AUCKLAND HOUSING 2018 CLASSIFICATION ANALYSIS

Sebastian Thomas, 22/07/2020

EXECUTIVE SUMMARY

This dataset is the Auckland Housing 2018 dataset, which was downloaded from the 2020 NZMSA Phase 1 GitHub repository. Data from the dataset '2018 Census Individual (part 1) total New Zealand by Statistical Area 1' from Statistics New Zealand along with data from 'NZDep2018 Statistical Area 1 (SA1) data' was also added to this dataset using APIs for this analysis.

The dataset initially consisted of 1555 observations, of which 515 were duplicates, and an additional 5 consisted of null values. After cleansing the dataset, I based my analysis on 1040 records along with 18 variables. The first variable is the ID number, followed by the number of bedrooms and bathrooms, address, land area occupied, the current value of the house, latitude, longitude, the statistical area it is situated within (SA1), the range of ages within that statistical area (from 0-60+ years), the scale of deprivation (NZDep2018), which is also the response variable, and finally, the NZ deprivation score. The response variable, NZDep2018, consists of 10 classes which range from 1 to 10, with 1 indicating least deprived areas and 10 indicating the most. The number of observations within each class is as follows:

Class	Observations
1	135
2	130
3	110
4	121
5	93
6	94
7	90
8	86
9	97
10	84

After exploring the data by calculating the summary statistics and by visualising the correlation between each of the variables, several positive relationships were found. Four machine learning algorithms were also implemented after splitting the dataset into training and testing datasets. The best model was found based on the combination of accuracy, precision, recall and F1 scores.

INITIAL DATA ANALYSIS

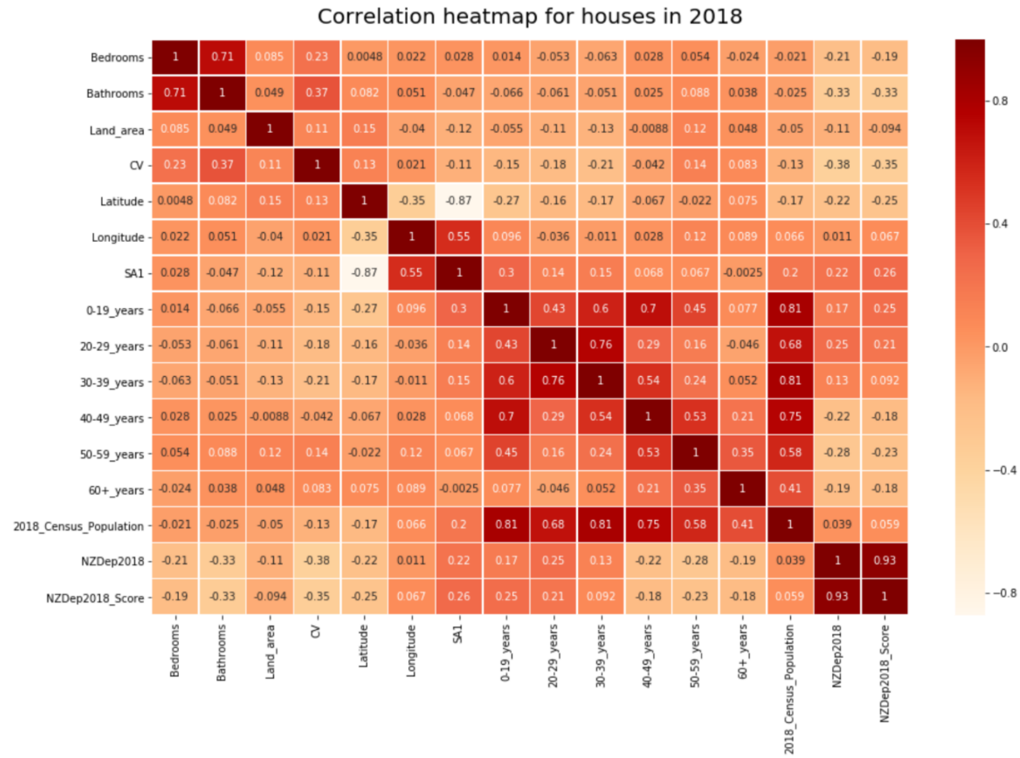
The initial data exploration began with summary and descriptive statistics.

Individual feature statistics which include the count, mean, standard deviation, min, lower quartile, median, upper quartile and max of the 1040 records and 18 variables are shown as follows:

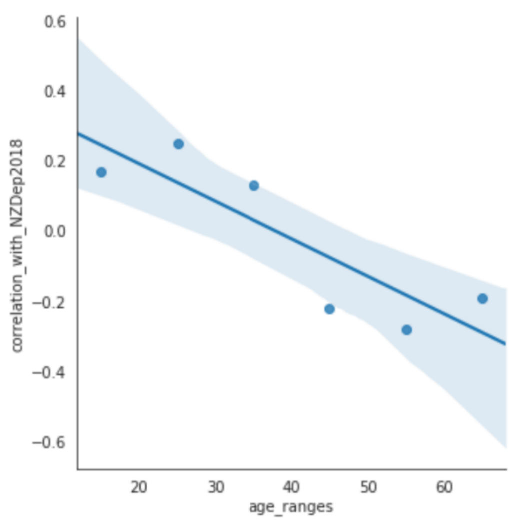
	count	mean	std	min	0.25	0.5	0.75	max
Bedrooms	1040	3.782692	1.171069	1	3	4	4	17
Bathrooms	1040	2.074038	0.994353	1	1	2	3	8
Land_area	1040	850.772115	1581.070983	40	323	570.5	825	22240
CV	1040	1381557	1163974	270000	780000	1080000	1600000	18000000
Latitude	1040	-36.89422	0.128469	-37.265021	-36.950487	-36.893455	-36.856094	-36.177655
Longitude	1040	174.79872	0.118222	174.317078	174.721131	174.797892	174.880943	175.492424
SA1	1040	7006326	2583.803	7001130	7004422	7006334	7008383	7011028
0-19_years	1040	47.538462	24.760576	0	33	45	57	201
20-29_years	1040	28.952885	21.038594	0	15	24	36	270
30-39_years	1040	26.982692	17.955181	0	15	24	33	177
40-49_years	1040	24.124038	10.978893	0	18	24	30	114
50-59_years	1040	22.580769	10.22477	0	15	21	27	90
60+_years	1040	29.313462	21.878873	0	18	27	36	483
2018_Census_Population	1040	179.780769	71.227962	3	138	174	207	789
NZDep2018	1040	5.066346	2.904714	1	2	5	8	10
NZDep2018_Score	1040	986.227885	93.536676	849	918	959	1030.25	1380

From this, it was observed that there are houses within Auckland that have 17 bedrooms and 8 bathrooms along with a value of 18 million, a staggering number. Something else that strikes out from this is the significant variation in Capital Value and the high value of the median house price within Auckland, both going over 1 million during 2018. This is evidence for the emerging Auckland housing crisis. In addition to this, the summary statistics also show that the maximum number of people in Auckland during 2018 who are 60+ years old are far higher than the number of people who are below 60, illustrating that Auckland must have had an ageing population during that year.

ANALYSIS OF CORRELATIONS AND PATTERNS IN THE DATA



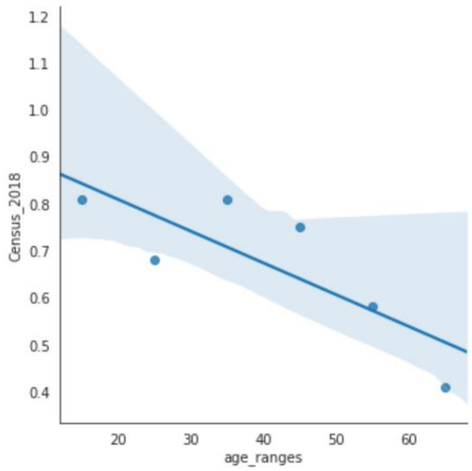
From the correlation heatmap, we can see that there is a slight negative correlation between the deprivation index and the age ranges.



So, this means that the older the age of the house, the more likely it is to be in a less deprived area. This does make sense as a large portion of the population are building homes in more deprived areas as the land there is of cheaper values. The older houses are valued at a slightly higher price than newer ones since they are probably regarded as antiques, making them more valuable than the newer ones. This theory is supported by other evidence in the heatmap, where there is a slight positive correlation between age and current value. The higher the age of the house, the greater its capital value.

There is also an extremely strong positive correlation between NZDep2018_Score and NZDep2018, which means, the higher the score, the more deprived those areas are.

A strong negative relationship can also be seen between the age ranges and the 2018_Census_population. The lower the age range within an area, the higher its population.



This is contrary to the argument made earlier regarding the summary statistics, where the maximum number of people in Auckland during 2018 who are 60+ years old are far higher than the number of people who are below 60. That number, 483, must have been an outlier and does not contribute much to the overall correlation between the two variables, 2018_Census_Poulations and 60+_years. Another outlier detected in the relationship between all the age ranges, and the census population is the correlation between the 20-29_years and 2018_Census_Population (0.68), which does not follow the negative trend stated above.

As from the heatmap, there are also strong positive relationships between Bathrooms and Bedrooms, and SA1 and Longitude, respectively. Moreover, there is also a strong negative relationship between SA1 and Latitude

MACHINE LEARNING MODEL ANALYSIS

In this analysis, four classification algorithms have been tested, which are Logistic Regression, Support Vector Machines, Decision Trees Classifier, and Random Forests Classifier.

All algorithms were trained with 70% of the data, with the remaining 30% being used for testing. Below are the results:

Model	Accuracy	Precision (weighted)	Recall (weighted)	F1-Score (weighted)
Decision Trees Classifier	97.8%	98%	98%	98%
Random Forests Classifier	90.1%	91%	90%	90%
Logistic Regression	21.8%	12%	22%	15%
Support Vector Machines	14.4%	35%	14%	6%

As shown from the table above, it is evident that the Decision Trees Classifier is the best model for this analysis since it has the highest combination of Accuracy, Precision, Recall, and F1 Scores.

The confusion matrix for this model is as follows:

```
[[41  0  0  0  0  0  0  0  0  0]
 [ 0 39  0  0  0  0  0  0  0  0]
 [ 0  0 27  0  0  0  0  0  0  0]
 [ 0  0  0 34  0  0  0  0  0  0]
 [ 0  0  0  3 25  0  0  0  0  0]
 [ 0  0  0  0  0 27  0  0  0  0]
 [ 0  0  0  0  0  2 30  0  0  0]
 [ 0  0  0  0  0  0  1 21  0  0]
 [ 0  0  0  0  0  0  0  0 32  0]
 [ 0  0  0  0  0  0  0  0  0 30]]
```

From the matrix, we can see that:

- The model can accurately predict positively for the classes 1,2,3,4,6,9,10
- The number of false negatives in classes 5, 7 and 8 are as follows:

Class	FN
5	3
7	2
8	1

- The total number of false positives is 0 for all the classes
- The total number of true negatives is as follows

Class	TN
1	0
2	0
3	0
4	0
5	3
6	0
7	2
8	1
9	0
10	0

Looking further into this analysis and investigating the cause of the large gap in the accuracy of each of the models, I found that due to the strong correlation between the NZDep2018 and NZDep2018_Score field, as evident in the heatmap earlier, the accuracy of the models are being skewed.

So, a decision was made to analyse the data without the NZDep2018_Score field, and the results were as follows:

Model	Accuracy	Precision (weighted)	Recall (weighted)	F1-Score (weighted)
Random Forests Classifier	34.0%	33%	34%	34%
Decision Trees Classifier	29.8%	31%	30%	30%
Logistic Regression	21.8%	37%	36%	36%
Support Vector Machines	14.4%	35%	14%	6%

As can be seen from above, the best model would now be the Random Forests Classifier, however, with significantly lower scores as compared to before. In this analysis, the gaps between the accuracies of the models do seem must closer to each other. Thus, I believe that this would be a more accurate representation of the machine learning model analysis. Further details about the Random Forests Classifier as follows:

Confusion Matrix:

```
[[22  8  4  1  0  2  4  0  0  0]
 [ 9 19  4  2  4  1  0  0  0  0]
 [ 6  7  6  3  1  2  1  1  0  0]
 [ 4  4  5 10  3  4  2  0  1  1]
 [ 1  7  4  5  3  1  2  1  4  0]
 [ 0  0  2  6  1  5  7  2  2  2]
 [ 0  0  2  1  4  5 11  3  6  0]
 [ 0  0  1  1  3  1  2  5  6  3]
 [ 0  1  1  0  2  0  2  6 12  8]
 [ 1  0  2  0  0  3  2  2  7 13]]
```

From the matrix, we can see that:

Class	True Positive	False Positive	False Negative	True Negative
1	22	21	19	40
2	19	27	20	47
3	6	25	21	46
4	10	19	24	43
5	3	18	25	43
6	5	19	22	41
7	11	22	21	43
8	5	15	17	32
9	12	26	20	46
10	13	14	17	31

CONCLUSION

This analysis has shown that the deprivation index, which is the response variable, cannot be confidently predicted from each of the other given variables, as evident by the performance of the machine learning algorithms. The accuracy of the highest performing model Random Forests Classifier, in this case, is 33%. The addition of the NZDep2018_Score field does improve the scores; however, skews the results significantly, thus making it inappropriate to use in this analysis.