Phase-3 Submission Report

Student Name: Alfiya Amreen T

Register Number: 510623104007

Institution: C. Abdul Hakeem College of Engineering and Technology

Department: Computer Science and Engineering

Date of Submission: 09-05-2025

GitHub Repository Link:

https://github.com/Alfiya586/house-price-prediction-3.git

---

1. Problem Statement

Accurate prediction of house prices is a crucial challenge in the real estate industry due to the influence of numerous factors

such as location, size, amenities, and current market dynamics. Traditional models often fail to handle the non-linearity and complex interactions present in housing data, leading to suboptimal pricing insights. This project addresses the issue by applying advanced supervised regression techniques to build a robust predictive model. The aim is to support buyers, sellers, and investors with data-driven insights, thereby enhancing real estate decision-making and pricing strategies.

2. Abstract

This project focuses on predicting housing prices using smart regression models by leveraging the Ames Housing Dataset. The objective is to overcome the limitations of traditional pricing methods that often miss complex relationships in data. The dataset underwent preprocessing, exploratory data analysis, and feature engineering to improve model quality. Various models like Linear Regression, Random Forest, and XGBoost were implemented and evaluated using RMSE, MAE, and R^2-score. Among these, XGBoost provided the most accurate predictions. The outcome is a predictive system capable of estimating house prices, assisting stakeholders in making informed real estate decisions.

3. System Requirements

Hardware: Minimum 4GB RAM, Intel i3 Processor or above

Software:

Python 3.10+

Jupyter Notebook / Google Colab

Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost, plotly

4. Objectives

Analyze influential features like area, number of rooms, amenities, and location

Preprocess and clean the dataset for high-quality input
Engineer relevant features to capture hidden patterns

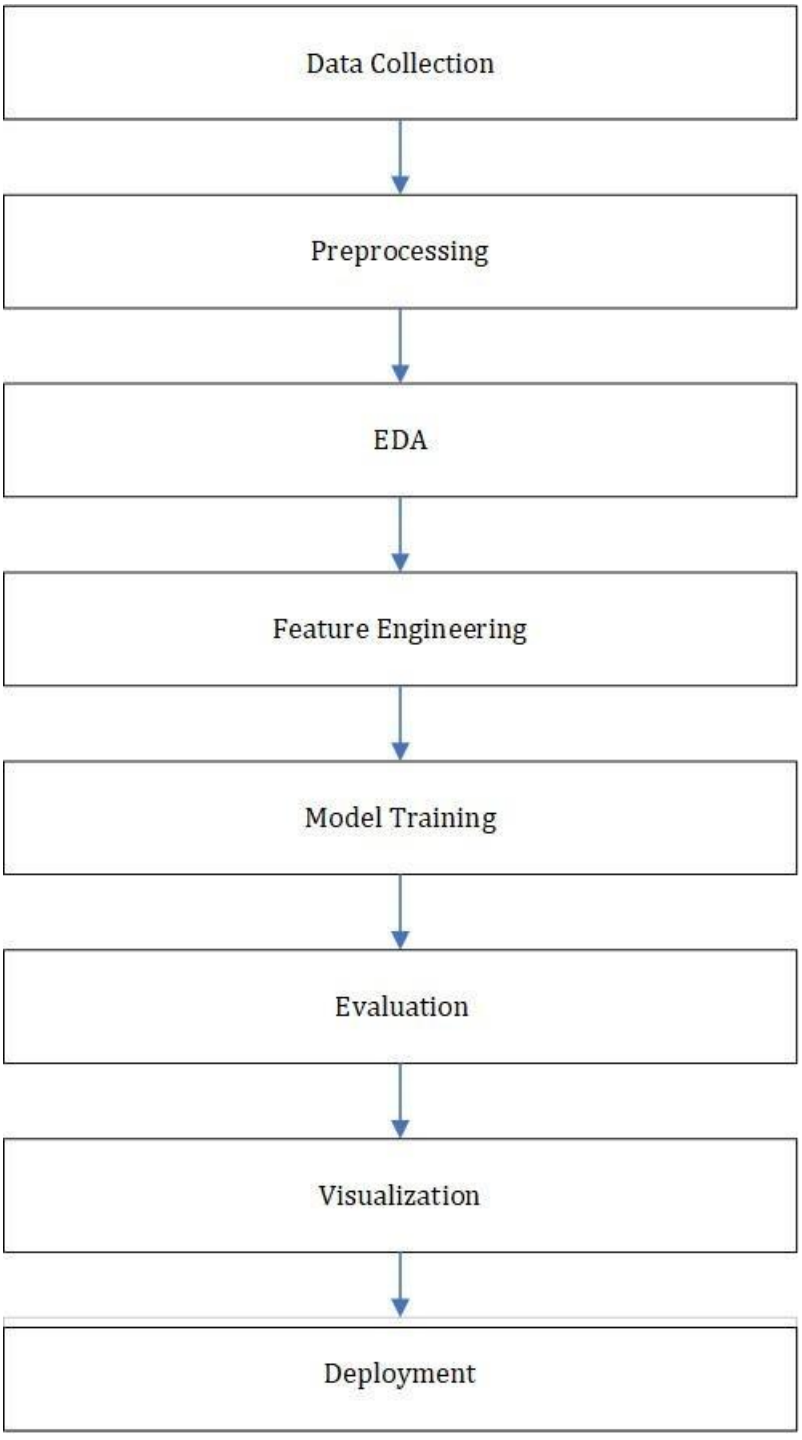
Develop and compare models: Linear Regression, Random Forest, XGBoost

Evaluate models using metrics like RMSE, MAE, and R^2

Identify the best model and present key insights using visualizations

---

5. Flowchart of Project Workflow



6. Dataset Description

Dataset Name: Ames Housing Dataset

Source: Kaggle (https://www.kaggle.com/datasets)

Type: Public, Structured

Size: ~2,930 records with ~80 features

Target Variable: SalePrice

	A	B	C	D	E	F
1	S.No	property_id	location_id	page_url	property_type	price
2	0	237062	3325	https://www.zameen.com/Property/g_10_g_10_2_ground_floor_corner_apartment_with_green_lawn_for_sale-237062-3325-1.html	Flat	10000000
3	1	346905	3236	https://www.zameen.com/Property/e_11_2_services_society_flat_available_for_sale-346905-3236-1.html	Flat	6900000
4	2	386513	764	https://www.zameen.com/Property/islamabad_g_15_house_is_available_for_sale-386513-764-1.html	House	16500000
5	3	656161	340	https://www.zameen.com/Property/islamabad_bani_gala_a_rare_minimalist_concept_in_a_quiet_location-656161-340-1.html	House	43500000
6	4	841645	3226	https://www.zameen.com/Property/dha_valley_dha_homes_islamabad_dha_valley_8_marla_home_for_sale-841645-3226-1.html	House	7000000
7	5	850762	3390	https://www.zameen.com/Property/ghauri_town_ghauri_town_phase_1_house_is_available_for_sale_in_ghauri_town_phase_1-850762-3390-1.html	House	34500000

---

7. Data Preprocessing

Handled missing values using mean/mode imputation

Removed duplicates and standardized column formats

Treated outliers using IQR method

Encoded categorical variables using One-Hot Encoding

Scaled features using Min-Max and Standard Scalers

```
RMSE: 25494370.485742256
R2 Score: -0.0608931929993044
```

	property_type	price	location	city	province_name
0	Flat	10000000	G-10	Islamabad	Islamabad Capital
1	Flat	6900000	E-11	Islamabad	Islamabad Capital
2	House	16500000	G-15	Islamabad	Islamabad Capital
3	House	43500000	Bani Gala	Islamabad	Islamabad Capital
4	House	7000000	DHA Defence	Islamabad	Islamabad Capital
5	House	34500000	Ghauri Town	Islamabad	Islamabad Capital
6	House	27000000	Korang Town	Islamabad	Islamabad Capital
7	Flat	7800000	E-11	Islamabad	Islamabad Capital
8	House	50000000	DHA Defence	Islamabad	Islamabad Capital
9	Penthouse	40000000	F-11	Islamabad	Islamabad Capital
10	Flat	35000000	Diplomatic Enclave	Islamabad	Islamabad Capital
11	Flat	48000000	Diplomatic Enclave	Islamabad	Islamabad Capital
12	House	400000000	F-6	Islamabad	Islamabad Capital
13	Flat	13500000	DHA Defence	Islamabad	Islamabad Capital
14	Flat	3600000	E-11	Islamabad	Islamabad Capital
15	Flat	5000000	E-11	Islamabad	Islamabad Capital
16	House	19000000	DHA Defence	Islamabad	Islamabad Capital
17	House	80000000	DHA Defence	Islamabad	Islamabad Capital
18	House	26900000	B-17	Islamabad	Islamabad Capital
19	Flat	1750000	PWD Housing Scheme	Islamabad	Islamabad Capital
20	House	55000000	G-11	Islamabad	Islamabad Capital
21	House	4500000	Bhara kahu	Islamabad	Islamabad Capital
22	Farm House	88500000	Bani Gala	Islamabad	Islamabad Capital
23	Flat	47000000	Diplomatic Enclave	Islamabad	Islamabad Capital
24	House	4500000	Garden Town	Islamabad	Islamabad Capital
25	House	6800000	Koral Town	Islamabad	Islamabad Capital
26	House	20000000	Soan Garden	Islamabad	Islamabad Capital
27	Flat	19400000	Blue Area	Islamabad	Islamabad Capital
28	House	100000000	F-6	Islamabad	Islamabad Capital
29	Flat	8000000	G-11	Islamabad	Islamabad Capital
30	Flat	6300000	E-11	Islamabad	Islamabad Capital

---

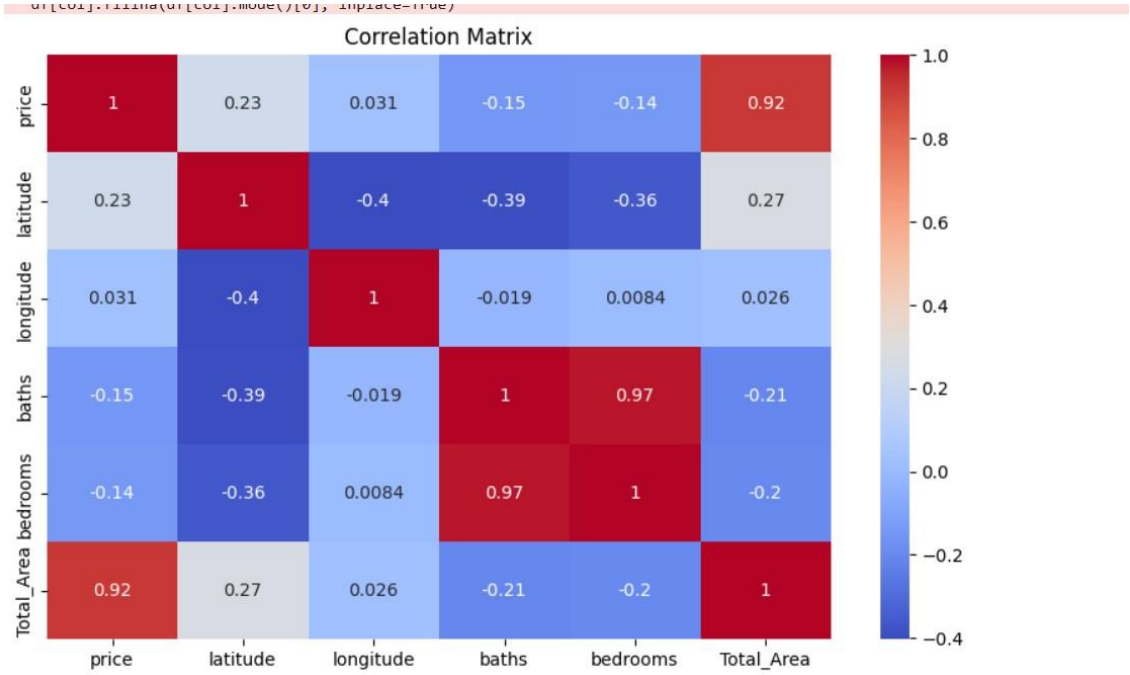
8. Exploratory Data Analysis (EDA)

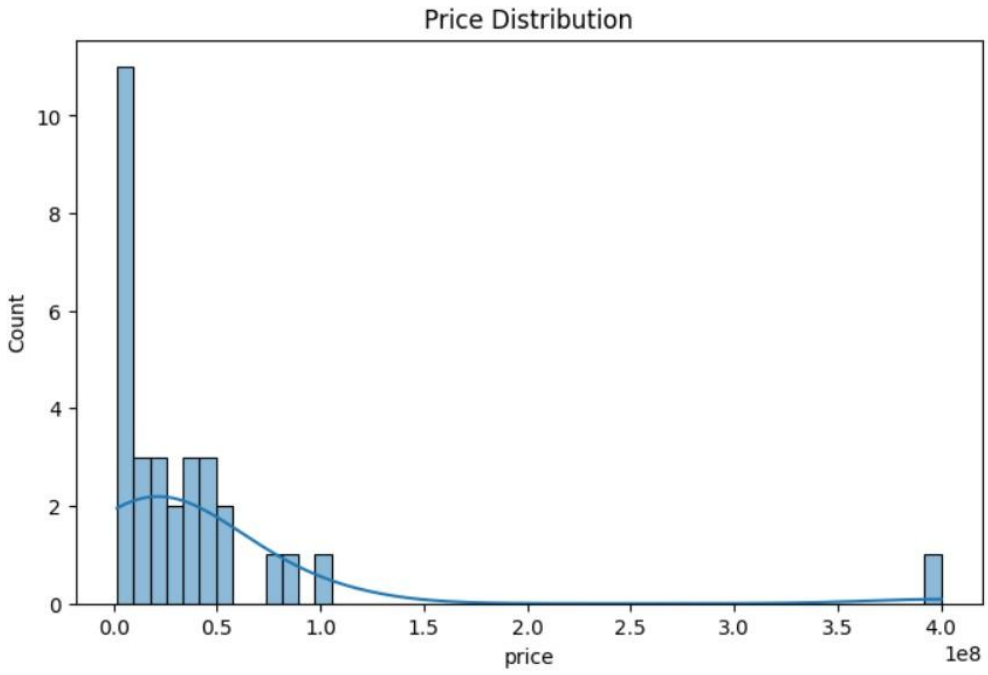
Univariate Analysis: Histograms and boxplots showed skewed distributions in price and area

Bivariate Analysis: Strong correlation between GrLivArea and SalePrice

Multivariate Analysis: Heatmaps showed multicollinearity; scatter plots revealed non-linear trends

Insights: OverallQual, GrLivArea, and Neighborhood are top influencing factors





---

9. Feature Engineering

Created: HouseAge = YearSold - YearBuilt, PricePerSqFt = SalePrice / TotalSqFeet

Encoded categorical features

Applied log transformation to reduce skewness

Removed low-importance features with high null values

---

10. Model Building

Models Used:

Linear Regression

Ridge & Lasso Regression

Decision Tree Regressor

Random Forest Regressor

XGBoost Regressor

Evaluation Metrics: RMSE, MAE, R^2-score

Cross-Validation: 10-Fold CV

Best Model: XGBoost due to handling of non-linearity and feature interactions

RMSE: 25494370.485742256
R2 Score: -0.0608931929993044

	property_type	price	location	city	province_name
0	Flat	10000000	G-10	Islamabad	Islamabad Capital
1	Flat	6900000	E-11	Islamabad	Islamabad Capital
2	House	16500000	G-15	Islamabad	Islamabad Capital
3	House	43500000	Bani Gala	Islamabad	Islamabad Capital
4	House	7000000	DHA Defence	Islamabad	Islamabad Capital
5	House	34500000	Ghauri Town	Islamabad	Islamabad Capital
6	House	27000000	Korang Town	Islamabad	Islamabad Capital
7	Flat	7800000	E-11	Islamabad	Islamabad Capital
8	House	50000000	DHA Defence	Islamabad	Islamabad Capital
9	Penthouse	40000000	F-11	Islamabad	Islamabad Capital
10	Flat	35000000	Diplomatic Enclave	Islamabad	Islamabad Capital
11	Flat	48000000	Diplomatic Enclave	Islamabad	Islamabad Capital
12	House	400000000	F-6	Islamabad	Islamabad Capital
13	Flat	13500000	DHA Defence	Islamabad	Islamabad Capital
14	Flat	3600000	E-11	Islamabad	Islamabad Capital
15	Flat	5000000	E-11	Islamabad	Islamabad Capital
16	House	19000000	DHA Defence	Islamabad	Islamabad Capital
17	House	80000000	DHA Defence	Islamabad	Islamabad Capital
18	House	26900000	B-17	Islamabad	Islamabad Capital
19	Flat	1750000	PWD Housing Scheme	Islamabad	Islamabad Capital
20	House	55000000	G-11	Islamabad	Islamabad Capital
21	House	4500000	Bhara kahu	Islamabad	Islamabad Capital
22	Farm House	88500000	Bani Gala	Islamabad	Islamabad Capital
23	Flat	47000000	Diplomatic Enclave	Islamabad	Islamabad Capital
24	House	4500000	Garden Town	Islamabad	Islamabad Capital
25	House	6800000	Koral Town	Islamabad	Islamabad Capital
26	House	20000000	Soan Garden	Islamabad	Islamabad Capital
27	Flat	19400000	Blue Area	Islamabad	Islamabad Capital
28	House	100000000	F-6	Islamabad	Islamabad Capital
29	Flat	8000000	G-11	Islamabad	Islamabad Capital
30	Flat	6300000	E-11	Islamabad	Islamabad Capital

	latitude	longitude	baths	purpose	bedrooms	Total_Area
0	33.679890	73.012640	2	For Sale	2	1089.004
1	33.700993	72.971492	3	For Sale	3	15246.056
2	33.631486	72.926559	6	For Sale	5	2178.008
3	33.707573	73.151199	4	For Sale	4	10890.000
4	33.492591	73.301339	3	For Sale	3	2178.008
5	33.623947	73.126588	8	For Sale	8	87120.000
6	33.579034	73.139591	8	For Sale	8	5445.000
7	33.698244	72.984238	2	For Sale	2	16879.562
8	33.540894	73.095732	7	For Sale	7	5445.000
9	33.679211	72.988787	5	For Sale	5	5445.000
10	33.728873	73.119628	3	For Sale	3	19329.821
11	33.728873	73.119628	2	For Sale	2	21235.578
12	33.731532	73.065696	0	For Sale	0	245025.000
13	33.538087	73.164536	5	For Sale	3	2722.510
14	33.698137	72.978215	1	For Sale	1	8439.781
15	33.698137	72.978215	2	For Sale	2	1089.004
16	33.508481	73.091826	3	For Sale	3	2722.510
17	33.541728	73.094103	7	For Sale	7	10890.000
18	33.694495	72.826653	6	For Sale	6	5445.000
19	33.570792	73.145256	0	For Sale	0	4083.765
20	33.671640	72.991655	7	For Sale	6	3811.514
21	33.737402	73.179159	3	For Sale	3	1361.255
22	33.713488	73.162680	3	For Sale	3	32670.000
23	33.728873	73.119628	2	For Sale	3	22869.084
24	33.636132	73.113921	4	For Sale	4	12795.797
25	33.602038	73.141966	4	For Sale	4	1089.004
26	33.569648	73.151522	5	For Sale	6	3267.012
27	33.713845	73.060970	1	For Sale	1	11706.793
28	33.724020	73.074524	5	For Sale	5	48460.678
29	33.675604	73.000367	2	For Sale	2	18240.817
30	33.698137	72.978215	3	For Sale	3	14429.303

---

11. Model Evaluation

Metrics:

RMSE: Lowest for XGBoost

MAE: Moderate error margin

R^2-score: ~0.91 for XGBoost

Visuals:

Residual plots

Model comparison bar chart

SHAP values (optional)

```
RMSE: 25494370.485742256
R2 Score: -0.0608931929993044
```

12. Source Code

```
# 1. Import Libraries import pandas as pd import
numpy as np import matplotlib.pyplot as plt import
seaborn as sns from sklearn.model_selection import
train_test_split from sklearn.preprocessing import
StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline from sklearn.impute
import SimpleImputer from sklearn.ensemble import
RandomForestRegressor from sklearn.metrics import
mean_squared_error, r2_score


# 2. Load Dataset
df = pd.read_excel("Forcasting house datasets.xlsx",
sheet_name="Sheet1")


# 3. Data Cleaning # Drop
unnecessary columns
df.drop(columns=['S.No', 'property_id', 'location_id',
'page_url', 'agency', 'agent'], inplace=True)


# Drop rows with missing target variable
```

```
df = df.dropna(subset=['price'])


# Fill missing values
num_cols = df.select_dtypes(include=['float64',
'int64']).columns
cat_cols = df.select_dtypes(include=['object']).columns

for col in num_cols:
    df[col].fillna(df[col].median(), inplace=True)

for col in cat_cols:
    df[col].fillna(df[col].mode()[0], inplace=True)



# 4. EDA (Exploratory Data Analysis)
# Plot correlations
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
# Plot price distribution plt.figure(figsize=(8,
5))
sns.histplot(df['price'], bins=50, kde=True)
plt.title('Price Distribution')
plt.show()



# 5. Feature Engineering X
= df.drop('price', axis=1)
y = df['price']


# Separate features by type
numerical_features = X.select_dtypes(include=['int64',
'float64']).columns.tolist() categorical_features
=
X.select_dtypes(include=['object']).columns.tolist()



# 6. Preprocessing Pipeline numeric_transformer
= Pipeline([
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline([
```

```
    ('imputer', SimpleImputer(strategy='most_frequent')),     ('onehot',
OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer([
    ('num', numeric_transformer, numerical_features),
    ('cat', categorical_transformer, categorical_features)
])



# 7. Modeling model
= Pipeline([
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=100,
random_state=42))
])


# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)


# Train the model
model.fit(X_train, y_train)
```

```
# Predict and Evaluate y_pred
= model.predict(X_test)
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R2 Score:", r2_score(y_test, y_pred)) print(df)
```

13.Future Scope

Implement real-time price prediction using a Streamlit web app

Integrate more external datasets for enhanced accuracy

Use deep learning models (e.g., neural networks) for comparison

---

15.Team Members and Roles

1.Jamal Be Fathima [510623104033]

Role: Team Lead & Model Building

Task: Led the project and implemented all regression models

2.Alfiya Amreen. T [510623104007]

Role: Data Collection & Preprocessing

Task: Handled dataset sourcing and cleaning

3.Farah Thasleem. S [510623104022]

Role: EDA & Feature Engineering

Task: Conducted EDA and created new features

4.Jansi Rani. K. S [510623104034]

Role: Model Evaluation & Report Preparation

Task: Evaluated models and compiled documentation
