MSDS 6372 Project 1 Report

Volodymyr Orlov, Jaime Villanueva, Jason Lin

Introduction

The United States energy consumption is one of the biggest in the world with residential sectors representing approximately 40 percent[1] of it. Here we explore data collected via *Residential Energy Consumption Survey*[2] conducted by the *US Energy Information Administration* in 2009. Using this data we perfom regression analysis of the data and model future average yearly energy consumption, in US dollars, of a household from various household's features and consumption habit of its tenants. In addition to that we also attempted to answer following questions:

- Q1
- Q2

Data Description

(Why we've chosen 2009 when we could take 2015?!!. Lets either somehow support it or use latest data).

The *US Residential Energy Consumption Survey* is a comprehensive data collection program that was conducted by the *US Energy Information Administration*[3].

The data file can be obtained from the www.eia.gov website and is formatted as a CSV file. The file contains responses from 12083 randomly selected sample of the population of 113.6 million US households. Response is coded using 940 variables, which can be broken down into following categories:

- 359 imputation flags.
- 85 measurements of various energy consumption metrics.
- 58 various quantitative measurements, collected from a household or its tenants.
- 438 categorical features of a household.

Exploratory Analysis

Fo our analisys and models we considered only 58 quantitative and 438 categorical features of a household and dropped 359 imputation flags as well as 85 measurements of energy consumption metrics. For our first model we've been primaraly interested in predicting total energy bill per household that has been recorded as a separate variable, in whole US dollars.

To better understand relationship between total yearly cost and household's features we looked at correlation matrix of all quantitative plus our dependent variable. The plot is depicted in Figure 1

By scanning through correlation matrix we've identified several good candidates for strong predictor variables: total square footage, total number of rooms, number of rooms heated, number of ceiling fans used, number of televisions used, heating degree days in 2009 above 65F. From Figure 2 you can see there is a clear linear relashionship between all of these variables and a response we are trying to predict.

In addition to that we've used our intuition to select following variables for further analysis:

[1] https://www.nap.edu/catalog/13360/effective-tracking-of-building-energy-use-improving-the-commercial-buildings

[2] https://www.eia.gov/consumption/residential/about.php

[3] https://en.wikipedia.org/wiki/Energy__Information__Administration

Objective 1

First goal of our analysis is to model total energy bill of a household using best subset of predictors. All our models are based on linear regression:

$$y = X\beta + \epsilon$$

Where X is a $x \times k$ matrix with k observation and n independent variables, y is a $n \times 1$ vector of observations, ϵ is a $n \times 1$ irreducible error vector and β be a $k \times 1$ vector of parameters that we want to estimate.

Model Selection

Given big number of predictors it made more sense for us to have two models:

1. A model for interpretation and hypothesis testing
2. A model for high predictability

For (1) we used our intuition to choose a best subset of predictors for interpretation. For (2) we have tested multiple combinations of quantitative and categorical predictors and selected the best model using AIC, BIC and CV criteria.

Predictive Model

To select a best predictive model we assembeled an ordered list of qualitative and quantitative predictors using scatterplots and our intuition as a guideline. The final list can be found in Table 1. Next, we have assembled 22 models starting from a simple model with just one predictor where each subsequent model uses increasing number of predictors from this table and measured BIC, AIC and CV metrics for each model. Resulting plots can be found in Figure 3. After looking at these plots we concluded that the best predictive model is model build from predictors marked with a star in Table 1.

Descriptive Model

Figures and Tables

Predictor	Description	Used in Best Model
TOTSQFT	total square footage	*
HEATROOM	number of rooms heated	*
TOTROOMS	total number of rooms	*
NUMCFAN	number of ceiling fans used	*
TVCOLOR	number of televisions used	*
WINDOWS	Number of windows in heated areas	*
MONEYPY	2009 gross household income	*
DIVISION	Census Division	*
YEARMADE	Year housing unit was built	*
POOL	Heated swimming pool	*
AGEHHMEMCAT2	Age category of second household member	*
STUDIO	Studio apartment	*
AGEHHMEMCAT3	Age category of third household member	*
REPORTABLE_DOMAIN	Reportable states and groups of states	*
LPGDELV	Propane delivered	*
PUGCOOK	Who pays for natural gas for cooking	*
SWIMPOOL	Swimming pool	*
USECENAC	Frequency central air conditioner used in summer 2009	*
PCSLEEP2	Sleep or standby mode for second computer when not in use	*
STORIES	Number of stories in a single-family home	*
AGEHHMEMCAT4	Age category of fourth household member	*
ROOFTYPE	Major roofing material	*
FUELPOOL	Fuel used for heating swimming pool	*
USECFAN	Frequency most-used ceiling fan used in summer 2009	*
INSTLWS	Caulking or weather stripping by this household	*
TIMEON2	Daily usage of second most-used computer	*
OVENUSE	Frequency of oven use	*
SIZRFRI2	Size of second most-used refrigerator	*
OVENFUEL	Fuel used by most-used stove	*
WHEATSIZ2	Secondary water heater size (if storage tank)	*
WASHLOAD	Frequency clothes washer used	*
REGIONC	Census Region	*
PCTYPE3	Third most-used computer - desktop or laptop	*
FUELHEAT	Main space heating fuel	*
REFRIGT2	Defrosting type of second most-used refrigerator	*
BATTOOLS	Number of rechargeable tools and appliances used	*
EDUCATION	Highest education completed by householder	*
AIA_Zone	AIA Climate Zone, based on average temperatures from 1981-2010	*
PERIODLP	Number of days covered by Energy Supplier Survey LPG	*
COMBODVR3	DVR built into the cable box or satellite box	*
TVONWE3	Third most-used TV usage on weekends	*
DRAFTY	Is home too drafty in the winter? (respondent reported)	*
DWASHUSE	Frequency dishwasher used	*
PELLIGHT	Who pays for electricity used for lighting and other appliances	*
SIZRFRI3	Size of third most-used refrigerator	*
WHEATAGE2	Secondary water heater age	*
DRYRUSE	Frequency clothes dryer used	*
AGECDRYER	Age of clothes dryer	
TYPEHUQ	Type of housing unit	

Table 1: Ordered list of variables considered for our predictive model. Predictors, selected for the best model are marked with *

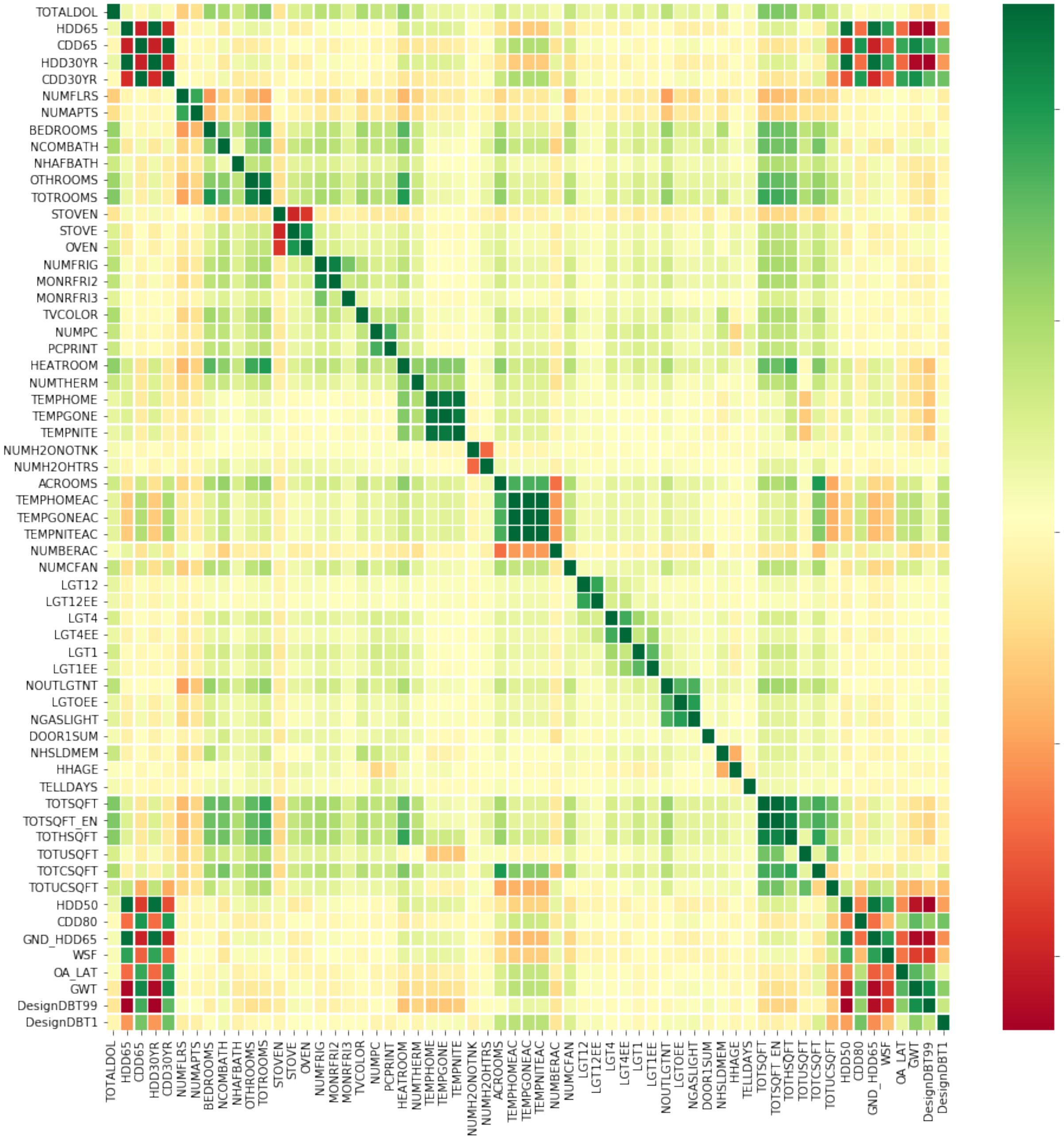

Figure 1: Correlation matrix of all quantitative household measurements, plus dependent variable.

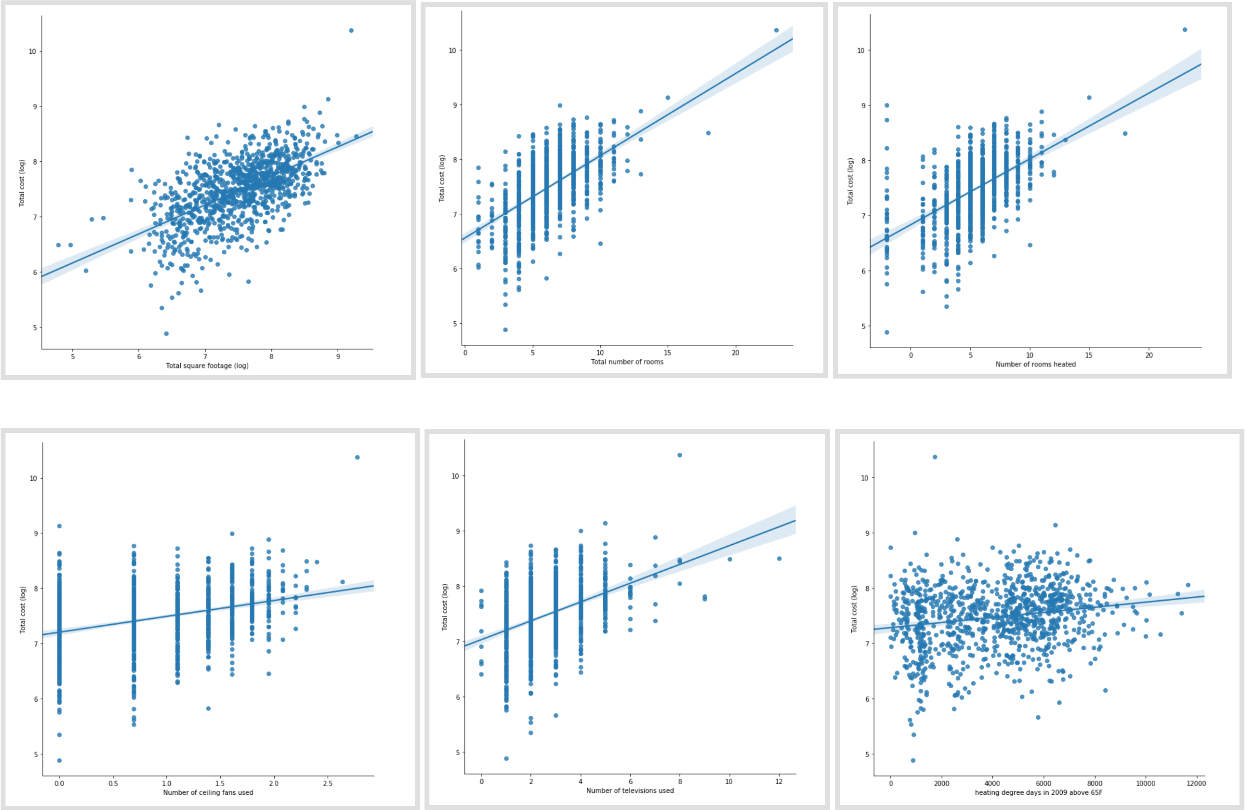

Figure 2: Scatterplots of selected predictors vs log(total cost).

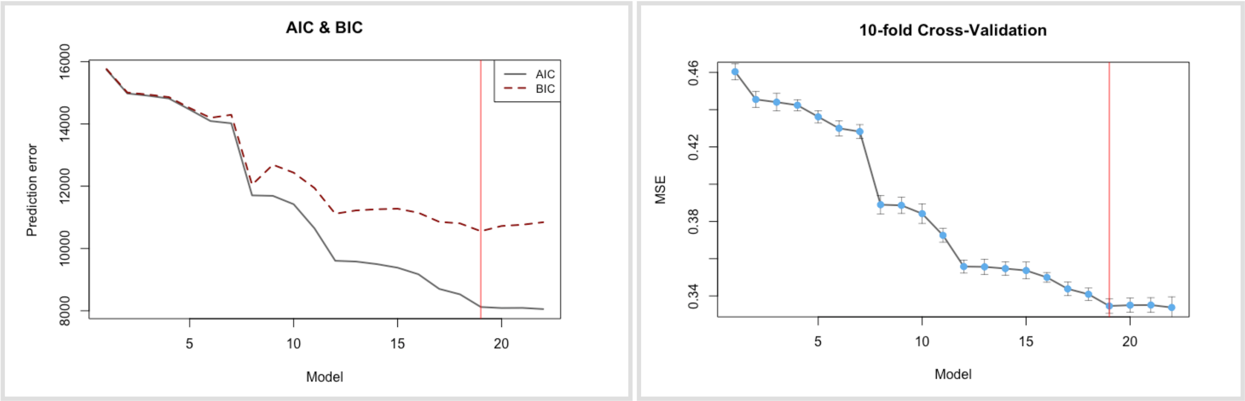

Figure 3: AIC, BIC and CV metrics for 22 predictive models. Vertical red line indicates selected model.