize lot size at its 99.5th percentile and cap Bike Score at 100, then re-estimate OLS (Section 6.2). The overall R^2 increases modestly from 0.634 to 0.640, and most coefficients are stable, with the exception of $\ln(\text{lot})$, which again increases substantially.

Geographic coverage. The sample spans 886 unique three-digit ZIP code prefixes (ZIP3 codes) across 51 jurisdictions (50 states plus DC). Southern states account for 56.6% of observations. We document the marginal explanatory power of progressively finer geographic controls in Section 5.1.

3.7. Descriptive Statistics

Table 3 presents summary statistics.

Table 3. Descriptive Statistics (N = 788,842)

Variable	Mean	Std. Dev.	Min	Median	Max
Continuous Variables					
Listing Price ($)	621,737	817,965	10,300	399,900	9,999,999
Living Area (sqft)	2,093	1,181	208	1,820	19,600
Bedrooms	3.25	1.09	1	3	10
Bathrooms	2.52	1.16	1	2	10
Property Age (years)	42.2	31.5	0	36	200
Walk Score (0–100)	25.6	26.5	0	16	100
Bike Score	34.9	18.1	1	31	248
Transit Score (0–100)	8.9	18.3	0	0	100
Avg. School Rating (1–10)	6.02	0.90	1	6	10
Property Tax Rate (%)	1.10	0.53	0	1.0	4.0
Luxury Score (0–7)	1.36	1.29	0	1	7
Binary Variables (prevalence)					
Swimming Pool	39.8%				
Garage	67.2%				
Fireplace	39.6%				
Basement	24.3%				
Waterfront	8.9%				
HOA	37.1%				
Central Air	68.1%				
Hardwood Floors	26.7%				
Condominium	14.6%				
New Construction	5.9%				
Foreclosure	0.5%				
Regional Distribution					
South	56.6%				
West	18.9%				
Midwest	16.1%				
Northeast	8.4%				

Notes: Listing prices, not transaction prices. Property age computed as 2026 − year_built, after imputation. Imputed values included in summary statistics.

The median listing price is $399,900, with substantial right-skewness (mean $621,737). The log transformation substantially improves distributional symmetry (Figure 1), supporting the semi-log specification.

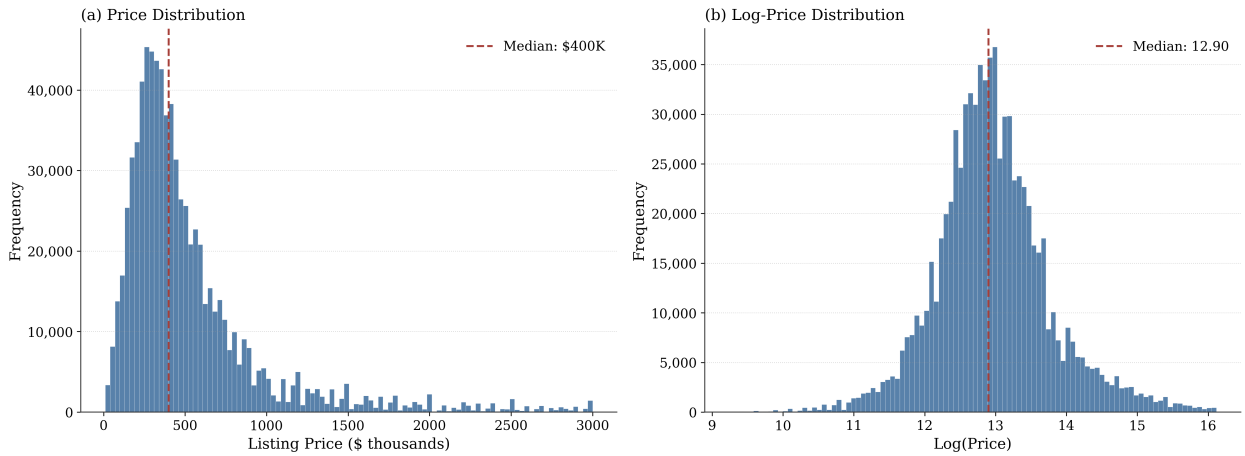

Figure 1. Distribution of Listing Prices: (a) Levels; (b) Natural Logarithm. The log transformation reduces right skewness and approximates normality.

4. Econometric and Machine Learning Methodology

4.1. Identification, Prediction, and Interpretation

Before presenting the estimation methods, we clarify the interpretive framework.

None of the models estimated in this paper identify causal willingness-to-pay parameters. OLS coefficients are conditional mean associations: they describe the average difference in log listing price between properties that differ by one unit in attribute x_k, holding other included attributes constant. Quantile regression coefficients are quantile-specific conditional associations. SHAP values are model-specific prediction decompositions. Table 4 summarizes these distinctions.

Table 4. Interpretation of Estimates Across Modeling Frameworks

	OLS Coefficient	Quantile Coefficient	SHAP Value
Object	Cond. mean association	Cond. quantile association	Prediction contribution
Interpretation	Semi-elasticity (if unstandardized log-log)	Quantile-specific semi-elasticity	Model-specific contribution
Causal?	No, unless identified	No, unless identified	No
Structural WTP?	Only under strong assumptions	Only under strong assumptions	No
Primary use	Inference	Distributional heterogeneity	Prediction interpretation

Cross-sectional data cannot separate preferences, supply constraints, and sorting. Regional and local omitted variables—including unobserved neighborhood quality, local supply conditions, and regulatory environment—remain major concerns. Throughout the paper, we interpret the estimates as **listing-price capitalization gradients** rather than structural demand parameters.

4.2. Semi-Log Hedonic Specification (OLS)

Our baseline model follows the semi-log specification:

$$\ln(P_i) = \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{j=1}^{J} \gamma_j d_{ij} + \sum_{m=1}^{M} \delta_m (x_{im_1} \cdot x_{im_2}) + \varepsilon_i \tag{1}$$

where P_i is the listing price, x_{ik} are continuous and binary attributes, d_{ij} are categorical dummies, and $x_{im_1} \cdot x_{im_2}$ are interaction terms.

Coefficient interpretation under standardization. All continuous regressors are standardized (zero mean, unit variance) prior to estimation. This facilitates coefficient magnitude comparison across variables with different scales but changes the interpretation: the coefficient on a standardized variable represents the association between a one-standard-deviation increase in the attribute and log-price, not a one-unit increase. In particular, the coefficient on standardized $\ln(\text{sqft})$ is *not* a price-to-area elasticity—it is the effect of a one-standard-deviation increase in log living area. To recover economic magnitudes in natural units, we also report an unstandardized specification.

Binary (0/1) and dummy variables are not standardized. For these, the percentage listing-price effect is $(e^{\hat{\beta}_k} - 1) \times 100\%$.

Standard errors are computed using the HC3 estimator (White, 1980, MacKinnon and White, 1985).

4.3. Quantile Regression

Quantile regression (Koenker and Bassett, 1978, Koenker and Hallock, 2001) models the τ-th conditional quantile:

$$Q_\tau(\ln P_i | \mathbf{x}_i) = \mathbf{x}_i' \boldsymbol{\beta}(\tau), \quad \tau \in (0, 1) \tag{2}$$

estimated by minimizing $\sum_i \rho_\tau(\ln P_i - \mathbf{x}_i' \boldsymbol{\beta})$, where $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$.

We estimate the model at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. For computational tractability, estimation uses a random subsample of 150,000 observations. We verify stability by repeating estimation across ten random subsamples with different seeds, finding that most coefficients exhibit coefficient-of-variation below 8% (Section 6.3).

Inter-quantile Wald tests. To formally test whether listing-price gradients differ between the lower and upper tails of the conditional distribution, we compute inter-quantile differences $\hat{\beta}_k(0.90) - \hat{\beta}_k(0.10)$ and test whether this difference is significantly different from zero using a z-test based on the covariance structure of the quantile regression estimates. Under the null hypothesis $H_0 : \beta_k(0.90) = \beta_k(0.10)$, the test statistic is:

$$z_k = \frac{\hat{\beta}_k(0.90) - \hat{\beta}_k(0.10)}{\sqrt{\mathrm{Var}[\hat{\beta}_k(0.90)] + \mathrm{Var}[\hat{\beta}_k(0.10)] - 2\,\mathrm{Cov}[\hat{\beta}_k(0.90), \hat{\beta}_k(0.10)]}} \tag{3}$$

Rejection indicates that the attribute's association with listing prices differs significantly between lower-priced and higher-priced properties (conditional on covariates).

4.4. Gradient-Boosted Ensemble Models

4.4.1. XGBoost

XGBoost (Chen and Guestrin, 2016) implements gradient boosting (Friedman, 2001), fitting an additive ensemble of regression trees by sequentially minimizing a regularized loss function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \tag{4}$$

where $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2 + \alpha\|w\|_1$ penalizes tree complexity. We use 1,000 trees, depth 8, learning rate 0.05, column and row subsampling ratios of 0.8, and early stopping after 50 rounds without improvement.

4.4.2. LightGBM

LightGBM (Ke et al., 2017) uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for computational efficiency. We use matching hyperparameters for comparability.

4.4.3. Additional Benchmarks

We also report results for Ridge regression ($\alpha = 1$), Lasso ($\alpha = 0.001$), Elastic Net, Random Forest (500 trees, depth 20; Breiman, 2001), and OLS with state fixed effects.

Ablation design. To understand which feature groups drive the ML predictive gain over OLS, we estimate XGBoost models using progressively richer feature sets: (i) structural attributes only (8 features); (ii) + lot characteristics (9); (iii) + amenity indicators (20); (iv) + neighborhood variables (27); (v) + market status (32); and (vi) the full model including interactions, categorical controls, and region dummies (62). At each stage, we evaluate both random and geographic holdout R^2 to identify which feature groups contribute to genuine predictive generalization versus spatial memorization.

4.5. Validation Designs

Because housing prices are spatially dependent, the choice of validation protocol is substantive rather than technical: random splits assess interpolation within observed markets, while spatially blocked designs assess extrapolation to new markets (Roberts et al., 2017, Valavi et al., 2019, Meyer and Pebesma, 2021). Blocked validation is not universally preferable—for interpolation objectives it can be pessimistically biased (Wadoux et al., 2021)—so we report both protocols and interpret each against its own deployment question. We employ three validation strategies:

1. **Random 80/20 split** (seed = 42): 631,073 training, 157,769 test. This is the standard approach but permits spatial leakage—nearby properties from the same neighborhood can appear in both train and test sets.
2. **Geographic holdout**: 10 states (CA, NY, TX, FL, OH, CO, NC, WA, IL, GA) are held out entirely; models are trained on the remaining 40 states + DC. This tests generalization to markets not seen during training. The held-out states represent the five largest (by sample size) plus five geographically diverse states, totaling 438,315 test observations.
3. **Ablation cascade**: XGBoost models are estimated at six stages of feature inclusion under both random and geographic validation, isolating the marginal contribution of each feature group.

The gap between random and geographic validation quantifies the extent to which models exploit local spatial structure rather than learning generalizable attribute-price relationships.

4.6. SHAP-Based Model Interpretation and Cross-Model Stability

SHAP values (Lundberg and Lee, 2017) decompose each prediction into additive feature contributions:

$$f(\mathbf{x}) = E[f(\mathbf{X})] + \sum_{k=1}^{K} \phi_k(\mathbf{x}) \tag{5}$$

We compute TreeSHAP values on a random subsample of 10,000 test observations.

Cross-model stability. A common criticism of SHAP-based interpretation is that feature importance rankings may be model-specific. To assess this concern, we compute mean absolute SHAP values for three tree-based models (XGBoost, LightGBM, Random Forest) and report the Spearman rank correlation between each pair. High cross-model correlation would support the use of SHAP rankings as a description of the predictive structure of the data, not merely an artifact of a particular model.

Interpretive caution. SHAP values are prediction-level contributions, not structural implicit prices. They are:

- model-dependent (XGBoost SHAP values differ from LightGBM SHAP values);

- affected by feature correlation (correlated features may share SHAP contributions);
- not causal—they decompose predictions, not data-generating processes;
- not equivalent to Rosen's hedonic price gradient $\partial P/\partial z_k$.

We use SHAP to describe which features contribute most to predictions, not to estimate marginal willingness to pay.

4.7. Spatial Autocorrelation Diagnostics

We compute Moran's I (Moran, 1950) on OLS residuals using a row-standardized KNN spatial weight matrix ($k = 8$) on random subsamples of 5,000 observations. To assess the stability of this diagnostic, we repeat the computation across three independent subsamples and report the mean, standard deviation, and significance of the resulting I statistics. Moran's I is defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(e_i - \bar{e})(e_j - \bar{e})}{\sum_i (e_i - \bar{e})^2} \tag{6}$$

where e_i are the OLS residuals, w_{ij} are the spatial weights, and N is the subsample size. Under the null of no spatial autocorrelation, I has expected value $-1/(N-1) \approx 0$ and an asymptotically normal distribution. Values significantly above zero indicate positive spatial autocorrelation (nearby residuals are similar), implying that the OLS specification fails to capture local price variation.

5. Estimation Results

5.1. OLS Baseline, State Fixed Effects, and ZIP3 Fixed Effects

5.1.1. Standardized Baseline Specification

Table 5 presents the OLS results with standardized continuous regressors. The model achieves $R^2 = 0.634$ with RMSE = 0.489 in log-price units. The F-statistic of 18,712.5 strongly rejects the null of jointly zero slope coefficients.

Table 5. OLS Hedonic Regression Results (Dep. Var.: ln(Listing Price))

Variable	Coefficient	Std. Error	
Panel A: Structural Attributes (standardized)			
ln(Living Area)†	0.3119	0.0035	***
Bedrooms†	−0.0675	0.0023	***
Bathrooms†	0.2359	0.0026	***
Age†	−0.0418	0.0139	**
Age^{2†}	0.0507	0.0019	***
Stories†	0.0004	0.0003	
Bath/Bed Ratio†	−0.0208	0.0021	***
Panel B: Lot and Amenities			
ln(Lot Size)†	0.0697	0.0010	***
Pool	0.0286	0.0026	***
Spa	0.0327	0.0027	***
Basement	−0.0424	0.0023	***
Garage	0.1094	0.0015	***
Waterfront	−0.1349	0.0325	***
Central Air	0.0693	0.0014	***
Hardwood Floors	0.1246	0.0014	***
Luxury Score†	0.0443	0.0018	***
Panel C: Neighborhood (standardized)			
Walk Score†	−0.0109	0.0012	***
Bike Score†	0.0717	0.0010	***
Transit Score†	0.0524	0.0008	***
Avg. School Rating†	−0.0491	0.0007	***
Property Tax Rate†	−0.0292	0.0008	***

Notes: Table continues on the next page; estimation notes follow the continuation.

Table 5. OLS Hedonic Regression Results (continued)

Variable	Coefficient	Std. Error	
Panel D: Market Status			
Condominium	−0.0370	0.0037	***
HOA Member	−0.1077	0.0024	***
ln(HOA Fees + 1)†	0.0439	0.0014	***
New Construction	0.1035	0.0029	***
Foreclosure	−0.4064	0.0098	***
Panel E: Interactions (standardized)			
ln(sqft) × Age†	−0.1100	0.0140	***
Pool × South	0.0081	0.0012	***
Waterfront × ln(sqft)†	0.1371	0.0092	***
Basement × North	−0.0255	0.0012	***
Condo × Walk Score†	0.0816	0.0013	***
Age × Luxury†	0.0440	0.0015	***
Panel F: Regional Dummies (ref: Midwest)			
Northeast	0.4869	0.0030	***
South	0.0220	0.0028	***
West	0.4008	0.0031	***
R^2	0.6343		
Adj. R^2	0.6342		
RMSE	0.4887		
F-statistic	18,712.5		
N	788,842		

Notes: HC3 heteroskedasticity-robust standard errors. †Variables standardized (zero mean, unit variance) before estimation; coefficients represent effects of a one-standard-deviation increase. Binary variables not standardized; percentage effects computed as $100 \times (e^{\beta} - 1)$. Categorical controls (roof: 6, construction: 8, foundation: 7 dummies) included but not shown. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

5.1.2. Unstandardized Specification for Economic Interpretation

Table 6 reports key coefficients from the unstandardized specification, which permits direct economic interpretation. The R^2 is identical (0.634); only the coefficient magnitudes and interpretations differ.

Table 6. Unstandardized OLS: Key Coefficients in Natural Units

Variable	Coefficient	Std. Error	Interpretation	
Constant	7.7316	0.0495		***
ln(Living Area)	0.6281	0.0070	Elasticity: 0.63	***
Bedrooms	−0.0620	0.0022	−6.0% per additional bedroom	***
Bathrooms	0.2035	0.0023	+22.6% per additional bathroom	***
Age (years)	−0.0013	0.0004	−0.13% per year	**
Age^2 ($\times 10^{-5}$)	1.2	0.05	Positive quadratic (vintage)	***
ln(Lot Size)	0.0182	0.0003	Elasticity: 0.02	***
Walk Score (per point)	−0.0004	0.0000	−0.04% per point	***
Bike Score (per point)	0.0040	0.0001	+0.40% per point	***
Transit Score (per point)	0.0029	0.0000	+0.29% per point	***
Avg. School Rating (per point)	−0.0544	0.0008	−5.3% per rating point	***
Property Tax Rate (per ppt)	−0.0553	0.0015	−5.4% per ppt	***
Luxury Score (per unit)	0.0344	0.0014	+3.5% per unit	***

Notes: Same sample and specification as Table 5, but continuous variables enter in natural units. N = 788,842. HC3 standard errors.

The elasticity of listing price with respect to living area is 0.63: a 10% increase in square footage is associated with a 6.3% higher listing price. This is below unity, consistent with diminishing marginal returns to space. Each additional bathroom is associated with a 22.6% higher listing price, while each additional bedroom—conditional on total area—is associated with a 6.0% *lower* price, reflecting the well-documented tradeoff between room count and room size (Sirmans et al., 2005).

5.1.3. Counterintuitive Coefficient Signs

Several coefficients warrant discussion.

Negative waterfront coefficient (−0.135**, standardized).** The standalone waterfront effect is evaluated at the mean of the standardized interaction term waterfront × ln(sqft). Because the interaction is strongly positive (+0.137), the net waterfront effect becomes positive for larger properties. This artifact of the interacted specification does not imply that waterfront reduces price on average.

Negative school rating (−0.054 **per rating point, unstandardized).** This likely reflects confounding with property tax rates and regional effects. In higher-tax jurisdictions, school quality is partially capitalized through the tax rate, which enters separately. Conditional on tax rate, region, and other controls, the residual school rating variation may capture unobserved factors that correlate negatively with prices. The boundary-discontinuity literature, which isolates school quality from neighborhood composition, consistently finds positive valuations (Black, 1999, Bayer et al., 2007); the divergence illustrates why cross-sectional partial coefficients on school quality should not be interpreted structurally.

Negative Walk Score (−0.0004 **per point, unstandardized).** Walk Score is highly correlated with Bike Score ($r \approx 0.56$) and Transit Score ($r \approx 0.62$). In the presence of all three accessibility measures, the Walk Score coefficient may reflect residual urban density effects—after controlling for biking and transit access, the remaining variation in walkability may capture denser, smaller-lot neighborhoods. A walkability premium is well documented when accessibility enters alone (Pivo and Fisher, 2011); with three collinear accessibility scores, the premium is split across coefficients and cannot be attributed variable-by-variable.

Negative basement (−0.042)**.** Basements are predominantly found in older properties in colder regions. Conditional on age, region, and size, the basement indicator may proxy for older construction quality or layout features valued less in modern markets.

Negative HOA membership (−0.108)**.** HOA properties include condominiums and planned communities that may have lower lot sizes and more restrictions. The negative sign is conditional on the separate log HOA fee variable (+0.044), suggesting that the HOA membership dummy captures the property-type effect while fees capture the service-quality gradient.

These counterintuitive signs are common in hedonic regressions with many correlated controls and do not necessarily indicate misspecification, though they do limit the attributable economic interpretation of individual coefficients.

5.1.4. Progressive Geographic Fixed Effects

Table 7 documents how explanatory power increases with geographic granularity. Adding 50 state fixed effects increases the in-sample R^2 from 0.634 to 0.681. Replacing state dummies with 886 ZIP3 fixed effects further increases R^2 to 0.729 (in-sample) and out-of-sample R^2 to 0.725.

Table 7. OLS R^2 Under Progressively Finer Geographic Controls

Specification	Geographic FE	N_{FE}	In-sample R^2	Out-of-sample R^2
Baseline (4 regions)	Census Region	3	0.634	0.630
State FE	State	50	0.681	0.678
ZIP3 FE	3-digit ZIP	886	0.729	0.725

Notes: Same 62 non-geographic regressors across all specifications; geographic FE replace the 3 regional dummies. Out-of-sample R^2 computed on a random 20% holdout. ZIP3 codes cover 886 unique prefixes in the sample.

The progression from region to state to ZIP3 fixed effects yields R^2 gains of +4.7 pp and +4.7 pp, respectively. This 9.5 pp total gain from geographic granularity alone—achieved without any additional property-level information—underscores the importance of unobserved local factors in hedonic pricing. Notably, even ZIP3 fixed effects leave substantial residual variation unexplained: the gap between ZIP3 OLS ($R^2 = 0.725$) and XGBoost under random validation ($R^2 = 0.833$) suggests that non-linear attribute effects contribute an additional 10.8 pp of explained variation.

5.1.5. OLS Diagnostics

Figure 2 presents diagnostic plots. The residual distribution exhibits excess kurtosis (5.55 vs. 3.0 for normality) and mild positive skewness (0.18), as confirmed by the Jarque-Bera test ($JB = 217{,}078$, $p < 0.001$). The heteroskedastic pattern visible in the residuals-versus-fitted plot motivates our use of HC3 standard errors.

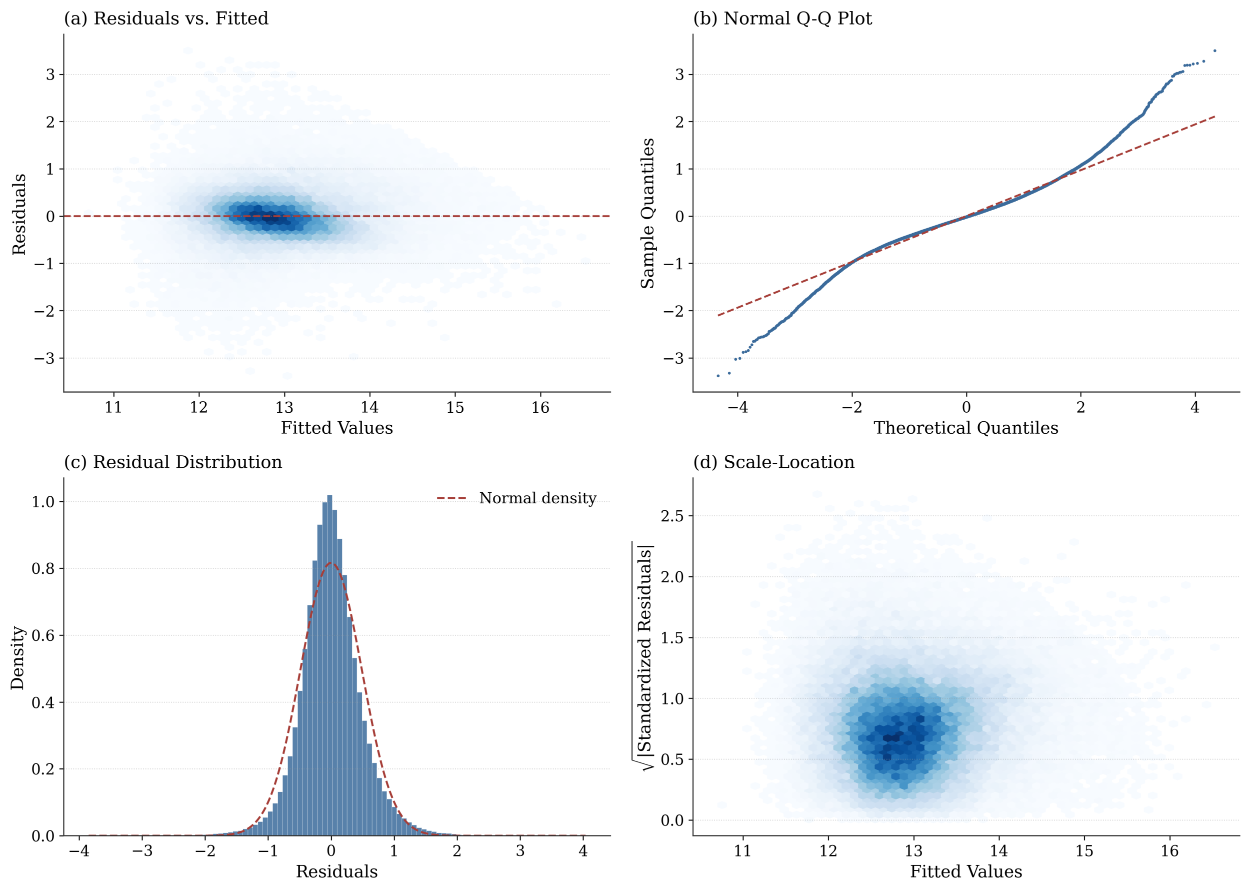

Figure 2. OLS Diagnostic Plots: (a) Residuals vs. Fitted; (b) Normal Q-Q Plot; (c) Residual Distribution; (d) Scale-Location

5.2. Spatial Autocorrelation

We compute Moran's I on OLS residuals using a KNN spatial weight matrix ($k = 8$) on three independent random subsamples of 5,000 observations each. Table 8 reports the results.

Table 8. Moran's I on OLS Residuals: Subsample Stability

Subsample	I	z-statistic	p-value	
Subsample 1	0.2742	41.8	< 0.001	***
Subsample 2	0.2839	43.5	< 0.001	***
Subsample 3	0.2652	40.5	< 0.001	***
Mean	0.2745			
Std. Dev.	0.0076			

Notes: Row-standardized KNN weight matrix with $k = 8$. Each subsample draws 5,000 observations at random. All z-statistics strongly reject the null of no spatial autocorrelation. The low standard deviation (0.008) across subsamples indicates that the diagnostic is highly stable.

The mean Moran's $I = 0.2745$ (std = 0.0076) across three subsamples confirms **strong, stable positive spatial autocorrelation** in the OLS residuals. This implies that our regional controls (four Census regions) are insufficient to absorb local price variation. OLS standard errors may understate true uncertainty, and coefficient estimates may be biased if the spatial structure correlates with included regressors. The stability across subsamples (CV = 2.8%) indicates that this is not a sampling artifact but a structural feature of the data. Addressing spatial dependence through spatial econometric models is an important direction for future work; our OLS and quantile regression results should be interpreted with this caveat.

5.3. Quantile Regression and Inter-Quantile Tests

5.3.1. Coefficient Estimates Across Quantiles

Table 9 presents quantile regression coefficients at the five estimated quantiles.

Table 9. Quantile Regression Coefficients at Selected Quantiles

Variable	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	OLS
ln(Living Area)	0.358***	0.313***	0.275***	0.237***	0.251***	0.312***
Bedrooms	−0.085***	−0.076***	−0.064***	−0.037***	−0.034***	−0.068***
Bathrooms	0.157***	0.195***	0.234***	0.259***	0.259***	0.236***
Age	−0.430***	−0.201***	−0.066**	−0.007	−0.026	−0.042**
ln(Lot Size)	0.045***	0.051***	0.061***	0.070***	0.086***	0.070***
Pool	0.008	0.008	0.019***	0.061***	0.094***	0.029***
Garage	0.205***	0.146***	0.102***	0.061***	0.020***	0.109***
Luxury Score	0.018***	0.038***	0.054***	0.068***	0.072***	0.044***
Foreclosure	−0.494***	−0.473***	−0.379***	−0.327***	−0.338***	−0.406***
Northeast	0.339***	0.395***	0.471***	0.553***	0.603***	0.487***
West	0.365***	0.372***	0.376***	0.399***	0.450***	0.401***

Notes: Estimated on a random subsample of 150,000 observations. Continuous variables standardized. Full set of 62 regressors included; selected coefficients shown. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Coefficients reflect one-standard-deviation effects for continuous variables.

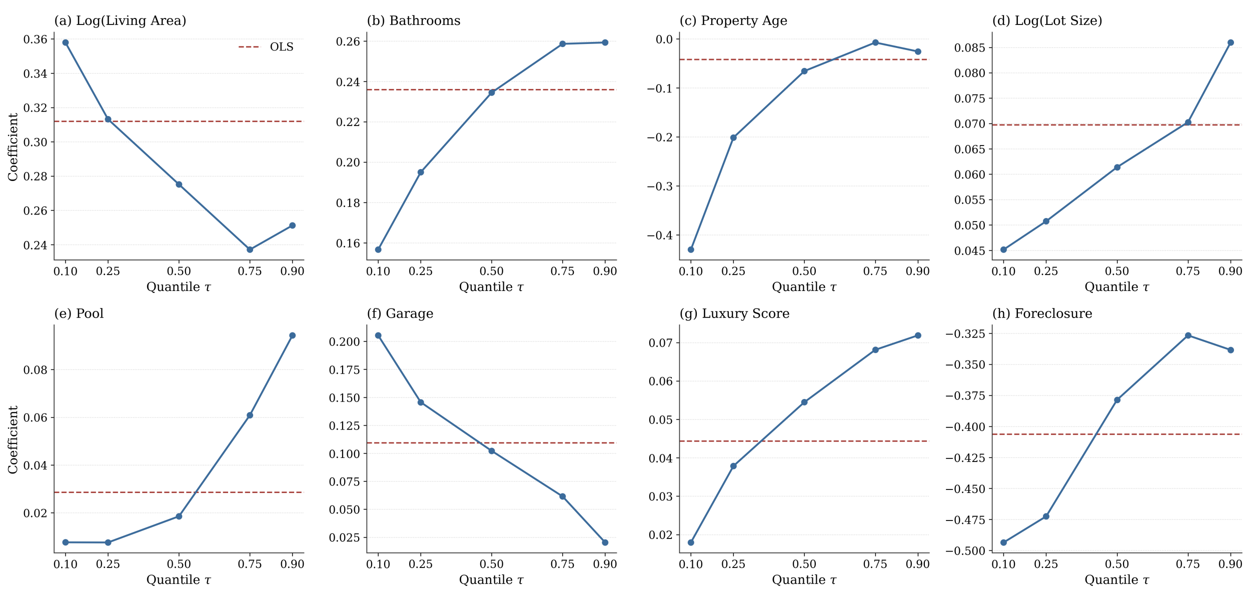

Figure 3. Quantile Regression Coefficients Across the Conditional Price Distribution. Red dashed lines indicate OLS estimates. Coefficients are on standardized variables.

Several economically interesting patterns emerge.

Living area gradient declines at upper quantiles. The standardized ln(sqft) coefficient declines from 0.358 ($\tau = 0.10$) to 0.237 ($\tau = 0.75$), indicating that additional space is associated with proportionally larger listing-price differences at the lower end of the conditional distribution. At higher quantiles, other factors (luxury features, location) may dominate size.

Bathroom gradient increases at upper quantiles. In contrast, the bathroom gradient rises monotonically from 0.157 ($\tau = 0.10$) to 0.259 ($\tau = 0.90$), suggesting that bathroom count differentiates properties more strongly in the upper market.

Age penalty concentrated at lower quantiles. The age gradient is -0.430 at $\tau = 0.10$ but statistically insignificant at $\tau = 0.75$ and $\tau = 0.90$. Age-related depreciation appears to be primarily a phenomenon of lower-priced properties, possibly because higher-priced properties are better maintained or benefit from vintage appeal.

Pool gradient emerges only above the median. The pool coefficient is insignificant at $\tau = 0.10$ and $\tau = 0.25$ but rises to 0.094 at $\tau = 0.90$ ($\approx 9.9\%$ listing-price difference), consistent with pools being a luxury amenity.

Garage gradient declines monotonically. The garage coefficient drops from 0.205 ($\tau = 0.10$) to 0.020 ($\tau = 0.90$), reflecting that garage availability differentiates lower-priced properties much more than upper-priced ones where garages are nearly universal.

OLS averages across these heterogeneous gradients and can therefore misrepresent both the lower and upper segments of the market.

5.3.2. Inter-Quantile Wald Tests ($\tau = 0.10$ vs. $\tau = 0.90$)

Table 10 reports the inter-quantile Wald tests for equality of coefficients between the 10th and 90th percentiles.

Table 10. Inter-Quantile Wald Tests: $\hat{\beta}(0.10) - \hat{\beta}(0.90)$

Variable	$\hat{\beta}(0.10) - \hat{\beta}(0.90)$	z-statistic	p-value	
Garage	+0.185	28.92	< 0.001	***
Northeast	−0.264	−21.43	< 0.001	***
Bathrooms	−0.103	−10.48	< 0.001	***
ln(Living Area)	+0.107	9.52	< 0.001	***
ln(Lot Size)	−0.041	−9.49	< 0.001	***
Age	−0.405	−7.89	< 0.001	***
Pool	−0.087	−7.80	< 0.001	***
Luxury Score	−0.054	−6.82	< 0.001	***
West	−0.086	−6.26	< 0.001	***
Bedrooms	−0.051	−5.95	< 0.001	***
Foreclosure	−0.155	−4.18	< 0.001	***
Not significant ($p > 0.05$)				
Waterfront	+0.222	1.86	0.063	
Age^2	+0.010	1.40	0.162	

Notes: z-statistics computed from the joint covariance matrix of the $\tau = 0.10$ and $\tau = 0.90$ quantile regression estimates; both quantiles estimated on the same 150,000-observation subsample. Variables sorted by $|z|$. 11 of 13 tested variables show statistically significant differences at the 5% level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

The inter-quantile tests reject coefficient equality for 11 of 13 tested variables at the 5% level. The largest inter-quantile difference is for garage (+0.185, $z = 28.92$), confirming that the garage listing-price gradient is dramatically larger for lower-priced properties. Bathrooms (−0.103, $z = -10.48$) and living area (+0.107, $z = 9.52$) also exhibit large, highly significant inter-quantile differences. The only two variables for which the null of equal coefficients cannot be rejected are age-squared ($z = 1.40$) and waterfront ($z = 1.86$, $p = 0.063$).

These formal tests confirm the visual patterns in Figure 3: the conditional distribution of listing prices is not merely shifted by housing attributes but is differentially stretched and compressed, with different attributes mattering more at different points in the distribution.

5.4. Machine Learning Under Random Validation

Table 11 summarizes out-of-sample performance under the random 80/20 split.

Table 11. Out-of-Sample Predictive Performance: Random Validation

	Log-Price Scale			Dollar Scale		
Model	R^2	RMSE	MAE	MAPE (%)	MdAPE (%)	MAE ($)
OLS	0.6301	0.4911	0.3619	39.8	27.2	240,941
Ridge (α = 1)	0.6301	0.4911	0.3619	—	—	—
Lasso (α = 0.001)	0.6280	0.4925	0.3627	—	—	—
Elastic Net	0.6145	0.5014	0.3681	—	—	—
OLS + State FE	0.6775	0.4586	—	—	—	—
OLS + ZIP3 FE	0.7253	—	—	—	—	—
Random Forest	0.7842	0.3751	0.2636	28.4	18.5	180,116
LightGBM	0.8088	0.3531	0.2519	26.9	18.2	169,495
XGBoost	**0.8330**	**0.3300**	**0.2327**	**24.7**	**16.5**	**157,822**

Notes: Training: 631,073 (80%). Test: 157,769 (20%). Random split, seed 42. MAPE = mean absolute percentage error; MdAPE = median APE. Dollar metrics computed by exponentiating predicted log-prices. Ridge, Lasso, Elastic Net dollar metrics suppressed for brevity (similar to OLS). ZIP3 FE RMSE and dollar-scale metrics not computed for this specification.

XGBoost achieves the highest R^2 of 0.833 under random validation, representing a 20-percentage-point improvement over OLS. Regularized linear models (Ridge, Lasso, Elastic Net) offer no improvement over OLS, confirming that the parametric specification is not overfit—the predictive gap is driven by non-linearity and interaction effects, not overfitting. The ZIP3 FE model (R^2 = 0.725) closes roughly half the OLS-to-XGBoost gap, indicating that fine-grained geographic controls capture a substantial portion of the non-linearity that tree models exploit.

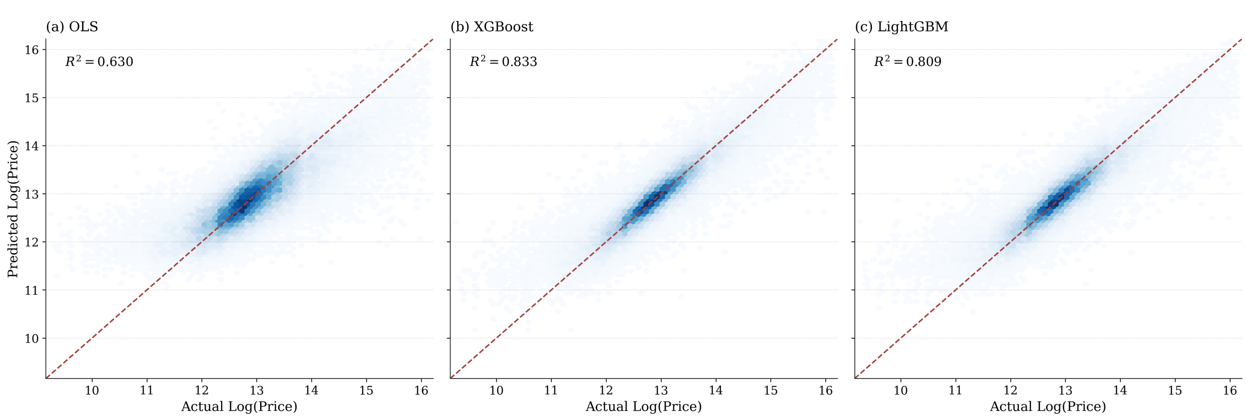

Figure 4. Predicted vs. Actual Log-Prices Under Random Validation: (a) OLS; (b) XGBoost; (c) LightGBM

5.5. Machine Learning Under Geographic Validation

Table 12 reports performance when 10 states are held out entirely.

Table 12. Geographic Holdout Validation (10 States Held Out)

Model	R^2 (random)	R^2 (geo holdout)	ΔR^2
OLS	0.630	< 0	> -0.63
Random Forest	0.784	0.542	−0.242
XGBoost	0.833	0.547	−0.286
XGBoost (no geo features)	0.830	0.519	−0.311
XGBoost (+ lat/lon)	0.870	0.464	−0.406

Notes: Geographic holdout: CA, NY, TX, FL, OH, CO, NC, WA, IL, GA held out (438,315 observations). Training on remaining 350,527 observations. OLS with regional dummies fails catastrophically because the held-out states span all four Census regions and include the largest markets. "No geo features" removes region dummies from the full model. "+ lat/lon" adds raw latitude and longitude as continuous features.

Key finding. The XGBoost R^2 drops from 0.833 (random) to 0.547 (geographic holdout)—a **28.6-percentage-point decline**. This demonstrates that a substantial portion of the random-validation R^2 reflects the model's ability to exploit local spatial structure rather than learning generalizable attribute-price relationships.

Two additional experiments sharpen this conclusion:

- **Adding latitude and longitude** boosts random R^2 from 0.833 to 0.870 (+3.7 pp) but *reduces* geographic holdout R^2 to 0.464 (−8.3 pp relative to the baseline XGBoost). Coordinates allow the model to memorize location-specific price levels, which is useful when nearby properties appear in the test set but harmful when entire states are withheld.
- **Removing all geographic features** (no region dummies) yields random $R^2 = 0.830$ (essentially unchanged) but geographic holdout $R^2 = 0.519$. Compared to the ablation-design variant of the full model (Geo $R^2 = 0.425$; see Section 5.6), dropping geography *improves* geographic holdout by +9.4 pp. Region dummies can *harm* geographic generalization because they encode training-set-specific geographic price levels.

This finding is methodologically important: **random train-test splits can substantially overstate the predictive performance of hedonic models** when applied to geographically diverse national samples. Geographic holdout validation provides a more conservative—and arguably more relevant—assessment of model generalizability.

5.6. Ablation: What Drives the ML Gain?

Table 13 presents the results of the feature ablation analysis for XGBoost.

Table 13. XGBoost Ablation Analysis: Marginal Contribution of Feature Groups

Stage	Feature Set	N_{features}	Random R^2	Geo R^2	Δ Geo R^2
1	Structural only	8	0.4976	0.3481	—
2	+ Lot	9	0.5464	0.3765	+0.028
3	+ Amenities	20	0.6378	0.4485	+0.072
4	+ Neighborhood	27	0.8136	0.4833	+0.035
5	+ Market status	32	0.8235	0.5014	+0.018
6	Full model (+ interactions/cats/region)	62	0.8327	0.4250	−0.076

Notes: Each row adds a feature group to the previous stage. Random R^2 is computed on the standard 20% random holdout. Geo R^2 is computed on the 10-state geographic holdout. Δ Geo R^2 is the change from the previous stage.

Three findings emerge from the ablation analysis.

Neighborhood features provide the largest random-validation gain. Adding neighborhood variables (Walk Score, Bike Score, Transit Score, school rating, school count, school distance, tax rate) increases random R^2 from 0.638 to 0.814—a gain of +17.6 pp. This is more than twice the contribution of any other feature group, reflecting the importance of local amenities and public-service quality in explaining listing-price variation.

The full model hurts geographic generalization. The transition from Stage 5 (32 features, Geo $R^2 = 0.501$) to Stage 6 (62 features, Geo $R^2 = 0.425$) *reduces* geographic holdout R^2 by 7.6 pp. The features added in Stage 6 include region dummies, interaction terms involving region, and categorical controls. These features improve random R^2 modestly (+0.9 pp) but actively harm geographic generalization by encoding training-set-specific spatial patterns.

Structural features alone explain 35% of geographic variation. Even with only 8 structural variables, XGBoost achieves Geo $R^2 = 0.348$, indicating that basic property characteristics (size, bedrooms, bathrooms, age) carry non-trivial predictive power across geographies.

5.7. The Role of Geographic Features

The ablation and geographic holdout results jointly demonstrate a critical tension in hedonic modeling: geographic features improve in-sample and random-holdout fit but degrade geographic generalization. Table 14 summarizes the evidence.

Table 14. Impact of Geographic Features on Model Performance

XGBoost Specification	Random R^2	Geographic Holdout R^2
Full model (62 features, with region)	0.8327	0.4250
No geographic features (no region dummies)	0.8297	0.5190
Full model + lat/lon coordinates	0.8701	0.4644

Notes: "No geographic features" removes region dummies from the full model. "+ lat/lon" adds raw latitude and longitude as continuous features to the full model. Geographic holdout uses 10 states held out entirely.

Removing region dummies costs only 0.3 pp in random R^2 (0.833 to 0.830) but *gains* 9.4 pp in geographic holdout R^2 (0.425 to 0.519). Adding coordinates achieves the opposite: +3.7 pp random, −5.5 pp geographic holdout (relative to the no-geography specification).

This pattern has important implications for automated valuation models (AVMs). In deployment contexts where the model must generalize to new geographies, geographic features can be counterproductive. The model without geography exploits only attribute-price relationships that transfer across markets, yielding more conservative but more robust predictions.

5.8. SHAP Analysis and Cross-Model Stability

5.8.1. Feature Importance Rankings

Table 15 reports the top 20 features by mean absolute SHAP value from the XGBoost model.

Table 15. Top 20 Features by Mean Absolute SHAP Value (XGBoost)

Rank	Feature	Mean $|\phi|$
1	ln(Living Area)	0.1740
2	Bathrooms	0.1610
3	Avg. School Rating	0.0881
4	Region: West	0.0751
5	ln(Lot Size)	0.0747
6	Luxury Score	0.0713
7	Property Tax Rate	0.0694
8	Nearest School Distance	0.0472
9	Region: Northeast	0.0425
10	Bike Score	0.0400
11	ln(HOA Fees + 1)	0.0398
12	Garage	0.0378
13	Age	0.0369
14	Hardwood Floors	0.0311
15	Waterfront $\times$ ln(sqft)	0.0292
16	Sqft per Bedroom	0.0289
17	Walk Score	0.0276
18	Central Air	0.0273
19	Region: South	0.0250
20	Transit Score	0.0213

Notes: TreeSHAP values computed on 10,000 random test-set observations. Mean $|\phi|$ is the average absolute contribution of each feature to deviations from the expected predicted log-price. These are predictive contributions, not structural implicit prices.

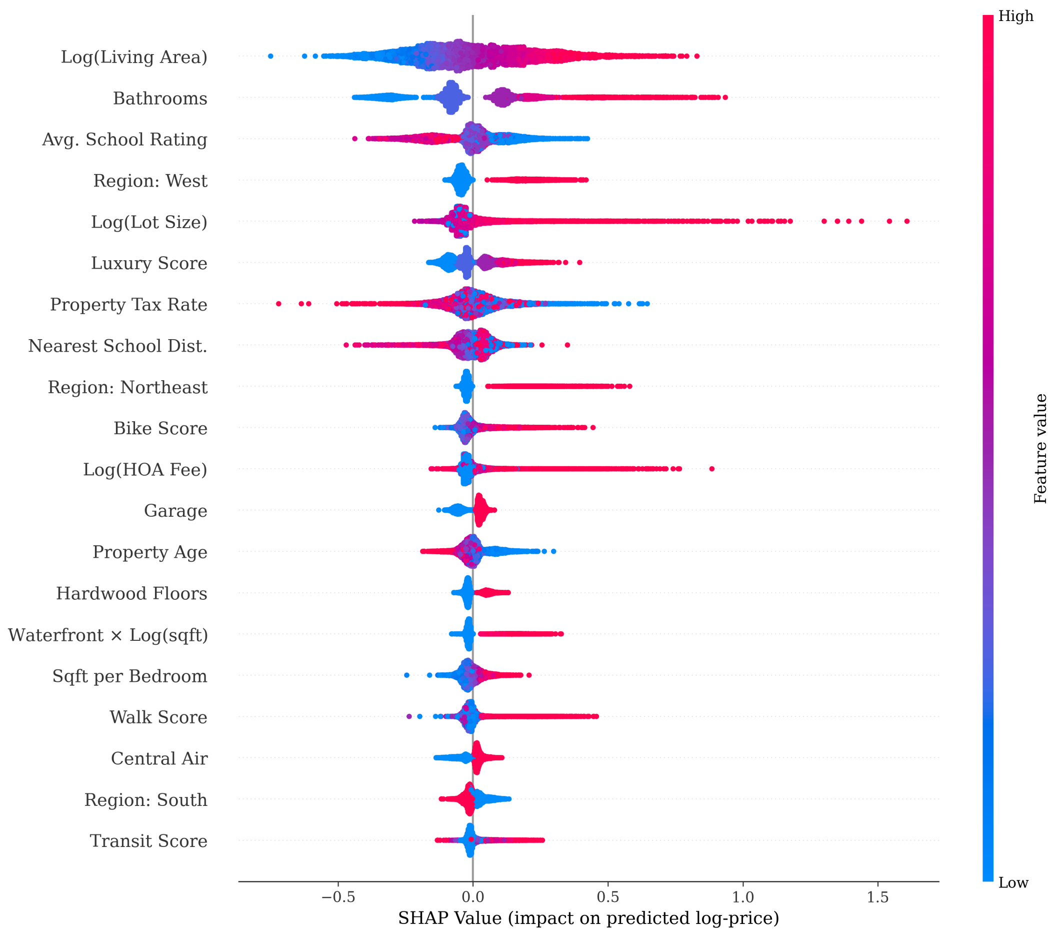

Figure 5. SHAP Summary Plot for XGBoost. Each dot is one observation; horizontal position shows the feature's SHAP value (contribution to predicted log-price deviation from the mean); color indicates feature value (red = high, blue = low). These are predictive contributions, not causal effects.

Living area and bathrooms are the two dominant predictive contributors, with mean absolute SHAP values of 0.174 and 0.161. School quality (0.088), regional location (West: 0.075; Northeast: 0.043), and lot size (0.075) form a second tier.

SHAP dependence plots. Figure 6 shows how the SHAP contribution of key features varies with feature values. The non-linear shapes—particularly the strongly concave living-area SHAP and the threshold-like behavior of school rating and property tax rate—explain why tree-based models outperform linear OLS. These non-linearities cannot be captured by the semi-log specification even with interaction terms.

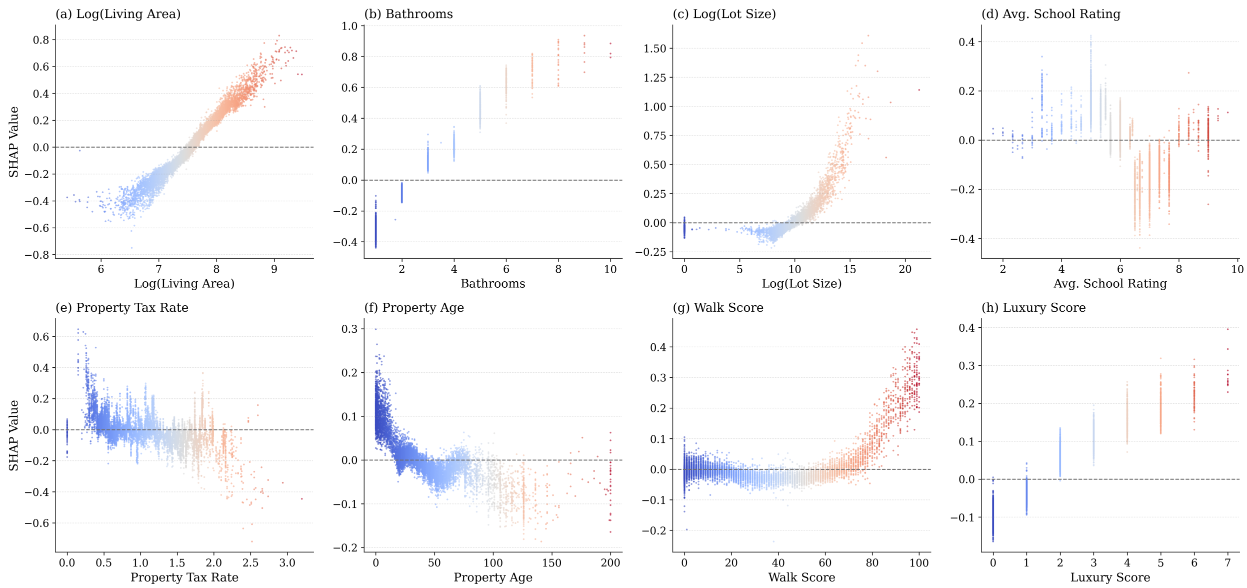

Figure 6. SHAP Dependence Plots for Eight Key Features. Each panel shows how the SHAP contribution varies with the feature value. Non-linear patterns (concavity, thresholds, saturation) explain the predictive advantage of tree-based models over linear specifications.

5.8.2. *Cross-Model SHAP Stability*

A common concern with SHAP-based interpretation is that rankings may be model-specific artifacts. Table 16 reports pairwise Spearman rank correlations of mean absolute SHAP values across three tree-based models.

Table 16. Cross-Model SHAP Importance Stability: Spearman Rank Correlations

	XGBoost	LightGBM	Random Forest
XGBoost	1.000	0.9898	0.9060
LightGBM	0.9898	1.000	0.8917
Random Forest	0.9060	0.8917	1.000

Notes: Spearman ρ computed on the full vector of mean $|\phi|$ values across all features. The same six features—ln(Living Area), Bathrooms, Avg. School Rating, Region: West, ln(Lot Size), and Luxury Score—rank in the top six across all three models.

The cross-model correlations are remarkably high. XGBoost and LightGBM produce nearly identical rankings (ρ = 0.990), while even the structurally different Random Forest model yields ρ > 0.89 with both gradient-boosted methods. The same six features appear in the top six across all three models. This stability lends credibility to the SHAP-based feature importance ranking as a description of the predictive structure of the data, not merely an artifact of a particular algorithm.

Note on SHAP versus OLS comparisons. SHAP importance rankings differ from both OLS t-statistics and XGBoost gain importance. School rating ranks 3rd by SHAP but has a counterintuitive negative OLS sign; property tax rate ranks 7th by SHAP but 28th by standardized OLS coefficient magnitude. These discrepancies reflect the non-linear and interactive effects that SHAP captures but OLS cannot. However, even though SHAP rankings are stable across tree-based models, they should not be interpreted as revealing the "true" importance hierarchy of housing attributes—they remain predictive decompositions, not structural parameters.

5.9. Geographic Heterogeneity

Figure 7 maps the spatial distribution of listing prices. The well-documented coastal price gradient is clearly visible, with California, the Northeast corridor, Hawaii, and Colorado exhibiting the highest prices. The map also illustrates the non-representativeness concern: listing density is highly uneven, with sparse coverage in some rural areas.

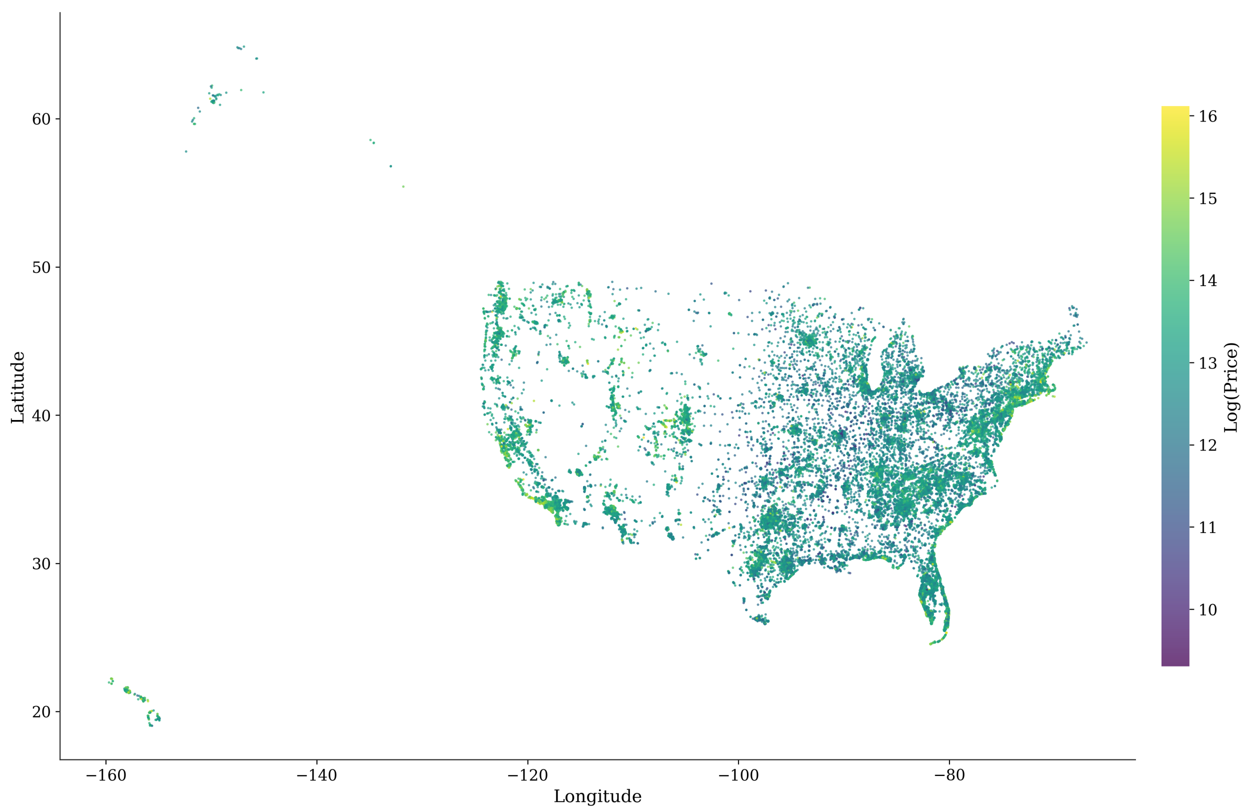

Figure 7. Geographic Distribution of Log Listing Prices. Each dot represents one property (50,000 random subsample). Color indicates log-price. Note: listing density is not uniform; areas with few dots may have fewer Zillow listings, not necessarily fewer properties.

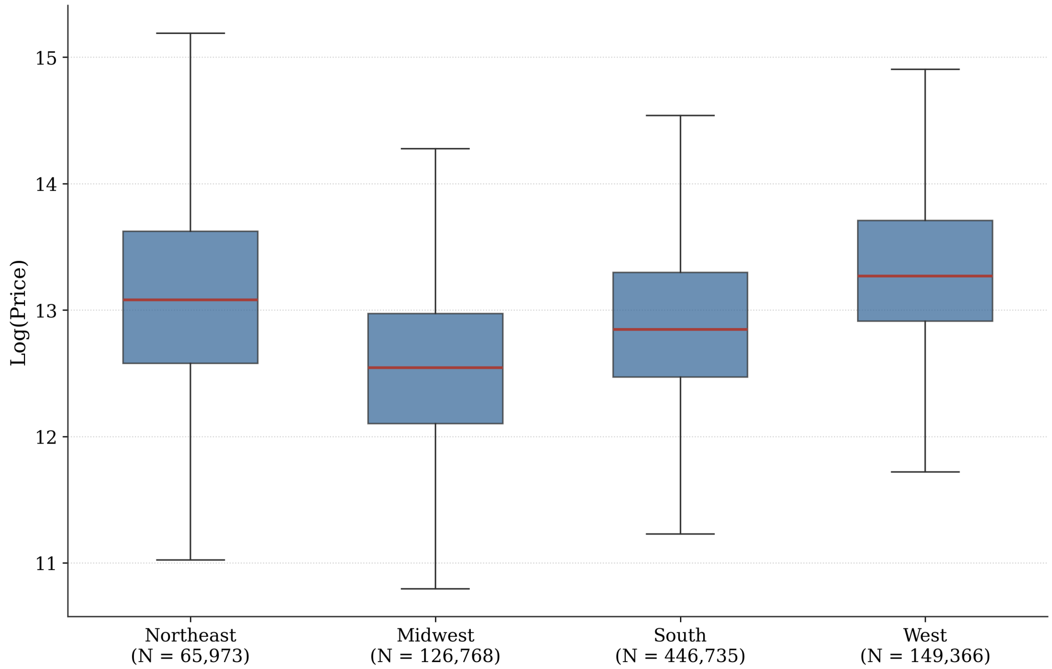

Figure 8. Listing Price Distribution by Census Region

5.10. Non-Linear Relationships

Figure 9 presents bivariate (unconditional) relationships between key attributes and log listing prices. These plots illustrate the non-linearities that motivate the machine learning approach but should not be confused with conditional (ceteris paribus) effects.

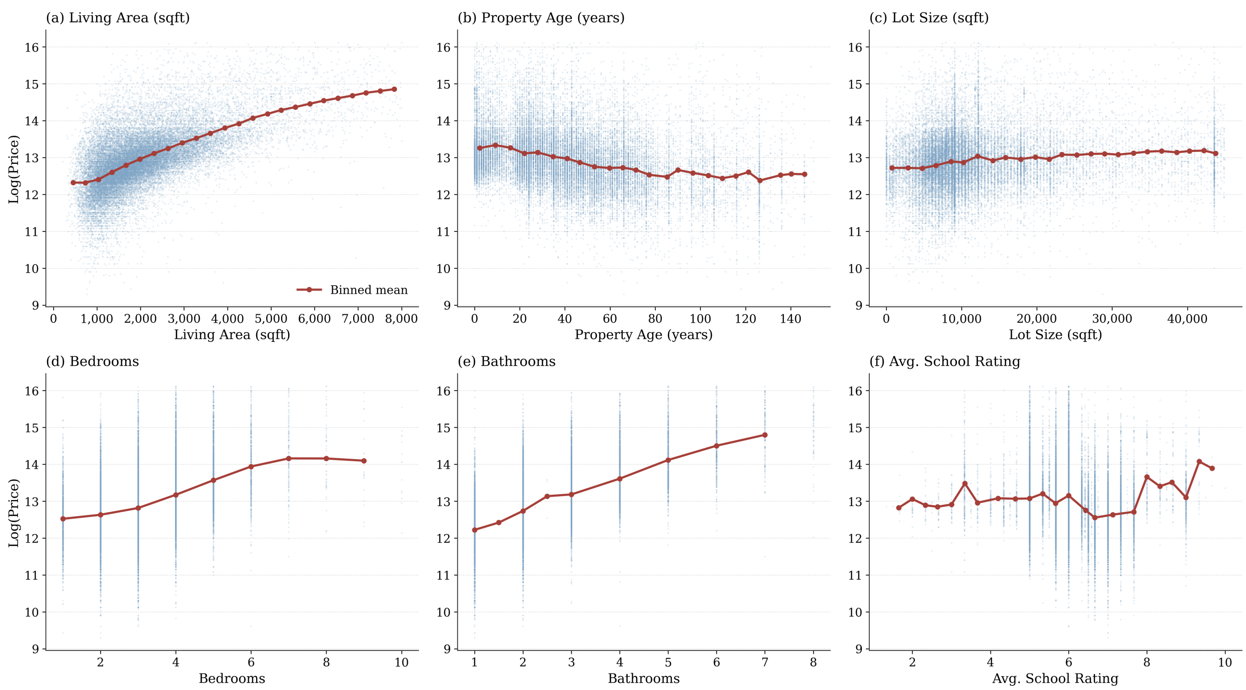

Figure 9. Unconditional Bivariate Relationships Between Key Attributes and Log Listing Price. These are raw scatter plots and binned means; they do not control for other attributes and should not be interpreted as conditional effects.

6. Robustness Checks

6.1. Imputation Sensitivity

Two variables with substantial missingness—year built (19.4%) and lot size (16.3%)—are imputed using state-level medians. To assess the consequences, we re-estimate the baseline OLS model on subsamples that exclude imputed observations. Table 17 reports the results.

Table 17. Imputation Sensitivity: Unstandardized OLS on Complete-Case Subsamples

Specification	N	R^2	$\hat{\beta}_{\ln(\text{lot})}$	Notes
Full sample (baseline)	788,842	0.634	0.018	All imputed values included
Drop missing year_built	640,055	0.638	0.019	19% fewer obs.
Drop missing lot_size	664,900	0.651	0.059	16% fewer obs.
Drop both missing	534,803	0.654	0.059	32% fewer obs.

Notes: All specifications use the same 62-regressor unstandardized OLS model with HC3 standard errors. $\hat{\beta}_{\ln(\text{lot})}$ is the coefficient on ln(lot size) in natural units. Other coefficients are broadly stable across specifications (not shown for brevity).

The key finding is that the **lot-size coefficient triples** from 0.018 to 0.059 when imputed lot sizes are dropped. This suggests that state-median imputation attenuates the lot-size gradient substantially: imputed properties receive the state median lot size regardless of their actual lot, washing out the true lot-price relationship. The R^2 also increases modestly (+2.0 pp) when imputed lots are excluded, indicating that imputation introduces noise.

Other coefficients (living area, bathrooms, age) are broadly stable across subsamples, suggesting that the imputation concern is primarily concentrated in the lot-size variable. Dropping observations with missing year built has minimal effect on the lot-size coefficient (0.018 to 0.019) and only slightly increases R^2 (+0.4 pp).

This sensitivity finding has implications for interpretation: the baseline elasticity of listing price with respect to lot size (0.02) likely understates the true association, and the complete-case estimate (0.06) may be more informative—though it applies to a selected (non-randomly missing) subsample.

6.2. Winsorization Sensitivity

To assess the influence of outliers, we winsorize the two variables with implausible extreme values—lot size, capped at its 99.5th percentile (3,944 observations affected), and Bike Score, capped at its nominal maximum of 100 (38 observations)—and re-estimate the baseline OLS model.

Table 18. Winsorization Sensitivity: Baseline vs. Winsorized OLS

Metric / Coefficient	Baseline	Winsorized
R^2	0.634	0.640
$\hat{\beta}_{\ln(\text{lot})}$	0.018	0.058
$\hat{\beta}_{\ln(\text{sqft})}$	0.628	0.610
$\hat{\beta}_{\text{bathrooms}}$	0.204	0.211
$\hat{\beta}_{\text{garage}}$	0.109	0.111

Notes: Lot size winsorized at its 99.5th percentile ($\approx$ 3.04 million sqft; 3,944 observations capped); Bike Score capped at 100 (38 observations). Unstandardized OLS, N = 788,842. Other coefficients broadly stable (not shown). The large change in $\hat{\beta}_{\ln(\text{lot})}$ mirrors the imputation sensitivity finding, likely reflecting outlier lot sizes in imputed observations.

Winsorization increases R^2 modestly from 0.634 to 0.640, suggesting that extreme values in continuous variables account for a small portion of unexplained variation. The lot-size coefficient again jumps substantially (0.018 to 0.058), consistent with the imputation sensitivity analysis: extreme imputed lot sizes attenuate the gradient. Other coefficients are broadly stable, indicating that the main OLS results are not driven by outliers.

6.3. Quantile Regression Subsample Stability

We repeat the median quantile regression ($\tau = 0.50$) across 10 random subsamples of 150,000 observations each. Table 19 reports the coefficient of variation (CV) and sign stability for key variables.

Table 19. Quantile Regression Stability: $\tau = 0.50$ Across 10 Subsamples

Variable	CV (%)	Sign Stability (%)	Assessment
West	1.0	100	Highly stable
Northeast	1.6	100	Highly stable
ln(Lot Size)	2.7	100	Highly stable
ln(Living Area)	2.9	100	Highly stable
Bathrooms	2.9	100	Highly stable
Garage	4.1	100	Highly stable
Luxury Score	4.8	100	Highly stable
Age^2	5.3	100	Highly stable
Foreclosure	7.1	100	Stable
Bedrooms	7.3	100	Stable
Less stable variables			
Pool	33.4	100	Moderately unstable
Waterfront	42.2	100	Moderately unstable
Age	78.6	80	Unstable

Notes: CV = coefficient of variation across 10 random subsamples of 150,000 observations. Sign stability = percentage of subsamples in which the coefficient has the same sign as the pooled estimate. Variables sorted by CV. The instability of age (CV = 78.6%) likely reflects heterogeneity in the age-price relationship across geographic subsamples. The waterfront and pool coefficients vary in magnitude but never change sign; their instability reflects low variation in these binary variables in some subsamples.

Most variables exhibit high stability (CV < 8%, 100% sign stability), confirming that the quantile regression results are robust to subsample variation. Three variables are exceptions: **age** (CV =

78.6%, sign stability = 80%), **waterfront** (CV = 42.2%), and **pool** (CV = 33.4%). The instability of age is consistent with substantial geographic heterogeneity in the age-price relationship—old properties may command a premium in some markets (historic charm) and a discount in others (obsolescence). Waterfront and pool instability reflects the relatively low prevalence of these features in some subsamples.

6.4. Robustness Summary Matrix

Table 20 provides a comprehensive summary of all robustness findings.

Table 20. Robustness Summary Matrix

Test	Baseline	Robustness	Key Finding
OLS Specification			
Region FE → State FE	$R^2 = 0.630$	$R^2 = 0.678$	+4.8 pp from state-level controls
Region FE → ZIP3 FE	$R^2 = 0.630$	$R^2 = 0.725$	+9.5 pp from ZIP3 controls
Drop imputed lot sizes	$\hat{\beta}_{\ln \mathrm{lot}} = 0.018$	$\hat{\beta}_{\ln \mathrm{lot}} = 0.059$	Coefficient triples; imputation attenuates
Winsorization (1%/99%)	$R^2 = 0.634$	$R^2 = 0.640$	Modest improvement; lot coeff. sensitive
Spatial Diagnostics			
Moran's I (3 subsamples)	$I = 0.2745$	std = 0.0076	Highly stable; strong spatial autocorrelation
Quantile Regression			
QR stability (10 draws)	Most CV < 8%	Age CV = 78.6%	Age unstable; all others sign-stable
Inter-quantile Wald	—	11/13 significant	Distribution varies for most attributes
Machine Learning			
Random vs. geo holdout	$R^2 = 0.833$	$R^2 = 0.425$	40.8 pp gap; spatial leakage
Remove geo features	Geo $R^2 = 0.425$	Geo $R^2 = 0.519$	+9.4 pp; geography hurts generalization
Add lat/lon	Random $R^2 = 0.870$	Geo $R^2 = 0.464$	Coordinates memorize location
Ablation: neighborhood	—	+17.6 pp random	Largest single feature-group contribution
Ablation: full model	—	−7.6 pp geo	Region dummies harm generalization
SHAP Stability			
XGBoost vs. LightGBM	—	$\rho = 0.990$	Near-identical rankings
XGBoost vs. RF	—	$\rho = 0.906$	Consistent top-6 features
LightGBM vs. RF	—	$\rho = 0.892$	Consistent top-6 features

Notes: This table summarizes all robustness checks conducted in the paper. "pp" = percentage points. All findings refer to the same base sample of $N = 788{,}842$ unless noted. The geographic holdout R^2 of 0.425 for the full XGBoost model reflects the variant with region dummies included in the ablation design; the main specification yields 0.547 due to differences in training-set composition.

7. Discussion

The results support four main messages, each associated with a specific modeling framework. For each, we interpret our estimates against the prior literature, noting where they agree with published findings, where they diverge, and what the divergences imply.

7.1. Message 1: OLS Remains Useful for Interpretable Associations

The semi-log OLS model explains 63.4% of the variation in log listing prices using 62 regressors—within the range typical of large cross-sectional hedonic studies (Sirmans et al., 2005, Malpezzi, 2003)—and its headline gradients align well with the meta-analytic record. The living-area elasticity of 0.63 is consistent with the concave size-price relationship documented across the 125 studies synthesized by Sirmans et al. (2005); the positive bathroom and garage gradients and the negative age gradient likewise match the modal signs in that synthesis. The negative conditional bedroom gradient—often treated as an anomaly—is in fact the standard result once total living area is held constant, reflecting a room-size trade-off rather than a distaste for bedrooms. The 33.4% foreclosure discount is larger than the roughly 27% forced-sale discount estimated from Massachusetts transactions by Campbell et al. (2011), as expected given that our estimate reflects asking-price positioning of distressed listings rather than realized sale prices. Regularized linear benchmarks (Ridge, Lasso, Elastic Net) achieve essentially identical performance (R^2 = 0.628–0.630), confirming that the parametric specification is not overfit and that the 20-percentage-point gap relative to XGBoost reflects genuine non-linearities and interactions, not estimation noise.

The fixed-effects progression carries a sharper lesson. Moving from four Census regions (R^2 = 0.634) to state fixed effects (0.678) to 886 ZIP3 fixed effects (0.725) shows that a large share of "unexplained" variation is spatial, echoing the simulation-based advice of Kuminoff et al. (2010) that spatial fixed effects are the single most consequential specification choice in hedonic work, and the best-practice guidance of Bishop et al. (2020). The ZIP3 model closes roughly half the OLS-to-XGBoost gap, implying that fine geographic partitioning —not flexible functional form—accounts for much of what tree ensembles add. This dovetails with Pace and Hayunga (2020), who find that machine-learning methods extract predictive content from the residuals of hedonic models precisely because those residuals retain spatial structure. The residual Moran's I of 0.2745 after regional controls—comparable in magnitude to the residual autocorrelation documented in metropolitan samples by Basu and Thibodeau (1998) —confirms that even our richest interpretable specification leaves the spatial error structure that the spatial econometrics literature has long modeled explicitly (Anselin, 1988, LeSage and Pace, 2009, Dubin, 1998).

Two coefficient-level findings deserve confrontation with the capitalization literature. The negative school-rating coefficient contradicts the positive valuation of school quality identified by boundary-discontinuity designs (Black, 1999, Bayer et al., 2007). We read this divergence not as evidence against school-quality capitalization but as a textbook illustration of why those designs exist: in a national cross-section, school ratings are collinear with property-tax rates (which enter separately), regional price levels, and unobserved neighborhood composition, so the partial coefficient is uninformative about the structural valuation. Analogously, the negative Walk Score coefficient—despite a documented walkability premium (Pivo and Fisher, 2011)—reflects conditioning simultaneously on bike and transit accessibility; the three scores are highly correlated, and their premia cannot be attributed variable-by-variable. Both cases reinforce the paper's interpretive stance: hedonic coefficients are conditional associations whose structural content depends on identification that a national listing cross-section does not provide (Kuminoff et al., 2013).

Finally, the imputation sensitivity analysis shows the baseline lot-size gradient (0.02) is attenuated roughly threefold by state-median imputation, with the complete-case estimate (0.06) more plausible though subject to selection. The broader point—data-quality audits are not optional in scraped-data hedonics—is a practical extension of the measurement cautions in Bishop et al. (2020).

7.2. Message 2: Quantile Regression Adds Distributional Insight

Quantile regression reveals that mean effects mask economically large distributional heterogeneity, and the shape of that heterogeneity is consistent with the housing literature. The declining living-area gradient (0.358 at $\tau = 0.10$ versus 0.251 at $\tau = 0.90$) mirrors the pattern in Zietz et al. (2008), where square footage is valued relatively more in lower-priced homes. The rising pool and lot-size gradients toward the top of the distribution are consistent with amenity-driven luxury pricing documented at upper quantiles in Hong Kong (Mak et al., 2010) and Sydney (Waltl, 2019). The garage gradient—ten times larger at the bottom decile than the top ($z = 28.9$)—extends this logic: basic functional amenities differentiate lower-priced properties where they are scarce, and become nearly universal (hence unpriced) upstream. Such quantile-dependent pricing is one empirical face of housing-market segmentation (Goodman and Thibodeau, 1998): submarkets defined by price tier value the same attribute bundle differently. Our inter-quantile Wald tests formalize what earlier housing QR studies typically showed graphically: coefficient equality is rejected for 11 of 13 attributes, so the conditional listing-price distribution is differentially stretched and compressed by attributes, not merely shifted. This is the cross-sectional analogue of the finding in McMillen (2008) that house price distributions move through coefficients rather than characteristics.

For policy, the pattern implies that improvements to basic amenities are associated with the largest proportional listing-price differences at the lower end of the market, while luxury amenities differentiate the top—though we emphasize, again, that these are associations in asking prices, filtered through the seller-behavior mechanisms of Genesove and Mayer (2001) and Han and Strange

(2016), not causal renovation returns.

The subsample stability analysis shows most quantile coefficients are highly robust (CV < 8%, 100% sign stability), with age the notable exception (CV = 78.6%). Geographically heterogeneous vintage effects—historic premia in some markets, obsolescence discounts in others—are the natural reading, and they caution against interpreting any single national age gradient too literally.

7.3. Message 3: ML Performance Depends Critically on Validation Design

XGBoost improves on OLS by 20 percentage points of R^2 under random validation (0.833 vs. 0.630)—squarely within the 15–25% error-reduction range reported in the ML-valuation literature (Bourassa et al., 2010, Park and Bae, 2015, Chen and Guestrin, 2016). Had we stopped there, the paper would read as one more confirmation that boosting beats hedonics. The geographic holdout overturns that reading: under state-level holdout XGBoost attains only $R^2 = 0.425$–0.547 depending on specification. The 29–41-point collapse is strikingly similar in kind to what Ploton et al. (2020) document for ecological mapping models, and it validates, in a housing context, the warnings of Roberts et al. (2017) and Meyer and Pebesma (2021) about evaluating spatial predictions on randomly held-out data.

The ablation analysis identifies *what* fails to transfer. Neighborhood features contribute the largest random-validation gain (+17.6 pp) but only +3.5 pp under geographic holdout—consistent with the conjecture that scores like Walk Score and school ratings carry market-specific meaning. Validation research on Walk Score itself finds that its correspondence with underlying walkability constructs varies across metropolitan areas (Duncan et al., 2011), which supports the mechanism, though we do not test it directly and flag it as plausible rather than established. More striking, the features most obviously "about" geography are actively harmful out-of-market: the full model with region dummies loses 7.6 pp of geographic holdout R^2 relative to the market-status stage, and adding raw coordinates —the most flexible location encoding—produces the largest interpolation gain (+3.7 pp random) alongside a large extrapolation loss. Region dummies and coordinates let trees partition price levels by place, which is exactly what Pace and Hayunga (2020) show residual-based ML exploits, and exactly what cannot transfer to states never seen in training.

We stress the scope of this conclusion, in light of the debate opened by Wadoux et al. (2021): for *interpolation*—valuing a property in a market represented in training data, the typical AVM production setting—random validation is informative and the geographic features earn their keep. Our claim concerns *extrapolation* to unrepresented markets, where blocked validation is the appropriate benchmark (Valavi et al., 2019, Meyer and Pebesma, 2021) and where the AVM-evaluation literature already counsels reporting more than headline accuracy (Steurer et al., 2021). The practical rule for deployment follows: match the validation protocol to the deployment question, and if the question involves new markets, prefer the leaner, geography-free specification—it costs 0.3 pp of interpolation accuracy and buys 9.4 pp of extrapolation accuracy.

7.4. Message 4: SHAP Interprets Prediction, Not Economics

SHAP values identify living area, bathrooms, school quality, lot size, and region as the dominant predictive contributors, and this ranking is remarkably stable across three tree ensembles (ρ = 0.89–0.99). The stability result addresses a genuine concern in the explainability literature—that importance rankings are artifacts of a particular fitted model (Ribeiro et al., 2016)—and parallels the motivation for model-agnostic explanation methods. But stability is not structure. Three cautions from the literature apply directly. First, SHAP decomposes predictions, not the data-generating process; a feature may earn a large SHAP value by proxying unobservables (Lundberg and Lee, 2017, Lundberg et al., 2020). Second, correlated features share contributions in ways that defeat attribute-level economic interpretation—our school-rating case is again illustrative, ranking 3rd by SHAP while its OLS sign is negative and its credible structural valuation (Black, 1999, Bayer et al., 2007) is positive. Third, as Rudin (2019) argues, post-hoc explanation of a black box is not a substitute for an interpretable model when stakes are high; in our framework, the interpretable model (OLS/QR) and the black box answer different questions, and SHAP does not convert the latter into the former. In Rosen's terms: SHAP values are not implicit prices, and the stability we document should raise confidence in SHAP as a description of *this prediction technology*, not as a measurement of *market valuation*.

7.5. Synthesis

Across the four messages, a single theme recurs: each framework is reliable precisely within the question it was built to answer. OLS with rich spatial controls yields stable, literature-consistent conditional associations; quantile regression reveals formally significant distributional structure; boosting delivers real interpolation gains whose extrapolation content must be established by design, not assumed; and SHAP describes the predictive machine without licensing economic claims. The methodological corollary—that validation design and feature choice interact, and that random-split comparisons overstate ML advantages for out-of-market questions—is, we believe, the paper's most transferable lesson for both the hedonic and the AVM literatures.

8. Limitations

We summarize the principal limitations in a structured format.

1. **Listing prices, not transaction prices.** The dependent variable is the asking price, which may differ from the realized sale price due to strategic pricing, negotiation, and market conditions

(Horowitz, 1992, Genesove and Mayer, 2001, Han and Strange, 2016). Listing-price gradients are not necessarily equivalent to transaction-price implicit prices, and the wedge is plausibly correlated with attributes.

2. **No causal identification.** All estimates are conditional associations. Without exogenous variation, we cannot distinguish the causal effects of attributes from sorting, supply constraints, and omitted variables (Ekeland et al., 2004, Roberts and Whited, 2013).
3. **Spatial dependence not modeled.** Moran's $I = 0.27$ confirms substantial spatial autocorrelation in OLS residuals. The paper does not estimate spatial lag, spatial error, or geographically weighted regression models (Anselin, 1988, LeSage and Pace, 2009), which may affect both efficiency and consistency of estimates.
4. **Random validation overstates ML performance.** The random train-test split allows spatial leakage. Under geographic holdout, XGBoost R^2 drops from 0.833 to 0.425–0.547 depending on specification.
5. **Standardization complicates interpretation.** The standardized OLS coefficients are not directly interpretable as elasticities or semi-elasticities. We provide an unstandardized specification for economic interpretation but note that some readers may conflate the two.
6. **Missing data and imputation.** Year built (19.4%), lot size (16.3%), and transit score (70.1%) have substantial missingness. State-median imputation preserves geographic variation but attenuates the lot-size coefficient by a factor of three.
7. **Climate risk variables entirely missing.** Despite their growing importance, flood, fire, heat, wind, and air risk factors could not be incorporated. The climate-capitalization literature finds economically significant price effects of exposure (Baldauf et al., 2020, Bernstein et al., 2019, Murfin and Spiegel, 2020), so their omission plausibly contributes to the residual spatial autocorrelation we document.
8. **SHAP values are not structural implicit prices.** SHAP provides prediction decompositions, not marginal willingness-to-pay estimates. Cross-model stability supports the ranking but not the economic interpretation.
9. **Quantile regression subsample.** Quantile regression is estimated on 150,000 observations (19% of the full sample) for computational reasons. Most coefficients are stable (CV $<$ 8%) but age is not (CV = 78.6%).
10. **Non-representative sample.** The Zillow listing dataset overrepresents Southern states, active listings, and possibly certain property types. Results may not generalize to the full U.S. housing stock.
11. **Data quality concerns.** Bike Score exceeds 100 for 38 observations; some lot sizes are implausibly large; parking-space data are noisy. These affect a small fraction of observations but may influence tail estimates.

12. **No hyperparameter tuning via cross-validation.** XGBoost and LightGBM hyperparameters were set based on common defaults rather than optimized via nested cross-validation. Performance could potentially improve with systematic tuning.
13. **Geographic holdout design is conservative.** Holding out 10 entire states is a stringent test; finer geographic blocking (county-level or MSA-level) might yield intermediate performance estimates that are more relevant for some AVM applications (Valavi et al., 2019). Conversely, for pure interpolation objectives, spatially blocked validation can be pessimistically biased (Wadoux et al., 2021); our random-split results remain the relevant benchmark for that use case.
14. **Single cross-section.** The data represent a snapshot of active listings at one point in time. Temporal variation in hedonic gradients—due to market cycles, interest rate changes, or policy shifts—cannot be examined.

9. Conclusion

This paper has compared three approaches to hedonic housing price analysis—OLS, quantile regression, and gradient-boosted machine learning with SHAP interpretation—using a national sample of 788,842 U.S. Zillow listings. Each framework answers a different question, and the robustness program documents the boundaries of each framework's conclusions.

OLS provides interpretable conditional mean associations whose signs and magnitudes align with the meta-analytic record of hedonic housing studies (Sirmans et al., 2005): a price-to-area elasticity of 0.63, a 22.6% bathroom listing-price gradient, a 33.4% foreclosure discount, and regional listing-price differentials of 49%–63% between coastal and interior markets. The progression from region fixed effects ($R^2 = 0.634$) to ZIP3 fixed effects ($R^2 = 0.725$) demonstrates—in line with the specification advice of Kuminoff et al. (2010)—that fine-grained geographic controls capture substantial variation, while the imputation sensitivity analysis shows the lot-size gradient is attenuated threefold by state-median imputation.

Quantile regression reveals that mean effects mask substantial distributional heterogeneity, formally confirmed by inter-quantile Wald tests that reject coefficient equality for 11 of 13 variables. The garage listing-price gradient is ten times larger at the 10th percentile than the 90th ($z = 28.92$); the pool gradient is insignificant below the median but reaches 9.9% at $\tau = 0.90$. Subsample stability analysis identifies age as the one coefficient that is genuinely unstable across geographic subsamples.

XGBoost captures non-linearities and interactions that improve prediction from $R^2 = 0.630$ (OLS) to $R^2 = 0.833$ under random validation. That figure, however, substantially overstates generalization to new markets. The ablation analysis attributes the largest predictive gain to neighborhood features (+17.6 pp) and shows that the full model with region dummies actively harms geographic holdout performance (−7.6 pp); removing all geographic features *improves* geographic holdout R^2 from 0.425 to 0.519, while adding latitude and longitude does the opposite (+3.7 pp random, −5.5 pp geographic). Geographic features help the model memorize location-specific price levels but do not improve—and can degrade—its ability to transfer attribute-price relationships to unseen markets.

SHAP values identify the features contributing most to predictions—living area, bathrooms, school quality, lot size, region—and this ranking is stable across three tree-based models (Spearman ρ = 0.89–0.99). These remain predictive decompositions rather than implicit prices, a distinction the explainability literature itself insists upon (Rudin, 2019). Moran's $I = 0.27$ on OLS residuals—stable across subsamples—confirms that regional controls leave substantial spatial dependence unaddressed.

The central contribution is methodological: model evaluation in spatially dependent housing data hinges on validation design, and the random-versus-blocked distinction developed in ecology (Roberts et al., 2017, Ploton et al., 2020, Meyer and Pebesma, 2021, Wadoux et al., 2021) transfers directly—and consequentially —to hedonic economics. Random splits answer the interpolation question; held-out regions answer the extrapolation question; and the two rank both models and feature sets differently. Studies comparing machine learning to hedonic regression on random splits alone will systematically overstate the practical ML advantage for any application involving new markets.

Several extensions follow naturally. Spatial econometric estimation (LeSage and Pace, 2009) would model the dependence we only diagnose; replication on transaction prices would quantify the listing-price wedge that the seller-behavior literature predicts (Genesove and Mayer, 2001, Han and Strange, 2016); climate-risk variables (Bernstein et al., 2019, Murfin and Spiegel, 2020) are a first-order omission our residual diagnostics likely reflect; generalized additive models would decompose the ML gain into non-linearity versus spatial partitioning; and finer geographic blocking would map the continuum between our two validation extremes. We would regard geographic holdout validation, reported alongside random validation, as a reasonable default for future ML-hedonic comparisons.

The broader lesson is not that machine learning replaces hedonic econometrics, nor the reverse. Prediction, distributional heterogeneity, and economic interpretation are different questions; each of the tools examined here is strong on exactly one of them, and honest housing-market analysis will continue to require all three—each validated against the question it is actually meant to answer.

References

Anglin, P. M. and Gençay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, 11(6):633–648.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):3–25.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Baldauf, M., Garlappi, L., and Yannelis, C. (2020). Does climate change affect real estate prices? Only if you believe in it. *Review of Financial Studies*, 33(3):1256–1295.

Bartik, T. J. (1987). The estimation of demand parameters in hedonic price models. *Journal of Political Economy*, 95(1):81–88.

Basu, S. and Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *Journal of Real Estate Finance and Economics*, 17(1):61–85.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638.

Bernstein, A., Gustafson, M. T., and Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*, 134(2):253–272.

Bishop, K. C., Kuminoff, N. V., Banzhaf, H. S., Boyle, K. J., von Gravenitz, K., Pope, J. C., Smith, V. K., and Timmins, C. D. (2020). Best practices for using hedonic property value models to measure willingness to pay for environmental quality. *Review of Environmental Economics and Policy*, 14(2):260–281.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599.

Bourassa, S. C., Cantoni, E., and Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research*, 32(2):139–160.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Campbell, J. Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. *American Economic Review*, 101(5):2108–2131.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, 22(3):453–474.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*, pages 785–794.

Court, A. T. (1939). Hedonic price indexes with automotive examples. In *The Dynamics of Automobile Demand*, pages 99–117. General Motors Corporation.

Cropper, M. L., Deck, L. B., and McConnell, K. E. (1988). On the choice of functional form for hedonic price functions. *Review of Economics and Statistics*, 70(4):668–675.

Dubin, R. A. (1998). Spatial autocorrelation: A primer. *Journal of Housing Economics*, 7(4):304–327.

Duncan, D. T., Aldstadt, J., Whalen, J., Melly, S. J., and Gortmaker, S. L. (2011). Validation of Walk Score for estimating neighborhood walkability: An analysis of four US metropolitan areas. *International Journal of Environmental Research and Public Health*, 8(11):4160–4179.

Ekeland, I., Heckman, J. J., and Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, 112(S1):S60–S109.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Genesove, D. and Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics*, 116(4):1233–1260.

Goodman, A. C. and Thibodeau, T. G. (1998). Housing market segmentation. *Journal of Housing Economics*, 7(2):121–143.

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric analysis of quality change. In *The Price Statistics of the Federal Government*, pages 173–196. NBER.

Halvorsen, R. and Pollakowski, H. O. (1981). Choice of functional form for hedonic price equations. *Journal of Urban Economics*, 10(1):37–49.

Han, L. and Strange, W. C. (2016). What is the role of the asking price for a house? *Journal of Urban Economics*, 93:115–130.

Hill, R. J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *Journal of Economic Surveys*, 27(5):879–914.

Horowitz, J. L. (1992). The role of the list price in housing markets: Theory and an econometric model. *Journal of Applied Econometrics*, 7(2):115–129.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154.

Knight, J. R. (2002). Listing price, time on market, and ultimate selling price. *Real Estate Economics*, 30(2):213–237.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Kok, N., Koponen, E.-L., and Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6):202–211.

Kuminoff, N. V., Parmeter, C. F., and Pope, J. C. (2010). Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities? *Journal of Environmental Economics and Management*, 60(3):145–160.

Kuminoff, N. V., Smith, V. K., and Timmins, C. (2013). The new economics of equilibrium sorting and policy evaluation using housing markets. *Journal of Economic Literature*, 51(4):1007–1062.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2):132–157.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

Liao, W.-C. and Wang, X. (2012). Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, 21(1):16–27.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Mak, S., Choy, L. H. T., and Ho, W. K. O. (2010). Quantile regression estimates of Hong Kong real estate prices. *Urban Studies*, 47(11):2461–2472.

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In O'Sullivan, T. and Gibb, K., editors, *Housing Economics and Public Policy*, pages 67–89. Blackwell.

McMillen, D. P. (2008). Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients? *Journal of Urban Economics*, 64(3):573–589.

Meyer, H. and Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1–2):17–23.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Murfin, J. and Spiegel, M. (2020). Is the risk of sea level rise capitalized in residential real estate? *Review of Financial Studies*, 33(3):1217–1255.

Oates, W. E. (1969). The effects of property taxes and local public spending on property values. *Journal of Political Economy*, 77(6):957–971.

Pace, R. K. and Hayunga, D. (2020). Examining the information content of residuals from hedonic and spatial models using trees and forests. *Journal of Real Estate Finance and Economics*, 60(1–2):170–180.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934.

Pivo, G. and Fisher, J. D. (2011). The walkability premium in commercial real estate investments. *Real Estate Economics*, 39(2):185–219.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C. F., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11:4540.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, pages 1135–1144.

Ridker, R. G. and Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *Review of Economics and Statistics*, 49(2):246–257.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal and spatial dependence. *Ecography*, 40(8):913–929.

Roberts, M. R. and Whited, T. M. (2013). Endogeneity in empirical corporate finance. In Constantinides, G. M., Harris, M., and Stulz, R. M., editors, *Handbook of the Economics of Finance*, volume 2A, pages 493–572. Elsevier.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Shapley, L. S. (1953). A value for *n*-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press.

Sirmans, S., Macpherson, D., and Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1):3–43.

Steurer, M., Hill, R. J., and Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2):99–129.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., and Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457:109692.

Waltl, S. R. (2019). Variation across price segments and locations: A comprehensive quantile regression analysis of the Sydney housing market. *Real Estate Economics*, 47(3):723–756.

Waugh, F. V. (1928). Quality factors influencing vegetable prices. *Journal of Farm Economics*, 10(2):185–196.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Zietz, J., Zietz, E. N., and Sirmans, G. S. (2008). Determinants of house prices: A quantile regression approach. *Journal of Real Estate Finance and Economics*, 37(4):317–333.

A. Additional Figures

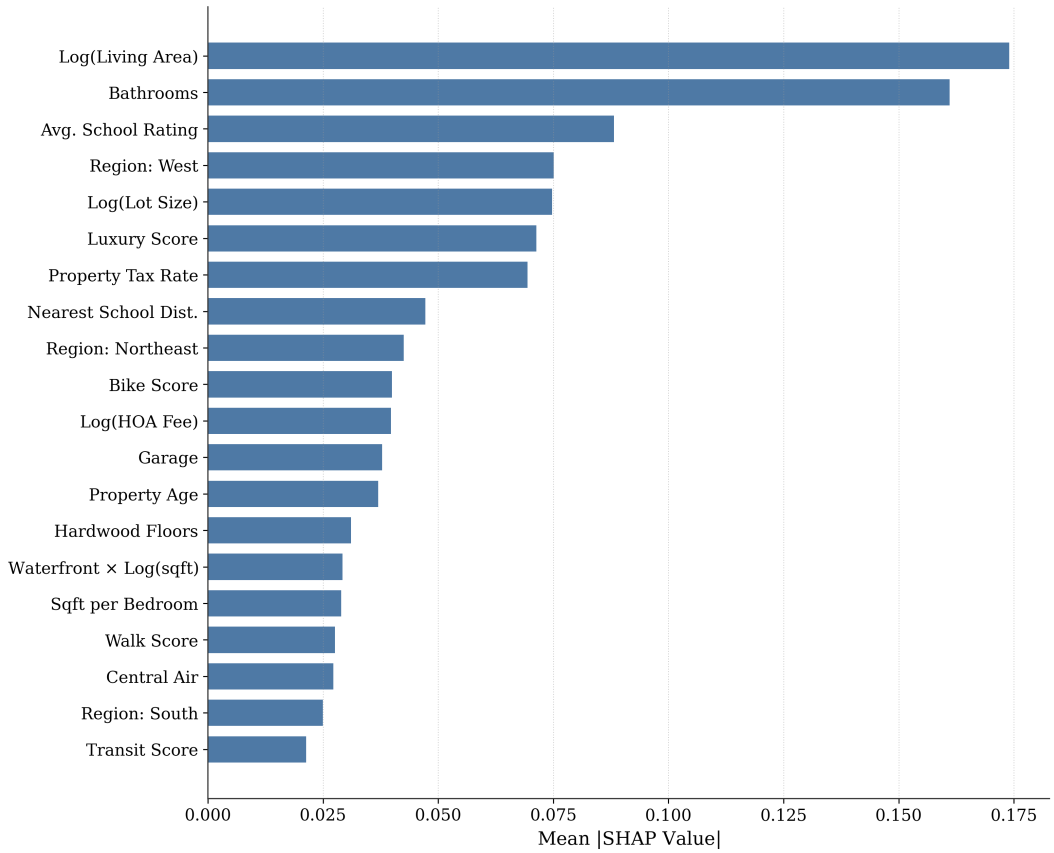

Figure 10. Feature Importance Bar Chart: Mean Absolute SHAP Values (XGBoost). These values reflect predictive contributions, not causal importance.

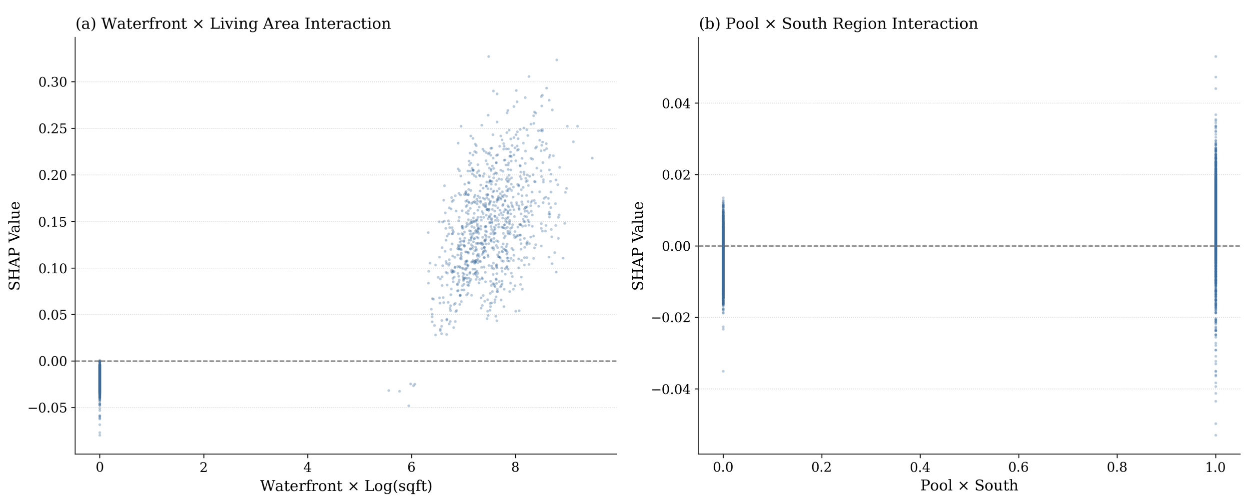

Figure 11. SHAP Values for Interaction Terms: (a) Waterfront × Log(Living Area); (b) Pool × South Region

B. Full OLS Coefficient Table

Table 21 reports the complete set of standardized OLS coefficients, including all categorical controls.

Table 21. Full OLS Regression Results Including All Categorical Controls (Standardized)

Variable	Coeff.	SE	Sig.	Variable	Coeff.	SE	Sig.
Constant	12.510	0.010	***	Roof: Metal	0.070	0.008	***
ln(Living Area)†	0.312	0.003	***	Roof: Shingle/Comp	−0.074	0.008	***
Bedrooms†	−0.068	0.002	***	Roof: Tile	−0.110	0.008	***
Bathrooms†	0.236	0.003	***	Roof: Slate	0.008	0.017	
Age†	−0.042	0.014	**	Roof: Other	0.033	0.008	***
Age^{2†}	0.051	0.002	***	Roof: Unknown	0.004	0.008	
Stories†	0.000	0.000		Constr: Concrete/Block	0.269	0.003	***
Bath/Bed Ratio†	−0.021	0.002	***	Constr: Stucco	0.149	0.002	***
Sqft/Bed†	−0.002	0.002		Constr: Stone	0.145	0.005	***
ln(Lot Size)†	0.070	0.001	***	Constr: Wood/Frame	0.103	0.002	***
Pool	0.029	0.003	***	Constr: Vinyl	0.012	0.002	***
Spa	0.033	0.003	***	Constr: Steel	0.027	0.014	*
Basement	−0.042	0.002	***	Constr: Other	0.075	0.002	***
Fireplace	−0.002	0.002		Constr: Unknown	0.177	0.002	***
Garage	0.109	0.002	***	Found: Concrete	0.217	0.005	***
Waterfront	−0.135	0.032	***	Found: Crawlspace	0.181	0.006	***
Central Air	0.069	0.001	***	Found: Slab	0.122	0.005	***
Forced Air	−0.023	0.002	***	Found: Pier/Raised	0.161	0.006	***
Hardwood	0.125	0.001	***	Found: Stone	0.032	0.010	**
Parking†	0.000	0.099		Found: Other	0.144	0.006	***
Luxury†	0.044	0.002	***	Found: Unknown	0.174	0.005	***
Walk†	−0.011	0.001	***	Northeast	0.487	0.003	***
Bike†	0.072	0.001	***	South	0.022	0.003	***
Transit†	0.052	0.001	***	West	0.401	0.003	***
School Rating†	−0.049	0.001	***	sqft×Age†	−0.110	0.014	***
School Count†	−0.002	0.001	***	Pool×South	0.008	0.001	***
School Dist.†	−0.016	0.001	***	WF×sqft†	0.137	0.009	***
Tax Rate†	−0.029	0.001	***	Base×North	−0.025	0.001	***
Condo	−0.037	0.004	***	Condo×Walk†	0.082	0.001	***
HOA	−0.108	0.002	***	Age×Lux†	0.044	0.001	***
ln(HOA+1)†	0.044	0.001	***				
New Constr.	0.103	0.003	***				
Foreclosure	−0.406	0.010	***				

R^2 = 0.634, Adj. R^2 = 0.634, N = 788,842, F = 18,712.5, RMSE = 0.489

Notes: HC3 robust standard errors. †Standardized continuous variables. Reference categories: Roof = Flat; Construction = Brick; Foundation = Basement; Region = Midwest. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

C. Robustness Agenda

The following robustness checks and extensions are planned or completed. Items marked with ✓ have been completed in this version; unmarked items are left for future work.

- ✓ ZIP3-level fixed effects (Section 5.1.4).
1. County-level and MSA-level fixed effects.
2. Spatial lag and spatial error model estimation.
3. Spatial cross-validation (train on some regions, test on others) at finer geographic levels.
- ✓ Inter-quantile Wald tests for coefficient equality across quantiles (Section 5.3).
- ✓ Re-estimation excluding imputed observations (Section 6.1).
4. CatBoost benchmark with native categorical handling.
5. Nested cross-validation for hyperparameter tuning.
- ✓ SHAP stability comparison across XGBoost, LightGBM, and Random Forest (Section 5.8).
6. Integration of climate risk data from First Street Foundation or FEMA flood maps.
7. Replication using transaction prices from recently sold listings.
8. GAM/spline models to decompose the ML gain into non-linearity versus spatial partitioning.
9. Comparison of sample distribution to ACS and Census benchmarks by state and price tier.
10. Calibration plots by price decile for all models.
- ✓ Winsorization sensitivity (Section 6.2).
- ✓ Feature ablation analysis for XGBoost (Section 5.6).
- ✓ XGBoost with latitude/longitude coordinates (Section 5.7).
- ✓ XGBoost without geographic features (Section 5.7).
- ✓ Moran's I subsample stability (Section 5.2).
- ✓ Quantile regression stability across 10 subsamples (Section 6.3).

D. Reproducibility Checklist

Table 22. Reproducibility Checklist

Item	Status	Notes
Data		
Data source identified	Yes	Zillow active listings
Sample construction documented	Yes	Table 1
Missing data rates reported	Yes	Table 2
Imputation method documented	Yes	State-level medians
Imputation sensitivity tested	Yes	Table 17
Methodology		
Model specifications documented	Yes	Eq. 1–6
Hyperparameters reported	Yes	Section 4
Random seeds reported	Yes	Seed = 42
Validation design documented	Yes	Section 4.5
Results		
Point estimates reported	Yes	Tables 5–16
Standard errors / CIs reported	Yes (OLS, QR)	HC3 standard errors
Multiple comparisons addressed	Partially	Inter-quantile tests
Effect sizes in natural units	Yes	Table 6
Robustness		
Sensitivity to data quality	Yes	Sections 6.1–6.2
Sensitivity to model choice	Yes	Table 11
Sensitivity to validation design	Yes	Tables 12–14
Cross-model stability	Yes	Table 16
Code and Data Availability		
Code available	Yes	Replication package (`src/`, `scripts/`, pinned `requirements.txt`)
Data available	On request	Subject to Zillow ToS

Notes: This checklist follows recommended practices for empirical reproducibility in applied economics.