INSIGHT SEEKERS

FINAL PROJECT

Homework Stage 3

By InsightSeekers

Risanto Darmawan
Asri Nur Azizah
Fransiska Angelina Widianti
Naufa Tasha Nabila

INSIGHT SEEKERS

Split Train & Test

Split Train and Test dilakukan dengan memisahkan feature yang akan digunakan sebagai data training dan data testing. Setelah split train and Test dilakukan, kami melakukan standardization dengan menggunakan StandardScaler pada feature cleanliness_rating, price_per_person, bedrooms, person_capacity, guest_satisfaction_overall, dist, metro_dist, accessibility, attr_index_norm, rest_index, lat, lng.

Hal ini dilakukan untuk menstandarisasi feture tersebut. Selain itu, terdapat kolom yang tidak lakukan standarisasi.

Modeling

Algoritma

Algoritma yang kami gunakan ada 8, yaitu

1. Linear Regression
2. Ridge
3. Lasso
4. ElasticNet
5. Decision Tree Regressor
6. Random Forest Regressor
7. Gradient Boosting Regressor
8. Extreme Gradient Boosting

URL Google Colab:

Klik disini

INSIGHT SEEKERS

Hasil Modeling Algoritma

	Linear Regression		Ridge		Lasso		ElasticNet		Decision Tree Regressor		Random Forest Regressor		Gradient Boosting Regressor		Extreme Gradient Boosting	
Model	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
MAE	0.0595	0.0597	0.0595	0.0597	0.4314	0.4331	0.4085	0.4089	3.8585e-12	0.0022	0.0009	0.0025	0.0165	0.0170	0.0059	0.0080
RMSE	0.0823	0.0830	0.0823	0.0830	0.5393	0.5402	0.5122	0.5118	3.637e-10	0.0211	0.0067	0.0179	0.0299	0.0307	0.0124	0.0228
R2	0.9787	0.9786	0.9787	0.9786	0.0908	0.0958	0.1797	0.1882	1.0	0.9986	0.9998	0.9990	0.9971	0.9970	0.9995	0.9983

Berdasarkan analisis, model **Extreme Gradient Boosting (XGBoost)** menunjukkan performa terbaik dengan nilai **R^2 yang tinggi (0.9995 Train, 0.9983 Test)** serta **MAE dan RMSE yang kecil**, mencerminkan akurasi prediksi yang sangat baik dan generalisasi yang optimal. **Random Forest** dan **Gradient Boosting** juga memiliki performa yang mendekati XGBoost, dengan perbedaan kecil antara Train dan Test, menunjukkan model yang andal tanpa overfitting. Sebaliknya, **Decision Tree** mengalami overfitting, terlihat dari R^2 Train sempurna (1.0) namun tidak diimbangi oleh generalisasi yang baik pada data Test. **Linear Regression** dan **Ridge** menawarkan performa sederhana yang konsisten, tetapi tertinggal dibandingkan model ensemble, sementara **Lasso** dan **ElasticNet** mengalami underfitting, dengan R^2 yang jauh lebih rendah. Secara keseluruhan, **XGBoost** adalah pilihan terbaik untuk memaksimalkan performa prediksi dengan risiko overfitting yang minimal.

INSIGHT SEEKERS

Modeling Algoritma (Eksperimen)

Kami melakukan eksperimen dengan melakukan perubahan Random State dari 42 menjadi 40. Hasil yang diperoleh terdapat perbedaan, yaitu sebagai berikut:

Random State 40 (Decision Tree Regressor)	Random State 42 (Decission Tree Regressor)
Hasil Model Training MAE: 1.7545484036333112e-12 Testing MAE: 0.0021885804071571347 Training RMSE: 2.522393147234491e-10 Testing RMSE: 0.020553900224080587 Training R2 Score: 1.0 Testing R2 Score: 0.9986910826175242 Hasil Hyperparameter Tuning Best Score (MAE): 0.0017928475929526403 Final MAE: 0.0017928475929526403 Final RMSE: 0.0144065688541882 Final R2 Score: 0.9993523830024894 Hasil Cross Validation Mean Cross-validation MAE: -0.0032818041697406306 Mean Cross-validation RMSE: -0.02373884736646354 Mean Cross-validation R^2: 0.998232896537577	Hasil Model Training MAE: 3.858522881116652e-12 Testing MAE: 0.002209187907527769 Training RMSE: 3.637089411923764e-10 Testing RMSE: 0.021085450254411796 Training R2 Score: 1.0 Testing R2 Score: 0.998622506666223 Hasil Hyperparameter Tuning Best Score (MAE): 0.0016329972655086738 Final MAE: 0.0016329972655086738 Final RMSE: 0.01297531132536231 Final R2 Score: 0.9994746694053979 Hasil Cross Validation Mean Cross-validation MAE: -0.0033010111870966355 Mean Cross-validation RMSE: -0.023980438987263535 Mean Cross-validation R^2: 0.9981959206051185

Random State 42 memberikan hasil yang sedikit lebih stabil** dibandingkan Random State 40, khususnya pada nilai Cross-Validation MAE dan RMSE yang lebih kecil. Namun, perbedaan performa kedua random state sangat kecil, dan keduanya menunjukkan model yang robust dan akurat.
Untuk kasus ini, tidak ada indikasi data leakage, dan overfitting pada training R^2 (1.0) masih dapat ditoleransi mengingat performa testing tetap sangat tinggi.

INSIGHT SEEKERS

Hyperparameter Tuning

Hyperparameter tuning yang kami gunakan adalah Randomized Search.

	Linear Regression	Ridge	Decision Tree Regressor	Random Forest Regressor	Gradient Boosting Regressor	Extreme Gradient Boosting
Best Score (MAE)	-	0.0596	0.0009	0.0015	1.3403e-06	0.0145
MAE	-	0.0596	0.0009	0.0015	1.3403e-06	0.0145
RMSE	-	0.0824	0.0111	0.0108	7.5641e-06	0.0257
R2	-	0.9787	0.9996	0.9996	0.9999	0.9979

Hasil hyperparameter tuning menggunakan Randomized Search menunjukkan bahwa model dengan performa terbaik berdasarkan nilai **MAE (Mean Absolute Error)** adalah **Gradient Boosting Regressor**, dengan nilai MAE terkecil sebesar **1.3403e-06**. Model ini juga memiliki nilai **RMSE** (Root Mean Squared Error) dan **R^2** yang sangat baik, yaitu **7.5641e-06** dan **0.9999**, menunjukkan akurasi prediksi yang sangat tinggi. Model **Random Forest Regressor** dan **Decision Tree Regressor** juga memberikan performa baik dengan nilai R^2 di atas **0.99**, namun MAE-nya sedikit lebih tinggi dibanding Gradient Boosting. Sebaliknya, model **Ridge Regression** memiliki performa yang lebih rendah dengan MAE di atas **0.0595**, menunjukkan bahwa model berbasis linear kurang mampu menangkap kompleksitas data dibanding model berbasis pohon seperti Gradient Boosting.

INSIGHT SEEKERS

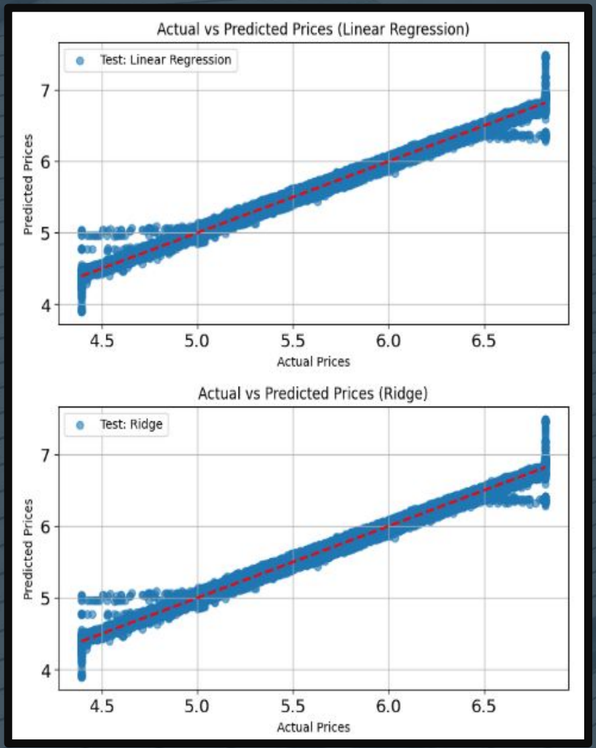
Model Evaluation

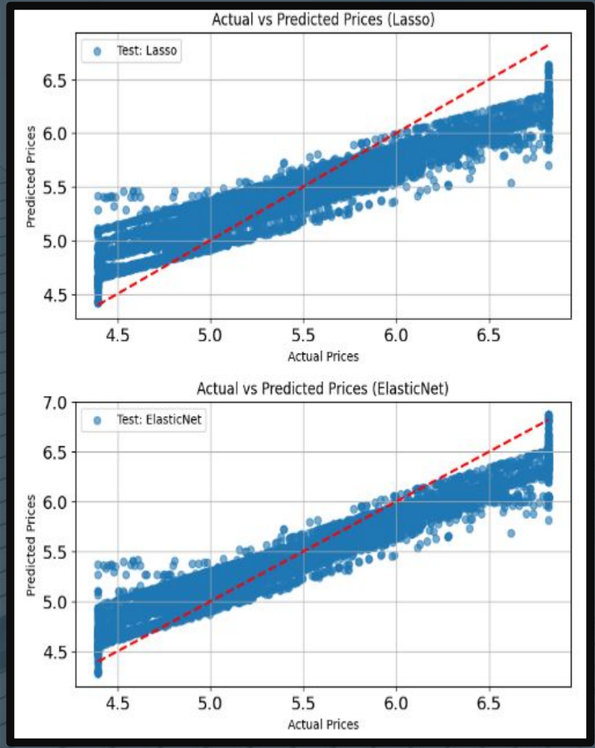
Model Evaluation yang kami gunakan adalah Cross Validation.

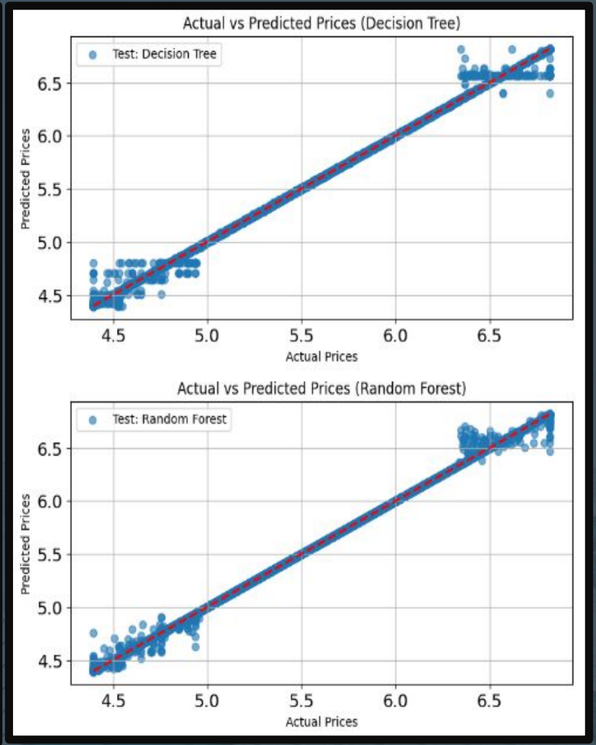
	Linear Regression	Ridge	Decision Tree Regressor	Random Forest Regressor	Gradient Boosting Regressor	Extreme Gradient Boosting
MAE	-0.0595	-0.0595	-0.0027	-0.0029	-0.0022	-0.0177
RMSE	-0.0824	-0.0824	-0.0235	-0.0202	-0.0195	0.0322
R2	0.9787	0.9787	0.9982	0.9987	0.9988	0.9967

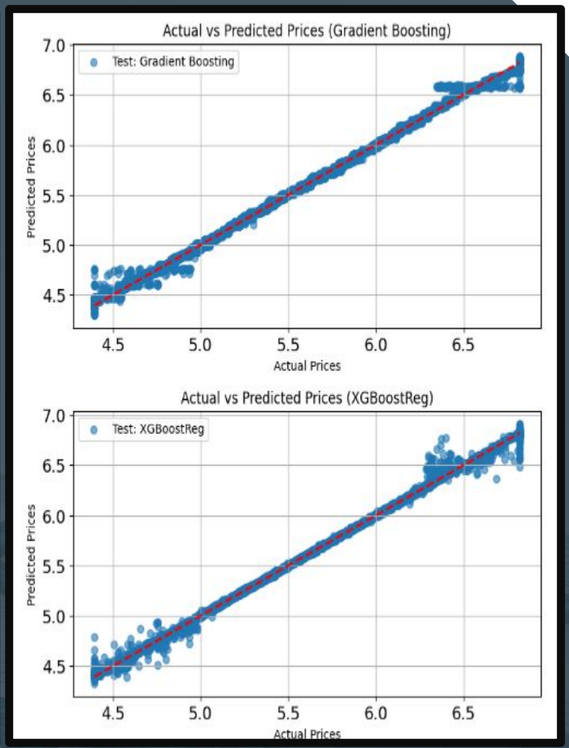
Hasil evaluasi model menggunakan cross-validation menunjukkan bahwa **Gradient Boosting Regressor** memiliki performa terbaik dengan nilai **MAE** paling kecil (**-0.0022**) dan **RMSE** terendah (**-0.0195**), disertai nilai **R^2** yang sangat tinggi (**0.9988**), menunjukkan kemampuan prediksi yang akurat dan stabil. **Random Forest Regressor** dan **Decision Tree Regressor** juga menunjukkan performa yang sangat baik dengan MAE dan RMSE rendah serta R^2 di atas **0.998**, tetapi sedikit lebih rendah dibanding Gradient Boosting. Di sisi lain, model berbasis linear seperti **Linear Regression** dan **Ridge Regression** memiliki performa lebih rendah, dengan MAE sebesar **-0.0595**, RMSE **-0.0824**, dan R^2 hanya **0.9787**, menandakan keterbatasan dalam menangkap pola kompleks pada data. **Extreme Gradient Boosting** memiliki R^2 tinggi (**0.9967**) tetapi performanya sedikit lebih rendah dibandingkan Gradient Boosting biasa. Ini mengindikasikan bahwa model berbasis ensemble, terutama Gradient Boosting, lebih unggul untuk data ini.

INSIGHT SEEKERS

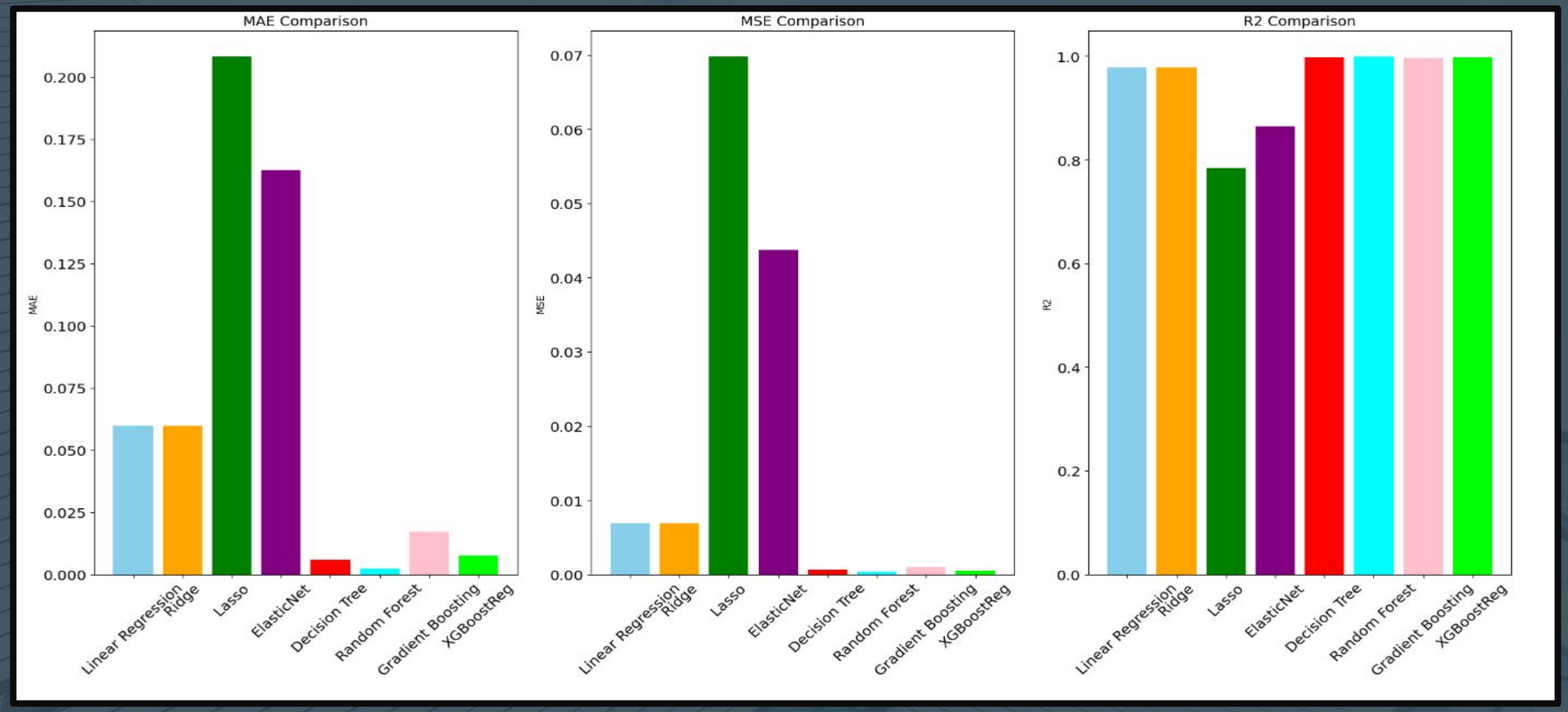








Model seperti **Gradient Boosting**, **Random Forest**, dan **XGBoostReg** menunjukkan distribusi titik yang sangat dekat dengan garis diagonal (garis merah putus-putus), yang mengindikasikan performa prediksi yang sangat akurat. Sementara itu, model berbasis linear seperti **Linear Regression**, **Ridge**, dan **Lasso** menunjukkan penyimpangan yang lebih besar, terutama pada bagian ujung distribusi nilai aktual. Hal ini menunjukkan bahwa model linear kurang mampu menangkap kompleksitas pola data dibandingkan model berbasis ensemble atau pohon keputusan seperti Gradient Boosting dan Random Forest. Secara keseluruhan, model Gradient Boosting tampaknya memberikan hasil terbaik dengan titik-titik yang hampir sempurna mengikuti garis diagonal, menandakan prediksi yang hampir identik dengan nilai aktual.



Dari grafik MAE dan MSE, terlihat bahwa model **Lasso** dan **ElasticNet** memiliki error yang jauh lebih tinggi dibandingkan model lainnya, menunjukkan underfitting. Sebaliknya, model ensemble seperti **XGBoostReg**, **Gradient Boosting**, dan **Random Forest** memiliki MAE dan MSE yang sangat kecil, mencerminkan akurasi prediksi yang tinggi. Pada grafik R^2, hampir semua model (kecuali Lasso dan ElasticNet) mendekati nilai 1, dengan **XGBoostReg**, **Gradient Boosting**, dan **Random Forest** menonjol sebagai model dengan performa terbaik. Model **Decision Tree**, meskipun memiliki R^2 tinggi, cenderung menunjukkan overfitting karena error yang relatif kecil hanya pada data train. Secara keseluruhan, XGBoostReg adalah pilihan terbaik dengan kombinasi error terendah dan R^2 tertinggi.

INSIGHT SEEKERS

FEATURE IMPORTANCE (EXTREME GRADIENT BOOSTING

	Feature	Importance
1	price_per_person	0.528395
3	person_capacity	0.452651
26	Shared room	0.004415
15	multi	0.001502
14	Entire home/apt	0.001373
11	lng	0.001240
9	rest_index	0.001201
12	host_is_superhost	0.001141
13	biz	0.000930
8	attr_index_norm	0.000924
6	metro_dist	0.000866
10	lat	0.000846
2	bedrooms	0.000716
5	dist	0.000710
4	guest_satisfaction_overall	0.000710
18	kota_barcelona	0.000558
16	kota_amsterdam	0.000516
0	cleanliness_rating	0.000503
27	Private room	0.000380
23	kota_paris	0.000211
25	kota_vienna	0.000075
20	kota_budapest	0.000061
24	kota_rome	0.000039
22	kota_london	0.000036
17	kota_athens	0.000000
19	kota_berlin	0.000000
21	kota_lisbon	0.000000
7	accessibility	0.000000

Evaluasi Feature yang Paling Penting

Berdasarkan hasil feature importance, fitur yang paling signifikan memengaruhi prediksi model adalah price_per_person dengan nilai penting sebesar 52.8%. Hal ini menunjukkan bahwa harga per orang merupakan faktor utama yang dipertimbangkan oleh pelanggan dalam memilih penginapan. Selain itu, fitur person_capacity (45.2%) juga memiliki pengaruh besar, mengindikasikan bahwa kapasitas akomodasi menjadi faktor penting kedua. Fitur lain seperti Shared room, Entire home/apt, dan host_is_superhost memiliki pengaruh yang lebih kecil tetapi tetap relevan.

Analisis dan Interpretasi

Fokus utama model terhadap price_per_person menunjukkan bahwa pelanggan sangat sensitif terhadap harga, sehingga harga menjadi indikator kuat dalam memprediksi keputusan mereka. Sementara itu, person_capacity menegaskan bahwa kebutuhan kapasitas penginapan, baik untuk keluarga maupun grup, adalah pertimbangan utama lainnya. Fitur lain seperti reputasi host (host_is_superhost) dan lokasi (metro_dist, lat) memiliki pengaruh yang lebih rendah, tetapi tetap berkontribusi pada preferensi pelanggan dalam memilih penginapan yang sesuai.

INSIGHT SEEKERS

BUSINESS INSIGHT AND RECOMMENDATION

Business Insight

Dari hasil evaluasi, terlihat bahwa harga per orang adalah faktor kunci dalam pengambilan keputusan pelanggan, sehingga penyesuaian strategi harga dapat memberikan dampak besar terhadap daya tarik listing. Kapasitas akomodasi juga menjadi peluang, terutama untuk menarik kelompok besar atau keluarga. Walaupun tidak sepenting harga, kualitas layanan seperti kebersihan dan reputasi host tetap penting untuk meningkatkan loyalitas pelanggan. Selain itu, lokasi yang strategis, seperti dekat transportasi umum, dapat menambah daya tarik listing tertentu.

Rekomendasi dan Action Items

Sebagai rekomendasi, perusahaan dapat mengoptimalkan strategi harga dengan memberikan diskon dinamis berdasarkan musim atau permintaan untuk meningkatkan daya tarik listing. Promosikan listing dengan kapasitas besar untuk memenuhi kebutuhan kelompok pelanggan tertentu. Selain itu, edukasi host untuk meningkatkan kualitas layanan, seperti kebersihan dan menjadi superhost, agar dapat meningkatkan tingkat kepuasan pelanggan. Fokuskan pemasaran pada properti yang berada di lokasi strategis untuk menarik lebih banyak pelanggan yang mempertimbangkan kemudahan akses.

INSIGHT SEEKERS

THANK YOU

Final Project Stage 3

RAKAMIN ACADEMY

InsightSeekers