CALIFORNIA HOUSING PRICE PREDICTION

MALIK ALRASYID BASORI

Purwadhika
Digital Technology School

Purwadhika
Digital Technology School

CONTENT

INTRODUCTION

The California real estate market is dynamic and influenced by factors like location, socioeconomic status, and proximity to the ocean.

Accurate house price predictions are crucial for investors, home buyers, and policymakers. This project uses the California Housing dataset, which includes data on geography, house age, rooms, population, and income.

The goal is to create a model that predicts median house values accurately, helping investors find high-value opportunities.

PREDICT ACCURACY

GOALS AND OBJECTIVES

To predict median house values with low-level errors model
to make data-driven decisions regarding real estate/property in California

Analyze Key Factors

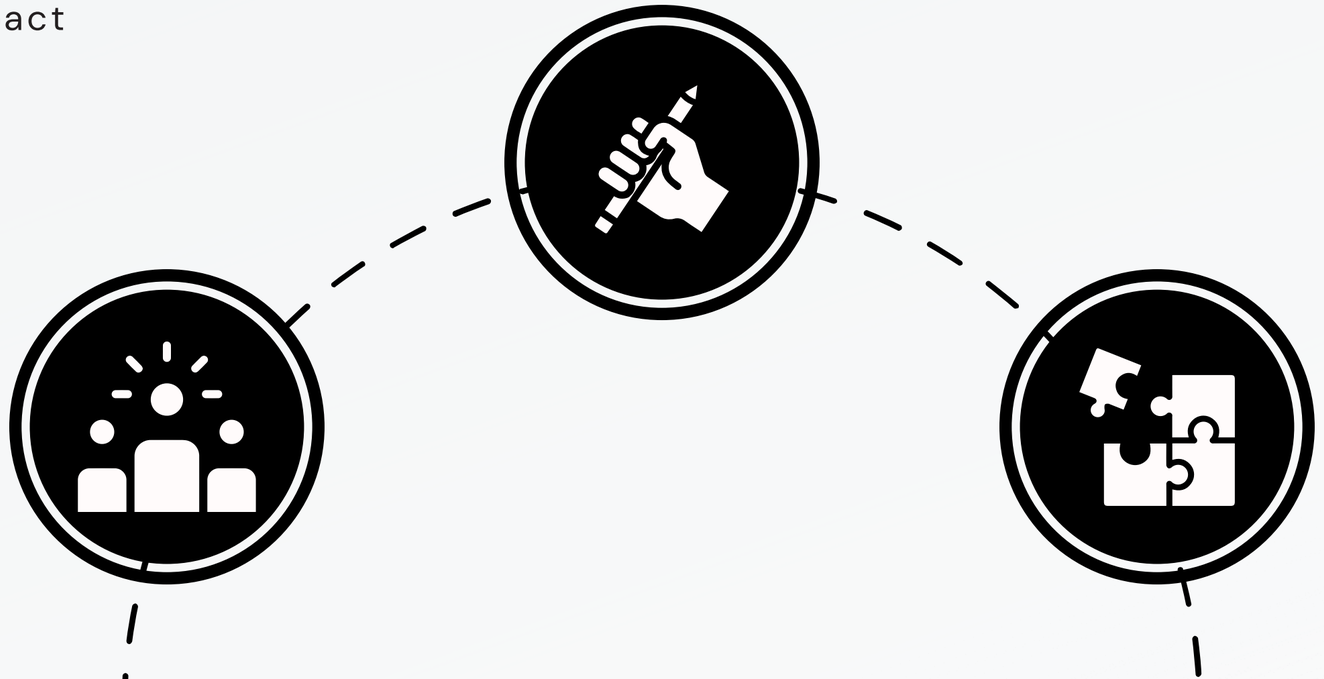
Perform feature importance analysis to understand which factors most significantly impact house prices

Develop a Predictive Model

Build and evaluate different regression models to predict the median house value in California based on the available dataset

Provide Actionable Insights

Translate model results into actionable insights for stakeholders such as real estate investors, home buyers, and policymakers



CALIFORNIA HOUSING DATASET

contains information on various housing and demographic characteristics of California districts, often referred to as block groups, the smallest geographical unit for which the U.S. Census Bureau publishes sample data.

Preview

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
-119.79	36.73	52.0	112.0	28.0	193.0	40.0	1.9750	INLAND	47500.0
-122.21	37.77	43.0	1017.0	328.0	836.0	277.0	2.2604	NEAR BAY	100000.0
-118.04	33.87	17.0	2358.0	396.0	1387.0	364.0	6.2990	<1H OCEAN	285800.0

Purwadhika
Digital Technology School

Rows/Column	: 14448/10
Target Variable	: Median House Value
Quantitative Feature	: 8
Qualitative Feature	: 1

UNIQUE FIDNING

Cleaning	Result	Action
Duplicates	0	None
Negative Value	14448	None
Missing Value	137	Filling when training model
Outliers	6 Column has Outlier (total_rooms, total_bedrooms, population, households, median_income, median_house_value)	Remove Multivariate Outlier when training model

UNIQUE FIDNING

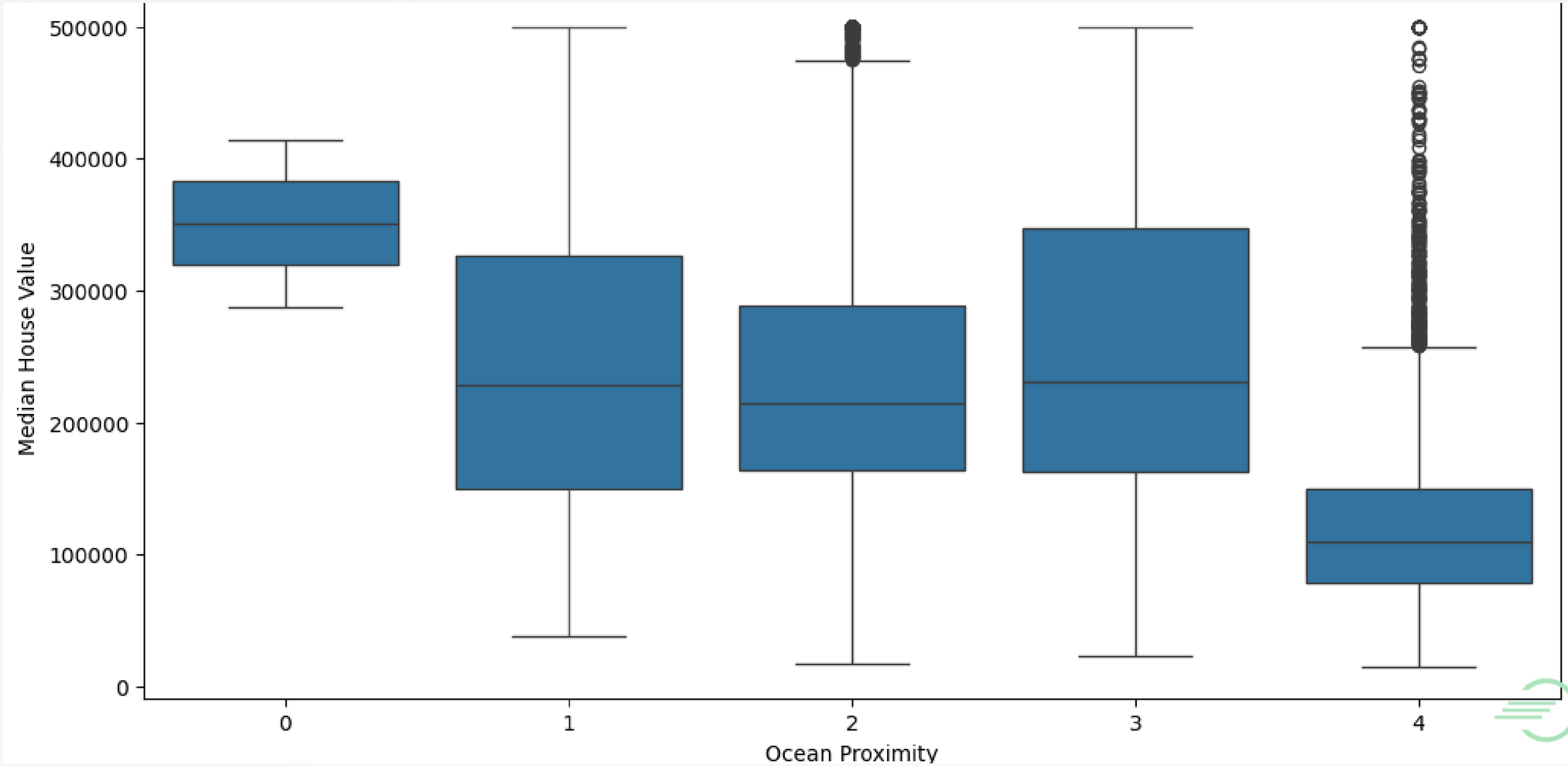
Correlation Matrix

Column Name	Correlation Value
Total Rooms	0.93, 0.85, 0.92
Total Bedrooms	0.93, 0.87, 0.98
Population	0.85, 0.87, 0.91
Household	0.92, 0.98, 0.91

Column Name	Correlation Value
Median House Value	0.69, -0.42
Median Income	0.2, 0.69
Ocean Proximity	0.2, -0.42

UNIQUE FIDNING

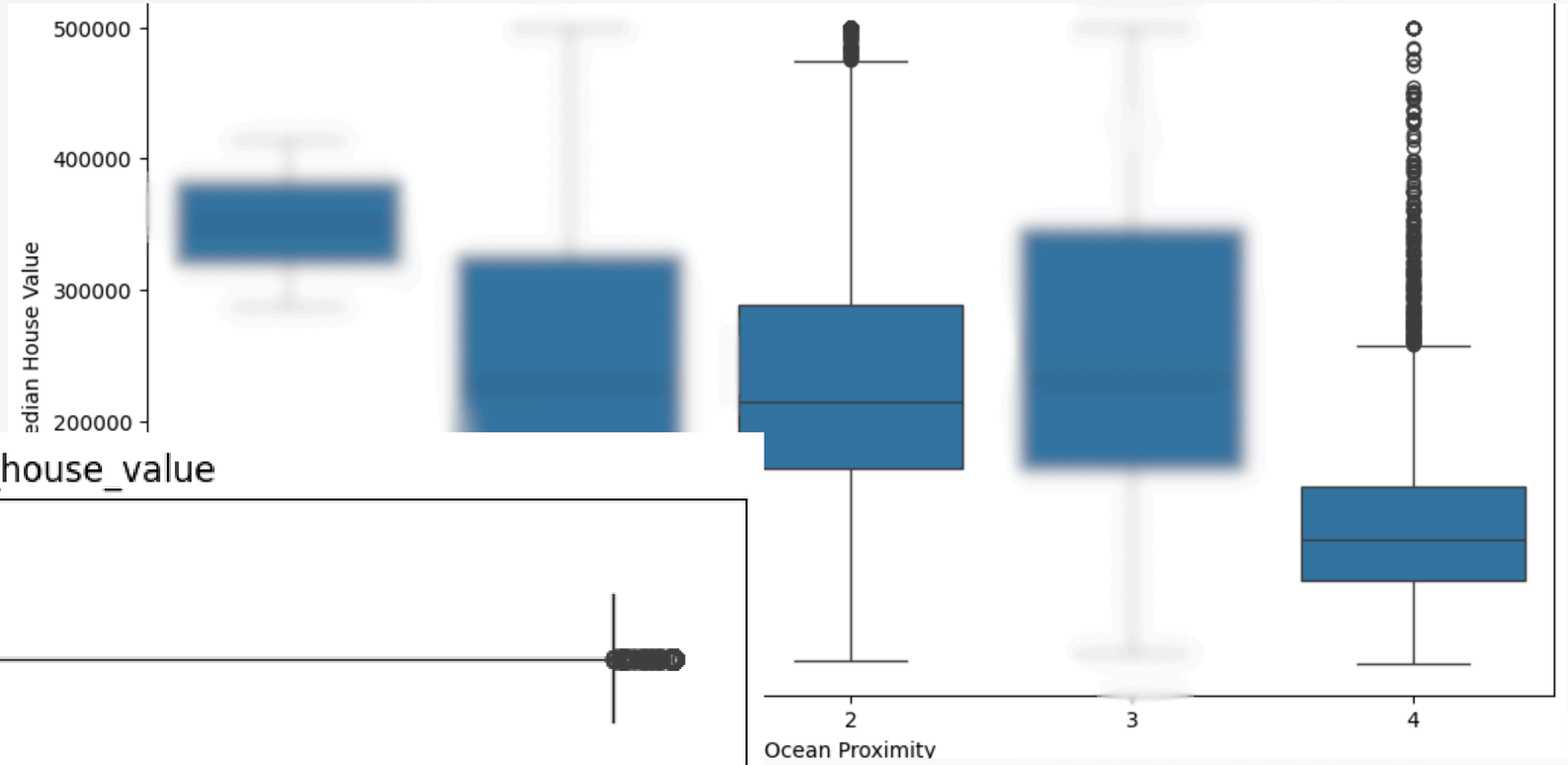
Ocean Proximity vs Median House Value

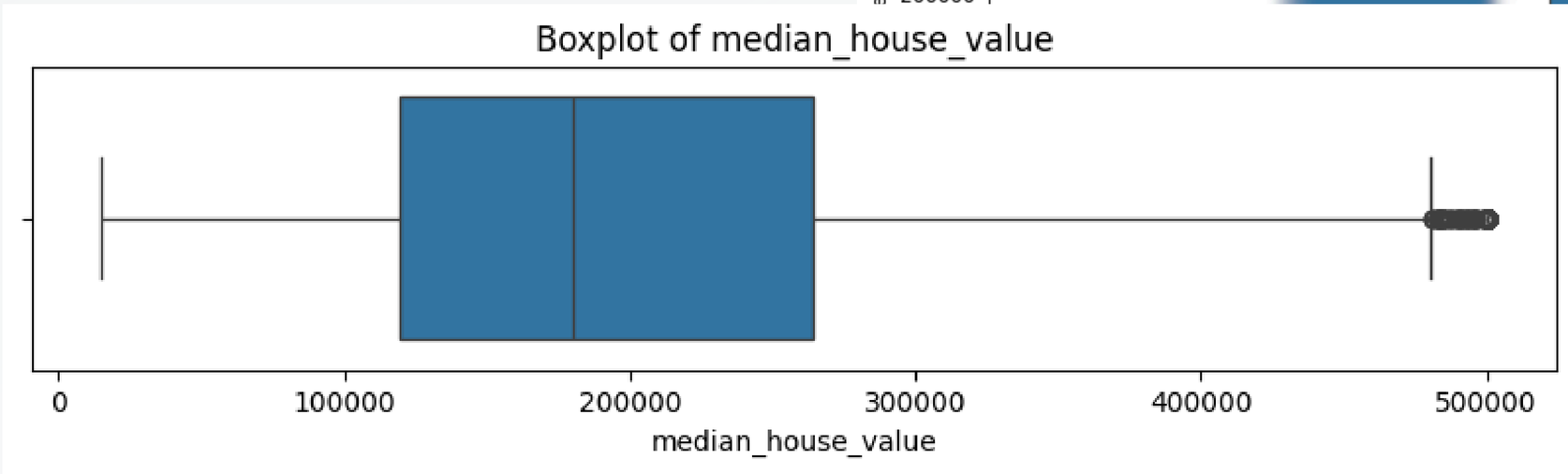


Column Value Order
0 = Island
1 = Near Ocean
2 = < 1H Ocean
3 = Near Bay
4 = Inland

UNIQUE FIDNING

Outliers





Purwadhika
Digital Technology School

DATA PREPROCESSING

01 FEATURE ENGINEERING

02 ENCODING

03 IMPUTE NULL VALUE

04 SCALING

Purwadhika
Digital Technology School

Preprocessing	Assessment	Action
Feature Engineering	Make a new column: • rooms_per_household = total_rooms / households • bedrooms_per_room = total_bedrooms / total_rooms • population_per_household = population'] / households	This will only applied to feature
Encoding	Using Custom Ordinal Encoding	This will only applied to Ocean Proximity Column
Impute Null Value	Using Simple Imputer with Median Value	This will only applied to Feature when training, but applied to both Feature and Target when saving the Best Model
Scaling	Using Standard Scaler	This applied to both Feature and Target

Purwadhika
Digital Technology School

MODELING

01 ALGORITHM & OUTLIER HANDLING METHOD SELECTION

02 HYPERPARAMETER TUNING

03 FEATURE IMPORTANCE

04 MODEL INTERPRETATION

Purwadhika
Digital Technology School

Preprocessing	Assessment	Action
Algorithm & Outlier Handling Method Selection	Outlier Method: Isolation Forest, Local Outlier Factor, One Class SVM, Elliptic Envelope, DBSCAN, IQR Method Model: Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, Random Forest, Gradient Boosting Regression, XGBoost Regression, Support Vector Regression	Iterate for Each Outlier Method and Model, and find for the lowest RMSPE and R2 Diff
Hyperparameter Tuning	Using Best Outlier Method and Best Model after Selection	Hyperparameter only applied to Model, Outlier Method using the default value.
Feature Importance	Explain Model Impact to Data	Try to Interpret relationship between Feature and Predicted Outcome
Model Interpretation	Illustration for Cost Efficiency Evaluation	Calculate R2 and RMSPE to find Return of Investment of the Model

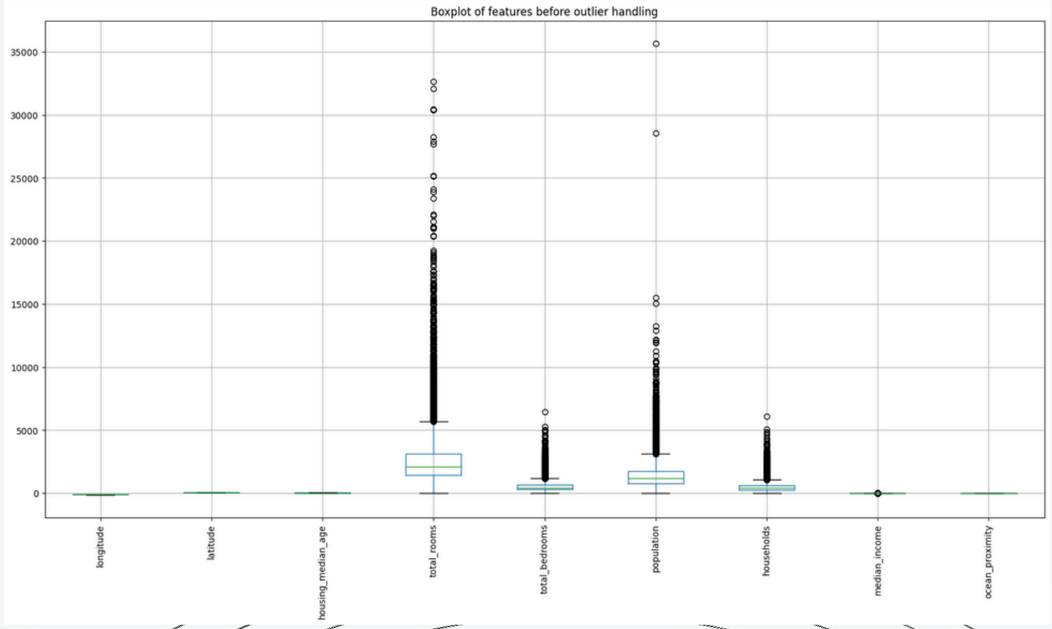
ALGORITHM & OUTLIER HANDLING METHOD SELECTION

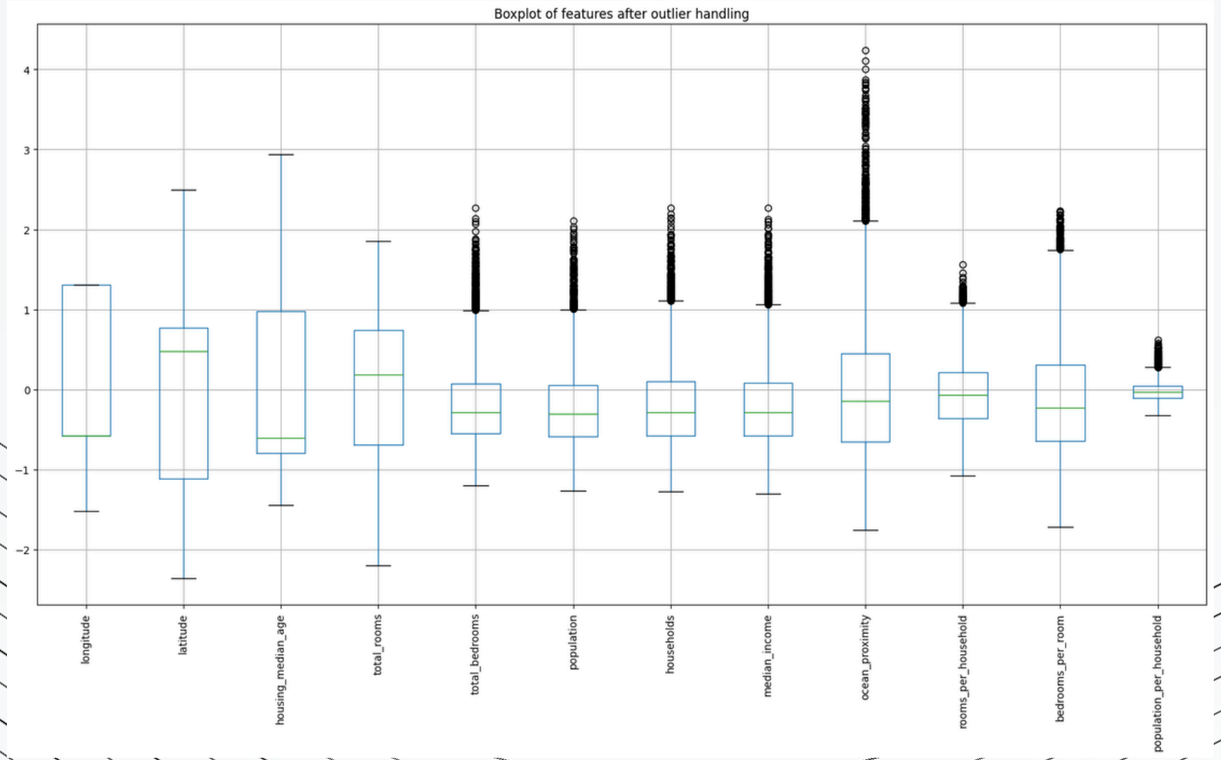
This is the Best Outlier Method & Best Model that fulfill the requirement of:
Test_R2 > 80%
Test_RMSPE < 35%
R2_diff < 5%

Outlier_Method	Elliptic Envelope
Model	Gradient Boosting Regression
Train_MSE	2182012801.33641
Train_RMSE	46712.019881
Train_MAE	33202.501762
Train_RMSPE	26.159174
Train_MAPE	18.369462
Test_MSE	2497834536.987638
Test_RMSE	49978.340679
Test_MAE	34929.489766
Test_RMSPE	30.116339
Test_MAPE	19.297088
Train_R2	0.822114
Test_R2	0.804733
R2_diff	0.017381
Num_Train_Outliers	2312
Num_Test_Outliers	578

Name: 29, dtype: object

ELLIPTIC ENVELOPE HANDLING OUTLIER





HYPERPARAMETER

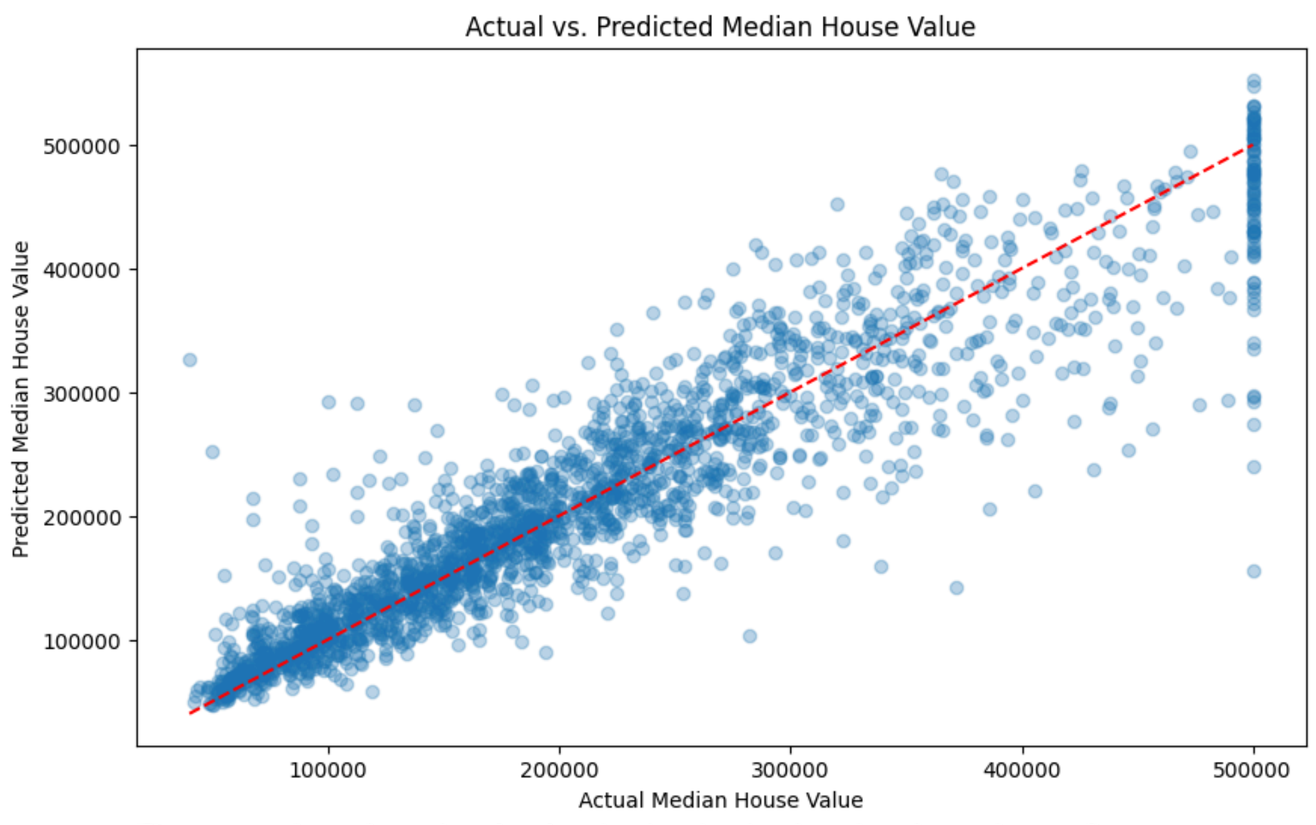
This is the Best Model Hyperparameter that fulfill the requirement of:
Test_R2 > 80%
Test_RMSPE < 30%
R2_diff < 5%

```
Outlier_Method                                                                                                                                          Elliptic Envelope
Model                                                                                                                                           Gradient Boosting Regression
Parameters        {'regressor__alpha': 0.85, 'regressor__learning_rate': 0.1, 'regressor__max_depth': 5, 'regressor__min_samples_leaf': 5, 'regressor__min_samples_split': 2, 'regressor__n_estimators': 100, 'regressor__subsample': 0.8}
Name: 40, dtype: object
```

```
Train_MSE      1378817538.415079
Train_RMSE          37132.432433
Train_MAE           25804.906705
Train_RMSPE            20.779541
Train_MAPE              14.27366
Test_MSE       1994117185.544048
Test_RMSE           44655.539248
Test_MAE            30636.637343
Test_RMSPE             27.396986
Test_MAPE              16.778728
Train_R2                0.887593
Test_R2                 0.844111
R2_diff                 0.043482
Name: 40, dtype: object
```

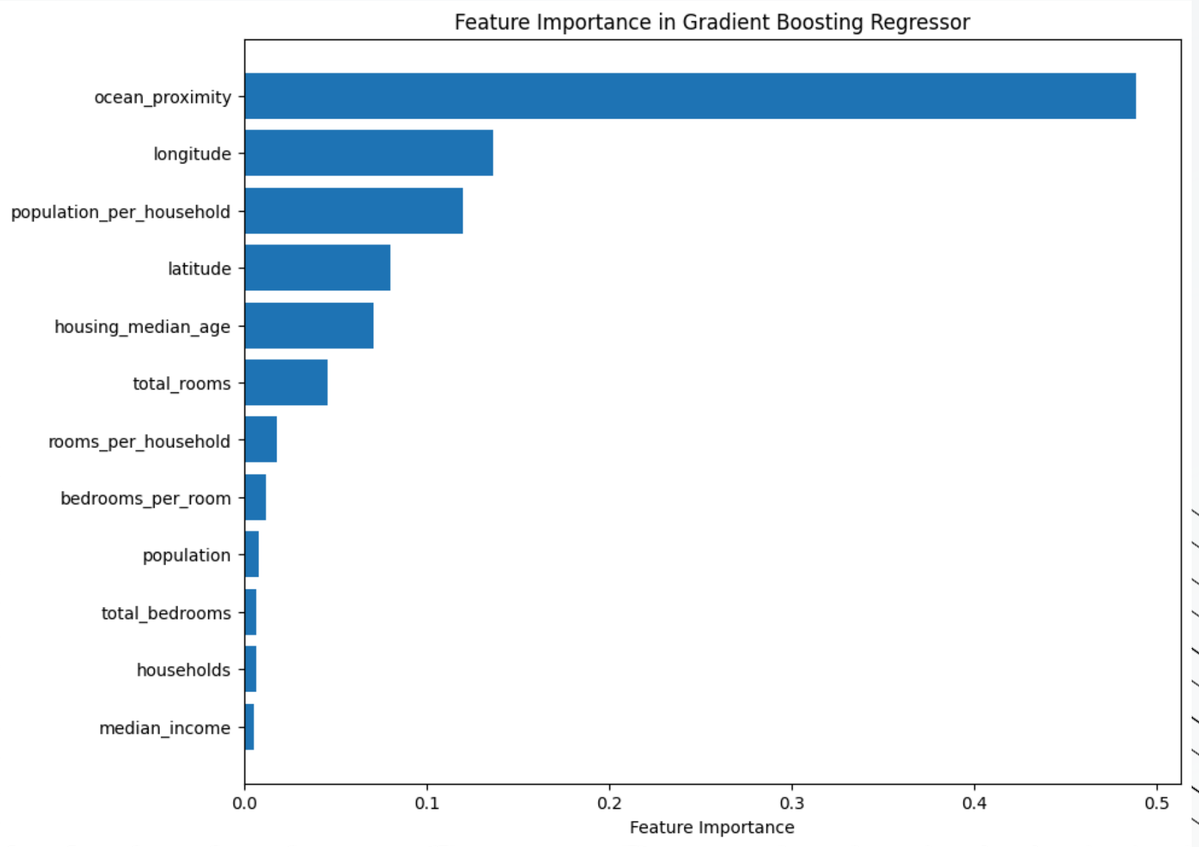
HYPERPARAMETER

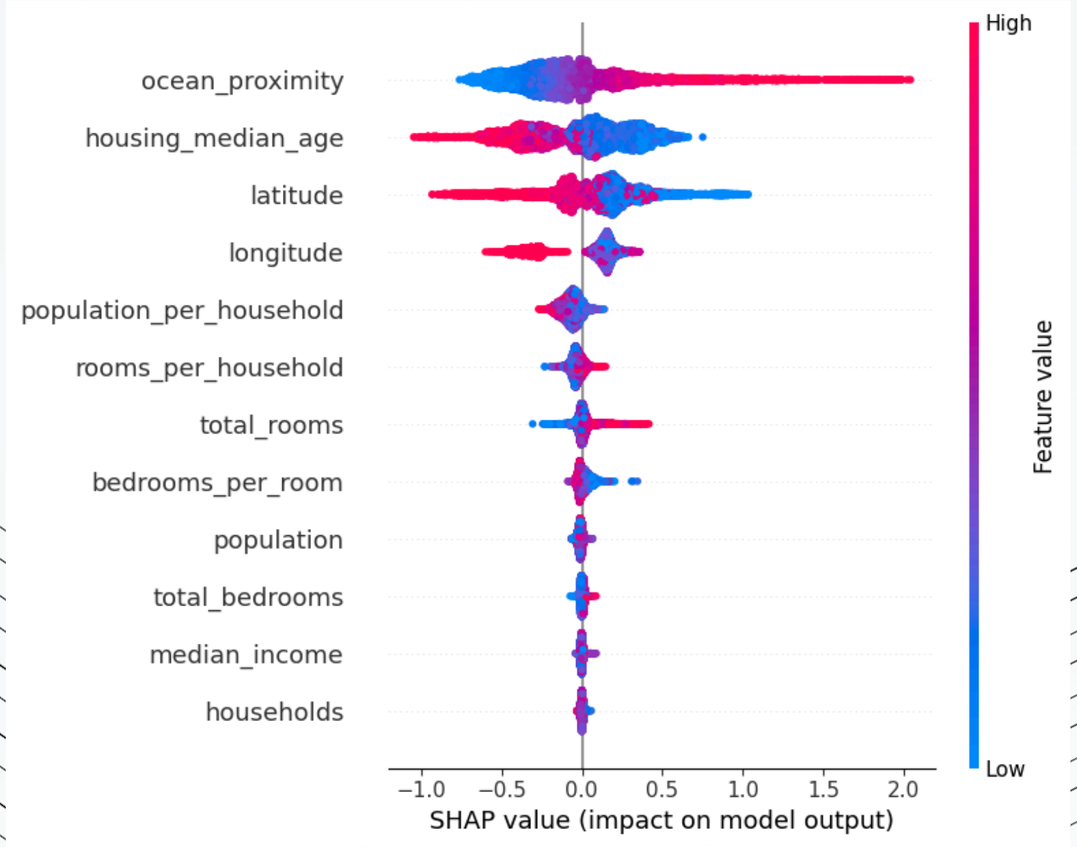
The scatter plot shows that the predicted median house values generally align well with the actual values, clustering around the red dashed line. There is increasing variability in predictions for higher house values, indicating potential model accuracy issues in that range. Additionally, there is a concentration of predicted values at the upper limit, suggesting a possible prediction cap in the model.



FEATURE IMPORTANT

Ocean Proximity become the most importance feature in this model and have a strong positive impact on housing prices. That mean the higher the Ocean Proximity (correspond to being closer to "INLAND" or “4” after Encoding) the housing prices tend to be higher





MODEL INTERPRETATION

Illustration Scenario:

R2 Cost Calculation:
- Better Pricing Decisions:
  - Scenario: Let's say your real estate business sells 100 houses per year, with an average price of $300,000.
  - Impact: If the model's high R2 enables more accurate pricing, you might reduce overpricing and underpricing.
- Revenue Increase:
  - Current Revenue: Without the model, let's assume your average house price is off by 5% due to less accurate pricing.
  - Model Impact: With the model, you might reduce this error to 2%. The 3% improvement (5% - 2%) can be translated into revenue increase.
- Calculation:
  - Average house price: $300,000
  - Total houses sold: 100
  - Potential revenue without model: 100 * $300,000 = $30,000,000
  - Improved pricing accuracy (3% of $300,000): $9,000 per house
  - Additional revenue: 100 houses * $9,000 = $900,000

RMSPE Cost Calculation:
- Cost of Prediction Errors:
  - Scenario: Let's say the average house price is $300,000, and your RMSPE is 28.40%.
  - Impact: An error of 28.40% translates to approximately $85,200 (28.40% of $300,000) per house.

- Improvement Potential:
  - If you reduce RMSPE from 28.40% to 20% when improving in the future, the error per house drops to $60,000 (20% of $300,000).
  - Savings per House: $85,200 - $60,000 = $25,200
  - Total Savings: For 100 houses, this translates to $25,200 * 100 = $2,520,000

MODEL INTERPRETATION

Illustration Scenario:

Calculation Metric:

ROI= ((Total Benefits - Total Costs) / Total Costs) * 100

Revenue Increase:
- Improved pricing accuracy (3% improvement): $900,000 additional revenue

Cost Savings:
- Reduced prediction error (from RMSPE improvement): $2,520,000 savings

Total Benefits = Revenue Increase + Cost Savings = $900,000 + $2,520,000 = $3,420,000

Total Costs (Illustration):
- Development Costs: $150,000
- Annual Operational Costs: $60,000
- Total Costs: $150,000 + $60,000 = $210,000

MODEL INTERPRETATION

Using ROI formula to calculate Cost Efficiency:

- Total Costs: $150,000 + $60,000 = $210,000
- Total Benefits: $900,000 + $2,520,000 = $3,420,000

$$\text{ROI} = \frac{\text{Total Benefits} - \text{Total Costs}}{\text{Total Costs}} \times 100\%$$

$$\text{ROI} = \frac{\$3,420,000 - \$210,000}{\$210,000} \times 100\% = 1528.57\%$$

The **1528.57% Return on Investment (ROI)** means that for every dollar you spend on developing and operating your house price prediction model.

MODEL INTERPRETATION

Dollar Interpretation:

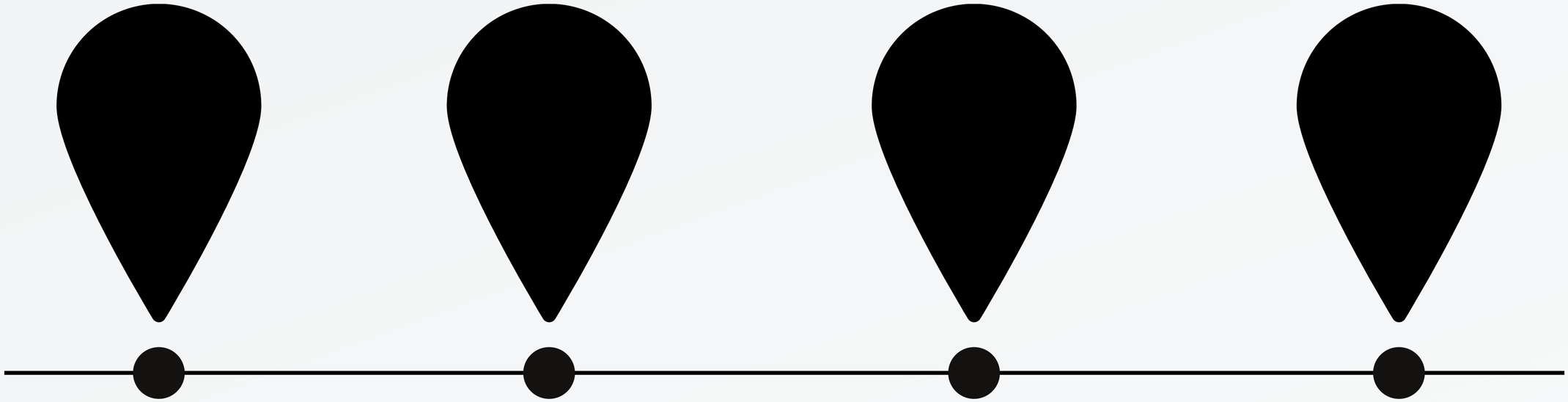
- Original Cost: $150,000 + $60,000 = $210,000
- Net Gain: $900,000 + $2,520,000 = $3,420,000

$$\text{ROI Percentage} = \frac{\text{Net Gain}}{\text{Total Costs}}$$

$$\frac{\$3,210,000}{\$210,000} = 15.29$$

This means for **every $1 spent**, you **get an additional $15.29** in net benefits.
So, the total **return is $16.29 for every $1 spent**
(original $1 + $15.29 net gain).

CONLCUSION



People who want to buy a house close to INLAND area tend to have a expensive house

People who buy new house also tend to have a more expensive house

Actual vs Predicted Median House Value from the model already giving great performance, make the model really applicable in actual scenario

With this model, people who using this are estimately gaining 1528.57% Return on Investment (ROI) or get an additional $15.29

RECOMMENDATION

Add more train data, so model can bring up more lower error on predicting target

Do more experiment on hyperparameter or trying another outlier handling to get more better model performance

Lowering the error of the model are the main objective to have more higher Cost Efficiency or Cost Return

MODEL DEPLOYMENT ON CLOUD