House price prediction model.

Authors:

Pulla Shradha Reddy

Abstract:

In this project, we develop a machine learning model to predict the price of houses in the state of California based on various features. Common features are the size of the house, number of bedrooms, and location. They are also many other hidden features that affect the price of the house. we collected our data from Kaggle. We used regression techniques such as linear regression and random forest regression to train the model. we used feature engineering to derive two more features the bedroom ratio and household rooms. we used metrics to evaluate the model. our results show that our model can accurately predict the price of the houses.

Introduction:

The house price prediction model is an important task in the real estate industry as it helps buyers, and private companies as it makes informed decisions about property transactions.

To develop a successful model first, we have to gather the data then we can clean and pre-process the data next, we can use various data mining techniques to build a predictive model.

Finally, we evaluate the performance of the model using metrics such as mean squared error, mean absolute error, and r- squared error. the overall goal of this project is to develop an accurate model that can help buyers and sellers to take the right decisions based on the predicted house prices.

Methods:

I) Data collection:

We have taken the dataset from Kaggle our dataset consists of 20,640 columns and 10 rows.

There are 10 different variables in our dataset.

Fig 1.1

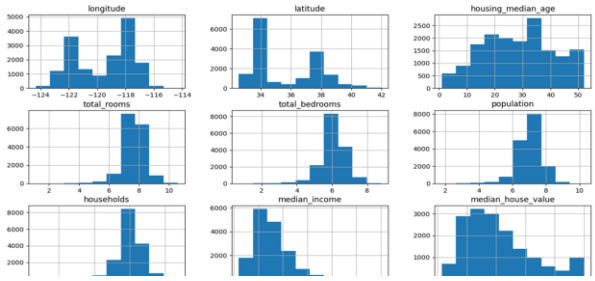
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income($10,000)	median_house_value	ocean_proximity
0	-122.23	37.86	41	880	129	322	126	8.3252	452600	NEAR BAY
1	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
2	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
3	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
4	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY

The longitude variable is the measure of how far west the house is located. The latitude variable is the measure of how far north the house is located. The median age of a house is the median age of a house within a block .lowe the number the newer the building. The toatal_rooms defines the total number of rooms within a block of houses. Total bedrooms define the total number of bedrooms within a block of houses. population variable defines the total number of people residing in a block of houses. The households variable defines the total number of households, a group of people residing in a home unit within a block of houses. Median income defines the median income of households within a block. Median house value is the median value of houses within a block it is measured in us dollars. Ocean proximity defines how near or farther the house is located from the ocean.

II) Data cleaning and pre-processing:

A histogram can be used to visualize the distribution of house prices in the dataset. It shows us how many houses fall within a particular price range. The X-axis shows us the variables in the dataset the y-axis is the frequency of houses.

Fig 2.1



Removing any missing and inaccurate values and transforming data into a suitable format for analysis. standardizing total rooms, total bedrooms, and population. Then converting categorical data into numerical data.

```
#   Column              Non-Null Count  Dtype
---  ------              --------------  -----
0   longitude           20640 non-null  float64
1   latitude            20640 non-null  float64
2   housing_median_age  20640 non-null  float64
3   total_rooms         20640 non-null  float64
4   total_bedrooms      20433 non-null  float64
5   population          20640 non-null  float64
6   households          20640 non-null  float64
7   median_income       20640 non-null  float64
8   median_house_value  20640 non-null  float64
9   ocean_proximity     20640 non-null  object
```

Fig 2.3

```
<1H OCEAN     7219
INLAND        5195
NEAR OCEAN    2115
NEAR BAY      1815
ISLAND           2
Name: ocean_proximity, dtype: int64
```

Regression analysis:

a) Linear regression:

To use linear regression analysis first we need to gather the data on the predictor variables and the house prices. when we performed the metrics we got an R-squared value of 0.65766777, MAE of 49715.1839 MSE of 4740761051.434472.

```
from sklearn.linear_model import LinearRegression
from sklearn import metrics
OLS = LinearRegression()
OLS.fit(X_train, y_train)
✓ 0.0s

▾ LinearRegression
LinearRegression()
```

```
y_pred=OLS.predict(X_test)
print(" The intercept is " +  str(OLS.intercept_))
print(" The coeffiients are " +  str(OLS.coef_))
print(" The R_sqaured value is " +  str(OLS.score(X_test, y_test)))
print("MAE is:", metrics.mean_absolute_error(y_test,y_pred))
print("MSE is:", metrics.mean_squared_error(y_test,y_pred))

✓ 0.0s

The intercept is -2146719.495872446
The coeffiients are [-2.72191872e+04 -2.61224400e+04  1.03631750e+03 -6.39771322e+00
 9.97707401e+01 -3.73545857e+01  4.99035484e+01  3.93656164e+04
-1.49036788e+05 -1.87787533e+05 -1.51883617e+05 -1.45673607e+05]
The R_sqaured value is 0.6576677709626819
MAE is: 49715.183929114246
MSE is: 4740761051.434472
```

b) Random forest regression:

First we need to gather data on predictor variables and corresponding house prices then we can fit a random forest regression model to the data, with the predictor variables as input and the house prices as output.

the random forest model creates a set of decision trees each of which predicts the house price based on a subset of features the final prediction is made by taking the average or majority vote of all the decision trees.

Random forest regression has many advantages over linear regression. One of the main advantages is that it can handle nonlinear relationships between predictor variables and house values also it can handle a large number of predictor variables and is less prone to overfitting.

when we performed the metrics we got an R-squared value of 0.8266384101819658, MAE of 2400784393 MSE of 48997.799883065345

```
from sklearn.ensemble import RandomForestRegressor

reg = RandomForestRegressor()

reg.fit(X_train, y_train)
```

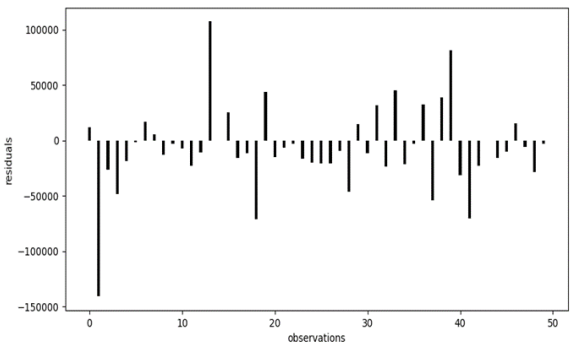
✓ 15.9s

```
RandomForestRegressor()
```

```
y_pred_r = reg.predict(X_test)
acc_rf = metrics.r2_score(y_test,y_pred_r)
print("R^2 values is :", acc_rf)
print("MAE is:", metrics.mean_absolute_error(y_test, y_pred_r))
print("MSE is:", metrics.mean_squared_error(y_test, y_pred_r))
print("RMSE is:", np.sqrt(metrics.mean_squared_error(y_test, y_pre
```

✓ 0.3s

```
0.8266384101819658
R^2 values is : 0.8266384101819658
MAE is: 31180.03259603213
MSE is: 2400784393.3809185
RMSE is: 48997.799883065345
```

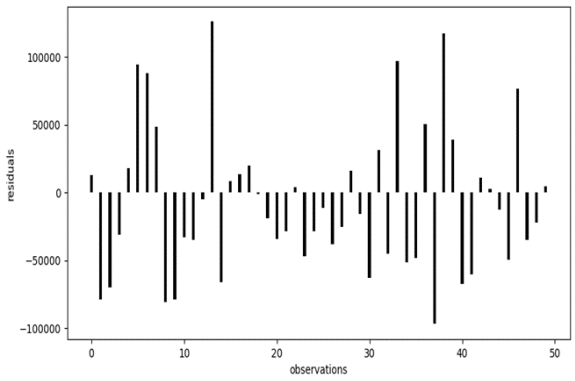


	PREDICTIONS	ACTUAL VALUES	error
15175	306784.03	328200	21415.97
15424	309734.27	156900	-152834.27
16212	111365	87200	-24165
15356	182558	141000	-41558
1899	123728	100800	-22928

Results:

For this project, we have chosen 20,640 rows and 10 columns of data consisting of data. As stated above, the project focuses on predicting the price values of houses based on the available predictors.

When we evaluated our model using linear regression the model consistently overestimates the price of houses with certain features.



	PREDICTIONS	ACTUAL VALUES	error
15175	315646.7385	328200	12553.26153
15424	235926.1562	156900	-79026.1562
16212	157003.3256	87200	-69803.32561
15356	172404.3402	141000	-31404.34025
1899	82931.27063	100800	17868.72937

We evaluated the model using random forest regression and got the following results.

Conclusion:

In conclusion, house price prediction is an important and complex task that can be tackled using various statistical and machine-learning techniques. Linear regression and random forest regression are used in house price prediction due to Their interpretability and ability to capture non-linear relationships between the predictor variables and the target variable.

In a house price prediction project, data cleaning and preprocessing are important steps to ensure that the data is properly formatted and contains all the relevant information needed to make accurate predictions. Feature engineering is also a crucial step, as it involves selecting and transforming the most relevant predictors to improve the accuracy of the model.

Evaluation metrics such as the mean squared error, root mean squared error, and

R-squared value can be used to assess the performance of the model and identify areas for improvement. It is important to carefully evaluate the error of the model and determine if there are specific areas where it is consistently underperforming.

Overall, a successful house price prediction model can have significant practical applications in the real estate industry, enabling buyers, sellers, and real estate agents to make more informed decisions based on accurate and reliable predictions.
Regenerate response.

References:

G. S. Ong, H. Y. Wong, and W. T. Ng, "House price prediction using machine learning: A review," in International Journal of Advanced Computer Science and Applications, vol. 10, no. 4, pp. 25-32, 2019.