AI-Assignment-04

Devyansh Chaudhary

2022156

Theory

Problem-01

$$H(S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

longterm Debt = A

$$IG(A) = H(S) - \left[-\frac{1}{2}\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right] - \frac{1}{2}\left[\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right]\right]$$

$$= 1 - 0.722 = 0.278$$

$$\boxed{IG(A) = 0.278}$$

Unemployed = B

$$IG(B) = H(S) - \left[-\frac{2}{10}\left[0 \times \log_2 0 + \frac{2}{2}\log_2 1\right] - \frac{8}{10}\left[\frac{5}{8}\log_2\frac{5}{8} + \frac{3}{8}\log_2\frac{3}{8}\right]\right]$$

$$= 1 - [0.76] = 0.24$$

$$\boxed{IG(B) = 0.24}$$

Credit Rating = C

$$IG(C) = H(S) - \left[\left[-\frac{3}{10}\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right]\right] + \left[-\frac{7}{10}\left[\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right]\right]\right]$$

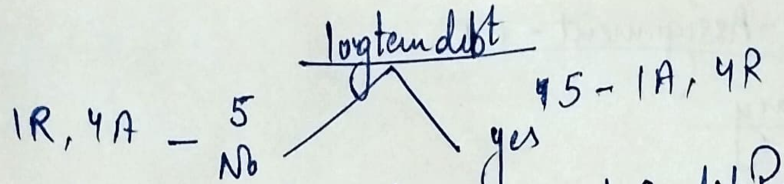
$$= H(S) - [0.275 + 0.689]$$

$$\boxed{IG(C) = 0.036}$$

[based on Infomation Gain first attribute to chose will be long term dept.]

$$IG(\text{Down Payment}) = H(S) - \left[-\frac{5}{10}\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right] - \frac{5}{10}\left[\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]\right]$$

$$\boxed{IG(\text{Down Paymnt}) = 1 - 0.97 = 0.029}$$



Unemployed CreditRating

4No – 4A 2Good – 1R, 1A

1yes – 1R 3bad – 3A

DownPayment

2No – 2A

3yes – 2A, 1R

~~IG(unemployed)~~

$$H(S) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= 0.72$$

$$IG(\text{Unemployed}) = 0.72 - \left[-\frac{4}{5}\left[\frac{4}{4}\log_2\frac{4}{4}+0\right]\right] - \left[-\frac{1}{5}\left[\frac{1}{1}\log_2 1+0\right]\right]$$

$$= 0.72$$

$$IG(\text{CreditRating}) = 0.72 - \left[-\frac{2}{5}\left[\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right] - \frac{3}{5}\left[\frac{3}{3}\log_2 1+0\right]\right]$$

$$= 0.72 + \frac{2}{5} = 0.32$$

$$IG(\text{DownPayment}) = 0.72 - \left[\cancelto{0}{-\frac{2}{5}\left[1\log_2 1+0\right]} - \frac{3}{5}\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right]\right]$$

$$= 0.72 - [0.55]$$

$$= 0.16$$

[Unemployed will be choosen with highest Info. Gain]

Unemployed Credit Rating

4No – 1A, 3R 1Good – 1A

1yes – 1R 4bad – 4R

DownPayment

3No – 1A, 2R

2yes – 2R

$$H(S) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= 0.72$$

$$IG(\text{Unemployed}) = 0.72 - \left[-\frac{4}{5}\left[\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4}\right] - \cancelto{0}{\frac{1}{5}\left[1\log_2 1+0\right]}\right]$$

$$= 0.72 - 0.64 = 0.07$$

$$IG(\text{CreditRating}) = 0.72 - \left[\cancelto{0}{-\frac{1}{5}\left[1\log_2 1+0\right]} - \cancelto{0}{\frac{4}{5}\left[1\log_2 1+0\right]}\right]$$

$$= 0.72$$

$$IG(\text{DownPayment}) = 0.72 - \left[-\frac{3}{5}\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right] - \cancelto{0}{\frac{2}{5}\left[\frac{2}{2}\log_2\frac{2}{2}+0\right]}\right]$$

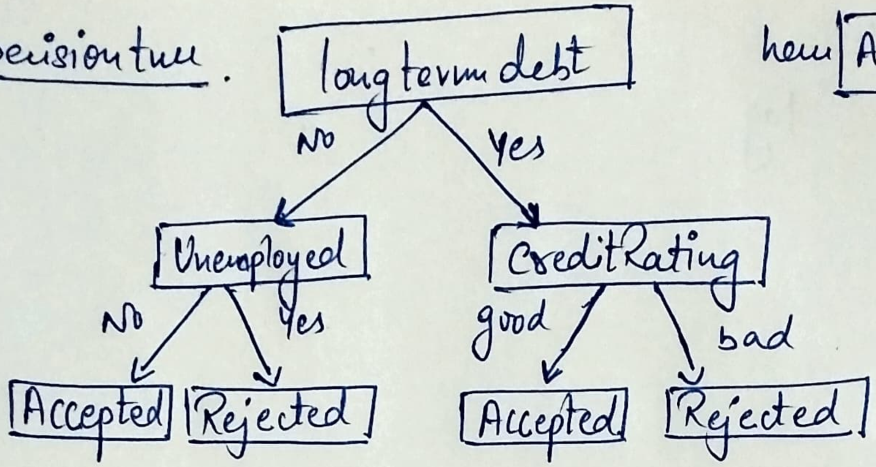
$$= 0.72 - 0.55$$

$$= 0.16$$

[Credit Rating will be choosen with highest Informt Gain]

Decision tree.

here [Accepted = Approve] ↓ on leaf nodes



[this is the final decision tree will look like using the given examples table].

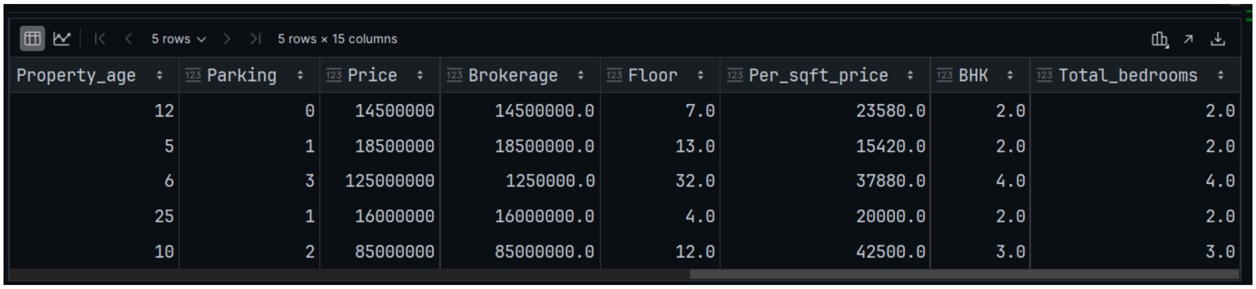
training error = 0 % all the given examples in training data are correctly classified using the given decision tree.

AI-Report-Coding-Assignment: 04

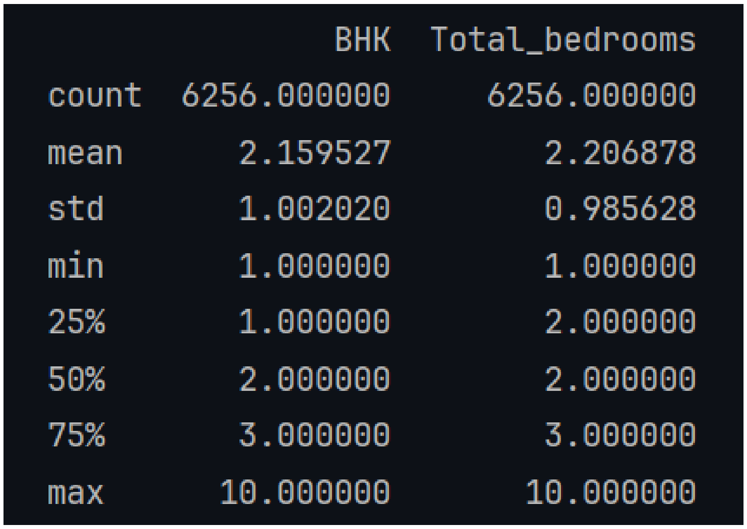
Devyansh Chaudhary
2022156

Problem 2: Data Preprocessing and Exploratory Data Analysis

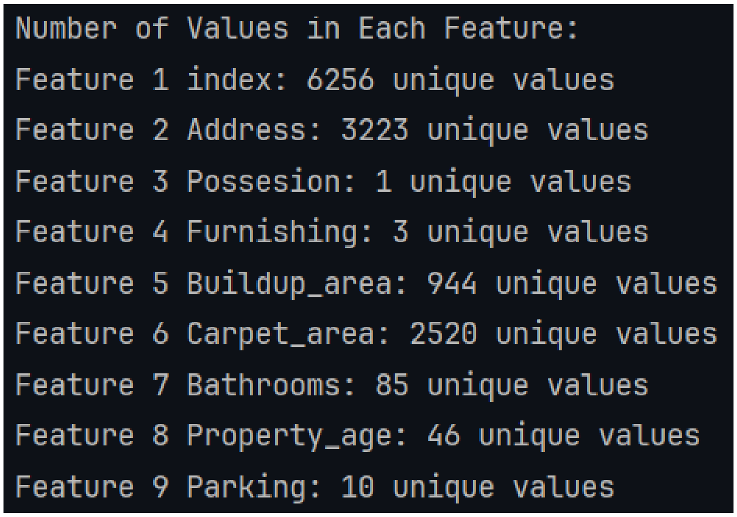
Task: 1 Understanding the Dataset



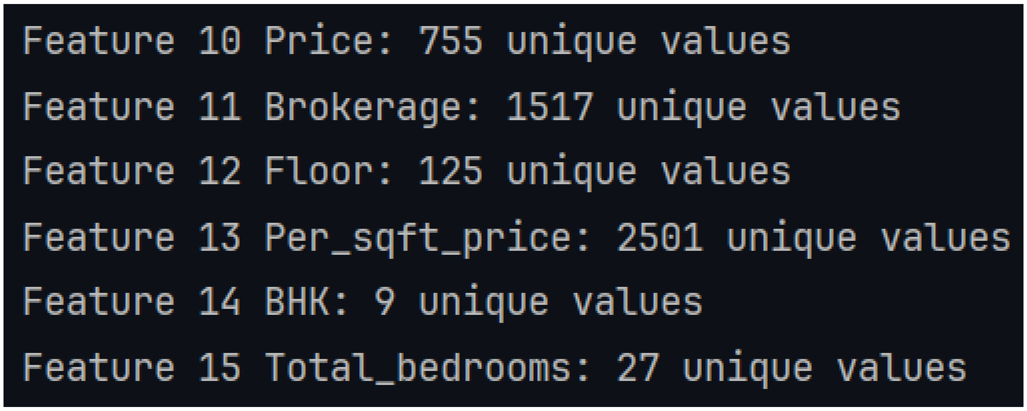
Property_age	Parking	Price	Brokerage	Floor	Per_sqft_price	BHK	Total_bedrooms
12	0	14500000	14500000.0	7.0	23580.0	2.0	2.0
5	1	18500000	18500000.0	13.0	15420.0	2.0	2.0
6	3	125000000	1250000.0	32.0	37880.0	4.0	4.0
25	1	16000000	16000000.0	4.0	20000.0	2.0	2.0
10	2	85000000	85000000.0	12.0	42500.0	3.0	3.0



```
               BHK  Total_bedrooms
count  6256.000000     6256.000000
mean      2.159527        2.206878
std       1.002020        0.985628
min       1.000000        1.000000
25%       1.000000        2.000000
50%       2.000000        2.000000
75%       3.000000        3.000000
max      10.000000       10.000000
```

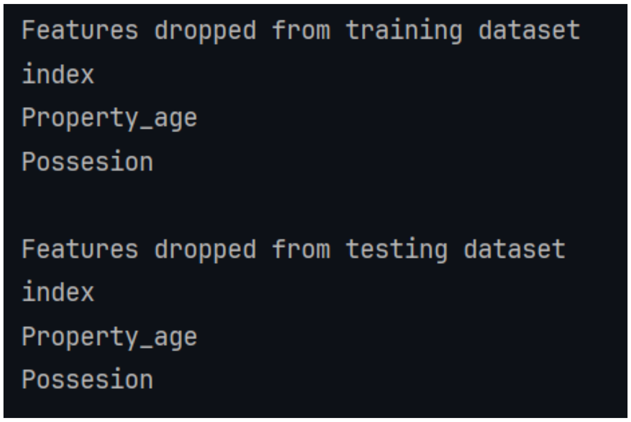


```
Number of Values in Each Feature:
Feature 1 index: 6256 unique values
Feature 2 Address: 3223 unique values
Feature 3 Possesion: 1 unique values
Feature 4 Furnishing: 3 unique values
Feature 5 Buildup_area: 944 unique values
Feature 6 Carpet_area: 2520 unique values
Feature 7 Bathrooms: 85 unique values
Feature 8 Property_age: 46 unique values
Feature 9 Parking: 10 unique values
```



```
Feature 10 Price: 755 unique values
Feature 11 Brokerage: 1517 unique values
Feature 12 Floor: 125 unique values
Feature 13 Per_sqft_price: 2501 unique values
Feature 14 BHK: 9 unique values
Feature 15 Total_bedrooms: 27 unique values
```

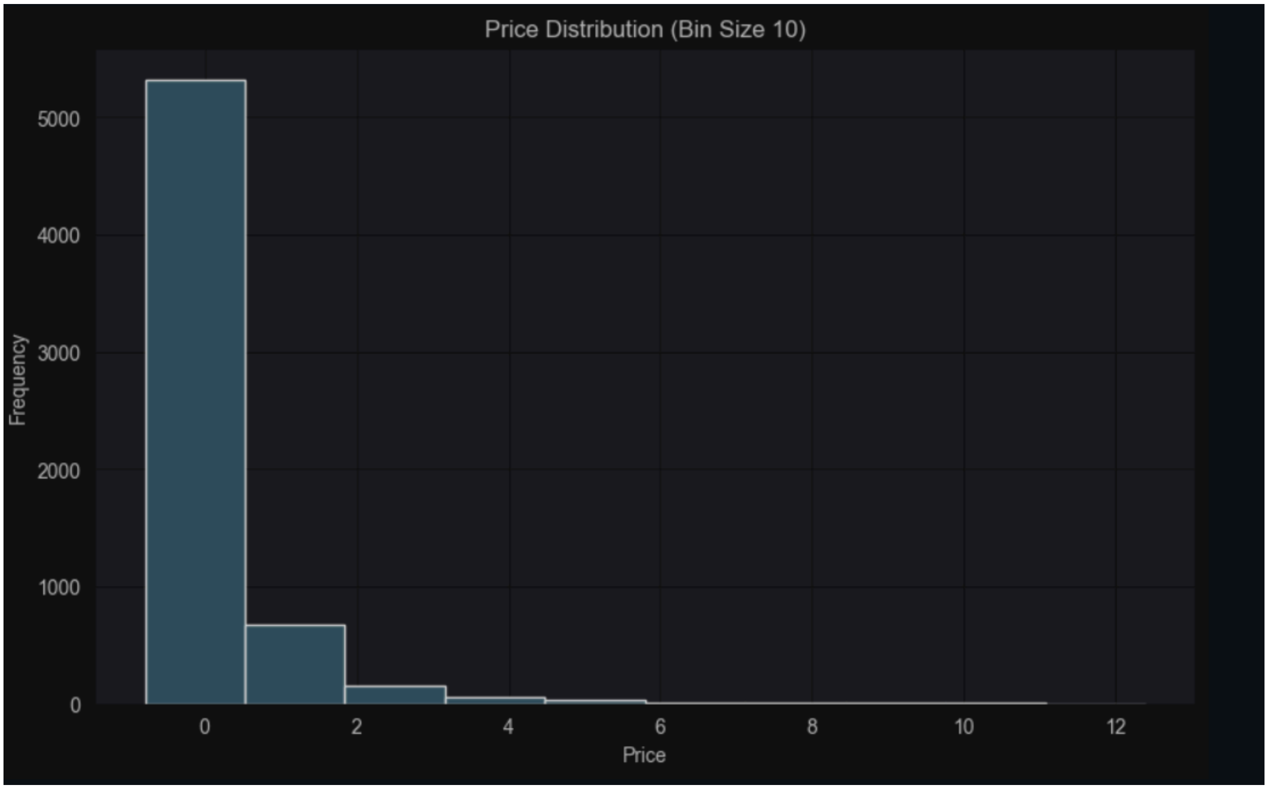
Task: 02 Drop Irrelevant Columns

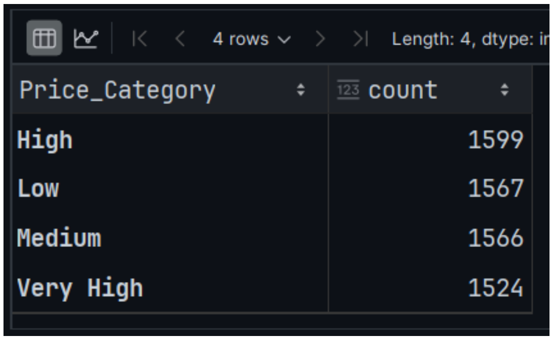


```
Features dropped from training dataset
index
Property_age
Possesion

Features dropped from testing dataset
index
Property_age
Possesion
```

These were columns dropped from the training and testing data, The column index is dropped because index has no relation with the price of property also Possession does not help in any way to predict the price of the property because our ultimate goal is to predict the price of the property with the help of given features. Property_age is dropped because of weak correlation found with the given conditions in the problem statement.

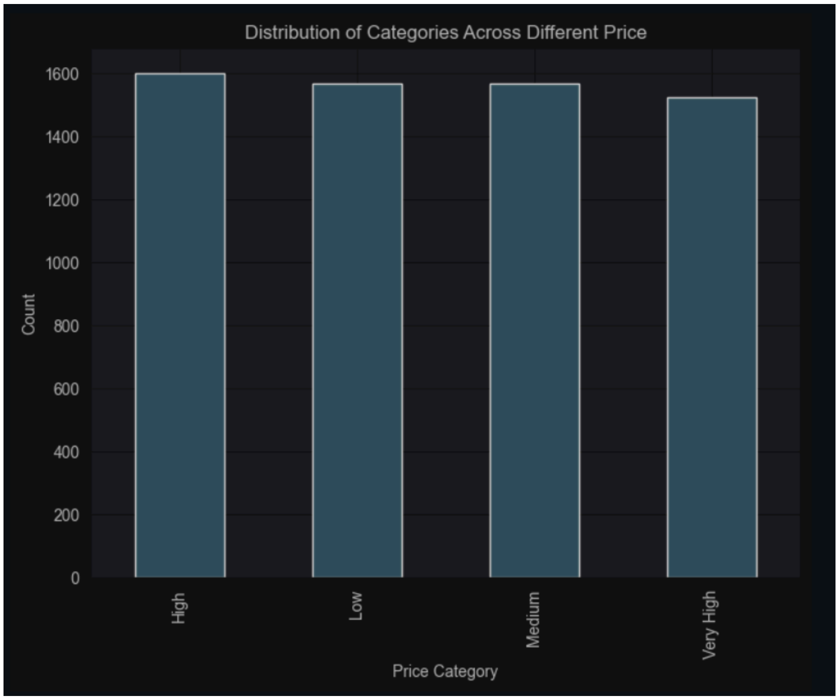
Task: 05 Target Variable Imbalance





Price_Category	count
High	1599
Low	1567
Medium	1566
Very High	1524

The price was somewhat Imbalance for different price Categories but in terms of bin 10 It was highly imbalanced. After Dividing it in price categories It is not that much imbalanced.



Task: 06 Handling imbalanced dataset

Undersampling:

Benefits: this helps in faster model training because of reduced datasize, memory efficient etc.

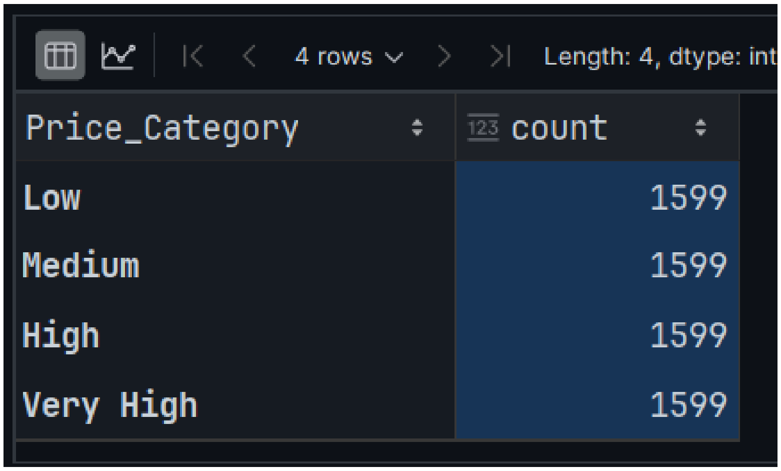
Limitations: We may lose some important examples on which model should have been trained for better performance.

Oversampling:

Benefits: Better generalization of the dataset, keeps all the valuable information inside the database.

Limitations: Increases model training time as model has to be learned on larger dataset now.

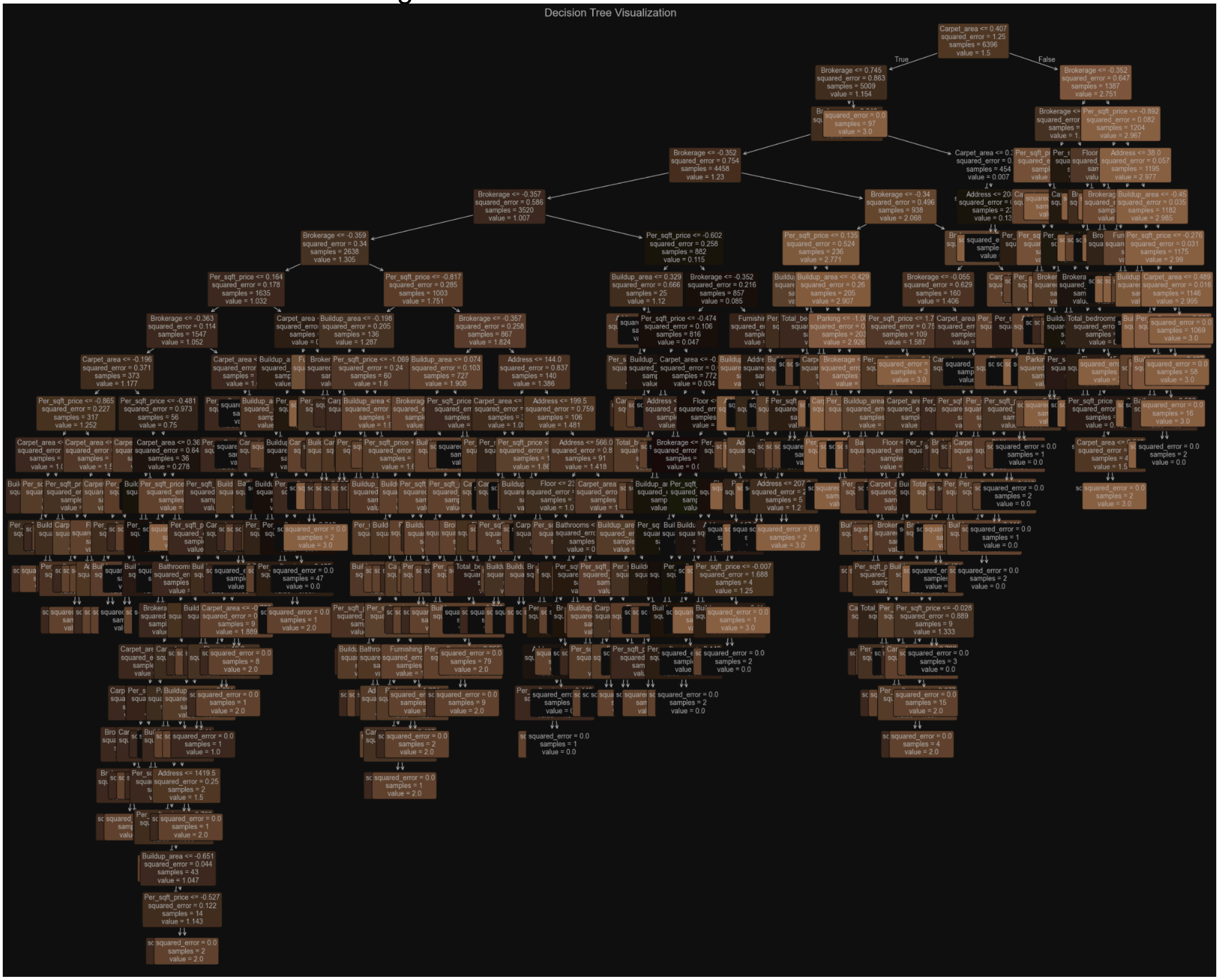
I did oversampling For this particular Dataset so the data Dont lose any Important information.

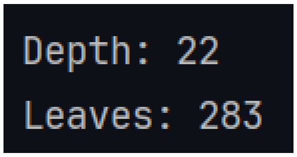


Price_Category	count
Low	1599
Medium	1599
High	1599
Very High	1599

Problem 3: Building The Decision Tree

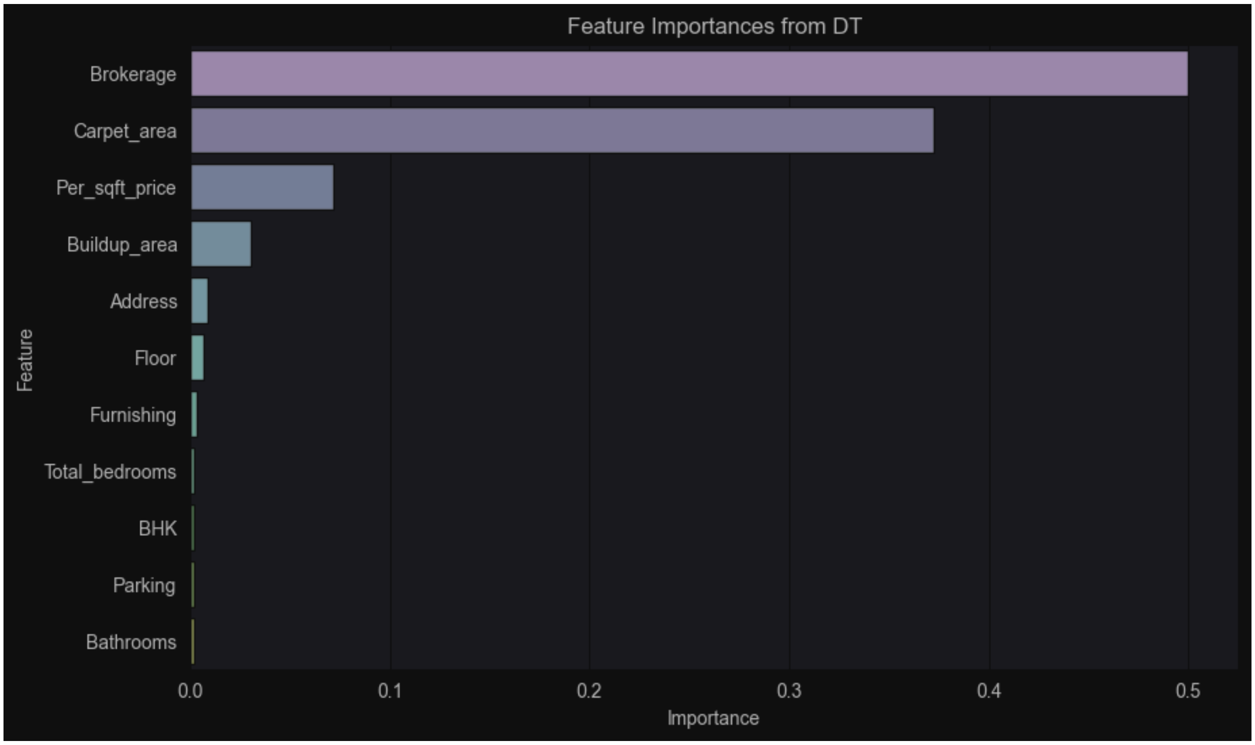
Task 1: Model Training





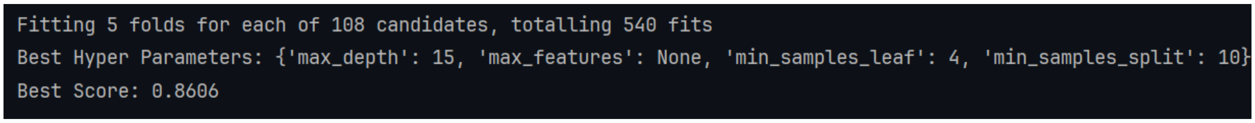
```
Depth: 22
Leaves: 283
```

Task 02: Feature Importance and Hyperparameter Tuning



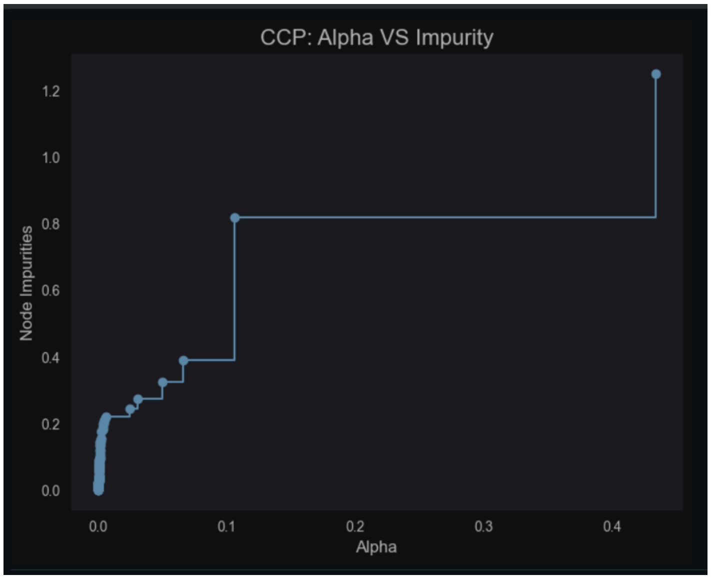
Some features like Brokerage, Carpet Area, per_sqft_price and Buildup_Area are important features because they determine the cost of building that property because more carpet area means more material used hence it costs more, similarly more brokerage required which is ultimately cost increases the overall price of that property.

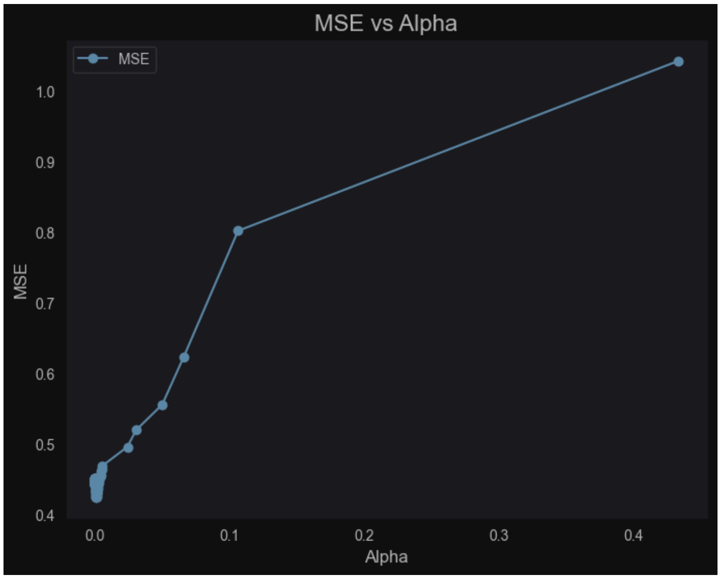
Yes, They do meet my expectations.

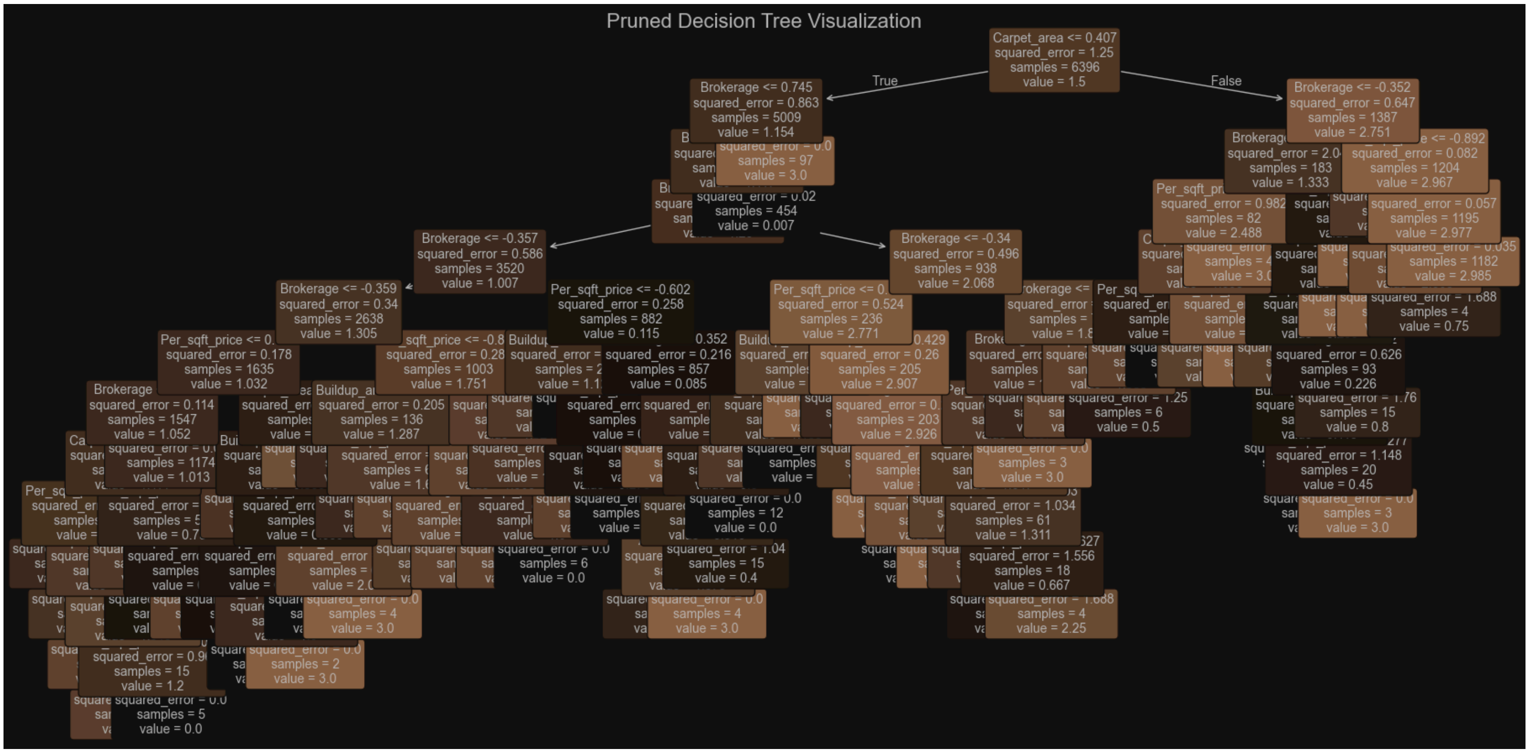


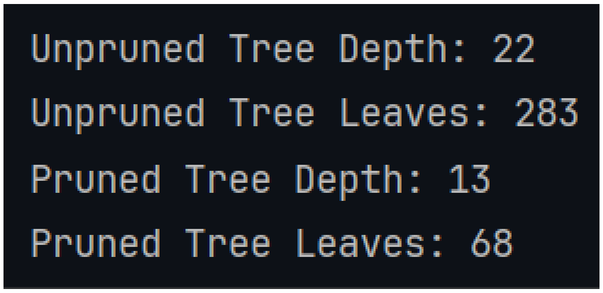
```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
Best Hyper Parameters: {'max_depth': 15, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 10}
Best Score: 0.8606
```

Task 03: Pruning the Decision Tree

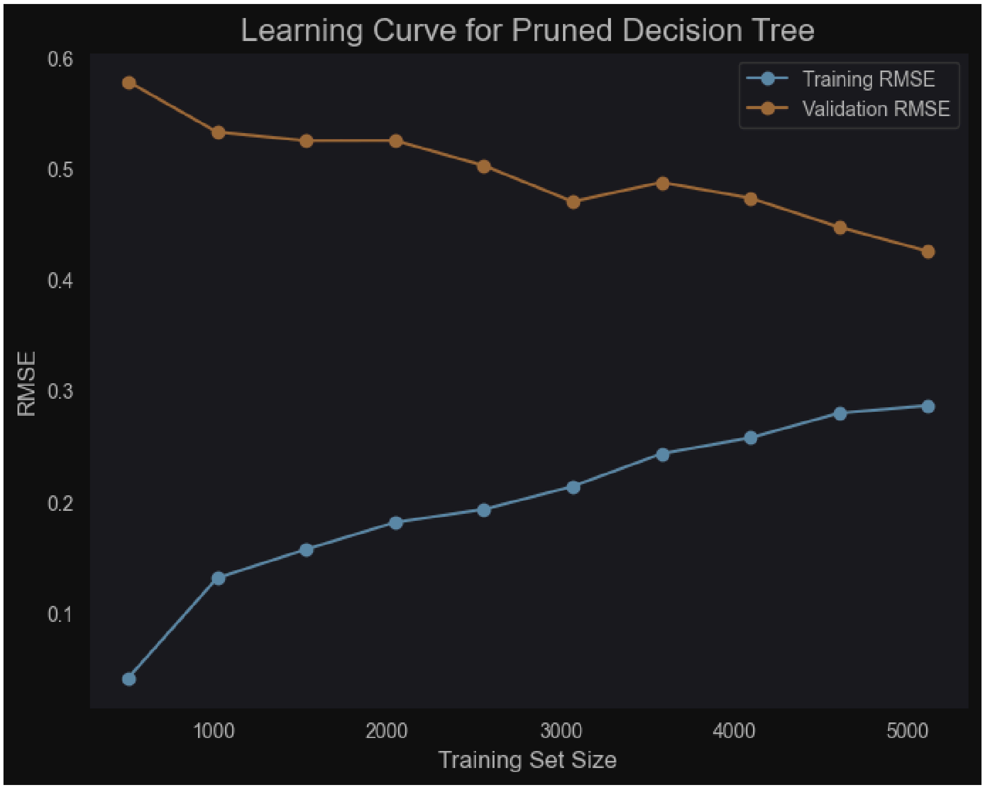






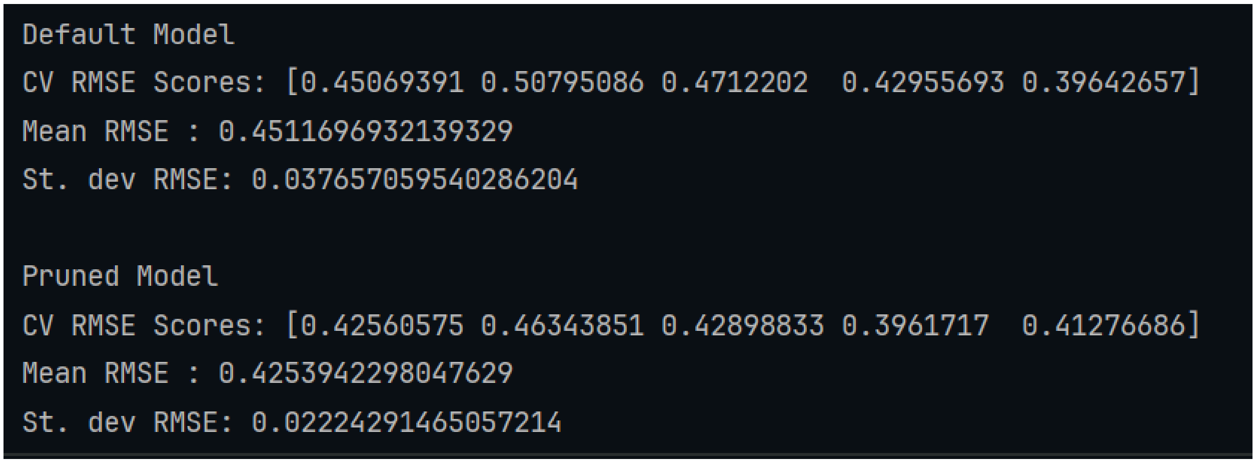


```
Unpruned Tree Depth: 22
Unpruned Tree Leaves: 283
Pruned Tree Depth: 13
Pruned Tree Leaves: 68
```



The tree we made previously was very large in depth and number of leaves that is the reason behind tree that it is overfitting the on the given dataset. To reduce the model’s complexity we pruned it first we determined the optimal value of alpha which is used to prune the tree using the chi squared technique. Now, after pruning the tree the overfitting has reduced as depth and number of leves reduced hence working better on dataset now. Also interpretation is more more simpler then previously because of extensively large decision tree.

Task 04: Handling the Overfitting



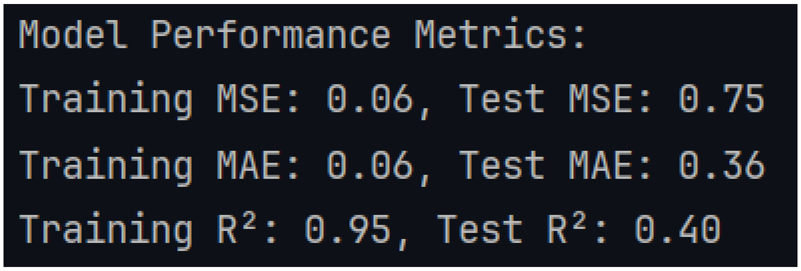
```
Default Model
CV RMSE Scores: [0.45069391 0.50795086 0.4712202  0.42955693 0.39642657]
Mean RMSE : 0.4511696932139329
St. dev RMSE: 0.037657059540286204

Pruned Model
CV RMSE Scores: [0.42560575 0.46343851 0.42898833 0.3961717  0.41276686]
Mean RMSE : 0.4253942298047629
St. dev RMSE: 0.02224291465057214
```

The pruned model can be seen that handled the overfitting well then before this can be seen using these results of Cross validation RMSE scores and st dev reduced which tells that the pruned tree is being generalized better than the default tree.

Cross-validation helps avoid overfitting in Decision Trees by testing the performance of the model on multiple splits of the data, which ensures that it generalizes well to unseen data. This prevents over-reliance on a single train-test split and provides an estimate of out-of-sample error. Validating performance across folds helps identify hyperparameters, for example, max depth, min samples split, that balance the complexity of the model and generalization.

Problem 4: Model Evaluation and Error Analysis

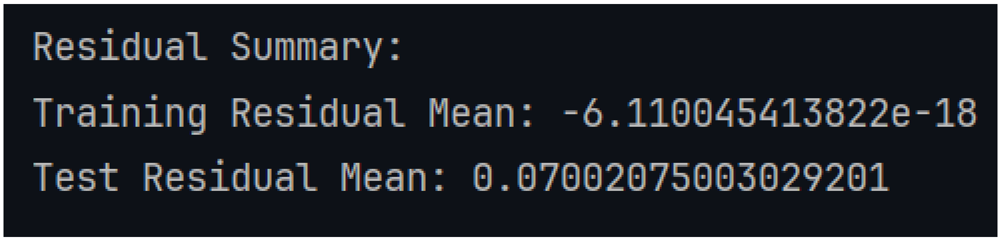
Task 01: Model Evaluation



```
Model Performance Metrics:
Training MSE: 0.06, Test MSE: 0.75
Training MAE: 0.06, Test MAE: 0.36
Training R²: 0.95, Test R²: 0.40
```

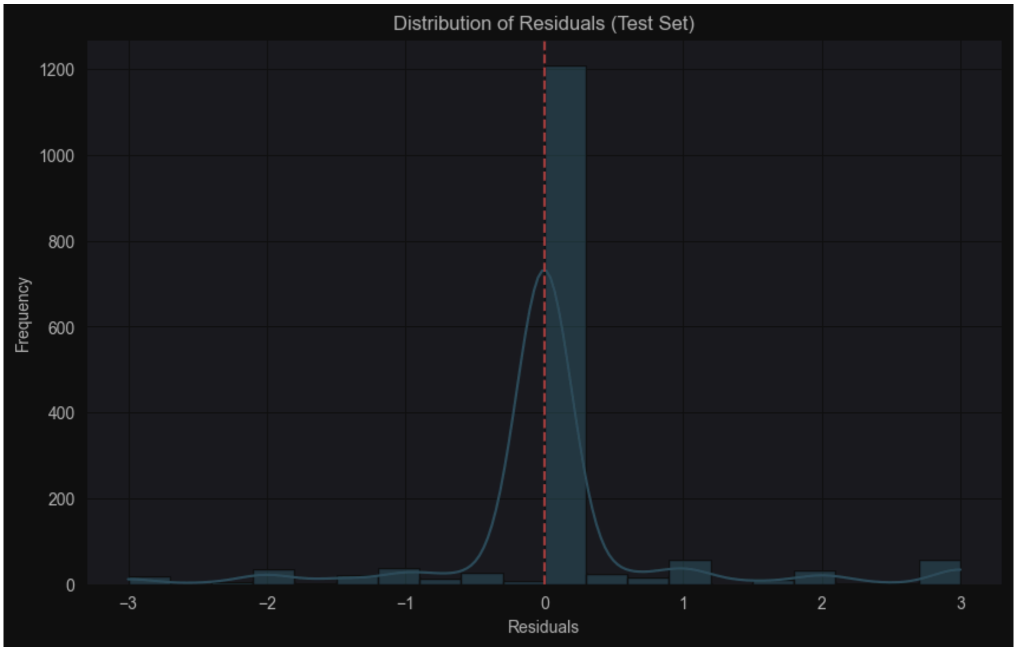
The model has overfitting that learns the training data superbly but fails to generalize over test data. It would therefore require regularization, or more hyperparameter adjustments that would reduce the model complexity.

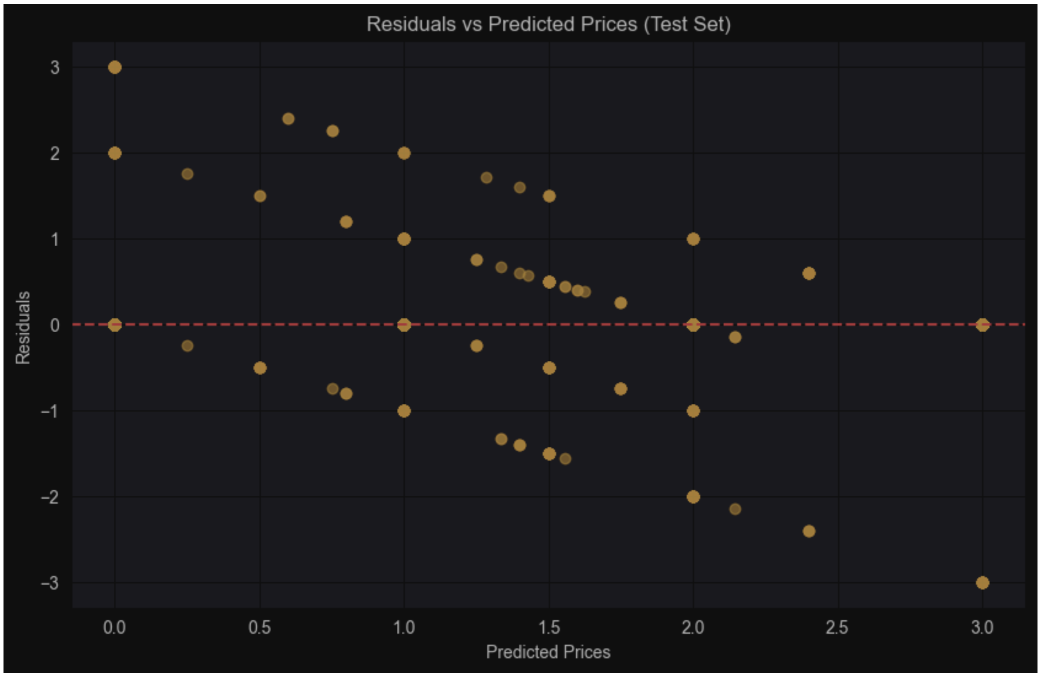
Task 02: Residual and Error Analysis



```
Residual Summary:
Training Residual Mean: -6.110045413822e-18
Test Residual Mean: 0.07002075003029201
```

Residual errors are very minimum therefore shows that the model has very low bias there is overfitting is also there.

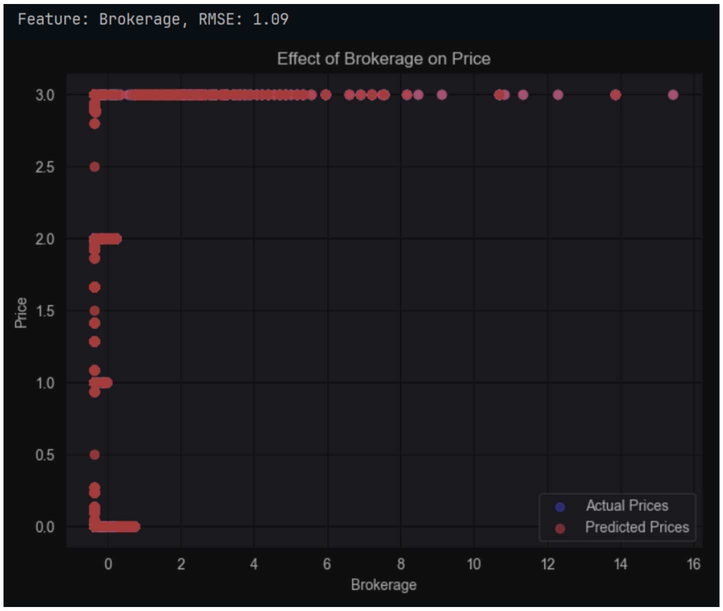




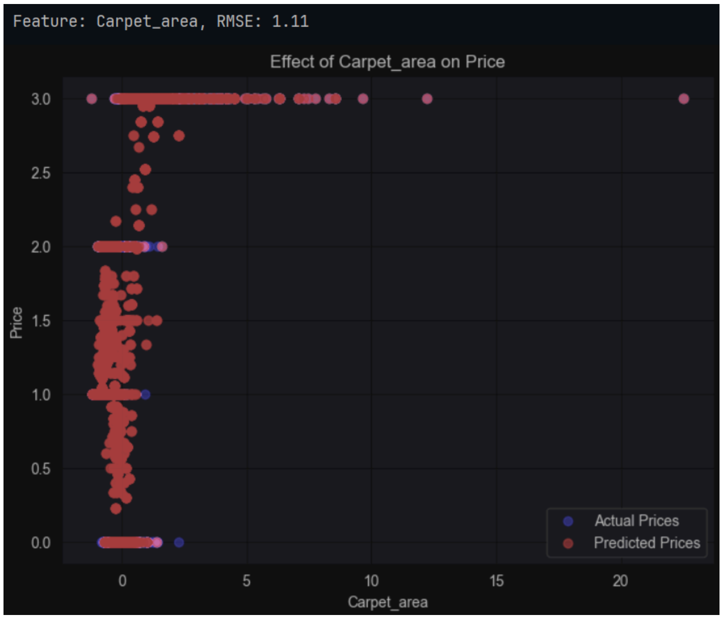
The residual analysis of the model shows that it is generally non-biased since the residuals are centered about zero. However, the sharp peak at zero indicateslow error, and the tails indicate possible outliers with significant deviations. However, it exhibits patterned clustering and deviation, particularly to the lower and higher points of the predicted price ranges, which indicates that the model is not doing well to some subsets of data may be due to complexity in a relationship between features or weak generalization to extreme cases of the data.

To overcome these problems, feature engineering can be used to extract the complex interactions or non-linearities in the data, and robust modeling techniques like gradient boosting or ensemble methods might improve the performance on outliers and edge cases.

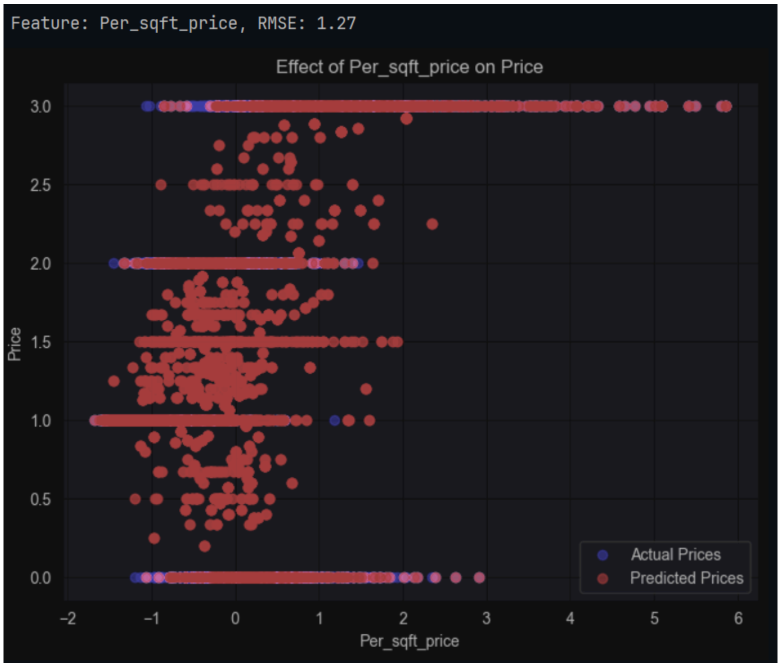
Task 3: Feature Importance based analysis



```
Feature: Brokerage, RMSE: 1.09
```



```
Feature: Carpet_area, RMSE: 1.11
```



```
Feature: Per_sqft_price, RMSE: 1.27
```

For Brokerage, most the price predictions are concentrated about fixed levels, with minimum variation. For Carpet Area, although the area increases vaguely related to higher prices, the model cannot capture these slight variations that may have occurred. Lastly, Per Square Foot Price is quite highly distributed, which reveals its huge but scattered influence on the price predictions. Per Square Foot Price, though a very influential variable, has the highest RMSE at 1.27, which may be due to its complex and scattered relationship with the target variable.