Housing Price Prediction Based on Multiple Linear Regression Model

Jia Lin
Student Number: 22117644
Postgraduate Diploma in Science in Data Analytics
National College of Ireland
x22117644@student.ncirl.ie

I. Introduction

Purchasing or investing a house is a crutial decision for both individuals and enterprises. Predicting housing prices has always been a vital challenge in data analytic field. In a common sense, the price of a house is determined by several factors, such as locations, surroundings, house facilities, neighbourhood, and so on. These factors have been generalised as house factors, environmental factors, transportation factors and regional socio-economic factors [1]. Researchers tried to create different models with high accuracy and least error to predict the housing price based on some of these factors.

The contribution of this project is creating a Multiple Linear Regression (MLR) model to predict housing prices based on a provided dataset. The IBM Statistical Package for the Social Sciences (SPSS) is used to analysis the data, conduct the regression procedure and build the model. In particular, Automatic Linear Modelling (ALM) [2] in SPSS is employed to evaluated the final model by comparing with other possible models based on the value of Adjusted R^2.

Some previous work has been done to implement the housing price prediction using statistical regression techniques are machine learning algorithms. (section II). A dataset in CVS format (section III) is provided to build the MLR model. Before creating the model, it is necessary to conduct data preprocessing (section IV). Procedures, like expunging any irrelevant attributes, flagging any erroneous aspects, identifying and fixing any missing value, are implemented to clean the data [3]. When the data is ready to use, the dataset is split into two dataset samples to accomplish the cross-validation process, a larger one for training, another smaller one for testing. In this paper, a MLR model for predicting housing price is built (section V). In order to generalise the sample model to the entire population, several assumptions must be met. Data is trained to satisfy these criteria and violations of any assumption are explored and improved. ALM is used to conduct variable selection and model selection (sectoin VI). Then the selected model is evaluated by applying it to the smaller dataset (section VII). The MLR model is a traditional approach to achieve the prediction, using methodology like Machine Learning (ML) or Deep Leaning (DL) are considered as future work (section VIII).

II. Literature Review

Previous study has been done to create more than one model based on the same dataset [4]. The simple linear. regression, multivariate regression and polynomial regression methods are use do to build a prediction of housing price. Another work payed more attention on identify the determinants of purchasing houses [5]. The multiple regression analysis (MRA) and its extension, hedonic regression analysis were demonstrated and two MLR models were built to illustrate the factors' priority on the process of investment a property. Using artificial intelligence approach to achieve the prediction is a trend, however, the statistics regressions are foundation. Multiple comparisons were made between MRA and artificial neural networks (ANN) [6]. MRA was recommended when the size of dataset is moderate. ANN had better performance, when the dataset is big enough. Machine learning algorithms used to implement prediction based on regression techniques. MLR uses least squares methodology, other regressions like Lasso and Ridge regression models, support vector regression and Extreme Gradient Boost Regression(XGBoost) algorithms are discussed and applied to accomplish the housing price prediction [7]. Machine learning techniques had better performance when applying their algorithms to a large dataset, in which data were collected on ages, in years [8].

III. Description of the dataset

The sample dataset is provided in a CSV format that presents the house prices and characteristics for Seattle and King County, WA (May 2014 - 2015)[1]. This dataset contains 21613 records of various properties and 21 variables.

The *price*, indicates the housing price, is the dependent variable (*DV*) and it is Scientific Notation type. Except the *id*, indicates the identification of each record, other variables are possible independent variables (*IV*s) candidates, off course, some of them need to be transformed, for instance, format of the *date* variable. Only the *data* is the String type, other *IV*s are Numeric type.

Nominal variables are qualitative variables in a model. There are 5 nominal variables whose details are listed in TABLE I. The *bedrooms*, indicates the number of bedrooms, should be a measure of Scale in SPSS, the same measurement of the

[1] https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/

bathrooms, indicates the number of bathrooms. The *waterfront* variable contains two values 0 and 1, which can be viewed as a dummy variable. The rest 3 variables are apparently Ordinal level of measurement which present ranks in order.

TABLE I
Nominal LEVEL OF MEASUREMENT

No.	IV Name	*Description*
1	*bedrooms*	Number of bedrooms
2	*waterfront*	1 if the property has a waterfront, 0 if not.
3	*view*	An index from 0 to 4 of how good the view of the property was
4	*condition*	Condition of the house, ranked from 1 to 5
5	*grade*	Classification by construction quality which refers to the types of materials used and the quality of workmanship. Buildings of better quality (higher grade) cost more to build per unit of measure and command higher value. Additional information in: KingCounty

In addition, there are some variables to present the size of a house, they may correlate from each other. For example, the *sqft_living*, indicates size of living area in square feet, may have relationship with *bedrooms*, *bathrooms*, *floors*, which indicate the number of the corresponding facilities respectively. And the *sqft_living* variable also may show connections with the *sqft_above* variable which indicates square feet above ground, and the *squft_living15* variable which indicates average size of interior housing living space for the closest 15 houses, in square feet. Other similar correlations between variables are concerned as well.

The location factor cannot be ignored when considering the housing price [9], hence, the *lat* and *long* which indicate the latitude and longitude are considered as significant variables in the model.

IV. PRE-PROCESSING OF THE DATA

Before conducting any data analytics, the data should be cleaned and ready for training. Normally, the data of type String, Numeric, in particular, the data of type Date are considered.

1) Change the type of *price* from Scientific Notation to Numeric, and set the decimals to 2.
2) Values of the *date* variable are transformed from String type to Date type in a proper format. By using function of CHAR.SUBSTR(*date*, 1, 8), for example, string “20141013T000000” is changed to “20141013”. Then creating a *date_sold* variable of type Date in the format of yyyy/mm/dd based on the modified date string, e.g. “20141013” is transformed to 2014/10/13.
3) Dealing with Nominal variables. The measure of *bedrooms* is modified to Scale, in order to keep the consistency with the *bathrooms*. The measure of variables *view*, *condition* and *grade* are converted to Ordinal to indicate the rank of specific features.

Then all variables are validated except *id*, based on the validation rule see Fig. 1. The result shows that the *waterfront* variable contains more than 95% constant 0, and the coefficient of variation of values of *zipcode* and *date_sold* less than 0.001(TABLE II and III). These three variables are cut out.

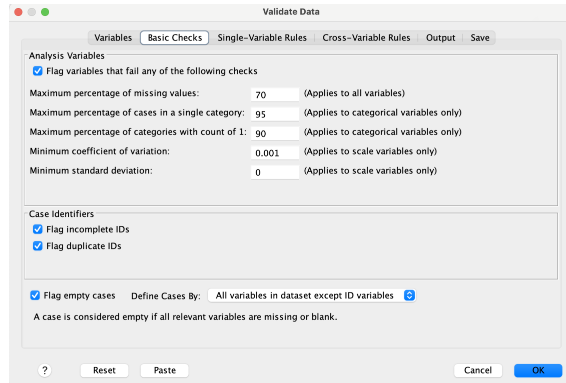

Fig. 1. Validation Rule

TABLE II
VARIABLE CHECKS BASED ON VALIDATION RULE

Variable Checks

Categorical	Cases Constant > 95	1 if the property has a waterfront, 0 if not
Scale	Coefficient Of Variation < 0.001	5 digit zip code

Each variable is reported with every check it fails.

TABLE III
VARIABLE CHECKS BASED ON VALIDATION RULE

Variable Checks

Scale	Coefficient Of Variation < 0.001	Format Date

Each variable is reported with every check it fails.

There is no empty case, no incomplete identification, but 176 groups of duplicated instances. The duplicated records indicate that the exactly same house was sale to 2 or 3 times on different prices. These instances are deleted and stored to another dataset called **Duplicated_IDs.sav**. After deleting the duplicated records, the original sample dataset has 21260 rows, and the pool of candidates *IV*s contains 17 variables.

In order to cross validate the MLR model, the dataset has been divided into two random samples, 80% for training and 20% for validation. The larger dataset contains 17050 records, and the smaller one is compose of 4210 instances.

Before creating the MLR model, the descriptive statistics table of *DV* and candidates *IV*s (TABLE IV), shows that *DV* and most *IV*s don’t have a normal distribution. By drawing the Boxplot Dialogs of each variable, most of them have outliers. Take the *DV* *price* for example, see Fig. 2, *price* has to be transformed to a new variable named *ln_price*. Values of *ln_price* are natural logarithm of values of *price* correspondingly. After the transformation, the new *DV* is normal distributed, see Fig. 3.

TABLE IV
DESCRIPTIVE STATISTICS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Sale price	17050	75000.00	7700000.00	539953.6096	366570.581	4.181	.019	37.917	.038
Number of bedrooms	17050	0	33	3.37	.932	2.352	.019	61.377	.038
Number of bathrooms	17050	0	8	2.12	.768	.531	.019	1.479	.038
Size of living area in square feet	17050	290	13540	2080.75	917.743	1.519	.019	5.843	.038
Size of the lot in square feet	17050	520	1651359	15157.43	41719.829	13.382	.019	303.093	.038
Number of floors	17050	1.0	3.5	1.497	.5406	.609	.019	-.486	.038
An index from 0-4, how good the view of the property	17050	0	4	.24	.767	3.363	.019	10.684	.038
Condition of the house, ranked from 1 to 5	17050	1	5	3.41	.648	1.031	.019	.537	.038
Classification by construction quality	17050	1	13	7.66	1.171	.780	.019	1.281	.038
Square feet above ground	17050	290	9410	1792.04	829.539	1.484	.019	3.768	.038
Square feet below groud	17050	0	4820	288.71	439.768	1.577	.019	2.739	.038
Year built	17050	1900	2015	1971.35	29.340	-.480	.019	-.648	.038
Year renovated. 0 if never renovated	17050	0	2015	83.94	400.664	4.564	.019	18.835	.038
Latitude	17050	47.1593	47.7776	47.559599	.1388244	-.477	.019	-.702	.038
Longitude	17050	-122.519	-121.315	-122.21328	.141426	.884	.019	1.083	.038
Average size of interior housing living space for the closest 15 houses, in square feet	17050	399	6210	1988.88	682.824	1.092	.019	1.563	.038
Average size of land lots for the closest 15 houses, in square feet	17050	651	871200	12832.93	27918.264	9.817	.019	162.037	.038
Valid N (listwise)	17050								

Fig. 2. Boxplot of *DV price*

The aim of making a MLR model is to generalise the sample model to the entire population. In order to accomplish this goal, several assumptions must be met. These criteria are used to refine candidates *IV*s, if any *IV* break any assumption, the corresponding variable is either considered to be transformed or it will be dropped. Variables selection will be execute step by step in the process of building the MLR model.

The first and most important assumption is Gauss-Markov

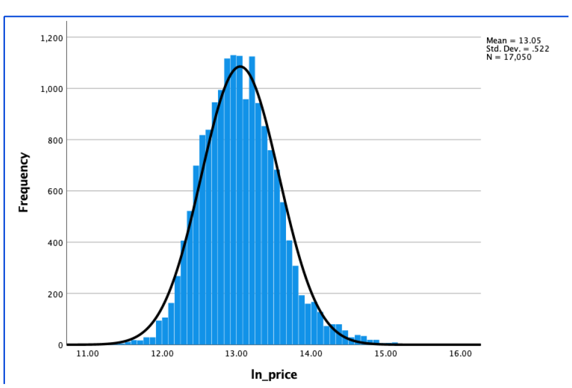

Fig. 3. Histogram of new *DV ln_price*

assumptions. The model is built on the Ordinary Least Squares(OLS) method. If every *IV* meets the Gauss-Markov assumptions, the OLS estimates are BLUE, in which BLUE stands for best linear unbiased estimator.

- Linearity. Correct functional form for our model.
- No multicollinearity between independent variables.
- Predictor variables must be independent of the error term.
- Homoscedasticity, errors have constant variance.
- Errors are normally distributed.
- No influential data points.

V. THE MULTIPLE LINEAR REGRESSION MODEL BUILDING PROCESS

After data pre-processing, in particular, the *DV price* has been transformed to *ln_price*, A MLR model is build as below:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \qquad (1)$$

in which,

- $\ln(Y)$: The response or dependent variable *ln_price*.
- X_p : The predictor or independent variable.
- β_0 : The intercept (constant term).
- β_p : The slope of each X.
- ϵ : The residuals (error terms).

Variable selection is in the procedure of identifying the diagnostics and check all assumptions:

Assumption 1: The relationship between the *IV*s and the *DV ln_price* is linear.

Scatter Dialogs are created between each *IV* and the *DV ln_price*. These scatter plots show these all *IV*s are acceptable at this stage.

Assumption 2: No multicollinearity between *IV*s.

Multicollinearity shows two or more *IV*s are functions of each other. If multicollinearity exists, these two predictors must be strongly correlated, although their correlation cannot determine that they have multicollinearity with one another. The correlation between *DV* and *IV*s (TABLE V) shows the *sqft_living* has a strong correlation with the the *sqft_above* with the value of 0.878. In addition, the correlations between *DV* and some *IV*s are very weak. Any correlation value less than .1 is considerd to be dropped in further analysis.

The variance inflation factor (VIF) test gives us a formal method to present how much common variance exists between *IV*s. If an *IV* has a VIF of 5 or above then it will likely be collinear with others. All VIF value in TABLE VI are smaller than 5. However, the Sig. value of the *sqft_lot15* is greater than .05, indicates this *IV* does't have big contributions. It can be dropped. In addition, a standardised coefficient beta value presents strength of the effect from each *IV* to *DV*.

By observing the corresponding correlations between each *IV* and *DV* (TABLE V), some *IV*s do represent very weak association to *DV*. These variables will be dropped at this stage. They are *IV*s called *bedrooms*, *sqft_lot*, *floor*, *condition* and *yr_renovated*.

Additionally, the parameters value of *sqft_above*, *sqft_basement* and *sqft_living15* are 0 (TABLE VI), the

TABLE V
SEGMENT OF CORRELATION BETWEEN *DV* AND *IV*S

Correlations

		ln_price	Number of bedrooms	Number of bathrooms	Size of living area in square feet	Size c in sq
Pearson Correlation	ln_price	1.000	.345	.547	.692	
	Number of bedrooms	.345	1.000	.516	.576	
	Number of bathrooms	.547	.516	1.000	.753	
	Size of living area in square feet	.692	.576	.753	1.000	
	Size of the lot in square feet	.097	.035	.083	.166	
	Number of floors	.304	.175	.499	.349	
	An index from 0–4, how good the view of the property	.347	.080	.183	.282	
	Condition of the house, ranked from 1 to 5	.037	.019	-.137	-.070	
	Classification by construction quality	.700	.360	.662	.761	
	Square feet above ground	.600	.478	.683	.878	
	Square feet below groud	.311	.302	.283	.431	
	Year built	.071	.155	.503	.312	
	Year renovated. 0 if never renovated	.116	.022	.056	.062	
	Latitude	.445	-.011	.018	.044	
	Longitude	.044	.130	.220	.236	
	Average size of interior housing living space for the closest 15 houses, in square feet	.615	.394	.564	.753	
	Average size of land lots for the closest 15 houses, in square feet	.090	.029	.084	.178	

MLR model wasn't able to find a linear relationship between the *DV* and these *IV*s. After practices on these variables, *sqft_above* and *sqft_living15* are transformed to their natural logarithm *ln_sqft_above* and *ln_sqft_living15*. The *sqft_basement* is dropped. After transforming and deleting some *IV*s, a new MLR model is built based on the left *IV*s.

TABLE VI
MODEL COEFFICIENTS

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	-55.698	2.244		-24.816	<.001		
	Number of bedrooms	-.012	.003	-.021	-4.329	<.001	.611	1.636
	Number of bathrooms	.073	.005	.107	15.701	<.001	.299	3.349
	Size of the lot in square feet	4.621E-7	.000	.037	6.714	<.001	.459	2.180
	Number of floors	.069	.005	.071	13.531	<.001	.500	1.999
	An index from 0–4, how good the view of the property	.075	.003	.111	26.910	<.001	.822	1.216
	Condition of the house, ranked from 1 to 5	.069	.003	.086	20.582	<.001	.803	1.246
	Classification by construction quality	.155	.003	.348	50.818	.000	.296	3.373
	Square feet above ground	.000	.000	.217	26.443	<.001	.206	4.850
	Square feet below groud	.000	.000	.119	22.679	<.001	.501	1.996
	Year built	-.003	.000	-.180	-31.100	<.001	.416	2.405
	Year renovated. 0 if never renovated	4.217E-5	.000	.032	8.109	<.001	.871	1.148
	Latitude	1.349	.015	.359	90.890	.000	.891	1.122
	Longitude	-.071	.017	-.019	-4.249	<.001	.668	1.496
	Average size of interior housing living space for the closest 15 houses, in square feet	.000	.000	.139	22.020	<.001	.346	2.892
	Average size of land lots for the closest 15 houses, in square feet	-1.748E-7	.000	-.009	-1.687	.092	.452	2.211

a. Dependent Variable: ln_price

Assumption 3: The residuals or error terms are independent.

The Durbin-Watson value (TABLE VII) is close to 2 (Durbin-Watson = 1.990), indicates that this assumption has been met.

Assumption 4: The variance of the residuals is constant. The output graph shows the ZRESID (standardised residuals) against ZPRED (standardised predicted values), see Fig. 4. This plot doesn't present a funnel shape, which indicates the homoscedasticity assumption has been met. Error terms in this model are similar at each point of the model.

TABLE VII
DURBIN-WATSON STATISTICS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.866[a]	.751	.750	.26084	1.990

a. Predictors: (Constant), ln_sqft_living15, Latitude, An index from 0-4, how good the view of the property, Year built, Longitude, Number of bathrooms, Classification by construction quality , ln_sqft_above

b. Dependent Variable: ln_price

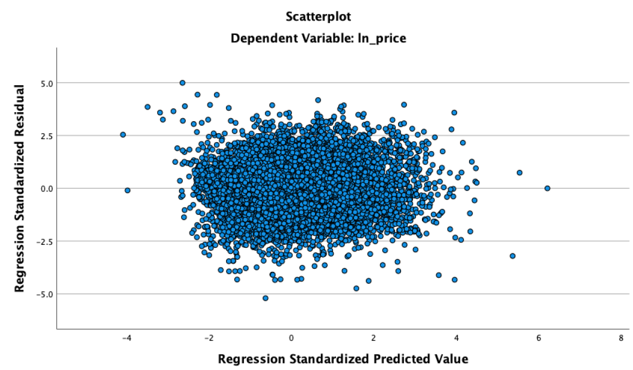

Fig. 4. The Output Graph of ZRESID Vs. ZPRED

Assumption 5: The residuals are normally distributed.

The P-P plot of this model shows this assumption has been met. Dots are very close to the diagonal line, the residuals are normal distributed, see Fig. 5.

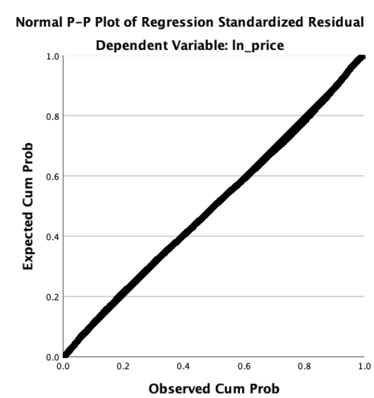

Fig. 5. The P-P plot for the Model

Assumption 6: There are no influential cases.

This assumption is met, the maximum of Cook's Distance value is under 1 (TABLE VIII), which indicates individual cases are not influencing the model.

Then the coefficients are determined (TABLE IX) and the final MLR model is as below, see equation (2):

TABLE VIII
RESIDUALS STATISTICS

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	11.2009	15.8569	13.0497	.45231	17050
Std. Predicted Value	-4.088	6.206	.000	1.000	17050
Standard Error of Predicted Value	.002	.022	.006	.002	17050
Adjusted Predicted Value	11.1980	15.8569	13.0497	.45232	17050
Residual	-1.35713	1.30286	.00000	.26078	17050
Std. Residual	-5.203	4.995	.000	1.000	17050
Stud. Residual	-5.204	4.998	.000	1.000	17050
Deleted Residual	-1.35783	1.30459	.00000	.26094	17050
Stud. Deleted Residual	-5.208	5.002	.000	1.000	17050
Mahal. Distance	.384	117.132	8.000	6.460	17050
Cook's Distance	.000	.010	.000	.000	17050
Centered Leverage Value	.000	.007	.000	.000	17050

a. Dependent Variable: ln_price

TABLE IX
FINAL MODEL COEFFICIENTS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-57.323	2.208		-25.965	<.001		
	Number of bathrooms	.128	.004	.188	32.239	<.001	.429	2.333
	An index from 0-4, how good the view of the property	.090	.003	.132	31.902	<.001	.861	1.161
	Classification by construction quality	.174	.003	.391	58.741	.000	.330	3.028
	Year built	-.004	.000	-.247	-49.613	.000	.592	1.688
	Latitude	1.345	.015	.358	89.469	.000	.916	1.092
	Longitude	-.082	.017	-.022	-4.955	<.001	.720	1.388
	ln_sqft_above	.189	.009	.155	22.173	<.001	.301	3.324
	ln_sqft_living15	.257	.010	.161	26.805	<.001	.406	2.460

a. Dependent Variable: ln_price

$$\ln(\hat{y}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_7 \ln(x_7) + \beta_8 \ln(x_8) \quad (2)$$

in which,

- $\ln(\hat{y})$: The estimate value of *ln_price*
- $\beta_0 = -57.323$
- $\beta_1 = .128$ and x_1 is value of *bathrooms*
- $\beta_2 = .090$ and x_2 is value of *view*
- $\beta_3 = .174$ and x_3 is value of *grade*
- $\beta_4 = -.004$ and x_4 is value of *yr_built*
- $\beta_5 = 1.345$ and x_5 is value of *lat*
- $\beta_6 = -.082$ and x_6 is value of *long*
- $\beta_7 = .189$ and $\ln(x_7)$ is value of *ln_sqft_above*
- $\beta_8 = .257$ and $\ln(x_8)$ is value of *ln_sqft_living15*

VI. AUTOMATIC LINEAR MODELLING

Automatic Linear Modelling (ALM) is introduced in Version 19 of IBM SPSS. It allows researchers to identify the best subset of predictors automatically [10]. The number of possible MLR models depends on the number of predictors. If there are p predictors, there will be 2^p possible models. Ideally, every model can be carried out and compute least squares for all possible subsets and choose the best one based on some criteria. ALM provide the function to identify the best subset of predictors, it provided three model fit options: Akaike's Criterion Information Corrected(AIC_c), Adjusted R^2, and Overfit Prevention Criterion (ASE).

In this paper, the Adjusted R^2 is used to be the evaluation criterion and the best subset is chosen as the model selection method. The result shows that the model with selected predictors has 75% of accuracy, see Fig. 6. And the ALM lists the top 10 possible models contains these *IV*s. The one with all of these predictors is the best one with the biggest value of Adjusted R^2, see Fig. 7.

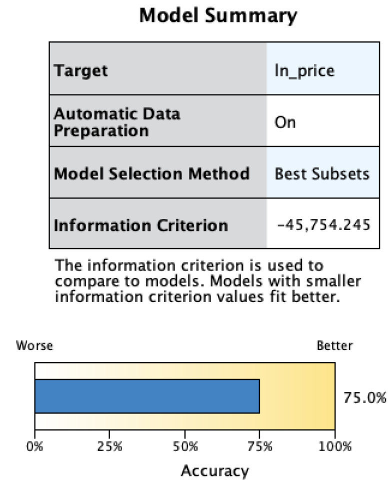

Fig. 6. The ALM Model Summary

Model ...

Target: ...

		Model									
		1	2	3	4	5	6	7	8	9	10
Adjusted R Square		.750	.749	.742	.742	.740	.740	.734	.733	.732	.731
Effect	view_transformed	✓	✓	✓	✓	✓	✓	✓	✓		
	bathrooms_transformed	✓	✓	✓	✓	✓	✓			✓	✓
	yr_built_transformed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	lat_transformed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	long_transformed	✓			✓	✓		✓		✓	
	ln_sqft_above_transformed	✓	✓			✓	✓	✓	✓	✓	✓
	ln_sqft_living15_transformed	✓	✓	✓	✓			✓	✓	✓	✓
	grade_transformed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The model building method is Best Subsets using the Adjusted R Square criterion.
A checkmark means the effect is in the model.

Fig. 7. ALM Model Selection

VII. CROSS-VALIDATION

Cross-Validation, also called Out-of-Sample Test, is used to generalise a statistical analysis to a independent dataset. After pre-processing, the original data set was divided to two sample datasets. The larger one, has 17050 records, is used to train the data and create a MLR model. The smaller one, which is composed of of 4210 instances, is used to evaluate the model. By setting a *filter_$* variable, whose value is either 1 or 0. The value 1, indicates the corresponding record is in the training sample set. Otherwise, it is a test sample instance.

By applying the MLR model to this test sample, we could compare the residuals statistics result between two samples (TABLE X). The mean and standard deviation of predicted values of two samples are similar to each other. The mean and standard deviation of residuals of these two dataset are close as well. By presenting their P-P plots (Fig. 8) side by side, two plots are similar to one another. The output graphs of ZRESID Vs. ZPRED are nearly the same shape (Fig. 9).

TABLE X
THE CROSS VALIDATION BETWEEN TRAINING AND TEST SAMPLES

Residuals Statistics[a,b]

	filter_$ = 1 (FILTER) = Not Selected (Selected)					filter_$ = 1 (FILTER) ~= Not Selected (Unselected)				
	Minimum	Maximum	Mean	Std. Deviation	N	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	11.8642	15.6091	13.0642	.46479	4210	11.1028	15.9036	13.0530	.45818	17050
Std. Predicted Value	-2.582	5.475	.000	1.000	4210	-4.220	6.109	-.024	.986	17050
Standard Error of Predicted Value	.005	.036	.012	.003	4210	.005	.043	.012	.003	17050
Adjusted Predicted Value	11.8622	15.6195	13.0642	.46482	4210	11.1028	15.9036	13.0530	.45818	17050
Residual	-1.39026	1.15004	.00000	.25958	4210	-1.37093	1.34511	-.00329	.26122	17050
Std. Residual	-5.351	4.426	.000	.999	4210	-5.276	5.177	-.013	1.005	17050
Stud. Residual	-5.364	4.432	.000	1.000	4210	-5.271	5.163	-.013	1.004	17050
Deleted Residual	-1.39720	1.15309	.00000	.26021	4210	-1.37093	1.34511	-.00329	.26122	17050
Stud. Deleted Residual	-5.382	4.442	.000	1.001	4210	-5.271	5.163	-.013	1.004	17050
Mahal. Distance	.631	80.978	7.998	6.016	4210	.388	113.165	8.025	6.442	17050
Cook's Distance	.000	.017	.000	.001	4210	.000	.045	.000	.001	17050
Centered Leverage Value	.000	.019	.002	.001	4210	.000	.027	.002	.002	17050

a. Dependent Variable: ln_price
b. Pooled Cases

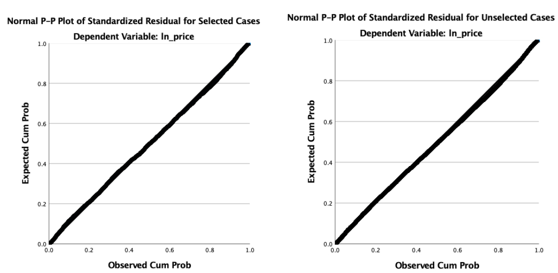

Fig. 8. P-P plots of two sample datasets side by side.

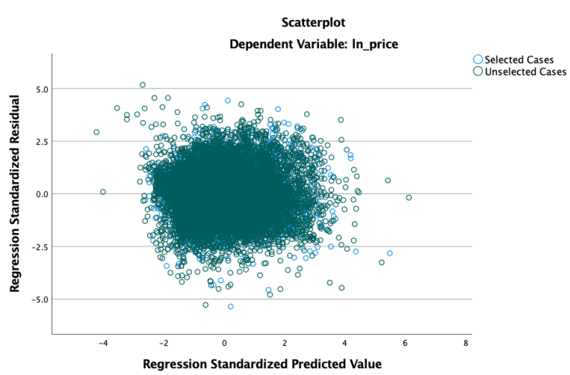

Fig. 9. Cross Validation of the Output Graphs of ZRESID Vs. ZPRED

Finally, by comparing the residuals' standard deviation, 0.25983, of the testing sample (TABLE XI) with the residuals' standard deviation, 0.26084, of the training sample (TABLE VII), they are very close, only minor change. This fact illustrates that the MLR model has been built on the training sample can be generalised to predict housing price.

TABLE XI
CROSS VALIDATION MODEL SUMMARY

Model Summary[b,c]

	R					Durbin-Watson Statistic	
Model	filter_$ = 1 (FILTER) = Not Selected (Selected)	filter_$ = 1 (FILTER) ~= Not Selected (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate	filter_$ = 1 (FILTER) = Not Selected (Selected)	filter_$ = 1 (FILTER) ~= Not Selected (Unselected)
1	.873[a]	.866	.762	.762	.25983	1.935	1.990

a. Predictors: (Constant), ln_sqft_living15, Latitude, An index from 0-4, how good the view of the property, Year built, Longitude, Number of bathrooms, Classification by construction quality , ln_sqft_above
b. Unless noted otherwise, statistics are based only on cases for which filter_$ = 1 (FILTER) = Not Selected.
c. Dependent Variable: ln_price

VIII. CONCLUSION AND FUTURE WORK

In terms of the result of ALM in SPSS, the MLR model created in this paper is good enough to predict housing price in general. The model could be refined by transforming more predictors, or consider about the interaction term of two variables. For example, the product of *lat* and *long*. *lat* and *long* are used to determine the location of a house which is the most important factor to affect housing price. The future work is using Machine Learning, even Deep Leaning techniques to improve the accuracy of the prediction model and deal with larger datasets

ACKNOWLEDGMENT

I would like to thank my lecturers Hicham Rifai and Tony Delaney for their patience and detailed explanation.

REFERENCES

[1] H. Kusan, O. Aytekin, and I. Ozdemir, "The use of fuzzy logic in predicting house selling price" Expert Systems with Applications, vol. 37, no. 3, pp. 1808–1813, 2010.

[2] H. Yang, "The case for being automatic: introducing the automatic linear modelling (LINEAR) procedure in SPSS statistics". Multiple Linear Regression Viewpoints, 39(2), pp.27-37, 2013.

[3] M. Haider, "Getting Started with Data Science: Making Sense of Data with Analytics". IBM Press, 2015.

[4] R. Manjula, J. Shubham, S. Sharad, and R. K. Pranav. "Real estate value prediction using multivariate regression models." In IOP Conference Series: Materials Science and Engineering, vol. 263, no. 4, p. 042098. IOP Publishing, 2017.

[5] A. M. Yusof, and S. Ismail, "Multiple regressions in analysing house price variations". Communications of the IBIMA, p.1, 2012.

[6] N. Nghiep, and C. Al, "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks". Journal of real estate research, 22(3), pp.313-336, 2001.

[7] J. Manasa, R. Gupta. and N. S. Narahari, 2020, "Machine learning based predicting house prices using regression techniques". In 2nd International conference on innovative mechanisms for industry applications (ICIMIA) (pp. 624-630). IEEE. March 2020.

[8] W. K. Ho, B.S. Tang and S.W. Wong. "Predicting property prices with machine learning algorithms". Journal of Property Research, 38(1), pp.48-70, 2021.

[9] Q. Zhang, "Housing Price Prediction Based on Multiple Linear Regression". Scientific Programming, 2021.

[10] T. C. Oshima, and T. Dell-Ross. "All possible regressions using IBM SPSS: A practitioner's guide to automatic linear modelling". In Georgia educational research association conference. 2016.