Machine Learning **Policy Insight** **Spatial Analysis**

California Housing Price Prediction

(1990 Baseline)

https://github.com/julianafoni/california-housing-1990-ML

By: Juliana Foni

Business Problem

Problem:

Why do housing prices vary dramatically across California?

- Coastal–inland inequality
- Strong socioeconomic differences
- Demand > supply in high-opportunity areas

Project Overview:

- Build a regression model to predict 1990 housing prices
- Identify the strongest predictors
- Create a reusable ML pipeline
- Provide policy insights using model interpretation

Dataset Overview

Dataset

```
Dataset shape (rows, columns): (14448, 10)

Column names:
['longitude', 'latitude', 'housing_median_age', 'total_rooms', 'total_bedrooms', 'population', 'households', 'median_income', 'ocean_proximity', 'median_house_value']

Data types:
longitude             float64
latitude              float64
housing_median_age    float64
total_rooms           float64
total_bedrooms        float64
population            float64
households            float64
median_income         float64
ocean_proximity        object
median_house_value    float64
dtype: object
```

Metric	Value	Description
Number of Observations (Rows)	14,448	Housing Districts
Number of Variables (Columns)	10	Features and Target
Granularity	Census District	Represents a Local Housing Market Cell
Data Type	Tabular, Moderately Large	Suitable for in-depth statistical analysis and ML model development

Data Issues Identified

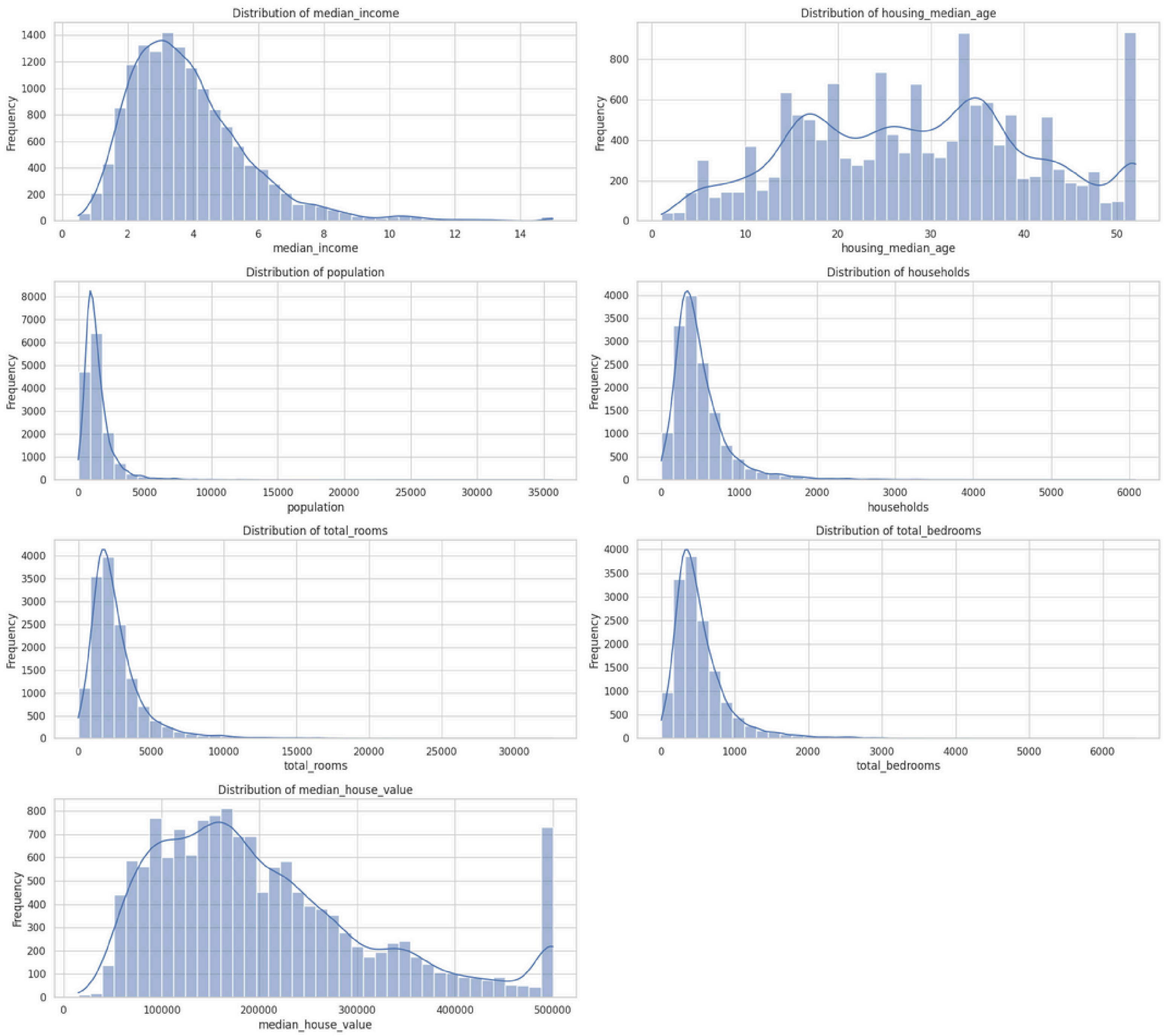
Problem:

- 137 missing values in total_bedrooms
- 137 missing in bedrooms_per_room
- 1 categorical variable
- Highly skewed income distribution
- Strong correlations between geographic variables

```
Number of duplicate rows: 0

Missing values per column:
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        137
population              0
households              0
median_income           0
ocean_proximity         0
median_house_value      0
dtype: int64

Rows before: 14448, after removing outliers: 12890
```



Preprocessing Pipeline

Step Performed:

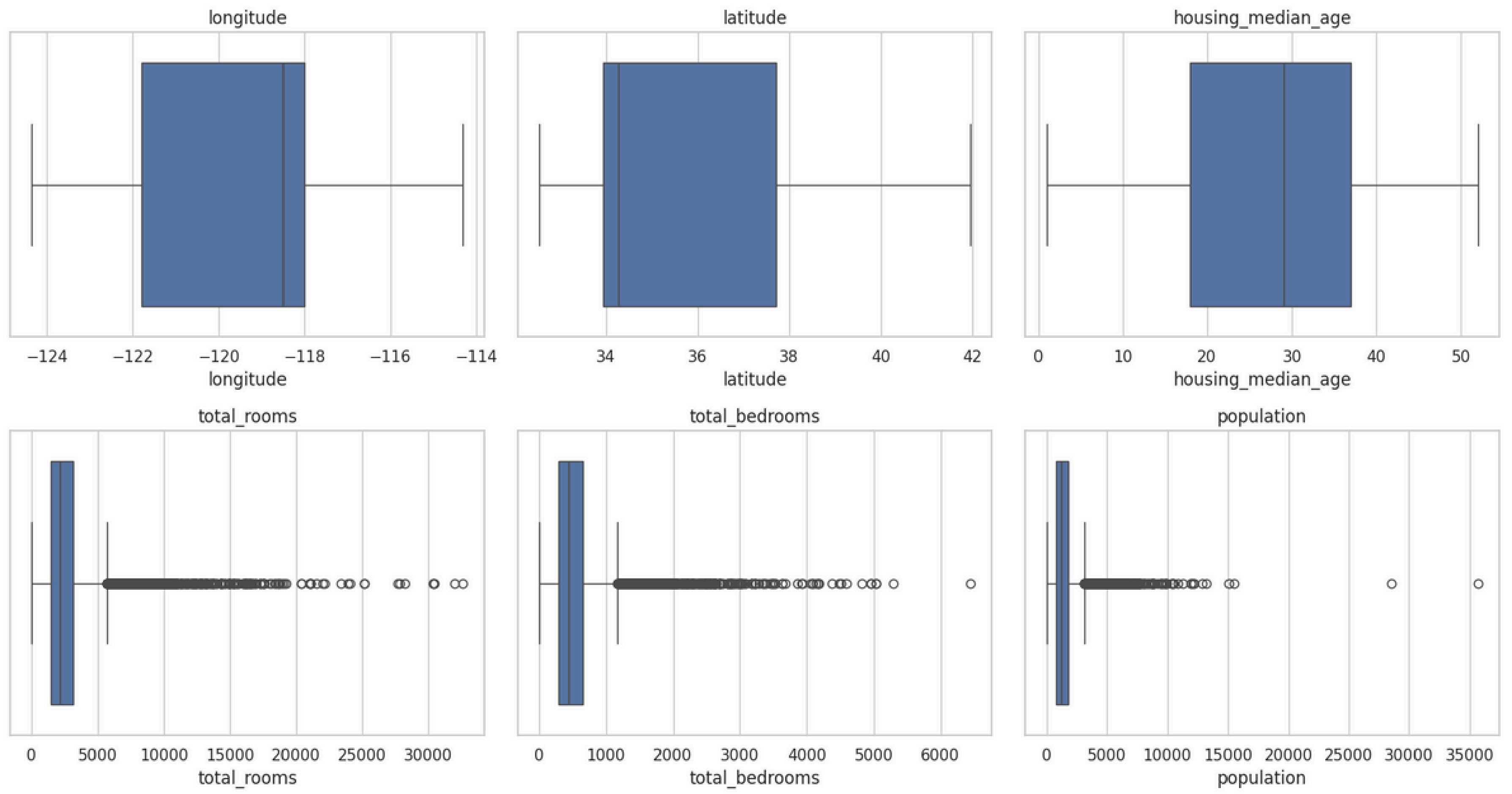
- Median imputation (numeric)
- One-hot encoding (ocean_proximity)
- Feature scaling (StandardScaler)
- Feature engineering:
  - rooms_per_household
  - population_per_household
  - bedrooms_per_room
- Train-test split (80/20)

Engineered Features

Engineered:

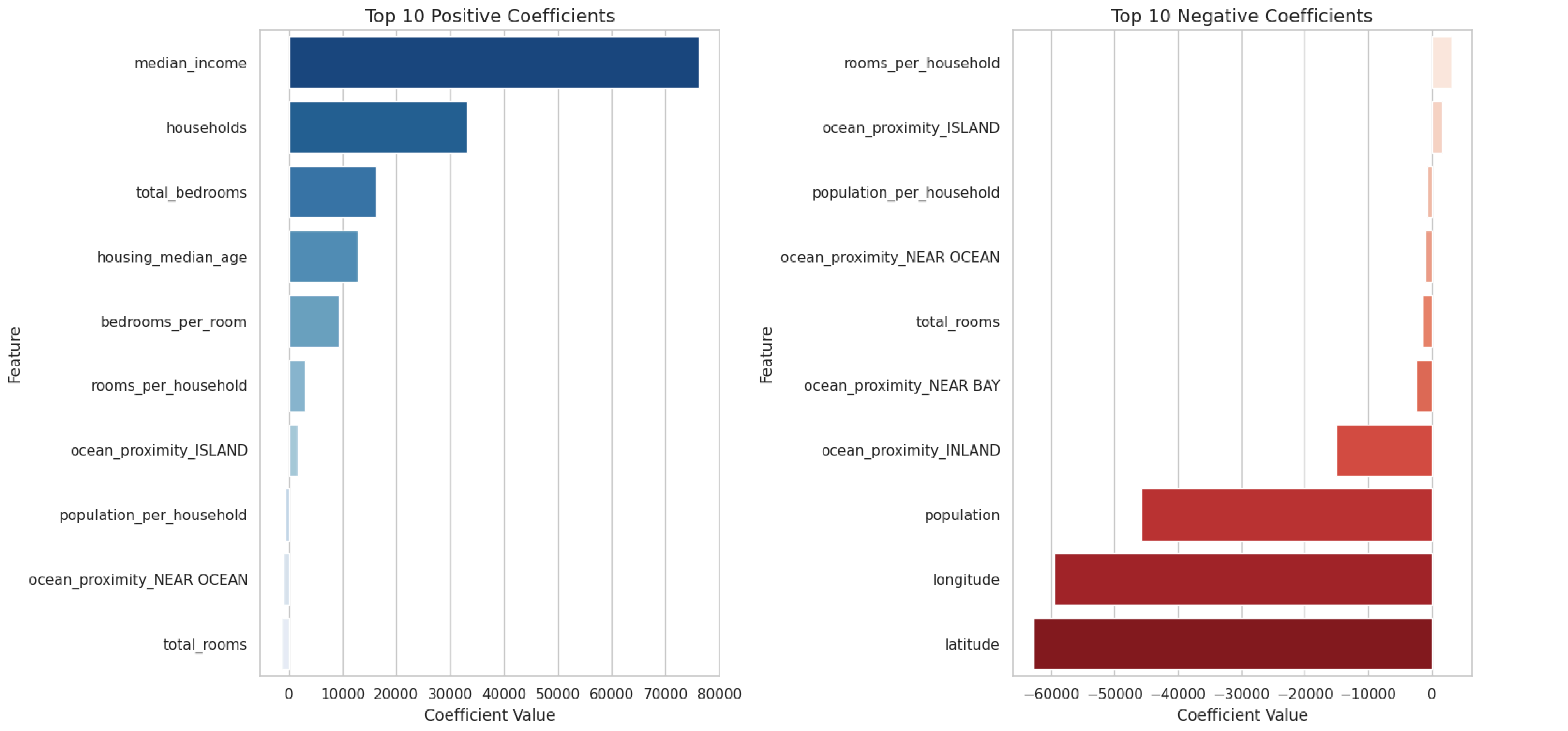
- rooms_per_household → housing spaciousness
- population_per_household → crowding indicator
- bedrooms_per_room → housing quality

Engineered Feature	Formula	Mean	Max
rooms_per_household	total_rooms / households	5.38	132.53
bedrooms_per_room	total_bedrooms / total_rooms	0.213	1.000
population_per_household	population / households	2.95	63.75



Model Tested

Regression Models:

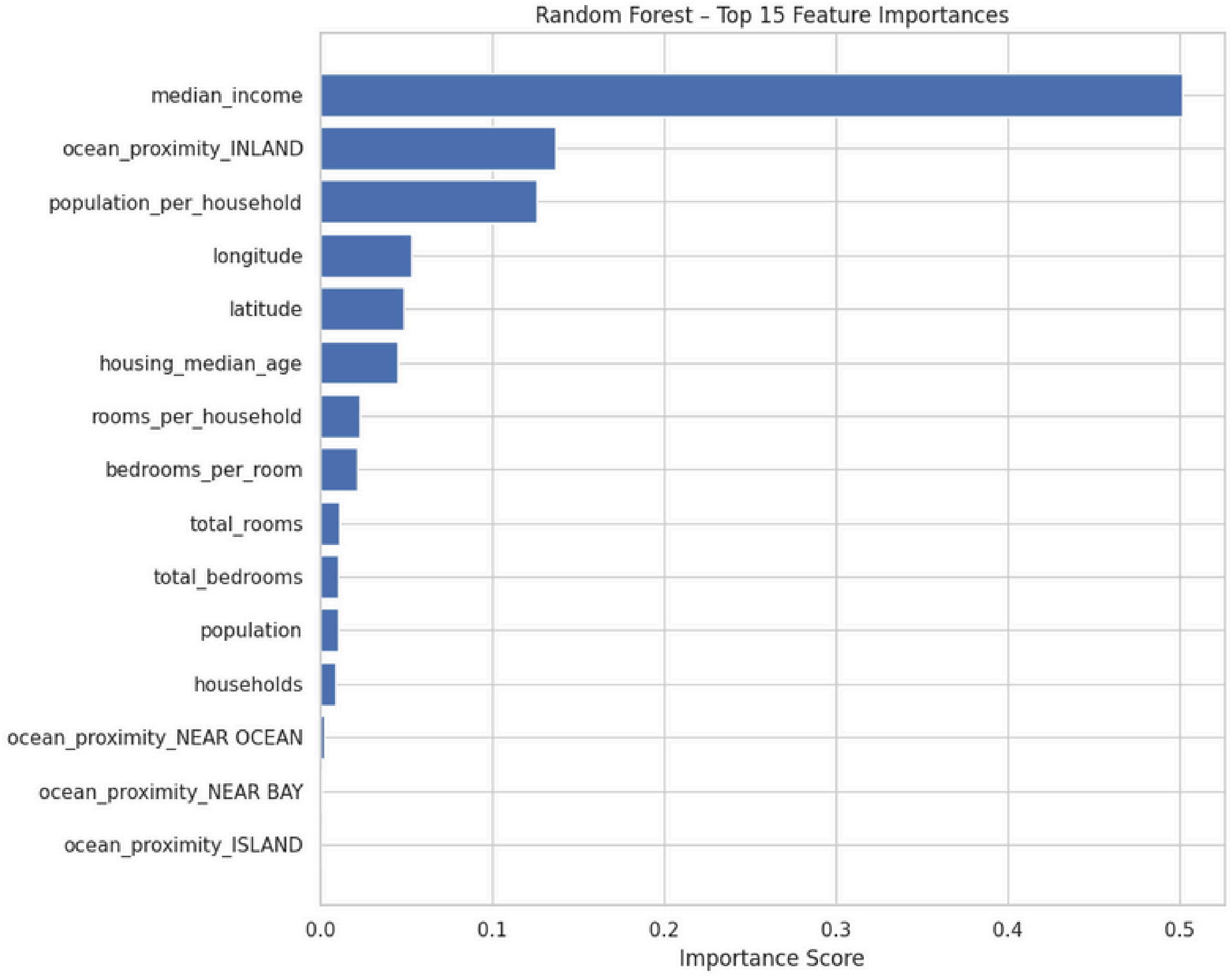


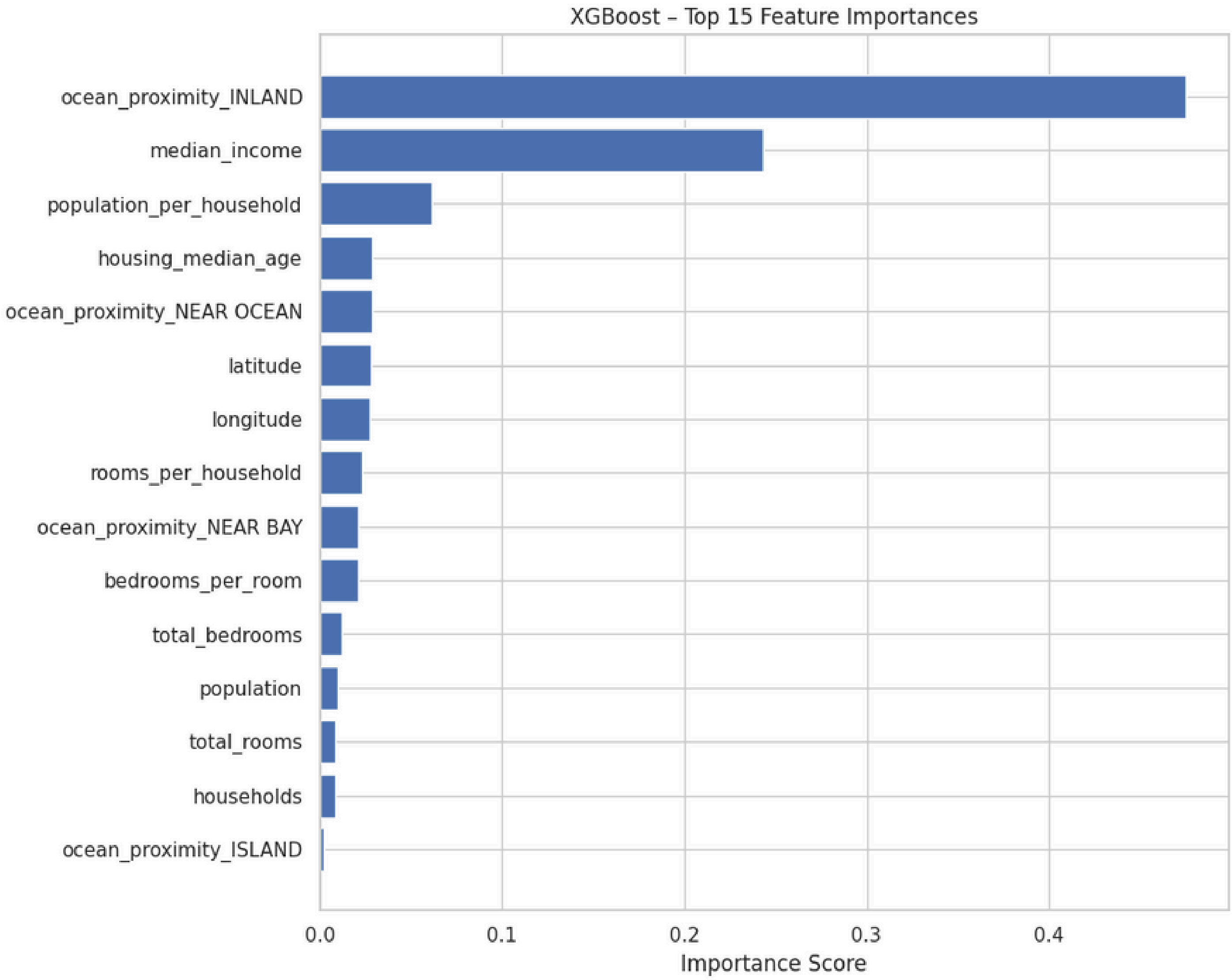
Actual Result:

Model Performance

Regression Model Performance Summary

	RMSE	MAE	R2
LGBMRegressor	46,760.52	31,187.96	0.8313
XGBRegressor	47,002.84	30,910.11	0.8296
RandomForestRegressor	52,160.18	34,306.56	0.7901
GradientBoostingRegressor	53,941.73	37,156.68	0.7755
KNeighborsRegressor	60,137.01	41,106.92	0.7210
DecisionTreeRegressor	65,334.66	43,036.66	0.6707
LinearRegression	66,846.16	48,934.86	0.6553





Why GXBoost Wins

Why?

- Captures nonlinear relationships
- Strong at interaction effects
- Resistant to multicollinearity
- Low cross-validation variance
- Best generalization

Model	RMSE	MAE	R2
LGBMRegressor	46,760.52	31,187.96	0.8313
XGBRegressor	47,002.84	30,910.11	0.8296
RandomForestRegressor	52,160.18	34,306.56	0.7901
GradientBoostingRegressor	53,941.73	37,156.68	0.7755
KNeighborsRegressor	60,137.01	41,106.92	0.7210
DecisionTreeRegressor	65,334.66	43,036.66	0.6707
LinearRegression	66,846.16	48,934.86	0.6553

Interpretation

Key Quantitative Findings

- Income = strongest driver of home value
- Geography (lon/lat) = coastal premium
- Crowding (pop_per_household) = lower home values
- Older districts = higher value
- Market already nonlinear in 1990 → why tree models dominate

Policy Recommendations

- Address income inequality - Programs, tax incentives, mixed-income housing
- Reform coastal zoning - Higher density housing allowed near coast
- Improve inland regions - Renovation, infrastructure, affordable housing
- Preserve historic districts - Balanced preservation + modernization
- Use ML for policy simulation- XGBoost = what-if analysis engine

Historical Interpretation

1990 dataset reveals:

- Coastal–inland inequality already long-established
- Income-driven segregation existed before tech boom
- Inland overcrowding signals early housing stress
- Valuable older housing stock near coast/urban center

What Stakeholders Gain

- Clear understanding of historical inequality
- Predictive tool for price forecasting
- Feature importance for targeted policy
- Ability to simulate scenarios
- Reusable ML pipeline

Thank you