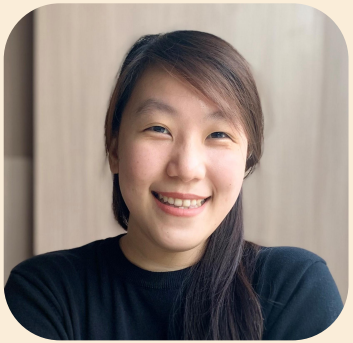
JCDSVL-05

PREDICTING PROPERTY VALUES IN PHILADELPHIA

Created By : BETA ENGINEERS

BETA ENGINEERS PROFILE

Yehezkiel Gabriel



Yohanna Inawati

Risdan Kristori

01

Problem Understanding

OPA

City of Philadelphia

Office of Property Assessment (OPA)

Determines what every piece of property within Philadelphia is worth.

Property Value Determination

1

Penentuan nilai properti harus **Adil**, **Akurat** dan **Dapat Dipahami**.

2

Harga Penjualan dan **Karakteristik Properti** akan dipertimbangkan untuk penentuan nilai properti.

Property Value Purposes

Manfaat Nilai Properti

1. Dasar pertimbangan masyarakat untuk menentukan harga transaksi properti.
2. Menentukan nilai pajak yang harus dibayarkan.

Hubungan Terhadap Pajak

1. Pajak digunakan untuk pendanaan sekolah umum.
2. Pajak terlalu tinggi akan mempengaruhi tarif pajak yang harus dibayarkan masyarakat.
3. Pajak terlalu rendah akan mempengaruhi harga transaksi properti.

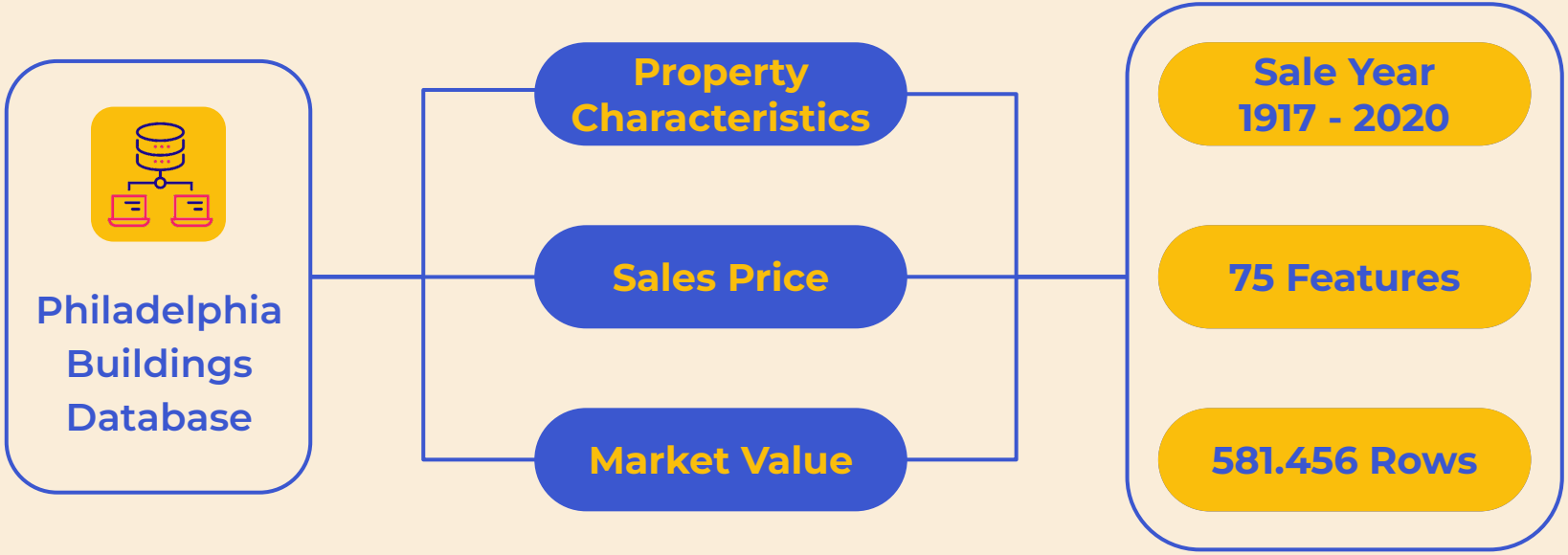
Problem Statements

1. Tipe properti apa saja yang memerlukan perhatian lebih dalam pelaksanaan quality control terhadap nilai propertinya?
2. Tipe zoning properti apa yang memiliki jumlah terbanyak dan memiliki nilai properti tertinggi di Kota Philadelphia?
3. Variabel apakah yang paling berpengaruh terhadap nilai properti di Kota Philadelphia?
4. Model machine learning apakah yang dapat membantu OPA dalam memprediksi nilai properti dengan baik?

02

Data Understanding

Dataset Information



Drop Features

Null Value > 50%

UNIQUE

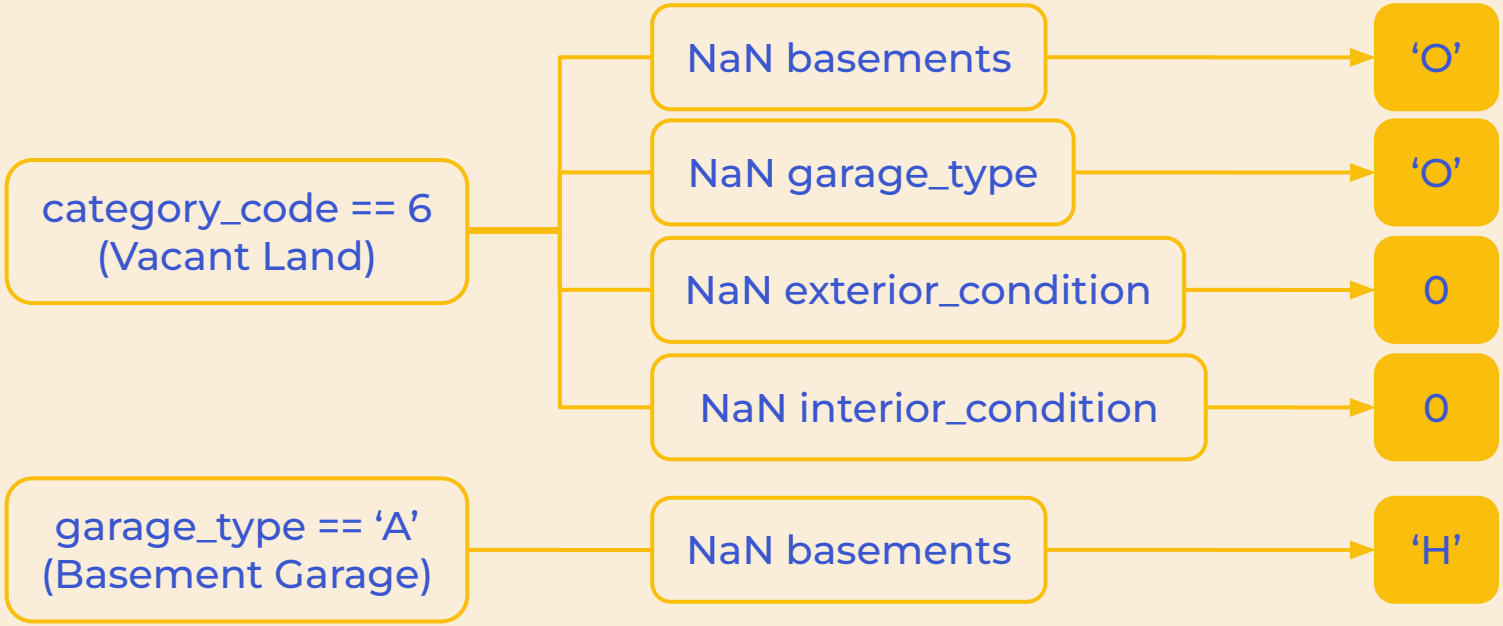
Date Features

Processed from Market Value

Description of Other Features

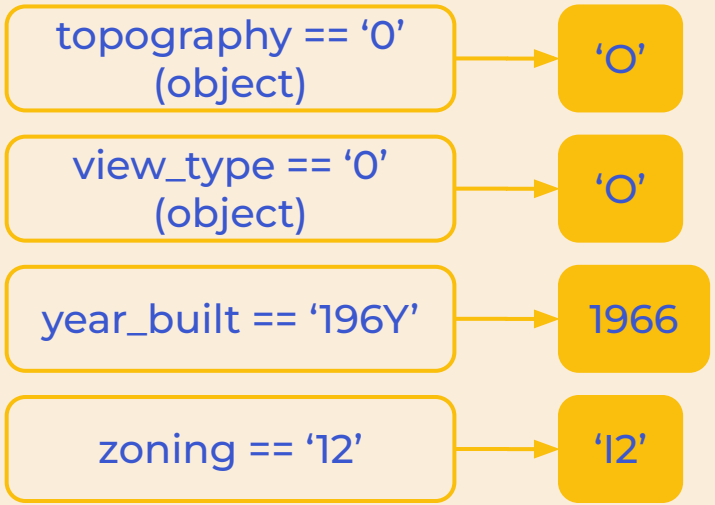
Data Cleaning (1)

Missing Value



Data Cleaning (2)

Change Value

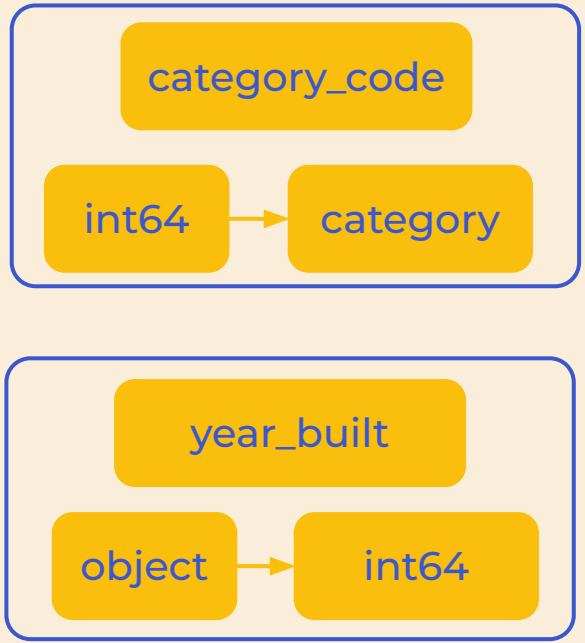


Drop Rows

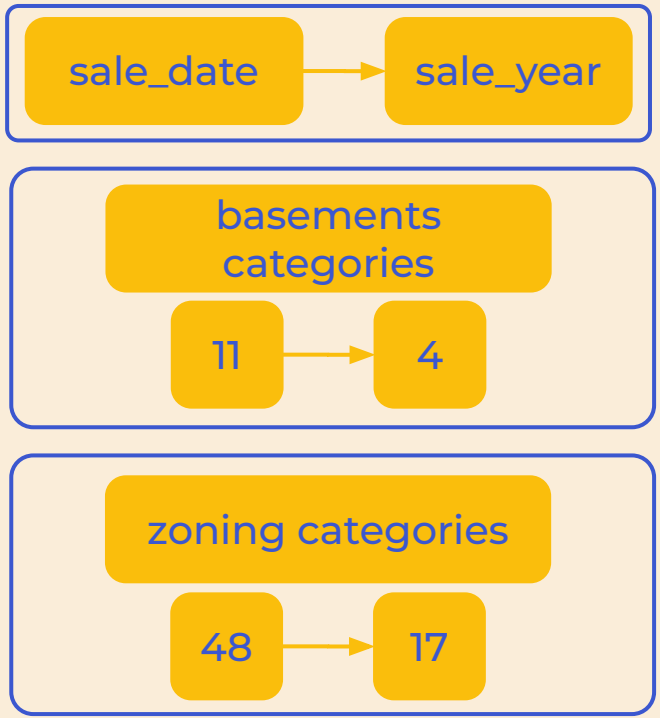
- market_value == 0
- sale_price <= 1
- total_area < 5
- year_built == 0
- zoning NaN
- sale_year < 1977

Data Cleaning (3)

Data Types



Data Manipulation

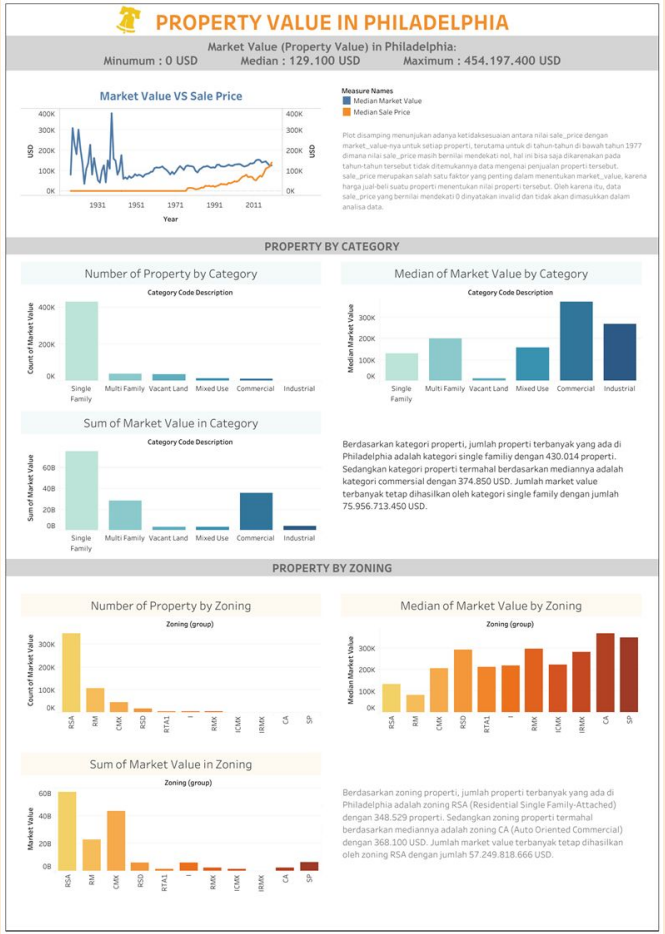


Used Dataset

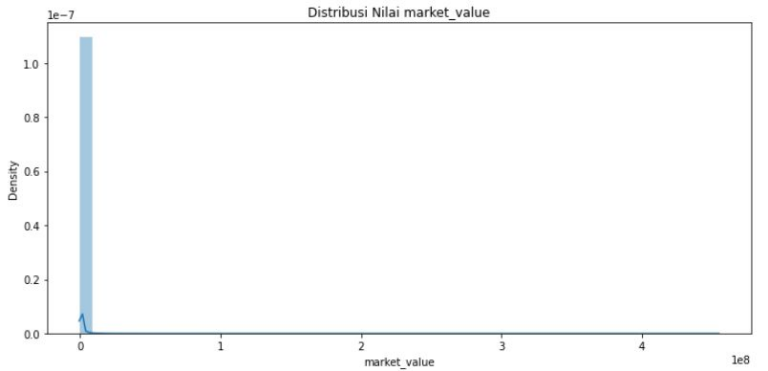
Sale Year	1977 - 2020
Features	before cleaning 75, after cleaning 28
Rows	before cleaning 581.456 after cleaning 355.154 (61% of data)
Target	market_value

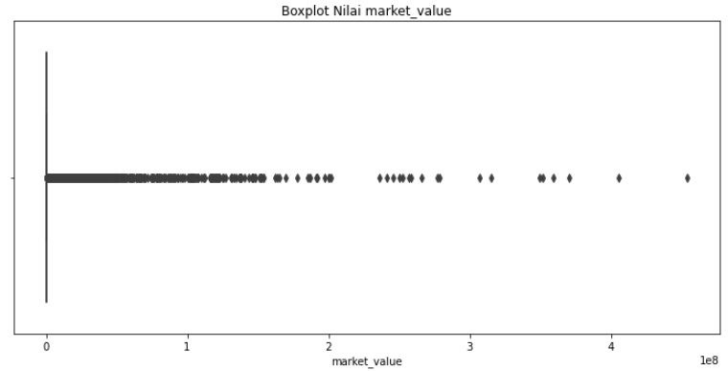
03 Findings and Solutions

Tableau Dashboard



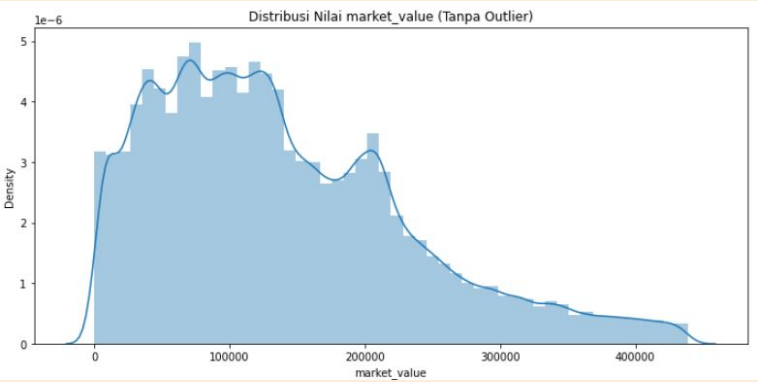
Distribusi Market Value

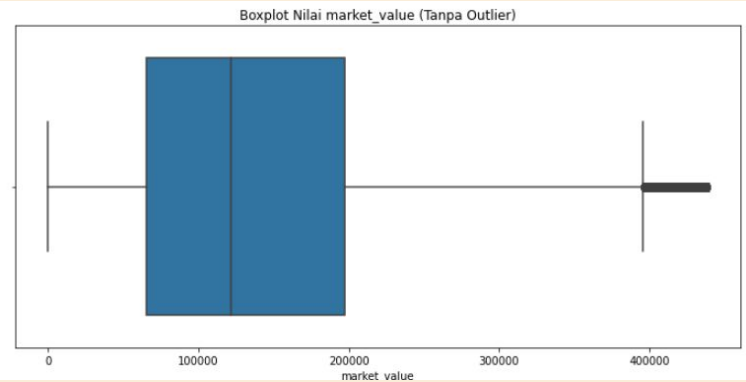




Nilai outlier belum tentu merupakan nilai anomali, bisa jadi dikarenakan fitur yang dimiliki properti tersebut memang menghasilkan harga yang sangat mahal (Contoh : Stadion).

Distribusi Market Value (dengan mengeluarkan nilai outlier)



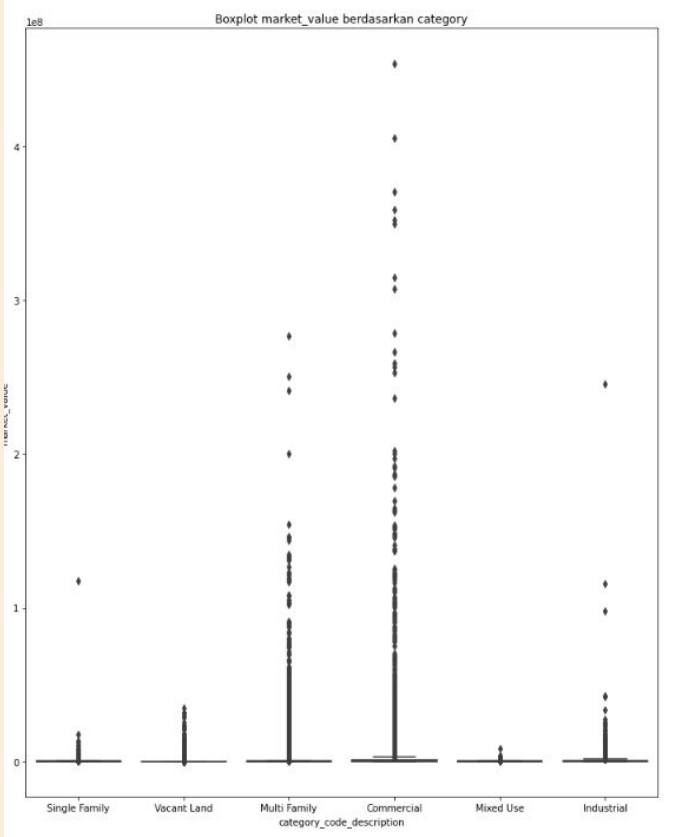


Berdasarkan plot distribusi market_value, dapat dilihat bahwa nilai properti cenderung condong ke kanan (positively skewed).

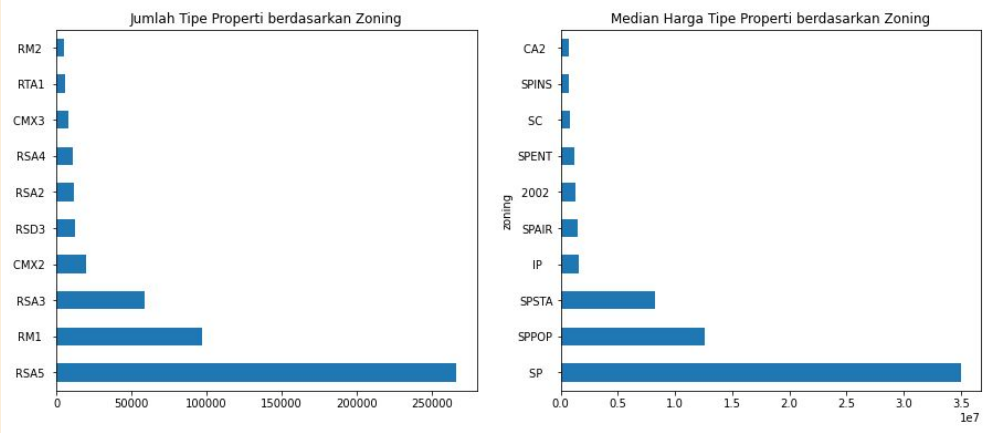
Persebaran Market Value Berdasarkan Category

Properti dengan kategori commercial memiliki nilai-nilai outlier yang tinggi daripada properti dengan kategori lain, hal ini wajar karena sifat properti kategori tersebut yang memiliki nilai bisnis.

Properti pada kategori vacant land dan mixed use memiliki harga cenderung lebih rendah.



Jumlah dan Market Value berdasarkan Zoning



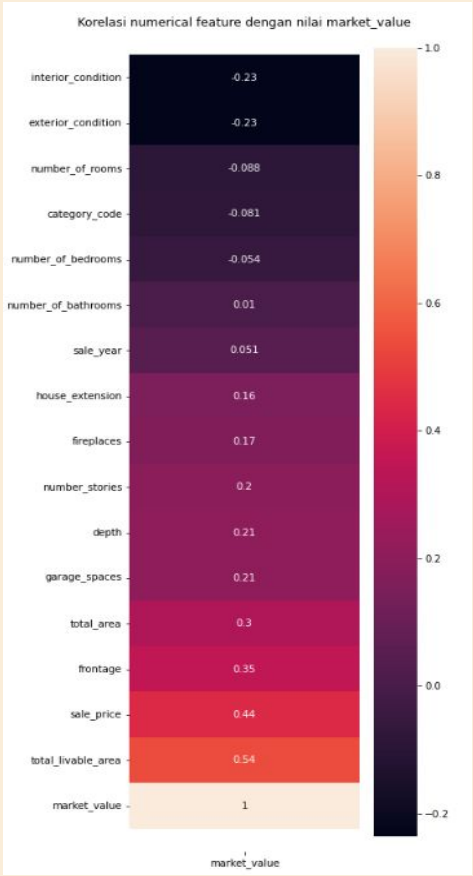
Mayoritas adalah tipe Residential (perumahan) dan diikuti dengan tipe Commercial. Sedangkan properti dengan median harga yang paling tinggi adalah Special Purpose (SP) dan Port Industrial.

Nilai Korelasi Antara Feature Numerik dengan Market_value

Beberapa fitur numerikal yang berkorelasi tinggi terhadap nilai properti adalah:

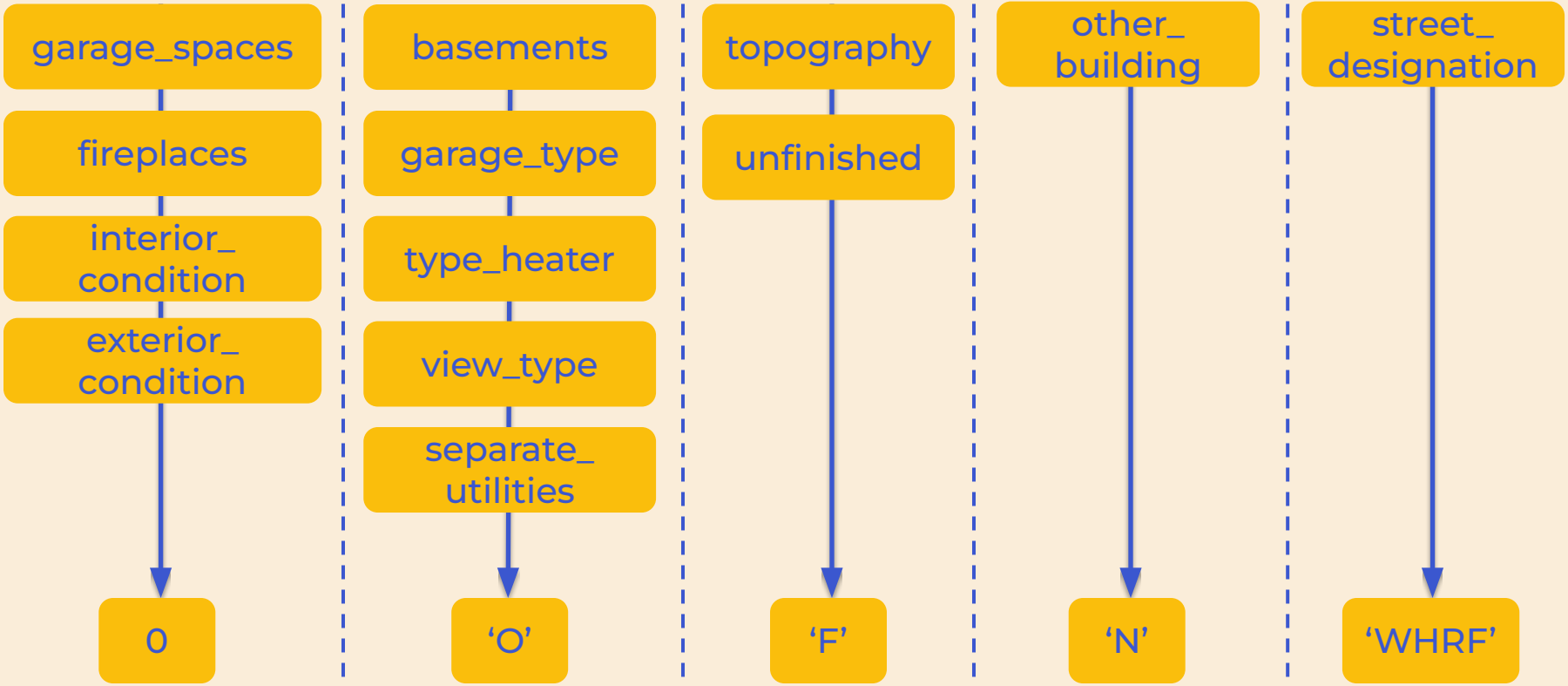
- total_livable_area (0.54)
- sale_price (0.44)
- frontage (0.35)
- total_area (0.30)

Fitur interior_condition dan exterior_condition juga memiliki korelasi yang cukup baik, tetapi bernilai negatif, hal ini dikarenakan kedua fitur tersebut mengurutkan nilai terbaik dari 1 hingga terburuk 9.

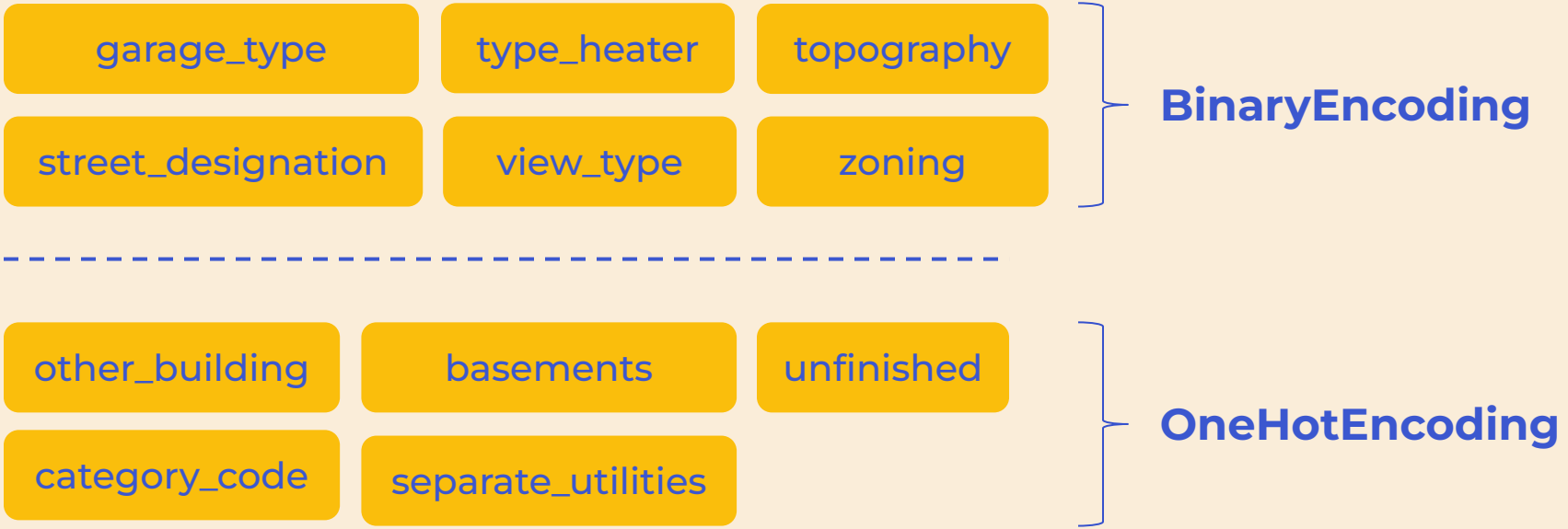


Pembuatan Model Machine Learning

Preprocessing (Imputer)



Preprocessing (Encoding)



Preprocessing (Scaler)

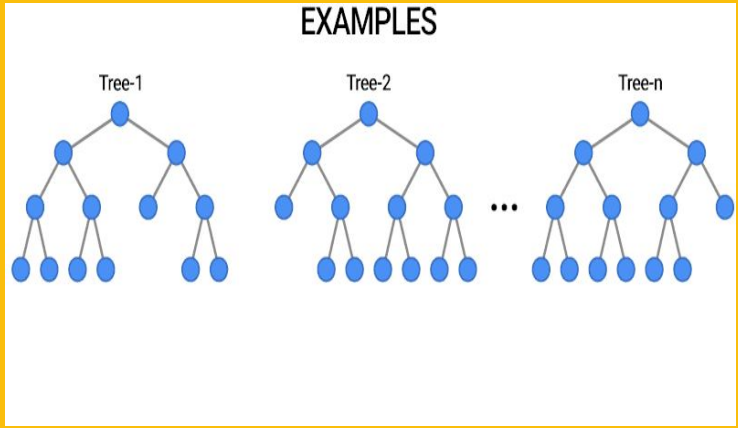
Modeling (Benchmark)

Pada tahap pemilihan benchmark model, kami memilih 3 model **(Linear Regression, Random Forest Regressor, XGBoost Regressor)** yang akan masuk ke tahap cross validation dan akan dinilai berdasarkan nilai MAPE.

	Mean MAPE	Std MAPE
Linear Regression	-1.182605e+14	2.365211e+14
Random Forest Regressor	-0.204273	0.002810
XGBoost Regressor	-0.240156	0.002057

Random Forest

Random Forest bekerja secara bagging dengan membuat beberapa tree dari sejumlah dataset dengan metode bootstrap yang selanjutnya akan menghasilkan prediksi berdasarkan majority vote.



XGBoost

Extreme Gradient Boosting menggabungkan beberapa set pembelajar (tree) yang lemah menjadi sebuah model yang kuat sehingga menghasilkan prediksi yang kuat.

Modeling (Predict to Test set)

	MAPE	R-Squared
Random Forest Regressor	0.238641	0.780251
XGBoost Regressor	0.361665	0.782011

Kemudian berikutnya akan dilakukan hyperparameter tuning untuk keduanya, sebagai bahan pertimbangan pemilihan model.

Hyperparameter Tuning

```
# XGBOOST

# Maximum depth of a tree
max_depth = list(np.arange(2, 30))

# Step size shrinkage used in update to prevents overfitting
learning_rate = list(np.arange(1, 100)/100)

# Number of gradient boosted trees
n_estimators = list(np.arange(100, 201))

# Subsample ratio of the training instances
subsample = list(np.arange(1, 10)/10)

# Subsample ratio of columns for each level
colsample_bylevel= list(np.arange(1, 10)/10)

# L1 regularization term on weights
reg_alpha = list(np.logspace(-3, 1, 10))
```

Percobaan pertama :
Random Forest Regressor menggunakan Randomized Search dengan n_iter = 25.
Hasil pengujian pada dataset test tidak menghasilkan score yang lebih baik.

Percobaan kedua :
XGBoost Regressor menggunakan Randomized Search dengan n_iter = 50.

Hyperparameter Tuning

Best_score: {**0.24369049364506817**}

Best_params:
{'model__subsample': **0.7,** 'model__reg_alpha': **0.007742636826811269**, 'model__n_estimators': **199**, 'model__max_depth': **27**, 'model__learning_rate': **0.28**, 'model__colsample_bylevel': **0.9**}

Predict to test set

	MAPE	R-Squared
XGBoost Regressor	0.231748	0.783908

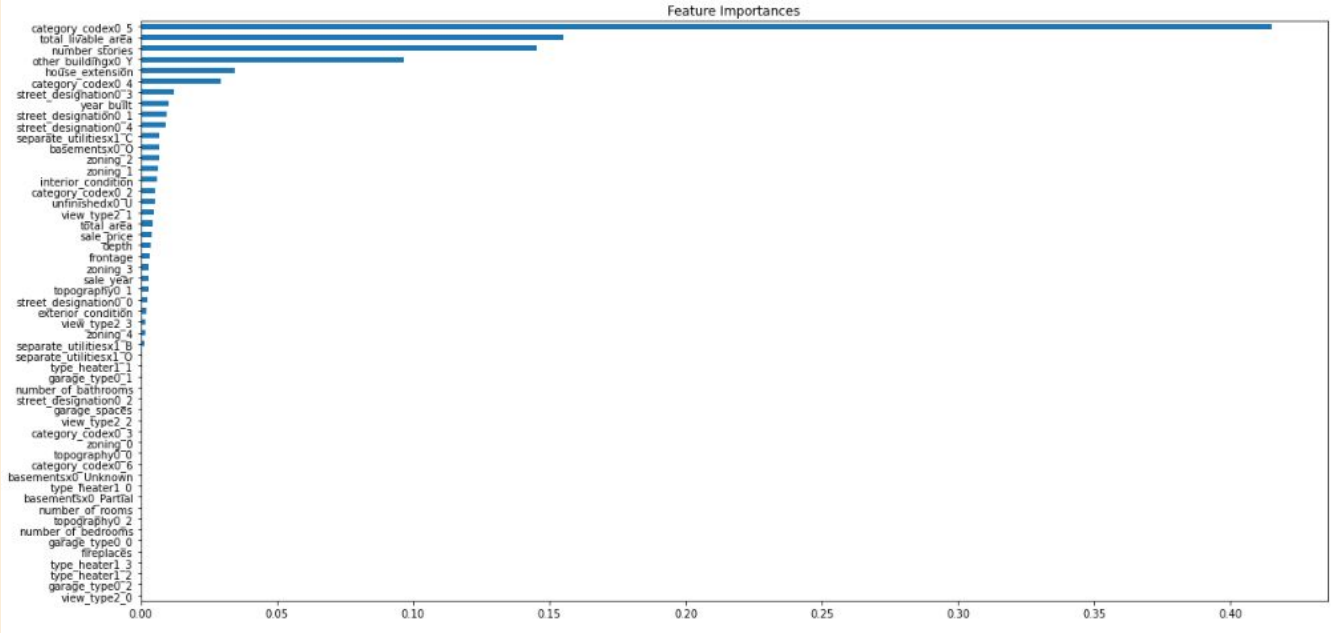
Comparison Score

	MAPE	R-Squared
XGBoost Regressor Before Tuning	0.361665	0.782011
RandomForest Regressor Before Tuning	0.238641	0.780251
XGBoost Regressor After Tuning	0.231748	0.783908

XGBoost after tuning memiliki nilai MAPE dan R-squared yang paling baik.

Oleh karena itu, model yang dipilih untuk masuk ke tahap features importance adalah XGBoost after tuning.

Features Importance



Beberapa feature yang paling penting adalah category_code (x0_5) dengan nilai 0.414936, diikuti oleh total_livable_area, number_stories, dan other_building (x0_Y).

Features Importance (2)

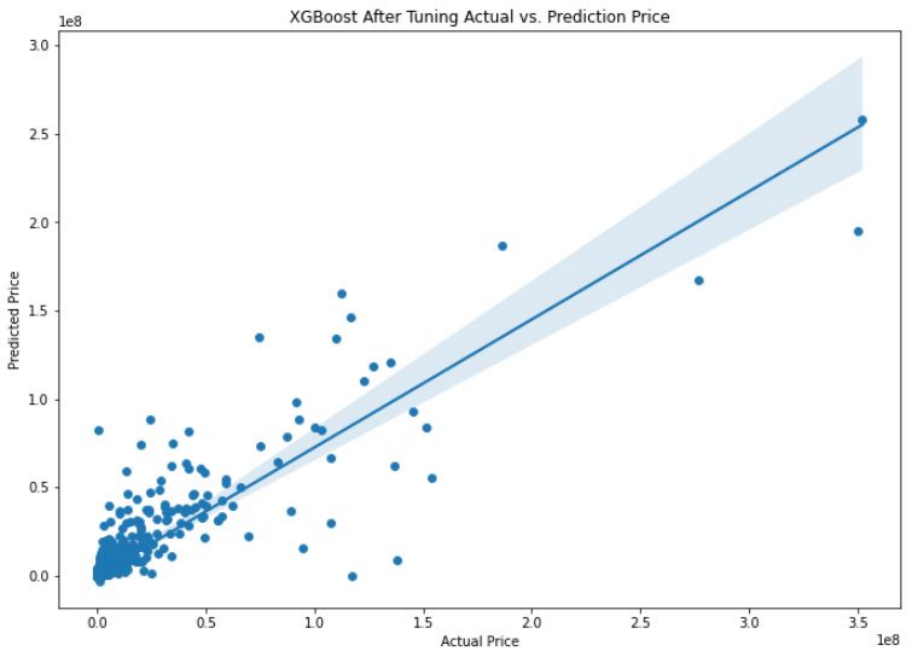
Feature yang dianggap tidak terlalu penting adalah :

1. garage_type
2. type_heater
3. fireplaces
4. number_of_bedrooms
5. number_of_rooms
6. view_type
7. topography

Berikutnya akan dicoba menjalankan ulang model XGBoost Regressor after tuning dengan feature selection. Hasil dari feature selection menunjukkan bahwa penerapan feature selection tidak menghasilkan score yang lebih baik.

No.	Drop Features	MAPE Score	RSquare Score
1	type heater, fireplaces, number_of_bedrooms, number_of_rooms (4 Feature)	0.241160	0.788057
2	type heater, fireplaces, number_of_bedrooms, number_of_rooms, garage_type (5 Feature)	0.239830	0.788639
3	type heater, fireplaces, number_of_bedrooms, number_of_rooms, garage_type, view_type (6 Feature)	0.239771	0.775811
4	type heater, fireplaces, number_of_bedrooms, number_of_rooms, garage_type, view_type, topography (7 Feature)	0.241250	0.778274

Final Model



Final model yang dipilih adalah XGBoost Regressor after tuning dengan score sebagai berikut:

	MAPE	R-Squared
XGBoost Regressor	0.231748	0.783908

04

Conclusion and Recommendation

Conclusion

- Properti berkategori **commercial, multi family, dan industrial** merupakan kategori properti yang memiliki nilai outlier terbanyak pada market_value nya.
- Tipe properti yang memiliki jumlah terbanyak di Kota Philadelphia adalah tipe zoning **RSA** (Residential Single Family) dan **RMA** (Residential Multifamily). Sedangkan, tipe properti yang memiliki nilai median market_value tertinggi adalah tipe zoning **Special Purpose**.
- Berdasarkan nilai korelasi spearman, 3 feature numerik yang memiliki korelasi tertinggi terhadap market_value adalah **total_livable_area, sale_price, dan frontage**.
  Sementara, berdasarkan model yang dibangun, 3 feature yang memiliki pengaruh paling dominan terhadap market_value adalah **category_code, total_livable_area, dan number_stories**.

Conclusion

- Model yang telah dibangun memiliki score **MAPE sebesar 23%**. yang berarti ketika model yang dibuat digunakan untuk memprediksi nilai properti pada rentang nilai seperti yang dilatih terhadap model **(market value: 1,300 USD - 352,143,800 USD)**, maka hasil prediksi yang dihasilkan oleh model memiliki kemungkinan tingkat kesalahan sebesar **23%** dari nilai aslinya. Dengan nilai MAPE sebesar 23% menjadikan model ini sebagai model yang menghasilkan nilai prediksi yang wajar.
- Model yang telah dibangun memiliki score **R-squared sebesar 78,3%** yang berarti model yang telah dibangun mampu menjelaskan faktor-faktor yang mempengaruhi market_value sebesar 78,3%.

Recommendation

Bagi pemerintah Kota Philadelphia:

- Dengan nilai market_value yang telah ditentukan, maka pemerintah Kota Philadelphia dapat memprediksi pendapatan yang dihasilkan dari pajak properti pada Kota Philadelphia. Hal tersebut dapat menjadi acuan bagi pemerintah untuk membuat anggaran Kota Philadelphia khususnya pada bidang pendidikan untuk periode selanjutnya.
- Dengan menetapkan nilai pajak properti berdasarkan nilai market_value sebuah properti yang telah sesuai dengan keadaan setiap properti, pemerintah dapat memperkuat kebijakan terkait pembayaran pajak serta mendorong masyarakat untuk taat membayar pajak. Hal tersebut diharapkan dapat meningkatkan persentase tingkat pembayaran pajak tahunan.

Recommendation

Bagi OPA:

- Berdasarkan hasil dari model yang telah dibangun, OPA dapat menjelaskan kepada masyarakat Kota Philadelphia mengenai faktor-faktor yang mempengaruhi nilai dari masing-masing properti sehingga masyarakat dapat memahami apakah properti miliknya memiliki nilai yang lebih rendah/tinggi dari yang seharusnya.
- Dalam melakukan quality control, OPA dapat memberikan perhatian lebih terhadap properti yang berkategori commercial, multi family, dan industrial karena kategori tersebut cenderung memiliki outlier yang tinggi.

Bagi Investor / Pelaku Bisnis:

- Nilai market value dapat digunakan sebagai acuan karakteristik penduduk pada area tersebut, sehingga Investor/pelaku bisnis dapat menentukan lokasi bisnis yang sesuai dengan target market mereka.

Recommendation

Untuk Model selanjutnya:

- Hasil benchmark model RandomForest menunjukkan hasil yang lebih baik dibandingkan dengan XGBoost. Penerapan Hyperparameter dengan parameter yang lebih beragam pada model Random Forest mungkin saja dapat menghasilkan model dengan score yang lebih baik.
- Menambahkan feature-feature baru yang dapat menjelaskan market_value khususnya pada properti dengan kategori industrial, commercial, dan multifamily.
- Score yang lebih baik mungkin dapat dicapai dengan menggunakan algoritma machine learning yang lain. Sehingga, perlu adanya prediksi menggunakan model regresi selain Linear Regression, Random Forest Regressor, dan XGBoost Regressor.

THANK YOU !

“LIFE IS ABOUT THE PEOPLE YOU MEET AND THE AWESOME THINGS YOU CREATE WITH THEM.”

Fellexandro Ruby