Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL27981

Housing Value Drivers: OLS vs Random Forest

Jingyue Zhang

Aquinas High School, Guiyang, China
zhangjingyuedr@gmail.com

Abstract. This study examines the determinants of California house prices using the 1990 census block-group dataset (n = 20,640) and a unified evaluation protocol with a fixed train/validation/test split (60/20/20). This study compares two linear hedonic baselines, OLS without distances (A1) and OLS with five geographic distances (A2), to a Random Forest (RF) trained on log prices and back-transformed to dollars via Duan smearing, while A1/A2 are estimated in levels (USD). Models are compared primarily by test-set R^2 in dollars. RF attains the highest accuracy (test $R^2 \approx 0.834$, RMSE≈$47k), outperforming A2 ($R^2 \approx 0.648$), while A2 improves modestly over A1, quantifying the incremental value of explicit geographic accessibility. Cross-method evidence converges on a stable core of determinants: median income and distance to coast rank first and second, followed by latitude/longitude and distances to major cities; structural count variables are comparatively weaker at the block-group level. Price-tier analysis (Low/Mid/High, defined at training P30/P70) shows stronger fit in the low tier and larger errors/heterogeneity at the high end. Robustness checks (Winsorization, Huber, dropping distances, higher-order lat/long polynomials) do not overturn conclusions; dropping distances degrades performance most. Overall, purchasing power and spatial accessibility jointly organize the cross-section of California house prices, with nonlinearities favoring RF for prediction and A2 for interpretation.

Keywords: Linear Regression, Random Forest, Housing Prices

1. Introduction

Housing is not only a basic human need but also a key pillar of economic activity, accounting for approximately 15–18% of U.S. GDP in recent years [1]. Unlike standardized goods, real estate values exhibit substantial heterogeneity and are jointly determined by multiple attributes of the property and its geographic location. The hedonic pricing framework models market price as a function of observable characteristics [2], with the implicit (marginal) price of each attribute given by the slope of the hedonic price function. In practice, factors such as dwelling size, age, the number of bedrooms and bathrooms, neighborhood quality, and transportation accessibility collectively determine market value.

A large empirical literature employs hedonic regressions to estimate the marginal contributions of individual attributes to house prices. Overall, prior studies confirm that certain structural characteristics (e.g., larger floor area, more rooms, additional bathrooms) consistently raise housing values, whereas other factors (e.g., proximity to hazards or high-crime areas) depress prices.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL27981

However, results are not uniform across all contexts—some attributes show mixed evidence across studies, indicating that their effects may depend on local market conditions or buyer populations [3]. This complexity makes it both challenging and important to identify which determinants are robust and under what conditions that robustness holds.

In recent years, machine learning (ML) has become increasingly prevalent in house price analysis, with a primary emphasis on improving predictive accuracy. Random Forests (RF) and Gradient Boosting (GB) capture nonlinearities and complex interactions, and multiple empirical studies show that, with moderate sample sizes and conventional feature sets, they often outperform linear regression [4-5]. However, these models have weak interpretability and are effectively "black boxes," creating an accuracy–interpretability tension vis-à-vis the mechanism-based explanations required in policy and practice; many studies prioritize prediction while neglecting explanation, thus offering limited guidance for decision making.

Moreover, most studies focus on mean effects, implicitly assuming that determinants operate uniformly across the price distribution. In reality, feature importance may vary by price tier; evidence from quantile regression shows that buyers of high-priced homes place different weights on attributes such as floor area and bathrooms than buyers of low-priced homes [6]. This suggests the need to test whether distinct pricing mechanisms exist across market segments (entry-level versus luxury), rather than adopting a one-size-fits-all approach. This study uses real California housing data to examine this field.

2. Research questions

Building on the above motivations and findings, this study focuses on the following three research questions (RQ):

RQ1 (Determinant Identification): At the full-sample level, which attributes are the most important and stable determinants of California house prices?

RQ2 (Heterogeneity): Do the key determinants exhibit systematic differences across price tiers (Low/Mid/High)?

RQ3 (Consistency/Robustness): Do different methods (statistical and machine learning) converge on the same "key determinants"?

RQ1 emphasizes identification (providing a clear ranking based on the best-performing model); RQ2 emphasizes heterogeneity (whether the relative roles of key determinants shift across price tiers) and RQ3 emphasizes testing (whether different methods yield the same key set). The three are complementary and serve distinct purposes.

3. Data and variables

3.1. Data source and unit of analysis

This study uses the publicly available processed California Housing Prices dataset from Kaggle (https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features/data), whose original data come from the 1990 U.S. Census. The unit of observation is the census block group. The dataset contains housing and demographic characteristics for 20,640 block groups in California and is a commonly used benchmark in both machine learning and hedonic pricing research.

DOI: 10.54254/2753-8818/2025.DL27981

3.2. Variable composition and measurement units

For subsequent model preparation, this study groups the fields into four categories:

Target variable: median house value (1990 USD).

Population and housing stock: median income, housing median age (years), total rooms, total bedrooms, population, households.

Geographic location: latitude, longitude (decimal degrees).

Geographic accessibility (distance controls): distance _to _coast, distance _to _la, distance _to _san_diego, distance_to_san_jose, distance _to _san _francisco (great-circle distance; meters).

3.3. Descriptive statistics and distributional features

This study draws on California's 1990 census block group–level housing data, comprising 20,640 observations. Quality-check results are reported in Table 1. In terms of geographic coverage, latitudes span 32.54–41.95° N and longitudes -124.35 to -114.31° W, fully covering California's boundaries. Owing to a historical upper cap, 27 house price records are top-coded at $500,000 (1990 USD).

Table 1. Data quality summary and descriptive statistics of median house value

Section	Variable	Value
Data Overview	N rows	20,640.0
	Top coded 500k	27.0
	Latitude min	32.5°
	Latitude max	42.0°
	Longitude min	-124.3°
	Longitude max	-114.3°
Descriptive stats	median house value mean	206,855.82$
	median house value sd	115,395.62$
	median house value p25	119,600.00$
	median house value p50	179,700.00$
	median house value p75	264,725.00$

The lower panel of Table 1 reports descriptive statistics for the target variable median_house_value (in 1990 USD). For the target variable, median_house_value (1990 USD), the sample mean is $206,856, with a standard deviation of $115,396; the quartiles are P25 = $119,600, P50 = $179,700, and P75 = $264,725, indicating a right-skewed distribution with a long upper tail. For household income (median_income, measured in $10,000 units), the mean is 3.87 ($38,700), the median is 3.53, and P25–P75 spans 2.56–4.74. The other structural variables—total_rooms, total_bedrooms, population, and households—also display pronounced right skewness, consistent with patterns commonly observed in real estate data.

DOI: 10.54254/2753-8818/2025.DL27981

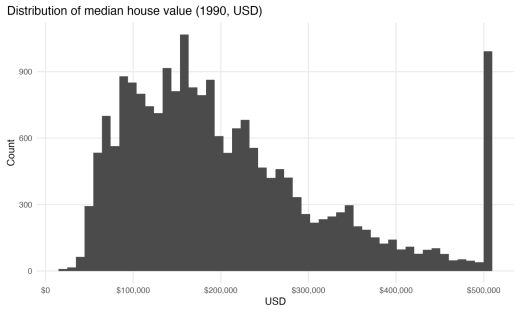

Figure 1. Distributing of median house value (1990, USD)

Figure 1 of median_house_value shows a clearly right-skewed distribution and a truncation spike at $500,000, corresponding to the aforementioned top-coded records. This distributional shape indicates that subsequent modeling should account for skewness and heteroskedasticity. The specific transformation and correction procedures will be detailed in the Methods section.

3.4. Feature engineering and transformations

To balance interpretability with numerical stability, this study adapts a minimal engineering strategy:

Random Forest target: set log(median_house_value) as the dependent variable, and apply Duan's smearing correction when converting predictions back to the dollar scale.

Right-skewed counts: for markedly right-skewed count-type features (total_rooms, total_bedrooms, households, population), apply log / log1p transformations to mitigate long tails (tree-based models are insensitive to scaling and generally more robust to extreme values).

OLS A1/A2: to preserve the economic meaning of coefficients, keep both the dependent variable and regressors in their original dollar/physical units (no standardization or normalization).

Geographic terms: keep latitude, longitude, and the five distance variables on their original scales; perform only missingness and range checks.

Overall, this study does not create complex derived features, ensuring that subsequent results are reproducible, interpretable, and mapped one-to-one to the specifications in the Method section.

3.5. Train/validation/test split

To ensure reproducibility, this study fixes the random seed at the sample level and adopt a 60% / 20% / 20% split, specifically, training set: 12,384; Validation set: 4,128; Test set: 4,128.

4. Methodology

4.1. Estimation strategy and evaluation metrics

All models are evaluated under the same random split (train/validation/test = 60%/20%/20%): estimation is conducted on the training set only; the validation set is used for specification and hyperparameter selection; the test set is strictly held out for final reporting to ensure reproducibility and to avoid data leakage. Evaluation metrics are reported on two scale:

Dollar scale (primary): report RMSE and R^2 . If a model is trained on the log scale (e.g., RF with z=ln y as the target), use Duan smearing to back-transform predictions to dollars.

Log scale (secondary): when necessary, provide RMSE (log) and R^2 (log) to facilitate comparisons of model fit stability.

DOI: 10.54254/2753-8818/2025.DL27981

Comparison and decision rule: models are primarily ranked by test-set R^2, while also reporting RMSE and the log-scale metrics to ensure cross-model comparability and consistency.

4.2. Model specifications

To address RQ1–RQ3, this study estimates three complementary models: two linear baselines (A1/A2) and one nonparametric ensemble (RF). Unless otherwise noted, models are fit on the training set, the validation set is used for tuning, and the test set is used once for evaluation.

4.2.1. OLS (A1: without geographic distances)

Dependent variable: median house price. Regressors: income, age, rooms, bedrooms, population, households, latitude, longitude (no distance controls).This model serves as a baseline that controls for location using latitude and longitude only.

4.2.2. OLS (A2: with five distance controls)

On top of A1, add distances to the coast and four major cities: distance_to_coast, distance_to_la, distance_to_san_diego, distance_to_san_jose, distance_to_san_francisco. This model captures explicit geographic accessibility and measures its incremental contribution to explanatory power.

4.2.3. Random Forest Regression (RF, log-y)

The Random Forest is trained on the log of house values. Predictions are back-transformed to the dollar scale using Duan's smearing correction, computed from training residuals only to avoid leakage. All reported RMSE and R^2 in dollars are based on these corrected predictions.

Features include income, age, log-transformed counts of rooms/bedrooms/population/households, latitude/longitude, and the five distance variables. This model captures nonlinearities and interactions, and, together with permutation importance and PDPs, provides interpretability.

4.3. Variable handling and feature engineering

4.3.1. Dependent variable and units

OLS (A1/A2): directly model median_house_value in dollar levels to keep coefficients economically interpretable (USD per unit).

RF: train on z=ln (MHV); apply Duan smearing to back-transform predictions to dollars before computing RMSE and R^2 .These addresses right skewness and heteroskedasticity.

4.3.2. Transformations and retention for covariates

Count variables (total_rooms, total_bedrooms, households, population):

RF: use logarithmic engineering (total_rooms_log, total_bedrooms_log1p, households_log1p, population_log1p) to smooth heavy tails and enhance split stability in trees;

OLS: retain original scales to preserve coefficient interpretability and ensure consistency with the hedonic regression literature.

Geographic variables: keep latitude, longitude, and the five straight-line distances (to coast / Los Angeles / San Diego / San Jose / San Francisco) in their original units (decimal degrees, meters). A2

DOI: 10.54254/2753-8818/2025.DL27981

includes these five distances explicitly in the linear regression to identify the incremental information from "accessibility"; RF uses both latitude/longitude and distances to capture nonlinear spatial gradients. Given pronounced spatial autocorrelation in house prices, omitting spatial/geographic controls can induce correlated residuals and biased coefficients [7]; retaining latitude/longitude and accessibility distances is therefore necessary.

Standardized coefficients: to compare effect sizes across variables in the linear model, this study additionally reports post-estimation standardized coefficients for A2.

4.3.3. Missing values, extremes, and top-coding

Core fields show no systematic missingness; due to the 1990 census cap at $500,000, there are 27 top-coded observations, all retained. The main specification applies neither Winsorization nor outlier removal.

4.3.4. Collinearity and information redundancy

Structural variables are correlated with population/households, but since the study prioritizes prediction and ranking, this study does not forcibly delete variables in the linear models; by reporting standardized coefficients and comparing RF permutation importance and PDPs, this study cross-validates the effect directions and relative magnitudes.

4.4. Interpretability and uncertainty assessment (VI & PDP)

To address the "key determinants and their consistency" in RQ1/RQ3, this study employs two tools on the final RF model, the first one is Permutation Importance (ΔR^2). On the test set, randomly permute a single feature repeatedly (≈60 times) and record the decline in R^2 (dollar scale) as ΔR^2; report the mean and standard deviation to quantify importance and uncertainty. A larger ΔR^2 indicates a more critical feature. This method is sign-agnostic, so it must be paired with PDPs for directional interpretation.

The second one is Partial Dependence Plots (PDP), which illustrate how the expected prediction changes with a focal variable, averaging over the distribution of other features [8]. For median_income, distance_to_coast, latitude, and longitude, construct PDPs to show the average change in predictions with respect to each variable. The model is trained on log(y); results are back-transformed to dollars via Duan smearing for direct monetary interpretation. Use ≈100 bootstrap replications to form 95% confidence bands, and extract the 10%–90% effect span to quantify magnitude.

4.5. Price-tier heterogeneity (RQ2)

To test whether the Low/Mid/High tiers are driven by different factors, this study takes the 30th and 70th percentiles of house prices on the training set ($q_{0.3}$, $q_{0.7}$) and, based on these thresholds, divide the sample into Low/Mid/High tiers. Quantile regression is a classic tool for characterizing heterogeneity in conditional distributions and can estimate the "shadow prices" of attributes at different price quantiles [9]. The same thresholds are then used to label the validation/test sets to avoid information leakage.

Within each tier, this study estimates A2 OLS independently (level. Each model is fit only on the training subsample of that tier, uses the validation subsample for robustness checks, and the test subsample for final evaluation.

DOI: 10.54254/2753-8818/2025.DL27981

Evaluation focuses on Test R^2 and also reports RMSE (USD); log-scale metrics are added when necessary. Results compare goodness of fit and coefficient differences by tier to reveal the drivers in each price segment and directly address RQ2.

4.6. Model comparison and selection

Primary metric: test R^2 ; also report $RMSE_{\$}$ and log-scale metrics.

Comparability: use a unified train/validation/test split to avoid data leakage.

Threshold: if the difference in test R^2 is within 0.02, treat performance as equivalent and prefer A2 (parsimonious and interpretable); if it exceeds this threshold, use RF as the primary predictive model, with A2 as the explanatory baseline.

Incremental value: the A2–A1 improvement measures the marginal contribution of the distance variables.

Consistency check: compare significant coefficients in A2 with RF permutation-importance and PDP rankings; if they converge, classify the factor as a robust key determinant.

5. Results

5.1. Predictive accuracy of the core models (dollar scale)

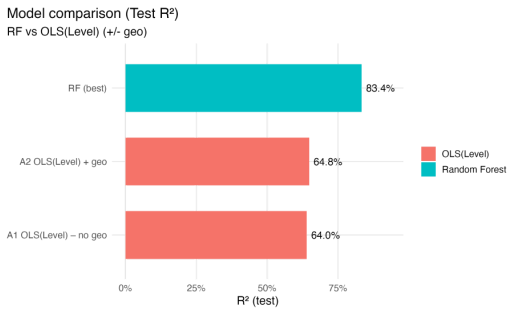

Figure 2. Model comparison (Test R2)

Table 2. The performance of the core model on the validation set and test

Family	Model	R2(Val)	RMSE(val,USD)	R2(Test)	RMSE(test,USD)
Random Forest	RF(best)	0.841	45774.34	0.834	47161.80
OLS(Level)	A2OLS(Level) + geo	0.647	68342.95	0.648	68568.84
OLS(Level)	A1 OLS(Level) -no geo	0.642	68777.75	0.640	69390.36

As shown in Figure 2, random Forest (RF) performs best. Table 2 demonstrates the details. Test R^2 =0.834, RMSE = $47,162, clearly outperforming the linear models. Relative to A2 OLS with geographic variables (R^2 =0.648, RMSE = $68,569), RF's R^2 is higher by $\approx$ 0.186, and RMSE is lower by $\approx$ $21,407 ($\approx$ 31%). Geographic information adds value to the linear model. A2 OLS versus A1 OLS (no geographic distances; R^2 =0.640, RMSE = $69,390) shows a slight gain: $\Delta R^2 \approx$ +0.008, RMSE decreases by $\approx$ $821 ($\approx$ 1.2%). Validation and test rankings are consistent (e.g., RF R^2 (val) = 0.841, A2 R^2 (val) = 0.647), indicating out-of-sample stability.In our sample and feature set, the nonlinear ensemble model (RF) clearly leads in predictive accuracy; within the linear framework, adding geographic features brings limited but steady improvement.

DOI: 10.54254/2753-8818/2025.DL27981

Subsequent chapters therefore use RF for importance and PDP interpretation, with A2 OLS retained as the interpretable and robustness benchmark.

5.2. RQ1: which features matter most

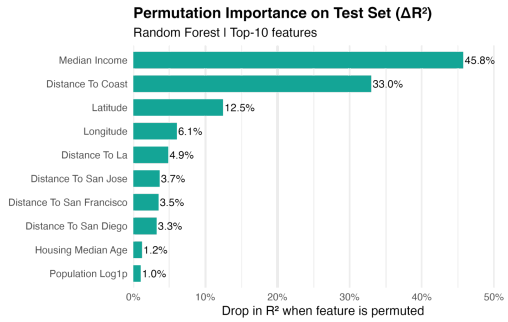

Figure 3. Permutation Importance on Test Set (ΔR2)

Holding other variables constant, by how much does the model's test-set R^2 fall (ΔR^2) when a single variable is permuted? A larger drop indicates a more "critical" variable. As shown in Figure 3, median_income stands out ($\Delta R^2 \approx 0$ .458, SD $\approx$ 0.011), the primary determinant of price. distance_to_coast is second ($\approx$ 0.330, SD $\approx$ 0.010), indicating a pronounced coastal premium. Latitude / longitude forms a third tier ($\approx$ 0.125 / 0.061 respectively), implying that coarse geographic position itself carries substantial price information. Distances to major cities—Los Angeles, San Jose, San Francisco, San Diego—each contribute ~3–5% in ΔR^2 (i.e., distance_to_la, distance_to_san_jose, distance_to_san_francisco, distance_to_san_diego), representing fine-grained accessibility gains. Structural variables such as housing_median_age and population_log1p contribute relatively little (around 1%).

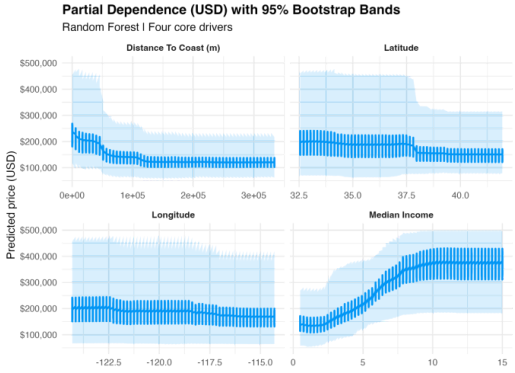

Figure 4. Partial dependence (USD) with 95% bootstrap bands

Median income monotonically increases with diminishing marginal effects at higher income levels, consistent with hedonic intuition; the PDP's 10–90% span indicates an impact on the order of hundreds of thousands of dollars. As shown in Figure 4, distance to coast: prices drop rapidly within 0–80/100 km, then level off, indicating that the coastal premium arises mainly in the near-coastal band. As for latitude, a clear "kink" appears around ≈37.5°, reflecting North–South differences (systematic price gaps between the Bay Area/Southern California and Northern California). For longitude, prices decline gradually from west to east, consistent with a spatial gradient of coast > inland.

DOI: 10.54254/2753-8818/2025.DL27981

Prices are driven by a combination of income and geography: income is the strongest single factor; coastal proximity and coarse latitude/longitude carry second-tier information; accessibility to core cities provides additional yet robust explanatory power. Structural count variables contribute, but in this dataset are markedly weaker than geography and income. This evidence is highly consistent with the classic hedonic pricing "location × purchasing power" framework and provides an anchor for subsequent RQ2 (price-tier heterogeneity) and RQ3 (cross-method consistency).

5.3. RQ2: price-tier heterogeneity, model fit and error structure across low/mid/high segments

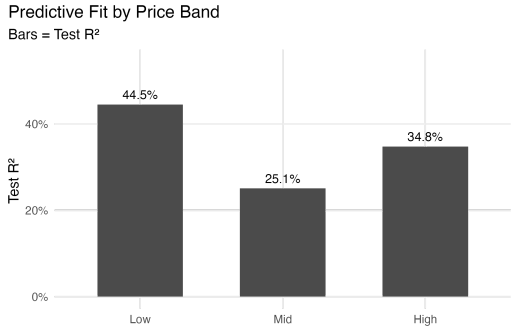

Figure 5. Predictive fit by price band

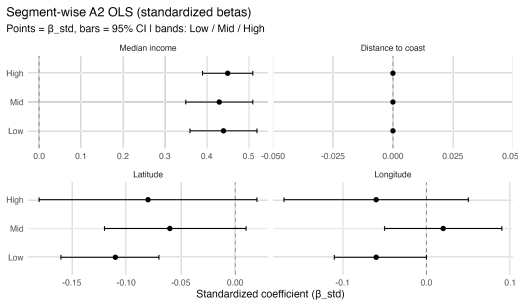

Figure 6. Segment-wise A2 OLS (standardized betas)

Explanatory power varies markedly by tier (see left panel). Low tier: Test R^2 ≈ 0.45 (RMSE ≈ \$17.7k), best fit. Mid-tier: R^2 ≈ 0.25 (RMSE ≈ \$26.9k), lowest explanatory power. High tier: R^2 ≈ 0.35 (RMSE ≈ \$71.8k), with much larger absolute errors, the R^2 values are visualized in Figure 5. Interpretation: the high-price tier is more influenced by unobserved granular factors (view/lot, subtle school-district differences, renovations, neighborhood brand, etc.), whereas the entry-level tier is more strongly driven by "basic attributes + accessibility." Core drivers across tiers (see right panel, β_{std} with 95% CI). As shown in Figure 6, income is a consistently strong positive factor: β_{std} ≈ 0.43–0.45 across tiers, CIs exclude zero. Purchasing power is a common driver in all tiers. Latitude is significantly negative, stronger in the Low tier; in Mid/High the negative effect is weaker and some CIs touch zero; consistent with the coastal–southern belt price gradient relative to Northern California. Distance to coast is negative but very small (rounded as −0.00); under linear OLS the effect is diluted by collinearity with latitude/longitude and city-distance variables and by linear form restriction. Longitude/city distances are generally weak and unstable across tiers, with many CIs touching zero; structural count variables (rooms/bedrooms/population/households) have limited influence at the block-group aggregation level.

DOI: 10.54254/2753-8818/2025.DL27981

So, the strengths of determinants differ by price tier. Income is a robust primary driver in all tiers; geographic gradients (latitude/coastal proximity) show clearer linear effects in the Low tier, whereas in the High tier they are more affected by unobserved heterogeneity and nonlinearities.

5.4. RQ3: do OLS and random forest converge on the same “key determinants”

Table 3. Comparison of variable Importance between OLS and RF models

Rank	OLS_Term	OLS_beta_std	OLS-p	RF-variable	RF_delta R^2	RF_SD
1	latitude	-0.860	0	median_income	0.458	0.011
2	distance_to-san_diego	0.703	0	distance_to_coast	0.330	0.010
3	median_income	0.628	0	latitude	0.125	0.004
4	population	-0.455	0	longitude	0.061	0.002
5	longitude	-0.433	0	distance_to_la	0.049	0.002
6	distance_to_la	-0.361	0	distance_to_san_jose	0.037	0.001
7	distance_to_san_ francisco	-0.305	0	distance_to_san_ francisco	0.035	0.001
8	distance_to-san_ jose	0.302	0	distance_to-san_diego	0.033	0.001
9	total_bedrooms	0.301	0	housing_median_age	0.012	0.001
10	households	0.253	0	population_log1p	0.010	0.001

Table 3 demonstrates the comparison of variable importance between OLS and RF models. Main findings show highly consistent "top four core drivers": median_income: OLS standardized coefficient +0.628 ($p \approx 0$); RF importance ≈ 0.458 — both methods rank it as the primary determinant. distance_to_coast: OLS +0.703 ($p \approx 0$); RF ≈ 0.330 — a "strong" geographic driver on par with income; PDPs show a pronounced coastal premium (see §7.2). latitude / longitude: OLS −0.860 / −0.443 ($p \approx 0$); RF ≈ 0.125 / 0.061 — both methods indicate that coarse lat/long gradients matter, but are weaker than income/coastal distance. City-distance variables are "secondary": distance_to_la / san_jose / san_francisco contribute about 3–5% ΔR^2 in RF; in OLS, signs and significance are affected by collinearity with coastal distance and lat/long, so directional interpretation should be cautious (overall information gain is limited). Scale/population variables are weaker: e.g., population_log1p, housing_median_age, total_bedrooms rank low in both methods (RF ≈ 1%; OLS $|\beta| < 0.31$), indicating limited marginal explanatory power that is readily absorbed by income and geography.

Despite different estimation frameworks and measurement scales (coefficients vs. ΔR^2), OLS and RF converge strongly on “what matters most”: income and geography (coastal proximity + latitude/longitude) form a stable, consistent core; accessibility to major cities follows; scale/population variables contribute the least

DOI: 10.54254/2753-8818/2025.DL27981

5.5. Robustness / sensitivity analyses

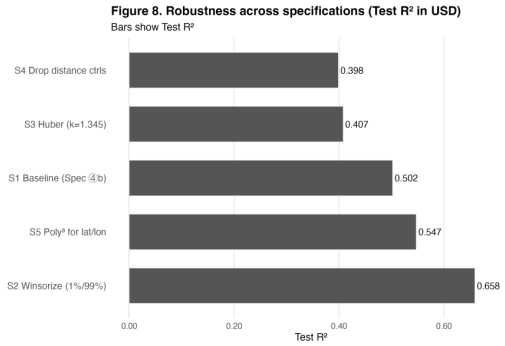

Figure 7. Robustness across specifications

This study introduces four robustness perturbations around the main linear baseline (A2: OLS with five distance controls + latitude/longitude + other covariates): S2: Winsorization (1%–99%). Mildly truncate extremes in the training set and apply the same thresholds to the validation/test sets. The cutoffs are derived solely from the training distribution (applied consistently to validation and test sets) to avoid data leakage, and dollar-scale metrics are always computed on back-transformed predictions. S3: Huber-RL. Replace least squares with the Huber loss to reduce outliers' influence on coefficients. The Huber loss uses threshold $\delta = 1.345\,\sigma$, with σ obtained as a robust scale (e.g., MAD $\times$ 1.4826) from training residuals. S4: Drop distance controls. Remove all distance variables (retain latitude/longitude) to test whether "distance to the coast/major cities" carries independent information. S5: Poly³ (latitude/longitude). Raise latitude/longitude from linear to cubic polynomials to test whether a more flexible spatial surface improves fit.

Mild truncation yields small gains. Figure 7 summarizes these robustness checks. S2's test R^2 (dollar scale) is slightly higher than the baseline, indicating that gentle treatment of heavy tails/outliers on the training side can enhance OLS stability, without altering the core ranking and conclusions. Huber is close to OLS. S3 is very similar to the baseline, suggesting that our conclusions are not "propped up" by a few extreme points. Dropping distances pronounced deterioration. S4's bars are visibly shorter (a sizable negative ΔR^2 relative to baseline), directly showing that "distance to the coast and accessibility to major cities" is a key information dimension not easily replaced by linear latitude/longitude. Higher-order spatial polynomials do not deliver meaningful gains. S5 is comparable to or slightly worse than the baseline, implying that, with the current sample size and variable set, linear lat/long plus explicit physical distances already capture the main spatial gradients; indiscriminate higher-order terms do not systematically improve generalization.

These robustness checks align with previous comments. Income and geographic accessibility (coast/cities) are first-order drivers of California house prices (S4's clear degradation corroborates this). The conclusions do not depend on extreme-value handling (S2/S3 agree) nor on higher-order spatial polynomials (S5 does not change the conclusions).

Therefore, our findings for RQ1 (relative performance ordering), RQ2 (price-tier heterogeneity), and RQ3 (cross-method consistency) are robust to methodological and data-processing choices.

DOI: 10.54254/2753-8818/2025.DL27981

6. Discussion

The cross-method consistent core variables are median_income and distance_to_coast, followed by latitude/longitude and distances to major metropolitan areas. Moving from A1 → A2 by adding distance controls yields $R^2 \approx$ +0.008, consistent with the RF importance ranking. In economic intuition, this reflects the combined force of "purchasing power — locational premium — coastal belt": income raises ability to pay; the coast and metropolitan corridors concentrate employment and public goods, forming a stable spatial gradient. PDP quantification: income from the 10th to the 90th percentile $\approx$ +$236k; near-to-far from the sea $\approx$ −$72k.

On the test set, RF attains $R^2 \approx$ 0.834, $RMSE_{\$} \approx$ 47k, clearly outperforming OLS (A2). The advantage lies in capturing nonlinearities, interactions, and thresholds, and in absorbing multicollinearity. Out-of-bag and test results both indicate that RF generalizes well rather than overfitting.

The quantile-based split shows: the low-price tier has higher explanatory power ($R^2 \approx 0.47$), while the high-price tier has larger errors ($RMSE_{\$} \approx$ 71k). This suggests that the high-end market is more affected by unobserved factors (school districts, views, renovations, scarcity). A one-size-fits-all model is suboptimal; stratified or segmented modeling is more robust.

Variants of A2 show: mild Winsorization and a cubic polynomial in latitude/longitude offer slight help; removing distance variables performs worst, indirectly demonstrating the necessity of geographic controls. This aligns with the RF importance ranking, indicating cross-method convergence.

Future research directions include adding features on more aspects, such as school quality, crime, commuting and environment; introducing SHAP/ICE explanations; adopting GAM/GWR/spatial models; using multi-period data for causal identification; and evaluating fairness and risk. And take a systematic review and practical guide as reference to model-agnostic interpretability methods such as PDP, ICE, and SHAP [10].

7. Conclusion

The main findings demonstrate that the formation mechanism of California house prices in 1990 reflects a typical income–coast–metropolitan-corridor structure. RF clearly outperforms OLS, indicating that nonlinearities and interactions are key; substantial heterogeneity exists within the price distribution. Robustness checks and cross-method validation strengthen confidence in the conclusions. In policy terms, focus on commuting improvements, stratified interventions, and data governance, while incorporating long-term risk considerations.

This study identifies several limitations and threats to validity: Top-coding: the $500k cap affects only 27 records but may understate the slope in high-price coastal areas; Aggregation level: block-group data weaken household-level differences; Missing variables: school quality, crime, commuting, etc. are absent; geographic distances are great-circle approximations; Temporal scope: a single 1990 cross-section limits extrapolation to the present; Endogeneity: income–price–location may be mutually influential; this study is prediction-oriented and does not make causal claims; Methodological scope: only OLS and RF are compared; broader ML/spatial econometric methods are not covered, potentially limiting the revelation of complex spatial dependence.

DOI: 10.54254/2753-8818/2025.DL27981

References

[1] National Association of Home Builders. (n.d.) Housing's Contribution to Gross Domestic Product. Retrieved from https: //www.nahb.org/news-and-economics/housing-economics/housings-economic-impact/housings-contribution-to-gross-domestic-product
[2] Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82, 34-55.
[3] Sirmans, G.S., Macpherson, D.A. and Zietz, E.N. (2005) The Composition of Hedonic Pricing Models. Journal of Real Estate Literature, 13, 1-44.
[4] Zietz, J., Zietz, E.N. and Sirmans, G.S. (2008) Determinants of House Prices: A Quantile Regression Approach. The Journal of Real Estate Finance and Economics, 37, 317-333.
[5] Čeh, M., Kilibarda, M., Lisec, A. and Bajat, B. (2018) Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. ISPRS International Journal of Geo-Information, 7, 168.
[6] Ho, W.K.O., Li, B.S.W. and Wong, S.W. (2021) House Price Prediction Using Machine Learning Algorithms: A Case Study of Hong Kong. Journal of Property Research, 38, 48-70.
[7] Basu, S. and Thibodeau, T.G. (1998) Analysis of Spatial Autocorrelation in House Prices. The Journal of Real Estate Finance and Economics, 17, 61-85.
[8] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29, 1189-1232.
[9] Koenker, R. and Bassett, G. (1978) Regression Quantiles. Econometrica, 46, 33-50.
[10] Molnar, C. (2022) Interpretable Machine Learning. Retrieved from https: //christophm.github.io/interpretable-ml-book/