University of Illinois Urbana-Champaign

Department of Statistics

Taiwan Housing Market

Author:
Yvonne Ho

A project report submitted for the course of

STAT-425

December 12, 2021

Contents

1 Introduction

The house price is influenced by many factors, such as location, level of convenience and age of house. In this project, different statistical modeling methods were used to analyze and predict the house price, from the dataset of real estates from Sindian District, New Taipei City, Taiwan, whcih can be found at the UCI Machine Learning Repository [1].

Table 1.1: Data Directory

Abbreviation	Decription
X1	Transaction Date between August 2012 and July 2013
X2	House Age (unit: year)
X3	Distance to the Nearest MRT Station (unit: meters)
X4	Number of Convenience Stores in the Living Circle on Foot (integer)
X5	Geographic Coordinate,Latitude (unit: degree)
X6	Geographical Coordinate, Longitude (unit: degree)
X7	Transaction Years
X8	Transaction Months
Y	House Price of Unit Area ((10000 New Taiwan Dol- lar/Ping; 1 Ping = 3.3 square meters))

Table 1.1 summarized all the variables, including their abbreviations and corresponding explanations. In this dataset, the house price is the response and the rest of the variables are the predictors.The transaction months (X7) and years (X8) were extracted from the transaction date (X1), according to its corresponding months and years, to analyze the relationships among the house price and the transaction date with associated months and years.

2 Preliminary Analysis

2.1 House Price in Different Time Formats

Figure 2.1 is a boxplot and shows the relationships bwtween house price and different transaction date formats.

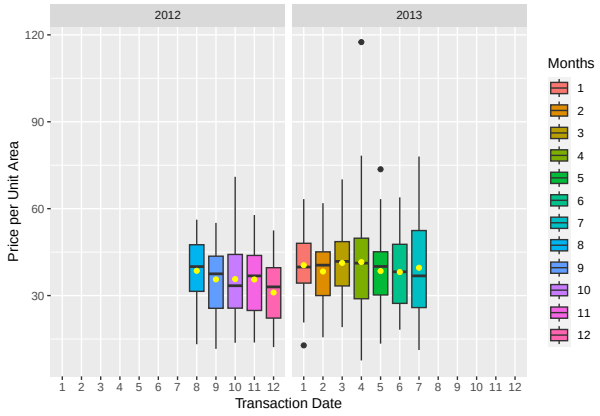

Figure 2.1: House Price vs Transaction Date

In Figure 2.1, the transaction date (X1) is a continuous data from August, 2012, to July, 2013, after grouping by the associated months and years. The mean of house price in each month with corresponding year is not significantly fluctuated, while its variances per month are different. Although two outliers are in January, April,

[1]UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

and May, 2013, the distribution of house price does not show particular pattern and trend in the transaction date. Also, the transaction month (X7) and year (X8) were extracted from the transaction date (X1), so the transaction date (X1) can concretely represent the transaction months (X7) and years (X8), and the generated variables would be ignored in the further analyses.

2.2 Multiple Comparisons of Pairwise Differences on Transaction Date

The transaction date (X1) combined the transaction months (X7) and years (X8). When it is regarded as a categorical data, the transaction date (X1) in Figure 2.1 shows how different house price changed by months from 2012 to 2013. In order to determine if means of house price were different in each month, Scheffe tests was applied.

2.2.1 Sheffe's Test

Figure 2.2 plotted results of the tests, based on 95% confidence intervals of different mean levels of transaction months with their corresponding transaction years.

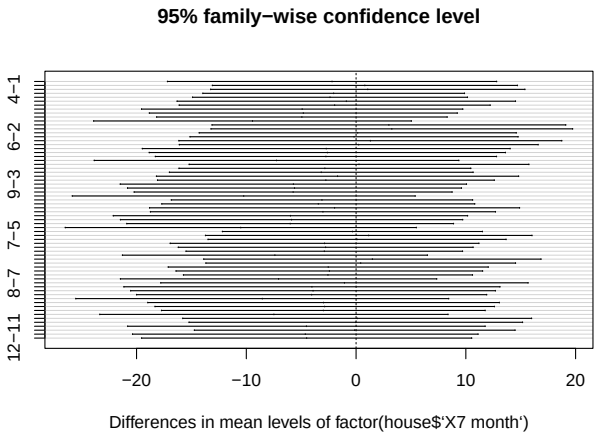

Figure 2.2: Mean Levels of Transaction Months by Years

In Figure 2.2, all the 95% confidence intervals include zero; there are no differences on average house price per month. As the transaction months (X7) and years (X8) were generated from the transaction date (X1), their correlations are strong. Thus, the transaction months (X7) and years (X8) were removed in regression analyses.

2.3 Descriptive Statistics

Table 2.1: Numerical Summary

X1	X2	X3	X4	X5	X6	Y
Min. :2013	Min. : 0.000	Min. : 23.38	Min. : 0.000	Min. :24.93	Min. :121.5	Min. : 7.60
1st Qu.:2013	1st Qu.: 9.025	1st Qu.: 289.32	1st Qu.: 1.000	1st Qu.:24.96	1st Qu.:121.5	1st Qu.: 27.70
Median :2013	Median :16.100	Median : 492.23	Median : 4.000	Median :24.97	Median :121.5	Median : 38.45
Mean :2013	Mean :17.713	Mean :1083.89	Mean : 4.094	Mean :24.97	Mean :121.5	Mean : 37.98
3rd Qu.:2013	3rd Qu.:28.150	3rd Qu.:1454.28	3rd Qu.: 6.000	3rd Qu.:24.98	3rd Qu.:121.5	3rd Qu.: 46.60
Max. :2014	Max. :43.800	Max. :6488.02	Max. :10.000	Max. :25.01	Max. :121.6	Max. :117.50

Table 2.1 numerically summarized each predictor's and response's distributions.

- The average house price per unit area (Ping) in Taiwan is 37.98 thousands New Taiwan Dollar.

- The number of convenient stores range from zero to ten.
- All the houses are less than 43.8 years. The average age of the houses is 17.7 years.

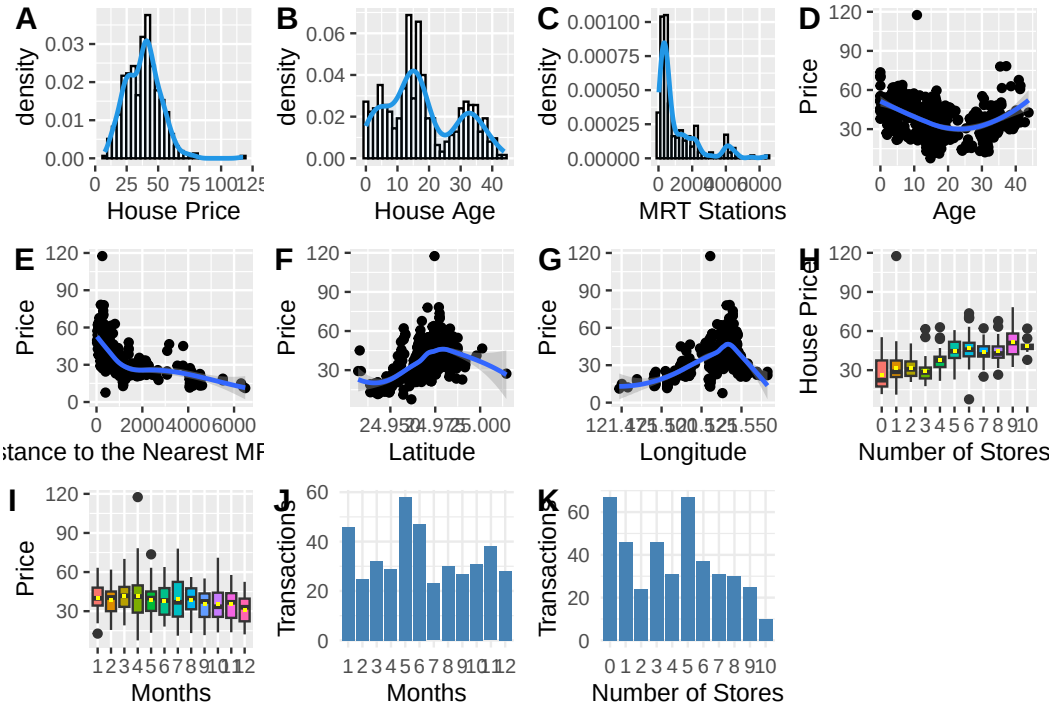

Figure 2.3: Characteristics among Each Variable

Figure 2.3 graphically summarized predictors' and response's characteristics.

A) The density of house price is similar to an unside-down bell curve, and slightly skewed to the right. The mean of house pirce is approximately round 38 thousands New Taiwan Dollar.
B) The density of house age is like a wavy line with two peaks. Most of houses are over 10 years old.
C) The density of distance to the nearest MRT station is significantly skewed to the right. Majority of the houses are close to the convenient stores.
D) The association between house age and price is little similar to a concave-up parabola.
E) The smaller distance to the nearest MRT station is, the higher house price will be.
F) The association between house price and latitude is little similar to a concave-down parabola.
G) The association between house price and longtitude is little similar to a concave-down parabola.
H) The more stores are nearby, the higher house price will be.
I) The median of house price is steady in each month. The house price had the largest variation in July 2013 and the smallest variation in January 2013.
J) The number of transactions are different in each month. The highest transactions occurred in May 2013; the lowest transactions was in July 2013.
K) Zero or five convenient stores around the houses had the much higher transactions. However, ten convenient stores in the living circle had lowest transactions.

2.4 Property Distribution

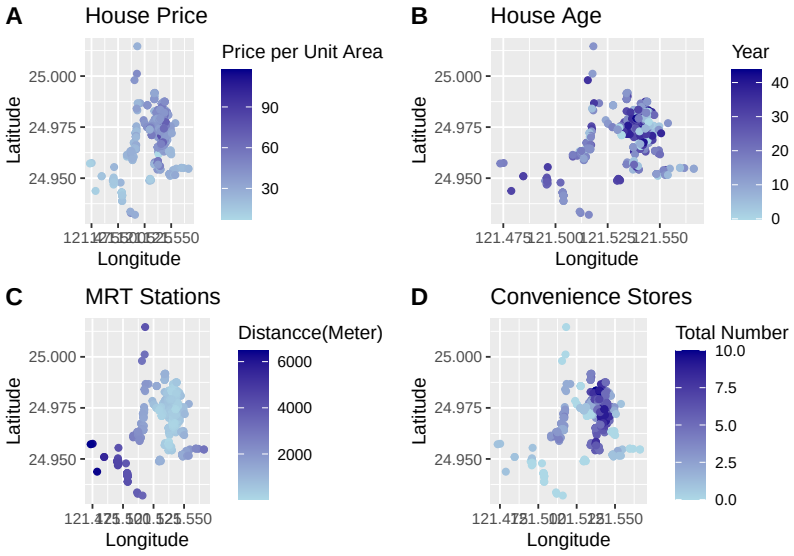

Figure 2.4: Property Lcoations

This map described associations between predictors and house price in Sindian District, New Taipei City, Taiwan. When the houses were located closer to downtown, number of convenient stores, house price and age were increasing, but the distance to the nearest MRT station was decreasing.

2.5 Correlations

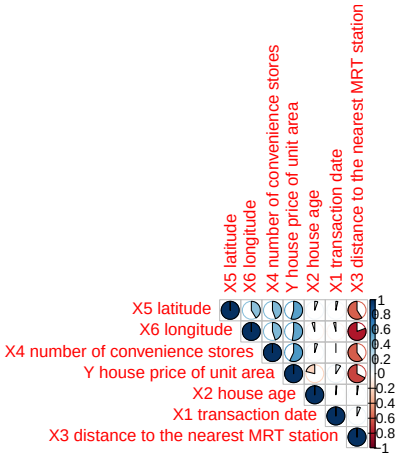

Figure 2.5: Correlation Matrix

Figure c visually summarized correlations among predictors and response. The distance to the nearest MRT station is significantly negatively correlated to the geographic coordinates, number of convenient stores, and house price, which is opposite to house age and transaction date. House price is moderately positively correlated to the geographic coordinates and number of convenient stores.

3 Multuple Linear Regression

The relationship between each explanatory variable and House Price were studied in multiple linear regressions through data transformation and elimination of the unnecessary variables at 5% significance level (α).

Full Model: $Price_i = \beta_0 + \beta_1 Date_i + \beta_2 Age_i + \beta_3 Distance_i + \beta_4 Stores_i + \beta_5 Latitude_i + \beta_6 Longitude_i + \varepsilon_i$, where i = 1,2,. . . ,n.

3.1 Diagnostic Analysis

3.1.1 Linearity

According to exploratory data analysis (Section 2), Figure 2.3 (D to I) weakly shows house price and predictors had a linear relationship betweem each other. After *Distance to the Nearest MRT Station* had a log transformation, it had a negative linear releationship with house price. Its transformation will be applied in Section 3.2 for further analysis.

3.1.2 Multicollinearity

Table 3.1: Variance Inflation Factor

	VIF
DistanceToMRT	4.322984
Longitude	2.926305
NumberOfStores	1.617021
Latitude	1.610225
TransactionDate	1.014655
HouseAge	1.014287

Variance of inflation factors (VIF) of all the predictors in Table 3.1 are between 1 and 5, so none of the predictors will be removed when VIF < 10. Thus, multicollinearity and potential dependence among residuals are not a concern.

3.1.3 Independence of Error

Table 3.2: Independence Tests for Errors

	Statistic	**P-value**	**Alternative Hypothesis**
Durbin-Watson Test	2.15273148361674	0.941481308696768	true autocorrelation is greater than 0

Based on Durbin-Watson Test in Table 3.2, its p-value is greater than 5% significance level, so the null hypothesis is rejected and the true autocorrelation is not greater than zero. Error correlation normally presents when data is related to the economic and other temporal activities. This dataset is also influenced by these temporal factors. thus, the errors are not independent.

3.1.4 Residuals

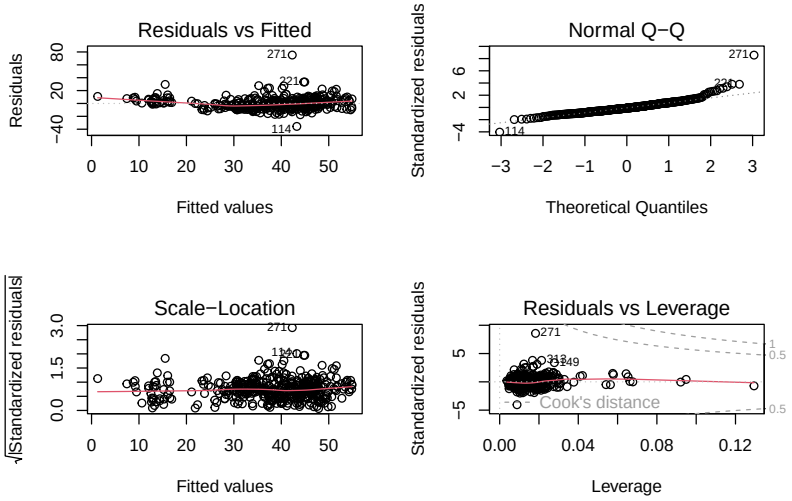

Figure 3.1: Residual Diagnostics

Figure 3.1 reflects different full-model residuals' characteristics. No specific partterns are shown in the top-left plot, so residuals' variances are similar. The residuals are normally distributed becasue most of the residuals are on the straight line in the QQ plot. The 271^{th} observation in the bottom-right plot (Table 3.1) is hardly identified as a high leverage, high-influential point or even outlier.

3.1.5 High Leverages

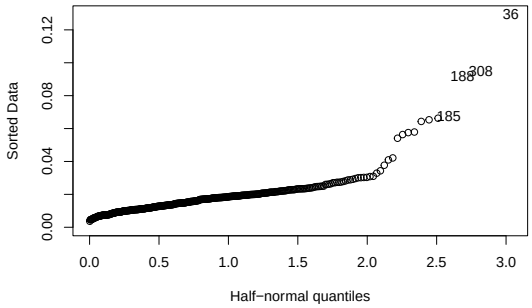

Figure 3.2: High Leverages

In Figure 3.2, high-leverage points are the points far away from the center of the whole dataset. The design matrix X from sample, $H = X^T[(X^TX)]^{-1}X$, since $h_i = H_{ii}, \sum_i h_i = p$, observations should be defined as high-leverage points when they have leverages more than $\frac{2p}{n}$ (twice the mean leverage) . In this case, the high leverage points are with leverages bigger than 0.03381643. There are 19 high-leverage points in total.

3.1.6 High-influential Points

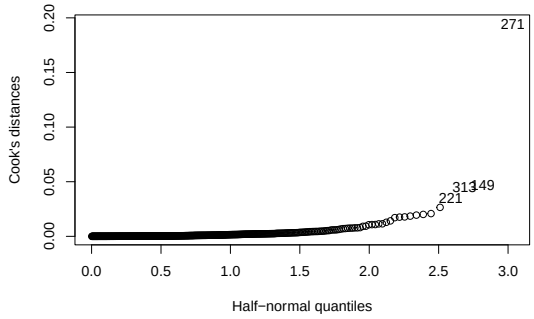

Figure 3.3: High-influential Points

Figure 3.3 shows the 271^{th} observation has the largest cook's distance, and its value of cook's distance is less than 1. Due to the rule-of-thumb, there is no high-influential points.

3.1.7 Outliers

Table 3.3: Outliers

	|Studentized Residual|
Studentized Residual	3.882142
No.271	9.451489

Outliers are the points in data that do not fit the model as the other data points. To control the type one error less than α, we use Bonferroni correction, and in this case, each data point is tested at level $\frac{\alpha}{n}$. As a result in Table 3.3, the 271^{th} observation, the only one outlier, is greater than the first absolute studentized residual, which is about twice as much as the residual. Hence, the 271^{th} observation would be omitted in later Section 3.2 .

3.1.8 Sequential ANOVA Model

Table 3.4: Models Comparison

c("Full Model", "Sequential Model")	**Res.Df**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
Full Model	323	21925.04	NA	NA	NA	NA
Sequential Model	298	14774.42	25	7150.62	5.769119	0

Following the Hierarchy Rule, additive and sequential models are compared to determine if the interactions between each predictor are involved. In the result, the p-value of the F test is statistically significant. Therefore, the interactions are not considered.

3.2 Transformation

Since not all the predictors and the response are linearly related to each other (Section 3.1.1), transformations of the response and predictors can improve the full model's performance.

3.2.1 Box-Cox Transformation

The optimal value of BoxCox transformation ($\hat{\lambda}$) in Figure 3.4 is exclusively between 0 and 0.5 and approximately equal to $\frac{2}{11} \approx 0.181818182$. The transformation of *House Price* is calculated and followed by the formula:

$y(\lambda) = \frac{y^{\lambda}-1}{\lambda}$, when $\lambda \neq 0$. Then *House Price* became *House Price*$(\frac{2}{11})$.

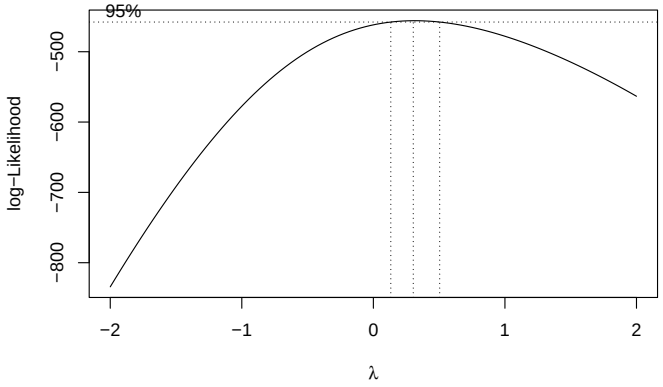

Figure 3.4: Box-Cox Analysis for House Price

3.2.2 Transformed Model

After transforming *House Price* (Section 3.2.1) and *Distance to the Nearest MRT Station* (Section 3.1.1) into logarithm, here is the transformed model: $\log(Price_i) = \beta_0 + \beta_1 Date_i + \beta_2 Age_i + \beta_3 \log(Distance_i) + \beta_4 Stores_i + \beta_5 Latitude_i + \beta_6 Longitude_i + \varepsilon_i$, where i = 1,2,. . . ,n.

The steps of diagnostic analysis (Section 3.1) had been applied again for the transformed model. All the result of diagnoses were met properties of a multiple linear regression model.

Linear Models Summary (Table 7.1) compared differences between full and transformed models.

- Latitude and longitude are significant predictors, which affect house price of unit area. This was also reflected in Linearity (Figure 2.3 F & G), the map (Section 2.4), and correlation (Figure 2.5).
- All the orignial and/or transformed predictors are significant.
- Transformed model ($R^2 = 70.3\%$) is more accurate than full model ($R^2 = 58.2\%$).

4 Ridge Regression

Ridge regression addresses highly correlated variables. The **Monte Carlo cross-validation** is used to randomly split the data into training and testing groups each time instead of fix **K** portions. When the entire K-fold cross-validation process is repeated enough times, it can averafe the error. The estimated testing error also becomes fairly stable, and not affected much by the random mechanism. Then the influence of randomness will be reduced. The transformed model (Section 3.2.2) was applied in ridge regression with 10-fold cross-validation.

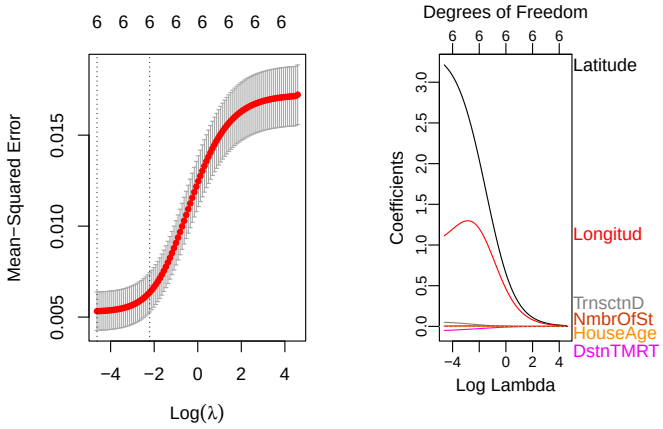

Figure 4.1: Minimum Lambda and Coefficents of Predictors

The left Figure 4.1 shows the minimum lamda in 10-fold cross-validation. The right Figure 4.1 reflects the coefficients of predictors. The coefficients of latitude and longitude are the largest, comparing with other predictors, so the geographical coordinates are the most important influential factors of house price.

5 Gradient Boosting Regression

Gradient Boosting is an ensemble machine learning algorithm that combines weak prediction models, usually decision trees, to create a powerful predictive model. By analyzing the variable importance measures in the full model without outliers, we can gain insights into which features have the most influence on the target variable.

Table 5.1: Variable Importance

Predictors	**Relative Influence**
DistanceToMRT	60.1039301
Latitude	20.9737694
HouseAge	11.8910304
Longitude	3.8500889
NumberOfStores	2.2988604
TransactionDate	0.8823207

The variable importance measures (Table 5.1) indicate the relative influence of each feature in the model's predictive performance. Higher importance scores suggest that the feature has a stronger impact on House Price. The analysis revealed several features that significantly influence House Price, including Transaction Date, House Age, Distance To the Nearest MRT Station, and Number of Convenience Stores.

6 Conclusion

Table 6.1: Models Comparsion

Model	**Training.Error**	**Testing.Error**	**R.Squared**
Linear Regression	262.0893	299.06949	0.5823850
Transformed Linaer Regression	1634.8426	1634.84261	0.7034690
Ridge Regression	1409.4140	1634.83490	0.7014334
Gradient Boosting Regression	286.2029	40.30768	0.7823600

In this housing market analysis, we compared multiple models to predict house prices. The results, as shown in Table 6.1, provide valuable insights into the performance and predictive accuracy of each model. Based on the comparison, the following refined conclusions can be drawn:

1. **Linear Regression**: The basic linear regression model achieved moderate predictive accuracy, with a training error of 262.0893 and a testing error of 299.0694. It has the advantage of simplicity and interpretability, allowing easy understanding of the impact of individual predictors on house prices. However, linear regression may struggle to capture non-linear relationships and complex patterns in the data. It is also sensitive to outliers and can be affected by multicollinearity among predictors.
2. **Transformed Linear Regression**: The transformed linear regression model addressed the limitations of basic linear regression by considering non-linear relationships between predictors and the target variable. With similar training and testing errors of 1634.8426, this model achieved a substantially higher R-squared value of 0.7035. It offers the advantage of capturing non-linear relationships without the need for complex algorithms. However, the transformation process requires domain knowledge and can introduce bias if not applied carefully.

3. **Ridge Regression**: The ridge regression model aimed to address potential multicollinearity issues among predictors. It achieved a training error of 1409.4140 and a testing error of 1634.8349, with a slightly lower R-squared value of 0.7014 compared to the transformed linear regression model. Ridge regression provides the advantage of mitigating multicollinearity and improving model generalization. It is suitable for housing market analysis where predictor variables may be highly correlated. However, it assumes a linear relationship between predictors and the target variable, limiting its performance when dealing with highly non-linear relationships.

4. **Gradient Boosting Regression**: Among the models examined, the gradient boosting regression model demonstrated the highest predictive accuracy. With a training error of 286.2029 and a testing error of 340.3077, it outperformed the other models in terms of minimizing errors. The remarkable R-squared value of 0.7824 suggests that the gradient boosting regression model effectively captures complex relationships between predictors and the target variable. It excels in handling non-linear relationships and complex interactions. However, gradient boosting models can be computationally intensive and require more time and computational resources for training and tuning.

In conclusion, the gradient boosting regression model emerges as the most effective model for predicting house prices in this analysis. Its superior performance, with the lowest errors and highest R-squared value, highlights its ability to capture complex patterns and relationships within the housing market data. However, it is important to consider the computational requirements and the potential need for additional resources when implementing gradient boosting models. Therefore, the gradient boosting regression model is recommended for further analysis and decision-making in the housing market.

7 Appendix

7.1 Multiple Linear Regression

Table 7.1: Linear Model Comparsion

	Full Model	Transformed Model
(Intercept)	-14441.983 *	-310.880 ***
	(6775.386)	(48.134)
TransactionDate	5.149 **	0.057 ***
	(1.557)	(0.014)
HouseAge	-0.270 ***	-0.002 ***
	(0.039)	(0.000)
DistanceToMRT	-0.004 ***	
	(0.001)	
NumberOfStores	1.133 ***	0.004 *
	(0.188)	(0.002)
Latitude	225.470 ***	3.388 ***
	(44.566)	(0.374)
Longitude	-12.429	0.941 **
	(48.581)	(0.328)
log(DistanceToMRT)		-0.057 ***
		(0.006)
N. obs.	414	330
R squared	0.582	0.703
F statistic	94.597	127.710

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.