

Predictive Modelling of Melbourne Housing Data

Anasah Wawem Christopher

October 21, 2025

Contents

1	Introduction	1
2	Data Structure	2
2.1	Data Overview	2
3	Data Preprocessing	2
3.1	Missing Value Treatment	2
3.2	Feature Engineering	3
4	Analysis And Results	3
4.1	Correlation Analysis	3
4.2	Ordinary Least Squares Model (OLS)	4
4.3	Ridge Regression	8
4.4	Least Absolute Shrinkage and Selection Operator (LASSO) Regression .	11
4.5	Principal Component Analysis (PCA) Regression	14
4.6	Random Forest Model	17
4.7	Model Comparison	20
5	Conclusion	20
6	Appendix	21

Abstract

This report provides a comprehensive analysis of Melbourne housing data utilizing some regression techniques. The data structure, data wrangling, preprocessing, and analysis are all covered in the report. This study aims to build regression models using ordinary least squares, ridge, lasso, random forest, and principal component analysis; compare, validate, and diagnose these models.

1 Introduction

One model may perform well with one set of data while performing poorly with another, as different models have different preferences for various types of data. It seems ideal to build multiple models on a given dataset and then select the best one. This study's goal is to develop and assess multiple regression models to accurately predict Melbourne

real estate prices by identifying the relevant contributing factors. The best model will be selected based on how well it generalizes to previously unobserved data. This study will help homebuyers, investors, and even those concerned about the housing market's situation to make informed decisions.

2 Data Structure

2.1 Data Overview

This is a real estate dataset on residential house prices in Melbourne. The data is from Kaggle.com and consists of 22 variables and 34,857 observations. 10 variables have at least 1 missing entry. Table 1 summarizes the structure of the dataset.

Table 1: Dataset Structure and Variable Description

Variable	Type	Unique Values	Description
Suburb	Character	351	Location suburb
Address	character	34009	Street address of the house
Rooms,	Integer	12	Number of rooms
Type	Character	3	Type of residential property (e.g, townhouse, unit)
Method	Character	9	Method of sale (e.g, vendor bid, property sold prior)
SellerG	Character	388	The real estate agency
Date	Character	78	The date property was registered/finalized
Distance	Character	216	Distance from central business district
Postcode	Character	212	Postcode of the area
Bedroom	Integer	16	Number of bedrooms
Bathroom	Integer	12	Number of bathrooms
Car	Integer	16	Number of car parking space
Landsize	Integer	1685	Total area of land plots in square meters
BuildingArea	Character	743	Total floor area of the building in square meters
YearBuilt	Integer	161	Year property was built
CouncilArea	Character	34	Local government area
Latitude	Numeric	13403	Geographical coordinate position (North-South)
Longitude	Numeric	14525	Geographical coordinate position (East-West)
Regionname	Character	8	Broader geographical region
Propertycount	Character	343	Total number of property in the suburb
ParkingArea	Character	8	Total parking space in square meters
Price	Integer	2872	Price of the price in Australian Dollars

3 Data Preprocessing

3.1 Missing Value Treatment

About 69.99 % of the total records have at least one missing entry. Table 2 shows the breakdown, and the treatment applied. The Little's Missing Completely at Random (MCAR) Test was performed, and the result shows that the missing values are not at random. This means that the probability of a missing entry is related to the unobserved

value itself or is systematically related to other observed variables in the data set. The multiple imputation method using the predictive mean matching strategy was used to impute the missing values.

Table 2: Variables with Missing Values and Treatment

Variable	Missing %	Treatment Method
Building Area	12.3%	Predictive Mean Matching
Year Built	7.8%	Predictive Mean Matching
Bedrooms	23.57%	Predictive Mean Matching
Bathrooms	23.60%	Predictive Mean Matching
Price	21.83%	Predictive Mean Matching
Car	25.04%	Predictive Mean Matching
Landsize	33.88%	Predictive Mean Matching
Latitude	22.88%	Predictive Mean Matching
Longitude	22.88%	Predictive Mean Matching
Propertycount	0.009%	Predictive Mean Matching

3.2 Feature Engineering

The following adjustments were made:

- The variables Distance, Postcode, BuildingArea, Price, and Landsize were converted to the right data type (numeric), and Propertycount was converted to an integer.
- All missing values in different forms, like "", ' ', , - were standardized to NA.
- Variables such as Suburb, Address, Date, SellerG, and CouncilArea were removed from the data set, since they are not of interest in the analysis.
- The clean data was partitioned into a ratio of 80:20 for the train and test datasets, respectively, using stratification.

4 Analysis And Results

4.1 Correlation Analysis

The correlation matrix for the variables is displayed in Figure 1. There is a strong positive relationship between Bedroom and Rooms. Bathroom and Rooms, Distance and Postcode as well as Bathroom and Bedroom have some moderate relationship. Car and Rooms as well as Car and Bedroom have a moderate relationship. Latitude and Longitude have a strong negative relationship.

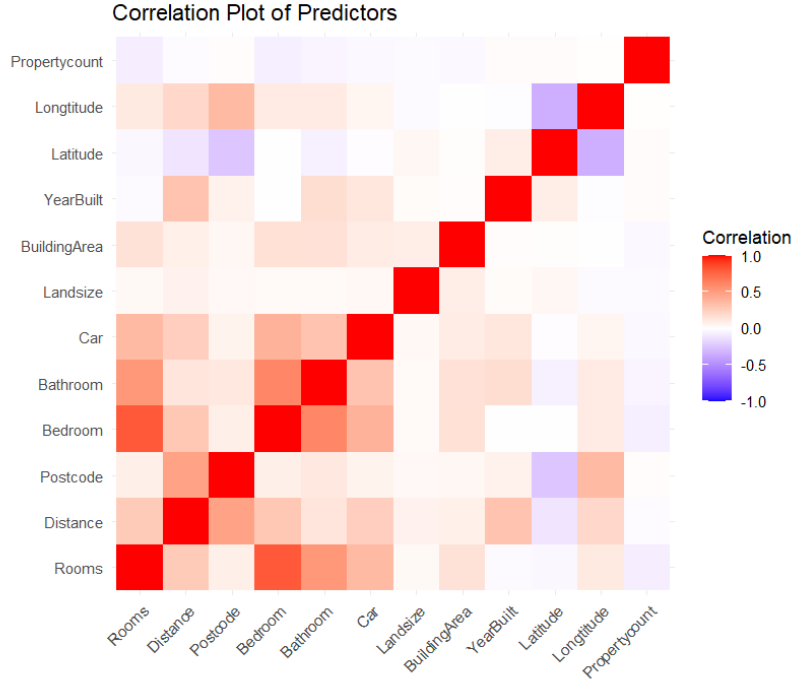


Figure 1: Correlation Matrix of Numerical Variables

4.2 Ordinary Least Squares Model (OLS)

Table 3 provides the summary of the regression model built using the ordinary least squares approach. The model was initially built, and the final model was selected using a stepwise function. The adjusted R-squared of the initial model and the final model selected by the step function were all same. This means that predictor removal could not improve the model. This OLS model will be the baseline model for the analysis. From the summary, 57.6 % of the variations in price is explained by the predictors, leaving 42.4 % unexplained. From the table, Method SA (Sold After), Method SN (Sold Not Disclosed), Method SP (Property Passed In), and Method SS (Sold Swap) using Method PI (Property Sold) as a reference are not significant at 5% significance level. Also, the regions of Eastern Victoria and South-Eastern Metropolitan, using Eastern Metro as a reference, are not significant at the 5% significance level. All other predictors are significant at the same significant level.

Figure 2 is the scatter plot of the predicted price and the actual price of the houses in Melbourne using the ordinary least squares regression model. There are inconsistencies in the points, where points are scattered at higher prices. The model made some good predictions at a lower price and systematic underpredictions at a higher price. Figure 3 is the leverage plot of the OLS model, which indicates the influential points on the model. Most of the points have low leverage, which are clustered around 0. However, some points are above the threshold of $3 \times \text{mean}$, including 2 points with extreme leverage. Figure 4 shows the quantile-quantile plot of the OLS model. From the plot, the residuals do not appear normally distributed, and this is confirmed by Anderson-Darling Normality Test in table 10. Table 4 provides the outlier test result of the OLS model. There are 10 observations with extreme outliers, with observation 32775 the most extreme and observation 28233 less extreme.

Table 3: Ordinary Least Squares Regression Model Summary

Variable	Coefficient	Std. Error	Sig.
Intercept	-71 530 000	4 667 000	***
Rooms	103 600	4966	***
Distance	-35 180	635	***
Postcode	239	34	***
Bedroom	77 020	5012	***
Bathroom	256 500	4948	***
Car	73 180	2939	***
Landsize	3	1	***
BuildingArea	67	9	***
YearBuilt	-4193	81	***
Latitude	-790 700	39 390	***
Longitude	341 900	31 990	***
Property Type (ref: h (House))			
Type t (Townhouse)	-167 800	9121	***
Type u (Unit)	-274 800	8269	***
Method (ref: PI (Property Sold))			
Method PN (Sold Prior)	65 910	28 610	*
Method S (Sold)	26 850	7875	***
Method SA (Sold After)	19 730	33 190	
Method SN (Sold Not Disclosed)	-23 780	15 190	
Method SP (Property Passed In)	-13 600	9867	
Method SS (Sold Swap)	34 830	82 980	
Method VB (Vendor Bid)	24 640	11 250	*
Method W (Withdrawn)	139 400	36 800	***
Region (ref: Eastern Metro)			
Eastern Victoria	48 690	36 370	
Northern Metropolitan	-106 300	10 420	***
Northern Victoria	279 200	38 340	***
South-Eastern Metropolitan	15 420	16 130	
Southern Metropolitan	154 900	10 230	***
Western Metropolitan	-145 800	12 410	***
Western Victoria	161 200	51 500	**
Model Statistics			
Observations	27,888		
R-squared	0.576		
Adjusted R-squared	0.576		
Residual Std. Error	437,100 (27859 df)		
F-statistic	1353 (p < 0.001)		

Note: Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05. Reference categories:
Property Type = h, Method = PI, Region = Eastern Metropolitan.

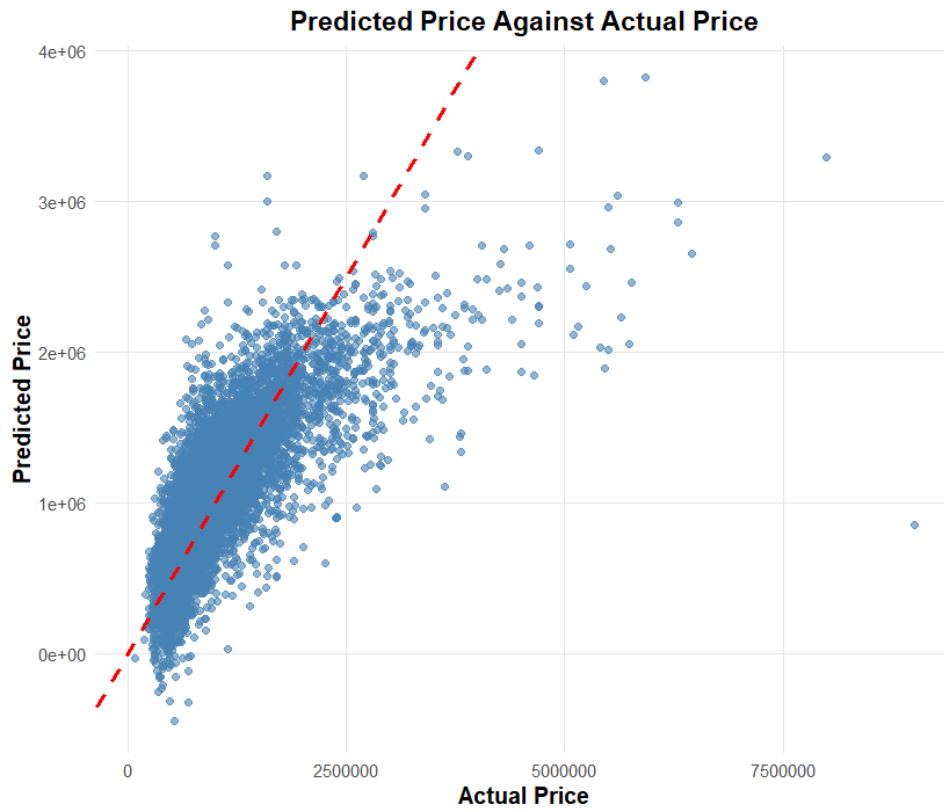


Figure 2: Plot of Predicted Price and Actual Price with OLS Model

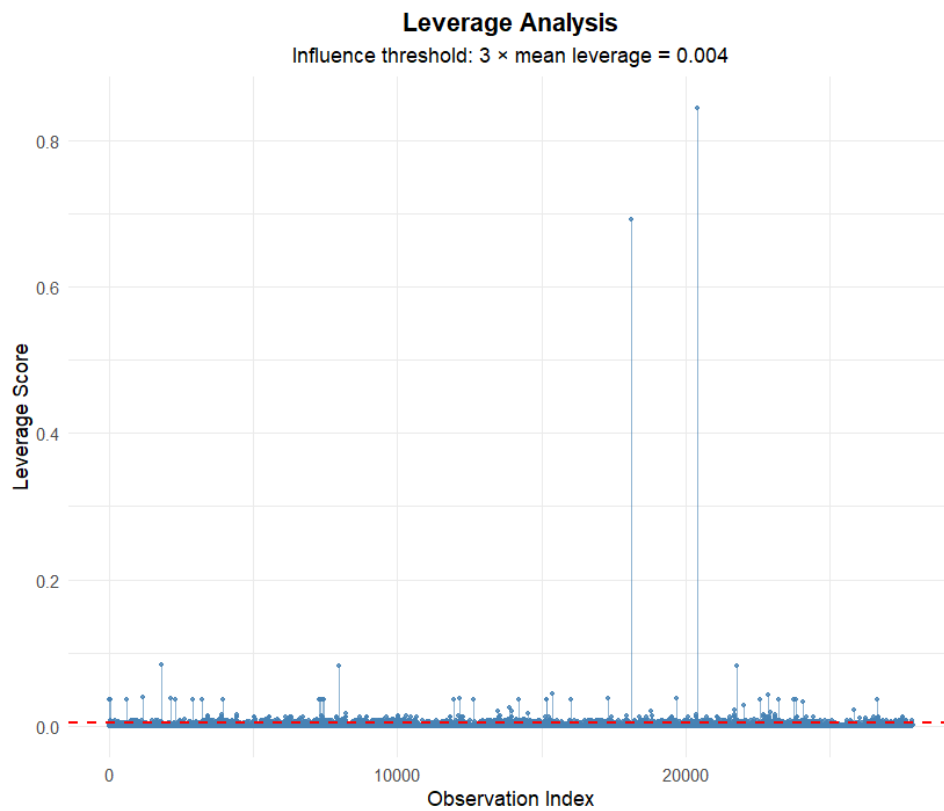


Figure 3: Leverage Plot of OLS Model

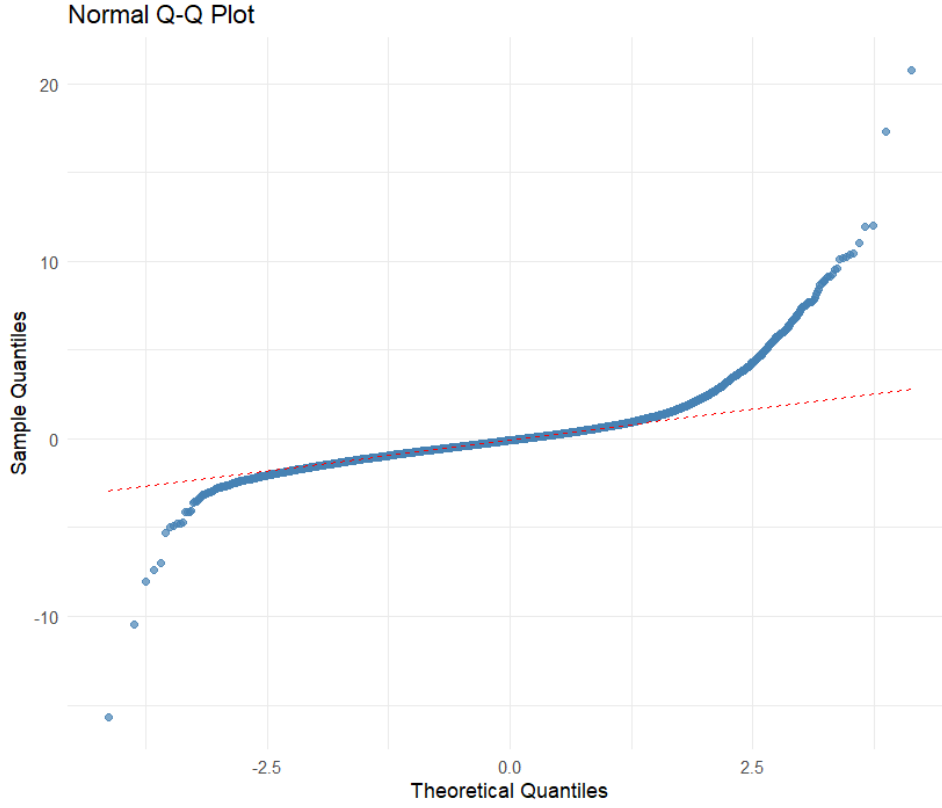


Figure 4: Quantile-Quantile Plot of OLS Model

Table 4: Outlier Test Results for OLS Model

	Observation	rstudent	Unadjusted p-value	Bonferroni p-value
1	32 775	20.774 02	3.9187×10^{-95}	1.0928×10^{-90}
2	22 413	17.289 47	1.2586×10^{-66}	3.5100×10^{-62}
3	25 559	-15.662 85	4.6644×10^{-55}	1.3008×10^{-50}
4	21 703	12.010 31	3.7881×10^{-33}	1.0564×10^{-28}
5	18 994	11.947 05	8.0921×10^{-33}	2.2567×10^{-28}
6	13 482	10.991 95	4.7714×10^{-28}	1.3306×10^{-23}
7	2288	-10.478 79	1.2067×10^{-25}	3.3654×10^{-21}
8	19 828	10.449 35	1.6446×10^{-25}	4.5865×10^{-21}
9	16 497	10.355 90	4.3695×10^{-25}	1.2186×10^{-20}
10	28 233	10.264 97	1.1215×10^{-24}	3.1277×10^{-20}

4.3 Ridge Regression

Table 5 shows the summary results of the coefficients from the ridge regression model. According to the algorithm, all the 36 predictors, including dummy variables, have their coefficients maintained, indicating that they significantly contribute to the model. Figure 5 shows the scatter plot of the predicted price and the actual price of the Melbourne houses using the ridge regression model. There is systematic underprediction like the OLS model, with points scattered at higher prices. There is a reasonable prediction at lower prices and a bad prediction at higher prices. There are several extreme outliers (very high actual prices with much lower predicted values).

The leverage plot of the ridge model is shown in Figure 6. The pattern is similar to that of OLS leverage but with slightly compressed leverage. Although most of the points are clustered around 0, there are some points above the threshold, including 2 extreme ones, which indicates influence on the model. Figure 7 shows the quantile-quantile plot of the ridge model. From the plot, the residuals do not appear normally distributed, and this is confirmed by Anderson-Darling Normality Test in table 10.

Table 5: Summary Results of Ridge Regression

Variable	Coefficient
Rooms	103,443.9
Distance	-35,119.6
Postcode	237.6
Bedroom	77,012.4
Bathroom	256,601.3
Car	73,186.5
Landsize	3.38
BuildingArea	66.92
YearBuilt	-4,191.1
Latitude	-792,508.1
Longitude	342,470.6
Property Type (ref: h(House))	
Type t (Townhouse)	-168,278.6
Type u (Unit)	-274,425.1
Method (ref: PI (Property Sold))	
Method PN (Sold Prior)	66,396.8
Method S (Sold)	26,651.4
Method SA (Sold After)	19,921.3
Method SN (Sold Not Disclosed)	-23,544.5
Method SP (Property Passed In)	-13,751.5
Method SS (Sold Swap)	34,189.8
Method VB (Vendor Bid)	24,539.6
Method W (Withdrawn)	139,607.2
Region (ref: Eastern Metro)	
Eastern Victoria	50,674.9
Northern Metropolitan	-103,405.1
Northern Victoria	279,367.6
South-Eastern Metropolitan	15,746.5
Southern Metropolitan	156,570.5
Western Metropolitan	-145,028.4
Western Victoria	159,412.2
Parking Area (ref: None)	
Carport	13,918.7
Detached Garage	-6,518.2
Indoor	3,907.6
Outdoor Stall	14,691.3
Parkade	330.8
Parking Pad	5,584.3
Underground	20,946.7
Propertycount	-0.52

Note: Regularized regression with L2 penalty.

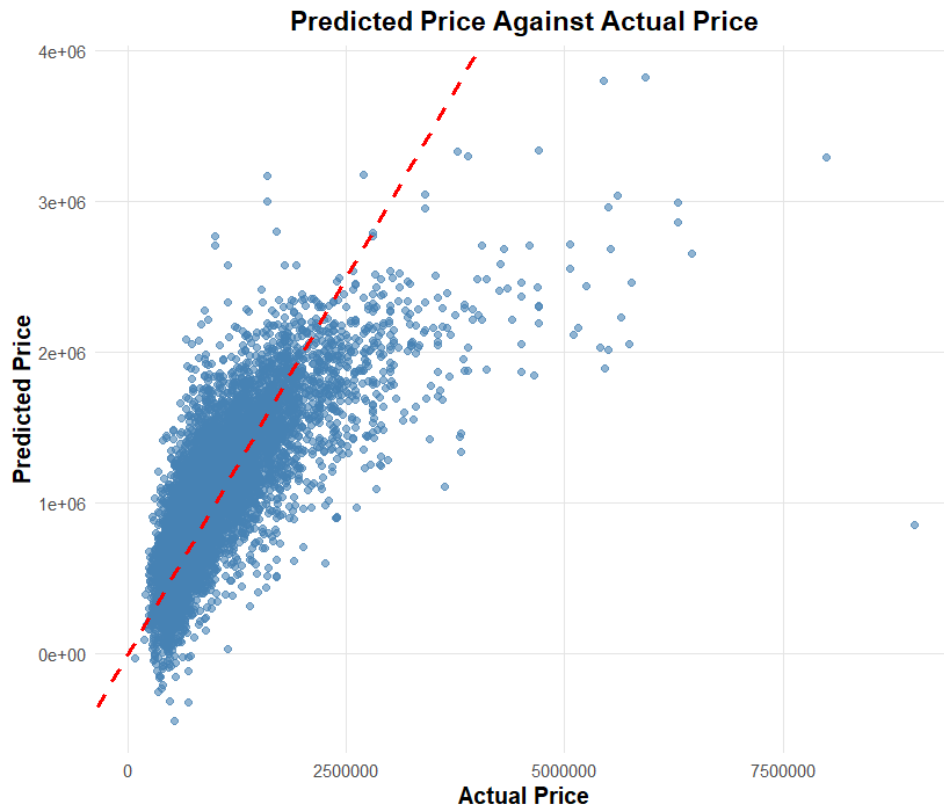


Figure 5: Plot of Predicted Price and Actual Price with Ridge Model

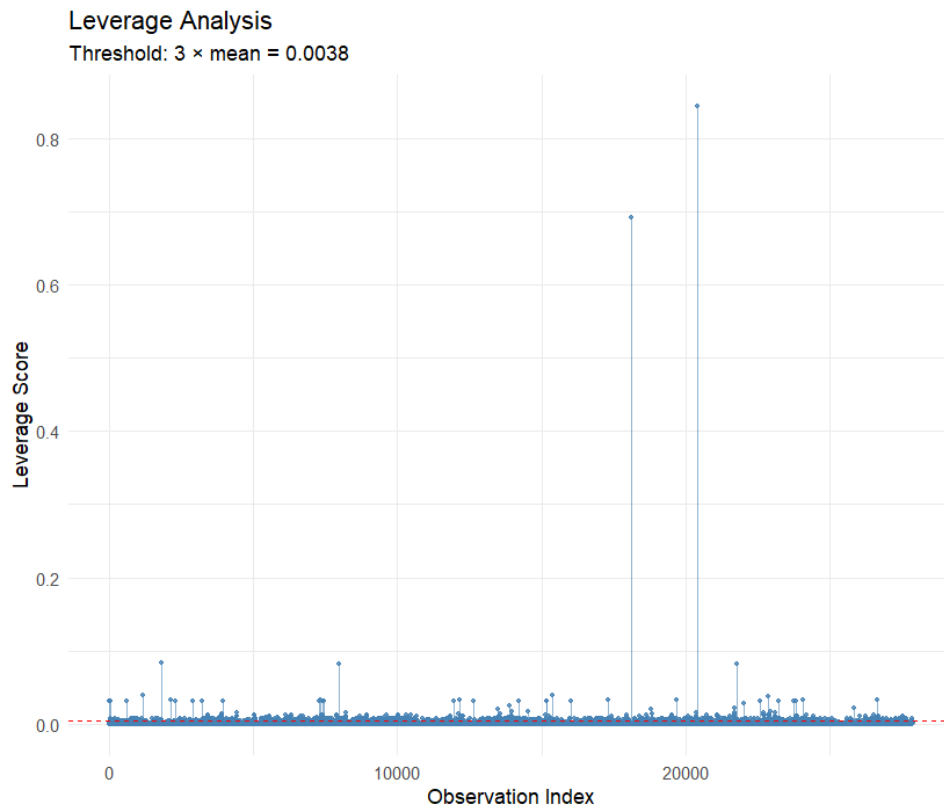


Figure 6: Leverage Plot of Ridge Model

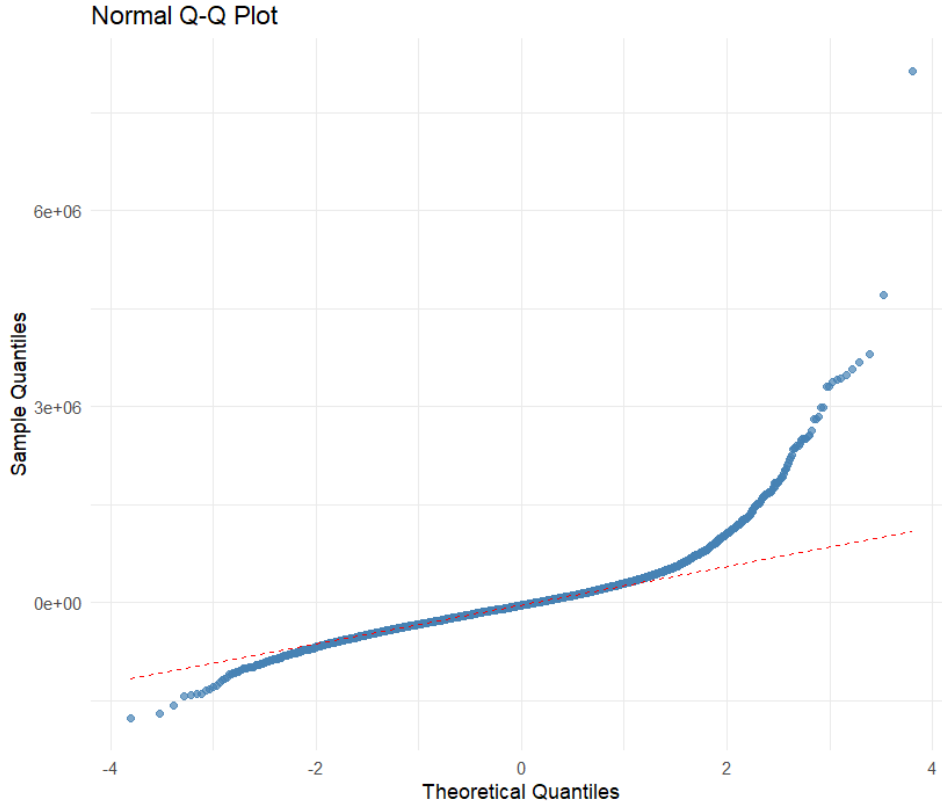


Figure 7: Quantile-Quantile Plot of Ridge Model

4.4 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

Table 6 below provides the summary results of LASSO regression coefficients. The algorithm shrunk 16 predictor coefficients to 0, indicating how insignificant they are to the model. The predictors are as follows:

Method: PN (Sold Prior), SA (Sold After), SN (Sold Not Disclosed), SS (Sold Swap), VB (Vendor Bid)

Region: Eastern Victoria, South-Eastern Metropolitan, Western Victoria

All ParkingArea types (7 categories) and Propertycount

Figure 8 shows the scatter plot of the predicted price using the LASSO model and the actual price of the Melbourne houses. There are extreme underpredictions of the prices, especially at higher prices compared to ridge and OLS regression, but with slightly more bias. This is likely due to the fact that more coefficients were lowered to 0. There are several extreme outliers (very high actual prices with much lower predicted values). Figure 9 shows the leverage plot of the LASSO model. The plot is identical to the leverage plot of the OLS and the ridge model despite dropping some features. This means that removing some features did not dramatically change the leverage. Figure 10 shows the quantile-quantile plot of the LASSO model. From the plot, the residuals do not appear normally distributed, and this is confirmed by Anderson-Darling Normality Test in table 10.

Table 6: Summary Results of Lasso Regression

Variable	Coefficient
Rooms	104,849.2
Distance	-30,348.0
Postcode	145.0
Bedroom	73,088.1
Bathroom	255,335.2
Car	65,984.0
Landsize	1.25
BuildingArea	49.48
YearBuilt	-4,219.7
Latitude	-726,290.2
Longitude	312,265.4
Property Type (ref: h (House))	
Type t (Townhouse)	-139,142.3
Type u (Unit)	-247,804.2
Method (ref: PI (Property Sold))	
Method S (Sold)	4,480.2
Method SP (Property Passed In)	-13,401.6
Method W (Withdrawn)	13,732.0
Region (ref: Eastern Metro)	
Northern Metropolitan	-74,655.9
Northern Victoria	138,669.3
Southern Metropolitan	181,145.1
Western Metropolitan	-119,693.2

Note: Regularized regression with L1 penalty.

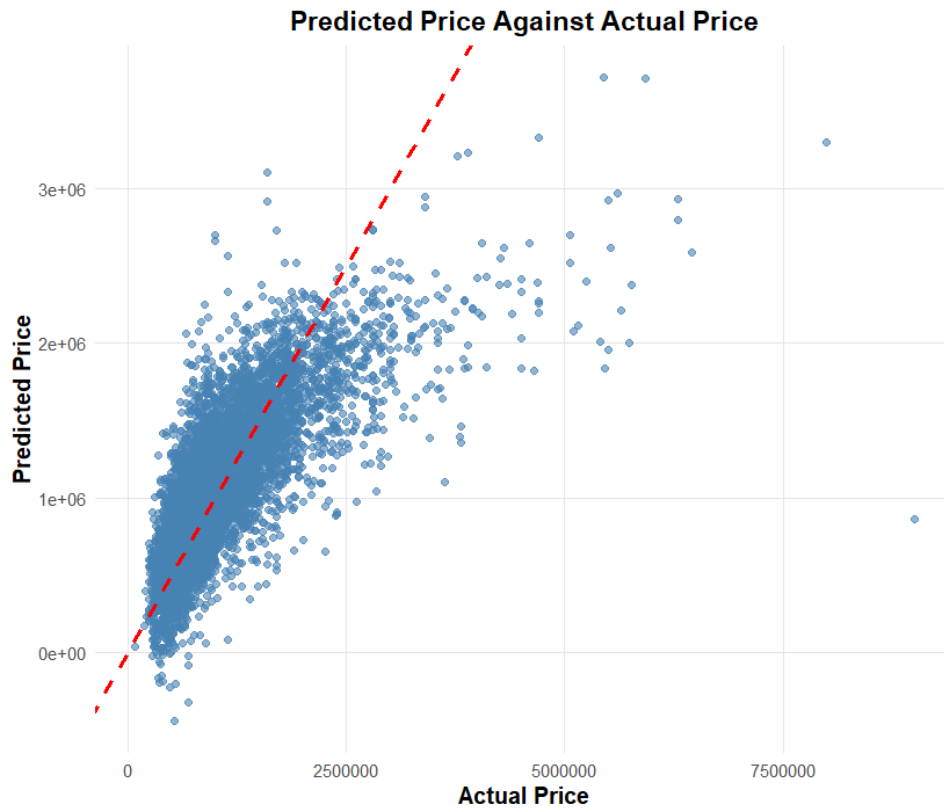


Figure 8: Plot of Predicted Price and Actual Price with LASSO Model

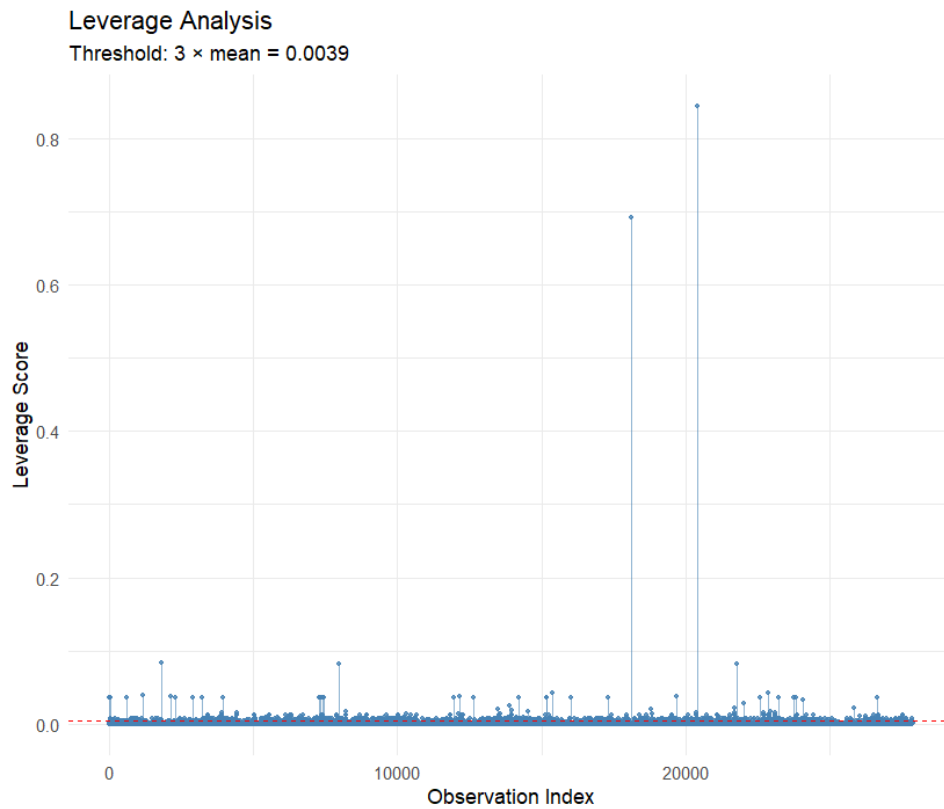


Figure 9: Leverage Plot of LASSO Model

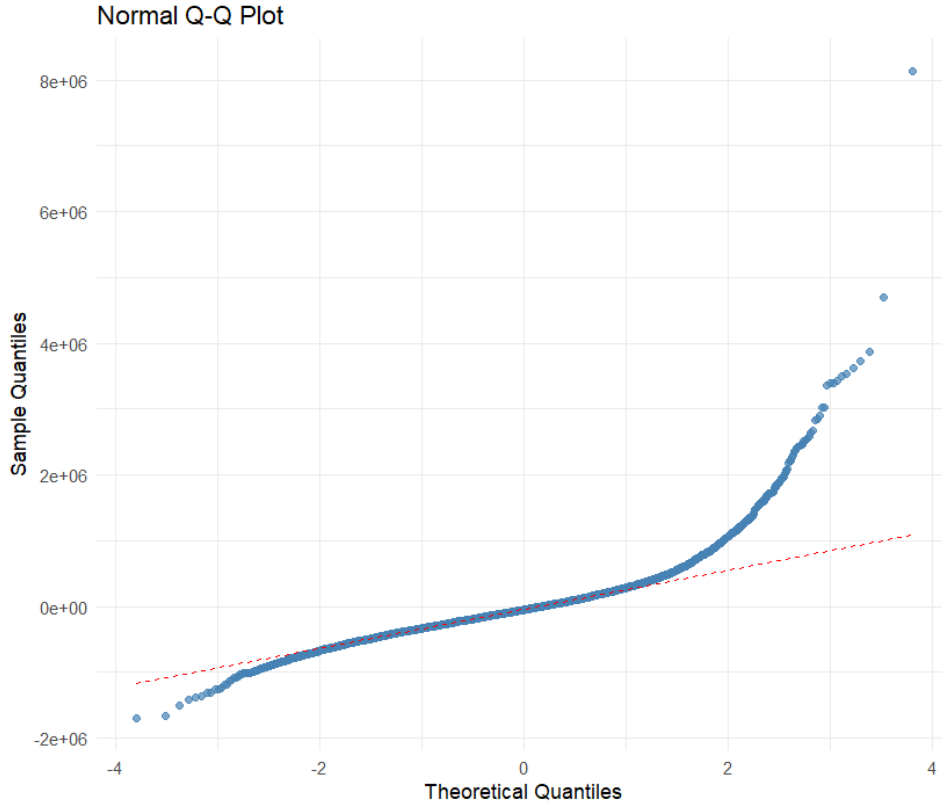


Figure 10: Quantile-Quantile Plot of LASSO Model

4.5 Principal Component Analysis (PCA) Regression

Table 7 shows the summary of the PCA regression. The optimal number of components using cumulative variance of 95% or more is 29. 55.5% of variability in the response variable is explained by the selected components. From the table, PC2, PC7, PC9, PC11, PC13, and PC17 are not statistically significant at 5% significance level. A step function was applied, and there was no improvement on the adjusted R-squared. Figure 11 shows the scatter plot of the predicted price using PCA model and the actual price of the Melbourne houses. Although the points are scattered at higher prices, the prediction looks better than ridge, LASSO, and OLS regressions. Again, there are some extreme outliers (very high actual prices with much lower predicted values).

Figure 12 is the leverage plot of the PCA model. There is more balanced leverage distribution, slightly different from that of leverage plots from OLS, ridge, and LASSO models. Although uncorrelated predictors were used to build the model, there are some influential points. Figure 13 shows the quantile-quantile plot of the PCA model. From the plot, the residuals do not appear normally distributed, and this is confirmed by Anderson-Darling Normality Test in Table 10.

Table 7: Summary of PCA Regression Results

Component	Coefficient	Std. Error	t-value	Significance
Intercept	1,083,822	2,681	404.21	***
PC1	-141,738	1,458	-97.23	***
PC2	-2,821	1,705	-1.65	.
PC3	-278,394	1,984	-140.35	***
PC4	94,778	2,089	45.37	***
PC5	-36,747	2,159	-17.02	***
PC6	-35,613	2,334	-15.26	***
PC7	1,069	2,428	0.44	
PC8	-25,562	2,437	-10.49	***
PC9	-2,585	2,480	-1.04	
PC10	80,400	2,491	32.27	***
PC11	2,814	2,552	1.10	
PC12	8,445	2,576	3.28	**
PC13	-1,887	2,577	-0.73	
PC14	7,528	2,587	2.91	**
PC15	-16,360	2,600	-6.29	***
PC16	-6,138	2,622	-2.34	*
PC17	4,369	2,629	1.66	.
PC18	22,068	2,650	8.33	***
PC19	-28,036	2,670	-10.50	***
PC20	13,685	2,678	5.11	***
PC21	-8,964	2,689	-3.33	***
PC22	31,298	2,702	11.58	***
PC23	14,305	2,757	5.19	***
PC24	-18,518	2,847	-6.50	***
PC25	-12,433	2,892	-4.30	***
PC26	-10,732	3,151	-3.41	***
PC27	111,662	3,629	30.77	***
PC28	23,686	3,942	6.01	***
PC29	-79,988	4,007	-19.96	***
Model Statistics				
Observations	27,888			
R-squared	0.555			
Adjusted R-squared	0.555			
Residual Std. Error	447,800 (27858 df)			
F-statistic	1200 (p < 0.001)			

Note: PCA regression using 29 principal components. Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05, . p < 0.1.

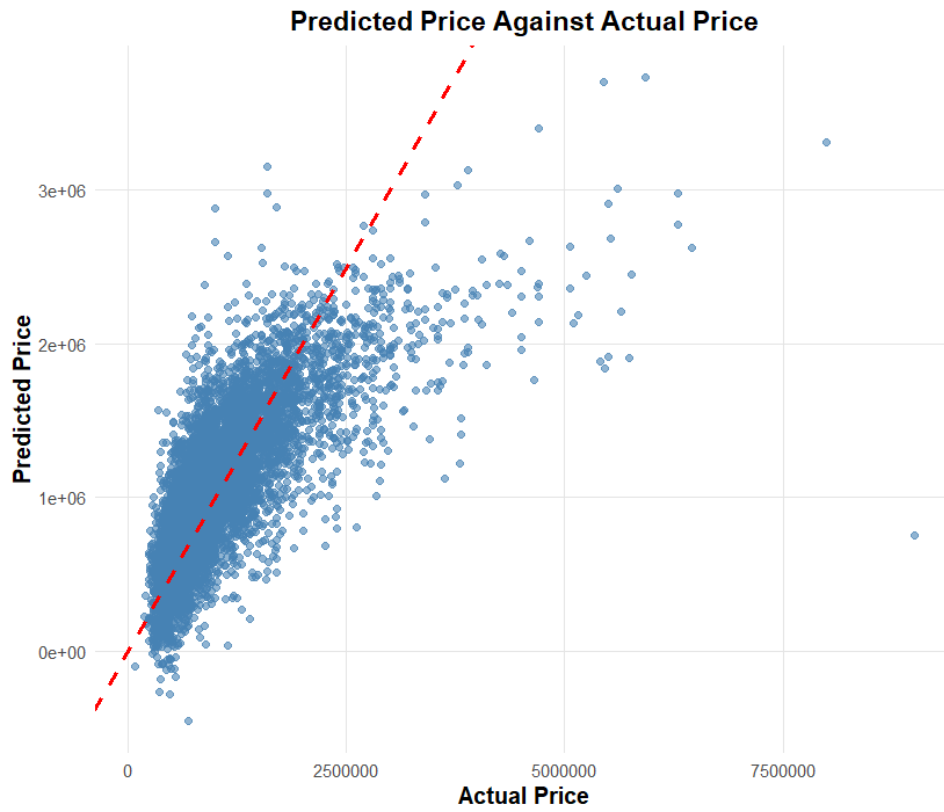


Figure 11: Plot of Predicted Price and Actual Price with PCA Model

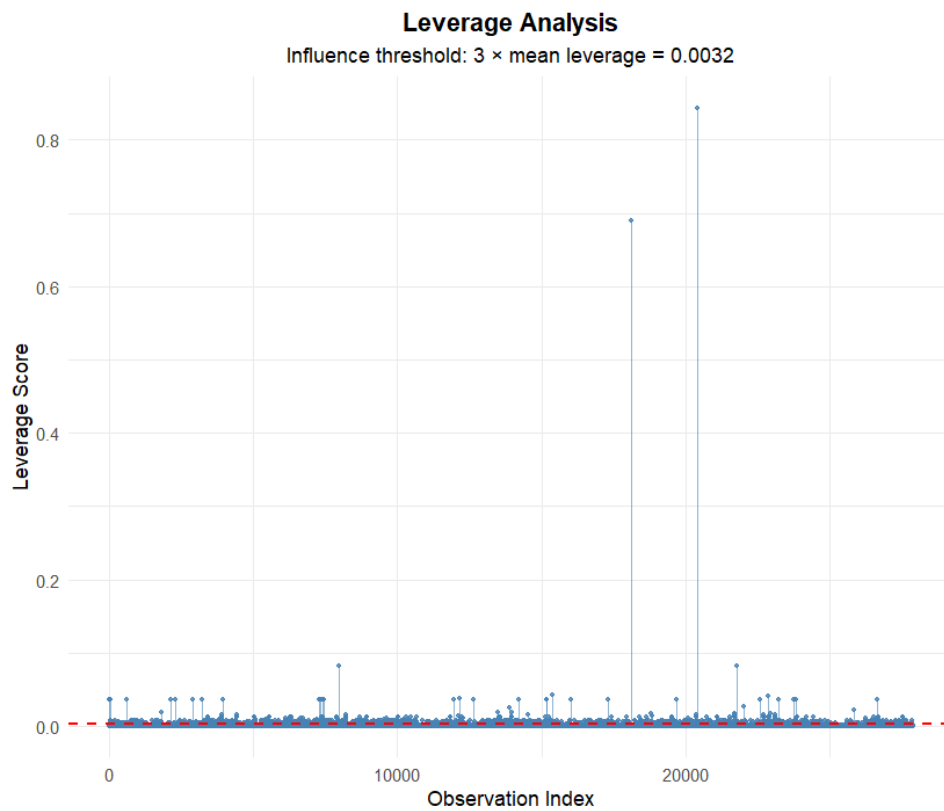


Figure 12: Leverage Plot of PCA Model

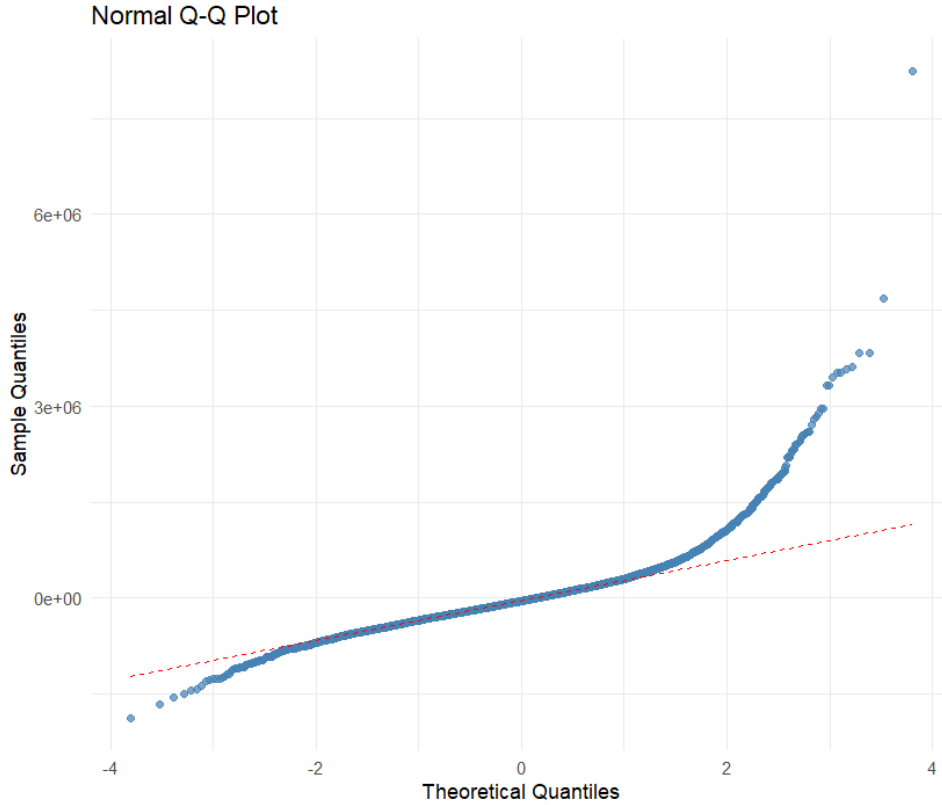


Figure 13: Quantile-Quantile Plot of PCA Model

4.6 Random Forest Model

Table 8 provides the summary results of random forest model. There were 500 decision trees and 5 random variables at each split. Figure 14 shows the scatter plot of the predicted price from the random forest regression model and the actual price of the Melbourne houses. This plot looks the best among all five plots. However, the reference line is almost centered between the points, with points scattered at higher prices. There are some extreme outliers (very high actual prices with much lower predicted values).

Figure 15 shows the plot of variable importance in the random forest model. Year-Built, Distance, and Latitude have the strongest influences, and Method, and ParkingArea having the weakest influence. Figure 16 shows the quantile-quantile plot of the random forest model. From the plot, the residuals do not appear normally distributed, and this is confirmed by Anderson-Darling Normality Test in table 10.

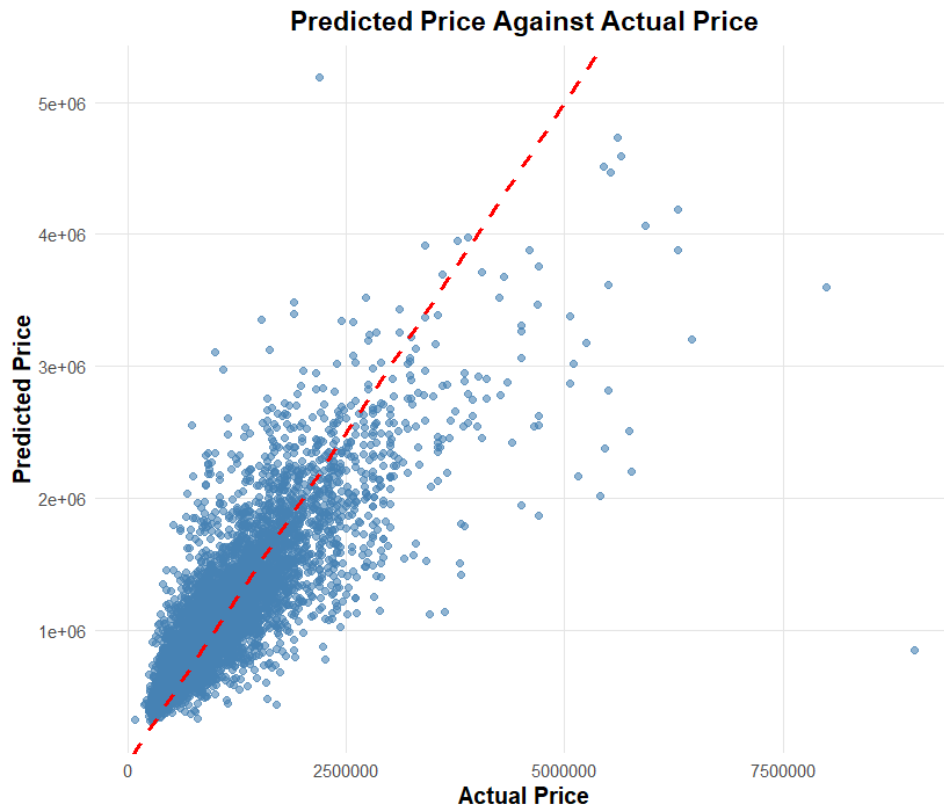


Figure 14: Plot of Predicted Price and Actual Price with Random Forest Model

Table 8: Random Forest Model Summary

Component	Description
Algorithm	Random Forest for Regression
Ensemble Size	500 decision trees
Feature Sampling	5 random variables considered at each split
Error Metric	Mean Squared Residuals: 1.388×10^{11}
Explanatory Power	69.17% of price variance explained
Training Data	27,888 observations

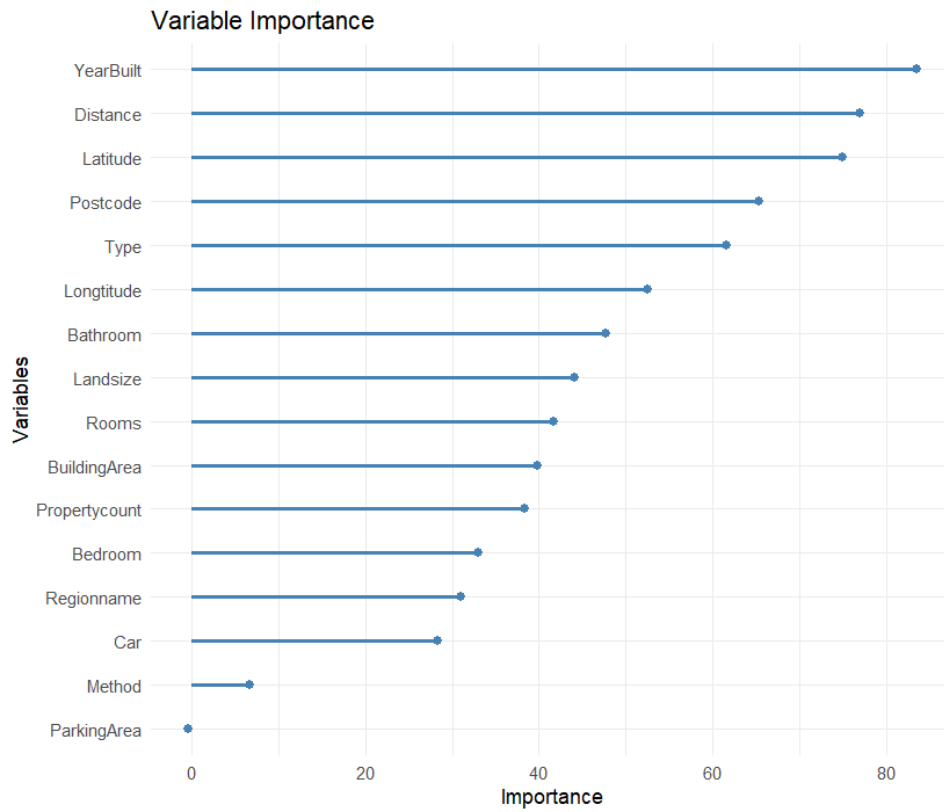


Figure 15: Variable Importance of Random Forest Model

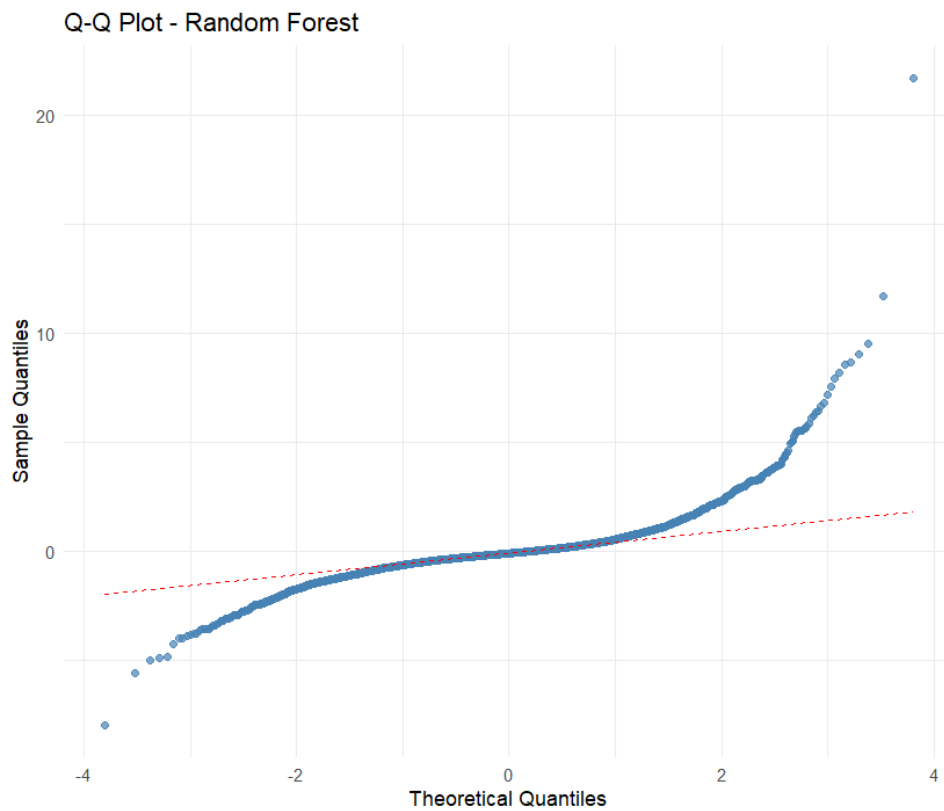


Figure 16: Quantile-Quantile Plot of Random Forest Model

4.7 Model Comparison

Table 9 compares the performance of different regression models using four metrics. These metrics are Mean Square Error (MSE), which measures the average squared difference between predicted and actual values; Adjusted R-squared, which measures the proportion of variance explained, adjusted for the number of predictors; Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC), both giving the likelihood of the model predicting well with new data. The Random Forest model has the smallest MSE, followed by OLS, Ridge, and LASSO. The PCA model has the highest MSE. The OLS model has the highest adjusted R-squared, followed by ridge, LASSO, and PCA. The random forest model has the smallest adjusted R-squared. In terms of AIC and BIC, the models follow the same order. Ridge has the smallest AIC and BIC, followed by LASSO and OLS. The PCA model has the highest AIC and BIC.

Table 9: Model Comparison Using Performance Metrics

Model	MSE	Adjusted R^2	AIC	BIC
OLS	192,751,253,776	0.5758	803590.6	803837.7
Ridge	192,753,511,657	0.5641	724449	724745.5
Lasso	193,505,929,063	0.5624	724597.3	724762
PCA Regression	202,534,492,230	0.5549	804934.1	805189.4
Random Forest	138,839,106,023	0.5409	-	-

Note: AIC/BIC is not applicable for random forest.

5 Conclusion

This analysis of Melbourne housing data provided several findings regarding predictive modeling techniques. The multicollinearity between predictor variables was substantial, as identified in the results, and this required strict model selection. Despite multicollinearity, PCA regression and standard regularized regression failed to attain the highest adjusted R-squared and lowest MSE. Regularized regression (Ridge and LASSO) was identified based on both BIC and AIC. The random forest model had the best predictive ability in predictive performance, with the lowest mean squared error and the maximum explanatory power for variability in results. The metrics for the model evaluation presented an inconsistent result: the ridge regression model had minimized the information criteria (BIC and AIC). Meanwhile, the Ordinary Least Squares had the largest adjusted R-squared. Visual examinations of predictive plots also validated the very good performance of the random forest model.

Diagnostic tests revealed strong limitations in all the models. Residual diagnostics never indicated evidence for normality assumptions, while outlier analysis revealed several influential cases that require further exploration. Despite these issues, the random forest model was found to be the most robust and efficient method to predict Melbourne house prices with a good balance of prediction accuracy and real-world use.

Future research should aim at overcoming the constraints noted by robust statistical techniques, exploring possible transformations of the variables, and examining influential cases so that the reliability and interpretability of the model are improved.

6 Appendix

Table 10: Anderson-Darling Normality Test Results for Model Residuals

Model	A Statistic	p-value	Normality Conclusion
OLS Model	628.46	$< 2.2 \times 10^{-16}$	Strongly Non-normal
Random Forest	293.63	$< 2.2 \times 10^{-16}$	Strongly Non-normal
LASSO	185.02	$< 2.2 \times 10^{-16}$	Strongly Non-normal
Ridge	167.60	$< 2.2 \times 10^{-16}$	Strongly Non-normal
PCA	155.53	$< 2.2 \times 10^{-16}$	Strongly Non-normal