Analysis of Social Deprivation Index

Michael Chung Ng

July 2020

Contents

Premable

The Social Deprivation Index is a measure of socio-economic status calculated for small geographic areas. The calculation uses a range of variables from the 2018 Census which represents nine dimensions of socio-economic disadvantage to create a summary deprivation score. The nine variables (proportions in small areas) in decreasing weight in the index are:

1. Income – People aged 18-59 receiving a means tested benefit
2. Employment – People aged 18-59 years who are unemployed
3. Income – People living in equivalised households with income below an income threshold
4. Communication – People with no access to a telephone
5. Transport – People with no access to a car
6. Support – People aged less than 60 years living in a single parent family
7. Qualifications – People aged 18-59 years without any qualifications
8. Living space – People living in equivalised households below a bedroom occupancy threshold
9. Owned Home – People not living in own house

The social Deprivation Index is used in the measurement and interpretation of socio-economic status of communities for a wide variety of contexts such as needs assessment, resource allocation, research and advocacy.

Note that the deprivation index, applies to areas rather than individals who live in those areas.

For the purpose of comparison, the Social Deprivation Index is presented as a scale, ranking small areas from the least deprived to the most deprived. `NZDep2018` correpond to the New Zealand deprivation index, with 10 as the most deprived and 1 as the least deprived.

Chapter 1

Executive Summary

We are interested in training a predictive model which can calculate the relative social deprivation index; we achieved this by taking factors from the 2018 NZ census which the social deprivation index is derived.

The dataset is `properties.csv` which contains an address, latitude, and longitude for each observation. From this, we can identify the NZ deprivation index from `nzdep2018.csv` by Statistical Area 1 (SA1). We performed API calls on the 2018 Census Individual total New Zealand by Statistical Area 1 to retrieve ethnicity populations, median income, active smokers, % of the population with a means-tested benefit, and % of the population that own or part-own their house. The final dataset after data cleaning gives us 1036 observations over 18 variables.

Our response variable in this model is the NZDep2018 which is the social deprivation index for an area measured on an ordinal scale of 1-10. The remaining factors are explanatory variables. Each observation row contains measurements on the ethnicity populations, median income, smokers, % of means-tested benefits, and % that own or part-own their house.

We explored the dataset by calculating the summary and descriptive statistics. We also created histogram visualisations for each variable to determine the normality of the dataset and identify any skewness, and we created a correlation matrix to find the highly correlated variables in the dataset.

After data exploration, we trained six predictive models on 10-fold cross-validation to identify the algorithm with the best predictive accuracy. We identified and selected the decision tree classifier as it had the best predictive accuracy across all six algorithms by a significant margin.

Chapter 2

Methods

2.1 Initial data exploration

All data collected was relative to the Statistical Area 1 (SA1) which is defined by Stats NZ.

In our initial data exploration, we calculated the summary/descriptive statistics for the dataset. The general descriptive statistics gives us: mean, standard deviation, minimum, lower quartile, median, upper quartile, and maximum, which can help with understanding the shape, spread, and patterns in the dataset.

We found 12 observations that had a value of -999; the reason Stats NZ has listed values of -999 is to suppress data in compliance with 2018 confidentiality rules. We dropped any observations with -999 to prevent it from impacting the results of our model. Any results with NaN, wherein no results were collected for that variable were dropped from the dataset. The reason why we opted to drop these observations instead of substituting the mean or median value of the whole variable is that we have a sufficient number of observations whereby dropping them would not adversely affect the dataset.

2.2 Analysis of Correlations and Patterns

Looking at the histogram plot in fig 2.2, we can see the general shape of each variable in the dataset. There does not appear to be any significant discrepancies or outliers within the dataset. There is a right skew on a few of the variables which we can resolve by scaling them. Overall, the dataset appears to be suitable and fit for use in modelling.

The correlation matrix in fig 2.1 gives us the correlation coefficient between sets of variables; from this, we were able to identify a few areas of interest on the correlation matrix. There is a strong correlation between the percentage of the population receiving a means-tested benefit & population of active smokers and the social deprivation index, therefore indicating areas with a higher percentage of the population receiving a means-tested benefit and active smokers would have a higher deprivation index score. We also found that the median total personal income and homeownership are both inversely correlated to the deprivation index, which indicates that areas with higher total personal income and homeownership have a lower deprivation index. A point of interest is that Maori and Pacific populations were disproportionately correlated with a higher deprivation score compared to other ethnic populations which may indicate that these populations are subject to greater relative socio-economic deprivation.

List of Figures

	CV	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	European Population	Maori Population
count	1.036000e+03	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000
mean	1.392359e+06	48.095560	29.250000	27.312741	24.411197	22.859073	29.681467	98.128378	21.043436
std	1.187959e+06	24.317099	20.881082	17.796615	10.705763	9.975691	21.718820	51.249225	21.704895
min	2.800000e+05	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	9.000000	0.000000
25%	7.800000e+05	33.000000	18.000000	15.000000	18.000000	15.000000	18.000000	60.000000	9.000000
50%	1.080000e+06	45.000000	25.500000	24.000000	24.000000	21.000000	27.000000	96.000000	15.000000
75%	1.605000e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	129.000000	27.000000
max	1.800000e+07	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	471.000000	165.000000

Pacific Population	Asian Population	MELAA Population	Median Income	Smokers	Benefit Percent	Home Ownership	NZDep2018_Score
1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000	1036.000000
25.647683	53.325290	4.062741	38519.305019	15.576255	0.113166	0.397008	986.243243
36.014790	53.648988	5.070796	11449.509995	11.427950	0.084110	0.149121	94.681131
0.000000	0.000000	0.000000	13100.000000	0.000000	0.000000	0.000000	849.000000
3.000000	21.000000	0.000000	30175.000000	9.000000	0.050000	0.287500	918.000000
12.000000	39.000000	3.000000	37400.000000	12.000000	0.090000	0.400000	958.000000
33.000000	72.000000	6.000000	45625.000000	21.000000	0.150000	0.520000	1032.000000
276.000000	591.000000	51.000000	82800.000000	78.000000	0.560000	0.760000	1380.000000

Figure 2.1: General descriptive statistics

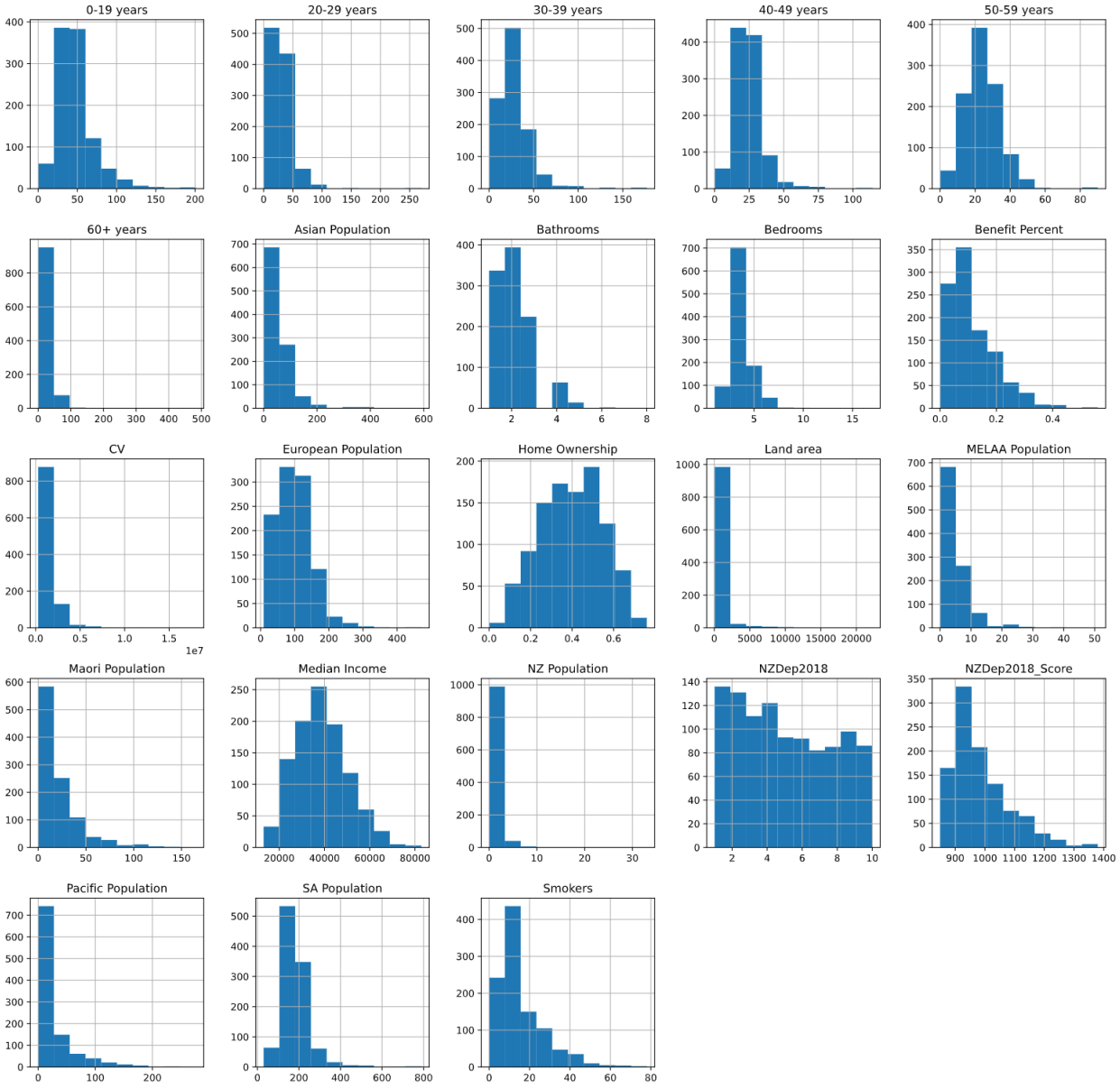

Figure 2.2: Histogram of each variable in the dataset

Chapter 3

Results

3.1 Model Selection

We want to create a predictive model which can the relative social deprivation index of a given area. The social deprivation index is an ordinal variable which means we can use a classification algorithm to identify which category (deprivation index) an observation belongs. We used 10-fold cross-validation on six classification algorithms to determine the best one to perform predictions.

Initially, prediction accuracy was poor across all six trained models. We suspected that some variables within the dataset were negatively affecting results, so we removed: `Bedrooms, Bathrooms, Land area, SA Population, NZ Population` . After removing these variables predictive accuracy was improved; the reason for this may be that these variables were confusing the model training.

Algorithm	Accuracy	Precision	F-Score
K-Nearest Neighbour	0.22010	0.20349	0.19426
Logistic Regression	0.40150	0.43350	0.42395
Random Forest Classifier	0.94309	0.90500	0.90336
Guassian Naive Bayes	0.45934	0.51085	0.47436
ADA Boost	0.35708	0.27766	0.30678
Decision Tree Classifier	0.97875	0.87078	0.98279

We can see that the decision tree classifier produces the best performance in terms of accuracy of 98.29% compared to all six algorithms in the comparison. We then plotted a confusion matrix to which is used to describe the performance of a classification model on a set of test data for which the true values are known. It allows us to visualise the performance of an algorithm. In fig 3.2, the confusion matrix shows us that all outputs are sitting on or close to the diagonal which indicates gives us strong evidence that we have a high performing model with high prediction accuracy.

List of Figures

```
[540]  M↓
       # Nearest neighbour
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(regressor, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.94 (+/- 0.08)

[549]  M↓
       # Logistic regression
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(logmodel, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.40 (+/- 0.07)

[572]  M↓
       # Random forest
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(RFClassifier, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.93 (+/- 0.08)

[553]  M↓
       # Gaussian naive bayes
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(gnb, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.46 (+/- 0.11)

[554]  M↓
       # ADA boost
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(abc, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.36 (+/- 0.07)

[573]  M↓
       # Decision tree classifier
       from sklearn.model_selection import cross_val_score

       scores = cross_val_score(dtc, x,y, cv=10)
       print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.98 (+/- 0.03)
```

Figure 3.1: Model Accuracies

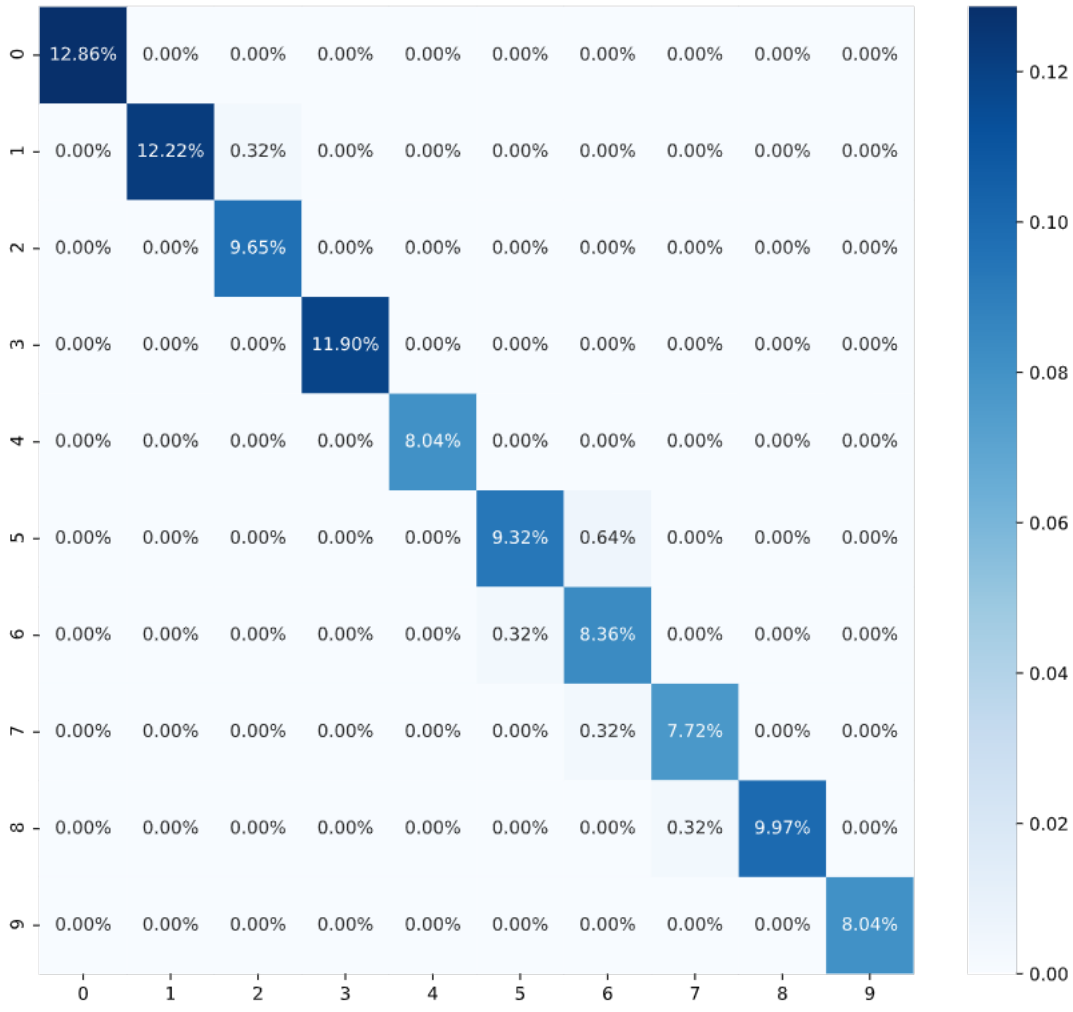

Figure 3.2: Confusion matrix for decison tree classifier

Chapter 4

Conclusion

The aim of this report was to produce a high predictive accuracy model which can predict social deprivation index. We used a base data of addresses to contruct a full dataset through API calls to collect data pertaining to ethnic populations, median personal income, active smokers, % of population on a means-tested benefit, and rate of homeownership. All API calls were made to the geospatial data management platform, Koordinates. Data was taken from the 2018 Census Individual (part1/part2) total New Zealand by Statistical Area 1.

Six predictive models were trained to predict the social deprivation index by factors from which the social deprivation index is derived from. We were able to create a predictive model using a decision tree classifier to get a predictive accuracy rate of 98.29% which is really good. It is also worth noting that only the two tree-based algorithms had a high accuracy rate; the four other models had a poor predictive accuracy $< 50\%$.