RANGGA AKHLI

CAPSTONE 3 & 4

PURWADHIKA'S DIGITAL TALENT INCUBATORS

RANGGA AKHLI

Predicting Apartment Prices in Daegu: A Machine Learning Approach

Presented for Capstone 3 and 4 Digital Talent Incubator
Data Science & Machine Learning - Purwadhika

RANGGA AKHLI

01

Business Problem

Opportunity to open up a new business model based on predicting apartment sales

02

Data Understanding

A general review of dataset collected

03

Data Preprocessing

An explanation on transforming raw data into a clean and usable format

04

Modeling

The stage where a chosen algorithm or multiple algorithms are applied to the preprocessed data to build and train predictive models.

05

Conclusion

Recommendation for Business and Model Limitation

06

Deployment Demonstration

A review where ML product is applied in GCS

RANGGA AKHLI

DAEGU OVERVIEW

4th largest city

Daegu is the fourth largest city in South Korea. It is known for its vibrant textile industry and rich cultural heritage, boasting historical sites like the Daegu Yangnyeongsi Museum of Oriental Medicine

$44,144 million

Robust industries are there, from textile, metal, machinery that contributed to South Korea's $44,144 million in 2021, marking this city as an important city in South Korea

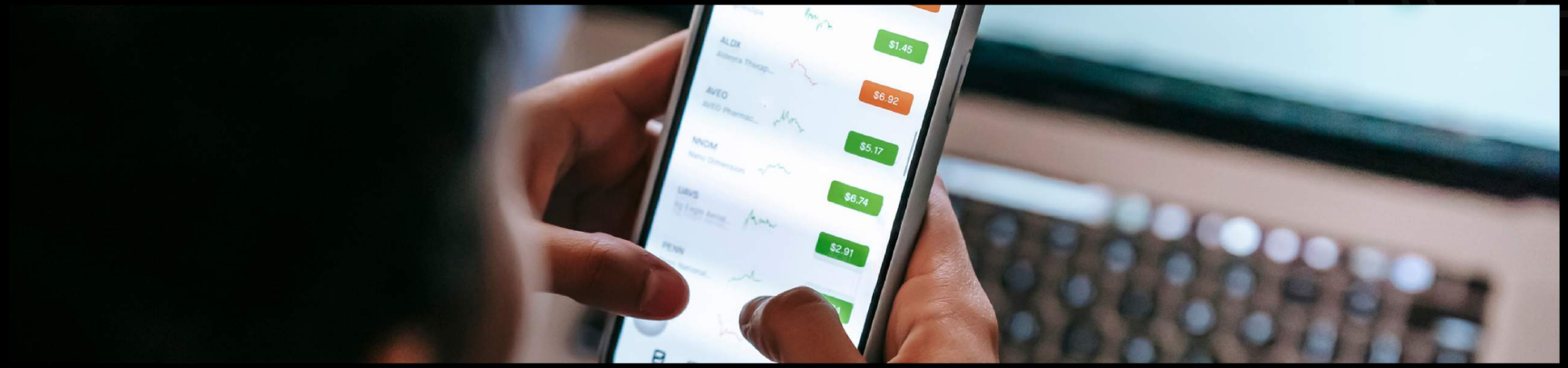
22,000 new jobs

As a Free Economic Zone (FEZ) since 2008, attracting over US$751 million in foreign investments. This status has transformed Daegu into a magnet for diverse industries has created over 22,000 new jobs

239,000 apartment

In 2022, the total number of apartment buildings in Daegu, South Korea, amounted to approximately 239,000. About 142,000 apartments were built with two to four stories, accounting for the largest share of all types of apartment buildings in Daegu that year.

RANGGA AKHLI



BUSINESS PROBLEM

Company XXX, an online real estate platform, aims to expand its operations in South Korea, specifically targeting Daegu. The platform offers a wide range of services including buying, selling, renting, and leasing accommodations, houses, apartments, and shops. Transactions are facilitated entirely through a mobile app for convenience. The platform prides itself on providing comprehensive property information, detailed property photos, and ensuring reliability.

It wants to optimize its operation by gathering and analyzing key apartment sales characteristics in Daegu.

RANGGA AKHLI

Company XXX Business Metrix

Services	Stakeholders	Needs	Machine Learning	Business Model
listing and promotion	real estate agents, property developer	reach wider audience	property recommendation, price prediction, market analytics	freemium model, commision from transaction
property discovery	homebuyers, renters	find properties matching their need	property recommendation, price prediction	targeted marketing
financial services	bank and mortgage providers	offer mortgage product to users	fraud detection, risk analysis	partnership and communication model
investment analysis	real resstate investor	find investiong opportunities	market analytics, price trend prediction	freemium model, commision from transaction
valuation services	property buyers and investors	estimate property values	price prediction, market analytic	freemium model, subscription

RANGGA AKHLI

2. DATA UNDERSTANDING

RANGGA AKHLI

GENERAL DATA CHARACTERISTICS

Dataset Info

4123 rows x 11 columns

Cheapest and Most Expensive

Apartment prices vary widely. The cheapest option, at 32,743 Won, has a corridor hallway and was built in 1992. The most expensive, priced at 585,840 Won, features a terraced hallway and was built in 2007. These examples highlight how hallway type and construction year influence pricing in Daegu.

Proximity & Higher Price

Proximity to subway stations significantly impacts apartment prices. Apartments located within 0-5 minutes of a subway station command a median price of 279,646 Won, highlighting a clear trend where closer proximity correlates with higher prices.

RANGGA AKHLI

3. DATA PREPROCESSING

01

MISSING VALUES

02

DATA DISCREPANCIES

03

OUTLIER DETECTION

04

DATA DUPLICATES

05

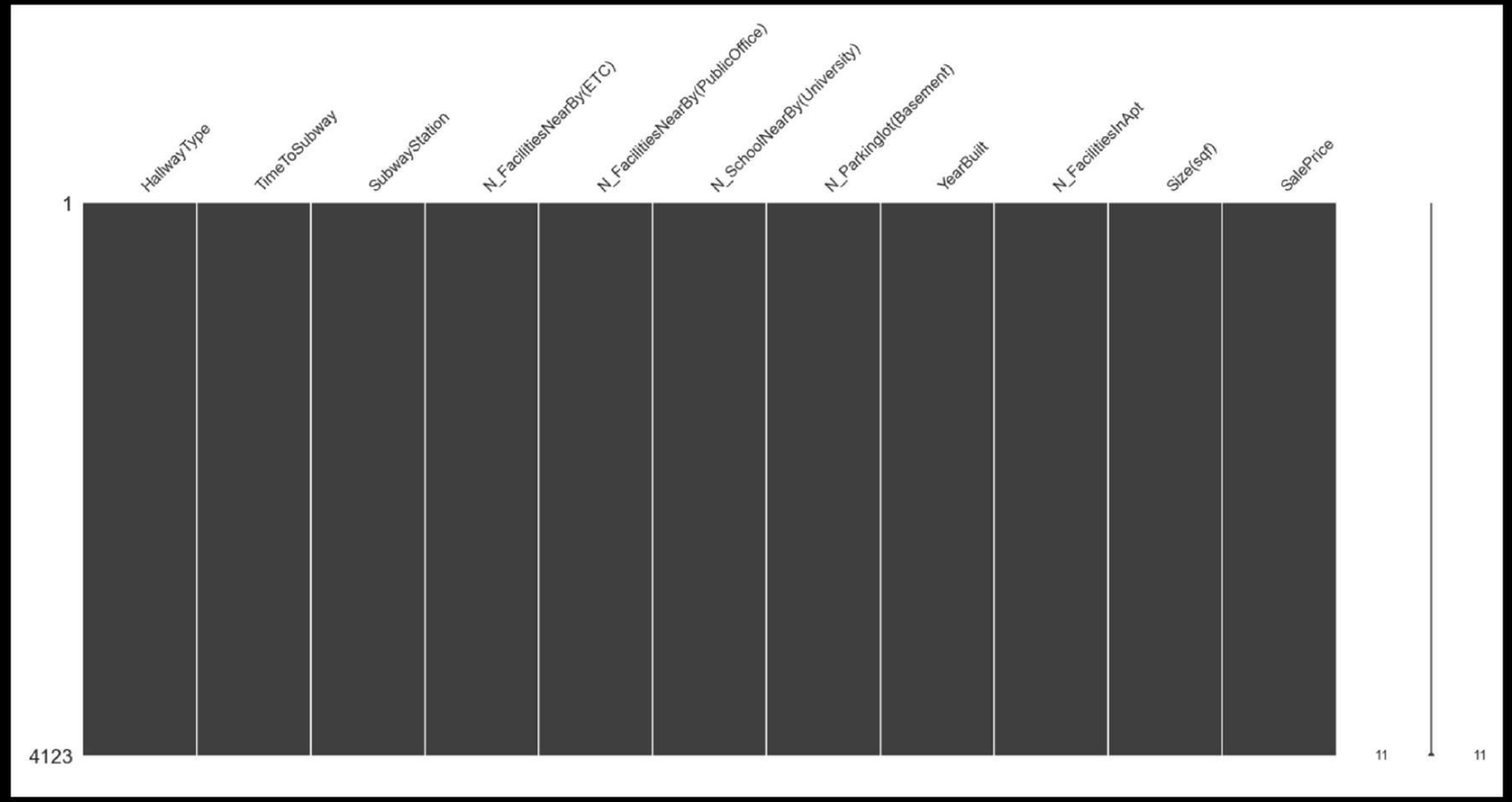
FEATURES ENGINEERING

DATA PREPROCESSING STRATEGY

A SCHEME

RANGGA AKHLI

MISSING VALUES



No missing value. The absence of missing values in the data is a positive aspect. It indicates that the data is complete and free from gaps that could potentially introduce biases or inaccuracies in the analysis. This clean data foundation provides a solid basis for reliable and trustworthy analysis.

RANGGA AKHLI

DATA CLEANING

Operation	Before	After	Explanation
Fix incorrect value	'no_bus_stop_nearby'	'no_subway_nearby'	Changed value from 'no_bus_stop_nearby' to 'no_subway_nearby' to accurately indicate the absence of nearby subway stations, correcting the value contextually.
Convert 'Size(sqf)' to 'Size(SqMetre)'	Size(sqf)	Size(SqMetre	In general usage, "square meters" (SqMetre) is more commonly used worldwide compared to "square feet" (sqft). Square meters are widely adopted in countries that follow the metric system, including many parts of Asia, Europe, and other regions globally. On the other hand, square feet is more prevalent in countries that use the imperial system, such as the United States and Canada. Specifically in the context of property measurements like apartments in Daegu, South Korea, square meters (SqMetre) would typically be the more commonly used unit for expressing size or area
Adjust two columns that relational: TimeToSubway and SubwayStationn			if the column TimeToSubway contains no_subway_nearby so the SubwaySattion should logically have no option rather than no_subway_nearby as well

31 rows

Outliers in the SalePrice column represent extreme values which could be high-value properties. Removing these could result in losing valuable insights about the distribution and range of property prices.

64 rows

Outliers in the Size(SqMetre) column could indicate measurement errors or unusually large/small properties. These outliers can skew the analysis and affect the reliability of results.

1382 rows

Duplicate rows can introduce bias and redundancy in the dataset. Removing duplicates ensures that each row represents a unique observation.

2558 rows

cleaned data is 2558 rows out of 4123

Split, transformation..

RANGGA AKHLI

DATA SPLITTING

- test size=25
- X train = 1918
- X test = 640

DATA TRANSFORMATION

- One Hot Encoding: HallwayType
- Ordinal: TimeToSubway
- Binary: SubwayStation
- Create a pipeline
- Scaling
- Transform fit

STORING INTO DATAFRAME

data cleaned

DATA READY TO USE FOR ML

RANGGA AKHLI

4. MODELING

RANGGA AKHLI

MODELING SCHEME

01 DEFINE BASELINE MODEL

model: multiple regression
evaluation method: adjusted R2, MAPE
data issues: multicolinearity, VIF check

02 DEFINE OTHER LINEAR AND NON LINEAR

Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, XGBoost, KNN, SVM, , MLP Regressor

03 CROSS VALIDATION

fold: n_split 5
shulffle=True
random state

04 BENCHMARK BASE COMPARISON

identify and choose two best model based on their performance and stability in adjusted R2 and MAPE, before and after tuning

05 SAVE BEST MODEL

Building relationships with potential customers, business partners and other stakeholders.

RANGGA AKHLI

Dep. Variable:	SalePrice	R-squared:	0.752
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	552.2
Date:	Wed, 17 Jul 2024	Prob (F-statistic):	0.00
Time:	20:10:18	Log-Likelihood:	-31368.
No. Observations:	2558	AIC:	6.277e+04
Df Residuals:	2543	BIC:	6.285e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.602e+04	8783.095	-4.101	0.000	-5.32e+04	-1.88e+04
HallwayType_mixed	-3.41e+04	5807.234	-5.871	0.000	-4.55e+04	-2.27e+04
HallwayType_terraced	1.51e+04	6026.664	2.505	0.012	3281.532	2.69e+04
TimeToSubway	5745.0767	2207.551	2.602	0.009	1416.296	1.01e+04
SubwayStation_0	2.47e+04	9791.437	2.522	0.012	5497.512	4.39e+04
SubwayStation_1	-4366.8767	3383.621	-1.291	0.197	-1.1e+04	2268.056
SubwayStation_2	2.335e+04	4081.695	5.720	0.000	1.53e+04	3.14e+04
SubwayStation_3	4228.9949	3731.580	1.133	0.257	-3088.251	1.15e+04
N_FacilitiesNearBy(ETC)	-2.677e+04	5796.390	-4.619	0.000	-3.81e+04	-1.54e+04
N_FacilitiesNearBy(PublicOffice)	1.952e+04	7389.417	2.642	0.008	5032.997	3.4e+04
N_SchoolNearBy(University)	-3869.5077	1.11e+04	-0.348	0.728	-2.57e+04	1.79e+04

...

MULTIPLE REGRESSION RESULTS

	variables	VIF
0	N_FacilitiesNearBy(ETC)	5.630283
1	N_FacilitiesNearBy(PublicOffice)	12.164557
2	N_SchoolNearBy(University)	16.694363
3	N_Parkinglot(Basement)	9.966609
4	YearBuilt	9.480796
5	N_FacilitiesInApt	13.360957
6	Size(SqMetre)	5.081688

RANGGA AKHLI

INSIGHTS FROM THE BASELINE MODEL...

75,2 %

the multiple regression score is 75,2 for R2 and 75,1 for adjusted R2. r2 score indicates that 75.2% of the variance in SalePrice is explained by the model. thus, it suggest a good fit

HallwayType

Apartments with mixed hallway types are priced **34,100** Won lower than the reference hallway type. Apartments with terraced hallway types are priced **15,100** Won higher than the reference hallway type.

Travel Time to Subway

Each increase in travel time to the nearest subway station is associated with a price increase of **5,745 Won.**

Nearby Facility Impacts

The presence of additional facilities (ETC) nearby decreases apartment prices by **26,770 Won.** Nearby public offices increase apartment prices by **19,520 Won.**

Statistical Significance

Some variables, such as proximity to subway station 1 and nearby universities, do not significantly impact apartment prices ($p > 0.05$).

Overall insights

Factors like hallway type, proximity to subway stations, and nearby public facilities significantly influence apartment prices. Certain variables may not need to be included in more complex models due to their lack of statistical significance.

Best four model comparison

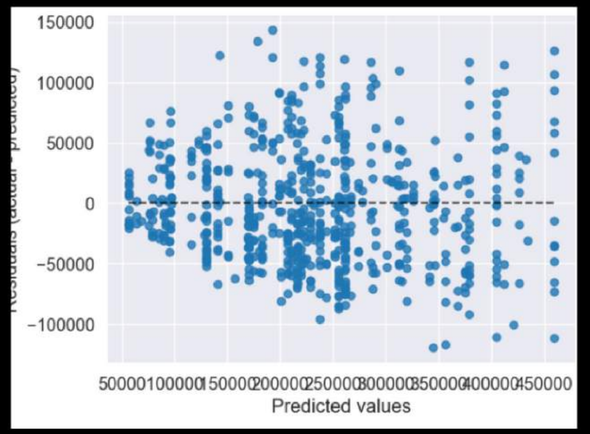
Model Name	Adjusted R2 validation mean
XGBoost	79,00
Random Forest	77,86
Gradient Boosting	77,85
Decision Tree	77,42

Evaluation after tuning two best model

Model Name	Adjusted R2 validation mean	Adjusted R2 Test Score (%)	Adjusted R2 Difference	MAPE Score
Random Forest Before Tuning	79,00	78.49	-0.61	19.53
Random Forest After Tuning	77,86	78.48	-0.64	19.57
XGboost After Tuning	77,85	78.67	-0.65	19.57
XGBoost before tuning	77,42	78.43	-0.68	19.55

RANGGA AKHLI

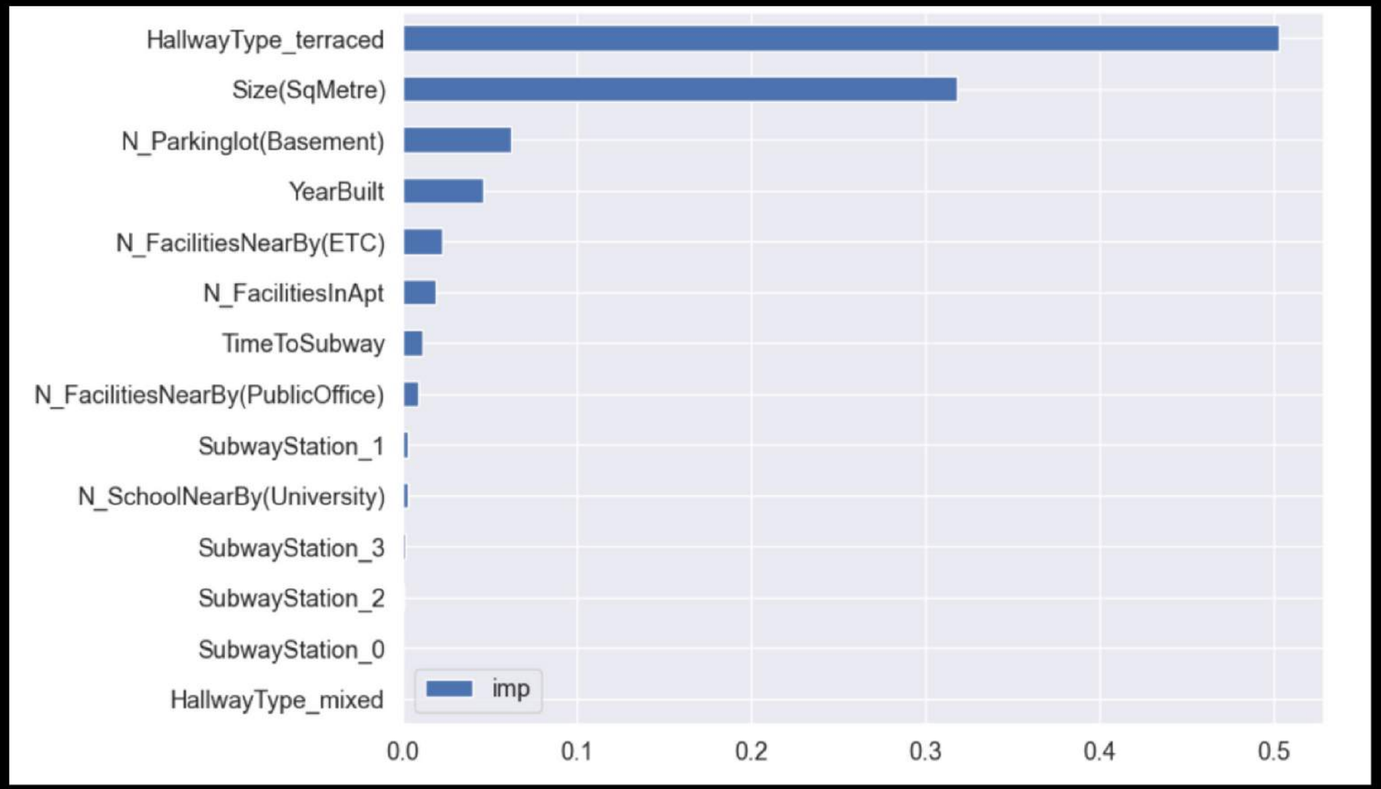
RANDOM FOREST MODEL



RANGGA AKHLI

Top 5 Feature Importance

Variables	Importance
HallwayType_terraced	0.503285
Size(SqMetre)	0.318106
N_Parkinglot(Basement)	0.061800
YearBuilt	0.046438
N_FacilitiesNearBy	0.022362



RANGGA AKHLI

5. CONCLUSION AND LIMITATION

RANGGA AKHLI

CONCLUSION

01

Model interpretation

Adjusted R^2 Value: An adjusted R^2 of around 79% signifies that the model can account for 79% of the variance in apartment prices using the input features. This high percentage reflects the model's robustness in capturing the relationships between the features and the target variable.

02

Use of Prediction

Given the high adjusted R^2, this model is recommended for predicting apartment prices in Daegu. It provides reliable predictions, helping property developers, real estate agents, and potential buyers make informed decisions

03

Model enhancement

Although the model performs well, the remaining 21% unexplained variance suggests there is room for improvement. Future iterations could explore additional variables or advanced feature engineering to capture more of the underlying patterns, such as economic indicators, future infrastructure projects, or detailed neighborhood characteristics.

04

Efficiency and Practicality

The small difference in adjusted R^2 between validation and test sets (0.61%) indicates consistent performance across different data sets. Regular validation and monitoring should be conducted to ensure the model maintains its accuracy and relevance over time, adapting to changes in the market.

RANGGA AKHLI

LIMITATION

01 Data Availability

The model's performance is dependent on the quality and completeness of the data used. Any missing or inaccurate data could impact the model's accuracy and generalizability.

02 Feature Selection

While the model accounts for many relevant features, there could be other influential factors not included in the dataset, such as economic conditions, government policies, or future infrastructure developments.

03 Temporal Validity

The model is based on historical data and may not account for future changes in market dynamics or unexpected events, such as natural disasters or significant economic shifts.

04 Geographical Specificity

This model is specifically tailored for Daegu and may not be directly applicable to other cities or regions without adjustments and retraining on local data.

05 Complexity of the Market

The housing market is influenced by a myriad of factors, including but not limited to buyer sentiment, interest rates, and cultural preferences. The model captures a substantial portion but not all of this complexity.

6. DEPLOYMENT

will be deployed in to Google Cloud Console

```
pipeline_model.fit(X_train, y_train)
with open('rangga_daeguapart_model.pkl', 'wb') as wtm:
 pickle.dump(pipeline_model, wtm)
```