

PREDICTING HOUSE PRICE USING MACHINE LEARNING

TEAM MEMBER

810621104007: BHAVANI.S

PHASE 2 SUBMISSION DOCUMENT

PROJECT: House price prediction



INTRODUCTION:

- ✓ The real market is one of the most dynamics and lucrative sectors, with house prices constantly fluctuating based on various factors such as location, size, amenities, and economics conditions. Accurately predicting house prices is crucial for both buyers and sellers, as it can help make informed decisions regarding buying, selling, or investing in properties.
- ✓ Traditional linear regression models are often employed for house price prediction. However, they may not capture complex relationships between predictors and the target variable, leading to suboptimal predictions. In this project, we will explore advanced regression

techniques to enhance the accuracy and robustness of house price prediction models.

- ✓ Briefly introduce the real estate market and the importance of accurate house price prediction. Highlight the limitations of traditional linear regression models in capturing complex relationship.
- ✓ Emphasize the need for advanced regression techniques like Gradient Boosting and XGBoost to enhance prediction accuracy.

CONTENT FOR PROJECT PHASE 2:

Consider exploring advanced regression techniques like Gradient Boosting or XGBoost for improved prediction accuracy.

DATA SOURCE:

A good data source for house price prediction using machine learning should be accurate, complete, covering the geographic area of interest, accessible.

Dataset Link:(<https://www.kaggle.com/datasets/vedavyasv/usa-housing>)

<u>Avg. Area Income</u>	<u>Avg. Area House Age</u>	<u>Avg. Area Number of Rooms</u>	<u>Avg. Area Number of Bedrooms</u>	<u>Area Population</u>	<u>Price</u>	<u>Address</u>
<u>79545.46</u>	<u>5.682861</u>	<u>7.009188</u>	<u>4.09</u>	<u>23086.8</u>	<u>1059034</u>	<u>208 Michael Ferry Apt. 674 Laurabury.</u>

						<u>NE 37010-5101</u>
<u>79248.64</u>	<u>6.0029</u>	<u>6.730821</u>	<u>3.09</u>	<u>40173.07</u>	<u>1505891</u>	<u>188 Johnson Views Suite 079 Lake Kathleen, CA 48958</u>
<u>61287.07</u>	<u>5.86589</u>	<u>8.512727</u>	<u>5.13</u>	<u>36882.16</u>	<u>1058988</u>	<u>9127 Elizabeth Stravenue Danieltown, WI 06482-3489</u>
<u>63345.24</u>	<u>7.188236</u>	<u>5.586729</u>	<u>3.26</u>	<u>34310.24</u>	<u>1260617</u>	<u>USS Barnett FPO AP 44820</u>
<u>59982.2</u>	<u>5.040555</u>	<u>7.839388</u>	<u>4.23</u>	<u>26354.11</u>	<u>630943.5</u>	<u>USNS Raymond FPO AE 09386</u>
<u>80175.75</u>	<u>4.988408</u>	<u>6.104512</u>	<u>4.04</u>	<u>26748.43</u>	<u>1068138</u>	<u>06039 Jennifer Islands Apt. 443 Tracyport, KS 16077</u>
<u>64698.46</u>	<u>6.025336</u>	<u>8.14776</u>	<u>3.41</u>	<u>60828.25</u>	<u>1502056</u>	<u>4759 Daniel</u>

						<u>Shoals</u> <u>Suite 442</u> <u>Nguyenbur</u> <u>gh, CO</u> <u>20247</u>
<u>78394.</u> <u>34</u>	<u>6.9897</u> <u>8</u>	<u>6.6204</u> <u>78</u>	<u>2.42</u>	<u>36516.3</u> <u>6</u>	<u>157393</u> <u>7</u>	<u>972 Joyce</u> <u>Viaduct</u> <u>Lake</u> <u>William, TN</u> <u>17778-</u> <u>6483</u>
<u>59927.</u> <u>66</u>	<u>5.3621</u> <u>26</u>	<u>6.3931</u> <u>21</u>	<u>2.3</u>	<u>29387.4</u>	<u>798869</u> <u>.5</u>	<u>USS Gilbert</u> <u>FPO AA</u> <u>20957</u>

DATA COLLECTION AND PREPROCESSING:

- ✓ Gather a dataset that contains the variable of interest, including the dependent variables (features).
- ✓ Ensure that your dataset is clean, free from missing values, and has been preprocessed.

FEATURE ENGINEERING:

- ✓ Create new features or transform existing ones to improve the models predictive power. This can involve techniques like one-hot encoding, features scaling, or creating interaction terms.

ADVANCED REGRESSION TECHNIQUES:

- ✓ **Ridge Regression**
- ✓ **Lasso Regression**
- ✓ **Elastic Net Regression**

- ✓ **Support Vector Regression (SVR)**
- ✓ **Decision Tree Regression**
- ✓ **Gradient Boosting (eg, XGBoost)**

MODEL SELECTION AND EVALUATION:

- ✓ Split your dataset into training and testing sets to evaluate model performance.
- ✓ Train various regression models on the training data and access their performance metrics(e.g, Mean squared error, R-squared, MAE etc.

MODEL INTERPRETATION:

- ✓ If interpretability is essential, analyze feature importance or coefficients to understand the impacts of each feature on the target variable.

DEPLOYMENT AND PREDICTION:

- ✓ Once you have a satisfactory regression model, deploy it in a production environment to make prediction on new data.

Gradient Boosting Regression:

Gradient Boosting Regression is a powerful machine learning techniques used for regression tasks.

1. IMPORT LIBRARIES AND LOAD DATA:

```
import pandas as pd  
  
import xgboost as xgb  
  
from sklearn.model_selection import train_test_split  
  
from sklearn.metrics import mean_squared_error, r2_score  
  
  
# Load your dataset (replace 'your_dataset.csv' with your data)  
  
data = pd.read_csv('your_dataset.csv')
```

Define your features (X) and target variable (y)

X = data.drop('Price', axis=1) # Features (excluding the target variable)

y = data['Price'] # Target variable

2.SPLIT DATA INTO TRAINING AND TESTING SETS:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

3.CREATE AND TRAIN THE XGBoost REGRESSION:

Create an XGBoost regressor instance

regressor = xgb.XGBRegressor (objective='reg:squarederror', random_state=42)

Train the model on the training data

regressor.fit(X_train, y_train)

CONCLUSION:

In the phase 2 conclusion, we will summarize the key findings and insights from the advanced regression techniques. We will reiterate the impact of these techniques on improving the accuracy and robustness of house price prediction.

