Auckland House Prices Analysis

Gavin Cheng
July 2020

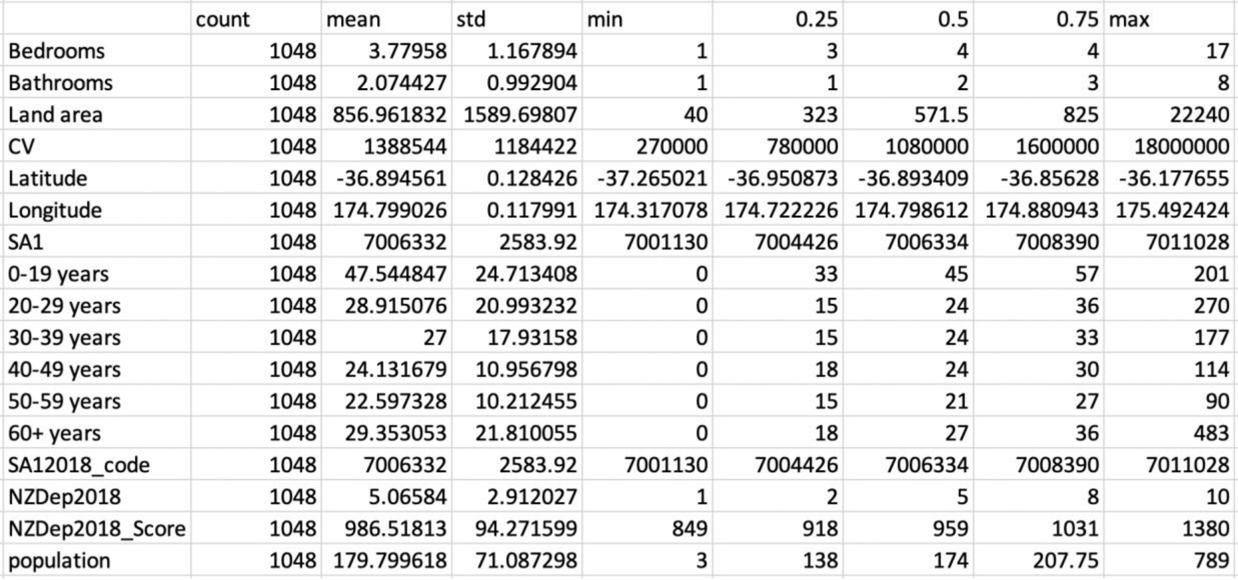
Executive Summary

We want to predict the house price, i.e., Capital Value (CV), with all of the explanatory variables. The explanatory variables include number of bedrooms, number of bathrooms, address, land area, latitude, longitude, SA1, number of people whose age is within some ranges in the SA1 unit area, suburb name, deprivation scale (NZDep2018), deprivation score (NZDep2018_Score), and population. These data come from the 2018 census in New Zealand.

Initial Data Exploration

We start our initial data exploration by checking null values in the data. We found three rows that contain null values and simply removed them. We then found the data type of Land area is object instead of int64 or float64. The problem was the inconsistent format of the column (some rows are like "123" and some are like "123 m^2"), and was resolved by extracting numbers from the column and converting the data type to float64.

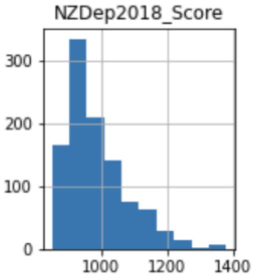
The next step is to check some summary and descriptive statistics. The descriptive statistics include minimum, maximum, mean, median, standard deviation, and distinct count for numeric columns.



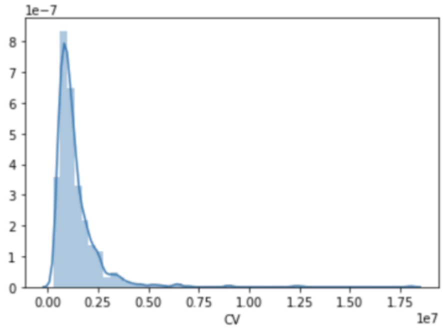
	count	mean	std	min	0.25	0.5	0.75	max
Bedrooms	1048	3.77958	1.167894	1	3	4	4	17
Bathrooms	1048	2.074427	0.992904	1	1	2	3	8
Land area	1048	856.961832	1589.69807	40	323	571.5	825	22240
CV	1048	1388544	1184422	270000	780000	1080000	1600000	18000000
Latitude	1048	-36.894561	0.128426	-37.265021	-36.950873	-36.893409	-36.85628	-36.177655
Longitude	1048	174.799026	0.117991	174.317078	174.722226	174.798612	174.880943	175.492424
SA1	1048	7006332	2583.92	7001130	7004426	7006334	7008390	7011028
0-19 years	1048	47.544847	24.713408	0	33	45	57	201
20-29 years	1048	28.915076	20.993232	0	15	24	36	270
30-39 years	1048	27	17.93158	0	15	24	33	177
40-49 years	1048	24.131679	10.956798	0	18	24	30	114
50-59 years	1048	22.597328	10.212455	0	15	21	27	90
60+ years	1048	29.353053	21.810055	0	18	27	36	483
SA12018_code	1048	7006332	2583.92	7001130	7004426	7006334	7008390	7011028
NZDep2018	1048	5.06584	2.912027	1	2	5	8	10
NZDep2018_Score	1048	986.51813	94.271599	849	918	959	1031	1380
population	1048	179.799618	71.087298	3	138	174	207.75	789

We notice that for number of bedrooms, mean is 3.8 and 75% percentile is 4, while the maximum is 17, which means 17 is an outlier. The presence of outliers could mislead the model that tries to fit all observed data points in the machine learning part later.

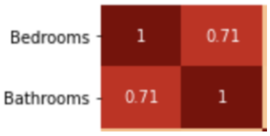
For the number of bathrooms, the maximum number 8 is an outlier as well. Same happens to land area, CV, xx-xx years, and population. This means there exists some luxury houses that are much larger and more expensive than the most of other houses. There also exists a few SA1 unit areas that are very crowded in terms of the number of people living in the area.



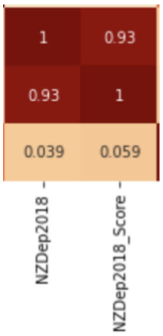
The plot above shows that the most of SA1 areas have a deprivation score between 900 and 1000, and there are fewer and fewer areas as we increase the deprivation score. This distribution is skewed to the right.



The plot above is the distribution of CV values. We can see it is skewed to the right, with most of the values below 2,500,000. This means for the most of the houses, you can purchase with a budget of $2,500,000 (assuming CV is roughly the same as the price in the market).



From the heatmap, we notice that the number of bedrooms is highly correlated to the number of bathrooms. This makes sense because when you have a larger house, you probably want more bedrooms as well as bathrooms.



NZDep2018 and NZDep2018_Score are highly correlated. This is because NZDep2018_Score is the specific score and NZDep2018 is the corresponding scale number.

We also notice that xx-xx years tend to correlate other xx-xx years. They are also correlated to Population, which makes sense since Population is roughly the sum of xx-xx years columns although not exactly the sum as the data is not complete.

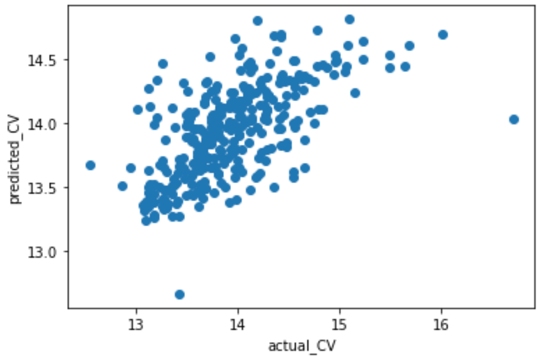
Since we want to predict the capital value of a house, we removed duplicate columns and some columns that are not helpful to the prediction in our opinion. The removed columns are SA12018_code, Latitude, Longitude, and Address.

Analysis

We use Root Mean Squared Error (RMSE) as our metric of performance. R^2 is a measurement of how well the data fits the model.

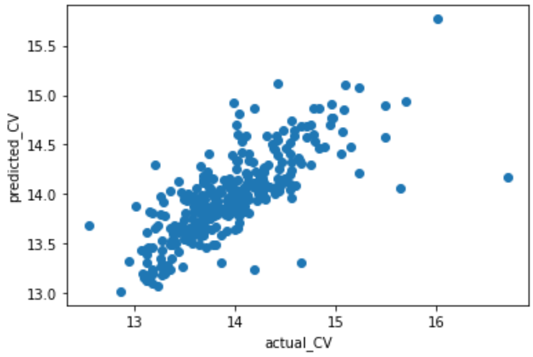
The result of the Linear Regression model:

R^2	0.394
RMSE	0.435



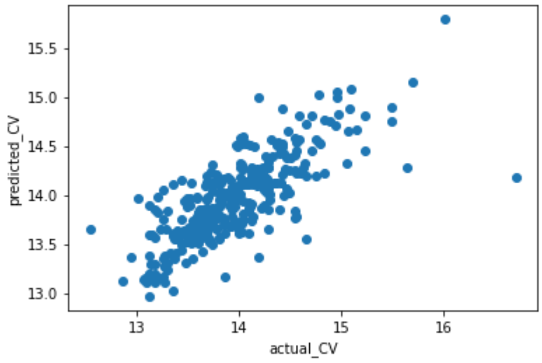
The result of the Random Forest Regression model:

R^2	0.585
RMSE	0.360



The result of the XGBoost model:

R^2	0.602
RMSE	0.353



Conclusion

This analysis has shown that the house prices can be predicted from its number of bedrooms, number of bathrooms, land area, SA1, number of people whose age is within some ranges in the SA1 unit area, deprivation scale (NZDep2018), deprivation score (NZDep2018_Score), and population.

We tried three models, and the best model is XGBoost with the lowest RMSE = 0.353.