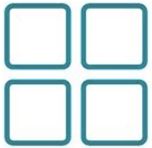
How effective are non-Linear Regression Models at Predicting Housing Prices?

Ziyad Amer, Chelle Davies

The Importance of Housing Price Predictions

Economic Indicator

Housing prices signal economic health and wealth



Complex Influences

Driven by property attributes, location, and economy

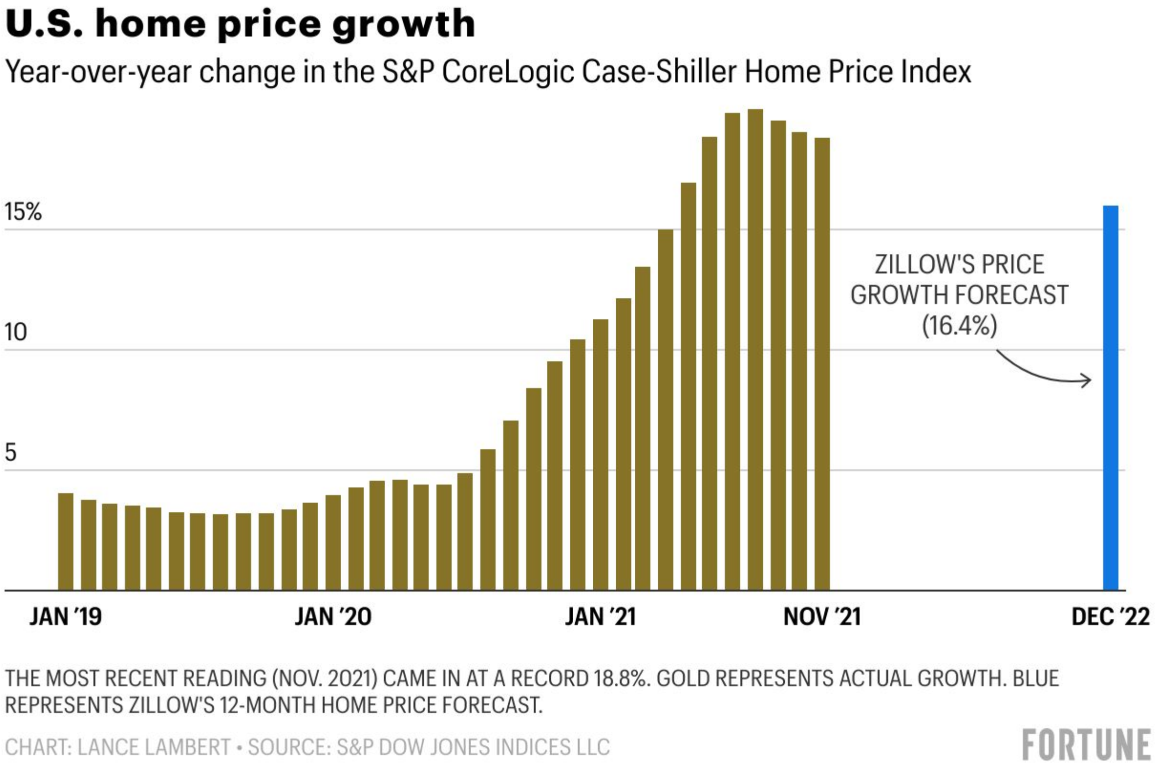
Impactful Decisions

Guides policymakers, investors, and buyers

Affordability

Rising Prices highlight equity and planning challenges

Issue with Existing Models



Approach and Results

Non-Linear Regression Models

Utilizes non-linear methods to predict housing trends, moving beyond traditional linear assumptions

Balances Historical Weighting

Ensures historical data does not dominate, providing a more adaptive forecasting model

Demographic/Block Features

Analyzes demographic and neighborhood trends at the granular block level for deeper insights

MAE: $26,856
R-squared: 0.81

Data Source: U.S. Census Housing Data, California

- **Target:** Median house value
- **Socioeconomic:** Median income, median housing age, block group demographics
- **Structural:** Number of rooms, number of bedrooms
- **Geographic:** Longitude and latitude, proximity to ocean

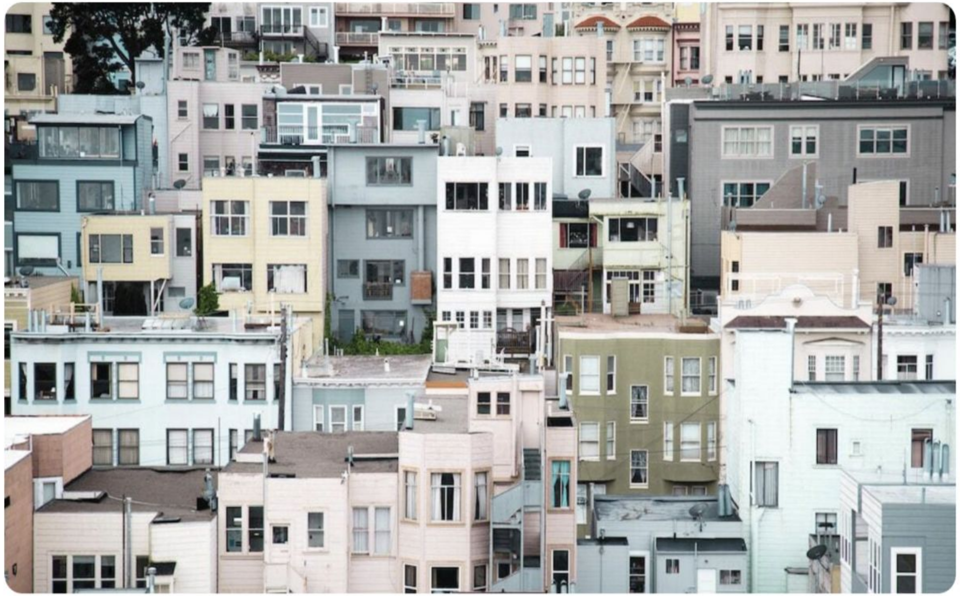
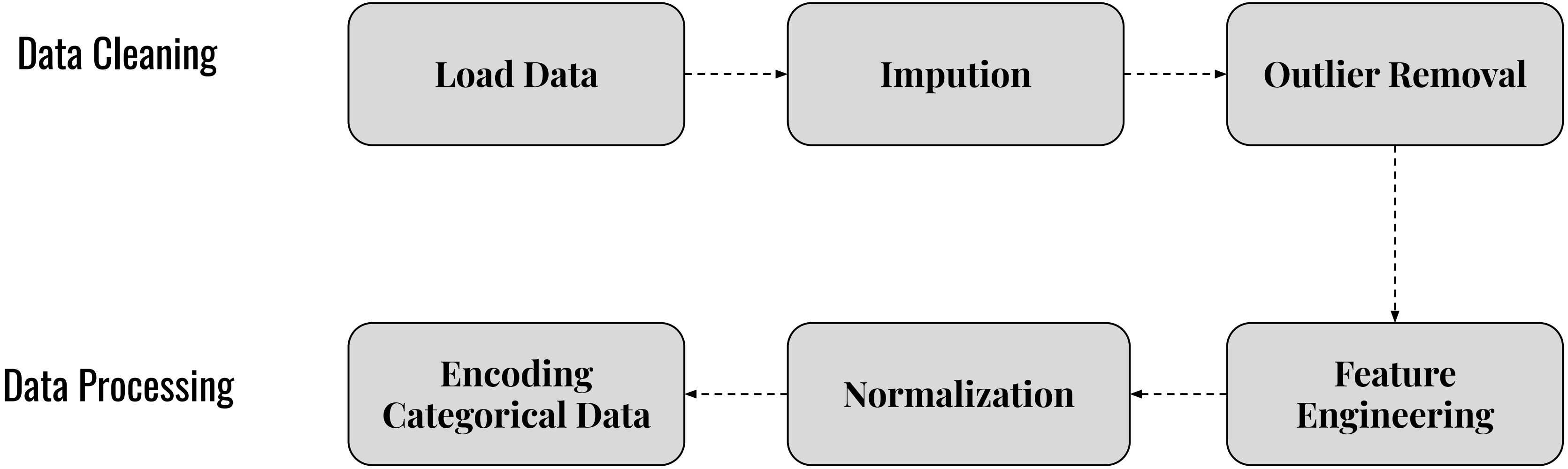


Photo by Sam Ellis on Unsplash

Data Preparation

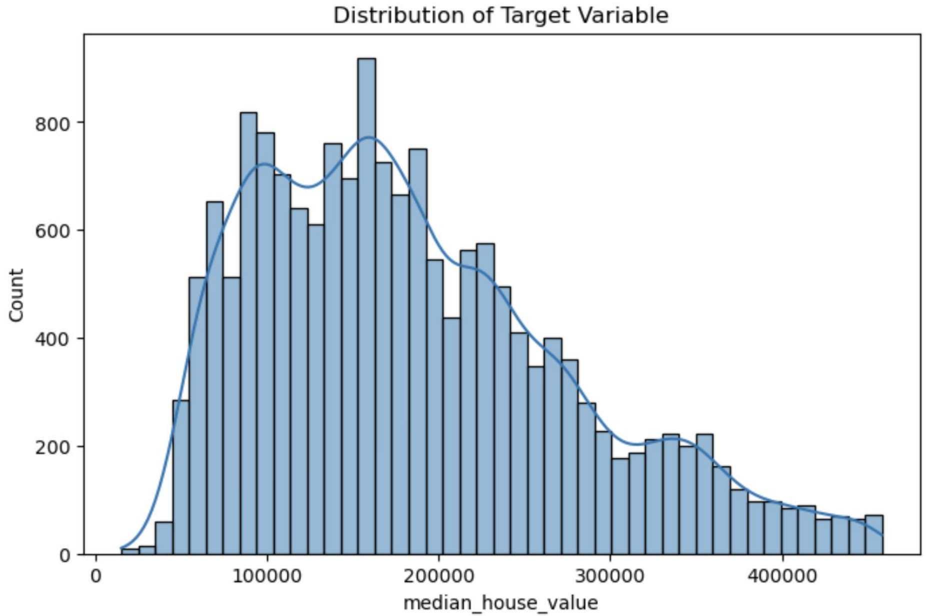


Relevant Statistics

Feature Skewness: Most variables are right-skewed, reflecting higher means than medians.

Correlation with Target:

- Median income shows highest correlation with house value (0.68)
- Ocean proximity plays a significant role in pricing patterns (0.51)



Modeling

Linear Regression

Exploratory Model

- Exploratory approach to establish initial insights and benchmark predictions.
- Identifying basic feature relationships.

Random Forest Regressor

Baseline Model

Using multiple decision trees to make predictions.

Strengths: Robust to overfitting, handles non-linear relationships well.

Weaknesses: Performance limited by default settings, less interpretable.

Gradient Boosting

Improvement Model

Sequentially builds trees to minimize error on previous iterations.

XGBoost

Improvement Model

Optimized implementation of Gradient Boosting.

Neural Network

Improvement Model

Multi-layer perceptrons tested as a more flexible approach.

Experiments

```
Fitting 3 folds for each of 10
candidates, totalling 30 fits

RandomizedSearchCV*(cv=3,
estimator=RandomForestRegressor(
random_state=42),

param_distributions={'max_depth'
: [10, 20, None],

'min_samples_leaf': [1, 2, 4],

'min_samples_split': [2, 5, 10],

'n_estimators': [50, 100, 200,
300]},

random_state=42,
scoring='neg_mean_squared_error'
, verbose=1)
```

******We got comparable results with GridSearch.*

n estimators	max depth	min samples split	min samples leaf	mean test score (10^9)	STD test score (10^8)
100	None	5	1	2.005070	1.358761
300	None	10	2	2.013681	1.342007
300	None	10	1	2.018342	1.353665
300	20	5	4	2.020988	1.356353
200	20	2	4	2.025034	1.323973
100	20	10	1	2.027694	1.357980
200	10	5	4	2.244064	1.222815
300	10	10	1	2.265825	1.191305
200	10	10	1	2.269705	1.168511
50	10	5	1	2.271323	1.145363

Conclusions

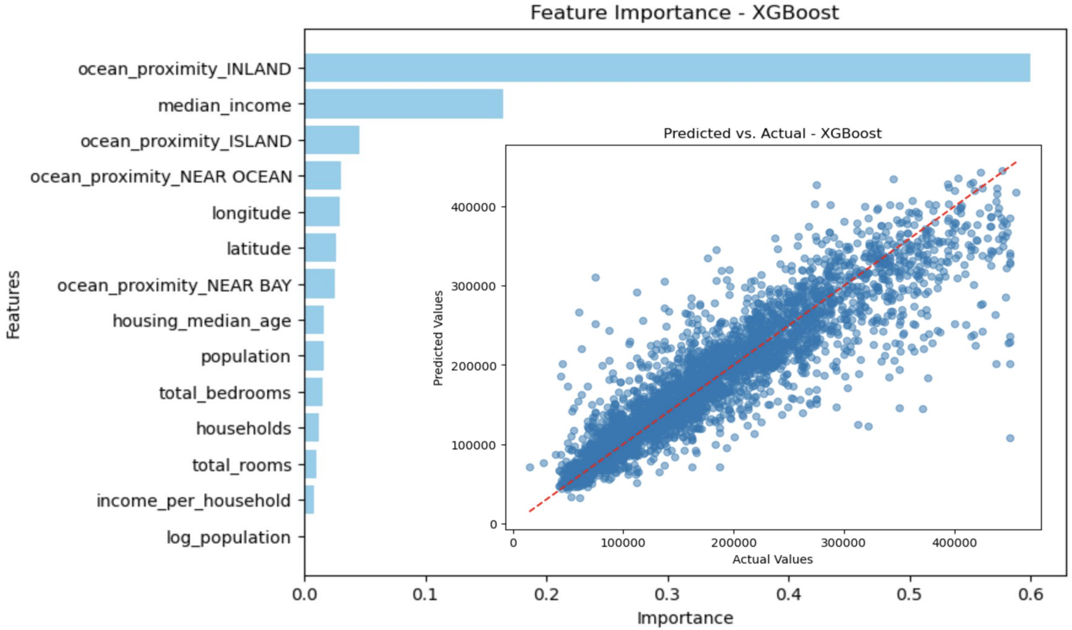
Model	MSE (10^9)	MAE	R^2 Score
Random Forest Regressor	2.018	30,977.54	0.7563
Gradient Boosting	2.15	32,898	0.741
XGBoost	1.60	26,865	0.806
Neutral Network	2.85	38,779	0.656

Key Learnings:

- **XGBoost** delivers the strongest predictive capability.
- Variables like **population**, **households**, and **median income** are effective in predicting housing prices but lack completeness.

Future Work:

- Enhance Feature Set
- Model Exploration
- Data Enrichment (temporal trends, spatial analyses)



Additional Feedback?

Chelle Davies (chelle.davies@ischool.berkeley.edu)

Ziyad Amer (zaa4uf@ischool.berkeley.edu)