LED/NON-LED CLASSIFICATION

EXPLORATORY DATA ANALYSIS & CLASSIFER RESULTS

OBJECTIVE AND OVERVIEW

The objective of a hypothetical LED/Non-LED Classifier platform is to take the daily peaks of active power, apparent power, and power factor data from any circuit and use it determine whether the circuit is LED or incandescent.

The main motivation behind developing this classifier is to add sophistication to International Electron's lighting opportunity analyses. Currently those analyses apply universal reductions to lighting circuits when estimating savings regardless of whether the fixtures on the circuit have existing LED or not. By creating an algorithm that can discriminate between LED and non-LED circuits when running these opportunity analyses, International Electron would be able to offer services like a lighting audit through its metering system, with additional benefits of ongoing metering.

Through our exploratory data analysis (hereby referred to as EDA) and classifier trials, we have found that there is minimal correlation between data points (particularly "aggregated" power factor) associated with non-LED & LED circuits. However, if there are large clusters of granular data, some classifiers do work despite the low correlation.

DATA SOURCES

ACTIVE POWER

To form a baseline understanding, the "Active Power" readings at fifteen-minute granularity are retrieved by calling the Redaptive API endpoint.

APPARENT POWER

To find the "aggregated" power factor, which is defined as the aggregated active power readings divided by the aggregated apparent power readings, "Apparent Power" readings at fifteen-minute granularity are also retrieved through the API.

POWER FACTOR

Presented as a ratio, the "Power Factor" readings were also called at a fifteen-minute granularity from the Redaptive API endpoint. Both the mean and the median will be found across each circuit to gather insights.

SITES FOR DATA

Store	SITE #	Address	City	Stat	Zip	Start Date	Complete Da	Retrofit "Status" #1	Retrofit "Status" #2	Notes	
2641	97fe489c-de54-4753-9615-3b52d9f4d0eb	CONFIDENTIAL INFORMATION				4/19/2021	4/21/2021	A		Other submetered: dip during dates, but peak bounces back	
2621	17870e9a-5c49-4498-afd3-d9ddf18d486f					4/29/2021	5/1/2021	A		Largest consuming system is "Other" ??	
2632	4250980f-0cb1-4091-9137-cfb1c5888738					5/5/2021	5/9/2021	A	B	B here refers to "Other" Small retrofit "shifts" for lighting No clear shift; retrofit syncs with cycle Other - Sign? What does that mean?	
2619	890cc779-b74f-4ce1-ba0b-17e1f6a2fd9b					5/17/2021	5/19/2021	A		Very straightforward site. Largest consuming system is Other Submetered	? refers to notes within *Retrofit "Status"* section
2625	a3cc34b9-ffed-42f9-b923-d66b3ef5e988					5/19/2021	5/20/2021	A	B?	? - slight dip; e.g. 3.2->2.7 Another straightforward site	
2635	72587baf-8c00-43ad-bc95-2bf45d1aeeac					5/20/2021	5/28/2021	A		Largest consuming system is Other Submetered, doesn't seem to have any lighting	
2636	1edf7f39-dfca-48c8-a536-512987e928e9					5/25/2021	5/25/2021	A		Largest consuming system is Other, most likely doesn't have any lighting	
2642	f0c5242c-b36a-43a5-851e-c57bfcafda65					5/28/2021	5/29/2021	A	B?	? - slight dip; e.g. 6.7->5 Largest consuming system is Other Submetered, might have some lighting but not much impact	
2633	95c772bb-2984-4eb8-917d-978f07799f213					5/30/2021	6/3/2021	A	B?	? - slight dip; e.g. 6.8->4.5 Largest consuming system is Other Submetered, potentially has lighting circuits	

The meter readings from [COMPANY REDACTED] sites across [STATE] were sifted through to find sites that had a dip in their peaks before and after the retrofit – labeled “A” as shown in the table to the left. These sites were used for individual EDA, as well as concatenated EDA, and finally for the classification trials as well. The dates chosen for “before retrofit” were 03/08/2021 – 04/19/2021, and the “after retrofit” dates were 06/04/2021 – 07/15/2021, as during these dates all sites were firmly either before or after their retrofit, and the dates were selected such that an equal amount of data would be collected on both ends.

APPROACH

First, for the individual EDA, all the readings for active power, apparent power, and power factor respectively were parsed & grouped either by the sum, mean, or median. Then the data is set to only consider dates before and after their retrofit timeline, such that the data collected on both ends were equal (i.e. both the “before retrofit” time period and “after retrofit” time period possessed the same amount of days). Next, only the peak grouped values for each day are chosen to analyze. This analysis is done through finding the correlation coefficient, the mean and median for the “before” and “after” sets to understand how much of a dip there is on an overall level, a histogram/KDE chart, a boxplot, and scatterplot. This same process is then done with all the sites combined (what was referred to as the “concatenated EDA”).

For the classifier, two different classification algorithms are trained: one to make predictions from readings (using the “Aggregated” Power Factor measurement for reasons explained later) and another from the dummy variable “non-LED(0)/LED(1)”. 4 classifiers were used: K-Nearest Neighbors, Logistic Regression, SVM, Random Forest. Classification was done on the [STATE] [COMPANY] sites noted above with 9:1 training set:test set ratio, and then on a [COMPANY 2] site as the test set and the training set being 90% of the [STATE] [COMPANY] data.

RESULTS

INDIVIDUAL EDA RESULTS

ACTIVE POWER

Store #	SITE #	Corr.	Before - Mean	After - Mean	Before - Median	After - Median
2619	890cc779-b74f-4ce1-ba0b-17e1f6a2fd9b	-0.540171	2.009555	1.068054	0.7921	0.4039
2621	17870e9a-5c49-4498-afd3-d9ddf18d486f	-0.581693	2.882924	1.6128	1.4308	1.0975
2625	a3cc34b9-ffed-42f9-b923-d66b3ef5e988	-0.592004	4.675577	2.531088	1.1502	0.685
2632	4250980f-0cb1-4091-9137-cfb1c5888738	-0.221882	1.414569	1.043054	0.7848	0.7301
2633	95c772bb-2984-4eb8-917d-978f0799f213	-0.288754	0.303209	0.176992	-0.2395	-0.0741
2635	72587baf-8c00-43ad-bc95-2bf45d1aeeac	-0.411572	11.414327	6.606125	2.9845	2.2584
2636	1edf7f39-dfca-48c8-a536-512987e928e9	-0.19911	1.059158	0.866985	0.474	0.4589
2641	97fe489c-de54-4753-9615-3b52d9f4d0eb	-0.187766	1.578948	1.177569	0.5193	0.4262
2642	f0c5242c-b36a-43a5-851e-c97bfcafda65	-0.337513	0.871837	0.367157	0	0.0217

AGGREGATED POWER FACTOR (ACTIVE/APPARENT)

Store #	SITE #	Corr	Before - Mean	After - Mean	Before - Median	After - Median
2619	890cc779-b74f-4ce1-ba0b-17e1f6a2fd9b	-0.078519	0.559350	0.520079	0.440746	0.379455
2621	17870e9a-5c49-4498-afd3-d9ddf18d486f	-0.26624	0.575711	0.502473	0.490293	0.426832
2625	a3cc34b9-ffed-42f9-b923-d66b3ef5e988	-0.317376	0.611742	0.487229	0.471966	0.344318
2632	4250980f-0cb1-4091-9137-cfb1c5888738	-0.497757	0.730525	0.537805	0.694952	0.585294
2633	95c772bb-2984-4eb8-917d-978f0799f213	-0.017292	-0.144418	-0.014997	-0.585225	-0.267101
2635	72587baf-8c00-43ad-bc95-2bf45d1aeeac	-0.355051	0.610589	0.454933	0.567553	0.290632
2636	1edf7f39-dfca-48c8-a536-512987e928e9	-0.082104	0.900006	0.884567	0.850099	0.826500
2641	97fe489c-de54-4753-9615-3b52d9f4d0eb	-0.103536	0.337195	0.289488	0.189688	0.144247
2642	f0c5242c-b36a-43a5-851e-c97bfcafda65	-0.191906	0.273140	0.213421	-0.003159	0.043030

POWER FACTOR (MEAN)

Store #	SITE #	Corr	Before - Mean	After - Mean	Before - Median	After - Median
2619	890cc779-b74f-4ce1-ba0b-17e1f6a2fd9b	0.293791	0.221604	0.262876	0.204432	0.232271
2621	17870e9a-5c49-4498-afd3-d9ddf18d486f	-0.136483	0.415306	0.37754	0.316957	0.246329
2625	a3cc34b9-ffed-42f9-b923-d66b3ef5e988	-0.223543	0.228627	0.171301	0.073669	0.02674
2632	4250980f-0cb1-4091-9137-cfb1c5888738	-0.025024	0.413354	0.413964	0.332567	0.345608
2633	95c772bb-2984-4eb8-917d-978f0799f213	-0.139645	-0.206173	-0.29221	-0.392025	-0.427913
2635	72587baf-8c00-43ad-bc95-2bf45d1aeeac	-0.34251	0.495151	0.399504	0.349778	0.306553
2636	1edf7f39-dfca-48c8-a536-512987e928e9	-0.022988	0.680694	0.655987	0.55174	0.513333
2641	97fe489c-de54-4753-9615-3b52d9f4d0eb	0.113054	0.190713	0.206672	0.157759	0.120123
2642	f0c5242c-b36a-43a5-851e-c97bfcafda65	-0.007455	0.378389	0.407217	0.258317	0.322283

POWER FACTOR (MEDIAN)

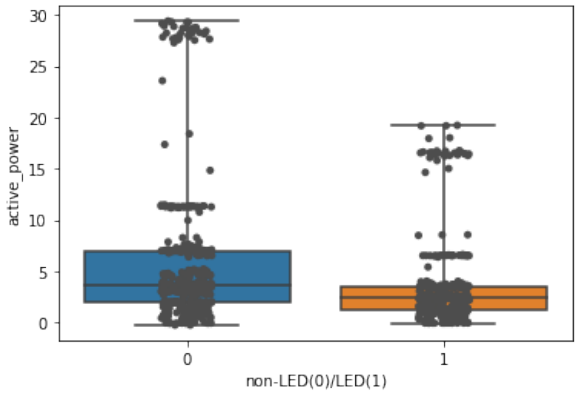
Store #	SITE #	Corr.	Before - Mean	After - Mean	Before - Median	After - Median
2619	890cc779-b74f-4ce1-ba0b-17e1f6a2fd9b	-0.129235	0.409716	0.352704	0.46335	0.336
2621	17870e9a-5c49-4498-afd3-d9ddf18d486f	-0.214649	0.43849	0.324187	0.2213	0.142
2625	a3cc34b9-ffed-42f9-b923-d66b3ef5e988	-0.189324	0.271867	0.233399	-0.068000	-0.037175
2632	4250980f-0cb1-4091-9137-cfb1c5888738	0.026152	0.566652	0.524271	0.53	0.47935
2633	95c772bb-2984-4eb8-917d-978f0799f213	-0.134058	-0.23661	-0.398458	-0.44435	-0.53965
2635	72587baf-8c00-43ad-bc95-2bf45d1aeeac	-0.319095	0.598979	0.449184	0.44	0.2813
2636	1edf7f39-dfca-48c8-a536-512987e928e9	-0.107757	0.854929	0.782763	0.8793	0.67165
2641	97fe489c-de54-4753-9615-3b52d9f4d0eb	-0.15171	0.095357	0.111523	0.08	0.11
2642	f0c5242c-b36a-43a5-851e-c97bfcafda65	-0.117823	0.592320	0.560044	0.440350	0.455850

Across the board, the correlation coefficients are of extremely low magnitude, and other than for Active Power, don't have a consistent a sign (i.e. positive or negative).

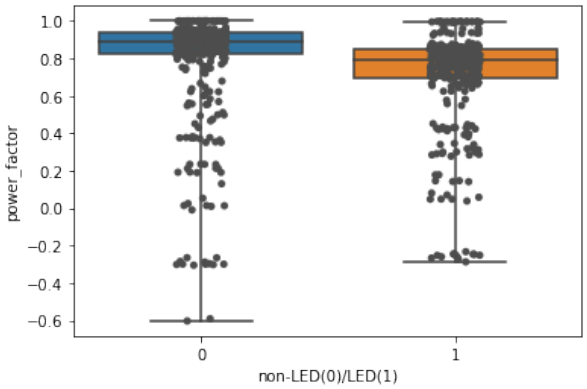
CONCATENATED EDA RESULTS

READING TYPE	CORRELATION COEFF.	BEFORE - MEAN	AFTER - MEAN	BEFORE - MEDIAN	AFTER - MEDIAN
Active Power	-0.22304	6.232404	3.518096	3.6937	2.378
Agg. Power Factor	-0.145739	0.787061	0.703717	0.893561	0.793673
Power Factor (Mean)	-0.054813	0.536006	0.504798	0.528565	0.494577
Power Factor (Median)	-0.093252	0.679257	0.679257	0.8317	0.7753

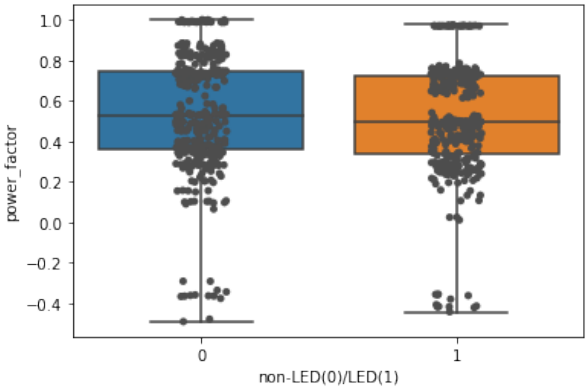
ACTIVE POWER



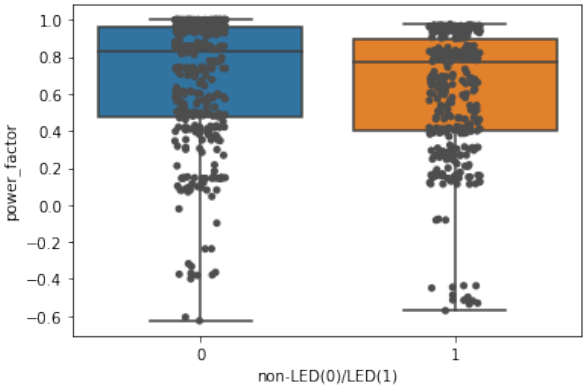
AGG. POWER FACTOR



POWER FACTOR (MEAN)



POWER FACTOR (MEDIAN)



Of the Power Factor readings, Aggregated Power Factor had the largest magnitude. However, cross the board, the correlation coefficients are of extremely low magnitude. Thus, for trial purposes, the Aggregated Power Factor readings were chosen for the classifiers, simply to see the results.

[STATE] [COMPANY] – CLASSIFICATION

Below are the confusion matrices showing the results of the classifiers for the [STATE] [COMPANY] sites:

K-Nearest Neighbors

n = 76	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	39	5	44
ACTUAL LED	4	28	32
	43	33	61/76 correct

Logistic Regression

n = 76	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	36	8	44
ACTUAL LED	21	11	32
	57	19	47/76 correct

SVM (Support Vector Machine)

n = 76	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	29	15	44
ACTUAL LED	12	20	32
	41	35	49/76 correct

Random Forest

n = 76	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	39	5	44
ACTUAL LED	4	29	32
	43	33	61/76 correct

As can be seen, the K-Nearest Neighbors & Random Forest Classifiers much better than was predicted by the low correlation scores. In contrast, the Logistic Regression & SVM classifiers performed a little higher than 50% - essentially giving coin toss results, which was our prediction due to how similar our data points on both the non-LED & LED values were (as shown in the concatenated EDA boxplot results). We believe that K-Nearest Neighbors is effective because of its ability to capture density at different points, and thus pick the correct trait despite overlapping data. Since the non-LED/LED distributions, as shown in the concatenated EDA boxplots, almost completely overlap each other, there is a lack of linearity, which leads to low magnitudes for the correlation coefficients and as well as low accuracy for the Logistic Regression & SVM classifiers as well – both of which rely on linearity for their classification. Random Forest combines many classifiers and consists of many decision trees, and thus also has an accuracy rate comparable to K-Nearest Neighbors.

The values above are subject to change dependent on the random sampling for the test set. However, the accuracy is always of a similar rate, and the accuracy rankings of these classifiers will remain the same

[COMPANY 2] – CLASSIFICATION

To verify that the high level of accuracy present in the K-Nearest Neighbors & Random Forest Classifiers was not only occurring because of circumstantial data from the [STATE] [COMPANY] sites, a [COMPANY 2] site's data was used as the test set, and 90% of the [STATE] [COMPANY] data was used as the training set – the approach for the training set is similar to the [STATE] [COMPANY] classification, but it was not the same training set due to different results when randomly sampling.

Below are the confusion matrices showing the results of the classifiers for the [COMPANY 2] Site:

K-Nearest Neighbors

n = 27	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	11	3	14
ACTUAL LED	9	4	13
	20	7	15/27 correct

Logistic Regression

n = 27	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	1	13	14
ACTUAL LED	0	13	13
	1	26	14/27 correct

SVM (Support Vector Machine)

n = 27	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	5	9	14
ACTUAL LED	3	10	13
	8	19	15/27 correct

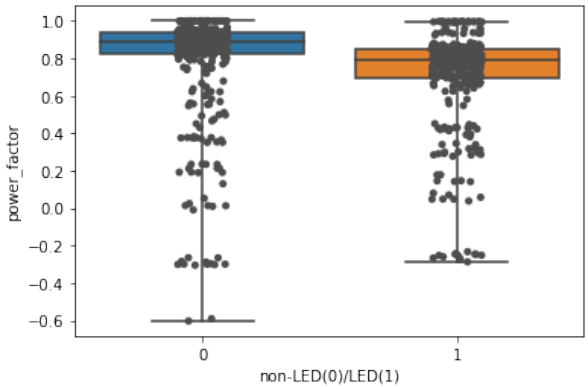
Random Forest

n = 27	PREDICTED NON-LED	PREDICTED LED	
ACTUAL NON-LED	10	4	14
ACTUAL LED	12	1	13
	22	5	11/27 correct

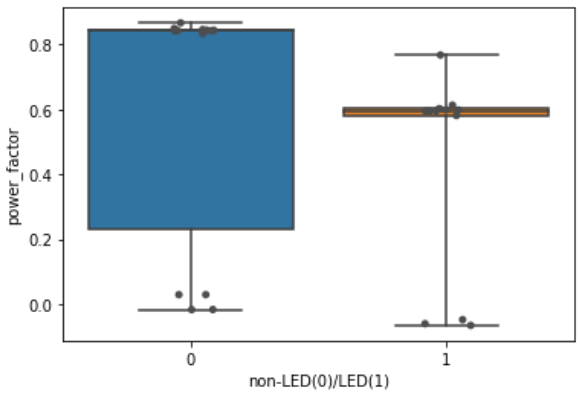
Unlike the [STATE] [COMPANY] sites, the accuracy for [COMPANY 2] was low across all classifiers. Additionally, while K-Nearest Neighbors and Random Forest predominantly predicted that the data points were non-LED, Logistic Regression and SVM largely predicted that the data points were LED, and in both cases were only correct about half of the time.

To understand why the drastic difference in accuracy occurred, we can compare the [COMPANY 2] site's distribution in comparison to the [STATE] [COMPANY]'s distribution (for Aggregated Power Factor).

[COMPANY]



[COMPANY 2]



Since the meter at the [COMPANY 2] site was only installed in May, and the retrofit occurred shortly after, there was not much "before retrofit" data to gather. Consequently, because we want balanced data sets for before and after the retrofit, our final dataset is quite small at only 27 data points. The initial hypothesis was that if the training set was large enough, the results would be accurate. However, by comparing the distributions we can see that the test sets also need to have sufficient data points for K-Nearest Neighbors to accurately "find the neighbors" and predict whether the data point was non-LED or LED. Similarly, since Random Forest utilizes multiple classifiers and decision trees, a lack of data for the test set does not allow the Random Forest classifier to perform with greater accuracy.

NEXT STEPS

The next steps moving forward to create an accurate classification system would be to:

1) Find clients outside of [COMPANY] that "pre-meter" – i.e. install an energy meter prior to the retrofit so that "before retrofit" data can be gathered for comparison
2) Find clients that have installed an energy meter and have never performed a retrofit.

The data from these sites is necessary to continue building the classification algorithms. There is an abundance of "after retrofit" data; however, to train and test the classifiers, "before retrofit" data is critical so that the classifiers are not biased towards the "after retrofit" classification.