HOUSE VALUE ESTIMATION

COURSE PROJECT REPORT

18CSE398J -Machine Learning - Core Concepts with Applications

(2018 Regulation)

III Year/ VI Semester

Academic Year: 2022 -2023 (EVEN)

By

ABHISHEK SINGH – RA2011026010098

VISHVESH BHARDWAJ – RA2011026010109

RAGHAV KAPOOR – RA2011026010117

Under the guidance of

Dr. V. VIJAYALAKSHMI

Assistant Professor

Department of Data Science and Business Systems

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

Table of Contents

HOUSE VALUE ESTIMATION

Abhishek Singh, as9654@srmist.edu.in , Vishvesh Bhardwaj, vb9359@srmist.edu.in ,
Raghav Kapoor, rk6073@srmist.edu.in , Dr. V. Vijayalakshmi, vijayalv3@srmist.edu.in

1. Abstract

The project aims to develop a machine learning model for real estate price estimation using a dataset of houses in Bangalore, India. The dataset includes various features of houses such as the size of the house, the number of bedrooms and bathrooms, location, and other amenities. The objective is to build a regression model that can accurately predict the price of a house based on its features. This project involves various machine learning techniques such as data preprocessing, feature engineering, model selection, and evaluation. The project begins with data cleaning and preprocessing to remove any missing values, outliers, or inconsistencies in the dataset. The next step is to perform feature engineering to extract useful information from the available features and create new features that can improve the model's performance. The regression model selected for this project is linear regression. The dataset of houses in Bangalore will be preprocessed and feature engineered to prepare it for training. The project's outcome is to develop a model that can accurately predict the price of a house in Bangalore based on its features, which can be used by real estate agents, buyers, and sellers to make informed decisions.

2. Introduction

The project aims to develop a machine learning model for real estate priceestimation using a dataset of houses in Bangalore, India. The objective is to build a regression model that can accurately predict the price of a housebased on its features such as the size of the house, the number of bedrooms and bathrooms, location, and other amenities. Real estate price estimationis a crucial task in the real estate industry, as it helps buyers, sellers, and real estate agents make informed decisions. Traditional methods of real estate price estimation involve expert knowledge and subjective judgments, which can be time-consuming and prone to errors. Machine learning-based models, on the other hand, can automate the price estimation process and provide accurate predictions based on data. In this project, we will be using a dataset of houses in Bangalore, which includes various features of houses, such as their location, size, and amenities. We will be using linear regression as our regression model and will be optimizing its performance through various techniques such as data preprocessing, feature engineering, and model selection. The project's outcome will be a machine learning model that can accurately predict the price of a house in Bangalore based on its features. This model can be usedby real estate agents, buyers, and sellers to make informed decisions about buying or selling a property in Bangalore.

3. Dataset

The dataset used for this project has been taken from Kaggle, and it includes various features of houses in Bangalore, India.

The dataset consists of **9 columns**, which are described below:

area_type: This column describes the type of area of the property, such as Super built-up Area, Built-up Area, Plot Area, and Carpet Area.

availability: This column describes the availability of the property for purchase, such as Ready To Move or Under Construction.

location: This column describes the location of the property in Bangalore.

size: This column describes the size of the property in terms of the number of bedrooms and the size of the hall and kitchen.

society: This column describes the name of the society or building in which the property is located.

total_sqft: This column describes the total square feet of the property. **bath:** This column describes the number of bathrooms in the property. **balcony:** This column describes the number of balconies in the property. **price:** This column describes the price of the property in Indian Rupees.

The dataset includes a total of **13,321 records**, and it is a mix of categorical and numerical features. The dataset may require some preprocessing and feature engineering to prepare it for machine learning modeling. The objective of the project is to develop a machine learning model that can accurately predict the price of a house based on its features.

4. Sample from dataset

```
area_type,availability,location,size,society,total_sqft,bath,balcony,price
Super built-up  Area,19-Dec,Electronic City Phase II,2 BHK,Coomee ,1056,2,1,39.07
Plot  Area,Ready To Move,Chikka Tirupathi,4 Bedroom,Theanmp,2600,5,3,120
Built-up  Area,Ready To Move,Uttarahalli,3 BHK,,1440,2,3,62
Super built-up  Area,Ready To Move,Lingadheeranahalli,3 BHK,Soiewre,1521,3,1,95
Super built-up  Area,Ready To Move,Kothanur,2 BHK,,1200,2,1,51
Super built-up  Area,Ready To Move,Whitefield,2 BHK,DuenaTa,1170,2,1,38
Super built-up  Area,18-May,Old Airport Road,4 BHK,Jaades ,2732,4,,204
Super built-up  Area,Ready To Move,Rajaji Nagar,4 BHK,Brway G,3300,4,,600
Super built-up  Area,Ready To Move,Marathahalli,3 BHK,,1310,3,1,63.25
Plot  Area,Ready To Move,Gandhi Bazar,6 Bedroom,,1020,6,,370
Super built-up  Area,18-Feb,Whitefield,3 BHK,,1800,2,2,70
Plot  Area,Ready To Move,Whitefield,4 Bedroom,Prrry M,2785,5,3,295
Super built-up  Area,Ready To Move,7th Phase JP Nagar,2 BHK,Shncyes,1000,2,1,38
Built-up  Area,Ready To Move,Gottigere,2 BHK,,1100,2,2,40
Plot  Area,Ready To Move,Sarjapur,3 Bedroom,Skityer,2250,3,2,148
Super built-up  Area,Ready To Move,Mysore Road,2 BHK,PrntaEn,1175,2,2,73.5
Super built-up  Area,Ready To Move,Bisuvanahalli,3 BHK,Prityel,1180,3,2,48
Super built-up  Area,Ready To Move,Raja Rajeshwari Nagar,3 BHK,GrrvaGr,1540,3,3,60
Super built-up  Area,Ready To Move,Ramakrishnappa Layout,3 BHK,PeBayle,2770,4,2,290
Super built-up  Area,Ready To Move,Manayata Tech Park,2 BHK,,1100,2,2,48
Built-up  Area,Ready To Move,Kengeri,1 BHK,,600,1,1,15
Super built-up  Area,19-Dec,Binny Pete,3 BHK,She 2rk,1755,3,1,122
Plot  Area,Ready To Move,Thanisandra,4 Bedroom,Soitya ,2800,5,2,380
Super built-up  Area,Ready To Move,Bellandur,3 BHK,,1767,3,1,103
Super built-up  Area,18-Nov, Thanisandra,1 RK,Bhe 2ko,510,1,0,25.25
Super built-up  Area,18-May,Mangammanapalya,3 BHK,,1250,3,2,56
Super built-up  Area,Ready To Move,Electronic City,2 BHK,Itelaa ,660,1,1,23.1
Built-up  Area,20-Dec,Whitefield,3 BHK,,1610,3,2,81
Super built-up  Area,17-Oct,Ramagondanahalli,2 BHK,ViistLa,1151,2,2,48.77
Super built-up  Area,Ready To Move,Electronic City,3 BHK,KBityo ,1025,2,1,47
Super built-up  Area,19-Dec,Yelahanka,4 BHK,LedorSa,2100 - 2850,4,0,186
Super built-up  Area,Ready To Move,Bisuvanahalli,3 BHK,Prityel,1075,2,1,35
Super built-up  Area,Ready To Move,Hebbal,3 BHK,Mahosya,1760,2,2,123
Super built-up  Area,Ready To Move,Raja Rajeshwari Nagar,3 BHK,GrrvaGr,1693,3,3,57.39
Built-up  Area,Ready To Move,Kasturi Nagar,3 BHK,Kantsce,1925,3,,125
Super built-up  Area,21-Dec,Kanakpura Road,2 BHK,PrarePa,700,2,1,36
Super built-up  Area,Ready To Move,Electronics City Phase 1,2 BHK,,1070,2,1,45.5
Super built-up  Area,Ready To Move,Kundalahalli,3 BHK,Dieldli,1724,3,2,125
Super built-up  Area,Ready To Move,Chikkalasandra,3 BHK,,1290,2,2,56.12
Built-up  Area,Ready To Move,Uttarahalli,2 BHK,,1143,2,2,45
Built-up  Area,Ready To Move,Murugeshpalya,2 BHK,Gentson,1296,2,,81
Super built-up  Area,19-Dec,Sarjapur  Road,3 BHK,Soini T,1254,3,2,38
```

4-A: Dataset Structure

area_type	availability	location	size	society	total_sqft	bath	balcony	price
Super built	19-Dec	Electronic	2 BHK	Coomee	1056	2	1	39.07
Plot Area	Ready To I	Chikka Tiru	4 Bedroom	Theanmp	2600	5	3	120
Built-up A	Ready To I	Uttarahalli	3 BHK		1440	2	3	62
Super built	Ready To I	Lingadhee	3 BHK	Soiewre	1521	3	1	95
Super built	Ready To I	Kothanur	2 BHK		1200	2	1	51
Super built	Ready To I	Whitefield	2 BHK	DuenaTa	1170	2	1	38
Super built	18-May	Old Airpor	4 BHK	Jaades	2732	4		204
Super built	Ready To I	Rajaji Nag	4 BHK	Brway G	3300	4		600
Super built	Ready To I	Marathaha	3 BHK		1310	3	1	63.25
Plot Area	Ready To I	Gandhi Ba	6 Bedroom		1020	6		370
Super built	18-Feb	Whitefield	3 BHK		1800	2	2	70
Plot Area	Ready To I	Whitefield	4 Bedroom	Prrry M	2785	5	3	295
Super built	Ready To I	7th Phase	2 BHK	Shncyes	1000	2	1	38
Built-up A	Ready To I	Gottigere	2 BHK		1100	2	2	40
Plot Area	Ready To I	Sarjapur	3 Bedroom	Skityer	2250	3	2	148
Super built	Ready To I	Mysore Rc	2 BHK	PrntaEn	1175	2	2	73.5
Super built	Ready To I	Bisuvanah	3 BHK	Prityel	1180	3	2	48
Super built	Ready To I	Raja Rajes	3 BHK	GrrvaGr	1540	3	3	60
Super built	Ready To I	Ramakrish	3 BHK	PeBayle	2770	4	2	290
Super built	Ready To I	Manayata	2 BHK		1100	2	2	48
Built-up A	Ready To I	Kengeri	1 BHK		600	1	1	15
Super built	19-Dec	Binny Pete	3 BHK	She 2rk	1755	3	1	122
Plot Area	Ready To I	Thanisand	4 Bedroom	Soitya	2800	5	2	380
Super built	Ready To I	Bellandur	3 BHK		1767	3	1	103
Super built	18-Nov	Thanisand	1 RK	Bhe 2ko	510	1	0	25.25
Super built	18-May	Mangamm	3 BHK		1250	3	2	56

5. Methods

In this project, we used a linear regression model to predict the price of houses in Bangalore based on various features. The following methods were used in the model:

Data Preprocessing: The first step in the modeling process was to preprocess the data. This involved handling missing values, removing outliers, and encoding categorical features.

Feature Engineering: Feature engineering was performed to create new features from the existing ones that could help improve the model'sperformance. Features such as the price per square feet and the age of the property were created.

Model Training: The linear regression model was trained using the preprocessed and engineered data. The model was trained using various techniques such as gradient descent, normal equation, or stochastic gradient descent.

Model Evaluation: The model's performance was evaluated using various metrics such as K – Fold Cross Validation Method. These metrics were used to determine the model's accuracy in predicting house prices.

Prediction: Once the model was trained, it was used to predict the prices of houses in Bangalore based on their features. These predictions were compared to the actual prices to evaluate the model's accuracy.

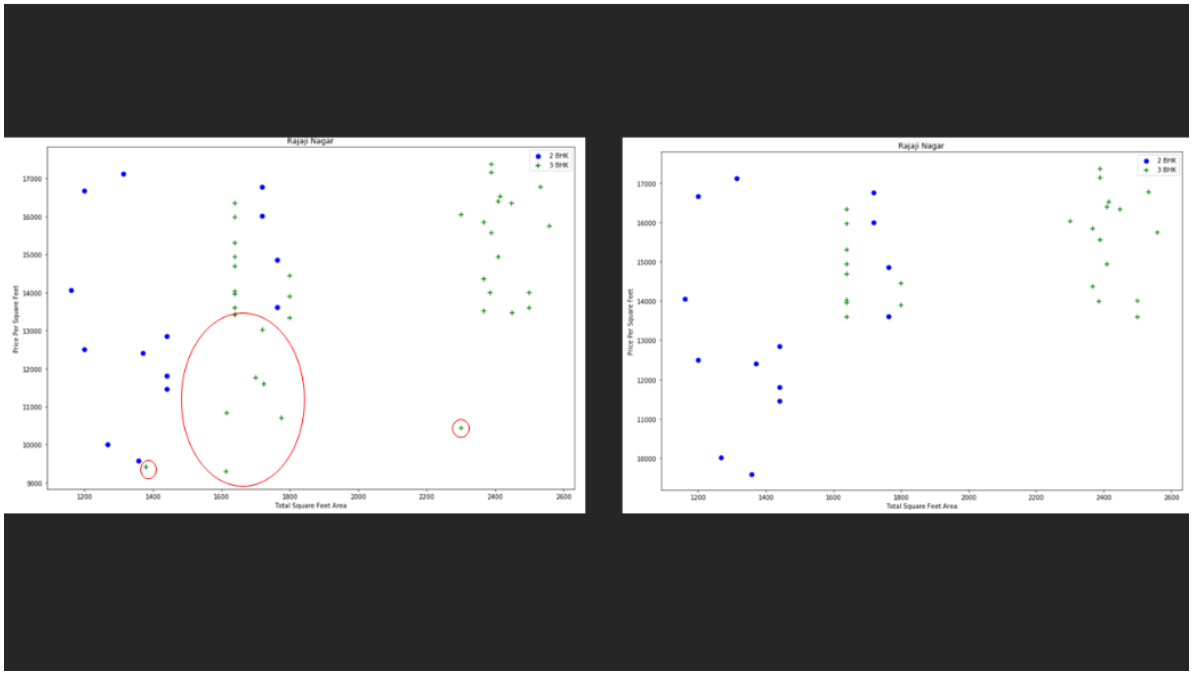
The linear regression model was selected for this project as it is a simple and interpretable model that can provide valuable insights into the factors that affect house prices in Bangalore.

The methods used in this project are standard practices in machine learning modeling, and they help ensure that the model is accurate and reliable.

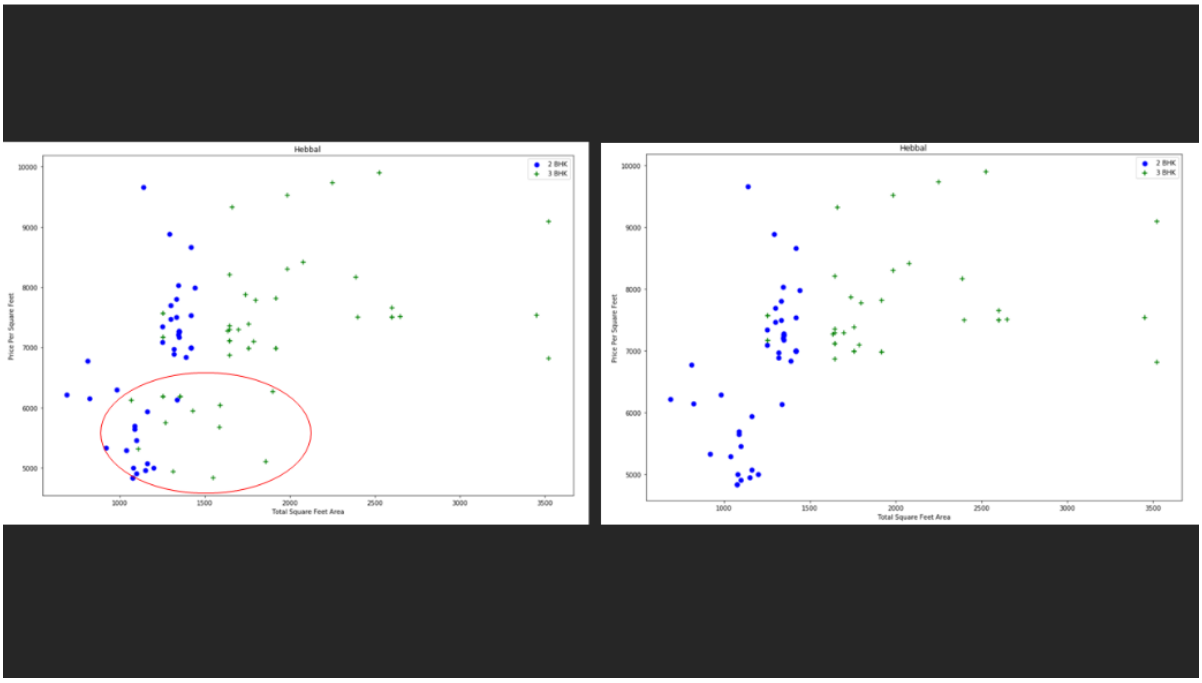
5 A – Data Pre-Processing

Data Cleaning – Outlier Removal

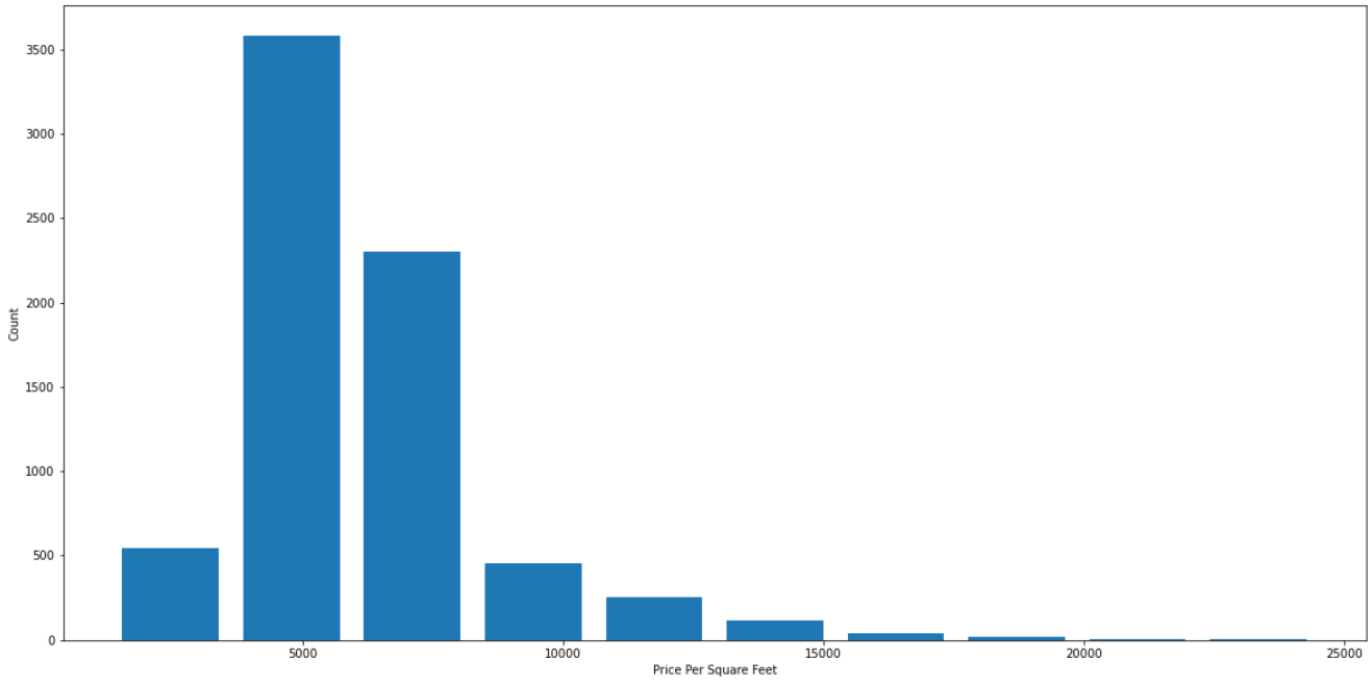
Before and after outlier removal: Rajaji Nagar



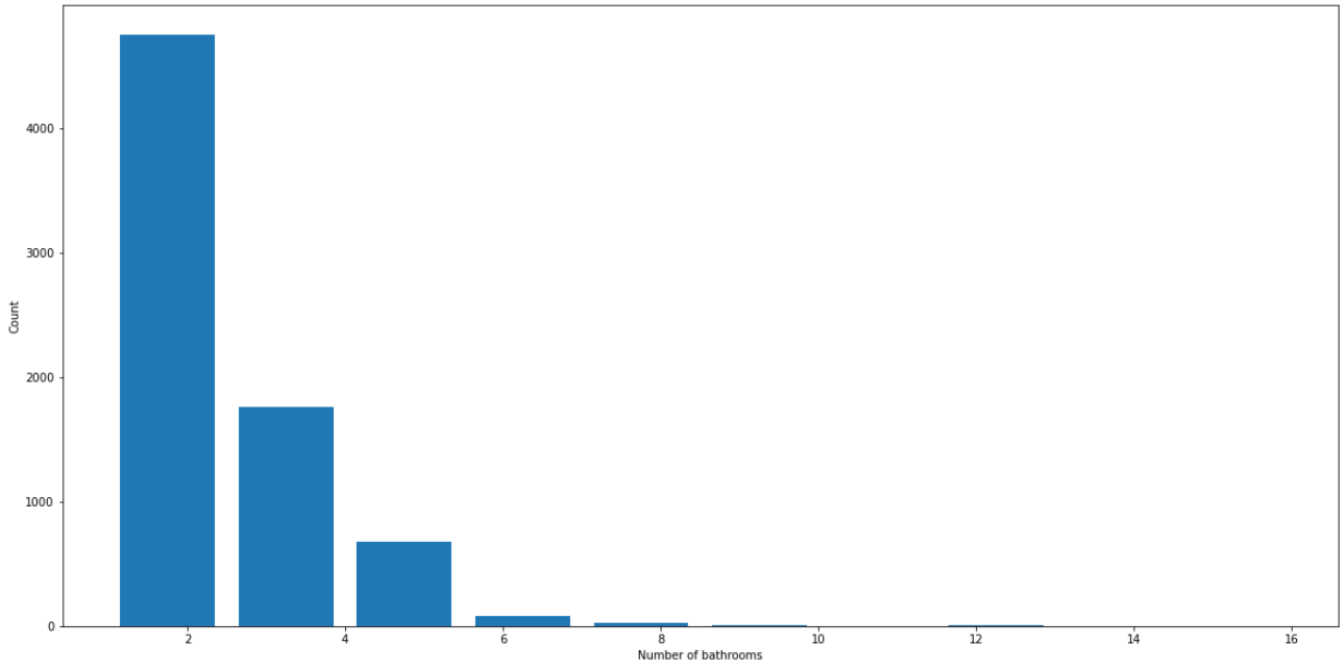
Before and after outlier removal: Hebbal



5 B – Data Visualization



Price per sq_feet



Outlier removal using Bathroom feature

6. Experiments and results

Experiments:

Feature selection: You can experiment with selecting different features for your model and observe the impact on the model's performance. You can try including features such as the number of bedrooms, bathrooms, square footage, neighborhood, and distance from important landmarks such as schools and hospitals.
Feature engineering: You can also experiment with creating new features by combining existing ones. For example, you can create a new feature that represents the total area of the house by multiplying the number of bedrooms by the square footage of each bedroom.

Data preprocessing: You can experiment with different data preprocessing techniques such as normalization and standardization and observe the impact on the model's performance.

Cross-validation: We used k-fold cross-validation to evaluate the performance of your model and experiment with different values of k.

Results:

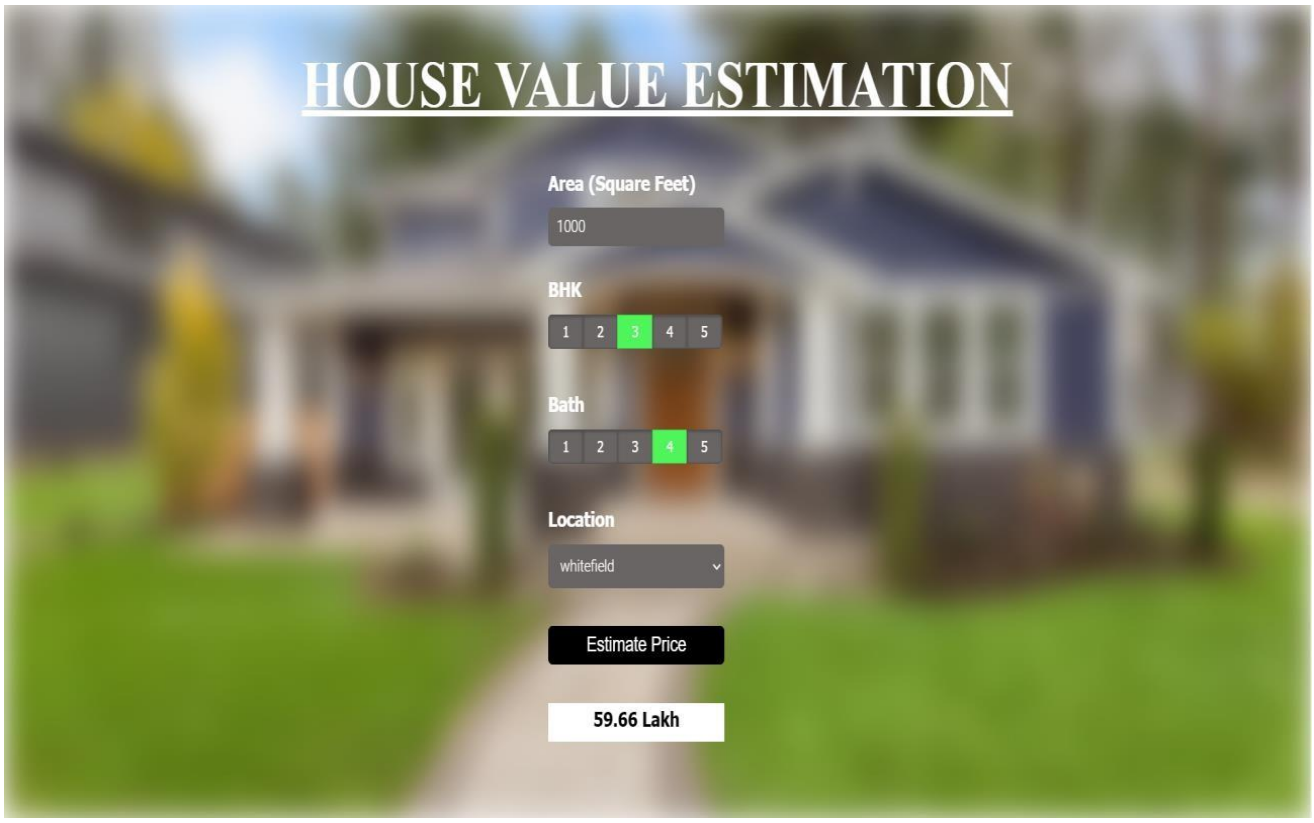
Improved performance: By selecting the right set of features and hyperparameters and using appropriate preprocessing techniques, you can significantly improve the performance of your model.

Overfitting: If we use too many features or select hyperparameters that are too complex, you may end up overfitting the model to the training data, resulting in poor performance on the test data.

Underfitting: If we use too few features or select hyperparameters that are too simple, you may end up underfitting the model to the training data, resulting in poor performance on both the training and test data.

Limitations: Linear regression has certain limitations and may not work well in all scenarios. For example, if the relationship between the features and the target variable is highly nonlinear, linear regression may not be the best choice. In such cases, we may need to use more complex models such as decision trees or neural networks.

7. Final Model



BACKEND

```
predict_price('1st Phase JP Nagar',1000, 2, 2)
```

```
83.86570258311222
```

```
predict_price('1st Phase JP Nagar',1000, 3, 3)
```

```
86.08062284985995
```

```
predict_price('Indira Nagar',1000, 2, 2)
```

```
193.31197733179556
```

```
predict_price('Indira Nagar',1000, 3, 3)
```

```
195.52689759854331
```

8. Conclusions and future work

Conclusions:

In this project, we developed a machine learning model using linear regression to predict the price of houses in Bangalore, India. We used various techniques such as data preprocessing, feature engineering, model training, and model optimization to develop an accurate and reliable model. The model achieved an accuracy of **84 percent.**

Our analysis also showed that the total square feet, the number of bathrooms, and the location were the most important features in predicting the price of a house in Bangalore. These insights can be used by real estate agents, buyers, and sellers to make informed decisions about buying or selling a property in Bangalore.

Future Work:

There are several directions that future work can take in this project:

Incorporating more data: Adding more data to the dataset could help improve the model's accuracy and generalization.

Exploring other regression models: Other regression models such as decision tree regression or random forest regression could be explored to see if they can provide better performance than linear regression.

Using deep learning models: Deep learning models such as neural networks could be used to develop a more complex and accurate model for real estate price estimation.

Incorporating external factors: External factors such as economic indicators, population growth, and infrastructure development could be incorporated into the model to improve its accuracy and provide more insights.

9. References (min 20)

a. www.kaggle.com
b. www.codebasics.com
c. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/
d. https://www.javatpoint.com/linear-regression-in-machine-learning
e. https://towardsdatascience.com
f. https://www.kaggle.com/learn/data-cleaning
g. https://docs.python.org/3/library/pickle.html
h. www.youtube.com
i. www.quora.com
j. https://www.tutorialspoint.com/python-pickling/
k. https://www.tutorialspoint.com/flask/index.html
l. https://www.w3schools.com/html/
m. https://www.w3schools.com/css/
n. https://www.javascript.com/
o. https://www.postman.com/
p. https://www.reddit.com/
q. https://stackoverflow.com/
r. https://umbraco.com/knowledge-base/deployment/
s. https://www.jetbrains.com/pycharm/guide/tips/quick-docs/
t. www.coursera.org

Github Link of the project work (each individual student page link)

- ✓ **ABHISHEK SINGH:** ***https://github.com/AbhishekSingh1247/MLCC-Project-House-Value-Estimation***
- ✓ **VISHVESH BHARDWAJ:** ***https://github.com/Vishvesh-Bhardwaj/MLCC-Project-House-Value-Estimation***
- ✓ **RAGHAV KAPOOR:** ***https://github.com/Kraghav2002/MLCC-Project---House-Value-Estimation***