Predicting the LIHEAP Funding Formula: Final Report

PPOL 564 - Data Science I: Foundations - Fall 2021

Alia Abdelkader

16 December 2021

Word Count: 2860

Github Repository: https://github.com/abdelkaderalia/FinalProjectPPOL564

1. Introduction

Funding for social safety net programs has long been a contentious topic, and the federal government uses various different program structures to allocate funds to those who might benefit. One of those program structures is a block grant, which uses a formula to allocate available funds to different parties based on a collection of data inputs. Opponents of block grants believe that this structure can be prohibitive to an equitable distribution of funds, because formulas may not be dynamic enough to capture all factors that are relevant in considering which parties should receive more funding. The aim of this analysis is to use LIHEAP as a case study and attempt to predict its funding formula, and analyze which formula inputs are most influential in determining what percentage of appropriated funds each state should receive. This report follows my efforts in data collection, wrangling, and analysis, followed by the machine learning component, and evaluation of results.

2. Problem Statement and Background

The Low Income Home Energy Assistance Program (LIHEAP) has been administered by the Department of Health and Human Services (HHS) since 1981 to provide low-income households with financial assistance to cover their energy bills and weatherization of their homes.

LIHEAP is one of only 21 federally funded block grants as of 2020.[1] The funding structure of a black grant dictates total sum of federal benefits is distributed by using a funding formula that allocates a percentage of the total funds appropriated by Congress to each grantee (U.S. state, territory, or indigenous tribe). States then distribute the funds to eligible households that apply for benefits.

The LIHEAP Formula was inherited from its predecessor program, the Low Income Energy Assistance Program (LIEAP), which only operated for one year in 1980. This "old" formula favored cold-weather states in terms of funding because it only used data from specific years before 1980, so funding percentages were static.

Congress' 1984 reauthorization of LIHEAP dictated that the formula should be updated yearly with recent population and energy data.[2] The exact LIHEAP formula calculation is not published, but there is awareness of the general metrics and data sources used as formula inputs. This project aims to call upon the same data sources used by LIHEAP and predict the percentage allocation to each state, with one caveat.

The 1984 LIHEAP statute outlined two "hold-harmless" measures to attempt a more equitable funding distribution.[3] Firstly, if the total LIHEAP appropriation for a given fiscal year exceeds $1.975 billion, then no

[1] ("Block Grants: Perspectives and Controversies" 2020)
[2] ("The LIHEAP Formula", 2019)
[3] ("The LIHEAP Formula", 2019)

state may receive less funding than they were allocated in 1984. Secondly, if the total appropriation exceeds $2.25 billion, then any state that would receive less than 1% under the formula calculation must receive whatever percentage they would have received of a $2.14 billion allocation. These two provisions, which are applied year to year, reduce the percentage for some states and increase it for others.

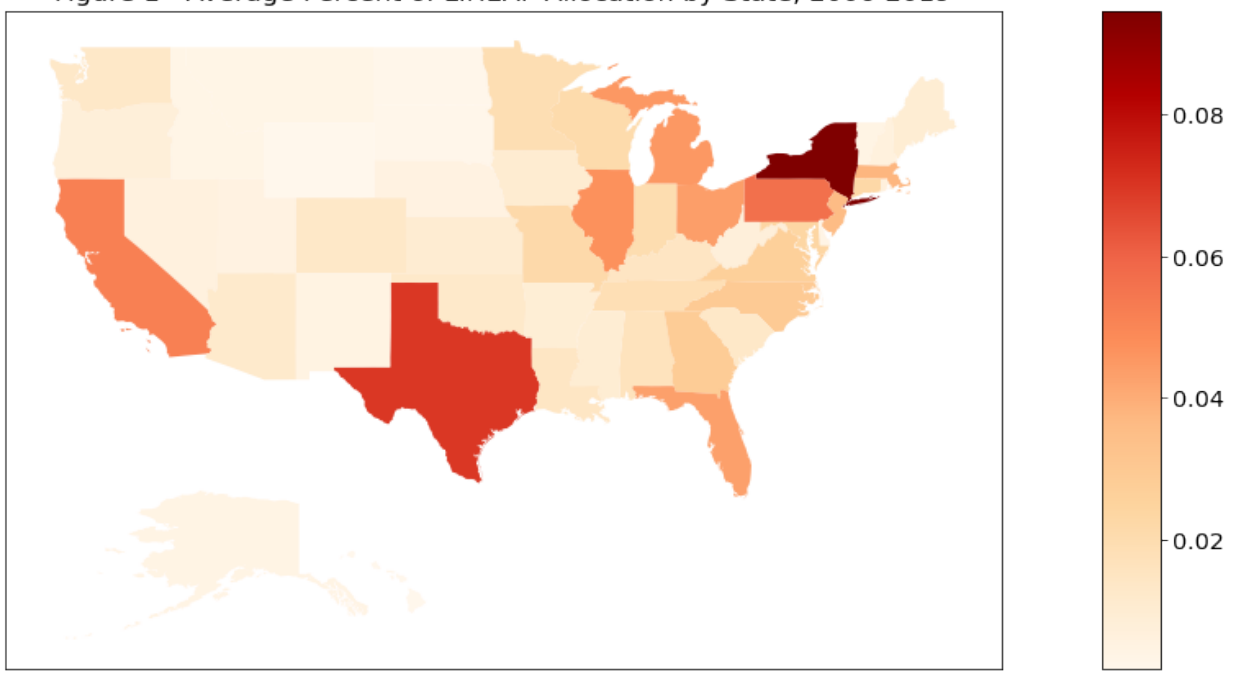
In my analysis, I have elected to exclude hold-harmless provisions and use the available data to predict the percentage that a state would have received without them. This allows me to more accurately assess the permutation importance of different variables in the prediction of the actual percentages. Hold-harmless provisions, which are triggered primarily by the size of the total allocation, reduce or increase allocations to an arbitrarily level that tells us very little about how the other data affected a state's allocation in a given year.

3. Data

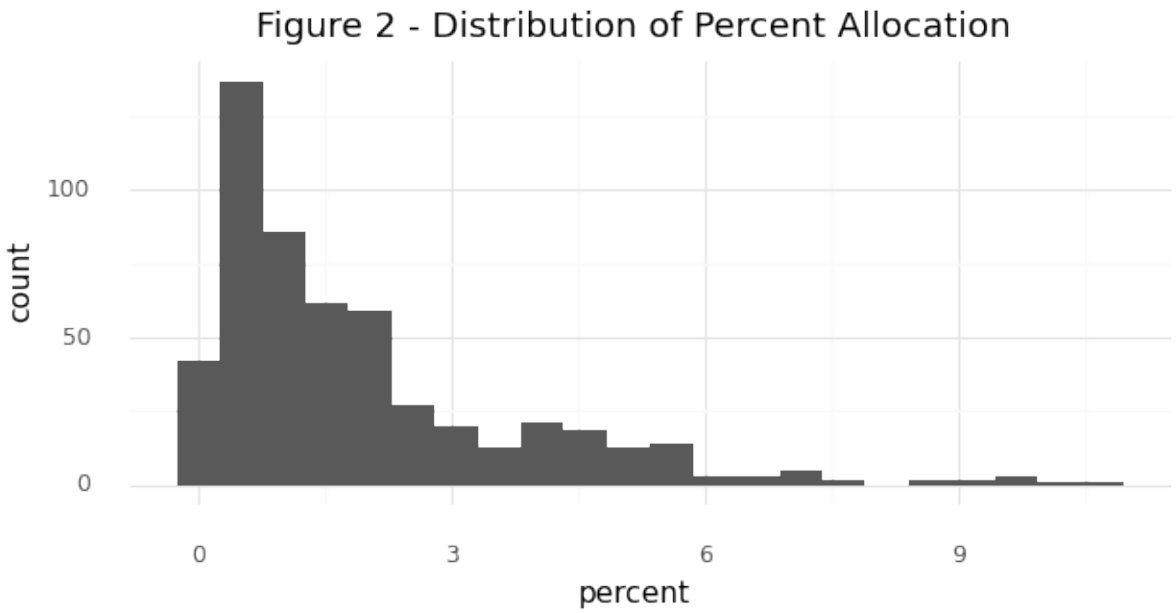
The Congressional Research Service report names several datasets that I pulled various predictors from. The unit of analysis is State-Year, and the final dataset includes all 50 states and Washington, DC for the years 2006 to 2019. There is no missingness in the data. The data wrangling process was very involved and required different steps for each dataset. I used **pandas** to manipulate and merge all dataframes from different sources, and the final dataframe used in my analysis had 714 observations and 36 predictors.

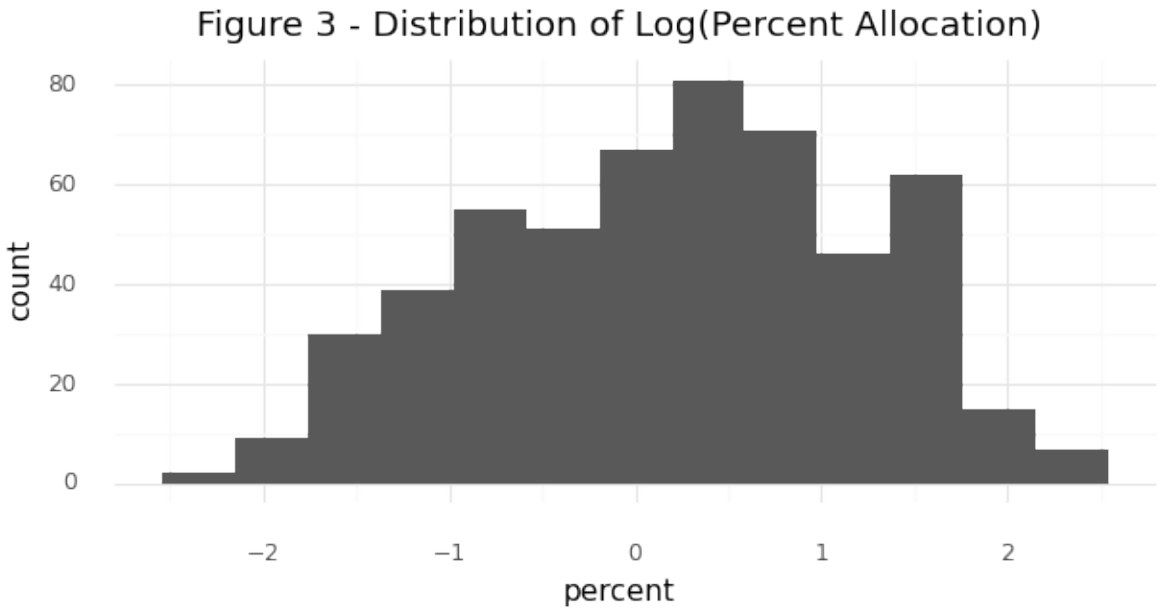
Dependent Variable The dependent variable is the percent of LIHEAP funds allocated to each state in a given year, before hold-harmless measures were applied. Because the actual LIHEAP formula is not available, I received this data directly from the LIHEAP Program, whom I work with as a scholar with the Massive Data Institute. A map of average percent allocation to states from 2006-2019 can be found in Figure 1 below. This plot shows that most states would receive far less than 1% of LIHEAP funds were it not for hold-harmless provisions. Predictably, large states like California and Texas would receive a proportionately higher percentage of the allocation.

Figure 1 - Average Percent of LIHEAP Allocation by State, 2006-2019



In order to combat right-skewness of the dependent variable, I applied a log transformation as shown in Figures 2 and 3 below. When tuning and comparing my models, I tested models that used both Percent_Allocation and log(Percent_Allocation), and the log transformation led to better model performance.





Independent Variables

Temperature Variables I used state-level data from the National Oceanic and Atmospheric Administration (NOAA) for variables relating to temperature, including Heating Degree Days (HDD), Cooling Degree Days (CDD)[4]. Also, from NOAA by way of the Department of Commerce, I pulled 1970-2000 data that provides a 30-year average of HDD weighted by state population.[5] I used **BeautifulSoup**[6] to scrape urls to download the "December" HDD and CDD files to pull the cumulative statistics for each year, used **urllib** to download each file, and used **pandas** to read each one as a fixed width file to add the data to a single data frame. This was a particular challenge and required a complex function to loop through the each file and manipulate the columns to create the desired **pandas** dataframe. Several corrections needed to be made to fix state names

[4]("Degree Day Statistics", 2021)
[5]("State, Regional, And National Monthly Heating Degree Days",2000)
[6](Richardson, 2014)

that were spelled incorrectly. *The code written to scrape and clean temperature data can be found by clicking here.*

I also created "Lag_HDD" and "Lag_CDD" variables, which is the state's HDD or CDD from two years prior. The 30-year weighted HDD was pulled manually from a single PDF for all 51 states; PDF scraping proved to be too laborious due to the complicated metadata of the file.

Initially, planned to include more data from NOAA, specifically on the count of extreme weather events in each State-Year that cause at least $1 billion in damage. I considered this because one of the sanctioned uses of LIHEAP funds is residential weatherization, meaning that households can apply for funds to cover the costs of weatherproofing their homes. In fact, LIHEAP actually allocates more funds for weatherization than the Department of Energy's Weatherization Assistance Program (WAP). However, the literature on the LIHEAP Formula makes no mention of any specific variables to account for variation in weatherization needs by state: this makes sense because the bulk of LIHEAP funds are spent on energy bills. Therefore, I did not include this variable in my analysis, as I was attempting to match the LIHEAP formula as closely as possible.

Energy Variables Using the the Department of Energy - Energy Information Administration's published data from the State Energy Data System (SEDS)[7], I pulled together variables on annual residential energy consumption, expenditure, and pricing by state (each one with its own data file), and manipulated them in **pandas**, looking up values by state and year in the source dataframe and filling in the output dataframe. *The code written to wrangle energy data can be found by clicking here.* A table of those variables can be found in Figure 4.

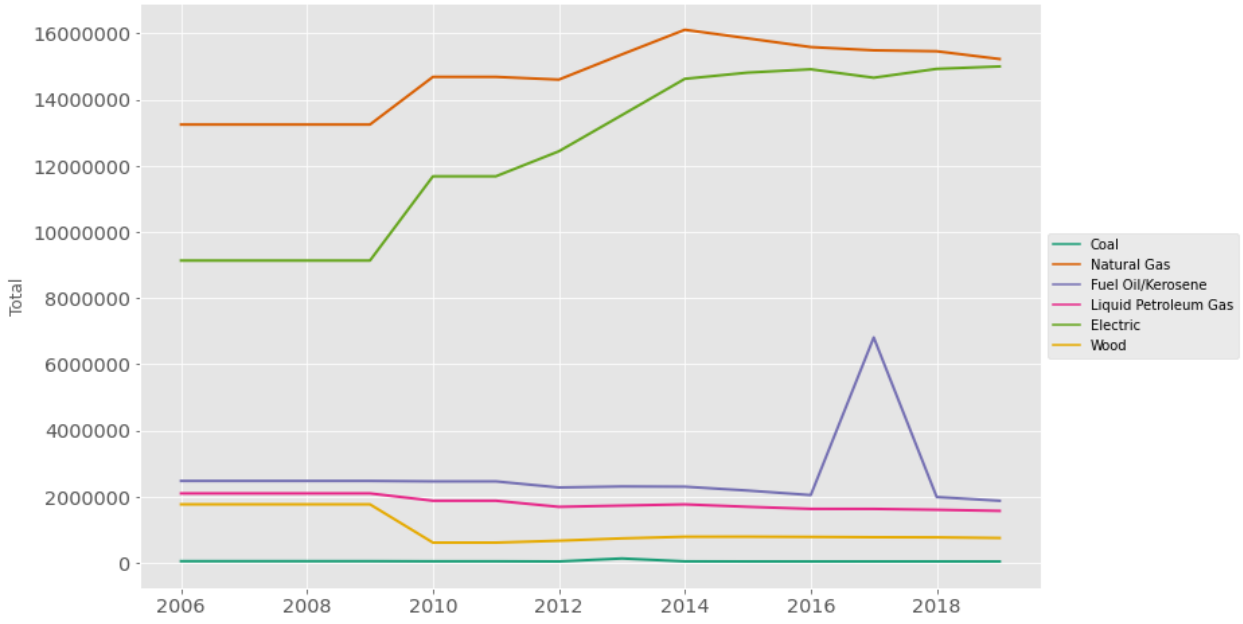
Figure 4 - SEDS Variables

Variable Name	MSN Code	Description	Unit
Coal_Consumed	CLRCB	Coal consumed by the residential sector	Billion Btu
Coal_Price	CLRCD	Coal price in the residential sector	Dollars per million Btu
Coal_Exp	CLRCV	Coal expenditures in the residential sector	Million dollars
FO_Consumed	DFRCB	Distillate fuel oil consumed by the residential sector	Billion Btu
FO_Price	DFRCD	Distillate fuel oil price in the residential sector	Dollars per million Btu
FO_Exp	DFRCV	Distillate fuel oil expenditures in the residential sector	Million dollars
EL_Consumed	ESRCB	Electricity consumed by (i.e., sold to) the residential sector	Billion Btu
EL_Price	ESRCD	Electricity price in the residential sector	Dollars per million Btu
EL_Exp	ESRCV	Electricity expenditures in the residential sector	Million dollars
LPG_Consumed	HLRCB	Hydrocarbon gas liquids consumed by the residential sector	Billion Btu
LPG_Price	HLRCD	Hydrocarbon gas liquids price in the residential sector	Dollars per million Btu
LPG_Exp	HLRCV	Hydrocarbon gas liquids expenditures in the residential sector	Million dollars
KE_Consumed	KSRCB	Kerosene consumed by the residential sector	Billion Btu
KE_Price	KSRCD	Kerosene price in the residential sector	Dollars per million Btu
KE_Exp	KSRCV	Kerosene expenditures in the residential sector	Million dollars
NG_Consumed	NGRCB	Natural gas consumed by (delivered to) the residential sector	Billion Btu
NG_Price	NGRCD	Natural gas price in the residential sector	Dollars per million Btu
NG_Exp	NGRCV	Natural gas expenditures in the residential sector	Million dollars
Total_Energy_Consumed	TERCB	Total energy consumed by the residential sector	Billion Btu
Total_Avg_Price	TERCD	Total energy average price in the residential sector	Dollars per million Btu
Total_Exp	TERCV	Total energy expenditures in the residential sector	Million dollars
Wood_Consumed	WDRCB	Wood energy consumed by the residential sector	Billion Btu
Wood_Price	WDRCD	Wood energy price in the residential sector	Dollars per million Btu
Wood_Exp	WDRCV	Wood energy expenditures in the residential sector	Million dollars

Also, from HHS I received their estimates of the number of Low-Income Households that use each fuel source, which they compute based on the American Community Survey Public Use Microdata Sample (PUMS). A plot of those estimates can be found in Figure 5 below. As shown, across all states, Natural Gas is the most commonly used fuel source, followed by Electric.

[7]("State Energy Data System (SEDS)", 2021)

Figure 5 - Estimated total of Low Income households using each fuel source



Eligible Households Data on the number of households that are federally eligible for LIHEAP was pulled from the LIHEAP Performance Management Data Warehouse.[8] This metric is computed based on five-year data from the American Community Survey, based on a maximum eligible income level of whichever is higher: 150% of the poverty guidelines (FPG) or 60% of a state's median income (SMI).

After pulling data from each source into its own data frame, I merged all of them together into one final dataframe, joining on State-Year as the identifier. State names were standardized, which generally required correcting other forms of DC to "District of Columbia" and converting state abbreviations to state names using a dictionary. All state names were converted to UPPERCASE to avoid any confusion.

4. Analysis

I used Python's **Scikit-learn** package to execute the machine learning models in this analysis.[9] After splitting the data 75/25 between training and test data, I used a machine learning pipeline to automate the workflow. This pipeline included pre-processing (which involved scaling the predictors), model tuning, cross validation, evaluating of model accuracy, and model selection. Cross validation was done with the k-Fold validation method; it is a key part of this process, as it allows us to select best model with only part of the training data as an input, reserving a small subset to use as a proxy for test data. After the user specifies a number for "k" (in this analysis, k=5), the k-Fold method splits the data into k groups, creating new "training" subsets, and then repeats this process and computes an average estimate across all groups and iterations.

All visualizations in this report were generated using **seaborn**[10], **plotnine**[11], **matplotlib**, **GeoPandas**[12], and **SKLearn**.

Modeling As previously stated, the dependent variable in this analysis was the natural log of the percent allocation to each state in a given year. The machine learning pipeline included 5 different types of models to predict this continuous outcome, which I will describe below.

[8] (LIHEAP Data Performance Management Data Warehouse, 2021)
[9] (Pedregosa, 2011)
[10] (Waskom, 2020)
[11] (Kibirige, 2020)
[12] ("GeoPandas Documentation", 2021)

The first model used was a Linear Model, which establishes a linear relationship between the combination of the predictors included in the specification and the response variable.

The second model used was Decision Tree Classifier, which uses various decision criteria based on the values of the model's predictors to go down different "branches of the tree. Based on these binary decisions (for example,"if the value of X is greater than 10, go down this branch"), each prediction is determined by the end node of the decisions made.

The third model was the Bagging Decision Tree Regressor, which is similar to the Decision Tree Classifier in that it builds many different decision trees and computes the mean of the output classes across all trees.

The fourth model was the K-Nearest Neighbors Classifier, which, based on specified parameter values of k, identifies k other points from within the training dataset that are deemed close or similar to the observation for which a prediction is being generated based on the values of the independent variables, and then assumes that the observation's outcome value will be similar to that of its "neighbors."

The fifth and final model was the Random Forest Classifier, which is also similar to a Decision Tree in that it builds many different trees using a random subset of predictors from the training data, and then generating a precition based on an average across all of the trees that were created.

In addition to testing and comparing different machine learning models, I also tested different model specifications to account for the fact that I had many different predictors and the best model may include a different combination of them. For the most part, I included all of the predictors that I collected in my models, except for *State*, *Year* and *State_Year*. I tested 4 different model specifications:

1. Used log(*Percent_Allocation* * 100) as Y, and used *Total_HDD* and *Total_CDD*, not lagged variables. Included all other predictors
2. Same as Specification 1, but excluded the variable *Total_Exp* to better evaluate permutation importance of other variables
3. Same as Specification 2, but did not apply a log transformation to *Percent_Allocation*
4. Same as Specification 2, but used *Lag_HDD* and *Lag_CDD*

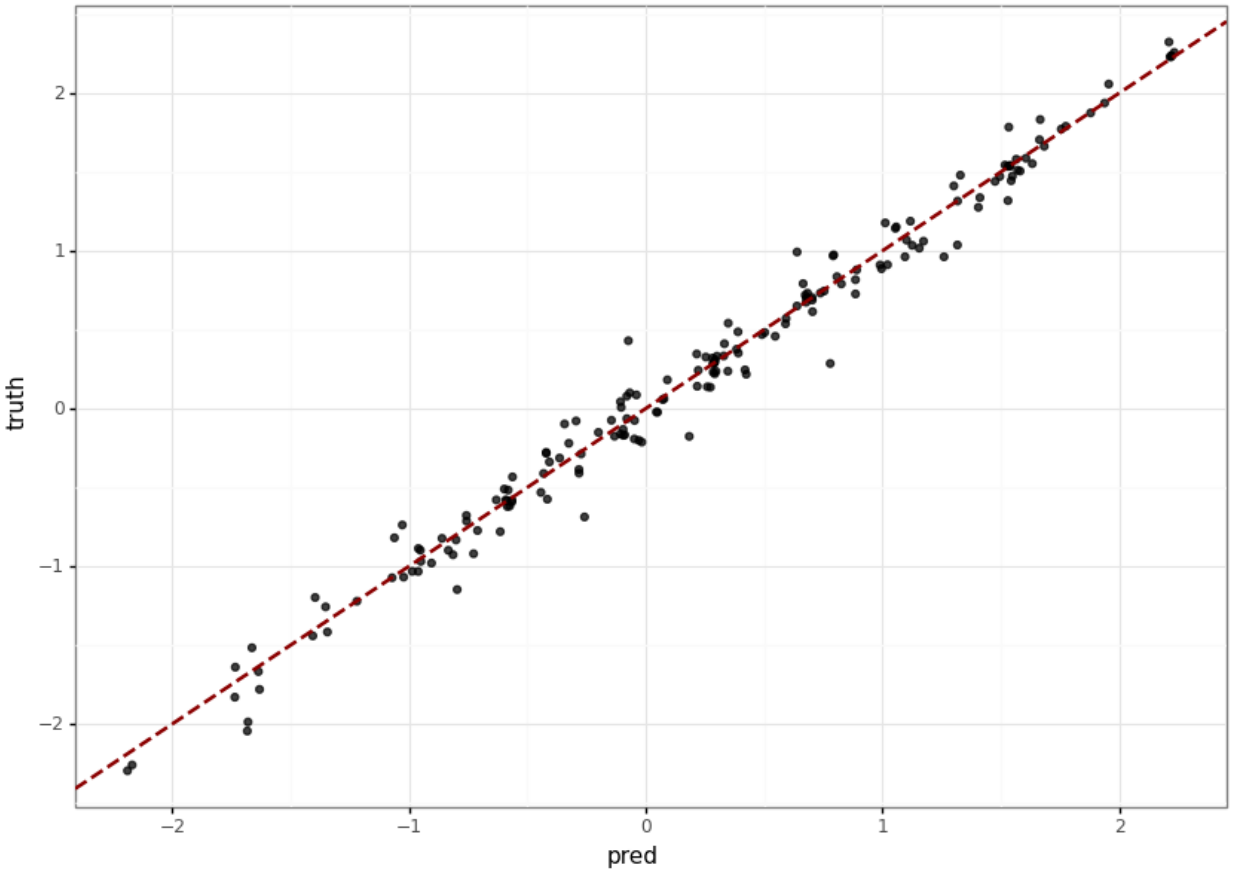
Total_Exp was excluded in the latter specifications because its permutation importance (as discussed later on) far exceeded that of all the other predictors, making it impossible to see which other variables were at all important.

I used mean squared error (MSE) as the metric to evaluate model performance; this method measures the average squares of errors, which are the differences between the model's fitted values and the true values of the observations in the data. The best performing model will have the lowest MSE.

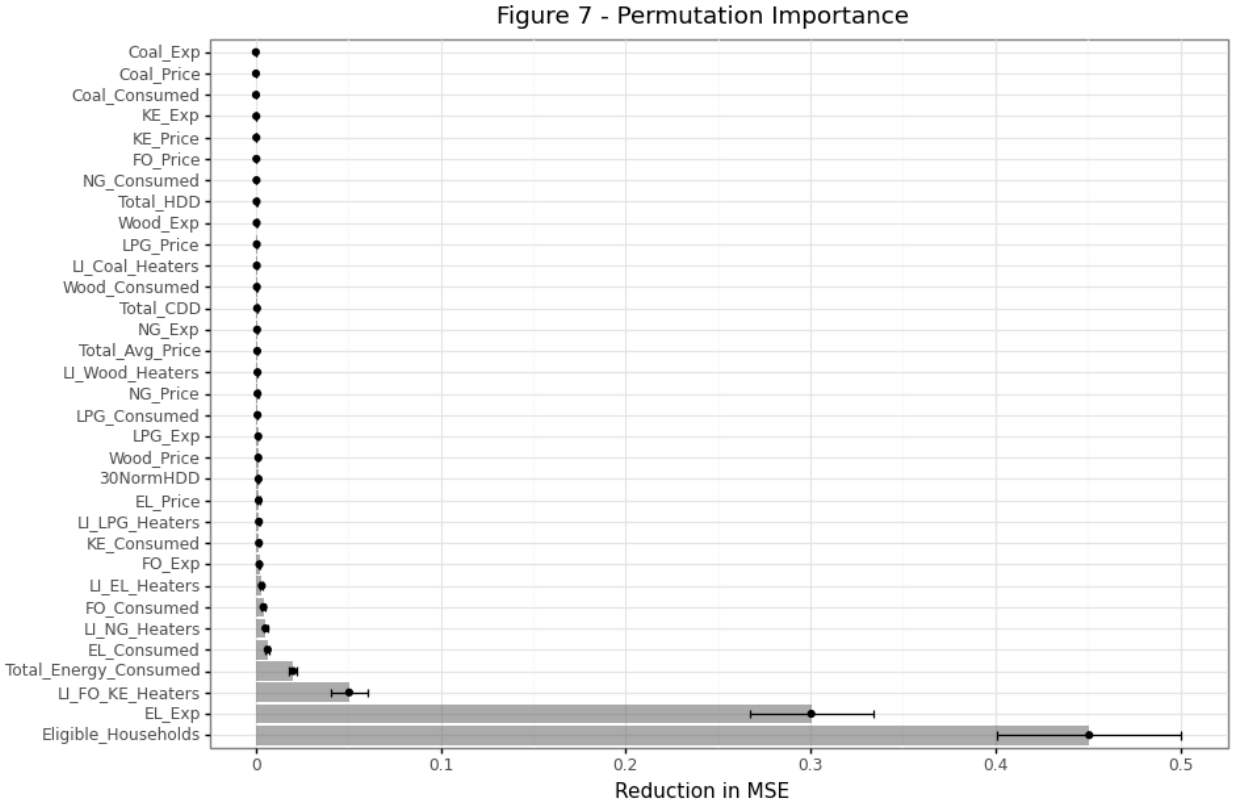
5. Results

The best performing model based on minimizing MSE was the Bagging Regressor, and the best model specification was Specification 2. Upon evaluating the results, we found that the MSE was 0.0167 and the R^2 value, which represents the proportion of variation in the data that can be explained by the model, was .9834. Figure 6 below shows the fit of the model. *The code for this model can be found by clicking here.*

Figure 6 - Best Model Fit



These results demonstrate that the predictors included in the model are strongly correlated to the response. Part of the aim of this project was to evaluate which of the predictors was most influential. A plot permutation importance of the model's predictors is shown in Figure 7 below.



As we can see, the most influential variable in the reduction of MSE was the number of eligible households. This is not a particularly surprising finding, as this statistic is likely to be correlated to state population and it is logical that states with more people will have more people living below the poverty line. However, some interesting findings emerge: that the next two most influential variables are electricity expense and the number of low income households using Fuel Oil/Kerosene as their heating source. I would have expected variables relating to Natural Gas to have higher importance than variables relating to Fuel Oil/Kerosene, given the earlier plot showing that Natural Gas was the fuel source used most commonly by low income households across the entire time span of the data. This is a finding that would definitely be worth investigating further in future analysis.

I also expected that variables related to temperature would have been more influential, given global trends relating to climate change and the expected effect that this would have on extreme weather and subsequent heating and cooling costs. This is another avenue of future analysis that could be worth pursuing.

The remaining predictors had very low permutation importance and were therefore not influential in predicting the percent allocation. I did consider removing these predictors from the specification, but ultimately decided to include them because, based on literature, the actual LIHEAP formula does include variables on energy consumption, pricing, and expenditure, and temperature data, in some form or other. Because I wanted to match the formula as closely as possible, it would not be accurate to remove them from the specification.

6. Discussion

Overall, this project was successful in predicting the LIHEAP Formula with high accuracy, despite not knowing the exact variables that the HHS uses in the formula. Based on the high predictive accuracy despite the relatively small sample size of 714 total observations, it is definitely possible that overfitting was an issue in this analysis. This occurs when the machine learning model fits exactly on the training data, rendering it unable to perform on any unseen data. Further evaluating the model's performance based on not only MSE but also variance could yield more evidence as to whether overfitting has occurred.

Several possible considerations for further analysis would be further investigation into the permuation importance of different predictors, as mentioned above, and the possible relationships between them. Also, testing these models on a much larger dataset could yield distinct results. I limited my frame of analysis to 2006-2019 for several reasons, including computational difficulties related to data collection & wrangling, and limitations to the expanse of data that were publicly available. As I continue my work with LIHEAP through the Massive Data Institute, I will consider building upon and expanding this model to include more observations and different predictors, based on data that may not be publicly available. Further analysis could also include an attempt to factor in hold-harmless measures, which were ignored for the purpose of simplifying this analysis. This could be represented as a dichotomous variable for indicating whether each of the two hold-harmless provisions were triggered based on the total allocation in that year, and by keeping the original percentage allocation as a predictor but using the ex-post hold-harmless percentage as a response. The hypothesis here would be that the hold-harmless variables and original percentage allocation would be the most influential variables in the model, but actually doing this analysis could confirm.

The findings of this analysis could lead to more informed policy discussions about LIHEAP, social safety net programs, and the allocation of federal funds in general. This is a policy area that I hope to explore even further during the remainder of my studies and my future career. I am hopeful that this analysis could be the start of an expansive body of work on the topic of grant-in-aid programs and how they can be designed more efficiently and effectively to ensure that government assistance actually reaches those who need it most.

Works Cited

"Block Grants: Perspectives and Controversies." (21 February 2020). Congressional Research Service. https://sgp.fas.org/crs/misc/R40486.pdf

"Degree Day Statistics." (2 December 2021) National Oceanic and Atmospheric Administration - National Weather Service. https://ftp.cpc.ncep.noaa.gov/htdocs/products/analysis_monitoring/cdus/degree_days/archives/

"GeoPandas Documentation" (16 October 20201). GeoPandas Developers. https://geopandas.org/en/stable/docs.html

Hassan Kibirige, et. al (5 August 2020). has2k1/plotnine: v0.7.1 (Version v0.7.1). Zenodo. http://doi.org/10.5281/zenodo.3973626 1717

LIHEAP Data Performance Management Data Warehouse. (2021) Department of Health and Human Services - Administration for Children and Families. https://liheappm.acf.hhs.gov/datawarehouse

"The LIHEAP Formula." (2 May 2010). Congressional Research Service. https://sgp.fas.org/crs/misc/RL33275.pdf

Michael Waskom, et al. (2020, September 8). mwaskom/seaborn: v0.11.0 (September 2020) (Version v0.11.0). Zenodo. http://doi.org/10.5281/zenodo.4019146

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Pytho nJournal of Machine Learning Research, 12, 2825–2830.

"PERCENT OF POVERTY GUIDELINES FOR LIHEAP COMPONENTS". (2021) Department of Health and Human Services - Administration for Children and Families. (2020)https://liheapch.acf.hhs.gov/tables/POP.htm

Richardson, Leonard. (2014). Beautiful Soup Documentation. https://beautiful-soup-4.readthedocs.io/en/latest/

"State Energy Data System (SEDS): 1960-2019 (complete)". (29 October 2021). Department of Energy - Energy Information Administration. https://www.eia.gov/state/seds/seds-data-complete.php?sid=US#Consumption

"State, Regional, And National Monthly Heating Degree Days" (2000). Department of Commerce. https://www.ncei.noaa.gov/data/climate-normals-deprecated/access/hcs/HCS_51.pdf