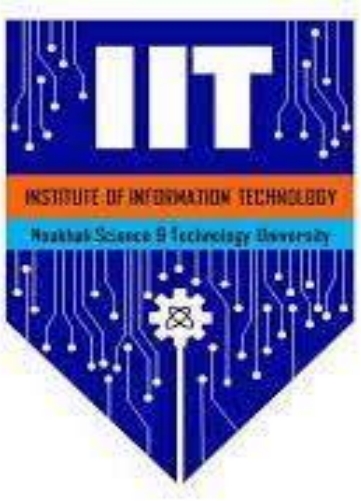
Noakhali Science and Technology University



Institute of Information Technology

Bachelor of Science in Software Engineering(BSSE)

Assignment : Report

Course Title :- Data Science Lab

Course code :- CSE - 3106

Submitted to :-

Nazmun Nahar

Lecturer

Institute of Information Technology

Noakhali Science & Technology University

Submitted by :-

Fardin Alam Alif *ID : MUH2025001M*

Md Sanwar Hossain *ID : MUH2025018M*

Md Foysal Mahmud *ID : MUH2025026M*

Date of submission:- 03 / 09 / 2023

Report's Title:

Smart House Price Predictor

Table of Contents

Overview

In this project, we aim to predict house prices using machine learning techniques. The primary goal is to develop a model that can accurately estimate the selling prices of houses based on various features. We'll follow a structured approach, starting from problem definition and data collection, all the way to model evaluation and selection.

1. Introduction

1.1 Problem Statement

The goal of this project is to predict the sales prices of houses in Bengaluru, India, using machine learning techniques. The project follows a structured approach to data analysis, model development, and evaluation.

1.2 Data Source

The dataset used in this project is sourced from Kaggle and contains information about various characteristics of houses in Bengaluru. It includes features such as area type, location, size, number of bathrooms, balcony count, and price.

1.3 Evaluation Metrics

The primary evaluation metric for this project is Root Mean Squared Logarithmic Error (RMSLE). This metric is suitable for regression problems and penalizes underestimation of house prices. Additionally, Mean Absolute Error (MAE) and R-squared (R^2) are used to assess model performance.

2. Data Preparation

2.1 Importing Libraries

In this section, necessary Python libraries, such as Pandas, NumPy, and Matplotlib, are imported to facilitate data analysis and manipulation.

2.2 Loading Data

The dataset is loaded from the provided CSV file, and its structure is examined using the **head()** and **info()** functions.

2.3 Initial Data Exploration

The initial exploration includes checking the shape of the dataset, identifying data types, and examining basic statistics using the **describe()** function. It also investigates the distribution of house prices through histograms and scatter plots.

2.4 Handling Missing Values

Missing values in the dataset are identified and addressed. Columns with significant missing data are analyzed, and strategies for handling missing values are applied.

2.5 Data Transformation

Data transformation steps are taken to convert object type columns into categorical types for model compatibility.

3. Data Analysis

3.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to gain insights into the dataset. Key visualizations, such as scatter plots and histograms, are used to analyze the relationships between features and house prices. Outliers are also identified and addressed during this stage.

3.2 Handling Missing Values

- Removed the 'availability' column as it was deemed unimportant.
- Filled missing values in the 'bath' and 'balcony' columns with their respective medians.
- Created binary columns indicating whether values were missing in 'bath' and 'balcony'.

3.3 Feature Engineering

- Converted non-numeric columns to categorical data.
- Converted categorical data to numeric codes.
- Engineered the 'bhk' feature from the 'size' column to represent the number of bedrooms.

3.4 Outlier Removal

- Removed outliers where the square footage per bedroom was less than 200.

4. Modeling

We experimented with two machine learning models:

4.1 Linear Regression

- Trained a Linear Regression model.
- Achieved an R^2 of 0.35 and a MAE of 33.57 on the test set.

4.2 Random Forest Regressor

- Employed a Random Forest Regressor.
- Used hyperparameter tuning with RandomizedSearchCV.
- Achieved an R^2 of 0.72 and a MAE of 19.89 on the test set, indicating superior performance.

4.3 Hyperparameter Tuning

- RandomizedSearchCV was used to search for the best hyperparameters for the Random Forest Regressor.
- Parameters such as 'n_estimators,' 'max_depth,' 'min_samples_split,' 'min_samples_leaf,' 'max_features,' and 'max_samples' were optimized.

5. Evaluation

5.1 Performance Metrics

The models are evaluated using various performance metrics, including MAE, RMSLE, and R^2, on both training and test datasets. The results are compared to select the best-performing model.

5.2 Model Comparison

The evaluation of our machine learning model will be based on the following metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Log Error (RMSLE)
- R-squared (R^2)

The goal is to minimize MAE and RMSLE and maximize R^2.

6. Conclusion

6.1 Summary of Finding

- The Random Forest Regressor outperformed Linear Regression in predicting house prices.
- The model achieved a reasonable R^2 score of 0.72 and a low MAE of 19.89 on the test set, suggesting its potential for accurate price predictions.
- Further fine-tuning and feature engineering could potentially improve model performance.

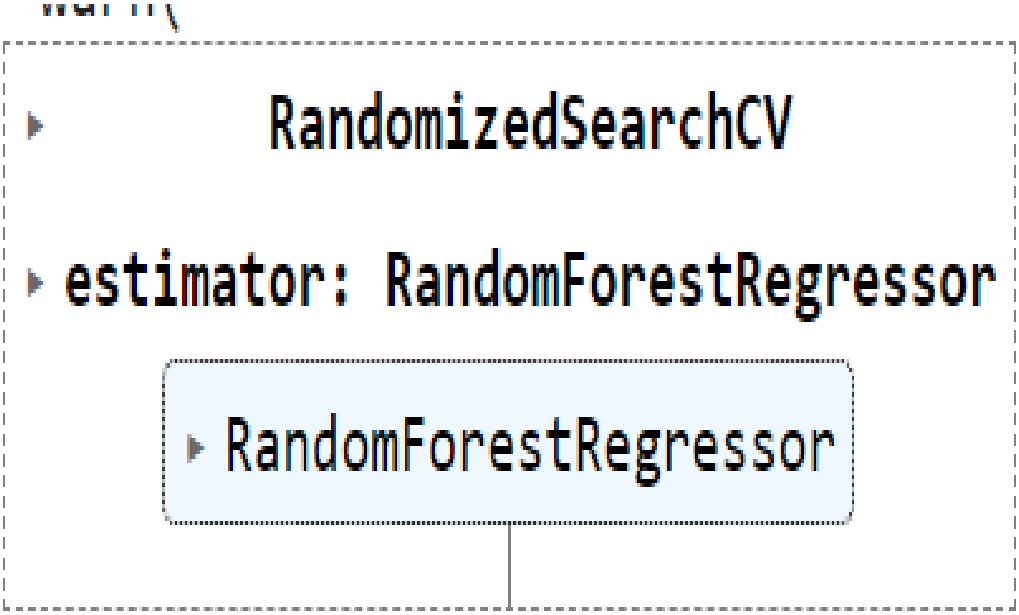
```
RandomForestRegressor
RandomForestRegressor(max_features=0.5, max_samples=3000, n_estimators=20)
```

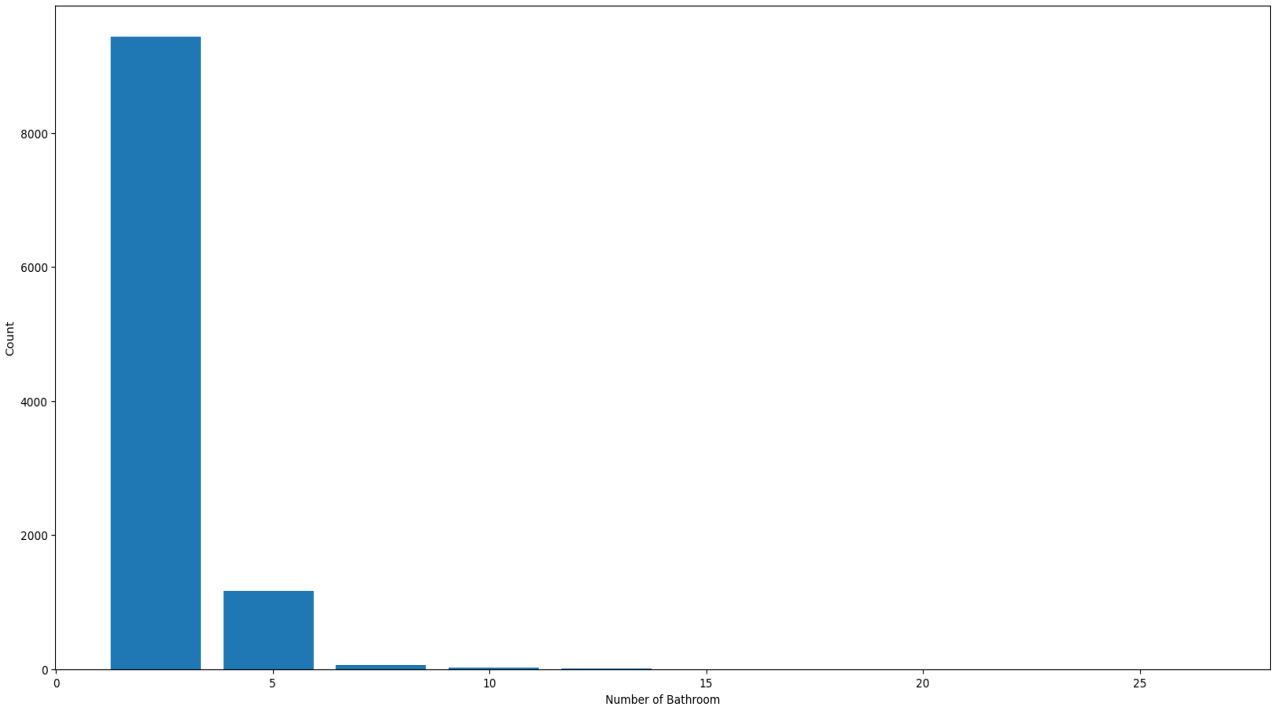
```
#score for ideal model
show_score(ideal_model)
```

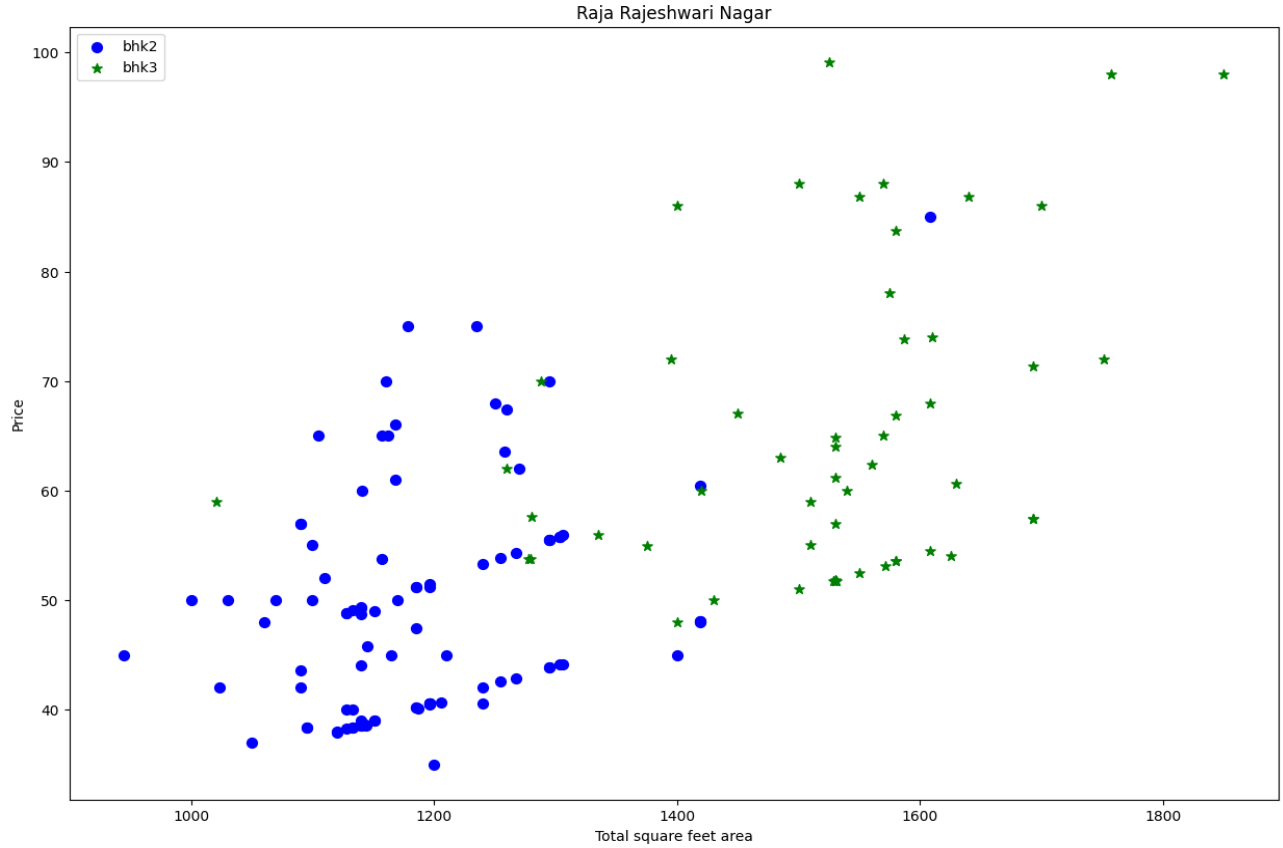
```
{'Training MAE': 14.413557051851408,
 'Test MAE': 19.89337139554667,
 'Training RMSLE': 0.1975671093979134,
 'Test RMSLE': 0.2757297174941152,
 'Training R^2': 0.8468940163705896,
 'Test R^2': 0.720085871934715}
```

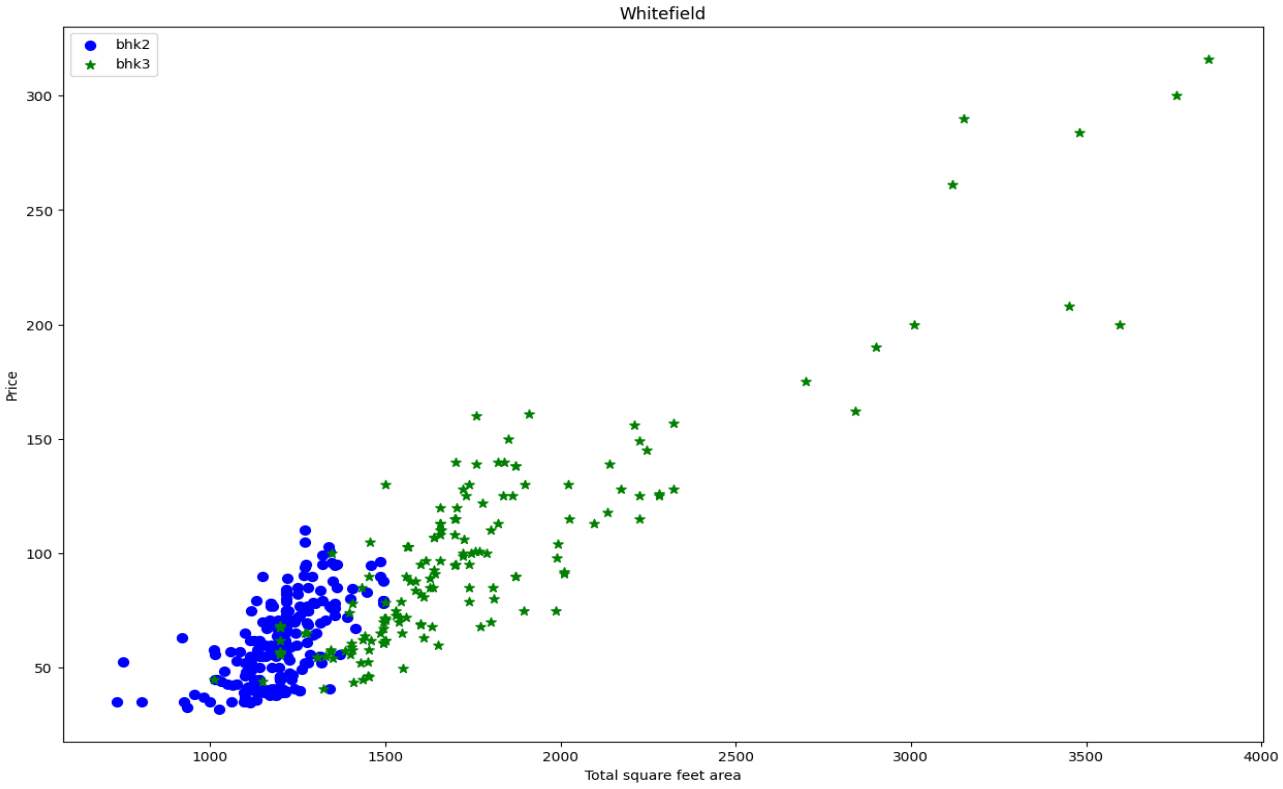
```
[ ] show_score(model)
```

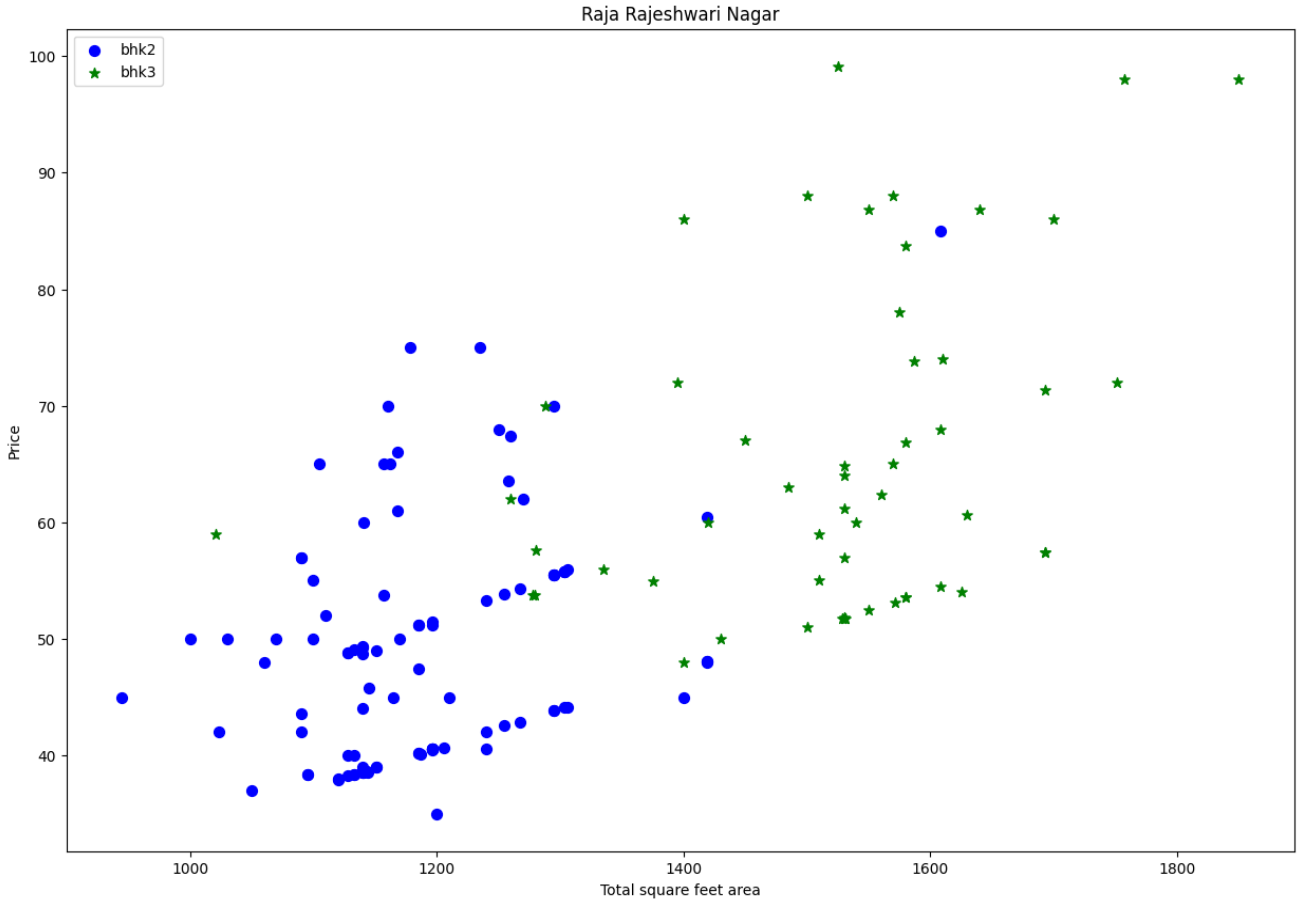
```
{'Training MAE': 18.972735932919353,
 'Test MAE': 21.291566053602356,
 'Training RMSLE': 0.24786726265973927,
 'Test RMSLE': 0.28250594831471315,
 'Training R^2': 0.7336886406653826,
 'Test R^2': 0.7044315177331714}
```


RandomizedSearchCV
estimator: RandomForestRegressor
RandomForestRegressor


Count
8000
6000
4000
2000
0
0
5
10
15
20
25
Number of Bathroom


Raja Rajeshwari Nagar
bhk2
bhk3
Price
Total square feet area
100
90
80
70
60
50
40
1000
1200
1400
1600
1800


Whitefield
bhk2
bhk3
Price
Total square feet area
300
250
200
150
100
50
1000
1500
2000
2500
3000
3500
4000



6.2 Future Work

- Experiment with more advanced regression algorithms.
- Gather additional relevant data to improve prediction accuracy.
- Explore feature scaling and normalization techniques to enhance model performance.
- Deploy the model in a real-world environment for house price predictions.

Word Count: [854]