Capstone Project 1: Project Proposal

Springboard Data Science Track
Ishmail Grady
March 2019

Problem Statement

The sociopolitical effects of climate change is shifting the focus of energy sector. Energy providers are facing pressure to be more efficient and environmentally sustainable. How energy providers communicate with their customers about their energy consumption and sustainability will play a crucial role in how their businesses are able to adapt to this changing market. This project will serve as a solution to help energy providers enhance their customer communication by providing a predictive model that will allow its customers to predict their energy consumption based on the characteristics of their household. This predictive model can then be incorporated into an online tool for customers that leverages the existing data of energy providers. This tool will give customers the ability to do the following:

- Predict their energy consumption in a more robust, and precise way using granular information specific to their household.
- Compare their actual consumption to predicted consumption in order to gauge how energy efficient their household is.
- Analyze how various household decisions can affect their energy use, thus providing a means to minimize their consumption.

Ultimately, providing this tool to customers will be beneficial for energy providers for the following reasons:

- It will allow providers to compete with other energy management solutions that can threaten the relevancy of their customer communications.
- Provides a means of collecting data on customers to build better datasets that for applications such as peak load management and targeted marketing for goods and services.

Data Source

The main dataset used to develop this model will come from the microdata of the 2015 Residential Energy Consumption Survey (RECS) Survey conducted by the U.S. Energy Information Administration. This survey is a national sample of housing units that are considered primary residences, as defined by the U.S. Census Bureau. The survey results contain data on 5,686 randomly selected households across the nation. This sample was statistically designed to represent 118.2 million households throughout the country. The RECS Survey contains documentation of their sampling methodology and sampling error. More information on how the sample was designed and how sampling error was computed, see the 2015 RECS Household Characteristics Technical Documentation Summary.

This dataset contains two main types of information: household characteristics and consumption & expenditures. Household characteristics data covers many areas such as appliances, electronics, space heating, household demographics, and more. Consumption & expenditures data contains information on the fuel type(s) used, the end uses of the fuel associated with the various household characteristics, and the dollar values of the energy used. The dataset contains over 700 variables in total; complete list of variables and descriptions can be found can be found in the Variable and Response Codebook. This data is relatively clean and publicly available. Special consideration will be given to the sampling error associated with the data, any imputed values, and the relative weights of each data point.

Methods & Deliverables

The goal of this project will be to develop a predictive model that will use household characteristics as input variables and consumption/expenditures as response variables. This model can later be incorporated into an online tool for customers that will them to input their household characteristics and generate a prediction of energy consumption and expenditures. This model will be a regression analysis using a suitable machine learning algorithm. It will be based on a set of features than can generalize the data with the best accuracy and efficiency.

The final deliverables of the project will be a codebase that will allow the user to reproduce the model and a slide deck that will detail key insights from the model and applications of the model.