2022

PREDICTIVE MODEL FOR HOUSE PRICE

BAN210 PREDICTIVE ANALYTICS – PROFESSOR JI QI

THI DIEM HUONG LE – ID 166792218

SENECA COLLEGE | FALL 2022

1. Objective

This paper is about creating a predictive model for house price from a range of variables including numeric variables and categorical variables.

Its objective is to run two models by Forward and Backward method based on Linear Regression algorithm.

The project outcomes will be helpful to organizations looking for an affordable real estate or understanding price range of a location or gaining insights for price negotiation with sellers.

2. Project workflow

To build an effective predictive model, we will go through the following steps:

1- Data uploading
2- Data cleaning and features engineering
3- Exploratory data analysis
4- Model building
5- Validation framework

3. Data uploading

The dataset is uploaded on SAS with a total of 31 variables, 22 numeric variables and 9 categorical variables. It has around 300 observations. SalePrice is target variables and others are independent ones.

4. Data cleaning and features engineering

Since we have a whole range of variables, we will deal with numerical variables first and then character ones. Next, we will handle missing values and outliers at last.

1- Numerical variables

We run proc corr to check correlation between numeric variables and target variable SalePrice and select the following 11 variables that have correlation higher than 50%:

Pearson Correlation Coefficient											
	Overall_Qual	Year_Built	Gr_Liv_Area	Garage_Area	Basement_Area	Full_Bathroom	Total_Bathroom	Age_Sold	Overall_Qual2	Log_Price	Bonus
Sale Price	0.73	0.61	0.65	0.57	0.68	0.58	0.60	-0.61	0.59	0.96	0.68

2- Categorical variables

Variable House_Style2 is selected for this part and dummy coding will be assigned to each category of this variable.

There are 5 types with below frequency.

	Frequency	Percent
1.5Fin	**28**	**9.33**
1Story	**198**	**66**
2Story	**40**	**13.33**
SFoyer	**13**	**4.33**
SLvl	**21**	**7**

It shows that there are several observations in each type, with 1Story with highest of 66 and SLvl with the lowest of 7, hence, we should keep all types in the model. Dummy codes are assigned:

- 1Story – 1
- 1.5Fin – 2
- 2Story – 3
- SFoyer – 4
- SLvl – 5

3- Missing values

We use proc means to check numeric variables and there are no missing values.

We use proc freq for remaining character variables and there are few missing values as followed:

Variable Name	Number of missing or NA values
Garage_Type_2	**32**
Masonry_Veneer	**2**
Lot_Shape_2	**1**
score	**300**

Mode will be used to replace these missing or NA values. These variables are categorical type so it does not make sense to replace with median or mean. For score variable, it will be dropped as it does not contain any data points.

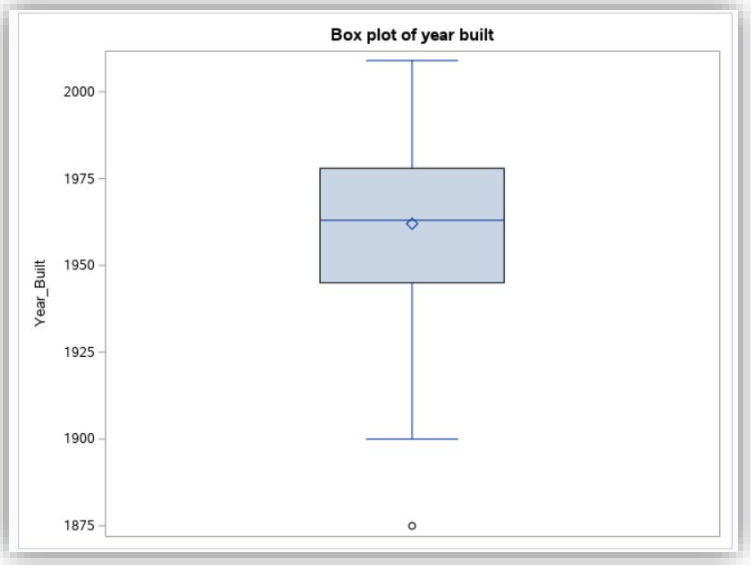
4- Drop variables

We will drop the following variables from dataset:

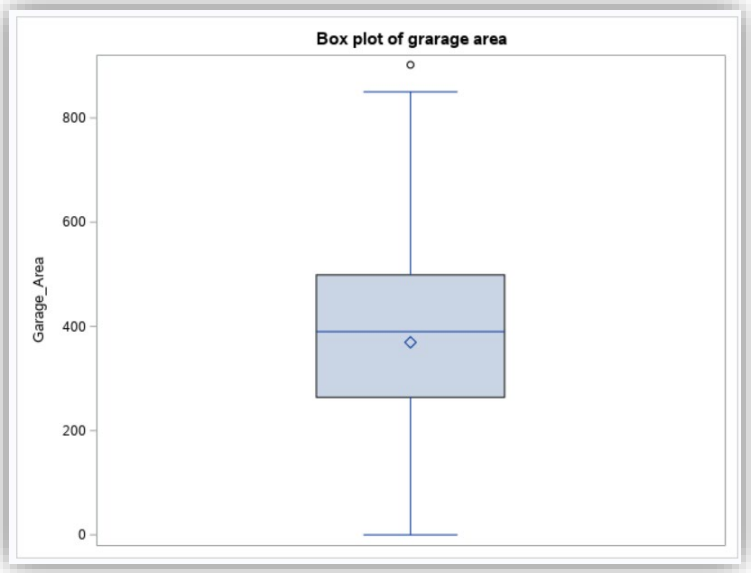
- PID: as it is just ID number
- House_Style: as it is included and already examined in variable House_Style2
- Score: as mentioned above that it does not have any values
- Log_price: as we don't want to leak this data while trying to predict price.

5- Outliers

We will use graphs to check if there are any potential outliers and then use Interquartile range method to exclude them. The following graphs show outliers in variables: Year_built, Gr_Lv_Area, Garage_Area.


Box plot of year built
Year_Built
2000
1975
1950
1925
1900
1875

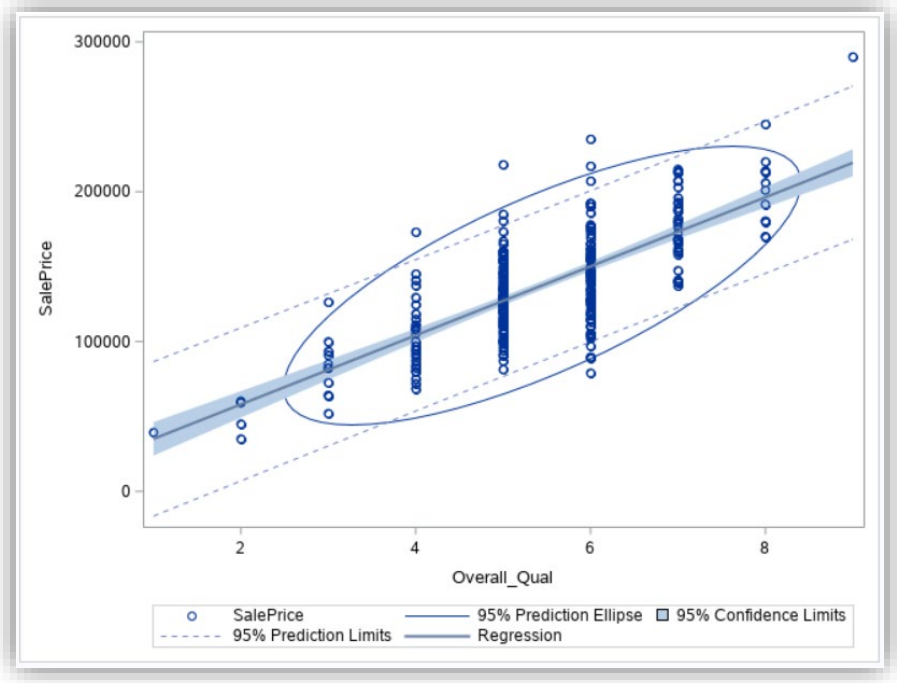

Box plot of living area
Gr_Liv_Area
1500
1250
1000
750
500


Box plot of grarage area
Garage_Area
800
600
400
200
0

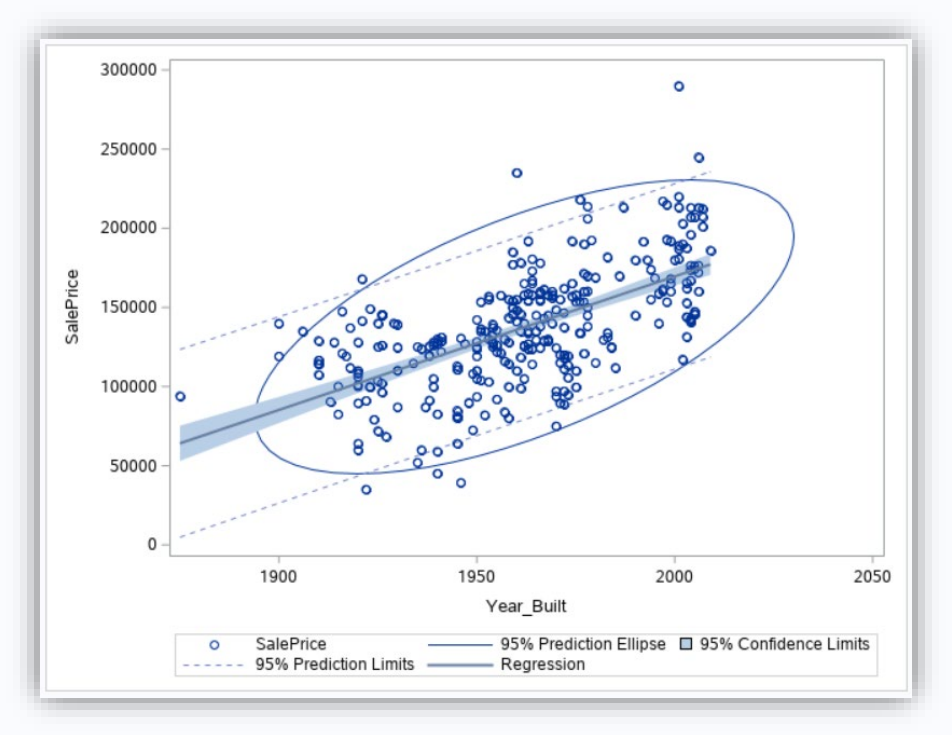
5. Exploratory data analysis

Now that the dataset is clean, we can explore our data using scatter plot and ellipse function for selected numerical variables.

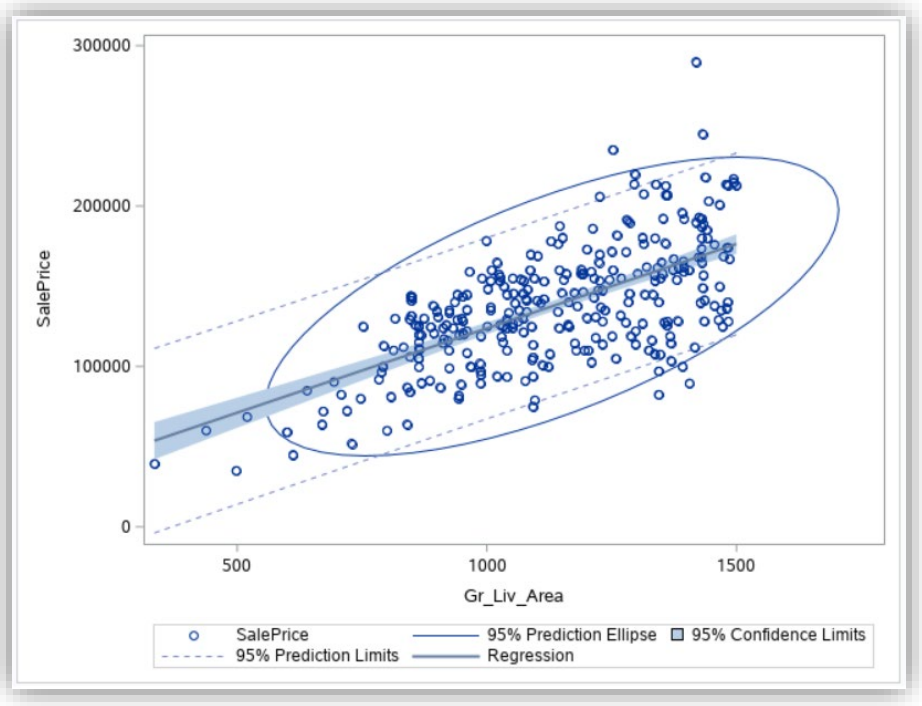
Correlation between Sale Price and Overall_Qual – Positive linear relationship



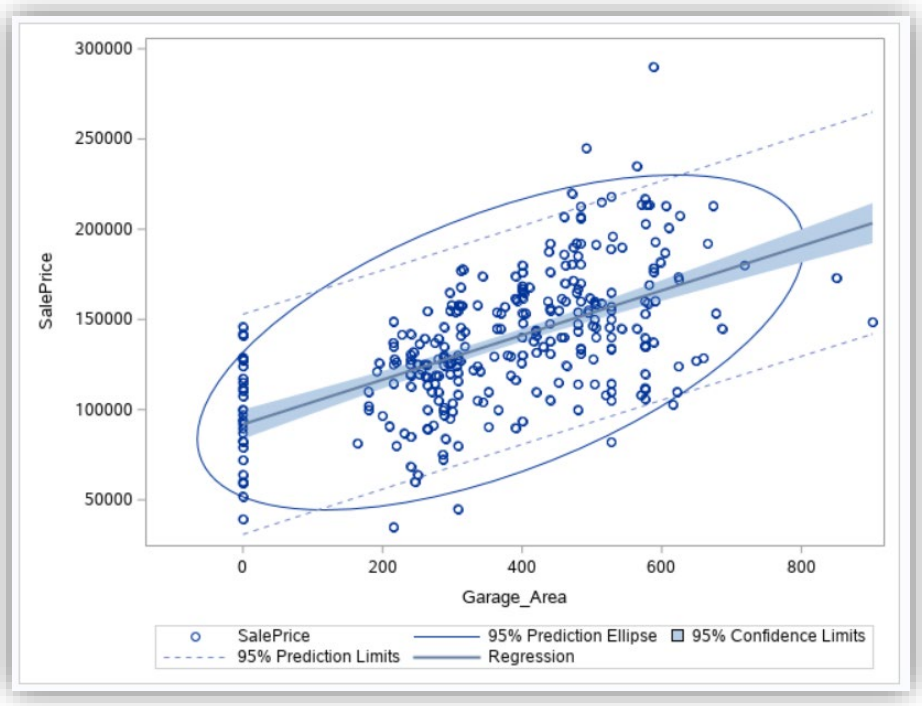
Correlation between Sale Price and Year_Built – Positive linear relationship



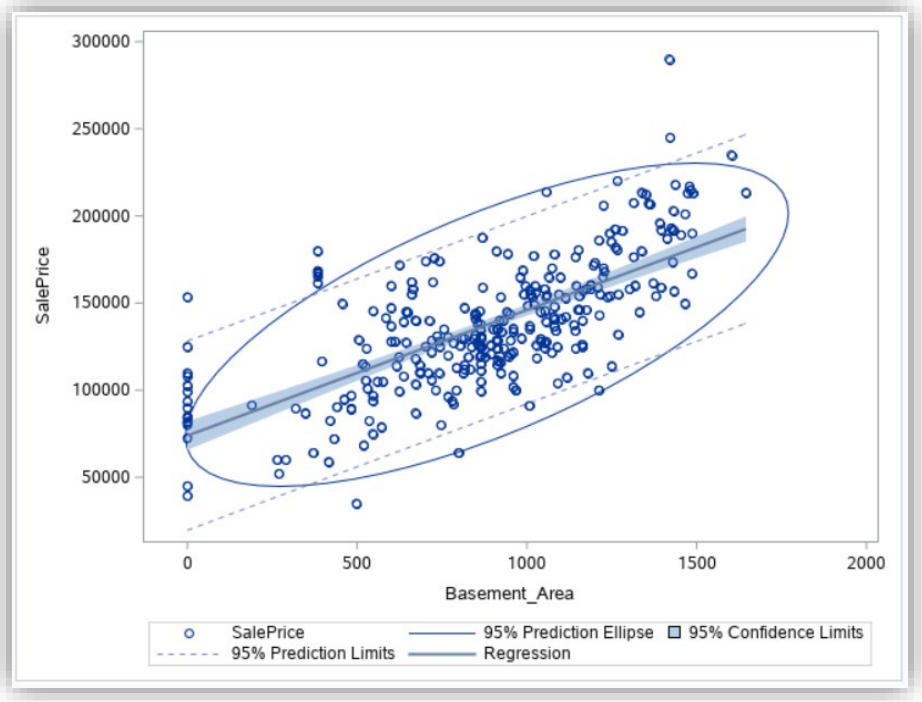
Correlation between Sale Price and Gr_Liv_Area – Positive linear relationship



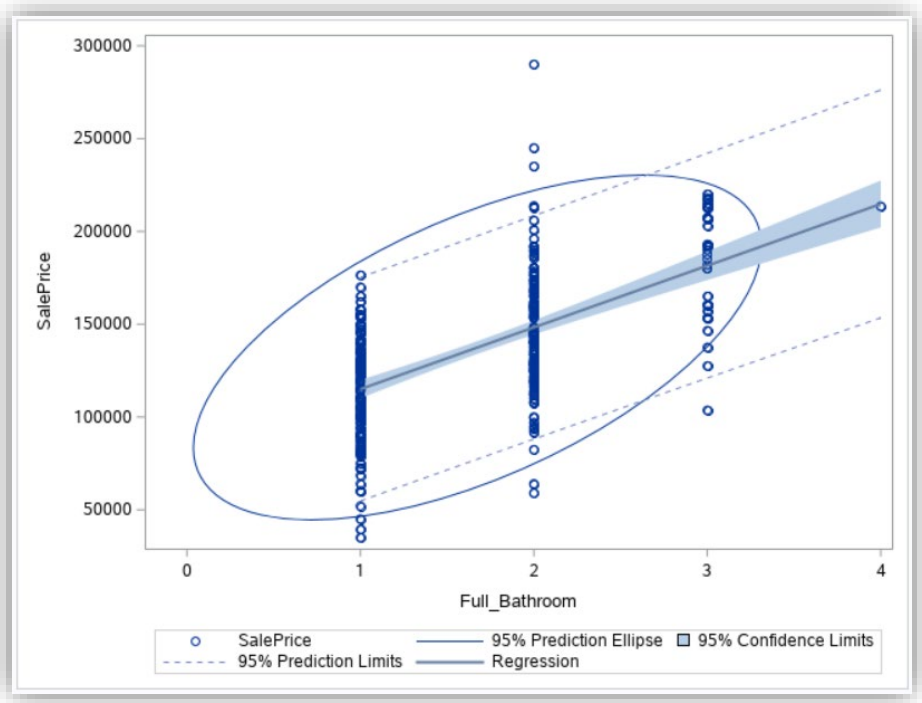
Correlation between Sale Price and Garage_Area – Positive linear relationship



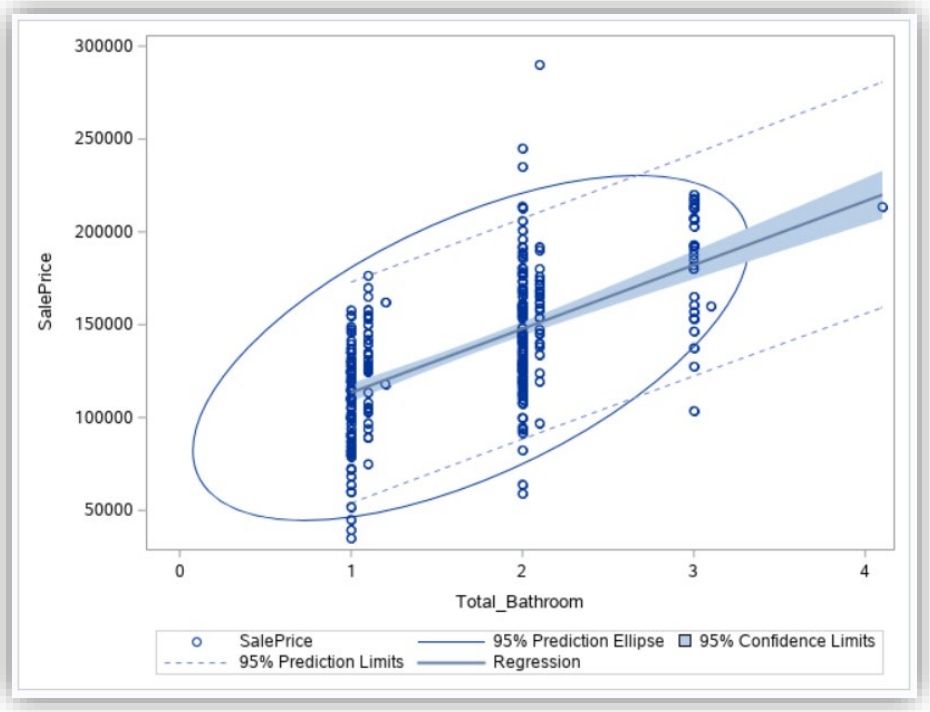
Correlation between Sale Price and Basement_Area – Positive linear relationship



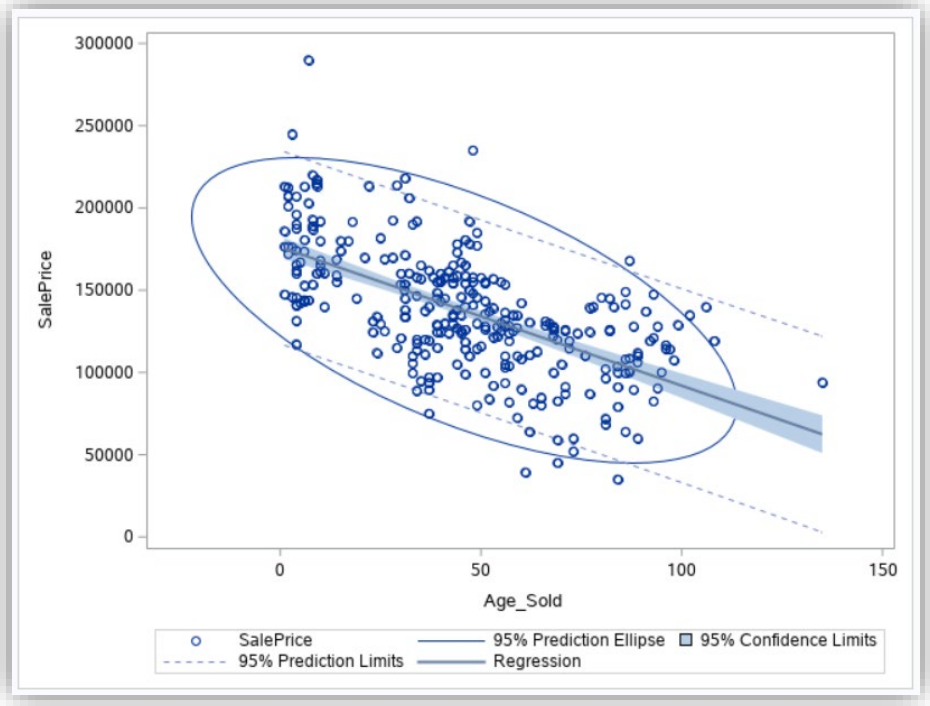
Correlation between Sale Price and Full_Bathroom – Positive linear relationship



Correlation Between Sale Price and Total_Bathroom – Positive linear relationship



Correlation between Sale Price and Age_Sold – Negative linear relationship



6. Dataset splitting

We split dataset into two parts by 80:20 ratio. 80% data (238 observations) is for training model and 20% (60observations) is for testing purpose.

7. Model building

We will use Linear regression algorithms and specifically, Forward and Backward to build two models and the one with higher R-square will be selected.

Model from **Forward** method is followed:

Variable	Parameter
Intercept	-414521
Overall_Qual	7617.38666
Year_Build	217.06169
Gr_Liv_Area	48.74540
Garage_Area	29.01368
Basement_Area	20.48334
Bonus	20924
House_Style2_2	-7226.36988
House_Style2_3	-3969.20786
House_Style2_5	1745.15722

The **first model** is statistically significant **with p-value <0.0001 and R-square = 0.8599**.

Model from **Backward** method is followed:

Variable	Parameter
Intercept	-389251
Overall_Qual	7696.35483
Year_Build	207.81594
Gr_Liv_Area	49.75345
Garage_Area	26.98722
Basement_Area	20.48407
Bonus	21168
House_Style2_1	-8380.72288
House_Style2_2	-11438
House_Style2_3	-6719.12753

The **second model** is statistically significant **with p-value <0.0001 and R-square = 0.8614**, which is higher and more reliable. Hence, we will choose **second model** over first one.

8. Validation framework

Now we can test the accuracy of the model by running it through testing dataset.

The result shows that selected model is able to predict up to 70% of house price with plus/minus error. Hence, dependent on the objective of a business, the model can be utilised a high level of accuracy or can be further trained to increase performance.

Prediction_Grade	Percent
Grade 1 (within 10% error compared to Actual Price)	70
Grade 2 (other than Grade 1)	30