Studying the Integration of Environmental Variables with Structural Housing Data: Impacts of Feature Engineering and Model Selection on Price Prediction

Chih-Yu Cheng | Chuchu Qi | Yuchen Huang | Yuzhe Cai (Leader)

1. Project Description

This project aims to study the impact of integrating environmental variables with property characteristics on housing prices prediction [1]in three California counties (Los Angeles, Orange and Ventura). We plan to train Linear Regression model, Random Forest [2], XGBoost, LightGBM, Hybrid Regression and Stacked Generalization Regression [3] to do predictions. Since Zillow doesn't provide the real housing price, our target is to predict the variable taxvaluedollarcnt (the sum of property valuation and tax) in the dataset. This is highly meaningful because it allows us to examine how a major company like Zillow approaches housing valuation, what factors it prioritizes, and how environmental variables influence predictive models. It also enables us to study how feature engineering affects model performance, and how the choice between single models and ensemble methods relates to prediction outcomes. It provides contributions like providing mature machine learning workflow, uncovering hidden patterns within the data and providing interdisciplinary study insights. Many related works are referenced in this work, including studies on how environmental pollution affects housing prices [4], and the use of hybrid or ensemble models for housing price prediction.

2. Dataset Description

(1) Zillow Prize: Zillow's Home Value Prediction [5] (Primary Dataset from Kaggle)

Scale: The dataset has 1.37GB (1048576 rows × 58 columns CSV).

Content: This dataset contains detailed property attributes for residential buildings in Los Angeles, Orange, and Ventura counties in 2016. The variables describe building structural characteristics, lot characteristics, and geographic attributes.

Why appropriate: This dataset offers large-scale and detailed housing data from Zillow.

(2) American Community Survey 2012-2016 5-Year Estimates [6]

Size: The dataset is 2.64 MB (8,035 rows × 58 columns CSV)

Content: This dataset contains various environmental variables summarizing the economic and environmental conditions of communities within the study area. This dataset can be integrated with housing data using postal codes, enabling multifaceted analysis.

Why appropriate: This dataset provides environmental data based on the U.S. Community Survey, making it suitable for analyzing spatial environmental risk factors.

(3) California Coastal Commission – Coastal Zone Boundary (CZB) dataset [7] and U.S. Census TIGER & Line 2010 ZIP Code Tabulation Areas (ZCTA5) [8]

Size: The dataset has 196 KB (1,267 rows × 12 columns CSV)

Content: These datasets provide geographical information for postal code areas. These datasets can be directly combined with housing data.

Why appropriate: Distance to the sea influences housing value and reflects accessibility to coastal amenities, ocean views, and exposure to coastal environmental risks.

3. Project workflow

3.1 Data Cleaning and Preprocessing

Before analysis, the Zillow dataset needs comprehensive cleaning and preprocessing. First, variables with more than 50% missing values were removed. Second, geographical coordinates were divided by 1,000,000 to convert them into correct values. Correcting zip codes was a significant challenge during the data preparation phase. All ZIP codes began with 96 or 97 (not valid ZIP codes for this study), as they were anonymized to protect user privacy. To address this, ZIP codes were rederived from geographic coordinates using the 2023 U.S. Census TIGER/Line ZCTA shapefile. Parcel points were processed in batches of rows. Each batch underwent a spatial joint to assign the correct ZIP code, and results were written to a csv file. After geocoding, environmental data were merged with the housing dataset based on ZIP codes.

To include coastline distance, the "Near" tool in ArcGIS was used to calculate the distance from each zip code centroid to the California coastline, providing distance data to the coastline for each zip code area.

Data processing also included handling missing values. Missing values in key spatial variables were interpolated using the Empirical Bayesian Kriging method to ensure completeness for subsequent analysis. For factors with weaker spatial correlation and fewer missing values, median imputation was applied.

3.2 Exploratory Data Analysis and Outlier Filtering

In the exploratory data analysis, the distributions of variables and their minimum and maximum values were visualized. This revealed three features (taxvaluedollarcnt, finishedsquarefeet12, and lotsizesquarefeet) requiring outlier filtering, all of which contained extreme values. Outliers include properties assessed at only a few dozen dollars (while similar neighboring homes are valued at several hundred thousand dollars), as well as land sizes reaching several million square feet. To address this, a log transformation was applied, which made the distributions closer to a normal-like shape, followed by the removal of values beyond n standard deviations from the mean. Model performance was evaluated under different filtering strategies, including no filtering, 6σ, 5σ, 4σ, and 3σ. The results show that the 4σ filtering yields the best overall model performance.

3.3 Feature Engineering and Complete Feature Lists

Then feature engineering was conducted. New property-level variables were created, including house age, area per room, and the bathroom/bedroom ratio. Spatial features were then introduced by computing ZIP level median value, median year and median area. Also, population density and coastal proximity are calculated. Property types were encoded using one-hot encoding to replace the original categorical field. The final dataset contains 31 features, compared with 22 in the original version, and model performance will be evaluated on both datasets for comparison in model evaluation section.

The original dataset contains 21 predictors. The predictors include property characteristics (finishedsquarefeet12, bedroomcnt, bathroomcnt, buildingqualitytypeid, lotsizesquarefeet, yearbuilt, propertylandusetypeid, ratio), spatial location (latitude, longitude) and environmental factors (PM2_5, Drinking_Water, Pesticides, Traffic, Groundwater_Threats, Imp__Water_Bodies, Solid_Waste, Pollution_Burden, POPULATION, SEA_DIST, SQMI). In the enhanced dataset, 10 new features are added, including house_age (replacing yearbuilt), area_per_bed, bath_per_bed, area_per_room, area_x_quality, coastal_flag, pop_density, and three ZIP-level group statistics (zip_median_value, zip_median_area, zip_median_year).

3.4 Exploratory Data Analysis

After constructing the spatial weight matrix using KNN ($k = 12$) and calculating Moran's I, the results are as follows: Moran's I ≈ 0.48 and p-value < 0.01. Moran's I is significantly positive, indicating that house prices show a distinct spatial positive autocorrelation: the area around high-priced regions remain high-priced, and the area around low-priced regions remains low-priced, not randomly distributed.

To avoid excessive computational load, 30,000 samples were randomly selected from the data for LISA analysis. The plots were plotted through color zoning.

The LISA classification results contain four types of aggregation: High - High: High-priced houses surrounded by high-priced neighbors. Low - Low: Low-priced houses surrounded by low-priced neighbors. High - Low: High-priced houses surrounded by low-priced neighbors. Low - High: Low-priced houses surrounded by high-priced neighbors. High-High hot spots are mainly concentrated in the city's core areas along major transportation and business resource-intensive regions.

Gi* statistics are used to identify hotspots, with high Z-values representing high-value aggregations and low Z-values representing low-value aggregations. The significant Hotspot group (high Z) covers the core urban area and areas with a concentration of high-value residential density. The cold spot group (low Z) is mainly distributed in the outer suburbs and areas with large land supply but weak housing prices.

3.5 Machine Learning Modeling

After the preceding steps, the feature-engineered dataset (31 columns and 1.7 million records) was used to train the Random Forest, XGBoost, and LightGBM models (80/20 train/test split using 5-fold cross validation). A controlled-variable strategy in parameter tuning was applied: one parameter was tuned while others were held constant. During tuning, model performance typically exhibited either increasing and then declining or rising until stabilizing. The parameter values were selected based on these performance peaks.

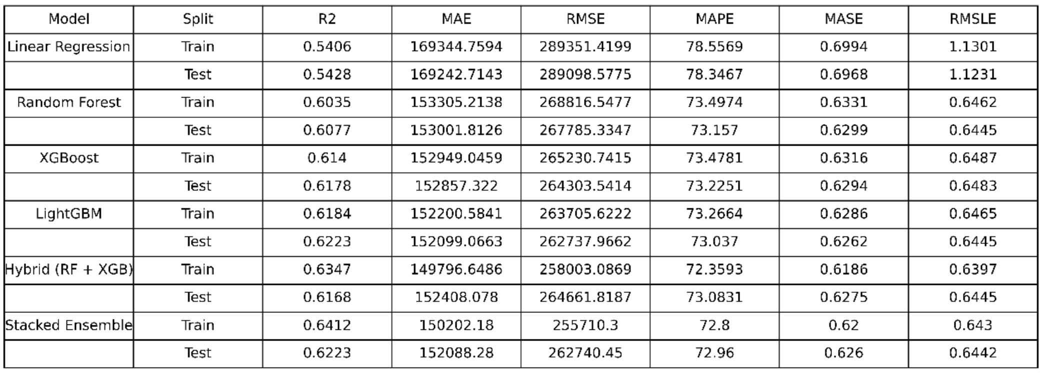
3.6 Parameter Tuning and Model Performance Results

For Random Forest, three key parameters were tuned. Among max_depth values from 15 to 30 step 5, max_depth = 25 produced the best performance. Among n_estimators from 100 to 400 step 100, n_estimators = 200 performs the best. Among min_samples_leaf from 50 to 250 step 50, min_samples_leaf = 50 performs the best. Under these parameters, Random Forest achieved an R^2 of 0.6035 in cross-validation and 0.6077 on the test set.

For XGBoost, three key parameters were tuned. Among max_depth values from 3 to 9 step 1, max_depth = 7 produced the best performance. Across min_child_weight values from 50 to 300 step 50, the strongest results appeared at min_child_weight = 250. For the number of boosting rounds, tested between 100 and 600 step 100, n_estimators = 300 yielded the highest accuracy. With these tuned settings, XGBoost achieved an R^2 of 0.6140 in cross-validation and 0.6178 on the test set.

For LightGBM, tuning focused on model complexity and sampling. Across num_leaves values from 20 to 80 step 10, num_leaves = 60 performed best. For min_child_samples, tested between 50 and 350 step 50, the optimal value was 350. Among subsample and colsample_bytree values from 0.5 to 1.0 step 0.1, 1.0 provided the strongest results, while for colsample_bytree, 0.6 outperformed the others. Under these parameters, LightGBM reached an R^2 of 0.6184 in cross-validation and 0.6223 on the test set.

After obtaining all tuned parameters for three models, Linear Regression, the Hybrid Ensemble, and the Stacked Ensemble are trained. The Hybrid Ensemble simply averages the predictions of RF and XGBoost, while the Stacked Ensemble combines RF, XGBoost, and LightGBM using a Linear Regression meta-learner. Linear Regression produced an R^2 of 0.5406 in cross-validation and 0.5428 on the test set. The Hybrid achieved stronger predictive power, reaching a train R^2 of 0.6347 and a test R^2 of 0.6168. The Stacked Ensemble has a train R^2 of 0.6412 and a test R^2 of 0.6223.



Model	Split	R2	MAE	RMSE	MAPE	MASE	RMSLE
Linear Regression	Train	0.5406	169344.7594	289351.4199	78.5569	0.6994	1.1301
	Test	0.5428	169242.7143	289098.5775	78.3467	0.6968	1.1231
Random Forest	Train	0.6035	153305.2138	268816.5477	73.4974	0.6331	0.6462
	Test	0.6077	153001.8126	267785.3347	73.157	0.6299	0.6445
XGBoost	Train	0.614	152949.0459	265230.7415	73.4781	0.6316	0.6487
	Test	0.6178	152857.322	264303.5414	73.2251	0.6294	0.6483
LightGBM	Train	0.6184	152200.5841	263705.6222	73.2664	0.6286	0.6465
	Test	0.6223	152099.0663	262737.9662	73.037	0.6262	0.6445
Hybrid (RF + XGB)	Train	0.6347	149796.6486	258003.0869	72.3593	0.6186	0.6397
	Test	0.6168	152408.078	264661.8187	73.0831	0.6275	0.6445
Stacked Ensemble	Train	0.6412	150202.18	255710.3	72.8	0.62	0.643
	Test	0.6223	152088.28	262740.45	72.96	0.626	0.6442

Figure 1: Model Performance (feature engineered)

After that, all the models were trained by the dataset before feature engineering and there are some interesting findings. After feature engineering, Linear Regression improved the most (+0.04 R²), showing that new features mainly help simpler models capture nonlinear patterns. Random Forest and LightGBM saw only slight gains (<0.002), while XGBoost slightly decreased (–0.001). Stacking matched the same performance showing feature engineering may have little help to advanced models.

3.7 Feature Importance

Feature Importance is based on feature engineered dataset. For Random Forest and XGBoost, the top five predictors already account for well over half of the total feature importance. In both models, finishedsquarefeet12 is the dominant driver, and area_x_quality together with buildingqualitytypeid consistently appear in the top five, highlighting the central role of house size and construction quality. Environmental variables in RF don't contribute so much, only having 2.7% importance in total, but environmental variables in XGBoost have 11.3% importance in total.

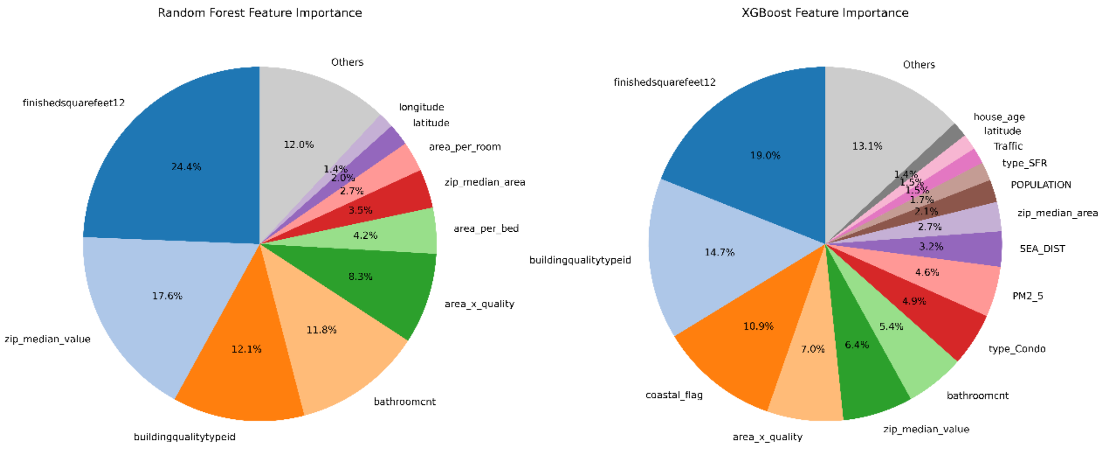


Figure 2: RF and XGBoost Feature Importance

LightGBM, by contrast, shows a different pattern: latitude and longitude emerge as the single most important factors, and overall importance is distributed more evenly across a broader set of features. Environmental indicators have 9.2% importance in total. Interestingly, PM2.5, water and traffic are frequently appearing in these three models' importance table. That shows the main concerns of people.

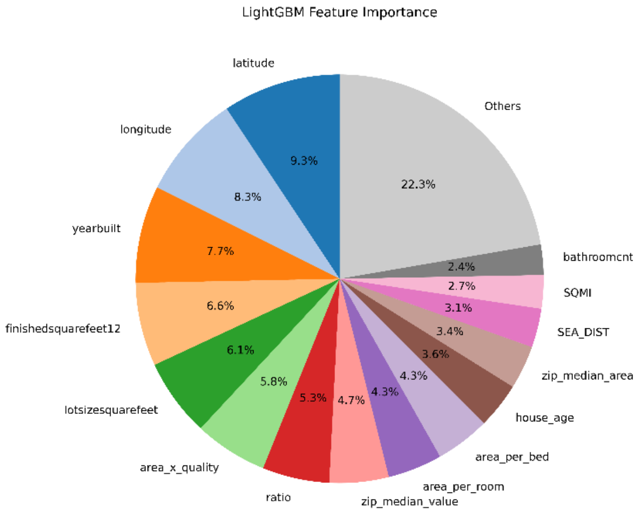


Figure 3: LightGBM Feature Importance

3.8 Conclusion

Feature engineering substantially improved Linear Regression, while tree-based models showed only minor changes, indicating that advanced algorithms already capture most nonlinear relationships and therefore gain little from additional features. Ensemble methods performed similarly to the strongest single model, offering no further improvement.

In feature importance, Random Forest and XGBoost exhibit highly concentrated contributions dominated by house size, construction quality, and related attributes, whereas LightGBM distributes importance more evenly and places greater emphasis on spatial factors. Environmental variables, though not primary drivers, consistently appear across all three models (PM2.5, water, and traffic) suggesting they offer meaningful supplementary information.

Overall, structural and spatial characteristics form the core determinants, while environmental indicators serve as complementary signals.

Team Assignments

Chih-Yu Cheng: Collect data, data preprocessing and help train models.

Chuchu Qi: Solve ZIP issue, complete EDA and ESDA and help train models.

Yuchen Huang: Collect data, data preprocessing and help train models.

Yuzhe Cai: Feature engineering; Model training, tuning, visualizing and evaluating.

References

[1] Jim, C. Y., and Wendy Y. Chen. “Impacts of Urban Environmental Elements on Residential Housing Prices in Guangzhou (China).” Landscape and Urban Planning 78, no. 4 (2006): 422–434. https://doi.org/10.1016/j.landurbplan.2005.12.003.

[2] Kim, Jeonghyeon, Youngho Lee, Myeong-Hun Lee, and Seong-Yun Hong. 2022. "A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices" Sustainability 14, no. 15: 9056. https://doi.org/10.3390/su14159056

[3] Truong, Quang, Minh Nguyen, Hy Dang, and Bo Mei. “Housing Price Prediction via Improved Machine Learning Techniques.” Procedia Computer Science 174 (2020): 433–442. https://doi.org/10.1016/j.procs.2020.06.111

[4] Li, Xiaodong, and Yanfang Liu. 2018. “Impacts of Urban Environmental Elements on Residential Housing Prices in Guangzhou (China).” Sustainability 10 (7): 2533. https://doi.org/10.3390/su10072533

[5] Zillow. 2017. Zillow Prize: Zillow’s Home Value Prediction (Zestimate) properties_2016.csv [Data set]. Kaggle. May 24, 2017. https://www.kaggle.com/competitions/zillow-prize-1/overview

[6] ProximityOne. 2018. 2018population.csv (American Community Survey 2016 — 2012–16 5-Year Estimates) [Data set]. January 18, 2018. https://proximityone.com/acs1216.htm

[7] California Coastal Commission. 2025. Coastal Zone Boundary (CZB) [Data set]. California Coastal Commission Open Data Portal. https://california-coastal-commission-open-data-1-3-coastalcomm.hub.arcgis.com/datasets/coastalcomm::coastal-zone-boundary/about

[8] U.S. Census Bureau. 2010. TIGER/Line Shapefile, 2010, 5-Digit ZIP Code Tabulation Area (ZCTA5) [Data set]. Data.gov. https://catalog.data.gov/dataset/tiger-line-shapefile-2010-5-digit-zip-code-tabulation-area-zcta5-nation