FLAT PRICE ESTIMATOR

Name:	KSHITIJ DURGE
Registration No./Roll No.:	21101
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	15 AUGUST, 2023
Date of Submission:	19 NOVEMBER, 2023

1 Introduction

The objective of this project is to predict the prices (in lakhs) of flats in various cities of India based on different factors. The factors as mentioned comprise of 9 features: 'Houses under construction', 'RERA statuts', 'BHK number', 'Square Feet', 'Ready to move', 'Resale', 'ADDRESS', 'Longitude' and 'Latitude'. Assessing the features, we find that the features 'BHK number' and 'Address' have categorical values in the data set. Categorical values in a data set cannot be used to train a machine learning model and hence, they need to be converted to numerical values for aiding the implementation of the model. Secondly, assessing the labels or the training data targets shows us that there are a number of outliers(labels with values that are inconsistent/ unusual as compared to the major portion of values) which happen to affect our results substantially. These major setbacks have been worked on through the course of this project, ultimately trying to get the best results possible through various regression models, while considering the numerous size of the data set and other time constraints and commitments.

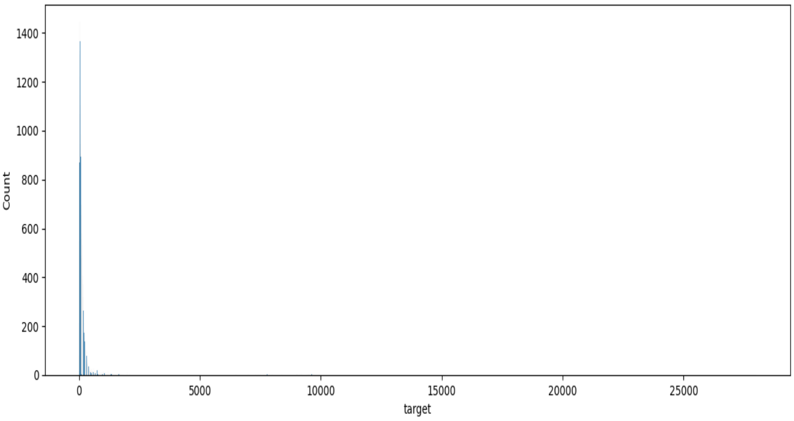

Figure 1: Overview of Data Set

2 Methods

- **Data Set Modification**: As cited in the Introduction section and Phase 1 report, the features 'ADDRESS' and 'BHK NO.' are assigned categorical values. To analyze the results for a raw dataset and test the relevance of both the features, I tried training the model by dropping both the features which lead to unsatisfactory results and an overall unstable model. Moving forward,

Table 1: Performance Of Different Classifiers Using All Features

Regressor	Mean Squared Error	Root Mean Squared Error	R2 Score
Adaptive Boosting	972941.114	986.3777	-1.5203
Decision Tree	357.2569	18.9012	0.376
Random Forest	517.113	22.7401	0.7058
Support Vector Machine	1816.174	42.6165	-0.0330
Linear Regression	821.2506	28.6574	0.5328

we need to evaluate the count of unique values in both the features which came out to be 6545 unique 'ADDRESS' values and unique 'BHK_NO.' values.

- 1) ENCODING: I implemented Ordinal Encoding on both the features which again did not return any fruitful results with RMSE values ranging from 200 to 900, accompanied by a decent R2 score. Moving on to One-Hot Encoding, every instance of the categorical features is converted to an individual feature and assigned binary values. One-Hot Encoding is preferred over Ordinal Encoding since it guarantees an appropriate representation of the data.

2) OUTLIER REMOVAL: We can infer from the plot of the data set, there are several label instances with values severely out of range of the most recurring values which eventually results in the incorrect learning of the specified parameters. To overcome this issue, we merge the training data and its labels and set a threshold value for the labels which will eventually cancel out the label values beyond the specified threshold value and only consider the values within the limit. Once the threshold is set, we tend to separate the training data and the target variables into two new data sets, from which the outliers are removed. 3) PIPELINE and GRID MODIFICATION: The initial models for feature selection and parameter scoring favoured classification models. To favour the implementation of a regression model, F Regression for feature selection and R2 score for parameter scoring were imported and implemented in the code.

- The data set contains features like 'RERA', 'Resale', 'Ready To Move' which have been pre-assigned binary values. We were instructed to use One-HOT Encoding on these features but encoding binary values does not seem to have any effect on the output of the model. Hence, One-Hot encoding was avoided on the specified features.

- Github Link to the python code for this model: https://github.com/kshiiitij/Flat-Price-Estimation.git

3 Experimental Setup

Using the cited methods, we trained the model with several regressors and their best parameters being 1. Linear Regresssor 2. Decision Tree Regressor: [rgrcriterion= 'squared error', rgrmaxfeatures= 'sqrt', rgrmaxdepth=60] 3. Random Forest Regressor: [rgrnestimators=500, rgrmaxdepth=300, rgrmaxfeatures='sqrt'] 4. Ridge Regressor: [rgrsolver='lbfgs'] 5. Supprt Vector Machine Regressor:[] rgrC=1, rgrkernel= 'linear'] 6.Lasso Regressor 7.AdaBoost Regressor

- Out of the seven Regressors, Linear Regressor and Random Forest Regressor proved to lay the best results.

- A notable observation being, before the Outliers were Removed the model laid a decent R2 score(close to 1) but the RMSE value was off the charts(ranging from 300 to almost a 1000). After the Outliers were Removed, the model saw a steep fall in the RMSE value(ranging from 15 to 40) but the R2 Score degraded to almost a half(0.5) for most Regressors.

4 Results and Discussion

Table 1 displays the results obtained by training the Regression model. Random Forest Regressor emerged as the top-performing algorithm, showcasing the highest scores across multiple metrics. Linear Regressor came in at second place with a very precise RMSE value but the R2 Score was not as intriguing. We can see similar happenings in the rest of the Regressors. The value of RMSE being accurate and R2 Score being far from the mark could mainly be because of Outlier Removal from the data set, as stated above.

5 Conclusion

We can conclude by saying that Random Forest Regressor proves to be the most effective and reliable model for training and implementation. We can observe the outcomes of unbalanced data points through the RMSE values and R2 Scores. Adding to this, we can infer that Outlier Removal is a crucial step in obtaining the observed contrasting result metrics. Parameter and Feature Selection is another crucial step in obtaining optimum results and achieving better precision through the model.

References

1. House price models Spatial analysis Relative location Space syntax Hedonic models by Axel Viktor Heymana , Dag Einar Sommervoll

2.Prediction of residential real estate selling prices using neural networks by PONTUS NILSSON .

3. Fellow Batchmate- Ankur Kumar(21043)