DAEGU APARTMENT PRICING PREDICTION

Nabila Ryrie

TABLE OF CONTENTS

BACKGROUND

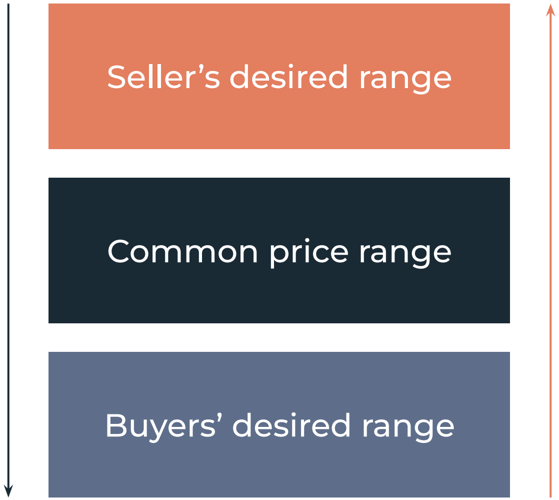
Overview of Daegu

Daegu is a bustling city located in South Korea, known for its rich history, vibrant culture, and economic significance. The city is known for its thriving economy, particularly in industries like textiles, manufacturing, and technology. As of 2023, Daegu has a population sum of 2,181 million people. In smaller cities like Daegu where land is limited, apartments are a popular solution for housing.

MAIN PROBLEM

MAIN PROBLEM?

Price too high and you lose buyers.



Price too low and you lose value.

GOALS

Implement machine learning algorithms to predict ideal housing prices based on each unit facilities.

CHALLENGE

Sellers need to set prices for apartments that are neither too high or too low according to the facilities each unit has.

APPROACH

Implement machine learning algorithms to predict ideal housing prices based on each unit facilities.

OUTCOME

Receive the optimal housing prices for sellers to benchmark on and negotiate.

DATA UNDERSTANDING

DATA UNDERSTANDING

	HallwayType	TimeToSubway	SubwayStation	N_FacilitiesNearBy(ETC)	N_FacilitiesNearBy(PublicOffice)	N_SchoolNearBy(University)	N_Parkinglot(Basement)	YearBuilt	N_FacilitiesInApt	Size(sqf)	SalePrice
0	terraced	0-5min	Kyungbuk_uni_hospital	0.0	3.0	2.0	1270.0	2007	10	1387	346017
1	terraced	10min~15min	Kyungbuk_uni_hospital	1.0	5.0	1.0	0.0	1986	4	914	150442
2	mixed	15min~20min	Chil-sung-market	1.0	7.0	3.0	56.0	1997	5	558	61946
3	mixed	5min~10min	Bangoge	5.0	5.0	4.0	798.0	2005	7	914	165486
4	terraced	0-5min	Sin-nam	0.0	1.0	2.0	536.0	2006	5	1743	311504

The dataset makes up the information regarding each apartment unit in the Daegu area.

Each row represents the unit of each apartment, along with the facilities that they have.

DATA UNDERSTANDING

Columns	Description
Hallway Type	Apartment type
TimeToSubway	Time needed to the nearest subway station
SubwayStation	The name of the nearest subway station
N_FacilitiesNearBy(ETC)	The number of facilities nearby
N_FacilitiesNearBy(PublicOffice)	The number of public office facilities nearby
N_SchoolNearBy(University)	The number of universities nearby
N_Parkinglot(Basement)	The number of the parking lot
YearBuilt	The year the apartment was built
N_FacilitiesInApt	Number of facilities in the apartment
Size(sqft)	The apartment size (in square feet)
SalePrice	The apartment price (Won)

DATA CLEANING

Dataset Info

The dataset has 11 columns and 4123 rows.

The data type is in accordance with the existing values of each column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4123 entries, 0 to 4122
Data columns (total 11 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   HallwayType                    4123 non-null   object 
 1   TimeToSubway                   4123 non-null   object 
 2   SubwayStation                  4123 non-null   object 
 3   N_FacilitiesNearBy(ETC)        4123 non-null   float64
 4   N_FacilitiesNearBy(PublicOffice) 4123 non-null   float64
 5   N_SchoolNearBy(University)     4123 non-null   float64
 6   N_Parkinglot(Basement)         4123 non-null   float64
 7   YearBuilt                      4123 non-null   int64  
 8   N_FacilitiesInApt              4123 non-null   int64  
 9   Size(sqf)                      4123 non-null   int64  
 10  SalePrice                      4123 non-null   int64  
dtypes: float64(4), int64(4), object(3)
```

DATA CLEANING

Duplicates

The dataset contains 1422 duplicates.

	Duplicates
HallwayType	3
TimeToSubway	5
SubwayStation	8
N_FacilitiesNearBy(ETC)	4
N_FacilitiesNearBy(PublicOffice)	8
N_SchoolNearBy(University)	6
N_Parkinglot(Basement)	20
YearBuilt	16
N_FacilitiesInApt	9
Size(sqf)	89
SalePrice	838

There are 1422 duplicated rows in total.

HANDLING DUPLICATES

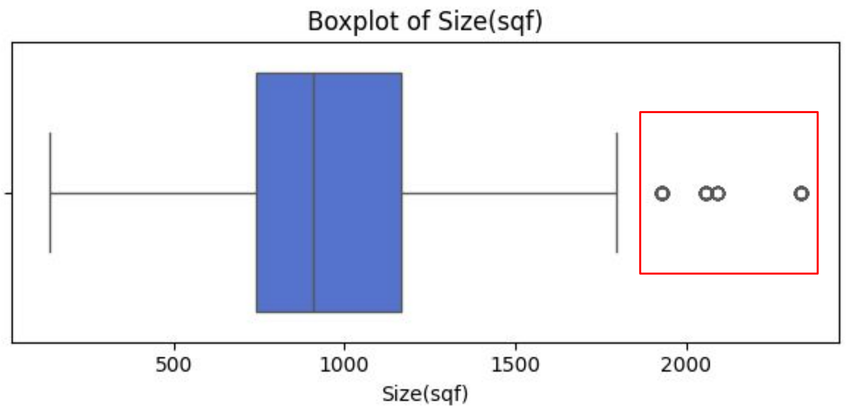
	HallwayType	TimeToSubway	SubwayStation	N_FacilitiesNearBy(ETC)	N_FacilitiesNearBy(PublicOffice)	N_SchoolNearBy(University)	N_Parkinglot(Basement)	YearBuilt	N_FacilitiesInApt	Size(sqf)	SalePrice
55	terraced	0-5min	Kyungbuk_uni_hospital	0.0	5.0	3.0	930.0	2013	7	910	263345
56	terraced	0-5min	Banwoldang	0.0	0.0	0.0	203.0	2014	10	914	371681
122	terraced	0-5min	Kyungbuk_uni_hospital	0.0	5.0	3.0	930.0	2013	7	644	149274
127	terraced	0-5min	Banwoldang	0.0	2.0	2.0	524.0	2007	4	1394	256637
133	mixed	15min~20min	Myung-duk	5.0	6.0	5.0	536.0	1993	4	644	168141
...	...	...	...	...	...	...	...	...	...	...	...
4113	terraced	5min~10min	Daegu	0.0	3.0	2.0	400.0	2015	7	644	300884
4114	corridor	10min~15min	Myung-duk	5.0	7.0	5.0	0.0	1992	3	355	86725
4115	mixed	15min~20min	Myung-duk	5.0	6.0	5.0	536.0	1993	4	1761	168141
4120	mixed	15min~20min	Myung-duk	5.0	6.0	5.0	536.0	1993	4	1761	168141
4122	terraced	0-5min	Kyungbuk_uni_hospital	0.0	3.0	2.0	1270.0	2007	10	868	250442

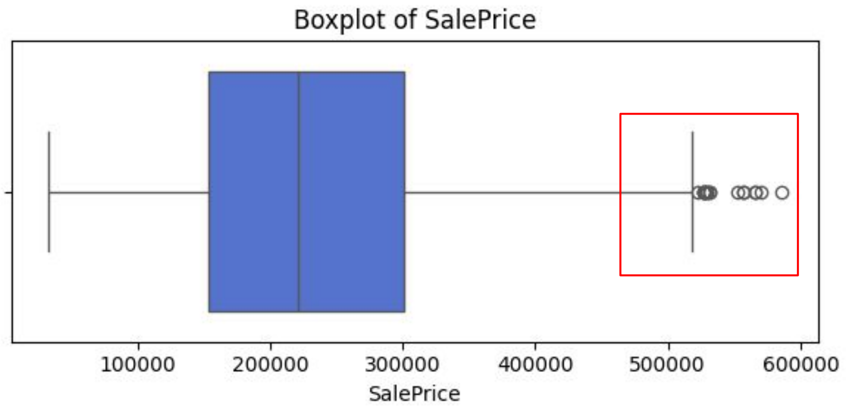
We can assume that data duplication occurs for apartments within the same building and having the same facilities, only in different units. Consequently, we can safely remove these duplicates as the duplicated rows can be represented by one row with the exact same values. Thus, our model can refer to that one row when making predictions.

DATA CLEANING

Outliers

There are outliers in the SalePrice column and Size(Sqf) column.





HANDLING OUTLIERS

Sale Outlier

Size(sqf)	SalePrice
1928	585840
1928	570796
1643	566371
1928	566371
1928	557522
...	...
135	35398
355	35398
355	34513
355	34070
355	32743

Size Outlier

Size(sqf)	SalePrice
2337	194690
2337	227433
2337	203539
2337	292035
2337	351769
...	...
135	60973
135	53274
135	62831
135	53982
135	35398

Outlier Information

Larger-sized units tend to command higher prices, and it's not uncommon to have availability of spacious apartments. Therefore, we've decided not to remove these outliers as they can be utilized to make predictions for larger sized unit pricing.

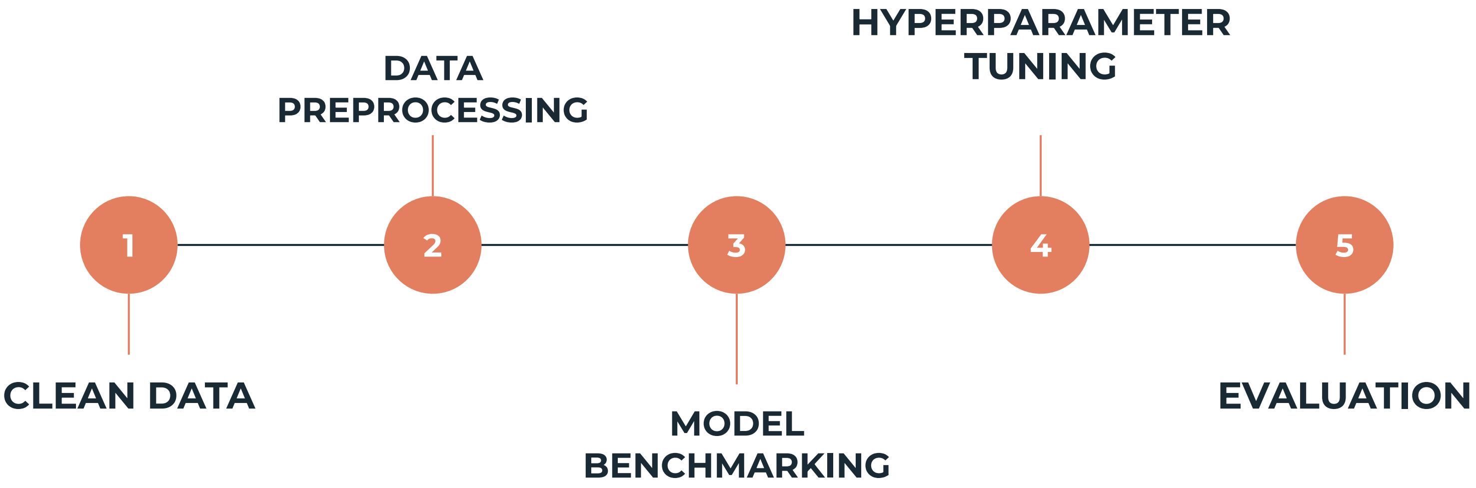
DATA CLEANING SUMMARY

DATA

- **NULL VALUES** — No null values, no need to drop
- **DUPLICATES** — Drop duplicates
- **OUTLIERS** — Can be kept, is in accordance with domain knowledge
- **FINAL CLEAN DATA** — Data with no duplicates

MACHINE LEARNING WORKFLOW

MACHINE LEARNING TIMELINE



DATA PREPROCESSING

DATA PREPROCESSING

Define Feature (X) and Target (Y)

X (Features): The characteristic or attributes of our data that we use as inputs for our model. In this case, our features are the facilities that our data has recorded.

Y (Target): The variable we would like to predict using our model. In this case, our target is the Sale Price of each apartment unit.

N_FacilitiesNearBy(PublicOffice)	N_SchoolNearBy(University)	N_Parkinglot(Basement)	YearBuilt	N_FacilitiesInApt	Size(sqf)	SalePrice
6.0	5.0	536.0	1993	4	2337	194690
6.0	5.0	536.0	1993	4	2337	227433
6.0	5.0	536.0	1993	4	2337	203539
6.0	5.0	536.0	1993	4	2337	292035
6.0	5.0	536.0	1993	4	2337	351769

Feature

Target

DATA PREPROCESSING

Define Feature (X) and Target (Y)

We will be using all the columns in our dataset for our model, with the SalePrice column being the target.

Columns	Description
Hallway Type	Apartment type
TimeToSubway	Time needed to the nearest subway station
SubwayStation	The name of the nearest subway station
N_FacilitiesNearBy(ETC)	The number of facilities nearby
N_FacilitiesNearBy(PublicOffice)	The number of public office facilities nearby
N_SchoolNearBy(University)	The number of universities nearby
N_Parkinglot(Basement)	The number of the parking lot
YearBuilt	The year the apartment was built
N_FacilitiesInApt	Number of facilities in the apartment
Size(sqft)	The apartment size (in square feet)
SalePrice	The apartment price (Won)

X Feature

HallwayType, TimeToSubway, SubwayStation, N_FacilitiesNearBy(ETC), N_FacilitiesNearBy(PublicOffice), N_SchoolNearBy(University), N_ParkingLot(Basement), Year Built, N_FacilitiesInApt and Size(Sqf)

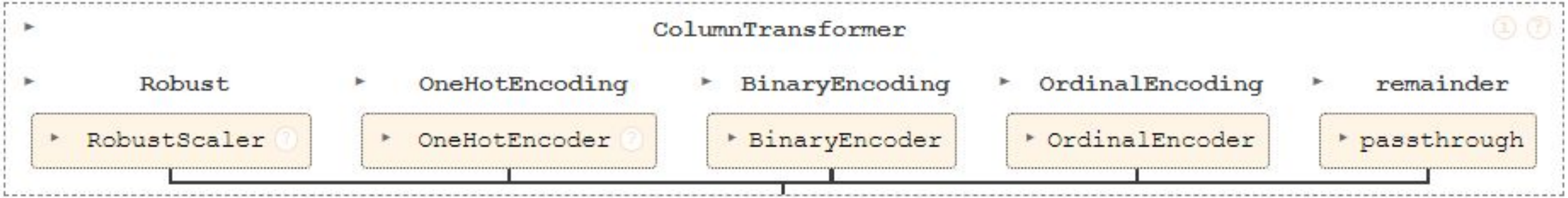
Y Feature

SalePrice

DATA ENCODING

Encoding

Encoding is the process of converting data from one form to another. In the context of machine learning, encoding typically refers to converting categorical data (data that represents categories or labels) into a numerical format that can be used by machine learning algorithms.



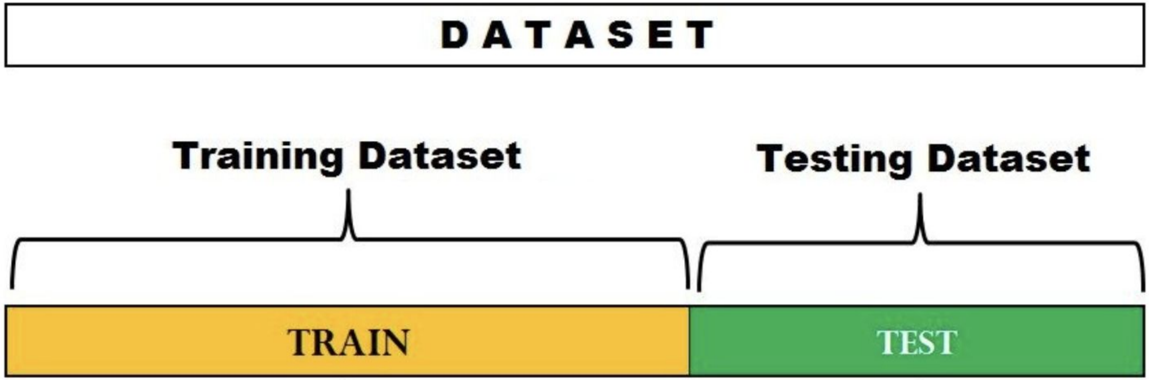
TimeToSubway - OrdinalEncoder
HallwayType - OneHotEncoder
SubwayStation - BinaryEncoder

DATA PREPROCESSING

Splitting Data

We will split the dataset into **two subsets: training data (train set) and test data (test set).**

The data will be split into two, **80% (train set) and 20% (test set)**.



MODEL BENCHMARKING

MODEL BENCHMARKING

Model	MAPE Mean	MAPE Std
Decision Tree	-0.190261	0.005464
Random Forest	-0.191263	0.005469
KNN	-0.204210	0.007920
Linear Regression	-0.221010	0.004781
SVM	-0.549291	0.028666

Two best models:

Decision Tree and Random Forest.

HYPERPARAMETER TUNING

PREDICT TO TEST SET

Model	MAPE
Random Forest	0.193559 (19.35%)
Decision Tree	0.197060 (19.70%)

Despite Decision Tree having better MAPE mean earlier, Random Forest had the better MAPE result in the test set.

HYPERPARAMETER TUNING

Hyperparameter tuning involves discovering the best set of values for a model's hyperparameters, which are predefined parameters governing the learning process in machine learning algorithms.

Random Forest Parameter

random_state = 2024

n_estimators = 300

min_samples_split = 5

min_samples_leaf = 1

max_features = 'sqrt'

max_depth = 20

criterion = 'absolute_error'

Decision Tree Parameter

random_state = 2024

min_samples_split = 2

min_samples_leaf = 1

max_features = 'sqrt'

max_depth = 20

criterion = 'absolute_error'

EVALUATION MODEL

Performance Comparison

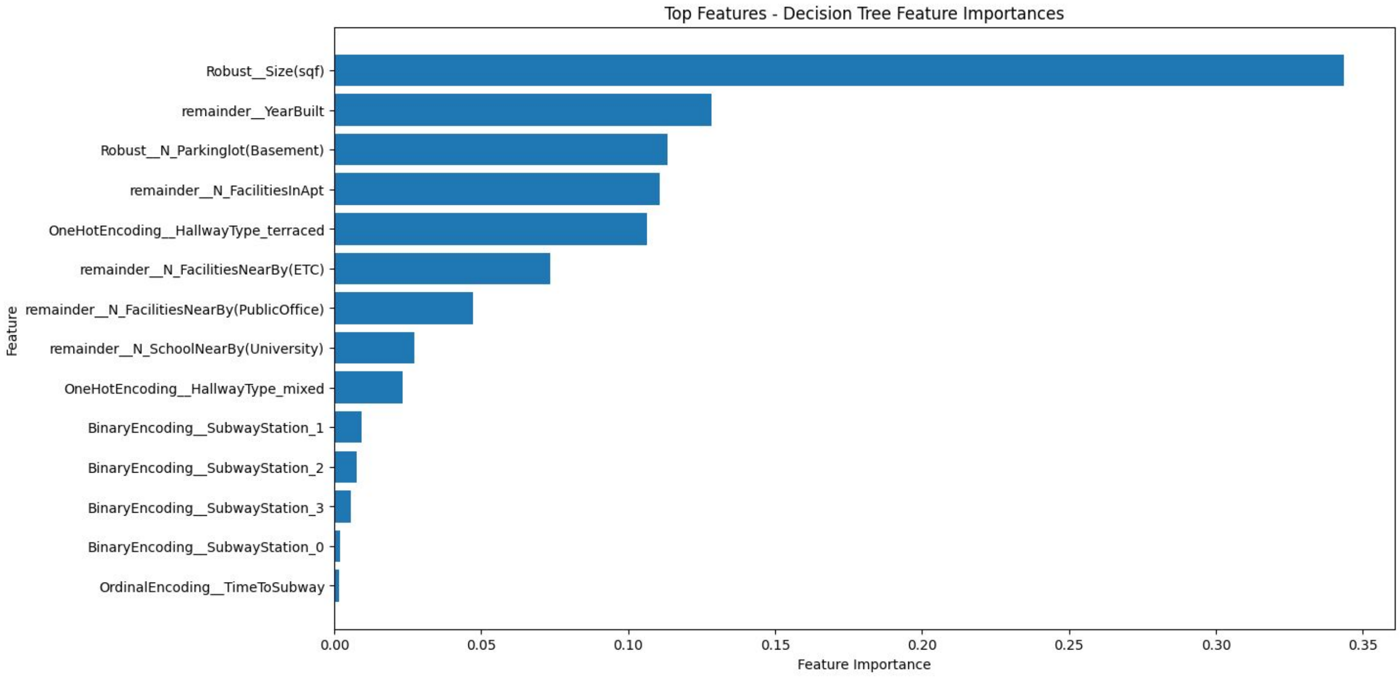
Performance comparison after hyperparameter tuning involves evaluating and comparing the effectiveness of machine learning models using optimized hyperparameters.

Model	MAPE Before Tuning	MAPE After Tuning
Random Forest	0.194098	0.183642
Decision Tree	0.197084	0.181389

This shows that there is 18.36% of error for Random Forest Model, and 18.13% error for Decision Tree Model.

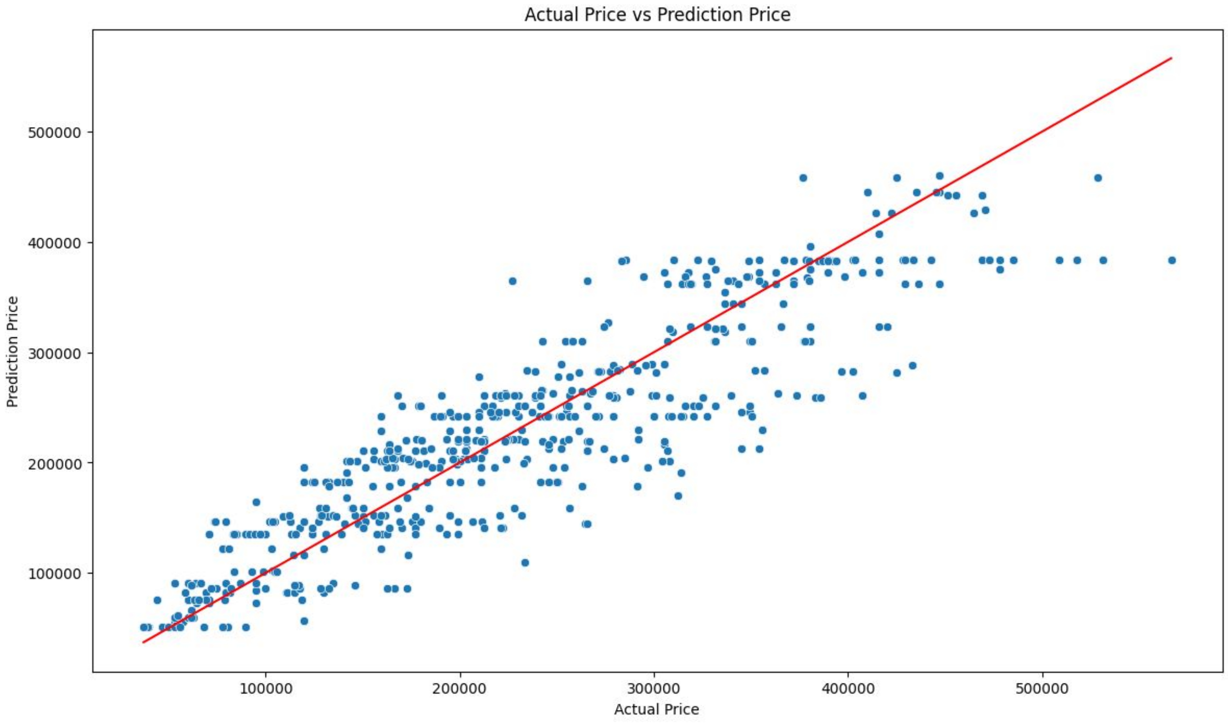
FEATURE IMPORTANCE AND REGRESSION PLOT

FEATURE IMPORTANCE



The size of the unit has the greatest impact on apartment pricing, followed by construction year and number of parking lots.

REGRESSION PLOT



We observe a strong correlation for prices below 450,000 won. However, above this threshold, the correlation between actual and predicted prices appears more varied.

CONCLUSION

1. Random Forest and Decision Tree performed the best, with MAPE means of -0.190 and -0.191 respectively.

2. Decision Tree had the best result after tuning, from 19.7% MAPE score to 18.3% MAPE score.

3. Decision Tree thus became the preferred model, which was used for the feature importance plot.

RECOMMENDATION

1. Enhance Model Features: we can add more features that could affect apartment prices in Daegu.

2. Fine-Tune Parameters: improves in boosting accuracy.

3. Integrate Model in Sales: incorporating the model in sales can assist real estate agents in setting optimal apartment prices based on negotiations with potential buyers.

4. Explore key features: further analyze the most influential features impacting apartment prices in Daegu.

THANK YOU!