House Price Prediction

Margareth Hamilton

Friday, November 8th 2024

Meet the Team

Alwy Bathia R.

Daniel Machsimus L.

Jason Hermawan

Background & Problem Statement

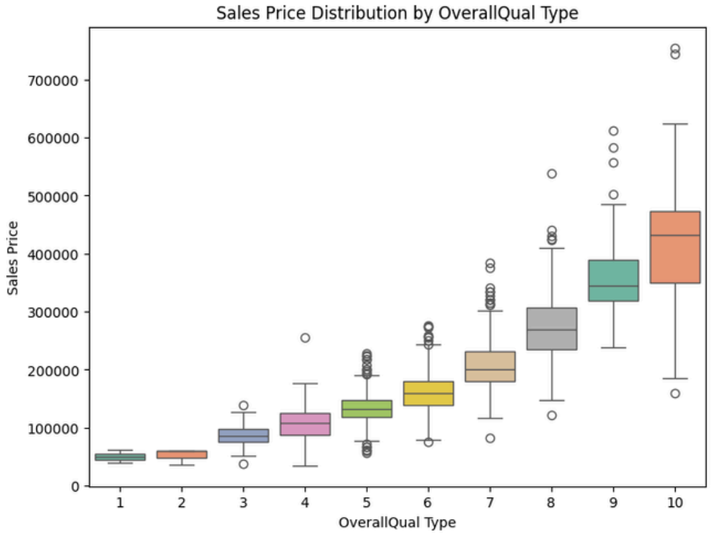
Dataset yang digunakan adalah **Ames Housing Dataset**. Dataset ini berisi berbagai **feature** terkait perumahan di daerah Ames, Iowa, US. Dataset ini memang sering digunakan untuk prediksi harga rumah.

Kebutuhan akan adanya **prediksi harga rumah** terus bermunculan dengan **pemenuhan kebutuhan primer** manusia. Tidak hanya diaplikasikan pada pembeli rumah, prediksi harga rumah dapat dimanfaatkan penjual (realtor), agen properti, dan investor dalam **penetapan harga jual rumah** yang **kompetitif** [1].

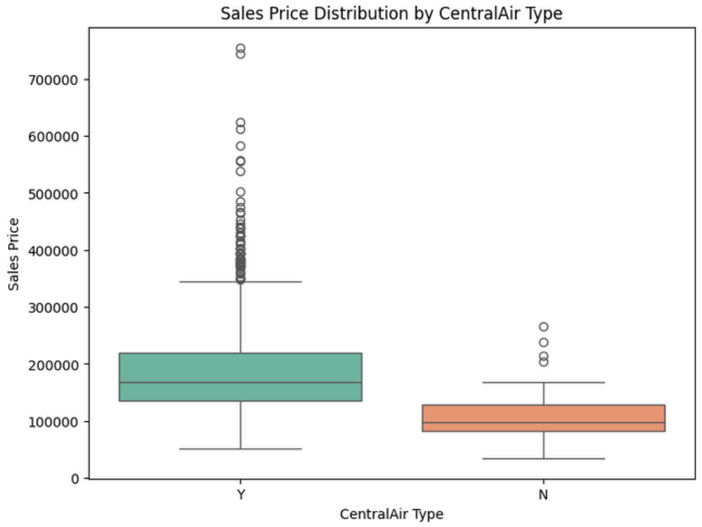
Machine Learning dapat digunakan sebagai *tools* untuk membantu dalam hal *decision-assisting* bagi banyak *stakeholder* [2].

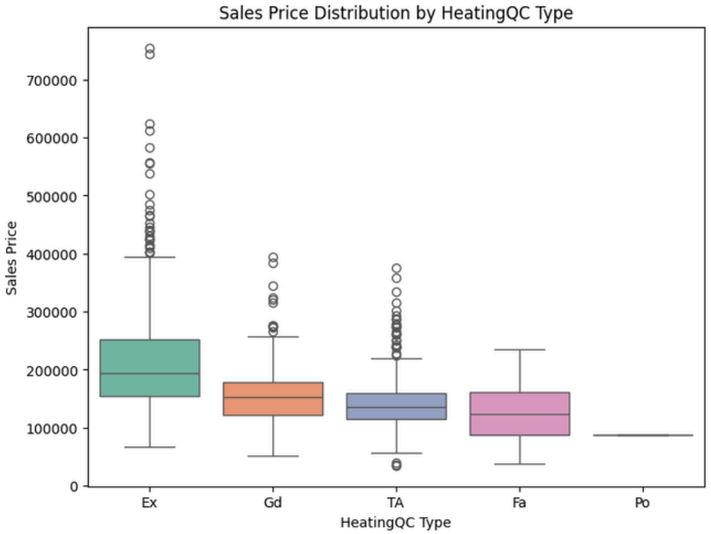
Exploratory Data Analysis **Explore Variables**

Terdapat 3 grafik yang menurut kami dilihat dari boxplotnya memiliki perbedaan antarkelas yang berbeda



Terdapat **26** dari total **81 feature** yang memiliki outlier.





Terdapat **26** dari total **81 feature** yang memiliki outlier.

Exploratory Data Analysis NA value

Persentase untuk Data yang terdeteksi ada nilai NA/None-nya

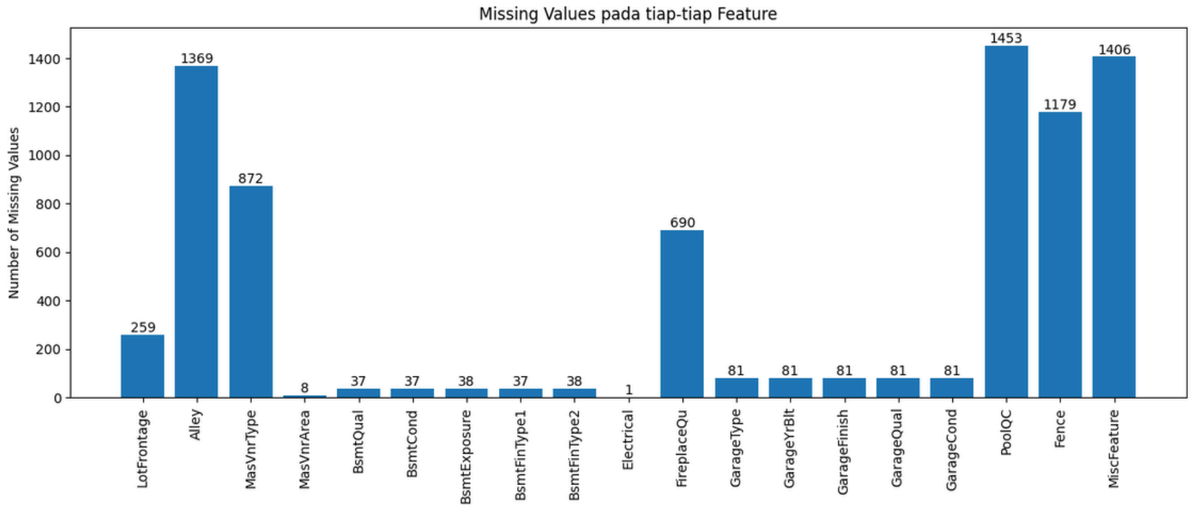
	0	1
LotFrontage	259	17.739726
Alley	1369	93.767123
MasVnrType	872	59.726027
MasVnrArea	8	0.547945
BsmtQual	37	2.534247
BsmtCond	37	2.534247
BsmtExposure	38	2.602740
BsmtFinType1	37	2.534247
BsmtFinType2	38	2.602740
Electrical	1	0.068493
FireplaceQu	690	47.260274
GarageType	81	5.547945
GarageYrBlt	81	5.547945
GarageFinish	81	5.547945
GarageQual	81	5.547945
GarageCond	81	5.547945
PoolQC	1453	99.520548
Fence	1179	80.753425
MiscFeature	1406	96.301370

Terdapat **19** dari total **81 feature** yang memiliki **nilai NA/null**.

Data yang memang variasi nilainya ada NA-nya:

- Alley (No Alley Access)
- PoolQC (No Pool)
- MasVnrType (None)
- BsmtQual (No basement)
- BsmtCond (No basement)
- BsmtExposure (No basement)
- BsmtFinType1 (No basement)
- BsmtFinType2 (No basement)
- FireplaceQu (No Fireplace)
- GarageType (No Garage)
- GarageFinish (No Garage)
- GarageQual (No Garage)
- GarageCond (No Garage)
- MiscFeature (None)
- Fence (No Fence)

Memang ada **15 feature** kategorikal memiliki variasi nilainya adalah nilai **NA/None**.



Untuk itu, ada **perlakuan** yang berbeda:
(1) perlakuan untuk **15 feature** yang memang bervariasi “NA” dan
(2) perlakuan untuk **4 feature** yang benar-benar hilang

Exploratory Data Analysis NA value

Untuk itu, ada **perlakuan** yang berbeda:
(1) perlakuan untuk **15 feature** yang memang bervariasi “NA” -> **imputasi dengan nilai spesifik**

Data yang memang variasi nilainya ada NA-nya:

- Alley (No Alley Access)
- PoolQC (No Pool)
- MasVnrType (None)
- BsmtQual (No basement)
- BsmtCond (No basement)
- BsmtExposure (No basement)
- BsmtFinType1 (No basement)
- BsmtFinType2 (No basement)
- FireplaceQu (No Fireplace)
- GarageType (No Garage)
- GarageFinish (No Garage)
- GarageQual (No Garage)
- GarageCond (No Garage)
- MiscFeature (None)
- Fence (No Fence)

Mengganti **nilai NA/None** menjadi variasi yang **sesuai** dengan **variasi None masing-masing feature**

Untuk itu, ada **perlakuan** yang berbeda:
(2) perlakuan untuk **4 feature** yang benar-benar hilang -> **imputasi**

Persentase untuk Data yang terdeteksi ada nilai NA/None-nya

	0	1				
LotFrontage	259	17.739726	Numerikal	→	Mean	
MasVnrArea	8	0.547945	Numerikal	→	nilai 0 (nol)	
Electrical	1	0.068493	Kategorikal	→	Modus	
GarageYrBlt	81	5.547945	Numerikal	→	nilai 0 (nol)	→ Feature Engineering

LotFrontage
Mean

GarageYrBlt
GarageYrBlt yang NA artinya rumah tidak memiliki garasi.

MasVnrArea
MasVnrArea yang NA artinya rumah tidak memiliki MasVnr

Electrical
Data kategorikal -> modus paling umum

Modelling

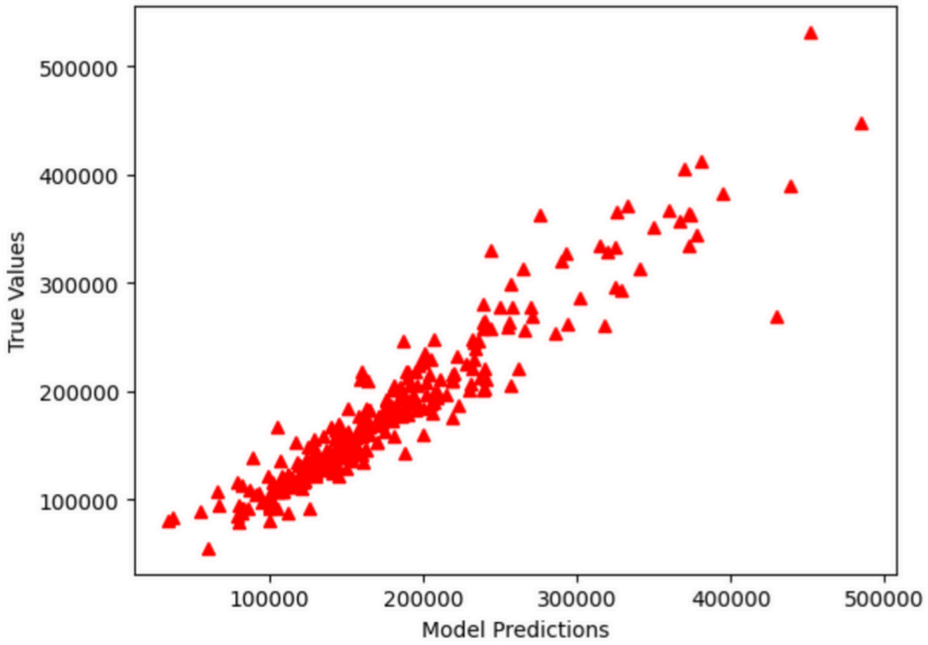
Pada Project ini kami menggunakan 2 Model

1. XGBOOST REGRESSOR

2. SUPPORT VECTOR REGRESSION

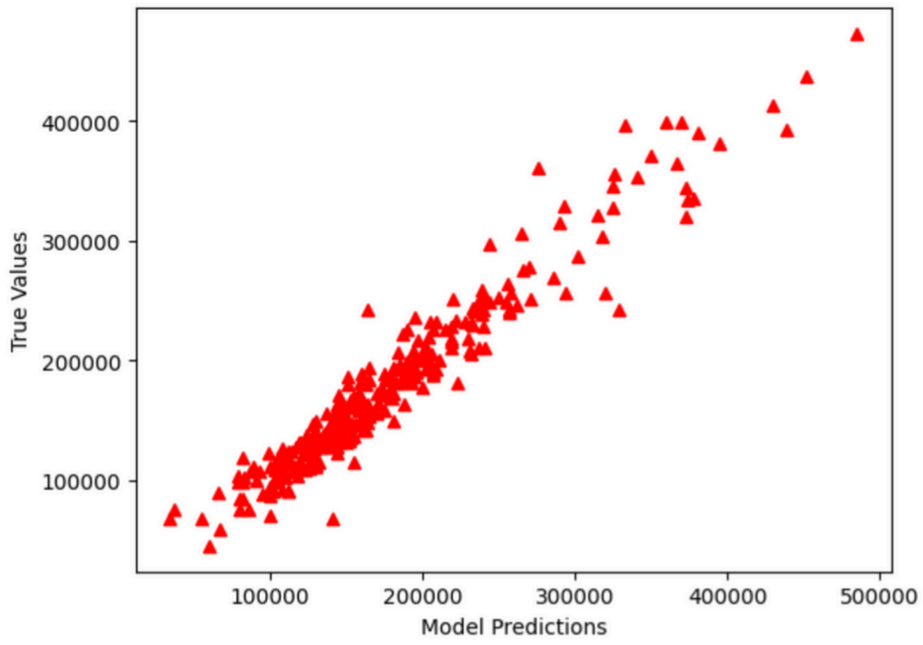
Modelling XGBOOST

BEFORE GRIDSEARCH



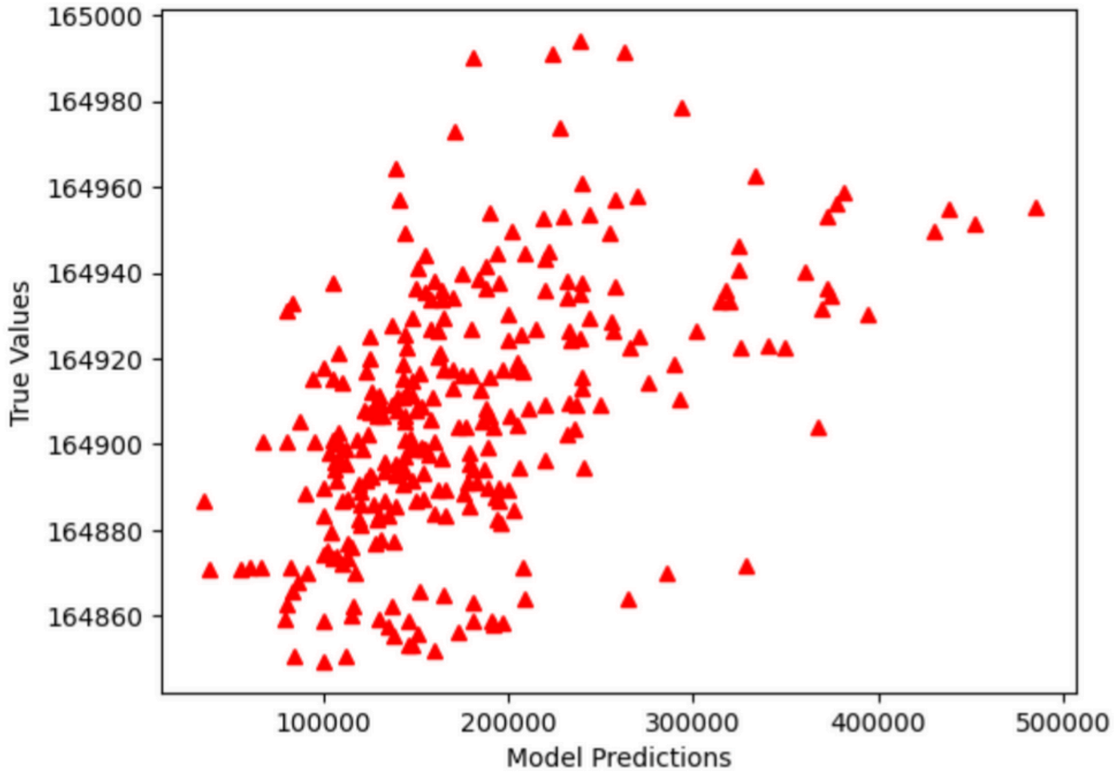

```
RMSE = 23574.021
MSE = 555734500.0
MAE = 15924.144
R2 = 0.9029815196990967
Adjusted R2 = 3.5665797970511695
```

AFTER GRIDSEARCH



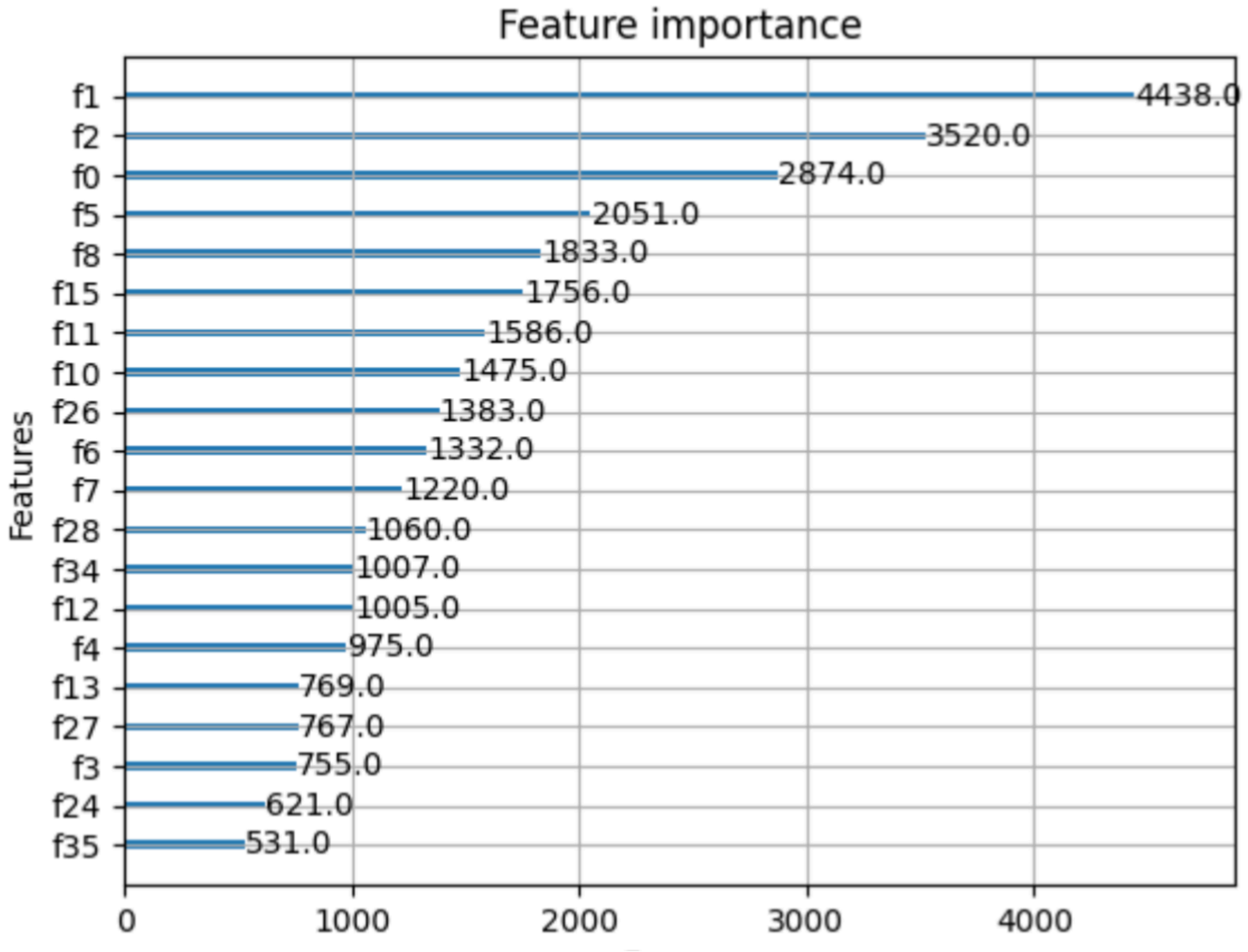

```
RMSE = 19438.764
MSE = 377865570.0
MAE = 13934.45
R2 = 0.9340333342552185
Adjusted R2 = 2.7451181574301287
```

MODELLING SVR




```
RMSE = 76681.091
MSE = 5879989724.456878
MAE = 55048.02728110719
R2 = -0.026511244663566913
Adjusted R2 = 28.155888381554362
```

SUMMARY



F0 :MSSubClass
F1:LotFrontage
F2: LotArea
F5:YearBuilt
F8: BsmtFinSF1

SUMMARY

-Perbandingan dari ke 3 Model ini terlihat cukup signifikan, XGBOOST sebelum menggunakan GRIDSEARCH memiliki performa yang lebih baik.

-Meskipun komputasi GRIDSEARCH cukup lama namun metode ini sangat membantu untuk menemukan parameter terbaik dan meningkatkan akurasi pada model XGBOOST

-Dari hasil feature Importance terlihat bahwa model dapat mengidentifikasi faktor apa saja yang berpengaruh terhadap harga jual rumah dan hasil ini sejalan dengan keadaan bisnis yang ada.