Model Behavior

February 2026

1 Project Overview

This project aims to predict California house prices using machine learning models with increasing levels of complexity. The target variable is LogClosePrice, which reduces skewness and improves model stability.

We compare three models Linear Regression (Baseline), Decision Tree and Random Forest Regressor.

Model performance is evaluated using R^2 (log scale and original dollar scale) to measure explained variance. MAPE to measure average percentage prediction error. MdAPE to measure typical (median) percentage error, which is more robust to outliers.

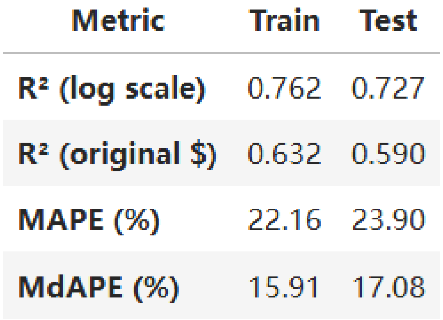
2 Model Behavior Analysis

2.1 Linear Regression Model

The baseline model is a Linear Regression using only PostalCode as the predictor, which is the feature with the highest single-feature R^2. The model predicts the logarithm of the closing price ($LogClosePrice$) to stabilize variance and better satisfy the assumptions of linear regression. The table is the univariate R^2 values for predicting the LogClosePrice.

Baseline Model Performance

The following table summarizes performance metrics on both training and test sets.



Metric	Train	Test
R^2 (log scale)	0.762	0.727
R^2 (original $)	0.632	0.590
MAPE (%)	22.16	23.90
MdAPE (%)	15.91	17.08

Figure 1: Linear Regression Model Evaluation

Training R^2 (log scale): 0.762. The model explains 76% of variance in the training data.
Test R^2 (log scale): 0.727. On unseen data, the model still explains 73% of variance, indi-

Table 1: Univariate R^2 Values for Predicting LogClosePrice

Feature	R^2
PostalCode	0.7621
City	0.6917
MLSAreaMajor	0.4690
LivingArea	0.3260
HighSchoolDistrict	0.3253
BathroomsTotalInteger	0.2849
BedroomsTotal	0.1470
Levels	0.0757
FireplaceYN	0.0672
Stories	0.0517
PoolPrivateYN	0.0296
Longitude	0.0272
AssociationFee	0.0179
CloseDate	0.0043
YearBuilt	0.0036
GarageSpaces	0.0031
NewConstructionYN	0.0029
Latitude	0.0016
ViewYN	0.0011
ParkingTotal	0.0008
LotSizeArea	0.0007
AttachedGarageYN	0.0001
LotSizeAcres	0.0000
LotSizeSquareFeet	0.0000

cating good generalization.
The lower adjusted (original scale) R^2 values suggest that using PostalCode alone limits predictive power.
MAPE (Mean Absolute Percentage Error): Measures the average percentage error of predictions. Training 22% and Test 24% indicate the model is sensitive to outliers; higher values indicate larger errors on expensive houses.
MdAPE (Median Absolute Percentage Error): Median percentage error, more robust to extreme prices. Error metrics remain consistent between training and test sets, making this model a stable but simple baseline for comparison with more advanced models.

Strengths and Weaknesses of Linear Regression

Strengths

- Interpretable coefficients
- Stable
- Fast to train
- Easy to explain

Weaknesses

- Cannot capture nonlinear relationships
- Requires manual interaction terms
- Sensitive to multicollinearity

2.2 Decision Tree Model

2.2.1 Decision Tree Predictor (Shallow: maxdepth=5)

We select predictors based on: Statistical signal strength (single-variable (R^2)), Real estate domain knowledge, Overfitting risk, Model interpretability. The goal is to explain variation in: LogClosePrice.

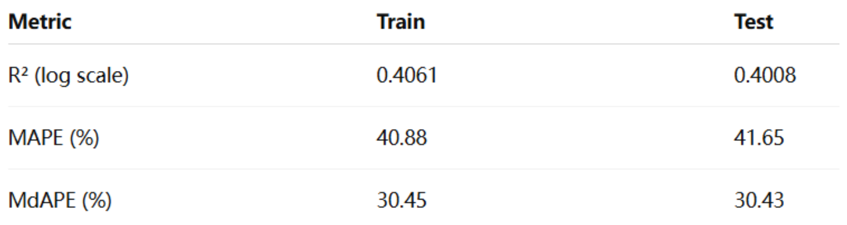
Feature Engineering Predictors We keep strong location features: PostalCode, MLSAreaMajor, HighSchoolDistrict; and strong structural features: LivingArea BathroomsTotalInteger BedroomsTotal. We exclude weak features with

$$R^2 < 0.05$$

to reduce noise and overfitting.

Model Performance

As shown in the following table:



Metric	Train	Test
R² (log scale)	0.4061	0.4008
MAPE (%)	40.88	41.65
MdAPE (%)	30.45	30.43

Figure 2: Decision Tree Model Performance

R^2 measures the proportion of variance explained by the model (1 = perfect, 0 = none). Observed values: Train R² (log): 0.4061, Test R² (log): 0.4008 indicate the model is underfitting.

MAPE and MdAPE Train MAPE: 40.88%, MdAPE: 30.45%; Test MAPE: 41.65%, MdAPE: 30.43% Shows that the Decision Tree performs worse than the baseline Linear Regression.

2.2.2 Decision Tree Predictor (Improved)

ZIP-Level Location Encoding One-hot encoding ZIP codes introduces thousands of sparse features and performs poorly in tree-based models. Instead, we encode location using the median log sale price per ZIP code calculated from the training data only. This provides a strong numeric signal for neighborhood price levels while avoiding high dimensionality. ZipMedianPrice acts as a learned neighborhood price prior, allowing the tree to focus on within-area structural differences.

Feature Predictors We select features that: Capture location and property size, Work well in tree-based models, Reduce overfitting risk. The selected features balance interpretability and predictive power. So we select ZipMedianPrice, LivingArea, BathroomsTotalInteger, BedroomsTotal.

Improved Decision Tree Model We apply regularization to improve generalization. These constraints reduce overfitting while preserving nonlinear patterns.

Model Performance We evaluated the improved decision tree model and get the following metrics:

Metric	**Train**	**Test**
R^2 (log scale)	0.8702	0.8473
MAPE (%)	16.42	17.77
MdAPE (%)	11.90	12.52

Comparison of the two Decision Tree models

Train and test R^2 of the shallow DT model are both low and very similar. The train and test errors are high. This indicates the model is underfitting. The model is too simple (tree too shallow) and cannot capture the underlying structure of the data.

The improved decision tree significantly improves predictive accuracy while maintaining good generalization. It achieves much higher R^2 values and the Errors (MAPE, MdAPE) are significantly lower.

Strengths and Weaknesses of Decision Tree Regression

Strengths

- **Capture Nonlinear patterns:** Housing prices often depend on threshold effects (e.g., $LivingArea > 2500$ sqft) and complex interactions (e.g., the effect of size varies by neighborhood).
- **Handles feature interactions automatically:** Unlike OLS, interaction terms do not need to be manually specified, as the tree structure naturally captures interactions through hierarchical splits.
- **Interpretability:** Decision trees can be visualized and interpreted through clear decision rules, making them easier to explain to stakeholders.

Weaknesses

- The shallow Tree model is underfitting.
- The shallow Tree cannot outperform a ZIP-level linear model

2.3 Random Forest Regressor

Feature Engineering Predictors All variables except: ClosePrice LogClosePrice. Encoding Categorical variables prevents multicollinearity.

Target Transformation Dependent Variable: LogClosePrice Natural log applied to ClosePrice Reduces right skew Improves model stability Helps satisfy regression assumptions.

Random Forest Model Training Model RandomForestRegressor Parameters: 500 trees $\sqrt{(}$features) sampled at each split Fixed random seed Parallel processing enabled Model learns non-linear relationships and feature interactions.

Metric	RF Train	RF Test
R^2 (log scale)	0.9821	0.8659
R^2 (original $)	0.9756	0.8121
MAPE (%)	5.72	15.98
MdAPE (%)	3.83	10.43

Model Performance

R^2 **(Log Scale):** Measures how well the model explains the variance in the log-transformed target variable.

- **Train:** Approximately 98% of the variance is explained.
- **Test:** Approximately 86% of the variance is explained, indicating strong generalization performance.

R^2 **(Original Dollar Scale):** Measures the proportion of variance explained in actual house prices.

- **Train:** Approximately 97.5% of the variance is explained.
- **Test:** Approximately 81% of the variance is explained.

MAPE (Mean Absolute Percentage Error): Measures the average percentage deviation between predicted and actual values.

- **Train:** About 5.7% average deviation.
- **Test:** About 16% average deviation.

MAPE is sensitive to outliers; higher values often indicate larger prediction errors for expensive homes.

MdAPE (Median Absolute Percentage Error): Measures the median percentage deviation and is more robust to extreme values.

- **Train:** Approximately 3.8% typical deviation.
- **Test:** Approximately 10% typical deviation.

MdAPE provides a clearer representation of the typical and realistic prediction error.

The model demonstrates strong learning capability on the training data and generalizes well to the test set, achieving a test R^2 (log scale) of approximately 0.87. However, the gap between training and testing performance (approximately 0.11) suggests mild overfitting.

Overall, the Random Forest model substantially outperforms the baseline linear regression model in terms of explanatory power and predictive accuracy.

Strengths and Weaknesses of Random Forest Regressor

Strengths

- High Predictive Accuracy

- Captures Nonlinear Relationships
- Robust to Multicollinearity
- Stable with Outliers
- Interpretability

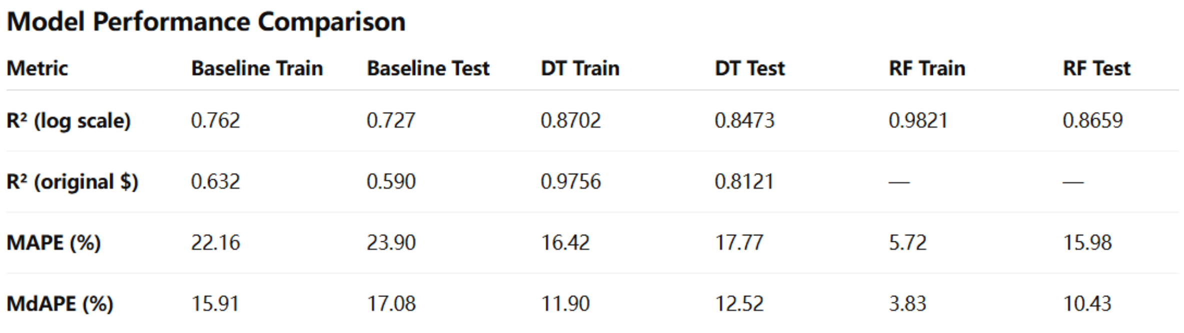
Weaknesses

- Mild Overfitting
- Cannot Extrapolate Beyond Training Range
- Computational Cost

3 Comparison of Models

Table 2: Model Comparison

Model	**Features**	**Target Variable**	**Key Characteristics**
Linear Regression (Baseline)	PostalCode (one-hot)	LogClosePrice	Simple linear model; captures only ZIP-level average effect. Log transformation stabilizes variance.
Decision Tree (Improved)	ZipMedianPrice, LivingArea, BathroomsTotalInteger, BedroomsTotal	LogClosePrice	Tree-based model capturing nonlinear interactions; uses ZIP-level median price encoding; regularized with `max_depth=6` and `min_samples_leaf=100`.
Random Forest Regressor	All features (one-hot encoded)	LogClosePrice	Learns non-linear relationships and feature interactions; `max_features='sqrt'`, `n_estimators=500`)



Model Performance Comparison

Metric	Baseline Train	Baseline Test	DT Train	DT Test	RF Train	RF Test
R² (log scale)	0.762	0.727	0.8702	0.8473	0.9821	0.8659
R² (original $)	0.632	0.590	0.9756	0.8121	—	—
MAPE (%)	22.16	23.90	16.42	17.77	5.72	15.98
MdAPE (%)	15.91	17.08	11.90	12.52	3.83	10.43

Figure 3: Models Performance Comparison

Decision Tree VS. Linear Regression

The improved Decision Tree model captures nonlinear relationships between features and log-prices, leading to higher explained variance; Explains 85% of variance in unseen data (Test R^2); Reduces median prediction error to 12.5% vs 17% for Linear Regression; Balances accuracy, interpretability, and robustness. Clearly outperforms the baseline Linear Regression on all key metrics.

Linear Regression: Easy to interpret, but only captures linear effects of ZIP codes.

Decision Tree: Extracted rules show how location and property size interact to determine prices.

Random Forest VS. Linear Regression

The Random Forest model generalizes learned from the training data and generalized pretty well on the test set, giving an R^2 value of 0.87, but does show signs of slightly overfitting to the training as shown by the gap of 0.11. The Random Forest classifier does perform significantly better than the baseline linear regression models.

Overall Ranking (Test Performance)

Random Forest > Decision Tree > Linear Regression

Linear Regression underfits, using only one feature.

Decision Tree achieves strong performance while remaining interpretable and it balances bias and variance well.

Random Forest delivers the highest accuracy but with increased model complexity and slight overfitting.