Phase-1 Submission Template

Student Name: Roy bennar.s

Register Number: 421223104069

Institution: karpaga vinayaga college of engineering and technology

Department: Computer science engineering

Date of Submission: 24.4.25

1.Problem Statement

In the real estate market, accurately predicting house prices is a critical yet complex challenge due to the influence of multiple dynamic variables such as location, size, neighborhood, market trends, and economic conditions. Traditional pricing methods often lack precision, leading to inefficient investments and decision-making. This project aims to develop a smart, data-driven regression model to forecast house prices with higher accuracy using historical data and advanced regression techniques.

2.Objectives of the Project

To build a predictive model that accurately estimates the price of a house based on various features.

To explore and compare multiple regression techniques (linear, polynomial, regularized, ensemble-based).

To identify the most influential features affecting house prices.

To evaluate and validate the model using appropriate performance metrics.

3.Scope of the Project

Data Collection and Preprocessing:

Feature Engineering

Model Building:

4.Data Sources

1. Kaggle Datasets

URL: https://www.kaggle.com/datasets

Example: "House Prices - Advanced Regression Techniques"

Description: Contains housing data from Ames, Iowa with 79 explanatory variables.

2. UCI Machine Learning Repository

URL: https://archive.ics.uci.edu/ml/datasets.php

Example: Boston Housing Dataset

Description: A classic dataset for regression tasks with 506 entries and 13 features.

3. Zillow Research

URL: https://www.zillow.com/research/data/

Description: Offers real estate data including home values, rent prices, and market trends (mainly for U.S. cities).

4. Open Government Data Portals

Example: data.gov (USA), data.gov.in (India)

Description: Provides public datasets including real estate transactions and property valuations.

5. Property Websites (via Web Scraping or APIs)

Examples: MagicBricks, 99acres (India), Realtor.com, Redfin (U.S.)

5.High-Level Methodology

1. Problem Understanding & Goal Definition

Define the objective: Predict house prices accurately using regression techniques.

Understand business context and the target variable (house price).

2. Data Collection

Gather data from sources like Kaggle, UCI, or Zillow.

Ensure the dataset includes features such as location, area, number of rooms, year built, etc.

3. Data Preprocessing

Handle missing values and outliers.

Convert categorical variables to numerical (e.g., one-hot encoding).

Normalize or scale data if required.

4. Exploratory Data Analysis (EDA)

Visualize distributions, relationships, and correlations.

Identify key variables affecting house prices.

5. Feature Engineering

Create new relevant features (e.g., price per square foot, age of house).

Select the most influential features using feature importance techniques.

6. Model Selection & Training

Apply multiple regression models:

Linear Regression

Ridge/Lasso Regression

Random Forest, Gradient Boosting, XGBoost, LightGBM

Tune hyperparameters using Grid Search or Random Search.

7. Model Evaluation

Use metrics like RMSE, MAE, and R^2 score to evaluate model performance.

Perform cross-validation to ensure model stability.

8. Model Comparison & Selection

Compare all models and choose the best-performing one based on evaluation metrics.

9. Deployment (Optional)

Develop a simple interface (web/app) to input house features and predict price.

Use tools like Flask or Streamlit for deployment.

10. Documentation & Reporting

Document the workflow, findings, and conclusions.

Present visualizations and model insights clearly

6.Tools and Technologies

- **Programming Language** – *python*

- **Notebook/IDE** – *Jupyter notebook, visual studio code*

- **Libraries** – *pandas, numpy, seaborn, matplotlib, scikit-learn, TensorFlow*

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

7.Team Members and Roles

1 V Thiruselvam -data collection and preprocessing
2. Rakesh C P-data exploration and development
3 Roy bennar S model evaluation and prediction
4.Simon benean D- deployment and presentation