FLAT PRICE ESTIMATION

Name:	**GOURI H**
Registration No./Roll No.:	20121
Institute/University Name:	IISER Bhopal
Program/Stream:	ECONOMIC SCIENCES
Problem Release date:	August 17, 2023
Date of Submission:	Nov 19, 2023

1 Introduction

The aim of this project is to estimate the flat prices in various cities across India. This project involves building a model by taking into account various attributes related to flat prices to predict possible prices for the same. Flat price will, therefore, be the target variable here. The data set is split into test data and training data. The training and test data contain 26506 and 2947 instances, respectively. Using the given values, we have to predict the flat prices. Also, the data set has no missing values. There are 9 different classes in the data set. This is a regression kind of problem as the given values are continuous.

2 Methods

2.1 Splitting Of Training Data

The training data is split into training and test data in 80:20 ratio.

2.2 Regression Methods Used

This is a regression problem for which I have explored the existing regression methods which is described as follows. For each model, R2 Score, Mean Absolute Error(mae) and Mean Squared Error(mse) is calculated. On the basis of these values, the best model is further selected.

- Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Support Vector Regression

2.3 Hyper parameter Tuning

GridSearchCV is one of the most widely used and basic hyper parameter tuning technique in which all feasible permutations of the hyperparameters for a specific model are used. Grid search is performed on Linear Regression, Decision Tree Regression and Random Forest Regression. In support vector regression, grid search takes a high computational time. For linear regression, the R2 score was reduced from 0.158 to 0.0787. This is a relatively low R2 score, suggesting that the linear regression model, even after hyperparameter tuning through grid search, may not be capturing a substantial portion of the variability in the target variable.

Hyperparameters are: 'fit$_intercept'$

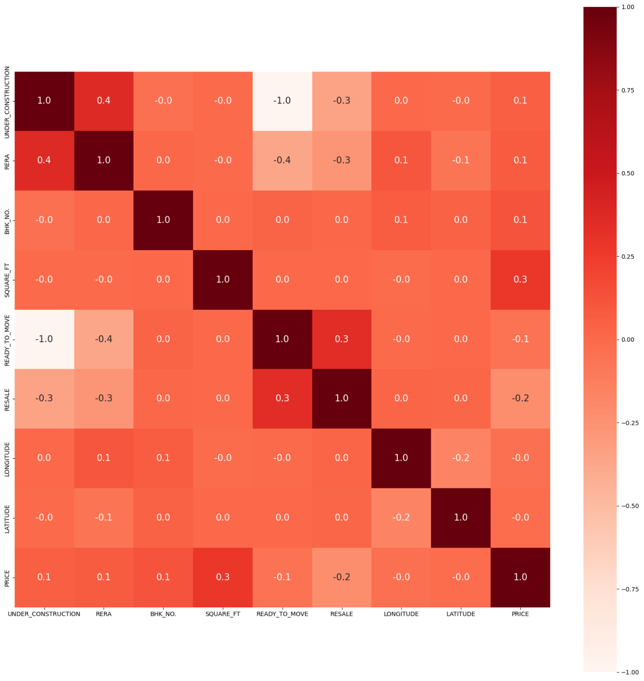

Figure 1: Heatmap

Decision tree, the accuracy was reduced from 0.93 to 0.92, upon performing grid search.

Hyperparameters are : $'\max_{d}epth' : None,' min_{s}amples_{l}eaf' : 6,' min_{s}amples_{s}plit' : 25$

Random Forest Regression gave the highest R2 score out of all, after the grid search. It gave an R2 score of 0.95

Hyperparameters are : $'\max_{d}epth' : None,' min_{s}amples_{s}plit' : 4,' n_{e}stimators' : 100$

2.4 Github Link

https://github.com/gourihari/FLAT-PRICE-ESTIMATION

3 Experimental Setup

The given datasets are analysed initially. To find the correlation between different attributes and flat pirices, a heat map is used (Figure 1). By analysing the heatmap, we can observe that the attribute square feet has more correlation with the price. Another (scatter) plot depicting the relation between price and square feet can be produced for better comprehension(Figure 2). Following this, we can include different plots between the variables(Figure 3). The baseline versions of all the models are trained on the dataset. It is then used to predict values for the test data splitted from the training dataset. Accuracy and errors are calculated using R2 score, mean absolute and mean squared errors. These values are then evauluated and analysed across all the four models to identify a model with least error and most accuracy. Table 1 gives the values of different models for the relevant parameters for evaluation.

GridSearchCV is also performed on the models, ideally on the models which gave promising results. With regard to SVR, the performance was very poor with a negative R2 score and considering the computational time for hyperparameter tuning, GridSearch was not performed for the same. The best hyper parameters were obtained upon repeated performance of gridsearch. These were then used for further tuning the model to obtain higher accuracy.

The values of R2 before and after hyperparameter tuning was analysed and it was observed that Random Forest Regression gave the highest R2 score and the least error, after GridSearch. Hence, the model was used to predict the prices for the test data. Although, decison tree also gave higher R2 score and lower errors, upon performing Gridsearch, the value of R2 reduced.

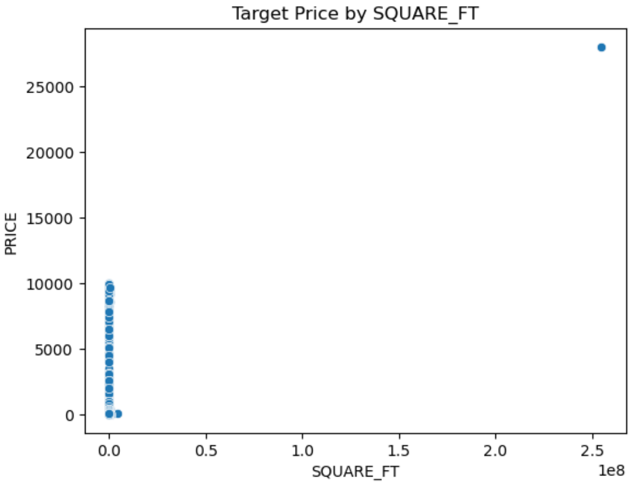

Figure 2: Scatterplot

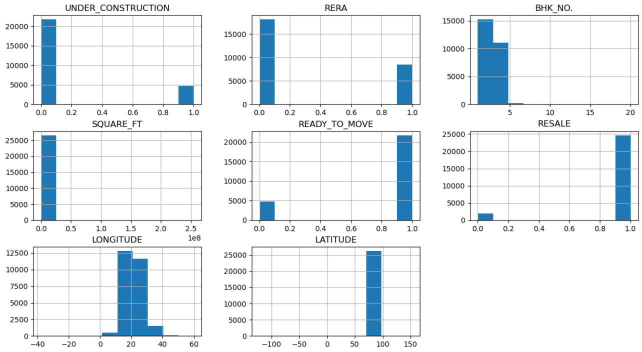

Figure 3: Histogram

Table 1: Performance Of Different Regression Methods

Regression Method	R2 Score	mae	mse
Linear Regression	0.078	146.36	419998.67
Decision Tree	0.948	38.2	23320.24
Random Forest	0.9536	33.9	21148.05
Support Vector Machine	-0.018	114.63	464126.53

Table 2: Performance Of Different Regression Methods After One Hot Encoding

Regression Method	R2 Score	mae
Linear Regression	1.0	11.2828227654088494e-12
Decision Tree	0.999884034806662	0.33605357479720827
Random Forest	0.9998584694966004	0.36675266930767775
Support Vector Machine	-0.018004624222385246	114.63371385257228

3.1 One Hot Encoding

One hot encoding was later performed separately, and a different result has been yielded. The results have been tabularized in Table 2.

Grid Search is then performed on Random Forest Regression and we obtain the
best hyper parameters:max$_{d}epth : Nonemin_{s}amples_{s}plit : 3n_{e}stimators : 100$
with a test score of 0.999839112552486, an increase from the previous R2 score. Similarly, One Hot Encoding was also performed on test data and some variables are dropped just like in the training data.

4 Results and Discussion

Upon using Random Forest Regression with the best hyper parameters for predicting test data variables, we can obtain a better and more accurate price for the corresponding flats given in the test data.

5 Conclusion

- Further work can be done on the same topic by analysing and effectively tuning different models. In this case, decision tree regression which gave a high R2 score can be further examined to make use of its predictions over the same test data. Although random forest regression was used for the prediction, it did face some limitations, such as large computational complexity etc.

- Feature engineering can be implemented which will help improve the results further, by increasing the efficiency.

- More tuning of parameters can be done and observed to identify the best among them. Also, including more parameters can also improve the efficiency and accuracy.

References

1. https://www.wikipedia.org/
2. https://www.kaggle.com/
3. https://github.com/
4. DSE317 Codes and notes in classroom