

DC PROPERTIES

TEAM OMEGA









Faykel Nicandro Hattu

- in linkedin.com/in/faykel-nicandro-hattu/
- github.com/faykelnicandro



Muhammad Rafi Amiruddin

- in linkedin.com/in/muhammad-rafi-amiruddin/
- github.com/Muhrafix/



Tamara Coglitore

- in linkedin.com/in/tamara-coglitore/
- github.com/tamaracoglitore/





TABLE OF CONTENTS





BUSINESS PROBLEM



DATA
UNDERSTANDING &
PREPROCESSING



PRICE PREDICTION



PROPERTY SEGMENTATION



CONCLUSION & RECOMMENDATION



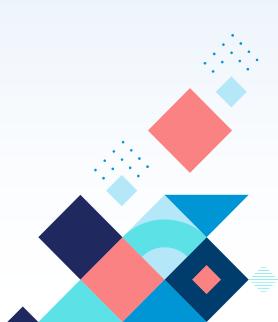
INTRODUCTION

Tempat tinggal merupakan salah satu kebutuhan utama untuk manusia. Harga properti berbanding lurus dengan prospek area yang memiliki aksesibilitas tinggi, terutama di kota-kota besar.





BUSINESS PROBLEM



BACKGROUND

- Washington DC merupakan Ibu Kota sekaligus kota pemerintahan di Amerika.
- Biaya hidup dan harga tempat tinggal cukup tinggi (rata-rata USD 647.000/tempat tinggal).
- Harga properti turut diatur oleh pemerintah agar tidak mengganggu . stabilitas ekonomi negara.





Banyak sekali perusahaan Real Estate yang berlomba untuk memenangkan pasar properti di Washington, salah satunya adalah OMEGA CORP. Dengan persaingan yang cukup tinggi, perusahaan harus mampu bersaing dengan harga yang sesuai dengan aturan pemerintah dan dengan pelayanan yang efektif.

Namun ada beberapa masalah yang dimiliki oleh perusahaan yaitu:

Subjektivitas surveyor yang dapat menyebabkan Underpriced dan Overpriced

Survey memakan waktu yang lama dan biaya yang tinggi

Proses marketing yang kurang efisien jika tidak ada segmentasi properti

Permasalahan tersebut dapat ditangani dengan lebih efisien menggunakan Machine Learning.

MENGAPA MACHINE LEARNING?



Tanpa Machine Learning:

- Subjektivitas surveyor dapat menyebabkan
 Overpriced dan Underpriced yang bisa menimbulkan kerugian kedua belah pihak
- 2. Ada biaya untuk melakukan survey, yang juga memakan waktu cukup lama
- 3. Proses marketing tidak tepat sasaran, menyebabkan proses marketing yang lama3. dan adanya potensi kehilangan customer

Dengan Machine Learning:

- Menghilangkan subjektivitas dalam penentuan harga sehingga mampu meminimalisir adanya Overpriced dan Underpriced, dengan regulasi pemerintah sebagai acuan
- 2. Mampu memangkas biaya survey dan waktu survey
- 3. Proses marketing lebih tersegmentasi yang menyebabkan meningkatnya efektivitas perusahaan



PRICE PREDICTION

- Membuat Model Machine Learning Regresi untuk price prediction yang memiliki Mean Absolut Error (MAE) serendah-rendahnya tanpa mengabaikan R-Squared atau Goodness of Fit.
- Mean Absolut Error merupakan rata-rata error yang didapatkan. Semakin rendah nilai MAE menunjukkan semakin rendah error pada model.
- R-Squared atau Goodness of Fit adalah nilai kesesuaian dari hasil prediksi dengan nilai aslinya. Hal ini menunjukkan akurasi model terhadap nilai sesungguhnya.

PROPERTY SEGMENTATION

• Membuat Model Machine Learning Clustering untuk segmentasi properti dengan menggunakan jumlah cluster sesuai dengan Silhouette Score tertinggi pada cluster di setiap model.











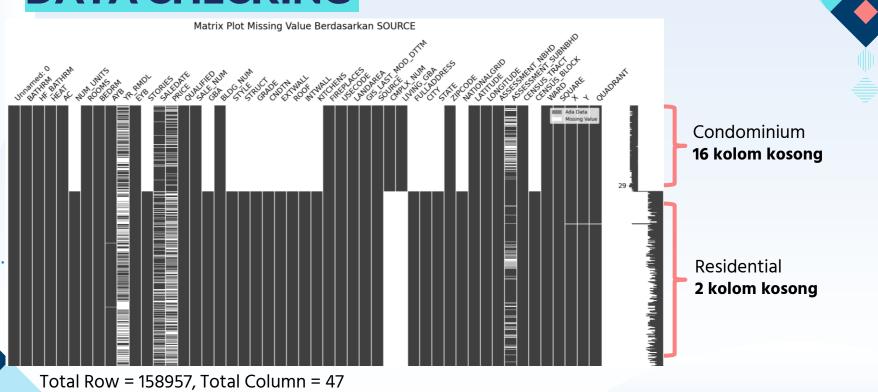
DATA UNDERSTANDING & PREPROCESSING

DATA SOURCE

Data yang digunakan diunduh dari <u>kaggle</u>. Data tersebut merupakan gabungan 3 data yang masing-masing mengandung informasi fitur residential, fitur condominium dan lokasi properti.



DATA CHECKING



DATA CHECKING

SORT BY SOURCE

JUMLAH DATA



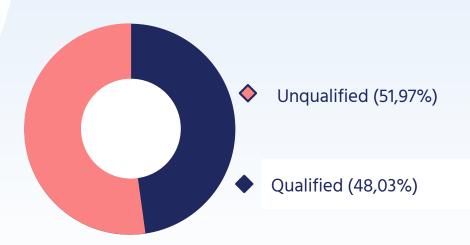
- Residential (67,12%)
- Condominium (32,88%)

- Dari perbedaan fitur berdasarkan missingness tersebut, dapat disimpulkan bahwa kedua tipe properti memiliki fitur yang berbeda
- Kedua tipe properti dinilai lebih baik memiliki model yang berbeda agar hasil prediksi lebih akurat
- Tipe Residential dipilih untuk digunakan dalam penelitian



DATA CHECKING

QUALIFIED



- Value pada kolom ini merupakan hasil penilaian pemerintah terhadap harga properti
- Dengan spesifikasi rumah yang hampir identik pada data Qualified dan Unqualified, perbedaan harga cukup jauh
- Data yang akan digunakan pada analisa hanya data yang Qualified berdasarkan pemerintah



U	Inq	ua	lif	ied

АУВ	WARD	GRADE	CNDTN	HEAT	AC	GBA	LANDAREA	QUALIFIED	PRICE
1800.0	Ward 2	Above Average	Good	Warm Cool	Υ	2520.0	1853	U	339500.0
1800.0	Ward 2	Above Average	Average	Warm Cool	Υ	2610.0	1971	Q	846000.0

Qualified

Dengan fitur yang hamper identik, perbedaan harga mencapai US\$506500



DATA DESCRIPTION





- BATHRM : Number of Bathroom
- HF_BATHRM : Number of Half Bathroom
- NUM_UNITS : Number of Unit
- ROOMS: Number of Room
- BEDRM : Number of Bedroom
- AYB: Year the property was build
- EYB: Most recent year the property have minor renovation
- GBA: Gross Building Area (square feet)
- SALEYEAR : Most recent year the property sold

- BLDG_NUM : Number of Building
- KITCHENS: Number of Kitchen
- FIREPLACES : Number of Fireplace
- LANDAREA: Land surface area (square feet)
- USECODE : Numeric code based on property function
- ZIPCODE: Numeric code based on location
- LATITUDE : Latitude
- LONGITUDE : Longitude
- SALEMONTH: Most recent month the property sold





CATEGORICAL

- HEAT : Type of Heater
- AC : AC Availability
- STYLE : House Style/Stories
- STRUCT : House Structure
- EXTWALL: Exterior Wall Structure and Material
- ROOF: Roof Material and Structure
- INTWALL: Interior Wall Structure and Material
- WARD : Location Segmentation in District

 QUADRANT : City Quadrant (Northwest, Southwest, Northeast, Southeast)

ORDINAL CATEGORIC

- GRADE: Property Grade
- CNDTN : Property Condition
- YR_RMDL: Most recent range year the property have a Structural Renovation

TARGET PREDICTION

PRICE : Property value in USD



- REMOVING UNUSED AND IRRELEVANT FEATURES
- HANDLING DATA TYPE
 SALEDATE, USECODE, ZIPCODE
- GENERATE NEW FEATURES

 SALEMONTH: Bulan dari kolom SALEDATE

 SALEYEAR: Tahun dari kolom SALEDATE
- REMOVING TYPO
 YR RMDL
- REMOVING COLUMNS WITH MISSING VALUE
 - ASSESSMENT_SUBNBHD

- REMOVING ROWS WITH MISSING VALUE < 1%
 <p>PRICE, BATHRM, NUM_UNITS, ROOMS, BEDRM,
 AYB, KITCHENS, FULLADDRESS, NATIONAL GRID,
 CENSUS_BLOCK, QUADRANT
- FILLING DATA ANOMALY WITH MODE
 AC, HEAT, EXTWALL, INTWALL, STRUCT, STYLE
- REMOVING ROWS WITH DATA ANOMALY
 - AYB > SALEYEAR
 - AYB > EYB
 - YR_RMDL > 2018
- REMOVING ROWS WITH IRRATIONAL VALUE PRICE



- BINNING COLUMN WITH MISSING VALUE YR_RMDL
- ONE HOT ENCODING AC, QUADRANT
- ORDINAL ENCODING
 YR_RMDL, GRADE, CNDTN
- BINARY ENCODING

 ZIPCODE, WARD, USECODE, HEAT, STYLE,

 STURCT, EXTWALL, ROOF, INTWALL
- SIMPLE IMPUTER WITH MODE
 AC, HEAT, EXTWALL, INTWALL

- SPLITTING DATA
 - Target = PRICE
 - Data train 80%, data test 20%







PRICE PREDICTION





1. DATA ANALYSIS





TREN HARGA PROPERTI



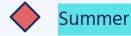


PRICE ANALYSIS







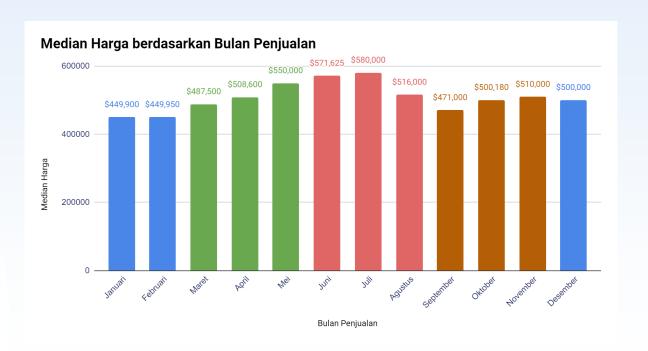








PRICE ANALYSIS











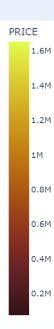




PRICE ANALYSIS







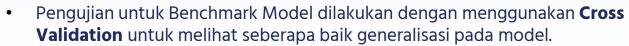


2. REGRESSION MODELING



BASE MODEL

	Model	MAE	МАРЕ	R-squared
0	Linear Regression	122861.016884	0.215980	0.702317
1	KNN	251120.005164	0.537944	0.440221
2	Decision Tree	117040.363544	0.194586	0.809011
3	Random Forest	77243.563093	0.128169	0.906752
4	Gradient Boosting	86234.019718	0.148705	0.892287
5	XGBoost	72148.333202	0.122755	0.919081



- Dengan menggunakan **data train**, model Random Forest dan XGBoost memiliki nilai MAE yang **paling rendah** dan nilai R-Square yang sama-sama **baik**.
 - Kedua model ini akan dilanjutkan ke pengujian menggunakan data test.







BASE MODEL

	MAE	MAPE	R-squared
XGBoost	74854.303014	0.131614	0.878719
Random Forest	77432.540281	0.129647	0.881790

- Dengan menggunakan **data test**, model Random Forest dan XGBoost memiliki nilai MAE dan R-Squared yang **cukup baik**.
- Kedua model ini akan dilanjutkan ke pengujian Hyperparameter Tuning untuk melihat parameter terbaik masing-masing model.



Hyperparameter Tuning dilakukan dengan menggunakan metode GridSearchCV agar mendapatkan hasil terbaik dengan mengeksplor semua parameter yang diuji.

Parameter yang digunakan untuk tahap ini antara lain:

1. Random Forest

Parameter	Value	Best Value
n_estimator	200, 250, 280, 300	300
max_depth	13, 15, 17	17
min_sample_split	2, 3, 5	2
min_sample_leaf	2, 3, 5	2

2. XGBoost

Parameter	Value	Best Value
n_estimator	100, 200, 250, 300	300
max_depth	2, 6, 9	9
learning_rate	0.01, 0.1, 0.3, 0.5	0.1



	MAE	МАРЕ	R-squared
XGBoost	74854.303014	0.131614	0.878719
Random Forest	77432.540281	0.129647	0.881790
XGBoost after Tuning	66680.164930	0.109857	0.859665
Random Forest after Tuning	77065.492001	0.124159	0.877516



• Analisa selanjutnya akan menggunakan model XGBoost setelah Tuning.

FEATURE SELECTION

Data **TRAIN** dengan Feature Selection

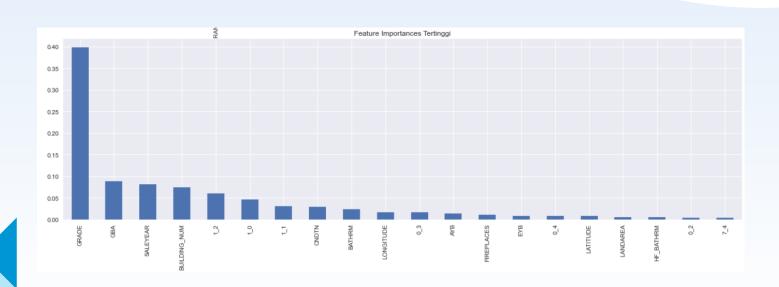
Data **TEST** dengan Feature Selection

Data **TEST tanpa**Feature Selection

Model	MAE	MAPE	R-squared
XGBoost FS	72139.432484	0.124084	0.927662
	MAE	MAPE	R-squared
XGBoostFS	76340.03574	0.135205	0.882712
	MAE	MAPE	R-squared
XGBoost	74854.303014	0.131614	0.878719

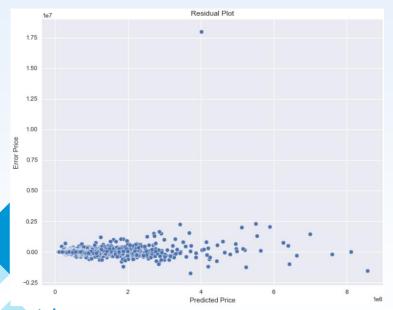
Dengan menggunakan Feature Selection K-Best Univariate dan melakukan beberapa percobaan, didapatkan K terbaik adalah 40 fitur. Nilai untuk data train lebih baik, namun pada data test kurang baik sehingga untuk analisa selanjutnya akan menggunakan data **tanpa** Feature Selection K-Best.

FEATURE IMPORTANCES



Dari Feature Importance Model based, didapatkan feature dengan pengaruh tertinggi ada di kolom GRADE yang diikuti oleh GBA dan SALEYEAR.

MODEL LIMITATION



- Pada harga prediksi dibawah USD 3,25 juta simpangan error masih dianggap baik (mendekati 0)
- Ditemukan kecenderungan peningkatan range error pada data dengan harga prediksi di atas USD 3,25 juta
- Terdapat 1 properti dengan harga outlier yang memiliki error hingga USD 17,5 juta

MODEL LIMITATION



Before Outlier Removal

MAE	MAPE	R-squared
66680.16493	0.109857	0.859665

After Outlier Removal

MAE	MAPE	R-squared
65784.206418	0.113625	0.933461

- Didapatkan hasil data tanpa outlier lebih baik dengan hasil:
 - ✓ R-Squared yang meningkat signifikan
 - ✓ MAE berkurang USD 900
- Model tanpa outlier akan digunakan untuk model fitting
- Dengan adanya kecenderungan peningkatan range error pada prediksi harga diatas USD 3.25 juta, akan dilakukan analisa lebih lanjut



MODEL LIMITATION

 Terdapat kecenderungan kelas 'Exceptional-B' sampai 'Exceptional-D' pada harga di atas USD 3,25 juta yang menunjukkan error yang cukup signifikan terjadi pada kelompok tersebut.

 Harga properti kelas 'Exceptional-B' sampai 'Exceptional-D' belum bisa diprediksi secara optimal menggunakan model ini.

Average	14010
Above Average	13368
Good Quality	9778
Very Good	4328
Excellent	1542
Superior	1246
Exceptional-A	287
Exceptional-B	62
Fair Quality	39
Exceptional-D	7
Exceptional-C	4
Low Quality	1
Name: GRADE, d	ltype: int64

Exceptional-A 82
Exceptional-B 73
Superior 54
Exceptional-C 23
Exceptional-D 15
Excellent 11
Very Good 1
Name: GRADE, dtype: int64

Harga di bawah USD 3,25 juta

Harga di atas USD 3,25 juta







PROPERTY SEGMENTATION_



1. CLUSTER MODELING





DATA PREPARATION

- Fitur yang akan digunakan untuk Clustering adalah GBA, SALEYEAR, GRADE dan PRICE.
- Kolom GBA, SALEYEAR dan GRADE didapatkan dari hasil pengujian Feature Importances.
- Kolom PRICE digunakan karena perusahaan ingin berpatokan terhadap harga/kemampuan financial customer dalam melakukan segmentasi properti.







- REMOVING OUTLIER PRICE, GBA
- ORDINAL ENCODING GRADE
- SCALLING WITH MINMAX SCALLER GBA, SALEYEAR, GRADE
- DATA SAMPLING 20000 SAMPLE USED





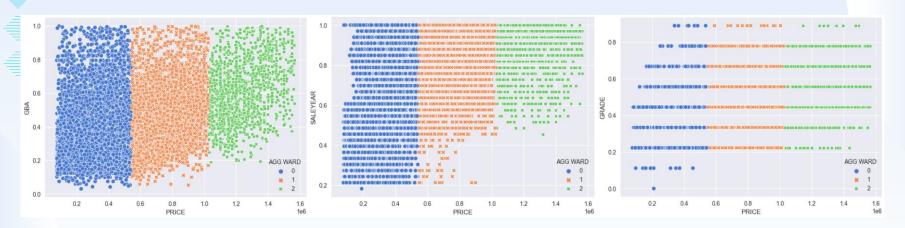


	Score	Jumlah Cluster
K-Means	0.587277	3
Ward	0.601539	3
Complete	0.577466	3
Average	0.598766	3



Setelah melalui percobaan dengan masing-masing model, jumlah cluster terbaik didapatkan dari Model **Agglomerative Ward** dengan Silhouette Score sebesar **0.601** pada **jumlah cluster 3**.

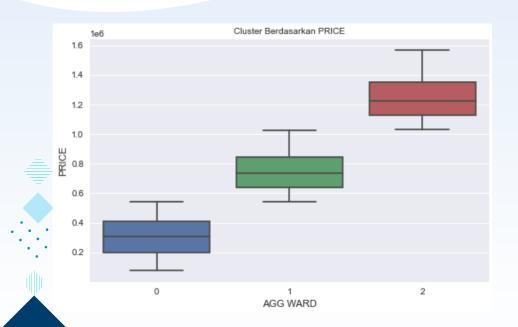
MODELING



Dari Scatterplot diatas, didapatkan bahwa PRICE dapat membagi GBA, SALEYEAR dan GRADE dengan sangat baik, yang ditunjukkan dengan garis perbatasan yang cukup tegas dari masing-masing cluster.

MODELING





Boxplot menunjukkan bahwa setiap cluster memiliki range harga yang berbeda. Sehingga distribusi harga akan digunakan sebagai acuan penentuan cluster dengan penjelasan sebagai berikut

- Cluster 0 = Lower memiliki range harga terendah
- Cluster 1 = Middle memiliki range harga tengah
- Cluster 2 = Upper memiliki range harga tertinggi



2. DATA ANALYSIS





CLUSTER ANALYSIS





- ♦ Lower (56,22%)
 - Middle (36,50%)
- Upper (8,28%)

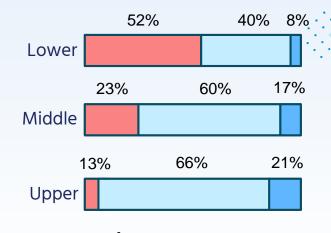


- Location
- Condition
- **Building Area**
- Renovation
 - >>> Year Remodeling
- Features
 - >>> Fireplaces, Half-Bathroom



CLUSTER ANALYSIS LOKASI ◆ Lower Middle Upper





- Average
- Good
- Very Good



CLUSTER ANALYSIS







Middle

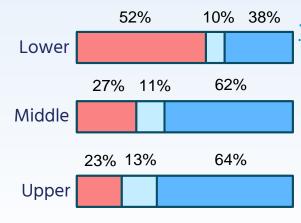
1440 - 1950 sqft



Upper

1800 – 2550 sqft





- No Remodeling
- 1971 2000
- > 2000

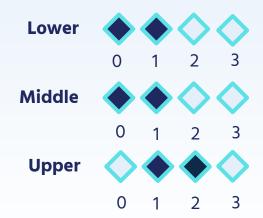




CLUSTER ANALYSIS

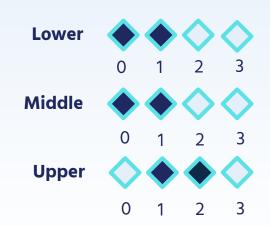


Distribution





Distribution





MODEL LIMITATION

- Model sangat sensitif terhadap data outlier, sehingga data dengan outlier akan dihapus.
- Model memiliki sifat kompleks, sehingga membutuhkan waktu dan device yang mumpuni untuk menangani jumlah data yang besar.





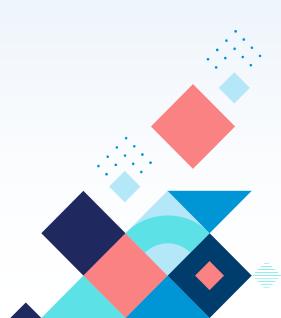




CONCLUSION & RECOMMENDATION



5.1 CONCLUSION





1. PRICE PREDICTION

Model terbaik yang dipilih adalah XGBoost dengan Tuning tanpa Outlier dengan nilai MAE sebesar 65784.21 dan nilai R-Squared sebesar 0.933461.

I. MEMINIMALISIR ERROR

АУВ	WARD	GRADE	CNDTN	HEAT	AC	GBA	LANDAREA	QUALIFIED	PRICE
1800.0	Ward 2	Above Average	Good	Warm Cool	Υ	2520.0	1853	U	339500.0
1800.0	Ward 2	Above Average	Average	Warm Cool	Υ	2610.0	1971	Q	846000.0

- Sebelum pemodelan
 Harga Qualified Harga Unqualified = USD 846000 USD 339500 = **USD 506500**
- Setelah pemodelan MAE = USD 65784.21
- Kerugian yang bisa dihindari adalah = USD 440715.79







1. PRICE PREDICTION

2. MEMOTONG BUDGET SURVEY

Harga rata-rata 1x survey = US\$500

Waktu rata-rata 1x survey = 3 jam + 1 jam untuk kelengkapaan laporan dan perjalanan

Survey/minggu = 40jam/minggu : 4 jam = 10x survey

Survey/tahun = $10 \times 52 \text{ minggu} = 520 \text{ survey}$

Biaya jasa surveyor per tahun = $520 \times 500 = US$260.000$

Perusahaan bisa menghemat sebesar **US\$260.000/surveyor** setiap tahun

3. FITUR-FITUR PENTING DAN LIMITASI

Fitur paling berpengaruh:

• GRADE : Property Grade

GBA : Gross Building Area (square feet)
 SALEYEAR : Most recent year the property sold

Limitasi:

 Harga properti kelas Exceptional B hingga Exceptional D belum bisa diprediksi dengan optimal



2. PROPERTY SEGMENTATION

Model yang dipilih adalah Agglomerative Ward dengan Silhouette Score 0.601 pada jumlah

1. HASIL SEGMENTASI

cluster 3.

CLUSTER	CHARACTERISTICS
LOWER	Tidak memiliki FIREPLACE, luas bangunan berada di rentang 1100-1650 sqft, kondisi properti average
MIDDLE	Telah melakukan proses remodeling, memiliki half bathroom, luas bangunan berada di rentang 1400- 1950 sqft, kondisi properti good
UPPER	Terdapat FIREPLACE, telah melakukan proses remodeling, memiliki half bathroom, luas bangunan lebih dari 1800 sqft, kondisi properti Good & Very Good, berlokasi di kuadran Northwest (Ward 2 & 3)

2. PROPERTY SEGMENTATION

2. BUSINESS APPROACH

- Menurunkan Lead Time hingga 67%, mengurangi 2 range pilihan dari 3 range pilihan yang ada
- Perusahaan bisa meningkatkan efisiensi dari Departemen Marketing dengan memanfaatkan 67% spare time

3. LIMITASI

- Sensitif terhadap outlier sehingga outlier harus mengalami penanganan terlebih dahulu
- Dengan kompleksnya model, maka perlu device yang lebih mumpuni untuk menghandle data yang besar



5.2 RECOMMENDATION







- Membuat pemodelan untuk properti Condominium
- Memastikan bahwa tidak terjadi kesalahan-kesalahan input data
- Menambahkan fitur lain yang mempengaruhi harga properti
- Melakukan improvement pada parameter tuning model price prediction dengan tuning yang ada sebagai benchmark
- Melakukan pemodelan price prediction ulang setelah dilakukan feature importance
- Dengan limitasi pada model yang dimiliki, hal yang bisa dilakukan yaitu:
 - Price Prediction Menambah jumlah data, khususnya pada GRADE Exceptional-B hingga Exceptional-D
 - Clustering Mencoba menggunakan pemodelan DBScan pada Properti Segmentation







- Menggunakan model machine learning price prediction yang telah dibuat
- Memberikan rekomendasi berdasarkan tren bulan yang dapat membantu untuk menyarankan waktu beli terbaik.
- Menyediakan informasi segmentasi properti yang disesuaikan dengan target market, berdasarkan karakteristik properti
- Menyarankan Department Marketing untuk membuat iklan berdasarkan kelompok rumah
- Membuat sebuah aplikasi dengan model yang telah dibuat untuk mempermudah prediksi dan segmentasi





THANK YOU!

OMEGA CORP

- Faykel Nicandro Hattu
- Muhammad Rafi Amiruddin
- Tamara Coglitore



Dashboard Descriptive Analytics

Link to Tableau Dashboard

