The University of Chicago Booth School of Business
BUSN 41201 — Big Data

# What Moves the Price of a Los Angeles Home?

A Hedonic and Causal Machine-Learning Analysis of LA County Housing, and a Lesson in Honest Out-of-Sample Evaluation

Pablo Zavala  Will Sigal

**May 2024**

**Abstract**

We study the price per square foot of 3,804 single-family homes across 152 Los Angeles County cities, joining each home's structural attributes to engineered neighborhood features — proximity to the coast, schools, parks, top-rated restaurants, recent foreclosures, and census-tract crime, income, and density. Following the toolkit of Taddy (2019), we pursue two goals and keep them deliberately separate: *prediction* (how accurately a learner can forecast price, and which learner prevails) and *causal inference* (how much a given attribute moves price, holding the rest fixed). On prediction, a wide learner set — penalized linear models, principal-component regression, CART, random forests, extremely randomized trees, gradient boosting (XGBoost, LightGBM, histogram GBM), kernel SVR, a neural network, and a stacked super-learner — competes under a single nested cross-validation protocol, each learner hyperparameter-tuned. The strongest, a tuned Stacked ensemble, reaches $R^2 = 0.621$ under random folds. Our central methodological point tempers that headline: the number flatters the models. As the held-out regions grow, skill decays smoothly — from $\approx 0.6$ under random folds, to $R^2 = 0.431$ under spatial-block folds, toward zero under leave-one-region-out — revealing that interpolation between near neighbors supplies a large share of the random-CV figure. On causation, cross-fitted Double/Debiased ML (Chernozhukov et al., 2018) with spatial block-bootstrap inference shows the coastal premium surviving full adjustment at roughly +29.8% per square foot (95% CI $[0.16, 0.33]$ in log points), whereas the apparent private-school "premium" dissolves once the estimate accounts for the neighborhoods that host private schools. Every figure, table, and number regenerates from committed code, free of proprietary data.

**Keywords:** hedonic prices; spatial cross-validation; double/debiased machine learning; gradient boosting; regularized regression; Los Angeles housing.
**JEL codes:** C21, C45, C52, R21, R31.

## 0. Executive summary

- **Question.** Price per square foot varies several-fold across Los Angeles County, often within a few miles. From 3,804 single-family listings enriched with structural and neighborhood features, we ask what explains that variation and how faithfully any learner can predict it, deploying the full Taddy (2019) toolkit together with a causal-ML extension.

- **Best predictive model.** Under random 5-fold cross-validation, a tuned Stacked ensemble leads the sweep with out-of-sample $R^2 = 0.621$, narrowly ahead of LightGBM and the tuned boosting and penalized-linear learners — though only slim margins separate the strong performers.
- **The honesty result (our headline).** That figure flatters every learner. Under *spatial* cross-validation — scoring on held-out geographic blocks — performance falls to roughly $R^2 = 0.431$, and under leave-one-region-out it decays toward zero. Skill erodes smoothly with extrapolation distance, which exposes much of the random-CV number as interpolation between near neighbors rather than transferable structure. Tellingly, the humble regularized linear model extrapolates across space better than the boosted ensembles that win the random-CV race. Anyone reporting a single cross-validated number on spatial data owes the reader this spectrum.
- **What drives price (association).** A hedonic model with robust standard errors and a permutation-importance analysis converge on one ranking: coastal proximity dominates, with neighborhood income, amenity density (restaurants, parks), and living area (negative, per square foot) following.
- **What drives price (causation).** Cross-fitted Double/Debiased ML places the coastal premium near +29.8% per square foot after partialling out every other attribute; the naive private-school "premium" collapses to zero once the estimate accounts for *where* private schools locate; and the foreclosure externality registers negative, though with modest precision.
- **Reproducibility.** The entire analysis regenerates offline from one cleaned table with a single command. The code writes every number cited in the prose directly to the manuscript's macros, so the paper and the results move together.

# 1 Introduction

Los Angeles ranks among the least affordable housing markets in the United States, and within the county the price of a home swings enormously over short distances — a beach bungalow and an inland tract house of identical size can differ several-fold in price per square foot. What accounts for such variation? Part lies in the house itself: its size, age, bathrooms, and stories. Part lies in the neighborhood: schools, parks, restaurants, safety, income, and — Los Angeles obliging — the miles separating the front door from the ocean.

The project grew from a straightforward question: gather everything measurable about a home and its surroundings, and discover how much of the price yields to explanation — and which factors carry the most weight. To that end, we assemble structural attributes from residential listings and engineer a battery of neighborhood features, spatially joining homes to schools, parks, restaurants, foreclosure filings, the coastline, and census-tract crime, income, and population density.

Following Taddy (2019), whose `gamlr`, cross-validation, and orthogonal-ML machinery anchor the Booth Big Data curriculum, we pursue two goals and keep them deliberately separate, because they demand different tools and answer different questions (Mullainathan and Spiess, 2017):

1. **Prediction.** How accurately can we forecast a home's price out of sample, and which model family prevails? The sweep compares penalized linear models (Tibshirani, 1996; Hastie et al., 2015), principal-component regression (Jolliffe, 2002), trees (Breiman et al., 1984), random and extremely randomized forests (Breiman, 2001), gradient boosting (Friedman, 2001; Chen and Guestrin, 2016; Ke et al., 2017), kernel machines, a neural network, and a stacked super-learner.

2. **Causation.** Holding other attributes fixed, how much does a specific feature — coastal proximity, private-school access, foreclosure exposure — move the price? Here the hedonic framework (Rosen, 1974; Harrison and Rubinfeld, 1978) supplies the association, and cross-fitted Double/Debiased ML (Chernozhukov et al., 2018, 2024) supplies a causal reading under explicit assumptions.

The prediction goal conceals a trap. Home prices exhibit strong *spatial autocorrelation*: nearby homes command similar prices (Tobler's first law, Tobler, 1970). Ordinary random cross-validation scatters a home's near neighbors across training and test folds, so the model merely interpolates between points it has already half-seen. The resulting $R^2$ can look impressive while revealing little about performance in a neighborhood the model meets for the first time (Roberts et al., 2017). We therefore evaluate every model twice — random folds and spatial-block folds — and trace the full spatial-optimism spectrum between the two regimes. The gap between those numbers constitutes, in our view, the paper's most important finding.

Every number below regenerates from committed code, and the road map runs as follows: Section 2 describes the data and its governance; Section 3 fixes the evaluation protocol; Sections 4–5 report predictive performance and interpret it; Section 6 turns to causal estimates; Sections 7–8 examine spatial and unsupervised structure; Sections 9–10 draw lessons, concede limits, and document reproducibility.

# 2 Data

## 2.1 Sources and assembly

The analysis table records one row per single-family residential listing in Los Angeles County, assembled from residential listing data (Zillow-derived) and filtered to the county and to single-family homes with a positive price. To each home we then attach neighborhood context through spatial joins and buffer counts:

- **Schools** (LA County GIS): counts within a fixed radius, plus an indicator for a private school within two miles.
- **Crime** (City of Los Angeles open data): violent-crime incidents aggregated to the home's census tract, as a count and a per-capita rate.
- **Census (ACS 2016–2020)** (U.S. Census Bureau, 2020): tract median household income and population density.
- **Parks** (LA GeoHub): total park acreage within two miles.
- **Restaurants** (top-1000 Yelp-rated): count within three miles, a proxy for amenity density and walkable urbanism.
- **Foreclosures** (2021 registered filings, LA City): count within one mile — the one-mile radius follows the foreclosure-externality literature (Immergluck and Smith, 2006).
- **Coast**: distance to the Pacific coastline and an indicator for sitting within half a mile of it.

After cleaning and validation (unique addresses, a positive and finite target, coordinates inside an LA-County bounding box), the table holds 3,804 homes across 152 cities. Table 9 in the appendix records the full generated data dictionary.

### 2.2 Data governance

We withhold all raw and home-level data from publication. Third-party terms govern the listing table, so it stays out of version control; only aggregated tables and figures reach the repository. For every source, `data/README.md` records provenance and licensing.

### 2.3 The target

We model log(price/square foot). Working in logs tames a heavy right tail (Figure 1) and lets coefficients read as approximate percentage effects; price per square foot, rather than raw price, offers the natural unit for comparing homes of very different sizes. We take the log exactly once — an easy step to duplicate by accident when a dataset already ships log-scaled columns. With the table assembled and the target fixed, the burden shifts to evaluation: Section 3 builds a protocol that scores every learner honestly.

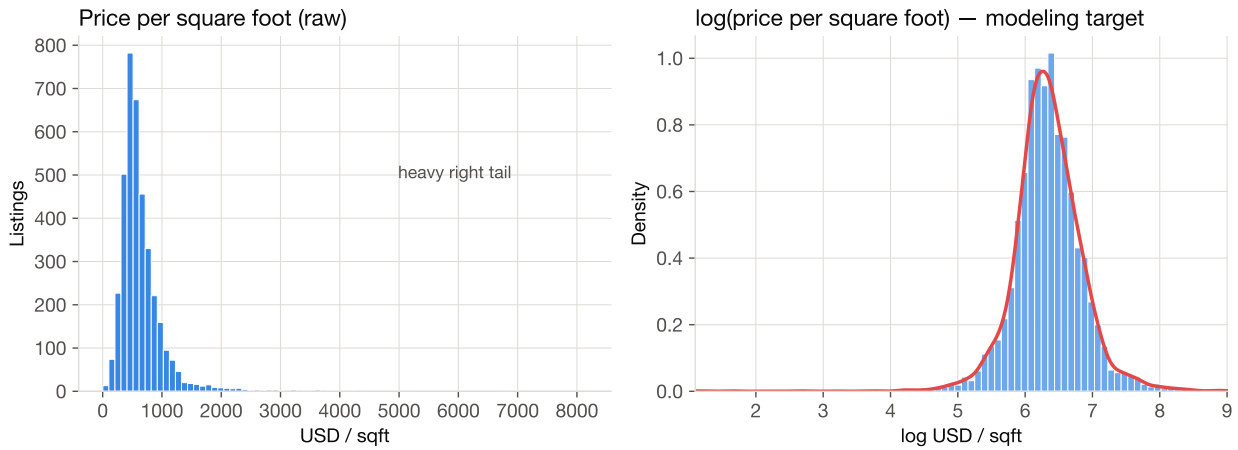


**Figure 1:** The raw price per square foot carries a heavy right tail (left); the log transform (right) approaches symmetry and serves as our modeling target. Figures throughout follow the design principles of Wilke (2019) and Tufte (2001), rendered in a colorblind-safe palette in the spirit of Wickham (2016).

## 3 Methods

### 3.1 Leakage-free preprocessing

Every transformation that learns from the data — imputation medians, scaling statistics, one-hot vocabularies, principal-component rotations — lives inside a single pipeline fit *only* on the training portion of each cross-validation fold. That discipline matters because the alternative — standardizing, imputing, or selecting a feature radius on the full dataset before cross-validating — lets test information leak into training. Within the pipeline, skewed positive features receive a $\log(1+x)$ transform before standardization, and the 152-level `city` identifier enters one-hot encoded with rare levels collapsed. All software builds on scikit-learn (Pedregosa et al., 2011).

### 3.2 A single evaluation protocol for every model

With preprocessing sealed inside the folds, we score all models on *identical* folds with the same metrics ($R^2$, RMSE, MAE), under two regimes:

- **Random** 5-fold cross-validation.
- **Spatial-block** 5-fold cross-validation, whose folds derive from $k$-means clusters of the projected coordinates, so training and test sets occupy largely separate geography (Roberts et al., 2017). (Cluster CV leaves block borders unbuffered, so it leans conservative while stopping short of a perfect spatial test; the spectrum below stresses it further.)

Throughout, we report *pooled* out-of-fold $R^2$ — a single $R^2$ computed from the concatenated held-out predictions, with a block-aware bootstrap standard error — in place of an average of per-fold $R^2$, which wobbles across spatial folds of very different size (a point raised in review). For the technique sweep, an *inner* randomized search tunes hyperparameters on the training portion of every outer fold (nested CV), so nothing gets tuned on the data that scores it (Kuhn and Johnson, 2013; Hastie et al., 2009); because the inner folds stay random, the tuned numbers lean, if anything, generous toward spatial transfer. Finally, to probe how far skill extrapolates, we trace a *spatial spectrum*: leave-one-block-out $R^2$ as the number of blocks shrinks and each held-out region grows.

### 3.3 The learner set

The baseline tier fixes sensible hyperparameters: mean baseline, OLS, ridge, LASSO, elastic net, PCR, CART, random forest, XGBoost. The tuned tier (nested CV) then adds $k$-NN, kernel SVR, extremely randomized trees, histogram gradient boosting, LightGBM, a multilayer perceptron, and a stacked super-learner with a ridge meta-model over the strongest bases. Boosting runs with fixed, regularized settings and forgoes test-set early stopping, which would let the boosted learners peek at the very data that scores them.

### 3.4 Hedonic inference

For association, we fit an ordinary least-squares hedonic model on the standardized design with HC3 heteroskedasticity-robust standard errors (MacKinnon and White, 1985), using reference-level (drop-first) encoding so the dummy set and the intercept stay linearly independent. Standardized coefficients then compare directly in magnitude. We present these estimates as associations rather than causal effects, and we flag multicollinearity with variance-inflation factors.

### 3.5 Causal machine learning

To move toward causation for a few policy-relevant attributes, we adopt the partially-linear model with cross-fitted Double/Debiased ML (Chernozhukov et al., 2018), the "orthogonal ML" of Taddy (2019, Ch. 6):

$$Y = \theta\, D + g(X) + \varepsilon, \qquad D = m(X) + v, \tag{1}$$

where $Y$ denotes log price/sqft, $D$ a treatment (e.g. coastal proximity), and $X$ all other attributes. Gradient-boosted trees learn the nuisances $g(X) = \mathbb{E}[Y \mid X]$ and $m(X) = \mathbb{E}[D \mid X]$; ordinary *random* five-fold cross-fitting — the Chernozhukov et al. standard, whose role consists of keeping each observation's own data out of its nuisance prediction rather than testing spatial transfer — precedes an orthogonalized (Robinson) partialling-out that delivers $\hat{\theta}$. Neyman orthogonality renders $\hat{\theta}$ first-order insensitive to nuisance error, and cross-fitting removes own-observation overfitting bias (Athey and Imbens, 2019; Chernozhukov et al., 2024). For inference, we deploy a *spatial block bootstrap* (resampling $k$-means geographic clusters), because the iid influence-function standard error understates uncertainty under the residual spatial autocorrelation we document in Section 7. These estimates remain observational: they identify a causal effect only under conditional ignorability

given $X$. Accordingly, we report each one beside the naive difference, the OLS-adjusted coefficient, and a control-set sensitivity, so the reader watches how far adjustment moves the number.

### 3.6 Spatial diagnostics

Finally, we test residuals for spatial structure with global Moran's $I$ (Moran, 1950; Anselin, 1988) under a row-standardized $k$-nearest-neighbor weight and a 999-permutation null, computed on genuinely out-of-sample residuals and validated against a known gradient. With the protocol set, we turn to results.

## 4 Predictive modeling

### 4.1 Baseline comparison

We begin with the standard learners at sensible defaults, all scored on the same folds (Table 1, Figure 2). Under random cross-validation the models cluster tightly: gradient boosting leads, the penalized linear models and the random forest follow, and CART and principal-component regression trail. Even ordinary least squares reaches $R^2 = 0.587$ — a reminder that, armed with 152 city fixed effects, a linear model resists improvement by much.

**Table 1:** Out-of-sample performance under a unified cross-validation protocol. Identical folds score every model; parentheses hold block-aware bootstrap standard errors of the pooled out-of-fold $R^2$.

| Model | $R^2$ random | $R^2$ spatial | RMSE random | MAE random |
|---|---|---|---|---|
| XGBoost | 0.608 (0.03) | 0.414 (0.10) | 0.321 | 0.191 |
| LASSO | 0.588 (0.03) | 0.436 (0.13) | 0.329 | 0.201 |
| Elastic Net | 0.588 (0.03) | 0.436 (0.13) | 0.329 | 0.201 |
| Ridge | 0.587 (0.03) | 0.434 (0.13) | 0.329 | 0.201 |
| OLS | 0.587 (0.03) | 0.431 (0.13) | 0.329 | 0.201 |
| Random Forest | 0.562 (0.03) | 0.309 (0.08) | 0.339 | 0.212 |
| CART | 0.446 (0.02) | 0.256 (0.08) | 0.382 | 0.256 |
| PCR (90% var) | 0.381 (0.03) | 0.300 (0.12) | 0.403 | 0.273 |
| Baseline (mean) | -0.000 (0.00) | -0.022 (0.05) | 0.513 | 0.369 |

*Note.* The target reads log(price per square foot). Random and spatial-block folds share the same data; the gap measures skill attributable to spatial interpolation.

### 4.2 The full technique sweep, hyperparameter-tuned

Next we let each learner tune itself under nested cross-validation, adding extremely randomized trees, histogram and LightGBM boosting, kernel SVR, a neural network, and a stacked super-learner (Table 2, Figure 3). Tuning helps at the margins yet leaves the story intact: the strong learners finish within a few points of one another, and a tuned Stacked ensemble tops the random-CV ranking at $R^2 = 0.621$. The stacked ensemble competes without dominating — when the base learners largely agree, stacking finds little disagreement to exploit. Table 8 in the appendix lists the selected hyperparameters.

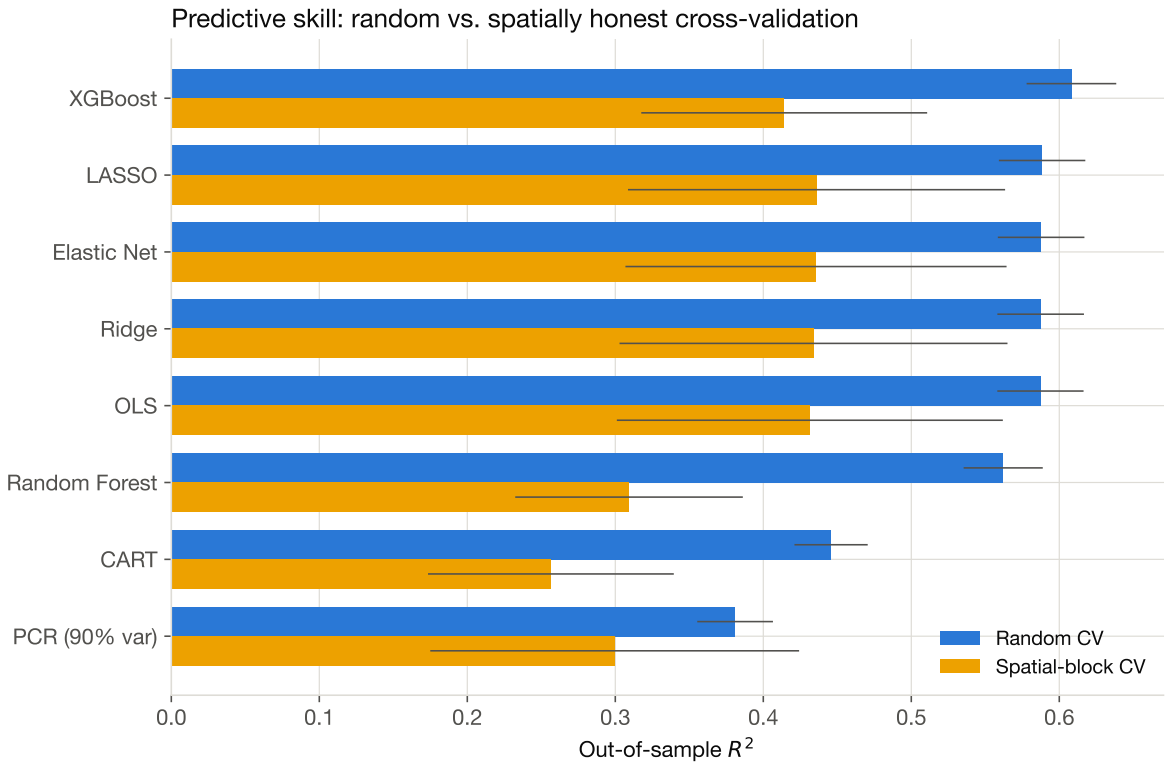


**Figure 2:** Out-of-sample $R^2$ under random (blue) versus spatial-block (orange) cross-validation. Every model loses ground once whole neighborhoods leave the training data.

**Table 2:** Hyperparameter-tuned learners under nested cross-validation. An inner randomized search tunes each learner on every outer training fold; the held-out outer folds supply the reported $R^2$ (pooled, with block-aware bootstrap standard errors in parentheses).

| Learner | $R^2$ random | $R^2$ spatial | RMSE random |
|---|---|---|---|
| Stacked ensemble | 0.621 (0.03) | 0.431 (0.09) | 0.315 |
| LightGBM | 0.615 (0.03) | 0.417 (0.08) | 0.318 |
| Random Forest | 0.611 (0.03) | 0.388 (0.08) | 0.320 |
| Hist Gradient Boosting | 0.611 (0.03) | 0.428 (0.09) | 0.320 |
| XGBoost | 0.610 (0.03) | 0.419 (0.08) | 0.320 |
| Extra Trees | 0.594 (0.03) | 0.351 (0.09) | 0.327 |
| SVR (RBF) | 0.582 (0.03) | 0.416 (0.10) | 0.331 |
| Elastic Net | 0.582 (0.03) | 0.448 (0.12) | 0.331 |
| KNN | 0.529 (0.03) | 0.239 (0.09) | 0.352 |
| MLP | 0.499 (0.03) | -0.047 (0.22) | 0.363 |

*Note.* Tuning explores elastic-net penalties, kernel/tree/boosting depths and rates, and network widths. The random-vs-spatial gap persists after tuning.

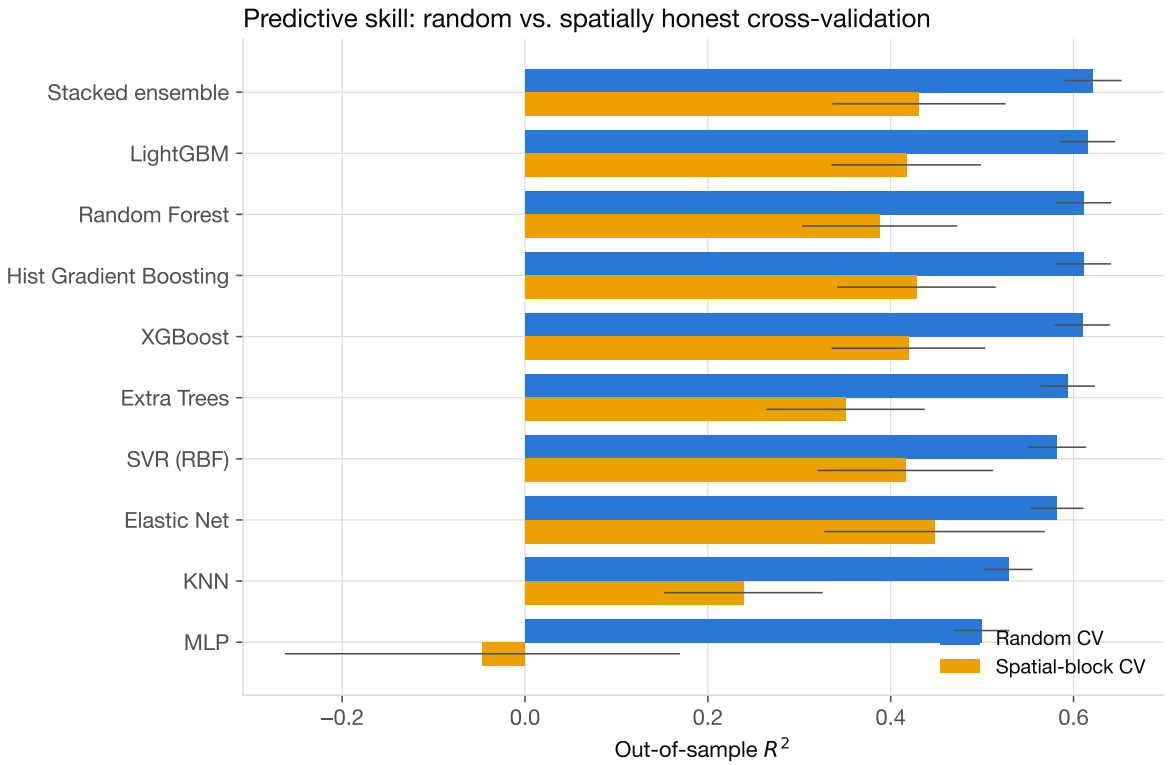


**Figure 3:** Hyperparameter-tuned learners under nested cross-validation, random (blue) versus spatial (orange). Tuning lifts the weaker learners toward the leaders yet leaves the random-versus-spatial gap intact.

### 4.3 The spatial reckoning

Now comes the honesty check. Every number above descends from random CV, and random CV asks the wrong question of spatially dependent data. Three regimes, each answering its own question, clarify the picture:

- **Random CV** ($R^2 \approx 0.6$): how well can we price a home whose comparable neighbors already sit in the training data? Quite well.
- **$k$-means spatial-block CV** ($R^2 \approx 0.35$–$0.45$): how well do we price homes when two of ten blocks stay held out? Performance drops sharply, yet real skill survives — the continuous neighborhood signal (income, coast distance) still transfers. Already at this regime the ranking begins to invert: the tuned elastic net posts the best spatial score (0.45), edging the stacked ensemble (0.431).
- **Leave-one-region-out** (Figure 4): as the held-out region grows to half the county, the flexible learners collapse — XGBoost's $R^2$ sinks to essentially zero ($\approx 0.01$) — while the regularized linear model degrades far more gently, holding near 0.28. What looked like the boosted models' edge amounted mostly to interpolation; only the linear signal travels to genuinely new ground.

Consequently, the honest summary reads "skill degrades with extrapolation distance, and unequally across learners," rather than "the models fail." Interpolation between nearby homes supplies a large share of the boosted models' random-CV number, and a single cross-validated $R^2$, reported without this spectrum, would overstate what transfers. The striking corollary: the humble regularized linear model (elastic net / LASSO) extrapolates across space best of all, out-generalizing the boosted ensembles that win the random-CV race precisely because it declines to memorize local structure.

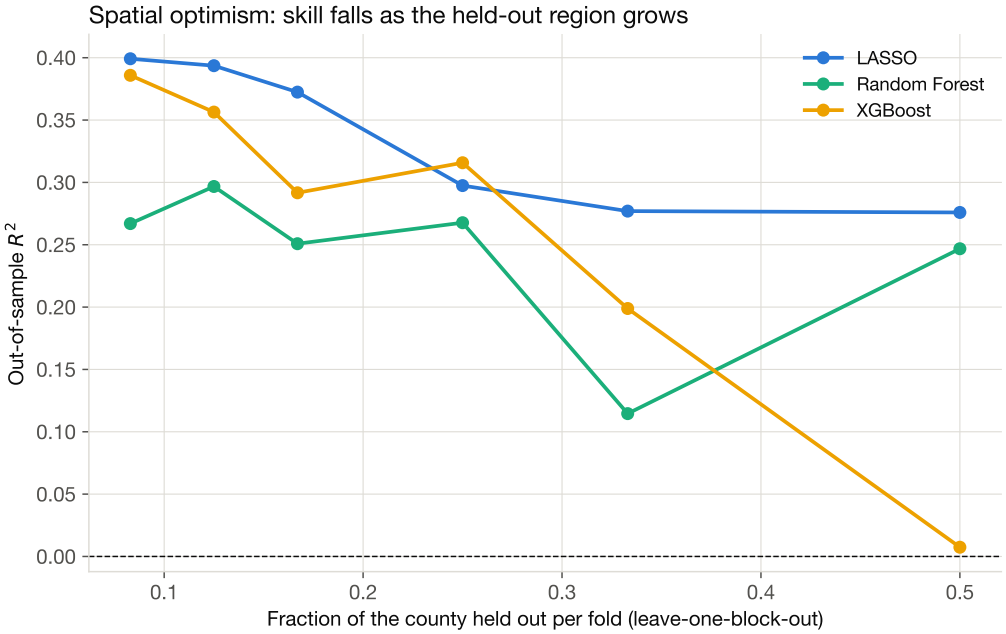


**Figure 4:** Spatial-optimism spectrum: out-of-sample $R^2$ as a function of the share of the county held out per fold. Skill decays from its random-CV level toward zero as extrapolation distance grows.

### 4.4 How much does the neighborhood identifier buy?

A natural question suggests itself: remove the cities and see how much predictive power we lose. We answer it cleanly. Dropping the 152-level `city` identifier lowers OLS random-CV $R^2$ by 0.093 (Figure 5). A tenth of the explainable variation rides on nothing more than which city a home occupies — coarse location, encoded as a fixed effect, performs real work that the measured neighborhood features only partly capture. That residual power of raw location foreshadows the spatial diagnostics of Section 7.

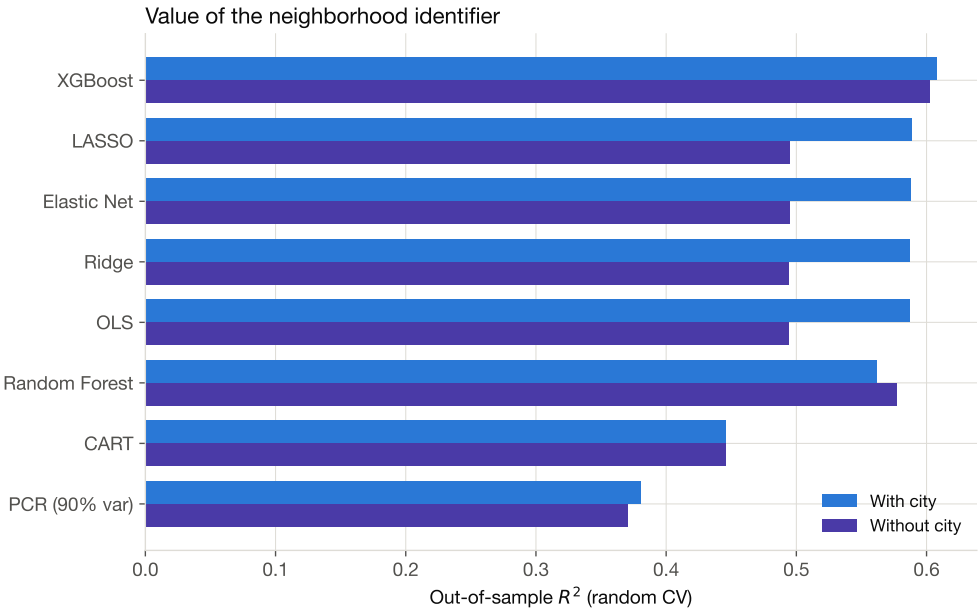


**Figure 5:** Predictive $R^2$ with and without the city fixed effect. The coarse location identifier contributes about 0.093 to $R^2$ on its own.

## 5 What drives price (association)

Prediction tells us *how well*; interpretation tells us *what matters*. Three lenses — hedonic coefficients, permutation importance, and partial dependence — converge on one answer.

### 5.1 Hedonic associations

The hedonic OLS narrates a coherent Los Angeles story (Figure 6, Table 3). Sitting within half a mile of the coast carries the single largest positive association; neighborhood income, nearby top-rated restaurants, private-school proximity, and park acreage all enter with positive, tightly estimated coefficients. On the negative side, living area exerts a strong downward pull — larger homes sell for less *per square foot*, the standard diminishing-returns result — while nearby foreclosures and distance from the coast likewise depress the price per square foot. Notably, the foreclosure estimate recovers, with the expected sign and robust standard errors, the neighborhood-externality effect that motivated our one-mile buffer (Immergluck and Smith, 2006).

These coefficients measure association rather than causation, and several predictors move together. Table 4 reports variance-inflation factors: the two crime measures and the two school-radius counts approach redundancy — a hazard worth flagging before any coefficient reading. We keep them for

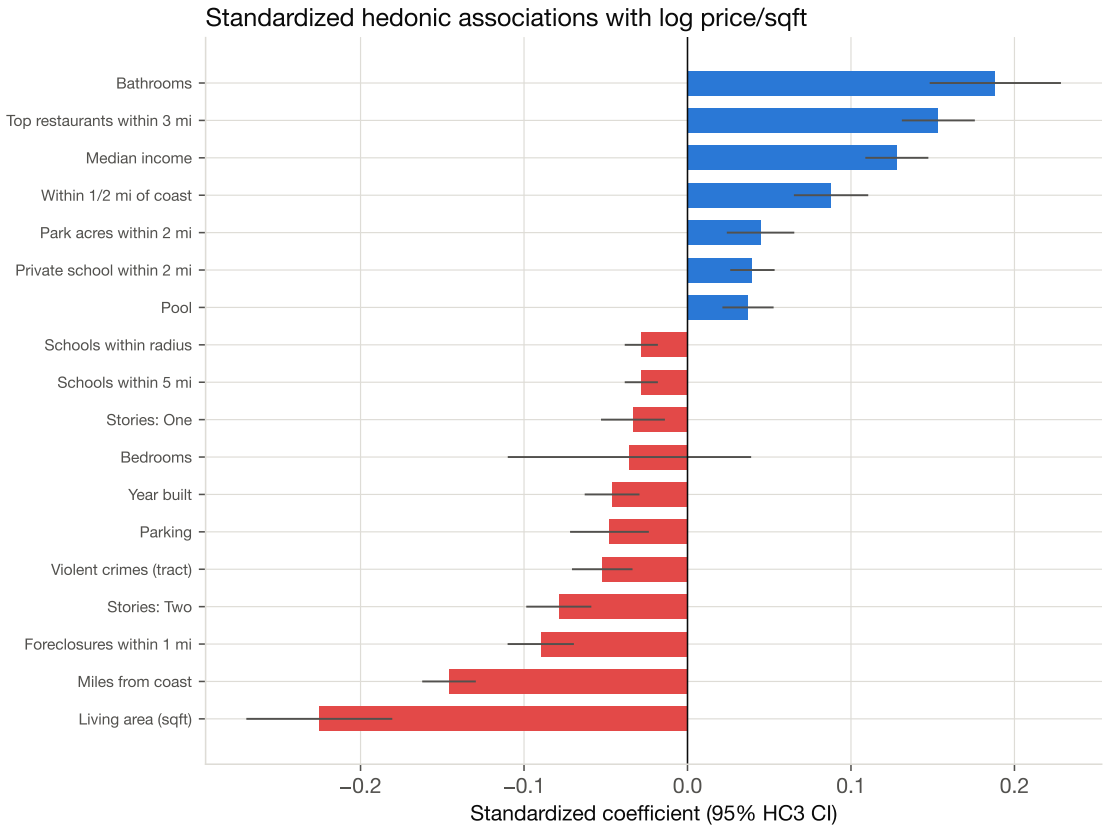


**Figure 6:** Standardized hedonic associations with log price per square foot (OLS, HC3-robust 95% intervals). Blue marks positive, red negative.

**Table 3:** Largest standardized associations with log price per square foot (OLS with HC3 robust standard errors).

| Attribute | Std. coef. | 95% CI | $p$ |
|---|---|---|---|
| Living area (sqft) | -0.225 | [-0.27, -0.18] | <0.001 |
| Bathrooms | 0.188 | [0.15, 0.23] | <0.001 |
| Top restaurants within 3 mi | 0.153 | [0.13, 0.18] | <0.001 |
| Miles from coast | -0.146 | [-0.16, -0.13] | <0.001 |
| Median income | 0.128 | [0.11, 0.15] | <0.001 |
| Foreclosures within 1 mi | -0.090 | [-0.11, -0.07] | <0.001 |
| Within 1/2 mi of coast | 0.088 | [0.07, 0.11] | <0.001 |
| Stories: Two | -0.079 | [-0.10, -0.06] | <0.001 |
| Violent crimes (tract) | -0.052 | [-0.07, -0.03] | <0.001 |
| Parking | -0.048 | [-0.07, -0.02] | <0.001 |
| Year built | -0.046 | [-0.06, -0.03] | <0.001 |
| Park acres within 2 mi | 0.045 | [0.02, 0.07] | <0.001 |
| Private school within 2 mi | 0.040 | [0.03, 0.05] | <0.001 |
| Pool | 0.037 | [0.02, 0.05] | <0.001 |
| Bedrooms | -0.036 | [-0.11, 0.04] | 0.350 |

*Note.* Associative rather than causal. Coefficients apply to standardized predictors, so magnitudes compare directly.

prediction, where regularization absorbs the overlap, flag them for interpretation, and revisit the strongest attributes causally in Section 6.

**Table 4:** Variance-inflation factors for numeric predictors (largest shown).

| Feature | VIF |
|---|---|
| Violent crimes (tract) | 14.54 |
| Violent crime rate | 14.19 |
| Bathrooms | 4.64 |
| Parking | 4.28 |
| Has garage | 3.98 |
| Living area (sqft) | 3.52 |
| Garage spaces | 2.59 |
| Median income | 2.17 |
| Population density | 2.15 |
| Bedrooms | 1.99 |
| Miles from coast | 1.66 |
| Top restaurants within 3 mi | 1.63 |

*Note.* VIF $> 10$ signals collinearity; the two school-count radii approach redundancy.

### 5.2 Permutation importance and partial dependence

A held-out permutation-importance analysis on the random forest corroborates the hedonic story (Figure 7): distance to the coast ranks as the single most useful predictor, with amenity density, neighborhood income, and living area following. The partial-dependence panels (Figure 8) then supply the *shape* of those relationships, traced from the stacked ensemble: price per square foot falls steeply with distance from the coast, climbs with tract income, and declines with home size — monotone, economically sensible curves, free of the artefacts a single mis-specified linear term can introduce. With the associations established, causation demands its own machinery; Section 6 supplies it.

## 6 Causal machine learning: treatment effects

Association stops short of causation. A coastal home commands a high price, but it also enjoys good schools, high-income neighbors, and restaurants; a bare coefficient on "coast" absorbs all of that. To learn what a given attribute contributes on its own, holding the rest fixed, we turn to cross-fitted Double/Debiased ML (Section 3.5). We cast three binary attributes and one continuous one as "treatments," partial out every other feature with gradient-boosted nuisances, and compare three estimates: the naive difference in means, the linear OLS-adjusted coefficient, and the orthogonalized Double ML estimate (Table 5, Figure 9).

Three findings stand out.

- **The coastal premium survives adjustment, at scale.** The naive within-half-mile gap looks enormous, but confounding inflates much of it. After Double ML partials out income, amenities, density, and structure, the adjusted coastal effect settles at $\hat{\theta} = 0.261$ log points (95% block-bootstrap CI $[0.16, 0.33]$), about +29.8% on price per square foot. The continuous specification agrees: each additional mile from the coast shaves roughly two percent off the price per square foot.

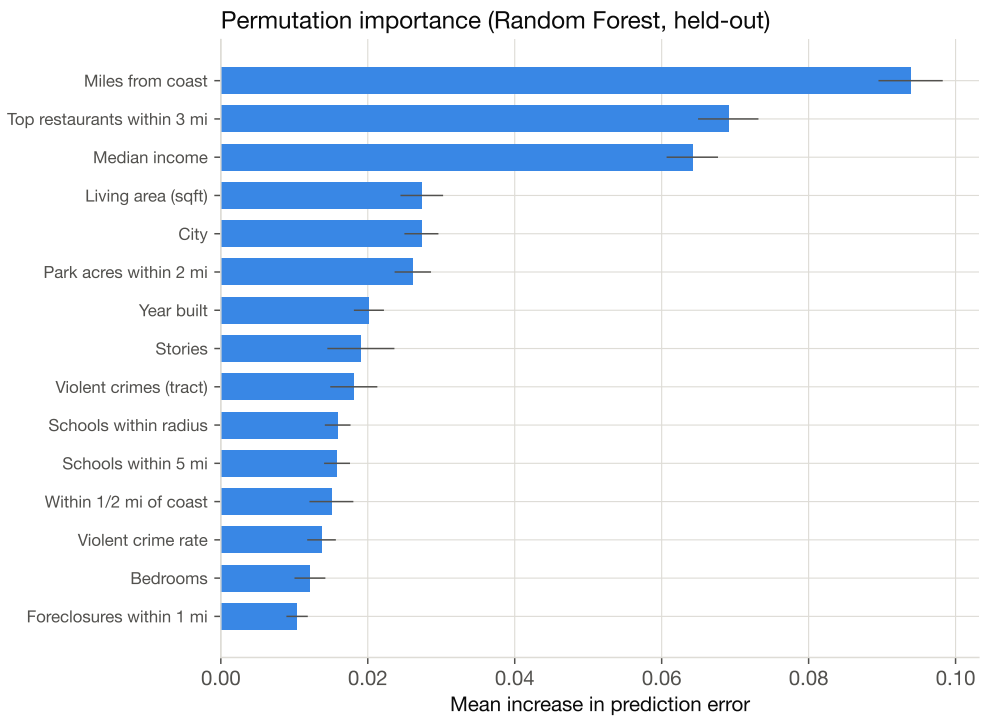


**Figure 7:** Random-forest permutation importance (held-out). Location and amenity/income features dominate the structural details.

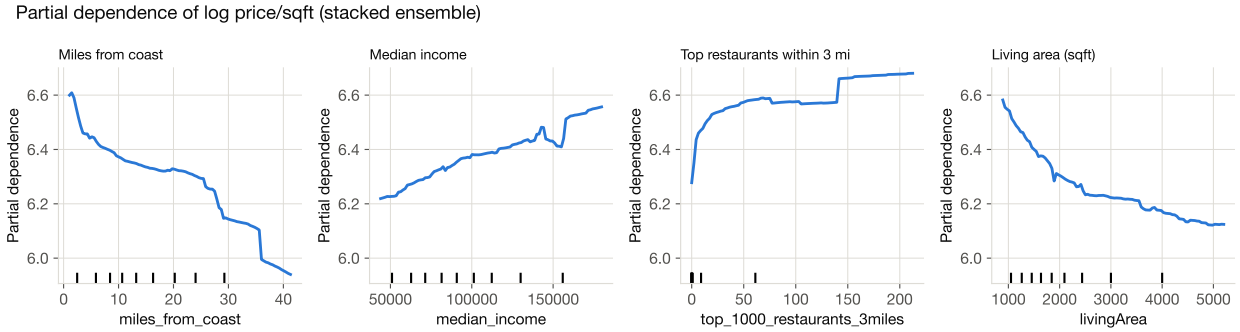


**Figure 8:** Partial dependence of log price/sqft on the leading drivers (stacked ensemble). The coastal gradient and the income and size effects run smooth and monotone.

**Table 5:** Treatment effects on log price per square foot: naive difference, linear-adjusted OLS, and cross-fitted Double/Debiased ML.

| Treatment | $n$ treated | Naive | OLS-adj. | Double ML | 95% CI | Effect |
|---|---|---|---|---|---|---|
| Within 1/2 mile of coast | 100 | 0.904 | 0.282 | 0.261 | [0.16, 0.33] | +29.8% |
| Private school within 2 mi | 3578 | 0.442 | 0.074 | 0.007 | [-0.11, 0.11] | +0.7% |
| Any foreclosure within 1 mi | 1660 | 0.265 | -0.047 | -0.072 | [-0.14, -0.01] | -7.0% |
| Distance to coast (per mile) | – | -0.022 | -0.021 | -0.018 | [-0.02, -0.01] | – |

*Note.* Double ML uses random cross-fitting with gradient-boosted nuisances (Chernozhukov et al. 2018); a spatial block bootstrap supplies the 95% CIs. "Effect" gives the implied percentage change for a binary treatment. These estimates adjust for confounding yet remain observational — causal only under ignorability given the controls — and the coastal treatment rests on a modest $n$ treated.

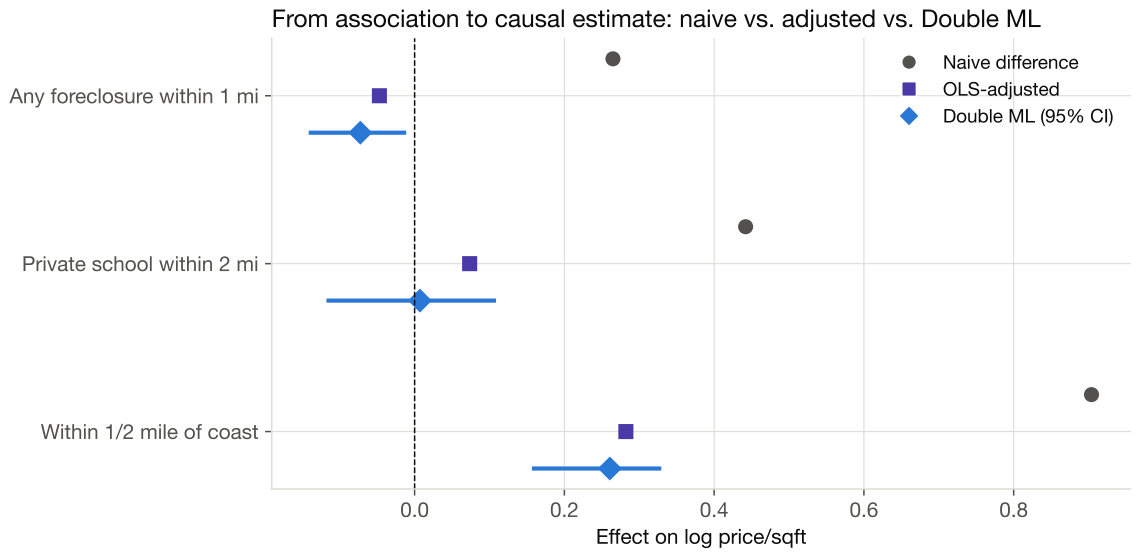


**Figure 9:** From association to causal estimate. For each treatment: the naive difference (grey), the linear-adjusted OLS coefficient (violet), and the cross-fitted Double ML estimate with its 95% confidence interval (blue). Adjustment moves the numbers substantially.

- **The private-school "premium" reduces to an artifact of location.** Naively, homes near a private school look far more expensive; yet private schools settle in already-expensive neighborhoods. Once the estimate adjusts for that geography, the effect shrinks to a statistical zero. Here stands exactly the confounding that a predictive model happily exploits and a causal question must refuse.
- **The foreclosure externality registers negative, at the edge of significance.** The naive sign runs positive (foreclosures cluster in dense, higher-priced areas), but adjustment flips it negative, consistent with Immergluck and Smith (2006); because the interval ends near zero, we read the estimate as suggestive.

**Which estimand? A control-set sensitivity.** A coefficient on "coast" depends on what the analysis holds fixed. With structural controls alone, we recover something close to the *total* coastal premium; adding amenities and then city fixed effects isolates the *direct* effect, net of the channels (income, restaurants, schools) through which the coast operates. Table 6 traces the path: the premium shrinks from roughly +79% (structural only) to +29.8% (full controls) yet persists throughout — coastal proximity carries a large price signal beyond the amenities that accompany it.

**Table 6:** Coastal-premium Double ML under progressively richer control sets.

| Control set | Double ML $\theta$ | 95% CI | Effect |
|---|---|---|---|
| Structural only | 0.585 | [0.22, 0.77] | +79.4% |
| + Neighborhood amenities | 0.291 | [0.20, 0.46] | +33.8% |
| + City fixed effects (full) | 0.261 | [0.16, 0.33] | +29.8% |

*Note.* Structural-only approximates the total premium; adding amenities and city fixed effects isolates the direct effect net of the channels coastal location works through.

We close by stressing the identifying assumption and its limits. These figures constitute

*confounding-adjusted observational estimates*, causal only under conditional ignorability — the premise that every confounder of treatment and price (ocean views, zoning, lot quality, school-district prestige) already appears among the controls. Double ML buys robustness to *functional-form* error in those controls (Chernozhukov et al., 2018), while omitted variables remain beyond its reach, and we advance neither an instrument nor a discontinuity design. Support also runs thin for the coastal treatment (a few hundred treated homes; see $n$ in Table 5), so the estimate leans on a modest slice of the county. We therefore read these numbers as the best adjustment the data allow, rather than settled causal parameters, and we return to them among the limitations.

# 7 Spatial structure

The spatial-CV collapse of Section 4.3 leaves a diagnostic fingerprint. Out-of-sample LASSO residuals remain significantly spatially clustered (global Moran's $I = 0.056$, $p = 0.001$; Figure 10, right): the model systematically over- or under-prices whole areas, and the map renders the "suburb versus urban divide" plain. The price surface itself (Figure 10, left) lays the spatial concentration of value bare; moreover, both maps draw on the actual census-tract geography rather than a proprietary basemap.

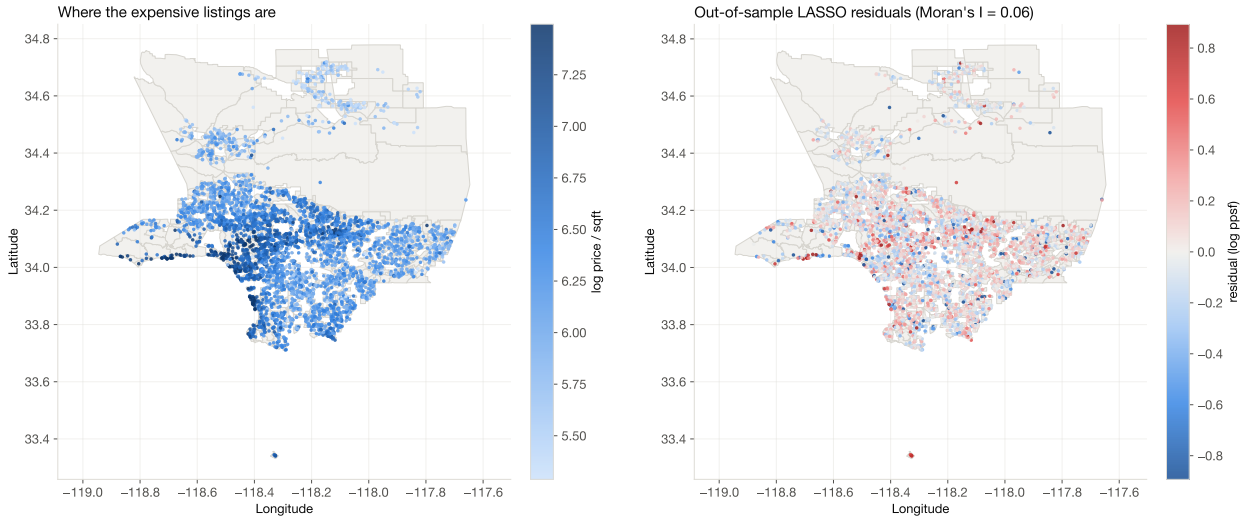


**Figure 10:** Left: log price per square foot across the county over census-tract polygons — value concentrates toward the coast. Right: out-of-sample LASSO residuals; the remaining clustering ($I = 0.056$) shows how much location the model leaves uncaptured.

That residual autocorrelation explains why the honest way forward lies in better spatial modelling rather than a fancier learner — an explicit spatial-error or spatial-lag specification (Anselin, 1988), or features encoding *relative* position (distance to the nearest comparable sale) in place of absolute coordinates. Before reaching for remedies, though, one more classical toolkit deserves its audit: the unsupervised analysis of Section 8.

# 8 Unsupervised structure and price tiers

## 8.1 Clustering

Left to choose its own number of clusters by silhouette, $k$-means prefers $k = 2$ — the data reject any finer segmentation into crisp market tiers. The dominant split separates the urban core from the suburbs (Table 7, Figure 11): one cluster sits denser and nearer the coast and commands a higher price *per square foot*; the other earns more, buys larger homes farther inland, and pays less per square foot for more space — a clean recovery of the suburb-versus-urban divide the maps of Section 7 display.

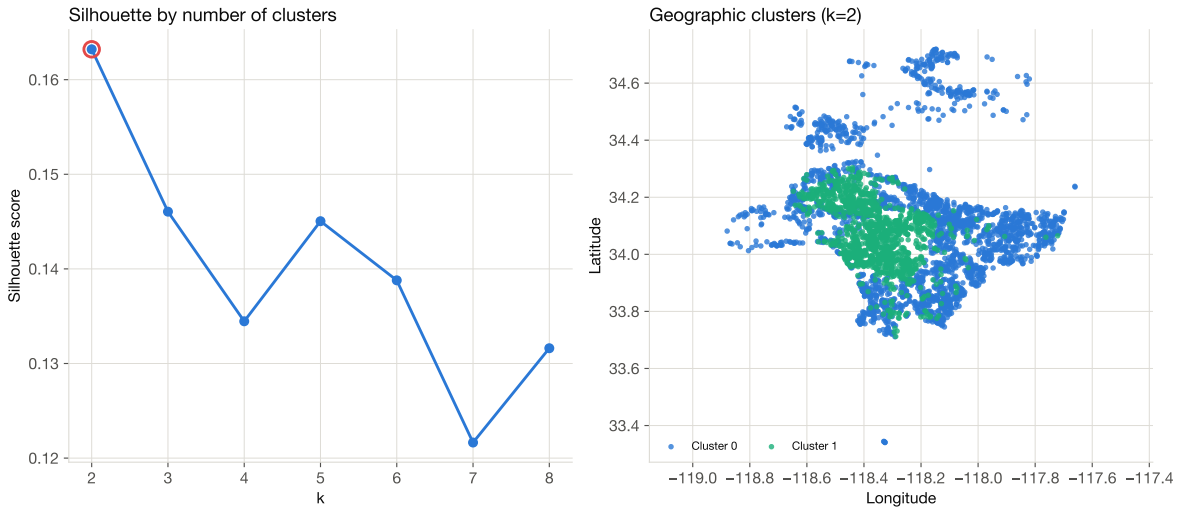


**Figure 11:** Silhouette by $k$ (left) selects $k = 2$; the resulting clusters (right) separate along an urban-core/suburban axis.

**Table 7:** k-means cluster profiles (feature means).

| Cluster | Median income | Viol. crime/person | Density/mi$^2$ | Schools | Coast (mi) | Living area | Year built | n | Mean log ppsf |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 104027.99 | 0.04 | 6443.69 | 122.76 | 18.44 | 2402.07 | 1961.97 | 2428.00 | 6.25 |
| 1 | 88210.99 | 0.02 | 11847.74 | 242.44 | 10.29 | 2251.69 | 1942.55 | 1376.00 | 6.51 |

## 8.2 Predicting the price tier

Framed as a four-way price-quartile classification, Gaussian Naive Bayes reaches 43.3% accuracy, comfortably above the 25% chance baseline yet modest. On identical folds, a random forest reaches 59.3% (Figure 12), and its errors land almost always in an *adjacent* quartile rather than a distant one: the model rarely mistakes a cheap home for an expensive one. These unsupervised and coarse-grained views echo the supervised ones — geography organizes the market — and the discussion now gathers the threads.

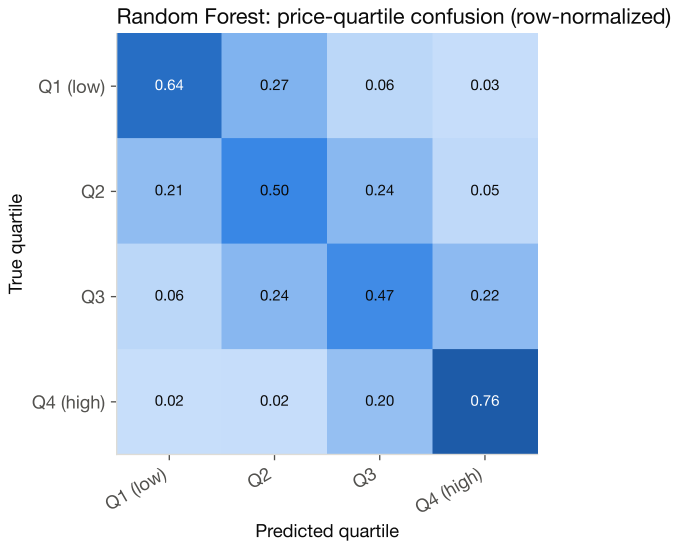


**Figure 12:** Row-normalized confusion matrix for random-forest price-quartile classification. Mass concentrates on the diagonal and the immediately adjacent cells.

# 9 Discussion

## 9.1 What we learned

Three findings stand out. First, the substantive one: the price of a Los Angeles home per square foot turns, above all, on *where* the home sits. Coastal proximity supplies both the strongest association and the most important predictor, and — crucially — it survives a causal adjustment that dissolves the apparent private-school premium. Neighborhood income, amenity density, and school access follow; once the analysis knows the location, the fine structural details of the house add surprisingly little.

Second, the methodological one, which travels well beyond this dataset: *on spatially dependent data, a single cross-validated $R^2$ can badly mislead.* Our best learner looked strong under random folds and progressively weaker as the held-out regions grew (Figure 4). Both ends of that spectrum deserve the label "correct," because they answer different questions. Random CV estimates how well we price a home whose comparable neighbors already inhabit the training data — realistic for an automated valuation model refreshed continuously. Leave-one-region-out estimates how well the learned relationships travel to a genuinely new area — the relevant question for policy (Kleinberg et al., 2015), or for any claim that the model has learned *what makes* homes valuable rather than *where* they sit. Reporting only the first overstates transferable knowledge.

Third, the value of separating prediction from causation (Mullainathan and Spiess, 2017). The most predictive features and the most causal ones differ: private-school proximity predicts well yet causes nothing, because it proxies for location; the coastal effect both predicts and (under our assumptions) causes. Double ML renders that distinction operational.

## 9.2 Limitations

We state the limits plainly.

- **Causal claims rest on ignorability.** Double ML removes functional-form bias in the controls; omitted-variable bias lies beyond its reach. Any confounder we failed to measure would tilt the

treatment effects.

- **Listing prices, rather than closed transactions.** The target reflects listing price per square foot from a single 2020–2021 vintage, short of verified sales.
- **Feature construction stays coarse.** Radius counts and tract aggregates serve as blunt proxies. Moreover, several engineered features carry known distortions that only the raw sources would let us recompute — metric buffers on unprojected coordinates, a noisy violent-crime label, centroid-based park counts. We repair the crime-imputation artifact (27% of rows) and project coordinates for our own spatial analysis, while the frozen feature columns retain those limitations.

### 9.3 What we would do next

The honest spatial result points toward better spatial modelling rather than a fancier learner: a spatial-error or spatial-lag specification (Anselin, 1988), geographically weighted regression, or causal forests (Wager and Athey, 2018) to map *where* the coastal and amenity effects run strongest. The bar stands where this analysis leaves it: a model whose out-of-sample skill survives the withholding of a whole neighborhood — a bar most off-the-shelf learners, on our evidence, fail to clear.

## 10 Reproducibility and data availability

Every figure, table, and inline number in this paper flows from the accompanying Python package and regenerates with a single command:

`make all` (runs the analysis and technique sweep, then typesets this paper)

The analysis runs entirely offline from one cleaned modeling table and operates free of API keys. Furthermore, the pipeline runs deterministically (a single fixed seed), dependencies stay pinned (`requirements-lock.txt`), and the code itself writes every number cited in the prose to `tables/macros.tex` and `tables/results_full.json`, so the manuscript stays locked to the results rather than drifting from them.

**Data availability.** We withhold all raw and home-level data from redistribution. `data/README.md` lists each source and its license; a collaborator holding the source files can rebuild the cleaned table with the ingestion code, supplying credentials through a local `.env`. `docs/METHODS.md` records the full methodology.

**Author contributions.** Pablo Zavala and Will Sigal designed the study (BUSN 41201, May 2024) and assembled the data; the committed Python package implements the analysis, the tuned technique sweep, the causal Double/Debiased ML, and the spatial cross-validation.

**Acknowledgment of tools.** The methods follow Taddy (2019) and the texts cited throughout; the figures follow the design guidance of Wilke (2019), Healy (2018), and Tufte (2001).

## References

Luc Anselin. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, 1988. ISBN 9789024737352.

Susan Athey and Guido W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019. doi: 10.1146/annurev-economics-080217-053433.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984. ISBN 9780412048418.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, 2016. ACM. doi: 10.1145/2939672.2939785.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. *Applied Causal Inference Powered by ML and AI*. 2024. doi: 10.48550/arXiv.2403.02467. https://causalml-book.org/.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

David Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. doi: 10.1016/0095-0696(78)90006-2.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton, FL, 2015. ISBN 9781498712163.

Kieran Healy. *Data Visualization: A Practical Introduction*. Princeton University Press, Princeton, NJ, 2018. ISBN 9780691181615.

Dan Immergluck and Geoff Smith. The external costs of foreclosure: The impact of single-family mortgage foreclosures on property values. *Housing Policy Debate*, 17(1):57–79, 2006. doi: 10.1080/10511482.2006.9521561.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, New York, 2nd edition, 2002. doi: 10.1007/b98835.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3149–3157, 2017.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–495, 2015.

Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, New York, 2013. doi: 10.1007/978-1-4614-6849-3.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950. doi: 10.2307/2332142.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40 (8):913–929, 2017. doi: 10.1111/ecog.02881.

Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974. doi: 10.1086/260169.

Matt Taddy. *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*. McGraw-Hill, New York, 2019. ISBN 9781260452778.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Waldo R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970. doi: 10.2307/143141.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2nd edition, 2001. ISBN 9780961392147.

U.S. Census Bureau. American community survey 5-year estimates (2016–2020). https://www.census.gov/programs-surveys/acs, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer, Cham, 2nd edition, 2016. doi: 10.1007/978-3-319-24277-4.

Claus O. Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, Sebastopol, CA, 2019. ISBN 9781492031086.

## A Selected hyperparameters

Table 8 records the configuration each learner selected under the inner randomized search of the nested sweep.

**Table 8:** Hyperparameters selected by inner cross-validation (full sample).

| Learner | CV RMSE | Selected hyperparameters |
|---|---|---|
| Elastic Net | 0.344 | alpha=0.0008, l1_ratio=0.2151 |
| KNN | 0.366 | n_neighbors=9, p=1, weights=uniform |
| SVR (RBF) | 0.337 | C=0.5415, epsilon=0.0057, gamma=0.0126 |
| Random Forest | 0.334 | max_depth=None, max_features=0.5, min_samples_leaf=2, n_estimators=700 |
| Extra Trees | 0.342 | max_features=0.5, min_samples_leaf=2, n_estimators=700 |
| Hist Gradient Boosting | 0.333 | l2_regularization=0.2481, learning_rate=0.017, max_iter=300, max_leaf_nodes=101 |
| XGBoost | 0.331 | colsample_bytree=0.7024, learning_rate=0.0115, max_depth=4, n_estimators=800, reg_lambda=0.7563, subsample=0.6807 |
| MLP | 0.378 | alpha=0.0857, hidden_layer_sizes=(64,), learning_rate_init=0.0001 |
| LightGBM | 0.331 | colsample_bytree=0.771, learning_rate=0.0109, n_estimators=400, num_leaves=65, reg_lambda=0.1156, subsample=0.8546 |

# B Data dictionary

Table 9 documents every modeling column — its type, missingness, cardinality, and an example value — generated directly from the analysis table.

**Table 9:** Data dictionary for the analysis table.

| Variable | Type | Missing | Unique | Example |
|---|---|---|---|---|
| X | int64 | 0 | 3804 | 18.0 |
| is_forAuction | int64 | 0 | 2 | 0.0 |
| event | string | 0 | 5 | Listed for sale |
| price | int64 | 0 | 1219 | 645000.0 |
| pricePerSquareFoot | int64 | 0 | 1123 | 548.0 |
| city | string | 0 | 152 | North Hollywood |
| yearBuilt | int64 | 0 | 128 | 1976.0 |
| streetAddress | str | 0 | 3804 | (redacted) |
| zipcode | int64 | 0 | 272 | 91605.0 |
| livingArea | int64 | 0 | 2188 | 1177.0 |
| bathrooms | int64 | 0 | 18 | 1.0 |
| bedrooms | int64 | 0 | 15 | 3.0 |
| parking | int64 | 0 | 2 | 1.0 |
| garageSpaces | int64 | 0 | 13 | 2.0 |
| hasGarage | int64 | 0 | 2 | 1.0 |
| levels | string | 0 | 14 | One |
| pool | int64 | 0 | 2 | 0.0 |
| spa | int64 | 0 | 2 | 0.0 |
| hasPetsAllowed | int64 | 0 | 2 | 0.0 |
| datePosted | str | 0 | 58 | 2021-07-13 |
| log_pricePerSquareFoot | float64 | 0 | 1123 | 6.306 |
| school_count | int64 | 0 | 455 | 183.0 |
| school_count_5miles | int64 | 0 | 455 | 183.0 |
| population | float64 | 1 | 1354 | 3686.0 |
| area_mi2 | float64 | 1 | 1565 | 0.651 |
| density_per_mi2 | float64 | 1 | 1565 | 5660.585 |
| violent_crime_count | float64 | 1020 | 407 | 3.0 |
| violent_crime_per_person | float64 | 1020 | 1147 | 0.001 |
| median_income | float64 | 1 | 1519 | 54773.0 |
| park_acres_within_2miles | float64 | 0 | 1007 | 171.562 |
| top_1000_restaurants_3miles | int64 | 0 | 303 | 0.0 |
| foreclosures_within_1mile | int64 | 0 | 65 | 18.0 |
| house_geometry | str | 0 | 3804 | (redacted) |
| census_geometry | str | 0 | 1566 | (redacted) |
| miles_from_coast | float64 | 0 | 3804 | 15.055 |
| half_a_mile_from_the_coast | int64 | 0 | 2 | 0.0 |
| urban_suburban | float64 | 1 | 2 | 1.0 |
| private_school_within_2miles | int64 | 0 | 2 | 1.0 |
| longitude | float64 | 0 | 3732 | (redacted) |
| latitude | float64 | 0 | 3741 | (redacted) |
| crime_missing | int64 | 0 | 2 | 0.0 |