

Prediksi Harga Rumah di California: Pendekatan Regressi Dengan Pengembangan Model XGBoost

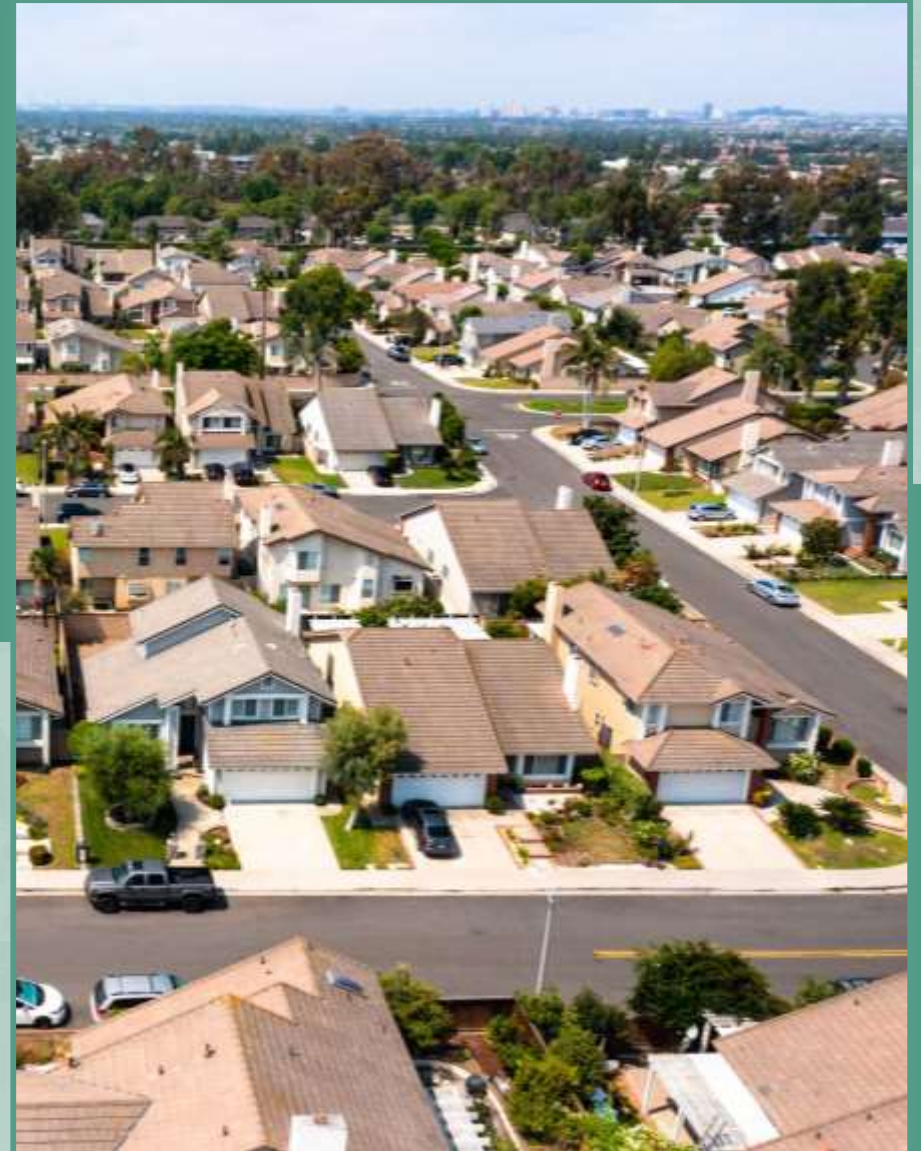


By : Rais Nugroho



Overview

- Business Problem Understanding
- Analytical Approach
- Metrics Evaluation
- Data Preprocessing
- Transformers and Pipeline
- Algorithms Models
- Result and Conclusion
- Recommendation
- Profit Estimation



Business Problem Understanding

➡ **Context**

**Problem
Statement** ⬅

➡ **Goals** ⬅



Context

Dataset California Housing berisi informasi mengenai properti perumahan di California dan digunakan untuk menganalisis faktor-faktor yang mempengaruhi nilai properti. Data ini meliputi atribut-atribut seperti koordinat geografis, usia rumah, jumlah ruangan, penduduk, dan pendapatan median. Dataset ini sering digunakan untuk memodelkan dan memprediksi nilai rumah berdasarkan berbagai fitur, serta untuk mengevaluasi faktor-faktor yang mempengaruhi harga properti di wilayah tersebut, menjadikannya cocok untuk mengajarkan dasar-dasar machine learning.



Problem Statement

Banyak developer dan individu yang sulit untuk mendapatkan harga rumah terbaik, dikarenakan minimnya informasi dan tidak adanya perhitungan dasar untuk menentukan harga rumah terutama di California. Tujuan dari analisis ini adalah untuk memahami faktor-faktor apa saja yang mempengaruhi nilai rumah di California dan bagaimana berbagai features mempengaruhi harga rumah. Dengan informasi ini, kita dapat membangun model yang memprediksi nilai rumah secara akurat.

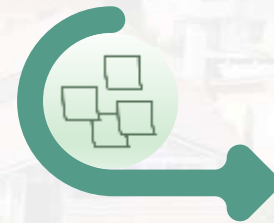


Goals

Tujuan utama kami adalah untuk memprediksi harga rumah dengan akurasi tinggi, sehingga dapat menyediakan informasi yang diperlukan untuk meningkatkan efisiensi dalam pengambilan keputusan investasi properti. Dengan demikian, kami berupaya membantu menghindari potensi kerugian yang dapat terjadi dalam proses investasi.



**Identifikasi Faktor-Faktor
yang Mempengaruhi Nilai**



**Pengembangan Model
Prediktif**



**Penerapan Insights Untuk
Pengambilan Keputusan**



Analytical Approach

Kami akan melakukan analisis regresi dengan variabel target berupa harga rumah dan fitur-fitur yang meliputi `housing_median_age`, `total_rooms`, `population`, `total_bedrooms`, dan lainnya. Tahapan pertama dalam analisis ini adalah memahami data beserta atribut yang ada. Selanjutnya, kami akan membersihkan data dari nilai yang hilang (missing values), duplikasi, dan anomali lainnya. tahapan selanjutnya, data diproses untuk modeling dan tuning agar mendapat akurasi yang baik.

Regression Taks



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
...
14443	-121.26	38.27	20.0	1314.0	229.0	712.0	219.0	4.4125	INLAND	144600.0
14444	-120.89	37.48	27.0	1118.0	195.0	647.0	209.0	2.9135	INLAND	159400.0
14445	-121.90	36.58	31.0	1431.0	NaN	704.0	393.0	3.1977	NEAR OCEAN	289300.0
14446	-117.93	33.62	34.0	2125.0	498.0	1052.0	468.0	5.6315	<1H OCEAN	484600.0
14447	-115.56	32.80	15.0	1171.0	328.0	1024.0	298.0	1.3882	INLAND	69400.0

14,448 Row, 10 Columns



Metrics Evaluation

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \\ \text{MAPE} &= \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\% \end{aligned}$$

Setelah melakukan evaluasi terhadap berbagai model, kami memutuskan untuk menggunakan Extra Gradient Boosting Regressor sebagai model utama, menggunakan model XGBoost baik digunakan pada data ini. Metrik utama yang kami pilih adalah Root Mean Square Error (RMSE) dan Mean Absolute Percentage Error (MAPE). Model dan metrics ini tahan terhadap outliers sehingga penggunaan untuk data california housing yang memiliki banyak outliers dianggap tepat.



Data Preprocessing



Data Cleaning

Melakukan pengecekan data duplikat, menghapus data yang hilang dan menghapus features yang memiliki VIF tinggi.



Data Outliers

Membuat data menjadi 2 versi, data yang memiliki outliers dan data yang sudah bersih dari outliers.



Variable

Pada pengujian model kali ini menggunakan 1 variable dependent dan 7 variable independent.

Transformers & Pipeline



Onehot Encoder

Dilakukan untuk merubah data kategorik



Standar Scaler

Dilakukan untuk menormalkan distribusi



MinMax Scaler

Dilakukan untuk membuat rentang 0 s/d 1



Robust Scaler

Dilakukan untuk normalisasi outliers



Cross Validation

Membagi data pada saat melakukan modeling





ALGORITHMS MODELS

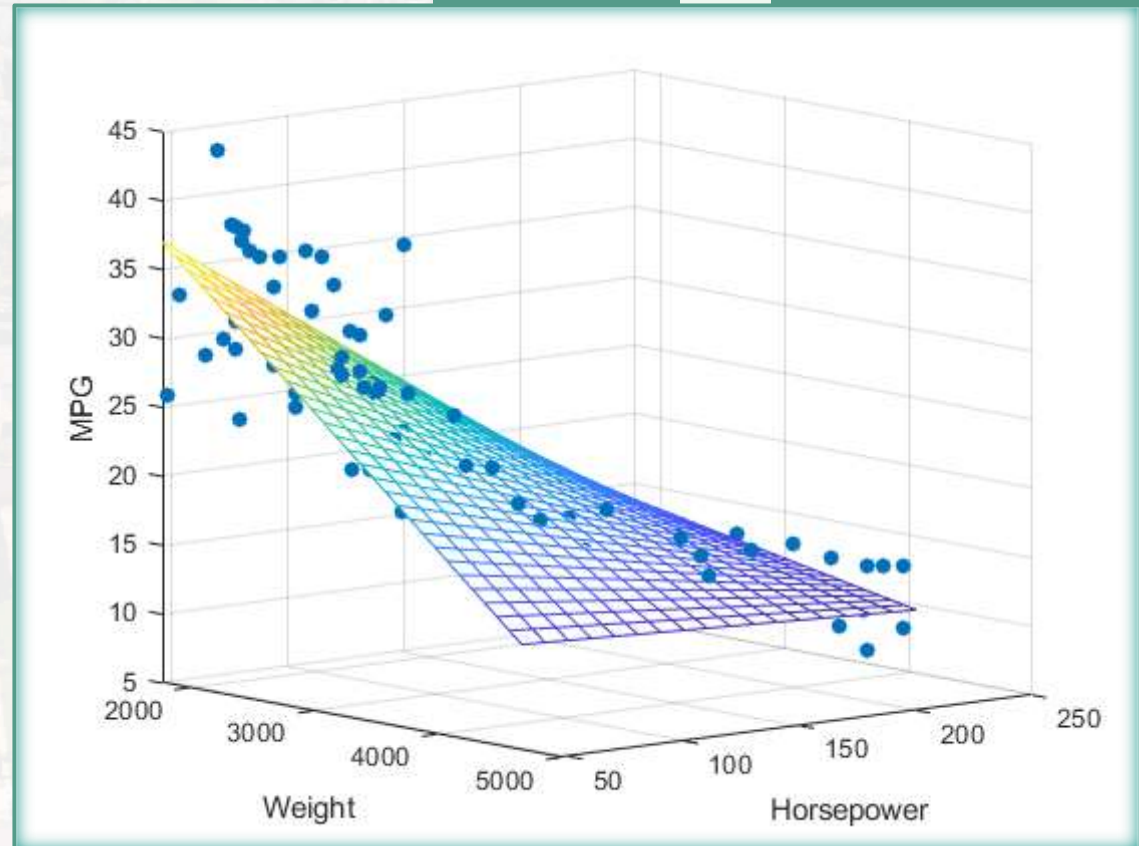


➤ Multiple Linear Regression

Multiple Linear Regression adalah ekstensi dari regresi linier sederhana yang melibatkan beberapa variabel independen. Model ini tergolong parametrik dan memiliki sejumlah asumsi dasar. Salah satu asumsi utama adalah bahwa hubungan antara variabel dependen (y) dan variabel independen (X_1, X_2, \dots, X_n) bersifat linear. Persamaan matematis dari multiple linear regression dapat dituliskan sebagai:

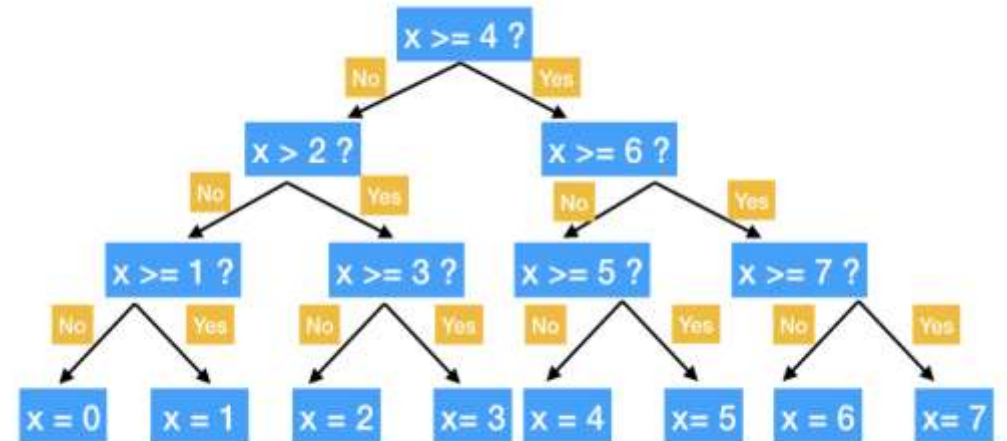
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

di mana β_0 adalah intercept, $\beta_1, \beta_2, \dots, \beta_n$ adalah koefisien yang terkait dengan masing-masing variabel independen, dan ε adalah error. Model ini bertujuan untuk meminimalkan jumlah kuadrat residu (sum of squared residuals), yang mencerminkan perbedaan antara nilai observasi dan nilai yang diprediksi.



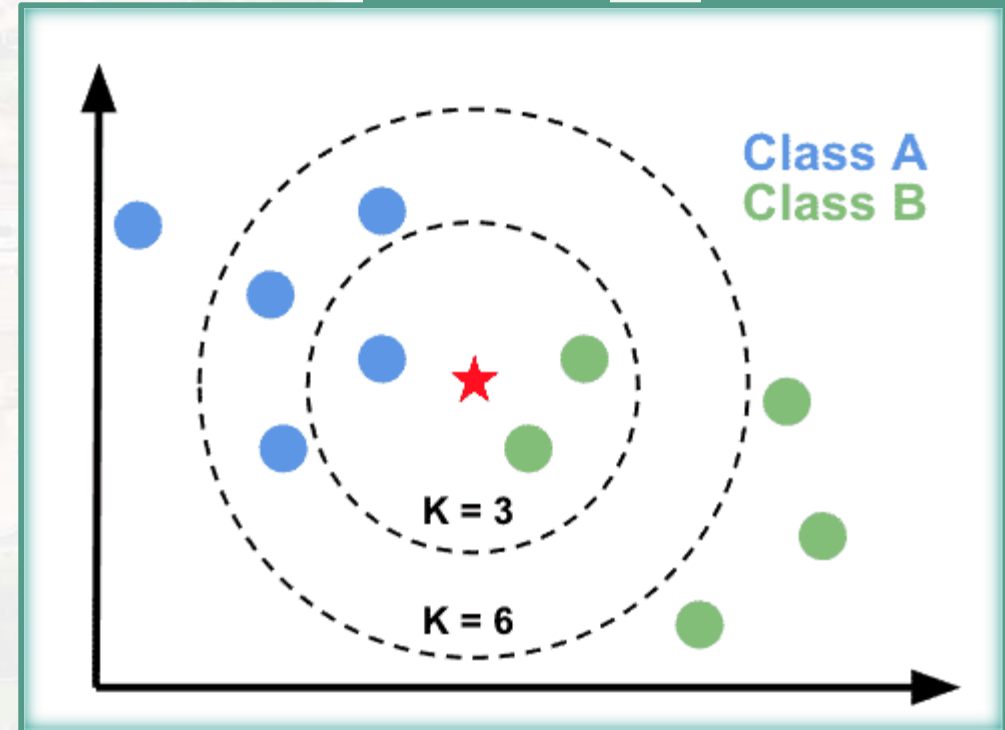
➤ Decision Tree

Decision Tree Regressor adalah model non-parametrik yang tidak terikat oleh asumsi tertentu, berfungsi untuk membuat prediksi dalam bentuk struktur pohon keputusan. Proses ini melibatkan pembagian data menjadi subkelompok yang lebih kecil dan lebih homogen berdasarkan fitur paling informatif, yang menghasilkan pengurangan maksimum dalam variasi (atau entropi, Gini impurity, dan sejenisnya dalam konteks klasifikasi). Setiap node dalam pohon mewakili fitur dalam dataset, sementara setiap cabang mencerminkan keputusan yang membagi data menjadi grup yang lebih homogen terkait variabel respon. Proses pemecahan ini berlanjut secara rekursif hingga kriteria penghentian terpenuhi, seperti kedalaman maksimum yang dicapai atau peningkatan minimum dalam homogenitas setelah pemecahan.



➤ K Nearest Neighbord

K Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin non-parametrik yang digunakan untuk klasifikasi dan regresi. Model ini berfungsi dengan mencari K titik data terdekat dalam dataset untuk memprediksi kelas atau nilai dari titik data baru. Proses ini melibatkan pengukuran jarak antara titik data yang tidak dikenal dan semua titik dalam dataset menggunakan metrik jarak seperti Euclidean atau Manhattan. KNN kemudian memilih K tetangga terdekat, dan keputusan diambil berdasarkan mayoritas kelas (untuk klasifikasi) atau rata-rata nilai (untuk regresi) dari tetangga tersebut.



➤ Voting Regressor

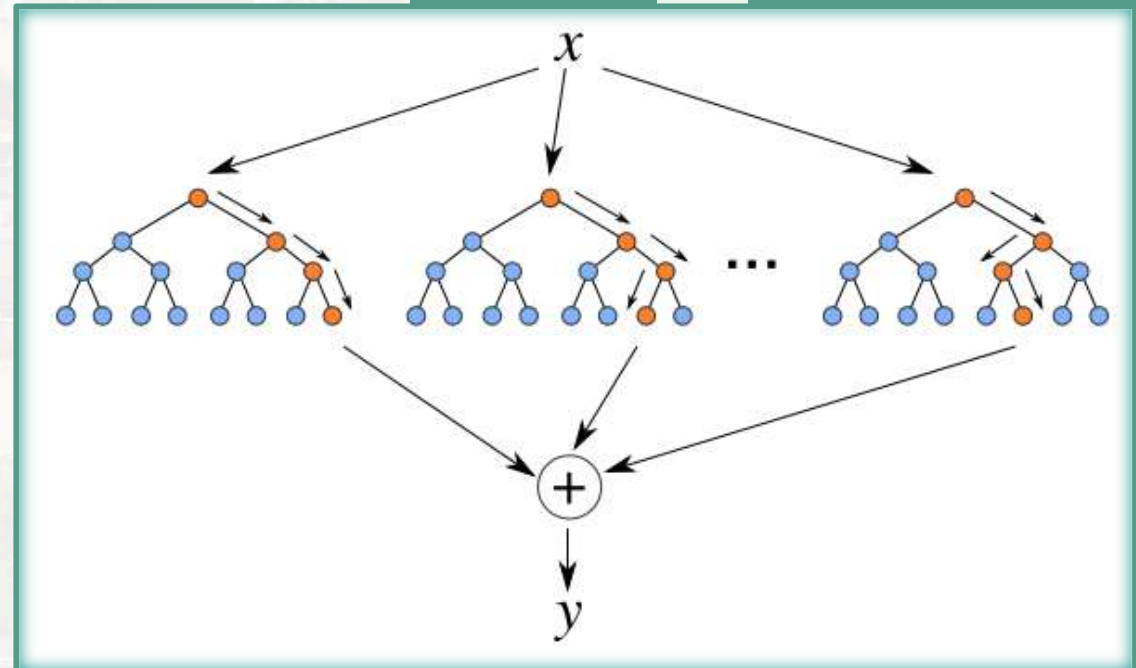
Voting Regressor adalah ensemble machine learning tipe various type, di mana kita dapat menggunakan beberapa tipe algoritma yang tidak setipe. Metode ini menggabungkan prediksi dari beberapa model regresi yang berbeda untuk meningkatkan keakuratan prediksi secara keseluruhan. Voting Regressor merata-ratakan hasil prediksi dari tiap model. Ini menghasilkan output akhir yang merupakan rata-rata prediksi dari semua model yang terlibat dalam training. Averaging membantu mengurangi varians dan bias yang mungkin dihasilkan oleh model individu, sehingga meningkatkan stabilitas prediksi keseluruhan. Voting Regressor dipilih karena seringkali menghasilkan prediksi yang lebih akurat daripada model individu.

➤ Stacking Regressor

Stacking Regressor adalah teknik ensemble dalam machine learning yang menggabungkan prediksi dari beberapa model regresi untuk meningkatkan akurasi. Dalam metode ini, model-model dasar (base learners) dilatih secara terpisah, dan kemudian prediksi mereka digunakan sebagai input untuk model meta (meta-learner) yang menghasilkan prediksi akhir. Dengan cara ini, stacking dapat menangkap kompleksitas dan pola yang mungkin tidak terdeteksi oleh model individual, seringkali menghasilkan performa yang lebih baik dibandingkan metode tunggal.

➤ Bagging Regressor With Random Forest

Bagging Regressor adalah metode untuk meningkatkan akurasi prediksi dalam model machine learning dengan cara melatih beberapa model secara terpisah pada bagian data yang acak dan kemudian menggabungkan prediksinya. Metode ini menggunakan sampling dengan pengembalian dari dataset utama untuk menghasilkan sampel-sampel baru, dan setiap model dilatih pada salah satu dari sampel tersebut. Setelah itu, hasil dari semua model diambil rata-ratanya untuk mendapatkan prediksi akhir. Teknik ini sangat efektif untuk mengurangi kesalahan prediksi yang sering terjadi karena overfitting, yaitu ketika model terlalu cocok dengan data latih hingga tidak efektif pada data baru.



➤ Gradient Boosting

Gradient Boosting adalah teknik machine learning yang digunakan untuk membangun model prediksi dengan menggabungkan beberapa model lemah (weak learners), biasanya decision trees, dalam urutan tertentu. Metode ini bekerja dengan cara membangun model secara bertahap, di mana setiap model baru dilatih untuk memperbaiki kesalahan dari model sebelumnya.

➤ Extreme Gradient Boosting

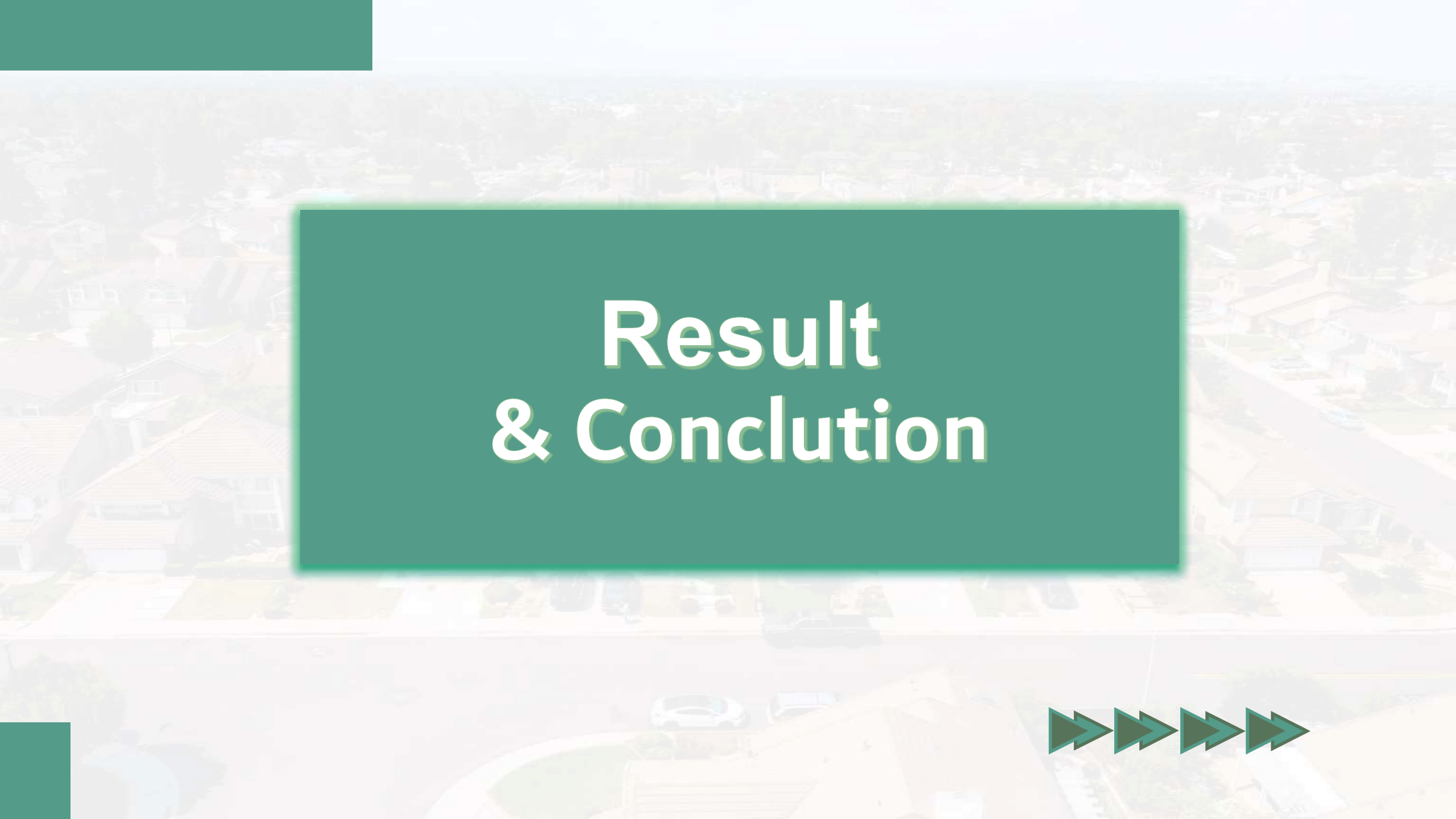
XGBoost (Extreme Gradient Boosting) adalah implementasi populer dari algoritma gradient boosting yang dirancang untuk efisiensi dan kinerja tinggi. XGBoost menggabungkan keunggulan gradient boosting dengan teknik optimasi yang canggih, termasuk regularisasi untuk mengurangi overfitting dan kemampuan untuk menangani missing values. Secara khusus, XGBoost menggunakan pendekatan paralelisasi dalam proses pelatihan, yang memungkinkan model untuk dilatih lebih cepat dibandingkan dengan implementasi gradient boosting tradisional.

➤ Ridge Regression

Ridge Regression (atau Tikhonov regularization) adalah metode parametrik dan digunakan untuk menganalisis data regresi dengan multikolinearitas, yaitu situasi di mana variabel independen sangat berkorelasi. Ini adalah perluasan dari regresi linier yang mencakup istilah regularisasi L2. Regularisasi adalah teknik yang digunakan untuk mencegah overfitting dengan menambahkan penalti pada ukuran koefisien dalam cost function regresi. Dalam ridge regression, cost function diubah dengan menambahkan istilah penalti yang proporsional dengan kuadrat besar dari magnitude koefisien (norma L2).

➤ Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) adalah metode parametrik dan merupakan jenis regresi yang melakukan regularisasi L1, di mana sebuah penalti yang setara dengan jumlah nilai absolut dari koefisien diterapkan. Tujuan dari Lasso adalah untuk memperoleh subset dari prediktor yang meminimalkan kesalahan prediksi dengan mengonstruksi model yang lebih interpretatif dengan mengurangi jumlah variabel.



Result & Conclusion



... Hasil model menunjukkan bahwa algoritma yang diterapkan pada rules based model tanpa machine learning menghasilkan kinerja yang lebih buruk dibandingkan dengan based model dengan machine learning.

Temuan ini mengindikasikan bahwa penggunaan rules model dapat memiliki pengaruh buruk terhadap kualitas model selama proses pelatihan dan pengujian.

	model	test_score_RMSE	test_score_MAE	test_score_MAPE	test_score_R2
0	Rule Based Non ML	138067.92	111301.49	0.77	-0.43



Rules Based
Tanpa ML

	model	test_score_RMSE	test_score_MAE	test_score_MAPE	test_score_R2
7	XGBoost	55933.80	38946.18	0.22	0.77



Based Model ML



... Hasil model menunjukkan bahwa algoritma yang diterapkan pada data tanpa outlier menghasilkan kinerja yang lebih buruk dibandingkan dengan data yang mengandung outlier.

Temuan ini mengindikasikan bahwa penghapusan outlier dapat memiliki pengaruh signifikan terhadap kualitas model selama proses pelatihan dan pengujian.

	model	test_score_RMSE	test_score_MAE	test_score_MAPE	test_score_R2
7	XGBoost	62325.22	43562.12	0.26	0.66



Tanpa Outliers

	model	test_score_RMSE	test_score_MAE	test_score_MAPE	test_score_R2
7	XGBoost	55933.80	38946.18	0.22	0.77



Dengan Outliers



... Hyperparameter tuning pada model Extreme Gradient Boost (XGB) ke data tanpa outlier menghasilkan kinerja tertinggi.

Model XGB memiliki score terbaik dari semua model, score ini di dapat dari beberapa experiment. Untuk menurunkan error agar lebih baik, maka dilakukan hyperparameter tuning, hal ini terbukti bahwa melakukan hyperparameter tuning menurunkan score error.

	model	Score RMSE	Score MAE	Score MAPE	Score R2
7	XGBoost	54160.90	37620.96	0.21	0.78



Setelah di Tuning

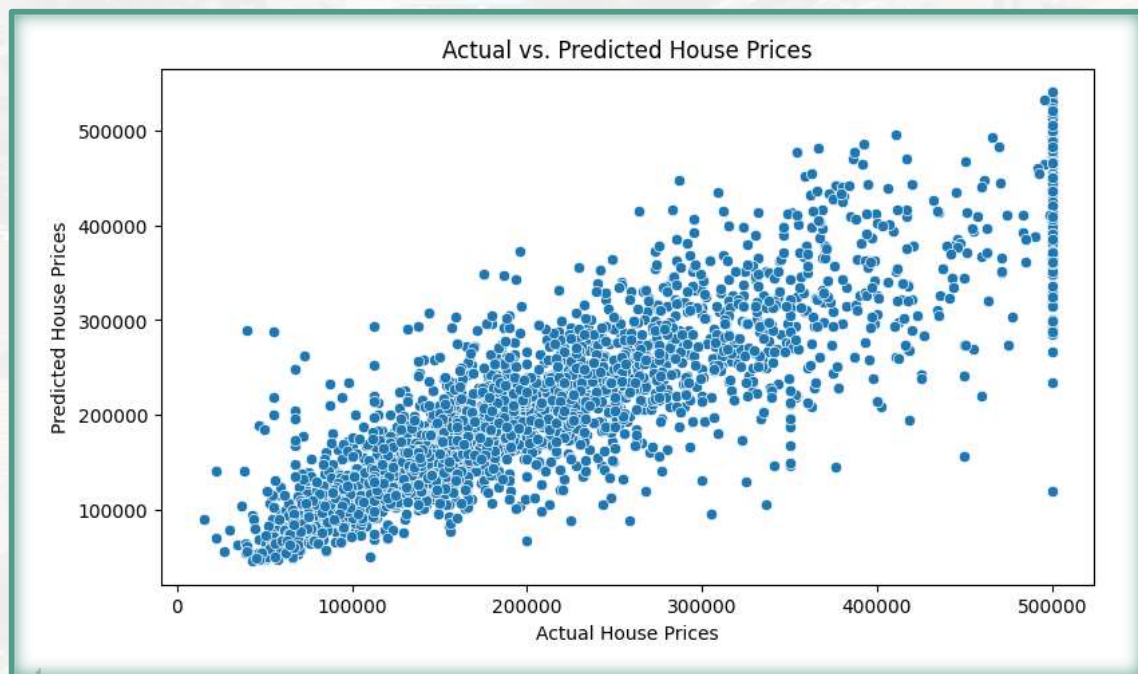
	model	test_score_RMSE	test_score_MAE	test_score_MAPE	test_score_R2
7	XGBoost	55933.80	38946.18	0.22	0.77



Tanpa Tuning



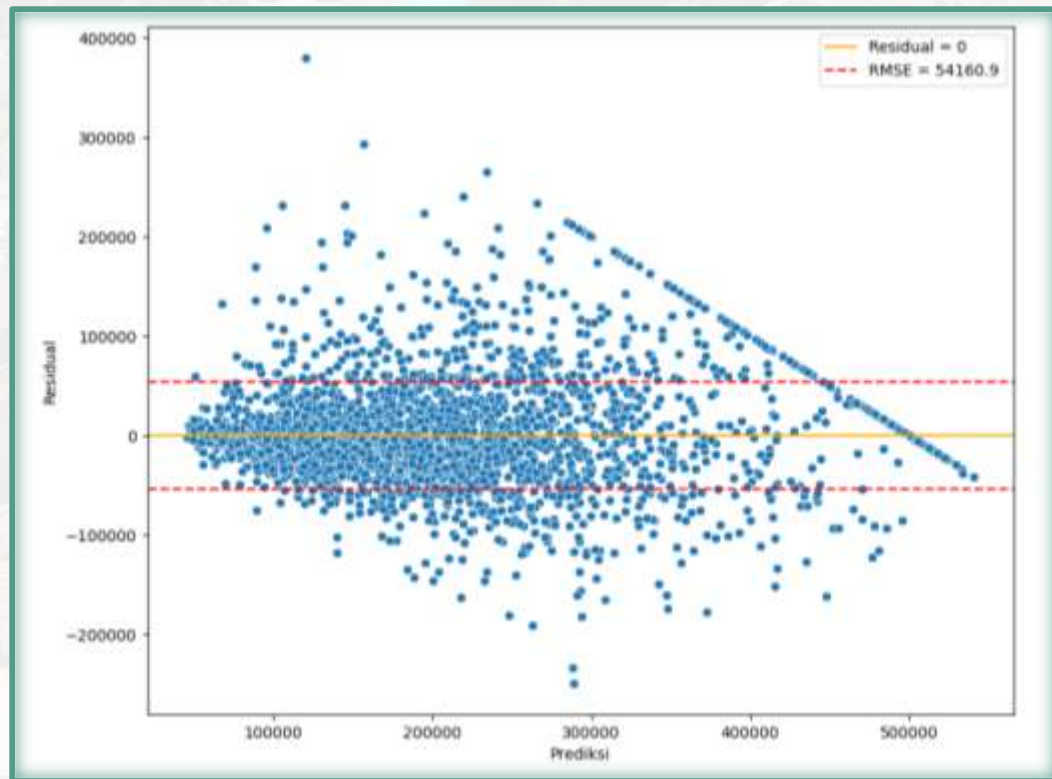
... Model akhir menghasilkan hasil dan akurasi yang memadai.



Plot ini menggambarkan hubungan antara harga rumah aktual dan harga rumah yang diprediksi oleh model. Titik-titik yang lebih dekat pada garis diagonal yang membentang dari kiri bawah ke kanan atas menunjukkan akurasi prediksi yang baik. Secara keseluruhan, terdapat tren positif yang jelas, mengindikasikan bahwa model memiliki kinerja prediksi yang relatif baik. Namun, penyebaran titik-titik, terutama pada kisaran harga rumah yang lebih tinggi, menunjukkan adanya variabilitas dalam akurasi prediksi di berbagai level harga. Hal ini mengindikasikan perlunya kalibrasi ulang model, khususnya untuk rentang harga yang lebih tinggi, atau evaluasi lebih lanjut terhadap dampak outlier terhadap kinerja model.



... Residuanya bersifat homoskedastis. Model berhasil memprediksi data tanpa bias sistematis.



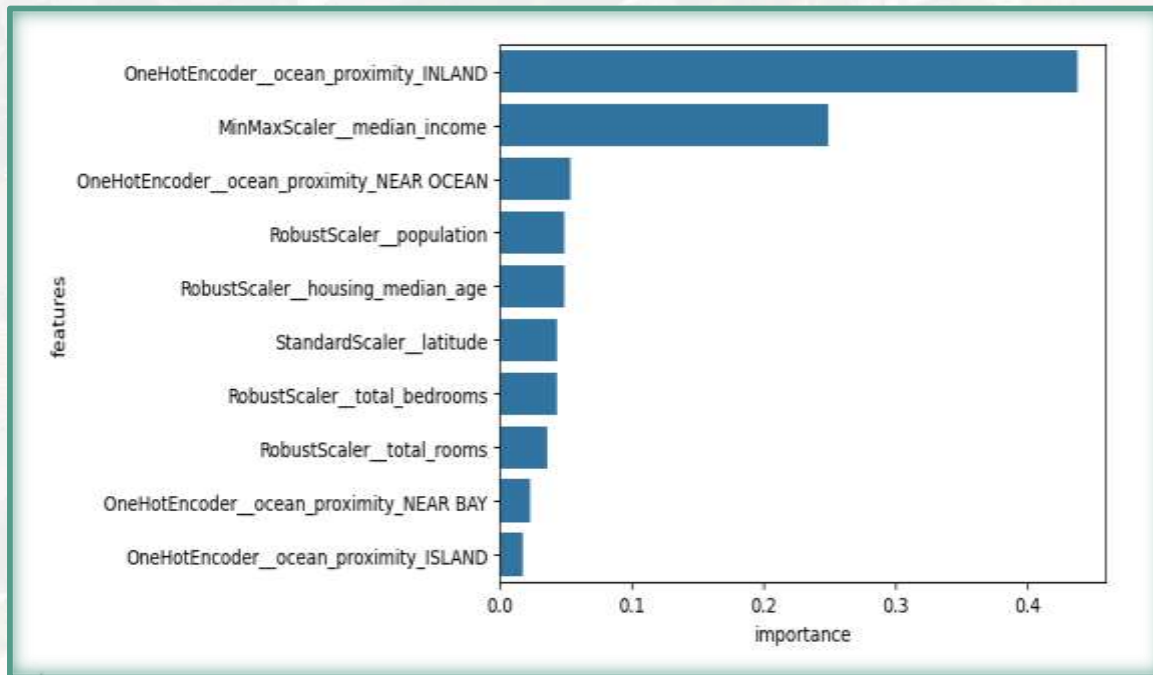
Estimasi yang Tidak Bias: Rata-rata error atau residu bernilai 0 menunjukkan bahwa model regresi tidak memiliki bias sistematis. Ini berarti model dapat memperkirakan harga rumah dengan baik, tanpa kecenderungan untuk selalu overestimate atau underestimate.

Heteroskedastisitas yang Tidak Ada: Dengan variabilitas error yang konsisten, kita dapat mengandalkan model untuk memberikan estimasi yang stabil di seluruh rentang data.

Batasan RMSE: RMSE sebesar 54,160.9 berfungsi sebagai batasan untuk residual. Jika residual berada di luar batas ini, bisa menunjukkan adanya outlier atau masalah lain yang perlu diperhatikan.



... Lokasi perumahan dan banyaknya penghasilan memiliki pengaruh tertinggi terhadap harga rumah.



Dalam analisis model prediksi harga rumah, hasil menunjukkan bahwa fitur paling berpengaruh adalah lokasi. Lokasi di dekat laut atau pedalaman berpengaruh signifikan terhadap harga, mencerminkan tren pasar real estate di mana properti dekat fasilitas wisata memiliki nilai lebih tinggi.

Selain itu, median_income juga memiliki pengaruh kuat dalam menentukan harga rumah, mencerminkan daya beli masyarakat. Fitur lain seperti populasi, umur rumah, dan jumlah kamar juga berkontribusi, meskipun dengan pengaruh yang lebih kecil. Misalnya, populasi yang lebih tinggi meningkatkan permintaan perumahan, sementara usia rumah mencerminkan kondisi dan daya tarik properti.





Conclution

Analisis dengan model XGBoost Regressor menunjukkan efektivitas dalam memprediksi harga rumah, dengan RMSE 54.160,9, MAE 37.620,9, dan MAPE 21%. Meskipun performa ini menunjukkan tingkat ketepatan yang relatif baik, nilai MAPE yang mencapai 21% mengindikasikan bahwa model mungkin mengalami kesulitan dalam situasi atau data tertentu. Kesalahan perkiraan sebesar USD 37.620 memberi panduan bagi pengembang dan investor untuk merencanakan strategi penetapan harga yang lebih akurat dan membuat keputusan investasi yang lebih realistis.





Recommendation



Rekomendasi Untuk Model



Explorasi Pada Fitur Lainnya

Untuk meningkatkan akurasi model regresi yang dibangun menggunakan XGBoost pada dataset California Housing, disarankan untuk melakukan eksplorasi lebih lanjut terhadap fitur tambahan yang dapat memberikan informasi lebih kaya.



Hyperparameter Tuning Lanjutan

Melakukan eksperimen hyperparameter tuning yang lebih mendalam dapat membantu menemukan kombinasi parameter optimal, sehingga dapat menurunkan nilai metrik seperti RMSE dan MAPE.



Feature Engineering Lanjutan

Selain itu, penerapan teknik feature engineering, seperti penggabungan variabel atau penghapusan variabel yang memiliki korelasi tinggi, dapat memperkuat model.



Rekomendasi Untuk Model



Identifikasi Outliers Lanjutan

Identifikasi dan penghapusan outlier yang berfungsi sebagai noise juga sangat penting, analisis lebih lanjut terhadap outlier dapat memberikan wawasan mengenai pengaruhnya terhadap model.



Pengujian Model Lainnya

Pertimbangkan untuk menguji model lain, seperti Elastic Net dan Support Vector Regression (SVR), yang mungkin menawarkan performa lebih baik dalam konteks ini.



Pengujian AB Testing

Terakhir, pelaksanaan A/B testing dengan berbagai versi model sangat dianjurkan untuk mengevaluasi efektivitas dan akurasi model dalam konteks operasional nyata.



Rekomendasi Untuk Bisnis



Mengingat RMSE menunjukkan bahwa kesalahan prediksi rata-rata dapat mencapai USD 54.160,9, disarankan agar perusahaan melakukan analisis pasar yang lebih mendalam untuk memahami faktor-faktor yang mempengaruhi harga rumah di berbagai lokasi.



Mengingat bahwa MAE menunjukkan kesalahan rata-rata sebesar USD 37.620 dalam estimasi harga rumah, disarankan agar pengembang dan agen real estate mempertimbangkan strategi penetapan harga yang lebih konservatif.



Dengan MAPE mencapai 21%, perusahaan sebaiknya melakukan segmentasi pasar lebih mendalam untuk mengidentifikasi area atau tipe properti yang mungkin memiliki karakteristik berbeda. Ini memungkinkan penyesuaian model yang lebih spesifik dan akurat untuk setiap segmen.



Rekomendasi Untuk Bisnis



Untuk mengoptimalkan keputusan investasi, investor dapat menggunakan informasi dari model ini untuk melakukan analisis risiko yang lebih baik termasuk perencanaan investasi untuk mengantisipasi kesalahan estimasi harga.



Mengingat adanya potensi kesalahan estimasi yang signifikan dalam situasi tertentu, disarankan agar pengembang menggunakan pendekatan berbasis data untuk melakukan penilaian pasar secara berkala.



Data dari prediksi harga ini dapat digunakan untuk merancang program promosi yang ditargetkan, seperti diskon untuk properti yang sulit terjual, sehingga dapat meningkatkan likuiditas dan menarik lebih banyak calon pembeli.



Profit Estimation

Sebagai contoh, penggunaan model ini dapat mengurangi risiko perbedaan harga hingga sekitar 20%.

Harga rumah yang tidak valid: USD 150.000

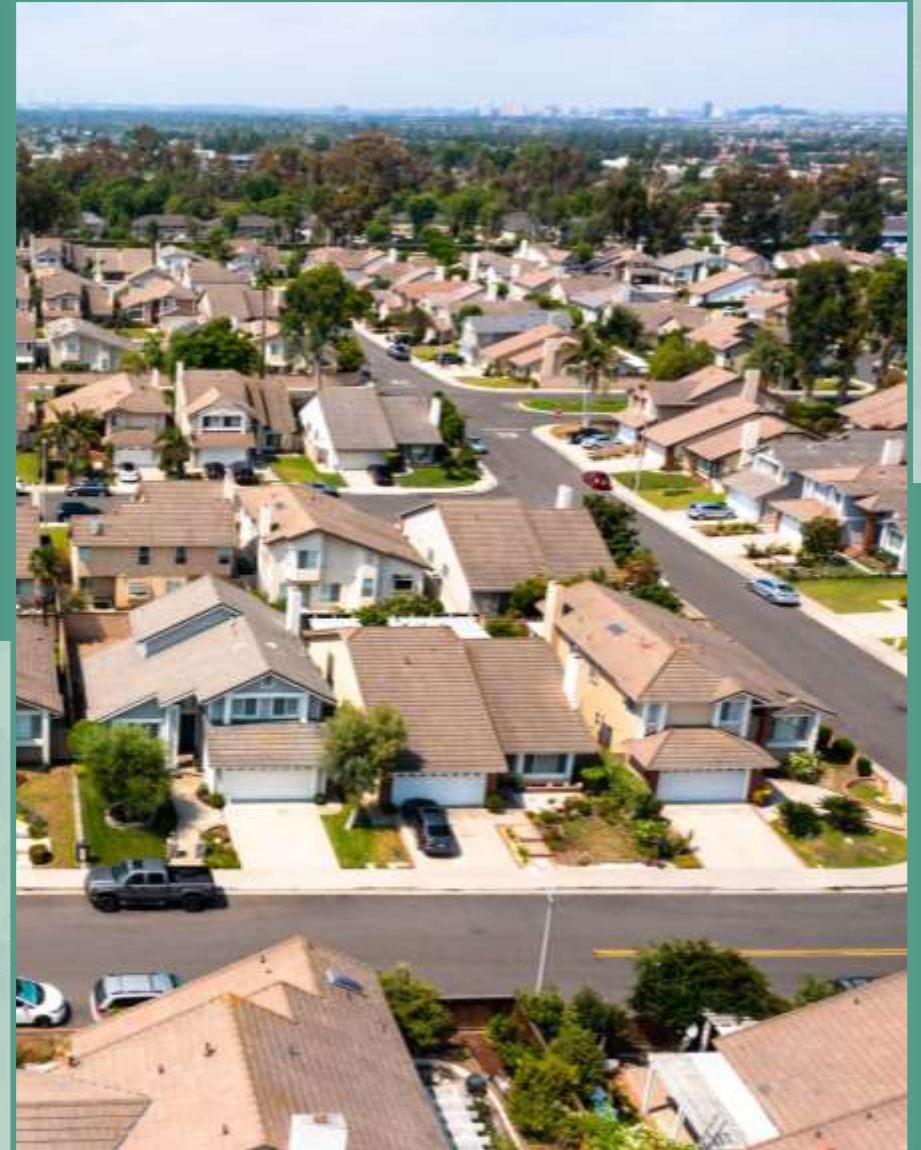
Dengan Model Prediktif

Pengurangan nilai sebesar 20% : USD 30.000

Maka, harga rumah yang diperoleh :
 $\text{USD } 150.000 - \text{USD } 30.000 = \text{USD } 120.000$

Harga Rumah Sebenarnya: USD 120.000

Kesimpulan: Dengan memanfaatkan model prediksi yang akurat, kita dapat meminimalkan risiko terkait validitas informasi harga rumah. Hal ini sangat penting untuk pengambilan keputusan investasi yang cerdas bagi para developer atau individu. Dengan informasi yang lebih tepat, developer atau individu dapat membuat keputusan yang lebih baik dan terinformasi dalam pasar properti di California.



Thank You

