Phase-1

Student Name : A.Kaviya

Register Number : 410723104031

Institution : Dhanalakshmi college of engineering

Department : Computer Science And Engineering

Date of Submission : 30.04.2025

1.Problem Statement

Forecasting house prices accurately using smart regression techniques in datascience.

2.Objectives of the Project

PROJECT OBJECTIVE:

To develop an accurate and reliable regression model for predicting house price using smart data science techniques. The goal is to enhance prediction accuracy through feature engineering, model comparison, and optimization.

Key Outcomes:

1. A trained regression model capable of accurately predicting house prices.
2. Identification of the most influential features affecting price.
3. Comparative performance analysis of different regression techniques.
4. Improved model accuracy through tuning and validation.
5. A deployable and interpretable solution for real-world use.

3.Scope of the Project

The primary objective of this project is to develop a data-driven system capable of accurately predicting house prices using smart regression

techniques in data science. The system leverages historical housing data and machine learning algorithms to model and forecast housing prices with high accuracy.

1. Data Collection and Exploration

Gather real-world housing data from public sources such as the Ames Housing dataset or Boston Housing dataset.

Perform Exploratory Data Analysis (EDA) to uncover trends,correlations, and patterns between features and housing prices.

2. Data Preprocessing

Clean the dataset by handling missing values, removing outliers, and correcting anomalies.

Encode categorical variables and normalize numerical data.

3. Model Development

Develop and compare a variety of regression models, including:

Linear Regression

Ridge and Lasso Regression

Decision Tree Regression

Random Forest Regressor

Gradient Boosting Models (e.g., XGBoost)

4. Model Evaluation and Optimization

Evaluate model performance using key metrics such as:

Mean Absolute Error (MAE)

Root Mean Squared Error (RMSE)

R-squared (R^2) score

5.Deployment (Optional)

Develop an interactive web-based tool using Flask or Streamlit.

Allow users to input property details and receive real-time price predictions.

6.Applications

Help homebuyers and investors assess property values accurately.

Assist real estate agents in pricing strategies.

Provide financial institutions with a tool for mortgage risk assessment and loan approvals.

4.Data Sources

Source:

Kaggle https://www.kaggle.com/datasets/thomasnibb/amsterdam-house-price-prediction ,and it is public dataset and it is a dynamic dataset.

Type: Public data.

Access: Downloadable via API or library functions.

Nature: Static for training and experimentation;can be extended to dynamic updates for demo.

5.High-Level Methodology

1. Data Collection:

The dataset was sourced from Kaggle, containing detailed information on house features and sale prices.

2. Data Cleaning:

We used key functions such as:

drop_duplicates() – to remove any repeated records.

isnull().sum() – to identify and assess missing values.

info() – to understand the structure and types of data.

Additional steps included handling outliers and encoding categorical variables as needed.

3. Model Building:

Libraries imported: pandas, NumPy, matplotlib, seaborn, scikit-learn, XGBoost, etc.

Steps followed:

Load and explore the dataset

Data preprocessing (scaling, encoding, feature selection)

Train-test split

Training multiple regression models (Linear, Ridge, Lasso, Random Forest, XGBoost)

Selecting the best-performing model

4. Model Evaluation:

Models were evaluated using the following metrics:

Mean Absolute Error (MAE)

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

R^2 Score

Adjusted R^2 Score

5. Visualization & Interpretation:

We used visual tools to better understand the data and model results:

Boxplots – for outlier detection

Histograms – to study feature distributions

Scatter Plots – to analyze feature relationships

Feature Correlation Heatmaps – to select influential predictors

HvPlot – for interactive visualizations

6. Deployment:

Planned deployment method: Web Application using Flask or Streamlit.

The trained model will be exposed via a simple UI for predicting house prices.

Hosting options considered: Heroku, GitHub Pages, or local deployment via Flask/Streamlit.

6.Tools and Technologies

- **Programming Language** – python.
- **Notebook/IDE** – Google Colab, Jupyter Notebook.
- **Libraries** – pandas, numpy, seaborn, matplotlib, scikit-learn, ,yfinance,TensorFlow/Keras.

- **Optional Tools for Deployment** – Streamlit, Flask, Gradio, FastAPI

7.Team Members and Roles

S.NO	NAMES	ROLES	RESPONSIBILITY
1	D.N.Abarna	Leader	Visualization and Interpretation
2	G.S.Harini	Member	Data collection, data cleaning
3	A.Kaviya	Member	Model Building and Feature Engineering
4	S.Keerthika	Member	Model evaluation,Training and Testing