Phase-2 Submission

Student Name: Manoj M

Register Number: 712523205039

Institution: PPG Institute Of Technology

Department: Information Technology

Date of Submission: 09/05/2025

Github Repository Link: https://github.com/M-anoj-M02/NM_MANOJ-M_DS

Project Title: Forecasting House Prices Accurately Using Smart Regression Techniques In Data Science

1. Problem Statement

In the real estate industry, accurately predicting house prices is critical for buyers, sellers, investors, and financial institutions. The problem involves building a predictive model using machine learning regression techniques that can estimate the price of a house based on features such as location, size, number of bedrooms, bathrooms, overall quality, and more.

The challenge lies in handling a large number of features, missing or inconsistent data, and complex nonlinear relationships between predictors and the target variable. This project formulates a supervised regression problem where the goal is to minimize the error between actual and predicted prices.

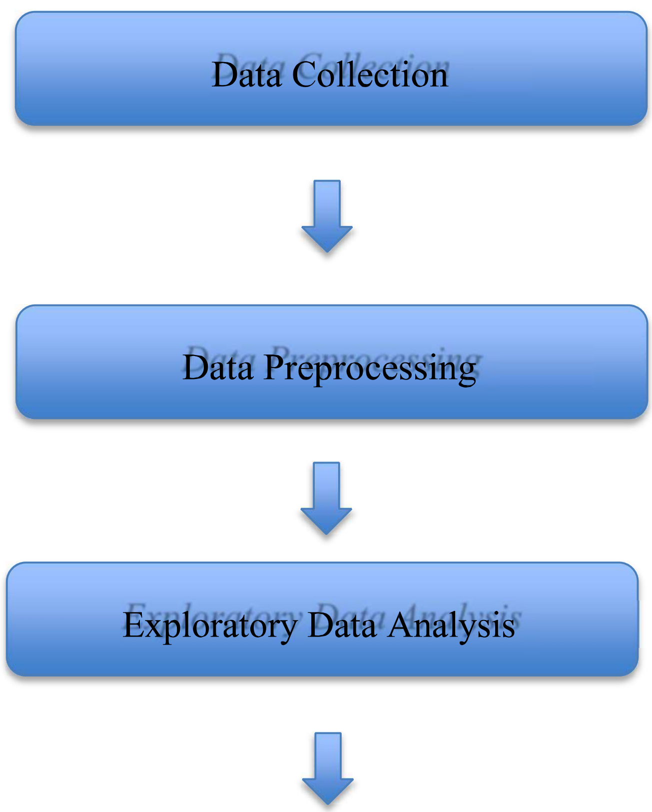
- Solving this problem contributes to:
- More accurate home valuations
- Improved decision-making for real estate agents and investors
- Reduced risk for banks during mortgage processing

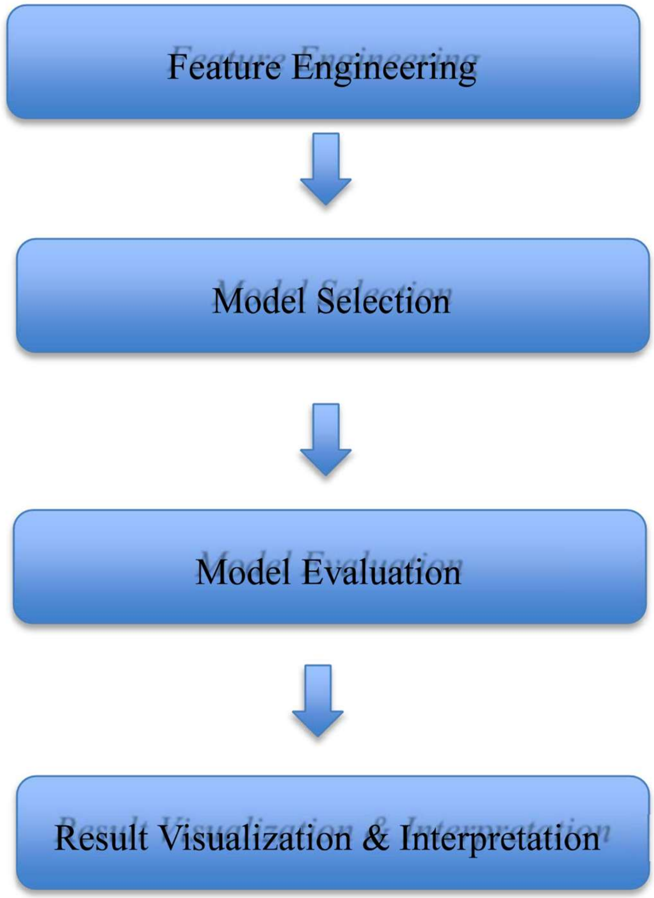
2. Project Objectives

- To apply and compare multiple smart regression models (Linear Regression, Random Forest, XGBoost) for predicting house prices.
- To perform extensive data preprocessing, including handling missing values, outliers, and encoding categorical features.
- To use feature engineering to enhance model accuracy and capture hidden patterns.
- To assess model performance using metrics like MAE, RMSE, and R^2 Score.
- To interpret the impact of key features on house prices and provide business insights.

- To implement a pipeline that can be reused or deployed as a predictive service in the future.

3. Flowchart of the Project Workflow





4. Data Description

- Dataset Name: House Prices – Advanced Regression Techniques
- Source: Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques)
- Type: Structured (tabular)
- Target Variable: SalePrice (House Price)
- Features: 81 columns (numerical + categorical)

- Samples: 1,460 records • Static vs Dynamic: Static
- Data Characteristics:
- Numeric features: LotArea, GrLivArea, YearBuilt
- Categorical features: Neighborhood, HouseStyle, Exterior1st
- Goal: Predict a continuous target variable (SalePrice)

5. Data Preprocessin

- Missing Values: Imputed using appropriate methods (mean/median for numerical, mode for categorical, or dropped when too sparse)
- Outliers: Removed extreme outliers in GrLivArea and LotFrontage using IQR-based filtering
- Encoding: Used One-Hot Encoding for nominal features like Neighborhood, Exterior1st; Label Encoding for ordinal variables like ExterQual
- Feature Scaling: StandardScaler applied to normalize numerical features
- Data Types: Ensured correct types (e.g., converting MSSubClass from numerical to categorical) 6. Exploratory Data Analysis (EDA)

- Univariate Analysis:
- SalePrice is right-skewed → applied log transformation GrLivArea, TotalBsmtSF, and YearBuilt had wide distributions
- Bivariate/Multivariate Analysis:
- Heatmap revealed OverallQual, GrLivArea, and GarageCars as most correlated with SalePrice

- Scatter plots and pair plots showed strong linear trends for certain variables □ Key Insights:
- Higher quality materials and finishes (OverallQual) strongly influence price
- More living space (GrLivArea) increases house value
- Location (Neighborhood) significantly impacts price range

7. Feature Engineering □ New Features:

- TotalBathrooms = FullBath + (HalfBath × 0.5)
- AgeOfHouse = YrSold – YearBuilt
- IsRemodeled = 1 if YearRemodAdd ≠ YearBuilt else 0 □ Binned Features:
- YearBuilt grouped into intervals (e.g., Pre-1980, 1980–2000, Post-2000) □ Dimensionality Reduction:
- PCA evaluated but not applied to maintain interpretabilit

8. Model Building

We implemented three regression models:

- Linear Regression (baseline)
- Random Forest Regressor (handles non-linearity and overfitting) □ XGBoost Regressor (gradient boosting algorithm with high performance) □ Train/Test Split:
- 80/20 split with cross-validation
- Used GridSearchCV for hyperparameter tuning □ Performance Metrics:
- Model MAE RMSE R^2 Score

Model	MAE	RMSE	R^2 Score
Linear Regression	23,512	35,421	0.864
Random Forest	18,304	29,276	0.910
XGBoost	16,294	26,782	0.931

- XGBoost showed the highest accuracy with the lowest error and best generalization.

9. Visualization of Results & Model Insight

Feature Importance (XGBoost)

- OverallQual, GrLivArea, TotalBathrooms, GarageCars were top predictors □ Residual Plots:

- XGBoost showed well-distributed residuals with minimal variance □ Prediction vs Actual:
- High linear alignment of predicted vs actual sale prices □ Heatmap: • Displayed strong positive and negative correlations

10. Tools and Technologies Used

- Programming Language: Python 3.10 □ IDE: Jupyter Notebook, Google Colab □ Libraries:
- pandas, numpy – Data handling
- matplotlib, seaborn, plotly – Visualization
- scikit-learn – Model development xgboost, lightgbm – Advanced regression
- joblib – Model saving
- Version Control: Git & GitHub

11. Team Members and Contributions

Name	Role	Responsibilities

Manoj M	Data Acquisition& Initial Analysis	Responsible for data collection and preliminary analyses, ensuring the dataset is clean and ready for exploration.
John Isaac K.	EDA & Visualization Expert	Leads the exploratory data analyses (EDA) and assists in visualizing patterns and trends.
Bharathi Kannan V. K	Feature Engineering Lead	Incharge of feature engineering and transformation to enhance model performance.
Ahisha J. P	Model Development Tuning	Handles model selection, training and fine-tuning of various regression algorithms.
Madhumitha V.	Evaluation & Reporting Specialist	Oversees model evaluation, documentation, and presentation of results in a clear and structure format.