Foundations of Data Science BCSE206L

Bangalore Apartment Price Prediction

Group 6

Agenda

- Introduction
- Overview of Lifecycle
- Discovery
- Data Preparation
- Model Planning
- Model Building
- Communication of Results
- Operationalize
- Conclusion

Introduction

Bangalore, India's "Silicon Valley," has witnessed a significant boom in recent years, attracting businesses, professionals, and homeowners alike. This surge in demand has also impacted the city's housing market, leading to rising property prices. This project aims to analyze the current state of Bangalore's housing market by delving into various factors that influence house prices.

Lifecycle

Discovery

Problem statement
Stakeholders
Data sources

Data Preparation

Data cleaning
Exploratory data analysis
Data normalisation

Model Planning

Variable selection
Deciding the most suitable machine learning model

Model Building

Splitting dataset
Building model
Calculate error

Results

Communicate results
Deploy the model
Maintain and scale model

Discovery

Problem statement and Domain

Predicting apartment prices in Bangalore, relevant to real estate or property market analysis

Stakeholders

People seeking to buy property, real estate agents, brokers, etc.

Identification of Data Sources

Bengaluru House Data by Amitabh Chakraborty on Kaggle

Initial Hypotheses

Apartment prices depend on size, bhk and amenities.

Data Preparation

- involves cleaning, transforming and organising raw data into a format suitable for analysis.
- ensures that the data used is complete, concise, accurate and relevant.

Data Preparation

Preparing a Sandbox

- creating a directory structure
- Jupyter Notebooks is the environment used for analysis

Performing ETLT

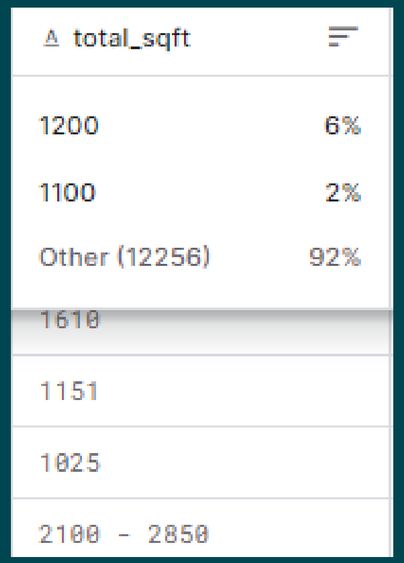
- Extract: Obtain the raw data from kaggle
- Transform: Cleanse, filter, and preprocess the data to make it suitable for analysis.
- Load: Store the transformed data in a structured format

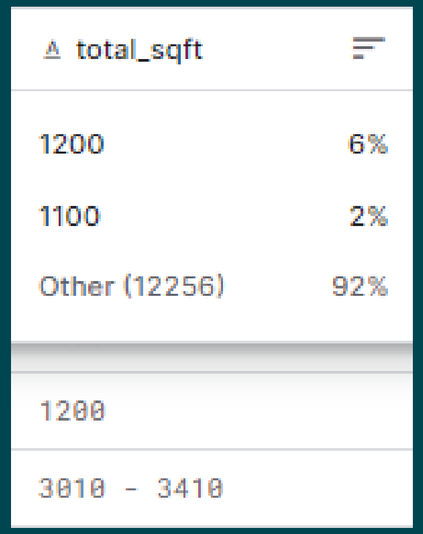
Data Cleaning

- Dropping rows containing null values :

Rajaji Nagar	4 BHK	Brway G	3300
Marathahalli	3 BHK		1310
Gandhi Bazar	6 Bedroom		1020
Whitefield	3 BHK		1800

- Taking mean of values which are in the form of range :





- Formatting the data :

```
array(['2 BHK', '4 Bedroom', '3 BHK', '4 BHK', '6 B
       '1 BHK', '1 RK', '1 Bedroom', '8 Bedroom', '
```

	location	total_sqft	bath	price	bhk
0	1st Block BEL Layout	1540.0	3.0	85.0	3
1	1st Block BEL Layout	1800.0	5.0	250.0	4
2	1st Block HBR Layout	2500.0	6.0	500.0	5
3	1st Block HBR Layout	600.0	1.0	45.0	1
4	1st Block HBR Layout	3150.0	4.0	150.0	4

- Locations with apartments < 10 are moved to new column named other

Data Preparation

Learning About the Data

- Explore the dataset to understand its structure, variables, and distributions.
- Identifying any patterns or anomalies

Data Conditioning

- Handling Outliers
- Removing outliers from columns
  - sqft : using price per sqft
  - bhk
- Columns not required are dropped

Data Preparation

Survey and Visualize

- Gathering relevant information, structuring it into a dataset,
- employing exploratory data analysis to uncover insights

Data Preparation

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

	location	total_sqft	bath	price	bhk
0	1st Block Jayanagar	2850.0	4.0	428.0	4
1	1st Block Jayanagar	1630.0	3.0	194.0	3
2	1st Block Jayanagar	1875.0	2.0	235.0	3
3	1st Block Jayanagar	1200.0	2.0	130.0	3
4	1st Block Jayanagar	1235.0	2.0	148.0	2

Visualization

- Data visualization is a powerful tool for interpreting and communicating complex data.
- It helps in uncovering patterns, trends, and insights that are not easily apparent from raw data alone.

Visualization

Key Considerations

Granularity and Range of Values

- Understanding the depth of data and checking value ranges for comprehensive analysis.

Population Representation

- Ensuring the dataset represents the broader population accurately.

Standardization and Normalization

- Maintaining consistent scales for meaningful comparisons.

Geospatial Consistency

- Ensuring accuracy and consistency in representing location-based data.

Visualization

Key Considerations

Granularity and Range of Values

- converting features like 'total_sqft' and 'bhk.'
- Ensures a comprehensive understanding of the dataset's depth and characteristics.

Population Representation

- Location Specific Factors
- Outlier removal based on location-specific price per square foot .

Visualization

Key Considerations

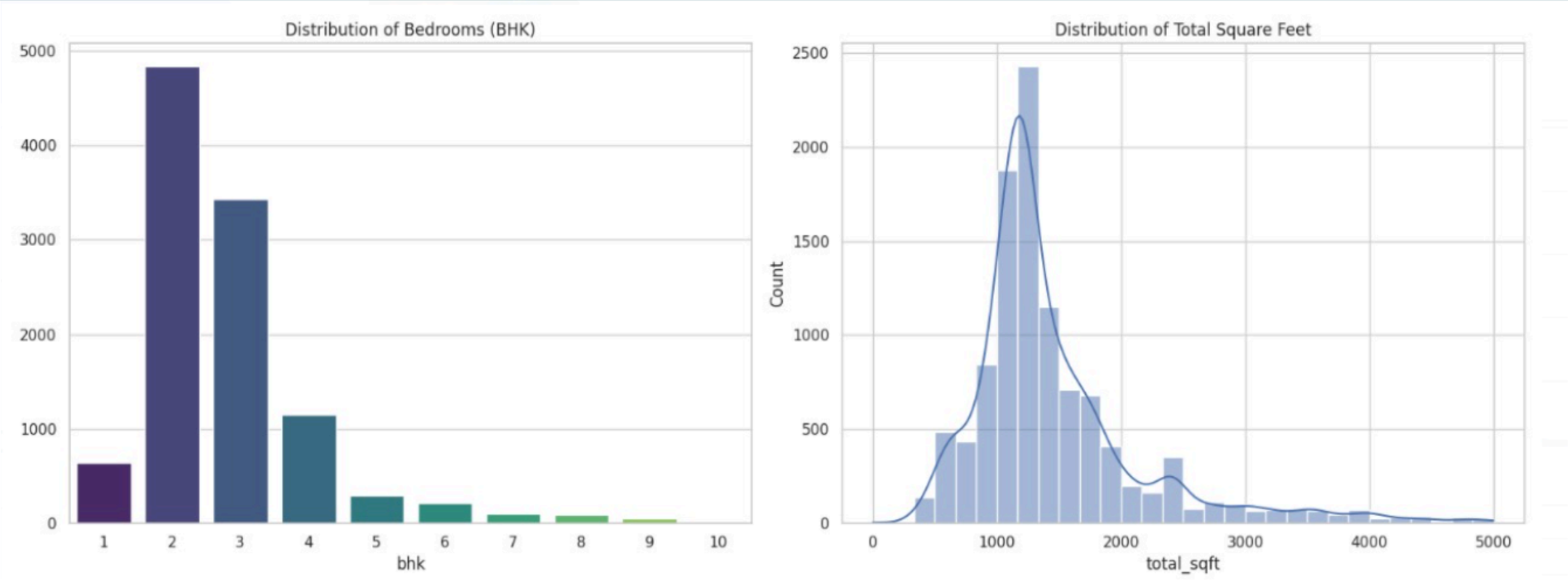
Standardization and Normalization

- Normalization ensures consistent scales for features like 'total_sqft,' improving comparability.
- Conversion of 'total_sqft' to numeric values for standardized analysis across different properties.

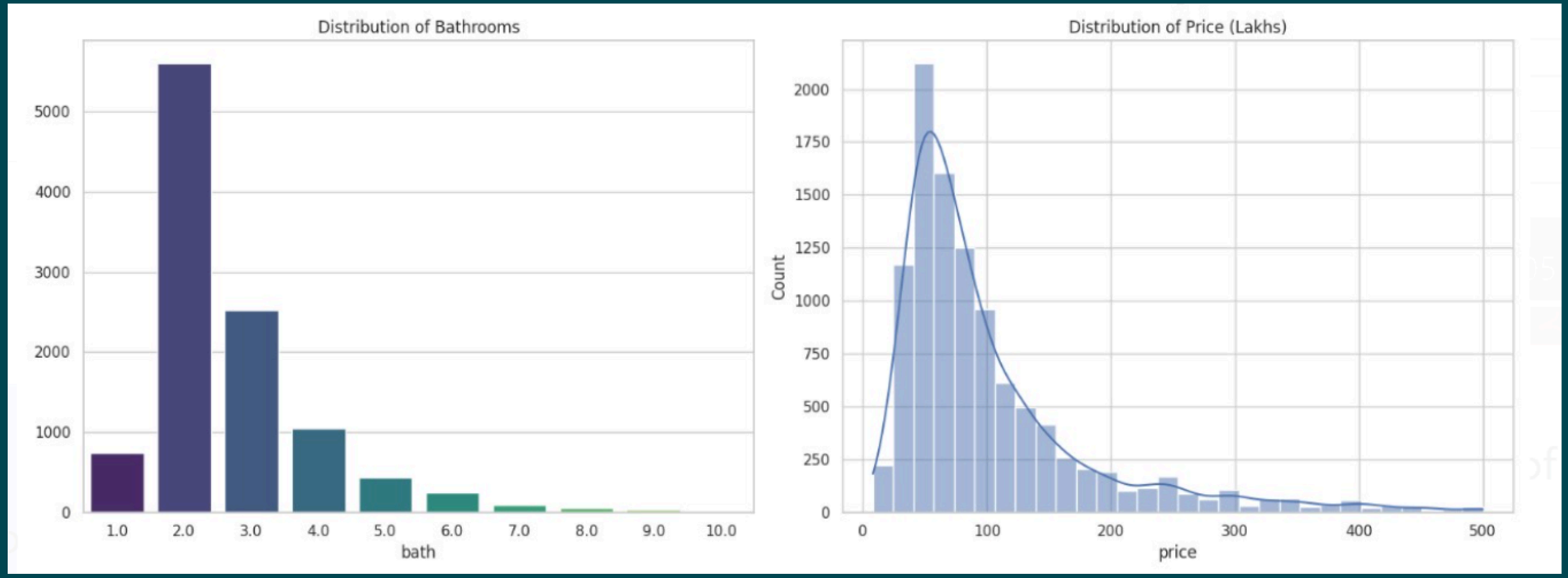
Geospatial Consistency

- crucial for accurate analysis of location-based data

Data Visualisation (EDA)



Data Visualisation (EDA)



Model Planning

Variable Selection

- Location
- Number of bedrooms
- Number of bathrooms
- Total square feet
- Price

Choice of Models

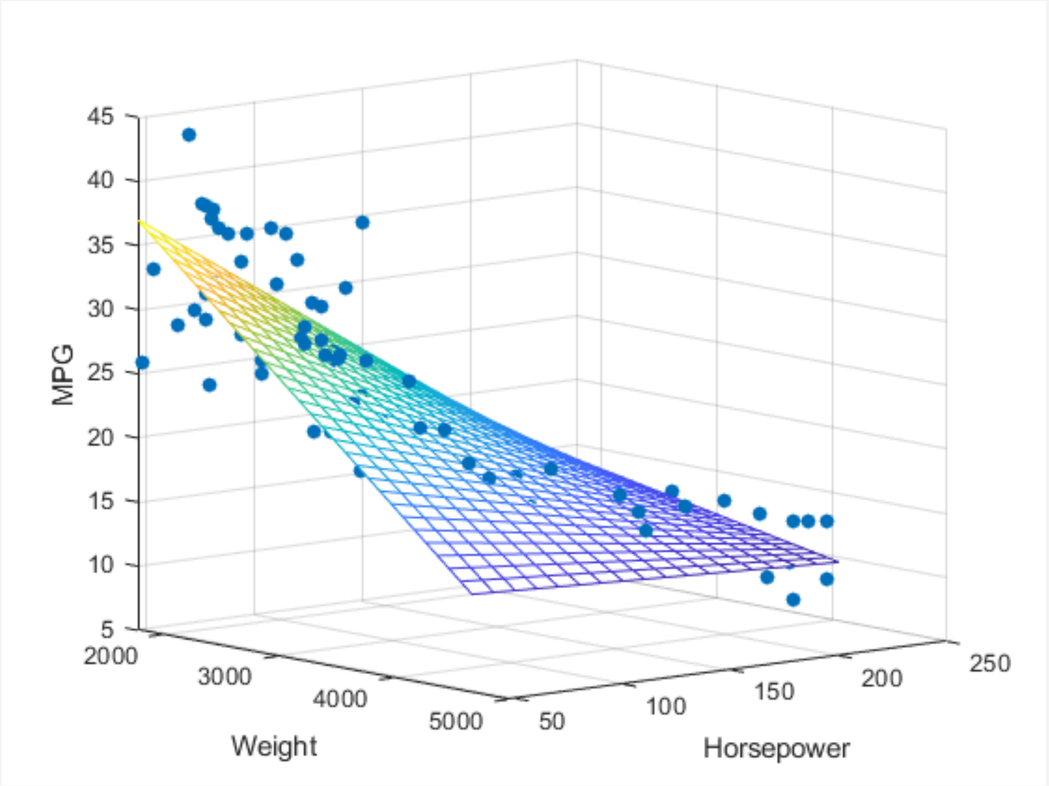
- Gradient Boosting
- Multiple Linear Regression
- Random Forest (most accurate)

Model Building

What is Multiple Linear Regression?

MLR examines how multiple independent variables are related to one dependent variable.

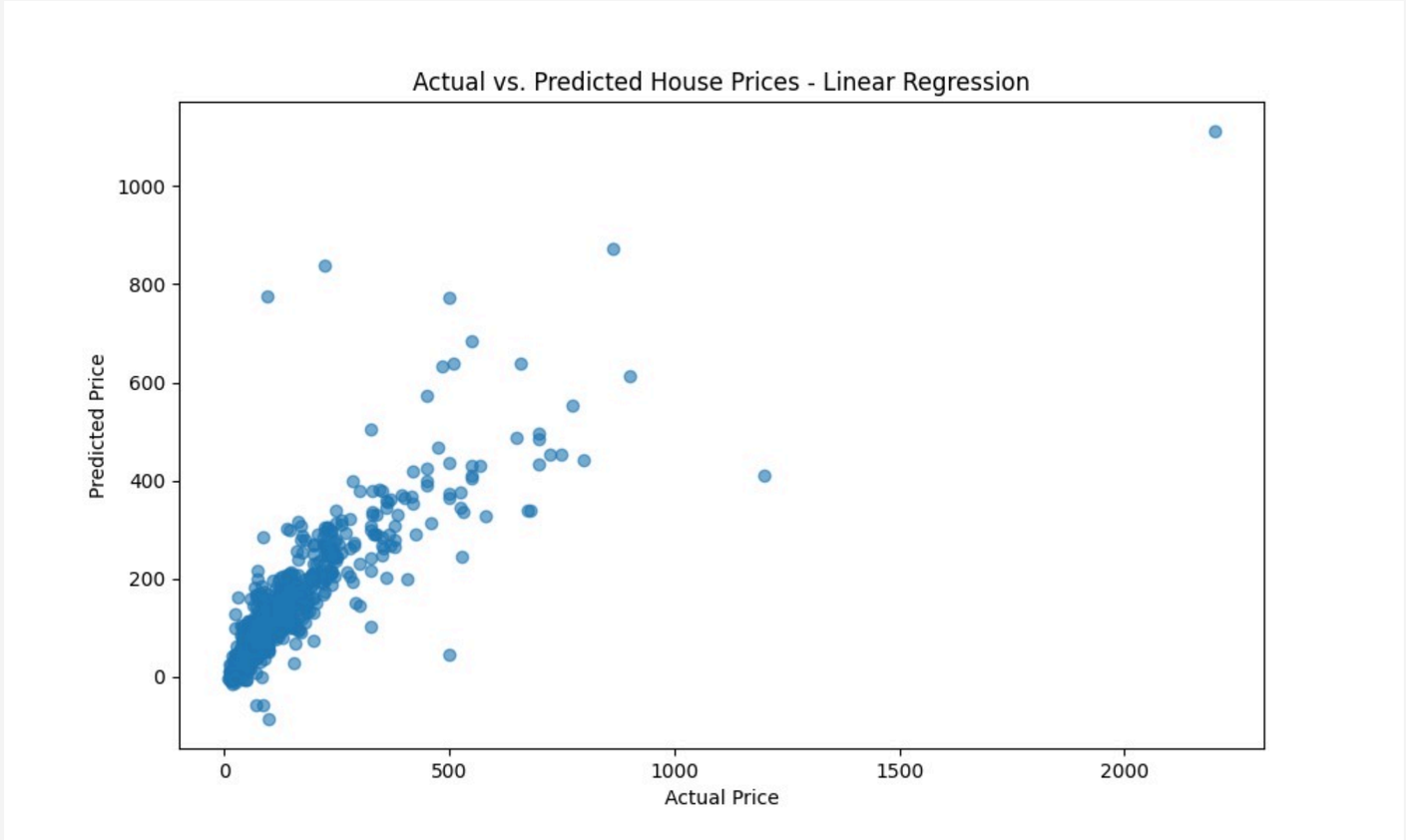
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$



Model Building

Linear Regression

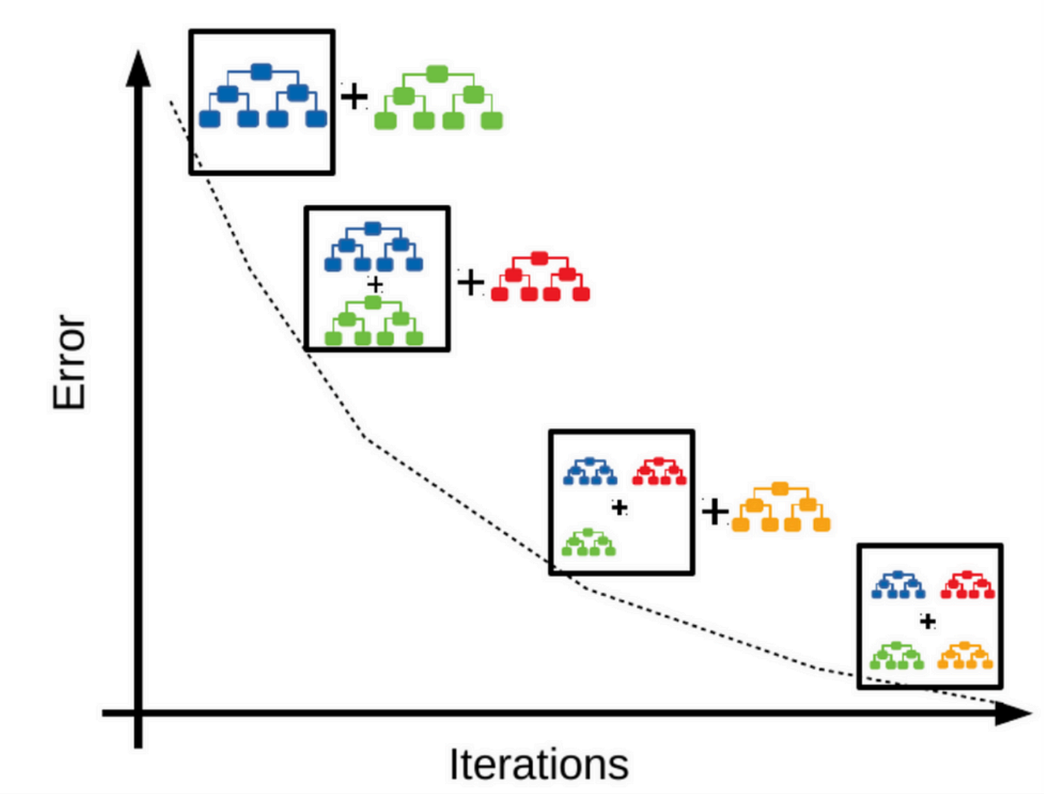
Accuracy score: 0.64



Model Building

What is Gradient Boosting?

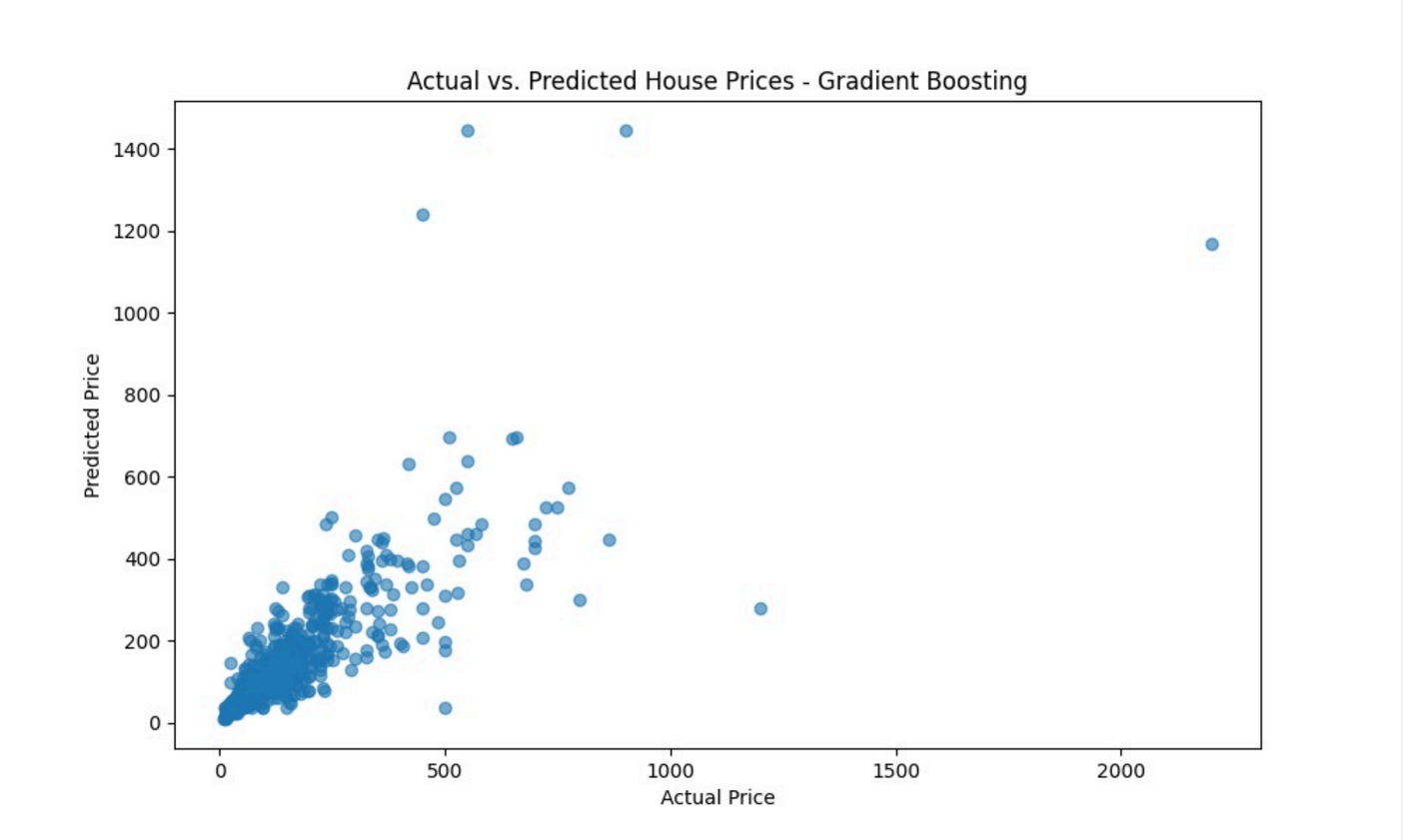
Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.



Model Building

Gradient Boosting

Accuracy score: 0.75

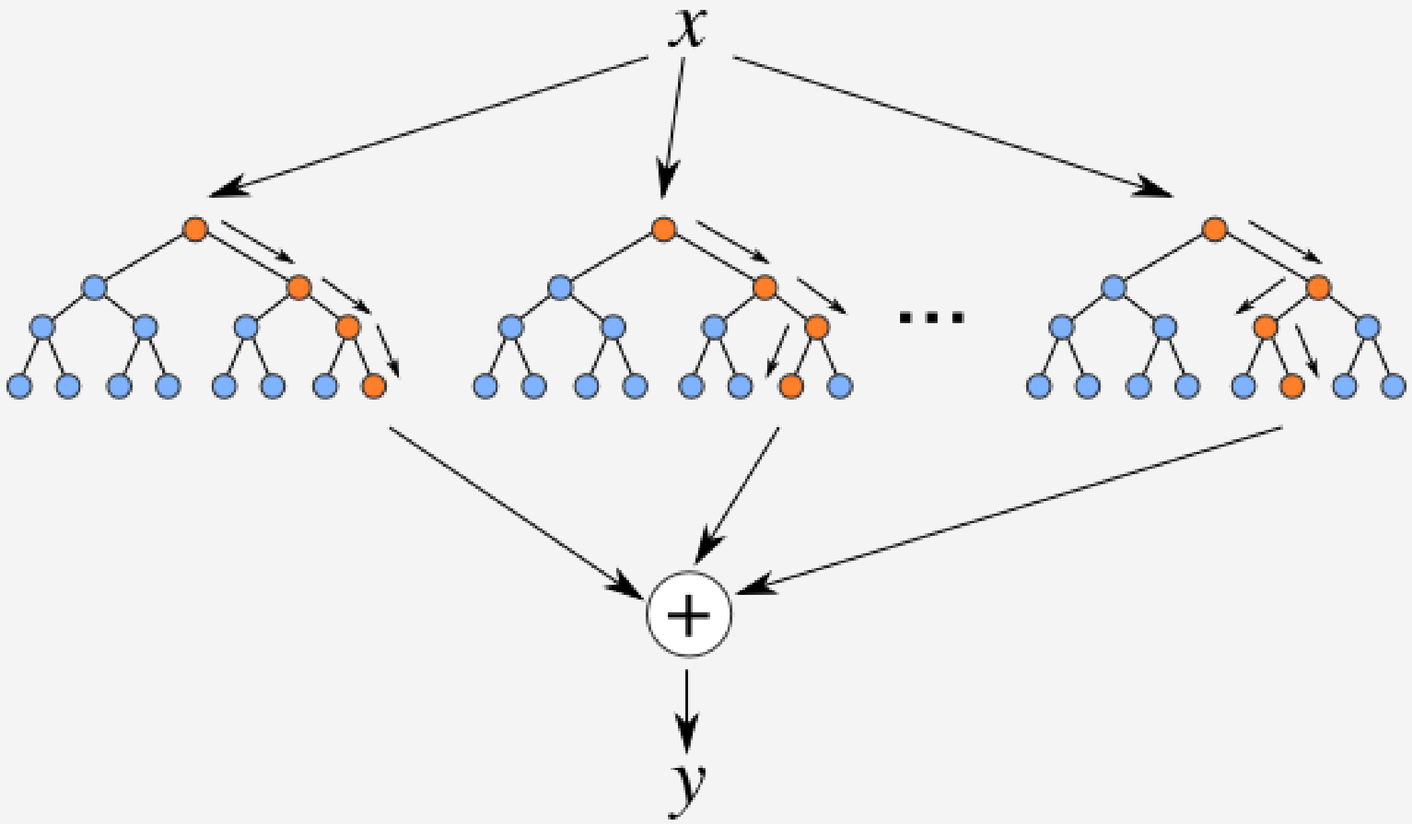


Model Building

What is Random Forest?

Every decision tree has high variance, but when we combine all of them in parallel then the resultant variance is low as each decision tree gets perfectly trained on the sample data, hence the output doesn't depend on one decision tree.

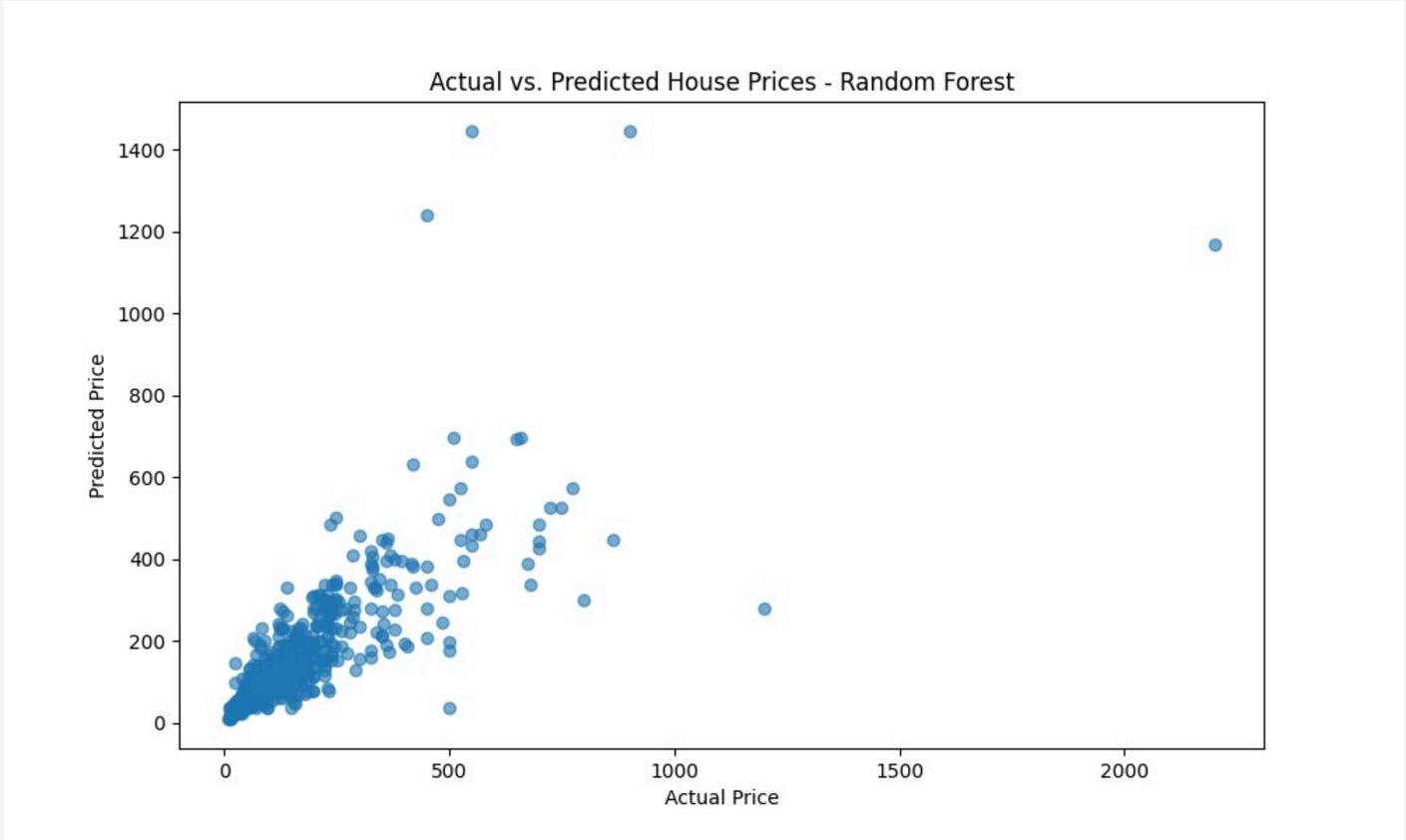
In the case of a regression problem, the final output is the mean of all the outputs.

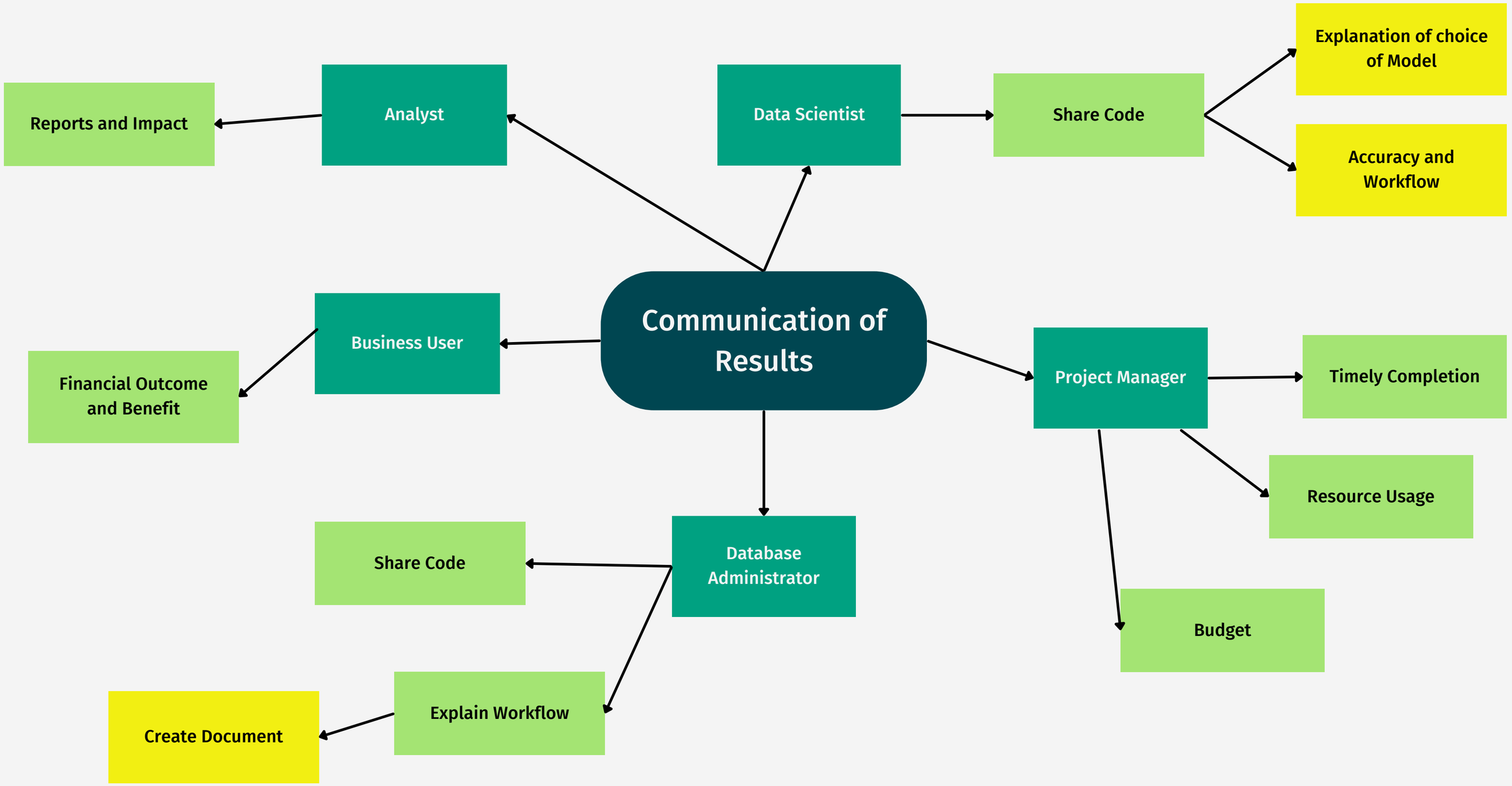


Model Building

Random Forest

Accuracy score: 0.81




Communication of Results
Analyst
Reports and Impact
Data Scientist
Share Code
Explanation of choice of Model
Accuracy and Workflow
Business User
Financial Outcome and Benefit
Project Manager
Timely Completion
Resource Usage
Budget
Database Administrator
Share Code
Explain Workflow
Create Document

Model Deployment

- Deploying the model
- Scalability
- Maintenance
- Feedback
- Security

Model Deployment:

- **Integration** web application where users can input apartment features to get price predictions.
- **Utilization** containerization tools like Docker to package the model along with its dependencies, ensuring consistency across different deployment environments.
- **Hosting** platforms like AWS or Google Cloud Platform to deploy the model, making it accessible to users.
- Implement **versioning** to track changes to the model over time and logging to monitor its performance in production.
- Set up **monitoring** and alerting systems to detect any issues with the deployed model, such as unexpected fluctuations in predicted prices.

Scalability& Maintenance

- **Monitoring** the performance of the deployed model to identify scalability bottlenecks, such as increased latency during periods of high demand.
- **Optimizing** the model architecture and code for scalability, ensuring that it can handle a growing user base and increasing data volumes without compromising performance.
- Regularly **retraining** the model with new data to ensure that it stays up-to-date and continues to provide accurate price predictions.
- Perform routine **maintenance** tasks, such as updating dependencies and optimizing model hyperparameters, to ensure optimal performance over time.

Feedback &Security

- Gather feedback from users on the accuracy of the model's price predictions and any additional features they would like to see.
- Incorporate user feedback into the model training process to improve its accuracy and relevance over time.
- Ensure compliance with relevant regulations, such as data privacy laws, when handling and storing user data.
- Educate team members on best practices for data security and compliance to minimize the risk of data breaches or regulatory violations.

Group 6

Saksham Gupta - 21BCE0161
Arnav Mahani - 21BCE0526
Saniya Vijayvargiya - 21BCE0808
Isha Nevatia - 21BCE2303
Khushi Teli - 21BCE2755
Rohan Khatua - 21BCE3982
Samikshha K - 21BDS0258