Identifying Electricity and Disconnection Patterns in Disadvantaged Communities

Mariam Germanyan, Casey McGonigle, Ayesha Yusuf
University of California, Berkeley

ABSTRACT

We are working with East Bay Community Energy (EBCE), the local electricity supplier in Alameda County, to identify electricity and disconnection patterns in disadvantaged communities. We want to investigate the attributes and patterns of EBCE customers with energy disconnections to predict if a customer is following a certain disconnection pattern in order to prevent future disconnections from happening. Ultimately, our project's objective is to assist EBCE customers in disadvantaged communities by identifying disconnections in their earlier stages and offering assistance. By warning at-risk customers and giving them resources to avoid disconnections, there should be fewer disconnections overall.

1. INTRODUCTION

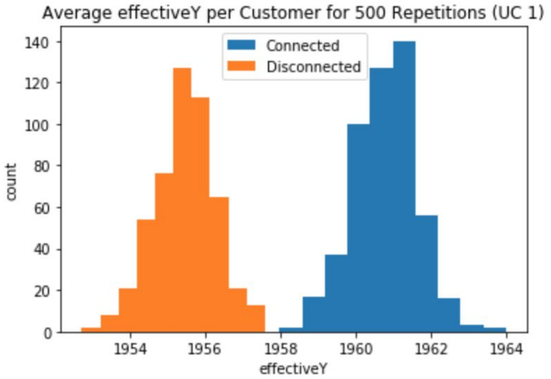
Our study covered customers from all 14 east bay cities that EBCE services: Albany, Berkeley, Dublin, Emeryville, Fremont, Hayward, Livermore, Oakland, Piedmont, San Leandro, Union City, Newark, Pleasanton, and Tracy. Using only public datasets and EBCE's own record of past disconnections, we built machine learning models to predict who is likely to disconnect in the future.

2. METHODS

The data came directly from EBCE. As a group, we worked with anonymized data from 2016-2019 consisting of approximately 97,000 connected and 33,000 disconnected EBCE customers. The data was nicely separated among disconnected customers and connected customers making it easier to work with and understand the data and its attributes. We also used parcel data (ie. square footage or the number of bathrooms in a building), census data (ie. median income or average age in a 'census tract'), and CalEnviroScreen data (ie. Poverty rate and Asthma rate for each census tract').

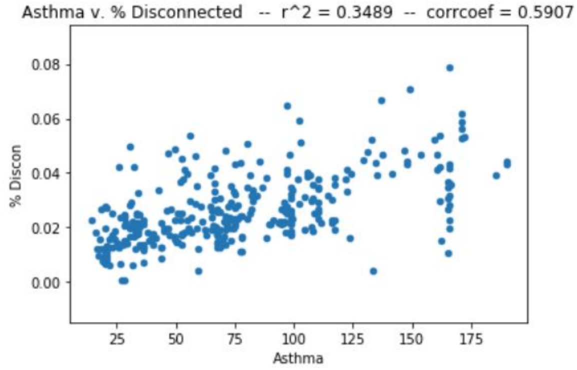
Our project involved several steps that led us to identify variables associated with disconnections; first, we explored the data to understand each feature's characteristics. We did this by creating visualizations of the features, specifically looking at differences between connected and disconnected customers.

(Figure 1)



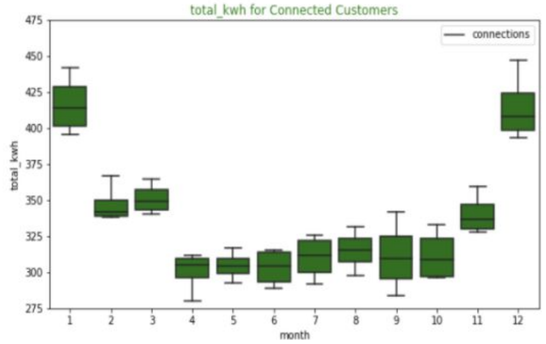
In figure 1, we notice that for customers living in single-family homes, those with a history of disconnections tend to live in older homes than customers that have never been disconnected.

(Figure 2)

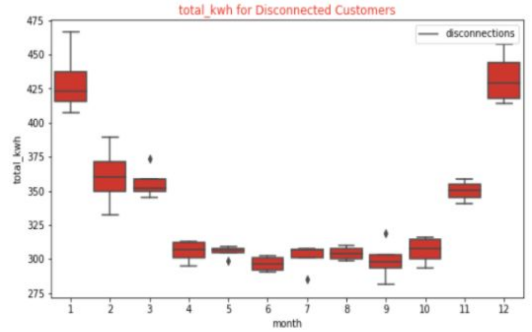


In figure 2, we plot asthma rate vs. % disconnected for each census tract. In other words, each point is a census tract. The X variable is the asthma rate for that census tract and the Y variable is the % disconnected for the census tract. We find the 2 variables are correlated with the correlation coefficient .5907. That's the highest correlation between any of our CalEnviroScreen variables and % disconnected, but it's not an exceptionally strong correlation.

(Figure 3a)



(Figure 3b)



In figure 3a, we observe that Connected Customers' variance in usage stays fairly constant from month-to-month. On the other hand in figure 3b, Disconnected Customers tend to have much lower variance in the summer than the winter.

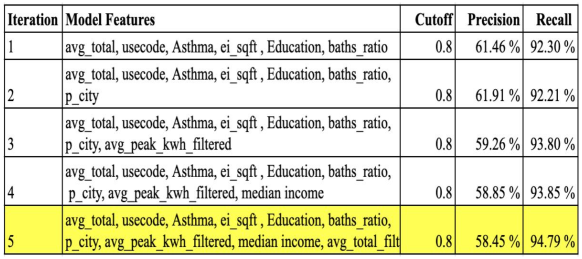
After visualizing the relationships between our features, we turned to Principal Component Analysis and Stepwise Regression to illuminate which features are the best predictors of disconnections. Once armed with many different 'iterations' of combinations of relevant features, we were ready to predict.

Finally, we built our machine learning models -- specifically Random Forests, Naive Bayes, and Logistic Regression -- to see which combinations of features, ML algorithm, and decision cutoff resulted in the best precision and recall scores.

3. RESULTS

These tables show the best precision and recall scores for each of our ML algorithms across different model features. Note: each algorithm excels using different features and (often) different decision cutoffs.

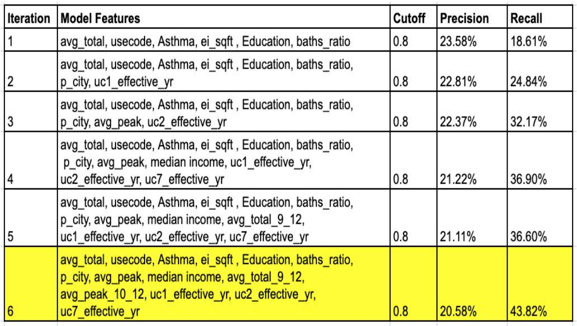
Random Forest



Iteration	Model Features	Cutoff	Precision	Recall
1	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio	0.8	61.46 %	92.30 %
2	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city	0.8	61.91 %	92.21 %
3	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak_kwh_filtered	0.8	59.26 %	93.80 %
4	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak_kwh_filtered, median income	0.8	58.85 %	93.85 %
5	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak_kwh_filtered, median income, avg_total_filt	0.8	58.45 %	94.79 %

Random Forest is easily our best overall model. Our 95% recall score indicates that 19/20 truly disconnected customers are correctly classified as disconnected. That said, this model casts a wide net: Only 58% (precision score) of the customers that are predicted disconnected are actually disconnected. We catch many non-disconnected customers in our net as well (note: this happens in all of our models).

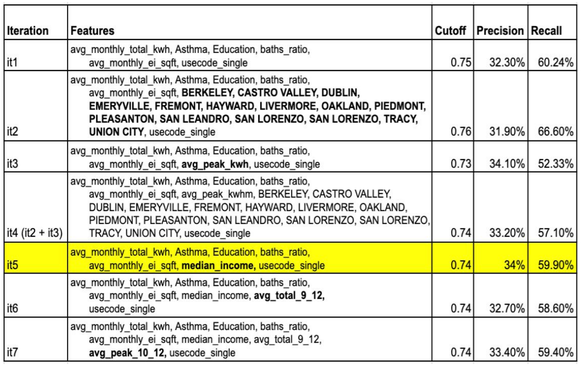
Naive Bayes



Iteration	Model Features	Cutoff	Precision	Recall
1	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio	0.8	23.58%	18.61%
2	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, uc1_effective_yr	0.8	22.81%	24.84%
3	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak, uc2_effective_yr	0.8	22.37%	32.17%
4	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak, median income, uc1_effective_yr, uc2_effective_yr, uc7_effective_yr	0.8	21.22%	36.90%
5	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak, median income, avg_total_9_12, uc1_effective_yr, uc2_effective_yr, uc7_effective_yr	0.8	21.11%	36.60%
6	avg_total, usecode, Asthma, ei_sqft , Education, baths_ratio, p_city, avg_peak, median income, avg_total_9_12, avg_peak_10_12, uc1_effective_yr, uc2_effective_yr, uc7_effective_yr	0.8	20.58%	43.82%

Naive Bayes is our poorest performing algorithm. It correctly predicts less than half of the truly disconnected customers correctly (recall). Moreover, nearly 1/5 of the predicted disconnected customers are truly non-disconnected (precision score = 20.58%).

Logistic Regression



Iteration	Features	Cutoff	Precision	Recall
it1	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, usecode_single	0.75	32.30%	60.24%
it2	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, **BERKELEY, CASTRO VALLEY, DUBLIN, EMERYVILLE, FREMONT, HAYWARD, LIVERMORE, OAKLAND, PIEDMONT, PLEASANTON, SAN LEANDRO, SAN LORENZO, SAN LORENZO, TRACY, UNION CITY**, usecode_single	0.76	31.90%	66.60%
it3	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, **avg_peak_kwh**, usecode_single	0.73	34.10%	52.33%
it4 (it2 + it3)	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, avg_peak_kwhm, BERKELEY, CASTRO VALLEY, DUBLIN, EMERYVILLE, FREMONT, HAYWARD, LIVERMORE, OAKLAND, PIEDMONT, PLEASANTON, SAN LEANDRO, SAN LORENZO, SAN LORENZO, TRACY, UNION CITY, usecode_single	0.74	33.20%	57.10%
it5	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, **median_income,** usecode_single	0.74	34%	59.90%
it6	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, median_income, **avg_total_9_12,** usecode_single	0.74	32.70%	58.60%
it7	avg_monthly_total_kwh, Asthma, Education, baths_ratio, avg_monthly_ei_sqft, median_income, avg_total_9_12, **avg_peak_10_12,** usecode_single	0.74	33.40%	59.40%

Logistic Regression is our middle-of-the-road algorithm here. It correctly predicts nearly 60% of truly disconnected customers. But again that net of predicted disconnected customers is wide; only 34% of our predicted disconnected customers are actually disconnected. The rest is wrongly-predicted non-disconnectors.

4. CONCLUSION

From the results above, we were able to get a better sense of which features differentiate disconnected customers from non-disconnected customers, which could be used to reach out to those disconnected customers to prevent future disconnections, as mentioned earlier.

Yet, there are some major areas for improvement in our models. To improve our precision and recall scores, we could explore different features that we simply didn't have the time to investigate (i.e. gas usage data or deeper dives into census data).

Moreover, our current models don't factor in time. They simply ask "is this person likely to disconnect?", not "Will this person disconnect in the next month if we don't help them first?". Adding a time component to our predictions would be a very valuable next step as well.