

A Predictive Model to Forecast Customer Adoption of Rooftop Solar

Vivian Sultan, Bader Alsamani, Norah Alharbi, Yazeed Alsuhaibany, Mansour Alzahrani

Claremont Graduate University

Center for Information Systems and Technology, Claremont, CA, USA

Email: vivian.sultan@cgu.edu

Abstract—This study was designed to explore the overarching research question “How can we forecast residential solar rooftop adoption?” In order to answer this question, we utilized big data tools to forecast residential solar rooftop adoption likelihood based on both customer demographics and expenditure data. A model was developed utilizing SPSS, Azure, and ArcMap 10.3 software to yield an accurate prediction of household profiles likely to adopt solar rooftops. The study revealed a total of ten variables that significantly impact homeowners’ decisions regarding whether or not to install solar rooftops. These ten variables accounted for around 23 or 24 percent of solar adoption variances. Importantly, this predictive analysis can be applied for all Energy Informatics use types; however, we focused on the residential parcel type located within LA County due largely to data availability. Future research is needed to expand on this analysis and to include the whole state of California, as to apply this analysis to other Energy Informatics cases, such as commercial properties and factories.

Keywords—component; renewable energy sources, photovoltaic solar systems, DERs

I. INTRODUCTION

Evidence indicates that worldwide energy demand is growing, significantly. Moreover, the current supply from fossil fuels and non-renewable energy sources is not sufficient to meet this ever-increasing demand. Additionally, research has shown, unequivocally, that fossil fuels cause irreparable damage to the environment and are directly responsible for climate change, increased levels of air pollution, and a dramatic rise in carbon emissions [1]. Therefore, the global community is currently plagued with two, overarching issues – a lack of energy to meet increasing demands and a climate crisis caused, largely, by fossil fuel use. One viable solution to solve both of these pressing problems is to supplement the current energy supply with renewable energy sources (RES). Arguably, the most feasible type of RES available in today’s marketplace is photovoltaic (PV) solar systems, which are commonly referred to simply as rooftop solar or solar energy [2].

Pertinent studies have been done by a wide range of subject-area experts to pinpoint the most likely customers to invest in renewable energy technology for their personal residences. Some of the factors, which help determine consumer likelihood, include household income level, house ownership, education level, and other financial factors (Islam,

2014; Long, 1993; Kasulis et al., 1981; Poortinga et al., 2003; Walsh, 1989; Held, 1983; Olsen, 1981; Heberlein and Warriner, 1983). Numerous studies have shown that these factors can and should be used as predictive models for future PV solar systems’ customers. For instance, Zhang et al. [3] created and, subsequently, demonstrated the effectiveness of their age-based model for PV rooftop solar adoption. In this project, we present another useful predictive model that accurately analyzes the available data using the Azure Machine Learning platform (Azure), alongside SPSS analysis.

II. DEFINING THE PROBLEM

In August 2014, the state regulatory body responsible for electric utilities, California Public Utilities Commission (CPUC), called for distributed energy resources (DERs) to be integrated into the electric distribution system. According to the Electric Power Research Institute [2], DERs are defined as “smaller power sources that can be aggregated to provide power necessary to meet regular demand.” An example of DERs is a solar rooftop. Then, in February 2015, CPUC issued a distribution resources plan (DRP) which called for transforming the current electric distribution system to a newer and more integrative system that allows for two-way energy flow between customers and their respective utilities companies [4]. This was vital because it gave customers an incentive to install renewable energy sources, since they can be directly tied into their power grid, thereby negating the need for batteries or other forms of energy storage devices.

Based on these new energy-related mandates, solar energy companies have predicted a surge in customer demand. Our research addresses the time-relevant question: “How can we forecast residential solar rooftop adoption?” The aim of the project is to explore big data and analytics tools to help forecast the residential solar rooftop adoption rates based on customer demographics and expenditure data. The study is particularly pertinent and inspiring since the predictions made based on evidence uncovered in this study could, potentially, become a valuable resource for policymakers who are seeking out more efficient and effective policies to promote solar adoptions. If particular patterns can be confirmed, then policymakers will, likely, be more acceptable to consider implementing and reinforcing them through laws and mandates. Ultimately, the result would be a higher rate of residential solar rooftop installations. Specifically, the main objective of our research

is to develop a predictive analysis based on household demographics and expenditure data to yield a projection of household profiles based on their likeliness of adopting solar rooftops. While the focus of this paper is residential solar rooftop adoptions, this model can, subsequently, be applied for all energy informatics on various levels, to include commercial, factories, and residential uses.

III. DATA SELECTION AND ACQUISITION

A. Solar Parcel Data

For this research, Los Angeles County's Solar Map and Green Planning Tool (<http://solarmap.lacounty.gov>) served as the primary data source. Its solar map was selected as the first data input. It includes the following pertinent information:

- Total roof area and area suitable for solar: This data is estimated for each parcel based on a 2006 countrywide solar radiation model; the model calculates and ranks incoming solar energy every 25 square feet using pitch, orientation, and shading from surrounding structures and vegetation to provide the most accurate estimates possible;
- Potential solar system size: This data is calculated from each parcel's optimal roof area and panel specifications; calculations are based on a 18.1 percent efficient SunPower 225 panel;
- Potential solar annual output: This data is calculated by multiplying potential system size by 5.3 hours of generation each day and a 75 percent performance ratios that takes into consideration alternating current conversion (AC), electrical losses, and other environmental factors such as soiling; and
- Potential Cost Savings: This variable is computed based on selected rates for each utility.

B. Customer Expenditure Data

The research relied on data obtained from the Environmental Systems Research Institute (ESRI) for customer expenditures and demographic data.¹ ESRI reports data at the block level, and the reported data relevant for this study includes total expenditures, average spending per household, and a Spending Potential Index (SPI).

C. Customer Demographic Data

ESRI was also utilized to obtain pertinent customer demographic data. Particularly, data from the following categories were obtained:

- Population demographic data: Included data such as age, sex, race, labor force, education level, marital status, civilian labor force number, and employment by industry and occupation;

- Household data: Included data such as total number of households, total number of family households, and average household size; and
- Income data: Included data such as total household income, per capital income, income by age, disposable income, and net worth.

The following is a detailed description of the steps taken to prepare the data to be analyzed:

- Step 1: Data were extracted and loaded into the ArcMap to show three map layers. The first map layer showed LA County solar data by parcel. The other two map layers showed customer expenditures and demographic data by block.
- Step 2: Parcel data from LA County's solar map were combined with customer expenditures and demographic data. Using the ArcMap model builder tool, the following model was specifically designed to spatially join the aforementioned three map layers and provide a table output in Dbf file format (See Figure 1 and Figure 2).
- Step 3: Unnecessary and null fields were eliminated from the dataset to reduce the file size to 2.78 GB.
- Step 4: The output file was saved as a geodatabase and exported as a txt file.
- Step 5: The text file was converted to CSV and loaded into the Microsoft Azure Machine Learning platform.
- Step 6: A SQL query was written to pull the current data pertaining to the installations of solar rooftops (JSON file) from the LA County's Solar Map and Green Planning Tool database.
- Step 7: The output JSON file was converted to a CSV file format using Python code (See Figure 3).
- Step 8: A second model was constructed using the ArcMap Model Builder to obtain an output map layer showing solar installations within LA County and their respective expenditure and demographics data by parcel. The Dbf output file was converted to a CSV file using the aforementioned steps. Finally, the installation data was merged with the entire dataset to identify which households have solar rooftops.

IV. MODEL DESIGN

After taking into consideration the literature review and the defined problem the research question was designed to answer, the model was designed using the following key variables:

- Dependent variable: likelihood of a household adopting solar energy panels; and
- Independent variables: (1) parcel information, such as: parcel age, parcel value, etc. (2) customer demographics, such as: household income, household size, etc. (3) expenditure data, such as: electricity usage, mortgage value, etc.

¹ ArcGIS online subscription is required to get access to this database. ESRI's United States customers' expenditure data is based on the latest Consumer Expenditure Surveys (CEX) from the Bureau of Labor Statistics.

This project relied on the following three tools to analyze and process data, as well as explore trends and patterns:

- Microsoft Azure Machine Learning (Azure): Azure has the capacity to store and analyze large datasets. Azure Machine Learning Studio is a GUI-based, integrated development environment for machine learning workflow. The predictive model created from this study was built using this particular model.
- SPSS: The standard and most widely used software package for complex statistical analysis.
- ArcGIS: A scalable and secure software-as-a-service program hosted by Esri. It can both process data and visualize the results.

According to Dasu and Johnson [5], data cleaning is best understood as the process by which anomalies and errors are removed from the case data. This allows data to be analyzed accurately and efficiently. For this research project, we employed several data cleaning methods, using both Azure and SPSS. The following is a breakdown of each of the methods we used, as well as our justification:

- Clean missing data: Designed to remove missing values and replace them with a placeholder, mean, or other value, when appropriate. Otherwise, the columns and rows with missing numbers are completely removed. After this step was conducted, we ended up with 94 variables and 713,000 records. Eventually, through this method, we were able to get our data down to 69,000 records.
- Descriptive statistics: Calculates the mean and median of the variable, standard deviations, and the minimum and maximum values. This tool is

useful for summarizing and analyzing data that can be used for meaningful applications.

- Frequency testing: Tests performed to locate ‘dirty data’ among the variables. This technique was utilized to help identify the distribution of the variables and detect unequal distribution patterns within each category.

According to Johnson and Wichern [7], data transformation is the application of a deterministic mathematical function to each and every point within a data set. To do so, where f is a function, each data point is replaced with a transformed value – $y_i = f(z_i)$. Typically, data transforms are utilized in order for data to appear to more closely meet the presuppositions of the statistical inference procedure that is to be applied. Also, data transforms are used improve interpretability or the aesthetic appearance of graphs. For our project, we transformed some columns to Z scores. Other columns, we divided. For instance, we divided the city identification column into four columns – three for each of the largest cities and one for all other cities (See Fig. 4).

The information obtained from the aforementioned tests revealed that there was a significant gap in the number of residences that installed solar panels (1,700) and the number that did not (711,300). Therefore, the results of the tests were not significant; therefore, we had to conduct a random sampling in order to even the records for analysis and comparative purposes. We used SPSS to obtain a random sample of 1,700 residences that did not have solar panels and combined this data with the data obtained from the 1,700 residences that did have solar panels. Subsequently, the same tests were performed and the results varied, significantly.

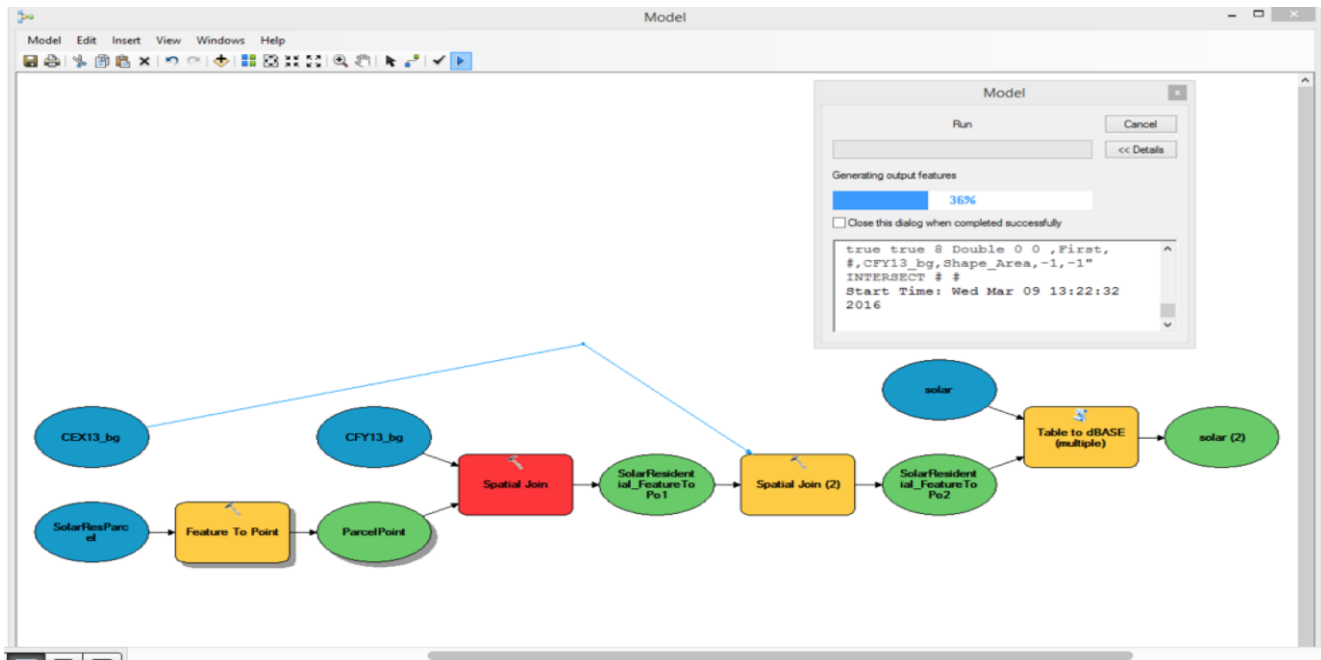


Figure 1: ArcMap Model Builder to add demographic and expenditure data to each parcel

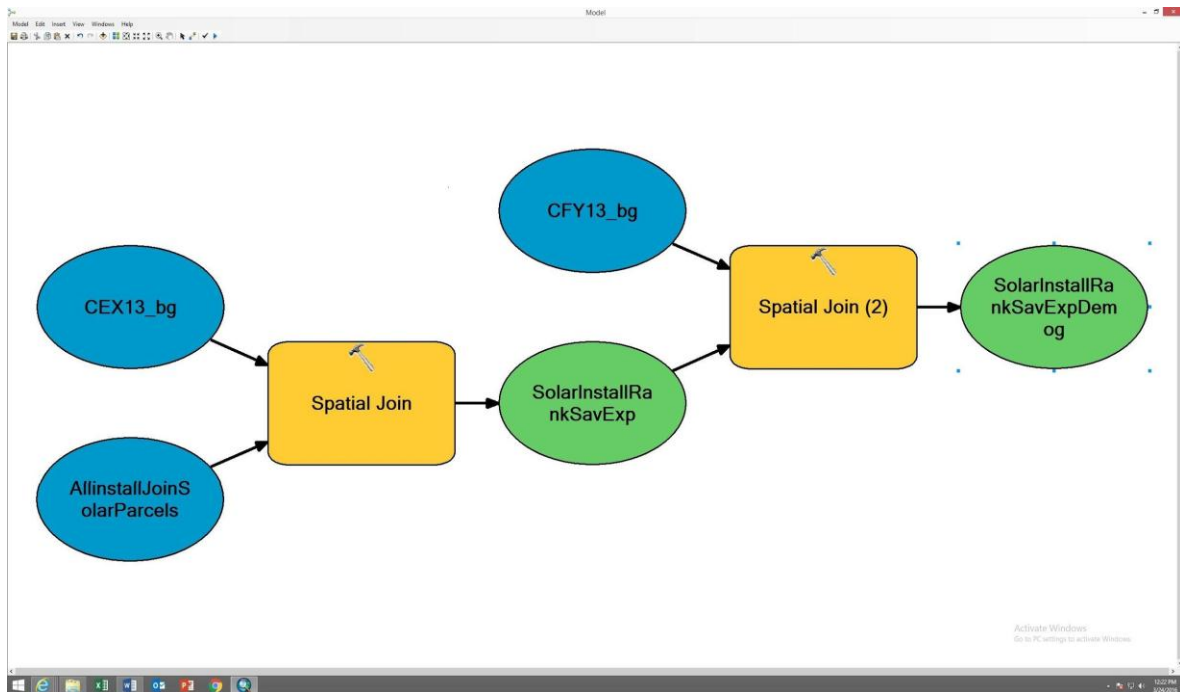


Figure 2: ArcMap Model Builder to add demographics and expenditure data to parcels with current solar installations

```

import csv
import json

x="""[
  { "pk": 22, "model": "auth.permission", "fields":
    { "codename": "add_logentry", "name": "Can add log entry", "content_type": 8 }
  },
  { "pk": 23, "model": "auth.permission", "fields":
    { "codename": "change_logentry", "name": "Can change log entry", "content_type": 8 }
  },
  { "pk": 24, "model": "auth.permission", "fields":
    { "codename": "delete_logentry", "name": "Can delete log entry", "content_type": 8 }
  }
]"""

x = json.loads(x)

f = csv.writer(open("test.csv", "wb+"))

# Write CSV Header, If you dont need that, remove this line
f.writerow(["pk", "model", "codename", "name", "content_type"])

for x in x:
    f.writerow([x["pk"],
                x["model"],
                x["fields"]["codename"],
                x["fields"]["name"],
                x["fields"]["content_type"]])
  
```

Figure 3: JSON File Converted to CVS Format Using Python Code

N	O	P	Q
Correlations			
LosAngelesCity	LongBeachCity	PasadenaCity	OtherCity
-.024	-.037	.077	-.008
.161	.031	.000	.626
3373	3373	3373	3373
-.006	-.044	.134	-.058
.710	.010	.000	.001
3373	3373	3373	3373
-.032	-.034	.010	.041
.060	.046	.579	.017
3373	3373	3373	3373
-.096	-.019	-.068	.143
.000	.269	.000	.000
3373	3373	3373	3373
-.105	-.087	-.179	.256
.000	.000	.000	.000
3373	3373	3373	3373
-.062	-.069	-.167	.199
.000	.000	.000	.000
3373	3373	3373	3373
-.051	-.021	.090	.000
.003	.228	.000	.995
3373	3373	3373	3373
-.063	-.017	.089	.010
.000	.328	.000	.560
3373	3373	3373	3373
-.095	-.017	.060	.058
.000	.336	.000	.001
3373	3373	3373	3373
-.055	-.013	.086	.002
.001	.462	.000	.893
3373	3373	3373	3373
-.073	-.018	.063	.037

Figure 4: The new four columns for the largest three cities and the fourth one for the rest of the cities

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
ParcelAge	Equal variances assumed	.237	.627	-6.096	3219	.000	-5.874	.964	-7.763	-3.985
	Equal variances not assumed			-5.945	1326.215	.000	-5.874	.988	-7.813	-3.936
ZscoreTotal4R	Equal variances assumed	3.286	.070	-.558	3219	.577	-.02268	.04067	-.10241	.05706
	Equal variances not assumed			-.751	2685.382	.453	-.02268	.03020	-.08190	.03654
zscoreforTotalaresuitableforsolarrooftopr1r2	Equal variances assumed	6.673	.010	-2.006	3219	.045	-.086862883	.0432923785	-.171746302	-.001979463
	Equal variances not assumed			-1.573	1002.508	.116	-.086862883	.0552364900	-.195255277	.0215295117
zscoreforTotalbuildingareassquarefeet	Equal variances assumed	105.456	.000	5.233	3219	.000	.1858647984	.0355181654	.1162242883	.2555053085
	Equal variances not assumed			8.305	3131.623	.000	.1858647984	.0223797143	.1419844049	.2297451919
zscoreforAverageHouseholdSize	Equal variances assumed	172.904	.000	15.610	3219	.000	.6190173199	.0396546613	.5412663772	.6967682627
	Equal variances not assumed			18.999	2091.965	.000	.6190173199	.0325819195	.5551209626	.6829136773
zscoreforAverageHouseholdIncome	Equal variances assumed	26.560	.000	-18.784	3219	.000	-.726293642	.0386659568	-.802106030	-.650481253
	Equal variances not assumed			-17.882	1255.142	.000	-.726293642	.0410756735	-.806878191	-.645709093
zscoreforAverageHomeValue	Equal variances assumed	22.414	.000	-19.576	3219	.000	-.753291528	.0384797702	-.828738861	-.677844196
	Equal variances not assumed			-18.815	1295.884	.000	-.753291528	.0400371811	-.831836322	-.674746735

Figure 5: Independent samples t-test

V. DATA ANALYSIS

We performed a number of statistical tests using SPSS and Azure. The following is a brief description of the overarching aim of each test and its justification.

- The Independent-samples t-test: This test aims to determine whether or not the mean of a dependent variable is the same in two, unrelated and independent groups. This test is used to determine whether or not the mean different between two groups is significantly different to zero, statistically speaking. These tests were utilized to examine the bivariate relationship between all continuous factors, individually, (twenty continuous variables) and the dependent variable (solar adoption) (See Figure 5). Interestingly, only one continuous factor (solar saving) showed no significant different with regard to the amount of possible total savings between people who installed solar rooftops and people who

did not ($t = -0.565$, $p = 0.572$). One possible explanation is that households might opt to adopt solar panel not to save money but instead to preserve the environment. The remaining nineteen continuous factors show significant difference between the two groups (i.e. those who did and those who did not adopt solar rooftops), with $p < 0.05$.

- Chi Square: This test is utilized to reveal whether or not there is a relationship between two categorical variables. For this study, it was used to examine the relationship between solar adoption groups and both the electric company (See Table1; chi-square = 502.683. $df = 6$, $p = .000$) and LA county cities (See Table 1; chi-square = 871.160. $df = 84$, $p = .000$). The chi-square tests showed a significant association between solar adoption groups and both the electric company and LA county cities.

TABLE I. CHI-SQUARE TESTS FOR LA COUNTY CITIES FACTOR

Chi-Square Tests for LA county cities			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	502.683^a	6	.000
Likelihood Ratio	432.345	6	.000
Linear-by-Linear Association	9.903	1	.002
N of Valid Cases	4000		

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	871.160^a	84	.000
Likelihood Ratio	906.352	84	.000
Linear-by-Linear Association	14.741	1	.000
N of Valid Cases	4000		

The variance inflation factor (VIF) and tolerance: These scores were computed for all factors to determine whether the assumption of multicollinearity was met. The results indicated that only five factors did not have multicollinearity, ($VIF < 10$ and tolerance > 0.10) (See Table 2 and Table 3). This was not surprising because, for instance, it make sense that average household size and average family size highly associated with each other. This, the average household size variable was included with the five factors (due to high correlation with the dependent variable) and average household size was excluded. As a result of these combinations, the multicollinearity issue for average family size is resolved after removing average household size variable. For the rest-excluded variable, they also have multicollinearity issue as they are highly associated with

each other and they all related to household income. Therefore, we decided to include only the average household income as it comprehensively represent the rest of excluded variables. As a result, the average household income has no multicollinearity issue after removing all related variables.

Hosmer and Lemeshow test contingency: The purpose of this test is to examine the sample size assumption. According to Soliman Abu-Bader [5], "sample size assumption in logistic regression requires that none of the cells should have expected value less than 1 and no more than 20% of all cells should have expected frequencies of less than five cases" [5]. The results are shown in Table 4 Significantly, none of the cells has an expected value less than 1 and smallest expected value is 6.086.

TABLE II. VIF AND TOLERANCE ANALYSIS FOR ALL FACTORS

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	ParcelAge	.739	1.353
	zscoreforTotalaresuitableforsolarrooftopr1r2	.981	1.019
	zscoreforTotalbuildingareasquarefeet	.920	1.087
	zscoreforTotalPopulation	.705	1.418
	zscoreforAverageHouseholdSize	.036	27.576
	zscoreforAverageFamilySize	.037	27.159
	zscoreforAverageHouseholdIncome	.000	6284.688
	zscoreforAverageDisposableIncome	.003	347.215
	zscoreforAverageNetWorth	.049	20.205
	zscoreforAverageHomeValue	.222	4.496
	zscoreforCollegeTuitionAverage	.002	558.635
	zscoreforInvestmentsAverage	.005	191.724
	zscoreforOriginalMortgageAverage	.001	1368.174
	zscoreforClosingCostsAverage	.002	410.526
	zscoreforPurchaseofOtherPropertyAverage	.006	173.532
	zscoreforVehicleLoansAverage	.001	1268.909
	zscoreforHomeImprovementAverage	.001	1911.022
	zscoreforElectricityAverag	.000	2388.736
	zscoreforElectricityOwnerAverag	.001	1448.412

TABLE III. VIF AND TOLERANCE ANALYSIS AFTER EXCLUDING FACTORS WITH MULTICOLLINEARITY ISSUE

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	zscoreforAverageHouseholdSize	.734	1.363
	ParcelAge	.900	1.111
	zscoreforTotalaresuitableforsolarrooftopr1r2	.984	1.016
	zscoreforTotalbuildingareasquarefeet	.935	1.069
	zscoreforAverageHouseholdIncome	.407	2.456
	zscoreforAverageHomeValue	.352	2.845

TABLE IV. HOSMER AND LEMESHOW TEST CONTINGENCY

Contingency Table for Hosmer and Lemeshow Test

		InstallYN10 = 0		InstallYN10 = 1		Total
		Observed	Expected	Observed	Expected	
Step 10	1	316	315.914	6	6.086	322
	2	313	305.995	9	16.005	322
	3	307	296.535	16	26.465	323
	4	299	284.106	23	37.894	322
	5	262	272.089	60	49.911	322
	6	246	256.206	76	65.794	322
	7	213	233.481	109	88.519	322
	8	200	204.972	122	117.028	322
	9	162	167.054	160	154.946	322
	10	96	77.648	226	244.352	322

TABLE V. FACTORS CONTRIBUTE TO THE PREDICTION MODEL

Factor name	Description
Z score for Average Household Size	This variable is the Average Household Size that was taken from (Esri)
Parcel Age	The age of the house. Was calculated by the year that the house was build in - 2016
Z score for Total area suitable for solar roof top r1r2	Average Household Size (Esri)
Z score for Total building area square feet	
Z score for Average House hold Income	This variable is the Average House hold Income that was taken from (Esri)
Z score for Average Home Value	This variable is the Average Home Value that was taken from (Esri)
Utility company	This variable is the utility company that the parcel belong to
LA county cities	This variable had two value (1/0) for those parcels that were in LA City had a value of 1 and the rest had the value of 0

TABLE VI. COX AND SNELL AND THE NAGELKERKE R²

		B	Wald	df	Sig.	R2
Step 10	Pasadena Water and Power(1)	-2.516	144.542	1	.000	.092 - .137
	zscoreforAverageHomeValue	.193	6.288	1	.012	.172 - .255
	zscoreforAverageHouseholdSize	-.546	67.045	1	.000	.183 - .271
	zscoreforAverageHouseholdIncome	.543	54.491	1	.000	.195 - .289
	zscoreforTotalbuildingareasquarefeet	-1.592	32.886	1	.000	.216 - .319
	zscoreforTotalaresuitableforsolarrooftopr1r2	.297	29.183	1	.000	.223 - .330
	LONGBEACH(1)	-1.039	31.094	1	.000	.229 - .339
	Southern California Edison(1)	.261	5.759	1	.016	.231 - .342
	ParcelAge	.006	7.290	1	.007	.233 - .345
	Glendale Water and Power(1)	-.579	4.928	1	.026	.234 - .346
	Constant	1.832	19.300	1	.000	

Testing the four assumptions of logistic regression: In order to conduct the next step, it was vital for our data to pass through four assumptions that are required for a valid

result, per Pearson's correlation. The four assumptions are as follows:

- Assumption 1: The dependent variable is to be measured on a dichotomous scale;
- Assumption 2: The one or more independent variables can be either continuous (i.e. an interval or ratio variable) or categorical (i.e. an ordinal or nominal variable);
- Assumption 3: There should be independence of observations; the dependent variable(s) should have exhaustive and mutually exclusive and categories; and
- Assumption 4: There should be a linear relationship between all continuous independent variable and the dependent variable's logit transformation.

For this research project, we used logistic regression to explore the research question and determine how likely particular households will be to adopt solar systems. Logical regression is designed to predict the probability than a particular observation will fall into one of the dependent variable's two groups [6]. It is based on the understanding that the one or more independent variables can be categorical or continuous.

Prior to conducting the logistic regression, independent t-test, chi square tolerance and VIF were utilized to be certain that we did, in fact, meet all of the four assumptions of logistic regression, and that we would only include variables that were appropriate to the model and would add significant contributions.

A forward stepwise likelihood ratio (forward LR) logistic regression analysis was then conducted to shed light on the research question and accurately predict the probability of a household adopting solar systems. As previously indicated, we decided on eight factors, two of which were categorical factors since they met the assumptions and appeared to be the most appropriate contributing factors for the prediction model. Table 5 shows the comprehensive list of used factors used in the logistic regression model. The result of logistic regression shows that six continuous factors emerged as significant predictors of solar adoption groups. These factors are: parcel age (wald = 7.290, p = .007); average household size (wald = 67.045, p = .000); total area suitable for solar roof top (wald = 29.183, p = .000); total building area square feet (wald = 32.886, p = .000); average household income (wald = 54.491, p = .000); and average home value (wald = 6.288, p = .012).

Specifically, parcel age, total area suitable for solar rooftops, and average home value all showed significant positive relationships to solar adoption, while average household size, total building area square feet, and average household income showed significant negative relationship to solar adoption. This demonstrates that, typically, the older the house is, the larger the area suitable for solar rooftop

installation, and the higher the home value is, the more likely it will be to adopt solar systems. Conversely, the larger the household size, the greater the total building area and the higher the household include, the less likely it is to adopt solar systems.

Concerning the categorical variable, both the utility company and the LA county cities showed significant predictions of solar adoption. Specifically, seven different utility companies were utilized and each on was represented as dichotomous factors. Only three utility companies showed significant differences between the solar adoption groups. These were the Glendale Water and Power (wald = 4.928, p = .026), Pasadena Water and Power (wald = 114.542, p = .000) and Southern California Edison (wald = 5.759, p = .000). This strongly suggests that Southern California Edison household customers are more likely to adopt solar panel while Pasadena Water and Power and Glendale Water and Power customers show a significant negative relationship to solar adoption. For LA county cities, there are eighty-nine identified cities in LA County. Thus, for simplicity's sake, only three cities were transformed into individual dichotomous factors and the rest were classified as 'other.' However, only Long Beach city was identified as significant predictor to the model (wald = 31.094, p = .000).

The results show (See Table 6) of Cox and Snell and the Nagelkerke R² indicate that Pasadena Water and Power accounted for 9.2 to 13.7 percent of the variance in solar adoption. Average home value accounted for 8 to 11.8 percent of the variance. Average household size accounted for 1.1 to 8.3 percent of the variance. Average household income, total building area square and total area suitable for solar rooftop added 1.2 to 1.8 percent, 2.1 to 3 percent and 0.6 to 0.9 percent, respectively, to the variance in solar adoption. Finally, Southern California Edison, parcel age accounted for and Glendale Water and power also added 0.2 to 0.3 percent, 0.2 to 0.3 and 0.1 percent, respectively.

VI. TESTING THE MODEL

The model successfully classified 94.0 percent of households that did not install the solar panels and 35.4 percent of total cases that did not install the solar panel. Overall, the model has a success rate of 79.4 percent. To evaluate the model, we divided the dataset into two: 70 percent of the dataset for training and building the model and 30 percent of the dataset for testing the built model. Model testing result show relatively similar accuracy of solar adoption prediction. Specifically, the tested correctly classified 92.0 percent of households that did not install the solar panel and 33.6 percent of household that did not install the solar panel. Overall, the model has success rate of 78 percent (See Fig. 6, Fig. 7, and Fig. 8).

$$\text{Probability of event, } P(\text{event}) = \frac{1}{1 + e^{(a + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_{10} \cdot x_{10})}} = \frac{1}{1 + 2.718^{(1.832 + -2.516 \cdot x_1 + .193 \cdot x_2 + -.543 \cdot x_3 + 0.543 \cdot x_4 + -1.592 \cdot x_5 + .297 \cdot x_6 + -1.039 \cdot x_7 + .261 \cdot x_8 + .006 \cdot x_9 + -0.579 \cdot x_{10})}}$$

InstallYN10 * PredictedValue Crosstabulation

Count

		Total
InstallYN10	0	586
	1	193
Total		779

Classification Table^{a,b}

Observed		Predicted		Percentage Correct
		InstallYN10		
Step 0	InstallYN10	0	1	
		2414	0	100.0
		807	0	0.0
Overall Percentage				74.9

a. Constant is included in the model.

b. The cut value is .500

Classification Table^a

Step 10	InstallYN10	0	2270	144	94.0
		1	521	286	35.4
	Overall Percentage				79.4

Figure 6. Testing the model

The team applied a decision tree on the ten aforementioned identified variables to see which houses from the uninstalled parcels had the highest prediction to install solar rooftop. First, LA city was chosen to represent this finding. Second, we took each variable and identified the average value for this variable on parcels that had solar panels installed. Then, we tested the average on the uninstalled parcels. To demonstrate this step, we took, for example, the variable ‘parcel age’ which had an average value of 62 years in the parcels with solar panels. Thus, we created a new variable and we assigned two values (0/1) for this variable. We applied an equation to the parcel age of the houses that did not install rooftop. If the parcel is younger than 62 years, the value will be ‘0’, and if older than 62 years, the value will be ‘1.’ Subsequently, we conducted the same process to all 10 variables. Based on this modeling system, the parcels with the most ‘1’ scores were considered the most likely to install solar panels. In other words, the more ‘1’ scores a parcel received, the more likely it was predicted to install solar panels.

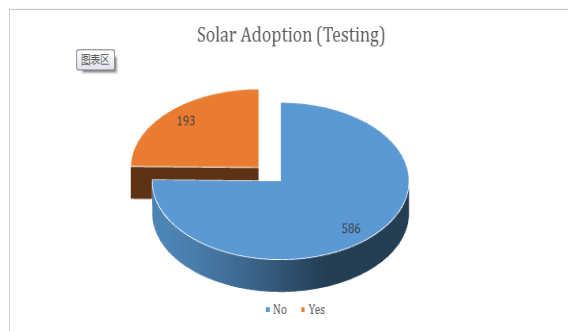


Figure 7. Testing the model

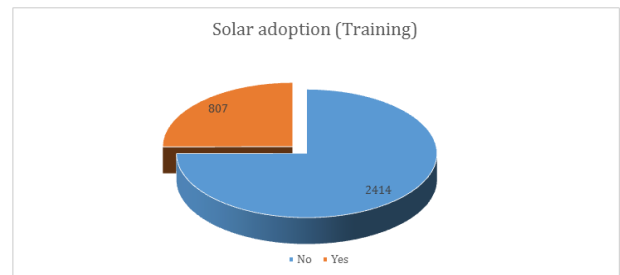


Figure 8. Testing the model

Furthermore, we decided to use GIS to view the above results. After diligently going through the whole modeling process, we ended up with a map containing all the parcels with solar rooftop panels installed and the predictions for the remaining parcels. There are three steps to build the GIS model. First, we collected all pertinent parcel data for LA County and combined this data with the project sample from the information table. This was accomplished using ArcMap and the “AssessorID” field, as well as the output stored in GeoDatabase, which used a Dbf file format. Second, we generated a selected parcel layer, which contained a total of 2392100 parcels located within LA County. Using the Future to Layer tool, we converted the future to point. This step made it very easy to both present and visualize the map (See figure 8). Third, we used Future tools to enable the layer to be split into different layers based on our selected options and fields, which were as follows:

- Parcels with solar rooftops already installed.
- Parcels with various predictions by percentage.
- The total of parcels within LA County.

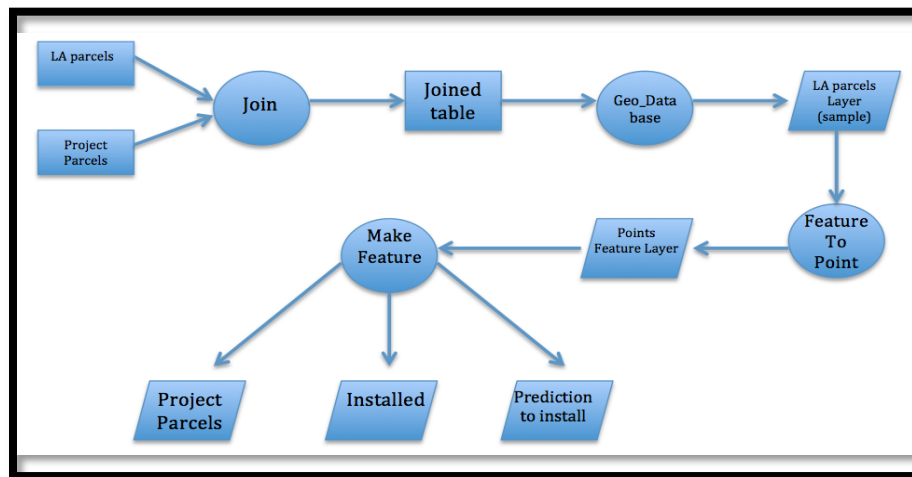
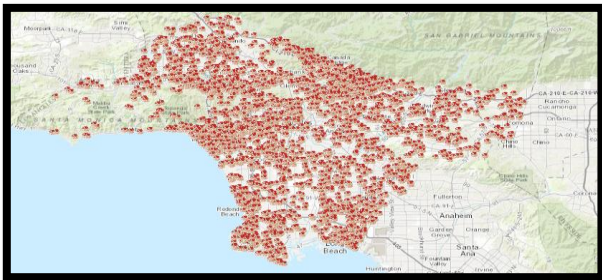
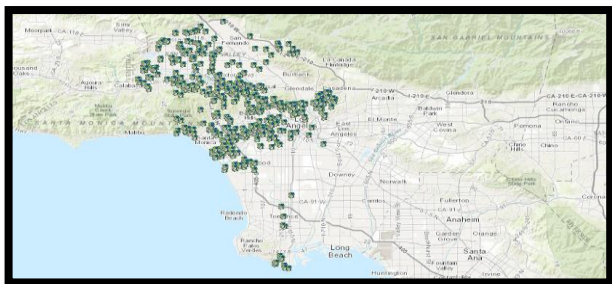


Figure 9. GIS Model Steps

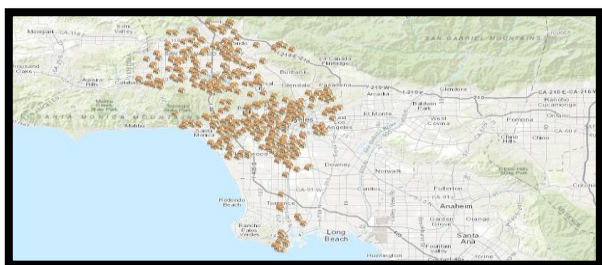
After creating the various layers using ArcGIS software, we were able to visualize our results. We utilized ArcGIS online to transfer the data onto the map and to make it accessible via the web.



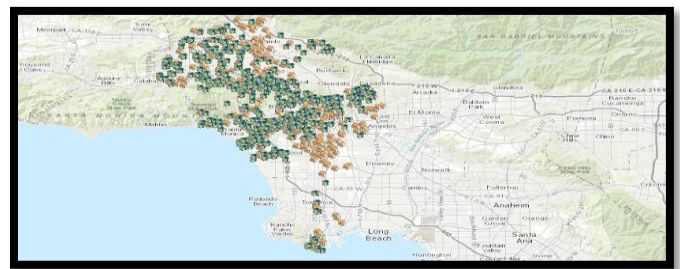
Map 3: The sample parcels for various cities within LA County



Map 4: The parcels with solar rooftops installed in the city of LA



Map 5: The parcels we predict will install solar rooftops in the city of LA



Map 6: The total of parcels within LA County

VII. CONCLUSION

This study set out to address the overarching research question “How can we forecast residential solar rooftop adoption?” In order to answer this overarching question, we explored big data tools to forecast residential solar rooftop adoption based on both customer demographics and expenditure data. A model was developed utilizing SPSS, Azure, Tableau, and ArcMap 10.3 software to yield an accurate prediction of household profiles likely to adopt solar rooftops. Importantly, this predictive analysis can be applied for all Energy Informatics use types; however, we focused on the residential parcel type located within LA County due largely to data availability. The study revealed a total of ten variables that significantly impact homeowners’ decisions regarding whether or not to install solar rooftops. These ten variables accounted for around 23 or 24 percent of solar adoption variances. Future research is needed to expand on this analysis and include the whole state of California, as to apply this analysis to other Energy Informatics cases, such as commercial properties and factories.

ACKNOWLEDGMENT

We would like to acknowledge the efforts of Dr. Brian Hilton who provided insightful comments and helped us with the project analysis. Also, we would like to thank Mr. Warren Roberts and the group working in CGU GIS lab for their support.

REFERENCES

- [1] McLamb, E. (2011). Fossils Fuels vs. Renewable Energy | Ecology Global Network. Retrieved May 9, 2016, from <http://www.ecology.com/2011/09/06/fossil-fuels-renewable-energy-resources/>
- [2] Electric Power Research Institute (2015). Distribution Resources Plan. Retrieved May 4, 2016, from <http://www.epri.com/Our-Work/Pages/Distributed-Electricity-Resources.aspx>
- [3] Zang, H. & Vorobeychik, Y. & Letchford, J. & Lakkaraj, K. (2015). Predicting Rooftop Solar Adoption Using Agent-Based Modeling. Retrieved May 9, 2016, from <http://www.aaai.org/ocs/index.php/FSS/FSS14/paper/viewFile/9126/9124>
- [4] California Public Utility Commission (2015). Distribution Resources Plan. Retrieved October 10, 2015, from <http://www.cpuc.ca.gov/PUC/energy/drp/index.htm>
- [5] Dasu, T. & Johnson, T. (2003). Exploratory data mining and data cleaning. New York: Wiley Interscience.
- [6] Abu-Bader, S. H. (2010). Advanced and multivariate statistical methods for social science research with a complete SPSS guide.
- [7] Johnson, R. & Wichern, D. (1982). Applied multivariate statistical analysis. Englewood Cliffs, N.J.: Prentice-Hall.