Predicting Kolkata Housing Prices using Machine Learning

Student Name: Tamoghna Dey

Course/batch and Institute name: B.A. Economics, Jadavpur University

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project focuses on predicting median house values in Kolkata using the Kolkata Housing dataset. The primary goal was to analyse house prices and predicting prices based on other factors. The project involved a comprehensive Exploratory Data Analysis (EDA) to understand data distributions and feature correlations. A Linear Regression model, Random Forest Regressor, Decision Tree Regressor were implemented. The model was trained on a portion of the dataset and subsequently evaluated on unseen test data. Performance was measured using Mean Squared Error (MSE) and R-squared test and a comparative analysis was made of the different results.

2. Introduction

Predicting housing prices is a crucial task in the real estate sector, offering valuable insights to buyers, sellers, and investors. This project demonstrates a practical application of machine learning to solve this real-world problem by modeling prices in the complex California market. The entire analysis was conducted in Python, utilizing key data science libraries such as Pandas for data handling and Matplotlib/Seaborn for visualization. The procedure involved data exploration, preprocessing, model training, and comparative evaluation to identify the most suitable algorithm for the dataset. The purpose was to build a robust regression model and document the end-to-end machine learning workflow.

During the first two weeks of the internship, I received training on the following topics:

- **Python for Data Science (Data structures, functions, oops, NumPy, Pandas)**
- **Data Visualization with Matplotlib and Seaborn**
- **Fundamentals of Machine Learning**
- **Regression**
- **Classification**
- **Fundamentals of LLM**
- **Communication skills**

---

3. Project Objective

The objectives for this project are outlined below:

- **To perform a comprehensive Exploratory Data Analysis (EDA) on the California Housing dataset to understand its features and their relationships.**
- **To prepare the data for machine learning by splitting it into training and testing sets.**
- **To build and train a linear regression model**

- **To evaluate and quantitatively compare the performance of each model using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics**
- **Check if the model is suitable for prediction.**

---

4. Methodology

The project was executed following a structured workflow from data collection to model evaluation.

Data Collection and Pre-processing:

1. **Data Source: The project used the standard California Housing dataset, which was collected from the sklearn.datasets library in Python.**
2. **Data Cleaning: There were various methods and steps used in data cleaning**
    - **Certain functions were loaded i.e. pandas, numpys and matplotlib as pd,np,plt respectively.**
    - **Synthetic Missing values (NaN) were inserted using the index function along with numpy.**
    - **Information about the data using the info() function was used to know about the information of the data in the dataset.**
    - **The Latitude and Longitude columns were dropped since they did not match with the dataset required for regression using the drop() function**
    - **Using the describe() function the statistics of the data in the numeric columns were acquired**
    - **The distribution of the median house Values along the dataset was plotted into a graph using the plt() function**
    - **The total area of the houses was plotted using a seaborn plot using the sns function, here total area was calculated as 'total rooms' + 'Total bedrooms'**
    - **The check for duplicated rows was done using the duplicate().sum() function and were removed if present.**
    - **Missig values were also checked using isna().sum() function**
    - **The missing values were handled using certain techniques: Replacing by mean, standard deviation, interpolation, interpolation using polynomial method, KNN imputation and KNN imputation on inversed data. Out of these KNN Imputed values were chosen for replacement.**
    - **A correlation heatmap between all the features was created using the sns.heatmap() function and then the features that were correlated to Median House Value to a certain extent were noted.**
    - **Also the distribution of each feature was plotted using an sns.pairplot() function.**
    - **`Median income`, `total rooms`, `housing median age` were the features that were selected for regression**

- **The sns.pairplot() function was again used to plot the distribution of different features corresponding to its median house value.**

3. **Data Splitting: The dataset was divided into a training set and a testing set, with 60% of the data used for training the models and the remaining 40% reserved for testing and evaluation.**

Model Development and Validation:

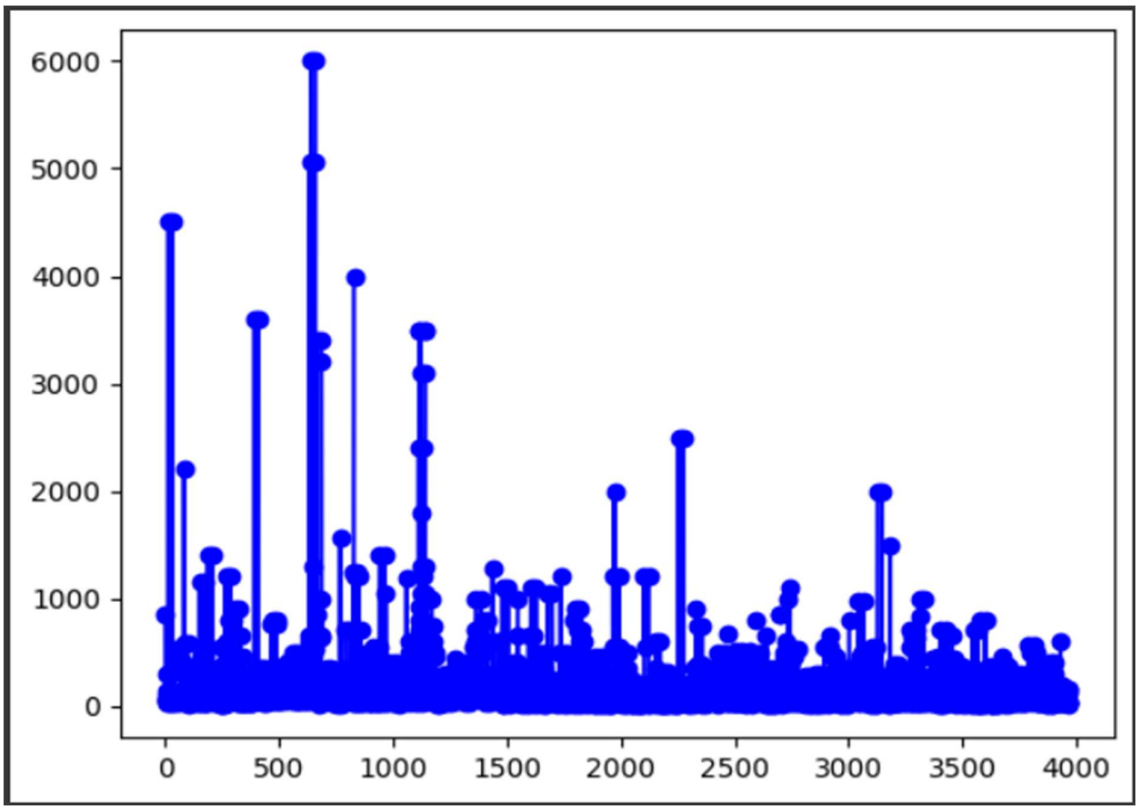
- **Model Selection: Three regression models were selected for this task. Linear Regression was used as a simple baseline.**
- **Validation: The train-test split methodology served as the validation strategy. Model was trained exclusively on the training data and the predictive performance was assessed on the unseen test data to provide an unbiased evaluation.**

The Python code developed for this project has been made available on GitHub.

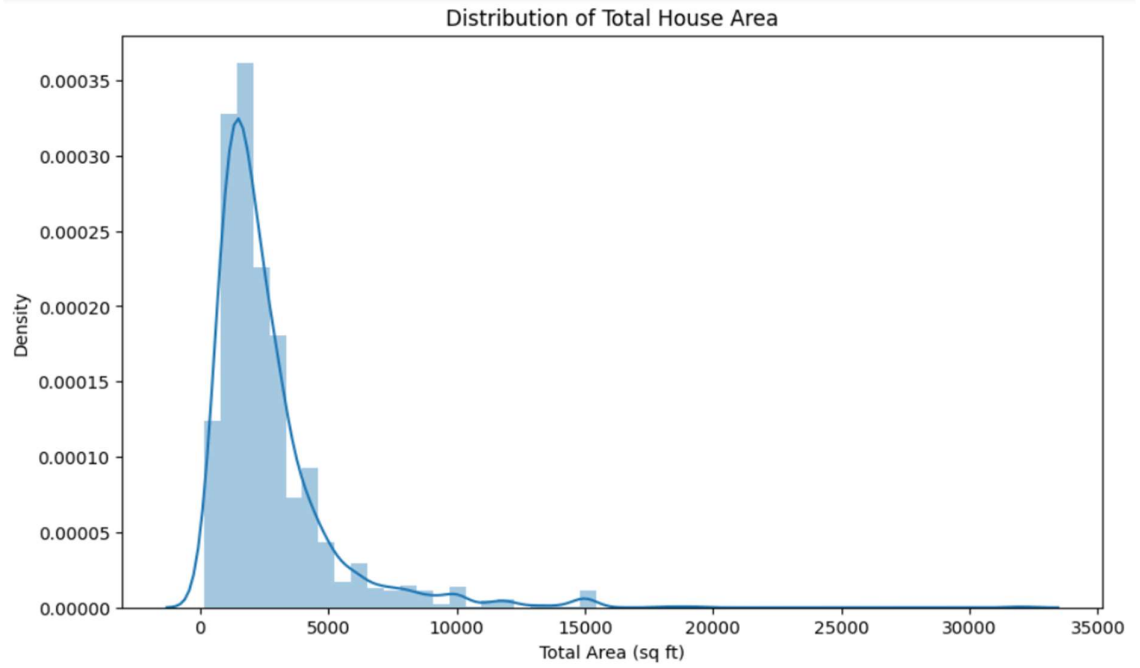
---

5. Data Analysis and Results

Descriptive Analysis: The EDA phase provided valuable insights into the dataset. Histograms were generated for all features, revealing the distribution of each variable. For instance, EMI rate, Total Square feet, Price per square feet were identified as having a right-skewed distribution. A correlation heatmap was also created, which visually summarized the relationships between variables. This heatmap clearly showed a strong positive correlation between these features and the target variable, Flat Price, which aligns with real-world expectations.

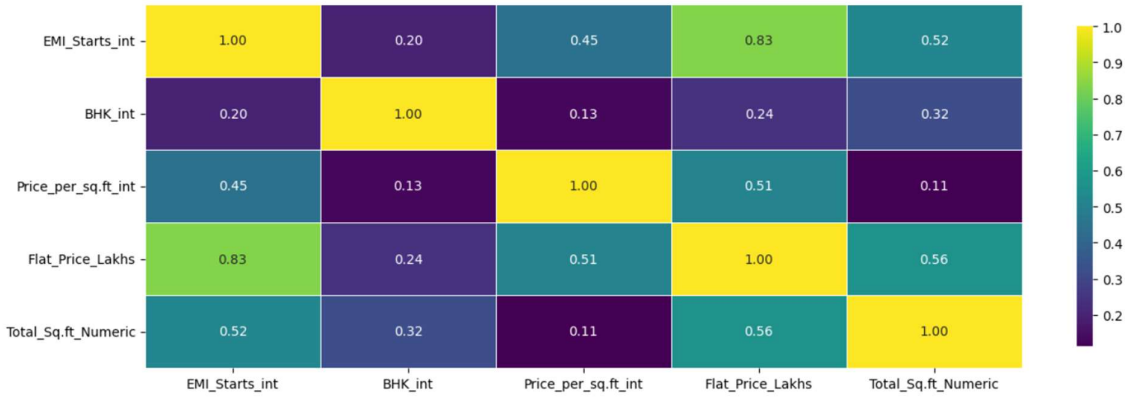


The above graph shows distribution of House Price in the dataset. The plot gives a visual representation of how the house prices are spread, showing the range and any potential outliers.



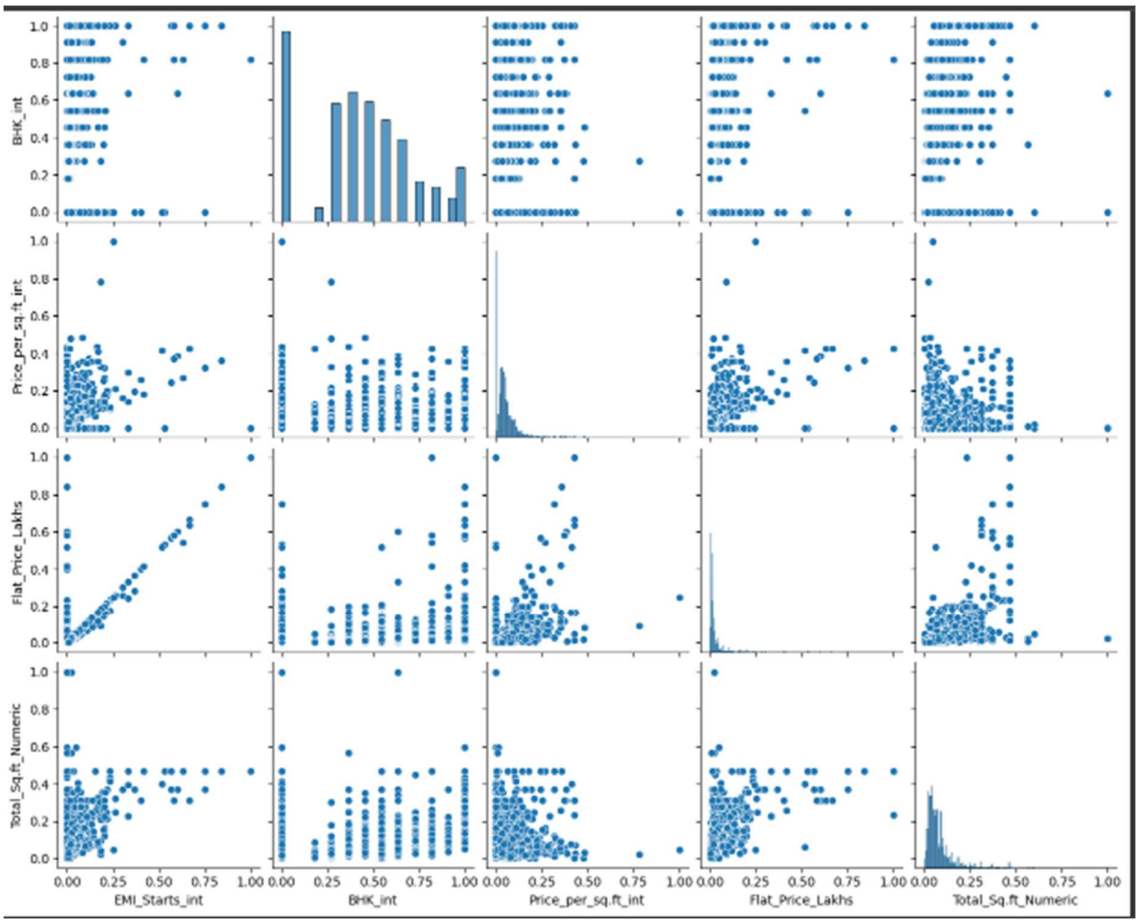
This plot is a histogram with a Kernel Density Estimate (KDE) overlay showing the distribution of the total house area (Total_Sq.ft_Numeric). The x-axis represents the total

area in square feet, and the y-axis represents the density. The bars of the histogram show the frequency of houses within different area ranges, while the KDE curve provides a smoothed representation of the distribution.



This heatmap displays the correlation matrix for the numeric features in the dataset after KNN imputation and scaling. The colour intensity indicates the strength of the correlation between pairs of features. A value close to 1 (darker colour) indicates a strong positive correlation, a value close to -1 (lighter colour) indicates a strong negative correlation, and a value near 0 indicates a weak correlation.

For example, we can see a strong positive correlation (0.83) between Flat_Price_Lakhs and EMI_Starts_int, suggesting that houses with higher prices tend to have higher EMI starts. We can also observe the correlation of each feature with itself, which is always 1.



This pairplot shows the pairwise relationships between the numeric features in the dataset after KNN imputation and scaling.

Each scatter plot in the non-diagonal cells represents the relationship between two different features. We can observe trends, clusters, or patterns that suggest a correlation between them. For instance, we might see if higher values in one feature correspond to higher or lower values in another.

Inferential Analysis and Model Results: A comparative analysis of the three models was conducted based on their performance on the test data. The results are summarized in the table below.

index	Model	MSE	R-squared
0	Linear Regression	0.003873291671305263	0.3049478287
1	Decision Tree Regressor	0.002719930061412777	0.5119155862
2	Random Forest Regressor	0.0020259573287890807	0.6364471979

The Random Forest Regressor achieved the lowest MSE and highest goodness of fit, indicating its superior prediction accuracy compared to the other two models.

---

6. Conclusion

The project successfully achieved its objective of building and evaluating machine learning models to predict Kolkata housing prices. The analysis concluded that the Random Forest Regressor was the best-performing model, as evidenced by its significantly lower Root Mean Squared Error of 0.002. This result highlights the effectiveness of ensemble methods in improving predictive accuracy for complex regression tasks.

For future work, it is recommended to explore hyperparameter tuning for the Random Forest model to potentially enhance its performance further. Additionally, experimenting with other advanced algorithms like Gradient Boosting could yield even more accurate predictions.

---

7. APPENDICES

1. **References: https://www.kaggle.com/datasets/kuntalmaity/house-price**
2. **Github link for the codes developed:**
    - **https://github.com/Tamoghna-Dey/Autumn-Internship-Project---Tamoghna-Dey.git**
3. **Any other Document Link**
    - **https://colab.research.google.com/drive/1F431vU1cQreaRiWLNfNJ237CS0LZeA8-#scrollTo=zOqW_k-F5uLJ**