Flat Price Estimation

Name: Parmar Pranali Ben
Registration No./Roll No.: 21361
Institute/University Name: IISER Bhopal
Program/Stream: ECS 308
Problem Release date: August 17, 2023
Date of Submission: November 19, 2023

Introduction:

A. Problem Statement:
Thousands of houses are sold everyday. There are some questions every buyer asks himself/herself like: What is the actual price that this house deserves? Am I paying a fair price?

B. Objective:
The objective of this project is to predict the prices (in lakhs) of flats in various cities of India based on different factors.

C. Best Possible Solutions:
a.Housing Expert
b.Intuition About House
c.Using Machine Learning

D. Introduction About Project:

House Price prediction are very stressful work as we have to consider different things while buying a house like the structure and the rooms kitchen parking space and gardens. People don't know about the factor what are the things which influence the house price. But by using Machine learning we can easily find the house which is to be perfect for us and helps to predict the price accurately.

E. Aim:
a. To identify relevant features of the given training data to estimate the flat prices (in lakh)(s in India.
b. To design novel feature selection framework or use the existing techniques with proper justifications to identify salient features of the given data.
c. To report the performance of different regression techniques on the training data to demonstrate that the proposed feature selection scheme is working well. Subsequently, execute the best framework on the test data and submit the estimated flat price in a text file.

F. Data Collection:
For this project we used the data that is available on google drive given for project. (https://drive.google.com/drive/folders/1G8KQYowaIYMcZ7KsEEpiubCbED5X72LU?usp=drive_linkclick here for data).There are 9 columns and 26505 Rows(of training data set) to estimate the price of house of test data(9 columns and 2947 rows). These are the major point about the data set.

UNDER-CONSTRUCTION
AREA
BHK-NO.
SQUARE-FT
READY TO MOVE
RESALE
ADDRESS
LONGITUDE
LATITUDE

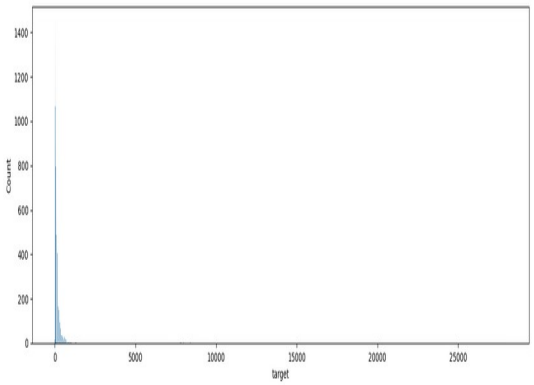

Figure 1: Dataset with outliers

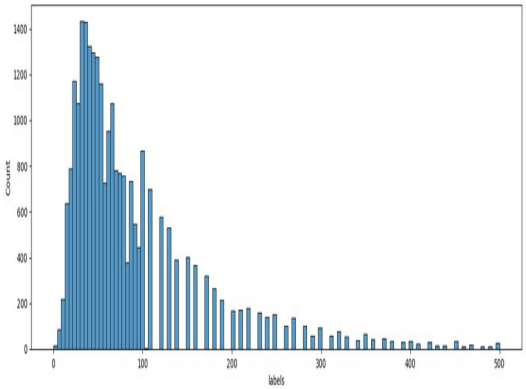

Figure 2: Dataset without outliers

Table 1: Performance Of Different Classifiers Using All Features

Classifier	MSE (with outliers)	MSE(without outliers)	RMSE(with outliers)	RMSE(without outliers)	R2-Score(with outliers)	R2-Score(without outliers)
Linear Reression	1717455.550166193	7455.550166193	1310.5172834290256	310.5172834290256	-2.5005022906813372	-2.5005022906813372
Ridge	489981.1977357764	981.1977357764	699.9865696824307	99.9865696824307	0.00132477670301400 79	0.0013247767030140 079
Lasso	451757.8688557755	757.8688557755	672.1293542583716	72.1293542583716	0.07923121817625156	0.0792312181762515 6
Decision Tree Regressor	20666.396280830086	66.396280830086	143.7581172693566	33.7581172693566	77.878779389579115	77.957877938957911 5
Random Forest Regressor	3517.1130188899009	517.1130188899009	22.740119148542316	22.7401191485423 16	70.7058536339918271	77.705853633991827 1
Support Vector Machine	20666.396280830086	66.396280830086	143.7581172693566	33.7581172693566	56.878779389579115	35.957877938957911 5
AdaBoost method	357.1130188899009	57.1130188899009	265.74011914854231 6	22.7401191485423 16	79.7058536339918271	45.705853633991827 1

Methods:

A .Data Cleaning:

We will perform data cleaning which involve following steps.
Our target variable is total-Price(in lakh)(s).

a. Impute/Remove missing values or Null values(NaN)(For bigger data we can do this but luckily we have no missing values.)
b. Remove unnecessary and corrupted data.(We do not hav any unnecessary data)
c. We remove most of the outliers to have more accuracy.

B. Feature Engineering:

We have done One-hot encoding for all necessary categorical values(ADDRESS). We have also done Ordinal coding which didn't gave the desired results.

C. Data Normalization:
Normalization (min-max Normalization) In this approach we scale down the feature in between 0 to 1 We have numerical column where we can apply min-max Normalization.

D. Choosing Best ML Model:
List of the model that we can use for our problem
a. Linear Regression model
b. Ridge model
c. Lasso model
d. Decision Tree model
e. Random Forest model
f. SVM(Support Vector Model)

Experimental Setup:
Githut code link:

Results and Discussion:

Random Forest model is giving best results by giving :-

```
RandomForestRegressor(max_depth=300,max_features='sqrt,n_estimators=300))
])
```
MSE: `517.1130188899009`

RMSE: `22.740119148542316`

R2-Score: `0.7058536339918271`

Conclusion:
We have concluded that there are more than one model(Linear Regression model, Ridge model, Lasso model, Decision Tree model, Random Forest model, SVM(Support Vector Model)) by which we can estimate flat price but by observing results I concluded that the best model will be Random Forest regressor model.

References:
1. Codes from Tanmay Basu sir