Model Building: Real Estate price prediction

Feziwe M Shongwe

melvin.shongwe@gmail.com

Introduction

Real estate is one of the most competitive markets in terms of pricing and the same tends to vary significantly based on a lot of factors which include location, Features and Condition and Supply and demand. Researchers have worked on developing models to accurately predict housing prices with high accuracy and least error, using various factors such as the area of the house, the number of bedrooms, etc. Previous studies show that this prediction can include a lot of factors especially for businesses that deal in real estate are likely to have billions of features to choose from as this can have several drawbacks such as heavy computations will be required. Lastly, most models tend to include highly correlated factors and uninfluential features in predicting the price which does not improve the accuracy of the model.

Approach to solve this problem

In this short study, I will use the features from the dataset to build an optimum model (Multiple regression model) for the prediction of the price for a specific house using 5 features. The features that I will be using includes date of purchase, house age, location, distance to nearest MRT station, and house price of unit area. I will follow the four basic steps for building a good model which consist of 4 phases (data collection and preparation Reduction of explanatory variables, Model refinement and selection, and Model validation). Under the collection and preparation phase, the dataset (see section 1 under appendix) that I will be using throughout the analysis is publicly available on the Kaggle website based on a study that was conducted in Sindian Dist, New Taipei City, Taiwan. My variables seem to be related to the response variables as they are influencing the prediction which makes me to employ the Explanatory observations studies. I split my data into two sets which is the training set (for modelling) and validating set (for validating the proposed model). The main reason for applying linear regression is because my main goal for this study, is to predict an accurate house price given different features which makes it multivariable linear regression as I will have more than one predictor variable and liner regression focuses on continuous variables which is applicable in this case also. Lastly, studies show that linear regression model has proven to be a reliable and scientific way to predict the future. Below is an information per attribute:

Predictor Variables:

No =the observation number (1 represents 1st house information)

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (Unit: degree)

X6=the geographic coordinate, longitude. (Unit: degree)

Response variable:

Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Full analysis

From my exploratory variables, I found that the predictor variable which represents number of observations is not useful in this model. Thus, I dropped it during the preliminary investigation stage. In addition, I decided to include an interaction term of the coordinated which is a product of the longitude and the latitude. During the preliminary investigation, I also obtained a scatter plot to determine relationship among my predictor variables (excluding the dropped variable).

From the Multicollinearity analysis (see figure 4), the geographical latitude is highly correlated with the interaction variable of the latitude and the longitude with a correlation coefficient of ($r = 0.978$. Similarly, to the geographical longitude is highly correlated with the distance of nearest MRT station with a Pearson correlation of coefficient of ($r = -0.800$).

The statistics for the multiple linear regression model (Full model see Figure 5), $R^2 = 0.583$, which indicates 58.3% of the variability in the predictor variables is explained by the response variable (House price of unit area). From the parameter estimates the statistics shoes that all the estimators have $p - values < 0.05$ which indicated that all these predictors will be included in the model as we reject the null hypothesis in saying the variable(s) equals zero for all of them. Lastly, the Analysis of variance (ANOVA) F-value is extremely large which indicated that not all of the coefficients equal to zero, furthermore the $p - value < 0.05$ which supports the rejection of the null hypothesis.

Using the model selection criteria (see Figure 6), $R_p^2 \; and \; Adjusted \; R^2$ are increasing from one model to another I am more interested whenever the increase rate is no longer large. Thus, the maximum $max(R^2) = 0.5755$ suggest 5 variables in the model (All the predictor variable). The mallows' C_p considers good subset if $C_p < p \; and \; E\{C_p\} \approx p$, in this case we have $p = 6$, it follows that C_p suggests 4-5 variables as good models for this prediction. Furthermore, the AIC_p suggest 5- variables should be included in the model.

Models:

$$\hat{Y} = -15654 + 4.84870X_1 - 0.30272X_2 - 0.00379X_3 + 1.08678X_4 + 1.95592X_5X_6$$

$$\hat{Y} = -11278 + 5.62372X_1 - 0.289515X_2 - 0.00539X_3 + 1.21885X_4$$

$$\hat{Y} = -17441 + 5.33620X_1 - 0.29741X_2 - 0.00518X_3 + 2.20344X_5X_6$$

Where:

$X1 = the \; transaction \; date \; (for \; example, \; 2013.250 = 2013 \; March, \; 2013.500 = 2013 \; June, \; etc.)$

$X2 = the \; house \; age \; (unit: \; year)$

$X3 = the \; distance \; to \; the \; nearest \; MRT \; station \; (unit: \; meter)$

$$X4 = the\ number\ of\ convenience\ stores\ in\ the\ living\ circle\ on\ foot\ (integer)$$

$$X_5X_6 = Interaction\ variable\ for\ the\ latitute\ and\ longitude$$

To verify the computations from the Models selection criterions I will then compute the Automatic search methods which are Forward Stepwise regression, Forward selection, and Backward elimination. Surprisingly, all the methods support the models that were suggested by the AIC_p & C_p as they also show that the model should include all the predictor variables. (See figures under section 4 in appendix)

Therefore, I can conclude that the model with 5 variables is a good model for this prediction.

$$\hat{Y} = -15654 + 4.84870X_1 - 0.30272X_2 - 0.00379X_3 + 1.08678X_4 + 1.95592X_5X_6$$

The parameters are still the same as stated above.

Model Validation

Starting by comparing the factors that are considers as good factors for these predictions, studies show that location, structural and Neighbourhood are good factors for this prediction. Thus, it corresponds with the factors I have computed using the multiple linear regression as my model consists of when was the transaction made, how old is the house since it was built, Distance to the nearest MRT station, Number of convenience stores and the location of the house using the longitude and latitude. Thus, my study supports the Neighbourhood and the location from the previous studies on these factors

Furthermore, I used the validation set to re-estimate the coefficients as it is one of the methods one can use to validate a model. The coefficients estimators using the validation set form the following regression model.

$$\hat{Y} = -20562 + 6.99335X_1 - 0.24294X_2 - 0.00314X_3 + 1.4432X_4 + 2.18937X_5X_6$$

Comparing the two models (Training set model and Validation model) I can clearly see that the model I have produced is valid and it will be applicable to different dataset as it is similar to the one, I formed using the validation set.

Summary and Conclusion

As stated under the validation section, the model is accurate and suitable model for this problem as it is validated, and studies suggest the factors that I found that their good factors for the house price are also similar with the ones I have chosen. Lastly, throughout this study there was no funding and my limitations I had was time as we have other modules and different schedule. I would recommend the next research about this to have a larger dataset and more predictor variables.

Appendix

1. *Datasets*

For full datasets visit the repository on GitHub: Click Here

The Full Dataset

Obs	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT s	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
2	2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
3	3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
4	4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
5	5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
6	6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
7	7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
8	8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
9	9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
10	10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1

Figure 1: First 10 Elements in the Dataset (Full)

The Training set for training my Model

Obs	Selected	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT s	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	1	2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
2	1	3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
3	1	5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
4	1	7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
5	1	8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
6	1	9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
7	1	11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4
8	1	13	2012.917	13	492.2313	5	24.96515	121.53737	39.3
9	1	14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8
10	1	15	2013.5	13.2	1164.838	4	24.99156	121.53406	34.3

Figure 2: First 10 Elements in The Training Set

The Validation set for testing my Model

Obs	Selected	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT s	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	0	1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
2	0	4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
3	0	6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
4	0	10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1
5	0	12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1
6	0	17	2013.25	0	292.9978	6	24.97744	121.54458	70.1
7	0	20	2012.667	1.5	23.38284	7	24.96772	121.54102	47.7
8	0	36	2013.5	13.9	4079.418	0	25.01459	121.51816	27.3
9	0	37	2012.917	14.7	1935.009	2	24.96386	121.51458	22.9
10	0	38	2013.167	12	1360.139	1	24.95204	121.54842	25.3

Figure 3: First 10 Elements in the validation set

2. *Multicollinearity*

Pearson Correlation Coefficients, N = 290 Prob > |r| under H0: Rho=0	X1	X2	X3	X4	X5	X6	INTER_X5_X6
X1 **the transaction date**	1.00000	0.01144 0.8462	0.09454 0.1081	-0.00431 0.9418	0.02834 0.6307	-0.07453 0.2057	0.00871 0.8826
X2 **The house age**	0.01144 0.8462	1.00000	0.01680 0.7758	0.01217 0.8365	0.03402 0.5639	-0.05804 0.3246	0.01752 0.7664
X3 **The distance to the nearest MRT station**	0.09454 0.1081	0.01680 0.7758	1.00000	-0.62367 <.0001	-0.54522 <.0001	-0.80023 <.0001	-0.66913 <.0001
X4 **The number of convenience stores in the living circle on foot (integer)**	-0.00431 0.9418	0.01217 0.8365	-0.62367 <.0001	1.00000	0.44169 <.0001	0.43111 <.0001	0.49335 <.0001
X5 **The geographic coordinate, latitude. (Unit: degree)**	0.02834 0.6307	0.03402 0.5639	-0.54522 <.0001	0.44169 <.0001	1.00000	0.35638 <.0001	0.97775 <.0001
X6 **The geographic coordinate, longitude. (Unit: degree)**	-0.07453 0.2057	-0.05804 0.3246	-0.80023 <.0001	0.43111 <.0001	0.35638 <.0001	1.00000	0.54446 <.0001
INTER_X5_X6 **Interaction of the corordinates(latitude and longitude)**	0.00871 0.8826	0.01752 0.7664	-0.66913 <.0001	0.49335 <.0001	0.97775 <.0001	0.54446 <.0001	1.00000

Figure 4: Multicollinearity between the Predictor variables

Number of Observations Read	290
Number of Observations Used	290

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	31456	6291.12013	79.38	<.0001
Error	284	22509	79.25780		
Corrected Total	289	53965			

Root MSE	8.90269	R-Square	0.5829
Dependent Mean	37.73069	Adj R-Sq	0.5755
Coeff Var	23.59534		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-15654	3827.73438	-4.09	<.0001
X1	the transaction date	1	4.84870	1.85189	2.62	0.0093
X2	The house age	1	-0.30272	0.04632	-6.53	<.0001
X3	The distance to the nearest MRT station	1	-0.00379	0.00063357	-5.98	<.0001
X4	The number of convenience stores in the living circle on foot (integer)	1	1.08678	0.22611	4.81	<.0001
INTER_X5_X6	Interaction of the corordinates(latitude and longitude)	1	1.95592	0.41526	4.71	<.0001

Figure 5: Multiple Linear regression (Full Model)

3. *Regression*

The REG Procedure
Model: MODEL1
Dependent Variable: Y

R-Square Selection Method

Number of Observations Read	290
Number of Observations Used	290

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Parameter Estimates					
						Intercept	X1	X2	X3	X4	INTER_X5_X6
1	0.4333	0.4314	99.8291	1354.8836	1362.22339	45.68560	.	.	-0.00712	.	.
1	0.3536	0.3513	154.1529	1393.0846	1400.42441	-14301	.	.	.	.	4.72515
1	0.3302	0.3279	170.0326	1403.3629	1410.70264	26.91593	.	.	.	2.62231	.
2	0.4896	0.4861	63.5119	1326.5509	1337.56052	50.65118	.	-0.28645	-0.00708	.	.
2	0.4774	0.4738	71.8211	1333.4034	1344.41300	38.60720	.	.	-0.00531	1.22554	.
2	0.4763	0.4727	72.5422	1333.9905	1345.00010	-6686.59525	.	.	-0.00510	.	2.21777
3	0.5367	0.5318	33.4800	1300.5069	1315.18638	43.46650	.	-0.29404	-0.00520	1.26673	.
3	0.5366	0.5317	33.5266	1300.5497	1315.22922	-6990.03585	.	-0.29659	-0.00496	.	2.31943
3	0.5258	0.5208	40.9043	1307.2525	1321.93202	-9974.73051	.	-0.31398	.	1.70217	3.29895
4	0.5728	0.5668	10.8552	1278.9387	1297.28808	-6186.19926	.	-0.30214	-0.00355	1.12120	2.05252
4	0.5503	0.5440	26.1847	1293.8339	1312.18334	-11278	5.62372	-0.29515	-0.00539	1.21885	.
4	0.5490	0.5426	27.1012	1294.7007	1313.05006	-17441	5.36628	-0.29741	-0.00518	.	2.20344
5	0.5829	0.5755	6.0000	1274.0218	1296.04105	-15654	4.84870	-0.30272	-0.00379	1.08678	1.95592

Figure 6:Subset Models

4. *Forward Stepwise Regression, Forward Selection and Backward Elimination*

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X3		The distance to the nearest MRT station	1	0.4333	0.4333	99.8291	220.24	<.0001
2	X2		The house age	2	0.0563	0.4896	63.5119	31.65	<.0001
3	X4		The number of convenience stores in the living circle on foot (integer)	3	0.0470	0.5367	33.4800	29.04	<.0001
4	INTER_X5_X6		Interaction of the corordinates(latitude and longitude)	4	0.0362	0.5728	10.8552	24.13	<.0001
5	X1		the transaction date	5	0.0101	0.5829	6.0000	6.86	0.0093

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X3	The distance to the nearest MRT station	1	0.4333	0.4333	99.8291	220.24	<.0001
2	X2	The house age	2	0.0563	0.4896	63.5119	31.65	<.0001
3	X4	The number of convenience stores in the living circle on foot (integer)	3	0.0470	0.5367	33.4800	29.04	<.0001
4	INTER_X5_X6	Interaction of the corordinates(latitude and longitude)	4	0.0362	0.5728	10.8552	24.13	<.0001
5	X1	the transaction date	5	0.0101	0.5829	6.0000	6.86	0.0093

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-15654	3827.73438	1325.55866	16.72	<.0001
X1	4.84870	1.85189	543.33006	6.86	0.0093
X2	-0.30272	0.04632	3384.54121	42.70	<.0001
X3	-0.00379	0.00063357	2837.30393	35.80	<.0001
X4	1.08678	0.22611	1830.95009	23.10	<.0001
INTER_X5_X6	1.95592	0.41526	1758.31314	22.18	<.0001

5. *Validation*

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-20562	6808.48465	-3.02	0.0031
X1	the transaction date	1	6.93335	3.13553	2.21	0.0289
X2	The house age	1	-0.24294	0.08251	-2.94	0.0039
X3	The distance to the nearest MRT station	1	-0.00314	0.00133	-2.36	0.0197
X4	The number of convenience stores in the living circle on foot (integer)	1	1.44320	0.40016	3.61	0.0005
INTER_X5_X6	Interaction of the corordinates(latitude and longitude)	1	2.18937	0.84649	2.59	0.0109

6. *Tools and Resources*

Kaggle link for the dataset : Click here

Computations were done using SAS

The computations and full results can be found under my repository in GitHub: Click Here

timeline

Task	Date
Understanding the Problem	27 September 2021 -5 October 2021
Brainstorming ideas	6 October 2021 – 9 October 2021
Planning	11 October 2021 – 14 October 2021
Implement	15 October 2021 – 21 October 2021
Documentation	21 October 2021 – 24 October 2021