Purwadhika

CALIFORNIA HOUSING PRICE

CAPSTONE PROJECT 3

Created By:
Banu Baskara Devar

Purwadhika

CONTENTS

BISNIS PROBLEM

Context
Sensus California 1990 dilakukan di tengah lonjakan transaksi jual beli rumah. Data sensus membantu memahami harga rumah di berbagai daerah, mendukung kebijakan perumahan, dan penilai dalam menentukan nilai jual yang tepat.

Problem Statement
Lonjakan penjualan rumah di California mendorong agensi properti untuk merekrut lebih banyak penilai properti, yang akan menimbulkan biaya tambahan dan potensi pengurangan staf jika permintaan menurun.

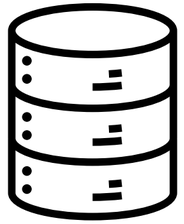
Goals
Pengembangan model machine learning untuk memprediksi harga median rumah di distrik-distrik California dengan tujuan mengurangi biaya dan mempercepat proses penilaian.

Metric Evaluation
MAPE sebagai metric utama, dengan tujuan meminimalkan persentase kesalahan prediksi.

DATA UNDERSTANDING

Dataset merupakan data sensus Callifornia 1990 yang dimana setiap baris merepresentasikan informasi terkait suatu distrik di California. Berikut adalah contoh data dalam dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
0	-119.79	36.73	52.0	112.0	28.0	193.0	40.0	1.9750	INLAND	47500.0
1	-122.21	37.77	43.0	1017.0	328.0	836.0	277.0	2.2604	NEAR BAY	100000.0
2	-118.04	33.87	17.0	2358.0	396.0	1387.0	364.0	6.2990	<1H OCEAN	285800.0
3	-118.28	34.06	17.0	2518.0	1196.0	3051.0	1000.0	1.7199	<1H OCEAN	175000.0
4	-119.81	36.73	50.0	772.0	194.0	606.0	167.0	2.2206	INLAND	59200.0

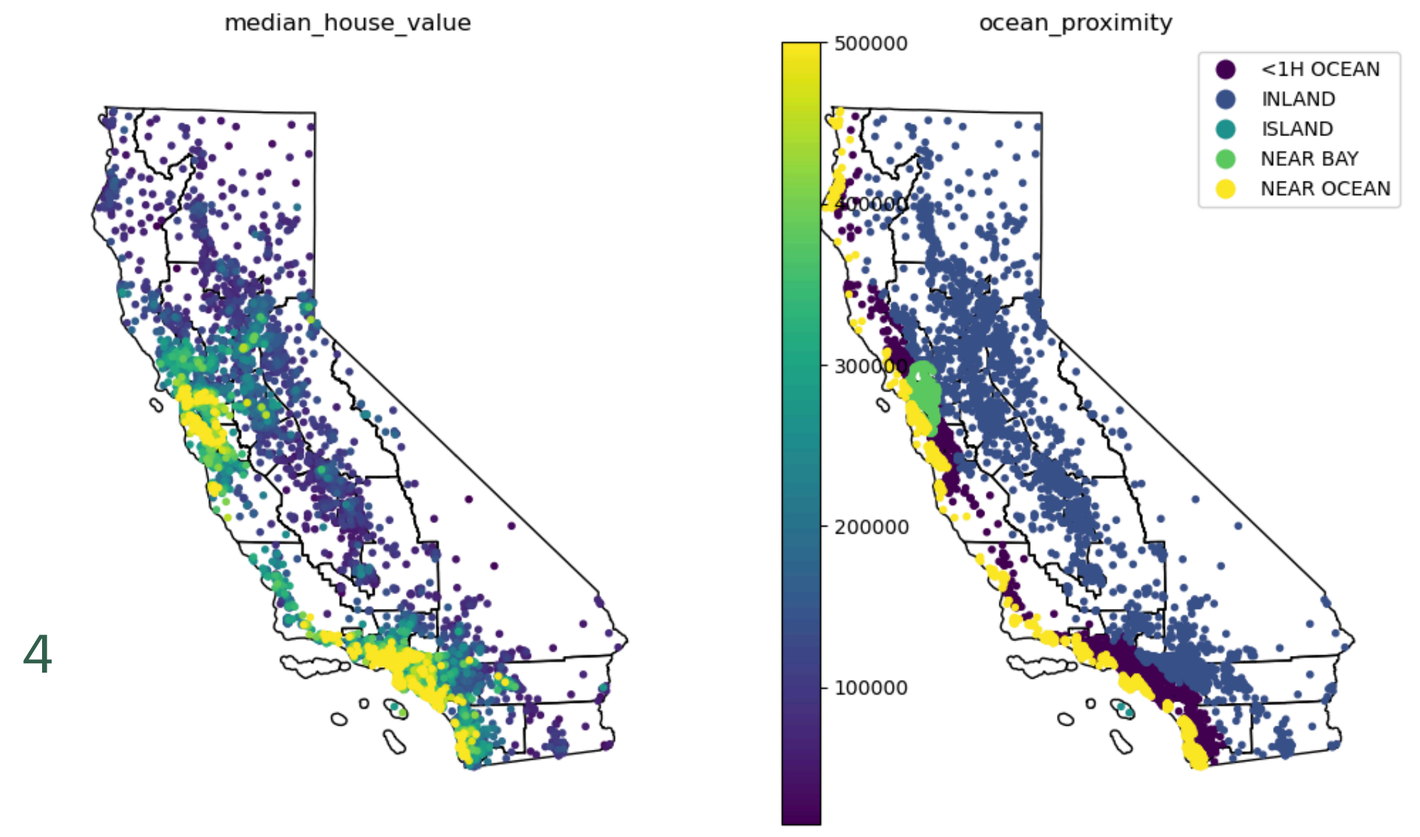


Dataset terdiri dari
14.477 baris dan 10 kolom

EXPLORATORY DATA ANALYSIS

Distributsi Data
Semua kolom yang ada memiliki data yang tidak terdistribusi normal

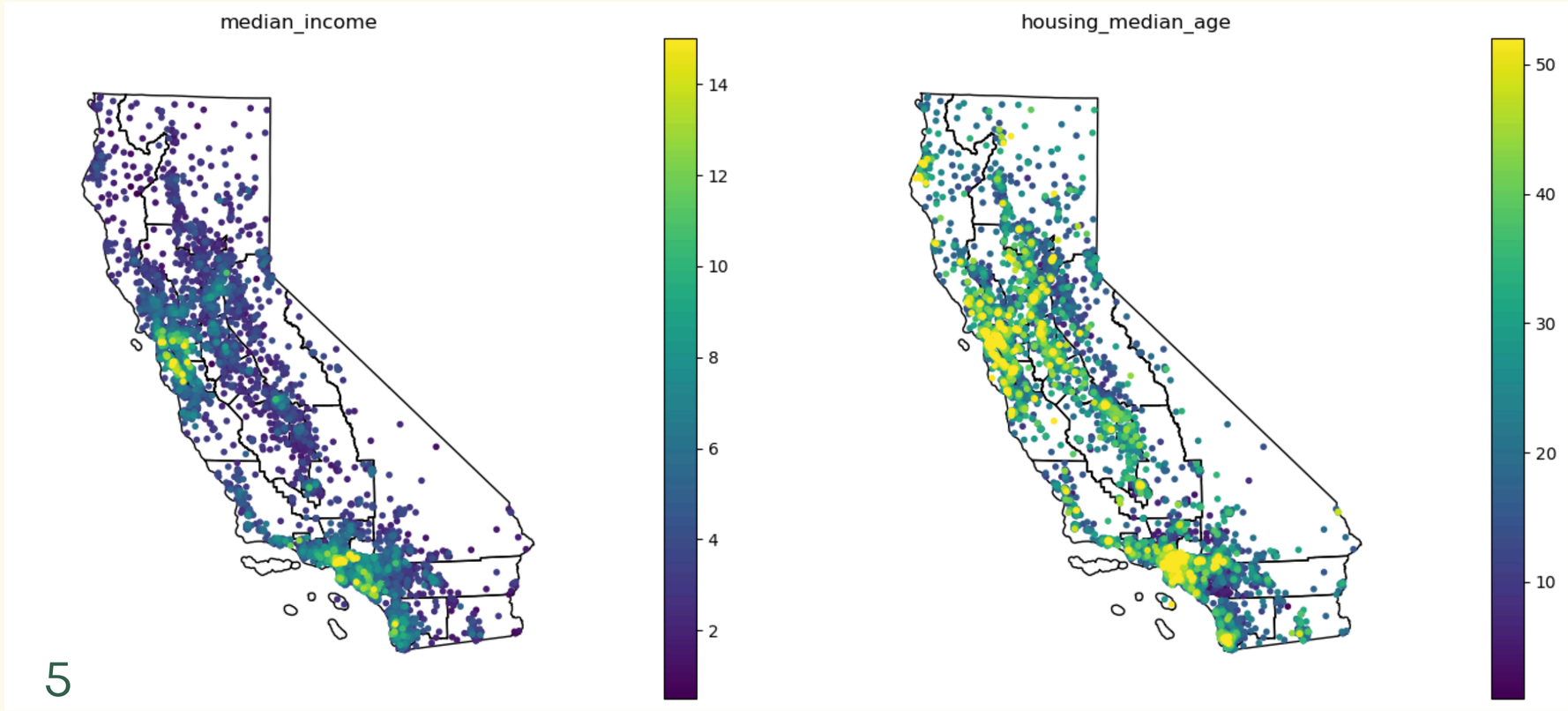
Geografi Plot
Visualisasi geografi untuk melihat persebaran data menggunakan latitude dan longitude



Median harga rumah cenderung semakin tinggi jika mendekati laut

EXPLORATORY DATA ANALYSIS

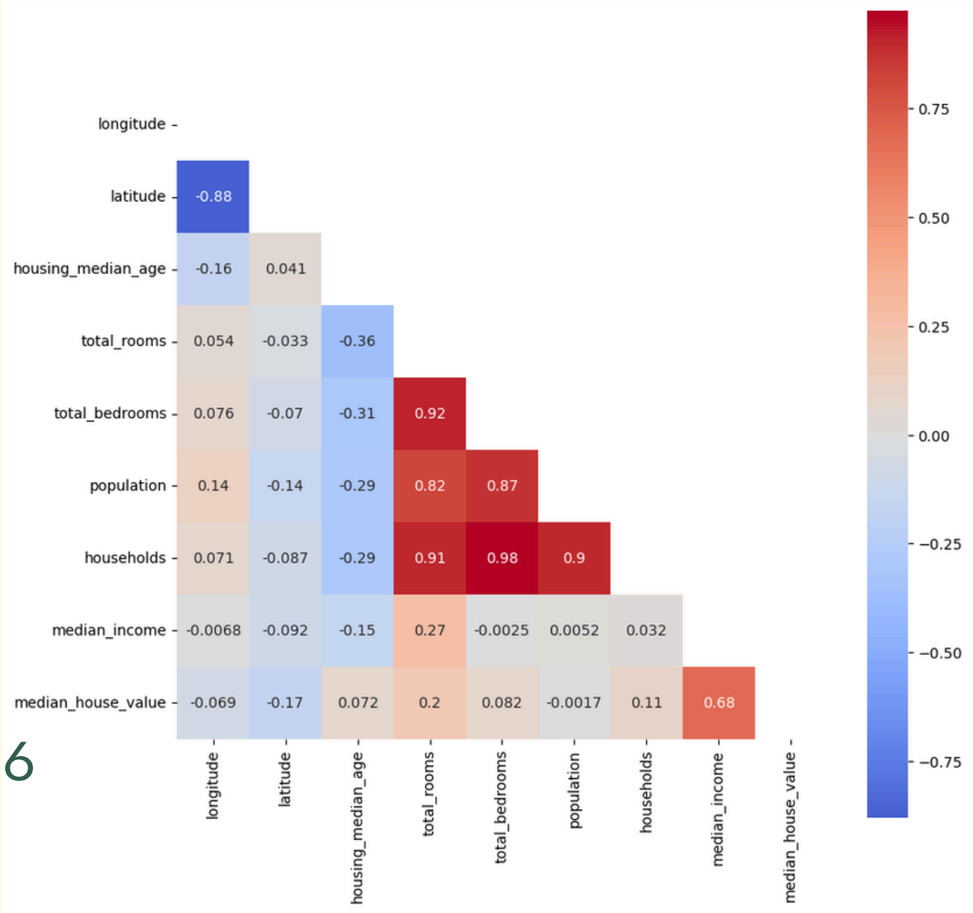
Geografi Plot
Visualisasi geografi untuk melihat persebaran data menggunakan latitude dan longitude



1. Median income cenderung semakin tinggi jika mendekati laut
Median umur rumah didominasi oleh rumah yang berumur lebih dari 40 tahun

EXPLORATORY DATA ANALYSIS

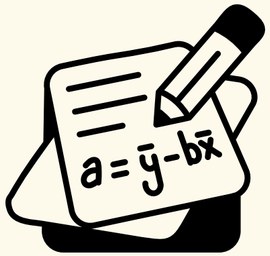
Correlation
Hubungan dari setiap fitur numerikal



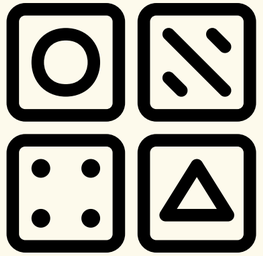
Fitur median_income memiliki korelasi paling kuat terhadap median house value

EXPLORATORY DATA ANALYSIS

Features vs Target



Numerical Features vs Target:
Hanya median_income yang memiliki pengaruh kuat terhadap target



Categorical Features vs Target:
ocean_proximity memiliki pengaruh terhadap median harga rumah

DATA PREPROCESSING

	Informasi	Handling
Missing Value	1%	drop data
Duplicated	0	tidak ada
Replace Value	housing_median_age dan ocean_proximity	housing_median_age =1 -> 2 ocean_proximity: INLAND & NOT INLAND
New Features	Membuat 3 Fitur Baru	population_per_household, room_per_household, bedroom_per_household
Remove Features	Membuang 3 Fitur Lama	total_rooms, total_bedrooms, population
Outliers	median_house_value 5 %	Membuang outlier
Collinearity	Terdapat 3 fitur	dibiarkan
Cardinallity	ocean_proximity = 2	tidak ada

MODELING

Pipeline Awal

	Informasi	Handling
Encoding	ocean_proximity	OneHotEncoder
Scaller	menyamakan skala fitur	RobustScaler
Model	Menggunakan Model Regresi	LinearRegression() KNeighborsRegressor() DecisionTreeRegressor() RandomForestRegressor() XGBRegressor() Ridge() Lasso() GradientBoostingRegressor() AdaBoostRegressor()

→

Model	Train MAPE	Test MAPE
XGBoostRegressor	0.1729	0.1813
RandomForestRegressor	0.1845	0.1897
GradientBoostRegressor	0.204	0.2098

MODELING - Hyperparameter Tuning

Model	Train MAPE	CV	n_iter
XGBoostRegressor	'model__max_depth': list(np.arange(3, 13)), 'model__learning_rate': [0.1, 0.01, 0.001], 'model__n_estimators': list(np.arange(80, 201, 10)), 'model__subsample': list(np.arange(8, 10)/10), 'model__gamma': [0.1,5] , 'model__colsample_bytree': list(np.arange(8, 10)/10)	5	100
RandomForestRegressor	‘model__n_estimators' : list(np.arange(80, 201, 10)), 'model__max_features' : ['sqrt','log2',None], 'model__max_depth' : range(3,21,3), 'model__min_samples_split': range(2, 21, 2), 'model__min_samples_leaf': range(1, 11, 2)	5	100
GradientBoostRegressor	'model__n_estimators' : range(10,101,10), 'model__max_features' : ['sqrt','log2',None], 'model__max_depth' : range(10,101,10), 'model__min_samples_split': range(2, 21, 2), 'model__min_samples_leaf': range(1, 21, 2)	5	100

MODELING

Tuning Summary

Model	Train MAPE	Test MAPE	MAPE After Tuning
XGBoostRegressor	0.1729	0.1813	0.1688
RandomForestRegressor	0.1845	0.1897	0.1716
GradientBoostRegressor	0.204	0.2098	0.1859

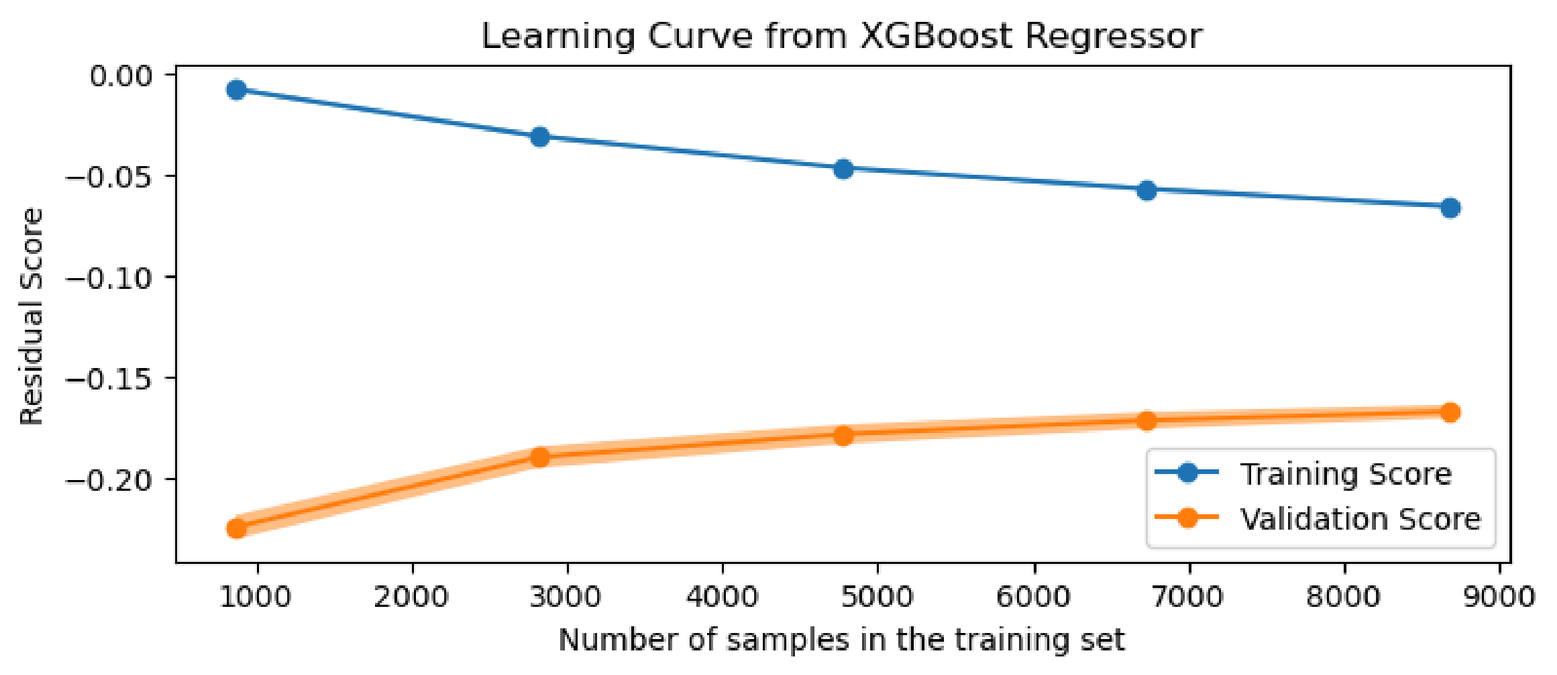
MODELING

Explanation About Best Algorithm

- XGBoost adalah sebuah perpustakaan gradient boosting yang dioptimalkan untuk efisiensi tinggi, fleksibilitas, dan portabilitas.
- Algoritma ini mengimplementasikan algoritma pembelajaran mesin di bawah kerangka Gradient Boosting. XGBoost menyediakan penguatan pohon paralel

Paramter	Value
model_subsample	0.8
model_n_estimators	150
model_max_depth	8
model_learning_rate	0.1
model_gamma	5
model_colsample_bytree	0.8

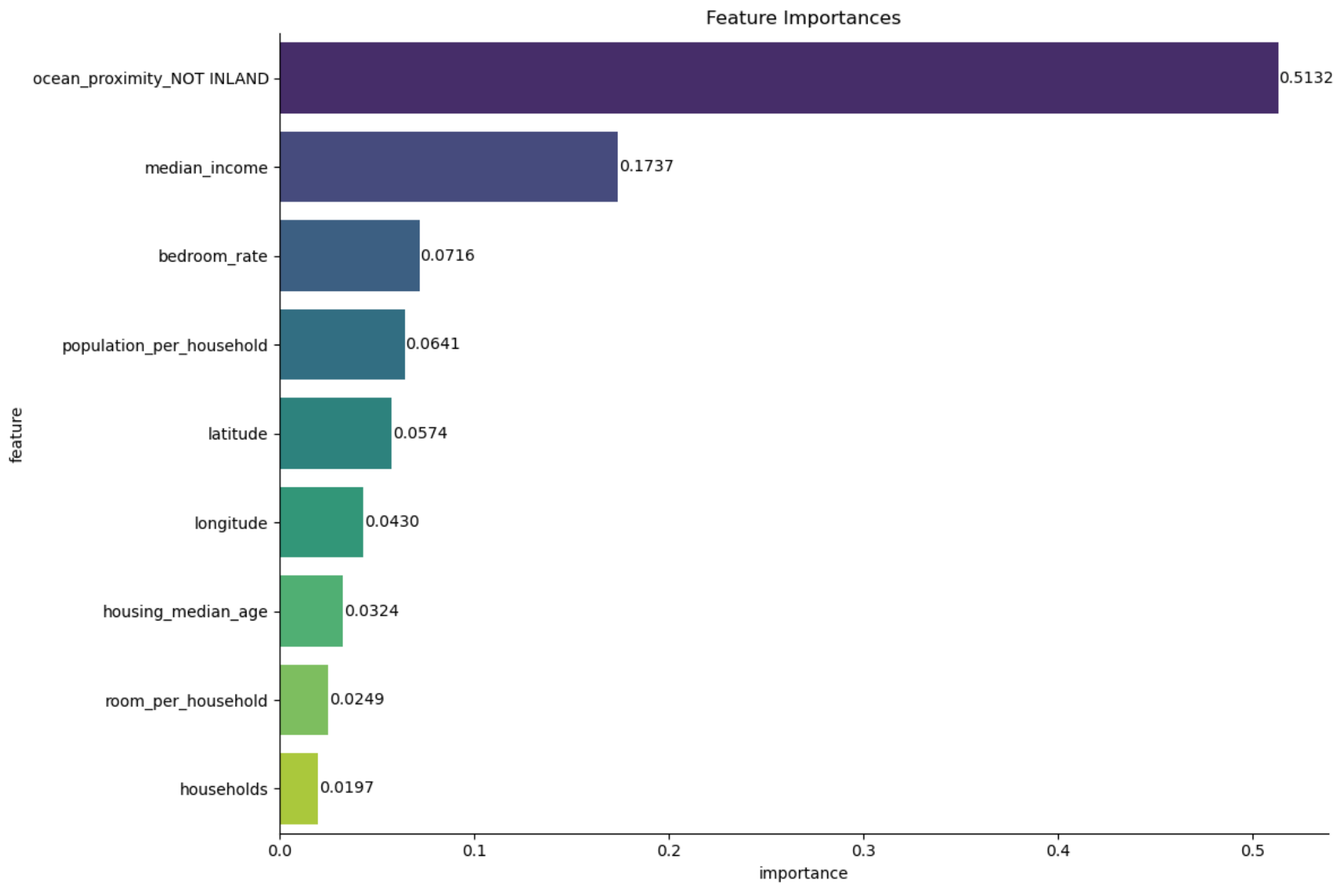
LEARNING CURVE



Kurva pembelajaran menunjukkan bahwa XGBoost Regressor mampu belajar cukup baik dari data pelatihan.

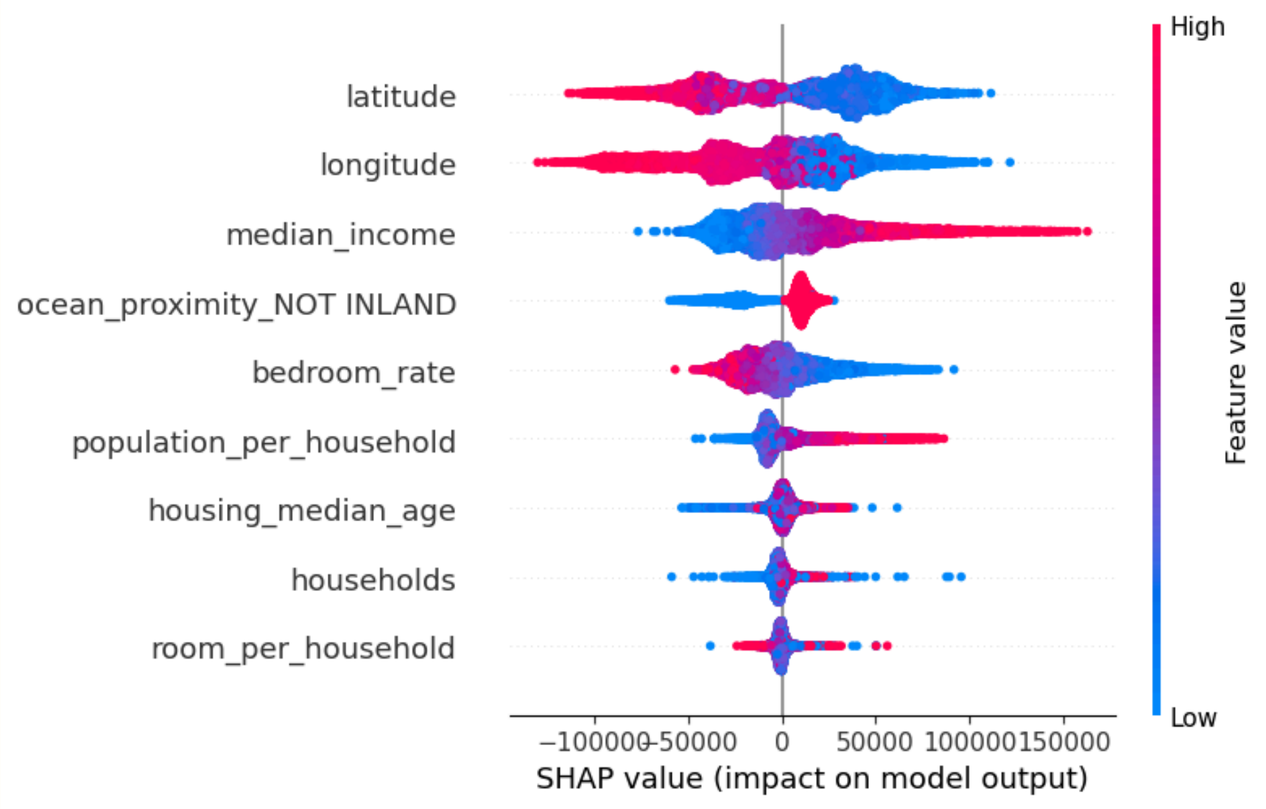
Perbedaan kecil antara skor pelatihan dan skor validasi menunjukkan bahwa model tidak terlalu cocok dengan data pelatihan, yang merupakan indikasi yang baik untuk generalisasi model ke data baru.

EVALUASI-FEATURES IMPORTANCES



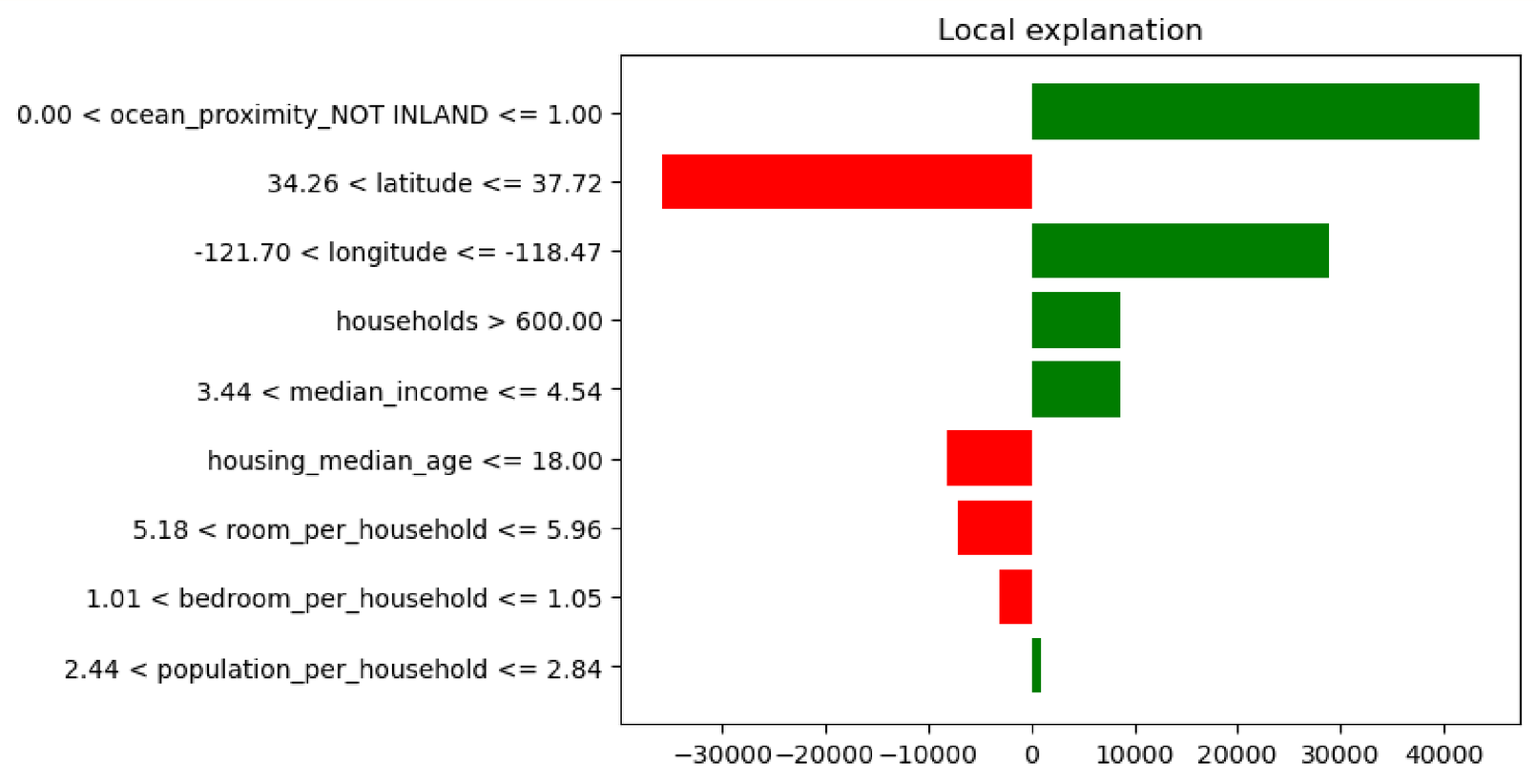
ocean_proximity_NOT INLAND memiliki pengaruh 50% dari model ini

EVALUASI-SHAP



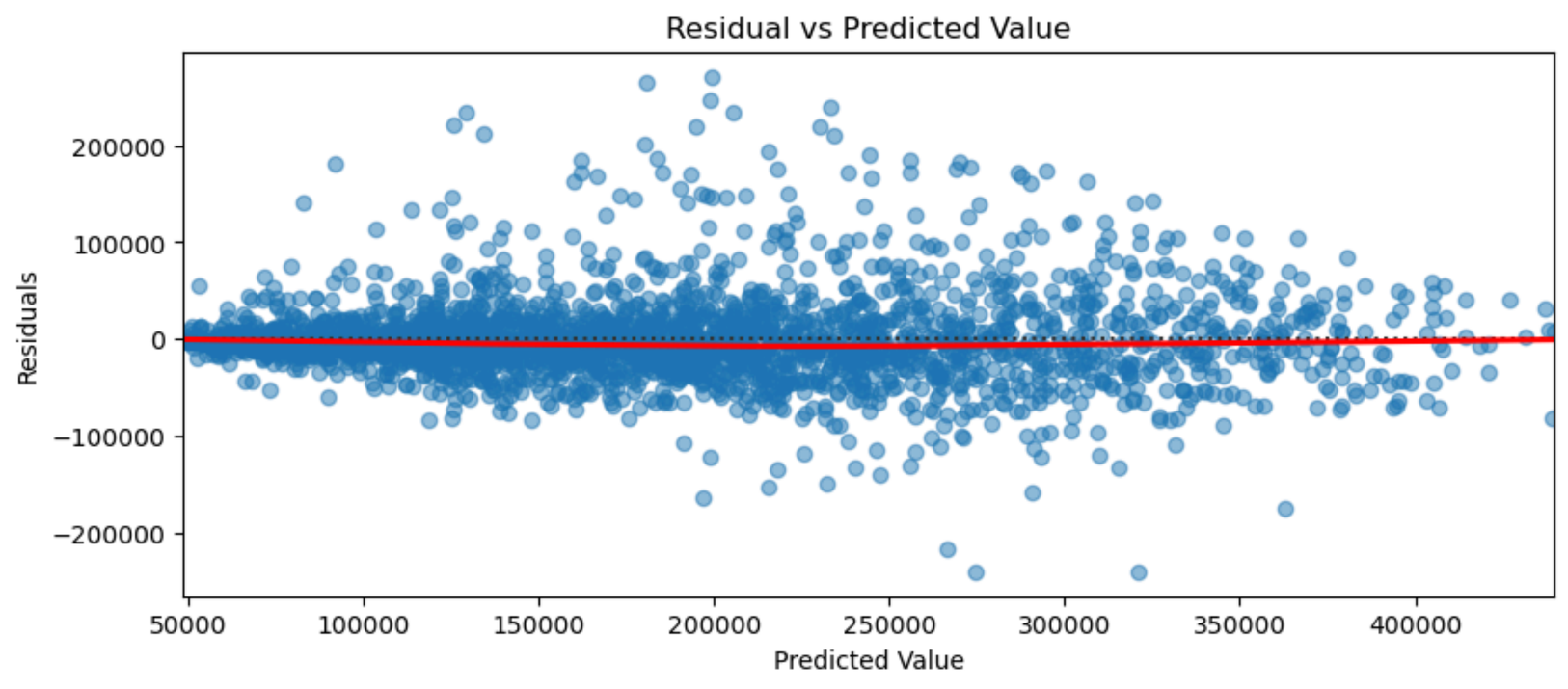
letak geografis menjadi fitur paling penting untuk menentukan median house value

EVALUASI-LIME



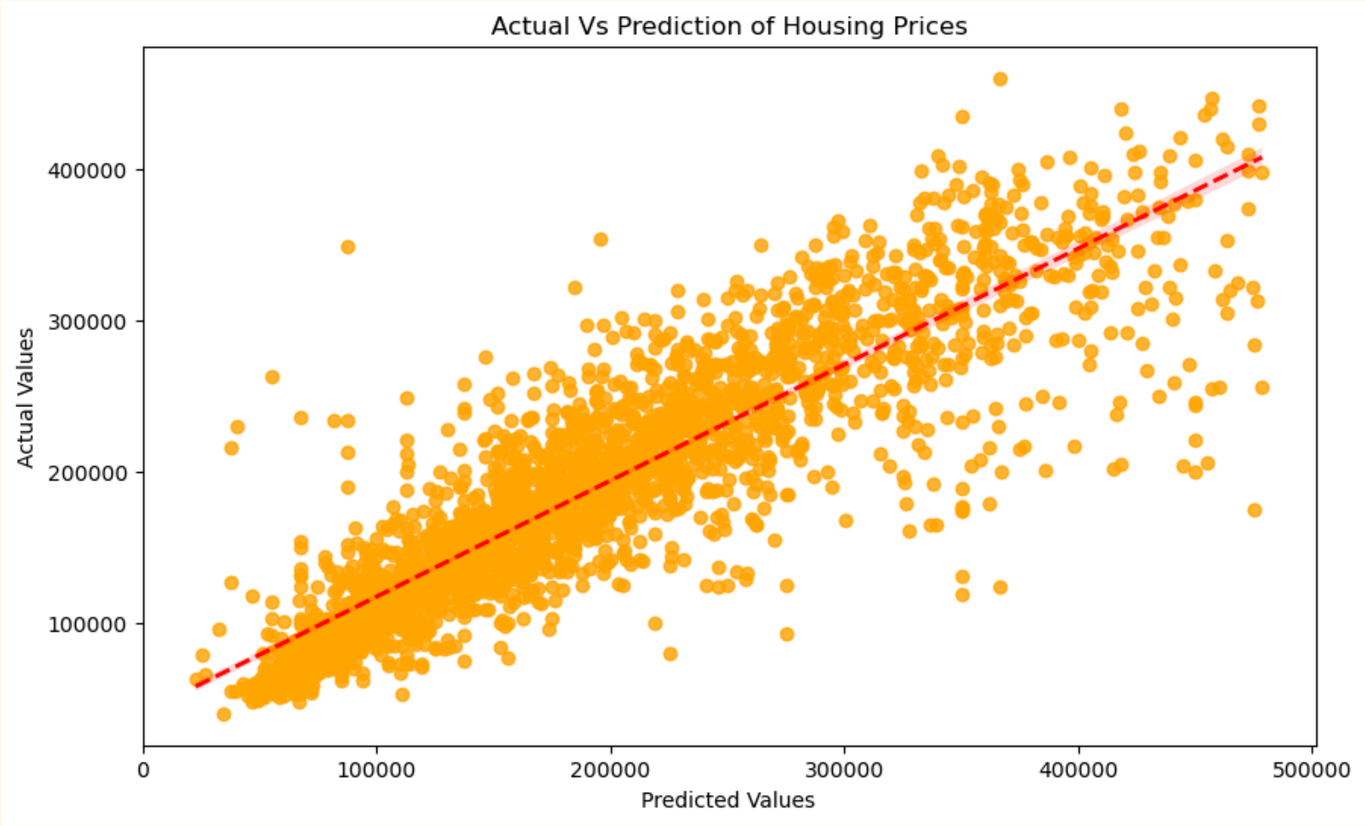
ocean_proxmity dekat laut membuat median house value menjadi mahal

RESIDUAL PLOT



Plot residual menunjukkan bahwa kesalahan prediksi tidak tersebar secara acak di sekitar garis horizontal. Ada pola lengkung yang menunjukkan bahwa kesalahan model tergantung pada nilai prediksi rumah yang diprediksi.

ACTUAL VS PREDICTED PROBABILITY



Pola ini menunjukkan bahwa model cenderung melakukan kesalahan yang lebih besar dalam memprediksi harga rumah yang lebih tinggi

COST BENEFIT ANALYSIS

Pada proses penghitungan biaya mengguanakan perhitungan 400 rumah yang didapatkan berdasarkan median nilai household dengan asumsi bahwa setiap household memiliki 1 rumah.

Jumlah Rumah/Bulan	400	400
Rumah yang Dapat Dinilai per Bulan	3	N/A
Biaya per Penilaian Rumah	USD 1.000	USD 100
Total Biaya Penilaian per Bulan	USD 3.000	USD 40.000
Jumlah Home Appraiser yang Dibutuhkan	133	N/A
Total Biaya per Bulan untuk Home Appraisal	USD 400.000	USD 150.000
Pendapatan dari Jasa Penilaian Rumah	USD 480.000	USD 480.000
Keuntungan		
- Total Keuntungan per Bulan	USD 80.000	USD 330.000
- Margin Keuntungan	16.67%	68.75%

COST BENEFIT ANALYSIS

	Home Appraisal	With Machine Learning
Keuntungan		
- Akurasi	✓	✗
- Biaya	✗	✓
- Waktu	✗	✓
- Kapasitas	✗	✓
- Konsistensi	✗	✓
- Kepercayaan	✓	✗
- Penanganan Kasus Kompleks	✓	✗
- Skalabilitas	✗	✓
- Keterbatasan Model	✗	✓
- Memerlukan Banyak Data	✗	✓

Secara keseluruahan lebih mengutungkan menggunakan machine learning:

- Biaya lebih murah
- Pengerjaan lebih cepat

LIMITASI MODEL

Batasan Model	Rentang Nilai
Longitude dan Latitude	Longitude: -124.35 hingga -114.31, Latitude: 32.54 hingga 41.95
Housing Median Age	2 hingga 52 tahun
Households	2 hingga 6082 rumah tangga
Median Income	0.5 hingga 15
Population per Household	0.75 hingga 599
Room per Household	0.85 hingga 132.5
Bedroom per Household	0.33 hingga 34.1

KESIMPULAN

Pemanfaatan machine learning untuk memprediksi median house value di California tahun 1990 memberikan keuntungan signifikan. Analisis cost-benefit menunjukkan pengurangan biaya dan peningkatan efisiensi dalam proses penilaian rumah. Meskipun demikian, implementasi machine learning perlu mempertimbangkan keterbatasannya.

Model XGBoostRegressor berhasil mencapai MAPE 16% dalam memprediksi median house value, menunjukkan tingkat kesalahan sebesar 16%. Fitur yang paling berpengaruh adalah 'ocean_proximity_NOT INLAND' dan 'median_income'.

REKOMENDASI

Model
Pemodelan yang dilakukan saat ini telah menghasilkan features importance yang didapatkan dari algoritma XGBoost dimana hal tersebut dapat menjadi sebuah insight untuk pengembang selanjutnya untuk menghapus fitur yang tidak relevan pada target untuk menghasilkan model yang baik.

Bisnis
Pada perusahaan sebaiknya mencoba mengumpulkan data baru yang mungkin saja bisa menjadi sebuah fitur penting untuk memprediksi harga median rumah, dan juga tidak perlu mengambil data yang tidak menjadi fitur penting dalam model saat ini untuk kedepannya.