

Machine Learning Regression Model: Prediksi Harga Apartemen Daegu

Fatimah Azzahra

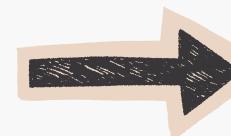
Gabriella Davintia

Tengku Arika Hazera

Executive Summary

Projek ini bertujuan mengembangkan model machine learning berbasis regresi untuk memprediksi harga apartemen di Daegu, Korea Selatan. Harga Apartment di Daegu dipengaruhi oleh berbagai faktor, seperti karakteristik bangunan, lokasi, dan fasilitas apartemen.

Predict Harga Apartment dengan menggunakan XGB Model untuk menunjukkan performa terbaik



Model akhir dapat digunakan untuk penetapan harga apartemen yang lebih kompetitif dan realistik

Table of Content

01

Introduction

02

Data Understanding

03

EDA

04

Modeling

05

Conclusion

06

Recommendation



INTRODUCTION

Fatimah Azzahra

Gabriella Davintia

Tengku Arika Hazera

Latar Belakang

- Dipicu oleh keterbatasan lahan permukiman di perkotaan, dimana manusia terus bertambah sedangkan ketersediaan lahan tidak
- Apartemen menjadi solusi hunian masyarakat modern
- Didukung oleh tingginya aktivitas bisnis dan mobilitas
- Harga apartemen dipengaruhi faktor internal dan eksternal

Business Problem

- Harga apartemen dipengaruhi faktor internal (tipe, luas unit, usia bangunan)
- Faktor eksternal meliputi lokasi dan fasilitas sekitar
- Penetapan harga sering didasarkan pada pertimbangan subjektif
- Harga terlalu tinggi → unit sulit terjual
- Harga terlalu rendah → risiko kerugian finansial



Goals

1. Menganalisis faktor internal dan eksternal utama yang memengaruhi harga apartemen di Daegu menggunakan data historis perumahan
2. Membangun model machine learning berbasis regresi yang mampu memprediksi harga apartemen berdasarkan karakteristik properti dan fitur yang berkaitan dengan lokasi
3. Meningkatkan akurasi penetapan harga dengan menyediakan estimasi harga yang objektif dan berbasis data sehingga mencerminkan kondisi pasar saat ini

Analytical Approach

1. Data Understanding
2. EDA
3. Data Preprocessing
4. Modeling
5. Interpretation Model



Metric Evaluation

- RMSE
- R-Squared
- MAE
- MAPE

Stakeholder

- Investor/Pemilik Unit Apartemen
- Pembeli Apartemen
- Marketing Team



DATA UNDERSTANDING

Fatimah Azzahra Gabriella Davintia Tengku Arika Hazera

DATA DICTIONARY

Fitur

Faktor Internal

HallwayType	Tipe Apartemen
N_Parkinglot (Basement)	Slot parkir basement dalam suatu Apartemen
YearBuilt	Tahun Apartemen dibangun
N_FacilitiesInApt	Jumlah fasilitas umum dalam suatu Apartemen
Size(sqf)	Ukuran unit Apartemen dalam satuan feet squared

Faktor Eksternal

TimeToSubway	Waktu yang dibutuhkan dari Apartemen ke Stasiun Subway terdekat
SubwayStation	Nama Stasiun Subway terdekat
N_FacilitiesNearBy (ETC)	Jumlah fasilitas umum terdekat
N_FacilitiesNearBy (PublicOffice)	Jumlah fasilitas umum berupa kantor layanan publik terdekat
N_SchoolNearBy (University)	Jumlah Universitas terdekat

 Object

 Float

 Int

4123 Baris Data
11 Kolom Data

Target

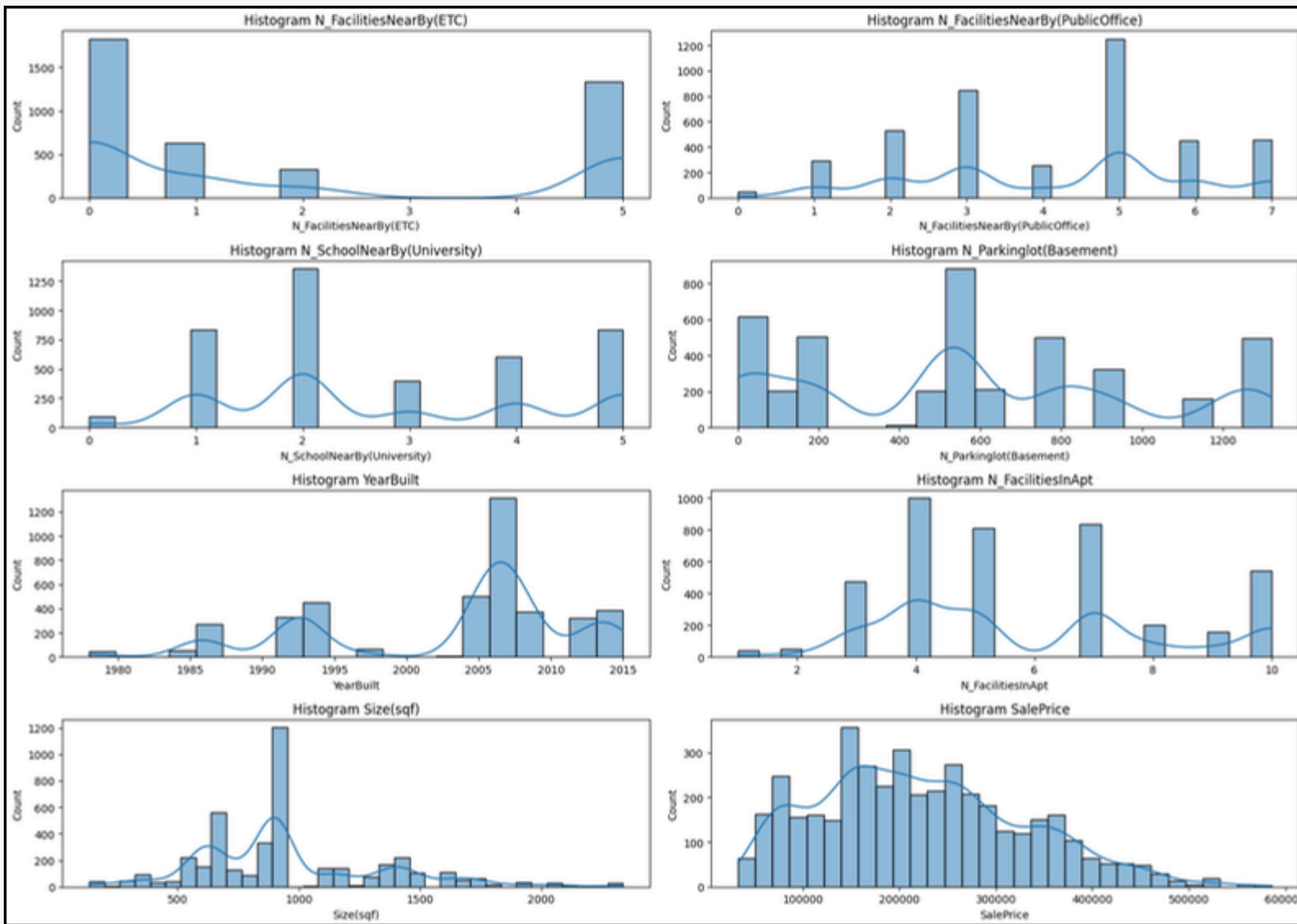
SalePrice	Harga unit Apartemen dalam satuan Won
-----------	---------------------------------------

Missing Value X
1422 Data Duplikat

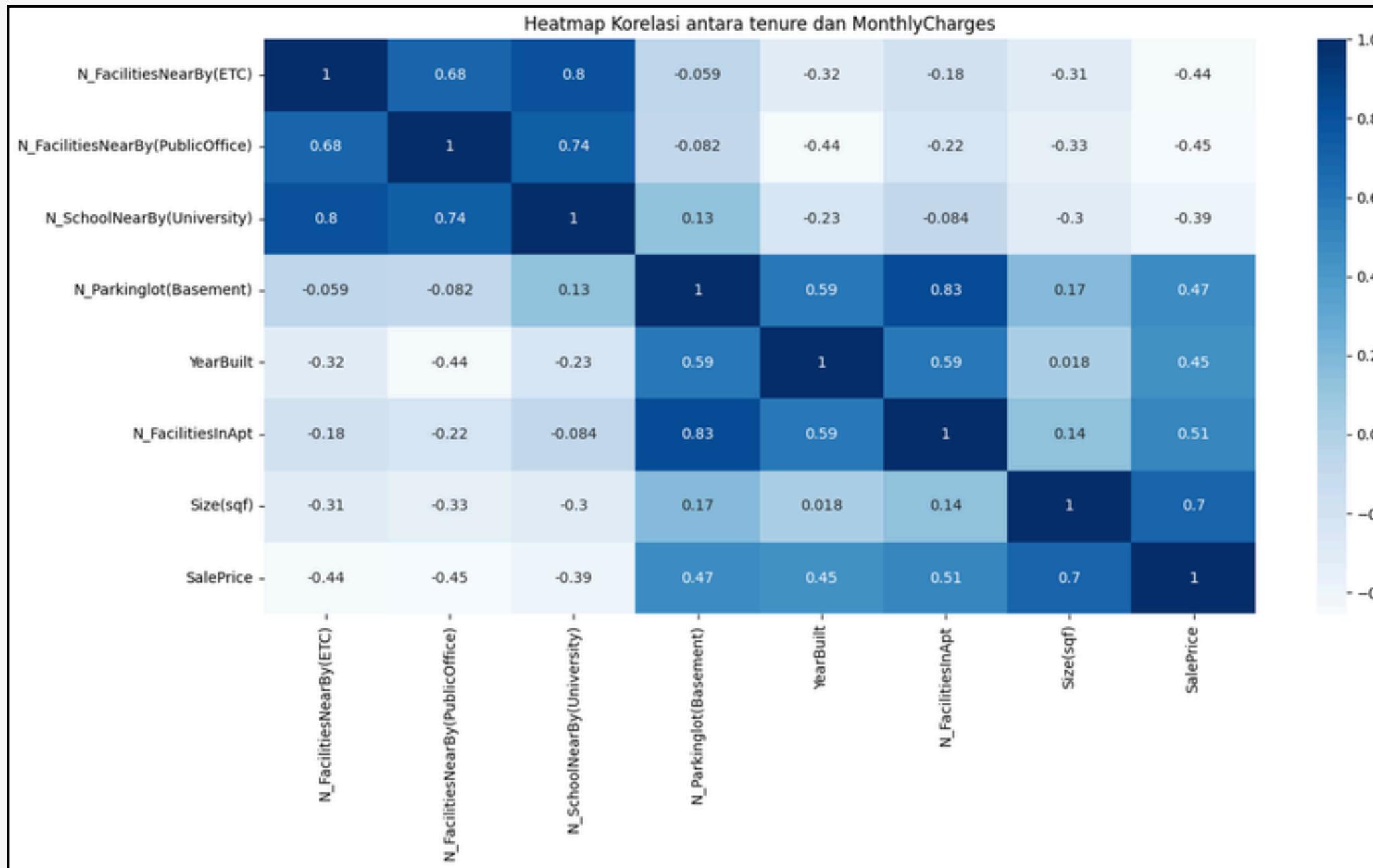
EXPLORATORY DATA ANALYSIS

Fatimah Azzahra Gabriella Davintia Tengku Arika Hazera

Histogram Data Numerik

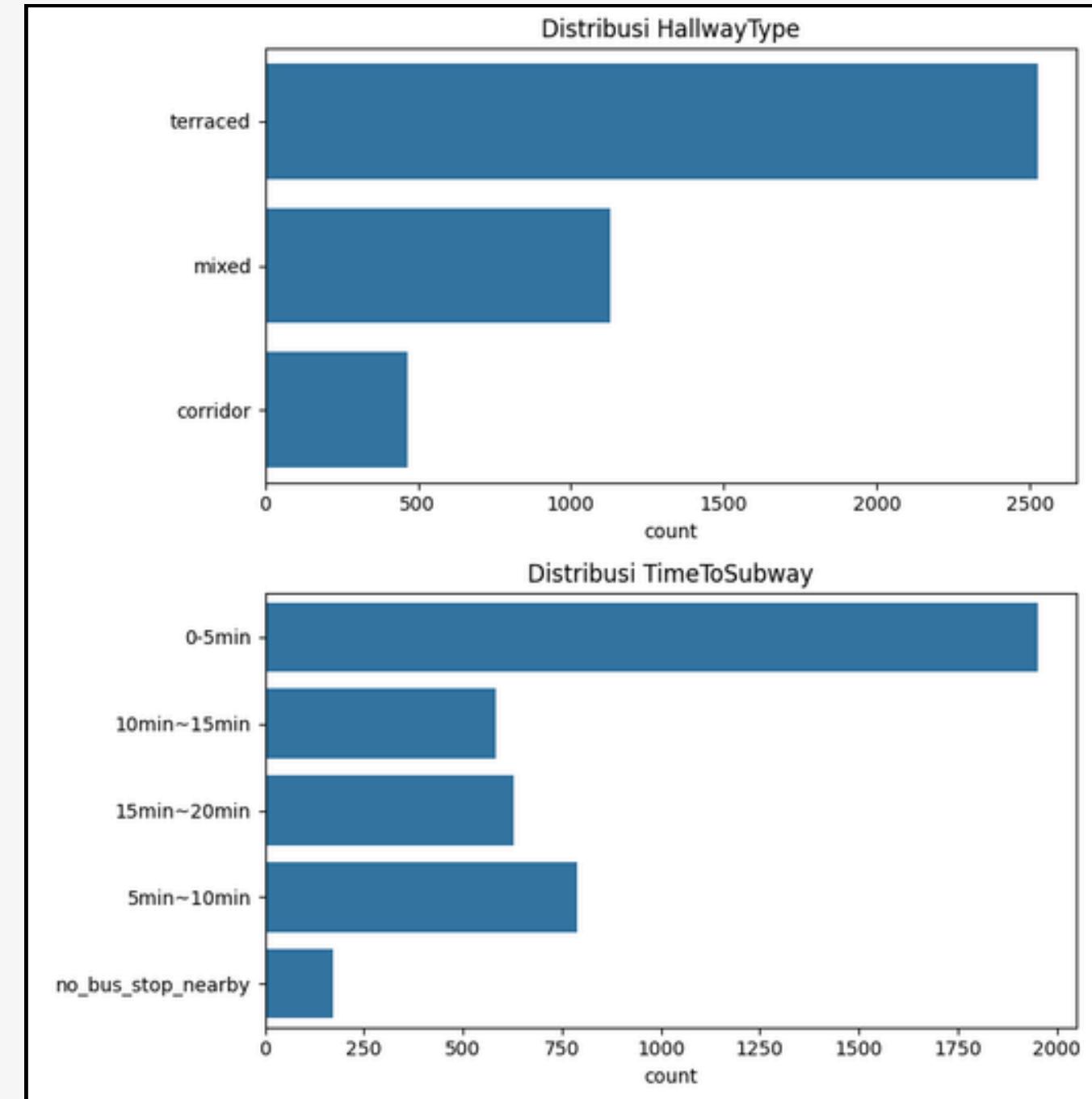
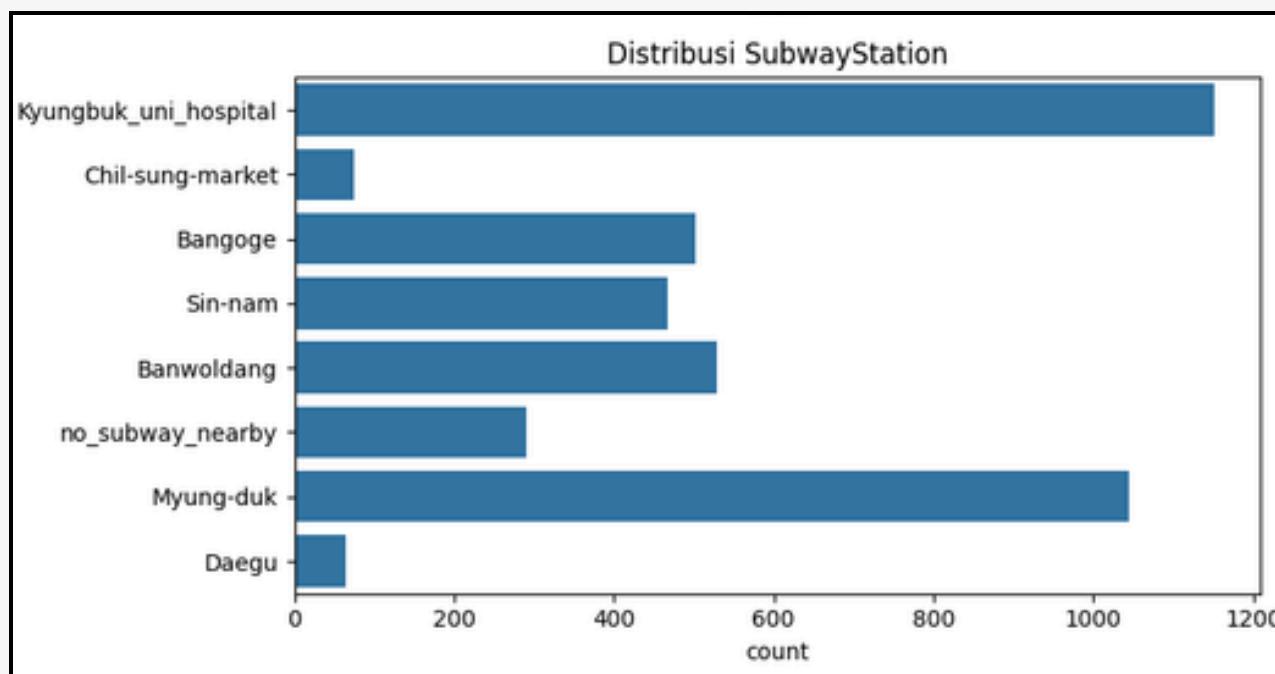


Heatmap Korelasi Data Numerik



Data Kategorik

	count	unique	top	freq
HallwayType	4123	3	terraced	2528
TimeToSubway	4123	5	0-5min	1953
SubwayStation	4123	8	Kyungbuk_uni_hospital	1152



DATA PREPROCESSING

Fatimah Azzahra Gabriella Davintia Tengku Arika Hazera

Steps

1. Data splitting

- `X = df.drop(columns=['SalePrice'])`
- `y = df['SalePrice']`

2. Train and test splitting

- Test size = 0.2

3. Encoding

- Data Kategorik → HallwayType dan TimeToSubway → One Hot (drop=first)
- Data Kategorik → SubwayStation → Binary Encoder

4. Scaling → Data Numerik → Standard Scaler

5. Baseline Model

- Linear Regression
- Ridge Regression
- Lasso Regression
- Gradient Boosting
- XGBoost



MODELING

Fatimah Azzahra Gabriella Davintia Tengku Arika Hazera

Model Benchmarking

Evaluasi hasil dari 5 kandidat algoritma yang digunakan

	Model	Mean_R2	Std_R2	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE
0	Linear Regression	0.739031	0.017493	-54807.250246	1205.448739	-43298.902326	510.743516	-0.224065	0.003527
1	Ridge Regressor	0.739033	0.017494	-54807.086097	1205.470019	-43298.868643	510.722984	-0.224065	0.003527
2	Lasso Regressor	0.724057	0.024881	-56324.687242	1831.132519	-44793.127567	746.580065	-0.232818	0.003928
3	GradientBoosting Regressor	0.834765	0.005846	-43634.930832	357.894600	-33999.560445	379.305073	-0.175119	0.005275
4	XGB Regressor	0.833774	0.005450	-43767.322656	315.176641	-33970.932813	269.840829	-0.175267	0.004865

Top 2 Best Model: XGB model dan Gradient Boosting Model

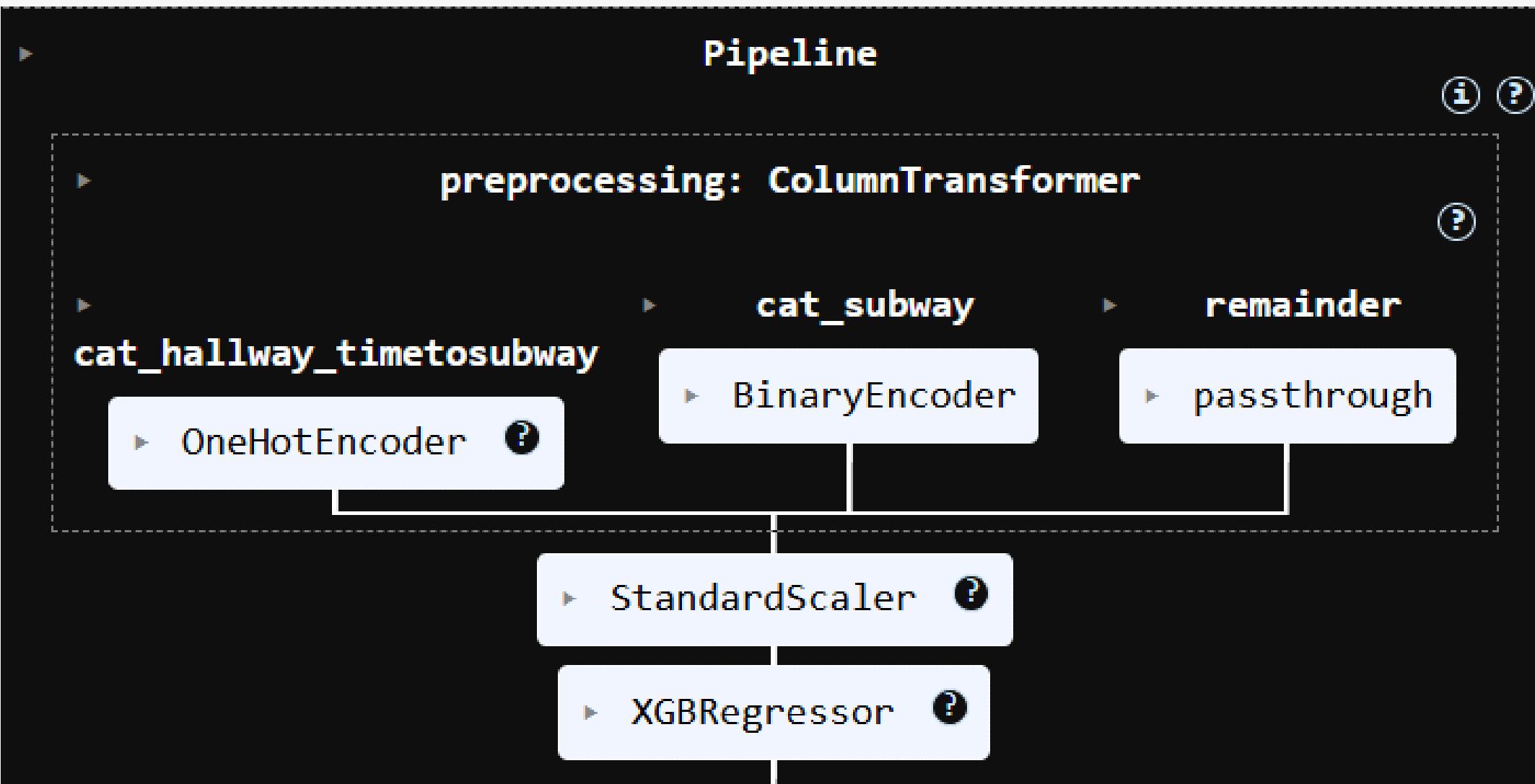
Benchmark 2 model terbaik berdasarkan performa pada test data

	R2	RMSE	MAE	MAPE
XGB	0.843193	41122.361022	32204.324219	0.178031
GradientBoosting	0.841995	41279.140747	32719.357255	0.181330



Hyperparameter Tuning

Fitting data training untuk mencari parameter terbaik



Hyperparameter Tuning

Score before tuning

	R2	RMSE	MAE	MAPE
XGB	0.843193	41122.361022	32204.324219	0.178031



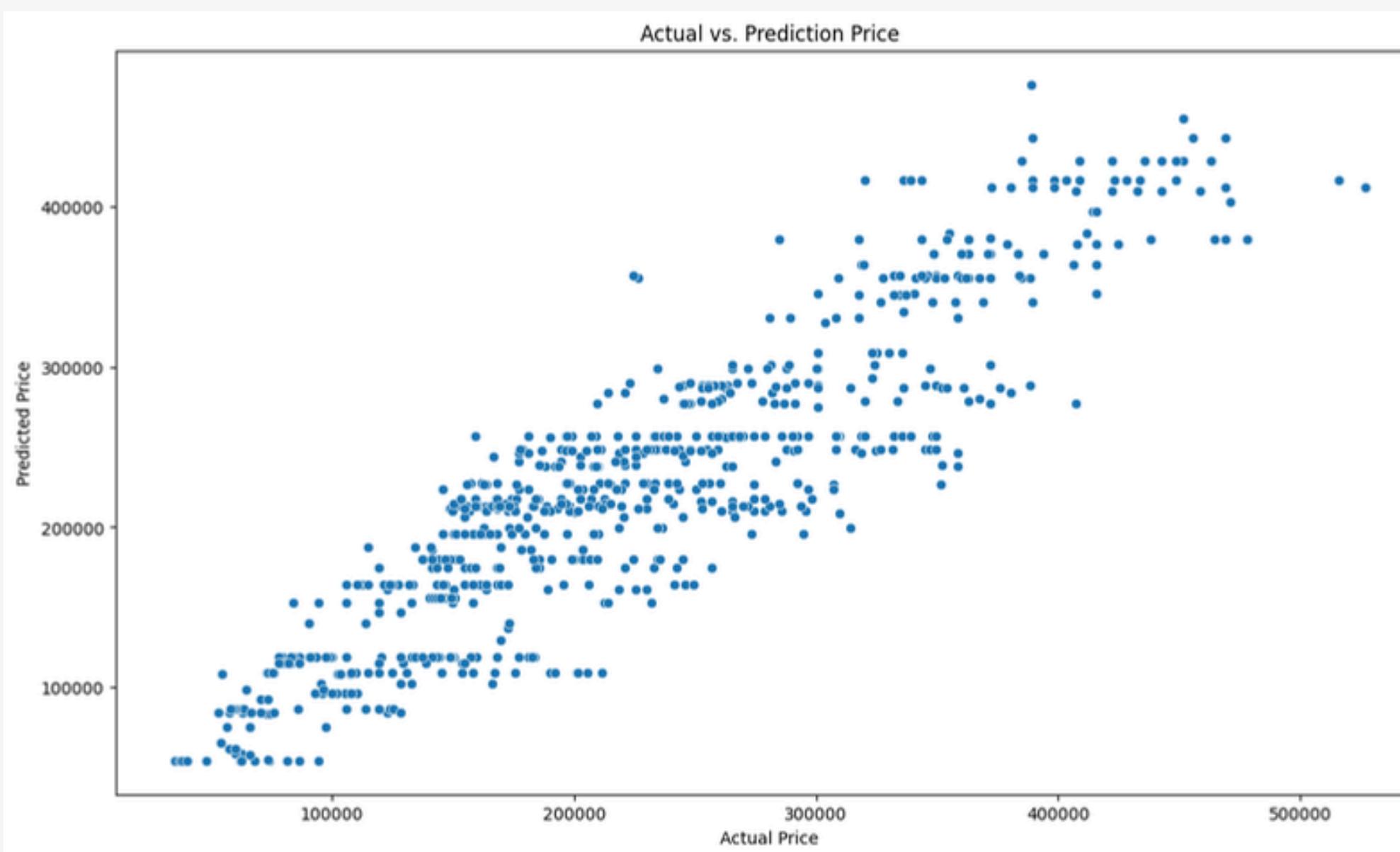
Score after tuning

	R2	RMSE	MAE	MAPE
XGB	0.84282	41171.224126	32251.728516	0.17808

Model XGB setelah dituning mengalami penurunan performa sehingga base model yang dipilih adalah model XGB tanpa hyperparameter tuning



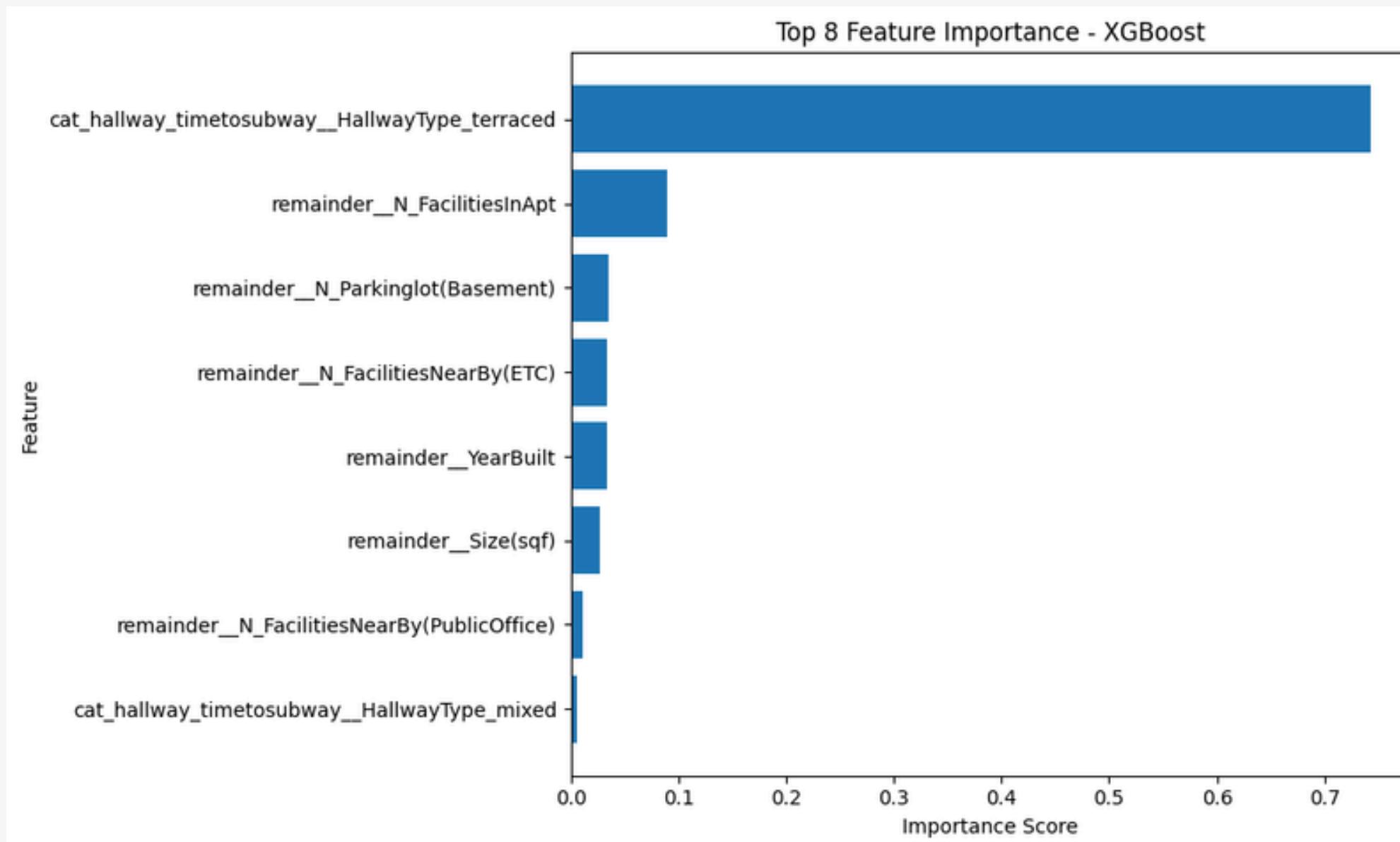
Actual vs Prediction Price



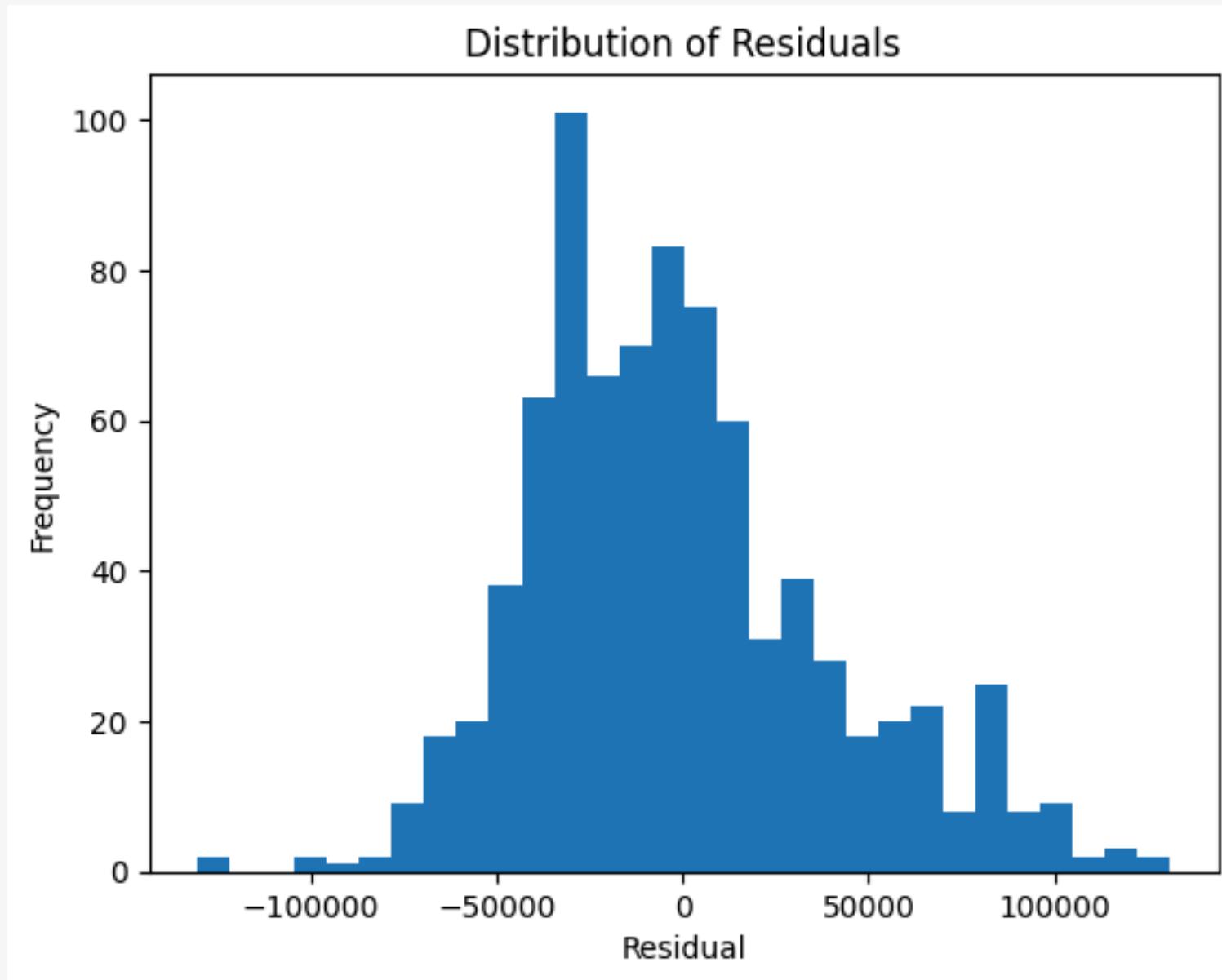
Sebaran titik yang cenderung mengikuti garis diagonal, menandakan bahwa prediksi model secara umum mendekati nilai aktual



Feature Importance



Distribution of Residuals



- Distribusi residual bersifat kurang lebih simetris dan terpusat di sekitar nol
- Indikasi bahwa kesalahan prediksi secara umum seimbang, dengan sedikit kecenderungan model melakukan underestimasi pada apartemen dengan harga tinggi



CONCLUSION & RECOMMENDATION

Fatimah Azzahra

Gabriella Davintia

Tengku Arika Hazera

Conclusion

- Projek ini mengembangkan model regresi machine learning untuk memprediksi harga apartemen di Daegu
- Model memanfaatkan faktor internal properti dan faktor eksternal berbasis lokasi
- Evaluasi dengan cross-validation menunjukkan **XGB Regression** sebagai model awal terbaik
- Hyperparameter tuning (RandomizedSearchCV) mengakibatkan **penurunan** akurasi prediksi:
 - RMSE meningkat
 - R² menurun pada data uji
- **Base model XGB** dipilih sebagai **model terbaik**
- model XGBoost → kemampuan yang cukup baik dalam menangkap pola hubungan antara harga aktual dan harga prediksi
- Namun terjadi penyebaran residual yang cukup lebar, terutama pada rentang harga menengah hingga tinggi. Adapun terjadi **underestimasi** pada apartemen dengan harga tinggi
- Pendekatan machine learning efektif sebagai alat berbasis data untuk penetapan harga apartemen yang lebih kompetitif dan realistik



Recommendation

1. Rekomendasi Data

- Menambahkan fitur temporal, seperti tahun transaksi, bulan, atau tren harga historis, untuk menangkap dinamika perubahan harga dari waktu ke waktu
- Memperinci karakteristik bangunan premium, seperti kualitas material, renovasi, atau status apartemen (luxury vs non-luxury), untuk mengurangi kesalahan pada segmen harga tinggi

2. Rekomendasi Model

- Melakukan eksperimen dengan model lain, seperti Random Forest atau LightGBM, yang dikenal efektif untuk data tabular dengan kompleksitas tinggi
- Mengembangkan model terpisah berdasarkan segmen harga (low, medium, high) agar model lebih fokus pada karakteristik masing-masing segmen



Recommendation

3. Rekomendasi Bisnis

- Siapa yang menggunakan model ini?
 - Tim pemasaran dan penjualan properti: untuk menentukan harga listing yang kompetitif
 - Investor dan analis properti: sebagai alat pendukung dalam menilai kewajaran harga dan potensi investasi
 - Manajemen platform properti: untuk memberikan rekomendasi harga otomatis kepada pemilik apartemen
- Bagaimana maintenance model ini?
 - Retraining model disarankan setiap 3–6 bulan
 - Retraining juga perlu dilakukan apabila terjadi:
 - Perubahan signifikan kondisi ekonomi (misalnya perubahan suku bunga)
 - Pergeseran tren permintaan properti
 - Masuknya data transaksi baru dalam jumlah besar
- Kapan model ini digunakan?
 - Saat penentuan harga awal (initial pricing) sebelum unit dipasarkan
 - Saat evaluasi ulang harga, jika unit belum terjual dalam periode tertentu
 - Saat analisis strategi pemasaran, untuk menentukan segmen harga dan target konsumen



Thank You

Fatimah Azzahra Gabriella Davintia Tengku Arika Hazera