Predicting House Prices with Machine Learning Techniques

Table of Contents

Executive Summary

Residential house sale prices have experienced a lot of uncertainties and some think that the housing market will change radically due to several factors both economic and housing characteristics. This report aimed to create a model that can predict future house sale prices to gain a competitive advantage within the housing market. Linear regression was used to predict house sales prices against variables that influence house sale prices. Out of 81 variables, 6 were chosen: YearBuilt, GrLivArea, GarageCars, TotalBsmtSF, YearRemodAdd, and OverallQual. All variables do not have any missing values or NA values. According to the correlation test, all six variables significantly correlate to SalePrice. Within the linear regression model, all variables positively impact SalePrice to increase with each respective unit with the model explaining roughly 77% of variation in the data. OverallQual influenced the increase in SalePrice the most out of all 6 variables followed by GarageCars. The F-statistic is also significant within the model, indicating that the model has a good predictive capability and is valid to make predictions. Further additions of economic considerations and external housing factors like neighbourhood may contribute to a more rounded prediction of house sale prices.

Introduction

House prices play an essential role in the economic position of a country (Miller et al., 2011). Naturally, house prices and values are expected to increase in the long term due to land scarcity(Manasa, 2020). However, many factors, such as the underlying economic elements or government policies, influence the price of houses (Algieri, 2013). Among these factors are the physical housing characteristics such as overall house quality and garage space. Residential house prices, such as those in New Zealand, have risen extraordinarily since the early 2000s and have been rising for the last two decades (Xin, 2009). Some think the housing market will change radically following the boom and bust cycle of house prices (Burnside et al., 2016 & Algieri, 2013). In order to gain advantageous insights into the housing market, a model for predicting house sale prices will be explored.

The report will cover the prediction of house sale prices using linear regression analysis, as regression techniques have been widely used in creating predictive pricing models (Manasa, 2020). Linear regression is a number of statistical processes for the determination of the relationship between a dependent variable and a number of independent variables (Freund et al., 2006). In this case, the model will try to predict the house sale price based on factors such as the remodel date.

The following will be covered in the report: descriptive statistics, correlation, modelling, discussion, limitations, and further steps.

Descriptive Statistics

Out of 81 variables, six were chosen from the available data to be used in the model. According to Sirmans et al. (2005), it is found that all six variables have been found to positively influence house prices in previous regression models.
The six variables that were used in the model are:

- YearBuilt - Original construction date of the house
- GrLivArea - Above grade (ground) living area in square feet
- GarageCars - Size of garage in car capacity
- TotalBsmtSF - Total square feet of basement area in square feet
- YearRemodAdd - The year the house was remodelled in
- OverallQual - Overall material and finish quality rating

None of the variables used had any missing values or NAs. The average house has a living area of 1,515 square feet and a basement area of 1,057 square feet. Moreover, it can fit two cars and has an average overall material and finishing rating of 5.

Most of the houses were built before the year 2000, but the remodelling for all the houses started after 1940. This includes houses built before 1900 which means that some of these houses were not modelled for at least 40 years, and the difference between the year a house was built and the year it was remodelled decreases with time.

Correlations

The correlation test is used to evaluate the association between two or more variables (Good, 2009). The test gives a deeper understanding of the degree and direction of that relationship in a quantitative manner. With the degree of correlation represented on a continuous scale along the range of negative one to positive one. To further elaborate on this quantitative correlation, a relationship between two or more variables ranging from zero to positive one, signifies a positive correlation and the closer to positive one, the stronger the relationship. And vice versa, when the relationship ranges from zero to negative one, it is negatively correlated and the closer to negative one, the stronger that correlation is.

Having highlighted the general idea behind correlation test, the test was performed on the house dataset with a focus on understanding the relationship between different variables and the SalePrice variable. Table 1.1 displays the variables of interest that had a strong correlation along with a positive relationship with the SalePrice.

Furthermore, the P-value is the probability of seeing the correlation result in table 1.1 given that the null hypothesis is true. In a correlation test, the null hypothesis states that there is no relationship between the variables. And as observed in table 1.1, the P-values of zero can be interpreted as a zero percent chance the results from the sample occurred by chance. Indicating significance and therefore can reject the null hypothesis and accept the alternative hypothesis that there is a linear relationship between the dependent variable (SalePrice) and the independent variables.

As such, the variables of interest highlighted in table 1.1 display a strong, positive and linear correlation with SalePrice followed by a significant P value. This provides further proof to proceed with these variables in building the linear regression model.

Model Results

The linear regression output is how the analyst gauges choosing and testing the validity of the model. The model output (Table 1.1) consists of Coefficient estimates, which are the constant values that are multiplied by the variable to calculate the sale price of a house. These coefficients also have p-values that represent the statistical significance of the variables. The model output also contains an f-statistic, multiple R-squared, and adjusted R-squared values to measure the overall validity of the model.

Our models output predicted the intercept coefficient as -$1,053,000, which is the predicted sale price of a house. This value does not make any practical sense as it is negative however because it is just the intercept value, we are confident that the other variables in the model will make our house sale price predictions above zero. The estimated coefficient for the first variable input in the model is 221.10. This indicates that for every one unit increase in the year, or for every one year newer a house is, our model expects an increase in sale price of $221.10. Every one unit increase in year starts from the oldest year a house was built in the dataset.

All other coefficients follow the same rule, that for every one unit increase in the variable, the coefficient is multiplied against it to determine a dollar value that goes towards calculating the sale price of a house. This process is called a linear regression equation, it allows the user to calculate the sale price of any house based on the models predictions for each variables coefficients. Once all of these calculations are made, the model outputs a predicted sale price for a house based on the parameters.

All of the variables inputted into the model are highly statistically significant. We can see this with the three stars next to all of the p-values. This tells us that the variables we have inputted into the model are very relevant variables to include in the model and make good predictions with. The f-statistic p-value is also highly statistically significant which tells us that the overall validity of our model is very good and we can use this to make good predictions about the sale price of houses.

Finally, there are two other measures of model validity that tell us our model is fit-for-purpose and is a suitable one to make predictions. The adjusted R-squared value is 0.77, or 77%. This means our model can explain 77% of the variation in the data for the response variable SalePrice. 77% is a very high value for behavioural/social models that are trying to predict human behaviour.

The other measure is the root mean squared error (RMSE). This value for our model was 39216.41, this is a metric that tells us the average distance between the actual values in the dataset versus the predicted values from the model.

Generally, the lower the RMSE the better fit the model is. The RMSE is particularly useful when comparing it to other models and observing which has the lowest RMSE. In our code, we compare the RMSE from the training and test sets of the data. Our test set RMSE was 37689.97, less than the training RMSE which tells us that our model slightly overfits the data. However, we are still confident in using our model because of its predictive accuracy and the other statistical measures that have tested the validity of it. Also, when building a model to predict numeric quantities such as dollar values, a model that overfits the data is better than underfitting the model because we want our model to have high predictive accuracy as we are trying to predict the sale price of a house.

Discussion

Our purpose is to build a linear regression model to predict housing prices in the future based on the previous dataset. We include six numerical variables in our model that are positively correlated with housing prices. According to the statistical result, the model performs well on the testing data. Moreover, using the linear regression model creates simplicity in terms of implementation and interpretation. However, there are some limitations to the model.

First, the current model fails to reflect the current housing market. For instance, the estimated selling price of a particular property in 2023 will be the same as in 2011. Therefore, we need to consider the time value of money and the rapid growth of housing prices. In contrast, the median price of New Zealand residential properties has increased by more than 100 percent since 2012 (Figure 1.1).

Second, the model ignores categorical variables, such as neighbourhood, zoning, and building type. However, those factors have important impacts on housing prices. For instance, houses in Remuera are always sold at higher prices than in Glenfield.

Moreover, the model overlooks the importance of economic factors such as economic growth, inflation, and interest rate. For example, New Zealand house prices have recently seen the biggest fall in the past five years. One of the crucial reasons is that the reserve bank increases the official cash rate to slow inflation (Figure 1.2). This policy caused a sharp rise in home loan interest rates to roughly 6%, comparing the mortgage rate to only around 2% in 2021 (Figure 1.3). An increasing home loan interest rate means home buyers borrow less money at higher costs. Therefore, the willingness and affordability to purchase houses are decreasing, causing a big fall in the housing market.

Conclusion

Overall, the test result of our model is good. Therefore, our model will successfully predict future house prices if we can overcome the above limitations. There are three recommendations for modifying our model.

First, we should include the growth rate of residential property when modifying our model. In particular, we should identify the pattern of house price changes and calculate the house's average growth rate. Second, we should convert categorical variables into numeric variables. For instance, we could rank the neighbourhood from one to ten. One represents the worst neighbourhood. Ten represents the best neighbourhood. Nevertheless, we need to be careful about how to rank neighbourhoods. Finally, we should add economic factors into our model, such as the economic state.

References

Algieri, Bernardina (2013). House Price Determinants: Fundamentals and Underlying Factors. Comparative Economic Studies, 55(2), 315–341. doi:10.1057/ces.2013.3

Burnside, C., Eichenbaum, M., & Rebelo, S. (2016). Understanding booms and busts in housing markets. *Journal of Political Economy*, *124*(4), 1088-1147.

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Elsevier.

Good, P. H. I. L. L. I. P. (2009). Robustness of Pearson correlation. *Interstat*, *15*(5), 1-6.

Janet Ge, Xin. (2009). Determinants of house prices in New Zealand. *Pacific Rim Property Research Journal*, *15*(1), 90-121.

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

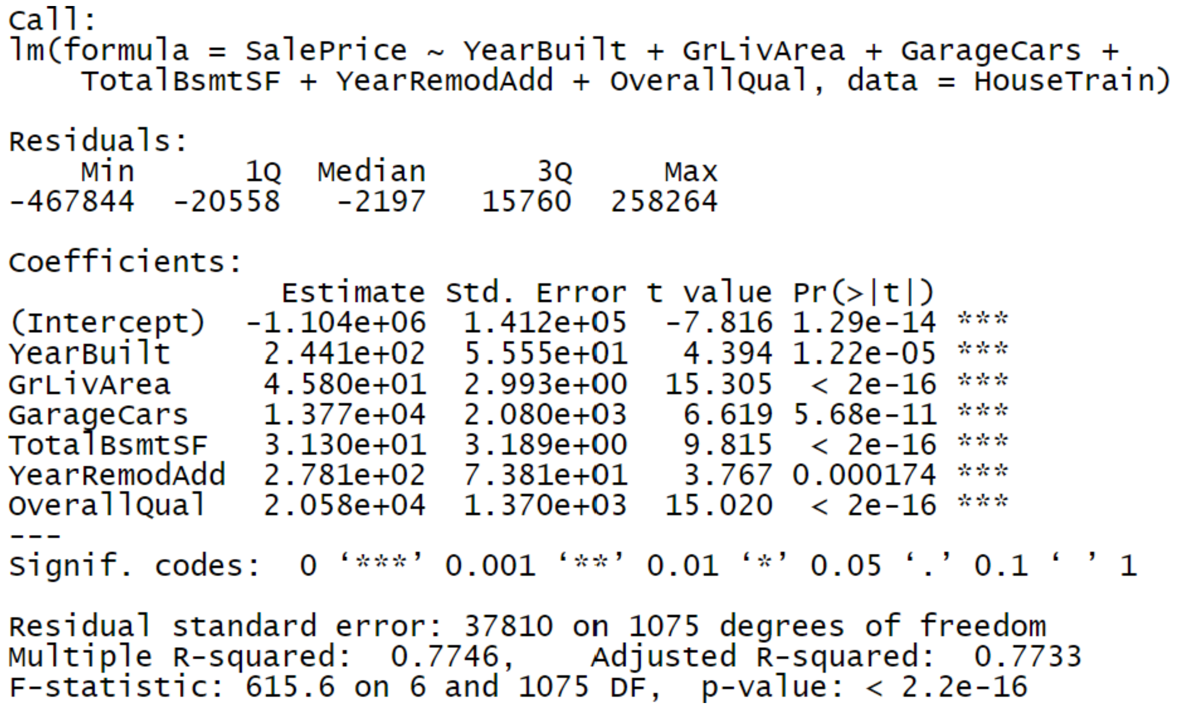
Miller, N., Peng, L., & Sklarz, M. (2011). House prices and economic growth. *The Journal of Real Estate Finance and Economics*, *42*(4), 522-541.

Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), 3–43. http://www.jstor.org/stable/44103506

Appendix

Variables	SalePrice	P-value
OverallQual	0.79	0.00
GrLivArea	0.71	0.00
GarageCars	0.64	0.00
TotalBsmtSF	0.61	0.00
YearBuilt	0.52	0.00
YearRemodAdd	0.51	0.00

Table 1.1: Correlation Test Result



```
Call:
lm(formula = SalePrice ~ YearBuilt + GrLivArea + GarageCars + 
    TotalBsmtSF + YearRemodAdd + OverallQual, data = HouseTrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-467844  -20558   -2197   15760  258264 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.104e+06  1.412e+05  -7.816 1.29e-14 ***
YearBuilt     2.441e+02  5.555e+01   4.394 1.22e-05 ***
GrLivArea     4.580e+01  2.993e+00  15.305  < 2e-16 ***
GarageCars    1.377e+04  2.080e+03   6.619 5.68e-11 ***
TotalBsmtSF   3.130e+01  3.189e+00   9.815  < 2e-16 ***
YearRemodAdd  2.781e+02  7.381e+01   3.767 0.000174 ***
OverallQual   2.058e+04  1.370e+03  15.020  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37810 on 1075 degrees of freedom
Multiple R-squared:  0.7746,	Adjusted R-squared:  0.7733 
F-statistic: 615.6 on 6 and 1075 DF,  p-value: < 2.2e-16
```

Table 2.1 : Linear regression model output of the training data.

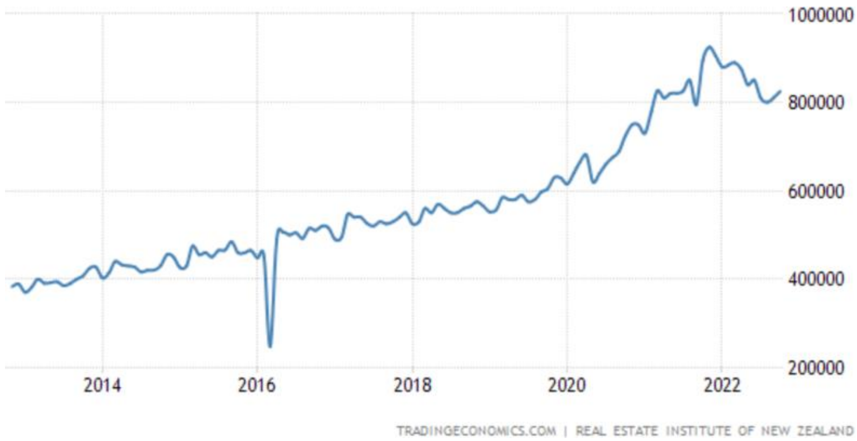


Figure 1.1: Median price of New Zealand residential property 2012 to 2022.

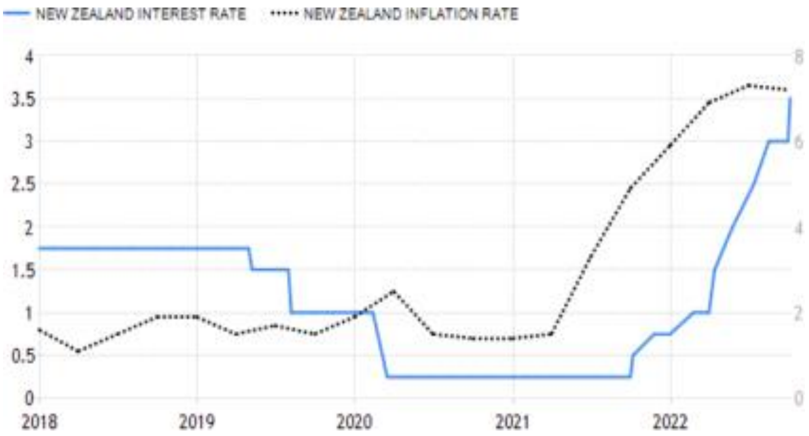


Figure 1.2: The relationship between New Zealand Interest Rate and Inflation Rate

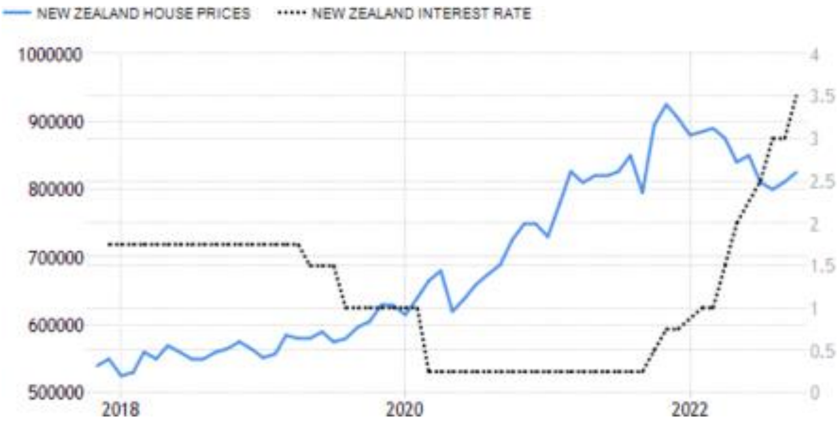


Figure 1.3: The relationship between New Zealand House Price and Interest Rate