Data Science Capstone

Linear Regression Model for Housing Price Prediction

By:

Rajan Gyawali

August 2020

1. Introduction/Business Problem

As a part of Data Science Capstone, I am going to implement the linear regression model for predicting the houses price. The linear regression technique is used for estimating linear relationships between various features and a continuous target value. An example scenario, could be one can estimate the selling price of the house based on different parameters like number of bedrooms, number of bathrooms, sq. ft., number of floors, number of years built, etc. if a linear regression model can be built on housing dataset. The linear regression is an example of supervised machine learning algorithm.

The business problems of this capstone project are:

1. How much more price one can sell his/her house with additional bedroom/bathroom?
2. How does the price of house differ with increase in square feet?
3. What is the impact of number of years built for the pricing of houses?

2. Data Collection/Preprocessing

The housing dataset is downloaded from the Kaggle and is attached in GitHub. It consists of a number of attributes like:

1. Price
2. Number of Bedrooms
3. Number of Bathrooms
4. Living Room Area
5. Number of Floors
6. Waterfront
7. View
8. Condition of the House
9. Year built
10. Year renovated, etc

The dataset is to be split into training and test samples. The training samples are used to fit the model of housing price prediction using linear regression. As a rule of thumb, 80% of the data will be used to train the model and remaining 20% will be used for testing the model.

3. Methodology

The housing price prediction is done using linear regression. Linear regression is a machine learning algorithm to predict the continuous value based on the number of different independent variables.

A basic linear regression model can be formulated as:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \dots \qquad (1)$$

The dataset contains different attributes but the attributes I used are as follows:

1. price
2. Bedrooms
3. Bathrooms
4. sqft_living
5. sqft_lot
6. floors
7. waterfront
8. view
9. condition
10. grade
11. sqft_above
12. sqft_basement
13. number_of_days_built

An analysis of data is done in terms of number of bedrooms. The plot of count of houses with number of bedrooms is shown in the figure below:

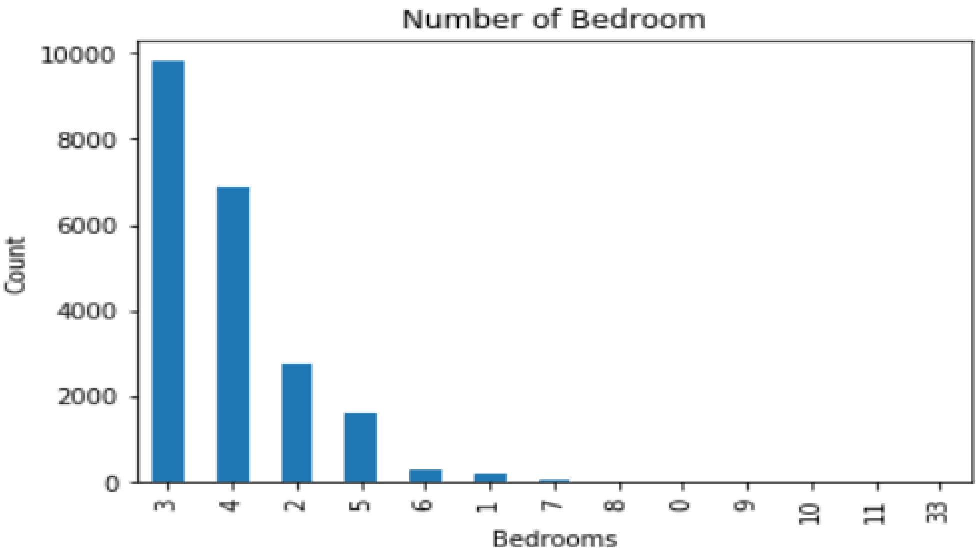

Figure: Plot of Number of Bedrooms Vs Count of Houses

Further, analysis of data is done in terms of number of bathrooms. The plot of count of houses with number of bathrooms is shown in the figure below:

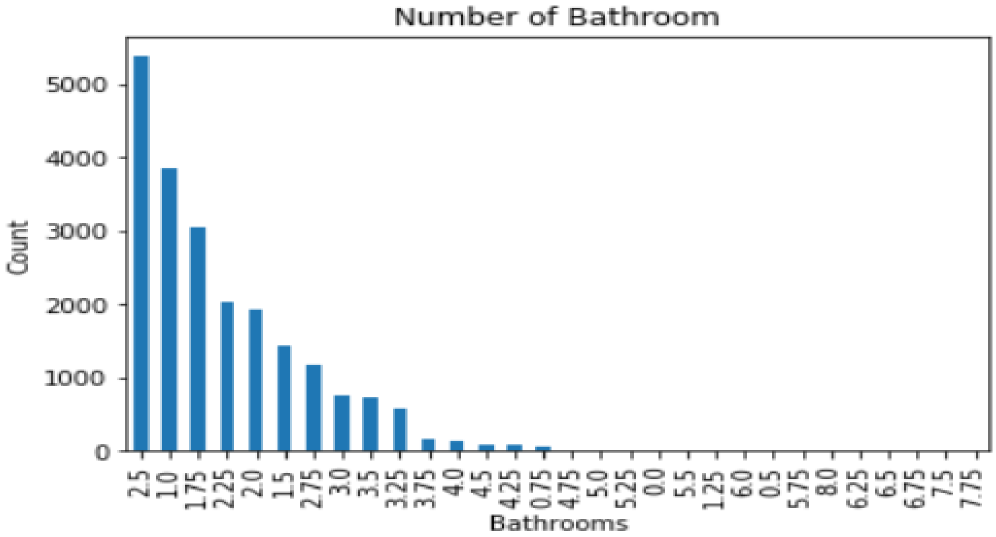

Figure: Plot of Number of Bathrooms Vs Count of Houses

Further, exploratory analysis of data was done with scatterplot between price of house and number of bedrooms/number of bathrooms.

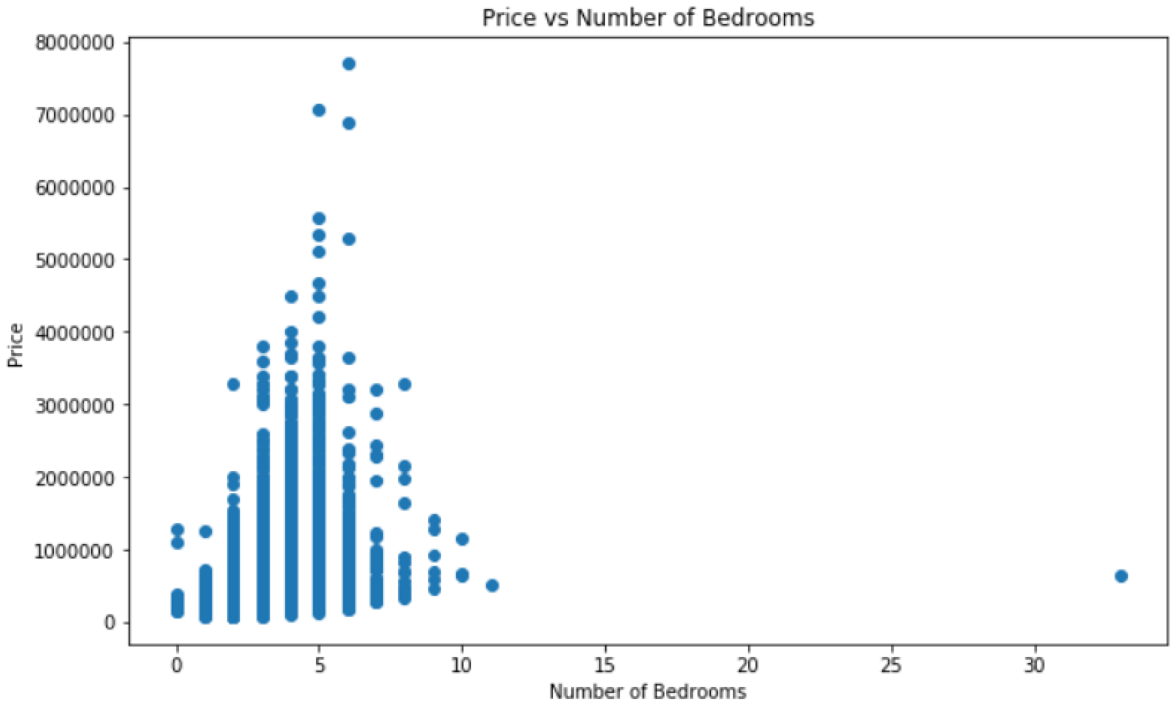

Figure: Scatterplot to indicate the relationship between price and number of bedrooms

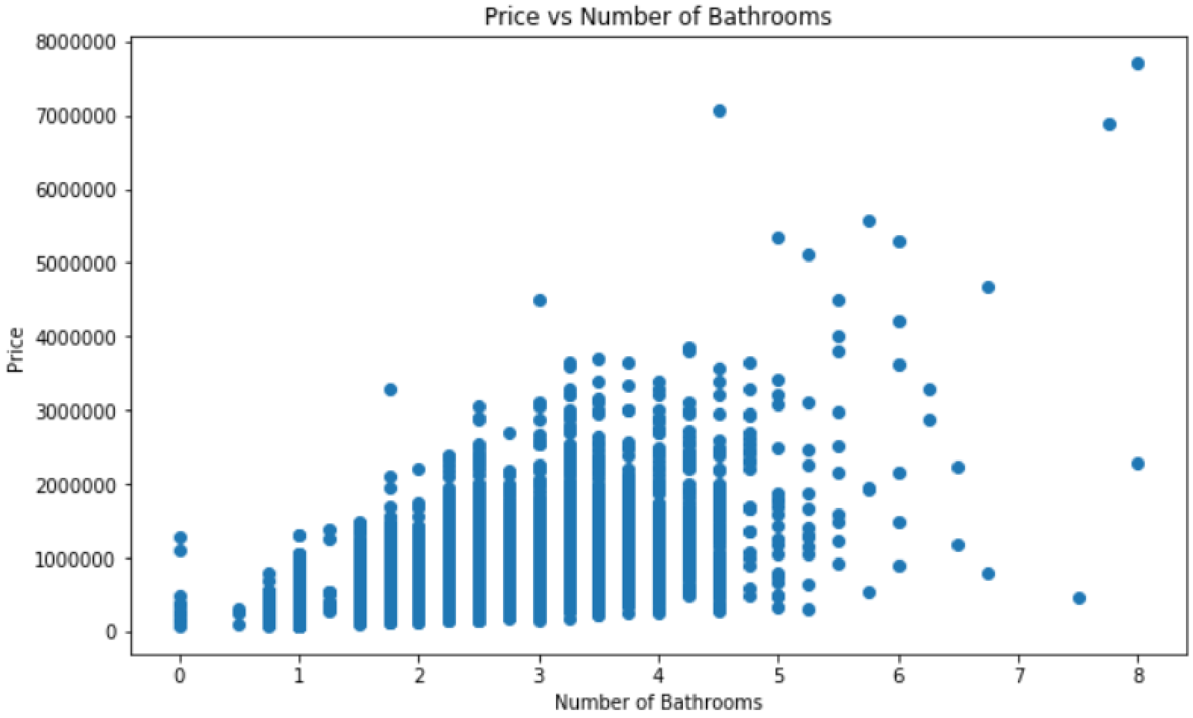

Figure: Scatterplot to indicate the relationship between price and number of bathrooms

Similarly, the relationship between price and square feet is analyzed with the scatterplot and the result is shown as below:

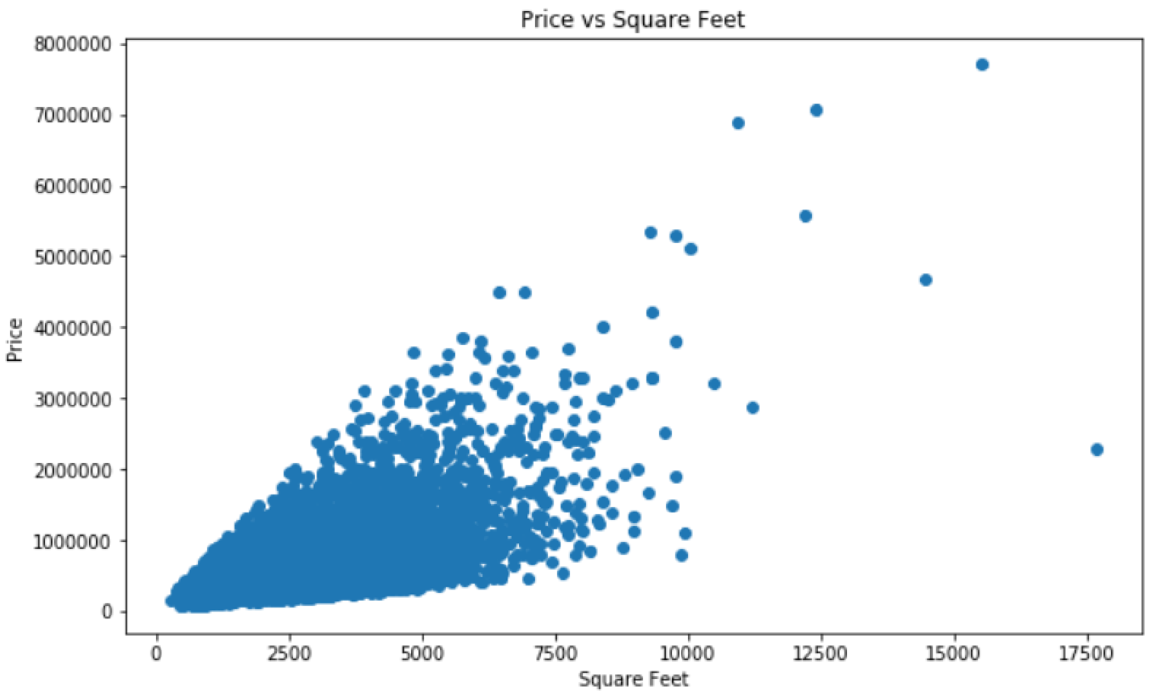

Figure: Scatterplot to indicate the relationship between price and square feet

Model Fitting and Prediction

The linear regression model is made for housing price prediction. I used scikit learn machine learning library to implement linear regression model. From the scikit library, *train_test_split()* is used to split the dataset into training and testing samples. The four set of samples namely *x_train, x_test, y_train and y_test* is developed. The *y_train* contains the *prices* for number of houses and *x_train* contains the other independent attributes for housing price determination.

From the overall dataset, 80% is converted to training set and rest 20% is converted to test set.

With these training samples, the machine learning model for linear regression is modeled and fitted with the scikit learn library.

The fitted model is tested with the test dataset and a regression score is calculated.

4. Results

The housing price prediction is done with the linear regression model. The sample results are shown as below:

Predicted Price	Actual Price
666735	735000
1455081	1150000
337335	350500
1183792	860000
191910	122000
750749	725000
644535	417000
551787	594950
511011	471000
820174	634950
518060	500000
770764	768000
486381	323000
418779	430000
609564	625000
609511	710000
455584	620000
1015401	665000
1609385	1600000
1109880	875000

The model is evaluated with the value of Regression Score and is found to be 0.69. This value is lower and indicates the housing prices are not predicted with higher accuracy. This is because of the less samples of data used for training. This accuracy can be increased by increasing the number of training samples.

5. Discussion

Linear regression model is highly helpful in determining the continuous target value. In this housing price prediction model, a number of dependent variables are used to determine the pricing of house. The model can be more useful in prediction of houses if the model can be trained with sufficiently large datasets.

This model can be further extended in predicting continuous value of target values.

6. Conclusion

The linear regression model is successfully implemented for prediction of housing prices with number of independent variables like number of bathrooms, number of bedrooms, square feet, number of floors, etc. With accurate tuning and parameter optimization, the more accurate prediction of housing price can be done.