Predicting solar installation rates in different locations using demographic data

Anthony Liu

March 1, 2020

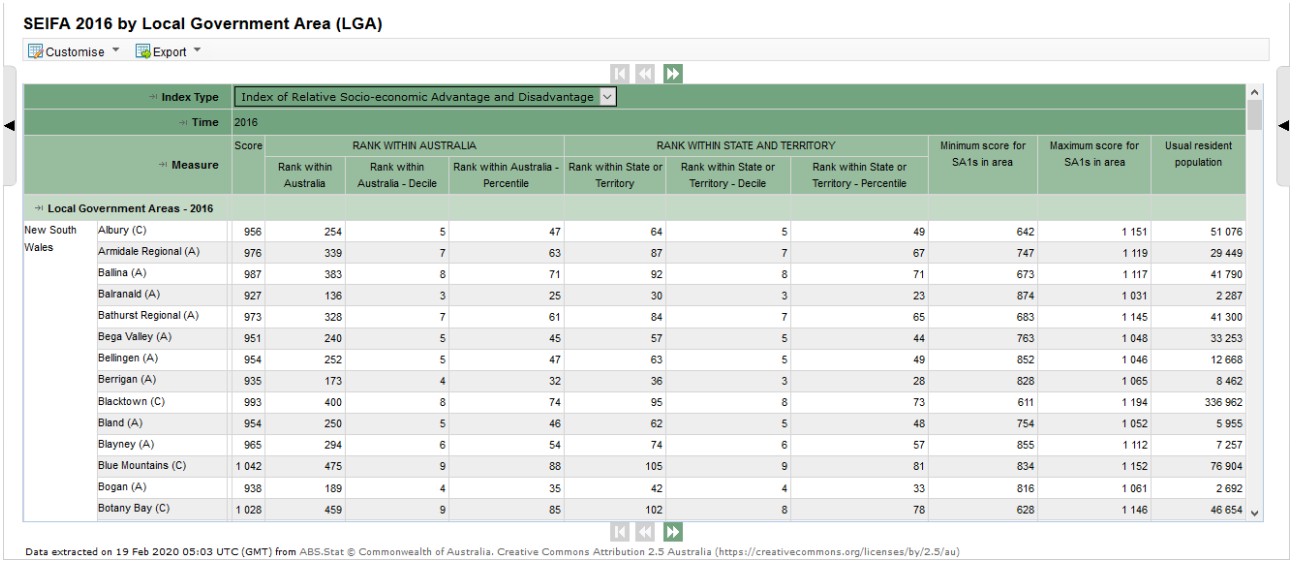
Contents

1 Introduction

The central problem that the analysis in this report will attempt to address is "Can we predict solar installation rates in the different local government areas of Australia using demographic data?". As the effects of global climate change are becoming better understood, the move towards renewable energy sources has become an important focus [1]. Solar energy is emerging as one of the most popular forms of renewable energy for reasons such as decreasing costs, environmental ethics, health, government incentives and accessibility [2, 3].

Australia is in a particularly privileged position to have some of the best solar energy resources in the world [4]. Being able to answer the proposed question may lead to insights which guide policy, investment or further research to better utilise this natural resource. Given that installation of solar constitutes a purchase and purchases are made by individuals under their own unique circumstances, there is good reason to believe that demographic factors (such as financial status, dwelling structure, education, etc.) may affect installation rates. In this analysis we limit our scope to prediction of solar installation rates in each local government area (LGA) based on available demographic data. Questions of inference such as which specific demographic factors are related to installation rates are not considered.

Parties who may be interested in this problem include policy makers, investors and researchers. Reliable predictions on installation rates may guide projections and highlight both the level and content of policy intervention justified in affecting solar uptake. Also, investors in the solar industry hold an advantage if they have reliable guidance on which local government areas in the future (accounting for demographic shifts) are likely to install solar. Finally, results from attempting to answer this question may lead to new lines of inquiry among researchers (e.g. if we can indeed predict installation rates then that provides justification to then do further research on which exact demographic factors have an effect).



SEIFA 2016 by Local Government Area (LGA)

Customise Export

Index Type: Index of Relative Socio-economic Advantage and Disadvantage

Time: 2016

Measure		Score	RANK WITHIN AUSTRALIA			RANK WITHIN STATE AND TERRITORY			Minimum score for SA1s in area	Maximum score for SA1s in area	Usual resident population
			Rank within Australia	Rank within Australia - Decile	Rank within Australia - Percentile	Rank within State or Territory	Rank within State or Territory - Decile	Rank within State or Territory - Percentile			
Local Government Areas - 2016											
New South Wales	Albury (C)	956	254	5	47	64	5	49	642	1 151	51 076
	Armidale Regional (A)	976	339	7	63	87	7	67	747	1 119	29 449
	Ballina (A)	987	383	8	71	92	8	71	673	1 117	41 790
	Balranald (A)	927	136	3	25	30	3	23	874	1 031	2 287
	Bathurst Regional (A)	973	328	7	61	84	7	65	683	1 145	41 300
	Bega Valley (A)	951	240	5	45	57	5	44	763	1 048	33 253
	Bellingen (A)	954	252	5	47	63	5	49	852	1 046	12 668
	Berrigan (A)	935	173	4	32	36	3	28	828	1 065	8 462
	Blacktown (C)	993	400	8	74	95	8	73	611	1 194	336 962
	Bland (A)	954	250	5	46	62	5	48	754	1 052	5 955
	Blayney (A)	965	294	6	54	74	6	57	855	1 112	7 257
	Blue Mountains (C)	1 042	475	9	88	105	9	81	834	1 152	76 904
	Bogan (A)	938	189	4	35	42	4	33	816	1 061	2 692
	Botany Bay (C)	1 028	459	9	85	102	8	78	628	1 146	46 654

Data extracted on 19 Feb 2020 05:03 UTC (GMT) from ABS.Stat © Commonwealth of Australia, Creative Commons Attribution 2.5 Australia (https://creativecommons.org/licenses/by/2.5/au)

Figure 1: A preview of the socio-economic indicators dataset on the ABS website, prior to exporting as a CSV file and performing data wrangling.

2 Data

To answer our problem we use two separate types of data: data regarding solar installation rates in each LGA and data regarding demographics in each LGA. We obtained our first dataset on solar installations from the Australian Photovoltaic Institute [5]. The Australian Photovoltatic Institute (APVI) is a not-for-profit company whose stated objective is to "Support the increased development and use of PV via research, analysis and information" [6]. The photovoltaic (PV) data by the APVI was compiled from Australian government body sources: PV installation data came from the Clean Energy Regulator [5, 7] and attached LGA data was from the Australian Bureau of Statistics (ABS) [5, 8]. Of particular interest to this analysis is the density variable, an estimate of the percentage of dwellings in each LGA which have installed solar systems.

Detailed demographic data on each LGA including socio-economic indicators, dwelling structure, age, sex, cultural background, education and employment was also obtained from the ABS [9]. Figures 1 and 2 provide a preview of how the socio-economic indicators and education/employment data look on the ABS website respectively. These entries are later wrangled into an appropriate format and act as individual features/predictors of a machine learning algorithm to predict density (i.e. solar installation rate in each LGA). We restricted our use of demographic information to only that recorded by the ABS in the year 2016 as this year contained the most complete data (the last census took place in 2016).

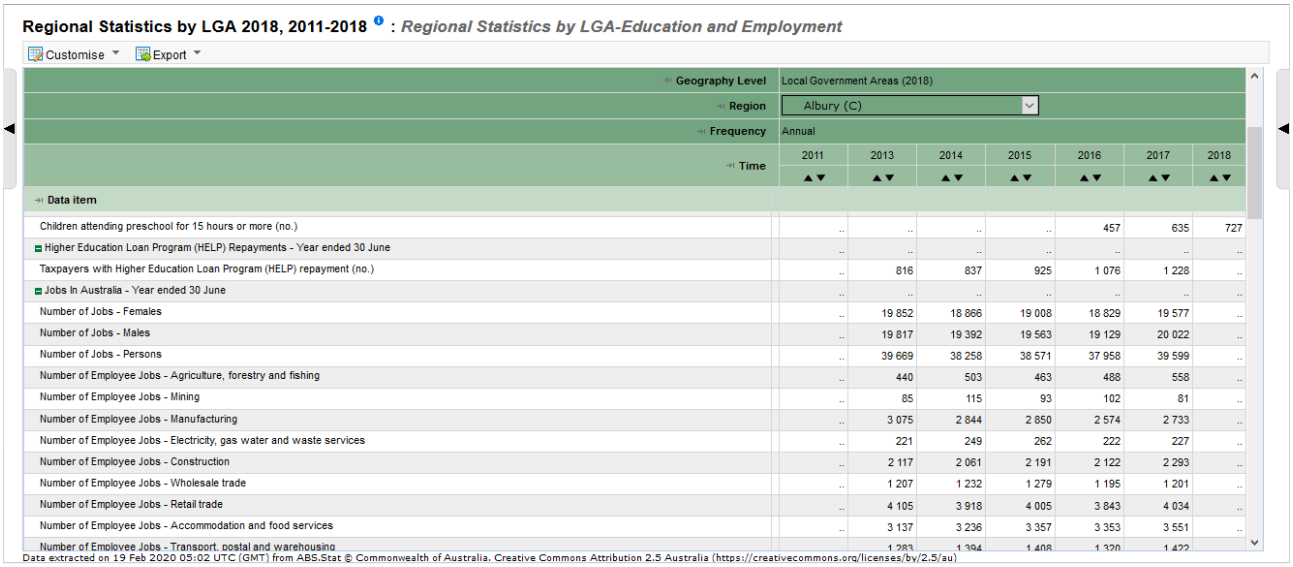
3 Methodology

3.1 Data wrangling

The data obtained from the APVI and ABS were loaded in their raw form using Python 3.7.4. The APVI data contained information including number of installations, total energy capacity, and breakdowns of installations by energy capacity (see Figure 3). As our problem was to predict solar installation rates in each LGA, only information pertaining to the density of installations was used in this analysis (see Figure 4).

The demographic data from the ABS in its raw form also contained excessive information but we defer the removal of excess information to our principal components analysis later on. We choose to temporarily include all demographic indicators and simply perform a transformation on the data so that it is in a format appropriate for merging with our solar installation data. Figures 5 and 6 demonstrate the transformation performed on the education and employment dataset from the ABS.

Similar transformations were performed on the SEIFA socio-economic indices, population and dwelling structure datasets from the ABS. These demographic datasets were then combined with the solar installation dataset to form a combined dataframe where the LGA was the index, the density (our response) was the first



Regional Statistics by LGA 2018, 2011-2018 : *Regional Statistics by LGA-Education and Employment*

Customise Export

Geography Level	Local Government Areas (2018)						
Region	Albury (C)						
Frequency	Annual						
Time	2011	2013	2014	2015	2016	2017	2018
Data item							
Children attending preschool for 15 hours or more (no.)	..	..	..	..	457	635	727
Higher Education Loan Program (HELP) Repayments - Year ended 30 June	..	..	..	..	..	..	..
Taxpayers with Higher Education Loan Program (HELP) repayment (no.)	..	816	837	925	1 076	1 228	..
Jobs In Australia - Year ended 30 June	..	..	..	..	..	..	..
Number of Jobs - Females	..	19 852	18 866	19 008	18 829	19 577	..
Number of Jobs - Males	..	19 817	19 392	19 563	19 129	20 022	..
Number of Jobs - Persons	..	39 669	38 258	38 571	37 958	39 599	..
Number of Employee Jobs - Agriculture, forestry and fishing	..	440	503	463	488	558	..
Number of Employee Jobs - Mining	..	85	115	93	102	81	..
Number of Employee Jobs - Manufacturing	..	3 075	2 844	2 850	2 574	2 733	..
Number of Employee Jobs - Electricity, gas water and waste services	..	221	249	262	222	227	..
Number of Employee Jobs - Construction	..	2 117	2 061	2 191	2 122	2 293	..
Number of Employee Jobs - Wholesale trade	..	1 207	1 232	1 279	1 195	1 201	..
Number of Employee Jobs - Retail trade	..	4 105	3 918	4 005	3 843	4 034	..
Number of Employee Jobs - Accommodation and food services	..	3 137	3 236	3 357	3 353	3 551	..
Number of Employee Jobs - Transport, postal and warehousing	..	1 283	1 394	1 408	1 320	1 422	..

Figure 2: A preview of the education and employment dataset on the ABS website, prior to exporting as a CSV file and performing data wrangling.

	LGA	lga_name	lga_state	capacity_lga	dwellings_lga	instals_lga	density_lga	capunder10	cap10_100	capover100	countunder10	count10_100	count100
0	10050	Albury (C)	NSW	27597.0	25547.0	5115.0	18.8	16674.0	7827.0	3096.0	4796.0	319.0	4.0
1	10130	Armidale Regional (A)	NSW	13380.0	12406.0	2690.0	20.3	9466.0	3914.0	0.0	2518.0	172.0	0.0
2	10250	Ballina (A)	NSW	29350.0	19436.0	7314.0	36.4	23756.0	5010.0	584.0	7072.0	242.0	3.0
3	10300	Balranald (A)	NSW	154436.0	1174.0	211.0	15.8	853.0	535.0	153048.0	185.0	26.0	2.0
4	10470	Bathurst Regional (A)	NSW	17636.0	17524.0	3383.0	17.9	12549.0	5087.0	0.0	3144.0	239.0	0.0

Figure 3: Solar installation data by the APVI in its raw form upon being loaded.

	LGA	Density (%)
0	Albury (C)	18.8
1	Armidale Regional (A)	20.3
2	Ballina (A)	36.4
3	Balranald (A)	15.8
4	Bathurst Regional (A)	17.9

Figure 4: Solar installation data by the APVI after cleaning. Only the LGA and Density (percentage of dwellings with installed solar) variables were retained.

	MEASURE	Data item	REGIONTYPE	Geography Level	LGA_2018	Region	FREQUENCY	Frequency	TIME	Time	Value	Flag Codes	Flags
0	PRESCH_2	4 year olds enrolled in preschool or in a pres...	LGA2018	Local Government Areas (2018)	10050	Albury (C)	A	Annual	2016	2016	580.0	NaN	NaN
1	PRESCH_2	4 year olds enrolled in preschool or in a pres...	LGA2018	Local Government Areas (2018)	10050	Albury (C)	A	Annual	2017	2017	613.0	NaN	NaN
2	PRESCH_2	4 year olds enrolled in preschool or in a pres...	LGA2018	Local Government Areas (2018)	10050	Albury (C)	A	Annual	2018	2018	634.0	NaN	NaN
3	PRESCH_3	5 year olds enrolled in preschool or in a pres...	LGA2018	Local Government Areas (2018)	10050	Albury (C)	A	Annual	2016	2016	224.0	NaN	NaN
4	PRESCH_3	5 year olds enrolled in preschool or in a pres...	LGA2018	Local Government Areas (2018)	10050	Albury (C)	A	Annual	2017	2017	268.0	NaN	NaN

Figure 5: The educational demographic data from the ABS in its raw form.

Data item	LGA	4 year olds enrolled in preschool or in a preschool program (no.)	5 year olds enrolled in preschool or in a preschool program (no.)	Advanced Diploma/Diploma (%)	Agriculture, Environmental and Related Studies (%)	Architecture and Building (%)	Bachelor Degree (%)	Certificate (%)	Children attending preschool for 15 hours or more (no.)	Children attending preschool for less than 15 hours (no.)	...
0	Adelaide (C)	81.0	15.0	8.1	1.0	3.5	24.4	8.1	69.0	16.0	...
1	Adelaide Hills (DC)	389.0	87.0	10.5	3.7	5.2	19.5	18.7	353.0	113.0	...
2	Adelaide Plains (DC)	104.0	22.0	6.3	3.8	5.8	4.6	25.0	91.0	31.0	...
3	Albany (C)	422.0	6.0	8.9	5.1	6.2	9.6	23.7	284.0	120.0	...
4	Albury (C)	580.0	224.0	8.5	2.7	6.3	11.3	23.8	457.0	330.0	...

5 rows × 87 columns

Figure 6: The education and employment demographic data from the ABS after transformation. There are 87 columns in total, corresponding to 86 predictors (LGA is not used as a predictor). Similar transformations are performed on the other demographic datasets from the ABS.

column, and all the predictors made up the remaining columns (see Figure 7). The dataframe is now in a "tidy" format where each row is an observation and each column is a variable, a standard which allows for easier data manipulation and analysis [10].

3.2 Exploratory data analysis

Within our exploratory data analysis, we computed summary statistics of the combined dataframe to look for any anomalies (see Figure 8). We also explored the effect on dataset size when dropping missing values. The dataset size at this point was 481 rows by 194 columns, so the number of observations (rows) we had were quite small considering the number of predictors we had. We found that dropping rows or columns with missing values resulted in too big of a decrease in dataset size. We also looked at the number of missing values each column had and found a majority only had a few missing values. There did exist other columns, however, which had so many missing values that imputing values into these would have significantly changed predictor data. We elected here to impute the median values for those 125 columns which had less than only 3 missing values, and drop the remaining columns.

3.3 Preprocessing: train-test split and principal components analysis

A train-test split of 70-30 was performed on the combined dataframe. We temporarily set aside our test set without referring to it in order to avoid introducing bias later when using the test set for model evaluation. Given that our dataset at this point still had a significant number of predictors relative to observations, we performed a dimension reduction in such a way so as to reduce the number of predictors while retaining most of their information. There is good reason to suspect that there are many redundant predictors in this dataframe given the large number of predictors and that predictors like the four different SEIFA socioeconomic indices roughly represent the same thing. Indeed, computation of the correlation matrix for the training set predictors as shown in Figure 9 confirms this.

Thus a principal components analysis was performed to isolate a subset of variables which captured most of the variability within the data. Prior to this the data was also scaled so that each column had mean 0 and standard deviation 1. We found that the first 20 principal components captured about 99.999% of variability in

LGA	Density (%)	IEO	IER	IRSAD	IRSD	4 year olds enrolled in preschool or in a preschool program (no.)	5 year olds enrolled in preschool or in a preschool program (no.)	Advanced Diploma/Diploma (%)	Agriculture, Environmental and Related Studies (%)	Architecture and Building (%)	...	Flat or apartment: In a one or two storey block	Flat or apartment: In a three storey block	House or flat attached to a shop, office, etc.	Improved home, tent, sleepers out	Not appliacable	Not stated	Semi-detached, row or terrace house, townhouse etc. with : One storey	Semi-detached, row or terrace house, townhouse etc. with : Two or more storeys	Separate house	Total
Albury (C)	18.8	961.0	960.0	956.0	971.0	580.0	224.0	8.5	2.7	6.3	...	563	112	55	7	91	82	3702	586	18127	23464
Armidale Regional (A)	20.3	1015.0	958.0	976.0	980.0	290.0	70.0	7.2	6.6	4.4	...	647	60	39	13	72	99	1036	189	10483	12738
Ballina (A)	36.4	999.0	998.0	987.0	1003.0	480.0	180.0	9.5	3.2	6.7	...	727	198	38	20	48	144	2532	1847	12760	19154
Balranald (A)	15.8	915.0	969.0	927.0	942.0	18.0	10.0	4.4	9.3	3.3	...	55	0	7	5	20	4	28	3	1044	1221
Bathurst Regional (A)	17.9	978.0	993.0	973.0	986.0	382.0	154.0	8.1	3.0	5.7	...	799	17	31	10	63	171	1495	434	14345	17431

5 rows × 194 columns

Figure 7: The combined dataframe containing Density, the response we are trying to predict in this analysis, in the first column, and demographic predictors in the remaining columns.

	Density (%)	IEO	IER	IRSAD	IRSD	4 year olds enrolled in preschool or in a preschool program (no.)	5 year olds enrolled in preschool or in a preschool program (no.)	Advanced Diploma/Diploma (%)	Agriculture, Environmental and Related Studies (%)	Architecture and Building (%)	...
count	481.000000	481.000000	481.000000	480.000000	480.000000	477.000000	414.000000	481.000000	480.000000	478.000000	...
mean	22.973181	972.503119	969.611227	962.362500	963.800000	519.758910	133.630435	7.401663	4.913542	5.242678	...
std	10.207851	74.714097	98.282038	85.529212	99.413466	1017.663145	217.730616	1.902498	3.787377	1.817519	...
min	0.000000	741.000000	484.000000	566.000000	404.000000	1.000000	1.000000	1.700000	0.500000	1.200000	...
25%	17.100000	929.000000	960.000000	932.000000	942.750000	35.000000	9.000000	6.200000	1.700000	4.200000	...
50%	22.700000	960.000000	986.000000	962.000000	980.000000	138.000000	41.000000	7.300000	3.800000	5.100000	...
75%	30.100000	996.000000	1014.000000	998.000000	1014.000000	516.000000	176.250000	8.900000	7.000000	6.200000	...
max	48.400000	1195.000000	1136.000000	1166.000000	1123.000000	13155.000000	1365.000000	13.200000	19.200000	20.200000	...

8 rows × 194 columns

Figure 8: Summary statistics of the combined dataframe (prior to imputing and removing missing values).

	0	1	2	3	4	5	6	7	8	9	...
0	1.000000	0.540561	0.884046	0.745616	0.162119	0.641033	-0.176344	-0.214650	0.898070	-0.554260	...
1	0.540561	1.000000	0.833816	0.932903	0.162930	0.618553	0.049693	-0.030803	0.359541	0.136643	...
2	0.884046	0.833816	1.000000	0.957171	0.220626	0.716136	-0.160641	-0.126520	0.776435	-0.251396	...
3	0.745616	0.932903	0.957171	1.000000	0.183891	0.692517	-0.058270	-0.068887	0.589250	-0.028256	...
4	0.162119	0.162930	0.220626	0.183891	1.000000	0.384494	-0.383059	0.135638	0.273399	-0.068304	...
...	...	...	...	...	...	...	...	...	...	...	...
119	0.182143	0.104779	0.199206	0.165639	0.852376	0.311083	-0.341198	0.147449	0.274690	-0.067282	...
120	0.309427	0.151289	0.311300	0.246646	0.704649	0.462840	-0.474388	0.094689	0.452213	-0.206812	...
121	0.310522	0.098097	0.278945	0.190631	0.877427	0.306464	-0.330373	-0.007656	0.419093	-0.291909	...
122	0.121564	0.184026	0.195590	0.186431	0.972214	0.409165	-0.380698	0.217367	0.216804	0.018499	...
123	0.228530	0.160036	0.259501	0.213448	0.975287	0.407616	-0.400762	0.142743	0.338172	-0.113698	...

124 rows × 124 columns

Figure 9: Correlation matrix for predictors in training set. Notice that there are many predictors with high correlation.

the training data. We elected to transform the training predictors dataframe so that the columns represented values of each principal component, and retain only these first 20 principal components to use as predictors.

3.4 Model building and selection

Having isolated the first 20 principal components to use as predictors, this was used along with the corresponding response variable (Density) to fit multiple machine learning models. Using the scikit-learn 0.22 Python library, we built 4 models in total: a Multiple Linear Regression model, a Boosting Regression model (AdaBoost), a Random Forest model, and a second Random Forest model where hyperparameter tuning was used. We then evaluated each model's performance using its R squared value when predicting on the test set which we set aside earlier. Identical transformations to that applied on the training set (using parameters of the training set) were applied to the testing set before evaluation. We then selected a final model using the testing set R squared values.

4 Results

A table comparing the performance of each model is presented in Figure 10. Recall that the R squared is a value between 0 and 1 and captures the proportion of variability in the response that can be explained by the set of predictors, and is roughly indicative of how confident we can be in our predicted responses. As expected, the testing set accuracy (an estimate of out-of-sample accuracy) is lower than the training set accuracy since the model is prone to bias to the dataset it was trained on. Furthermore, we find that the testing set accuracy is *significantly* lower than the training set accuracy in the Boosting and Random Forest models. This suggests high levels of bias and overfitting towards the training data, which can be attributed to the relatively small dataset size given that the Boosting and Random Forest algorithms require large amounts of data to be reliable.

We find that our Random Forest Regression models significantly outperform both the Multiple Linear Regression and Boosting Regression models. The use of hyperparameter tuning for the Random Forest model leads to only a slight increase in test set R squared but perfect fit on the training set (suggesting very strong levels of overfitting). We also want to avoid the use of slight improvements in the test set R squared to justify the model with hyperparameter tuning because we would then be introducing bias from information in the

Model	Train set R squared	Test set R squared
Multiple Linear Regression	0.2703166789751422	0.15717867705481747
Boosting Regression (AdaBoost)	0.6810832282921078	0.2939429088986316
Random Forest Regression	0.9222982930546839	0.3645988762013138
Random Forest Regression with Hyperparameter Tuning	1.0	0.3701463941497942

Figure 10: A comparison of the performance of each machine learning model on the training and test sets using the R squared metric.

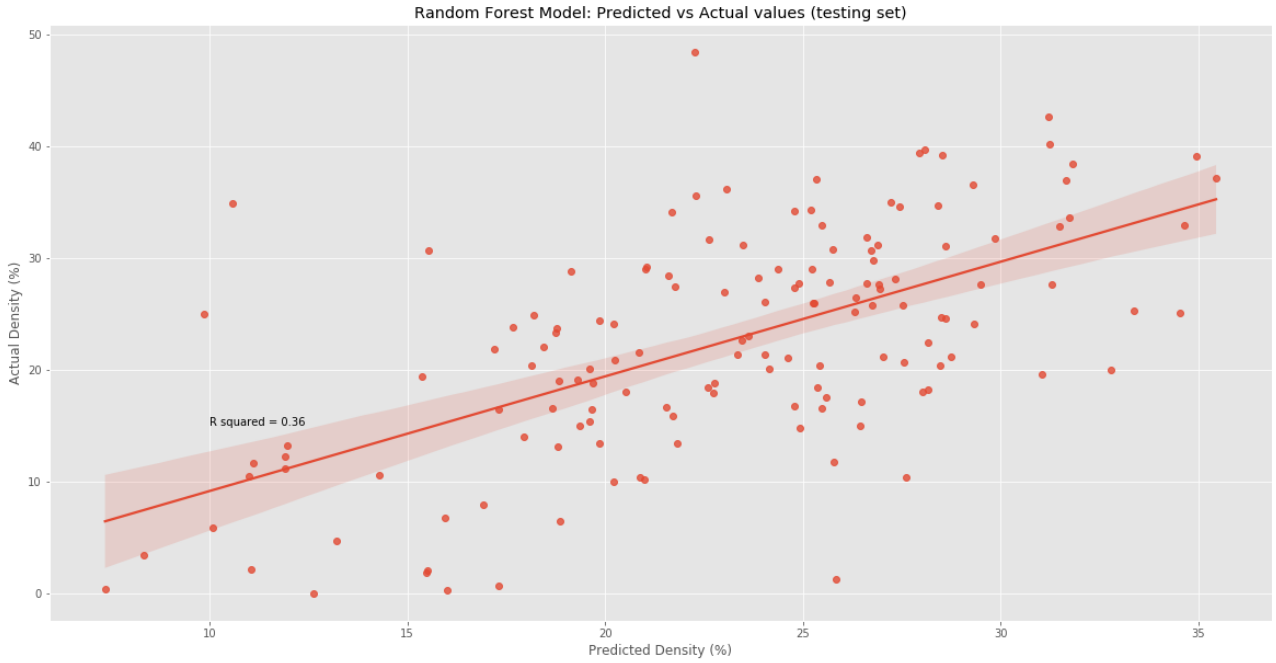


Figure 11: A regression plot comparing the predicted density of solar installation on the x-axis to the actual values on the y-axis.

testing set. Thus, we select the first Random Forest Regression model as our final model with an estimated out-of-sample R squared of 0.36. A plot of the predicted responses from this model compared to the actual responses is given in Figure 11.

5 Discussion

5.1 Interpretation of analysis

The analysis within this report was primarily to build a machine learning algorithm to predict solar installation rates (density) in each LGA based on demographic data and we have done this. The question of which demographic factors in particular are relevant in predicting solar installation rates is not answered, and as was mentioned is beyond the scope of this analysis. In this analysis we merely considered the question of using a large part of the combined information by demographic data to make predictions. In fact, the use of principal components analysis in our analysis obscures with our ability to answer questions pertaining to influential factors since it introduces interpretability issues. Where interpretability is concerned, an improvement that could be made is to use alternative dimension reduction techniques such as selecting the demographic variables with highest correlation with density or selecting those which have the lowest p-value when performing simple linear regression with density as the response. This would allow us to then evaluate the variable importance within our Random Forest model.

Insofar as this analysis is concerned, the results should be interpreted as:

- There is indeed a relationship between the solar installation rates and demographics of a LGA.
- A machine learning algorithm can indeed be built to at least roughly predict the average solar installation rates of LGAs with similar demographics. The model that we built in particular captures around 36% of the variability in solar installation rates via demographic information.

5.2 Implications of results

The developed model at current may be of only limited utility to policy makers or investors but could still potentially find use by researchers. The model does not have high enough predictive accuracy to reliably guide policy or investment decisions. Also, because the model was trained on demographic information from the 2016 census, its predictive ability is likely only limited to time periods around 2016. The model cannot account for social changes pertaining to solar installation, influential policy decisions, technological breakthroughs and other factors which may abruptly cause the model to be outdated. Instead, the density estimates can be used as *indicators* regarding the relative likelihood to install solar within a LGA keeping all other things constant. Researchers could still potentially use this model as a tool to run simulations under different demographic scenarios to guide parts of their research (while it has yet to be outdated), or as a basis to build more highly predictive and time-robust machine learning models.

5.3 Recommendations

- Conduct qualitative research to elicit what are the most relevant demographic or even individual factors affecting solar installation.
- Perform further data collection on a finer scale to create larger datasets for computational analysis, preferably after conducting the qualitative research so data of greater relevance is collected.

6 Conclusion

We were able to build a Random Forest Regression model that explains around 36% of the variability in solar installation rates across different LGAs via demographical information. Through this model we demonstrated that there is indeed a relationship between solar installations rates and demographics. The built model is of insufficient accuracy for use in policy or investment but could be of use for researchers under marginal circumstances.

References

[1] https://www.ipcc.ch/report/renewable-energy-sources-and-climate-change-mitigation/

[2] https://www.pewresearch.org/fact-tank/2016/10/05/americans-strongly-favor-expanding-solar-power-to

[3] https://www.seia.org/solar-industry-research-data

[4] https://www.ga.gov.au/scientific-topics/energy/resources/other-renewable-energy-resources/solar-energy

[5] Australian PV Institute (APVI) Solar Map, funded by the Australian Renewable Energy Agency, accessed from pv-map.apvi.org.au on 19 February 2020. Dataset can be downloaded from https://pv-map.apvi.org.au/historical#4/-26.67/134.12.

[6] http://apvi.org.au/about-us/

[7] http://www.cleanenergyregulator.gov.au/RET/Forms-and-resources/Postcode-data-for-small-scale-insta

[8] https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.003July%202016?OpenDocument

[9] Australian Bureau of Statistics 2020, viewed 26 February 2020, http://stat.data.abs.gov.au.

[10] Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1-23.