AI A4

THEORY

Date ..../..../.....

Q1) ~~Root~~ Entropy $H(S) = -\left(\frac{5}{10}\log_2 \right.$

Entropy $H(S) = -\sum_{i=a} p_i \log_2(p_i)$

Root Entropy $H(S) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right)$

$= 1$

Attr: Long-term Debt,

$H(LTD = Yes) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722$

$IG(S, LTD) = H(S) - \sum_{v} P(v) \cdot H(LTD = v)$

$H(LTD = No) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722$

$IG(S, LTD) = 1 - \left(\frac{1}{2}\cdot 0.722 + \frac{1}{2}\cdot 0.722\right) = 0.278$

Attr: Unemployed

$H(U = No) = -\left(\frac{5}{8}\log\frac{5}{8} + \frac{3}{8}\log\frac{3}{8}\right) = 0.954$

$H(U = Yes) = -(1\log 1 + 0\log 0) = 0$

$\therefore IG(S, U) = 1 - \left(\frac{8}{10}\cdot\frac{4}{5}\cdot 0.954 + 0\right) = 0.237$

Attr. Credit Rating

$H(CR = Good) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) = 0.49$

$H(CR = Bad) = -\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right) = 0.985$

$IG(S, CR) = 1 - \left(\frac{3}{10}\cdot 0.49 + \frac{7}{10}\cdot 0.985\right) = 0.164$

Spiral

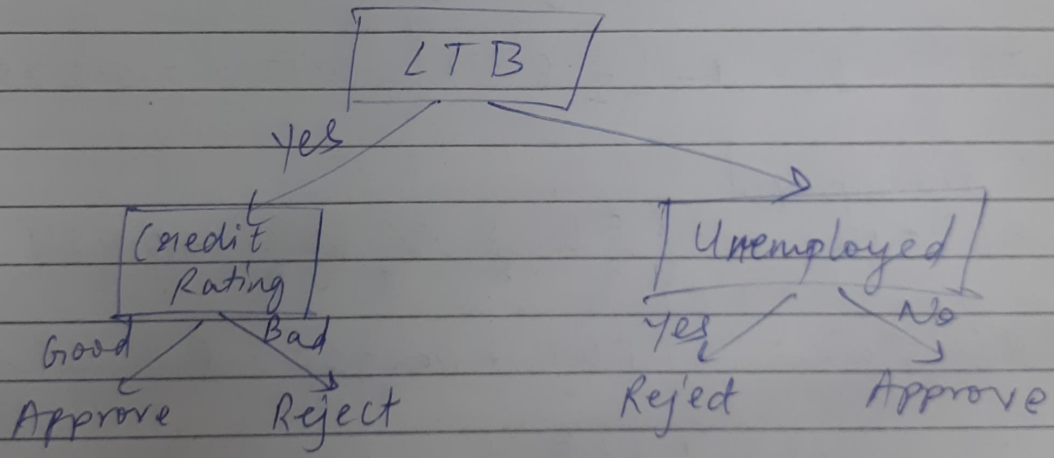
Similarly,

IG(S, Payment) = 0.029

∴ First Split on Debt (LTD)

Debt = Yes	Debt = No
IG(U) = 0.0722 IG(R) = 0.721 IG(P) = 0.17	IG(U) = 0.721 IG(R) = 0.321 IG(P) = 0.17
∴ Next Node is Credit Rating No further Split Needed ∵ H(Rating) = 0	∴ Next Node is Unemployed No further Split Needed ∵ H(Unemployed) = 0

∴ Decision Tree:



Training Error = 0

∵ No misclassifications.

Q2)
Unique values in each column

```
index              7820
Address            3725
Possesion             1
Furnishing            3
Buildup_area       1038
Carpet_area        2893
Bathrooms           104
Property_age         46
Parking              10
Price               832
Brokerage          1785
Floor               132
Per_sqft_price     2801
BHK                   9
Total_bedrooms       32
dtype: int64
```

Statistical Analysis

```
Statistical Analysis:
                 count          mean           std       min           25%  \
index           7820.0  4.873975e+03  2.766650e+03       1.0  2.497750e+03
Buildup_area    7820.0  1.116096e+03  7.222222e+02     180.0  6.500000e+02
Carpet_area     7820.0  8.620953e+02  5.733111e+02     150.0  4.738816e+02
Bathrooms       7820.0  1.973660e+00  9.005205e-01       1.0  1.000000e+00
Property_age    7820.0  7.471611e+00  7.217703e+00       1.0  2.000000e+00
Parking         7820.0  1.303581e+00  7.970482e-01       0.0  1.000000e+00
Price           7820.0  3.038559e+07  3.719014e+07  780000.0  1.050000e+07
Brokerage       7820.0  1.131909e+07  3.102861e+07       0.0  9.999900e+04
Floor           7820.0  1.993028e+01  1.396096e+01       2.0  1.000000e+01
Per_sqft_price  7820.0  2.340171e+04  1.300058e+04    1440.0  1.560000e+04
BHK             7820.0  2.154923e+00  9.999399e-01       1.0  1.000000e+00
Total_bedrooms  7820.0  2.201048e+00  9.798746e-01       1.0  1.000000e+00

                      50%          75%          max
index          4.908500e+03      7267.25       9546.0
Buildup_area   9.435000e+02      1322.00      15000.0
Carpet_area    7.077226e+02      1050.00      14000.0
Bathrooms      2.000000e+00         2.00         10.0
Property_age   5.000000e+00        10.00         99.0
Parking        1.000000e+00         2.00          9.0
Price          1.920000e+07  35000000.00  500000000.0
Brokerage      2.500000e+05  10700000.00  500000000.0
Floor          1.600000e+01        23.00         99.0
Per_sqft_price 2.143000e+04     28850.00     100000.0
BHK            2.000000e+00         3.00         10.0
Total_bedrooms 2.000000e+00         3.00         10.0
```

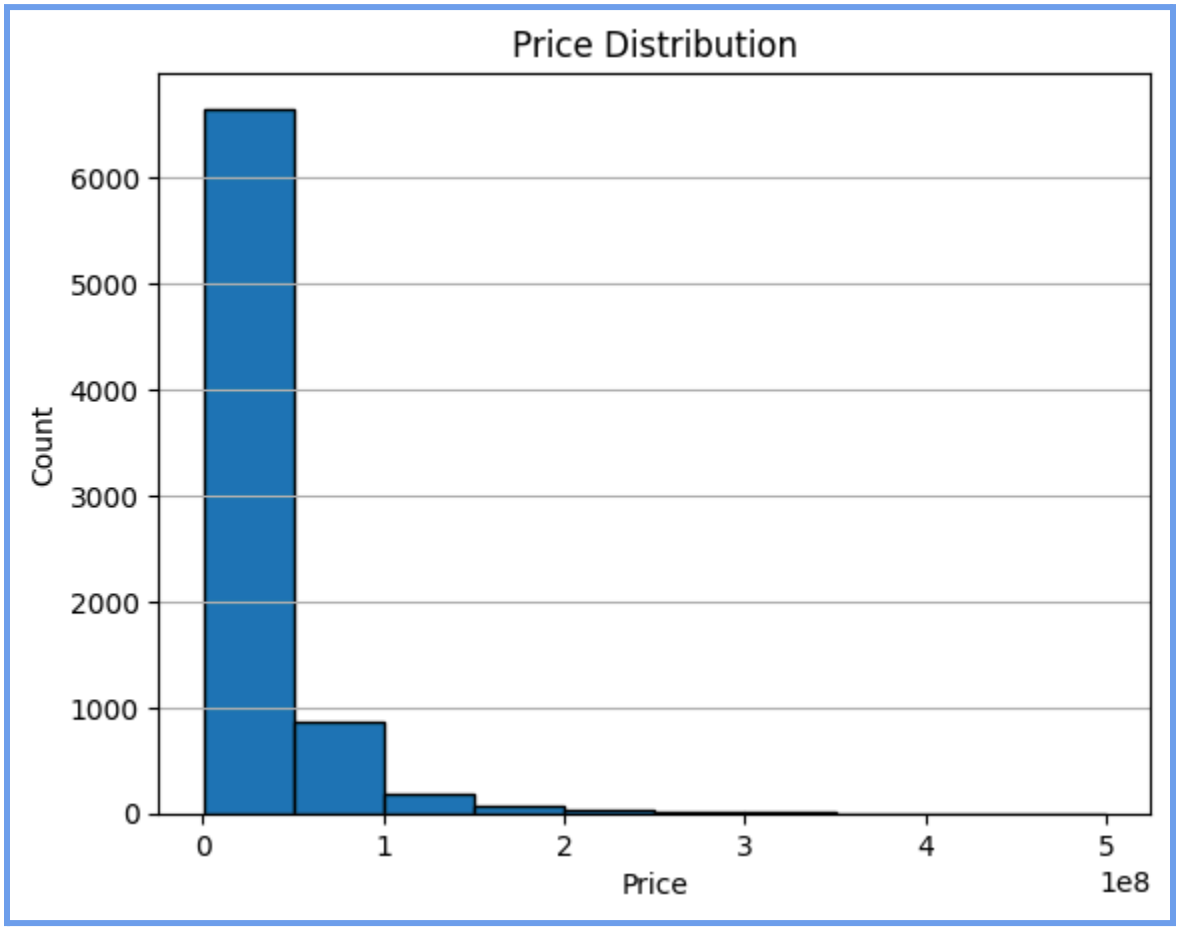
Dropping Irrelevant Columns: “index”, “Property_age” because of very less correlation with Target column “Price” and column “Adress” because it had too many values and column “Possesion” as it had only one value

Scaling was pretty much the same, did not improve by much

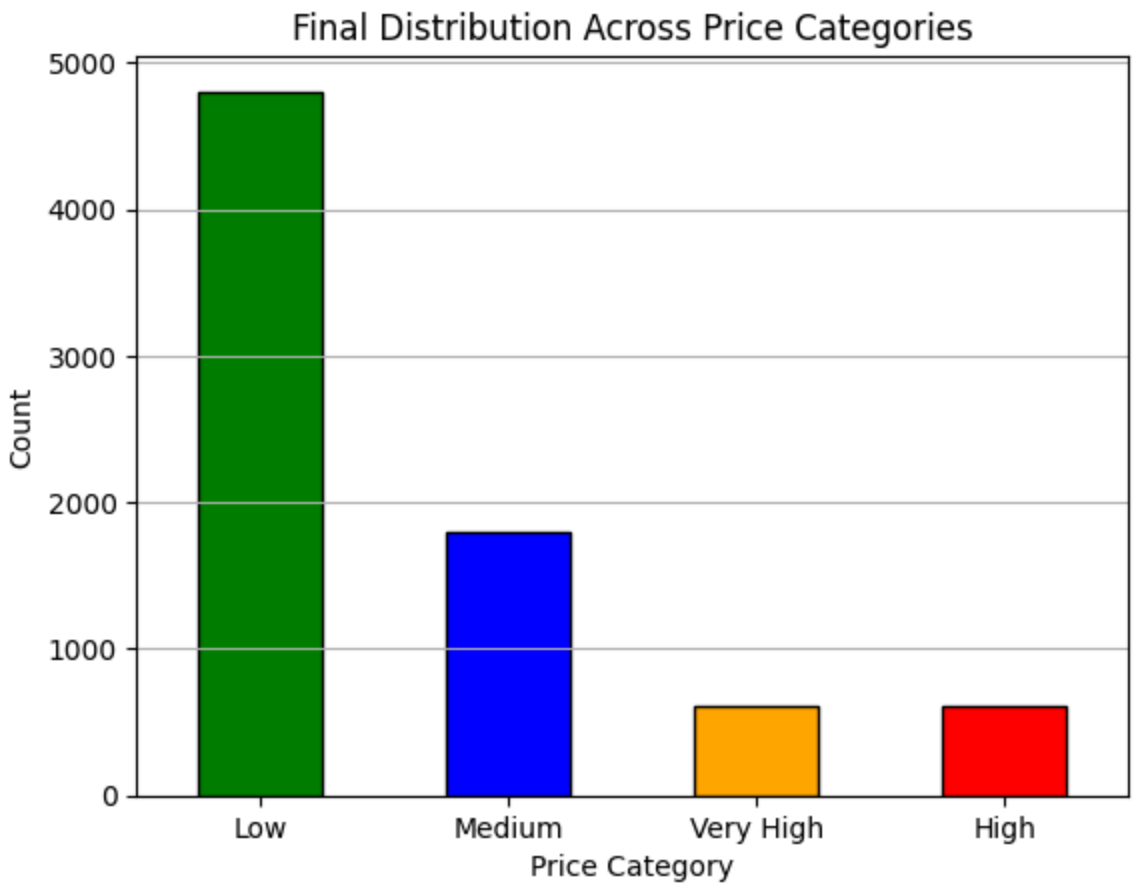
```
Decision Tree Regressor on Original Data
MSE Train: 3192439897317.046, MSE Test: 7859160104188.375
RMSE Train: 1786740.0195095665, RMSE Test: 2803419.359316115
R2 Train: 0.9977774847286836, R2 Test: 0.9932728184831365
Adjusted R2 Train: 0.9977739258571522, Adjusted R2 Test: 0.9932295011520557
MAE Train: 250733.73013566426, MAE Test: 929711.6368286443
```

```
Decision Tree Regressor on Scaled Data
MSE Train: 3190970046701.833, MSE Test: 7988000906911.587
RMSE Train: 1786328.6502493971, RMSE Test: 2826305.168751525
R2 Train: 0.9977785080104191, R2 Test: 0.9931625350107035
Adjusted R2 Train: 0.9977749507774494, Adjusted R2 Test: 0.9931185075477975
MAE Train: 250383.39194373402, MAE Test: 936024.3499573742
```

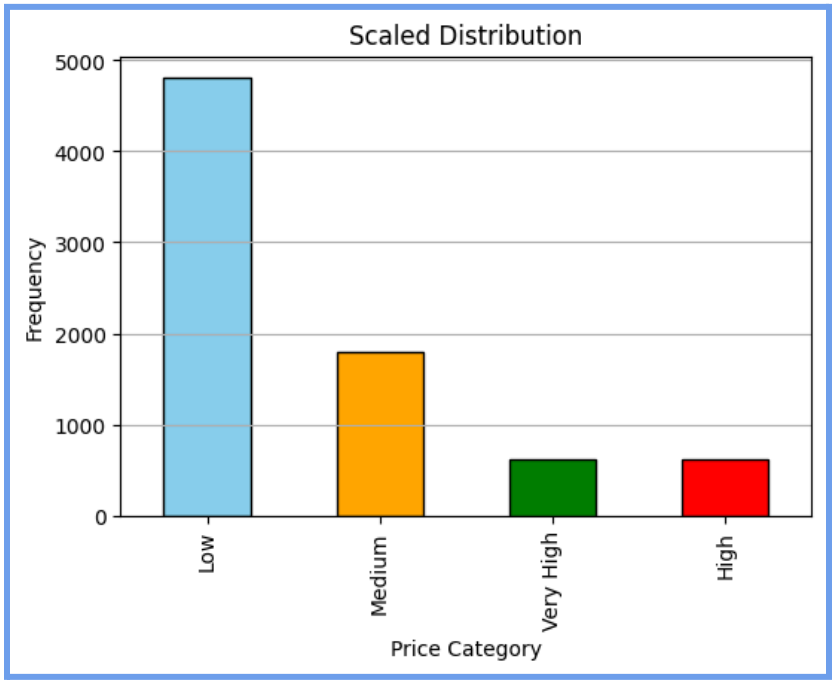
Target Imbalance:

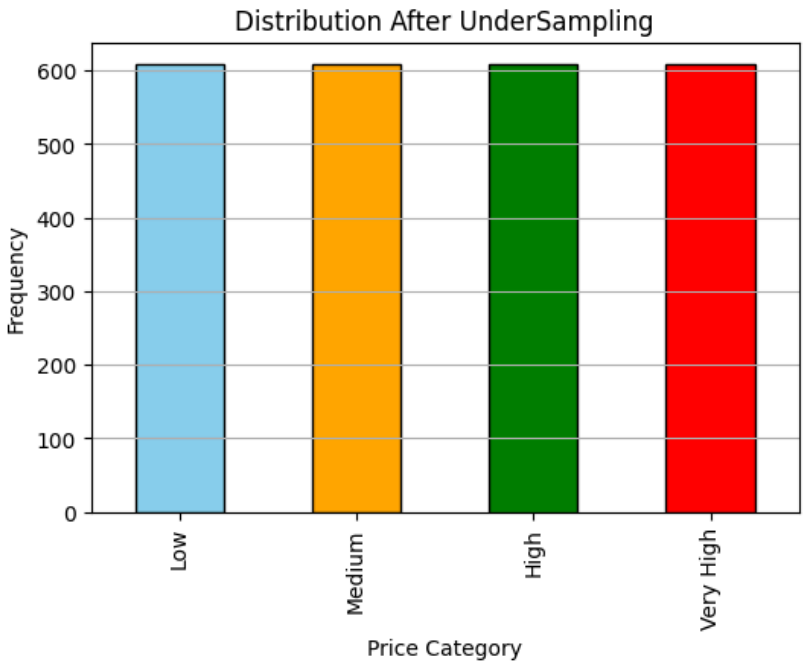


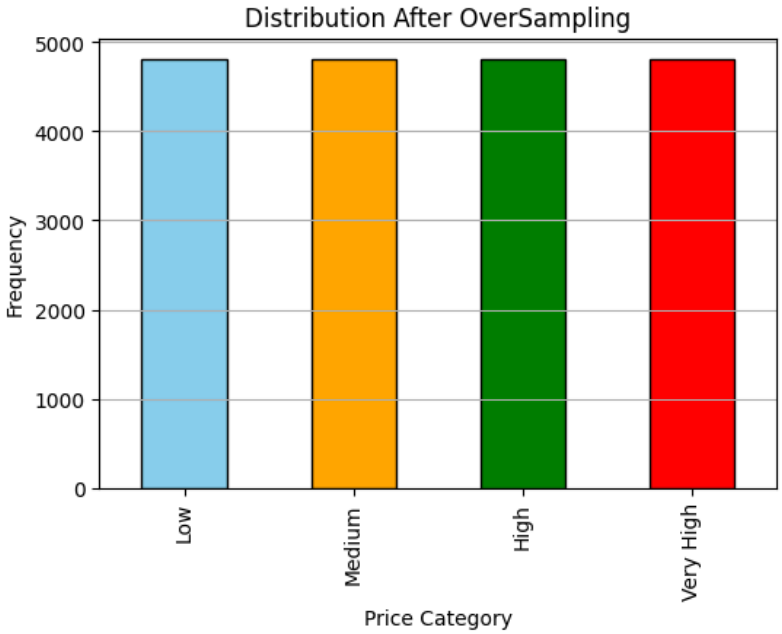
Creating Price Categories



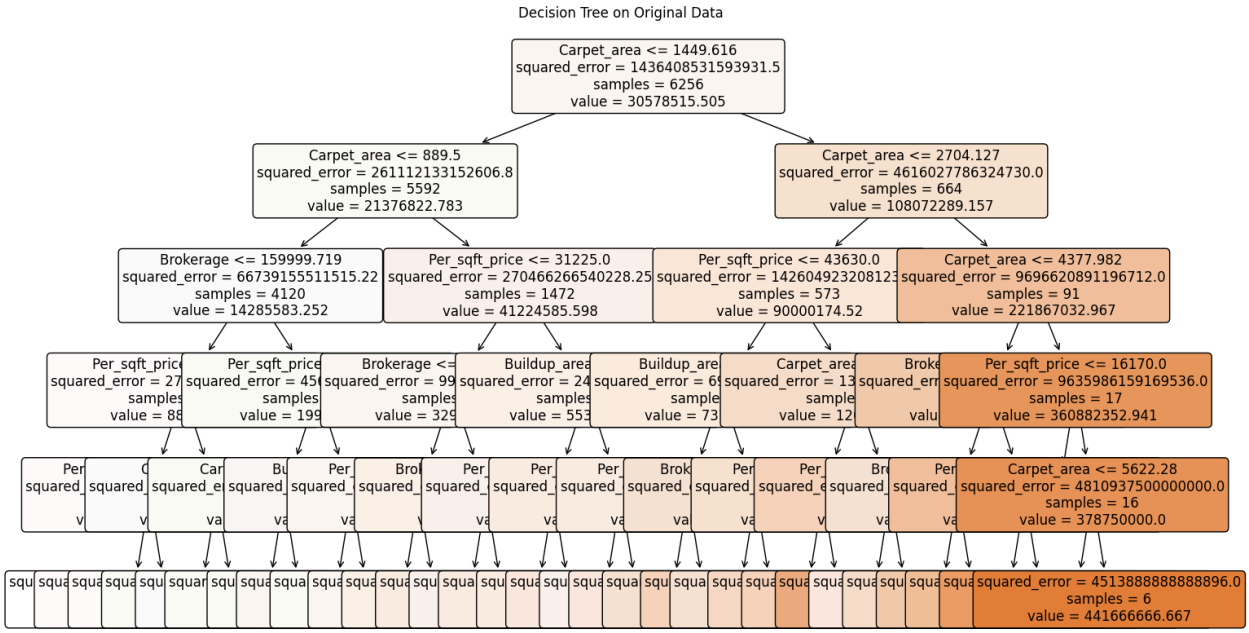
Over and Under Sampling performed



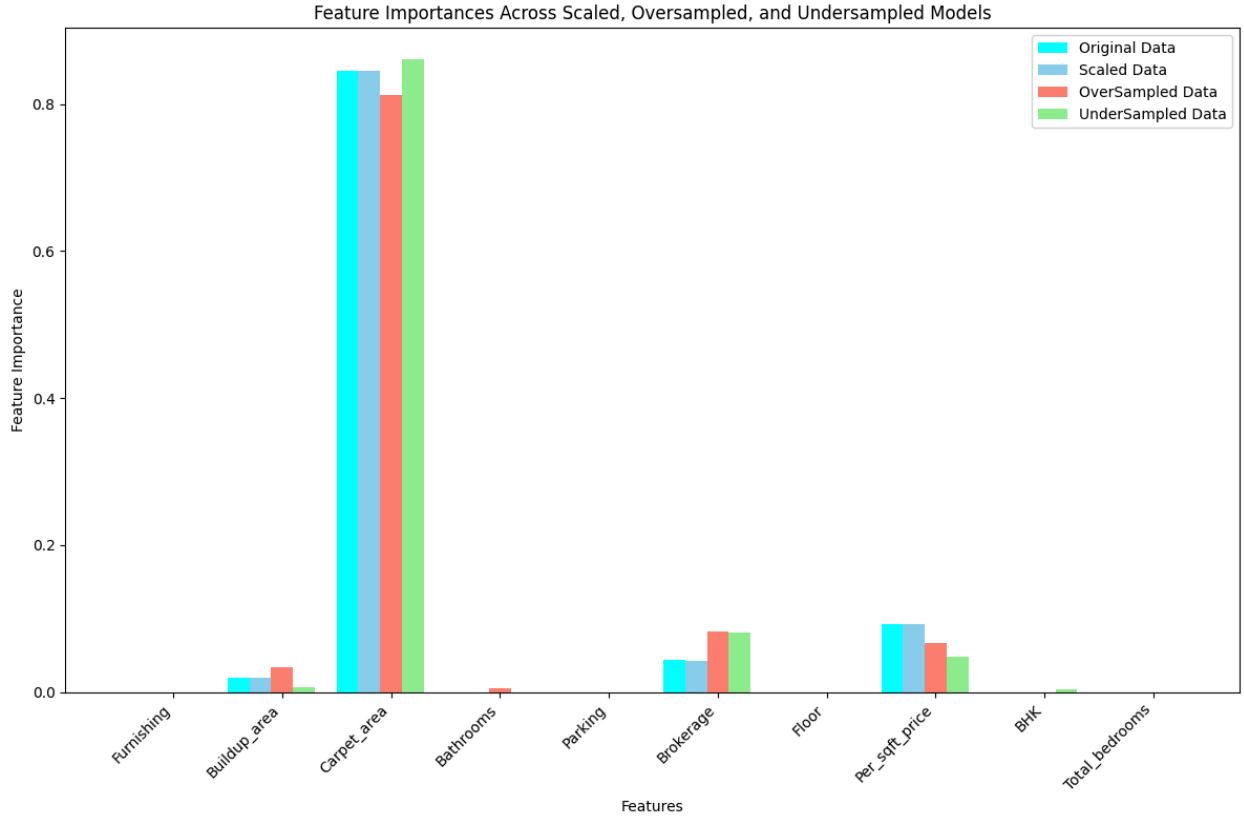




Q3)
Decision Tree on Original Data

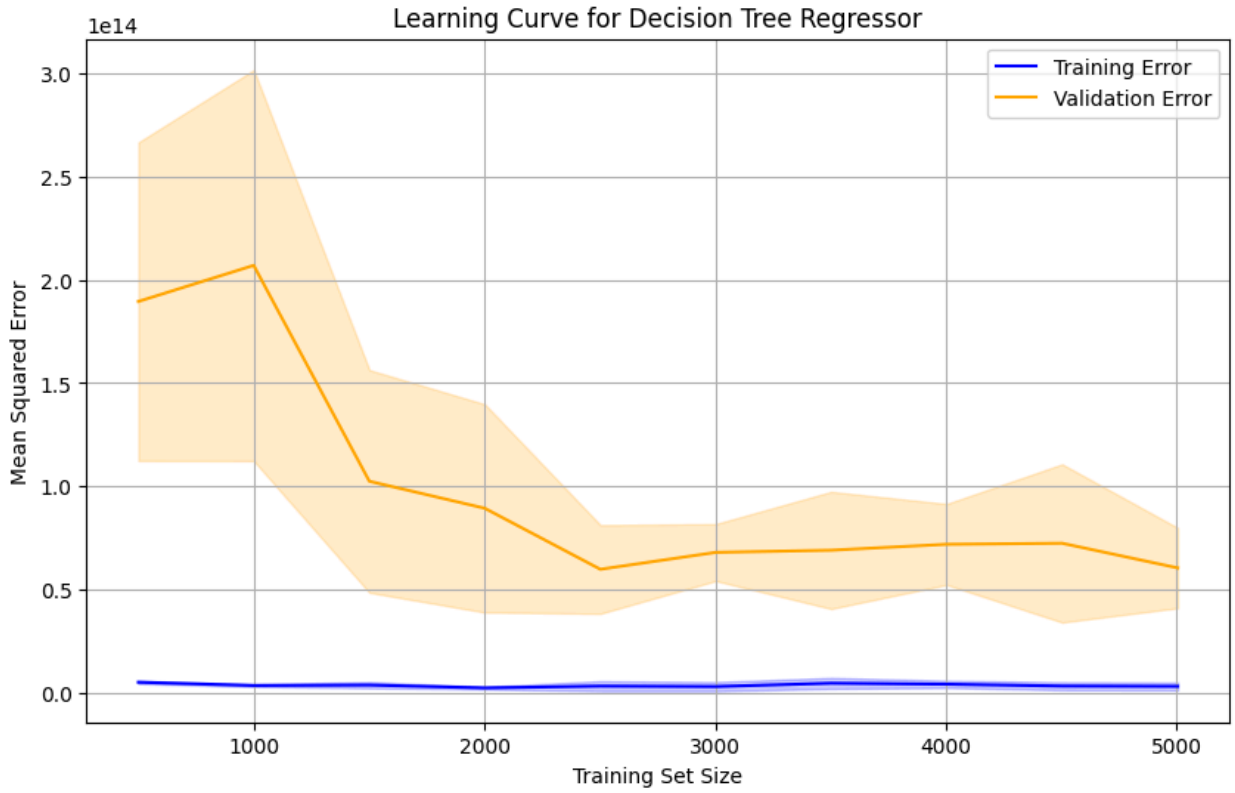


Feature Importance



Correlation and variability is high for features like Carpet_area and it is expected so because if the area is high the Price must be high as well

Learning Curve:



Q4)

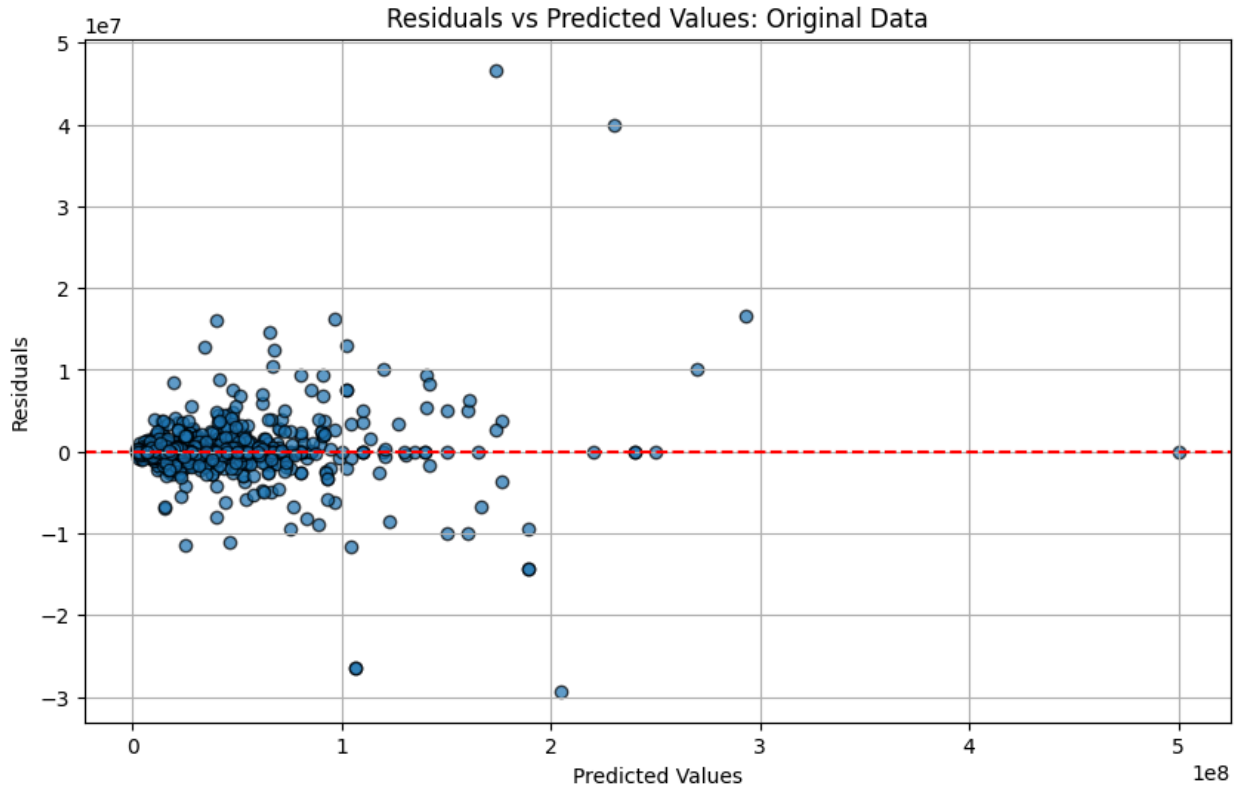
```
Decision Tree Regressor on Original Data
MSE Train: 3192439897317.046, MSE Test: 7859160104188.375
RMSE Train: 1786740.0195095665, RMSE Test: 2803419.359316115
R2 Train: 0.9977774847286836, R2 Test: 0.9932728184831365
Adjusted R2 Train: 0.9977739258571522, Adjusted R2 Test: 0.9932295011520557
MAE Train: 250733.73013566426, MAE Test: 929711.6368286443
```

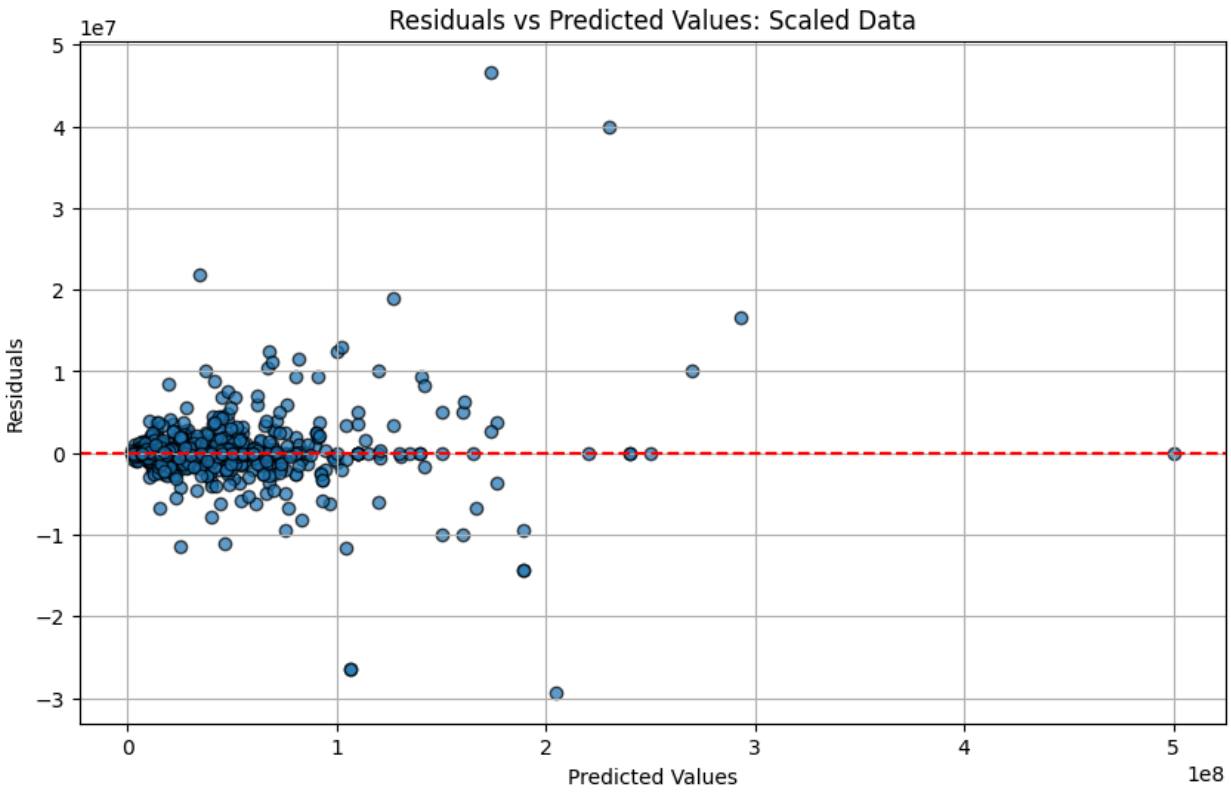
```
Decision Tree Regressor on Scaled Data
MSE Train: 3190970046701.833, MSE Test: 7988000906911.587
RMSE Train: 1786328.6502493971, RMSE Test: 2826305.168751525
R2 Train: 0.9977785080104191, R2 Test: 0.9931625350107035
Adjusted R2 Train: 0.9977749507774494, Adjusted R2 Test: 0.9931185075477975
MAE Train: 250383.39194373402, MAE Test: 936024.3499573742
```

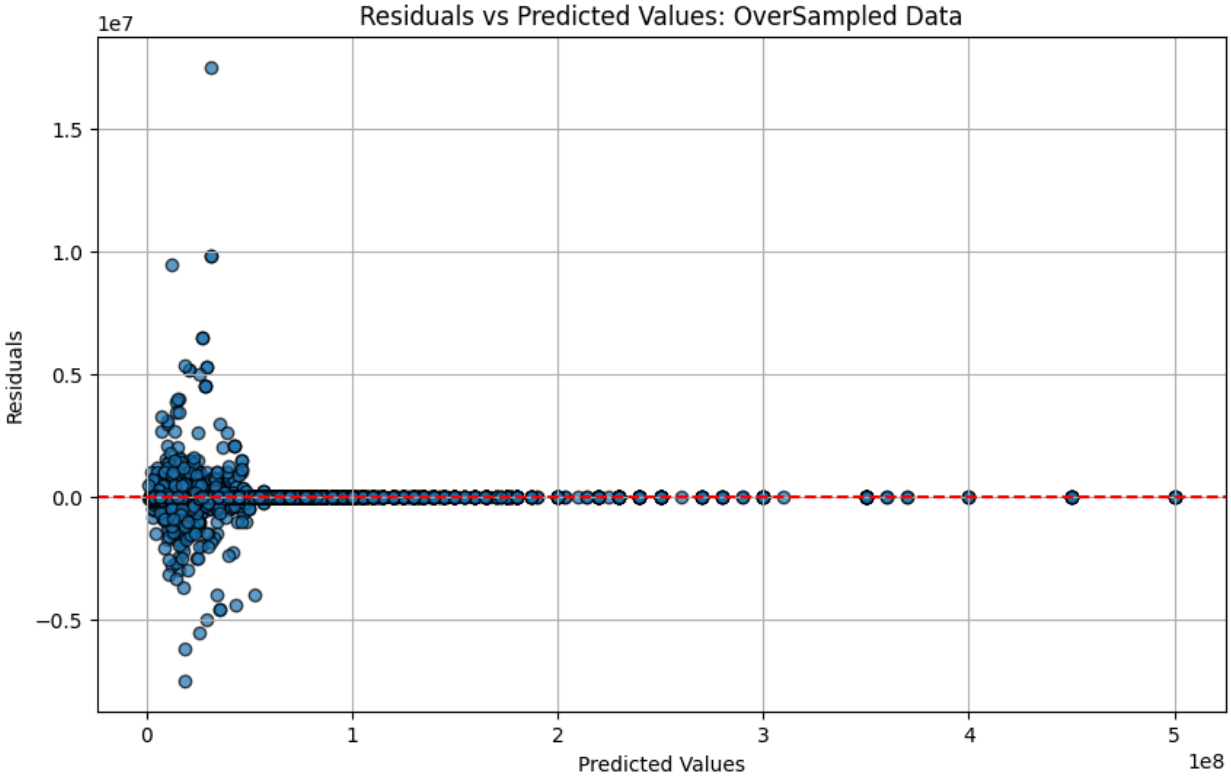
```
Decision Tree Regressor on Over-Sampled Data
MSE Train: 26675034919.138054, MSE Test: 444262059203.01483
RMSE Train: 163324.93661146192, RMSE Test: 666529.8636993056
R2 Train: 0.9999911201289802, R2 Test: 0.9998615873572727
Adjusted R2 Train: 0.9999911143496955, Adjusted R2 Test: 0.9998612262494128
MAE Train: 30435.527228581337, MAE Test: 158169.59330558448
```

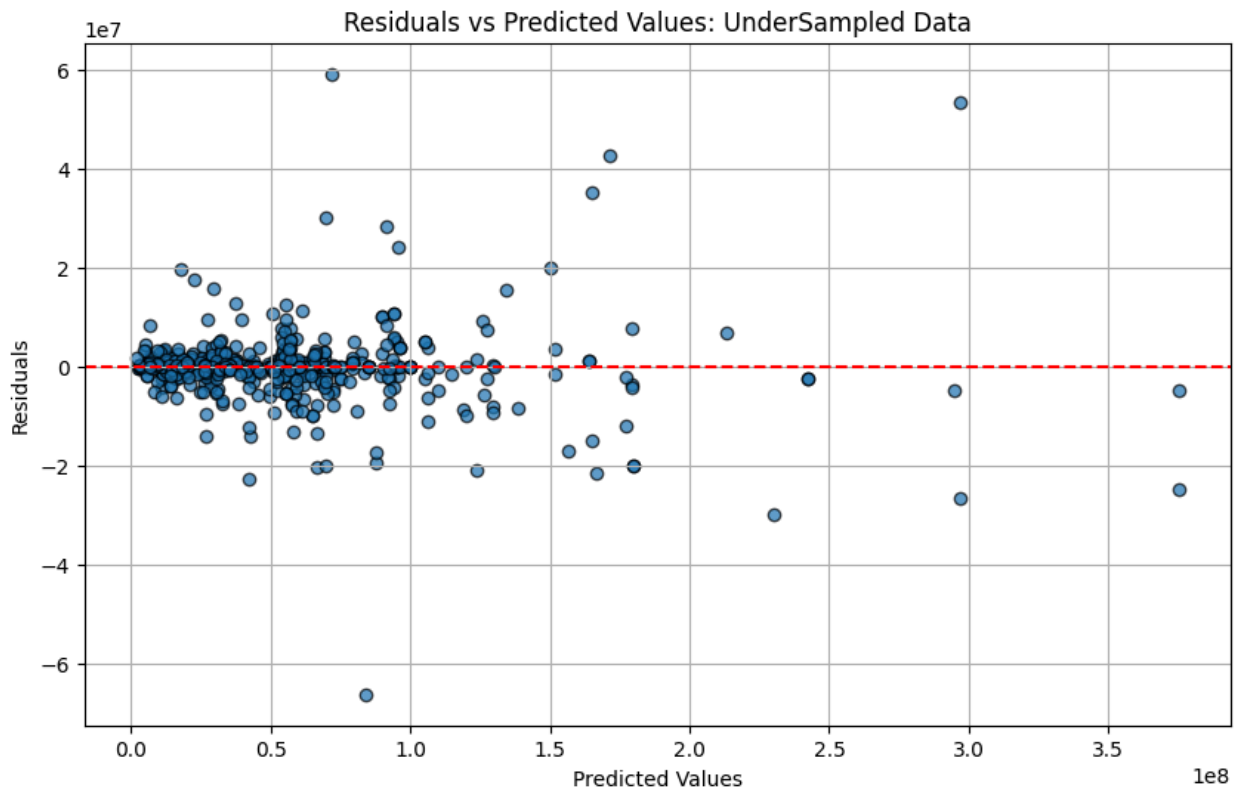
```
Decision Tree Regressor on UnderSampled Data
MSE Train: 19135900186589.547, MSE Test: 63479796426876.57
RMSE Train: 4374459.987997324, RMSE Test: 7967420.939480766
R2 Train: 0.9938207370239615, R2 Test: 0.9739193845447782
Adjusted R2 Train: 0.9937887863363914, Adjusted R2 Test: 0.9733714724553828
MAE Train: 1385049.8714652956, MAE Test: 3806381.930184805
```

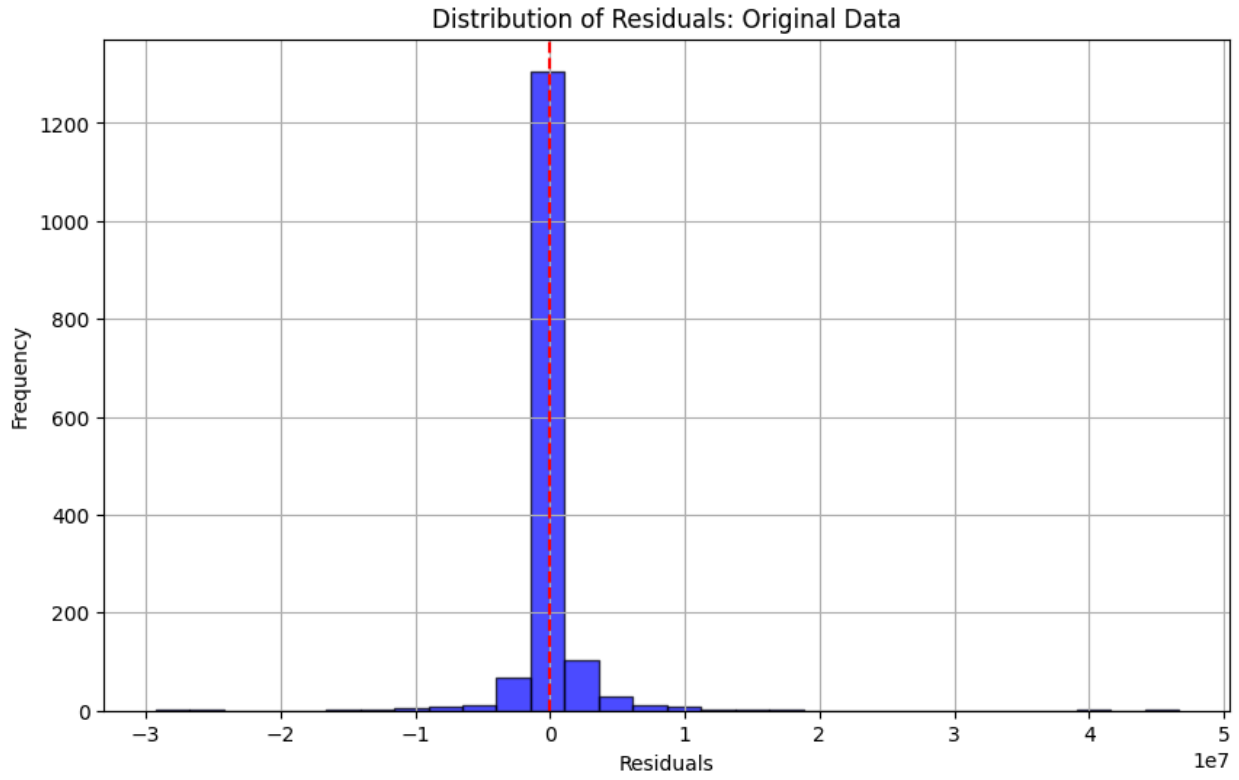
Residual plots:

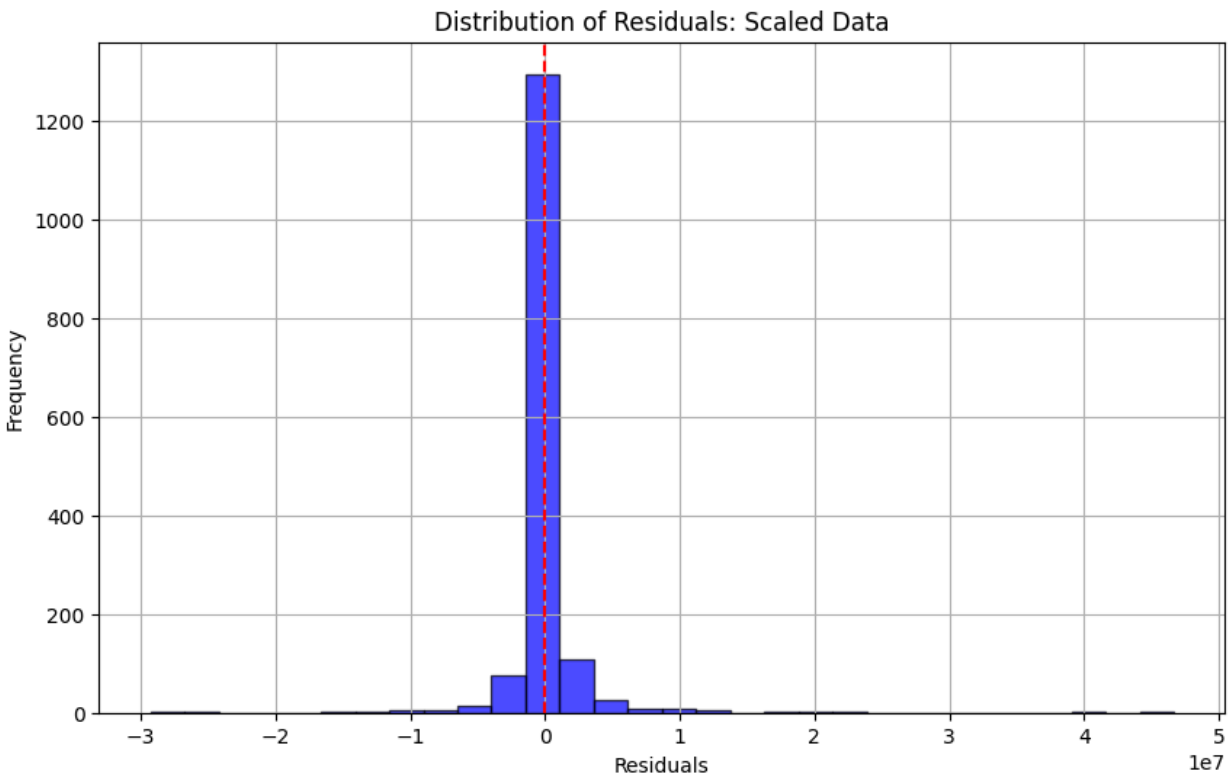


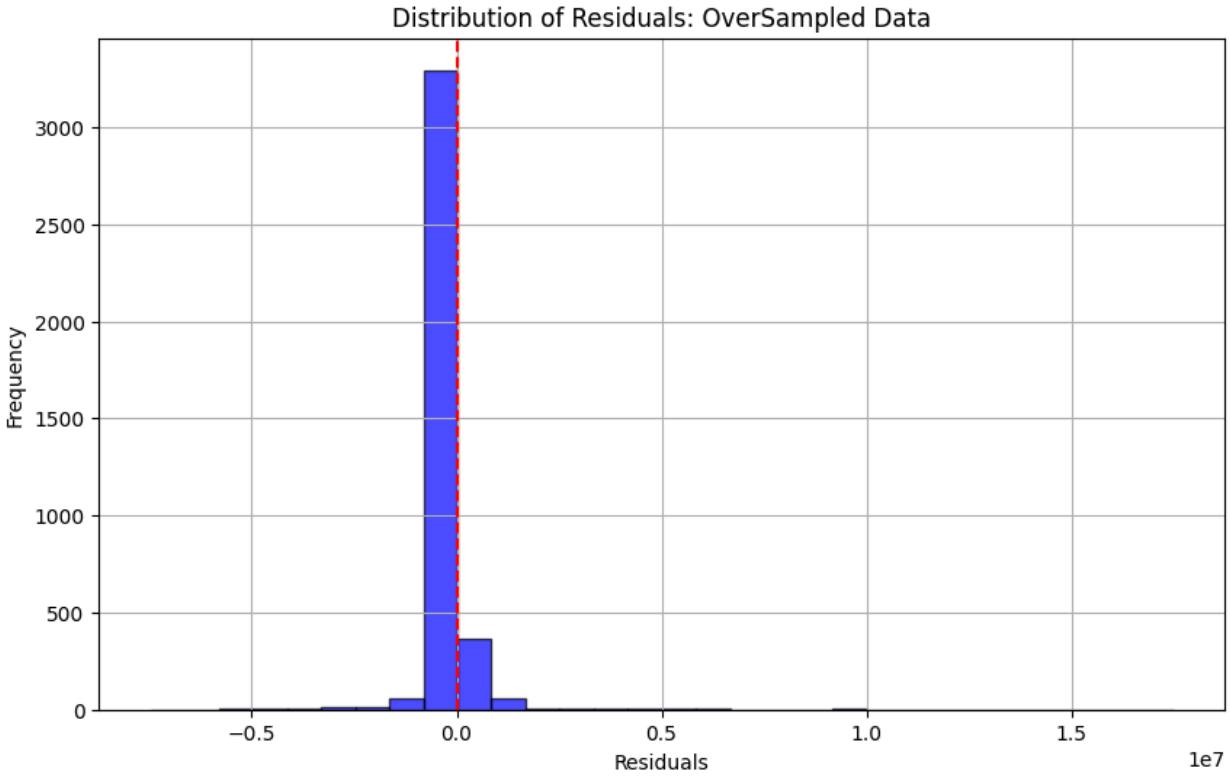

Residuals vs Predicted Values: Scaled Data
1e7
5
4
3
2
1
0
−1
−2
−3
Residuals
0
1
2
3
4
5
1e8
Predicted Values

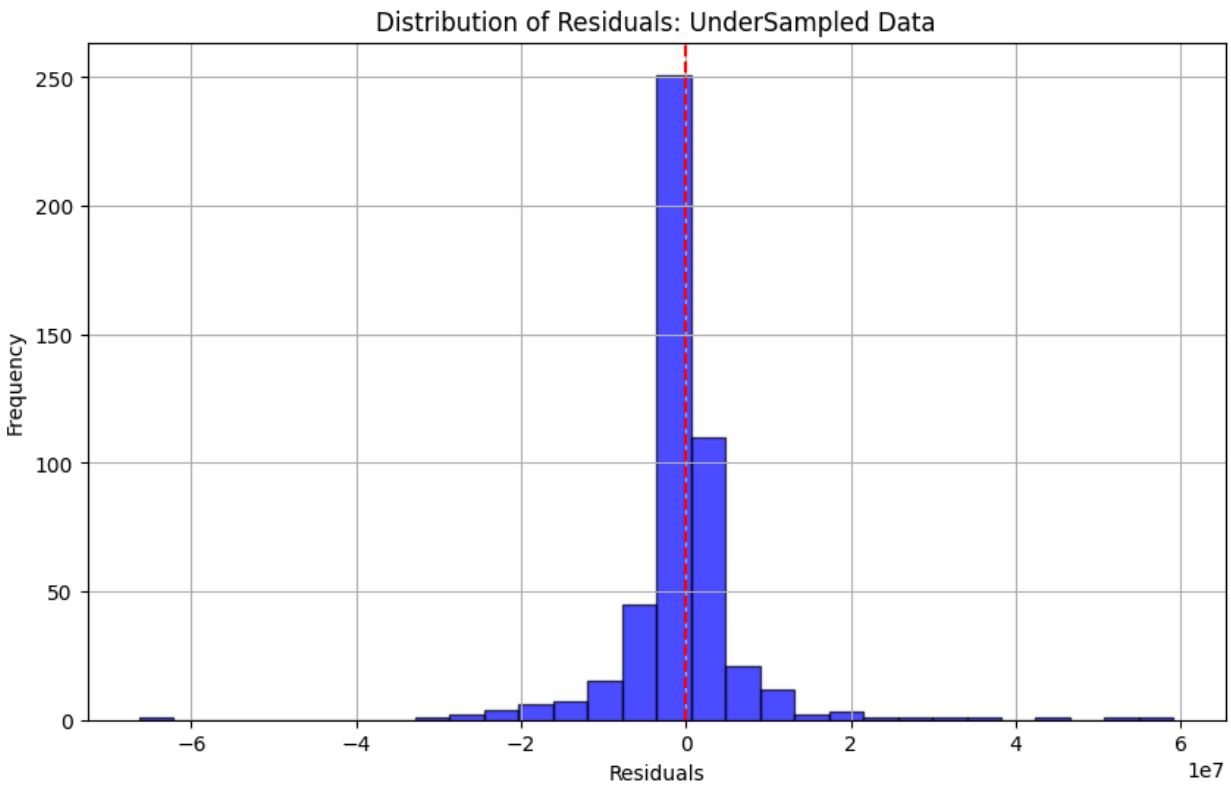

Residuals vs Predicted Values: OverSampled Data
1e7
1.5
1.0
0.5
0.0
−0.5
Residuals
0
1
2
3
4
5
1e8
Predicted Values


Residuals vs Predicted Values: UnderSampled Data
1e7
6
4
2
0
−2
−4
−6
Residuals
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Predicted Values
1e8

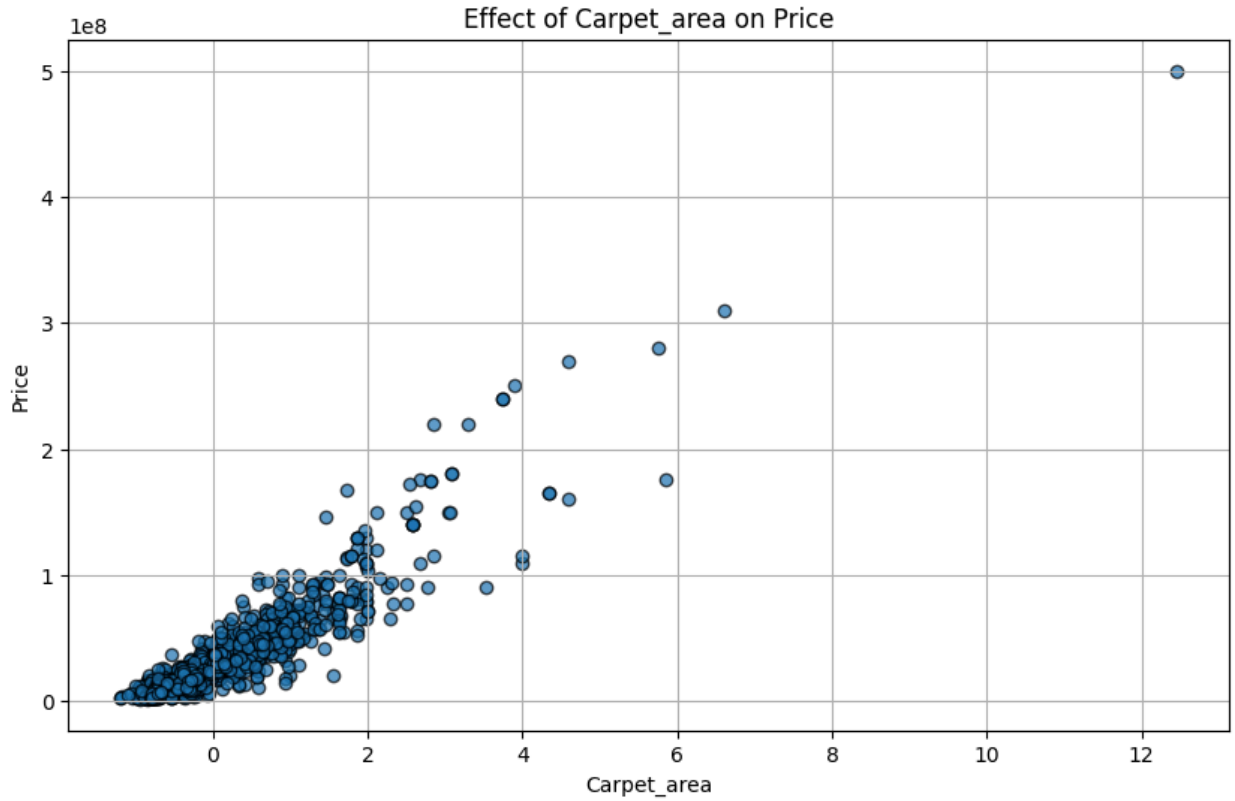

Distribution of Residuals: Original Data
1200
1000
800
600
400
200
0
Frequency
−3
−2
−1
0
1
2
3
4
5
Residuals
1e7

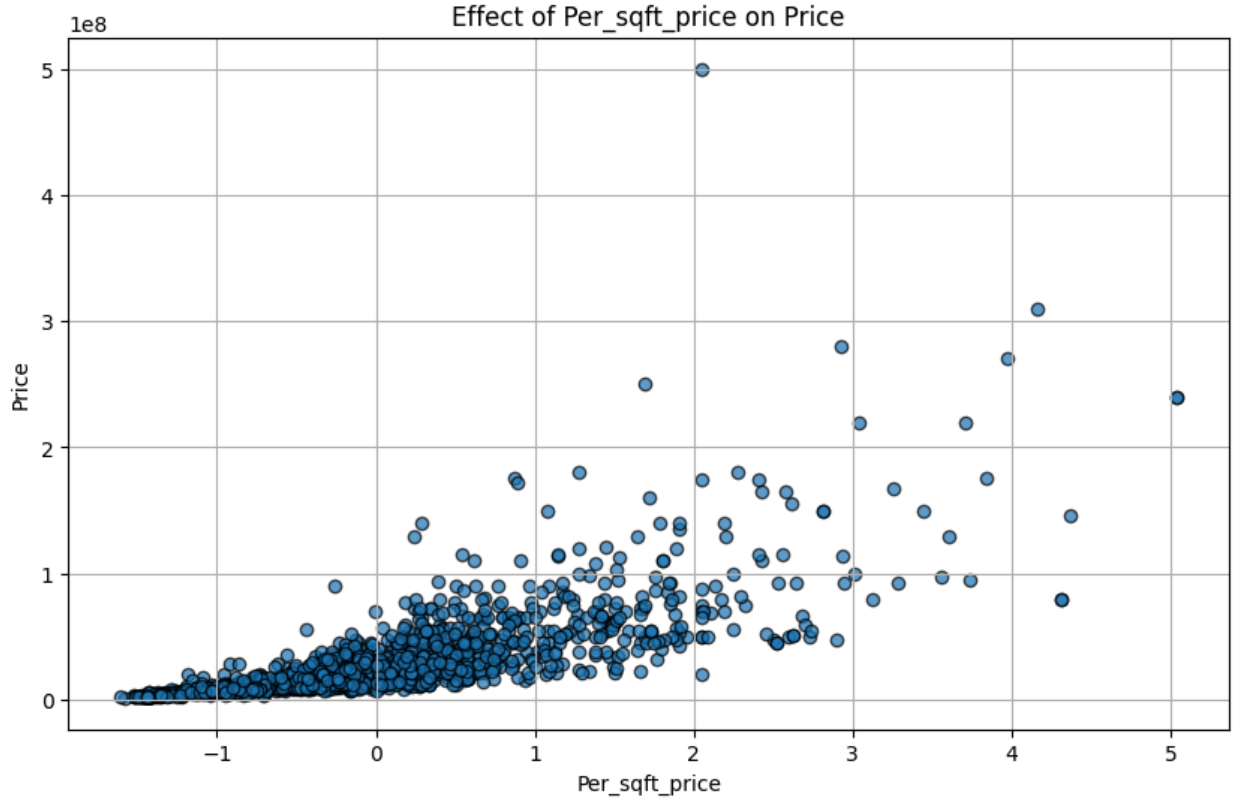

Distribution of Residuals: Scaled Data
Frequency
Residuals
1e7
0
200
400
600
800
1000
1200
−3
−2
−1
0
1
2
3
4
5

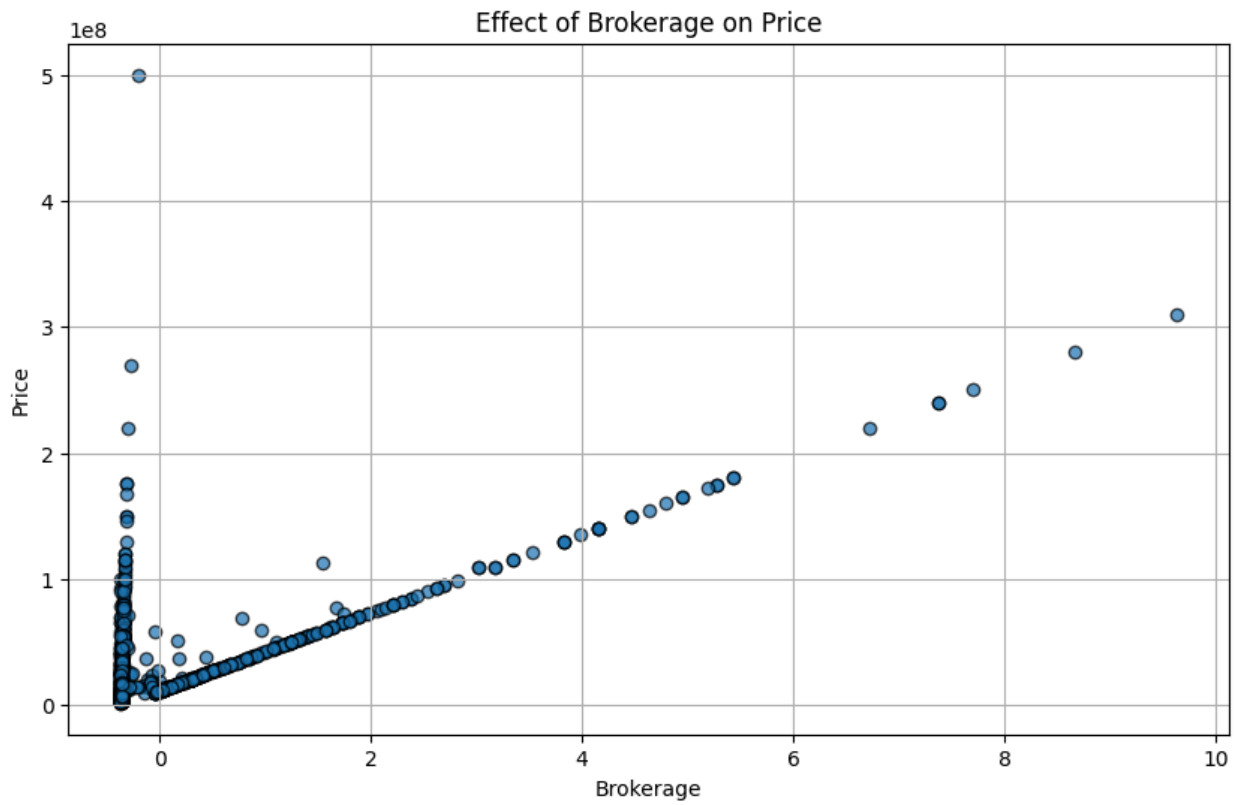

Distribution of Residuals: OverSampled Data
Frequency
Residuals
1e7
0
500
1000
1500
2000
2500
3000
−0.5
0.0
0.5
1.0
1.5



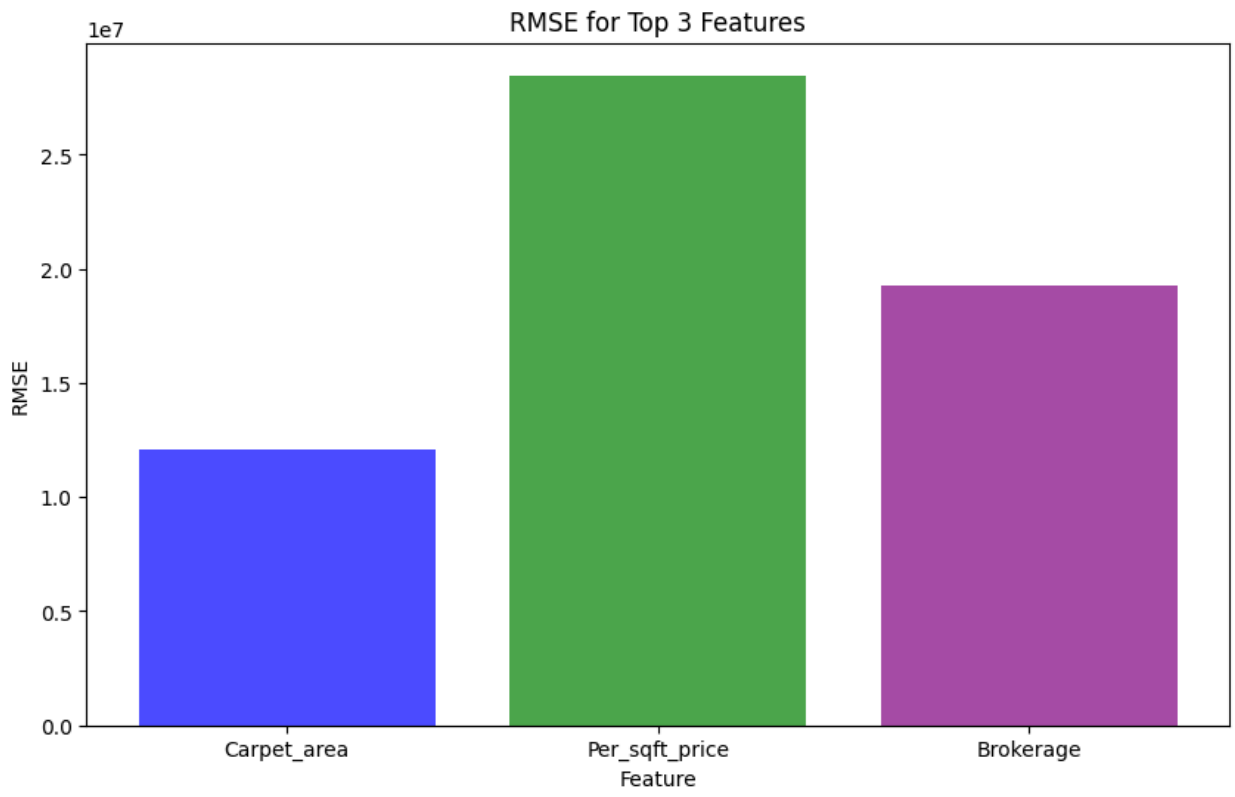
Effect of top features on Price


Effect of Carpet_area on Price
1e8
Price
Carpet_area
0
1
2
3
4
5
0
2
4
6
8
10
12


Effect of Per_sqft_price on Price
1e8
Price
Per_sqft_price
0
1
2
3
4
5
−1
0
1
2
3
4
5



RMSE for Top 3 Features


RMSE for Top 3 Features
1e7
2.5
2.0
1.5
1.0
0.5
0.0
RMSE
Carpet_area
Per_sqft_price
Brokerage
Feature