David Logsdon
GADS4 - Final Project Write-Up

Customer Distributed Generation Adoption Predictor

I. Problem Description

Con Edison, my employer, is the distribution utility servicing electricity, gas, and steam within New York City and Westchester. One of the most rapidly development and potentially disruptive trends affecting utilities such as Con Edison is the increasing willingness of our customer base to make the jump from power consumer to power producer through the adoption of distributed generation (DG). For our customer base DG takes two common forms: combined heat and power (CHP) and solar photovoltaics (PV). CHP is primarily adopted by larger commercial buildings or multifamily buildings have thermal needs that can take advantage of the generator's heat output--more than 100 of our customers have installed a total of 150 MW of CHP. The primary drivers for CHP adoption are high electric rates, available incentives, low natural gas prices, and a desire for resiliency and continuity during grid outages. Solar PV has been growing at astounding rates recently--annual growth for 2013 is on track to exceed 80%--more than 1,400 of our customers have installed a total of 27 MW of solar PV. This number was only 10 MW at the start of 2012; the primary drivers of solar PV adoption are high electric rates, robust available incentives, environmental stewardship, and decreasing panel prices.

As a utility it is important to our planning processes to be able to accurately predict DG adoption. Adoption curves over time will have an impact on our customers' rates and on our need for infrastructure spending. We have a decent understanding of overall rates of adoption and of market and technical potentials for both of the primary DG technologies. However, the location of DG adoption is very key to utility planning. DG in the right network (Con Edison has 64 distribution networks that divide our service territory) can actually allow us to defer infrastructure investment. Some networks have load relief needs and some do not. If we could take an entire network's customer base and apply an adoption prediction to understand which customers are most likely to install DG we could have a better understanding of true market potential and we could shape future incentive programs to have the greatest benefit to the most customers.

I view this project as an on-going one that this class has given me the initial tools to tackle. In order to get a working model in place by the deadline, I have scaled this project to fewer databases (two total) than will be included in the final database and have limited it exclusively to solar PV adoption.

Hypothesis

I believe that with a sufficient amount of customer data I will be able to construct a model that will accurately predict solar PV adoption with an AUC of .75 or greater. I believe the data I have available on customers (outlined in the next section) will be sufficient to this task as I will have a large amount of data on customers electric usage, buildings, and

location available both for customers who have and have not installed solar. I hypothesize that there is enough correlation between these variables, in aggregate, with solar PV adoption to give an accurate prediction of customer likelihood to adopt.

Data Set Descriptions

Two primary databases were used in creating the solar PV adoption predictor.

PV MASTER DATABASE

This is a database of customers who have already installed solar. Primarily all I need to get from this database is my dichotomous variable, “Has_PV.” If a customer is in the PV Master Database this dichotomous variable is 1 as they have installed solar PV already. If they do not appear in the PV Master database, Has_PV = 0. My approach is to compare the PV Master Database with my second database, the Rooftop Potentials database, based on customer account codes. If an account number appears in both, I set a new variable, Has_PV to 1 and add it to the Rooftop Potentials database. There are 1,400 entries in the PV master database.

ROOFTOP POTENTIALS DATABASE

This database contains 34,901 rows (customers) and 24 columns representing characteristics related to 3 categories of customer information: building information, electricity usage information, and locational information. I will not outline the characteristics individually here as they will be outlined later in the discussion on model creation and evaluation, but to describe the database further: it was created specifically to gauge the solar potential of different buildings so it includes estimates of how much solar each

building's roof would be able to accommodate.

MODELLING APPROACH

INITIAL MODEL SELECTION

As I wanted to classify the probability of customer adoption with the dichotomous variable Has_PV, I selected Logit Regression, and I used the statsmodel.api package for Python to implement it. This is certainly not the only modelling approach that could be taken. An alternative approach that I would like to try in the future and compare to the logit model I arrived at would be Naive Bayes, which should prove a valid approach to such a classification problem and would allow experimentation with the priors.

DATA MUNGING

As expected, data munging was the most time consuming task in the entire project. Account numbers can appear in either 14 or 15 digit form so I had to standardize those between the two databases in order to compare and create my dichotomous Has_PV variable in the Rooftop Potentials database. I was disappointed to learn only 156 customers who had installed solar appeared in the Rooftop Potentials database. I suppose this is because I created the database looking only at larger, commercial customers. A topic worth looking in to further would be expanding the Rooftop Potentials database to represent even smaller residential customers to create a more complete model. There were many other data formatting requirements to get all of the columns in to a format that the logit model would accept. I found Pyhton and Pandas to be very powerful in performing

these operations; I learned to love the lambda function feature.

BUILDING TEST/TRAIN SETS FOR CROSS VALIDATION

In order to create an accurate model and test it for accuracy I needed to use cross-validation, creating a test and train set for which I knew the true Has_PV value. My train dataset contained 106 customers who have installed solar PV (65% of the total and an additional 106 customers who have not. My test dataset contained 56 customers who have installed solar PV (35% of the total) and an additional 56 who have not.

BUILDING THE LOGIT MODEL

On the initial pass of the logit model, I didn't want to pre-eliminate any variables, by running the model with all variables, I would be able to see key characteristics for each that could be used to build the most robust model. Here is the initial result summary, which fails to resolve:

Dep. Variable:	Has_PV	No. Observations:	211
Model:	Logit	Df Residuals:	187
Method:	MLE	Df Model:	23
Date:	Tue, 13 Aug 2013	Pseudo R-squ.:	0.5311
Time:	20:11:45	Log-Likelihood:	-68.584
converged:	False	LL-Null:	-146.25
		LLR p-value:	1.269e-21

	coef	std err	z	P>|z|	[95.0% Conf. Int.]	
Yr_kWh	-4.527e-06	2.72e-06	-1.665	0.096	-9.86e-06	8.01e-07
Max_kW	-0.4352	3.618	-0.120	0.904	-7.526	6.655
LF	4.4792	2.516	1.780	0.075	-0.452	9.411
min load	1.8446	14.470	0.127	0.899	-26.516	30.205
kw export	0.2043	2.03e+04	1.01e-05	1.000	-3.98e+04	3.98e+04
NumBldgs	-0.0952	0.143	-0.665	0.506	-0.376	0.186
BldgArea	8.472e-05	2.58e-05	3.288	0.001	3.42e-05	0.000
NumFloors	-1.0041	0.287	-3.494	0.000	-1.567	-0.441
BldgFront	-0.0067	0.005	-1.353	0.176	-0.016	0.003
BldgDepth	-0.0169	0.006	-2.830	0.005	-0.029	-0.005
AreaPerFloor	0.0035	0.007	0.481	0.630	-0.011	0.018
roof capacity	-0.3878	0.811	-0.478	0.632	-1.977	1.201
Zone_H	16.3462	4.75e+07	3.44e-07	1.000	-9.3e+07	9.3e+07
Zone_I	-0.2857	2.197	-0.130	0.897	-4.592	4.021
Zone_J	-1.5453	1.815	-0.851	0.395	-5.103	2.012
Boro_QN	-0.4704	1.000	-0.470	0.638	-2.430	1.489
Boro_BX	-1.4673	1.669	-0.879	0.379	-4.738	1.803
Boro_MN	-1.8803	1.773	-1.061	0.289	-5.355	1.595
Boro_SI	-0.5728	1.676	-0.342	0.732	-3.857	2.711
Boro_WS	14.7596	1174.483	0.013	0.990	-2287.184	2316.703
Cis_Lat	1.5726	6.362	0.247	0.805	-10.897	14.042
Cis_Long	0.8557	3.498	0.245	0.807	-5.999	7.711
Block	3.315e-05	8.41e-05	0.394	0.693	-0.000	0.000
Lot	-3.425e-05	0.000	-0.161	0.872	-0.000	0.000

From this output I immediately noticed that some of the locational data was not being correctly interpreted; the latitude (Cis_Lat) and longitude (Cis_Long) were being interpreted linearly/directionally and the Block and Lot info was too granular to accurately influence the model. I did not have time to do so for this assignment, but in the future I'll go back and create dummy variables to properly bin these aspects for analysis. The LL-NULL value is also very poor because this is a poor model and does not disprove the null hypothesis of no relation between the independent and dependent variables. In order to screen this down to only the values making a significant contribution to the model, I weed out all parameters with p values greater than .1. This is throwing out data that might be able to contribute to the model with closer attention or in another format--this is something else

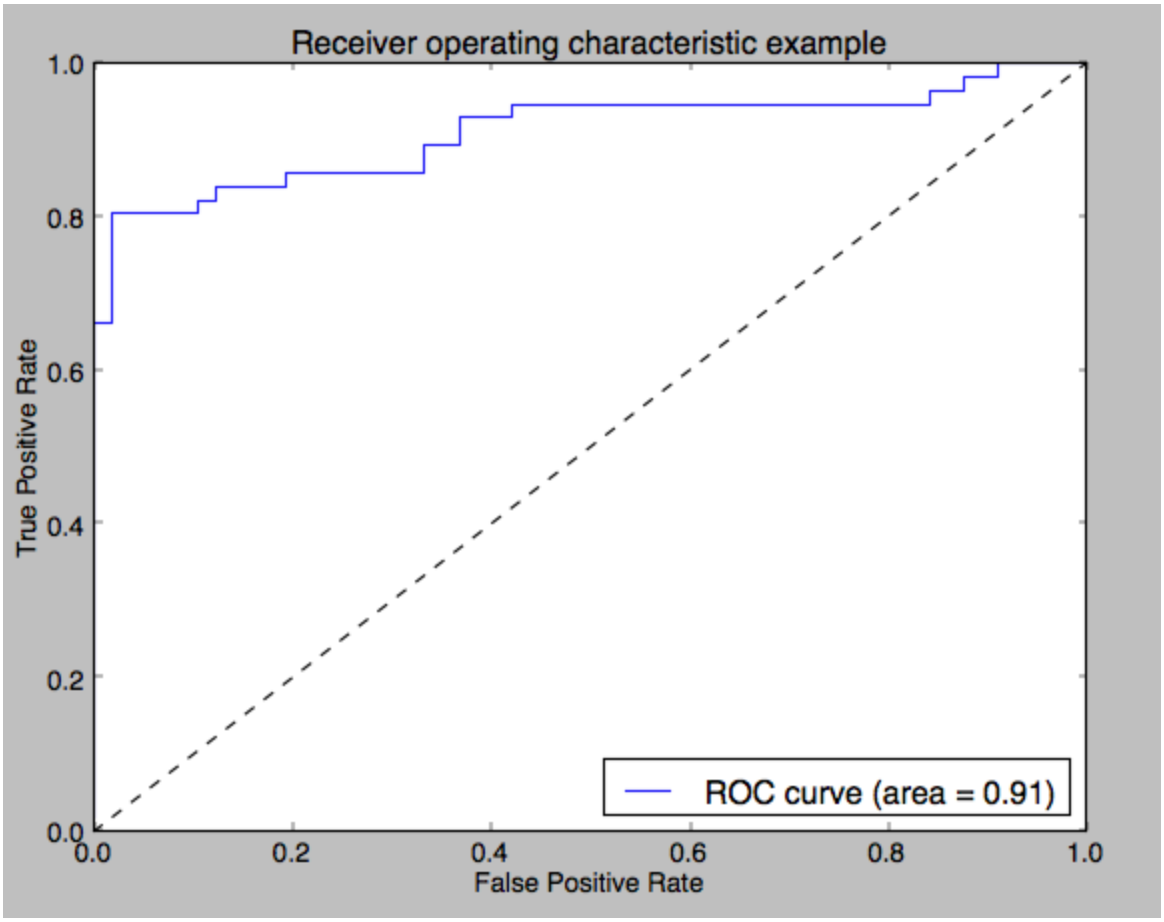
to look at more in the future. Here is the new model output, with only significant parameters remaining:

Logit Regression Results

Dep. Variable:	Has_PV	No. Observations:	211
Model:	Logit	Df Residuals:	206
Method:	MLE	Df Model:	4
Date:	Tue, 13 Aug 2013	Pseudo R-squ.:	0.3483
Time:	22:27:27	Log-Likelihood:	-95.307
converged:	True	LL-Null:	-146.25
		LLR p-value:	3.896e-21

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
Yr_kWh	3.514e-07	2.91e-07	1.207	0.228	-2.19e-07	9.22e-07
LF	1.8311	0.762	2.403	0.016	0.338	3.324
BldgArea	0.0001	2.36e-05	4.486	0.000	5.97e-05	0.000
NumFloors	-1.3127	0.277	-4.746	0.000	-1.855	-0.771
BldgDepth	-0.0117	0.003	-4.089	0.000	-0.017	-0.006

Applying this new model to my test set for cross validation yields a very satisfactory AUC score of .91:



NEXT STEPS AND IMPLEMENTATION

Constructing this model was a good learning experience, but in doing so I identified a number of areas for further exploration that will yield a more robust model: 1.) a deeper question that warrants exploration is: am I construction an adoption prediction or an installation detector? Maybe the model is only picking up indicators of existing installations in usage characteristics. Timing needs to be explored. 2.) explore further improvements from trying Naive Bayes or regularization; 3.) Try using weighted data if subject matter expertise dictates a field should be more significant; 4.) Use dummy variables and binning to incorporate higher-level location data as the current model incorporates no locational parameters; 5.) build a more complete training set using other databases, this training set

includes only 156 out of 1,400 customers who have installed solar; 6.) after a solar PV model is perfected I would like to begin work on one for CHP using similar methodology.

Ultimately I would like to refine the model, present it to internal management, and have it incorporated in to our customer information and mapping databases for assessment and program development.