Engaging Cybersensitives and Cyberawares in Energy Efficiency

Produced for the California Energy Commission by Indicia Consulting in fulfillment of Task 6 Deliverable 1 of the project Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency Program Design Funded by EPIC PON 14-306

July 31, 2018

Contents

Abstract

Our goal for Task 6 in the project "Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency" is the development of critical insights for supporting residential engagement in energy efficient behaviors. In this sub-report, "Engaging Cybersensitives and Cyberawares in Energy Efficiency Part 1: Decision-tree Models," we built two decision tree models to identify cybersensitives (and other segments) within a given population. This identification will allow for the estimation of segment size within a population. We constructed an ethnographic decision tree model and a classification and regression tree model based on our collected ethnographic and quantitative datasets. We offer a preliminary estimate of the impact of cybersensitives on energy consumption. We will carry out refinement of the model using a synthetic population constructed from American Census Survey and American Time Use Survey provided by the Network Dynamics Simulation Science Laboratory at Virginia Tech University. We discuss the pros and cons of applying these methods. We anticipate that by building and sharing these models, other entities such as the California IOUs, could make use of them to better segment their audiences, and target them with appropriate programs and incentives to save energy.

1.Introduction

1.1. Goals

Our goal for Task 6 in the project "Cybernetic Research across California: Documenting Technological Adoption and Behavior Change across Diverse Geographies and Populations to Inform Energy Efficiency" is the development of critical insights for supporting residential engagement in energy efficient behaviors. The sub-set of consumers we chose to focus upon is a group we have termed 'cybersensitives' due to their responsiveness to energy information provided via device. This process is also known as 'feedback' in the parlance of social and behavioral scientists who study the cybernetic relationships between humans, their technology, and information (ACEEE, 2010). Our overall project hypothesis, which is based upon an extensive review of energy efficiency and behavior literature, is that cybersensitivity (i.e., people who are emotionally responsive to information delivered via device) is a personality trait which distinguishes its possessor from other members of demographic cohorts such as age, gender, and income strata. We assert that a focus by utilities on this sub-set would be productive. Energy savings would be improved through a combination of energy efficiency investment and behavior change, if this sub-set of consumers, or segment, were properly targeted and messaged.

Our current task objective is to recommend an alternative energy efficiency model using decision trees. Two basic types of decision tree models are ethnographic and machine learning or computational. The former builds the model from qualitative observations and questioning based on ethnographic observations, and the latter codes machine learning algorithms based on (generally) quantitative data.

> "In decision tree modeling, an empirical tree represents a segmentation of the data that is created by applying a series of simple rules. These models generate set of rules which can be used for prediction through the repetitive process of splitting." (Tso and Yao, 2005)

There are two basic types of questions which decision trees normally address: they can predict the membership of someone in a category, or they can predict the behavior of a group member with respect to certain decisions. We designed our decision-trees to predict the membership of someone in a segment, and from there allow us to extrapolate their prevalence in a given population.

Our plan is to build one of each: an ethnographic decision tree model (EDTM), and a machine learning, Classification and Regression Tree (CART) model based on the ethnographic data and survey data respectively. We will use the percentage of segment members from the cybersensitive population as predicted by the CART to estimate their impact on energy consumption. With the data collected from these efforts, and using figures drawn from the literature around behavior-based residential energy efficiency programs, the team will show how our recommended targeting of cybersensitives would provide higher rates of energy savings. Our belief is that a grounded understanding of some of the different ways people use energy will help the State of California increase the yield of negawatts, or energy not expended. This yield increase will be delivered through utility behavior-change programs.

2.Background

The PON for this round of funding required that we, "[i]dentify specific metrics for social, cultural, and behavioral factors available in population statistics that can be projected forward 5 to 10 years, and therefore included within improved energy demand forecasting models prepared by Energy Commission staff and future energy efficiency "potential and goals" studies funded by the CPUC."

In this report, we use decision tree modeling to segment individuals into categories of cybersensitivity based on different attitudes and behaviors. We believe that the characteristics that make up cybersensitivity are independent of geography and demographic variables including age, gender, and income. As such, our classification system should be durable, and applicable across California.

In this paper, we discuss the construction and application of decision trees for modeling the identification and enumeration of our segments. Researchers define decision trees as "classification systems that predict or classify future observations based on a set of decision rules" (IBM, 2012).

> "Formal decision modeling allows the researcher to, "specify the interrelationships among factors, the relative importance of each factor, and the conditions under which some factors assume precedence... (Mukhopadhyay 1984:241)." (Bauer and Wright, 1996).

Decision rules are generally a set of binary Boolean decisions or criteria, such as true/false or yes/no, which the model then branches into a hierarchical tree structure (IBM 2012). Often, researchers use decision tree models to represent a decision process, such as the car buying model example shown in Figure 1.

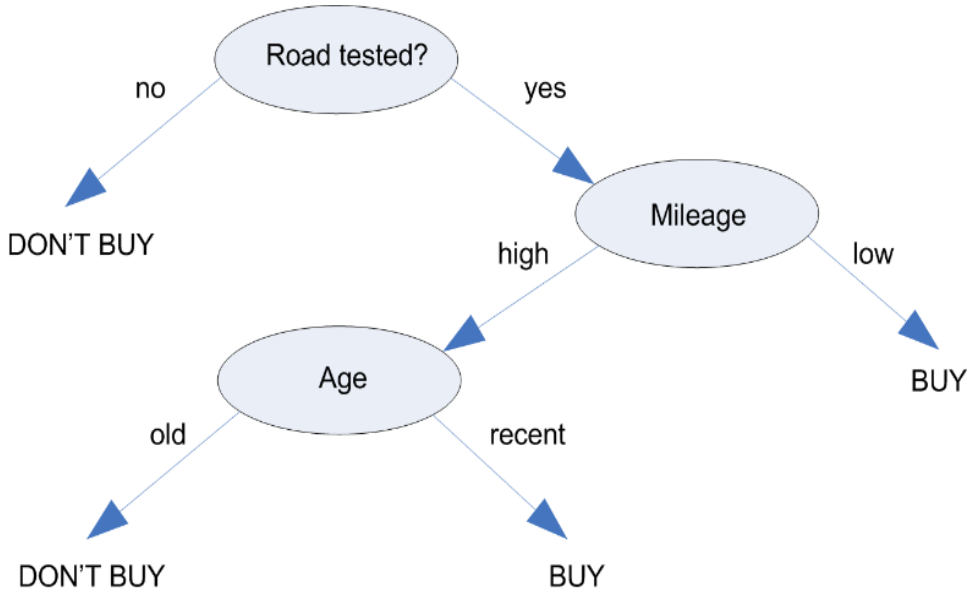


Figure 1. An example decision tree model illustrating the decisions to buy a used car (IBM, 2012).

Two basic types of classes of strategies for decision tree modeling are:

1. Ethnographic decision tree modeling on qualitative data
2. Machine learning decision tree modeling on quantitative data

Both decision tree types build a structure based on the data through a fluid set of procedures, although each utilizes different data) and thus employs similar but different protocol. Both forms of decision tree modeling are inductive since they build the model uniquely from the data. "In brief, the object of the game is to frame criteria, order or arrange them into a tree-like structure, and then test and revise the tree model" (Gladwin et al., 2001).

Most studies that model decision behavior through decision trees, whether ethnographic or machine learning in their approach, build their model on their sample data set and then test that data off a larger sample of the entire population[1] (such as Ryan 2006, Wright 1996, Bell 2018, Chapnick 1984, Murtaugh 1984). In most instances that means constructing the decision tree model based on a local set of data that they are studying and then testing that data on a national survey. Their sample data provides a space to explore in depth the features with which to build their model, and the national survey data becomes a way to make sure that their results generalize to the overall population and do not reflect specific features of that group. Here the model describes the specific data so accurately that it fails to generalize well to the overall population. In particular, decision tree models can become too large – that is, have too many branches – which can reflect considerations that are specific to that small group but do not occur for the population as a whole.[2] Testing against a larger, more general group and advanced modeling techniques (such as random forests for machine learning models) provide the primary way to address the potential for overfitting.

2.1. The Ethnographic Decision Tree Model

For this project we used ethnography to provide the rules that form the basis for the ethnographic decision tree model we report on in this paper. We will be using these rules to refine our assumptions about consumer behavior around electricity in California:

> "The method is called ethnographic decision tree modeling because it uses ethnographic fieldwork techniques to elicit from the decision-makers themselves their decision criteria, which are then combined in the form of a decision tree, table, flowchart, or set of 'if-then rules' or 'expert systems ‘which can be programmed on the computer." (Gladwin, 1989).

EDTMs, "are qualitative causal analyses that predict real, episodic behaviors, rather than—as does so much social research—the intent to behave in a certain way." (Ryan and Bernard, 2006). Researchers analyze and consolidate multiple informants’ explanations of their decision-making processes into an overall decision tree model. "There are direct and indirect eliciting methods, but both require the ethnographic model builder to look for contrasts in decision behavior, ask the informant to explain the contrast (e.g., ‘Why did you decide to evacuate with Hurricane Andrew but not with Hurricane Erin?’) and then test that explanation on another informant" (Gladwin et al., 2001).

---

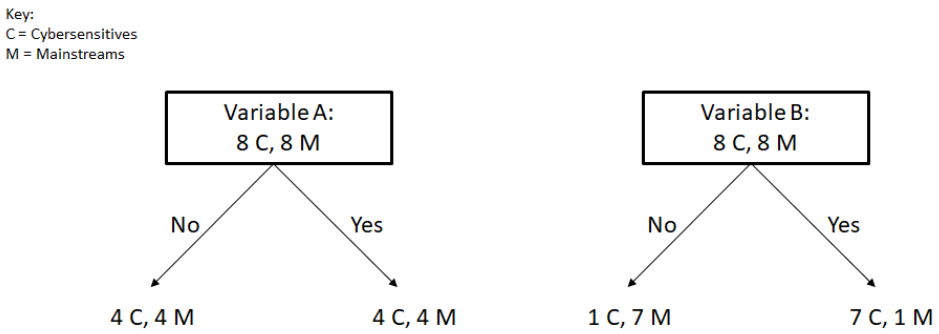
[1] Often drawn from survey data

[2] As an analogy, think about someone who sought to describe your morning routine. They described *your* decision making processes in the morning so accurately that the description is no longer accurate for anybody else, because it only describes the details of your routine. Sometimes, for the model to be accurate across a large group (or a whole population), it must be somewhat vague and not overspecific.

2.2. Classification and Regression Tree (CART) Model

Classification and Regression Trees (CART) refer to a set of machine learning, computational-based strategies that are the most common form of quantitative decision tree modeling.[3] CART models typically involve three basic steps: pre-development, model development, and pruning through testing and/or ensemble methods. Data cleaning refers to the process of organizing the data for the model, including developing independent and dependent variables, splitting the data into training and testing sets, resampling the data, etc. The next step is todevelop the decision tree model. The most important consideration in this process is the method used to determine branching: that is, the equation and algorithm used to determine when to split the data into smaller branches and the variable to use to split it. After creating the initial model, developers test the accuracy model with cross-validation techniques and based on those results prune (or strategically adjust) the tree further. Various ensemble methods like random forests, bootstrapping, boosting, and bagging, are particularly effective for pruning. This refines the nodes on the branches to ensure better accuracy and reliability (on both the data set and on potentially new potential data sets respectively).

Like most machine learning models, CART modeling requires balance between accuracy and reliability. A perfectly accurate model can account for every individual in the dataset and correctly predict it. For CART, that means adding as many levels or branches as possible until it classifies correctly. The potential issue with this is that such a large model overfits, becoming over-representative of the initial data and not representative of the population as a whole. A model is reliable, however, if it is robust enough to be able to accurately classify or predict cases from new datasets, samples, surveys, and/or studies. Typically, data scientists find the healthy balance between the two by developing the decision tree model on the initial model on their sample and testing it on another larger set of data. Another way to address this is through ensemble methods. Random forests is an extremely popular type of ensemble method for this.



In CART decision tree modeling, the most important algorithmic procedure is branch splitting: that is the process to decide when to split into a new branch and which variable to choose to do so. The goal is to choose the variable that splits the given data most distinctly into separate, homogenous groups. For example, if suppose for within a given pool of data, there are two remaining variables – Variable A and Variable B – each with two options – "Yes" and "No" (see the Figure above). If it splits by Variable A, it will now have two groups for yes and no – the "yes" group with four cybersensitives and four with mainstreams, and the "no" group with four cybersensitives and four with mainstreams as well; whereas, if

[3] Classification decision tree modeling develop trees with categorical and ordinal data, and regression tree modeling with continuous data. Because our data is primarily categorical or ordinal, we built a classification tree.

its splits by Variable B, it will have a "yes" group with seven cybersensitives and one mainstream and a "no" group with one cybersensitives and seven mainstreams. Splitting by Variable B is more beneficial, since doing so splits the two groups that (mostly) splits the cybersensitives and mainstreams into two different groups, unlike the Variable A, which only splits them into equal-sized groups. This would indicate that Variable B is a better potential variable to use to classify them in terms of cybersensitive and mainstream, given that for Variable B, most of each group fall in one side or the other (a.k.a. into as homogenous a group as one can), not straight down the middle.

There are two popular functions used to calculate how heterogenous (that is, varied) the resulting subgroups would be: the gini index function and entropy function.

The Gini Index is currently the most popular function used to decide branch splitting.[4] The Gini index functions calculates how heterogenous (that is, varied) the resulting subgroups would be if the tree split according to that variable. Minimum values are best: less is more. The range is not important, but it is a proportion, so it ranges from 0 to 1 (similar to a percent). It uses the squared proportional make-up of each type in the new split, according to the following equation: [5]

$$Gini\ Index = 1 - \sum_{t=0}^{t=k} P_t^2$$

Since this calculates how *heterogenous* a given variable's grouping would be, the feature variable from the data (like gender, how often they use a given technology, etc.) with the most *homogenous* grouping (that is, exclusively one type), that is splits most distinctly into separate groups, will have the *minimum* Gini index score and will be chosen as the next split in the graph.

After splitting, the decision tree model will repeat this process until all groups are of one type, or if a maximum depth is set, until that branch is reached. For example, in the situation above, after splitting by Variable B, the subgroups are still not completely homogenous: the "yes" and "no" group contain seven members of one group and one remaining of the other. This is closer to homogeneity than the 50-50 split before, but not yet *exclusively* one type. Thus, the model would repeat the process for each subgroup under Variable B: for the "yes" subgroup, it would search the other feature variables that best splits it (a.k.a. minimizes Gini index) and branch accordingly; and it would do that for the "no" subgroup as well.

Thus, it recursively splits each new subgroup using the Gini index function until all subgroups are the same type: in this example, all cybersensitive or all mainstream. As an alternative ending criterion, a maximum depth can be set in order to maintain a certain size in which case if that branch reaches that depth, it ceases splitting, whether or not it has a completely homogenous grouping. For example, if the maximum depth is ten, the algorithm will automatically stop after the tenth branch even if the subgroups are not completely homogenous. This recursive branching method is the heart of how CART modeling works.

---

[4] Entropy is another popular function for branch splitting, which uses a more complicated logarithmic approach. We tested the entropy function and found Gini index to be more accurate for our data (see Section 3.3.2.1 for details). Hence, we presented Gin index as our branch splitting example above.

[5] This equation comes from http://dni-institute.in/blogs/cart-decision-tree-gini-index-explained/.

2.3. Modeling Cost Effective Potential

In the PON for this project, we were tasked with, "[r]ecommend[ing] alternative formulations of energy efficiency potential models that clearly identify the portion of cost-effective potential that is achievable by segmenting populations in different ways, relying upon means such as variables descriptive of culture and behavior among subpopulations."

The decision trees discussed above will function to identify and enumerate the proportion of cybersensitives and cyberawares in any given population. We discuss how implementers of utility programs, both energy efficiency and marketing in general, can use the Ethnographic Decision Tree Model as presented in this paper to conduct their own segmentation exercise. Once the proportions of the various segments has been established, we recommend evaluating the segments in terms of their energy consumption. In our Task 5 report, *Cybersensitive Electricity Consumption Patterns*, we showed that cyber-segments (the combined set of cybersensitives and cyberawares) appear to use less energy than other segments. Because our sample size is small, we recommend that people with larger datasets make calculations using appropriate electricity consumption data. In this report, we offer a preliminary calculation that we consider to be representative of the overall gap in terms of energy consumption among our segments, for the two utility territories we looked at. A more refined model will be provided in the final version of this report, post testing the CART with a synthetic population.

Because we lack data from specific energy efficiency program implementation, we are choosing to infer impact on California using the concept of negawatts. Rocky Mountain Institute founder Amory Lovins coined the term 'negawatt' in 1990 to describe the power of the "watt you don't use while still receiving the same goods and services from the watts you do use" (Lovins, 1990). In the future, using the segmentation schema, and models, we have developed, it will be possible to estimate exact effects from differing energy efficiency uptake among segments. Because our sample size is so small, and our numbers are unreliable, we are offering an impact estimate instead of energy savings potential.

3.Methods

This section details the process of constructing the two models: the EDTM and the CART. The goal is to construct decision trees which faithfully reproduce behavior in such a way as to sort our population into one of the five segments: cybersensitives, cyberawares, mainstreams, low mainstreams, and nulls. It is important to note, that for this research, we are building an EDTM decision tree models, but we are not testing it. Testing ethnographic decision tree models, as an example, is an integral part of the overall process in constructing a reliably predictive model (Gladwin, 1989). However, this phase requires an entirely new sample of previously unexposed respondents, who nevertheless match the original sample as closely as possible (Gladwin, 1989). That phase would require an attendant need for resources in terms of time, labor, and monetary incentives to drive recruitment. Ideally, there would be two additional rounds of recruitment and testing, with two different samples: one to initially test assumptions and refine the model, and one to validate the findings. However, that lies beyond the scope of this project.

3.1. Datasets

Over the course of this project, we surveyed approximately 400 individuals living in the territories serviced by California investor owned utilities, Pacific Gas and Electric, Southern California Edison, and

San Diego Gas and Electric. We collected demographic data, and answers to questions about attitudes towards technology, device purchase and usage, and energy awareness and consumption.

EDTMs "are built from interviews with a relatively small sample of people (20-60) and are usually tested on a similarly small and local sample" (Ryan and Bernard, 2006). Therefore, we collected ethnographic data, including audio and video recordings from a sample of 45 households in both Northern and Southern California, during the period 2015 to 2017.

Finally, we collected energy data in several different forms. We collected up to 36 months of energy data from the utility accounts of a different but overlapping set of 23 of 45 participating households. We also collected energy data from large, anonymized, and aggregated sets provided by the local utilities, PG&E and SCE, for Marin and Long Beach, California respectively. These included:

- Anonymized residential customer data by zip code for Marin County (PG&E Public Data Sets: Electric Usage by Zip)
- Anonymized residential customer data by zip code for City of Long Beach (SCE Quarterly Customer Data Reports)
- California Public Utilities Commission (CPUC) 2016 Residential Electric Usage and Bill Statistics by Climate Zone

3.2. Ethnographic Decision Tree Model (EDTM)

An EDTM provides a tool for predicting the behavior of a certain persona type under a given set of circumstances. Our EDTM uses the questions and answers from the survey and interview contexts to build a formal model representing how people interact with their devices and consumer energy. This model 'sorts' people, based on their answers, into one of four segments—mainstream, null, cybersensitive, or cyberaware. Thus, our EDTM should:

- Identify ‘mainstreams’ who were not engaged with their energy consumption information.
- Distinguish those with technical skills from those who engage with information on a deeper level.
- Distinguish cyberawares from cybersensitives via the emphasis on tracking information (as opposed to merely receiving and responding to it).

3.2.1. Process

Ethnographic decision tree modeling is a tool used to help explain decision-making made by groups, rather than by individuals. The objective is to represent the decision-making process made by members of a group in response to a certain set of circumstances, e.g. freshmen deciding about a meal-plan at college. There are many sources that detail the process for building an Ethnographic Decision Tree Model, the most important being Christina Gladwin’s *Ethnographic Decision Tree Modeling* (1997). Here we will discuss briefly the steps involved in building the model, but for a detailed understanding we recommend her book and other literature on the subject, which we have cited.

The first phase of EDTM development always consists of conducting a series of ethnographic interviews with the members of the group under scrutiny. The interviews are designed in such a way as to elicit the process of decision-making in the words of the group themselves. In the second phase of EDTM development, the ethnographer(s) reviews the verbatim responses, and organizes the steps in the decision-process as captured in the ethnographic data. The third phase of EDTM development is to run the now

diagrammed set of choices past another, similar, set of group members, to see where the ethnographer may have misunderstood or missed pertinent information. If the EDTM model is generally predictive (the literature suggests that a minimum 80% of cases should be properly accounted for by a well-done model) then the modeler can stop, although ideally the process of refinement can continue *ad infinitum*.

In the book, *Ethnographic Decision Tree Modeling* (1997), Gladwin gives the illustrative example of using an EDTM to understand the decision-making process of a college student buying a meal plan. In the example, her goal is to understand, "under what circumstance will that freshman include breakfast when purchasing their meal plan?"

Gladwin starts by interviewing a set of college freshmen, asking them open-ended questions about their experience purchasing a meal plan. The interview guideline includes questions about the value of breakfast, the students' schedules, and access to resources like refrigerators. These questions can also elicit quantitative data, such as, 'how much did breakfast cost?'

The information given to her by the college freshmen will include the set of 'rules' guiding their decision (such as "I don't eat breakfast" or "McDonald's is cheaper.") The next step is to then organize those rules into a set of if-then statements[6] and order them in such a way as to reflect the proper order of decisions (e.g., someone who answers 'no' to did you buy the breakfast plan might skip questions which concern those who did purchase it). The next step is to test the decision tree on a similar set of freshmen and determine if it is reflective of reality. After this step, the ethnographer uses errors and outliers to refine the questions within the model. The goal is to have accounted for enough variables to explain the behavior of at least 80% of the group.

In our project, we conducted in-depth interviews (IDIs) that contained questions organized into three categories. We termed these categories Psych, Device, and Energy. The Psych category was concerned with emotional aspects of technology and energy usage. The Device category focused on purchase and usage behaviors. The Energy category looked at energy consumption, awareness, values, and habits.

As discussed in our Task 3 report, *Psychosocial Drivers of Technology Engagement Among Cybersensitives*, we found that our participants clustered into five consistent groupings according to how they responded to questions. Cybersensitives and cyberawares gave many responses for multiple codes across the three topic categories. Conversely, Nulls gave minimal responses to few codes across the topic categories. Mainstreams fell in between.

In our Task 4 paper, *Cybersensitive Response to Technology*, we analyzed these clusters statistically, and showed that for the Energy and Device categories, their responses were predictive of membership in their segments (while for the Psych category the evidence was inconclusive). We also correlated the participants' membership status in the cluster with their answers to our recruitment survey (outlined in Task 2 paper, *Preliminary Ethnographic Report on Cybersensitives and Technology Detailing the Fieldwork and Early Findings*). Overall there is a strong case for the relationship between how cybersensitive someone is, and how they will respond to a given question.

For Task 6, and the building of the EDTM we returned to the transcripts from the IDIs, looking for verbatim examples of patterns of behavior for inclusion in our EDTM. What were specific things that cybersensitives and cyberawares reported doing that was different from the rest of their cohort? We also

[6] If-then statements often form the basis for programming, being commands that 'if' condition X is met, 'then' action Y should occur.

returned to the survey and picked out the questions that had the strongest evidence for assessing differences among the segments.

For example, we assert that cybersensitives have different patterns of energy consumption than non-cybersensitives (as illustrated in our Task 5 report, *Cybersensitive Electricity Consumption Patterns*). Their consumption is, on average, lower than those of other members of their cohort. Therefore, it would make sense that anyone who did not report even attempting to 'save energy' or otherwise engage with their utility would be unlikely to be a cybersensitive. Thus, our first question for building our decision tree would be, "Have you tried to save energy?" Someone who answers 'no' to this question is unlikely to demonstrate the characteristics of a cybersensitive (such as engagement with energy information) and so we would tag them as 'mainstream' and they would exit the decision-tree[7].

If they answer 'yes' to the question 'Have you tried to save energy?" then we need to establish how they acquired information about saving energy. The next few questions help us establish this:

1. Do you read your monthly electricity bill?
2. Did any information from utility change your energy behavior?
3. Do you get information about saving energy from other sources (internet?)

Again, if they are interested in saving energy, but did not pursue it in any meaningful way, it is unlikely that they are cybersensitives, and they will be tagged as 'mainstream' and exit the query. The next few questions identify potential nulls.

4. Technology is easy for me
5. My friends ask my advice about tech issues
6. I have a lot of devices (laptop, phone, tablet)
7. I feel uncomfortable when separated from my devices

One can be cybersensitive without a deep understanding of technology. Conversely, people can be sophisticated in their understanding about technology, without evincing cybersensitivity. These questions allow people to give us enough context without prematurely winnowing them from the pool. It is helpful to remember that the EDTM is not a survey and would be administered face-to-face in another IDI. That means that people would be able to give some context, clues, even corrections, as the questions are asked.

From the ethnographic observations, we established a few common aspects to cybersensitive lifestyles (psychographic element), one of which was the pursuit of multiple, in-depth hobbies and a sort of zest for life/inquisitiveness about things in general. The next questions help them to self-identify:

8. I have several hobbies
9. I try new things all the time

[7] If we were asking freshmen about their meal plan at college and someone were to answer that they do not go to college, they would similarly be disqualified. It is not that they might not have interesting or valid opinions, but they lack membership in the group under investigation.

At the same time, based on their answers to survey questions, we know these folks report immersing themselves in the details of the things they are interested in. The next two questions help identify people who may be more superficial in terms of their engagement with new ideas and interests:

10. I quickly move on to the next new thing
11. I spend a great deal of time on details

The next set of questions helps us to flesh out the mode of communication and engagement they are most comfortable with. Since we are interested in people receiving information on their devices, we ask about applications and platforms.

12. Do you communicate with friends and family primarily via phone call?
13. Do you communicate with friends and family primarily via email?
14. Do you communicate with friends and family primarily via video call (Skype/Facetime)?
15. Do you communicate with friends and family primarily via text/WhatsApp?
16. Do you communicate with friends and family primarily via social media (Facebook)?

The next section deals with engagement with the utility, but also with devices, platforms, and applications, the modes of engagement with energy consumption, billing, and information about both. This will help further identify 'mainstreams' who slipped through earlier questions. People who do not engage with their utility online/via device or who receive information about their energy consumption solely through paper means, will be tagged as likely not cybersensitive.

17. Do you receive your electric bill via email?
18. Do you pay your electric bill online?
19. Do you log in to pay your electric bill?
20. Is your monthly payment automated?

The final set of questions help distinguish cyberawares from cybersensitives:

21. Do you use any device to track your energy?
22. Do you have any devices that track other things, like health/fitness?
23. Have those devices changed your behavior?

At the end of the model, the questions posed should have:

- Winnowed out mainstreams at various points along the way
- Distinguished nulls from cybersensitives and cyberawares
- Distinguished cybersensitives from cyberawares

We should have four pots containing cybersensitives, cyberawares, mainstreams, and nulls.

3.3. Classification and Regression Trees (CART)

We then combined our survey results with the quantitative coding data from our in-depth interviews, and from this quantitative data built a CART model. As noted in the Background, CART modeling follows three basic steps – pre-development, development, and pruning – and we will outline our methodology for each part below. In summary, during pre-development, we cleaned and resampled the data; during development, we built the model, optimizing between several different parameters; and during pruning, we built developed a random forest model to create the most accurate result.

3.3.1. Pre-Development

3.3.1.1. Data cleaning

We imported the responses from the surveys stored in an Excel file, and converted all non-numerical data into numerical data using Python coding.

3.3.1.2. Resampling

We used resampling to enlarge the dataset: resampling refers to ways to add dummy cases that maintain the same distribution and other data patterns as the given data. CART algorithms work better on a larger dataset, usually of a couple hundred, Resampling develops a larger case that still maintains the integrity of the data (its distribution and basic patterns across several dimensions) . Decision trees can vary widely when a small dataset is used, since slight variations in data disproportionately affect its decision-making process. Resampling provides a way to create a larger set of data that is still representative of the initial dataset. We developed a sample set of 300. This created a "sandbox" set of data from which we could develop the decision tree model.

The method we used for resampling was bootstrap resampling. This works by randomly adding the dataset values with replacement (in other words, values can be selected more than once) from the initial dataset until we reach 300. We then tested the means and standard deviations of each column to ensure that the resample is still representative of the original data, which it was.

3.3.2. Decision Tree Development

3.3.2.1. Test several parameters.

When building decision tree models, there are several parameters or options to construct it. One must choose a branching function (equation used to decide when to split the tree) and set a maximum depth (a maximum number of levels or branches for a path). We tested two different types of splitting functions – Gini index function and entropy function – and tested several maximum depths. Table 1 shows the accuracy for each branch splitting function type by maximum depth. Gini index and entropy are the two most commonly used functions for deciding when to split branches (see Section 2.1 for details on splitting). Maximum is the most number of branches allowed: once the tree reaches the maximum depth, the building algorithm stops splitting that section into new branches. Accuracy is measured by the percentage of cases the model is accurately able to predict when conducting Leave-One-Out Cross-Validation (LOOCV), which goes through each value in the data, builds the model without that value, and tests to see whether that built model is accurately able to predict that value. Here are two important considerations:

- Gini index function is consistently more accurate than the entropy function. Hence, we used the Gini index function to build our model.
- The accuracy levels off at a maximum depth of eight, which likely is because the model stops branching after eight cases and thus eight is the most natural depth.

Table 1. Accuracy by Branch Splitting Technique and Maximum Depth of Decision Tree Levels

Maximum Depth	Gini Index Accuracy	Entropy Accuracy
1	25.3%	33.3%
2	31.3%	35.3%
3	33.6%	43.6%
4	45.3%	47.6%
5	62.6%	57%
6	68.6%	59%
7	69.6%	66.6%
8	76%	72%
9	75.3%	72.6%
10	74.3%	74%
11	76.3%	74.6%
12	76.6%	74.6%
13	76.3%	74.6%
14	76.3%	74.6%
15	76.3%	74.6%

3.3.3. Build decision tree.

We developed the initial decision tree. We used the Gini method (see Appendix A) and a maximum depth of 8. The LOOCV accuracy score is 76%.

3.3.4. Pruning: Random Forests

We employed an ensemble method called random forests to address overfitting. Ensemble methods combine several models together to improve the results, based on a collaborative approach. Random forest generation is a way to compare several different CART decision trees to improve the accuracy. The random forest is a new model built from these separate decision tree models. Our random forest model has an accuracy of 100%.

Basically, a random forest algorithm produces several CART decision trees (called forests, because they build many trees) based on randomly selected subsets of data points within the sample. For each data point, the random forest model then classifies based on the mode classification among all the trees, that is as the attribute most frequent classification. For example, if most trees constructs classify person X as a cybersensitive, it classifies him/her as a cybersensitive.

Because random forests compare several different trees, they allow for significantly improved accuracy in their classifications, known to possess exceedingly accurate results. By creating and testing between several CART different decision trees, random forests address potential tendency towards variation and overfitting inherent in decision trees, since they pick up the wide patterns between many specific decision tree models, filtering out both any irregular variation or narrow, over focus of a particular model by looking at the patterns of the models as a whole. Here we built the random forest model (using the Gini index function to generate the trees). Our random forest produced one thousand randomly generated decision models

We used the same resampled sample set discussed in Section 3.3.1.2 to build the model since that provides the largest and most flexible model. We then tested it for accuracy on both the resampled population and the original sample. The former provides an internal test of the model's own ability to predict its own behavior, and the latter provides an initial external test of a similar but different model. Both have an accuracy of 100%, meaning that they were able to accurately predict the cyber status of all individuals. (Accuracy is measured by the percentage of individuals the model can accurately predict). The random forest model also produced a single decision tree with accuracy of 100%, a helpful single tree in of itself because it is the most accurate single decision tree generated (see in Section 4.2 for a diagram of it).

3.4. Testing

Testing our decision tree models – both EDTM and CART – is important in order to demonstrate the reliability of our data among the general population but is beyond the scope, resources, and time of this current Task. We plan to test our CART data on a synthetic population sample of Prince William County, Virginia, produced by Virginia Tech University in the future. Their synthetic population draws upon data from the American Census Survey (ACS) and the American Time Use Survey (ATUS), resampled to match the population of the county. The population is of roughly comparable in size to the geographic territories we focused on in our project, Marin County, and the City of Long Beach. As mentioned above, a full, rigorous test of the EDTM is not possible without additional time and money for recruiting at least one, if not two, new ethnographic population samples and conducting in-depth interviews with them, but we can use the CART to simulate a population (and thus indirectly test the EDTM because they are both constructed using the same data).

3.5. Inferring/Predicting Energy Savings

We used the proportions of each group predicted by the model with the estimates for average energy usage for each group in our study from Task 5 to predict energy savings. We offer a preliminary estimation for how much energy cybersensitives and cyberawares save compared with the other groups in the population based on the above decision tree/random forest model.

In the CART model, cybersensitives make-up about 18.18% of the population. Based on estimates in Task 5, we will approximate cybers serviced in PG&E as using 1,452 kWh of energy per year as opposed to 4,959 kWh for non-cybers, a difference 3,507 kWh per cybersensitive household per year.

Using the estimate of 18.18%, if one multiplies 637.5726 (0.1818x4,959) to the size of population in question, then one can estimate the difference in terms of electricity consumed by the cyber segments versus non-cyber segments in that population. PG&E and SCE possess 5.2 million and 14 million customers respectively.

Testing the model against a larger sample population would help verify and/or refine the figure. We intend to do this by testing our model against technology consumption and energy usage data of Prince Williams County, VA provided by Network Dynamics Simulation Science Laboratory at Virginia Tech University.

4.Results

In this section we present the two decision-trees, one ethnographic, the other machine learning. Both decision-tree models can identify members of the segments under discussion, and from there estimate their prevalence in any given population.

4.1. Ethnographic Decision Tree Model

Following the method outlined above, we constructed an Ethnographic Decision Tree Model (Figure 2), which energy efficiency program managers can administer to determine if someone is a cybersensitive, cyberaware, mainstream, or null.[8]

[8] We have no ability to distinguish 'Low' mainstream at this time.

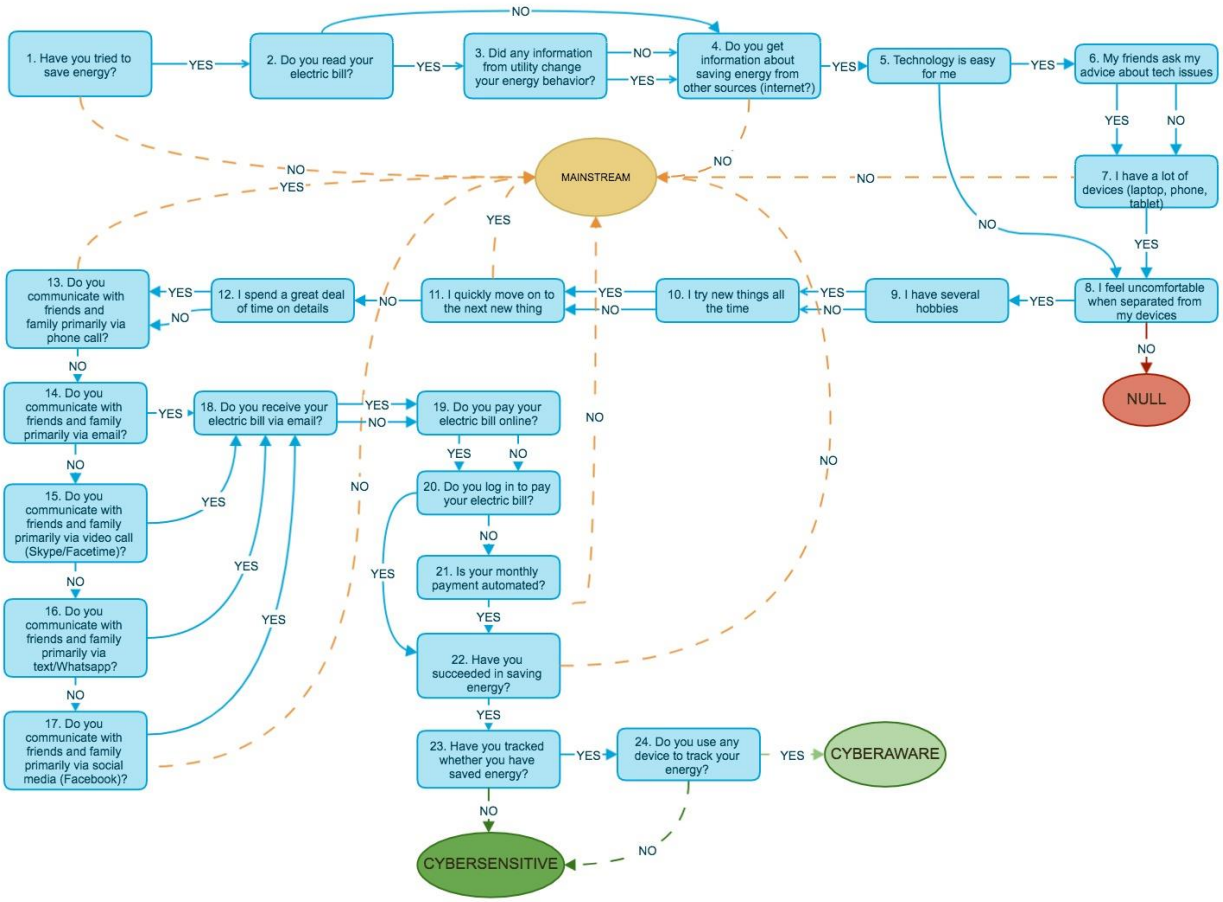


Figure 2. Finished Ethnographic Decision Tree Model (EDTM)

A decision tree works particularly well because the relationships among these groups are not amenable to linear representation. Although in prior reports we have grouped, or color coded the segments in a manner that suggests a linear or logarithmic relationship, that is an artifact of data visualization. Our research suggests that the relationships among the various levels of cybersensitivity are more complex. For example, while we initially hypothesized that cyberawares would be a 'less intense' form of cybersensitive, we have not found evidence to support this hypothesis. We do not currently have enough evidence to suggest that cyberaware engagement differs in *intensity* from that of cybersensitives – who may pursue energy efficiency measures/energy conservation methods, from a more holistic, lifestyle-oriented perspective, or as something that is important to civic or environmental duty. Instead, cyberawares are a variant of cybersensitive, who are prone to *tracking* energy consumption (and other forms of information such as health and fitness) through their personal devices.

Similarly, our so-called 'nulls' are in many ways very similar to cybersensitives/cyberawares, rather than occupying the opposite end of a spectrum. While they registered very low in terms of emotional engagement with either devices or energy consumption (the opposite of cybersensitives) they otherwise share many traits around device purchase and usage behavior, which can be easily misleading when trying to sort consumers into segments. Nulls tend to be technologically sophisticated, and may work in related fields, such as computer programming, or information technology. Nulls share many of the same characteristics with respect to technology as the cyber segments, but they responded drastically differently

to our questions in the realm of psychology and energy consumption[9]). We pay attention to nulls because they would otherwise give us potential false positives for cybersensitivity/cyberawareness. Thus, when seeking to identify cybersensitives and cyberawares, a focus on attitudes and emotions is important to distinguish them, because solely focusing on technology consumption and usage behaviors will be insufficient.

Finally, the 'mainstreams' are the most unlike any of the other segments. While they may report having an interest in saving money or energy, they lack demonstrable engagement with their energy consumption, and do not meaningfully pursue avenues of education around their energy consumption (or other forms of information), whether delivered by their utility or otherwise. They do not have deep technology knowledge, nor do they possess a commitment to tracking their energy use, health, or finances.

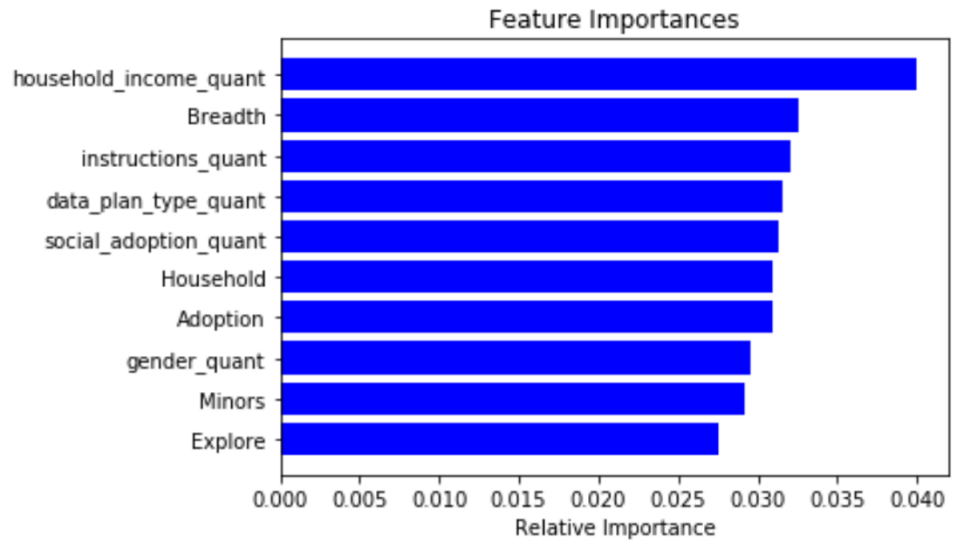
In Part 2: Recommendations, we will discuss how we believe how assigning people to these segments will prove useful to designers and implementers of residential energy efficiency programs, with a focus on behavior-based efficiency programs.

4.2. CART Model

As discussed above, our CART model has an accuracy of 76%, in terms of classifying the cybersensitivity of members of a cohort; but our more refined random forest model has an accuracy of 100%. Because random forests create many different CART trees, each with varying degrees of accuracy, displaying all the trees is not practical or significant: our model, for example, produced one thousand decision trees, most of which are in of themselves not very accurate and none of which contribute by themselves significantly in the how the model classifies individuals. Thus, for this type of modeling, data scientists are most interested in which variables have been the most significant in helping to filter individuals, in other words, which have been most central variables when trying to classify all individual data points among all the trees (Bell et al, 2018). The graph below shows the 10 most important variables. These, according to the random forest model, are the most important criteria in determining an individual's cyber status.

Figure 3. Random Forests Feature Importances

[9] *Psychosocial Drivers of Technology Engagement Among Cybersensitives*



Because our random forest model produced a CART decision tree with 100% accuracy, a rarity since develops these trees probabilistically, we have included a copy of that particular tree below. Even though as a single tree, it is one among one thousand in the random forest modeling, it demonstrates the most accurate CART decision tree model developed so far.

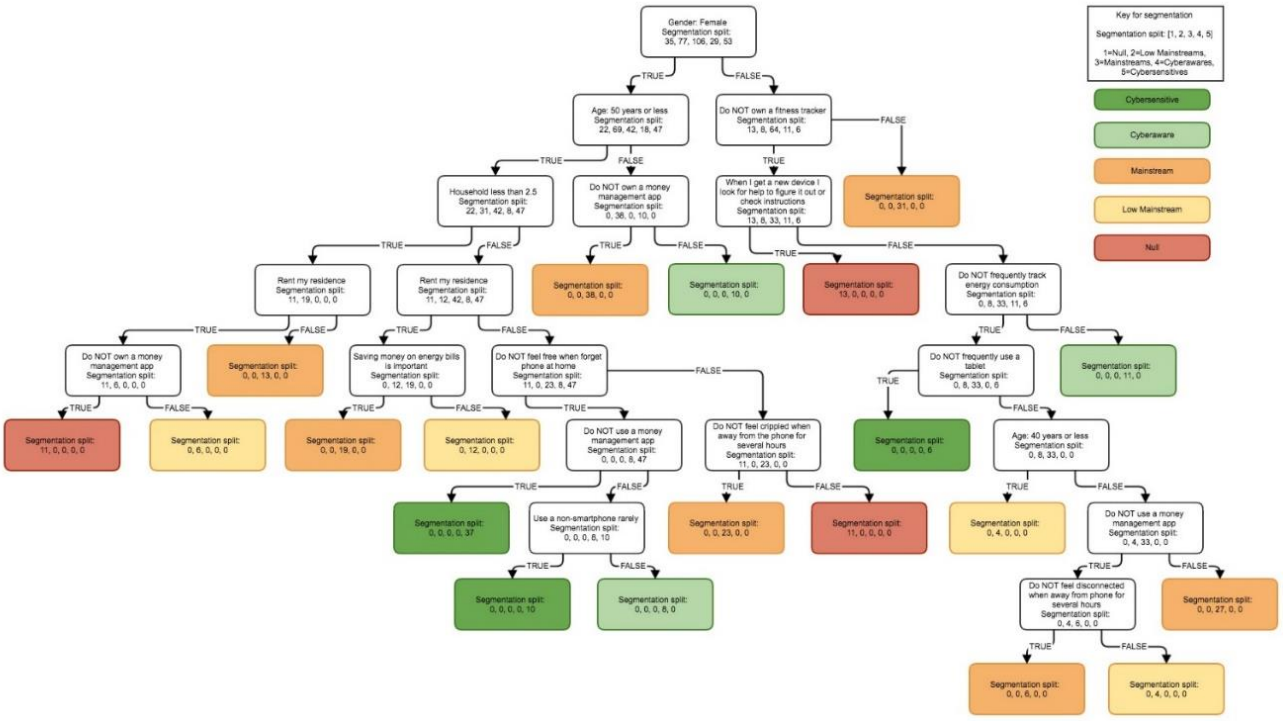


Figure 4 CART Decision Tree – Most Accurate CART Tree Generated

4.3. Impact of cyber segments on electricity consumption

The CART model predicts approximately 18% of the population will fall into cybersegments. In our Task 5 report, *Cybersensitive Electricity Consumption Patterns*, we saw a difference in energy consumption among our (very small) sample. For illustrative purposes only, we used that figure for calculating the 'negawatts' currently enjoyed by the state of California due to the presence of cybersensitives and cyberawares. Our conclusion is that, for the utility territories of PG&E and SCE, cybersensitives and cyberawares probably produce negawatts in the range of 17,309,687,040 kWh per year. In other words, energy usage in California would be much higher without the presence of cybersensitives and cyberawares. This is a preliminary estimate, using the data we currently have. A larger future sample would help solidify these figures.

5.Discussion

In looking for patterns of behavior around electricity consumption, it helps to go beyond traditional, attitudinal surveys to see what people actually do in their homes as they engage with the services that require electricity such as cooling, heating, and drying. The PON for this funding cycle recognized this and stipulated that:

> "[F]unded projects will use field work obtaining data from actual experiments to achieve the following objectives: Improve understanding of the role of culture and behavior (distinct from cost-effectiveness considerations) in energy efficiency uptake, and identify specific product/service marketing techniques that may influence these factors."

Our project has used ethnographic fieldwork to collect data around the role of culture and behavior in energy efficiency uptake, "[b]ecause the researcher uses ethnographic eliciting techniques to specify decision criteria, he or she avoids making unrealistic behavioral assumptions and armchair propositions about how people in the real world make important decisions." (Gladwin, 1989)

We have developed a schema for classifying consumers of electricity in terms of their distinct psycho-graphic and behavioral profiles. We have provided a model for replicating this work, in the form of an ethnographic decision tree model. In Part 2, we will discuss our specific recommendations in terms of what types of energy efficiency programs will be best suited to take advantage of these findings, and also explore the marketing technique of segmentation. In this section we discuss some of the pros and cons of using these models, and how other researchers might apply them.

5.1. EDTM pros/cons for this application

Constructing and deploying an Ethnographic Decision Tree is a method which allows energy customers (with the help of an ethnographer or other fieldworker) to self-segment as a cybersensitive, cyberaware, mainstream, or null. We will discuss the role of segmentation with respect to energy efficiency programs in Part 2, but we believe that segmentation allows for utilities to deliver messages to customers in a manner that increases their impact and decreases soft costs.

We want to stipulate that this method is not perfectly predictive. Human beings are complex, and because humans live in social groups like households, predicting energy consumption behavior at the household level from a few dozen questions is an imperfect art at best. However, we believe that we have

demonstrated the relevance of certain personality traits and psychological drivers in shaping broader energy consumption patterns.

While we recognize that this method is labor intensive, we believe that this could be mitigated through incorporation into a home energy audit, which would assist the managers of home energy audits to make and test some predictions concerning uptake of recommendations and energy efficiency investments.

5.2. CART pros/cons for this application

The biggest advantages of quantitative decision tree models are that results are easy to visualize and understand intuitively (unlike many other machine learning processes with obscure, unintelligible processes). CART models require little initial assumptions or requirements on the data, not assuming its type: categorical, ordinal, continuous, etc. Unlike many other methods, it does not require all data to be continuous and can handle a mixture of data types. They do not require normalization or extensive data cleaning to prepare the data, typically required by machine learning methods for which all data must be of one type and/or of the same scale.

CART models also provide a flexible, easily manipulatable foundation for random forests and other methods that combine several different models/approaches. They are easy to create and because of its branching style, it is easy to edit or prune one branch or part of the tree, and this makes it easy to build and synthesize the results from several decision trees coherently. See the discussion on random forests below [HG3] for more details.

Conversely, CART models can have a problem called overfitting. Here the model describes the specific data so accurately that it fails to generalize well to the overall population. Decision tree models can become too large – that is, have too many branches – which can reflect considerations that are specific to that small group but do not occur for the population. As with the EDTM, the CART (as of this writing) will be an untested model -- requiring more/larger sets of data to run before we can be satisfied that it reliably replicates real world behaviors. However, as with the EDTM, we believe that the model's construction will be robust enough that another entity, with their own resources and access to data, would be able to deploy the model and test it for use as one of a suite of tools for better understanding and predicting consumer behavior.

6.Conclusion

6.1. Summary of what we did

We collected data through a survey of approximately 400 people, and in-depth interviews with members in 45 households in two California IOU territories (PG&E and SCE). Using this data, we identified meaningful differences in how participants answered certain questions. Those questions formed the basis for a set of if-then statements, which we organized into an Ethnographic Decision Tree Model (EDTM). We used this model as the foundation for building a Classification and Regression Tree (CART) through machine learning processes. We then used the CART model to provide a baseline for estimating the impact of cybersensitives on energy consumption in the two utility territories. Both models require an additional step of testing against new samples. The EDTM will not be tested directly during this project, because testing requires the recruitment of a similar sample for in-depth interviews. However, we will test the CART using a synthetic population supplied by partners at Virginia Tech University. The testing of

the CART model will also indirectly support the validity of the EDTM, because they derive from the same datasets. This testing will continue into the Fall of 2018, and we will report out our conclusions in the Final Project Report.

6.2. What we found

Our Ethnographic Decision Tree Model provides a means whereby any interested party can segment a population according to cybersensitivity. The questions could be administered during a home energy audit, or even by survey. Applying the EDTM will organize any given population into four segments: cybersensitive, cyberaware, mainstream, and null. We offer recommendations for how best to incorporate segmentation with residential energy efficiency programs in Part 2: Recommendations.

The CART model appears to have a predictive accuracy of 76%. It produces a cybersensitive/cyberaware population of 18%. This is close to our goal of 80% accuracy, and close to our original hypothesis that cybersensitives and cyberawares make up two deciles, or 20% of any given[10] population.

Beyond constructing the models this paper reports on, this paper is an important contribution to qualitative methods, because "[F]ew publications have reported in detail the process whereby ethnographic interviews, survey interviews, and statistical analyses can be integrated with decision modeling to predict behavior." (Bauer and Wright, 1996). Even fewer have constructed an ethnographic decision tree model, and a classification and regression tree model, using the same datasets.

[10] Anglophone populations. We have not looked at literature or data outside of Anglophone North America and Europe.

References

Barros, Rodrigo C., De Carvalho, André C. P. L. F, and Alex A. Freitas. *Automatic Design of Decision-Tree Induction Algorithms*. Cham: Springer International Publishing, 2015.

Bauer, Mark, and Anne Wright. "Integrating Qualitative and Quantitative Methods to Model Infant Feeding Behavior among Navajo Mothers." *Human Organization* 55, no. 2 (1996): 183-92.

Bell, Andrew, Jennifer Zavaleta Cheek, Frazer Mataya, and Patrick Ward. "Do As They Did: Peer Effects Explain Adoption of Conservation Agriculture in Malawi." *Water* 10, no. 1 (2018): 51.

Dahan, Haim, Shahar Cohen, Lior Rokach, and Oded Maimon. *Proactive Data Mining with Decision Trees*. New York, NY: Springer New York, 2014.

Gladwin, Christina H., Hugh Gladwin, and Walter Gillis Peacock. "Modeling Hurricane Evacuation Decisions with Ethnographic Methods" International Journal of Mass Emergencies and Disasters 19, no. 2 (2001): 117-143

Gladwin, Christina H. *Ethnographic Decision Tree Modeling*. Newbury, CA: Sage, 1997.

Gladwin, Hugh, and Michael Murtaugh. "Test of a Hierarchical Model of Auto Choice on Data from the National Transportation Survey." *Human Organization* 43, no. 3 (1984): 217-26.

Harris, Marvin. "Why a Perfect Knowledge of All the Rules One Must Know to Act like a Native Cannot Lead to the Knowledge of How Natives Act." *Journal of Anthropological Research* 30, no. 4 (1974): 242-51.

Huang, Hsiu-Li, and Mei Chang Yeh. "Introduction to Ethnographic Decision Tree Modeling." *Journal of Nursing* 53, no. 3 (2006): 60-68.

IBM Corporation. (2012). *Decision Tree Models*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm

Khawaja, M. Sami, PhD, and James Stewart, PhD. *Long-run Savings and Cost-effectiveness of Home Energy Report Program*. Report. Waltham, MA: Cadmus Group, 2017. 1-20.

Klein, B., & Schlagenhauf, T. (2018). *What are Decision Trees*. Retrieved from Python Machine Learning Tutorial: https://www.python-course.eu/Decision_Trees.php

Lee, Saro, and Chang-Wook Lee. "Application of Decision-Tree Model to Groundwater Productivity-Potential Mapping." *Sustainability* 7, no. 10 (2015): 13416-3432.

Lovins, Amory. "The Negawatt Revolution." *The Conference Board Magazine*, 27 no. 9 (1990): 18-23.

Madadipouya, Kasra. "A New Decision Tree Method for Data Mining in Medicine." *Advanced Computational Intelligence: An International Journal* 2, no. 3 (2015): 31-37.
Mathews, Holly. "Predicting Decision Outcomes: Have We Put the Cart before the Horse in Anthropological Studies of Decision Making?" *Human Organization* 46, no. 1 (1987): 54-61.

Mingers, John. "An Empirical Comparison of Selection Measures for Decision-tree Induction." *Machine Learning* 3, no. 4 (1989): 319-42.

Mukhopadhyay, Carol. "Testing a Decision Process Model of the Sexual Division of Labor in the Family." *Human Organization* 43, no. 3 (1984): 227-42.

Murtaugh, Michael. "A Model of Grocery Shopping Decision Process Based on Verbal Protocol Data." *Human Organization* 43, no. 3 (1984): 243-51.

Navega, David, Catarina Coelho, Ricardo Vicente, Maria Teresa Ferreira, Sofia Wasterlain, and Eugénia Cunha. "AncesTrees: Ancestry Estimation with Randomized Decision Trees." *International Journal of Legal Medicine* 129, no. 5 (2014): 1145-153.

Neville, Padraic G. *Decision Trees for Predictive Modeling*. Report. SAS Institute Inc. 1994. 1-24.

Plattner, Stuart. "Economic Decision Making of Marketplace Merchants: An Ethnographic Model." *Human Organization* 43, no. 3 (1984): 252-64.

Rokach, Lior, and Oded Z. Maimon. *Data Mining with Decision Trees: Theory and Applications*. New Jersey: World Scientific, 2015.

Ryan, Gery W., and H. Russell Bernard. "Testing an Ethnographic Decision Tree Model on a National Sample: Recycling Beverage Cans." *Human Organization* 65, no. 1 (2006): 103-14.

Subbiah, Rajesh, Animitra Pal, Eric K. Nordberg, Achla Marathe, and Madhav V. Marathe. "Energy Demand Model for Residential Sector: A First Principals Approach." *IEEE Transactions on Sustainable Energy* 8, no. 3 (2017): 1215-224.

Tso, Geoffrey K.f., and Kelvin K.w. Yau. "Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks." *Energy* 32, no. 9 (2007): 1761-768.

Thorve, Swapna, Samarth Swarup, Achla Marathe, Young Yun Chun Baek, Eric K. Nordberg, and Madhav V. Marathe. "Simulating Residential Energy Demand in Urban and Rural Areas." Report. Department of Computer Science, Department of Agricultural and Applied Economics Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute, Virginia Tech. Blacksburg, VA, 2018.

Wilf, Eitan. "Toward an Anthropology of Computer-Mediated, Algorithmic Forms of Sociality." *Current Anthropology* 54, no. 6 (2013): 716-39.

Woodside, Arch G. *Case Study Research: Core Skill Sets in Using 15 Genres*. Bingley, UK: Emerald, 2017.

Yu, Zhun, Fariborz Haghighat, Benjamin C.m. Fung, and Hiroshi Yoshino. "A Decision Tree Method for Building Energy Demand Modeling." *Energy and Buildings* 42, no. 10 (2010): 1637-646.

Yu, Zhun, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. "A Systematic Procedure to Study the Influence of Occupant Behavior on Building Energy Consumption." *Energy and Buildings* 43, no. 6 (2011): 1409-417.

Appendix A

Decision Tree Modeling

The model uses decision trees to classify our study participants* according to their cyber status (cybersensitive, cyberaware, mainstream, low mainstream, and null) based on the survey and interview data. The goal of the model is to classify/predict what cyber status of an individual.

Data Cleaning

In [1]:

```
# Imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cross_validation import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
from sklearn import model_selection
from scipy import signal
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: De
precationWarning: This module was deprecated in version 0.18 in favor of the
model_selection module into which all the refactored classes and functions ar
e moved. Also note that the interface of the new CV iterators are different f
rom that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

In [2]:

```
# Loading the datasets from their respective files
filepath = 'C:\\Users\\swphi\\Documents\\Indicia\\Epic\\Task 4\\'
file1 = pd.ExcelFile(filepath + 'Master EPIC dataset - copy.xlsx')
file2 = pd.ExcelFile(filepath + 'all survey responses - copy.xlsx')

cyber_status = file1.parse('Cyber Status')
psych_codes = file1.parse('Psych')
energy_codes = file1.parse('Energy')
device_codes = file1.parse('Device')
surveys = file2.parse('Interview Responses')
key = file2.parse('Question Key')
```

In [3]:

```
# Merging all the datasets into one data for the model
data = cyber_status.merge(psych_codes, left_on = 'Name', right_index = True)
data = cyber_status.merge(energy_codes, left_on = 'Name', right_index = True)
data = cyber_status.merge(device_codes, left_on = 'Name', right_index = True)
```

```
data = cyber_status.merge(surveys, left_on = 'Name', right_on = 'Name')
```

In [4]:

```
# Converting non-numerical data into numbers so that it will run through the
decision tree packages

# Converts it to a quanitative variable based on the given key
def quantize(x, key):
    for k in key:
        if x == key[k]:
            return k
    return x

'''
Key for Variable Edits:
rent_or_own_quant: 0 if the user rents and 1 if the user owns
gender_quant: 0 if female and 1 if male
age_quant: Goes to smallest age in the range
household_income_quant: Goes to the smallest income and -1 if prefer not to a
nswer
community_quant: 0 if rural, 1 if suburban, 2 if urban
instructions_quant: 0 if wait for someone else, 1 if use instructions, 2 if f
igure out on your own
region_quant: 0 if southern california, 1 if northern california
smartphone_usage_quant: 0 if never, 1 if a few times a month, 2 if a few time
s a week, 3 if once a day
nonsmartphone_usage_quant: 0 if never, 1 if a few times a month, 2 if a few t
imes a week, 3 if once a day
gaming_console_usage_quant: 0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
laptop_usage_quant:0 if never, 1 if a few times a month, 2 if a few times a w
eek, 3 if once a day
ipod_usage_quant: 0 if never, 1 if a few times a month, 2 if a few times a we
ek, 3 if once a day
fitness_tracker_usage_quant: 0 if never, 1 if a few times a month, 2 if a few
times a week, 3 if once a day
health_fitness_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a
few times a week, 3 if once a day
home_automation_system_usage_quant: 0 if never, 1 if a few times a month, 2 i
f a few times a week, 3 if once a day
home_security_usage_quant: 0 if never, 1 if a few times a month, 2 if a few t
imes a week, 3 if once a day
energyConsumption_app_usage_quant: 0 if never, 1 if a few times a month, 2 if
a few times a week, 3 if once a day
```

```
money_management_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a few times a week, 3 if once a day
tablet_usage_quant: 0 if never, 1 if a few times a month, 2 if a few times a week, 3 if once a day
music_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a few times a week, 3 if once a day
gaming_app_usage_quant: 0 if never, 1 if a few times a month, 2 if a few times a week, 3 if once a day
data_plan_type_quant: 0 if they do not have a plan, 1 if 1-5 GB plan, 2 if more than 5 GB, 3 if unliminited, -1 if unsure what plan they have
energy_awareness_quant: 0 if not aware, 1 if generally aware, 2 if always aware
social_adoption_quant: 0 if one of the last to adopy any new technology, 1 if wait until somewhat widely adopted, 2 if one of the first ones to adopt
'''

quant_keys = {'Rent or Own': ['rent_or_own_quant', 'Do you currently rent or own your primary residence?', {0: 'Rent', 1: 'Own'}],
              'Gender': ['gender_quant', 'Please indicate your gender below:', {0: 'Female', 1: 'Male'}],
              'Age': ['age_quant', 'Please indicate your age below:', {25: '25-34', 35: '35-44', 45: '45-54', 55: '55-64', 65: '65-74'}],
              'Household Income': ['household_income_quant', 'Which of the following ranges best indicates your annual household income?', {2000: '$20,000 to $49,999', 50000: '$50,000 to $99,999', 100000: '$100,000 to $149,999', 150000: '$150,000 to $199,999', 200000: '$200,000 or more', -1: 'Prefer not to answer'}],
              'Community': ['community_quant', 'How would you describe the type of community you reside in?', {0: 'Rural community', 1: 'Suburban community', 2: 'City or Urban community'}],
              'Instructions': ['instructions_quant', 'Which of the following statements best describes you?', {0: 'When I get a new device, I usually wait for someone else to help me figure out how to use it', 1: 'When I get a new device, I jump right into the instructions and learn how to use it in detail', 2: 'When I get a new device, I figure out how to use it on my own, and look at the instructions only if I get stuck'}],
              'Region': ['region_quant', 'Region', {0: 'Southern California', 1: 'Northern California'}],
              'Smartphone Usage': ['smartphone_usage_quant', 'Mobile phone with internet capability:How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
              'Non-Smartphone Usage': ['nonsmartphone_usage_quant', 'Mobile phone without internet capability:How often, if ever, do you use or access the
```

following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Gaming Console Usage': ['gaming_console_usage_quant', 'Gaming console (such as Playstation, X-box, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Laptop Usage': ['laptop_usage_quant', 'Laptop:How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'ipod usage': ['ipod_usage_quant', 'Portable digital music player (such as iPod, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Fitness Tracker Usage': ['fitness_tracker_usage_quant', 'Health / fitness tracker (such as Fitbit, Garmin watch, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Health/Fitness App Usage': ['health_fitness_app_usage_quant', 'Health / fitness tracking apps on your mobile phone (such as RunKeeper, Strava, etc.):How often, if ever, do you use or access the following?', {0: 'Neve r', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Home Automation System Usage': ['home_automation_system_usage_quant', '', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Home Security Usage': ['home_security_usage_quant', 'Home security or home monitoring system (such as ADT, web-enabled surveillance devices, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Energy Consumption App Usage': ['energy_consumption_app_usage_quant', 'MEnergy consumption tracking devices, apps or services:How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Money Management App Usage': ['money_management_app_usage_quant', 'Money management applications (such as Mint, Quicken, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Tablet Usage': ['tablet_usage_quant', 'Tablet (such as iPad, etc.):How often, if ever, do you use or access the following?', {0: 'Never', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once a day'}],
'Music App Usage': ['music_app_usage_quant', 'Music apps on my phone:How often, if ever, do you use or access the following?', {0: 'Never',

```
1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least once
a day'}],
              'Gaming App Usage': ['gaming_app_usage_quant', 'Gaming apps on
my phone:How often, if ever, do you use or access the following?', {0: 'Never
', 1: 'A few times a month or less', 2: 'A few times a week', 3: 'At least on
ce a day'}],
              'Data Plan Type': ['data_plan_type_quant', 'What type of the da
ta plan do you have on your mobile phone?', {-1: 'I have a data plan, but not
sure about the type', 0: 'I do not have a data plan', 1: '1-5 GB per month',
2: 'More than 5 GB per month', 3: 'Unlimited data plan'}],
              'Energy Awareness': ['energy_awareness_quant', 'Consider the le
vel of energy consumption in your household. Which of the following statement
s would you agree with the most?', {0: 'I am not aware of the level of energy
consumption in my household. I generally do not participate in making changes
around the house to reduce our energy consumption.', 1: 'I am generally aware
of some aspects of energy consumption in my household. I do not monitor all a
spects of energy usage in great detail, but participate in making changes to
our energy consumption whenever it is convenient.', 2: 'I am fully aware of,
and monitor the level of energy consumption in my household. I have made and
continue to make many changes wherever possible to our energy usage, and lead
the charge in this aspect in my household.'}],
              'Social Adoption': ['social_adoption_quant', 'Which of the foll
owing statements best describes you?', {0: 'I am usually the last one to adop
t any new technology', 1: 'I wait for new technologies to be somewhat widely
adopted before adopting them myself', 2: 'I am usually among the first ones t
o buy the latest electronic devices'}]
              }

for q in quant_keys:
    current = quant_keys[q]
    data[current[0]] = data[q].apply(quantize, args = (current[2],))
    data[current[0]] = data[current[0]].fillna(0)
    key = key.append(pd.Series(current, index = key.columns), ignore_index =
True)
```

In [5]:

```
# Develops the X and y matrices for model development

dep_vars = ['region_quant', 'Smartphone', 'Non-Smartphone', 'Gaming Console '
,
            'Laptop', 'iPod', 'Fitness Tracker',
           'Home Automation System', 'Health/Fitness App', 'Home Securtiy',
           'Energy Consumption App', 'Money Management App', 'Tablet ',
           'Music Apps', 'Gaming Apps ', 'smartphone_usage_quant',
```

```
            'nonsmartphone_usage_quant', 'gaming_console_usage_quant', 'laptop
_usage_quant',
            'ipod_usage_quant', 'fitness_tracker_usage_quant', 'home_automatio
n_system_usage_quant',
            'health_fitness_app_usage_quant', 'home_security_usage_quant',
            'energy_consumption_app_usage_quant', 'money_management_app_usage_
quant',
            'tablet_usage_quant', 'music_app_usage_quant', 'gaming_app_usage_q
uant', 'data_plan_type_quant',
            'Relieved', 'Happy', 'Anxious', 'Unaffected', 'Concerned',
            'Bored', 'Free', 'Excited', 'Crippled', 'Stressed', 'Unproductive'
, 'Frustrated',
            'Disconnected', 'Adoption', 'Breadth', 'Explore',
            'Advice', 'Bills', 'Energy Saving', 'Fun', 'Easy', 'Technology New
s',
            'social_adoption_quant', 'instructions_quant', 'gender_quant', 'ag
e_quant',
            'household_income_quant', 'Household', 'Minors', 'community_quant'
, 'rent_or_own_quant']

X = data[dep_vars]
y_qual = data['Cyber Status']

cyber_key = {0: 'Null', 1: 'Low Mainstream', 2: 'Mainstream', 3: 'Cyberaware'
, 4: 'Cybersensitive'}
y = y_qual.apply(quantize, args = (cyber_key,))
y_quant = y
```

Resampling

In this section, we use bootstrap resampling to create a larger version of the sample data with a size of 300. Decision tree models vary impulsively on smaller datasets, and this will help address this. Resampling allows one to create a distribution, which is larger (or sometimes smaller) than the orginal, but which still represents the overall distribution.

In [6]:

```
X_resample = signal.resample(X, 300)
X_resample = pd.DataFrame(X_resample, columns = X.columns)
X_resample = X_resample.round()
y_resample = signal.resample(y_quant, 300)
y_resample = y_resample.round()
y_resample = pd.DataFrame(y_resample, columns = ['Cyber Status'])
```

In [7]:

```
# Develops a qualtiative version the y_resample for models that require a qua
litative version
def qualitize(x, key, minimum, maximum):
    if x < minimum:
        return key[minimum]
    if x > maximum:
        return key[maximum]
    return key[x]

y_resample_qual = pd.DataFrame(columns = ['Cyber Status'])
y_resample_qual['Cyber Status'] = y_resample['Cyber Status'].apply(qualitize,
args = (cyber_key, 0, 4,))
```

I then tested the means and standard deviations of each column to ensure that the resample is still representative of the original data. They are pretty much the same with only slight varaition several decimal places in (the one exception to this is income, which because it contains 5 to 6 digit values means that its slight variation is proportionally larger. It is still a slight variation). This indicates that the resampling still matches the sample's distribution.

In [8]:

```
test = pd.DataFrame(columns = ['Original Mean', 'Resampled Mean', 'Original S
td', 'Resampled Std'])
test['Original Mean'] = X.mean()
test['Resampled Mean'] = X_resample.mean()
test['Original Std'] = X.std()
test['Resampled Std'] = X_resample.std()
test['Mean Difference'] = abs(test['Original Mean'] - test['Resampled Mean'])
test['Std Difference'] = abs(test['Original Std'] - test['Resampled Std'])
display(test)
```

Decision Tree Development

We will first determine the most accurate decision tree model to construct. In particular, we will determine the best type of splitting function - gini index or entropy - and what are maximum depth is. Below are the accuracies when conducting Leave-One-Out Cross-Validation (LOOCV) for each type of function type for each depth.

Two conclusions from the table:

1) Gini index are consistently more accurate than entropy.

2) The accuracies start to level off at a maximum depth of 7, inicating that 7 is the best maximum depth

Thus we will use the gini index with a maximum depth of 7.

In [10]:

```
loocv = model_selection.LeaveOneOut()
```

```
accuracy = pd.DataFrame( columns = ['Maximum Depth', 'Gini Index Accuracy', '
Entropy Accuracy'])

for depth in range(1, 34):
    score = [depth]

    clf_gini = tree.DecisionTreeClassifier(criterion = "gini", random_state =
100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_gini, X_resample, y_resampl
e_qual, cv=loocv)
    score.append(results.mean())

    clf_entropy = tree.DecisionTreeClassifier(criterion = "entropy", random_s
tate = 100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_entropy, X_resample, y_resa
mple_qual, cv=loocv)
    score.append(results.mean())

    accuracy.loc[len(accuracy)] = score

accuracy
```

Out[10]:

In [11]:

```
accuracy.to_csv('Decision Tree LOOCV Results v2.csv')
```

Build initial decision tree.

Here is the primary decision tree. I uses the gini index function and has a maximum depth of 8. It's LOOCV accuracy score is 76%.

In [12]:

```
max_depth = 8
dtree = tree.DecisionTreeClassifier(criterion = "gini", random_state = 100
, max_depth=max_depth, min_samples_leaf=1)
dtree.fit(X_resample, y_resample)
tree.export_graphviz(dtree, out_file='tree.dot', feature_names=X.columns,
filled=True, rounded=True )
```

In [13]:

```
results = model_selection.cross_val_score(dtree, X_resample, y_resample_qu
al, cv=loocv)
print('LOOCV Accuracy Score: ' + str(100*results.mean()) + '%.')
```

```
LOOCV Accuracy Score: 76.0%.
```

Random Forests

Here I build a random forest with the resampled set. I then tested it on both the resampled set and original sample. For both, it has an accuracy of 100%. (Accuracy is measured by the percentage of individuals the model is able to accurately predict.)

In [14]:

```
rf = RandomForestClassifier(n_estimators = 1000)
rf.fit(X_resample, y_resample)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: DataConve
rsionWarning: A column-vector y was passed when a 1d array was expected. Plea
se change the shape of y to (n_samples,), for example using ravel().
```

Out[14]:

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

In [15]:

```
pred = rf.predict(X_resample)
correct = 0
for x in range(len(X_resample)):
    if pred[x] == y_resample['Cyber Status'][x]:
        correct += 1
score = correct/len(X_resample)
print('Accuracy on resampled set: ' + str(100*score) +'%')
```

```
Accuracy on resampled set: 100.0%
```

In [16]:

```
y[0]
```

Out[16]:

```
0
```

In [17]:

```
pred = rf.predict(X)
correct = 0
for x in range(len(X)):
    if pred[x] == y[x]:
        correct += 1
score = correct/len(X)
print('Accuracy on original sample: ' + str(100*score) +'%')
```

```
Accuracy on original sample: 100.0%
```

Plot the model.

The graph below shows the 10 most important variables, which have been most significant variables when trying to classify each individual among all the trees. A larger, full table with the relative

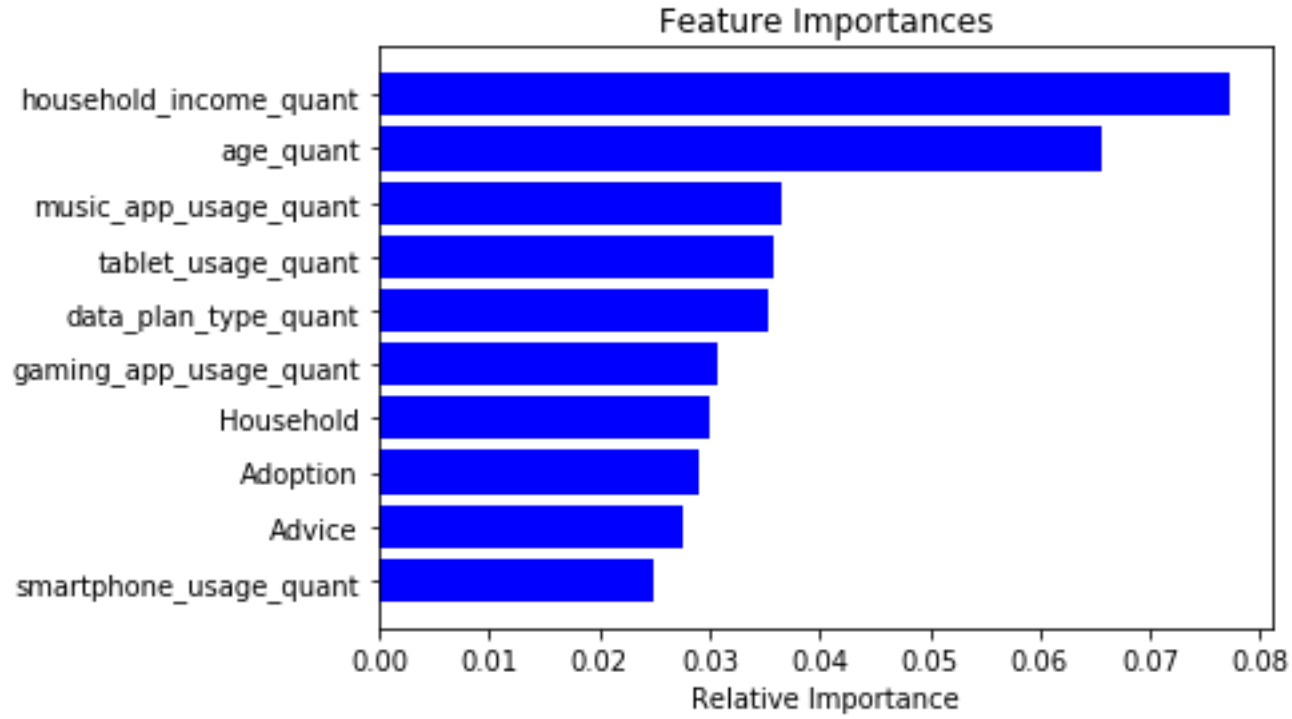
significance of all the variables is provided below that. (I made it latter table for your records; it is too much information for a final report, especially since most of it is extraneous. The grap provides the most important variables to discuss.)

In [18]:

```
# The code for this graphing approach is based on a similar model from http:/
/www.agcross.com/2015/02/random-forests-scikit-learn/
features = X.columns
importances = rf.feature_importances_
indices = np.argsort(importances)
l = len(indices)
top_indices = indices[l-10:l]
plt.figure(1)
plt.title('Feature Importances')
plt.barh(range(len(top_indices)), importances[top_indices], color='b', align=
'center')
plt.yticks(range(len(top_indices)), features[top_indices])
plt.xlabel('Relative Importance')
```

Out[18]:

```
Text(0.5,0,'Relative Importance')
```



In [19]:

```
def column_name(column_number, df):
    return df.columns[column_number]

def get_importance(index, list_of_importances):
    return list_of_importances[index]
```

```
def print_full(x):
    pd.set_option('display.max_rows', len(x))
    print(x)
    pd.reset_option('display.max_rows')

importances_ranking = pd.DataFrame({'Column Number': indices})
#output.Predicted = output.Predicted.apply(qualitize, args = (cyber_key, 0
, 4,))
# 'Column Name': column_name(X, indices), 'Relative Importance': importanc
es})
importances_ranking['Column Name'] = importances_ranking['Column Number'].
apply(column_name, args = (X,))
importances_ranking['Relative Importance'] = importances_ranking['Column N
umber'].apply(get_importance, args = (importances,))
print_full(importances_ranking.sort_values('Relative Importance', ascendin
g = False))
```

	Column Number	Column Name	Relative Importance
60	56	household_income_quant	0.073520
59	55	age_quant	0.067818
58	27	music_app_usage_quant	0.036126
57	29	data_plan_type_quant	0.034547
56	26	tablet_usage_quant	0.033653
55	28	gaming_app_usage_quant	0.030865
54	57	Household	0.030570
53	43	Adoption	0.030422
52	46	Advice	0.027976
51	15	smartphone_usage_quant	0.025409
50	18	laptop_usage_quant	0.025143
49	44	Breadth	0.022487
48	19	ipod_usage_quant	0.022474
47	53	instructions_quant	0.021726
46	25	money_management_app_usage_quant	0.021551
45	58	Minors	0.021307
44	51	Technology News	0.021102
43	45	Explore	0.020535
42	60	rent_or_own_quant	0.019648
41	21	home_automation_system_usage_quant	0.019525
40	49	Fun	0.019162
39	22	health_fitness_app_usage_quant	0.018548
38	50	Easy	0.018303
37	47	Bills	0.017428
36	13	Music Apps	0.016551
35	48	Energy Saving	0.015454
34	14	Gaming Apps	0.014398
33	17	gaming_console_usage_quant	0.013776
32	59	community_quant	0.013560
31	11	Money Management App	0.013262
30	20	fitness_tracker_usage_quant	0.013014
29	5	iPod	0.012586
28	52	social_adoption_quant	0.010916
27	34	Concerned	0.010694
26	42	Disconnected	0.010602
25	36	Free	0.010070
24	3	Gaming Console	0.010000

```
23              33                          Unaffected              0.009874
22              24  energy_consumption_app_usage_quant              0.009765
21               0                          region_quant            0.009690
20               8                 Health/Fitness App               0.009564
19              54                          gender_quant            0.009377
18              23          home_security_usage_quant               0.009216
17              41                           Frustrated             0.009004
16              32                              Anxious             0.008769
15              30                             Relieved             0.008455
14              12                              Tablet              0.008248
13               7             Home Automation System               0.007824
12              40                          Unproductive            0.006298
11               6                    Fitness Tracker               0.005899
10               1                           Smartphone             0.005579
9               10             Energy Consumption App               0.005430
8               31                                Happy             0.005138
7                9                      Home Securtiy               0.005135
6               38                           Crippled              0.004793
5                4                             Laptop              0.003623
4               39                           Stressed              0.003595
3               16          nonsmartphone_usage_quant               0.003097
2               37                            Excited              0.002992
1                2                     Non-Smartphone              0.002963
0               35                              Bored               0.000944
```

Energy Saving Estimator

In [20]:

```
print('Break down by Cyber Status according to Our Model:')
pred = rf.predict(X)
output = pd.DataFrame({'Actual': y_qual, 'Predicted': pred})
output.Predicted = output.Predicted.apply(qualitize, args = (cyber_key, 0, 4,
))
output.Predicted.value_counts()
```

```
Break down by Cyber Status according to Our Model:
```

Out[20]:

```
Mainstream        11
Low Mainstream     9
Cybersensitive     6
Null               4
Cyberaware         3
Name: Predicted, dtype: int64
```

In the model, cybersensitives make-up about 18.18% of the pouplation.

Based on estimates in Task 5, we will approximate cybers serviced in PG&E as using 1,452 kWh of energy per year as opposed to 4,959 kWh for non-cybers, a difference 3,507 kWh per cybersensitive.

Using the estimate of 18.18%, if one multiplies 637.5726 (0.1818x4,959) to the size of poulation in question, then one can estimate the amount of energy saved by Cybersensitives in that population.

PG&E and SCE possess 5.2 million and 14 million customer respectively, making an initialy estimate of energy savings of 17,309,687,040 kWh per year.

In [21]:

```
(5200000+14000000)*0.1818*4959
```

Out[21]:

```
17309687040.0
```