Analysis and Report

R.Riddell

2023-10-09

Introduction

The housing market changes can have ripple effects through large sections of the economy. The effects can be seen in areas like consumer spending or directly impact the construction industry. Having a broader understanding of the factors that affect housing prices can better equip people to make decisions when entering or exiting the market RBA-2019. This report will examine house prices concerning broader factors surrounding the housing market between 2016 - 2018 and seek to shed some light on the key drivers at that time.

Data

The data in this report is from publicly available sources, including the ABS (Australia Bureau of Statistics) and kaggle. The data can be found at ABS DATA, Melbourne Housing, Postcode/ LGA data. The data was aggregated to create a single data set that is used for this exploration and analysis.

The 2018 Melbourne Housing data was used as the central data source. The data was then enriched with local area data that was obtained from the ABS. The two data sets couldn't be directly connected as the area identifier on the ABS data was the LGA (Local Government Area), which wasn't in the housing data. A linking table was used, this table came from an ABS source and gives the LGA code for each postcode with this information. The ABS data can be mapped to the postcode field by the LGA code and name. This can be linked to the housing data using a fragment of the council area and the postcode. This method has some risks, as the same postcode can be linked to multiple LGAs. This risk has been mitigated using the Label and Council Area values from the ABS and Housing data, respectively.

The additional data from the ABS was centred around the Economy/Industry and Education/Employment based on the LGA. This information was collected at the same time as the Melbourne Housing data. When multiple years of was data avaiable, the data was averaged over 2016 - 2018. There could be additional work to look at this data in direct relation to the sale data and if lead/ lag from those values could be a predictor or a response. The LGA grouped all values, and the additional variables added were:

ECONOMY AND INDUSTRY, Local Government Area, 2011, 2016-2022

Total house transfers - Taken directly from the report
Total Motor Vehicles - Taken directly from the report
Age of Motor Vehicles - Calculated from columns under Registered moto vehicles - Year of

manufacture - at 31 January (CS - CU). The calculation was done by getting the total of the three columns and then calcualting the percentage of vehicles that were less than five years old.

EDUCATION AND EMPLOYMENT, Local Government Area, 2011, 2016-2022

Total Number of Jobs - Taken directly from the report(Number of employee jobs - Total (AH))
Most popular job occupation - Calculated from columns under Occupation of employed persons - Persons aged 15 years and over - Census (CD - CL). The extraction of this was done by taking the label that correspondes to the highest percentage selected in each LGA.
Most popular job category - Calculated from columns under Jobs in Australia - year ended 30 June (O - AG). The extraction of this was done by taking the label that correspondes to the highest number in each LGA.

NA Treatment

After assessing all the columns that had more than 20% null values, the decision was made to drop them. It didn't make any clear sense to impute 0 into these columns; the values were then assessed against the whole data set and the other data points in their suburb. The range was reasonably significant on both counts, so mean imputation didn't seem practical. The observations that included no final price were also dropped. As the price is the aim of the investigation, it was decided not to create target values synthetically. Seven additional values were dropped as they had some NA values in the Postcode/ Property Count. This has left the data set with 25198 observations and 15 variables.

The columns dropped were BuildingArea, YearBuilt, Landsize, Car, Bathroom, Bedroom2, Lattitude, Longtitude, Code, Address, SellerG, Postcode, Suburb, CouncilArea.

Exploratory Analysis

Distribution

The numeric values were analysed using histograms. All were relatively normal. The average number of house transfers shows a normal distribution, with some areas having significantly more (figure 1). The ratio of vehicles under five years shows some right-hand skewness, implying that some areas have a notable proportion of vehicles under five years old (figure 2). There is no apparent reason to drop any of these values based on their distribution.

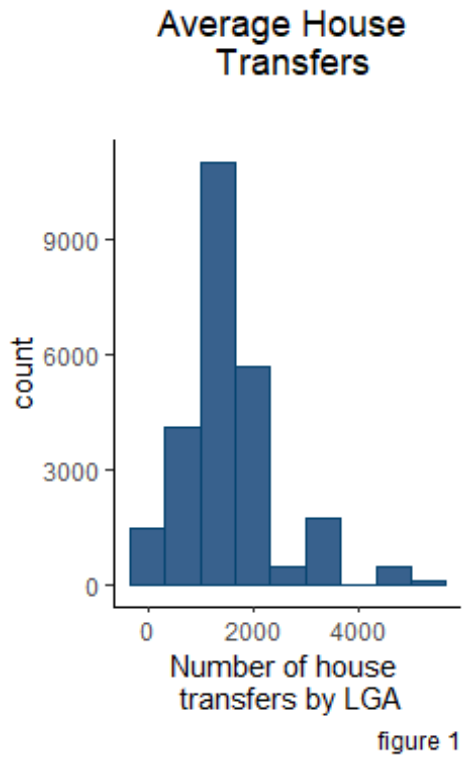


figure 1

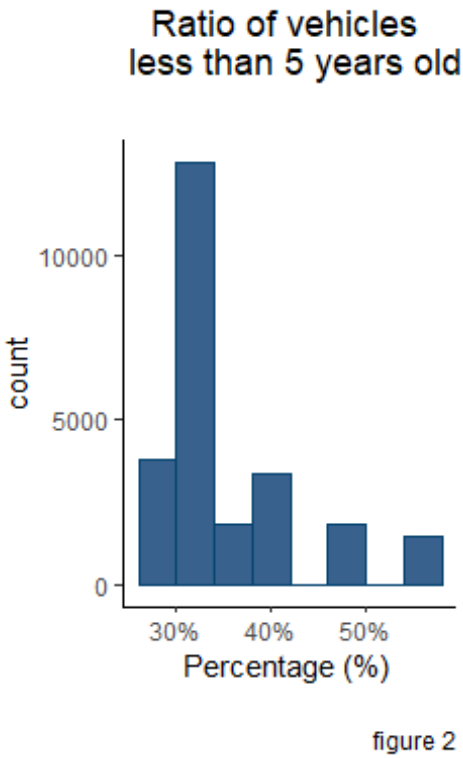


figure 2

The distribution of some of the categorical variables was assessed against the price. In some cases, the distribution was significant across each category and showed some potential outliers. In particular, the most common occupation shows that “Professionals” have significantly more buying potential in some cases, but their median is somewhat similar to the other professions (figure 3). When looking closer though, the data set is largely skewed by Professionals as they account for 86% of the responses. The Region Name also appears to be a factor in the price, with the regions with the higher medians having the most significant distribution. This may point to the fact that a house in every region is valued at the same price, but it could be assumed that it is of a different quality, size or age (figure 4). When analysing the distribution of values in the data set, three regions provide roughly 80% of the data: Southern Metropolitan, Northern Metropolitan and Eastern Metropolitan.

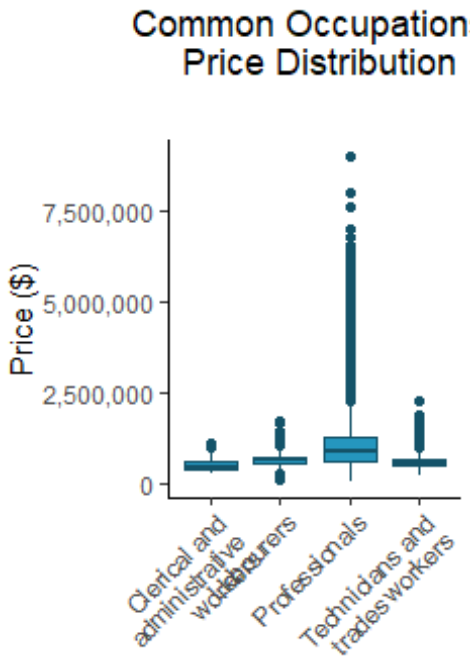


figure 3

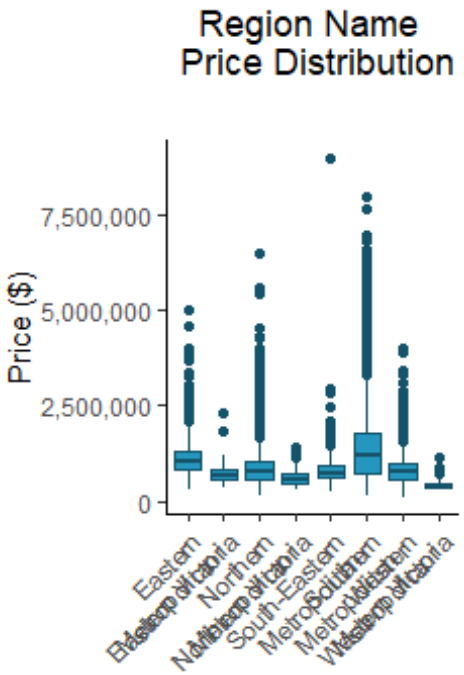


figure 4

Relationships

The correlation coefficient of the numeric values against the price was calculated. The Rooms variable showed the highest positive correlation with a result of 0.46, showing that a property with more rooms may result in a higher sale price. The strongest negative correlation was the Distance from the CBD; this shows a slight relationship that the sale price will be less when a house is further from the CBD. Purely from the correlation coefficients, the Property count and average total jobs have been excluded as they are between -.10 and .10.

Assumption

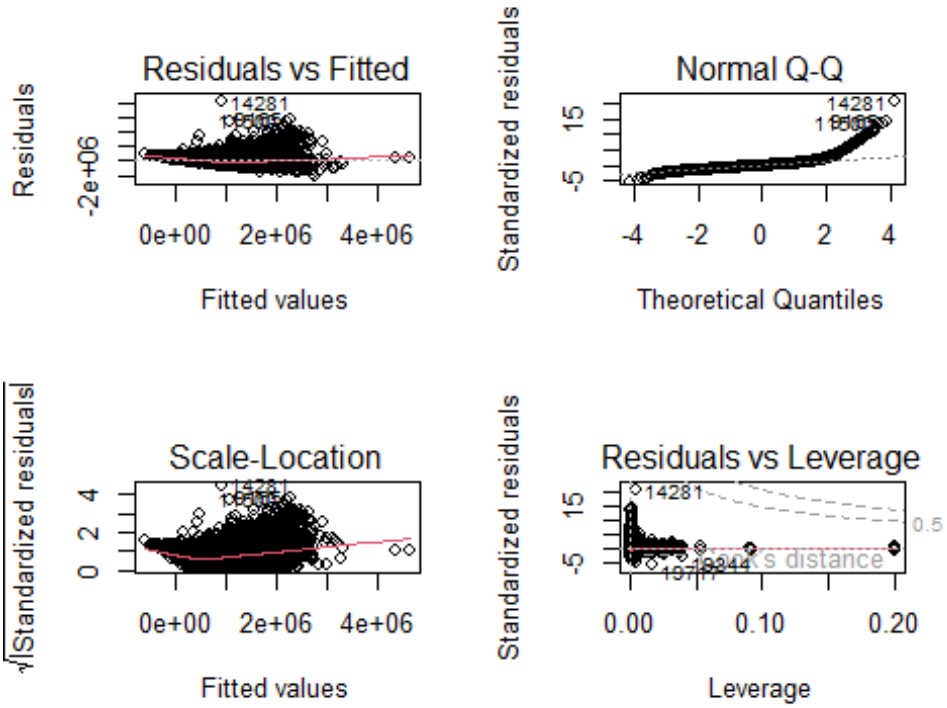
Before attempting modelling, it is essential to check the data to ensure it fits the models' assumptions. This process helps to select the models that are most appropriate to the data, it will provide a more valid result and the model should be more reflective of the data and therefore more accurate. As a starting point, a linear model has been created with the price as the target variable, and it will be tested against the assumption for the linear model.

Continous Variable

To test the data structure and makeup, a linear regression model is used. This means the target variable must be continuous.

Outliers

Using Cook's Distance and examining the residual plots, we see significant issues with how a linear model represents the data. With some, the Residual plot has an obvious pattern where it should be more random. The Q-Q plot does not follow the line and has a significant upward curve.



As the residual plots showed the data was not well represented by linear modelling, it doesn't make sense to continue testing the linear modelling assumptions against this data. Other tests that could be completed are for homoscedasticity, near zero variance or multicollinearity.

This information has assisted in understanding what regression modelling techniques may be suitable; as such, the models chosen to test will be random forest regression and KNN regression. KNN regression assumes that the closer a data point is to another, the more similar it will be. Random forest regression (decision trees) assumes that data can be split into subsets reasonably distinctly.

Modelling

The data is split randomly into two groups of testing and training; the testing group is 20% of the observations, and the training is 80%. The training data is then used to build the models; once models are built, the testing data can be used to assess the RMSE and R squared. In the model-building phase, there is also repeated cross-validation. This divides the data into several partitions, with some partitions not used for model building. The

unused partitions can then be used as an internal test, and the mean performance is reported across the iterations.

Results

Random Forest

We can extract a variable importance graph after applying the regression model to the data. This graph shows the variables the model found to be most important in determining price. The variables reported with the highest importance for this data were the number of rooms, type of house being a unit and how many house transfers there had been in the larger area. These three variable are very important in predicting price, and it could be valuable to investigate the specific relationship between rooms, units house transfers and price to better understand why there is a strong link.

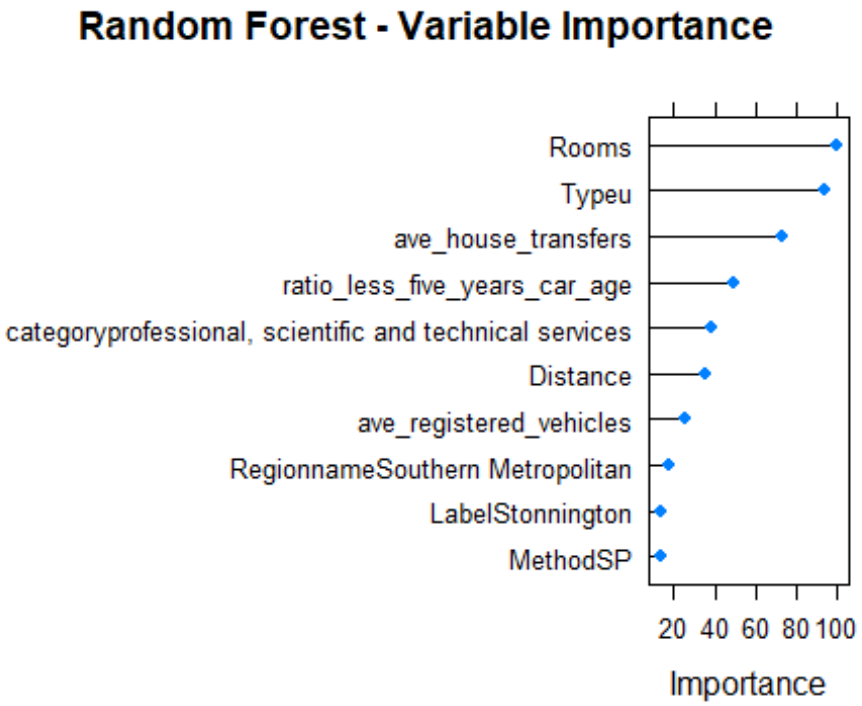


figure 5

KNN Regression

The KNN model showed a decrease in RMSE as the number of neighbours was increased. This could show some overfitting with low numbers as it relies too heavily on the data adjacent to each point. The elbow plot indicates that the KNN model performed best with a k value of nine. As the k value increases, the RMSE seems to also increase, suggesting that k of nine will be the best value. More hyperparameter tuning could be attempted with a k of

ten could be tested as the tuning went from nine to 11.

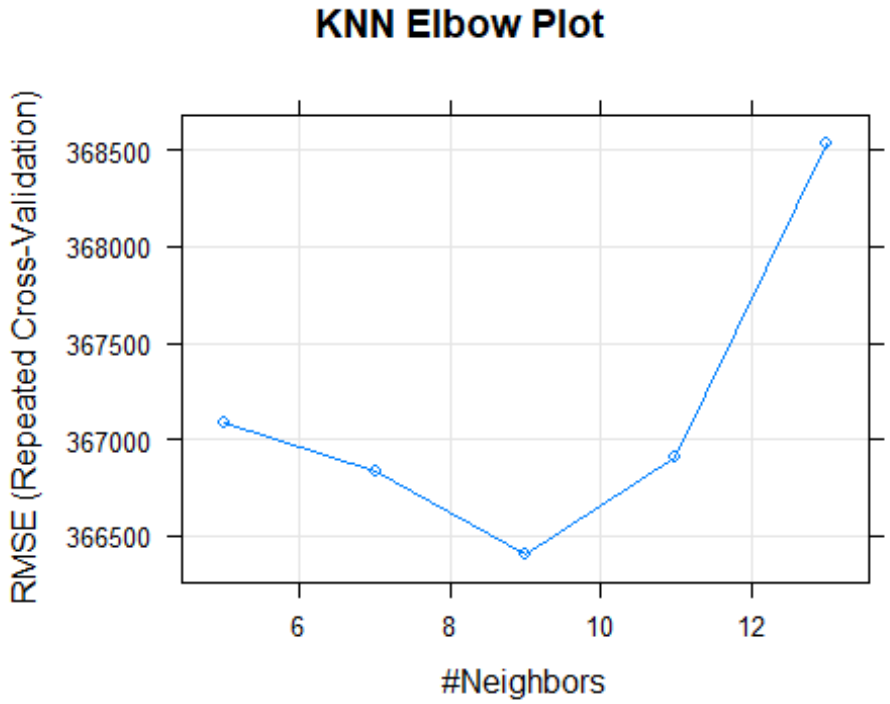


figure 6

Comparison

When comparing the two models RMSE and R squared values the random forest returns worse with and RMSE of 388,809 and an R squared of 0.62. In comparison, the results from the KNN method were 367,543 and an R squared of 0.64. This shows that neither model is particularly well suited to the data, and our values could better predict house prices.The KNN does have better performance but both could still be improved with further feature engineering and hyper parameter tuning.

Conclusion

With neither model showing solid results and the best RMSE of nearly ~$360,000, it can be concluded that this data set is not an accurate predictor of house prices. When assigning the value that the new variables have created, we can investigate the variable importance from the random forest model. Interestingly, the number of house transfers over that period was the third most important when making the predictions. An interesting angle to investigate could be how a quiet market vs a busy market affects house prices. The fourth most important variable was the ratio of vehicles under five years old. A further analysis into the vehicles type (Car, Campervans, Motorcycles) could be interesting in unpacking how and when people spend their money and if there is more of a correlation between vehicles of certain types/ ages associated with areas with higher selling houses. The areas where the most popular job category were Professional, scientific and technical services, which was also seen as relevant predictor and was the fifth most important. The number of

registered vehicles was also seventh in importance, and the data suggests that once the number of registered vehicles exceeds 150,000, house prices will be, on average, less than 1 million dollars. Around 100,000 registered vehicles, with 35% being less than five years old, show a strong cluster with average prices over 1.5 million dollars. The variables for the most popular occupation stayed in the data set but didn't have much of a relationship to house price. The total number of jobs had such a small correlation it was dropped earlier in the analysis.

As a final test, the KNN model was built on a data set that excluded the ABS features. This data set returned an RMSE of 354,666 and an R sqaured of 0.64. This result shows that the ABS data may add some additional detail to the modelling process but does not improve the evaluation metrics. However, the distinction between the three sets of evaluation metrics is inadequate to give a clear picture of the drivers around house pricing.