Moving Beyond the Buzz Words – Simplified Examples Provide Context to the "Mysteries" of "Data Science

Noel Stevens TRC, Bilsay Varcin TRC, Wendy Lu TRC, and Alex Valentino TRC

ABSTRACT

Utilities historically relied on savings from large businesses, while achieving the bulk of small business savings through lighting retrofits. As lighting savings disappeared and large businesses are increasingly saturated, utilities are looking for ways to improve savings to small and medium businesses (SMB). However, the sheer number and diversity of SMB firms confounds finding high valued prospects.

Over the past decade, utility program staff, regulators, and other stakeholders have heard much chatter regarding using "big data" and "data science" to improve program performance to these Hard-to-Reach (HTR) customers. Retailers and direct marketers use tools such as journey mapping, propensity modelling, segmentation analysis, lifetime-value analysis, and persona development to develop target marketing strategies that improve profit margins 5% - 10% with far less data than utilities have at their disposal. If retailers can do it to build multi-billion-dollar industries, so can utilities. However, we need to move beyond the buzzwords to non-data scientists, who struggle to understand what all this "fuzzy math" and technical jargon can apply it to their work.

This paper provides simplified examples of data science concepts to help non- data scientists gain a basic appreciation of key data science tools. We will demonstrate how to reconstruct past program and marketing history to describe customers in terms of their recency, frequency, and magnitude of past behavior. This information provides insight into customer awareness, acceptance, and need for subsequent energy services which can be used to predict participation and isolate customers with highest expected return on marketing dollars to make programs perform more effectively.

INTRODUCTION

Historically, most utilities primarily focus on customers' annual or peak energy consumption as these parameters directly impact the utility's core source of revenue and costs. Prior to the advent of energy service offerings (e.g., Demand Side Management, Demand Response Services, Renewable Energy Services, Electric Vehicle Services, or Distributed Generation Services), use of consumption and/or demand to classify and market to customers allowed utilities to effectively understand what they needed to know about customers to sufficiently manage the demand and supply of power (electricity and or natural gas) and service core utility functions: demand and supply planning, cost analysis, rate design; and regulatory support.

As utilities seek to grow their energy service offerings, they recognize a need for greater customer intelligence to inform the integration of service offering teams with marketing, customer service, and core utility operations. While all customers need power, not all customers need energy service offerings. Program Administrators, Implementation Staff and Contractors,

and other stakeholders recognize that customer intelligence will help develop and evaluate outreach efforts by allowing them to focus outreach and marketing efforts on those customers who are aware of, procurement practices that favor, and need for energy services, while conserving program resources by not marketing to customers who do not need supplemental services, or have no propensity for participating.

This is a paradigm shift, away from a conventional regulated utility model that requires only limited customer intelligence for core utility services, to a more data-driven approach to identifying customers who need energy service offerings. In recognition of this paradigm shift, Program Administrators and Implementers may look to Data Science to help identify new or repeat customers. Advances in data processing speed and storage, advanced metering, and the ability to track web traffic and click behavior has created a lot of "Buzz" over terms like "Advanced Analytics," "Big Data," "Predictive Analytics" and more recently "Machine Learning," and "Artificial Intelligence." For the typical program staff, implementer, or evaluator, these terms probably sound like a whole lot of jargon that is unlikely to change very much about what we do to achieve our goals. Their concern is that these ideas will be the next "Big Thing" that does not improve their ability to address ever increasing goals with finite populations of customers and limited solutions to meet those goals.

If retailers can use data science to achieve corporate goals without knowing the time of day people turn on the lights, or without knowing each customer's entire transaction history across all firms in the industry, utility programs certainly could use the tools to predict participation given there are no other suppliers and we can essentially tell the time of day that someone uses their product as well as typically knowing all purchases the customer made for energy efficiency services.

In this paper, we intend to show that data science is the "real deal." We restructure 20 years of program participation (2003 – 2023) history and 3 years of monthly consumption data from Public Service Electric and Gas Long Island's (PSEG LI's) electric customers to construct new variables used to isolate repeat customers. We show how these variables serve as powerful predictors that we use to develop predictive models characterize customers in terms of their likelihood of re-engaging with PSEG LI's programs in the future. We show how to characterize customers who are more and less likely to participate using customer attributes most utilities have in their existing data. We will employ common language when discussing these concepts to take the mystery out of data science and provide the reader with more practical applications of these tools.

OBJECTIVES

This paper addresses the following research objectives:

1. Reconstruct past program and marketing history into to describe customers in terms of their recency, frequency, and magnitude (R, F, M) of past behavior.
2. Develop load profiles for each account using historical consumption data to define customers in terms of their heating, cooling, and baseloads.
3. Demonstrate how to construct predictive models using RFM and load profiles variables to isolate customers' likelihood of participating in subsequent program activities.
4. Explain data mining concepts used to characterize and describe customers according to known attributes.

JOURNEY MAPPING – RECONSTRUCTING PROGRAM HISTORY INTO [R]ECENCY, [F]REQUENCY, AND [M]AGNITUDE CAN ANSWER MANY PROCESS-EVALUATION QUESTIONS

Utility data systems are designed for revenue and regulatory reporting, forecasting, and billing. The challenge with data structured for these purposes is they do not provide insights into the customer's purchase behavior, procurement practices, awareness of service offerings, or energy needs. Each record is typically stored separate from all other transactions, so it is difficult to analyze customer behavior over time. To be useful for explaining customer behavior (past and future), we often must transform typical billing and program tracking data into a more useful format that takes the entire customer history into account.

Customers often do not make investment decisions independent of transactions they made previously. Viewing program data for each account records who installed a given measure on a certain date and their deemed savings. It does not provide any information about how often a customer participates, , and the magnitude of the project (and expenditure) made. All this information is likely needed to determine what may impact the customer's decision to invest today. These decisions are often a function of the customers energy and technology needs, budget constraints, procurement practices, energy management strategy, as well as program awareness and acceptance. In this section, we discuss three key concepts to restructuring program data that provide insights into customer awareness, acceptance, procurement practices, and energy management strategies, as well as energy and technology needs, [R]ecency, [F]requency, and [M]agnitude, or RFM.

RECENCY

Recency is arguably the most important metric derived from past program history that defines a customer's awareness, acceptance, and need for energy services. This metric measures the amount of time since a customer's most recent activity. By activity, we could mean a participation event, marketing event, or any other activity we wish to track. In marketing theory, previous customers (or participants) are considered the best prospects for future purchases. Recency implicitly accounts for a number of factors that can be difficult to capture explicitly through specific variables, such as relevance of your product or service to current business priorities or policies, mindshare (the extent to which your product or service is at the forefront of their mind), customer acceptance, and current priorities. The amount of time since the most recent participation may often impact the technology investment budget, expectation around future benefits from participating, as well as need for new technology of different types. Thus, utilities may be able to use recency to create customer segments like those that are often used in marketing analytics to segment customers due to their predictive nature:

- Active customers – Participated in the past 12 months,
- Aging customers – Participated the last 13 to 70 months (2 to 5 years),
- Inactive customers – Participated in the in the past, but 71 months ago or more,
- Prospect customers – Customers who have never participated.

FREQUENCY (CADENCE).

The frequency in which a customer engages (or participates) is a function of many things that are intrinsic to that customer such as energy needs, the presence of an energy management plan, budget constraints, and other factors unique to that customer. Some customers may often participate in a utility's programs because energy savings is important to their overall cost structure, or because technology investments impact their productivity, or because they have an energy management strategy. We can think of frequency as a strong indicator of procurement practices that favor energy efficiency and likely awareness of energy conservation as a key initiative or value.

Because the size and recency of previous investments often impacts the frequency that future investments occur, it is common to view recency and frequency metrics together. For example, a recent whole building retrofit, updated HVAC system, or entire custom process upgrade may impact the available budget for further upgrades. Alternatively, a building audit with minor prescriptive measure upgrades may have stimulated interest in further improvements. Examining the frequency in which a customer engages with their utility can help identify engaged customers versus those who are disengaged. Frequent participants may represent those with energy management concerns and plans, while those who were less frequently engaged may represent customers that could participate again if certain barriers are overcome, such as additional financing options. Combining Recency and Frequency variables provides an additional set of variables to describe customers. Examples of these variables include -

• Number of electric participation events in the past 12 months, 5 years, or lifetime
• Number of midstream electric participation events past 12 months, 5 years, or lifetime
• Number of non-lighting electric participation events past 12 months, 5 years, or lifetime
• Number of prescriptive gas participation events past 12 months, 5 years, or lifetime
• Number of custom gas participation event past 12 months, 5 years, or lifetime

MAGNITUDE

Magnitude is a measure of the size of the opportunity a customer represents that most closely ties to the "revenue" side of the "value" of the opportunity. Magnitude is a function of the total energy consumption and the number of measures installed through each participation event. For demand response program, Magnitude could measure the total peak load averted. Similar to Frequency variables, Magnitude variables, we can combine Magnitude variables with Recency variables to measure the total or average kWh saved over the past 12 months, 5 years, or lifetime. Magnitude can be used to measure the following metrics:

- Total savings from electric participation events past 12 months, 5 years, or lifetime
- Total savings from midstream electric participation events past 12 months, 5 years, or lifetime
- Total savings from non-lighting electric participation events past 12 months, 5 years, or lifetime
- Total savings from prescriptive gas participation events past 12 months, 5 years, or lifetime
- Total savings from custom gas participation event past 12 months, 5 years, or lifetime

IMPORTANCE OF SETTING AN "EVENT DATE"

To compute RFM, it is essential to first select an Event Date, which is a "point -in-time" from which all variables are constructed for the analysis. It defines the Pre- and Post-conditions for the analysis and establishes the date for which cumulative and historical information is measured (i.e., It is the reference date from which all RFM variables are computed). For recency variables, this is the date from which we count days or months to determine how long it has been since the customer last participated, and whether the customer is Active, Aging, Inactive, or a Prospect. In the sections that follow, the Event Date is necessary for defining "Pre-" and "Post-" conditions. If we want to construct a model, we will define the RFM in terms of the event date, and then identify response transactions as those occurring after the "event date."

EXAMPLE OF RFM

Table 1 provides an example of using RFM analysis to define a customer's entire participation history at a point in time. For an example of this analysis, we selected January 1, 2020 as our event date, which is the date that PSEG LI launched their Commercial heat pump program. The table presents several RFM variables for three different customers including the number of events and the savings by end use for each customer by recency period.

Table 1. Example of RFM metrics for three separate customers

Metric	Recency	Measure	Customer 1	Customer 2	Customer 3
Frequency - Participation events	Lifetime	Lighting	10	4	3
		HVAC	0	1	0
		Kitchen	0	0	1
		Other non-lighting	0	3	2
	Past 12 months	Lighting	1	0	0
		HVAC	0	0	0
		Kitchen	0	0	0
		Other non-lighting	0	0	1
	Past 13 to 70 months	Lighting	8	4	3
		HVAC	0	1	0
		Kitchen	0	0	0
		Other non-lighting	0	0	0
	Greater than 70 months	Lighting	1	0	0
		HVAC	0	0	0
		Kitchen	0	0	0
		Other non-lighting	0	3	2
Magnitude – kWh Saving	Lifetime	Lighting	54,234	20,092	127,758
		HVAC	0	59,000	0
		Kitchen	0	0	6,599
		Other non-lighting	0	45,972	447,278
	Past 12 months	Lighting	7,328	0	0
		HVAC	0	0	0
		Kitchen	0	0	6,599
		Other non-lighting	0	0	0
	Past 13 to 70 months	Lighting	44,681	20,092	127,758
		HVAC	0	59,000	0
		Kitchen	0	0	0
		Other non-lighting	0	0	0
	Greater than 70 months	Lighting	2,225	0	0
		HVAC	0	0	0
		Kitchen	0	0	0
		Other non-lighting	0	45,972	447,278

LOAD DISAGGREGATION AND CUSTOMER'S ENERGY NEEDS

Regression analysis is a statistical process that identifies whether there is a relationship between two or more variables. Regression analysis that identifies the relationship between energy consumption and weather (heating degree-days (HDD) or cooling degree-days (CDD)) has historically been used to forecast weather normalized load, or estimate weather normalized savings from DSM programs. While monthly consumption data can be useful for identifying gross change.

Figure 1 depicts the basic concept of load disaggregation at temperatures between T0heating and T0Cooling. At T0the customer has only base load (i.e. no heating or cooling load). At, T0heating they begin to turn on their heat and generate a heating load, and at T0Cooling they begin to turn on their cooling and generate a cooling load. The orange triangle represents the total heating loads, while the blue triangle and green rectangle represent the cooling and baseloads, respectively.

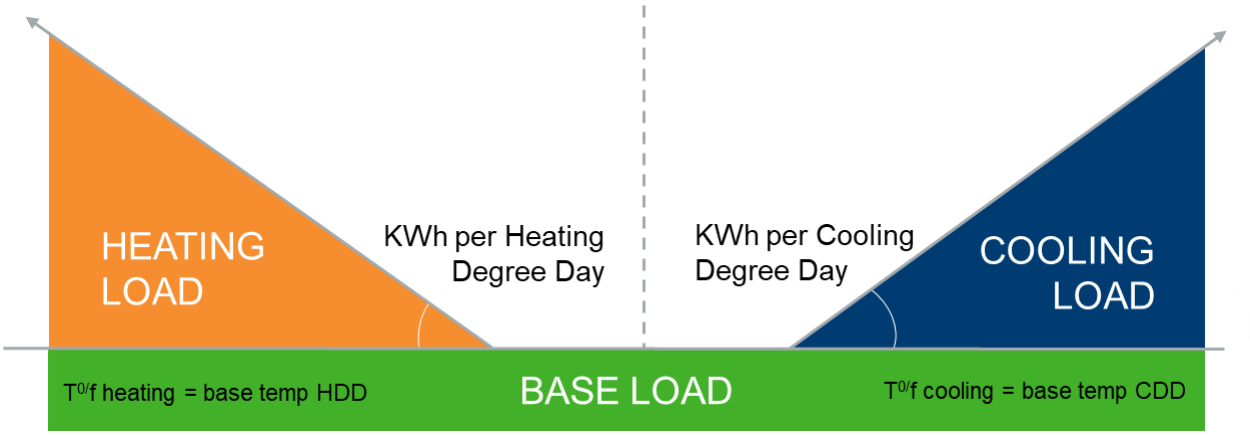

Figure 1. Visual representation of load disaggregation

Data scientist can use results from models like this to provide additional variables with explanatory power. Customers who have excessive electric or gas heating loads provide evidence that a customer may benefit from heating or cooling upgrades. Tracking those loads over time can suggest a customer's system may be malfunctioning, failing or there were facility upgrades or expansions that may provide opportunities for program support. Knowing whether a customer has electric or gas heating and or base loads sufficient to represent water heating, can suggest opportunities for fuel switching. In cold climate, absence of an electric or gas heating load and/or substantial base load may suggest the presence of delivered fuels such as oil or propane. This information can be used to identify fuel switching opportunities for electrification measures.

This information can help identify the level of customer engagement and depth of energy needs and past savings achieved. We can then potentially further calculate the lives of previously installed measures and the remaining useful life of those measures. Such analysis can help identify customers who may be in the market for supplemental measures and/or replacement of previously installed technology. It can also help identify those who may represent good candidates for heat pumps, HVAC controls, demand control ventilation, or solutions that impact temperature sensitive loads. Those with high base loads could also represent good candidates for

water heating or process equipment. Load disaggregation data can also interacted with RFM data to further describe past behavior and its impact on likely future participation.

Table 1. Example of normalized annual heating, cooling, and baseloads (kWh)

Customer	Customer size	Baseload	Heating load	Cooling load	Total load
1	Small		195,000	1670	444,800
2	Large		557,200	3,000,600	3,707,800
3	Large		0	6,930	1,387,600
4	Micro		0	820	20,900
5	Micro		13,650	220	46,000
6	Micro		120	1,320	2,500
7	Micro		11,370	0	25,150

SEGMENTATION ANALYSIS

It is often desirable to group customers with similar characteristics together into "segments." Segmentation can improve the effectiveness of program designs, outreach and marketing efforts by isolating customers with certain needs and developing targeted programs, marketing, or outreach strategies that address each segment's specific needs.

In addition to identifying customers by recency variables, we can use RFM variables to group customers according to their purchase behavior. By viewing each customer's past participation by end use, we can use various metrics to define customer segments that may depict a customer's future purchasing behavior. For example, we can identify the recency and frequency of each customer's past participation by end use, and explore the specific measure installed, project costs and savings. This information can help identify the level of customer engagement, their depth of energy needs and their past savings achieved. We can further calculate the lives of previously installed measures and the potential remaining useful life of those measures. These analyses can help identify customers who may be in the market for supplemental measures and/or replacement of previously installed technology. For example, we could define the following segments:

- Aware and Engaged – Customers who participate often and recently.
- Steady not Engaged – Customers who have had accounts for a long time and with a relatively consistent year-over-year consumption trend but limited to no program participation.
- New and Growing – Customers who recently opened their account in the past 3 years and annual consumption increased at least 10% than industry average.

DEVELOP PROPENSITY MODELS TO IDENTIFY CUSTOMERS WITH HIGHEST PROPENSITY TO PARTICIPATE

Propensity modeling is a statistical approach to identifying customers with the highest expected value in response to a given event. The event could be a change in program strategy, introduction of a new program, a marketing campaign, or any other activity that takes place at a point-in-time and where customers can be separated into distinct groups of those who did and did not experience the event. The purpose of a propensity model is to estimate the probability or likelihood that customers will do something (i.e. participate or participate in a certain program) after the Event Date. To estimate this probability, we consider 1) an Event, 2) predictor

variables that define each customer right up to the Event, and 3) a measure of Response (i.e. what you are predicting).

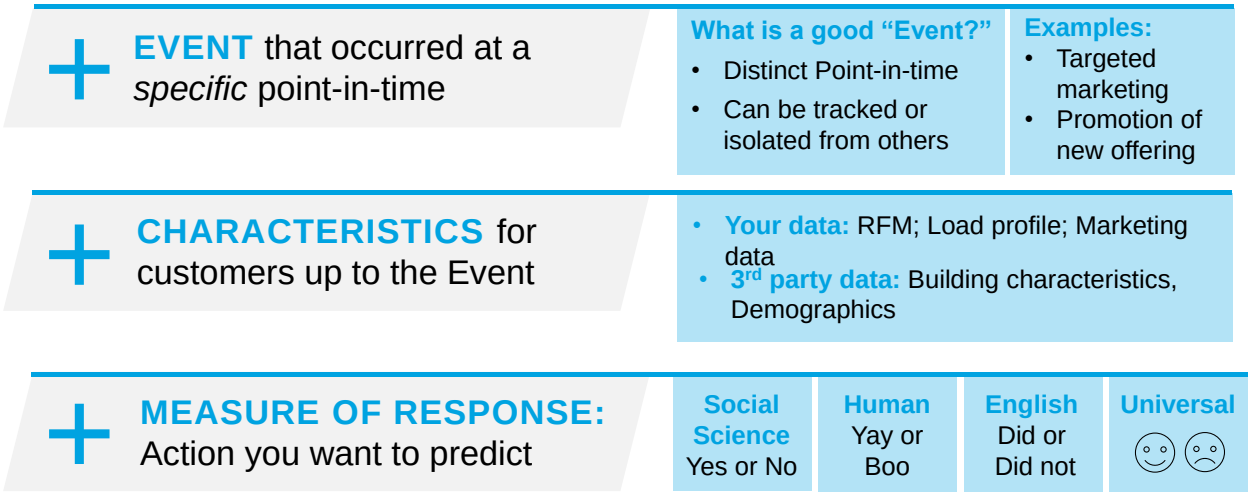

Figure 2. Requirements for a propensity model

To effectively create a propensity model, it is important that all variables used to describe the customers characterize them up to the Event Date. It is also important that all response behavior is measured after the Event Date. Figure 2 provides a simplified visual depiction of how we can use RFM and load disaggregation to develop a propensity model. These models are typically estimated using logistic regression analysis, but more recently, Data Scientists began using machine learning (ML) to estimate them as ML allows data scientists to develop and test many models to find most predictive set of variables.

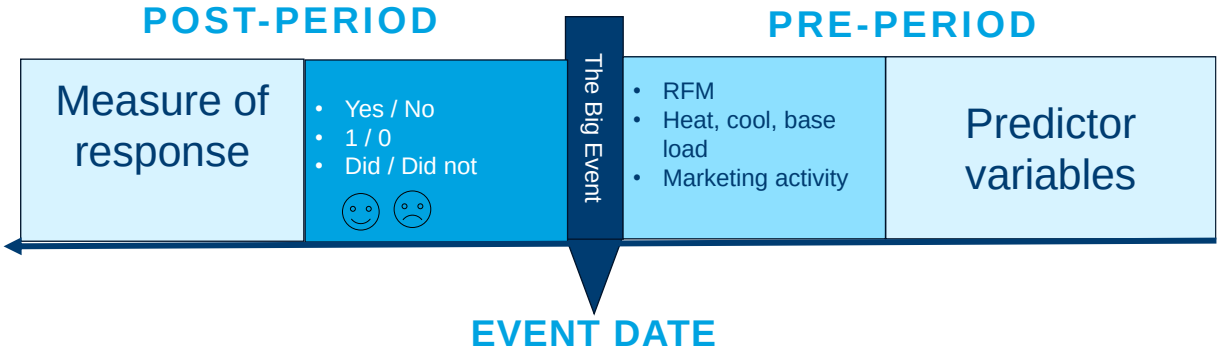

Figure 3. Visual representation of a propensity model

As seen in Table 3, TRC developed a propensity model to estimate the likelihood of customers re-engaging with PSEG LI's programs after 2020. The precise model definition is outside the scope of this paper; however, we do present the model results to demonstrate the predictive power of RFM and load disaggregation variables we discussed above. The model clearly shows the predictive power of RFM and load disaggregation variables. Moreover, we can separate customers according to their consumption size range, presence of cooling load, and whether they are in a disadvantaged community.

Table 2. Propensity model to estimate likelihood of participation in PSEG LI's Commercial programs after January 1, 2020

Coefficient	Estimate	P-Value
RFM Variables		
Active Period Variables		
Number of participations <- 12 months: Large business	1.46989783	0.0476
Number of participations <- 12 months: Micro business	-3.51653865	0.0000
Number of participations <- 12 months x log(Savings from those events)	-0.13143784	0.0247
Aging Period Variables		
Aging customer with cooling	-1.91720028	0.0000
Number of participation events 13 to 70 months: Micro business	-4.56675139	0.0000
Number of participation events 13 to 70 months for customers with cooling	1.14463849	0.0000
Number of participation events 13 to 70 months x log(Savings from those events)	-0.04527091	0.0973
Inactive Period Variables		
Number of participations > 70 months: Micro business	-4.33922962	-
Number of participations > 70 months: Small business	-0.22805221	0.0000
Consumption Variables		
Log normalized annual consumption: Large Business	0.09664578	0.0002
Log normalized annual consumption: Micro Business	0.32311213	0.0000
Log normalized annual consumption: Small Business	0.07716983	0.0079
Disadvantaged Community		
Micro-business in a disadvantaged community	-0.80184362	0.0563

After estimating this model based on historical data, we recalculated the RFM variables using the most recent date in the program tracking data as the event date. We then estimated the likelihood of customers participating in the future by applying the estimated model to the most recent RFM data. Figure 4 shows customers rank ordered by the estimated probability of participating. After rank ordering each customer, we group them into deciles and show the average probability of participating within each decile. This allows us to identify groups of customers who have a probability of participating high enough to justify marketing expenditures.

The figure shows you customers in Deciles 1 and 2 have substantially higher likelihood of participating than the remaining deciles. Customers in Deciles 8 – 10 are all less than 10% likely to reengage with the program.

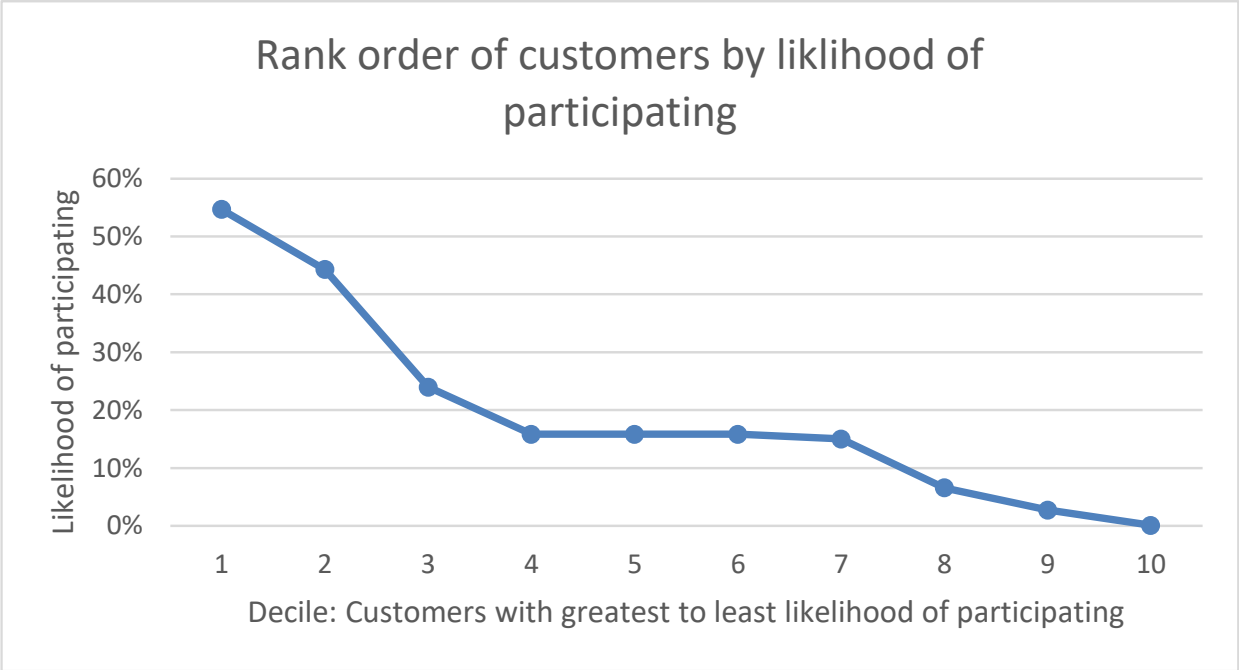

Figure 4. Rank Order of customers by likelihood of participating post November 30th, 2023

The value of this type of modeling is it allows you to identify customer attributes that characterize customers who are most likely to respond to program offerings, marketing efforts, or other forms of outreach activities. By knowing the probability of response, we can apply an average expected savings to customers in each decile. Comparing the value of expected savings to marketing costs allows for the allocation of resources to target customers for whom the value of expected savings is less than the cost to market to them. Saved funds can then be allocated to other activities that may prove more fruitful in attracting customers.

A WORD ON MACHINE LEARNING AND AI

Recently, focus of Data Science has shifted to machine learning and Artificial Intelligence. Because the purpose of this paper was to provide basic understanding of data science, we limited our analytic approaches to more simplified techniques. Our intent was to develop an understanding of how to transform existing data into variables that can be used to identify likely program participants, not the modeling techniques themselves. Both Machine

Learning and AI use similar techniques to those discussed in this paper. The key differences is these methods allow researchers to explore a far greater number of model variations than we can do using conventional model building tools.

LIMITATIONS AND FUTURE RESEARCH

This study was focused on developing predictive tools using data already available to the program team and fairly simplified modeling. This data did not include any information regarding the building characteristics of customers in PSEG LI's territory such as square footage, building use, vintage, heating, cooling, and water heating systems. The analysis also so did not include demographic information such as industry, number of employees, revenue, building or ownership. The analysis included only 3 years of consumption history, which precludes us from identifying changes to heating, cooling, and base loads that may have occurred in past participation events and limited our ability to identify whether there is evidence of system degradation prior to participation. Finally, the weather normalization models were limited to using a base temperature of 65 degrees for computing heating and cooling degree days.

TRC is in the process of expanding upon this analysis to address the limitations. First, we are incorporating information reporting building characteristics of each account in PSEG LI's territory. This will allow us to consider characteristics such as number of building units, building ownership, square footage, presence of natural gas, or delivered fuels for heating, water heating, or other uses. We also will enhance the load disaggregation to allow for variable temperature set points for heating and cooling.

Finally, we are looking to expand the number of years used for load disaggregation to explore whether there is a discernable trend in use per HDD or CDD that may indicate systems are failing. This information could allow us to isolate such buildings, and which could play an important role in predicting future participation.

SUMMARY AND CONCLUSIONS

When we move beyond all the jargon and fancy math, it should be clear that Data science is not Rocket science. Much of it is common sense. The idea behind "Big Data" is not to overwhelm people, rather it is to structure lots of things that people do to tease out patterns in their behavior that can help isolate the people most likely to participate, provided you reach them.

Through this study, we showed how to transform past participation data into new variables that reveal insights about customer's actual decision-making process. We showed how tracking a customer's recency, frequency, and magnitude of past participation can reveal their purchase patterns, provide insight into their awareness and acceptance of programs and energy efficiency. Thus, combining RFM data with load disaggregation information can provide key insights into a customer's energy needs. By combining all this information, we can potentially develop more sophisticated and informative customer segmentation strategies. We can also use data to construct propensity models that isolate customers and the characteristics of customers most likely to participate.