Adam Gough
ECON 1680
Professor Handlan
10 May 2024

Household Energy Consumption in Machine Learning

Introduction

The research seeks to understand the factors influencing energy consumption in the United States, focusing on answering the following questions: How do factors such as average household income, average household size, year of construction, and regional characteristics influence average yearly energy consumption for a household? How may this correspond to remote work transition and energy prices in more recent years? As energy efficiency technologies improve and remote work becomes more common[1], understanding these dynamics becomes crucial for addressing climate change impacts at a household level. Additionally, the transition to clean energy on a global scale[2] requires constant analysis and evaluation of current energy consumption habits, and the goal is to help measure and predict such trends utilizing the data available.

While existing literature identifies correlated predictors of energy consumption, it often lacks specific weighting of these factors at the household level. The study aims to fill this gap by examining U.S. households, given the country's significant influence on global energy trends. The data used is from the US Energy Information Administration, the US Bureau of Labor Statistics, and the US Survey of Working Arrangements and Attitudes. The methodology includes employing Random Forest regression to model the complex interactions of household characteristics with energy use and LASSO regression to pinpoint the most significant predictors and prevent overfitting. This approach will help compare the effectiveness of these models in understanding energy consumption patterns.

Ultimately, the study revealed that there is a clear relationship between household characteristics and consumption levels. The regressions were able to predict with high accuracy what future consumption trends for households would look like. Additionally, LASSO regression seemed to be more suitable, rather than the Random Forest regression, for the data based on standard metric measurements that were noted.

Data Sources

The data sources used are all publicly available on three separate sources, in the form of excel spreadsheets, as mentioned in the introduction: the EIA, the SWAA, and the USBLS. The data is structured slightly different across each source, and as such there was necessary concatenation to use the data.

The data from the EIA is separated across the years 1997, 2001, 2005, 2009, 2015, and 2020 by region, and within each year and region there are total yearly averages separated by household characteristics that measure annual average household energy consumption. The household

characteristics that were analyzed, out of the ones present in the data, were household income, household size by square foot, construction year, and sub-region in the US. The data from the SWAA is more straightforward, in that it contains the percent of the workforce that works at home in the US tracked annually across the years 1969 to present day. Finally, the USBLS tracks gas and electricity prices across the years 1978 to 2024. Seeing as though the data used from the EIA dates to 1997, this is the earliest the data dates back to.

The variables that are used in this research are:

- Average annual household energy consumption by household income, size (square foot), year of construction, and sub-region
- Average price of gas in the US
- Average price of electricity in the US
- Percent of work-from-home individuals (WFH) in the workforce

Methodology

In this research, Random Forest and LASSO regressions are utilized to analyze factors influencing household energy consumption. Random Forest is ideal for managing multicollinearity in the dataset of interrelated features such as household income and size, by building multiple decision trees and merging their outputs without assigning individual variable weights, thus preventing overfitting. Complementarily, LASSO regression introduces a regularization term that penalizes the absolute size of regression coefficients, effectively reducing the number of features by setting less important coefficients to zero, which aids in feature selection and enhances model interpretability. LASSO is particularly known for helping to prevent overfitting, and in this case, there is a lot of data that is averaged where some averages that are used as the regressors may not contribute as significantly to the predictive power of the model. Both methods are implemented using Python's scikit-learn library, with model performance evaluated through metrics like MSE, MAE, and RMSE to ensure accuracy and effectiveness in predicting energy consumption. In this way, the two different methodologies are compared to see which may be a best fit for the data.

The main parameter of interest in LASSO regression is the dependent variable which is ultimately the prediction of household energy consumption levels (in millions of BTU) given the independent variables used for the analysis. For the Random Forest regression, the parameter of interest is the final feature importance score for each feature on a decision tree. This value is obtained by averaging the feature importance over all the tress using normalization and the random forest modeling functions. These two methodologies will highlight not only the significance of the coefficients, but also the overall accuracy of the model itself.

To work with the structure of the data and to clearly understand how each variable contributes to energy consumption, four different household characteristics are used, listed in the research question, to create a total of eight different regressions: 4 LASSO and 4 Random Forest. Within these regressions, energy prices and work from home data are included as regressors to further analyze how these factors influence energy consumption.

Results

There are a handful of different results that can be drawn from the analysis. To start, the four LASSO regressions had an average R-squared value of around 0.8039, with the income regression having the highest R-squared value out of the four regressions. For the four Random Forest regressions, the R-squared value had an average of 0.7366. The MSE for LASSO regressions was on average lower for the RF regressions, indicating that the predicted values for LASSO on energy consumption were more accurate than those in the RF regressions. Out of all the coefficients in each of the regressions, the income regressor had the largest set of coefficients, indicating that this was the biggest predictor of energy consumption in the US when taken everything else into account. By running each of the regressions individually, it exposes the values that represent each coefficient.

When looking at the visuals, Figure 1 and Figure 2 depict the four LASSO and Random Forest regressions, respectively, overlayed on top of each other, where the actual energy consumption is represented by the dots and the predicted energy consumption is represented by the line. Visually, it is clear to see that the LASSO regression does a better job of fitting the best predicted line, as it intersects a clearer set of points on the graph.

Another significant result was evaluating the effects of the work from home data as well as the energy price data. The coefficients for these two regressors in each regression were significant, but not as large as the other coefficients that were included in each regression. On average, the coefficient for the energy price data was larger, negatively, than the work-from-home coefficient. Given that the energy data was split between electricity and gas prices, the electricity price had a larger effect than the gas prices on consumption, a more inverse relationship on energy consumption. An example of this can be seen in Figure 3, where the coefficients in this scenario are from the house size regression. The work-from-home coefficient was usually always in a lower positive range, indicating the positive relationship with energy consumption.

Conclusion

In this research project, LASSO and Random Forest regression models are employed to explore the impact of various household characteristics on energy consumption. The models demonstrated strong performance, validating the significance of the included variables. Particularly notable was the robustness of the LASSO model, which adeptly managed the available dataset despite its complexities.

Despite challenges with cleaning the data and preparing it for regression, the project yielded insightful conclusions. The regressions exposed the fact that the four household characteristics evaluated are significant predictors of energy consumption, affirming the initial research question. For instance, larger households and incomes were associated with increased energy consumption, likely due to larger living space and accessibility to amenities.

While the remote-work percentage was initially hypothesized to have a substantial impact on energy consumption, the actual effect observed was smaller than expected. In contrast, energy prices exhibited a larger influence, underscoring their critical role in household energy dynamics. This outcome suggests that while remote work does alter energy use, its effect is overshadowed by economic factors such as cost.

The research highlighted the complexity of energy consumption behaviors, driven by a multitude of factors, each contributing variably. This project underscores the importance of considering a wide array of variables to truly understand and predict energy consumption patterns. The findings serve not only as a basis for further research but also provide actionable insights that could inform policy decisions aimed at improving energy efficiency and managing consumption more effectively.

Bibliography

1. "Economic Release Finder." *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 10 Sept. 2014, www.bls.gov/regions/.

2. Haan, Katherine. "Remote Work Statistics and Trends in 2024." *Forbes*, Forbes Magazine, 9 Jan. 2024, www.forbes.com/advisor/business/remote-work-statistics/.

3. Iea. "The Energy World Is Set to Change Significantly by 2030, Based on Today's Policy Settings Alone - News." *IEA*, 24 Oct. 2023, www.iea.org/news/the-energy-world-is-set-to-change-significantly-by-2030-based-on-today-s-policy-settings-alone.

4. "U.S. Energy Information Administration - EIA - Independent Statistics and Analysis." *EIA*, www.eia.gov/consumption/residential/data/2009/index.php. Accessed 12 May 2024.

5. "U.S. Survey of Working Arrangements and Attitudes (SWAA)." *WFH Research*, wfhresearch.com/data/. Accessed 12 May 2024.

Graphs and Tables

Figure 1

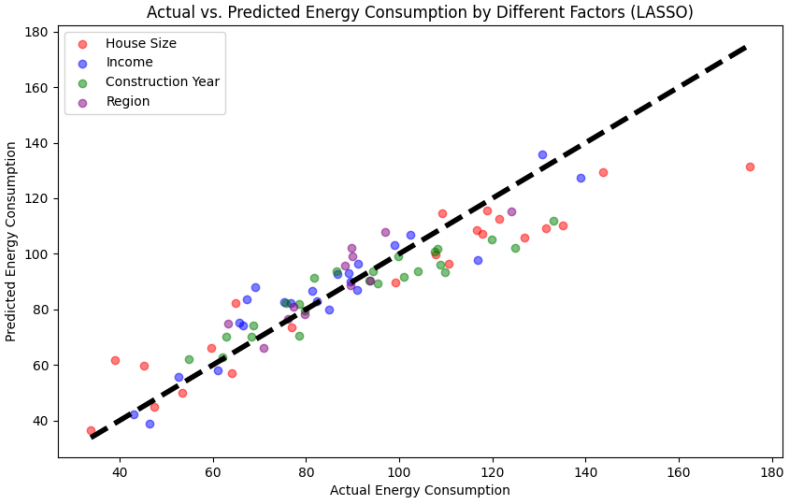

Figure 2

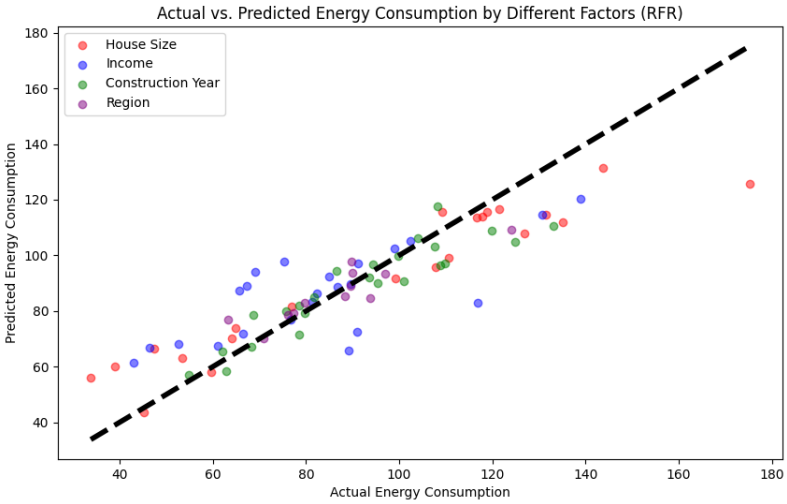

Figure 3

Feature	Coefficient
Year	-1.664351
Average Electricity Price	-6.312181
Average Gas Price	-0.635983
WFH_share	1.074474
Region_Midwest	15.462846
Region_Northeast	16.784083
Region_Northeast	0.620485
Region_South	-10.123923
Region_West	-7.439349
Region_south	-37.937455
House Size_1,000 to 1,499	-19.625060
House Size_1,500 to 1,999	-0.366150
House Size_2,000 to 2,499	5.572937
House Size_2,500 to 2,999	16.325607
House Size_3,000 or more	39.936332
House Size_3,500 to 3,999	17.889284
House Size_4,000 or More	36.961028
House Size_500 to 999	-28.102955
House Size_Less than 1,000	-41.947006
House Size_Less than 500	-39.075842