MACHINE LEARNING CASE STUDY

Housing Mispricing Model

Feature Attribution for Lending and Collateral Valuation

Roger Estrada · Machine Learning in Business · AI Strategy · 2026

A Random Forest regression model explained 82% of the variation in California home values (R^2 0.821, test error of $48,720), and all three attribution methods ranked median income as the top driver. The model turns valuation weights from individual judgment into a ranked, defensible scorecard.

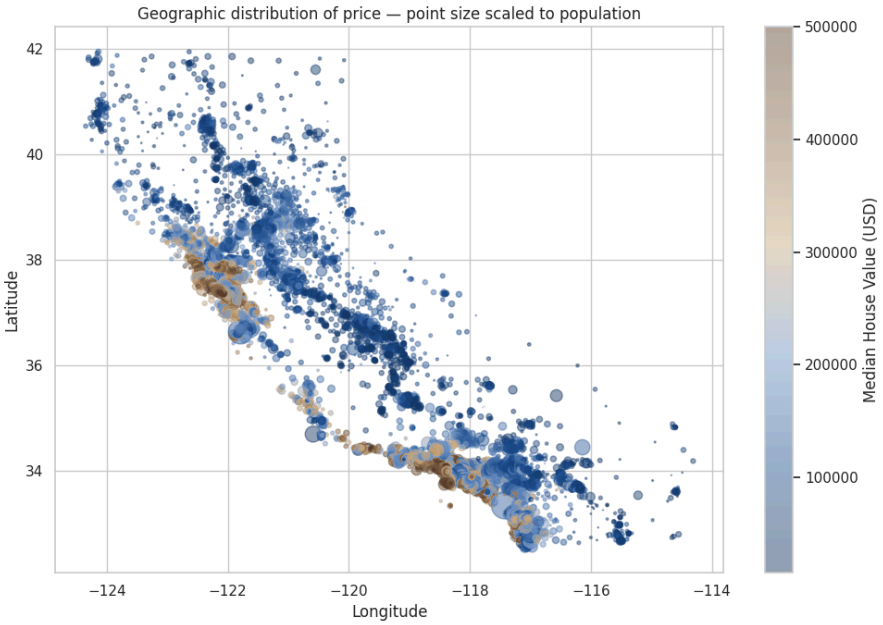
Business problem

Two analysts valuing the same property can differ by 10–20%, because each weights location, age, and income differently. Across hundreds of decisions a year, that inconsistency erodes portfolio performance. It also leaves no documented basis for the weights, which is a problem under fair-lending and fair-housing review.

The question this model answers: which housing attributes most strongly drive median home value, and how should they be weighted in an underwriting scorecard?

Approach and results

Data. 20,640 California census block groups from 1990 (a methodology proof, not a current market model). Each row is a neighborhood, not a single property. Price rises with income and follows the coast.

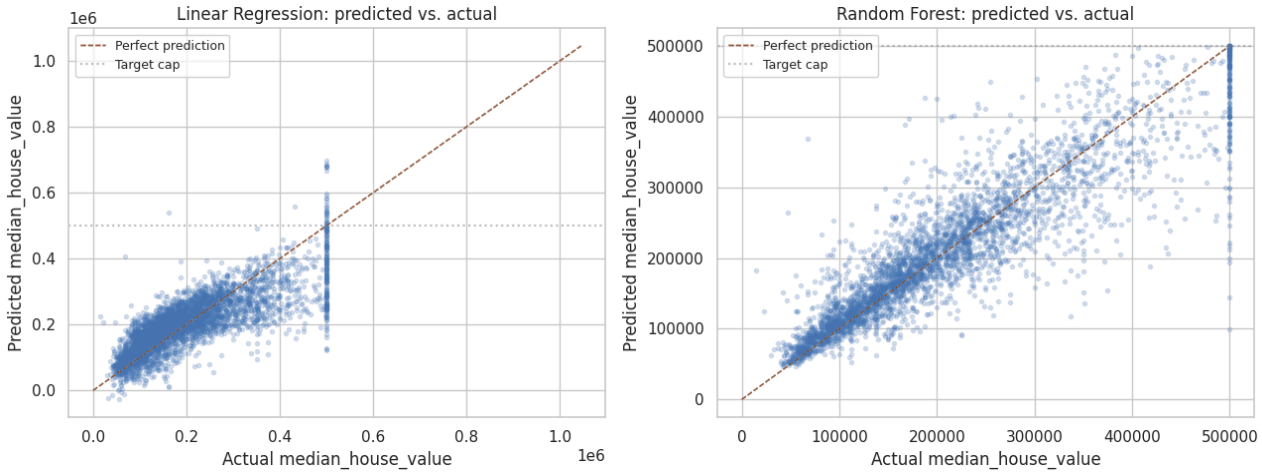


Preparation. Four raw size counts (total rooms, total bedrooms, population, households) correlate with each other above r = 0.85, so they were replaced with three ratios: rooms per household, bedrooms per room, and population per household. Missing bedroom counts (207 rows, 1%) were filled with the median of their ocean-proximity group. The 965 rows capped at $500,001 (4.7%) were flagged as a known limit. The data was split 80/20, stratified by income.

Models. Linear Regression gives signed, interpretable coefficients. Random Forest (200 trees) captures non-linear geography. Success criterion, set before training: R^2 of at least 0.75.

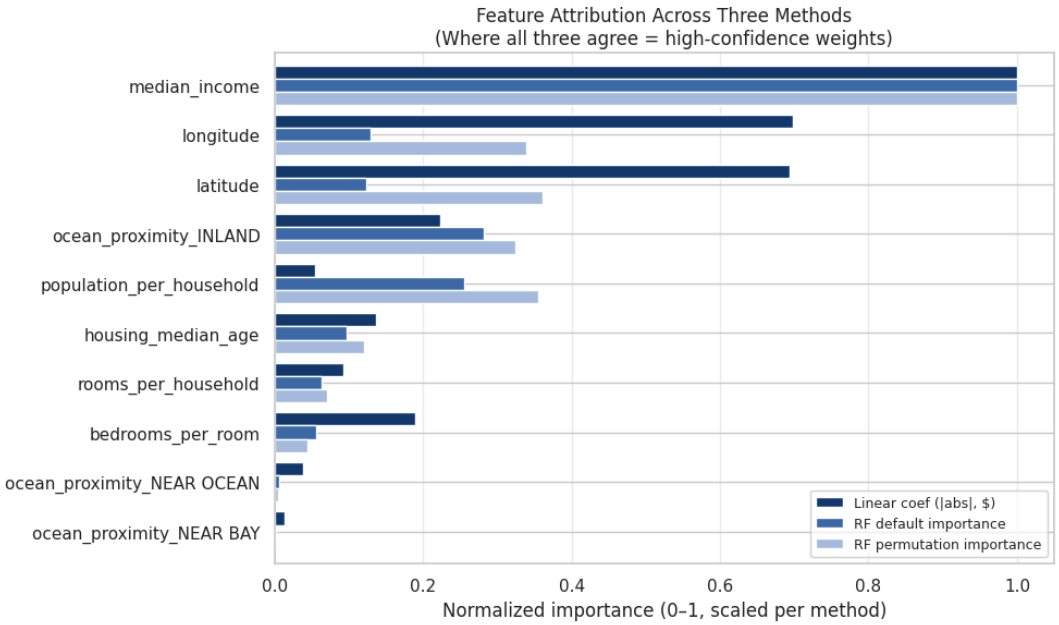
Model	Test RMSE	Test MAE	Test R^2
Linear Regression	$71,734	$51,397	0.611
Random Forest	$48,720	$31,602	0.821

Random Forest passed the 0.75 threshold. The $23,000 RMSE gap between the two models comes from non-linear coastal price structure that a straight line cannot express.



Attribution. Three methods were compared: linear coefficients, Random Forest default importance, and Random Forest permutation importance. Median income ranked first in all three. One standard deviation of income adds about $77,458 to predicted value in the linear model.

Feature	Linear rank	RF default rank	RF permutation rank	Average rank
Median income	1	1	1	1.00
Longitude	2	4	4	3.33
Latitude	3	5	2	3.33
Inland location	4	2	5	3.67
Population per household	8	3	3	4.67
Housing median age	6	6	6	6.00
Rooms per household	7	7	7	7.00



Where the methods disagree (population per household, bedrooms per room), both results are reported, because the gap separates direct effects from interaction effects.

Enterprise application

The same attribution approach supports collateral valuation, home-equity and mortgage underwriting consistency, and portfolio exposure review. It gives the underwriting team a weighted scorecard and gives acquisitions a way to screen inbound deals on the top few drivers.

The notebook quantifies the value for a California-focused fund making 50 acquisitions a year at a $400,000 average price. Heuristic underwriting carries about a 15% pricing error ($60,000 per deal). The Random Forest's average error is $31,602 per property, about 8%. The difference is $28,398 per deal, or about $1.42M a year in avoided overpayment, plus roughly 400 analyst hours saved by screening on the top three features. These figures are a modeled business case built on stated assumptions, not a measured production result.

Governance and limits

- **Vintage.** The 1990 data makes this a methodology proof. A production model needs current data and a refresh schedule.
- **Censored target.** The $500,001 cap hides premium-segment values and distorts importance rankings for those markets.
- **Granularity.** Predictions are neighborhood medians, not individual property prices. The weights transfer to property-level work; the predicted prices do not.
- **Fair-lending exposure.** Location features are legitimate for market segmentation but risky in borrower-facing pricing. Production use needs a documented use-case boundary and a fair-lending audit.

- **No time dimension.** Pair this model with a separate appreciation model before projecting returns.