Report Progress

Stage 3

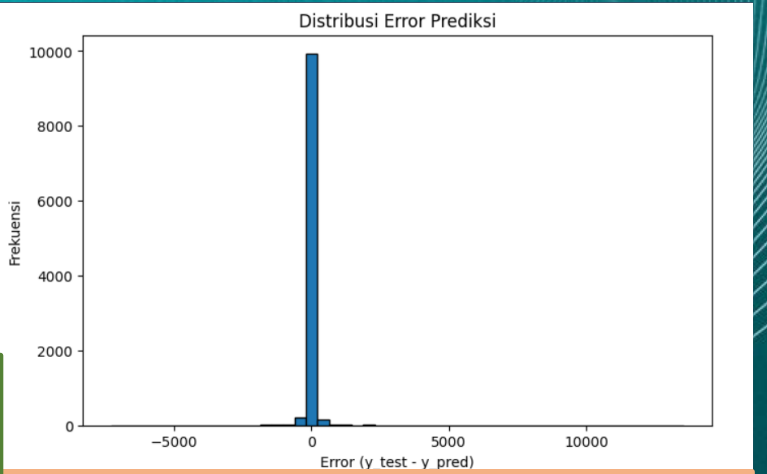
Kanva – Group 5
Anggota
•Veraldo Efraim
•Novisna Lintang Negari
•Alexander Panggabean
•Kevin William Markus Simbolon
•Adila
Mentor
Dino Febriyanto

Final Model Testing

Metrik Evaluasi	Nilai	Interpretasi
Mean Squared Error (MSE)	**73,145.61**	Rata-rata kuadrat error antara prediksi dan nilai asli, semakin besar semakin buruk.
Root Mean Squared Error (RMSE)	**270.45**	Kesalahan rata-rata dalam unit harga akomodasi, semakin kecil semakin baik.
Mean Absolute Error (MAE)	**57.83**	Rata-rata kesalahan absolut, menunjukkan bahwa rata-rata model meleset sebesar 57 unit harga.
R^2 Score	**0.49**	Model hanya mampu menjelaskan **49%** variabilitas harga akomodasi. Sisanya (51%) tidak bisa dijelaskan oleh model.

1. Analisis Hasil Evaluasi Model

Berdasarkan pengujian model yang telah dilatih sebelumnya dalam **Stage 2**, model yang diuji adalah **Random Forest Regressor**, yang telah dioptimalkan menggunakan **Randomized Search CV**.



2. Analisis Histogram Distribusi Error

- **Sebagian besar error terkumpul di sekitar 0**, artinya banyak prediksi yang cukup akurat.
- **Beberapa error besar (>5000 unit harga)** menunjukkan bahwa model **gagal memprediksi beberapa harga dengan baik (outliers)**.
- **Distribusi error tidak simetris**, yang berarti model masih memiliki bias terhadap kategori harga tertentu.

3. Kesimpulan

- Model ini menggunakan **Random Forest Regressor**, yang merupakan model terbaik berdasarkan **Stage 2**.
- Hasil pengujian menunjukkan bahwa model memiliki **kesalahan yang cukup tinggi**, dengan RMSE sebesar **270**.
- **R^2 Score rendah (0.49)** menunjukkan bahwa model belum cukup akurat dalam menjelaskan variabilitas harga.
- Model masih perlu **banyak perbaikan**, terutama dalam menangani **outlier** agar prediksi harga akomodasi lebih akurat. Beberapa cara untuk memperbaikinya:

1. **Feature Engineering:** Tambahkan fitur tambahan seperti **rating pelanggan, musim, atau tren harga**.
2. **Hyperparameter Tuning:** Lakukan pencarian parameter lebih lanjut menggunakan **Grid Search**.
3. **Gunakan Model Lain:** Model seperti **XGBoost atau LightGBM** bisa lebih baik dalam menangani pola non-linear.

Discuss Model Evaluation

Hasil Evaluasi Model & Analisa Underfitting/Overfitting

Metrik Evaluasi	Training	Testing	Interpretasi
Mean Squared Error (MSE)	**7,277.91**	**73,145.61**	Error meningkat drastis di testing, indikasi **overfitting**.
Root Mean Squared Error (RMSE)	**85.31**	**270.45**	Model lebih akurat di training, tetapi buruk di data uji.
R^2 Score	**0.93**	**0.49**	Model sangat baik di training, tetapi buruk di testing (indikasi overfitting).

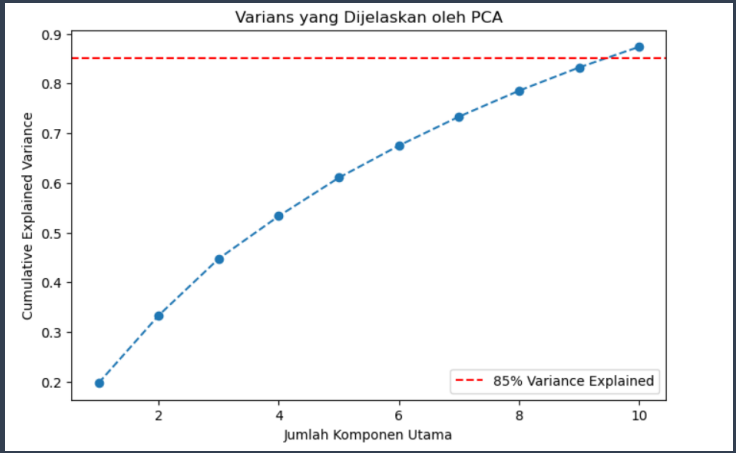
- **Model memiliki kesalahan prediksi yang cukup besar di dataset uji, terlihat dari MSE = 73,145.61 dan RMSE = 270.45.**
- **R^2 Score hanya 0.49, artinya model hanya bisa menjelaskan 49% dari variasi harga akomodasi, sedangkan sisanya 51% dipengaruhi oleh faktor lain yang tidak dapat dijelaskan oleh model.**
- **Model terlalu baik dalam mempelajari data training, tetapi buruk dalam memprediksi data baru.**
- **Model mengalami overfitting, karena:**

➢ Training R^2 Score = 0.93, Testing R^2 Score = 0.49 (Selisih besar = 0.43).

➢ Error meningkat drastis dari training ke testing, yang menunjukkan model hanya bekerja baik pada data yang sudah dikenali.

➢ Jika model generalisasi dengan baik, perbedaan R^2 Score seharusnya tidak lebih dari 0.1-0.2.

Error Analysis (berdasarkan PCA)

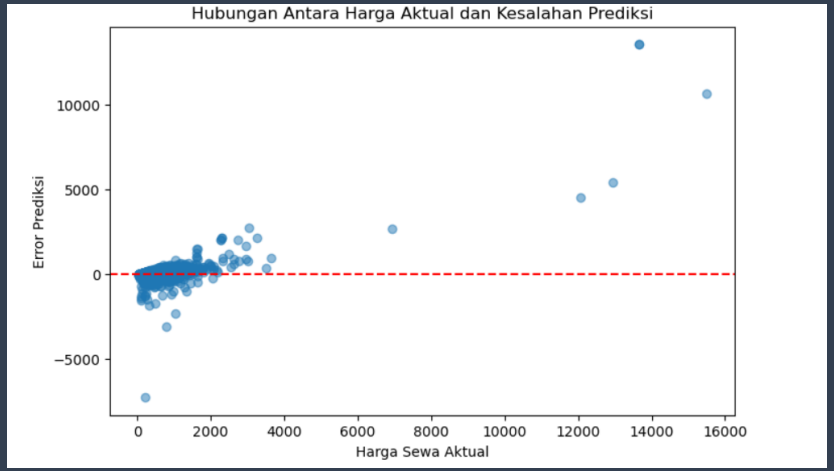



```
(array([0.19747125, 0.13588842, 0.1141402 , 0.08568564, 0.07712099,
        0.06461017, 0.05820975, 0.05213249, 0.0467502 , 0.04149715]),
 array([0.19747125, 0.33335967, 0.44749987, 0.53318551, 0.6103065 ,
        0.67491667, 0.73312642, 0.78525891, 0.83200911, 0.87350626]))
```

Hasil analisis PCA di atas menunjukkan:

1. Komponen pertama menjelaskan ~19.7% varians dalam data.
2. 10 komponen utama menjelaskan sekitar 87.4% varians, yang berarti sebagian besar informasi dalam data bisa direpresentasikan dengan 10 komponen saja.
3. Jumlah komponen optimal adalah sekitar 8-9 komponen, agar tetap menangkap >85% varians tanpa kehilangan terlalu banyak informasi.

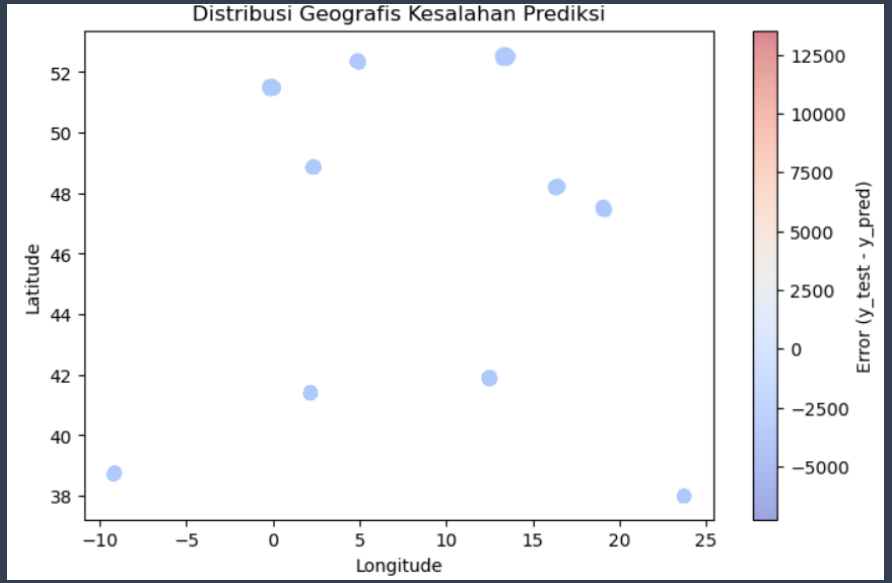
Error Analysis



1. **Sebagian besar prediksi cukup akurat untuk harga sewa yang lebih rendah**, terlihat dari titik-titik yang berkumpul dekat garis nol.
2. **Error prediksi cenderung meningkat seiring dengan kenaikan harga sewa aktual**, menunjukkan bahwa model mungkin kurang mampu memprediksi harga sewa yang lebih tinggi dengan akurasi yang baik.
3. **Terdapat beberapa outlier**, di mana kesalahan prediksi sangat besar, baik dalam bentuk overestimasi (di atas garis merah) maupun underestimasi (di bawah garis merah).
4. **Model mungkin perlu perbaikan atau penyesuaian**, seperti normalisasi data, penggunaan fitur tambahan, atau model yang lebih kompleks untuk meningkatkan akurasi pada harga sewa yang lebih tinggi.

Secara keseluruhan, meskipun model bekerja cukup baik untuk harga yang lebih rendah, masih terdapat kelemahan dalam memprediksi harga sewa yang lebih tinggi, yang perlu diperbaiki untuk meningkatkan performa model.

Error Analysis



1. **Kesalahan prediksi bervariasi di berbagai lokasi geografis**, ditunjukkan oleh perbedaan warna dan ukuran titik pada peta.
2. **Titik dengan warna lebih merah menunjukkan kesalahan prediksi yang lebih besar dalam bentuk overestimasi**, sedangkan **warna lebih biru menunjukkan kesalahan dalam bentuk underestimasi**.
3. **Beberapa lokasi memiliki error yang jauh lebih besar dibandingkan lokasi lainnya**, yang dapat mengindikasikan bahwa model mengalami kesulitan dalam memprediksi harga di wilayah tertentu.
4. **Pola distribusi kesalahan bisa menjadi indikasi bahwa model belum mempertimbangkan faktor geografis dengan optimal**, sehingga mungkin diperlukan fitur tambahan yang lebih spesifik terkait lokasi.

Adjustments

Summary

Kesimpulan
- **Pastikan X_train_scaled sudah didefinisikan sebelum digunakan.**
- **Kurangi kombinasi hyperparameter untuk mempercepat pencarian.**
- **Gunakan verbose=2 untuk melihat progres pencarian.**
- **Setelah GridSearch, gunakan model terbaik (best_model) untuk evaluasi.**

Github:
https://github.com/Efraim5/Stage3-Group5-Kanva/


Thank you!
Raih Mimpi #TanpaBatas