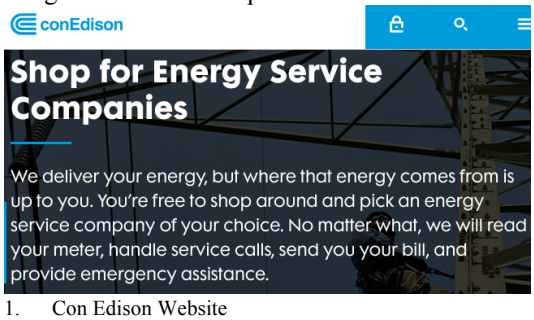
Predicting User Switches to Renewables in Age of Consumer Energy Choice

Michael Niamehr, Masuma Mansur

***Abstract*—Simple and K-Fold logistic regression methods are applied to a case of consumer renewable energy choice. It is possible to predict customer switches to renewable ESCOs based on six features ranging from energy prices to political affiliation – at an accuracy 70% for both methods, though it is known heuristically that the K-Fold interpretation will yield more accurate results on larger datasets and that much more than 70% accuracy can be achieved using target data that is natural, not synthesized. The limiting factor for this paper is dataset availability, but the concepts covered remain valid. With more data can come a long-term study on individual consumers' choices. This paper raises the idea that solar penetration's best bet could be boosting consumer choice for renewables.**

I. Introduction

THE smart grids of tomorrow will likely have renewable electricity generation deeply ingrained into local and national power systems, but even the grids of today allow for some flexibility in generation types. More and more, utility companies like Con Edison allow their customers to choose the sourcing of their electricity. In some cases this means simply paying a premium cost for the utility to purchase the equivalent of their monthly energy usage (or in some cases only 2% of it) from a 'green' generation company. However, advances in mandated competition in the electricity market have led to Con Edison and other utilities like them paving the way for "Consumer Energy Choice." Utility customers can select an Energy Service Company (ESCO) to be their sole provider of electricity. There is a plethora of 'green' energy providers boasting tremendous sales and theoretical reductions in CO_2 emissions from customers switching to their services; however, what exactly causes a customer to switch to renewable energy, and at when does the cost premium outweigh the desire to help save the environment?



1. Con Edison Website

The key issue with modeling such an uncertain prediction lies in the fact that customer motivations are not always entirely logical the way all linearized models are. One can assume that when a user is offered renewable energy at the same cost as non-renewable energy, *of course*, the user is going to select renewable, but can that be said for certain? No. For the purposes of this study, that assumption will be taken for granted along with a few other assumptions, but for large-scale applications of this prediction, many, many model class features are required to accurately predict a customer's behavior. The models explored in this paper consider a binary approach to customer choice. An essay by Markus Peters under the compilation "Machine Learning Algorithms for Smart Electricity Markets" provides the inspiration for this approach: while this considers the technical feasibility of modeling such a decision prediction, it does not – nor does any current public research – provide a thorough model nor applied features for selecting a model. At a time when nearly every institution and corporation involved with the electricity industry has a stake in increasing renewable energy penetrations, concretely determining the threshold at which users of different backgrounds will switch to renewable will prove to be a phenomenal boost to penetration levels. An accurate large-scale model could prove to be the perfect method to setting goals for renewable affordability. For example, if the results yield that on average the price difference between renewable and non-renewable is sufficiently small, might there be other motivations for costumers to stick with non-renewable?

The infamous duck curve shows the dangerous side of renewable penetration: during peak solar production hours the net demand post- renewable-depletion is at a significant spike that must be accounted for by ramped-down-and-then-up variable generators; however, accurate solar penetration prediction on a scale of years can mitigate these stresses on the grid, and ultimately by increasing renewable injections – motivated by an increase of customers paying premiums for solar energy – the duck curve will both drop and, due to high predictability of these injections, will inevitably flatten. Some consider the promotion of microgrid technology to be the most promising boost to solar penetration and reduction of reserve requirements; however, it can be argued that, due to the economies of scale, the already in-place large solar farms offer a more promising solar injection from an environmental point of view. These large-scale farms cannot operate symbiotically like they do in small-scale microgrids, they must be motivated by consumer premiums, which this paper concludes are feasibly minimal.

II. Model

This paper takes a machine learning lens to the problem at hand: how might we train a machine learning model to predict the threshold for which a user will refuse renewables under certain environmental and dispositional circumstances? Machine learning is the process by which an algorithm is generated automatically using some selection of 'training' data points and 'test' data points. This algorithm applies 'weights' to each feature considered. For example, supplying a machine learning package with a lot of data on the weight of mailed shipping boxes along with their heights, widths, and depths may yield a correlation between the weight and each of the three dimensions that differs from dimension to dimension across all boxes; the algorithm determined may be that the weight of the boxes follows a trend suggesting that it is more or less equivalent to two times the height plus three times the width plus 5 times the depth. Of course, these features are overly simplified.

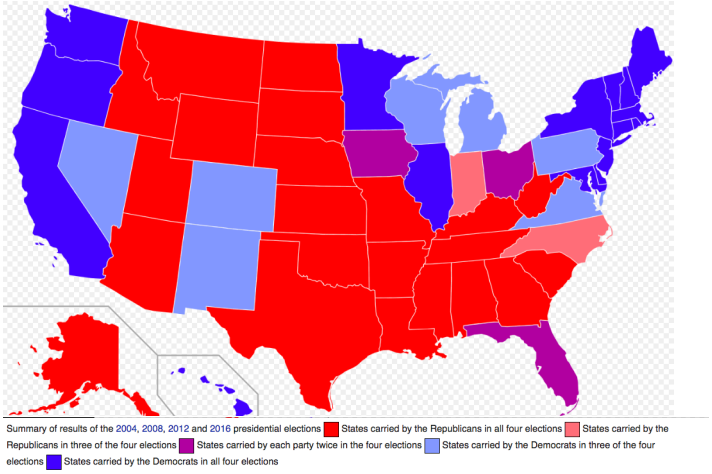
For this paper, the regression packages are brought into python along with a few packages supplementing them:

```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import linear_model, preprocessing
from sklearn.model_selection import KFold
from sklearn.metrics import precision_recall_fscore_support
```

1. Python Jupyter Notebook

The need for a variety of features – classes like the height, width, and depth of the aforementioned example – comes from the fact that while there is a direct relationship between the cost difference (renewable minus non-renewable) of the electricity and the decision of a user to stick with non-renewable energy, this is not a 100% correlation; there must be other factors involved – these factors are divided proprietarily into a set of natural logistics and a set of morality markers. While a more robust study may be able to survey and psycho-analyze the motives of individuals in regards to the environment with a sufficient number of people, this study does not. Instead, morality is simplified by the socioeconomic principles of income and political leaning on a larger scale of state averages.

The data in this model is taken for the 50 states of the US and comes in 7 forms, from a variety of sources. First is net generation of solar energy per state, per month of 2014 to 2018 aggregated. Next is average retail price of electricity overall, per state, per month of 2014 to 2018 aggregated. Average retail price of solar-sourced electricity; with limited data on nationwide consumer choice options, this can be obtained by modifying the previous data with a 1.4x multiplier in cold months, 1.5x multiplier in warm months, and 1.6x multiplier in hot months. This 1.5x average multiplier is a value conceived based on local NYC price comparisons and does not represent what could be a much more robust look into price comparisons. However, for the purpose of this model, the multiplier method suffices. Also taken into account in this model is average monthly temperature. Finally, the socioeconomic features which set to describe a user's moral motivations comes in two forms. One is a weight assigned to statewide political affiliations based on ratio of republican candidates voted into election in the past 4 presidential elections, aggregated uniquely from the following heat map:



2. https://en.wikipedia.org/wiki/File:Red_state,_blue_state.svg

25 (Red on map). = republican voted in 4 of 4 elections
50 (Pink on map).. = republican voted in 3 of 4 elections
75 (Purple on map) = republican voted in 2 of 4 elections
100 (Cyan on map). = republican voted in 1 of 4 elections
125 (Blue on map). = republican voted in 0 of 4 elections

The motivation for this feature and this weighting system comes from a study by Pew Research Center (a non-partisan study) that found that Democratic/Democratic-leaning individuals were consistently, over the course of the last 6 years, almost twice as likely to prioritize alternative energy sources over their Republican peers. Therefore, moral obligation to supporting renewables can at least in part be modeled by political affiliation. Along with this feature, some method for scaling the price differences of non-renewable versus renewable according to individual socioeconomics must be adopted. Here the selected feature is median income in each state.

What is the shipping box weight – the target variable – must be selected for this model as well. In this scenario, the target variable is a binary decision value where 0 means the user does not switch to renewable electricity and 1 means the user does switch to renewable electricity. With these selections in place, there are 6 possible classes or features for each monthly data point of the target variable. Therefore, the datasets must come in the shape of # of months of data for 50 states (#months rows by 50 columns), but in practice, the classifier model calls for one column of concatenated data from all feature datasets.

This model takes into account the fact that there is fairly limited public data to supply a machine learning training set. Using the method of class selection, one could automatically select the highest-correlated features, therefore allowing the model to include thousands of features that would only increase solution accuracy. However, even these six features have been painstakingly cultivated for the limited data out

there. The inclusion of additional features like consumer age may contribute to better guesses, but due to privacy concerns and proprietarily gathered data in companies, these features are the best of what is out there.

III. Solution Technique

In practice, the data used for this model training comes from disappointingly divergent sources. Net generation of solar energy (in gigawatthours) per state, per month of 2014 to 2018 is aggregated uniquely from data found on http://eia.gov. Average retail price of electricity (in cents per kilowatthour) overall, per state, per month of 2014 to 2018 is aggregated uniquely from data found on http://eia.gov. Average retail price of solar-sourced electricity (in cents per kilowatthour) is obtained by modifying the previous data with the aforementioned 1.4x, 1.5x, and 1.6x multipliers. Average monthly temperature (in degrees Fahrenheit) is duplicated each year as historical data is not available per state and is aggregated uniquely from individual datasets from each state from https://ncdc.noaa.gov. The political affiliation weights are determined arbitrarily as a scale from 25 to 125 and are applied to each state and copied uniformly to each month and year based on the heat map. Median household income per state is duplicated for each month and year and is collected from https://www.kff.org/other/state-indicator/median-annual-income/. Finally, the target binary decision variable is generated by viewing a heat-map of relative features for each state and performing logical guesses for each user response:

	AL		TEMP
INCOME	47221	Jan-14	37.3
POLITICS	25	Feb-14	47.3
TEMP	79.5	Mar-14	52.3
		Apr-14	62.4

3. Sample of the heat map used, created in excel by applying color sampling from maximum to minimum values of income, politics, and temperature over all 50 states (Alabama shown) as well as the temperature series per state over all months.

The above 'heat map' was simply the reference point for making the logical guesses. A more robust study could survey costumers in each state to conclusively determine if they have switched to renewables, since decisions by costumers is not always logical.

	AL	AK	AZ	AR	CA	CO	CT	DE	FL	GA	...
2014-01-01	0	0	0	0	0	0	0	0	0	0	...
2014-02-01	0	0	0	0	1	0	0	0	0	0	...
2014-03-01	0	0	0	0	1	1	1	0	0	0	...
2014-04-01	0	0	0	0	1	1	1	0	0	0	...
2014-05-01	0	0	0	0	1	1	1	1	1	0	...

4. Sample of the target variable per state when loaded into python's pandas package as a dataframe.

To begin processing the data, once all the aforementioned packages are loaded into python, the data must be loaded in as well. For this model, the data is loaded in as a dataframe object, from the pandas package built into python. These dataframes are created by reading an attached excel spreadsheet which shows the 57 months data has been collected for as 57 rows and the 50 states for which data points are collected.

```
sol_gen = pd.read_excel('SOLAR_GEN.xlsx', index_col=0)
sol_gen1 = sol_gen.fillna(sol_gen.mean())
```

```
all_prices = pd.read_excel('PRICES.xlsx', index_col=0)
all_prices1 = all_prices.fillna(all_prices.mean())

sol_prices = pd.read_excel('SOLAR_PRICES.xlsx', index_col=0)
sol_prices1 = sol_prices.fillna(sol_prices.mean())

temperature = pd.read_excel('TEMPERATURE.xlsx', index_col=0)
temperature1 = temperature.fillna(temperature.mean())

politics = pd.read_excel('POLITICS.xlsx', index_col=0)
politics1 = politics.fillna(politics.mean())

income = pd.read_excel('INCOME.xlsx', index_col=0)
income1 = income.fillna(income.mean())
```

5. Sample of the target variable per state when loaded into python's pandas package as a dataframe.

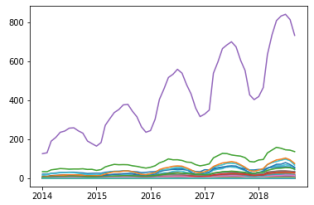
The first step of this solution is to manipulate the data to be processed correctly. pd.read_excel() loads in the data, but it does not take into account discrepancies like missing data and dimensional restraints. Before the data can be brought into the relevant regression and classification models, the data must be adjusted and manipulated to fit into the criteria necessary for smooth modeling.

	AL	AK
2014-01-01	0.02315	NaN
2014-02-01	0.02426	NaN

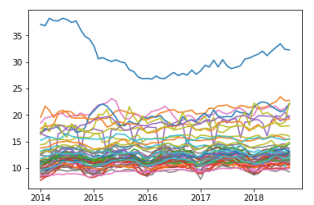
6. Sample of the missing data found in solar generation values for Arkansas.

These gaps can be remedied using many methods. The method chosen for this paper is to fill in all missing data labeled NaN (Not-a-Number) with the state's mean value. This is done simply using the fillna() command shown above. When the data is loaded in for each class and the missing data is remedied, the classes are ready to be remapping. At this stage, all data comes in the form of 57 by 50 dataframes. These must be reshaped before they can be processed; however, in this state they can be easily graphed:

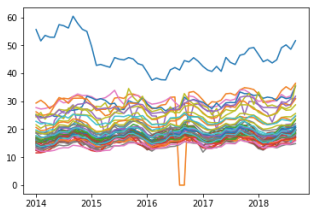
```
plt.plot(sol_gen1);
```



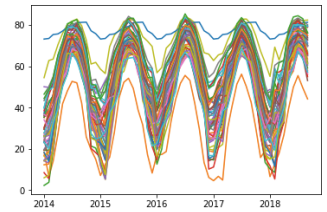

```
plt.plot(all_prices1);
```



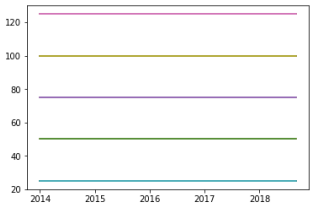
```
plt.plot(sol_prices1);
```



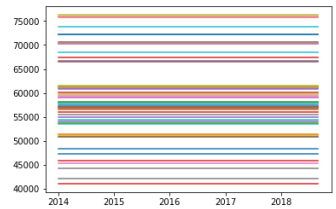
```
plt.plot(temperature1);
```



```
plt.plot(politics1);
```



```
plt.plot(income1);
```



7. Data from each feature plotted as values against years with each state representing a different color line.

Note that when viewing these graphs, a second discrepancy becomes clear – a zero-value for solar energy prices in one of the states. Similar to the NaN values in solar generation, due to imperfections in the data, there is some data for the retail prices of electricity that is shown to have a value of 0. Of course, in actuality, there can never be an electricity price of $0, so these gaps must be filled similarly:

```
sol_prices2 = sol_prices.replace(0,sol_prices.mean())
plt.plot(sol_prices2);
```

8. Data from each feature plotted as values against years with each state representing a different color line.

Note also that the plots for political affiliation and for income are flat lines at different scales. That is because these datasets were constructed by duplicating data over time. Finally, the data can be put into a form recognizable to the machine learning systems. The data directly from their sources is categorized by month and by state. Preserving this structure is not necessary for the purposes of this project because the features of politics, temperature, and income are both based on the state and more relevant data points for predicting energy choice than just the state. Anyway, to operate binary classification, the data must be restructured so that rather than 57 rows of months and 50 columns of states for each feature, there are 57*50=2850 rows of data (with 1 column) for each feature. Then the data for the 6 different features must be concatenated into a single array using the 'vstack' or vertical-stack function.

```
sol_gen1_1 = sol_gen1.values.flatten()
all_prices1_1 = sol_gen1.values.flatten()
sol_prices2_1 = sol_gen1.values.flatten()
temperature1_1 = sol_gen1.values.flatten()
politics1_1 = sol_gen1.values.flatten()
income1_1 = sol_gen1.values.flatten()

income1_1.shape
```

```
X = np.vstack([sol_gen1_1,all_prices1_1,sol_prices2_1,temperature1

X.shape
```

```
(6, 2850)
```

9. Data manipulations

This yields a 'feature matrix' that is 6 rows by 2850 columns. To make more sense of the data, the transpose is taken so that there are 2850 rows and 6 columns. The target variable is loaded and manipulated similarly. Then, finally, the data can be inputted into the classifiers. First is a simple logistic regression model stored in the linear_model package as a Logistic Regression object. The model is trained using all data points:

```
logreg = linear_model.LogisticRegression(C=1e5)
logreg.fit(X1,y1)
```

10. Logistic Regression

Stored in this 'logreg' object is the predict function which predicts the target variable using the linear regression model and trained data. Running the predicted target values and comparing them to the known target values yields the accuracy (the average similarity between the known and predicted target values) of the model:

```
yhat1 = logreg.predict(X1)
accuracy1 = np.mean(yhat1==y1)
accuracy1
```

```
0.7063157894736842
```

11. Logistic Regression Results

Accuracy is found simply with the np.mean function taking an average of the identical data points between the predicted target yhat1 and the known target y1. An accuracy of 0.706 out of 1.0 leaves a lot to be desired, but it is clear when solving a more complex, traditionally more accurate model and obtaining similar results, that this is not a shortcoming of the model type but of the quantity of data points and the quality of the synthesized target variable data that would not in practice be synthesized.

The next method for regression is K-Fold Logistic Regression. It is important to note though that in this case the training data and the test data are one and the same. Therefore, it is recommended to perform k-fold cross-validation and randomly permute sets of training and test data. This is done with the sklearn.model_selection package that can be used to run the shuffle randomly with the argument 'shuffle=True'. In this case the shuffle is run with 10-fold cross-validation. This package also allows for the convenient retrieval of the precision, recall, and f1-score – variables that are indicators to success in statistics but are not practical parameters for gauging the desirability of a model – and the error rate can be calculated by taking the average number of discrepancies between the known and the predicted target values over each fold:

```
from sklearn.model_selection import KFold
from sklearn.metrics import precision_recall_fscore_support
nfold = 10
kf = KFold(n_splits=nfold,shuffle=True)
prec = []
rec = []
f1 = []
err_rate = []

for Itr, Its in kf.split(X1):
    Xtr = X1[Itr,:]
    ytr = y1[Itr]
    Xts = X1[Its,:]
    yts = y1[Its]

    logreg.fit(Xtr, ytr)

    yhat = logreg.predict(Xts)

    preci,reci,f1i,_=precision_recall_fscore_support(yts,yhat,average='binary')

    prec.append(preci)
    rec.append(reci)
    f1.append(f1i)
    err_rate.append(np.mean(yts != yhat))

prec = np.mean(prec)
rec = np.mean(rec)
f1 = np.mean(f1)
err_mean = np.mean(err_rate)

print('Precision = {0:.4f}'.format(prec))
print('Recall = {0:.4f}'.format(rec))
print('f1 = {0:.4f}'.format(f1))
print('Error rate = {0:.4f}'.format(err_mean))
```

```
Precision = 0.6930
Recall = 0.1515
f1 = 0.2471
Error rate = 0.2940
```

```
accuracy2 = np.mean(yhat==yts)
accuracy2
```

```
0.7052631578947368
```

12. The motivation and guidelines for this code are provided by NYU Tandon's Professor Sundeep Rangan.

The next model used was Support Vector Machine. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. Different kernels were implemented for the SVM model, namely, linear, RBF and polynomial and parameters gamma and C were selected via cross validation using for loops. The selection of the parameters depend on the problem and decent performance of the SVM requires that one select these parameters carefully. Generally, one tries different values of gamma and C and selects the pair of values with the lowest test error rate. In our code, we tried to use 6 values for C and gamma as specified in the arrays C_test and gam_test. For each C and gamma in these arrays, we fit a model on the training data and measured the accuracy on the test data. Then, the C and gamma that resulted in the best accuracy was printed.

C = 10
gamma = 0.6

Kernel	Accuracy
Linear	0.7277
RBF	0.7586
Polynomial (d = 3)	0.7494

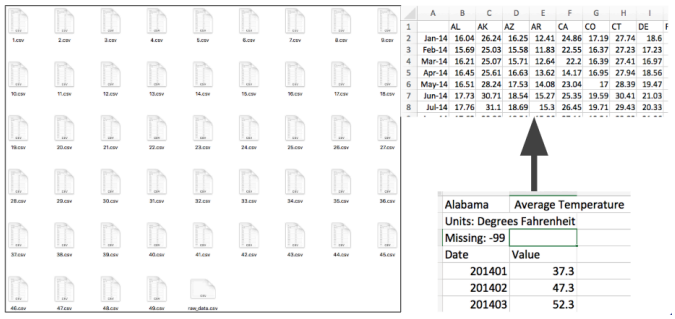
While trying different degrees for the polynomial kernel, it was noticed that the accuracy started falling after a certain degree. This could be because of over fitting. The optimal degree turned out to be 3.

In general, SVM is very good when there are many features, which is not the case in our dataset, but it shows that this model can be used when this problem is extended in the future on the availability of more relevant features.

Note again that due to the limitations in the data and the synthetic nature of the known target values, accuracy is not drastically affected by model choice. However, if this practice is recreated for larger sample sizes, additional features, and real target data, the SVM model is much more ideal.

IV. Case Study

As mentioned before, the largest setback to this initiative is a lack of data available. The U.S. Energy Information Administration (EIA) is the biggest contributor of constructive data for this paper. The EIA "collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment" but it does not incorporate this data into relevant comparisons. Data is only plotted against time individually. And the data is extremely difficult to extract in a usable fashion. Nevertheless, it was the most reliable and usable source of data out of all the other sources and it provided a baseline of data on solar generation and overall electricity cost all in one table. The more challenging dataset to work with is that of the National Centers for Environmental Information. This data comes extremely piecemeal. Data cannot be extracted without first loading slowly a visual graph of temperature over time. Furthermore, the data can only be extracted one state at a time. Therefore, the aforementioned excel sheet could not be built automatically. A set of 50 .csv files needed to be imported individually to the master sheet and reconstructed to obtain a formatted sheet similar to that of the solar generation and retail price sheets. This process should not be necessary for datasets that pride themselves on accessibility.



13. The process by which the data reaches a usable form

The political affiliation weights in this solution type are 25, 50, 75, 100, and 125 are chosen on an arbitrary scale. While other reports might assign binary variables to each type of affiliation – in what is called one-hot coding, similar to unit commitment problems – these weights are a much easier to work with dataset than 5 individual binary variables that would result in an additional 4 features. The limitation of taking weights instead of binaries for this case study does not limit its accuracy significantly. However, on a larger set of

data, the arbitrary nature of the weights could affect accuracy. Finally, income data is gathered from the Kaiser Family Foundation, a non-profit, non-partisan organization focused on provided data on national health issues, as well as the U.S. role in global health policy for the betterment of society. The data available on this site is not on a time period. Instead the data is duplicated over all months like the political data is. While this is again not ideal for accurate data, this source was the best and most accessibly source for any such data at all. The values are representative of median annual income. Median values are ideal for such a problem because average values are subject to extremes that alter the accuracy of the prediction model and cannot be easily isolated. The annual nature of the income data is not a limited factor, however, to the accuracy. Though the data can simply be divided by 12, this was not done for the dataset because the relevance come from relative income from state to state – much like in the political affiliation case, these can be arbitrarily scaled as long as they are scaled identically.

V. Conclusion

Clearly, it is very possible to predict customer switches to renewable ESCOs, as expected. Also, while the accuracy for the models of machine learning were similar around 72%, it is known heuristically that the SVM will yield more accurate results on larger datasets and that much more than 70% accuracy can be achieved using target data that is natural, not synthesized. The limiting factor to this prediction is dataset availability, but the concept remains valid. Additionally, it heuristically clear that there are many factors other than price difference that influence consumer choice. This paper does not attempt to cover all of them, though. A more robust study may consider a larger number of features; maybe there is more to socioeconomics worth delving into like culture or age of a consumer. What is clear, though, is that much more data needs to be publicly available for further studies to take place with any increase is significance. With more data can come a long-term study on individual consumers' choices. This can come in the form of surveys or information gathered by utilities or government bodies. Most importantly, this paper poses the question of how much more would you pay for solar energy? It could just be within the limits of a nearby green generator.

VI. References

[1] https://www.kff.org/other/state-indicator/median-annual-income/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[2] https://www.ncdc.noaa.gov/cag/statewide/time-series

[3] http://www.rssweather.com/climate/Hawaii/Honolulu/

[4] https://en.wikipedia.org/wiki/File:Red_state,_blue_state.svg

[5] https://www.eia.gov/electricity/data/browser/#/topic/0?agg=1,0&fuel=8&geo=vvvvvvvvvvvvo&freq=M&start=201709&end=201809&chartindexed=0&ctype=map&ltype=pin&rtype=s&pin=&rse=0&maptype=1

[6] http://www.pewresearch.org/fact-tank/2017/01/23/two-thirds-of-americans-give-priority-to-developing-alternative-energy-over-fossil-fuels/ft_17-01-20_energypriorities_repdem/