Analysis of the Melbourne Real Estate Market

Principal Investigator: Rufus Petrie (rpetrie@wisc.edu)

In this paper, I examine the real estate market in Melbourne, Australia. Until late 2017, housing prices in Australia had been rising rapidly. However, this growth tapered off towards the beginning of 2018, and prices started declining significantly later in the year. Most of my family lives around Melbourne, and I like the idea of living there, but I also wouldn't want half of my home equity to blow up overnight. Therefore, I want to get a better sense of the real estate market and how prices behave. In particular, I examine how prices vary between different regions of Melbourne, what factors play the largest role in determining housing prices, and the degree to which models can predict these prices. In this paper, I find that real estate listings near the CBD (Central Business District) and its eastern suburbs have the most extreme prices, and postcode income per capita and building sizes alone can account for a large amount of variation in prices.

Data Overview

For this project, I combine three separate datasets. The first comes from Kaggle, where a member continually scrapes data about the Melbourne housing market from online listings. This data has a variety of information about each listing and includes variables like price, number of bedrooms/bathrooms/parking spaces, age of building, etc. Because the general economic status of the area plays a large role in determining housing prices, I augment this data with tax information from the Australian government. In particular, I find the income per capita and working population for each postcode that appears in the listings data. Finally, I use GIS data from the Australian government so that I can get a

spatial understanding of the other data. Combining these three datasets at the postcode level gives a solid overview of the Melbourne housing market.

Table 0.1 in the appendix contains a full description of these variables. Note that all price variables in this table are recorded in Australian dollars (A$1 $\approx$ $0.72 as I write this), and all distances and size variables are recorded in kilometers and square meters, respectively. From this table, we see that the average listing in the Melbourne real estate market goes for about A$1,050,000. At the extremes, one listing went for about A$85,000, and another listing went for over eleven million Australian dollars. Table 0.2 in the appendix describes the correlation of each variable with listing price. As we would expect, variables like amount of rooms and postcode income per capita have strong positive correlations with listing price, and variables like year built and distance to the CBD have negative correlations with price.

Examining prices at the postcode level also gives more insight into the housing market. Figure 0.1 in the appendix breaks down the average price of real estate listings by postcode. As we can see, postcodes towards the center of this figure have the highest average listing prices. This area corresponds with the downtown/CBD area of Melbourne. Additionally, we see that prices remain reasonably high to the east of the CBD, an area that corresponds to the nicer suburbs of Melbourne. As you move away from these areas, prices gradually start to decline.

We can also get a better sense for the structure of this market by looking at clustered neighborhoods. In particular, I use K-means, an algorithm that turns some relevant variables into standard normals and sorts them into K groups that minimize within-group mean squared distance from the center of the group. This method gives a solid way of comparing regions and makes it so that variables with different scales don't take

precedence over each other. Figure 0.2 in the appendix depicts the Melbourne real estate market colored into four different clusters. As we can see, the cluster that corresponds to the CBD is completely connected, whereas the other clusters are broken into a bunch of separate parts. This provides further evidence that the real estate listings near the CBD have some qualitative difference from the rest of the listings.

Methodology

To better understand the role that each variable plays in determining prices, I generate a variety of models that attempt to predict housing prices. To judge their performance, I split the data into training sets to select models and test sets to evaluate the performance of those models. For every test/train split, I allocate 80% of the data to the training set and 20% to the test set. To generate these splits, I set a seed in Python so that I never use any of the test data before I evaluate my models. For the models that require tuning, I select the parameters that lead to the best 10-fold cross validated MSE on the training set. Finally, I compute the MSE of these models on the test set to judge how they would perform on new data. For all of my models, I used all available variables and dropped NA values.

For my models, I start by estimating a simple linear regression using all available covariates. To judge whether this overfit the data, I also fit a lasso and compared their performances. For the lasso, I scaled the covariates to standard normals and used the value of lambda that provided the best training set accuracy. Because the lasso led to a trivial increase in predictive performance, I then moved onto some more flexible algorithms. In particular, I performed a k-nearest-neighbors regression using the standard normal scaled covariates. For the final KNN regression, I ended up using ten neighbors.

Finally, I fit a random forest to the un-transformed data and adjusted the number of trees and amount of covariates to get the best performance. The final model used six maximum features and grew 128 trees. I also tried fitting some neural networks, but I couldn't find find a way of scaling the data that would give reasonable results.

Results

Table 0.3 in the appendix contains the results of my models. As we can see, the random forest had the best performance, followed by the k-nearest neighbors regression, lasso, and OLS. In particular, the final random forest had a test MSE of 7.4653×10^{10} out of sample and explained over 82% of the variation in the test set. Because MSE's are difficult to interpret at this scale, I also calculated the mean absolute deviation for the test set precitions, which equalled 160,146.71. In other words, when estimating the prices of real estate listings, the random forest missed by about A$160,146.71 on average. Therefore, while this model may have explained a good amount of variation in the data, we wouldn't necessarily be confident in its predictions when buying or selling houses in Melbourne.

Although the random forest does not have a concrete equation or coefficients to interpret, we can still get a sense of how it explained variation through its variable importances. These importances roughly correspond the reduction in MSE associated with each variable across all the trees in the forest. Table 0.4 in the appendix contains the variable importances for the final random forest. From this table, we see that building area and postcode income per capita had significantly higher importances than any of the other variables. Furthermore, some of the variables that have high correlation with listing price don't necessarily drive much of a reduction in MSE. In particular, the variables describing the amounts of rooms and parking spots had fairly low importance for

the final random forest. Therefore, some of these variables probably contain redundant information for determining prices, and future attempts of these models should probably include a greater variety of variables.

Conclusion

In conclusion, I find that the CBD and its eastern suburbs have the most extreme prices of anywhere in the greater Melbourne Area. While some of my models explained a good amount of variation in prices, they still had shaky performances in real terms. In particular, postcode income per capita and building size alone explain a great deal of the variation in housing prices; however, the estimates from my best model missed the mark by over A$160,000 on average. Future models could improve upon this by either finding a better way to deal with some of the extreme housing prices or getting more covariates. Additionally, nothing in this analysis suggests why prices have been declining so rapidly in the Melbourne area. Future research could expand on this by using the times of listings or sales to judge the evolution of the market over time. Future research could also attempt to quantify the degree to which foreign money enters the market, as foreign investment has led to large price increases in other real estate markets over the past few years.

Appendix

Table 0.1: Summary Statistics

Variable	N	μ	σ	Min	Max
Price	27246	1.050×10^6	6.414×10^5	8.500×10^4	1.120×10^7
Income per Capita	27246	6.961×10^4	2.249×10^4	3.999×10^4	2.046×10^5
Rooms	27246	2.992	0.955	1.000	16.000
Bedrooms	20806	3.046	0.955	0.000	20.000
Bathrooms	20800	1.592	0.701	0.000	9.000
Distance to CBD	27246	11.280	6.787	0.000	48.100
Parking Spots	20423	1.715	0.994	0.000	18.000
Land Size	17982	593.489	3757.266	0.000	433014.000
Building Area	10656	156.835	449.223	0.000	44515.000
Year Built	12084	1966.609	36.762	1196.000	2019.000
Property Count	27244	7566.781	4492.382	83.000	21650.000
Working Population	27246	15056.770	8223.934	395.000	55342.000

Table 0.2: Correlation of Variables with Price

Variable	Correlation
Rooms	0.465
Income per Capita	0.443
Bedrooms	0.430
Bathrooms	0.430
Parking Spots	0.202
Building Area	0.101
Land Size	0.033
Property Count	-0.059
Working Population	-0.172
Distance to CBD	-0.211
Year Built	-0.333

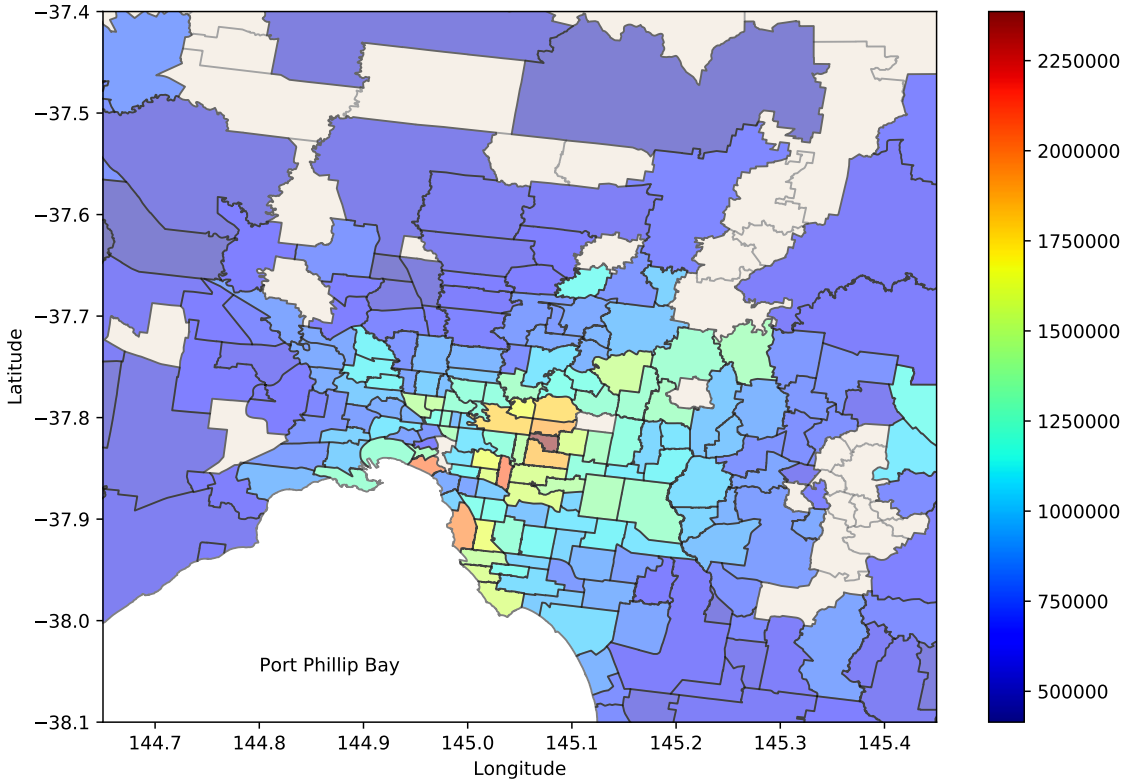
Table 0.3: Model Results

Model	Test MSE	Test R-squared
OLS	1.4839×10^{11}	0.6482
Lasso	1.4835×10^{11}	0.6483
KNN (K=10)	1.2023×10^{11}	0.7150
Random Forest	7.4653×10^{10}	0.8241

Table 0.4: Variable Importances for Random Forest

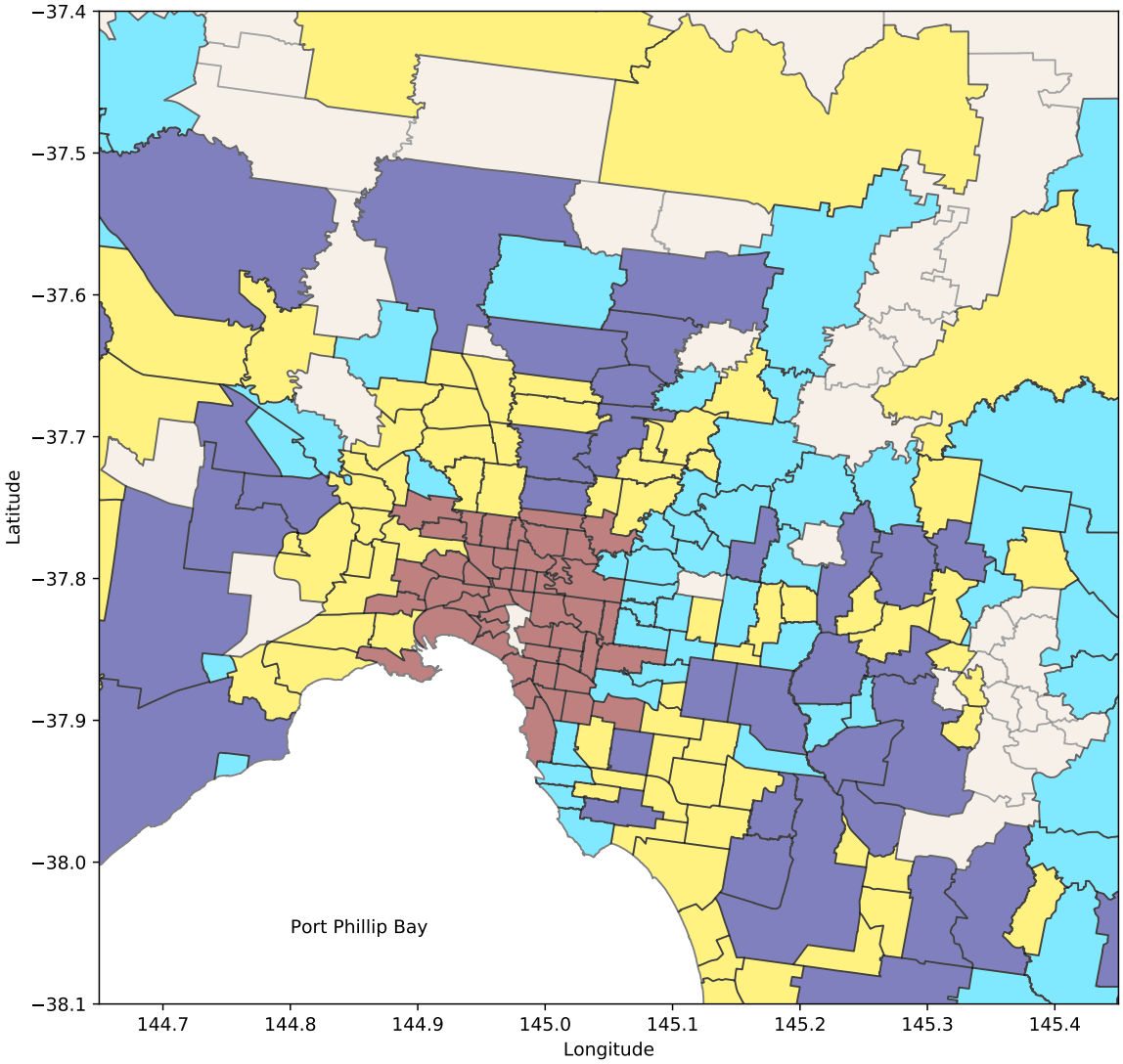
Variable	Importance
Income per Capita	0.268
Building Area	0.244
Year Built	0.096
Distance to CBD	0.070
Land Size	0.065
Rooms	0.063
Bathrooms	0.046
Home	0.032
Working Population	0.031
Bedrooms	0.026
Property Count	0.025
Unit	0.019
Parking Spots	0.014
Townhouse	0.001

Figure 0.1: Average Real Estate Price by Postcode



Note: tan areas correspond to postcodes with no observations.

Figure 0.2: Clustered Postcodes



Note: tan areas correspond to postcodes with no observations.