Prediction of The Accommodation Price in Tehran, Iran

Niloufar Saeedi[1] and Hassan Kabirian[2]

***Abstract*— The price of accommodation is rising rapidly in Iran, especially in Tehran, due to the huge current inflation. In this project, we use seven main features of an accommodation to see how any of these in relation to each other, and separately affect the total estimated price of a house, flat, or apartment. Our results will help the home-seekers and real-state holders estimate and validate an accommodation with specific features.**

I. INTRODUCTION

We use the dataset collected by Mohammadreza Karimnejad [1]. The features include Area in square meters, Number of bedrooms, Has Parking or not, Has elevator or not, Has warehouse or not, The region where the house is placed, and Price in Toman and USD.

II. PREPROCESSING

A. Column 'Address'

We had 192 unique values of the names of the neighborhoods. We used a Python script to extract each neighborhood's geographical longitude and latitude and added these two columns instead of the previous one.

B. Column 'Price(USD)'

We dropped the column Price(USD) as it was highly dependent on the column 'Price'(in Toman) and useless in our predictions.

C. Column 'Area'

We converted the column Area's datatype from string to float.

D. Outliers of the column 'Area' and 'Price'

We detected and removed the outliers using the box plot in Figure 1. We did the same with the column 'Price'.

[1]N. Saeedi is with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran nilousaeedi@gmail.com

[2]H. Kabirian is with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran hassan.hk.ka@gmail.com

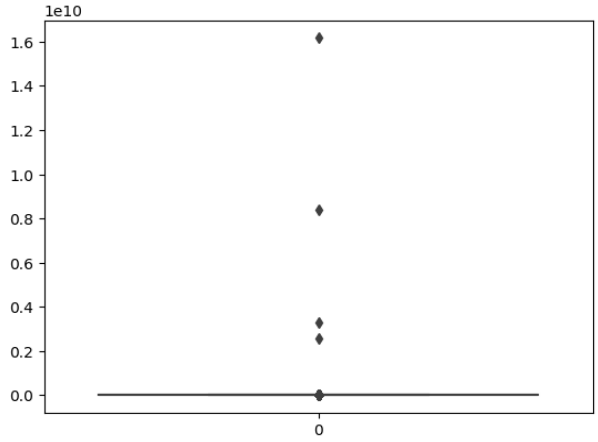

Fig. 1. The box-plot of the values 'Area'

III. VISUALIZATION

A. Outlook

To get an outlook on all the attributes after some preprocessing, we provided the Figure 2. The columns 'Area', 'Price', 'Latitude', and 'Longitude' include continuous values. The column 'Room' has 6 discrete values from 0 to 5. The three columns 'Parking', 'Warehouse', and 'Elevator' have Boolean values True and False.

B. The relationships

The scatter plot of the three variables 'Longitude', 'Latitude', and 'Price' is illustrated in Figure 3. The XY coordinate plane of Figure 3 can be interpreted as the surface of Tehran.

The scatterplot of the continuous variables is illustrated in Figure 4. A straightforward relationship is observed between the 'Area' and the 'Price'.

The heatmap of the correlation of all of the variables is drawn in Figure 5. The highest correlations respectively belong to 'Area' and 'Price', then 'Area' and 'Room', then 'Price' and 'Room', and then 'Parking' and 'Warehouse'.

REFERENCES

[1] M. Karimnejad, "House price (tehran, iran)," 2021. [Online]. Available: https://www.kaggle.com/datasets/mokar2001/house-price-tehran-iran?resource=download

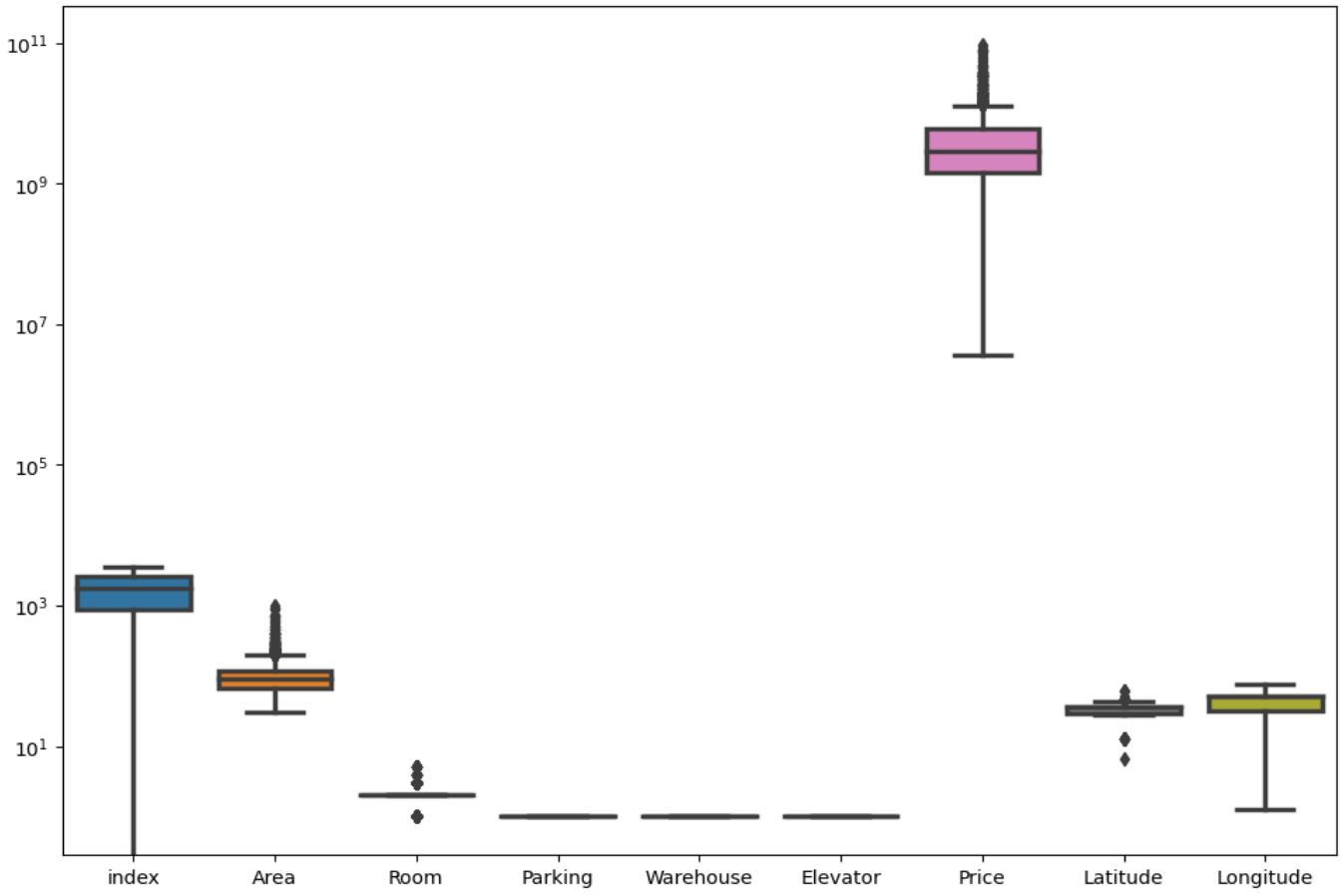

Fig. 2. The box-plot of all the values

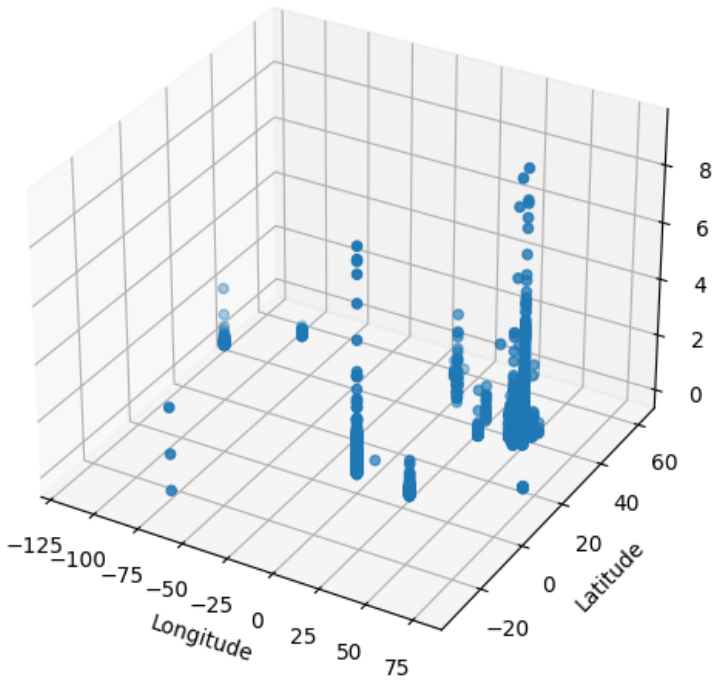

Fig. 3. The scatter plot of the three variables 'Longitude', 'Latitude', and 'Price'.

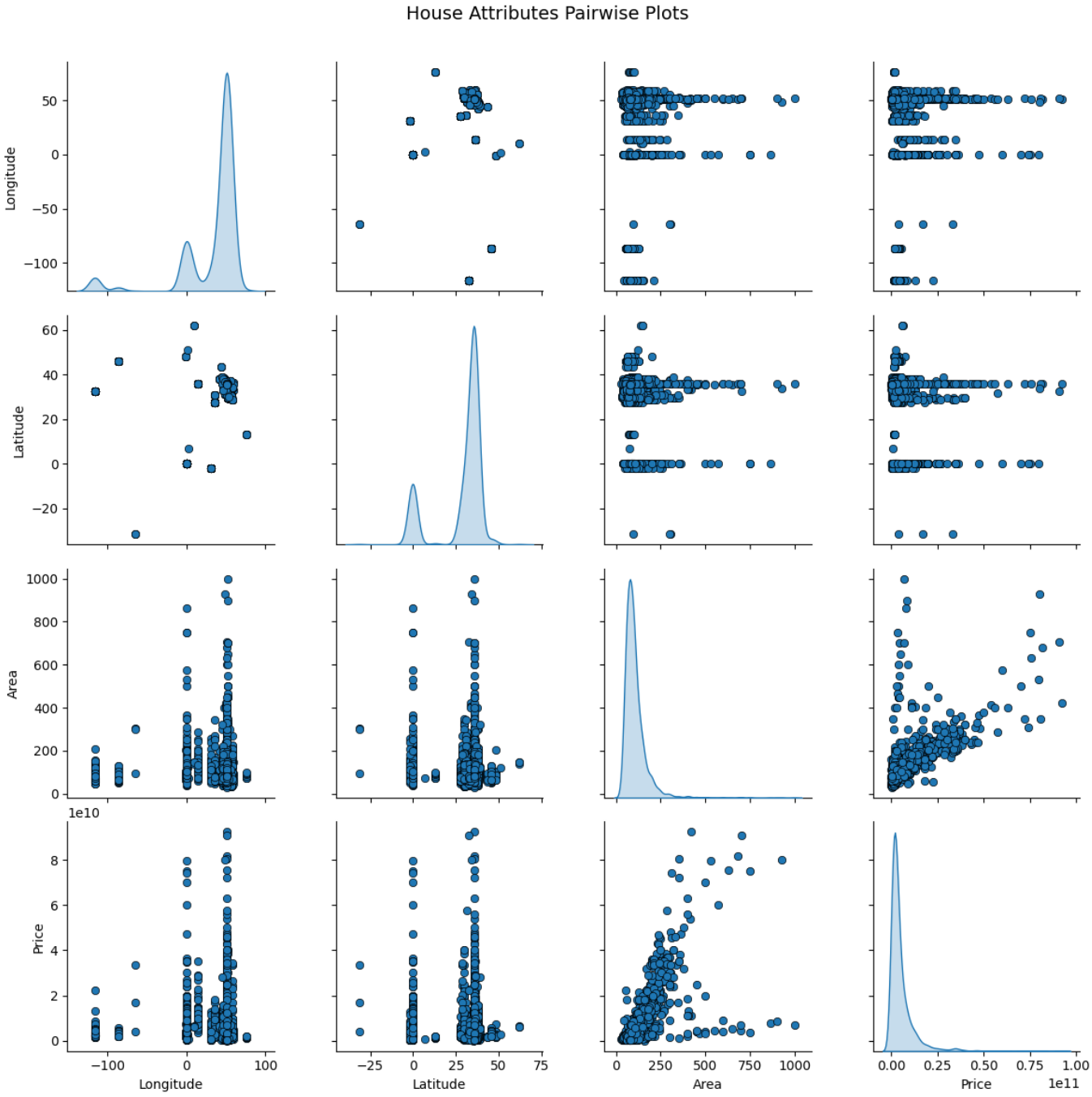

Fig. 4. The scatterplot of the variables 'Longitude', 'Latitude', 'Area', and 'Price'

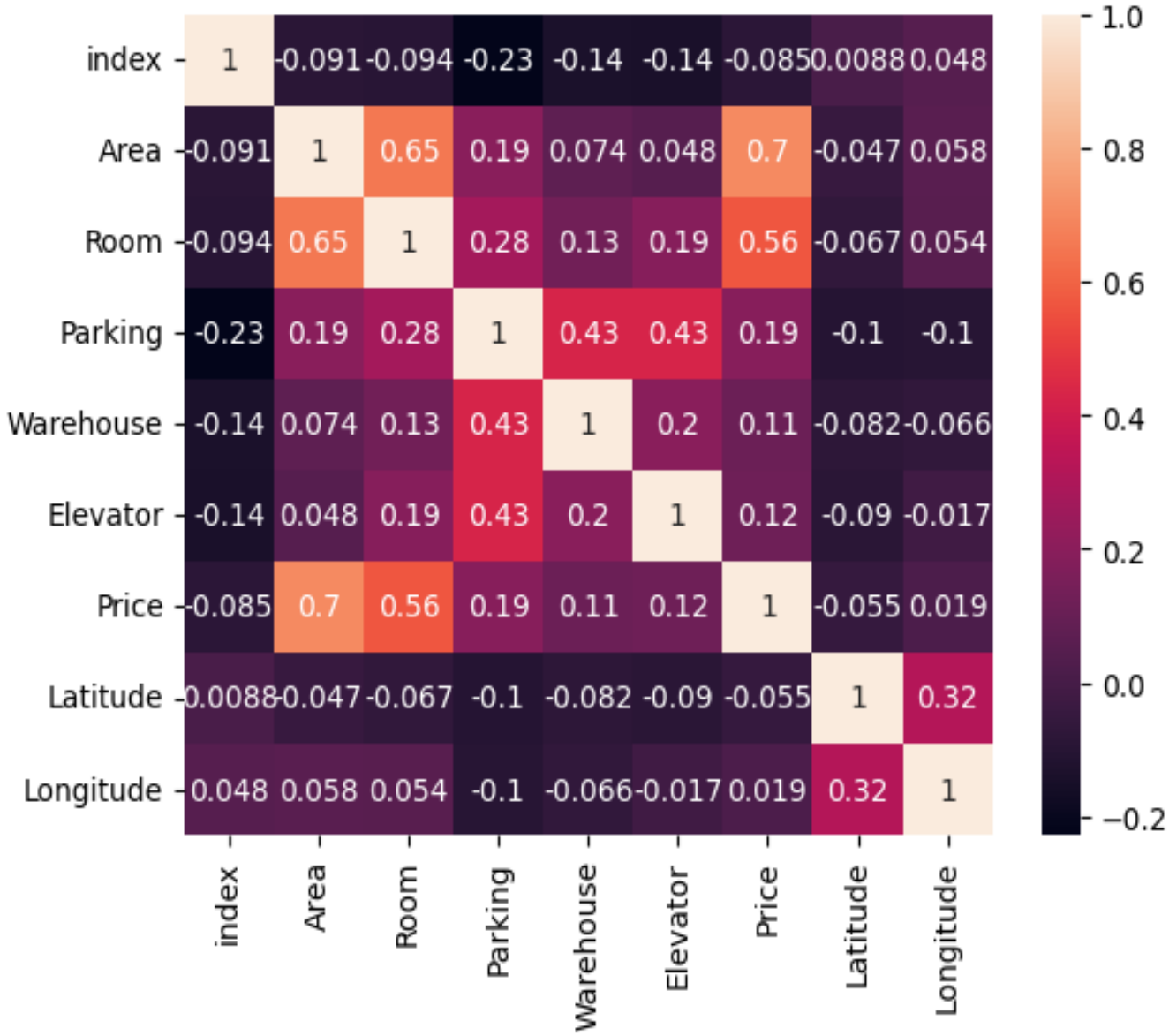

Fig. 5. The Correlation Heatmap

IV. Trained Models

A. Linear Regression

We set this as our baseline model. The result is as below.
Mean absolute error = 2853217587.57
MSE = 3.0358853027861082e+19
Median absolute error = 1657705882.45
Explain variance score = 0.5
R2 score = 0.5

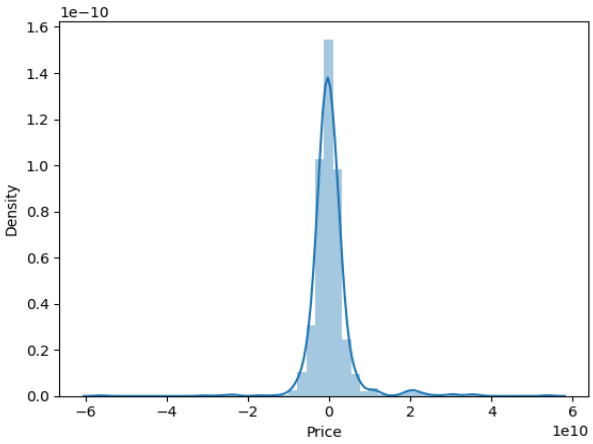

Fig. 6. Test labels and predictions of the Linear Regressor

B. Decision Tree Regressor

MAE: 1879711395.5457826
MSE: 2.8096122601455677e+19
RMSE: 5300577572.4401655

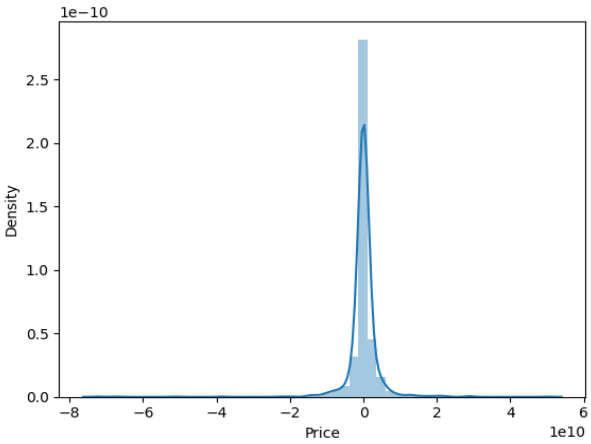

Fig. 7. Test labels and predictions of Decision Tree

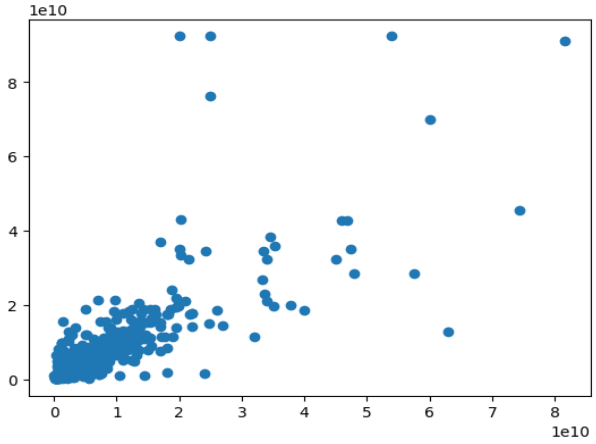

Fig. 8. The scatter plot of test labels and predictions of decision tree

Results from GridSearchCV:
MAE: 2495150056.1213694
MSE: 2.9038014698149446e+19
RMSE: 5388693227.318609

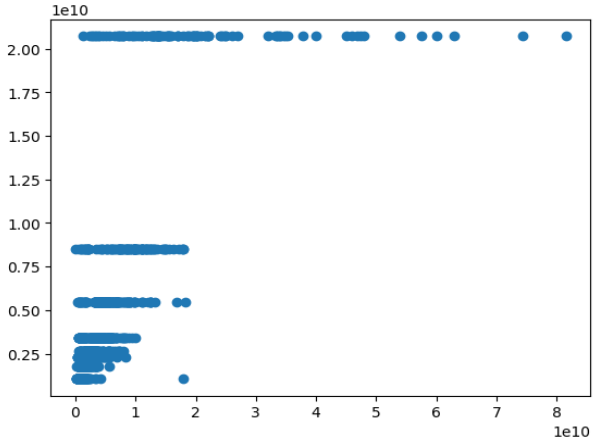

Fig. 9. Test labels and predictions of decision tree with tuned hyper-parameters

C. Gradient Boosting Regression

Results from Grid Search:
The best score across ALL searched parameters:
0.6945930819108665
The best parameters across ALL searched parameters:
$'learning\_rate'$: 0.02,$'$ max_depth' : 6,
$'n\_estimators'$: 100,$'$ $subsample'$: 0.5
Mean absolute error = 1847184299.21
MSE = 1.736066767490709e+19
Median absolute error = 807307243.54
Explain variance score = 0.71
R2 score = 0.71

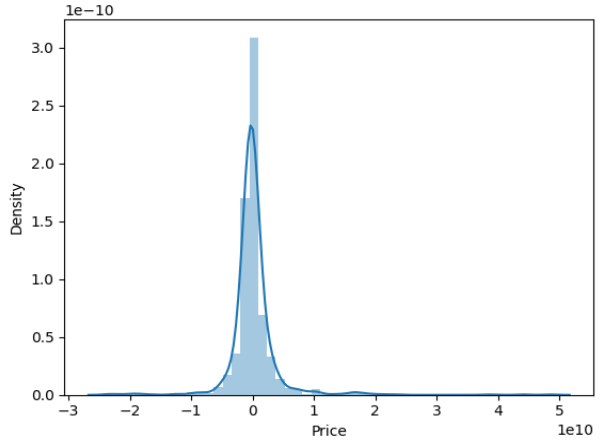

Fig. 10. Test labels and predictions of Gradient Boosting Regressor

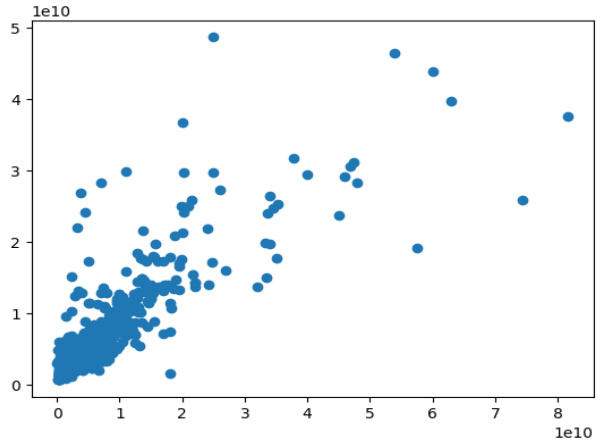

Fig. 11. The scatter plot of test labels and predictions of the Gradient Boosting Regressor

1) Deleting the columns 'Longitude' and 'Latitude' for tuned GBR: results in increasing MSE.
Mean absolute error = 2205304765.48
MSE = 1.86274525788012e+19
Median absolute error = 1104915912.45
Explain variance score = 0.69
R2 score = 0.69
2) Bining the 'Area' Attribute using qcut in the tuned GBR model: increases MSE.
Mean absolute error = 1792536405.99
MSE = 1.902683220577537e+19
Median absolute error = 639280755.95
Explain variance score = 0.68
R2 score = 0.68

D. Random Forest Regressor

MAE: 1590432063.4294627
MSE: 1.7045589846839994e+19
RMSE: 4128630505.0028386
1) Tuned with Randomized Search CV: $'n\_estimators'$: 600,$'min\_samples\_split'$: 10,$'min\_samples\_leaf'$: 1, $'max\_features'$:$'$ $sqrt'$,$'max\_depth'$: 110,$'bootstrap'$: $True$
MAE: 1538718937.5463989
MSE: 1.3838159626627074e+19
RMSE: 3719967691.610651

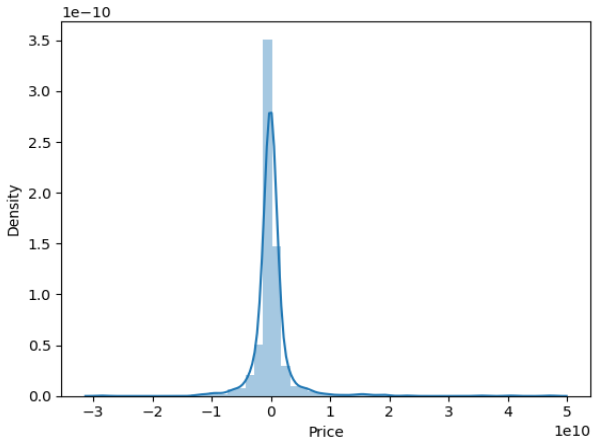

Fig. 12. The plot of test labels and predictions of tuned Random Forest

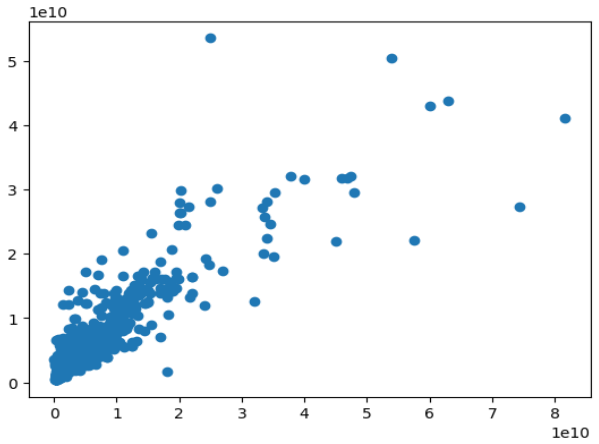

Fig. 13. The scatter plot of test labels and predictions of Tuned Random Forest

V. Conclusion

Random Forest performs best in our data. This can be due to the fact that the accommodations are somehow in clusters. Some data is predicted very far from their actual value in all models. We can consider them as noise data.

VI. Suggestions

- Use local predictions based on the attribute 'Address'.
- Detect the noise data to see how they affect the tuned models' hyper-parameters.
- Try hybrid models.