Predicting the Potential for Energy Efficiency Retrofits in Single-Family Homes: An Exploration of Data Targeting Mechanisms

by

Kaitlin Ryan Goldstein
A.B. Environmental Engineering
Brown University, 2008

M.S.E. Architectural Engineering
University of Texas at Austin, 2010

SUBMITTED TO THE DEPARTMENT OF ARCHITECTURE IN PARTIAL FULFILLMENT TO THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN BUILDING TECHNOLOGY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2014

© 2014 Kaitlin Ryan Goldstein. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereby after created.

Signature of Author: __________________________________________

Certified by: ________________________________
Leslie Norford
Professor of Building Technology
Thesis Supervisor

Accepted by: ________________________________
Takehiko Nagakura
Associate Professor of Design and Computation
Chair of the Department Committee on Graduate Students
Predicting the Potential for Energy Efficiency Retrofits in Single-Family Homes: An Exploration of Data Targeting Mechanisms

by

Kaitlin Ryan Goldstein

Thesis Committee

Leslie Norford
Professor of Building Technology
Thesis Supervisor

Sarah Williams
Assistant Professor, DUSP
Thesis Reader

Roy Welsch
Professor of Statistics & Management Science, Sloan
Thesis Reader
Predicting the Potential for Energy Efficiency Retrofits in Single-Family Homes: An Exploration of Data Targeting Mechanisms

by

Kaitlin Ryan Goldstein

Submitted to the Department of Architecture on August 8, 2014 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Building Technology

ABSTRACT

Historically, the lack of data on the United States' housing stock has been one of the primary barriers to market penetration of residential energy efficiency retrofits. Without knowledge of the homes and customers to reach, outreach has been untargeted and inefficient. As such, a study was performed to determine whether the potential for residential energy efficiency retrofit could be determined in the absence of utility data. The first phase of the research investigated the best pre-retrofit gas consumption metric to predict post-retrofit savings. Energy intensity (weather normalized total gas consumption per square foot) was chosen from four distinct metrics as the best corollary to energy savings. The second phase attempted to predict the pre-usage metric from phase one using only home characteristics and demographics, and the most predictive variables were determined. Data mining techniques were then explored to predict retrofit candidacy using energy intensity as a proxy. After showing that this was difficult to predict even when utility data was available, the progression to the third phase was reconsidered but explored. The models did not perform as expected for three reasons: 1) the marketing variables were not clean/accurate enough 2) the marketing variables did not explain enough of the variance in energy intensity and, 3) the connection between energy intensity and retrofit candidacy was not sufficiently well defined. While a definitive model of retrofit candidacy in the absence of utility data was not found, the research completed offers: 1) a mechanism by which to connect retrofit savings data to homes that have not yet undergone retrofit 2) an in-depth look at using publicly available variables to predict home energy consumption and, 3) a detailed examination of the connection between retrofit potential and raw gas utility data.

Thesis Supervisor: Leslie Norford
Title: Professor of Building Technology
ACKNOWLEDGMENTS

I cannot thank my thesis advisor, Les Norford enough for his unwavering support and kindness and patience during this process. I would also like to thank my Mom and Dad, without whom none of this would be possible. I would also like to thank the countless friends and creatures (Colonelk and Zeke) who have held me up and picked me back up. When I've reached out, there's always been a hand. For my friends: Josh, especially, and MinWah, who is an inspiration to every woman. For all my running buddies who helped me through things before the sun had risen. I would also like to thank my collaborator Michael Blasnik for his support and guidance.
# Table of Contents

ABSTRACT ........................................................................................................................................ 5

ACKNOWLEDGMENTS...................................................................................................................... 7

1. INTRODUCTION AND RESEARCH BACKGROUND ................................................................. 9
  1.1 MOTIVATION ............................................................................................................................ 9
  1.2 INTRODUCTION ........................................................................................................................ 10
  1.3 LITERATURE REVIEW ............................................................................................................ 14

2. PHASE ONE: PRE-RETROFIT METRIC MOST INDICATIVE OF SAVINGS .................... 23
  2.1 METHODOLOGY FOR PHASE ONE ....................................................................................... 23
  2.2 RESULTS FOR PHASE ONE ...................................................................................................... 26
  2.3 CONCLUSIONS FOR PHASE ONE .......................................................................................... 29

3. PHASE TWO: PREDICTION OF PRE-RETROFIT METRIC ................................................ 31
  3.1 INTRODUCTION TO PHASE TWO ......................................................................................... 31
  3.2 MOTIVATION FOR PHASE TWO ........................................................................................... 31
  3.3 VISION FOR USE OF PHASE TWO ........................................................................................ 32
  3.4 OVERVIEW OF PHASE TWO PROCEDURE: DATA PREPARATION .................................. 33
  3.5 PHASE TWO MODEL TESTING AND DEVELOPMENT ....................................................... 64

4. PHASE THREE: APPLICATION TO HOMES WITHOUT ENERGY DATA ....................... 84
  4.1 INTRODUCTION TO PHASE THREE ...................................................................................... 84

5. CONCLUSIONS AND RECOMMENDATIONS ...................................................................... 86

WORKS CITED .................................................................................................................................. 90

APPENDIX A: COMMENTED USAGE DATA CLEANING CODE .............................................. 92

APPENDIX B: B-PRISM .................................................................................................................. 94

APPENDIX C: EXPERIAN CLEANING CODE ........................................................................... 103
1. INTRODUCTION AND RESEARCH BACKGROUND

1.1 MOTIVATION
The market potential for residential energy efficiency retrofits is extremely high. Buildings represent 40% of the United States’ primary energy consumption (U.S. Department of Energy, 2009). According to the Rockefeller Foundation, (Rockefeller Foundation, 2012), building energy retrofits represent a $279 billion dollar investment opportunity with potential savings over 10 years of $1 trillion dollars. The avoided carbon amounts to 600 million metric tons of CO₂ per year or 10% of the total U.S. emissions in 2010. In addition, of the 3,000 trillion Btu of energy savings potential, the residential sector alone represents almost 2,000 trillion Btu of this possibility, and almost $200 billion of the $300 billion investment opportunity.

The residential sector is also the most fragmented, with small diffuse opportunities that are difficult to aggregate to a substantial scale. While single-family homes are 88% of the total residential housing stock and use 80% of the total energy, they are also the smallest consumers per square foot of real estate; the usage is extremely diffuse. In addition, McKinsey (McKinsey and Company, 2009), notes that 71% of end-use potential in the residential sector is tied up in improving the building shell and upgrading the heating and cooling equipment in existing homes.

In order to solve energy efficiency, energy efficiency in buildings must be solved. To solve energy efficiency in buildings, energy efficiency in single-family homes must be solved. In order to succeed, this must be accomplished at great scale, but the solutions are well known and do not vary substantially across the sector. Fix the shell, and upgrade HVAC systems. Historically, the pure lack of data on the United States’ housing stock has been one of the primary barriers to market penetration of residential energy efficiency retrofits. Without knowledge of the homes and customers to reach, outreach has been untargeted and inefficient. As such, what is
needed is a model to help those on the ground perform better, more targeted outreach for retrofit measures, leading to higher audit to retrofit conversion rates, and to help inform the very initial steps of the energy audit process. This research does not pretend to approach the equally critical factor of homeowner engagement. Figuring out whether or not a home has the potential for energy savings is just the first step to completing the retrofit. The homeowner has to both agree to, and then pay for the retrofit (or some of the retrofit) once his or her home has been identified.

1.2 INTRODUCTION
The ultimate goal of this research was to predict a home’s candidacy for retrofit using only a combination of demographic and home-characteristics data available for the entirety of the U.S. residential housing stock. This is important, as utility data is almost always protected for privacy and thus unavailable to assist in targeting where energy efficiency retrofits will be successful. However, in order to build and verify this model, it was necessary to have actual data on what homes consume and how much homes that have undergone retrofits have saved as a result of the work. As a very small fraction (less than one percent) of homes in the US have actually completed retrofits, the best way to predict for retrofit potential is to use a sample of homes where energy data is available to build a model for pre-retrofit consumption or energy intensity. Further, in order to verify that this energy intensity is indeed indicative of retrofit savings potential, another (albeit much smaller) dataset is necessary for which both pre-retrofit and post-retrofit data is available.

Available to the research, as diagrammed in what follows, was 6.5 years of gas billing data (monthly bills from 2007-2013) for about 1 million customers in Ohio. This data essentially lists the number of therms consumed per month, per customer. Approximately 13,000 of these homes underwent an energy efficiency retrofit in 2011. The work completed, cost of work completed, projected savings, and change
in characteristics as a result of the retrofit were all also available for these homes. For each of the following research steps, “pre-use” and “post-use” refers to a weather normalized annual gas consumption metric which is calculated through a billing analysis of monthly utility bills in combination with recent and long-term weather data. A pre-use and post-use metric refers to a particular formula used to characterize the consumption before and after retrofit. For example, weather normalized annual consumption could include just the component of the gas usage attributable to heating demand, or it could also include baseload. Furthermore, each of these values could be further normalized by the area, volume, or surface area of the home to yield “energy intensity”, or average therms consumed per year per square foot of home. Further detail on the billing analysis methodology, variants in the classic methodology utilized in this research, and the various different ways to quantify energy intensity pre-retrofit are discussed in the methodology section.

The proposed research workflow took the following form:

0) Learn the basic building blocks of the analysis procedure
   a. Billing analysis technique (monthly gas bills \(\rightarrow\) yearly usage)
      i. Bayesian PRISM method (B-PRISM)

1) Determine the pre-retrofit-use metric most indicative of retrofit savings potential- how does pre-retrofit usage predict post-retrofit usage?
   a. Use a sub-set of homes for which both pre- and post-retrofit consumption is available (~7,000 homes)
   b. Test various pre-retrofit-use metrics and their correlation with retrofit savings (utility data post-retrofit)
   c. Choose metric with best fit and explanatory power

2) Test various models to predict this chosen pre-use metric using home characteristics and demographics- how do home characteristics and demographics predict home energy consumption (pre-usage)?
   a. Use a large dataset of representative homes where energy data is available (~1,000,000 homes)
b. Test multiple data mining and modeling techniques for predictive power (*utility data pre-retrofit*)

3) Scale this model up to homes for which no energy data is available
   a. Use chosen modeling technique to predict chosen pre-use metric
   b. Validate model using homes left out of model development as test sample
   c. Quantify scalability of model to other regions/demographics/home-stocks (independent sample of homes in different region with pre-use/post-use/retrofit data available).
Figure 1: Research workflow diagram.
The Phase Two data preparation process took longer than originally expected, and the data was found to be less accurate and consistent than desired. During this phase it was also found that predicting energy intensity from the publicly available variables was not accurate enough to distinguish between candidacy groups (high potential, low potential). Phase three was reconsidered at this juncture, but a methodology was proposed to complete the analysis.

1.3 LITERATURE REVIEW
In order to take raw energy consumption data and turn this into useful information regarding how a home uses energy on an annualized and weather normalized basis, a billing analysis must be completed. This analysis combines long-term historical weather data, recent weather data corresponding to the usage period, and the monthly bills. With this information, an estimate of the gas used for baseload (cooking/water-heating), an estimate of the gas used for home-heating, and an indoor balance point temperature (related both to the indoor set-point and internal loads) are extracted. Furthermore, a metric for the amount of therms used per heating degree day (an aggregated measure indicating indoor-outdoor temperature differences) can be scaled utilizing the long-term weather data to estimate the energy consumed for heating energy in the typical year (normalized annual heating consumption). An estimate of total gas consumption in a normalized year can also be estimated by adding the yearly baseload estimate to the normalized annual heating consumption to get a normalized annual consumption. The historic development of this type of analysis and the ways in which the analysis used in this research differs from the original is discussed in what follows.

The PRISM method, developed in the late 1970's and greatly expanded by Margaret Fels in 1986, was originally designed to evaluate the effectiveness of home energy retrofit programs. The method utilizes a set of customer utility bills in addition to daily averaged outdoor temperatures both before and after retrofit to generate a weather-adjusted indicator of a home's performance called the Normalized Annual
Consumption (Fels M. F., 1986). This statistical method is also applicable when simply evaluating a home's candidacy for retrofit and results in three critical numbers describing the home's energy usage: the reference temperature, the baseload, and an indicator of the home's heating system and envelope performance, the heating slope. The model plots the rate of gas consumption, $f$, against the average daily outdoor temperature. The amount of energy $f$ needed to keep the home at the set point temperature may be described as:

$$f = \frac{L(T_{in} - T_{out})}{\eta} - \frac{Q}{\eta}$$

where $L$ represents the "lossiness", an indicator of the infiltration rates and insulation in the home, $Q$ represents the internal and solar gains within the home which counteract these losses in the heating season, and $\eta$ is the total heating system efficiency (equipment and distribution). $T_{in}$ and $T_{out}$ are simply the indoor temperature, and the average outdoor daily temperature, respectively. This equation may be simplified to:

$$f = \beta(t - T_{out})$$

where $t$ is the change-point temperature,

$$\beta = \frac{L}{\eta}$$

where $\beta$ is the heating slope, and

$$t = T_{in} - \frac{Q}{L}$$

The heating slope is the rate at which the home needs to be heated to maintain the desired indoor temperature while the reference temperature indicates the highest outdoor temperature at which the heating system first turns on. If fuel is utilized in the home for purposes apart from space heating such as cooking and hot water
heating, this enters the equation as the y-intercept, $\alpha$, of the fuel consumption assuming it remains relatively consistent in demand throughout the year.

$$f = \alpha + \beta (r - T_{out})$$

While each of these three variables, the reference temperature, heating slope and baseload are covariant and errors in the estimation of one variable will lead to errors in the other estimations, the model has been proven to be accurate (Fels & Keating, 1993) and consistent, particularly when utilized to compare the performance of the building before and after retrofit. The model has also been utilized to diagnose the efficacy of various measures in the program, separating out the baseload, thermostat settings, and shell improvements (Fels & Goldberg, 1986).

The second step in the PRISM algorithm is to define the Normalized Annual Consumption, (NAC), which is an indicator of what the fuel use in the building would have been in an average year of weather. This provides a means by which to compare the same building before and after retrofit in different heating seasons, and to compare one building to another where bills are available across different periods of time. It should be noted that while the fuel consumption might vary significantly in a given house from year to year, if computed correctly, the heating slope should remain constant regardless of the weather as long as no other changes are made to the building.

PRISM is one of many “variable-base degree day” (VBDD) methods. These methods essentially test multiple balance point temperatures for goodness of fit and choose the balance point with the highest $R^2$ in a plot of Use per day against Heating Degree Days per day. The energy consumed in a home may be attributed to two primary sources, the efficiency of the home, and the behavior of the people living inside of the home. As the physical potential for home retrofit is based purely on the former, it is critical to understand what portion of the consumption may be attributed to human behavior. These VBDD models are thus beneficial as they utilize a variable
balance point temperature towards a model solution. As such, the customer thermostat setting, which is an indicator of human behavior, may be partially disentangled from the purely shell and efficiency based losses within the home.

An AHRAE Research Project (ASHRAE 1050-RP) expanded upon the work of Fels and Goldberg to include inverse models applicable to commercial buildings where many of the base assumptions in PRISM tend to fail. The toolkit contains variable-base degree-day methods such as PRISM, but also includes change-point models where base loads are allowed to vary, and simultaneous heating and cooling may be accounted for. The toolkit also contains multivariate reference and variable-base degree-day models where other independent variables might serve to better explain the variance in fuel usage. The work also includes an algorithm that decides the best reference model to fit the data (Kissock, Haberl, & Claridge, 2003) (Kissock, Haberl, & Claridge, 2002). The Inverse Modeling Toolkit is available and the manual is included within the final ASHRAE report.

Examples of the applications of these models to the validation of predicted retrofit savings are plentiful (Fels & Keating, 1993) (Haberl, Thamilseran, Reddy, Claridge, O'Neal, & Turner, 1998) (Kissock, Haberl, & Claridge, 2003) (Nadel & Keating, 1991). Each of these papers utilizes the NAC before and after retrofit to compare the weather normalized fuel consumption before and after retrofit. NAC is the metric of choice when evaluating program efficacy, as it is very stable and consistent, with an average of 3% standard error (Fels M. F., 1986). In the past, program evaluation has been the primary application of these variable base degree-day and reference methods, and little was written regarding retrofit targeting utilizing the methodology.

However there have been numerous papers looking at the potential for utilizing these techniques for targeting purposes, specifically within the commercial sector (Reddy, Deng, & Claridge, 1999) (Reddy A., 1989) (Reynolds, Komor, & Fels, 1990) (Kissock, Reddy, & Claridge, 1998). There have also been two more recent examples
of work very similar to that proposed here, applying the methodology to the residential sector with a few key differences. Marcus Bianchi presented research from the National Renewable Energy Laboratory (NREL) in 2010 (Bianchi & Casey, 2010), which looked to assess the potential for a variable base degree-day method to identify good targets for energy efficiency retrofits. In this paper the authors simulate the utility data for approximately 300 homes of various user, base-load and envelope profiles and attempt to then correctly identify each signature utilizing the VBDD method. They also compare the accuracy of utilizing the Heating Slope as a proxy for code compliance against both Fuel Usage and Fuel Usage per Heating Degree Day. They find that in the simulated homes, the Heating Slope correctly identifies 100% of the homes above and below code compliance while the Fuel Usage and Normalized Fuel Usage fail to capture the nuances of base-load, human behavior, and shell or equipment performance. As such, homes that were code compliant but had very high set points and baseloads were incorrectly classified. The authors mention that the strength of the VBDD method is its capacity to disaggregate weather independent usage from the energy used to condition the space. Finally, they examined the accuracy of the VBDD model to identify the user characteristics (set point), baseload consumption, and building characteristics (shell and heating system efficiencies). They found that they could correctly identify homes that were simulated as having high users and high baseloads as well as those that needed envelope and/or heating system upgrades.

Kelly Kissock has also published in the residential targeting arena (Kissock & Raffio, 2007). In this paper the authors utilize the reference method to target 300 low-income homes for specific retrofit solutions. They do this by calculating the NAC and individual model coefficients for multiple buildings across multiple billing periods using multiple years of utility bills for each building. By visiting the homes after the analysis, they were able to validate their predictions. The authors find similar results in the utility of heating slope as a proxy for envelope and heating system efficiency upgrade needs. While NAC was a good indicator of retrofit potential, targeting using NAC alone missed 20% of the homes with the highest heating slopes.
Furthermore, 60% of the homes with the highest baseloads and 80% of the homes with the highest balance point temperatures would have been missed had NAC been used as the lone indicator of retrofit potential. After visiting the homes, they found that they had accurately predicted at least one of the primary problems in each home between 80 and 100% of the time. The range of accuracies came from utilizing each of the various coefficients (heating slope, balance point, baseload, and NAC) as their primary targeting metric.

These two papers prove the VBDD or reference model to be an effective means by which to target home for residential energy efficiency retrofits. Furthermore, each author was able to accurately decompose the primary components of home fuel usage into balance point, base load, and envelope/furnace efficiencies. This indicates that the method can serve as a means by which to ascertain specific information about the most likely work that needs to be accomplished where utility bills are available for the home.

There is however, a lack of literature on the development of models to predict the heating slope of an unvisited home for which utility bills are unavailable. The literature available on home energy retrofit predictions utilizing auxiliary data is focused heavily on predicting overall gas consumption that may or may not be an effective indicator of potential for retrofit.

Both Kolter and Livingston (Kolter, 2011) (Livingston, 2011) develop models that utilize a combination of demographic and building characteristic data to predict energy consumption of a home that has not yet received an energy audit. The Kolter paper focuses on the contextualization of this information to identify outliers that might be clear cases for retrofit and utilizes utility data for 6,500 buildings in Cambridge, MA. The Livingston paper develops a nonparametric regression and explores the use of a smooth backfitting estimator to understand residential energy consumption. The work relies upon the aggregate 2005 Residential Energy Consumption Survey (RECS) microdata.
Another prolific area of research is the prediction of end-use consumption to better understand not only how much energy is used, but also how it is utilized. Zeke Hausfather, now of Efficiency 2.0, developed a detailed regression model also using aggregate RECS data to predict end-use consumption at the zipcode level for the entire United States (Hausfather, 2012). Howard and Parshall also look at utilizing regression techniques to model the spatial distribution of end-use energy consumption and do so across multiple building types (Parshall, 2012). This work utilizes the aggregate building data from both RECS and CBECS (the commercial building survey corollary to RECS). These papers are focused on large-scale targeting of efforts for retrofit (zipcode level resolutions) and the prediction of large-scale demand needs across fuel types.

While all of these are critical areas of exploration, there is a gap in the prediction of both retrofit potential on a home-by-home basis in the absence of utility bills and in the prediction of a metric that is specific to building retrofit and not just total energy usage. While beginning by targeting the highest users is a logical approach, total consumption is not the only predictor of retrofit potential. Utilizing only this metric for targeting skews the selection to the largest homes, misses where baseload or behavior might be the underlying issue, and does not offer any indication of the relative savings that might be achieved through upgrade. In addition, while predicting end-use distribution is critical to predicting loads, space heating accounts for the highest fraction of energy consumption in heating-dominated climates and is also the greatest area of savings potential for retrofit. Finally, while aggregate energy data such as RECS and CBECS is incredibly useful for predicting large-scale patterns, it is too coarse to capture nuances in the housing stock of a particular region. As this research looks to predict actual building characteristics (those that are indicative of retrofit potential) these nuances are critical and the spatial resolution of the patterns needs to be very high.

By first investigating the metric most predictive of savings rather than using total energy consumption this work tried to shed light on building-specific potential for
retrofit. In addition, while the dataset was extremely dense but confined to a small region of the United States, very local patterns could be analyzed and a home-by-home metric established. While this resolution might be achieved at the expense of model generalizability, it is proposed that a similar analysis could be re-run for any region of the United States where high resolution utility data is available. As any utility in the country has billing data for their customers the work is repeatable for any gas utility willing to participate. Furthermore, as utility bills might not always be available for a given region, the primary goal of this work was to develop a model to predict for the heating slope in the absence of a set of home utility bills utilizing a combination of home and demographic variables available for purchase in any region of the United States.

There are two other very unique characteristics of the research performed here. This research utilizes an updated version of the classic PRISM methodology that was developed by a research collaborator, Michael Blasnik. This method takes a Bayesian approach to the estimation of the reference temperature. The probability of any given reference temperature is calculated based upon prior knowledge of reference temperatures in residential homes. Since the PRISM method essentially loops through each given reference temperature and solves the regression for each temperature in turn, the temperature with the highest R-squared value is generally chosen regardless of its likelihood. Blasnik takes a more informed approach that looks at the number of readings that were used to calculate the metric, the R-Squared of the fit, and the likelihood of the reference temperature. When referring to this method in subsequent sections of the paper, "" will indicate the “Bayesian Prism” method developed by Blasnik. The differences between B-PRISM and PRISM will also be further discussed in Section 3.4. The second characteristic that sets this work apart is the sheer volume of variables available for data exploration as well as the magnitude of the dataset itself. There are bills for one million residential customers over the course of five years enriched with hundreds of possible explanatory variables. While most of these variables were found to be overshadowed in importance by a few key variables that explained most of the
variance, there are variables that remained in the final model that would never have otherwise been correlated with retrofit potential and move the model slightly closer to accurate prediction. In addition, the ability to examine a whole host of variables helps to inform our physical understanding of what drives and does not drive energy efficiency potential. This has the secondary effect of informing which variables might be collected in a pre-screening phone call, or even where to place emphasis during the audit data collection process. In the end, while the methods applied are statistical in nature, the purpose of the model is to guide physical intuition.

This represents a substantial contribution to the field of energy efficiency research as residential retrofit represents the single largest sector of untapped potential for savings. There are two primary and substantial barriers to unlocking this potential: understanding which homes have the highest potential for upgrade; and understanding how to engage the demographic that not only has high retrofit potential but will also be interested in moving forward with the audit. This research will approach only the first barrier, as the second cannot be overcome until the first has been fully addressed.
2. PHASE ONE: PRE-RETROFIT METRIC MOST INDICATIVE OF SAVINGS

2.1 METHODOLOGY FOR PHASE ONE

For this portion of the research, looking to quantify the pre-use metric most predictive of retrofit potential, a billing analysis was performed on a sub-group of homes that received retrofits in 2011. These 7,000 homes had energy bills available both before and after the retrofit and billing data from 2007-2012.

Before running any of the analyses, the dataset was cleaned and filtered. This filtering discarded observations with poor utility data, such as two readings for a full year, missing values, or non-active meters, multiple different readings for the same date or estimated readings; and those with poor estimates from the B-PRISM model, such as those with: non-physical values, such as negative baseloads or negative heating demands, too few readings per year, poor regression fits, or >65% change in usage. This reduced the number of homes from about 7,000 to 3,000.

After cleaning the dataset, all analyses were performed on the bills for each independent year of data except for the year when the retrofit took place in 2011. Furthermore, multiple metrics of usage were calculated to test the correspondence between pre-use and savings. Each of these individual metrics was then regressed against its corresponding change before and after the retrofit. The four weather normalized annual energy consumption metrics looked at were:

1) **heating** energy consumption divided by total **floor** area \((\frac{E_{\text{heating}}}{A_{\text{floor}}})\),
2) **total** energy consumption divided by total **floor** area \((\frac{E_{\text{total}}}{A_{\text{floor}}})\),
3) **heating** energy consumption divided by total home **surface area** \((\frac{E_{\text{heating}}}{A_{\text{surface}}})\),
4) **total** energy consumption divided by total home **surface area** \((\frac{E_{\text{total}}}{A_{\text{surface}}})\).

The rationale behind these metrics is that the retrofits in the dataset had three primary components: air-sealing, wall insulation, and attic insulation. These retrofits should have the largest effect on weather dependent usage, or heating, and
have little effect on the baseload. Furthermore, air and heat escape through all the surfaces of the building (not just the floors) and thus surface area should better coincide with consumption. However, surface area is much harder to calculate and less well known than floor area. As such, the uncertainty in the measurement could overwhelm the signal. However, these hypotheses needed to be tested.

Initially, the idea was to calculate an energy intensity in 2012, subtract that intensity from that of 2010 (before the retrofit) and regress that against the 2010 energy intensities, (2012-2010) vs. 2010. This proved to be a more complicated exploit than initially intended because pre-use is endogenous to post-use, as will be later illustrated with regression results. The definition of endogeneity is that there is a correlation between the variable of interest (the predictor) and the error term in the response variable. When two variables are regressed against one another, the errors are assumed to be independent. When we regress savings against pre-use, we assume that the relationship between these two variables is attributable to the retrofit. However, there are other factors that might lower one's energy consumption from one year to the next unrelated to either weather or work on the house. The error term, "reduction in energy intensity not attributable to the retrofit" masks the signal that is associated with the retrofit. For example, if in 2010, a window was broken that no one noticed, or a thermostat setting malfunctioned, or a couple of new members of the household came for an extended stay, or someone was injured and had to work from home for that heating season, or the family decided to go winter in Florida, the home energy consumption for the entire winter would significantly change but that change would be completely unrelated to the retrofit work performed on the home. As the pre-use data (2010) factors into the calculation of the savings (2012-2010), the 2010 data will be correlated with the error term (savings unrelated to retrofit) of the savings. The goal is to choose the pre-year that is representative of long-term energy use and free of the nuances (positive savings signal) not attributable to the work performed on the home. There is a trade-off here between choosing two years that are close in time such that no long-term changes have occurred (for instance, the birth of a new
child), and one that is too close such that the error term in the estimation is correlated (short-term changes from one year to the next). An important distinction to make is that the "best" pre-year is not necessarily for all homes. This work looked to determine which pre-year was the "best" (free of short-term nuances) for the sample as a whole. One way to understand this is to think about how likely it is that short-term changes (a broken window) gets fixed in one year, two years, three years. The general idea is that we want to make sure the short-term problem was fixed but no long-term changes occurred in that chosen interval.

The initial approach to this problem was to evaluate each pre-use metric without actually using the pre-use data within the savings metric. Two ways of doing this are to compare realization rates (actual savings divided by modeled savings) and cost-effectiveness (therms saved divided by money spent) with each of the metrics of interest. When this line of inquiry was pursued, there was almost no signal indicating the relationship between the pre-use metric and these savings proxies. This line of analysis was thus abandoned, as the end goal was to find the best pre-use metric.

Given the availability of yearly pre-use data from 2007 to 2010, it was believed that evaluating the pre-use metrics for 2009, 2008 or 2007 rather than 2010 would effectively remove the endogeneity problem, and thus allow the inquiry to proceed. To test this hypothesis, regressions were run for each of the metrics, plotting the difference (2010-2012) against 2010, 2009, 2008 and 2007. Further, in order to prove that the endogenous component entered through the inclusion of the reference year, and that this was not unique to 2010, the same regressions were run plotting the difference (2009-2012) against 2009, 2010, 2008 and 2007. This was run for the full cleaned and filtered dataset (~3000 homes).
2.2 RESULTS FOR PHASE ONE

In what follows the result for one metric, \((E_{heating}/A_{floor})\), will be presented but the reader should note the results look quite similar for all other pre-use metrics. This portion of the analysis is not attempting to choose the best metric, but instead to understand the best pre-year to evaluate the metrics. Table 1 presents the results of the endogeneity analysis for the chosen representative metric. To simplify the visual presentation, the regression coefficients and constants have been removed.

<table>
<thead>
<tr>
<th></th>
<th>2012 vs. 2010</th>
<th>2012 vs. 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>2008</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>2009</td>
<td>0.27</td>
<td>0.42</td>
</tr>
<tr>
<td>2010</td>
<td>0.41</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 1 R-Squared Regression results comparing pre-use baseline comparison year to post retrofit savings for \(E_{heating}/A_{floor}\)

As discussed before, the goal is to get rid of the short-term biases (endogeneity) while balancing long-term changes. What we would expect to see is a very high \(R^2\) when the pre-year is the year right before the retrofit, 2010, (lots of short term bias or signal not associated with the retrofit). The next highest \(R^2\) we would expect to see is to use the year right before that, 2009 (avoiding long-term changes).

Running a single-factor analysis of variance on the between-group source of variation yields a p-value \(< 0.05\) (0.007). This essentially conveys that there is a significant difference between the correlation between pre-use and the change in energy use after retrofit using different years as the base pre-year, (between 2007 and 2010) while keeping the savings estimate constant. Furthermore, this p-value jumps to almost 0.3 when the year of interest is excluded. This indicates that there is a significant difference between running the analysis with and without the year used in the savings calculation as the pre-year. Another way to think about this is that we removed at least some of the endogeneity by removing the year used to calculate savings as the pre-use baseline pre-year (either 2009 or 2010). Figure 2

![Figure 2](image)

**Figure 2: R² of Regressions for savings against pre-use for weather normalized annual heating consumption over total floor area (E_{heating}/A_{floor}) as a function of pre-year.**

This result confirms that the endogenous component of the relationship between pre-use and the change in use is significantly reduced when the year of interest is excluded from the analysis. In addition, the degree of the relationship that can be attributed to the correlation between the pre-use and the error in the savings (the endogenous component) can be estimated. Comparing each of the years, 2009 and 2010 including and excluding the years of interest, the average difference between the R² values is 0.15, indicating that 15% of the variance may be attributed to
endogeneity. Taking this result into consideration moving forward, the rest of the analysis will focus on evaluating the metrics based on (2012-2010) vs. (2009).

With this analysis of endogeneity completed, the original task of the research could proceed: quantify the most appropriate pre-use variable to characterize the potential for retrofit, or retrofit savings. Again, the four weather normalized annual energy consumption metrics examined were:

1) **heating** energy consumption divided by total floor area \(\frac{E_{heating}}{A_{floor}}\),
2) **total** energy consumption divided by total floor area \(\frac{E_{total}}{A_{floor}}\),
3) **heating** energy consumption divided by total home surface area \(\frac{E_{heating}}{A_{surface}}\),
4) **total** energy consumption divided by total home surface area \(\frac{E_{total}}{A_{surface}}\).

One of the serious limitations of the analysis is in the calculation methodology of the home surface area. All that is available with respect to home geometry is the number of floors and the summed floor area of the home. There is no information on home aspect ratio (length vs. width), or the height of each floor (generally standard at 8-9ft). Furthermore, the calculation for surface area becomes increasingly important as the number of floors gets higher. Attic insulation remains constant for homes with equal areas per floor, but the amount of total leakage, and the amount of wall insulation are a function of the total surface area through which heat and air can travel. As such, two different aspect ratios were evaluated, 1:1, and 1:1.5, and the analysis was repeated for each metric individually by number of stories. Almost all of the homes were 1 story, 1.5 stories, or 2 stories.

Table XXXX below shows that while the results are interesting in the differentiation of accuracy by floor, the primary conclusion is simply that using floor area to normalize the results always yields higher \(R^2\) values than using surface area.
<table>
<thead>
<tr>
<th>2012 vs. 2010</th>
<th>Aspect Ratio: 1.0</th>
<th>Aspect Ratio: 1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{heating/Afloor}$</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>$E_{total/Afloor}$</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>$E_{heating/Asurface}$</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>$E_{total/Asurface}$</td>
<td>0.20</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 2: R-Squared Values for each metric of interested where the metric savings from 2010 to 2012 are plotted against the baseline pre-year of 2009 for two different assumptions of home aspect ratio.

This result holds true regardless of assumed aspect ratio and number of floors. The result is 0.11 higher when using an aspect ratio of 1.5 and 0.08 higher when using an aspect ratio of 1.0, when considering all categories of floors (1 floor, 1.5 floors, and 2 floors). The aspect ratio of 1.5 yields slightly higher (~2%) results for homes with more than 1 story, and is identical for homes of one-story. Note that all p-values for all regressions are < 0.001. These results make it clear that despite the logic behind using surface area as a metric, too much variability is introduced in the assumptions made to calculate the surface area using number of floors and total floor area alone.

The normalized annual heating consumption yields a slightly lower $R^2$ value than the normalized annual total consumption. As such, it is concluded that the simplest metric, with the least amount of inherent added assumptions, is the normalized annual total gas consumption divided by total home floor area is the best metric to evaluate savings potential.

2.3 CONCLUSIONS FOR PHASE ONE
The goal of this phase was to quantify which pre-usage metric was most indicative of retrofit savings. In order to do so, the problem of endogeneity in using pre-usage data to predict post-use savings was explored and a simple solution found. The analysis then proceeded from this point, using data from two years before the
retrofit took place as baseline pre-usage data thus removing the endogenous component of the signal. It was found that the highest $R^2$ values came from using the simplest pre-use metric: normalized annual total gas consumption divided by total home floor area. This was true regardless of the aspect ratio used to calculate surface area, or the number of floors in the home. As such, the next phases of research will proceed from this point, attempting to build a model to predict for this metric.
3. PHASE TWO: PREDICTION OF PRE-RETROFIT METRIC

3.1 INTRODUCTION TO PHASE TWO
This section details the process and results from step two of what was envisioned as a three-step research process. To reiterate, the overall research goal was to predict a home's candidacy for retrofit using only a combination of demographics and home-characteristics available for any home in the United States. As stated previously, this must be done in multiple connected steps with datasets of increasing levels of detail.

Phase Two of the research asked how do home characteristics and demographics alone predict for pre-retrofit energy consumption, or simply, can these variables alone identify the top energy users without any utility information at all? This was accomplished by looking at 750,000 homes that had not undergone retrofit but for which utility data was available. These homes also had a significant number of other marketing variables available to the research that were assessed for their predictive power.

3.2 MOTIVATION FOR PHASE TWO
The motivation for trying to develop a model that uses only demographics and home characteristics to model potential for retrofit is to maximize the total energy savings per dollar spent in a given utility program. Right now, as mentioned in the introduction, too much time is spent auditing homes that do not have potential. The most effective retrofit programs are those with targeted programs. Furthermore, but not addressed as a part of this work, is the fact that too much time is spent visiting homes with residents who will not move forward with work. The Warm Choice program in Ohio is a prime example of an effective program that uses targeting to succeed. Warm Choice is a low-income weatherization program that is in its 10th year of operation and has consistently saved approximately 500 customers per year over 25% in heating costs. The program targets customers who are 80% below the median income and offers significant rebates to this customer base for home energy improvement. In addition, priority is given to the highest
residential energy users, and those that spend the greatest percentage of their income on their energy needs. In addition to providing priority to the homes that use the most and need the most, the program targets the elderly, persons with disabilities and family with children.

Using utility data to target retrofits (particularly the highest users for a given comparison group of homes) is ideal, but not always possible given privacy concerns. Furthermore, while certain states require utilities to annually save a percentage of their energy portfolio through energy efficiency, these programs often operate on a deemed savings model. This disincentives actual savings, and incentives the maximum number of measures installed, each with a calculated “deemed savings.” As such, since utilities can recover program costs from state-funds or ratepayers, and meet the energy efficiency standards required of them, there is a lack of incentive to design programs that actually save the most energy. In the ideal, effective programs using utility data would be designed hand-in-hand with the utility. This work operates under the assumption that this is difficult to do and that an ideal, targeted program could work either with or independently of the utility.

3.3 VISION FOR USE OF PHASE TWO

Current utility energy efficiency programs offer the same audit to every customer. There is no audit “triage” and the same amount of time and money is spent at homes that clearly do not have customers and/or homes with potential. With a model that tiered priority for those customers with the greatest home potential, more time and money could be spent on the high potential customers, going through the home in detail. Or, the audit could be done in conjunction with the work: a truck shows up with the auditor and the retrofit contractors at the same time after a phone call with customer indicating the level of incentive available. A short visit could be scheduled with those customers with a low-likelihood of savings. In short, the auditor would be there to confirm suspicion and direct work, or simply, to install a few light bulbs and showerheads. Having a sense of the likely highest savers would also help to
structure more effective policies, programs and incentives. For example, if a customer called in asking for an audit, a few questions on the phone and a good model could save a tremendous amount of time and money for all involved. Efficiency contractors could also be empowered to operate independently of the utility. Given the chance to visit only the homes needing a retrofit, each job would be more likely to yield a paycheck. Finally, while this work would never claim to operate in place of an audit, work of this nature is the first step toward the evolved commercial equivalents, FirstFuel or Retroficiency model of a low- or no-touch audit.

3.4 OVERVIEW OF PHASE TWO PROCEDURE: DATA PREPARATION

This work was dominated (as seen below) in preparing the dataset for analysis. The procedure for this phase of the research took the following form, each step of which will be further discussed in the sections to follow.

1) Clean the raw utility usage data
2) Merge the utility usage with weather data for nearest weather station
3) Run the rolling B-PRISM program on all utility data available (2007-2013)
   a. Run B-PRISM on just the most recent years (2011-2013)
   b. Run B-PRISM on year by year basis
   c. Compare the results of the rolling PRISM evaluation with year by year PRISM
4) Clean the PRISM results, drop outliers
5) Clean the Experian data
   a. Data dictionary from Experian to recode numeric indicators into text
   b. Filter the data and add variables of interest
   c. Questions of accuracy in the Experian data: area
6) Merge the cleaned Experian data and the cleaned PRISM results
7) Clean the final file
   a. Remove homes that received audit or retrofit
b. Drop outliers, non-physical results

8) Re-examine the choice of dependent variable: NAC/Area

9) Run various data-mining algorithms to determine variables of importance in predicting NAC/Area, NAC, NAHC/Area, NAHC

10) Of critical variables, determine co-linearity and/or co-dependence
   a. Remove redundant variables

11) Choose final set of independent and dependent variables for modeling

12) Test various modeling techniques for predicting dependent variable using final set

1) **Clean the raw utility usage data**

Available to this phase of the research was a large utility billing file for approximately 1,000,000 customers located in the Columbia Gas of Ohio service territory. This file included:

1) an anonymized unique identifier for each customer
2) an anonymized unique identifier for each home (premise)
3) date of the reading
4) usage for the billing period
5) an indicator of whether the meter reading was estimated or actual
6) an indicator of the premise status (whether current customer)

The way that the data is originally provided in its raw form is the number of therms per month the customer was billed for that month. This does not always mean that a meter reading occurred on the billing date. For example the dataset could include a number of therms for January, February, and March but February could be estimated by the utility based on prior usage and the usage that has already occurred during that year.
The cleaning process (code provided in Appendix A) involved removing duplicate observations, conflicting meter readings for the same date, and off-cycle readings (readings that occurred more than 40 days or less than 20 days from one another indicating the customer potentially not in the home or issues with meter). The only reason that a reading might be off-cycle is if someone new moved into the home which began a new billing cycle, or the home was vacant for a period between customers. Otherwise, this is an indication that there is something wrong with the utility meter.

When the usage was estimated and not the actual usage for the given month, the usage was accumulated across the estimated meter reading. This essentially means that if the usage in March was estimated it gets added to the next meter reading that is not estimated. So if the meter said 100 therms in March, there is an estimated bill for April at 50 therms, and the meter said 300 therms in May, the code would drop the estimated meter reading after adding the usage to the May bill, leaving 100 therms in March and 350 in May.

This leaves a file with only actual meter readings, which is essential when using PRISM to calculate the usage per heating degree day.

2) Merge the utility usage with weather data for nearest weather station

The next phase of the procedure involves mapping the cleaned usage data to a file containing the zipcode and associated weather station with that zipcode. This is necessary because when the PRISM code runs, it maps the name of the weatherstation to a weather file in a sub-directory, and pulls the daily average temperatures for that location. With the daily average temperatures the program can calculate the number of heating degree days elapsed between the meter readings.
3) Run the rolling B-PRISM program on all utility data available (2007-2013)
   a. Run B-PRISM on year by year basis
   b. Run B-PRISM on just the most recent years (2011-2013)
   c. Compare the results of the rolling and year-by-year B-PRISM analysis

PRISM and Refinement to B-PRISM
As previously described, PRISM, the Princeton Scorekeeping Method, is a technique that regresses the energy use per day of a home against the number of heating degree days elapsed during one meter reading and the next. The slope of this line is the Use/HDD, and multiplied by a long-term averaged number of heating degree days across an entire year yields the weather-normalized heating usage for an average year. The intercept of this line is gas use per day when there is no heating need (the baseload use/day) that might be attributed to dryers, water heaters, and stovetop ranges. Since heating degree-days are calculated based on a reference temperature which is assumed the balance-point of home, the model is extremely sensitive to the temperature chosen as the reference. The magnitude of difference between two different balance points is shown in Figure 3.

The balance point is the temperature outside of the home at which the heating system kicks-on to maintain a steady set-point temperature defined by the occupants of the home (generally around 65F or 70F during the winter). The balance point temperature differs from the set-point temperature defined by the occupants due to internal gains within the home (people and electronics) and any passive solar heating. Since the model is sensitive to the balance point and this number differs from home to home, PRISM is a variable-base degree-day method, meaning multiple possible balance-points are tested for fit in the regression and the temperature that yield the best fit (highest R²) is taken to be the home-balance point.
Figure 3: Data for a Boston home showing the different linear fits pre and post retrofit for two different assumed balance point temperatures.

The B-PRISM method, coded by Michael Blasnik, and provided in Appendix B, differs in that it does not rely upon $R^2$ alone to define the best balance-point temperature. Instead, a score is developed for each balance point that essentially details the likelihood of the estimated model, or a generalized $R^2$. The likelihood is a function of a given Bayesian prior, the number of readings used in the model, and the $R^2$ of the fit. The Bayesian prior is essentially what we know from experience, or in this case, from fitting the model to hundreds of thousands of homes. To choose the prior relying on the data being fitted, the average balance point can be calculated without the Bayesian method and then iteratively put back into the more sophisticated model. Using the prior shifts the unlikely (those with high standard deviations) balance points in the direction of the prior, essentially penalizing temperatures that are significantly different than what one might expect to find. In words, the “score” or likelihood of the model given the balance point in consideration is the likelihood...
of the model given the Bayesian prior and its standard deviation multiplied by the
likelihood of the linear regression using the given balance point. Or,

Likelihood of given balance point =

$$\log(T_{balSD}N(T_{balprior}, T_{balanced}, T_{balSD})(1 - R^2)^{-\frac{N_{reads}}{2}}$$

Where,

N= Normal Distribution Function
T_{balSD} = standard deviation of the balance point temperature
T_{balprior} = prior
T_{balanced} = estimated balance point in question
N_{reads} = number of readings utilized

Classically, the PRISM method relies solely on the R^2 and the number of readings to
determine the balance point. This leads to balance points that might be non-physical
but fit a model well. By using a Bayesian method, prior experience is guiding a
statistical decision.

Rolling B-PRISM or PRISM year by year

There are two ways that the PRISM model might be applied to multiple years of
data. PRISM is generally used to evaluate the effectiveness of retrofit and one year of
data pre-retrofit is regressed, and one year of data after the retrofit is regressed and
the results are compared. If more than one year of data is available, the model is
more likely to be correct because the regression is based upon more points overall.
The more readings available, the less likely any given reading significantly
influences the result. However, the further back or forward one goes in time from
the time of interest (generally the date of the retrofit) the more likely it is that
something else has changed in the home, a baby was born, a teenager takes longer
showers, or new residents moved in. As such, the choice of the ideal number of
readings to use to generate the weather normalized annual consumption is not as clear as using as much data as possible. In the case of this research, the Experian data was updated right before it came to me on July 2013, a snapshot in time. Because this portion of the analysis does not rely upon pre/post data, the July 2013 data pull represents the date at which the rest of the data is current (the demographics and home-characteristics). The utility data available goes from January 2007-March 2013, and so the question proposed was whether to use all of the data available, the most recent few years, or to analyze individual yearly PRISM results and take the result closest to the data pull. The way this was answered was to look at the individual most recent yearly results and see if there were large fluctuations from one year to the next indicating something significant had changed in the home not associated with weather conditions. B-PRISM was run on 2011 and 2012 and if the range of NAC values divided by the average value of NAC was greater than 35% the observation was dropped. Finally, B-PRISM was run on a rolling basis for 2011-2013 for those observations that remained after dropping observations with too much fluctuation between 2011 and 2012.

4) **Clean the B-PRISM results, drop outliers**

The results of the B-PRISM analysis were cleaned based on weather criteria, model goodness of fit, and by taking out physically unrealistic values for residential homes. Any home that had less than four meter readings per year during the period of analysis was dropped, both for the rolling analysis and the yearly analysis.

Based on the goodness of fit of the model, any home meeting the following criteria was dropped:

- \( R^2 \) fit less than 0.7
- NAC standard errors greater than 20% of the NAC

The period spanned by the data spanned needed to have enough weather variation to indicate that all seasons were fairly represented. This is important because each point on the graph represents an individual use/day and hdd/day. If the number of
heating degree days per day is similar for each reading period, (lack of weather variation), the line is fit through a vertical stack of clustered points. Mathematically, the difference in the number of heating degree days per day between the coldest and warmest data points used in the model needed to be greater than or equal to the yearly long-term average heating degree days per day. In addition the presence of the coldest months of the year had to be demonstrated. If the coldest months are not present, there is no guarantee that a heating season, where the highest use/day would be seen, is present in the data. Without a heating season, there is not a good representation of how the home performs on extreme weather days. This translated to 50% of the yearly long-term average heating degree-days needing to be present in the analysis period. For example, for Boston, at around 5000 HDD/year on a long-term averaged basis, 2500 HDD would need to be present during the span of time included. With a mean value of about 850 therms/year for a normalized annual consumption value. pre-cleaning observations that had NAC more than five standard deviations away from the mean (NAC > 3500) were dropped as well as those with negative values for normalized annual heating consumption. This value is so high to offer the possibility that a home is indeed an extreme user of energy (the exact homes the analysis is trying to identify). Any observation with a negative baseload estimate was also dropped. Finally, the two datasets, including observations without extreme variation between 2011 and 2012 and those with good PRISM results from running the analysis on the rolling years of 2011-2013, were compared and those that met both criteria were kept in the final dataset.

The original file of individual homes provided had almost 1.5 million customers. After running B-PRISM year by year and removing observations with too great a fluctuation between 2011 and 2012, and removing results that did not meet the criteria listed above for 2011-2013, 1.3 million homes were still left for the analysis.

5) **Clean the Experian data**
   a. **Data dictionary from Experian to recode numeric indicators into text**
   b. **Filtering the data and adding variables of interest**
c. Questions of accuracy in the Experian area

As a general Stata reference, numeric missing values are represented by very large positive numbers. As such any actual number is less than a missing value. In order to exclude all missing values, the code would read "less than missing" or "<." where the period represents a missing value. To include missing values, the value of the variable is greater than or equal to missing.

Many times, recoding of a variable occurs to flag certain observations meeting specified criteria. A variable is created, which originally is simply a list of null values. The variable is then replaced, based on the criteria specified, making sure another value has not already been assigned based on previously detailed criteria.

Data dictionary from Experian to recode numeric indicators into text

Working with the Experian data was one of the most significant and difficult tasks throughout the research. The original set of variables from Experian included the following (presented in Table 3).
<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Customer ID</td>
<td>30 Family/General-Mag Byr Cat</td>
<td>59 Dwelling Type</td>
</tr>
<tr>
<td>2</td>
<td>Number of Adults in Household</td>
<td>31 General-Contribtr Cat</td>
<td>60 Length of Residence</td>
</tr>
<tr>
<td>3</td>
<td>Aged Parent In Household</td>
<td>32 Health &amp; Inst-Contribtr Cat</td>
<td>61 Dwelling Unit Size Code</td>
</tr>
<tr>
<td>4</td>
<td>Presence of Child Age 0-18</td>
<td>33 Political-Contribtr Cat</td>
<td>62 Combined Homeowner</td>
</tr>
<tr>
<td>5</td>
<td>Number of Children (&lt;18)</td>
<td>34 Religious-Contribtr Cat</td>
<td>63 Estimated Current Home Value</td>
</tr>
<tr>
<td>6</td>
<td>Presence of Child Age 0-3</td>
<td>35 Sweepstakes</td>
<td>64 Home Stories</td>
</tr>
<tr>
<td>7</td>
<td>Presence of child Age 4-6</td>
<td>36 Do It Yourselvers</td>
<td>65 Home Bath</td>
</tr>
<tr>
<td>8</td>
<td>Presence of Child 7-9</td>
<td>37 News &amp; Financial</td>
<td>66 Home Bedrooms</td>
</tr>
<tr>
<td>9</td>
<td>Presence of Child 10-12</td>
<td>38 Photography</td>
<td>67 Home Exterior Wall</td>
</tr>
<tr>
<td>10</td>
<td>Presence of Child Age 13-15</td>
<td>39 Odds &amp; Ends-Mail Responder</td>
<td>68 Home Heat Ind</td>
</tr>
<tr>
<td>11</td>
<td>Presence of Child Age 16-18</td>
<td>40 Miscellaneous-Mail Responder</td>
<td>69 Home Fireplaces</td>
</tr>
<tr>
<td>12</td>
<td>I1 Mail Responder-Indiv</td>
<td>41 Education - Individual 1</td>
<td>70 Home Air Conditioning</td>
</tr>
<tr>
<td>13</td>
<td>Multi-Co Direct Mail Resp - HH</td>
<td>42 Home Business Indicator</td>
<td>71 Year Built</td>
</tr>
<tr>
<td>14</td>
<td>Multi Buyer</td>
<td>43 I1 Business Owner Flag</td>
<td>72 Date of Warranty Deed</td>
</tr>
<tr>
<td>15</td>
<td>Gift/Gadget-DM Merc Byr Cat</td>
<td>44 I1 Gender Code</td>
<td>73 Transaction Amount</td>
</tr>
<tr>
<td>16</td>
<td>Col/Spc Food-DM Merc Byr Cat</td>
<td>45 I1 Birth Date CCYMM</td>
<td>74 Type of Purchase</td>
</tr>
<tr>
<td>17</td>
<td>Books-DM Merc Byr Cat</td>
<td>46 I1 Combined Age</td>
<td>75 Mortgage Amount</td>
</tr>
<tr>
<td>18</td>
<td>Garden/Farm-DM Merc Byr Cat</td>
<td>47 Enh Est HH Income Amount</td>
<td>76 Home Building Square Footage</td>
</tr>
<tr>
<td>19</td>
<td>Crafts/Hobby-DM Merc Byr Cat</td>
<td>48 Ethnic Insight Match Flag</td>
<td>77 Home Property Indicator</td>
</tr>
<tr>
<td>20</td>
<td>Female Orient-DM Merc Byr Cat</td>
<td>49 Ethnicity Detail</td>
<td>78 Home Swimming Pool Ind</td>
</tr>
<tr>
<td>21</td>
<td>Male Orient-DM Merc Byr Cat</td>
<td>50 Language</td>
<td>79 Home Base Square Footage</td>
</tr>
<tr>
<td>22</td>
<td>Upscale-DM Merc Byr Cat</td>
<td>51 Religion</td>
<td>80 Home Land Square Footage</td>
</tr>
<tr>
<td>23</td>
<td>General-DM Merc Byr Cat</td>
<td>52 Ethnic Group Code</td>
<td>81 Total Enhancement Match Type</td>
</tr>
<tr>
<td>24</td>
<td>Health &amp; Fitness-Mag Byr Cat</td>
<td>53 e-Tech Group</td>
<td>82</td>
</tr>
<tr>
<td>25</td>
<td>Culinary Interest-Mag Byr Cat</td>
<td>54 Country of Origin</td>
<td>83</td>
</tr>
<tr>
<td>26</td>
<td>Garden/Farming-Mag Byr Cat</td>
<td>55 Enh Est Household income V4</td>
<td>84</td>
</tr>
<tr>
<td>27</td>
<td>Religious-Mag Byr Cat</td>
<td>56 Household Composition Code</td>
<td>85</td>
</tr>
<tr>
<td>28</td>
<td>Male &amp; Sport Orien-Mag Byr Cat</td>
<td>57 Occupation - Group 1</td>
<td>86</td>
</tr>
<tr>
<td>29</td>
<td>Female Orient-Mag Byr Cat</td>
<td>58 I1 Occupation</td>
<td>87</td>
</tr>
</tbody>
</table>

Table 3: Experian variables provided for analysis

Experian uses modeling techniques to enrich data when there is no confirmed record of the individual or household in question. The enrichment could be performed with a match on a household, an individual, or a geographic range. When the enrichment is done by geography, Experian takes the average data available for the geographic area where the home is located.
Experian also has various ways of encoding the accuracy of the enriched data. For example, the presence of a child and the presence of a child between a certain range of ages in the household could be “K” known, or “I” inferred. If the presence of the child is inferred, Experian has used a model to predict the likelihood that a child of that age range is present. Data that was inferred was kept as long as there was a household or individual rather than geographic match.

In the table I1 stands for “Individual 1” or, the assumed head of the household. Experian generally has details on more than one individual in the house, and a full data enrichment would include the data on each of the individuals, signified I1-I8. For each of the variables discrete decisions had to be made by the author about the level of detail kept in the dataset, how the data would be stored (as the file was quite large), and how to make the results physically meaningful. This was necessary such that when the modeling proceeded, it would be possible to have a sense of how logical the results were. For example, if religion came out as extremely important in predicting energy consumption, knowing that religion A5 uses more energy than B12 is hard to verify.

In addition to the data, a 75 page “enrichment detail” was provided through Experian. This detail mapped the relationship between the numeric or alphanumeric codes provided in the dataset and the actual textual meaning of those codes. All data was stored numerically, but showed up in the final dataset in logical textual information in two different ways,

1) By defining labels for each variable,
   a. for example, 0 "None" 1 "One" 2 "More than One"
2) By creating matching files (or lookup tables) that merged the large file with a file that contained the Experian code in one column and the associated text in another (see Figure 4 for an example)
   a. The text to alphanumerical matching came from the Enrichment detail
b. Once matched, the textual data was turned back into numerical codes through an encode command, but the text was kept for viewing.
c. When the dataset was exported, the option was given to export the text or export the numerical encoding associated with the text.
Experian
A world of insight

NISOURCE CORPORATE SVCS CORP
2591254- JULY 2013 ENRICHMENT

Enrichment Detail Report
Filtered By: Lists: All
Marketing Element: All
Broker: All
Group Level: Job

0108R - Religion
Ethnic Insights Religion indicates the likely religion.

Valid Values:
B = Buddhist
C = Catholic
E = Ethiopian Orthodox
G = Greek Orthodox
H = Hindu
I = Muslim
J = Jewish
K = Sikh
L = Lutheran
M = Mormon
O = Eastern Orthodox
P = Protestant
S = Shinto
X = Not Known or Unmatched

0108T - e-Tech Group
e-Tech derived code that groups ethnicities in general categories.

Valid Values:
A = African American
B = Southeast Asian
C = South Asian
D = Central Asian
E = Mediterranean
F = Native American
G = Scandinavian
H = Polynesian
I = Middle Eastern
J = Jewish
K = Western European
L = Eastern European
M = Caribbean Non-Hispanic
N = East Asian
O = Hispanic
Z = Uncoded

Figure 4: Example extract from Experian enrichment file
In what follows, each of the variables will be detailed more fully, and how they were handled in the coding will also be described where data handling is not fully evident from the text provided in Table 4.

An elderly parent in the house was coded either as Yes, and stored with a value of 1, or left empty as missing. The number of children and number of adults was already a numerical variable, and the number of total occupants in the household was simply these two numbers added together.

As previously described, the presence of children of different age ranges (variables 6-11 in Table 3) could be yes/no/unknown and the answer inferred or known. This was coded as coded with a 1 or a 5 preceding the letter Y, N, or U. The data was recoded to be in a binary yes/no format, where a string search command essentially looked at the string provided and searched “Y”. If the string Y was not present, a zero was returned, if the string did contain Y a 1 was returned. All 0 values were recorded as no and all 1 values were recorded yes, regardless of whether the data was inferred or confirmed and the data stored in the numerical binary format.

The individual mail responder variable (Variable 12) indicated whether the head of the household responded to direct mail and is either “Yes” (1) or "Unknown" (missing). Multi-Co direct mail responder, the next variable, indicated whether the head of the household responded to direct mail from multiple companies, multi-buyer indicated whether the head of the household purchased from direct mail across multiple categories. Variables 15-23 indicate direct mail merchandise buyer categories, indicating the number of responses to direct mail; variables 24-30 indicate the number of direct mail responses to each magazine buyer category; 31-34 represent the number of responses by direct mail in each of the contribution categories; and finally, 35-40 again, direct mail responses to each of the areas of interest. Each of these was listed as numeric values 0-9, and each was recoded as 0 "None" 1 "One" 2 "More than One".
Education could be less than high school, high school, some college, college, or graduate degree, and indicated the education level of the head of the household. This variable was stored with 1 being "High School Diploma" and the integers proceeded upwards to 4 "Graduate degree". However, 5 signified "Less than a high school diploma" and was recoded to be zero. This variable was derived from self-reporting surveys and/or occupation, or predicted through modeling. Education was stored preceded by a 1 or a 5 indicating the level of accuracy of the estimate where 1 indicated “extremely likely” and 5 indicated “likely.” In order to get the values coded to be usable for modeling, the modulus of the original value was taken with respect to 10, leaving the remainder of the division by 10. For example, “likely less that a bachelors” would be 53, and the modulus of 5 with respect to 10 would be 3, “Bachelor Degree.”

Home business indicator is a flag for a home business, coded as yes or unknown (missing). Business owner flagged whether the head of the hold owned their own business, again, coded or yes or unknown (missing).

The gender code indicated whether the head of the household was male or female, unknown, or both in the case that the head of the household is a “Mr. and Mrs.” combination on records. Unknown values and those stored as “both” were set to missing. Birth date is the birthday of the head of the household, combined adult age is the age of the head of the household. Both of these variables could be exact or estimated and both were considered in the analysis; unknown values were set to missing.

Head of Household income was recoded to be the Experian value multiplied by $1000 as all incomes were stored numerically in the thousands. The mortgage of the home was also refined to reflect its original encoding in the thousands.

Similarly, the ethnic insight match flag indicated whether there was a match to the ethnic insight tool which processes and analyzes last names, and connects them to
177 ethnicities, 33 groups of ethnicities, probable religions, and language preferences. This is not used as the sole indicator of these categories but serves as one cross-reference that checks zipcode and first name and updates if more accurate estimates are possible. The ethnicity detail included 191 possible ethnicities, and ranged from Botswanan to Jewish. The ethnic group had 33 categories and was wider ranging, including all of Western Europe in one single category. The categories were not based on the percentage of the population falling into each as for example, "Hawaiian" and "African" each had its own category. Language had 84 possible languages in the match file ranging from "Igbo" to "Chinese." Religion had 14 categories and country of origin had 22 possibilities. E-Tech group was another categorical ethnicity descriptor that was slightly broader than the ethnic group and had only half as many possibilities with primarily geographic indicators such as "South Asian" and "Middle Eastern."

Household income had the same format as the income of the head of the household. The household composition was an indicator of the family members living in the household and the sex of the adults. The occupation group included only 6 categories, "Professional/Technical," "Sales/Service," "Farm Related," "Blue Collar," "Other" or "Retired" while the full occupation variable had 46 much more descriptive indicators from "Professional Driver" to "Dental Hygienist."

Dwelling type indicated whether the residence was single family, multi-family or condo of less than 5 units, multi-family unknown, unknown or a post-office box. The length of residence indicated how many years the residents had lived there and dwelling unit size indicated how many units were present ranging from single-family to 101 or more units. Combined homeowner indicated whether the residents owned the home, rented the home, probably rented, or the percentage certainty the resident probably owned the home. Home Stories, was the number of stories in the home, bath, the number of baths, bedrooms the number of bedrooms. Both the number of stories in the home and the number of bathrooms were stored as the
actual number multiplied by ten. These variables were thus divided by 10, and rounded to the nearest half (35 would now be represented as 3.5).

Home exterior wall indicates the type of material on the exterior wall of the home and the home heat indicator indicates whether a furnace, electric, heat pump, etc. is used to heat the home. "Home fireplaces" indicates the number of fireplaces present in the home, home air conditioning whether central air, a window unit, etc. is present to cool the home. The year built came from tax assessor databases or was modeled by Experian. The first position of the value indicated the confidence of the estimate as 1 (Extremely likely) or 5 (Likely). Positions 2-5 contained the actual year built estimate. As such, in order to be left with just the year of interest, the integer in the 10,000’s position was extracted by dividing by first dividing by 10,000 and rounding to the nearest integer. For example if the Experian value were 51956, diving by 10,000 would leave 5.1956, and the nearest integer would be 5. Next, this value multiplied by 10,000 was subtracted from the Experian value, (51956 – 5*10,000) leaving just the year of interest, 1956. The certainty of the estimate was not considered in the analysis.

Type of purchase indicated whether the home was sold new or resold or whether this was unknown. Home building square footage estimated the total square footage of the home in hundreds, the base square footage the square footage in the hundreds of one single home story, and land square footage the total square footage in the hundreds of the land area upon which the home was built and allocated to the lot. The property indicator showed whether the home was residential, a condo, an apartment, a commercial condo, agricultural or a mobile home. Finally the swimming pool indicator showed whether a swimming pool was present otherwise set to unknown.

Matching files were created to encode variables with more than approximately 10 levels. These included the occupation code, country of origin, ethnic group code, language, and ethnicity detail.
Filtering the data and adding variables of interest

Once the data was matched to logical text descriptors, filtering proceeded based upon physical criteria. In homes that were designated single family but had more than 4 stories, the number of stories was replaced with null. All homes with a stored value less than $1000 had the value replaced as null.

Any home where information was unavailable from Experian or there was a geographic match rather than an individual or household match was dropped because of the uncertainty associated with the data enrichment. There were 106,000 homes in the Experian dataset that were dropped based upon this condition.

Variables of interest for the data-mining modeling were also created. For example, a variable called vintage recoded the values of year built into pre-world war II, (built before 1950), 1946-1979, and 1980+. The year 1979 is significant as a marker in construction because of the second oil crisis, the first of which occurred in 1973. The first oil crisis was the result of an oil embargo to the Unites States from the Arab world, and the second, the Iranian Revolution that decreased the overall oil supply by approximately 4%. Construction practices (and building codes) changed after this period to reflect a greater emphasis on energy efficiency. A second variable created indicated simply whether the home was built after or before the energy crisis and was stored as a binary “Yes”/ “No”.

A variable was generated to indicate whether the home was a single family, and any home that was not a single-family was removed from the analysis. Experian data was available for approximately 1.1 million homes and over 80% of them were designated single-family.
Considering area

There were three areas in the original Experian dataset: home base square footage, home land square footage, and home building square footage. The home building square footage estimated the total square footage of the home. As the analysis had available a dataset with both actual audit data and the estimated Experian data through the HPS dataset, the accuracy of the home building square footage was tested. It was found that there were significant errors in the home building square footage estimate (Figure 5 below) where the HPS data was larger by a significant margin over 80% of the time. The distribution of the errors was also concerning: whereas there was no difference between the areas some of the time, the normal distribution of the difference was significantly right-shifted and centered about approximately 1000ft². As such, multiple manipulations of the three area variables were tested for their proximity to the “actual” area. It was found that the area closest to that in the HPS dataset was the base square footage of the home multiplied by the number of stories estimate. The average difference between this estimation and the audit data was 100ft² but even this modified variable ranged between over -2000ft² and 2000ft². In addition, even though the area errors were consistently positive, indicating that the square footage from the audit was almost always higher, this was not always the case. The ideal would be to add a consistent shifted area back on to the Experian estimate to correct the inaccuracy but the errors were not consistent enough to find this shift.
Figure 5: Distribution of differences between the original and modified Experian area estimates relative to the audit area estimate. The y-axis is scaled such that the sum of the area of the bars is equal to 100%.

The possible inaccuracy in the area data is a critical. Area is one of the most important predictors of energy consumption in the home (as will be seen as the variables of importance are chosen). In addition, as the dependent variable for the model is energy intensity, (normalized annual energy consumption divided by area) the error in this area gets passed on to the variable of interest. As audit data is unavailable for all but 6,000 homes, there is no mechanism by which to catch these significant possible estimation errors in the Experian dataset except to filter the areas on a physical basis. Furthermore, the inaccuracies in the area call into
question the accuracy of all of the Experian data. As the modeling proceeded, this question of data accuracy became increasingly salient.

6) Merge the cleaned Experian data and the cleaned PRISM results
Once these two datasets had been prepared, the estimated yearly energy use data from the PRISM model and the recoded Experian data were merged. In the Experian dataset, all data was given based on a Customer number while in the monthly energy data, all data was given based on a Premise (home) number. In order to merge these files, a crosswalk file had to be generated that matched each customer to a premise. It was possible that a given customer was matched to more than one home (people move). If more than one home was associated with a given customer, this observation was dropped as the Experian data contains both home characteristics and demographics. This merged file contained all the recoded Experian data and one value for the normalized annual consumption that was performed on a rolling base, taking all of the energy data available from 2010-2013.

7) Clean the final file
   a. Remove homes that received audit or retrofit
   b. Drop outliers, non-physical results

The HPS homes that were used in Phase 1 of the analysis received an audit or retrofit during 2011. As such it was essential that these homes were removed from the analysis as this portion of the analysis focused on "potential" for retrofit. The HPS home premises were matched to the large file and any matches were removed from the analysis. Any home that was missing an area value and/or had an estimated area less that 500ft² or more than 7000ft² was dropped.

8) Re-examine the choice of dependent variable: NAC/Area

One of the issues that came up repeatedly in Phase 2 was whether energy intensity was the right variable to use as a dependent variable in the modeling. The question
at hand was, "Does this home have potential for retrofit?" where the definition of potential for retrofit meant that the home energy consumption was unusual given what was known about the home.

The concept behind using the intensity metric was to normalize for area, essentially asking the question, "Given the home's area, is the energy consumption unusual?" However, this was complicated by the fact that energy intensity did not actually normalize for area. As shown in Figure 6, smaller homes tend to have higher energy intensities, and there is less value with respect to overall energy savings in retrofitting a small home.
Figure 6: Home energy intensity vs. area showing that dividing by area does not normalize for the home area, or, that energy intensity is a function of area.

One suggestion for dealing with this fact was to use the knowledge that the relationship of the envelope of energy consumptions with floor area was exponential and take both a mean centered area and a mean centered area squared when examining the relationship between energy intensity and area in the models. This would work to better normalize energy consumption because as can be seen in the graph above, the relationship between energy intensity is area is exponential. The reason to mean center the area is both to remove the co-linearity between the two terms, and also to have the y-intercept of the regression model make intuitive sense. A correlation matrix between area and area squared indicates almost perfect col-linearity: the coefficient is 0.95. When the area is mean centered on the other
hand, the coefficient drops to 0.65. If the model predicts for energy intensity using area, the y-intercept will be the energy intensity when the home area is equal to zero. When the area is mean-centered on the other hand, the y-intercept is the energy intensity at the area mean. The result is more physically meaningful: what is the energy intensity of the home when the home is an average size?

Furthermore, there are certain fixed costs associated with each retrofit such that these retrofits tend to be less cost-effective. If the goal is to save raw energy, then one would solely target big homes that use a lot of energy. However, there are other factors at play in choosing what makes a home a good candidate for retrofit. One of the most obvious is the need of the homeowner, based on the percentage of income going to energy costs. While it might make more sense to predict raw energy and take just the highest users, this is not the way candidates are chosen. The right variable to predict is also complicated by the fact that retrofits save energy in different ways. The savings from shell upgrades, including insulation and air sealing, scale with the surface area of the home. Heating system replacements on the other hand scale savings directly with energy consumption. Homes will save a set percentage of their overall load by upgrading from an 85% efficient furnace to an 89% efficient furnace. Since energy consumption scales with the size of the home, as shown in Figure 7, bigger homes will almost always save more raw energy.
Figure 7: Relationship between raw energy use and area showing that large homes tend to use more energy (graph created from HPS dataset)

These issues will be discussed further in later sections. However, it was decided that the best variable to choose (energy intensity vs. raw energy) was the one that the final models did the best job of predicting. In other words, multiple modeling techniques would be tested with the ultimate set of variables, and the metric the models predicted best would be chosen as the variable to model as the dependent variable.

9) Run various data-mining algorithms to determine variables of importance in predicting NAC/Area, NAC, NAHC/Area, NAHC

10) Of critical variables, determine co-linearity and/or co-dependence
    a. Remove redundant variables
To define which variables should appear in the final model, regression, classification trees, regression trees and random forest were run on the dataset to predict both a continuous normalized annual energy consumption and the energy intensity. After these methods were run and the variables of importance identified, the co-linearity of the significant variables was examined, as shown in Figure 8. The accuracy of the model (the percentage of the variance of the dependent variable explained by the independent variables) is evaluated based on the assumption that the independent variables are independent, meaning that one variable is not just a simple linear combination of the others (for example the mortgage of the home and value of the home).

Figure 8: Scatterplot matrix visualizing the co-linearity of home price, mortgage amount and estimated home value in the Experian dataset.
Regression trees and classification trees use recursive binary splitting to subdivide the data. An algorithm decides the splitting variables, points, and the overall tree structure. The choice is based on the splits that serve to subdivide the data into the most distinct groups possible as a function of their relation to the dependent variable. For example, in a model predicting continuous energy consumption, homes that were built before 1980 could behave differently in their energy consumption than homes built after 1980. In this case, the tree would split the dataset at 1980.

The choice of which variables come first (at the top of the tree) is based upon this metric of distinctness. If the variable is ordered, the algorithm splits at a given value. If the variable is categorical the algorithm subdivides based upon the category. A random forest generates hundreds, thousands or tens of thousands of these trees, testing handfuls of variables at a time. The variables that appear repeatedly in the trees as most predictive rise to the top of the list.

Since the sample was so large, these data mining techniques were run on sub-samples of approximately 10,000 homes each. It was found that the same variables came up repeatedly as important regardless of whether energy intensity or raw energy was predicted. Furthermore, these same variables came up as important in predicting (within the HPS data) whether the home went forward with an audit or retrofit.
In Figure 9, mc_area denotes a mean centered area and mc_area_sq, a mean centered area squared. The percentage of the variance explained in predicting the energy intensity without examining redundancy in the variables provides is 47%.
Figure 10: Random Forest node purity plot examining the variables of importance in predicting raw energy.

In predicting raw energy, the percentage of the variance explained, again not removing the possible redundancy in the variables, is approximately 29.5%, as shown in Figure 10.

Basic regressions were then run with all the variables, first in predicting raw energy, and then in predicting energy intensity using both the mean centered area and the mean centered area squared in the independent variable list. The $R^2$ in predicting energy intensity was indeed slightly better, with a value of 0.46 while the $R^2$ in predicting raw energy was 0.38.
In choosing the variables of importance, the t-scores and p-values were also examined for each variable in the regression results.

The final choice of variables of interest (prior to examining variable redundancy) was based on a combination of the variable importance plots through random forest and the regression t-scores (and associated p-values). The t-score indicates whether the slope of the regression line is significantly different from zero. This value is both a function of the coefficient estimation and the standard error, how much the coefficient might differ from sample to sample. One needs to identify both the significance of the coefficient and the standard error of the estimate as it could be quite high but differ significantly with each handful of the dataset. In multiple regression as examined here, the coefficient on the variable indicates that for every one unit increase in the independent variable there an increase or decrease in the dependent variable equal to the coefficient, holding all other variables constant. The p-value indicates the probability of finding a t-score that far from zero if the null hypothesis is true. This can also be viewed as the statistical significance of the independent variable in explaining the dependent variable.

Once the set of significant variables was compiled, the independence of the cluster of variables was examined using the variance inflation factor. The variance inflation factor indicates how much the variance of a given regression coefficient is inflated because of colinearity with other variables. The variance inflation factor is 1 when the other predictor variables are completely orthogonal to one another. This orthogonality is determined by attempting to predict one variable given the other independent variables in the model. The rule of thumb for the VIF is that if it is greater than or equal to approximately 5-10, there is colinearity between this variable and the others. Variables with high VIF's were examined and the possible culprits of colinearity examined (for example the home mortgage and home value). Once the colinearity was determined, the variable that gave the most amount of information was kept and the other or others dropped.
Adding more variables to the model also removes many of the observations available to the analysis because of missing values. As the final list of variables were chosen, variables were removed one by one, and the number of observations added back into regression was recorded as well as the goodness of fit and root mean square error.

11) Choose final set of independent and dependent variables for modeling

The final non-collinear variables chosen, as shown in Table 4, were the year the home was built, the home value, the number of bathrooms, the number of bedrooms, the number of stories in the home, the age of the head of the household, the household income, the number of occupants, the reduced categorical ethnicity, the home heating index (the type of heating system), education level of the homeowner, the mean centered area of the home and the mean centered area squared. The goal of using the energy intensity as the dependent variable was again to determine whether the home's energy consumption was usual given its area. Since energy intensity is an exponential function of area (see Figure 6), including both terms in the regression should help control for this relationship.

<table>
<thead>
<tr>
<th>Variables Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year Built</td>
</tr>
<tr>
<td>Mean Centered Area</td>
</tr>
<tr>
<td>Mean Centered Area^2</td>
</tr>
<tr>
<td>Number of Stories</td>
</tr>
<tr>
<td>Number of Bedrooms</td>
</tr>
<tr>
<td>Number of Baths</td>
</tr>
<tr>
<td>Ethnicity Category</td>
</tr>
<tr>
<td>Home Value</td>
</tr>
<tr>
<td>Home Heating Type</td>
</tr>
<tr>
<td>Education Level</td>
</tr>
<tr>
<td>Number of Occupants</td>
</tr>
<tr>
<td>Age of Head of House</td>
</tr>
<tr>
<td>Household Income</td>
</tr>
</tbody>
</table>

Table 4: Final variables included in the model after balancing missing values, collinearity and significance
The energy intensity was much better predicted by the available variables and thus was chosen as the dependent variable to model. In full disclosure the model may be better for the same reason that the energy savings was better correlated with pre-use energy when the savings base year and the pre-use year of the data were the same. This is due to the error terms being correlated, or, that when area is inaccurate, so is the energy intensity.

12) **Test various modeling techniques for predicting dependent variable using final set**

3.5 PHASE TWO MODEL TESTING AND DEVELOPMENT

Regression and Residuals
Once the final variables were chosen for the model, and the final dependent variable was chosen to be energy intensity, multiple modeling techniques were attempted to predict energy intensity given the variables of interest.

One of the first questions was whether a binned area needed to be explored because the relationship between the independent and dependent variables was a function of different housing comparison groups entirely. In other words, perhaps homes that are 1000-1500ft² need to be modeled completely differently than homes that are 5000ft²-7000ft². If the question is whether the home uses an unusual amount of energy given what we know about the home, (year built, number of stories, size) keeping all the homes together in one pile assumes the relationships are the same for the entire dataset.

Examining the results of a simple regression analysis using the variables identified from the full dataset, it was clear that the residuals were not normally distributed (Figures 11 and 12). The error in the estimation of the energy intensity was a function of the energy intensity. Furthermore, the homes that we would be
interested in targeting (the homes that have high energy intensities) were the most poorly predicted. In what follows, all graphs will have energy intensity in [therms/ft\(^2\)/yr] and area in [ft\(^2\)].

Figure 11: Relationship between magnitude of the residual (error in prediction of energy intensity) vs. energy intensity
Figure 12: Plot of residuals vs. inverse normal distribution, where the green line indicates normality.
Furthermore, despite using both the mean centered area and mean centered area squared, these residuals were a function of the area of the home. Homes with higher energy intensities were more poorly predicted, and those homes also tended to be smaller as seen in Figure 13.

To alleviate the dependence on area, the homes were divided by quartile, <25 percentile (<1,000ft$^2$), 25-50% (1000-1400ft$^2$), 50-75% (1400-2000ft$^2$), and >75% (>2000ft$^2$). The regression was re-run within these area groups, and as shown in Figure 14, the strong relationship between energy intensity and area was lessened. However, the residuals were still not normally distributed (Figure 15), and there was a higher proportion of homes with large positive residuals (Figure 16) than would be expected with normality.
Figure 14: Homes between 1400 and 2000ft$^2$, examining the relationship between energy intensity, the regression residual predicting energy intensity, and area.
Figure 15: Homes between 1400 and 2000ft² showing the relationship between the residuals and a normal distribution of the residuals
The takeaways from this analysis were twofold: 1) If energy intensity is the target variable, the sample needs to be sub-divided into area groups prior to running the model to alleviate the dependence on area and, 2) The regression performs poorly in predicting the homes of interest: those with high energy intensities given the value of the independent variables.

As such, for each of the models in what follows the same model was run independently on each area sample,

Group 0: <25 percentile (<1,000ft²)
Group 1: 25-50% (1000-1400ft²)
Group 2: 50-75% (1400-2000ft²)
One important observation to note is that the regression results were significantly affected by dividing the sample into area groups. When the regression was run on the full sample of data points, the \( R^2 = 0.44 \), whereas when it was run for each of the individual area groups, the \( R^2 \) dropped to approximately 0.22.

In addition, area and hence mean centered area were not continuous variables in the dataset. There were very discrete areas recorded in increments of about 100ft\(^2\) for the areas, and within each area group these were not normally distributed. Not much can be done about this human coding issue. A human being guesses the size of the home and this guess (or interpolation as the case may be) has a profound impact on the modeling results.

**Robust Regression**

One method to be examined in identifying candidates was robust regression. Robust regression essentially fits a regression model and down-weights those observations with high residuals (at the tails of the normal distribution). This makes for a better fitting regression model because the points at the extreme are considered less important to predict accurately. This method also allows some flagging of sample observations to be performed based upon the weights assigned to each observation.

For each area group, the relationship between the weights (coded by the size of the points in Figure 17), the robust regression residuals and the energy intensity were similar. Down-weighted observations were those that occurred at the tails (smaller circles), and about twice as many down-weighted observations (those with weights less than the median value) were under-predicted vs. over-predicted.
Figure 17: Homes 1400-2000ft$^2$ Model Residual vs. Observed Value of energy intensity, robust regression weight coded by size of points.

**Predicting Candidacy**

The desire was to predict the homes that use an unusual amount of energy per square foot within their comparison group of area, given the value of the variables utilized in the regression. These homes would be those that were both under-predicted within the regression, (had a positive residual), and were given a very low weight within the robust regression model (poorly predicted). To flag the observations into different categories, the following logic was used:

1) Well predicted homes = homes with weights above the median value
2) Poorly predicted homes, Poor candidates = weights below the median and negative residual
3) Poorly predicted homes, Good candidates = weights below the median and positive residual

The question was then whether a classification method could predict the candidacy group for a given observation. Considering just the group of homes from 1400-2000ft\(^2\) the ability to classify two different groups of homes was examined. The first, was the distinction between the robust regression groups as listed above, and the second, just a simple cut-off value of the 90% percentile in energy intensity. The results of this classification were negative: approximately 85% of the time random forest could not distinguish between the good candidates, bad candidates and the well predicted homes as defined above. Furthermore, looking at using the variables to simply divide the homes in the area group by the value of energy intensity, there was nothing unique to classify the very high (>90%) homes.

While splitting the dataset into area groups makes sense physically, the dramatic reduction in the model accuracy makes classification into the assumed candidacy groups all the more difficult. One way of thinking about this is that the candidacy groups are defined by how well energy intensity can be predicted. If the model performed poorly, (the residuals of the regression were large), the energy intensity was not well predicted while controlling for the variables available.

Next examined was whether the weighting variable assigned for the full dataset based on the robust regression was indicative of actual energy savings in the HPS dataset. The HPS homes were re-merged with the full sample of homes, and the robust regression analysis was re-run with these homes added back in.

In Table 5 the average value of energy intensity is given by candidacy results and area grouping from the robust regression. In Table 6 the same grouping division is shown but with average energy savings post retrofit. In Figure 18 these results (energy savings vs. energy intensity) are plotted with the candidacy group indicated by color for the area group that ranges from 1400-2000ft\(^2\). It is shown that while the
division is not perfect (and not clean-cut), there is indeed a difference between the energy intensity and energy savings for each of the area groups and within the candidacy group defined by the robust regression.

<table>
<thead>
<tr>
<th>Area Group</th>
<th>Average</th>
<th>Candidate</th>
<th>Candidate</th>
<th>Not Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1000</td>
<td>0.8</td>
<td>1.3</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1000-1400</td>
<td>0.7</td>
<td>1.1</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>1400-2000</td>
<td>0.6</td>
<td>0.9</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>&gt;2000</td>
<td>0.4</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

Means of Energy Intensity [Therms/ft²/yr]

Table 5. Values of energy intensity by area group and robust regression group.

<table>
<thead>
<tr>
<th>Area Group</th>
<th>Average</th>
<th>Candidate</th>
<th>Candidate</th>
<th>Not Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1000</td>
<td>84</td>
<td>164</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>1000-1400</td>
<td>74</td>
<td>140</td>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>1400-2000</td>
<td>50</td>
<td>126</td>
<td>-25</td>
<td></td>
</tr>
<tr>
<td>&gt;2000</td>
<td>17</td>
<td>78</td>
<td>-24</td>
<td></td>
</tr>
</tbody>
</table>

Means of Energy Savings [Therms]

Table 6. Values of energy savings by area group and robust regression group.
While the robust regression division might not be perfect, it was more likely that the homes identified as candidates actually saved more than those clearly identified as "not candidate." However, the distinction between the "average candidates" and "candidates" was less clear-cut. The trend above was also much clearer for smaller homes, and increasingly less clear as the homes got larger. Looking at these results with homes for which savings data is available again calls into question whether the right metric was used to identify candidacy. However, in any model it is unrealistic to assume that the division would perfectly predict candidacy into perfect piles of positive and negative. But if we were looking to first predict for energy data that is not available, and then rely on these results to predict candidacy, this data might not be rich or accurate enough to do so. If even when energy data and savings data is available it is still difficult to divide homes into clear candidacy groups, relying on a model to simulate these values first does not make sense based on the results.
Cost-sensitive analysis
While using robust regression to define candidates was not perfect, a cost-sensitive classification method was also examined to weight the importance of accurately classifying certain observations over others. In this case, it is much more costly to perform a full audit of a home that is very unlikely to be a candidate. The misclassification cost of a no as a yes is high. It is also costly to misclassify a home with potential as not having potential as the opportunity to save on this home is then neglected. Overall, it is less costly to misclassify a maybe as a no and a no as a maybe than it is to distinguish between the definitely yes and definitely no.

A misclassification cost matrix $c(i,j)$ lists the observed values in the row, and the predicted value in the column. A correct classification $c(i,i)$ can have a negative cost (we gain something by classifying the observation correctly) or it can have no cost at all. On the other hand, $c(i,j) \geq 0$ when $i \neq j$ or there is an associated cost with misclassification. To compare one classification model to another, the cross product between the confusion matrix and the cost matrix can be taken after the model has been run. In this case, the model does not change based on the cost matrix but the ultimate model chosen depends on both classification accuracy and misclassification cost.

On the other hand, it is more desirable to consider the cost matrix during the actual classification process (while an observation gets dropped down the tree). This can be accomplished by first developing a set of classification rules in the absence of considering cost, but defining ultimate classes while minimizing the cost of misclassification.

The software TANAGRA was utilized to test whether a cost-sensitive learning algorithm could help to correctly classify the most critical observations. TANAGRA first develops the learning algorithm (classification tree) and tests the algorithm on a test data set left out of algorithmic development. The result is a confusion matrix
that simply defines the number of observations correctly classified in each category. In this case the classification development was a given home’s candidacy for retrofit based upon the robust regression categorization.

The misclassification cost matrix is then manually entered, and the cost-sensitive learning algorithm balances both the misclassification error rate and the cost associated with misclassifying the observation. In this case, multiple cost matrices were tested that attempted to put a price tag on misclassification. To think about this matrix in a quantitative manner, a couple of specific examples were envisioned based on the average costs and savings of retrofits. (Table 7). Correctly classifying high potential homes and very low potential homes was given a negative cost, while misclassifying high potential as low, and low as high was associated with a positive costs. There was no cost associated with correctly classifying a “maybe” home, and a small cost associated with incorrectly classifying maybe as no and maybe as yes.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Likely not a Candidate</th>
<th>Predicted</th>
<th>Maybe Candidate</th>
<th>Probably Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likely not a Candidate</td>
<td>-2</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Maybe Candidate</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Probably Candidate</td>
<td>5</td>
<td>2</td>
<td>-5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed</th>
<th>Likely not a Candidate</th>
<th>Predicted</th>
<th>Maybe Candidate</th>
<th>Probably Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likely not a Candidate</td>
<td>NN</td>
<td>NM</td>
<td>NY</td>
<td></td>
</tr>
<tr>
<td>Maybe Candidate</td>
<td>MN</td>
<td>MM</td>
<td>MY</td>
<td></td>
</tr>
<tr>
<td>Probably Candidate</td>
<td>YM</td>
<td>YY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cost = $5*NY + 5*YN + 2*YM + MN + MY + NM -5*YY -2*NN$
Profit = Cost*-1

Table 7. Example of a misclassification cost matrix.

Unfortunately, with the dataset utilized, and numerous iterations of the cost-matrix, the misclassification rate of high potential and low potential homes was still too high to proceed.
Total Resource Cost Test

Since the evaluation of a cost-sensitive learning algorithm in the context of the results from the robust regression did not yield clear results, the analysis moved on to consider the metric of cost-effectiveness.

Utility impact evaluations look primarily at two different measures of program effectiveness: Total Resource Cost (TRC) and Utility Resource Cost (URC). The primary difference between these two indicators is the audience of who benefits. Each indicator includes programmatic and other costs to administer the work but UTC includes the cost of the rebates within the calculation. TRC is the total amount put into the project, regardless of subsidy, as distinguished from the avoided costs incurred as a result of the program. The avoided costs can be calculated as the avoided cost to the utility or the avoided cost to the customer.

As shown in Figure 18, avoided costs were calculated based on price projections both for the customer (from the Energy Information Administration) and the utility (provided directly by the utility). The discount rate (the reduction in the value of savings over time) for the utility was set at 5.66% and for the customer at 10%. The value of a therm of savings per dollar spent was then the sum of the discounted savings over the assumed lifetime of the measure. The lifetime of the installed measures (air sealing and insulation) was assumed to be approximately 25 years. Each year, the measure is assumed to save a certain amount of therms which is devalued based on time from the present, but balanced based on the assumed rise in energy costs.
<table>
<thead>
<tr>
<th>Year</th>
<th>Year</th>
<th>Projected Utility Natural Gas Rate ($/MCF)</th>
<th>Discount Rate 5.66%</th>
<th>Proj Cust. Gas Rate ($/MCF)</th>
<th>Discount Rate 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2011</td>
<td>$8.20</td>
<td>$8.20</td>
<td>$10.78</td>
<td>10.78</td>
</tr>
<tr>
<td>2</td>
<td>2012</td>
<td>$9.26</td>
<td>$8.77</td>
<td>$9.91</td>
<td>$9.01</td>
</tr>
<tr>
<td>3</td>
<td>2013</td>
<td>$9.67</td>
<td>$8.67</td>
<td>$9.45</td>
<td>$7.81</td>
</tr>
<tr>
<td>4</td>
<td>2014</td>
<td>$9.97</td>
<td>$8.45</td>
<td>$9.73</td>
<td>$7.31</td>
</tr>
<tr>
<td>5</td>
<td>2015</td>
<td>$10.13</td>
<td>$8.13</td>
<td>$10.03</td>
<td>$6.85</td>
</tr>
<tr>
<td>6</td>
<td>2016</td>
<td>$10.29</td>
<td>$7.82</td>
<td>$10.33</td>
<td>$6.41</td>
</tr>
<tr>
<td>7</td>
<td>2017</td>
<td>$10.55</td>
<td>$7.58</td>
<td>$10.64</td>
<td>$6.00</td>
</tr>
<tr>
<td>8</td>
<td>2018</td>
<td>$10.81</td>
<td>$7.36</td>
<td>$10.96</td>
<td>$5.62</td>
</tr>
<tr>
<td>9</td>
<td>2019</td>
<td>$11.08</td>
<td>$7.14</td>
<td>$11.28</td>
<td>$5.26</td>
</tr>
<tr>
<td>10</td>
<td>2020</td>
<td>$11.36</td>
<td>$6.92</td>
<td>$11.62</td>
<td>$4.93</td>
</tr>
<tr>
<td>11</td>
<td>2021</td>
<td>$11.65</td>
<td>$6.71</td>
<td>$11.97</td>
<td>$4.62</td>
</tr>
<tr>
<td>12</td>
<td>2022</td>
<td>$11.94</td>
<td>$6.51</td>
<td>$12.33</td>
<td>$4.32</td>
</tr>
<tr>
<td>13</td>
<td>2023</td>
<td>$12.23</td>
<td>$6.32</td>
<td>$12.70</td>
<td>$4.05</td>
</tr>
<tr>
<td>14</td>
<td>2024</td>
<td>$12.54</td>
<td>$6.13</td>
<td>$13.08</td>
<td>$3.79</td>
</tr>
<tr>
<td>15</td>
<td>2025</td>
<td>$12.85</td>
<td>$5.95</td>
<td>$13.47</td>
<td>$3.55</td>
</tr>
<tr>
<td>16</td>
<td>2026</td>
<td>$13.18</td>
<td>$5.77</td>
<td>$13.88</td>
<td>$3.32</td>
</tr>
<tr>
<td>17</td>
<td>2027</td>
<td>$13.50</td>
<td>$5.60</td>
<td>$14.29</td>
<td>$3.11</td>
</tr>
<tr>
<td>18</td>
<td>2028</td>
<td>$13.84</td>
<td>$5.43</td>
<td>$14.72</td>
<td>$2.91</td>
</tr>
<tr>
<td>19</td>
<td>2029</td>
<td>$14.19</td>
<td>$5.27</td>
<td>$15.16</td>
<td>$2.73</td>
</tr>
<tr>
<td>20</td>
<td>2030</td>
<td>$14.54</td>
<td>$5.11</td>
<td>$15.62</td>
<td>$2.55</td>
</tr>
<tr>
<td>21</td>
<td>2031</td>
<td>$14.91</td>
<td>$4.96</td>
<td>$16.09</td>
<td>$2.39</td>
</tr>
<tr>
<td>22</td>
<td>2032</td>
<td>$15.28</td>
<td>$4.81</td>
<td>$16.57</td>
<td>$2.24</td>
</tr>
<tr>
<td>23</td>
<td>2033</td>
<td>$15.66</td>
<td>$4.66</td>
<td>$17.07</td>
<td>$2.10</td>
</tr>
<tr>
<td>24</td>
<td>2034</td>
<td>$16.05</td>
<td>$4.52</td>
<td>$17.58</td>
<td>$1.96</td>
</tr>
<tr>
<td>25</td>
<td>2035</td>
<td>$16.45</td>
<td>$4.39</td>
<td>$18.11</td>
<td>$1.84</td>
</tr>
</tbody>
</table>

$161.16 / MCF Saved  
$115.47 / MCF Saved  
$15.75 / therm Saved  
$11.29 / therm Saved

Assumptions:  
Gas rates from Columbia Gas of Ohio  
Discount Rate Utility 5.66% (Provided by Utility)  
Discount Rate Residential 10% (Loan interest)  
Gas Rates Residential Ohio from EIA, with 3% assumed increase  
www.eia.gov/dnav/ng/ng_pri_rescom_dcu_SOH_a.htm  
1 MCF = 10.23 Therms

Figure 18: Avoided Cost Calculation Spreadsheet
The Total Resource Cost was then calculated to be the savings (in therms) post-retrofit multiplied by the dollars per therm saved divided by the sum of the cost of each of the retrofit measures performed on the home.

\[
\text{TRC} = \frac{(\text{Savings} \times \$ / \text{Therm Saved})}{\text{Total Retrofit Cost}}
\]

The relationship between energy intensity and TRC was examined for homes that received full retrofits. The goal was to discern certain cut-off values for energy intensity in each area group signifying a cost-effective retrofit. If this value could be discerned, it could be fed into a model that predicts whether the observation falls to the right or left of this sign. The difficulty, as seen in Figure 19, is that the relationship between TRC and energy intensity was not always clear.
Cost-effective retrofits tended to take place in homes that used more energy per square foot, but the relationship was weaker than originally expected. Furthermore, the cost-effectiveness of the retrofit does not take into account the fact that "retrofit potential" has two components: how much the home saved and how much work was completed. It is easier to achieve a cost effective retrofit for low-cost jobs but more desirable to achieve a greater percentage of overall energy costs in each home. The TRC does not reflect the size of the retrofit. Finally, if the TRC is to be a useful metric, the model developed had to do a reasonable job at discerning the difference between homes at the cut-off values of energy intensity. When the model was run with the available variables there was no capacity to distinguish the energy intensity between TRC groups (>1 & < 1).
Audit vs. Retrofit

Also examined was what distinguished the homes that received retrofits vs. those that received only audits. It was originally assumed that the energy intensities associated with homes that received only an audit could be correlated to retrofit potential. Unfortunately, within the sample of homes available, the homes that received only audits actually used marginally more energy on average than those that received both an audit and retrofit. When divided into groups by area, the same phenomenon was observed, or the energy consumption between the groups was virtually indistinguishable. This speaks to the fact that retrofits occur not simply because of the characteristics of the housing stock but also because of the occupants. Within the assisted living category, 75% of homes went on to receive retrofits while in the non-assisted category only 45% went on to receive retrofits. This was true despite the fact that the assisted living homes used less energy on average than those that were not. As such, attempting to correlate cut-off energy intensities based on the audit/retrofit outcome was not possible.

Evaluating the Model

What all of this work came back to was trying to prove that homes with potential could be distinguished from those without based upon a prediction of energy consumption per square foot. The end goal was not only to predict potential based on energy consumption per square foot, but also to predict the consumption per se using the variables available (and the relative accuracy thereof).

To lend perspective to the analysis, using the variables of importance identified, and extracting just the 95th% and above, and the 5th% and below, the test data was classified for the homes 1400-2000ft². This led to a classification error rate of about 17%: 30% of the lowest users were classified as the highest and 10% of the highest were classified as the lowest. The error rate was similar for the next smallest area grouping of homes from 1000ft² to 1400ft² with a classification error rate of 16.5%,
misclassifying the lowest users as the highest 22% of the time, and the highest users as the lowest 11% of the time.

This basically says that not only is the data not good enough to distinguish between potential candidates and definite yes/no's but that the variables could not distinguish between the very highest and very lowest users within each area group about 15% of the time.
4. PHASE THREE: APPLICATION TO HOMES WITHOUT ENERGY DATA

4.1 INTRODUCTION TO PHASE THREE
As stated multiple times at this point, the original thesis goal was to predict energy usage data and hence potential for retrofit utilizing a set of publicly available variables. It has been shown that with the Experian dataset and the set of variables identified, not only is it difficult to accurately predict energy intensity, but it is difficult to distinguish between the very highest and very lowest users. It is also difficult to associate a particular energy intensity with a "capacity to retrofit" indicator.

In the case that the model had been good enough to clearly identify candidates however, it would also have to identify these candidates knowing nothing more than the values of the variables in the model. For example, while useful in identifying model outliers, robust regression relies directly on having the dependent variable available to calculate the difference between actual and predicted.

To move forward in proving the possibility of predicting retrofit without utility data, a matching scheme would have to search the combinations of the variables of homes identified as candidates and match to the closest observations in a test data set. The test data set would not have been used to develop the original robust regression model, but the observations would also have the data needed to run the regression. These homes in the test data set would then all be flagged as "not candidates". The homes used to build the model would be flagged if they met the residual criteria defined in the robust regression section (weights < median, residual > 0). The rest of the homes used to build the model would be flagged as "not candidates". A logistic regression would then be run on the homes left out of model development, where the candidates were "1" and all homes that were not candidates or in the test set were "0".

A propensity score, or "probability that the observation = 1" would result from this model. Once the data had been ordered by propensity score, the test homes closest
to the propensity value of the candidate homes would be extracted. These homes
would then be analyzed to see whether they were indeed candidates by running the
robust regression on the test homes and seeing whether the same observations
were identified as candidates as those through the propensity score matching.
If the same observations were flagged, then this would prove that the energy data
was unnecessary to flag candidates. This model is highly dependent upon how good
the original robust regression technique was at identifying poorly classified homes
as candidates. It is also dependent upon the linkage between the robust regression
results and the actual “potential for retrofit” metric. The diagram below shows the
outline of this technique involving many divisions and subdivisions of the dataset.

Figure 20: Diagram detailing Propensity Score Matching Technique to identify likely
retrofit candidates without using energy data
5. CONCLUSIONS AND RECOMMENDATIONS

This research looked at whether residential home demographics and characteristics from a publicly available marketing database could be used to predict the potential for home energy efficiency retrofits.

Available to the research was a database of 7,000 homes with monthly gas utility data before and after a retrofit was completed. For another 1,000,000 homes that had not undergone retrofit, monthly gas bills were available to benchmark home performance. For each of these datasets, about 100 other variables of information regarding the home and its residents was also available for use.

The first phase of the research looked at the connection between multiple pre-retrofit benchmark metrics and the savings attained after retrofit for the 7,000 homes. After examining four metrics it was found that energy intensity, or weather normalized annual gas consumption per square foot, was most closely correlated with post-retrofit savings.

The next phase of the research was dominated by preparing the larger dataset for analysis. Once the dataset was fully cleaned, filtered and recoded, an exploratory data analysis determined the variables of importance in the prediction of energy intensity. It was determined that the year the home was built, the income of the residents, the number of occupants, the area of the home, the education level and ethnicity of the residents, the age of the head of the house, the value of the home, the home heating type, the number of stories in the home and the number of bedrooms in the home were most indicative of energy intensity.

The next step was to determine the connection between retrofit candidacy and energy intensity, the variable found to be most indicative of savings potential. It was found that the relative magnitude of energy intensity was highly correlated to how well the value could be predicted in a regression model. Robust regression was thus
utilized as a technique to divide the dataset into “well-predicted homes” and “poorly predicted homes” based upon the weight (predictability) assigned through the regression. Poorly predicted homes could either be over- or under-predicted and candidates were defined as homes where energy intensity was highly under-predicted. Multiple classification techniques were then explored to try to bucket the observations by the robust regression category. When this categorization did not succeed, a simple cut-off value of energy intensity binned by area was defined. The classification methods could not distinguish the highest users (>90%).

Also examined was a cost-sensitive approach to the classification algorithm. This methodology weights the relative importance of correct classification by group. In this case, it was more important to accurately classify the homes with good potential and poor potential. The relative importance of correct classification (or the cost of misclassification) was linked to financial implications. This line of inquiry is one to be further explored but did not increase the classification accuracy for the most critical homes.

When none of the classification methods proved fruitful, the total resource cost test was examined as a way to define cut-off values for energy intensity using the dataset for which pre/post data was available. There was not a clear linkage between TRC and energy intensity, and in fact, the TRC was found to bias potential towards smaller jobs. Finally, the connection between whether homes received audit or retrofit in the sample with pre/post data was also examined to determine cut-offs for energy intensity. This did not yield clear results as the distinction between these homes was not based solely upon energy consumption but also on social determinants.

Finally, a simple grounding model was explored: with the variables available and their relative accuracy, the work looked to distinguish between the top 5% of EUIs and lowest 5% of EUIs. When the error rate was found to be upwards of 17% in classifying between the extremes, moving onto Phase 3 of the analysis was called
into question. However, a methodology was proposed and developed through which homes without energy data might be matched through propensity score to similar homes where energy data was available.

The accuracy and consistency of the data provided by Experian was called into question throughout the analysis. Models are only as good as the data available to build them. It is hypothesized that one driving reason for the lack of clear results is that the marketing data was simply not accurate enough. A secondary reason is that the variables available did not explain enough of the variance in home gas consumption to distinguish between retrofit candidates.

Again and again, the question arose, “how does home energy consumption define potential for retrofit?”. In the end, using the variables available and the accuracy thereof, distinguishing even the top users from the bottom users was unsuccessful. This does not mean that better data could not do a better job, or even that better clearer models that connect potential for retrofit to energy could not be found. However the following is proposed as an alternative approach, and left for future research: 1) Work with rather than around or in spite of utilities to identify candidates 2) Once utility data is available even for a subset of homes, divide them into comparison groups based upon their vintage, size, and value (the clear top three predictors). 3) Create a tiered outreach and audit process based upon the extremity of home energy intensity for the given comparison group.

Programmatic decisions could then be defined both by social factors and the energy intensity of the home (target older, smaller homes with residents on assisted living who use more energy than expected given the other homes within their comparison group). Predicting both potential and raw energy data was not possible with the Experian data. However, based upon the Phase One results and the exploratory data analysis of Phase Two, the groundwork was laid to reformulate the driving question.
If utility data is available, can access to a few more key variables help direct and design a retrofit triage? In future work, a full examination of all targeted utility retrofit programs would be needed. A further look of the driving forces behind utility data transparency would also be relevant. Finally a look into the added value of targeting, holding all else constant, would move the field forward in significant ways. Much of this work could be completed with the datasets utilized in this research. However, a stress on programmatic design principles and a significant control group would be necessary to fully describe what works and how well and how also to implement the elements of a successful targeted program.
WORKS CITED


APPENDIX A: COMMENTED USAGE DATA CLEANING CODE
written by Michael Blasnik, commented and applied by the Author

*combine all the years of utility data into one single file
  cd "/Users/krg/Desktop/Michael/usage/raw_usage/"
  use use_2007_kg, clear
  forvalues year=2007/2013 {
    append using use_'year'_kg
  }
  save use_2007_2013_kg, replace

* drop pure duplicates
  bysort premise date useasis : keep if _n==1

* create tomiss to flag readings that should be set to missing values
  gen byte tomiss=0

* set to missing if premise status isn't active in current or prior meter reading
  by premise (date) : replace tomiss=1 if index("ACDEIP",premstat) |
  (index("ACDEIP",premstat[_n-1]) & _n>1)

* set to missing if multiple conflicting meter reads for same date
  by premise date : replace tomiss=1 if _N>1

* add column of dayscycl indicating number of days between reads
  bysort premise (date): gen dayscycl = (date-date[_n-1])

* set to missing if off cycle reading -- <20 days or >40 days
  replace tomiss=1 if !inrange(dayscycl,20,40) & dayscycl!=.

* set to missing if dates don't agree with elapsed days
  bysort premise (date): replace tomiss=1 if (date-date[_n-1])!=dayscycl &
  useasis<. & _n>1
  drop premstat dayscycl

* identify estimated meter readings
  gen byte est= index("1267B", readtype)>0
  drop readtype
  bys premise date (est): keep if _n==1

* set usage to missing as flagged
  replace use=. if tomiss==1
  drop tomiss

* accumulate usage across estimated readings
  bys premise (date): replace use=use+use[_n-1] if est[_n-1]==1
* drop the estimates
  drop if est==1
  drop est

* drop meter readings if usage is missing and next reading usage is missing
  (provides no info)
  bys premise (date): drop if useasis==. & useasis[[_n+1]]==. & _n<_N
  rename useasis use

save "~/Users/krg/Desktop/Michael/usage/use_2007_2013_kg_clean", replace
APPENDIX B: B-PRISM
Code written by Michael Blasnik and applied throughout by the Author

program define prismho_new_debug

version 12.1

!* MB 20-May-2013

* prismho_new, saving(mytest) by(psid_no) hdd(weather/allhdds_long)

    syntax [if] [in], saving(str) by(str) [ STation(str) HDD(str) TH1(int 60)
THMIN(int 40) THMAX(int 75) STH1(int 8) debug(str) use(str)
noDAYweight]

* station is variable name with weather station identifier that matches into hdd file

* hdd is the name of the file with the long weather data: station, tref, date, hdd40, hdd41.. hdd70 where hdd vars are cumulative sums

* also requires a file with same name but ending with suffix "lt" that has long term average daily HDD for each tref: station, tref, lthdd

* debug: saves a detailed dataset with Tref search info - scores, etc

* nodayweight doesn’t have an impact yet

* if working with Columbia usage data use these two lines to create weather station variable

    decode wthrstatn, gen(station)

    replace station=lower(station)

* restrict usage data file to if and in conditions

    if "!if\`in\"==" keep \`if \`in'
if ":use\"==" local use "use"
if "\station\"==" local station "station"

    keep \`by\`use date \`station'

* define some temporary files and names for program use

    tempfile usefile idfile minusefile hdd_min
tempname out

* set up file to post regression results

qui postfile `out' _group nreads rxr base sebs tref hpdd sehpd covar hddpct hddspan using `saving', replace

noi di as res _n "preparing data..."

quietly {

* drop cases without weather station assignments

    count if `station'=="

    if r(N)>0 {

        noi di as err _n "Warning `r(N)' observations have no weather station listed and are being dropped" _n

        drop if `station'=="

    }

* create _group variable to identify each unique case regression to run (map "by" groups to integers)

    bysort `by': gen long _group=_n==1

    replace _group=sum(_group)

* calculate elapsed days and useperday

    bysort _group (date): gen int days=date-date[_n-1]

    gen useperday=use/days

* now just get hdd60 to see the coverage of weather data and get accurate # period counts

    gen byte tref=60

    merge m:1 `station' tref date using `hdd', keep(match master) keepusing(hdd)
count if _m==1

if r(N)>0 {
    noi di as err "Warning - weather data doesn't cover all meter reading periods"
    noi di as res "tabulation of _merge result by weather station:"
    noi tabulate `station' _merge
}

drop if _merge<3

drop _merge tref

bysort _group (date): gen hddpd=(hdd-hdd[_n-1])/days if useperday<.

* now drop observationbs outside the range of usage/weather data
  by _group: drop if hddpd==. & hddpd[_n+1]==.
  by _group: drop if hddpd==. & hddpd[_n-1]==. & _n>1
  by _group : gen int num_reads=sum(hddpd<.)
  by _group: replace num_reads=num_reads[_N]

drop hdd hddpd

save `usefile', replace

* create file with original "by" identifiers, station, # periods, sum of use, sum of days, begin and end dates
  keep `by' _group `station' num_reads date useperday days

* change useperday back to use for developing raw sum for idfile
  replace useperday=useperday*days

collapse (sum) sumuse=useperday sumdays=days (min) begdate=date (max) enddate=date, by(`by' _group `station' num_reads)
save `idfile'

* calc number of groups to regress

    sum _group, meanonly
    local ngroups=r(max)

* create most compact version of usage file - drop by vars, usage

    use _group `station' useperday date days num_reads using `usefile'

* drop cases with no weather station or less than 3 observations

    drop if `station'=="" | num_reads<3
    save `minusefile', replace

* now just grab the stations and range of dates from the hdd file that are needed

    use `station' date using `usefile'

* get min and max dates of usage file

    sum date, meanonly
    local mindate=r(min)
    local maxdate=r(max)

* now get list of stations and merge with weather file

    keep `station'
    drop if `station'=="
    bysort `station': keep if _n==1
    merge 1:m `station' using `hdd', keep(match master)

* should all match because we got rid of unmatched weather periods previously

    assert _merge==3
    drop _merge
keep if date>='mindate' & date<='maxdate'

keep if tref>='thmin' & tref<='thmax'

save 'hdd_min'

* now lets start the actual analysis

* count how many chunks of 50,000 cases each to do -- makes sure it fits within memory for most machines

local chunks=ceil('ngroups'/50000)

if `chunks'>1 noisily display as result _n "Analyzing data in `chunks' chunks of 50,000 cases each" _n

forvalues chunk = 1(1)`chunks' {

use if int(_group/50000)==(`chunk'-1) using `minusefile', replace

* replicate usage data for all values of Tref for all cases

expand `thmax'-`thmin'+1

bysort _group date: gen byte tref=`thmin'+_n-1

* use fixed Tref=th1 if <5 mdata points

drop if num_reads<5 & tref!=`th1'

* get weather data and calc HDD/day

merge m:1 `station' tref date using `hdd_min', keep(match master)

drop _merge `station'

bysort _group tref (date): gen hddpd=(hdd-hdd[_n-1])/days if useperday<.

* keep usable observations and needed variables

drop if useperday==. | hddpd==.

98
keep _group date useperday hddpd tref num_reads days

* calc r2 in pieces: numerator, denominator, can’t do in one step

* state vectorizes the calculations so sum() is a running cumulative sum across rows with by group

\[
\text{bysort}_\text{group} \text{ tref (date): gen double } r2 =
\text{num}\_\text{reads} \ast \text{sum(hddpd*useperday)} - \text{sum(hddpd)} \ast \text{sum(useperday)}
\]

\[
\text{by}_\text{group} \text{ tref (date): gen denom=}((\text{num}\_\text{reads} \ast \text{sum(hddpd}^2) - \text{sum(hddpd)} \ast \text{sum(useperday}}^2) \ast (\text{num}\_\text{reads} \ast \text{sum(useperday}^2) - \text{sum(useperday)} \ast \text{sum(useperday}}^2))
\]

* deal with \(r^2=1\) which yields missing when doing score calculation

\[
\text{replace } r2=\text{min(.9999,}r2^2/\text{denom)} \text{ if } r2<.
\]

* set \(r2\) to final \(r2\) for each Tref

\[
\text{by}_\text{group} \text{ tref (date): replace } r2=r2[\_N]
\]

\[
\text{drop if } r2==. \mid \text{denom==}.
\]

* calculate Byesian Tref score = likelihood(Tref) * likelihood(rxr)

\[
\text{gen double score=}((r2>0)\ast\text{normalden}(\text{tref}-\text{th1}'/(\text{sth1}'))\ast((1-r2)^{-\text{(num\_reads}/2}))
\]

* save debug file with details if requested

\[
\text{if } "\text{debug}'"!=" \\
\{ \\
\text{merge } m:1 \text{ group using } \text{idfile}, \text{keep(match master) keepusing('by')}
\]

\[
\text{drop '_merge}
\]

\[
\text{save } \text{debug'} \'_\text{chunk}', \text{replace}
\]

\[
\text{drop 'by'}
\]

\}

* keep data for best Tref only (highest score) for each case
drop if score==.
bysort _group (score): keep if tref==tref[_N]

* keep only what we need for regression analysis
keep _group tref num_reads useperday hddpd days

* calculate data quality info: hddspan and hddpct
bysort _group (hddpd): gen hddspanx=hddpd[_N]-hddpd[1]
by _group: gen hddpctx=sum(hddpd*days)
by _group: replace hddpctx=hddpctx[_N]

* now run regressions using best Tref data for each case using observation numbers
* j is the starting row number for this case, end is the ending row number for this case

local j=1
while `j'<=_N {
    local end=`j'+num_reads[`j']-1
    capture regress useperday hddpd [aweight=days] in `j'/`end'
    if _rc==0 {
        scalar base=_b[_cons]
        *scalar sebs=_se[_cons]
        mat v=e(V)
        scalar covar=v[1,2]
        *scalar htsl=_b[hddpd]
        *scalar seht=_se[hddpd]
* scalar nreads=e(N)
   post `out' (_group[j]) (`e(N)') (`e(r2)') (_b[_cons])
   (_se[_cons]) (tref[j]) (_b[hddpd]) (_se[hddpd]) (covar) (hddpctx[j])
   (hddspanx[j])
   }
   local j=`end'+1
   }
   noisily display "done with chunk `chunk' of `chunks' at $S\_TIME"
   }
   postclose `out'
   noisily display

   use `saving', replace
   * get original identifiers, weather stations, etc.
     merge 1:1 _group using `idfile'
     drop _group _merge
   * get long term average degree days for each station / tref
     merge m:1 `station' tref using `hdd\_lt', keep(match master) keepusing(lthdd)
     drop _merge
   * calculate standard error of NAC using coef std errors and covariance
     gen senac=365.25*sqrt((lthdd\_A2*sehpdd\_A2+2*covar*lthdd+sebs\_A2))
   * adjust various values to a full year
     replace sebs=sebs*365.25
     replace base=base*365.25
replace lthdd=lthdd*365.25
replace hddpct=hddpct/lthdd
replace hddspan=hddspan*365.25/lthdd
gen nahc=hpdd*lthdd
gen senahc=sehpdd*lthdd
gen nac=base + nahc
gen cvnac=senac/nac

* set friendly display formats

    format base sebs nac senac nahc senahc lthdd %7.0f
    format begdate enddate %td
    format cvnac rxr hddpct hddspan %5.3f

* should drop num_reads -- should be identical to nreads -- check this out first

    order 'by' nreads nac base nahc cvnac rxr tref begdate enddate hddpct hddspan lthdd 'station' sebs sehpdd
    replace nreads=num_reads if nreads==.
drop num_reads covar
save 'saving', replace

}
APPENDIX C: EXPERIAN CLEANING CODE

use experian.dta
rename _all, lower
rename enhesthhincemountv4 hhincom
rename numberofadultsinhousehold adult_cnt
rename numberofchildren18orless child_cnt
rename presenceofchildage018v3 child_code

gen byte havechildren=strpos(child_code,"Y")>0 if inlist(substr(child_code,2,1),"Y","N")
gen occupants=adult_cnt+child_cnt
* shorten some very long names
rename presenceofchild* poc*
rename *dmmercbyrcat *dm
rename *magbyrcat *mag
rename *ccyymmdd *
rename homebuildingsquarefootage area
replace area=100*area

gen area_clean=area if inrange(area,750,6000)
rename yearbuilt yearbuiltcode

gen yearbuilt=yearbuiltcode-10000*int(yearbuiltcode/10000)
gen byte vintage =1 if yearbuilt<1950
replace vintage =2 if vintage==. & yearbuilt<1980
replace vintage =3 if vintage==. & yearbuilt<2014
lab def vintage 1"pre WWII" 2"1946-1979" 3"1980+"
lab values vintage vintage

gen byte post1979=yearbuilt>=1980 if yearbuilt<.

gen byte singlefamily=dwellingtype=="S" if dwellingtype!="

rename educationindividual1 education_code

gen education = mod(education_code,10)

replace education =0 if education=5

lab def education 0 "<HS" 1"HS" 2"Some College" 3"College" 4"Grad Degree"

lab values education education

** doesn't look like the data I have includes marital status

**rename maritalstatus1 marital_status

**gen byte married=strpos(marital_status,"M")>0 if inlist(marital_status,"1M","5M","5S")

rename homebedrooms bedrooms

rename homebath baths

replace baths=round(baths/10,.5)

rename homestories stories_code

gen stories=round(stories_code/10,.5)

replace stories=. if stories>4 & singlefamily==1

rename transactionamount purchaseprice_code

gen house_price=1000*purchaseprice_code

rename estimatedcurrenthomevalue house_value

replace house_value=. if house_value<1000
rename multicodirectmailresphh directmail hh
rename i1mailresponderindiv directmail_ind
gen byte directmail=inlist( directmail hh,"M","Y") if directmail hh!="
replace directmail=1 if directmail_ind=="Y"
gen byte owner=inlist(combinedhomeowner,"7","8","9","H") if
!inlist(combinedhomeowner,"","U")
gen byte female_hh_new=exp_i1gendercode="F" if exp_i1gendercode!="
gen age hh=real(substr(i1combinedage,2,2)) if i1combinedage!="
gen ethnicity_tech = "Unknown"
local etechcodes "A B C D E F G H I J K L M N O Z"
local etechdesc "AfricanAmerican SE_Asian S_Asian C_Asian Mediterranean
NativeAmerican Scandinavian Polynesian MiddleEastern Jewish W_European
E_European Caribbean_NonHisp E_Asian Hispanic Uncoded"
local i=1
foreach abbrev of local etechcodes {
    local desc: word `i' of `etchdesc'
    qui replace ethnicity_tech="desc" if etechgroup="abbrev"
    local i=`i'+1
}
rename religion religion_code
gen religion = "Unknown"
local relcodes "B C E G H I J K L M O P S X"
local reldesc "Buddhist Catholic Ethiopian_Orthodox Greek_Orthodox Hindu Muslim
Jewish Sikh Lutheran Mormon Eastern_Orthodox Protestant Shinto NotKnown"
local i=1
foreach abbrev of local relcodes {
    local desc: word \i\ of `reldesc'
    qui replace religion="`desc"" if religion_code=="`abbrev"
    local i=`i'+1
}
rename occupationgroup1 occupation_grp
rename i1occupation occupation_code
merge m:1 occupation_code using occupationcodes, keep(match master)
drop _merge

*note that K stands for Known, and I for Inferred.
gen byte occup_grp_known=substr( occupation_grp,1,1)=="K"
gen byte occup_grp=real(substr( occupation_grp,2,2))
lab def occup_grp 1 "Prof/Tech" 2 "Sales/Svc" 3 "Farm" 4 "Blue Collar" 5 "Other"
   6 "Retired"
lab values occup_grp occup_grp

gen hhld_comp="U"
local hhldcomp_codes "A B C D E F G H I J U"

local hhldcomp_desc "SF SM MF FC MC MM FF MMC FFC U"
local i=1
foreach abbrev of local hhldcomp_codes {
    local desc: word \i\ of `hhldcomp_desc'
    qui replace hhld_comp="`desc"" if householdcompositioncode=="`abbrev"
    local i=`i'+1
rename mortgageamount mortgage

replace mortgage=1000*mortgage

rename _all exp_=
replace exp_agedparentinhousehold = "" if exp_agedparentinhousehold == "U"
encode exp_agedparentinhousehold, gen(exp_numagedparentinhousehold)
* replaced with encoded form

drop exp_agedparentinhousehold
drop exp_i2persontype exp_i2namefirst exp_i2namemiddleinitial exp_i2namelast
exp_i2nametitleofrespect exp_i2namesurnamesuffix exp_i2gendercode
exp_i3birthdateccymmm exp_i3persontype exp_i3namefirst exp_i3namemiddleinitial
exp_i3namelast exp_i3nametitleofrespect exp_i3namesurnamesuffix
exp_i3gendercode exp_i3combinedage exp_i3birthdateccymmm exp_i4persontype
exp_i4namefirst exp_i4namemiddleinitial exp_i4namelast exp_i4nametitleofrespect
exp_i4namesurnamesuffix exp_i4gendercode exp_i4combinedage
exp_i4birthdateccymmm exp_i5persontype exp_i5namefirst exp_i5namemiddleinitial
exp_i5namelast exp_i5nametitleofrespect exp_i5namesurnamesuffix
exp_i5gendercode exp_i5combinedage exp_i5birthdateccymmm exp_i6persontype
exp_i6namefirst exp_i6namemiddleinitial exp_i6namelast exp_i6nametitleofrespect
exp_i6namesurnamesuffix exp_i6gendercode exp_i6combinedage
exp_i6birthdateccymmm exp_i7persontype exp_i7namefirst exp_i7namemiddleinitial
exp_i7namelast exp_i7nametitleofrespect exp_i7namesurnamesuffix
exp_i7gendercode exp_i7combinedage exp_i7birthdateccymmm exp_i8persontype
exp_i8namefirst exp_i8namemiddleinitial exp_i8namelast exp_i8nametitleofrespect
exp_i8namesurnamesuffix exp_i8gendercode exp_i8combinedage
exp_i8birthdateccymmm
* replaced with binary

drop exp_child_code
gen byte havechildren03=strpos(exp_pocage03v3,"Y")>0 if
inlist(substr(exp_pocage03v3,2,1),"Y","N")
drop exp_pocage03v3
gen byte havechildren46=strpos(exp_pocage46v3,"Y")>0 if
inlist(substr(exp_pocage46v3,2,1),"Y","N")
drop exp_pocage46v3
gen byte havechildren79=strpos(exp_pocage79v3,"Y")>0 if
inlist(substr(exp_pocage79v3,2,1),"Y","N")
drop exp_pocage79v3
gen byte havechildren1012=strpos(exp_pocage1012v3,"Y")>0 if
inlist(substr(exp_pocage1012v3,2,1),"Y","N")
drop exp_pocage1012v3
gen byte havechildren1315=strpos(exp_pocage1315v3,"Y")>0 if
inlist(substr(exp_pocage1315v3,2,1),"Y","N")
drop exp_pocage1315v3
gen byte havechildren1618=strpos(exp_pocage1618v3,"Y")>0 if inlist(substr(exp_pocage1618v3,2,1),"Y","N")
drop exp_pocage1618v3

label define yn 0 "No" 1 "Yes"
label values havechildren03 yn
label values havechildren46 yn
label values havechildren79 yn
label values havechildren1012 yn
label values havechildren1315 yn
label values havechildren1618 yn

label variable havechildren03 "Presence of Children 0-3"
label variable havechildren46 "Presence of Children 4-6"
label variable havechildren79 "Presence of Children 7-9"
label variable havechildren1012 "Presence of Children 10-12"
label variable havechildren1315 "Presence of Children 13-15"
label variable havechildren1618 "Presence of Children 16-18"

label define yu 0 "Unknown" 1 "Yes"

encode exp_homebusinessindicator, gen(exp_homebusinesspresent)
recode exp_homebusinesspresent (1=0) (2=1)
label values exp_homebusinesspresent yu
drop exp_homebusinessindicator

encode exp_i1businessownerflag, gen(exp_businessowner)
recode exp_businessowner (1=0) (2=1)
label values exp_businessowner yu
drop exp_i1businessownerflag

encode exp_directmail_hh, gen(exp_numdirectmail_hh)
recode exp_numdirectmail_hh (3=1)(2=.) (1=2)
label define hh 1 Yes 2 Many
label values exp_numdirectmail_hh hh
drop exp_directmail_hh

encode exp_multibuyermin551551, gen(exp_nummultibuyer)
drop exp_multibuyermin551551

label variable exp_numdirectmail_ind "Individual Made Purchase by Direct Mail"
label variable exp_nummultibuyer "Individual Made Purchase by Direct Mail in Multiple Categories"
rename exp_numdirectmail_hh exp_levelsdirectmail_hh
label variable exp_levelsdirectmail_hh "Household Level Direct Mail Responder"
recode exp_giftgadgetdm (2/9 = 2)
label define dm 0 "None" 1 "One" 2 "More than One"
label values exp_giftgadgetdm dm

recode exp_generalcontribtrcat (2/9 = 2)
recode exp_healthinstcontribtrcat (2/9 = 2)
recode exp_politicalcontribtrcat (2/9 = 2)
recode exp_religiouscontribtrcat (2/9 = 2)

label values exp_generalcontribtrcat dm
label values exp_healthinstcontribtrcat dm
label values exp_politicalcontribtrcat dm
label values exp_religiouscontribtrcat dm

recode exp_sweepstakes (2/9 = 2)
recode exp_doityourselfers (2/9 = 2)
recode exp_newsfina nancial (2/9 = 2)
recode exp_photography (2/9 = 2)
recode exp_oddsmailresponde r (2/9 = 2)
recode exp_miscellaneousmailresponder (2/9 = 2)

label values exp_sweepstakes dm
label values exp_doityourselfers dm
label values exp_newsfina nancial dm
label values exp_photography dm
label values exp_oddsmailresponde r dm
label values exp_miscellaneousmailresponder dm

recode exp_giftgadgetdm (2/9 = 2)
recode exp_colspcfooddm (2/9 = 2)
recode exp_booksdm (2/9 = 2)
recode exp_gardenfarmdm (2/9 = 2)
recode exp_craftshobbydm (2/9 = 2)
recode exp_femaleorientdm (2/9 = 2)
recode exp_maleorientdm (2/9 = 2)
recode exp_upscaledm (2/9 = 2)
recode exp_generaldm (2/9 = 2)
recode exp_healthfitnessmag (2/9 = 2)
recode exp_culinaryinterestmag (2/9 = 2)
recode exp_gardenfarmingmag (2/9 = 2)
recode exp_religiousmag (2/9 = 2)
recode exp_maleorientmag (2/9 = 2)
recode exp_femaleorientmag (2/9 = 2)
recode exp_familygencermag (2/9 = 2)

label values exp_giftgadgetdm dm
label values exp_colspcfoooddm dm
label values exp_booksdm dm
label values exp_gardenfarmdm dm
label values exp_craftshobbedm dm
label values exp_femaleorientdm dm
label values exp_maleorientdm dm
label values exp_upscaled dm dm
label values expgeneral dm dm
label values exp_healthfitnessmag dm
label values exp_culinaryinterestmag dm
label values exp_gardenfarmingmag dm
label values exp_religiousmag dm
label values exp_malesportorienmag dm
label values exp_femaleorientmag dm
label values exp_familygeneralmag dm

drop exp_education_code
drop exp_i1birthdateccyyymm
drop exp_i1combinedage
drop exp_i2combinedage
replace exp_hhincom = exp_hhincom*1000
drop exp_ethnicinsightmatchflag
drop exp_etechgroup

merge m:1 exp_ethnicitydetail using "'/Users/krg/Desktop/Michael/premises/fullethnicitydetail.dta",
keepusing(ethnicdetail)
drop if _m ==2
drop exp_ethnicitydetail _merge
drop exp_religion_code

merge m:1 exp_language using "'/Users/krg/Desktop/Michael/premises/language.dta",
keepusing(exp_languagedetail)
drop if _m==2
drop _m
drop exp_language

merge m:1 exp_ethnicgroupcode using "'/Users/krg/Desktop/Michael/premises/ethnicgroup.dta",
keepusing(exp_ethnicgroupdetail)
drop if _m == 2
drop exp_ethnicgroupcode _merge
merge m:1 exp_countryoforigin using 
    "/Users/krg/Desktop/Michael/premises/countryoforigin.dta",
        keepusing(exp_countryoforigin_detail)
    sort _m
    drop if _m==2
    drop _m
    drop exp_countryoforigin

merge m:1 exp_occupation_code using 
    "/Users/krg/Desktop/Michael/premises/occupationcodes.dta",
        keepusing(exp_occupation_text)
    drop if _m == 2
    drop _m

drop exp_occupation_grp exp_occupation_code

drop exp_enhe Steph householderincomev4

rename exp_i1gendercode exp_gender_hh

drop exp_householdcompositioncode

replace exp_dwellingtype = "Single Family" if exp_dwellingtype == "S"
replace exp_dwellingtype = "Multi Family and Condo (<5 Units)" if exp_dwellingtype == "A"
replace exp_dwellingtype = "Multi Family Unknown" if exp_dwellingtype == "M"
replace exp_dwellingtype = "Unknown" if exp_dwellingtype == "U"
replace exp_dwellingtype = "Post Office Box" if exp_dwellingtype == "P"

gen exp_dwellingunits = "Unknown"
local dwellcodes "A B C D E F G H I U"
local dwelldesc "1_Unit 2_Units 3_Units 4_Units 5_9_Units 10_19_Units 20_49_Units 50_100_Units 101_more_Units Unknown"
local i=1
foreach abbrev of local dwellcodes{
    local desc: word `i' of `dwelldesc'
    qui replace exp_dwellingunits="\"desc\"" if exp_dwellingunitsizecode="\"abbrev\"
    local i=`i'+1
}

drop exp_dwellingunitsizecode

drop exp_combinedhomeowner

drop exp_dateofwarrantyy deed
replace exp_typeofpurchase = "New" if exp_typeofpurchase == "N"
replace exp_typeofpurchase = "Resale" if exp_typeofpurchase == "R"
replace exp_typeofpurchase = "Unknown" if exp_typeofpurchase == "U"

replace exp_homeswimmingpoolind = "Unknown" if exp_homeswimmingpoolind == "U"
replace exp_homeswimmingpoolind = "Yes" if exp_homeswimmingpoolind == "Y"

replace exp_totalenhancementmatchtype = "Household Match" if
exp_totalenhancementmatchtype == "H" | exp_totalenhancementmatchtype == "F"
replace exp_totalenhancementmatchtype = "Person Match" if
exp_totalenhancementmatchtype == "P" | exp_totalenhancementmatchtype == "I"
replace exp_totalenhancementmatchtype = "Geographic Match" if
exp_totalenhancementmatchtype == "G" | exp_totalenhancementmatchtype == "E"
replace exp_totalenhancementmatchtype = "No Match" if
exp_totalenhancementmatchtype == "N" | exp_totalenhancementmatchtype == "" 

drop exp_stories_code
drop exp_purchaseprice_code
drop exp_yearbuiltcode

drop if exp_totalenhancementmatchtype == "Geographic Match"

replace exp_gender hh = "" if exp_gender hh == "U"
replace exp_gender hh = "" if exp_gender hh == "B"
encode exp_gender hh, generate(exp_gender_num hh)
label define gen 1 Female 2 Male
recode exp_gender_num hh (2=1) (3=2)
label values exp_gender_num hh gen
drop exp_gender hh
rename exp_gender_num hh exp_gender hh

encode exp_dwellingtype, generate(exp_dwellingtype_new)
recode exp_dwellingtype_new (4=1) (3=0) (1=3) (2=2)
label define restype 0 "Post Office Box" 1 "Single Family Home" 2 "Condo or <5 Unit Multi" 3 "Multi-Family Unknown"
label values exp_dwellingtype_new restype
drop exp_dwellingtype
rename exp_dwellingtype_new exp_dwellingtype

label define sidetype2 1 "Aluminum"
abel define yn 0 "No" 1 "Yes"
label values exp_havechildren yn
drop exp_homebasesquarefootage exp_homelandsquarefootage exp_area_clean
label values exp_post1979 yn
label values exp_singlefamily yn
label values exp_directmail yn
label values exp_owner yn
drop exp_occup_grp_known
label values exp_female_hh_new yn