

HOUSE PRICE PREDICTION



SUBMITTED BY

Group 3

Akarsh S Nair

AM.EN.U4AIE21008

Alfy Alex

AM.EN.U4AIE21011

Nayan M.K

AM.EN.U4AIE21048

Shyamdev Krishnan J

AM.EN.U4AIE21060

Sabarinath B

AM.EN.U4AIE21080

Introduction:

House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. While there are people who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood. According to, housing markets have a positive impact on a country's currency, which is an important national economy scale. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location.. According to , numerous international organizations and human rights have emphasized house importance.

Houses are profoundly rooted in the economic, financial, and political structure of each country. Nevertheless, it is reported that the fluctuation of house prices has always been an issue for house owners, buildings and real estate, Besides stating that houses have become unaffordable as there is substantial price growth in several countries in the housing sector.

House price prediction can be done by using multiple prediction models (Machine Learning Model) such as Linear regression , logistics regression, support vector regression, and more. But in this project we are trying to predict the price of the house using the techniques called linear regression , were we take the each attribute in the dataset that we got from Kaggel (USA_housing.csv) and predict the price of the house and see how each attribute is having influence over the prediction of the price of house.

Objective:

The objective of this project is to find how the price of the house can be or is predicted using various models(Linear Regression Models) using the attributes like Avg. Area Income, Avg. Area House Age ,Avg. Area Number of Rooms and etc.As a result of this prediction the people can easily find the approximate price of the house for buying or selling their house based on the various factors as mentioned above.

Data:

The dataset has been selected from the kaggle named "USA_housing.csv" which was the data collected from the analysis done in the USA considering factors like Avg. Area House Age ,Avg. Area Number of Rooms , etc.This data has been provided by Kaggle.

The data consist of almost 500 rows and 7 columns

We are going to use the USA_Housing dataset. Since house price is a continuous variable, this is a regression problem. The data contains the following columns:

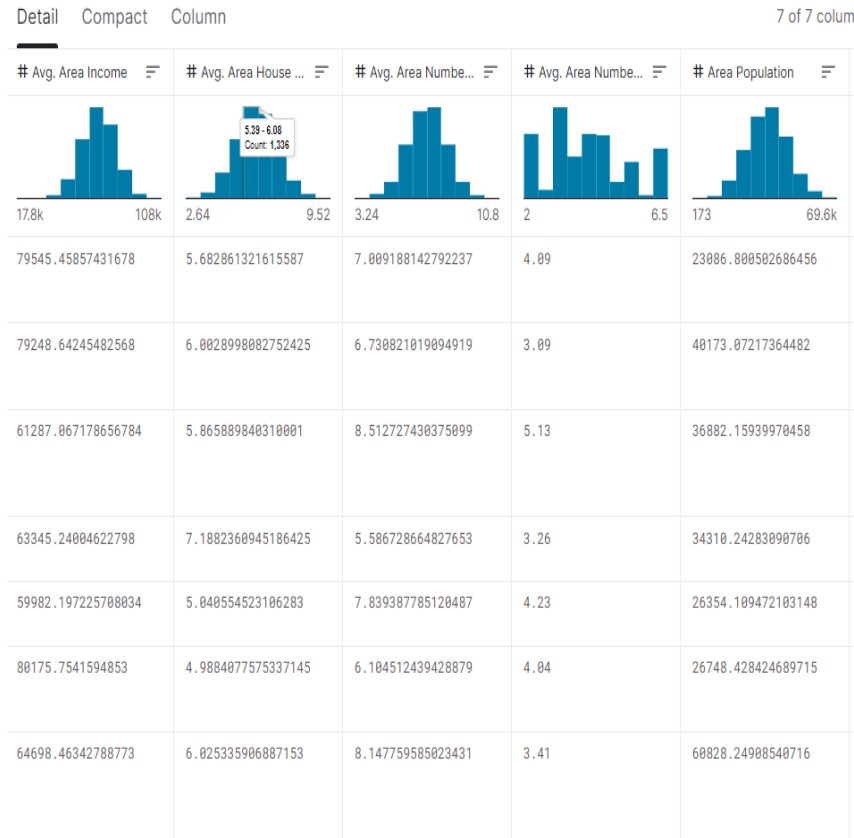
- 'Avg. Area Income': Avg. Income of residents of the city house is located in.
- 'Avg. Area House Age': Avg Age of Houses in same city
- 'Avg. Area Number of Rooms': Avg Number of Rooms for Houses in same city
- 'Avg. Area Number of Bedrooms': Avg Number of Bedrooms for Houses in same city
- 'Area Population': Population of city house is located in
- 'Price': Price that the house sold at
- 'Address': Address for the house

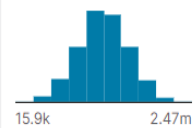
Among them we are going to predict the price of House.Here we have one dependent value and six independent variables.

Kaggle dataset:

<https://www.kaggle.com/code/gopalchetttri/usa-housing-machine-learning-linear-regression/data>

USA_Housing.csv (726.21 KiB)



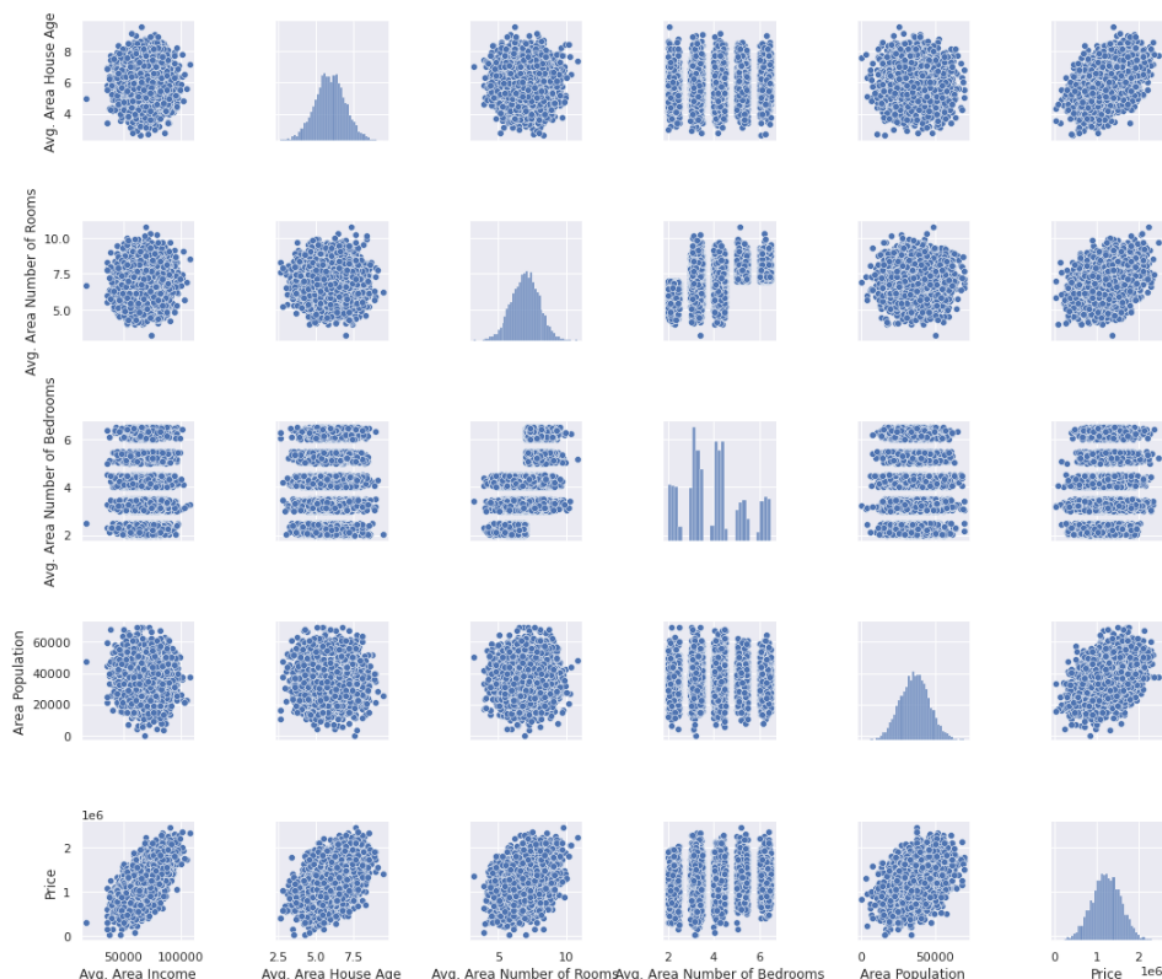
# Price	Address
	5000 unique values
15.9k	2.47m
1059033.5578701235	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
1505890.91484695	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
1058987.9878760849	9127 Elizabeth Stravenue Danieltown, WI 06482-3489
1260616.8066294468	USS Barnett FPO AP 44820
630943.4893385402	USNS Raymond FPO AE 09386
1068138.0743935304	06039 Jennifer Islands Apt. 443 Tracyport, KS 16077
1502055.8173744078	4759 Daniel Shoals Suite 442 Nguyenburgh, CO 20247
1573936.5644777215	972 Joyce Viaduct Lake William TN

Data Analysis:

❖ Getting the count of Each attributes and the data type:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
#   Column                                Non-Null Count  Dtype  
---  -
0   Avg. Area Income                      5000 non-null   float64
1   Avg. Area House Age                   5000 non-null   float64
2   Avg. Area Number of Rooms              5000 non-null   float64
3   Avg. Area Number of Bedrooms           5000 non-null   float64
4   Area Population                        5000 non-null   float64
5   Price                                 5000 non-null   float64
6   Address                               5000 non-null   object  
dtypes: float64(6), object(1)
```



Methodology:

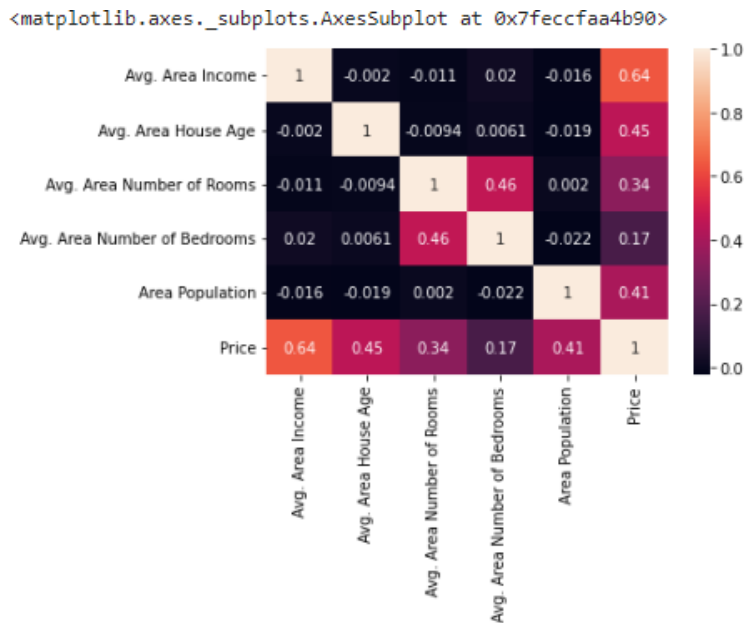
1.Preparing or Preprocessing the DataSet:

Here we are processing the data and getting the data of each , checking if there is a null value and removing the null value if there is any .As a part of feature selection, we have dropped the attribute address from our data set using python. This attribute will not be used in any of the prediction models which we are developing.

- `df.info()` - to get the information about the data, the rows,columns, etc

- df.columns- to get information in columns of the data.
- df.describe() - Used to describe the dataset

Find the relationship of each value using a heat map



HeatMap

2 Model Evaluation:

	coeff
Avg. Area Income	21.528276
Avg. Area House Age	164883.282027
Avg. Area Number of Rooms	122368.678027
Avg. Area Number of Bedrooms	2233.801864
Area Population	15.150420

Interpreting the coefficients:

- Holding all other features fixed, a 1 unit increase in Avg. Area Income is associated with an increase of \$21.52.
- Holding all other features fixed, a 1 unit increase in Avg. Area House Age is associated with an increase of \$164883.28.
- Holding all other features fixed, a 1 unit increase in Avg. Area Number of Rooms is associated with an increase of \$122368.67.
- Holding all other features fixed, a 1 unit increase in Avg. Area Number of Bedrooms is associated with an increase of \$2233.80.
- Holding all other features fixed, a 1 unit increase in Area Population is associated with an increase of \$15.15.

3 Preparing Data For Linear Regression

Linear regression has been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model.

As such, there is a lot of sophistication when talking about these requirements and expectations which can be intimidating. In practice, you can use these rules more as rules of thumb when using Ordinary Least Squares Regression, the most common implementation of linear regression. Try different preparations of your data using these heuristics and see what works best for your problem.

- **Linear Assumption.** Linear regression assumes that the relationship between your input and output is linear. It does not support anything else. This may be obvious, but it is good to remember when you have a lot of attributes. You may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).
- **Remove Noise.** Linear regression assumes that your input and output variables are not noisy. Consider using data cleaning operations that let you better expose and clarify the signal in your data. This is most

important for the output variable and you want to remove outliers in the output variable (y) if possible.

- **Remove Collinearity.** Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated.
- **Gaussian Distributions.** Linear regression will make more reliable predictions if your input and output variables have a Gaussian distribution. You may get some benefit using transforms (e.g. log or BoxCox) on your variables to make their distribution more Gaussian looking.
- **Rescale Inputs:** Linear regression will often make more reliable predictions if you rescale input variables using standardization or normalization.

4. Splitting the data into Test and Training data for prediction:

We are splitting the data into training and testing after data preprocessing.. This is done to train the model which is being used and to measure parameters like MSE, R2 score, RMSE, etc. Typically, while separating a data set into a training dataset and testing dataset, most of the data is used for the training process, and a smaller portion of the data is used for testing. After a model has been made by using this training set, test the model by making predictions against the test Set. We had divided the as 40% testing data and 60% of them as training data. Set the random state to 101.

5. Models used for the prediction

The features used for prediction is Area Income, Avg. Area House Age, Avg. Area Number of Rooms, Area Population.

Models used:

- Prediction of House Price based on Avg. Area Number of Rooms
- Prediction of House Price based on Avg. Area Number of Bedrooms

- Prediction of House Price based on Area Population
- Prediction of House Price based on Avg. Area Income and Avg. Area House Age
- Prediction of House Price based on Avg. Area Income and Avg. Area Number of Rooms
- Prediction of House Price based on Avg. Area Income , Avg. Area Number of Rooms and Area population
- Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Rooms
- Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Bedrooms
- Prediction of House Price based on Area population and Avg. Area Number of Rooms
- Ridge Regression: Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated.

```
# Fit a ridge model on the training set
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler

model = Ridge(alpha=1)
scaler = StandardScaler()
pipe = make_pipeline(scaler,model)
pipe = pipe.fit(X_train,y_train)

# Predict on test set
predictions = pipe.predict(X_test)

# Evaluate the model using the test data
mse = mean_squared_error(y_test, predictions)
print("MSE:", mse)
rmse = np.sqrt(mse)
print("RMSE:", rmse)
r2 = r2_score(y_test, predictions)
print("R2:", r2)
```

- Polynomial regression : It provides a great defined relationship between the independent and dependent variables.

```

from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree=3)
X = poly_reg.fit_transform(X)
#feature scaling
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X = sc_X.fit_transform(X)
#split
from sklearn.model_selection import train_test_split
X_train ,X_test, y_train ,y_test = train_test_split(X, y, test_size=0.2, random_state = 0)
#ols
from sklearn.linear_model import LinearRegression
lin_reg=LinearRegression()
lin_reg.fit(X_train,y_train)

```

Mean absolute error,Mean squared error ,Root mean squared error and r2_score were found for each model and predication were done and the graph of the predicted and actual values were displayed

Regression Evaluation Metrics

Here are three common evaluation metrics for regression problems:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Comparing these metrics:

- MAE is the easiest to understand, because it's the average error.
- MSE is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- RMSE is even more popular than MSE, because RMSE is interpretable in the "y" units.

R2_Score:

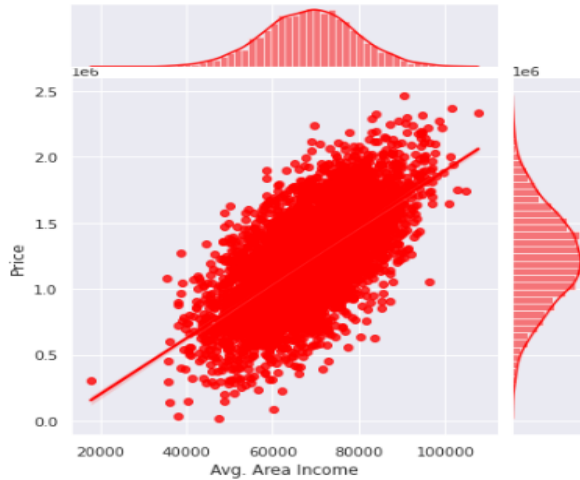
It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s).

Result:

Prediction of House Price based on Avg. Area Income

MSE: 73727682715.874
MAE: 217170.5274585681
RMSE: 271528.41972043
Accuracy: 41.983465594074666

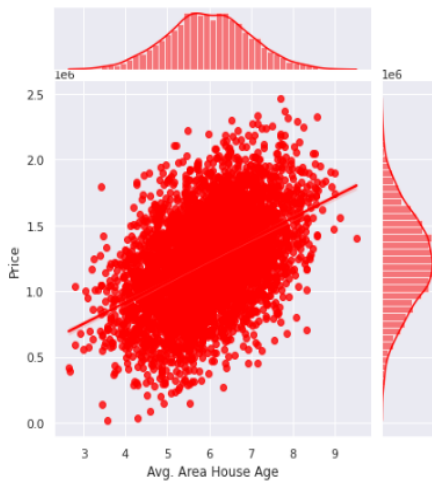
```
sns.jointplot(x='Avg. Area Income',y="Price",data=df,kind="reg",color="red");
```



Prediction of House Price based on Avg. Area House Age

MSE: 101283502165.68831
MAE: 254773.46835667192
RMSE: 318250.69075445586
Accuracy: 20.29970871601656

```
sns.jointplot(x='Avg. Area House Age',y="Price",data=df,kind="reg",color="red");
```



Prediction of House Price based on Avg. Area Number of Rooms

MSE: 112986824904.71535
MAE: 266659.4702288811
RMSE: 336135.1289358423
Accuaracy: 11.090328991320286

Prediction of House Price based on Avg. Area Number of BedRooms

MSE: 123029978141.4039
MAE: 280018.15310426796
RMSE: 350756.2945143022
Accuaracy: 3.1873416216267825

Here the r2 score for the prediction based on Avg No. of Bedroom is very less which shows that the effect of Bedroom on the price prediction is very less and it is the worst model.

Prediction of House Price based on Area Population

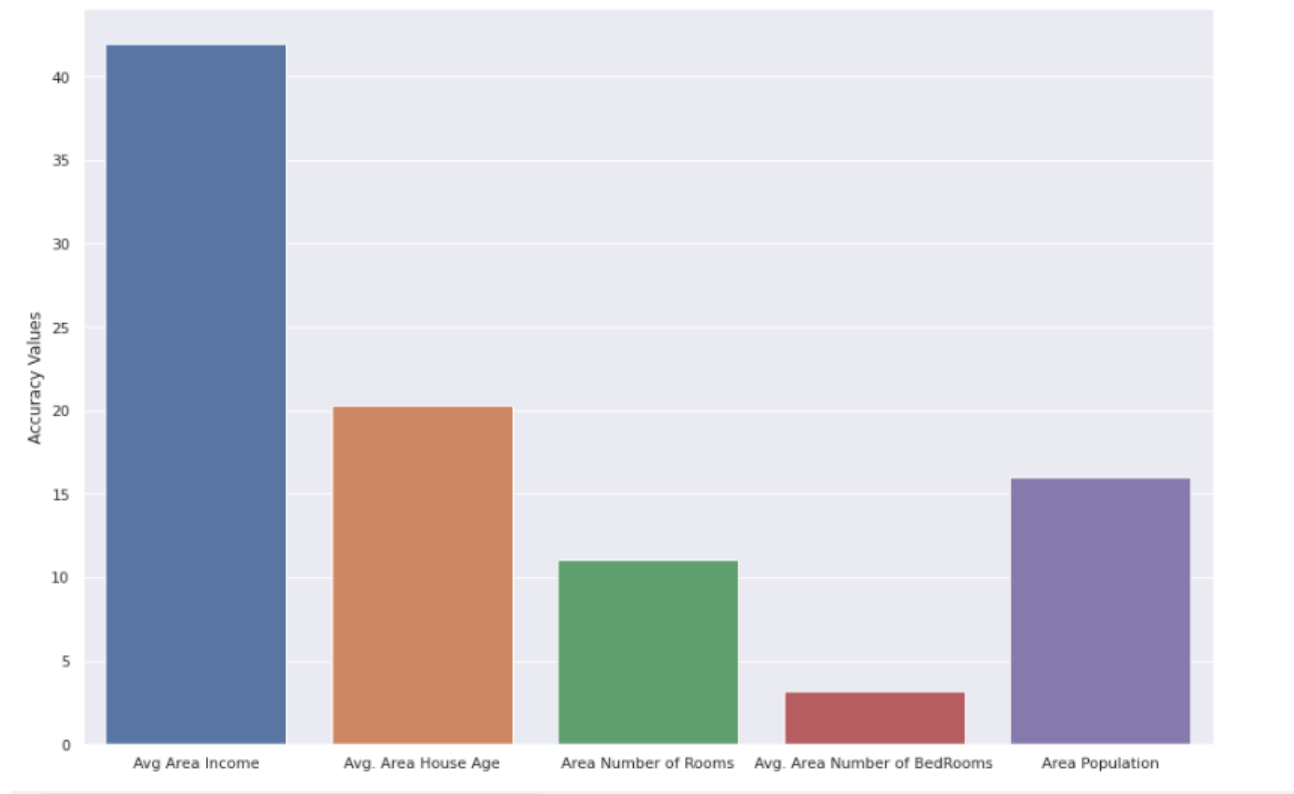
MSE: 106670239812.20892
MAE: 261509.89496165712
RMSE: 326604.10256487736
Accuaracy: 15.983234588698558

The r2_score was less when we take only the take only one attributes for the prediction

Prediction of House Price based on Avg. Area Income and Avg. Area House Age

MSE: 47127723506.06781
MAE: 172548.73873406454
RMSE: 217089.206332484
Accuaracy: 62.91505318566004

Bargraph showing the dependency of the each variables in the prediction of the house



Prediction of House Price based on Avg. Area Income and Avg. Area Number of Rooms

MSE: 60007208781.58539
MAE: 195827.56171990032
RMSE: 244963.68870015285
Accuaracy: 52.78014763739718

Prediction of House Price based on Area population and Avg. Area Number of Rooms

MSE: 92894358391.66727
MAE: 240735.9796764639
RMSE: 304785.75818378926
Accuaracy: 26.901151084378906

Prediction of House Price based on Avg. Area Income , Avg. Area Number of Rooms and Area population

MSE: 39129069872.215034
MAE: 158284.1101037989
RMSE: 197810.69200681502
Accuaracy: 69.20921769287554

Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Rooms

MSE: 33057956205.59746
MAE: 145660.05694526603
RMSE: 181818.4704742548
Accuaracy: 73.9865952252602

Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Bedrooms

MSE: 10459278321.46933
MAE: 82248.33498864669
RMSE: 102270.61318614126
Accuaracy: 91.76956255444573

Ridge Regression

MSE: 10459381769.466032
RMSE: 102271.11894110689
R2: 0.9176948115090721

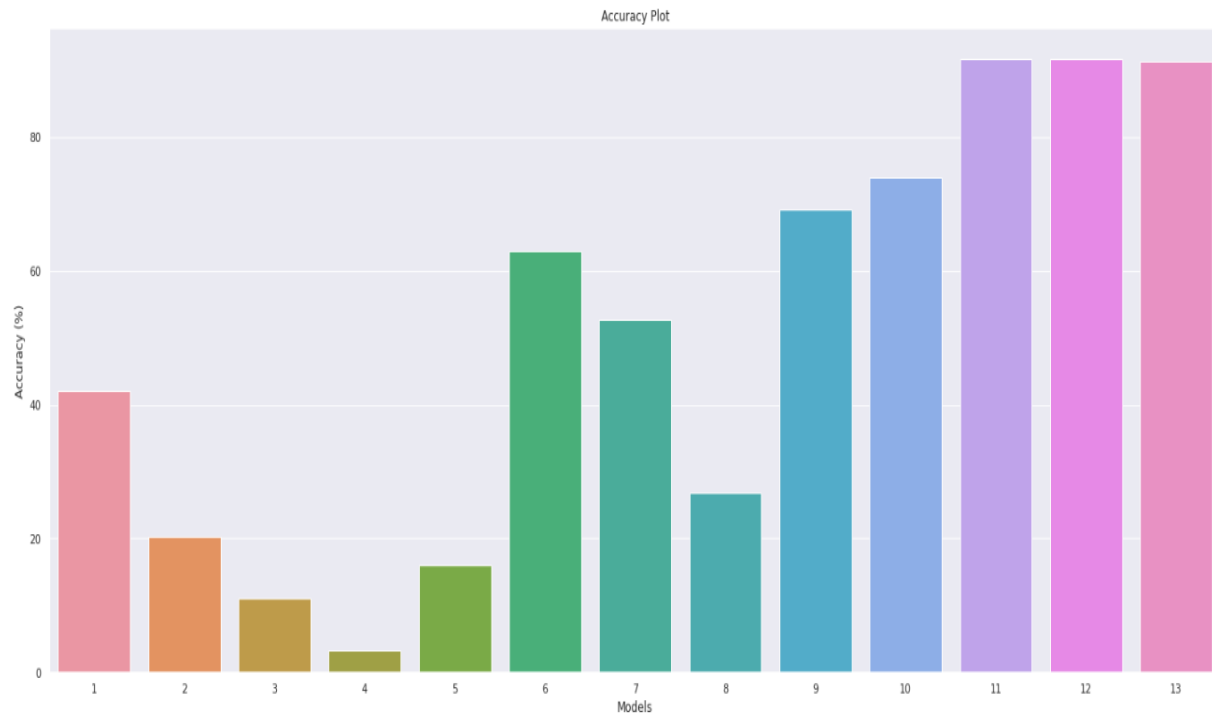
Polynomial regression

MSE: 4177430364388.6387
MAE: 221281.360908129
RMSE: 2043876.3084855792
Accuaracy: -3279.8302682252606

Best Model: 11

Accuracy: 91.76956255444573

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`
FutureWarning



From this bar graph we can infer that the accuracy is more for the 11th model(Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Bedrooms)which we did. In this Graph the Y axis is Accuracy and X axis shows each model that we used.

Discussion:

From each of the model done the good accuracy was 91.76948115090721 , 91.38384159252797 ,91.76956255444573 , 73.9865952252602 for Ridge Regression,Polynomial regression,Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Bedrooms andPrediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of

Rooms respectively. Among each of the 13 models Prediction of House Price based on Avg. Area Income , Avg. Area House Age , Avg. Area Number of Rooms and Avg. Area Number of Bedrooms has given us more accuracy.

As the count of input independent variables increases, The r^2 score is also increasing and can have a better prediction .

It is found that the r^2_{score} for the prediction of the house price for a single attribute is very less and is very very less for Prediction of House Price based on Avg. Area Number of Bedrooms (3.18) which shows that Number of Bedrooms has not that much effect on the House Price prediction and accuracy increases when more number of attributes are taken for prediction .

A low R-squared value indicates that your independent variable is not explaining much in the variation of your dependent variable.

Appendix:

Colab code:

<https://colab.research.google.com/drive/19oNbTKX1Kv292grpDMkxmKH5fNKJMY2Y?userstoinvite=dhruv.r.krish%40gmail.com&actionButton=1#scrollTo=QRsT98q68gUv>

Reference :

- <https://www.kaggle.com/code/kaiyungtan/usa-housing-linear-models/notebook>
- <https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f>