EAST BAY COMMUNITY ENERGY

Identifying Electricity and Disconnection Patterns in Disadvantaged Communities

Advisors: Kevin Li and Diego Ponce

Mariam Germanyan
Senior, Data Science

Casey McGonigle
Junior, Data Science

Ayesha Yusuf
Senior, Data Science

The Problem

We want to investigate 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.

Our project's objective is to assist EBCE customers in disadvantaged communities by identifying disconnections in their earlier stages and offering assistance.

Parameters

Worked with anonymized data from 2016-2019 consisting of approximately 97,000 connected and 33,000 disconnected EBCE customers.

- **Customer Parcel Data**
  - Residential attributes based on the property separated by status of connection
- **Customer Energy Usage Data**
  - Monthly customer level data identifying kwh usage
- **Public Census Data**
  - Income data based on local census data
- **Cal Enviro Screen**
  - Environmental data based on a census tract

Exploration Phase 1

Explore the data and understand each feature's characteristics.
Create visualizations of features where disconnections and connections are separated.
Investigate relationships and trends between different variables.

Kilowatt Usage and Variability lower in Summer months

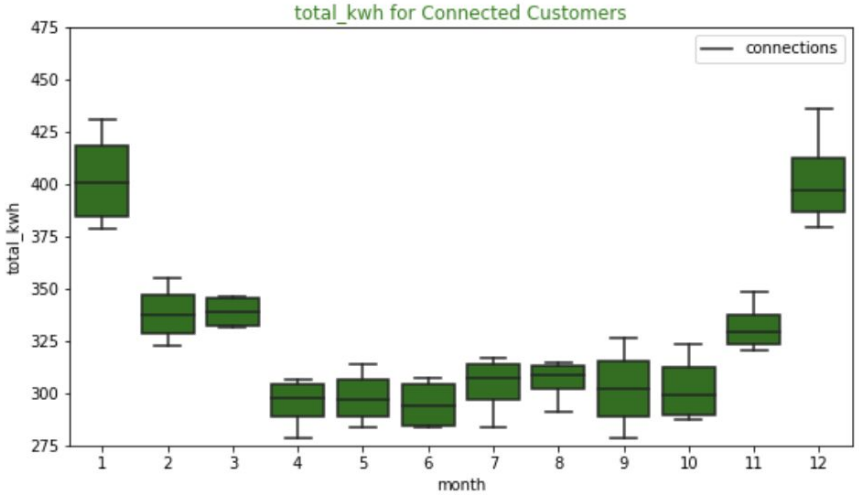


Figure 1a

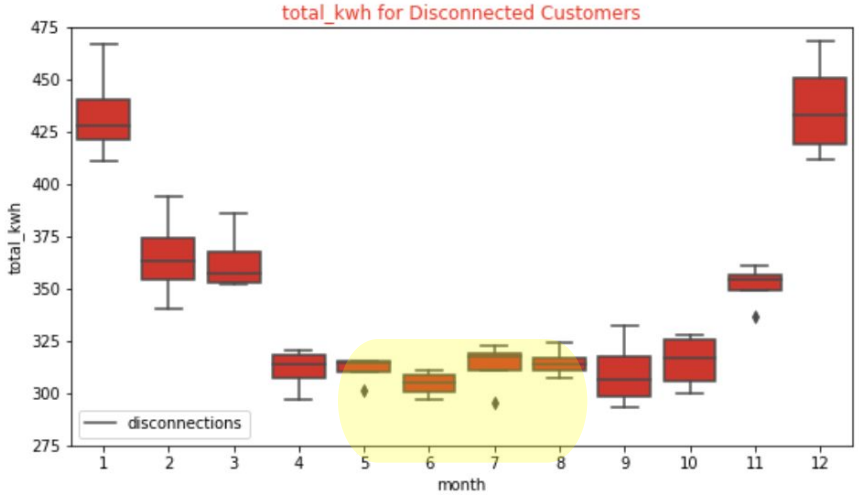


Figure 1b

- ➔ Connected Customers have similar intra-month variability across months. (Figure 1a)
- ➔ Disconnected Customers show less variability for each kwh usage in the summer months. (Figure 1b)

Average Kilowatt Usage during peak hours

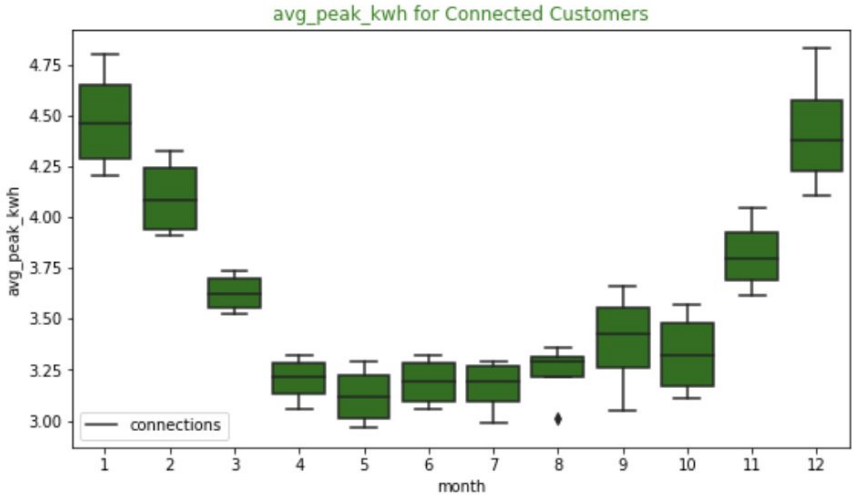


Figure 2a

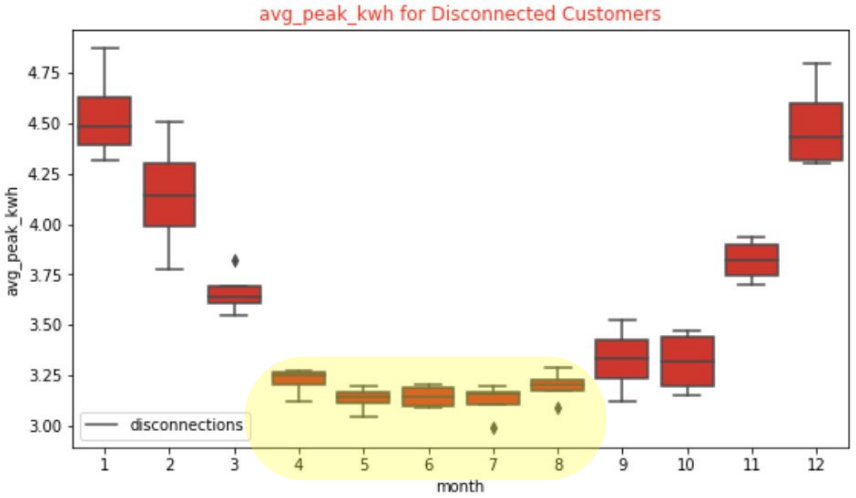


Figure 2b

➔ Avg peak kwh spread is smaller for disconnected customers with respect to the same month in connected customers during summer months.

Correlation Between our Environmental Variables

Figure 3

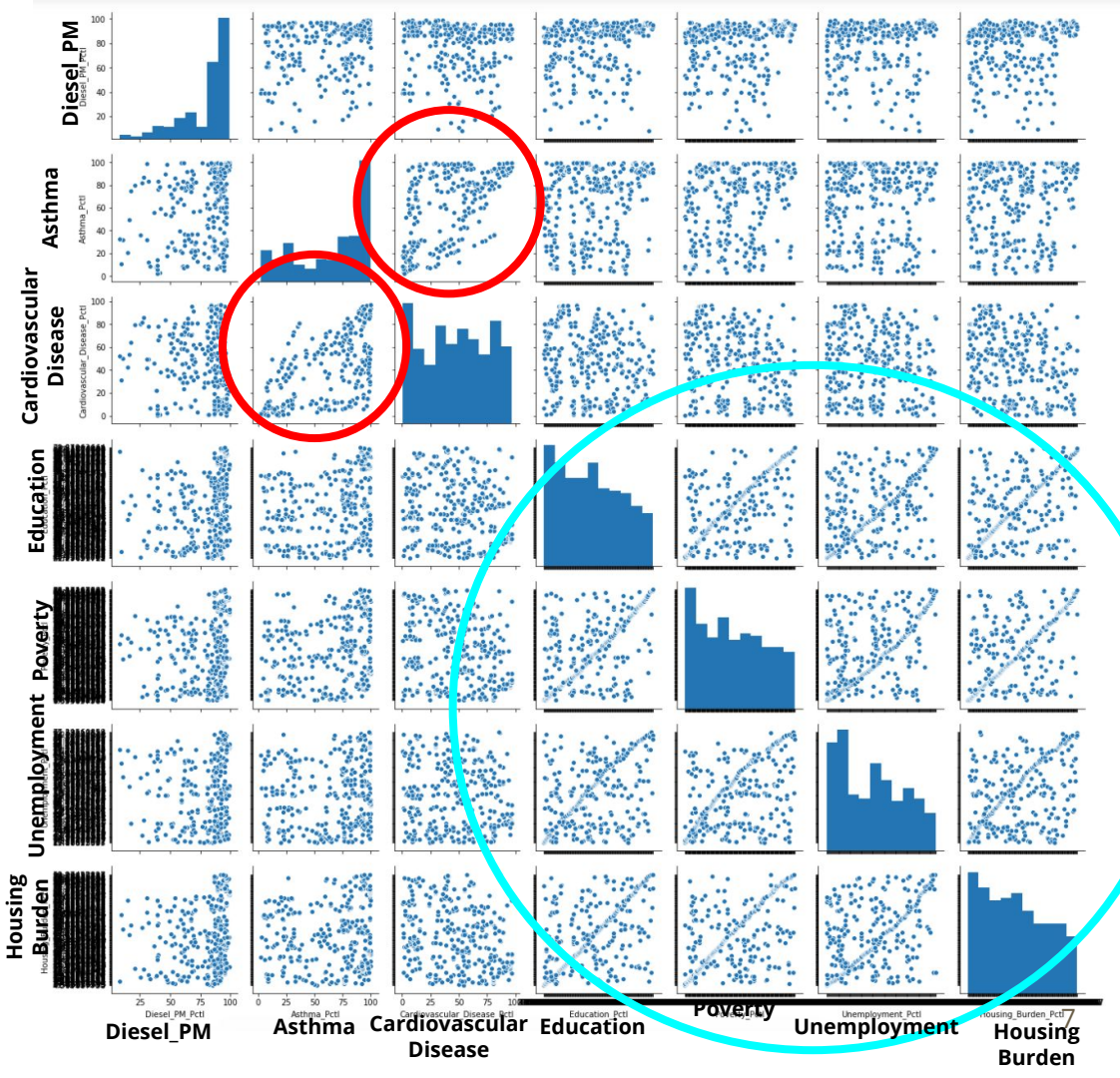
This plot helps us answer the question: "Which variables are correlated to each other?"

→ Cardiovascular Disease, Asthma (Red)
Correlation coefficient: .5018

→ Nothing else shows correlation
Though some census tracts are highly correlated (Blue) between Education, Poverty, Unemployment, Housing Burden, most census tracts show no correlation (unsure why)

→ We will only include 1 of Cardiovascular Disease and Asthma in our model b/c we only want independent features

→ Created a pairplot of 13 variables but only included 7 here for legibility



Environmental Variables are correlated with Disconnections

This top visualizations plots the Poverty Rate vs. the Percent of people that have been disconnected for a given census tract

→ in other words, each point is a census tract, the X variable is the Poverty Rate for that census tract and the Y variable is the %Disconnected for the census tract

The bottom plot does the same as the top plot, but “Asthma” replaces “Poverty”

Note: Asthma had the highest r^2 and corrcoef of any of our variables against % Disconnected. Even so, r^2 of .34899 and corrcoef of .5907 are not very high -- None of our variables are strongly correlated to disconnections

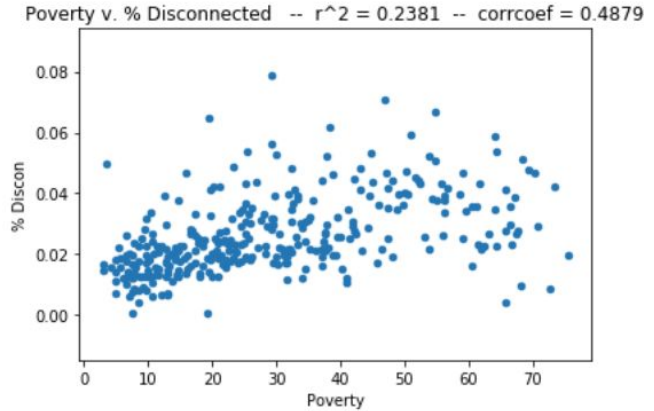


Figure 4a

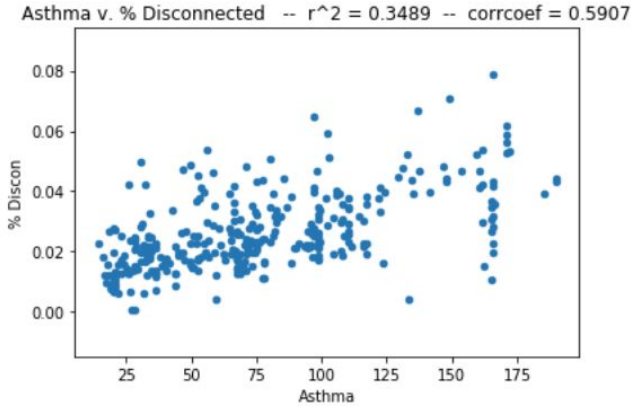


Figure 4b

Environmental Data on a City-By-City Basis

Like the last slide, this top visualization plots Poverty vs. % Disconnected. However, this time we've grouped by the cities -- that is, we've averaged all the census tracts in a given city into 1 point. Both the Poverty and % Discon are the means of their respective census tract data.

Again, the second plot does the same thing but with Asthma instead of Poverty

**note: we didn't include r^2 and corrcoef because they're artificially inflated with the averaging (which eliminates noise and gets our points closer to a linear relationship)

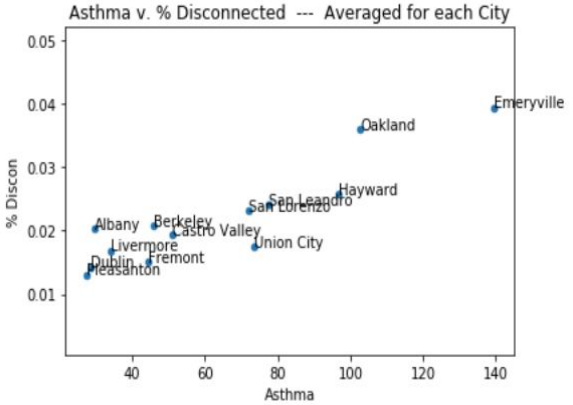


Figure 5a

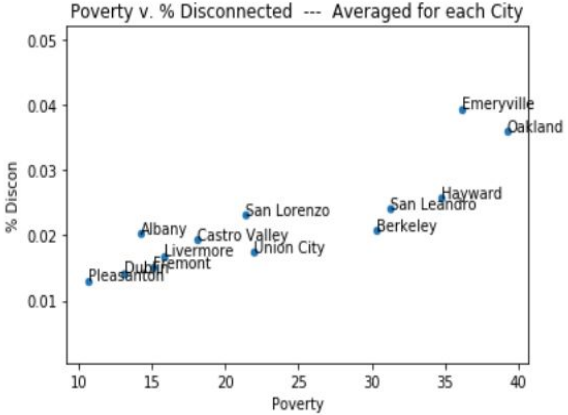


Figure 5b

Trends in Year Built Vary by Use Code

Plotted distributions by taking a random sample of customers and calculating their mean Year Built, repeating this for 500 different samples.

Use codes:

- ➔ 1: Single Family Residential Homes
- ➔ 2: Multiple Residential, 2-4 Units and Mobile Homes
- ➔ 7: Multiple Residential, 5+ Units

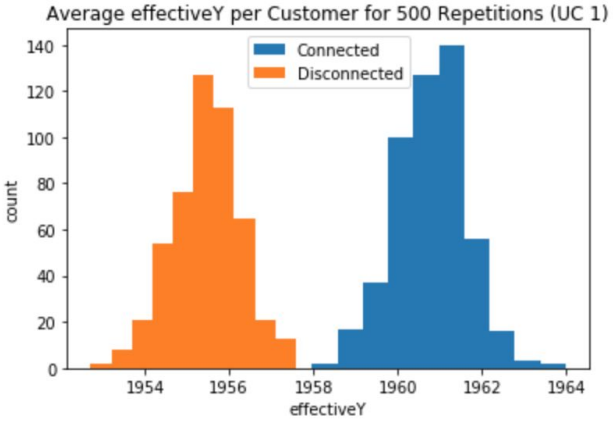


Figure 6a

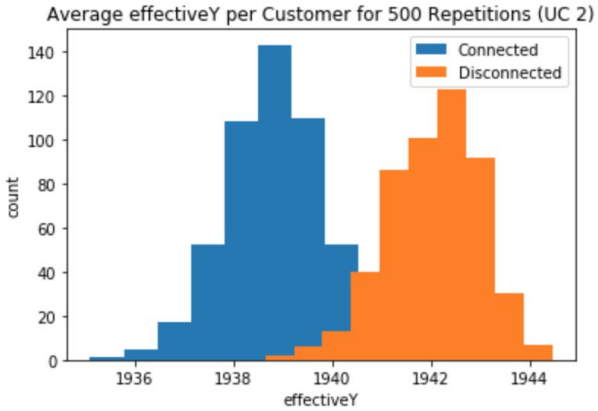


Figure 6b

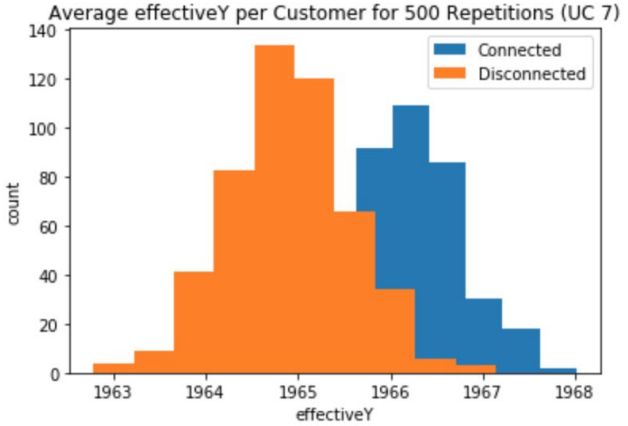


Figure 6c

Plots differ between use codes

- ➔ Use code 1 and 7: connected customers slightly tend to live in newer homes. (Figures 6a and 6c)
- ➔ Use code 2: disconnected customers live in newer homes. (Figure 6b)

Connected Customers Have Greater Average Baths per Unit

Plotted distributions by taking a random sample of customers and calculating their mean baths per unit, repeating this for 500 different samples.
*disqualified customers with 0 units, kept customers with baths per unit less than or equal to 5

➔ We see that connected customers have a higher average baths per unit than disconnected customers.

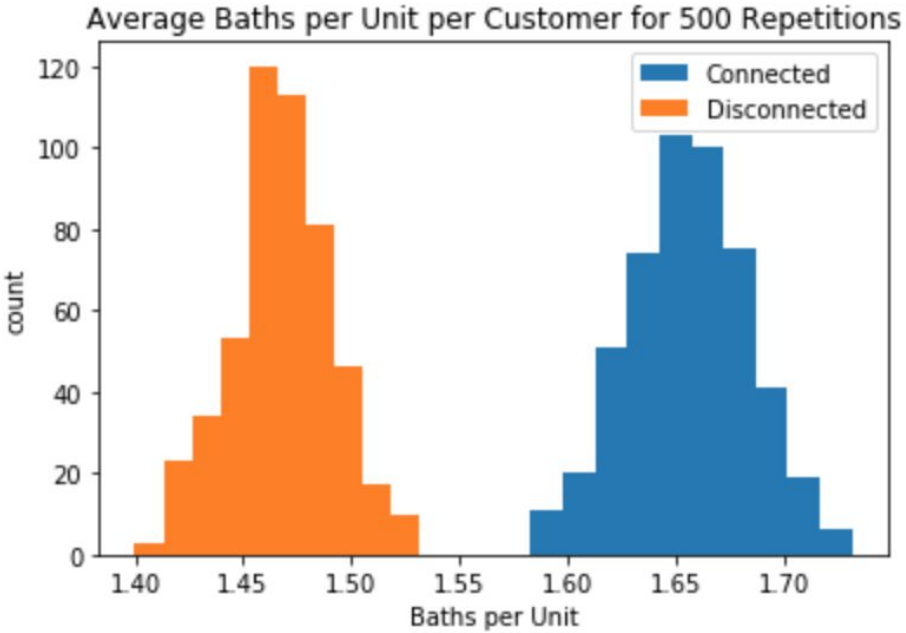


Figure 7

Exploration Phase 2

Stepwise Forward, Stepwise Backward,
Principal Component Analysis

Feature Engineering / Figuring out which variables are important

Stepwise Forward Regression

Forward Selection chooses a subset of the candidate predictor variables for the final model. Start with no variables in the model. For all predictors not in the model, compare the p-value to the threshold to determine if they are added to the model.

- ➔ Validate what attributes were useful for our machine learning algorithms
- ➔ Effective in feature extraction for Random Forest, but had better results with a combination of the features.

RESPONSE VARIABLE	MODEL	AGGREGATE LEVEL	PSEUDO R^2	THRESHOLD	FEATURES (OUTPUT)
Number of Disconnections (low <= 2 /high >3)	Logit	Customer	0.02767	0.05	Asthma, median_income, total_kwh
Number of Disconnections (low <= 2 /high >3)	Probit	Customer	0.02811	0.05	Asthma, median_income avg_daily_kwh
Number of Disconnections Per Season	Poisson	Customer Per Season	0.01181	0.05	Asthma, median_income, p_city, households
Number of Disconnections Per Season	Negative Binomial	Customer Per Season	0.008668	0.05	Asthma, median_income, p_city
Disconnections amongst all customers (binary)	Logit	Customer	0.03530	0.05	Poverty, Asthma, code, total_kwh
Disconnections amongst all customers (binary)	Probit	Customer	0.03519	0.05	Poverty, Asthma, codes, total_kwh

Stepwise Backward Regression

An automated method that can help identify candidate variables early in the model specification process. Removes independent variables one at a time using the variable's statistical significance (p-value).

Useful for finding relevant variables to use in the modeling stage.

Features it helped identify as significant: avg_daily_kwh/total_twh, Asthma, median_income

RESPONSE VARIABLE	MODEL	AGGREGATE LEVEL	PSEUDO R^2	THRESHOLD	FEATURES (OUTPUT)
Number of disconnections (low/high, binary)	Logit	Customer	0.02797	0.05/0.1	Median income, Avg daily kwh, Asthma
Number of disconnections (low/high, binary)	Probit	Customer	0.02775	0.05/0.1	Median income, Avg daily kwh, Asthma
Number of disconnections per customer per season (multi-class)	Poisson	Customer/Seasonal	0.01246	0.05	Household, Housing_Burden, Poverty, Median_income, Asthma, p_city
Number of disconnections per customer per season (multi-class)	Negative Binomial	Customer/Seasonal	0.009146	0.05	Housing_Burden, Poverty, median_income, Asthma, p_city
Disconnections amongst all customers (binary)	Logit	Customer	0.05157	0.05	Season, usecode_single, median_income, Asthma, Linguistic_Isolation, Education, total_kwh, avg_daily_kwh
Disconnections amongst all customers (binary)	Probit	Customer	0.05170	0.05	Season, usecode_single, median_income, Asthma, Linguistic_Isolation, Education, total_kw

Principal Component Analysis (PCA)

Principal Component Analysis reduces the number of variables in our model. Essentially, PCA manipulates and combines our features into “Principal Components” on a completely different scale. These are far less interpretable than our normal features, but they are guaranteed to be independent of each other and get rid of unimportant variables.

Ultimately, our PCA was not useful for 2 main reasons shown in the 2 provided graphs:

→ Scree Plot: cumulative explained variance is too spread out. In other words, we need a lot of PCs to explain variance in the data

→ Heatmap: Every variable is represented in the first 6 PCs. In other words, our PCA didn’t get rid of unimportant variables

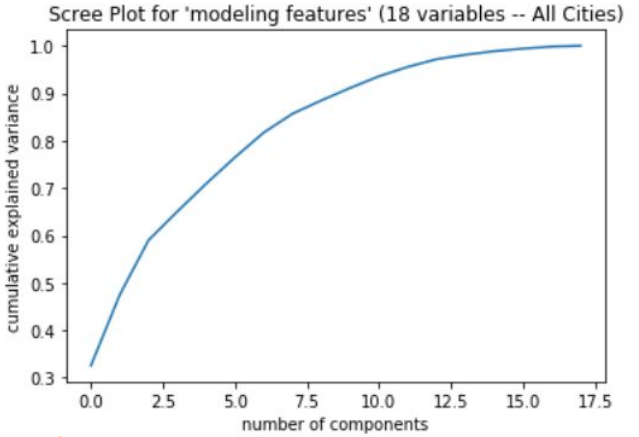
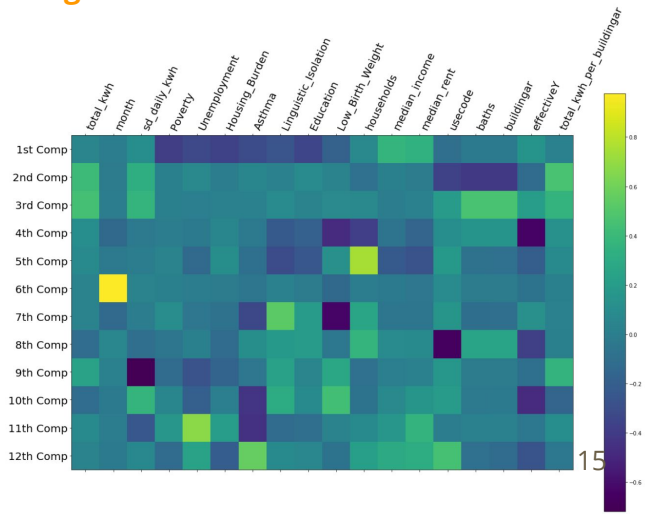


Figure 8



Modeling

Random Forest, Naive Bayes, Logistic Regression

Predicting who will be disconnected

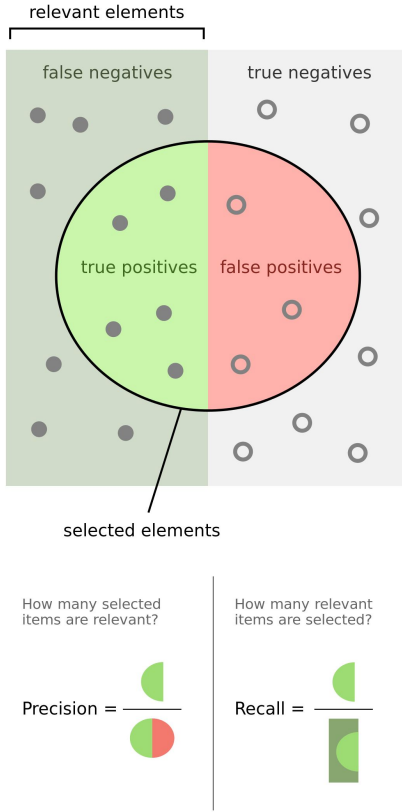
Measuring Success Rates of our Machine Learning Algorithms

Models were evaluated on unseen data. **Methodology:** For each customer, we are predicting the probability that they are disconnected based on the features we use. If that probability exceeds our cutoff, we classify them as disconnected, and non-disconnected otherwise.

To evaluate our performance, we'll use two metrics: **Precision & Recall**

Precision: True Positive / All Predicted Positive
A = # of customers predicted to be disconnected that were actually disconnected
B = # of customers predicted to be disconnected that were actually non-disconnected
Precision = A / (A + B)

Recall: True Positive / All Actual Positive
A = # of customers predicted to be disconnected that were actually disconnected
C = # of customers predicted to be non-disconnected that were actually disconnected
Recall = A / (A + C)



Random Forest

A classification algorithm consisting of many decision trees that creates an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Works its way down a decision tree depending on the cut off to reach a decision.

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 %

- ➔ Parameters of the model were finalized on what resulted in better training accuracies.
- ➔ Models were trained and tested on both standardized and non-standardized data.
- ➔ Effective for recall because we want to be able to correctly classify disconnected customers.

Naive Bayes

Naive Bayes finds the likelihood probability for each feature in each class and uses Bayes Rule to calculate which class has a higher probability. It assumes that a particular feature in a class is independent of the other features.

Parameters of NB include defining a set of prior probabilities for each class (connected/disconnected).

➔ I found that (0.6/0.4) yielded the best outcomes.

Ultimately chose iteration corresponding to best average precision/recall performance.

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%

Logistic Regression

Logistic Regression is a linear binary classifier that takes in many features and returns a 1 or 0 (yes or no)

In our project's context, logistic regression answers the question "Based on all of our features, is this a disconnected customer?"

Ultimately iteration5 (median_income + basic features) yielded our best Precision/Recall combination

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%

Summary of Results

Machine Learning Algorithms tested on 20% of the data:

- Random Forest
  - Accuracy: 89.93%, Precision: 58.45%, Recall: 94.79%
- Naive Bayes
  - Accuracy: 76.48%, Precision: 20.58%, Recall: 43.82%
- Logistic Regression
  - Accuracy: 60.20%, Precision: 34%, Recall: 59.9%

Best Features: avg total kwh usage, median income, baths per unit, Asthma, Education, usecode

Next Steps

- ➔ Explore more features we did not get a chance to explore in local census data.
- ➔ Create new features that have a clear distinction between connections and disconnections.
- ➔ Predict which customers are possible candidates with high probability of getting disconnected next month.
  - ◆ Add a time component to our predictions (predict when, not just who, will be disconnected)

Questions?