MENINGKATKAN KETEPATAN PENILAIAN HARGA RUMAH UNTUK OPTIMASI STRATEGI PENJUALAN PROPERTI DENGAN ALGORITMA MACHINE LEARNING

Adrian Irsanda Boestamam

TABLE OF CONTENT

BUSINESS PROBLEM

Ketergantungan pada pendekatan manual dalam menentukan harga rumah. Penilaian harga sering kali bergantung pada pengalaman agen yang dapat berbeda tergantung tiap agen

Faktor penentu harga yang kompleks dan saling mempengaruhi. Karakteristik properti sulit dianalisis tanpa alat bantu analitik

Tidak adanya sistem otomatis yang mampu memberikan rekomendasi harga yang cepat dan akurat. Hal ini dapat memperlambat proses pengambilan keputusan dalam memberikan saran ke klien

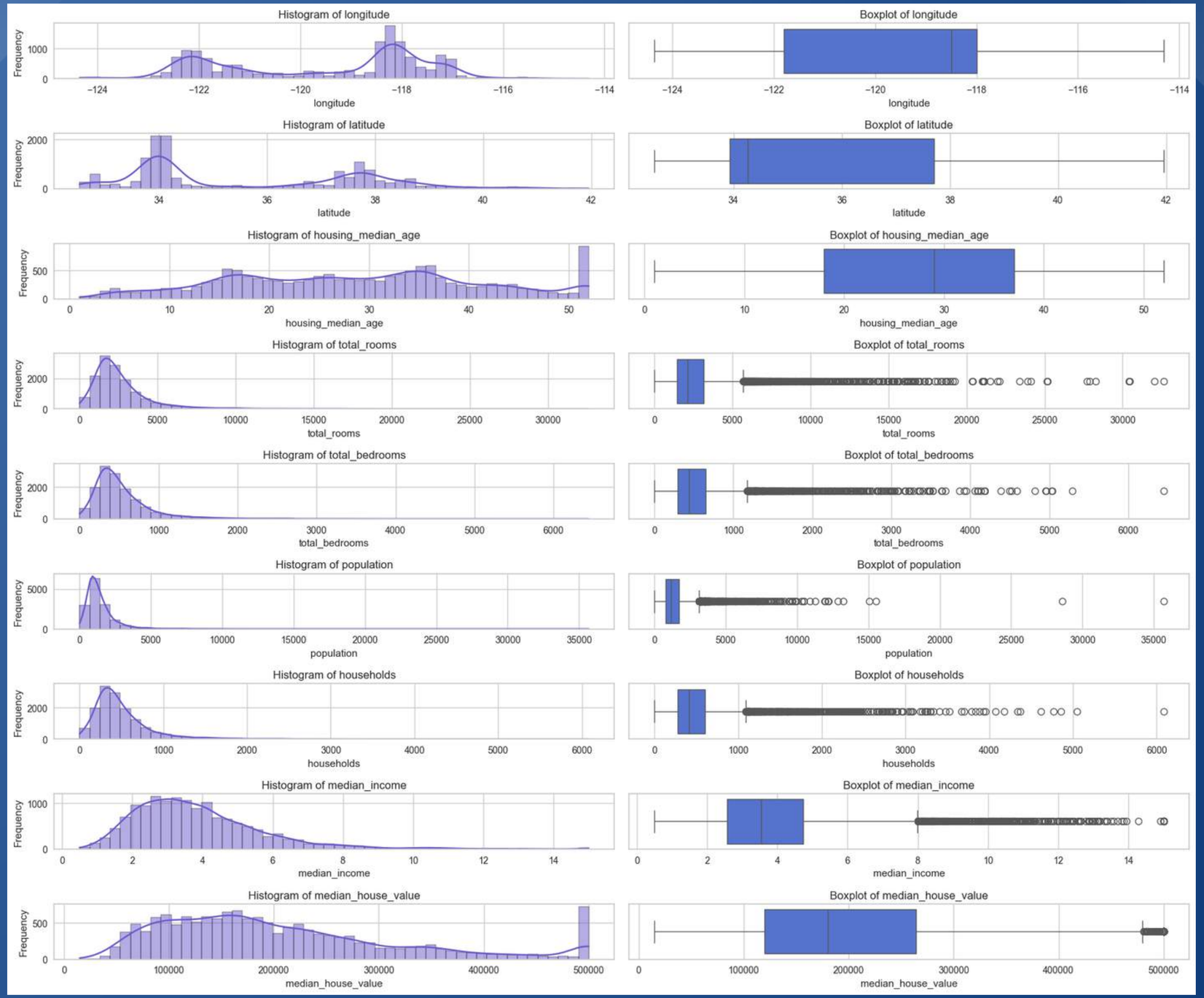
DATA UNDERSTANDING

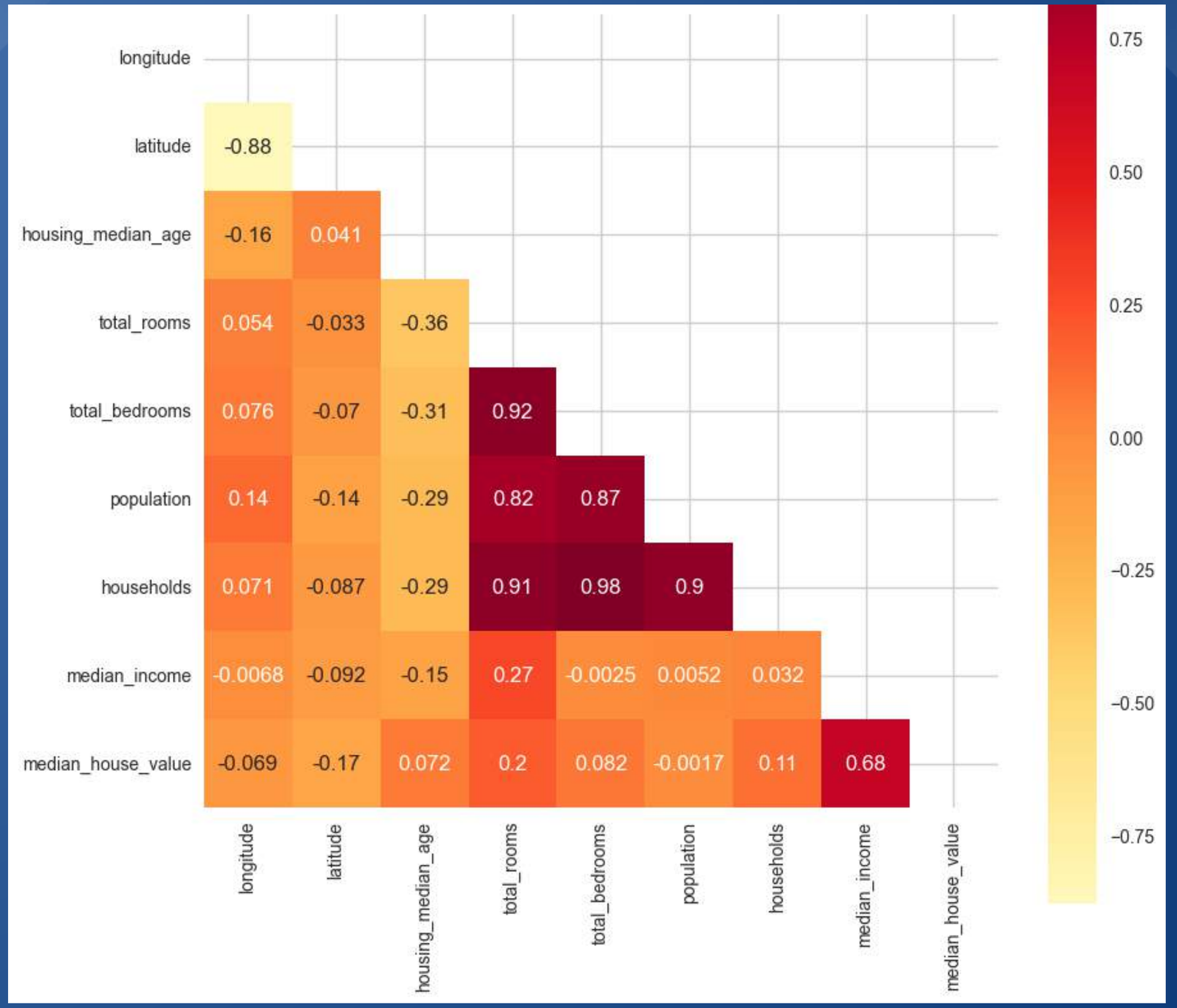
Dataset ini merupakan data sensus California tahun 1990. Dataset ini berisikan tentang informasi letak geografis, distribusi pendapatan, nilai rumah, kepadatan penduduk, karakteristik hunian dan tipe kategori rumah. Dataset terdiri dari 14448 baris dan 10 kolom.

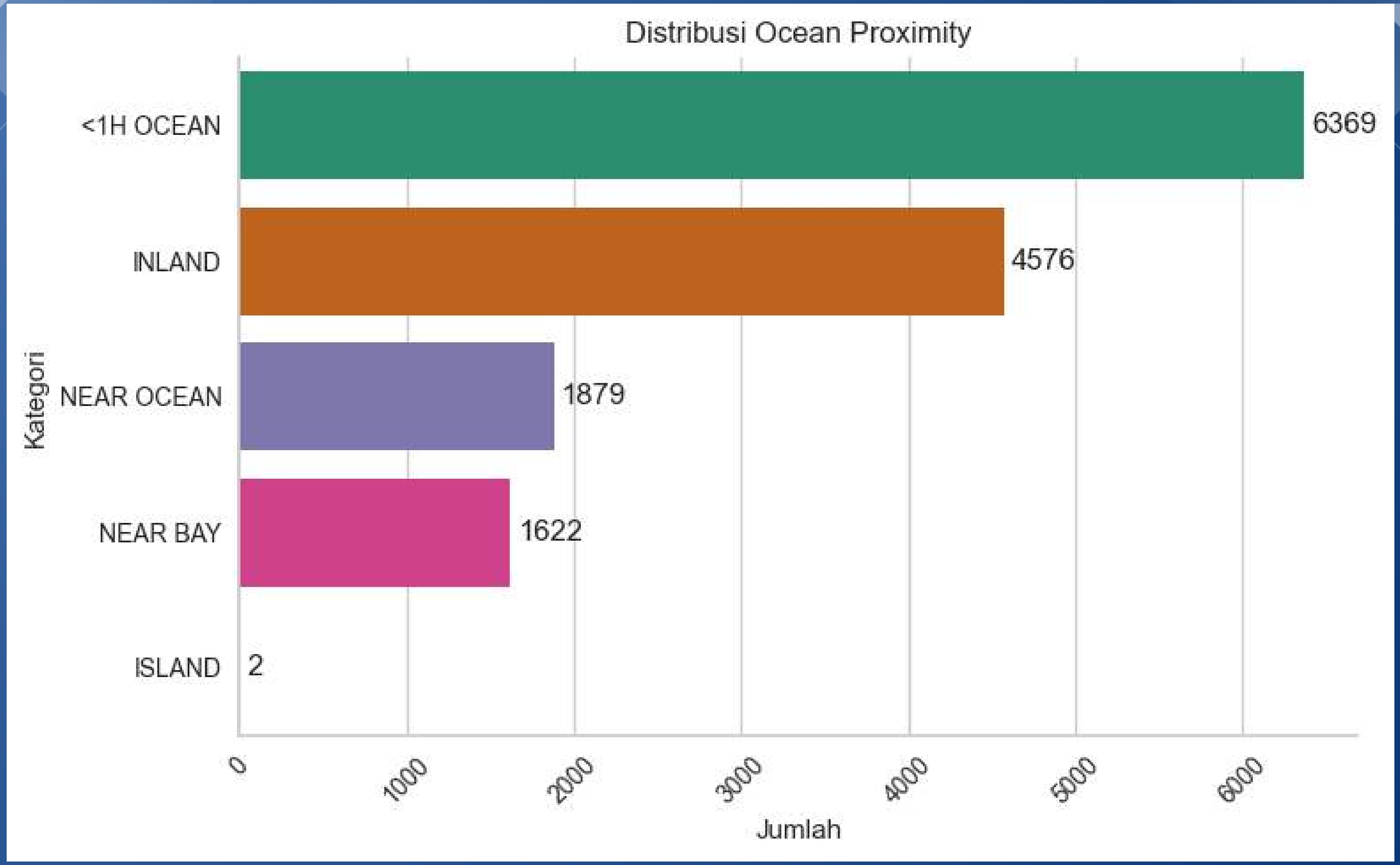
Longitude	Garis bujur lokasi rumah (dalam derajat)
Latitude	Garis lintang lokasi rumah (dalam derajat)
House Median Age	Usia median bangunan rumah di area tersebut (dalam tahun)
Median House Value	Nilai median rumah di area tersebut (dalam USD)
Median Income	Pendapatan median penduduk di area tersebut (dalam satuan puluhan ribu USD)
Total Rooms	Jumlah total kamar di area tersebut
Total Bedrooms	Jumlah total kamar tidur di area tersebut
Population	Jumlah total penduduk di area tersebut
Households	Jumlah total rumah tangga di area tesebut
Ocean Proximity	Kategori lokasi rumah terhadap laut

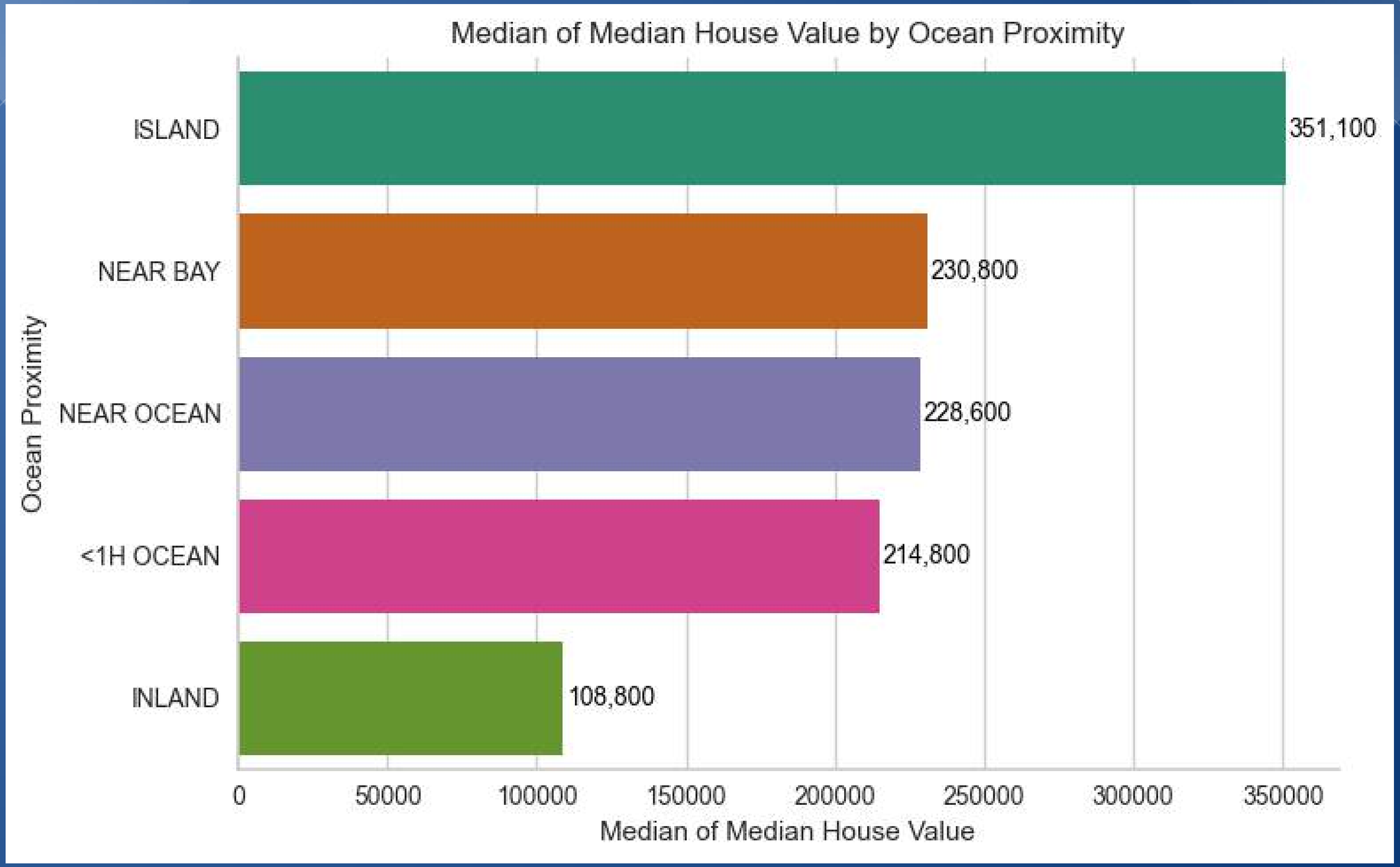
EXPLORATORY DATA ANALYST

- Numerical Columns
- Categorical Column


Histogram of longitude
Boxplot of longitude
Histogram of latitude
Boxplot of latitude
Histogram of housing_median_age
Boxplot of housing_median_age
Histogram of total_rooms
Boxplot of total_rooms
Histogram of total_bedrooms
Boxplot of total_bedrooms
Histogram of population
Boxplot of population
Histogram of households
Boxplot of households
Histogram of median_income
Boxplot of median_income
Histogram of median_house_value
Boxplot of median_house_value
Frequency
longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value


longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value
-0.88
-0.16 0.041
0.054 -0.033 -0.36
0.076 -0.07 -0.31 0.92
0.14 -0.14 -0.29 0.82 0.87
0.071 -0.087 -0.29 0.91 0.98 0.9
-0.0068 -0.092 -0.15 0.27 -0.0025 0.0052 0.032
-0.069 -0.17 0.072 0.2 0.082 -0.0017 0.11 0.68
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Distribusi Ocean Proximity
<1H OCEAN
INLAND
NEAR OCEAN
NEAR BAY
ISLAND
6369
4576
1879
1622
2
Kategori
0
1000
2000
3000
4000
5000
6000
Jumlah


Median of Median House Value by Ocean Proximity
ISLAND
NEAR BAY
NEAR OCEAN
<1H OCEAN
INLAND
351,100
230,800
228,600
214,800
108,800
Ocean Proximity
0
50000
100000
150000
200000
250000
300000
350000
Median of Median House Value

DATA CLEANSING AND FEATURE ENGINEERING

Missing Values

Drop < 1%

Duplicated Data

0 Duplicated Data

Handling Rare Values

Remove ISLAND
(only 2 rows)

New Feature

Population per Households
Rooms per Households
Bedrooms per Households

Remove Feature

Population
Households
Total Rooms
Total Bedrooms

Handling Outlier

Remove Outlier

PRE-PROCESSING

DEFINE X AND Y

Memisahkan antara target dan feature. Kolom median house value dijadikan sebagai target dan kolom lainnya dijadikan sebagai feature

TRAIN TEST SPLIT

Memisahkan data train dan data test dengan membagi proporsi 80:20 dan menggunakan random state 20

ENCODING SCALING

- Encoding : ocean proximity
- Scaling : logitude, latitude, housing median age, median income, population per household, room per household, bedroom per household

- Encoding = one hot encoder
- Scaling = robust scaler

MODELING

- Define Models
- Benchmark Models

DEFINE MODELS

- Linear Regression
- Ridge
- Lasso
- KNeighbors Regressor
- Decision Tree Regressor
- Random Forest Regressor
- XG Boost
- Gradient Boosting
- Ada Boost Regressor

BENCHMARK MODELS

Model	RMSE	MAPE	MAE
XG Boost	40529.03	0.16	27901.88
Random Forest	42920.09	0.17	29554.38
Gradient Boosting	44642.07	0.18	31977.82

Memilih model XG Boost sebagai model terbaik

MATRICS EVALUATION

01

RMSE (Root Mean Squared Error) adalah akar dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Digunakan untuk mengukur besar kesalahan dalam satuan yang sama dengan data

02

MAE (Mean Absolute Error) adalah rata-rata dari nilai absolut selisih antara nilai aktual dan nilai prediksi. Digunakan untuk memberikan gambaran seberapa jauh prediksi dari nilai sebelumnya

03

MAPE (Mean Absolute Percentage Error) adalah rata-rata dari persentase kesalahan absolut terhadap nilai aktual. Digunakan untuk mengukur kesalahan dalam bentuk persentase

EVALUATION

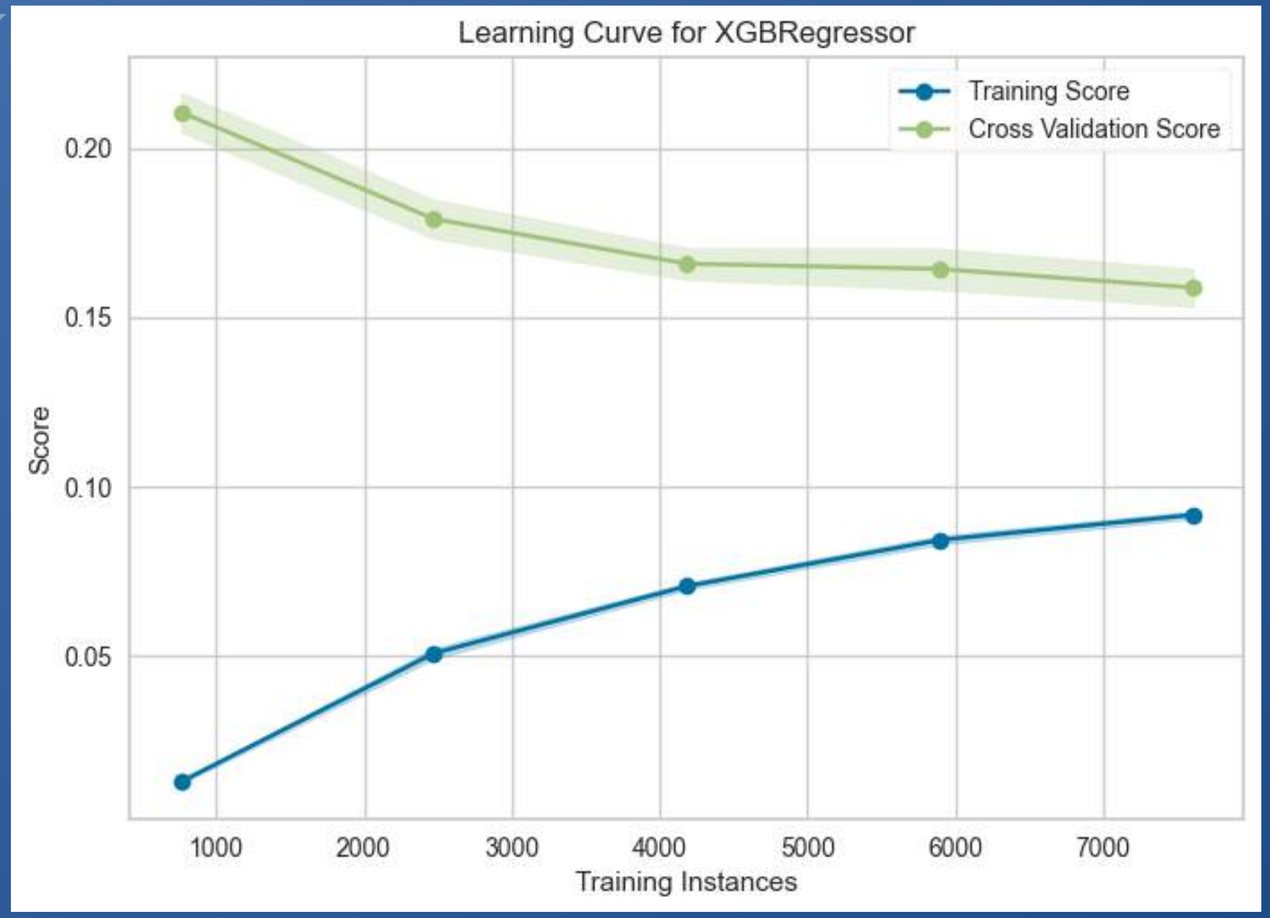
- Hyperparameter Tuning
- Tuning Comparison
- Learning Curve
- Residual Plot
- Actual vs Prediction
- Feature Importances

HYPERPARAMETER TUNING

model__n_estimator	**200**
model__max_depth	**5**
model__learning_rate	**0.2**
model__subsample	**1**
model__colsample_bytree	**1.0**
model__min_child_weight	**1**
model__gamma	**0**
model__reg_alpha	**0**
model__reg_lambda	**1**

TUNING COMPARISON

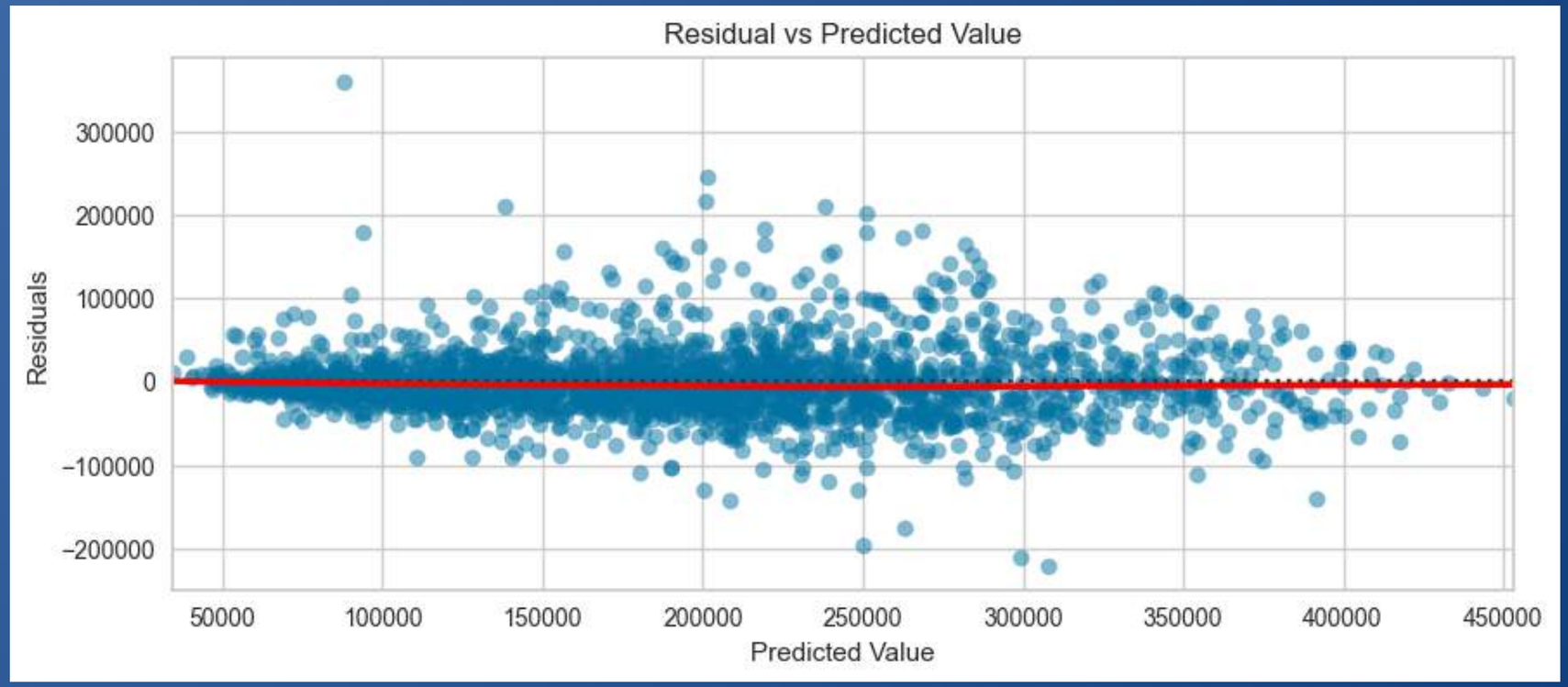
Model	RMSE	MAPE	MAE
Before Tuning	41307.33	0.159	27640.22
After Tuning	40667.97	0.158	27292.80



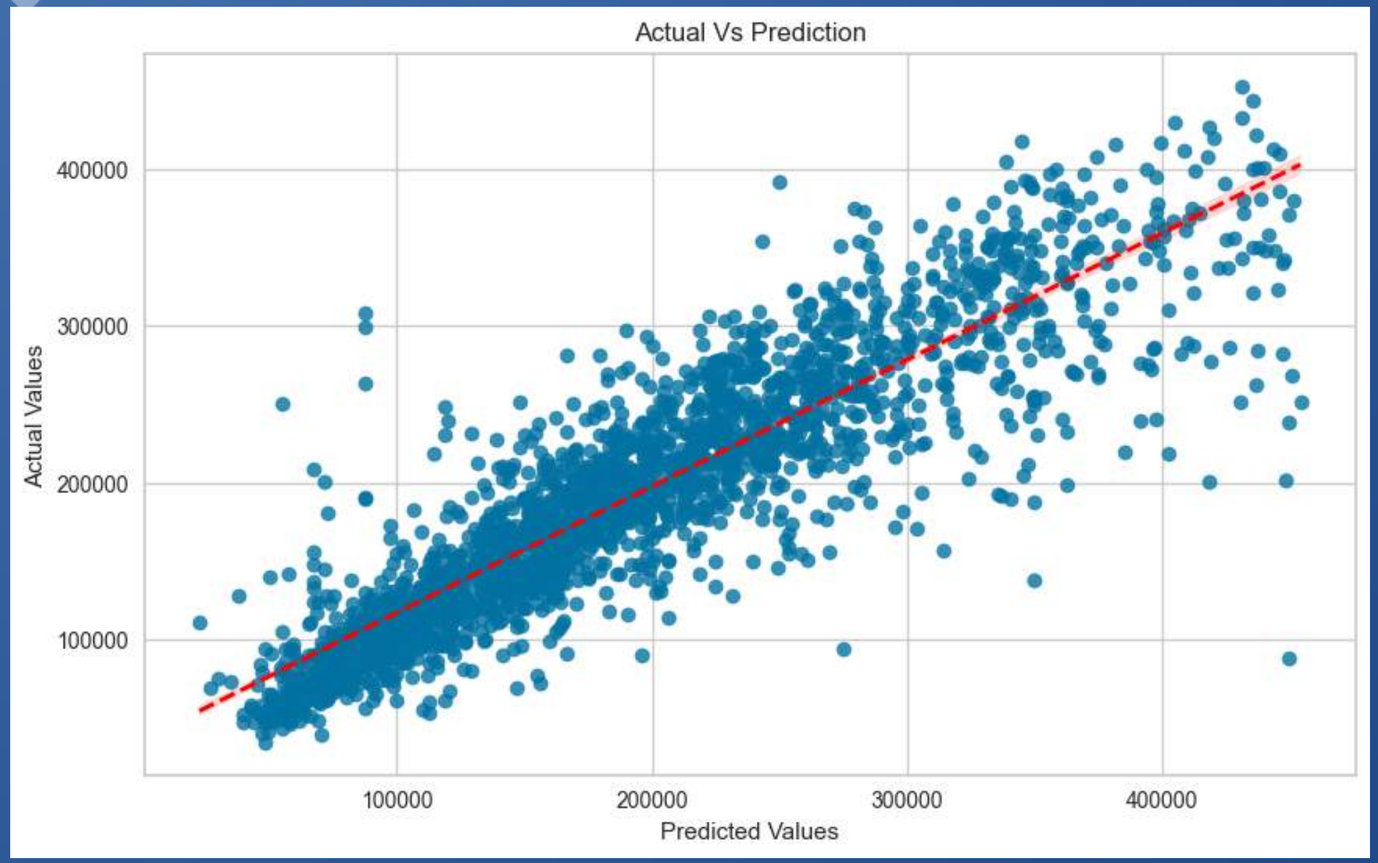
LEARNING CURVE

- Training Score
  - Model cocok untuk data train kecil
  - Nilai training error bertambah seiring bertambahnya data.
- Cross Validation Score
  - Performa meningkat seiring bertambah data
- Gap semakin sempit dengan bertambahnya data
- Menunjukkan overfit saat data sedikit, namun menjadi lebih baik saat data lebih banyak

RESIDUAL PLOT

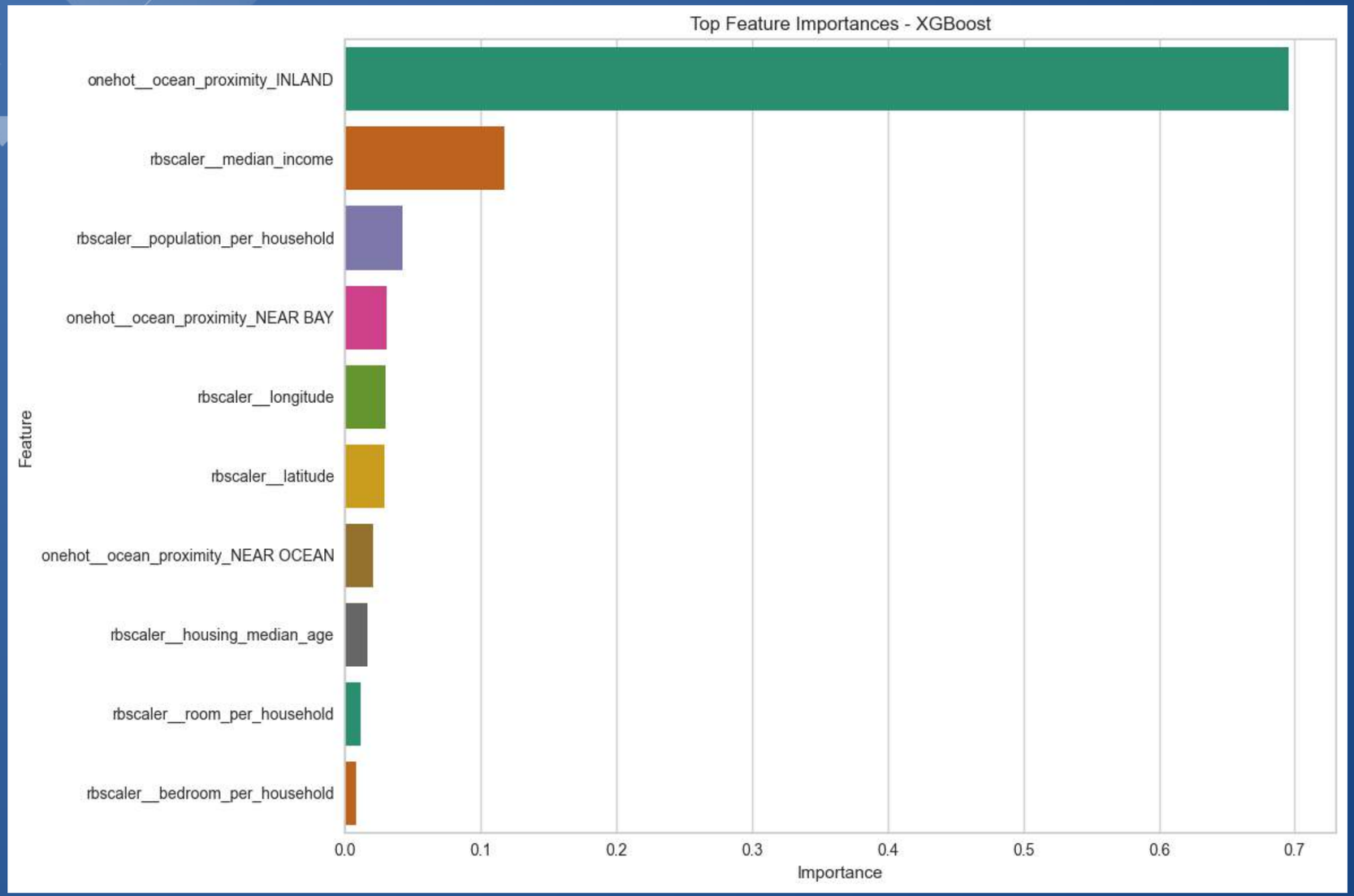


- Distribusi terpusat di angka 0
- Tidak ada pola U atau V = model cukup baik



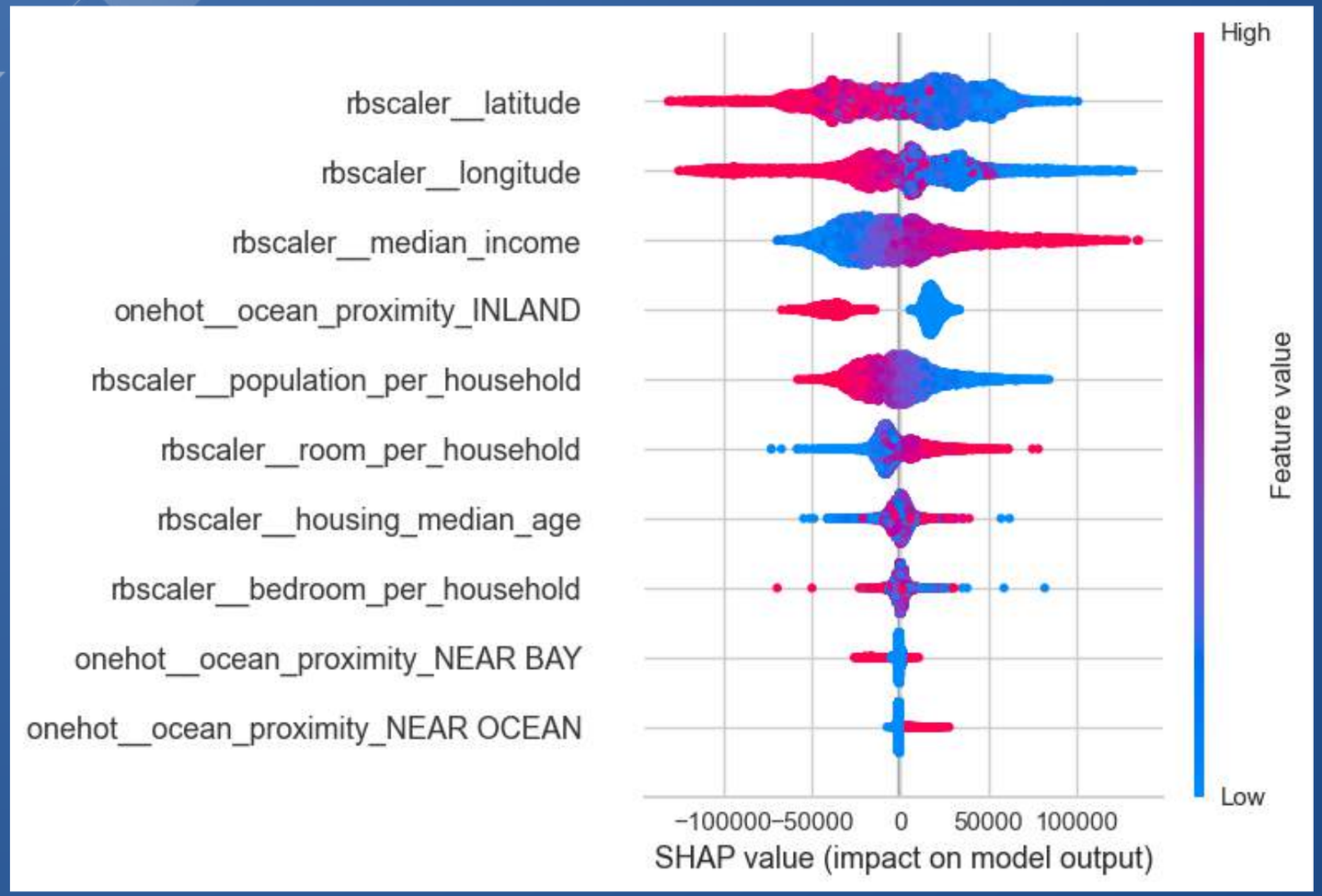
ACTUAL VS PREDICTION

- Distribusi titik dekat garis merah
- Pola linear yang jelas
- Sebaran residual lebih besar pada nilai tinggi
- Semakin padat titik-titik terhadap garis diagonal maka semakin baik model memprediksi



FEATURE IMPORTANCES

Ocean proximity INLAND merupakan feature yang paling dominan dengan kontribusi hampir 70% dari total importance. ini menunjukkan bahwa lokasi properti di area INLAND (tidak dekat laut) sangat berpengaruh terhadap harga rumah



MODEL INTERPRETATION

- Longitude dan latitude memiliki pengaruh besar terhadap penentuan harga
- Pendapatan berpengaruh terhadap penentuan harga rumah
- Ocean proximity INLAND memiliki harga lebih murah dibandingkan dengan rumah yang dekat dengan laut
- Kepadatan rumah tangga mengurangi penentuan harga rumah

COST ANALYST

MAPE

Ini menunjukkan bahwa rata-rata kesalahan prediksi model terhadap harga rumah hanya sekitar 15,89% dari nilai sebenarnya.

MAE

Prediksi harga rumah meleset sekitar $27 ribu dari harga sebenarnya

LIMITASI MODEL

- Fitur yang dapat digunakan
  - longitude
  - latitude
  - housing median age
  - median income
  - population per household
  - room per household
  - bedroom per household
- Model hanya dapat digunakan di wilayah california saja

CONCLUSION AND RECOMMENDATION

CONCLUSION

Model XG Boost menunjukkan performa prediktif yang cukup baik dengan MAPE 15.89%, RMSE $40.667 dan MAE $27.293

Fitur seperti median income, latitude dan ocean proximity memiliki pengaruh signifikan terhadap nilai rumah

RECOMMENDATION

Fokus terhadap fitur dengan pengaruh tinggi

Lakukan update dan pelatihan ulang model secara berkala

THANK YOU