Suniksha Gupta

Final Report:
Predicting House Prices

I. Introduction

a.) Problem Statement

Housing price prediction has become one of the most significant applications of data science and machine learning due to its direct relevance to the real estate market, financial decision-making, urban planning, and policy development. Real estate prices are influenced by a complex interaction of factors such as location, infrastructure, economic conditions, property characteristics, and market demand. Because these variables interact in non-linear and unpredictable ways, traditional statistical methods often fail to capture the true dynamics of the housing market. This creates a strong need for advanced, data-driven predictive modelling.

Machine learning techniques offer powerful tools to analyse historical housing data, recognize hidden patterns, and build predictive models that can estimate property prices with high accuracy. Accurate price prediction benefits a wide range of stakeholders. For buyers and sellers, it helps in making informed financial decisions. Real estate agents and developers can use predictive insights for strategic planning, while financial institutions rely on such models for loan approvals and risk assessment. Governments and policymakers can also utilize predictive analytics to understand housing trends, optimize resource allocation, and design housing policies.

b.) Goal

In this project, we aim to develop a robust and reliable machine learning model capable of predicting housing prices using a variety of features such as property size, number of rooms, location, age of the house, and other structural or environmental attributes. The project follows a complete data science pipeline that includes data cleaning, preprocessing, exploratory data analysis (EDA), model development, hyperparameter tuning, and performance evaluation. Multiple machine learning models, including Linear Regression, Decision Trees, Random Forests, and Gradient Boosting methods are implemented and compared to identify the most accurate model.

This project not only helps in understanding the technical workflow of regression-based machine learning problems but also provides hands-on experience in handling real-world data challenges such as missing values, outliers, categorical variables, and skewed distributions. Additionally, this work reflects how machine learning can be applied in real-life scenarios, making it an excellent example of practical data-driven decision-making.

Overall, the project demonstrates the potential of machine learning to transform complex housing market data into valuable predictive insights, contributing to a deeper understanding of both the dataset and the broader real estate landscape.

c.) Dataset Description

- **Source:** *https://www.kaggle.com/datasets/juhibhojani/house-price/data*
- **Number of rows:** *187531*
- **Number of features:** *21*
- **Target variable:** Price (in rupees)

Key Features in the Dataset: Carpet Area (sq ft), Super Area (sqft), Bedrooms, Bathrooms, Balcony, Location, Car Parking, Furnishing, facing, overlooking, Society, Status, Floor.

II. Data Cleaning and Wrangling:

I began the data wrangling process by **filtering the dataset**, reducing it from **187,531 rows to 37,501 rows**. This initial filtering step removed almost **20% of the data** that contained incomplete records, duplicated entries, or irrelevant information. Narrowing down the dataset ensured that only meaningful and usable entries were carried forward for further analysis.

Next, I focused on eliminating **inconsistencies and inappropriate values**. Several entries had unrealistic values, text where numbers were expected, or property information that did not align with any valid category. Cleaning these irregularities helped improve both the reliability and interpretability of the dataset.

A major issue involved **incorrect data types**. For example, the column *Amount (in rupees)* contained non-numerical entries such as **"Call for Price"**, which made it impossible to perform numerical operations.

After filtering out invalid entries, I standardized the *Amount* values, as the dataset contained prices in multiple units such as **crores**, **lakhs**, and **rupees**. I converted all entries into a **single unit (rupees)** using a lambda function that detected the unit and applied the correct multiplier. This ensured uniform pricing data across the entire dataset.

I also removed columns that contained only text descriptions or had no meaningful values. Dropping such columns reduced noise and kept the dataset focused on features relevant to predictive modelling.

Another issue involved the **Carpet Area**, which appeared in different units such as *sqft* and *sqyrd*. To maintain consistency, I converted all values into sqft using a lambda transformation. This step standardized the measurement unit and prevented scale-related discrepancies.

During inspection, I found several columns stored in **incorrect formats**, particularly features that should have been numeric but were read as objects due to symbols, text, or mixed formatting. I cleaned these columns by removing non-numeric characters and converting them into the appropriate **float** or **integer** types. Having correct data types is crucial for both exploratory analysis and model training.

Lastly, I addressed **missing values**. Depending on the nature of the feature: Numerical columns were imputed using the median, Categorical columns were imputed using the mode, Columns with excessive missing data were dropped entirely. This ensured that no missing values would disrupt model training or bias the results.

III. Exploratory Data Analysis:

The Exploratory Data Analysis phase focused on understanding the structure, quality, and underlying patterns of the cleaned housing dataset. This step helped uncover key trends in property prices, relationships among features, and areas requiring further preprocessing before modelling.

Removing Duplicate Entries: The EDA began with eliminating duplicate records from the dataset to ensure that insights and model training were not influenced by repeated entries. Removing duplicates improved dataset integrity and provided a more reliable foundation for subsequent analyses.

Understanding the Distribution of Housing Prices: To analyse the target variable, *Price (in rupees)*, its distribution was visualized and evaluated. The price distribution was **highly right-skewed**, indicating a concentration of properties in the mid-price range with a smaller number of luxury or high-value properties. This skewness suggested that high-end properties could disproportionately affect average pricing and impact model performance. A **log transformation** was applied to the price column shown in Figure 1, which significantly normalized the distribution. This improved the suitability of the target variable for regression models by reducing the effects of extreme values.

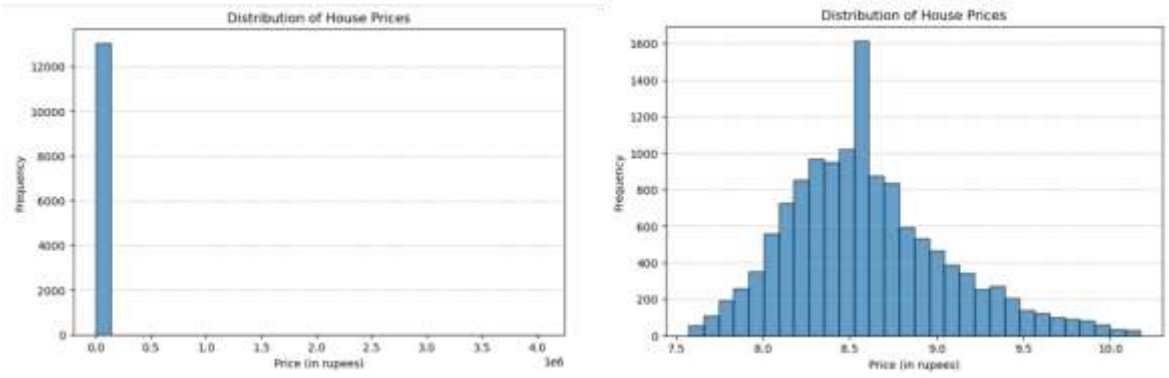

Figure 1: Distribution of Price in Rupees before and after log transformation

Handling Missing Values: A detailed inspection of missing values was performed across all features. Numerical columns were imputed using either the **mean** or **median**, depending on skewness. Categorical columns were imputed using the **mode**, ensuring category consistency. After systematic imputation, the dataset was brought to **zero missing values**, enabling robust visualizations and correlations without bias.

Location-Based Price Analysis: Locations play a major role in real-estate pricing. To understand geographical patterns. The dataset was **grouped by location** to compute the *average price* for each area. A **horizontal bar chart** was plotted to visualize price variations across locations. The **top 10 most expensive** and **top 10 most affordable** locations were extracted shown in Figure 2.

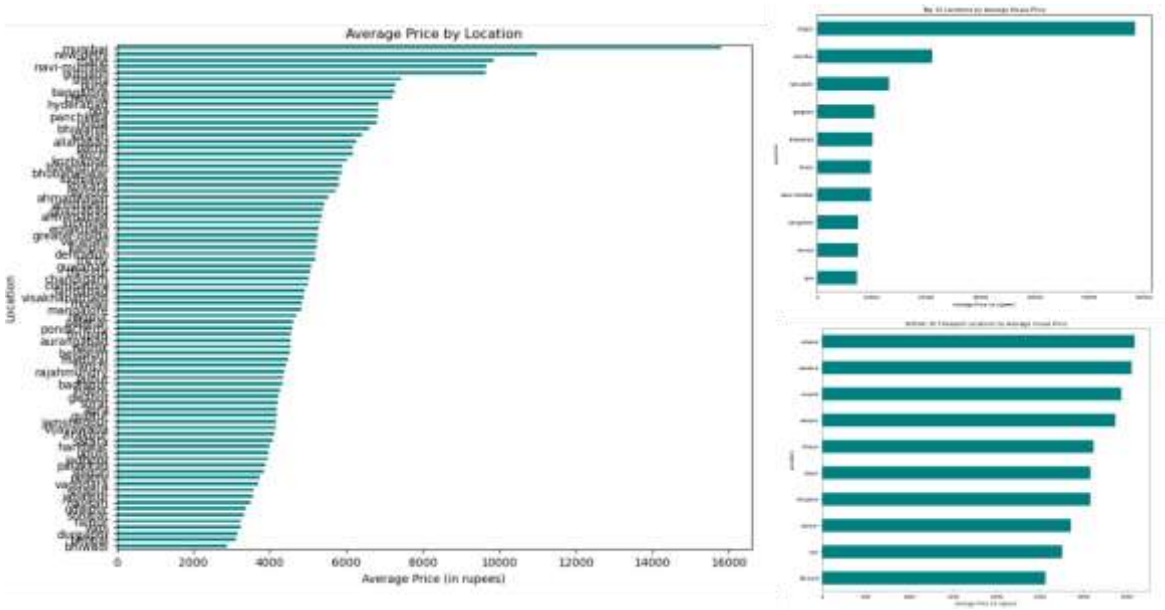

Figure 2: Horizontal Bar chart for location-based price analysis, on the right-hand side: Top 10 expensive and bottom 10 cheapest locations are shown

Premium neighbourhoods showed significantly higher mean prices. Budget locations exhibited lower average prices, indicating potential for cost-effective investment. These findings demonstrated strong regional price segregation and highlighted the impact of location on housing affordability.

Analysis of Numerical Features: Key numerical features such as **Carpet Area (sqft), Super Area (sqft), BHK, Number of Bathrooms,** and **Number of Balconies** were analysed to understand their influence on price.

Correlation Heatmap: A correlation matrix was generated to identify how numerical features relate to each other and to the target variable, Figure 3.

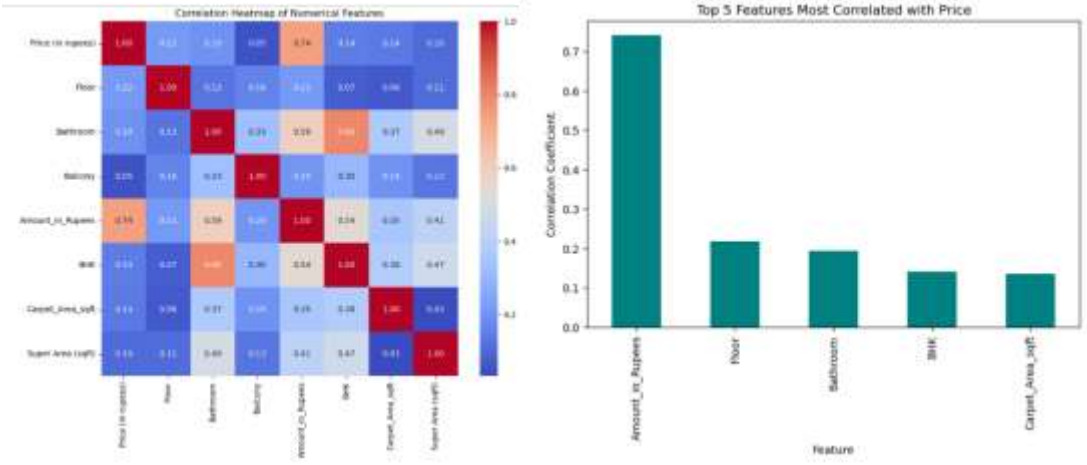

Figure 3. Correlation heatmap of Numerical Features, Top 5 most correlated features

Interpretation: *Amount_in_Rupees* is highly correlated with the target, confirming consistency between price representations. Floor, Bathroom, and BHK show meaningful relationships with price. The generally low correlations across features imply that housing prices depend on multiple factors simultaneously rather than a single dominant variable.

Feature Distribution Analysis

To further explore numerical variables, **histograms**, **KDE plots**, and **boxplots** were generated. **Carpet Area**, **Super Area**, and **Price** displayed right-skewed distributions with long tails. **BHK distribution** showed that *2BHK and 3BHK* units formed the dominant segment of the dataset. The presence of extreme values reinforced the need for transformations and outlier handling before modelling.

Outlier Detection and Removal: Outliers were identified for all major numerical features using statistical methods and visualization tools such as histogram. Outliers were removed to improve model performance and prevent extreme values from skewing predictions. Post-removal, distributions became smoother and more model-friendly shown in Figure 4.

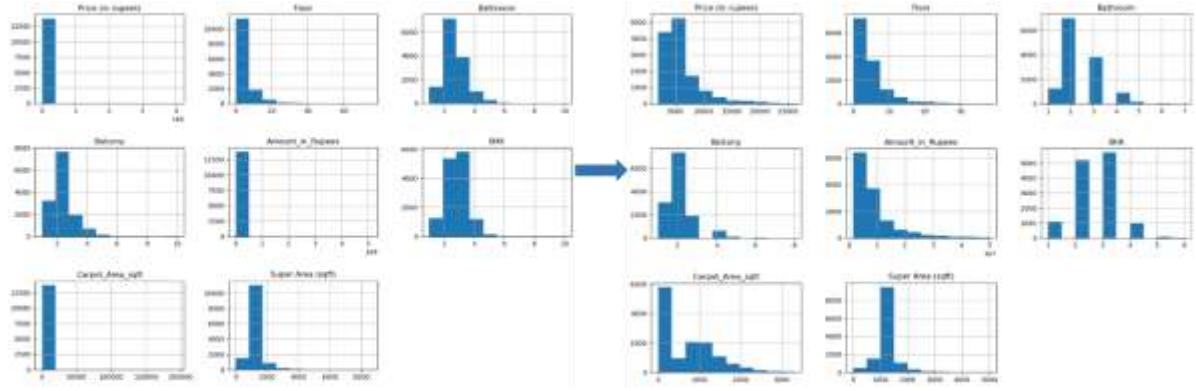

Figure 4. Histograms of Numerical Feature before and after outlier removal

Summary: Overall, the EDA revealed that: Housing prices are highly skewed and vary strongly with property size, number of bathrooms, and location. Multiple features contain skewness and outliers, reflecting both budget and luxury property segments. Location remains one of the strongest determinants of average property prices. The dataset, after imputation and outlier removal, is now balanced, consistent, and ready for feature engineering and modelling.

IV. Modelling

With the dataset prepared, five machine learning algorithms were selected for training: Linear Regression, Ridge Regression, Random Forest Regressor, Gradient Boosting Regressor, and AdaBoost Regressor. These models represented a mix of linear, regularized, and ensemble-based approaches and were chosen to explore a diverse range of predictive capabilities. Each model was trained on the prepared dataset and subsequently evaluated using a consistent set of performance metrics including R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics allowed for a comprehensive assessment of how accurate and stable each model was in predicting housing prices.

The model accuracy score for all five models along with the cross-validation training and test score came out to be:

	Algorithm	Model Accuracy Score
0	Linear Regression	0.717949
1	Random Forest	0.903235
2	Ridge Regression	0.717984
3	Gradeint Boosting	0.907254
4	Ada Boost	0.767587

	Algorithm	CV Train Score	CV Test Score
0	Linear Regression	0.734474	0.717397
1	Random Forest	0.905040	0.875657
2	Ridge Regression	0.734351	0.716784
3	Gradeint Boosting	0.911595	0.882512
4	Ada Boost	0.773633	0.758521

Table1. Model Accuracy Comparison and Model Train–Test Scores

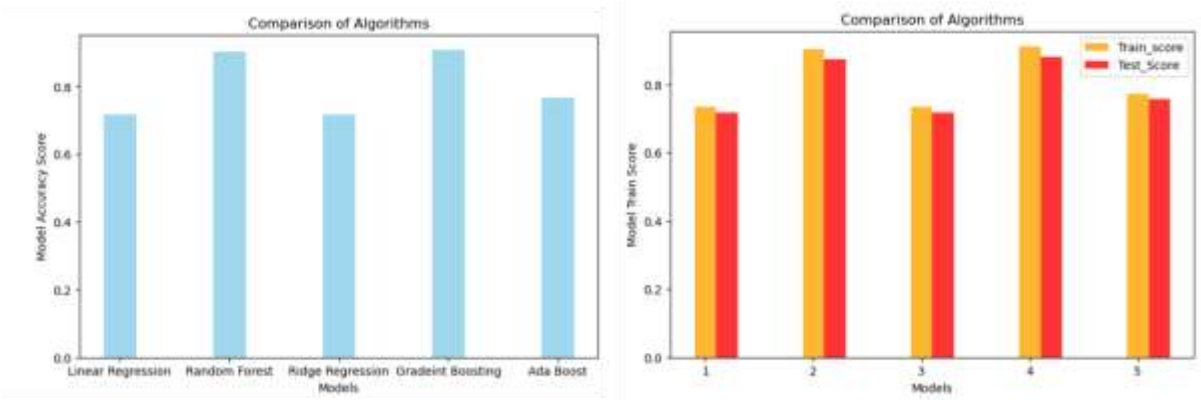

Figure 5. Graphical representation of model accuracies and comparison

The initial results revealed noticeable differences in model performance. Linear Regression and Ridge Regression achieved similar accuracy scores of approximately 0.718, indicating that simple linear models were unable to fully capture the non-linear relationships present in the housing data. AdaBoost performed modestly better, achieving an accuracy of around 0.768. Ensemble methods showed the strongest results, with Random Forest achieving a score of 0.903 and Gradient Boosting slightly outperforming it with 0.907. When evaluating train–test scores, Random Forest and Gradient Boosting demonstrated strong generalization with minimal overfitting, while Ridge Regression and Linear Regression showed weaker test performance relative to their training results. These comparisons highlighted that ensemble models were better suited for capturing the complex patterns in the dataset, with Random Forest and Gradient Boosting emerging as the top candidates for further optimization.

To enhance the model's performance, hyperparameter tuning was performed using Random Search. This method explored a large range of parameter combinations efficiently, identifying the optimal set of hyperparameters for the Random Forest model. The best configuration included 500 estimators, a maximum depth of None, log2 as the feature selection strategy for splits, and minimum sample split and leaf parameters of 2 and 1 respectively. This tuned model achieved a cross-validation R^2 score of 0.9003, indicating strong consistency across folds and confirming that Random Forest was a robust performer on the dataset.

Parameters chosen for Random Forest Regressor Hyperparameter tuning by applying Random Search:

```
param_dist = {
    'n_estimators': [100, 200, 300, 500],
    'max_depth': [None, 10, 20, 30, 40],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt', 'log2']
}
```

Best configuration after Random Search:
Best Parameters: {'n_estimators': 500, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': None}
Best CV R^2 Score: 0.9003323804260168

After refitting the model using the optimized hyperparameters, the final evaluation metrics were computed. The tuned Random Forest model achieved an **MAE of 0.0913, an MSE of 0.0219, an RMSE of 0.1480, and a final R^2 score of 0.8983**. These metrics demonstrate that the model was able to predict housing prices with high accuracy while maintaining low error across different scales of measurement. The combination of strong cross-validation performance, high accuracy, and low prediction error made the tuned Random Forest model the most reliable and effective predictive model among all those tested.
Overall, the machine learning modelling phase successfully transformed the cleaned dataset into a fully operational predictive pipeline. Through systematic encoding, feature engineering, scaling, model training, evaluation, and hyperparameter tuning, the final Random Forest model was able to deliver precise and consistent results. This stage laid the foundation for deploying the model or using it for further analysis in real-world housing price prediction scenarios.

V. Key Findings from the project:

The analysis and modelling conducted in this project reveal several important insights into the factors governing housing prices, as well as the effectiveness of different machine learning approaches for price prediction. The exploratory data analysis showed that housing prices are heavily right-skewed, reflecting a small population of luxury properties that significantly raise the upper end of the price spectrum. Features such as **location**, **carpet area**, **number of bathrooms**, **BHK**, and **floor level** emerged as the strongest predictors of price. These findings underline that both structural characteristics and geographical advantages play essential roles in determining property value.
The machine learning results further reinforced these insights. Linear models like Linear Regression and Ridge Regression delivered moderate performance, with accuracy scores around **0.718**, indicating their limitations in capturing the non-linear interactions inherent in housing data. Ensemble models, on the other hand, performed significantly better. **Random Forest** achieved an accuracy of **0.903**, while **Gradient Boosting** slightly outperformed it with **0.907**, demonstrating that tree-based ensembles are much better suited for this prediction task. Cross-validation results confirmed that these models generalize well, with relatively small gaps between training and testing performance. After hyperparameter tuning, the optimized model achieved an **R^2 of ~0.90**, along with low MAE, MSE, and RMSE values—indicating strong predictive reliability and robustness.
Overall, the findings of this study confirm that the housing market is influenced by a combination of location-specific and property-specific factors, and that powerful ensemble models can effectively capture these complex relationships. The final model developed here provides a reliable tool for price estimation and can support strategic decision-making in real estate applications.
Based on these insights, several recommendations can help stakeholders extract the maximum value from this work. First, the client can **deploy an automated property price estimation system** using the final optimized model. This would benefit real estate agents, buyers, and sellers by providing immediate, data-driven pricing guidance and reducing uncertainty during negotiations. Second, developers and investors can use the model's learned feature importance to make **more informed investment and construction decisions**—for example, prioritizing regions with higher average valuations or designing homes with features that command stronger price premiums. Finally, the findings can support **targeted marketing strategies**, allowing real estate agencies to tailor their property promotions based on which attributes are most valued by different buyer segments.
Together, these findings, conclusions, and recommendations position the model as a valuable analytical tool that can improve pricing accuracy, enhance investment planning, and support better customer engagement in the real estate domain.