PREDICTING HOUSE PRICE USING MACHINE LEARNING

Phase 3 Submission Document

Project: House Price Prediction

Introduction:

- The real estate market is one of the most dynamic and lucrative sectors, with house prices constantly fluctuating based on various factors such as location, size, amenities, and economic 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.

- Emphasize the need for advanced regression techniques like Gradient Boosting and XGBoost to enhance prediction accuracy.

Phase 3: Development Part 1

In this part you will begin building your project by loading and preprocessing the dataset. Start building the house price prediction model by loading and preprocessing the dataset. Load the housing dataset and preprocess the data.

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)

Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
79545.45857	5.682861322	7.009188143	4.09	23086.8005	1059033.56	208
79248.64245	6.002899808	6.730821019	3.09	40173.07217	1505890.91	188
61287.06718	5.86588984	8.51272743	5.13	36882.1594	1058987.99	9127
63345.24005	7.188236095	5.586728665	3.26	34310.24283	1260616.81	USS
59982.19723	5.040554523	7.839387785	4.23	26354.10947	630943.489	USNS
80175.75416	4.988407758	6.104512439	4.04	26748.42842	1068138.07	06039
64698.46343	6.025335907	8.147759585	3.41	60828.24909	1502055.82	4759
78394.33928	6.989779748	6.620477995	2.42	36516.35897	1573936.56	972 Joyce
59927.66081	5.36212557	6.393120981	2.3	29387.396	798869.533	USS
81885.92718	4.42367179	8.167688003	6.1	40149.96575	1545154.81	Unit 9446
80527.47208	8.093512681	5.0427468	4.1	47224.35984	1707045.72	6368
50593.6955	4.496512793	7.467627404	4.49	34343.99189	663732.397	911
39033.80924	7.671755373	7.250029317	3.1	39220.36147	1042814.1	209
73163.66344	6.919534825	5.993187901	2.27	32326.12314	1291331.52	829
69391.38018	5.344776177	8.406417715	4.37	35521.29403	1402818.21	PSC 5330,
73091.86675	5.443156467	8.517512711	4.01	23929.52405	1306674.66	2278
79706.96306	5.067889591	8.219771123	3.12	39717.81358	1556786.6	064
61929.07702	4.788550242	5.097009554	4.3	24595.9015	528485.247	5498
63508.1943	5.94716514	7.187773835	5.12	35719.65305	1019425.94	Unit 7424
62085.2764	5.739410844	7.091808104	5.49	44922.1067	1030591.43	19696
86294.99909	6.62745694	8.011897853	4.07	47560.77534	2146925.34	030 Larry
60835.08998	5.551221592	6.517175038	2.1	45574.74166	929247.6	USNS
64490.65027	4.21032287	5.478087731	4.31	40358.96011	718887.232	95198
60697.35154	6.170484091	7.150536572	6.34	28140.96709	743999.819	9003 Jay
59748.85549	5.339339881	7.748681606	4.23	27809.98654	895737.133	24282

Import Libraries:

Import the necessary libraries for your project. You'll likely need libraries such as Pandas, NumPy, and Scikit-Learn

Source Code:

```
import pandas as pd import
numpy as np from
sklearn.model_selection
import train_test_split
```

Load the dataset:

Load your housing dataset into a Pandas DataFrame. You can typically load data from a CSV file using the pd.read_csv() function

Source Code:

```
# Load the dataset

data = pd.read_csv('housing_data.csv')  # Replace 'housing_data.csv' with the actual filename.
```

Explore the data:

Take a look at the data to understand its structure and content. You can use functions like data.head(), data.info(), and data.describe() to get an overview.

Source Code:

```
# Display the first few rows of the dataset print(data.head())

# Get information about the dataset
print(data.info())

# Summary statistics of the data print(data.describe())
```

Data Preprocessing:

- Handling Missing Data:

Check for missing values in the dataset and decide how to handle them (e.g., by filling missing values with the mean, median, or using other techniques).

Source Code:

```
# Handle missing values (if any)
data.fillna(method='ffill', inplace=True)  # Example: Forward fill missing values
```

Feature Engineering:

Create new features or transform existing ones if necessary. For example, you might want to create features like the age of the house, the price per square foot, etc.

Categorical Data:

If your dataset contains categorical variables, you may need to encode them using techniques like one-hot encoding.

Source Code:

```
# Example one-hot encoding
data = pd.get_dummies(data, columns=['categorical_column'])
```

Scaling:

Normalize or standardize numerical features to ensure that they are on a similar scale.

Source Code:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data[['numerical_column1', 'numerical_column2']] =
scaler.fit_transform(data[['numerical_column1', 'numerical_column2']])
```

Split the data:

Split your dataset into training and testing sets. This is typically done to assess the model's performance on unseen data.

Source Code:

```
# Define the target variable (house prices) and features (predictors)
X = data.drop('price', axis=1)
```

```
y = data['price']

# Split the 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)
```

Conclusion and Future Work:

We have loaded and preprocessed the dataset. Next step we can proceed to build and train the house price prediction model in the next phase of the project.

Project Conclusion:

In Phase 3 of the project, the housing dataset was loaded and preprocessed. Key steps included handling missing data, feature engineering, categorical data encoding, and feature scaling. Exploratory analysis was performed to gain insights into the data's structure and quality. The dataset was then split into training and testing sets for model evaluation. These essential preprocessing steps have paved the way for subsequent phases, where the house price prediction model will be developed and assessed.

Thank You

Done by –

Rahul Kumar

Rajnish Kumar