# Residential Electricity Energy Consumption Prediction using Machine Learning Methods

By: Nehemie Joseph, Yashaswini Thokala, Jaya Keerthi Varagani

## Abstract:

In early 2022, the International Energy Agency warned of an energy crisis. In early February, the Guardian posted a graphic representation that depicts the electricity prices increasing by 12% within two years, starting with 2019 as the baseline. The goal of the research paper is to use various machine learning techniques to calculate how electricity is calculated. Once calculated, use the findings to back up a fictitious startup, Nucleus, that empowers homeowners about electricity usage via the product selling of Smart meter. The proposal is that the presence of smart meter forms an educated customer and acts as a deterrent on electricity prices. The source of this proposal is the 2015 Residential Energy Consumption Survey from the United States' Energy Information Administration. The data contained over 700 independent variables across over 5,000 observations that represent the United States population of 118 million in 2015. To decrease the number of independent variables, domain knowledge and Variance Inflation Factors were used to weed out multicollinearity, leaving little under 150 predictor variables. Based on 6 machine learning techniques, Random Forrest Regressor predicted the best energy consumptions at 80%- smart meter present. The top 10 features importance discovered that common household characteristics and regional location plays a significant role in energy consumption. Improvements within the prediction models and cleaning processes are elaborated.

## Introduction:

As technology grows and global climate change intensifies, residential energy use, also called home energy use, has been the topic of discussion. Specifically, the development of smart meters that predawn the birth of the internet has evolved in calculating the consumption of household items. Countries, such as the United Kingdom, have incorporated this technology in plans of creating a smart grid. The intention is to push the envelope into having an efficient and cheaper country for emergency purposes while educating the population.

Fast forward to 2022, some factors such as the COVID pandemic and the Russian- Ukraine war brewed a storm of skyrocketing electricity price, reported by the Guardian -12% increase. In 2020 the International Energy Agency warned of a global energy crisis, citing a real-time analysis to backup the change of electricity jumped to an average 4.5%  across seven continents in 2021. Therefore, the need for the general population to monitor electricity usage to either save money or become an educated consumer has grown. With the creation of smart meters, the idea of finding individual usage of electricity will act as a deterrent and educate consumers electricity consumption. Other benefits, such as creating a smart home that will be able to detect abnormal appliance trends to investigate to amplifying the usage of auto HomeTech.

Yet, despite the available technology, the reality is that majority of United States residential do not have a smart meter. Which begs the question, does having a smart meter make a difference in consumption of electricity? What are the significant factors that impact electricity consumption?

## Research Questions and Source:

In this study, we employ machine learning methods to predict residential electricity consumption using various household and weather-related factors. To narrow down our study, we are investigating these research questions:

1. Is there a regional-division significance that needs to be calculated when creating an algorithm to determine residential consumption?

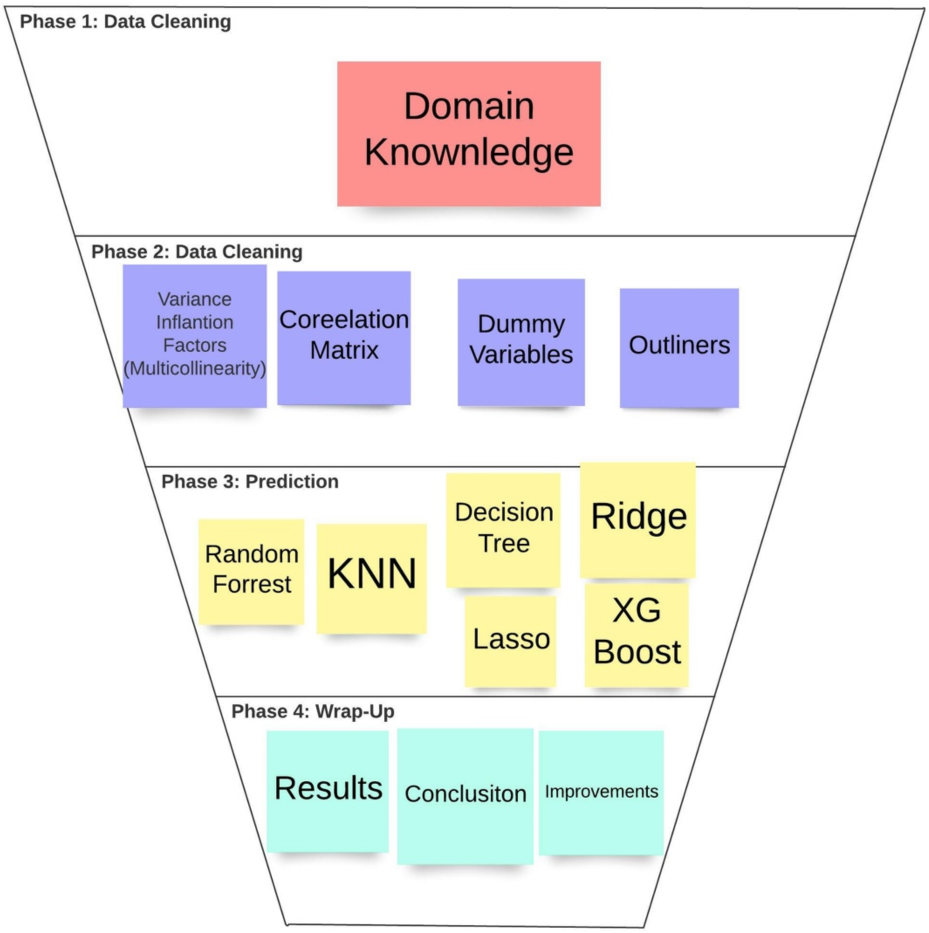
2. How do common household characteristics like the type of housing unit or the number of rooms in the household affect residential consumption?

3. Do factors such as race, income, education, or any other population classifications need to be addressed for residential consumption impact?

We will be using a dataset of more than 5,600 households randomly selected by the U.S. Energy Information Administration (EIA) in 2015. The attributes include household characteristics, final consumption, expenditures, weather, and more, totaling 759 independent variables. The dataset represents 118.2 million households across the United States using one of the four types of energy consumption: electricity, natural gas, propane, and fuel oil/kerosene.

Link: https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata

## Design and Implantation:

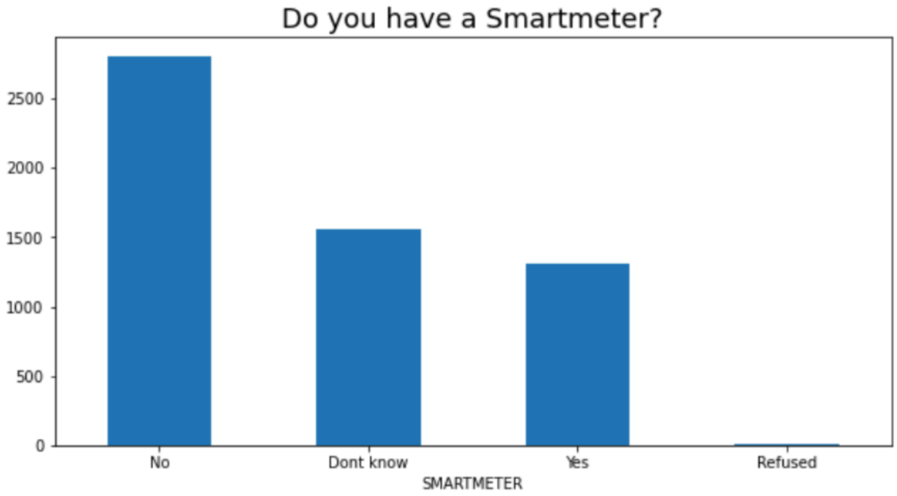


The research was broken into four phases: Phase 1 & 2 - Data Cleaning, Phase 3 - Prediction and Phase 4 - Wrap-Up.

**Phase 1: Data Cleaning**

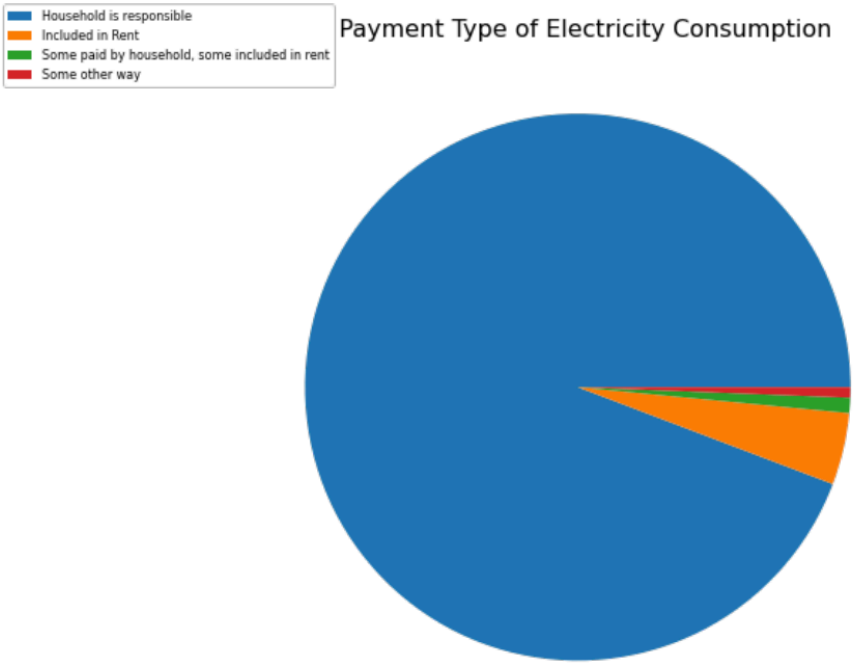
After establishing a basic knowledge of electricity consumption through reading the codebook a general gimp of an action plan was created. With this domain knowledge, descriptive visualization were created to understand the sample. Afterwards, we determine what research questions were feasible.

The first part was to discover how many households have a smart meter.



Estimating that approximately 80-90% of the population do not have a smart meter suggests an imbalance dataset. Therefore, the question: does having a smart meter make a difference in consumption of electricity? Cannot be answered. Performing a correction for the imbalance dataset is not possible since it is highly skewed beyond fix and the dataset is small

The second part to discover is if payment is handled by the household or is the responsibility removed? Our target audience within the research community are residents that are responsible for payments because as a result the value to save money aligns as their responsibility, bill conscious. The pie chart below depicts the sample responsibility in electricity.

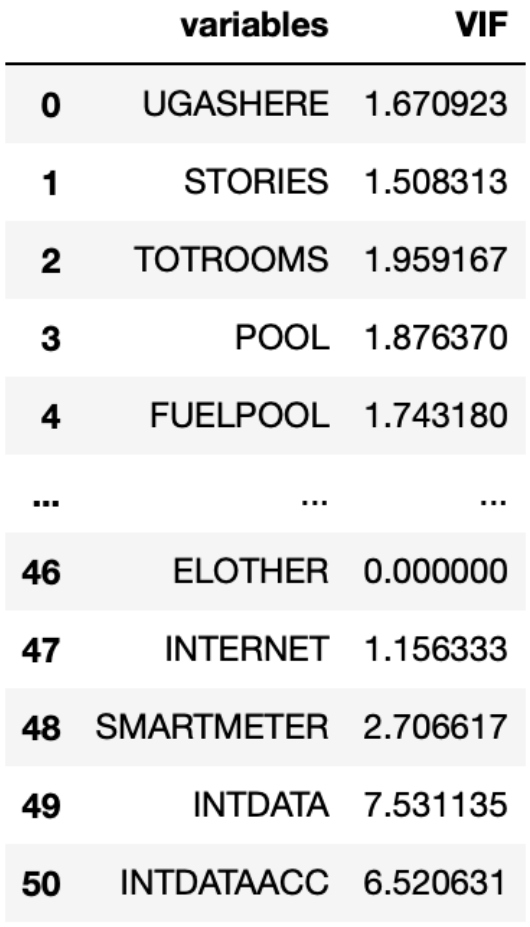


More than 90% of the sample are households that are responsible for payment of electricity consumption.

After evaluating the descriptive outputs and codebook, initial removal of variables was removed. Anything related to smart meter information was kept because of the business concept in mind.

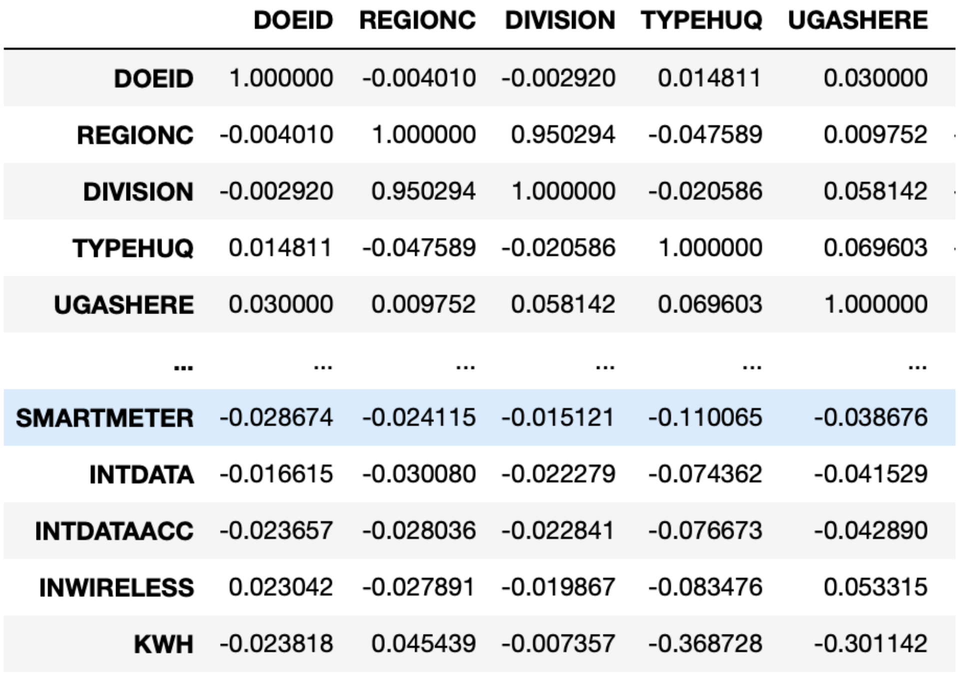
**Phase 2: Data Cleaning**

Due to the high number of independent variables, multicollinearity became apparent. Multicollinearity occurs when there are multiple variables, two or more, are highly correlated within the model. To address this issue, we used a combination of Variance Inflation Factors (VIF) and Correlation matrix to narrow the variables. VIF indicates multicollinearity if the value is high. As shown below, a typical value below 10 is accepted. However, for this research a value of 15 or below was acceptable.



| | variables | VIF |
|---|---|---|
| 0 | UGASHERE | 1.670923 |
| 1 | STORIES | 1.508313 |
| 2 | TOTROOMS | 1.959167 |
| 3 | POOL | 1.876370 |
| 4 | FUELPOOL | 1.743180 |
| ... | ... | ... |
| 46 | ELOTHER | 0.000000 |
| 47 | INTERNET | 1.156333 |
| 48 | SMARTMETER | 2.706617 |
| 49 | INTDATA | 7.531135 |
| 50 | INTDATAACC | 6.520631 |

VIF should not be trusted alone. To double check that important variables were not dropped or if more needed to be dropped, we ran a correlation matrix against the independent variable.



| | DOEID | REGIONC | DIVISION | TYPEHUQ | UGASHERE |
|---|---|---|---|---|---|
| **DOEID** | 1.000000 | -0.004010 | -0.002920 | 0.014811 | 0.030000 |
| **REGIONC** | -0.004010 | 1.000000 | 0.950294 | -0.047589 | 0.009752 |
| **DIVISION** | -0.002920 | 0.950294 | 1.000000 | -0.020586 | 0.058142 |
| **TYPEHUQ** | 0.014811 | -0.047589 | -0.020586 | 1.000000 | 0.069603 |
| **UGASHERE** | 0.030000 | 0.009752 | 0.058142 | 0.069603 | 1.000000 |
| **...** | ... | ... | ... | ... | ... |
| **SMARTMETER** | -0.028674 | -0.024115 | -0.015121 | -0.110065 | -0.038676 |
| **INTDATA** | -0.016615 | -0.030080 | -0.022279 | -0.074362 | -0.041529 |
| **INTDATAACC** | -0.023657 | -0.028036 | -0.022841 | -0.076673 | -0.042890 |
| **INWIRELESS** | 0.023042 | -0.027891 | -0.019867 | -0.083476 | 0.053315 |
| **KWH** | -0.023818 | 0.045439 | -0.007357 | -0.368728 | -0.301142 |

After having the remaining 50 variables, corrected steps to run a model through were taken. Outliners were removed and object type variables were dummy transformed.

**Phase 3: Prediction**

The data set was split into 80% training data and 20% testing data. The dependent variable is KWH, total electricity usage. A total of 6 models were ran.

- **Random Forest** – A random forest is a supervised algorithm which is created using several decision tree algorithms. It used ensemble learning, i.e., a method that puts together several classifiers to finally provide solutions to complex problems. The outcome of this algorithm is based on the predictions made by decision trees. This is done by taking the average of predictions made by different trees. The higher the number of trees, the higher the precision of outcome.

- **XGBoost** - XGBoost is a software library which can be used to implement optimized gradient boosting machine learning algorithms. It stands for eXtreme Gradient Boosting and provides parallel tree boosting and is now considered one of the top machine learning libraries for regression, classification, and ranking problems.
- **K-fold Cross Validation** - K-fold Cross Validation is a statistical method to test the skill of machine learning models by dividing the data into k (k > 1) different folds and testing the models several times, using each of these folds.
- **Decision Tree –** A decision tree is another example of a supervised machine learning example where the data is repeatedly split based on a certain parameter. A decision tree contains mainly two entities – decision nodes and leaves.
- **Lasso Regression –** Least Absolute Shrinkage & Selection Operator, also known as Lasso Regression is a type of regression formula that is used for the regularization of data models and feature selections to get a better statistical model by improving the prediction accuracy and interpretability of the model.
- **Ridge Regression –** Ridge regression is an example of another form of regression that focuses mainly on estimating the coefficients of multiple regression models in situations where the independent variables are highly correlated. This method overcomes the issues of overfitting of a model and deals with multicollinearity.

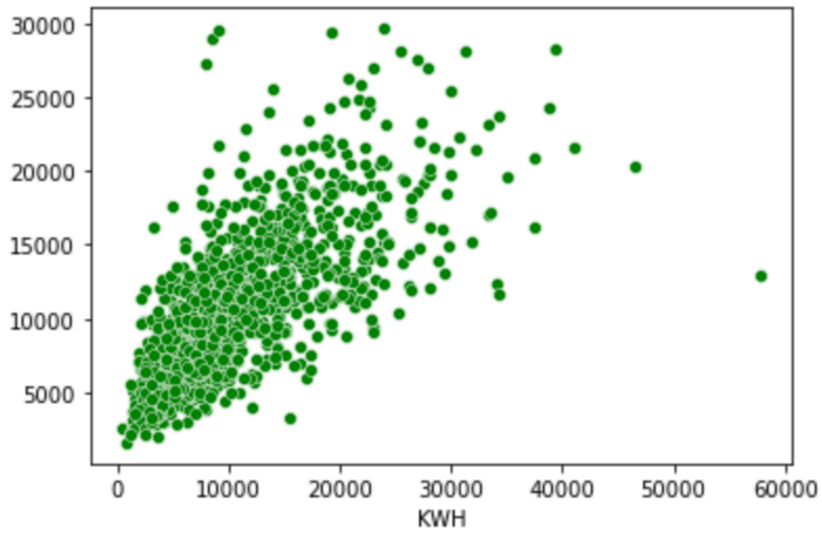
We will be using Python along with the various Python libraries to implement the techniques required to solve our research questions.

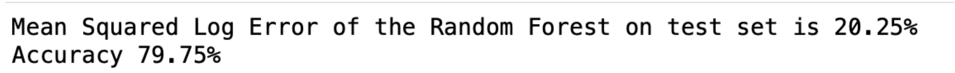
## Results:

Out of the 6 models, 4 models were able to run successfully without causing a value error .
Down below are the four models:

**Random Forest**

```
<AxesSubplot:xlabel='KWH'>
```

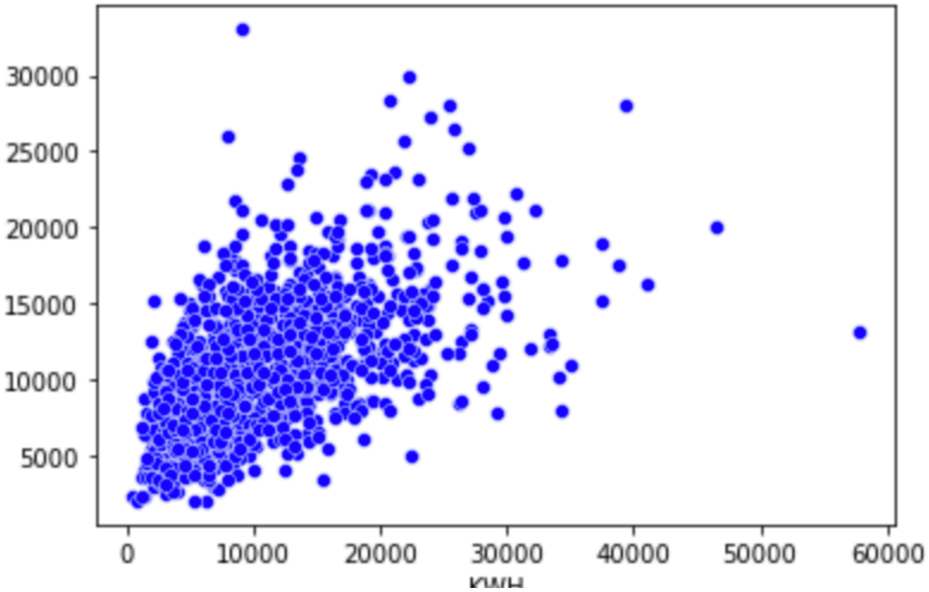


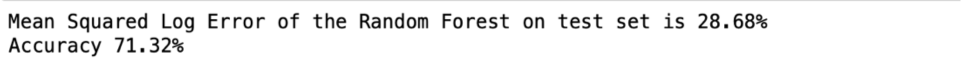


```
Mean Squared Log Error of the Random Forest on test set is 20.25%
Accuracy 79.75%
```

## KNN

```
<AxesSubplot:xlabel='KWH'>
```

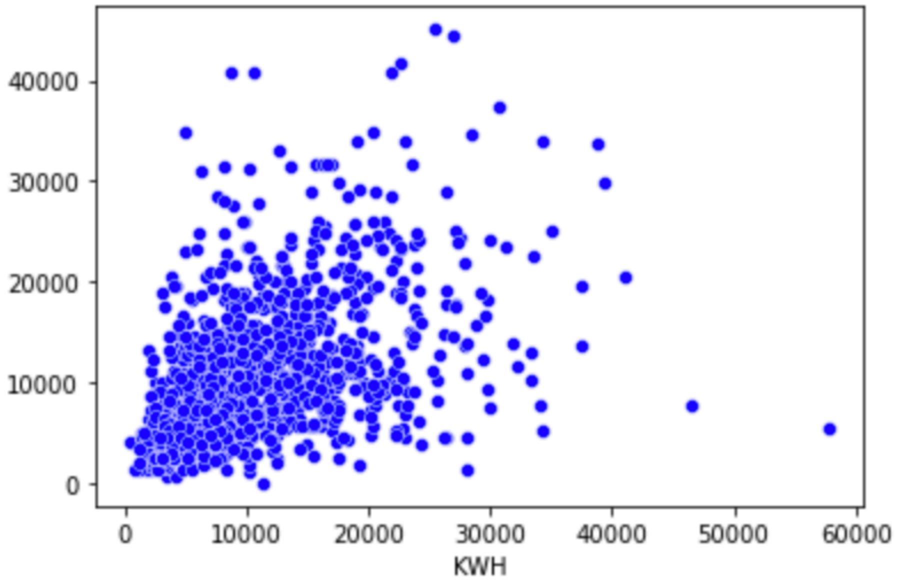




```
Mean Squared Log Error of the Random Forest on test set is 28.68%
Accuracy 71.32%
```

## Decision Tree

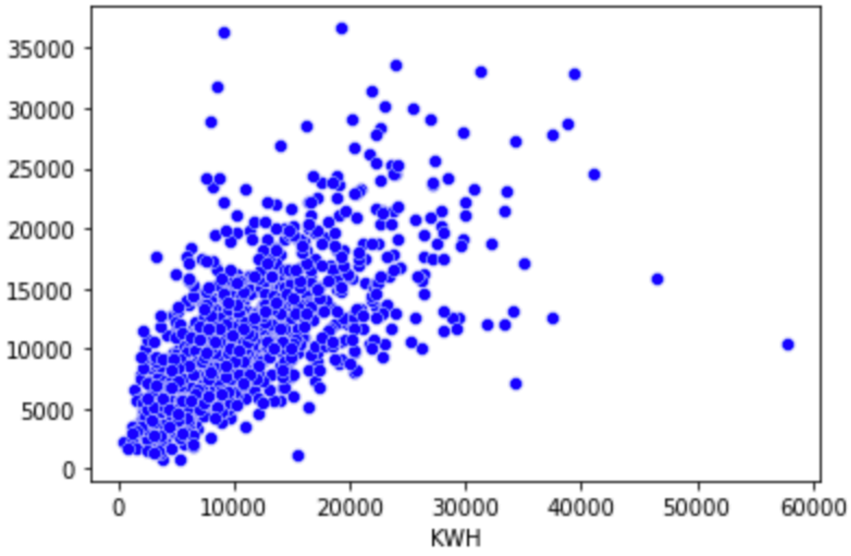
```
<AxesSubplot:xlabel='KWH'>
```

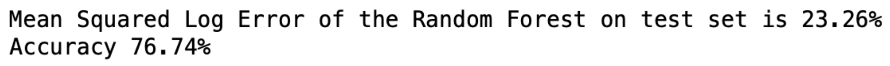


```
Mean Squared Log Error of the Random Forest on test set is 44.84%
Accuracy 55.16%
```

## XGBoost

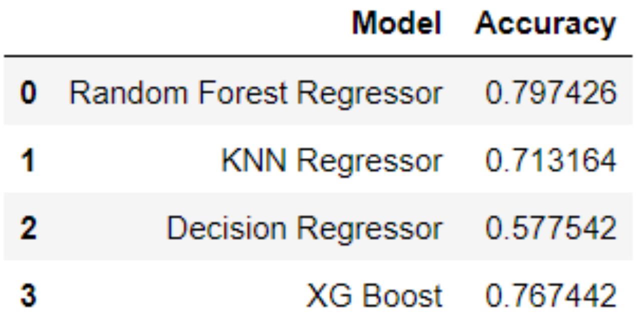
```
<AxesSubplot:xlabel='KWH'>
```





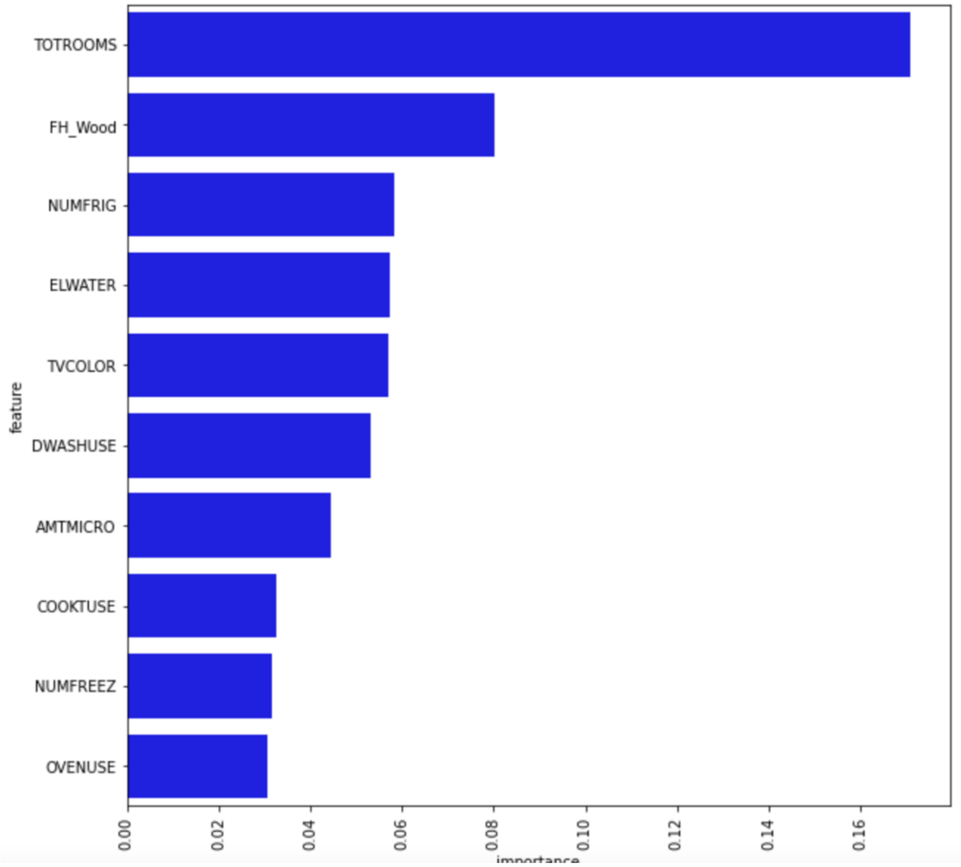
```
Mean Squared Log Error of the Random Forest on test set is 23.26%
Accuracy 76.74%
```

A common occurrence is that all four models suffer from outliners. The removal of these outliners was attempted with little success.



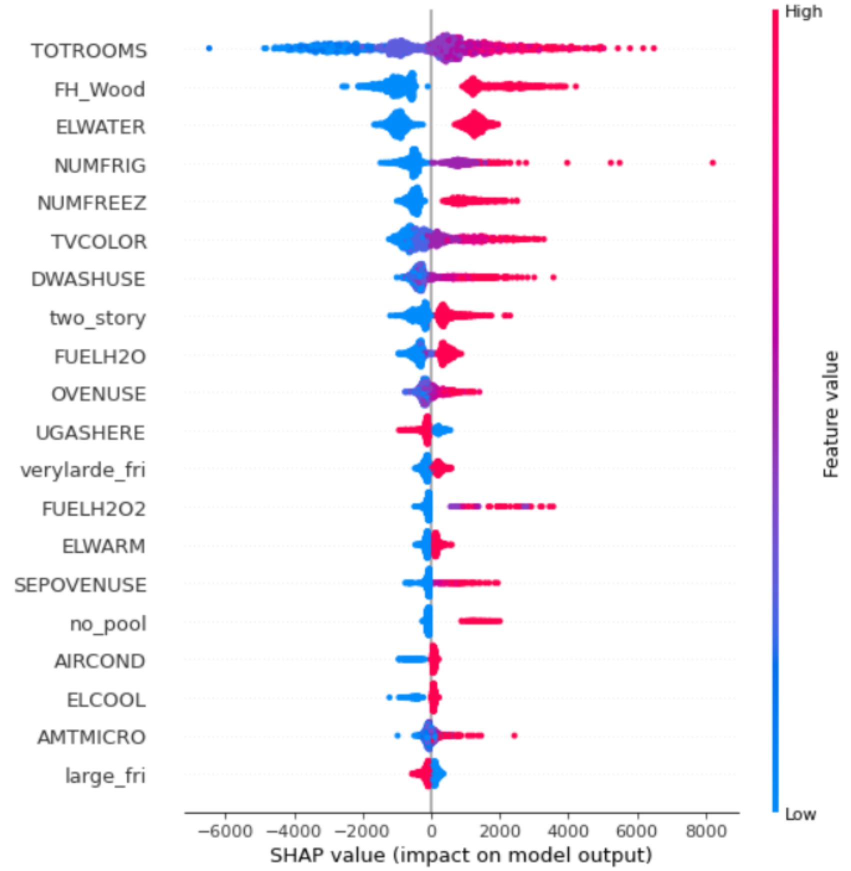
| | Model | Accuracy |
|---|---|---|
| 0 | Random Forest Regressor | 0.797426 |
| 1 | KNN Regressor | 0.713164 |
| 2 | Decision Regressor | 0.577542 |
| 3 | XG Boost | 0.767442 |

Based on accuracy, the random forest regressor is chosen with 80% accuracy. A feature importance selection was then generated.



In order of importance: TOTROOMS- total number of rooms in housing unit, excluding bathrooms, FH_WOOD - Wood used as fuel for main space heating, NUMFRIG- number of refrigerators used, ELWATER- electricity used in water heating, TVCOLOR- number of televisions used, DWASHUSE- frequency of dishwasher use, AMTMICRO-frequency of microwave use, COOKTUSE- frequency of use of cooktop part of the stove, NUMFREEZ-number of separate freezers used, OVENUSE- frequency of use of oven part of stove.

To understand the effect of these independent variables a Shap Value visualization was generated. The order of importance is in decreasing order. Observing the right-hand side visualization, if there is red then the independent variable increases the dependent variable and if blue then it decreases the value.



Not in order of importance: TOTROOMS- increases dependent value, FH_WOOD - increases dependent value, NUMFRIG- increases dependent value, ELWATER- increases dependent value, TVCOLOR- increases dependent value, DWASHUSE- increases dependent value, AMTMICRO- increases dependent value, NUMFREEZ- increases dependent value, LARGE_FRI- decrease dependent value and, UGASHERE-decreases dependent value.

## Conclusion:

Based on feasibility, only one umbrella question could be answered: What are the significant factors that impact electricity consumption? Using the breakdown of the research questions:

1. Is there a regional-division significance that needs to be calculated when creating an algorithm to determine residential consumption?

Yes, there is. This was determined when the variable REGIONC was kept after going through multiple VIF rounds. Later when generating the feature importance, the second most important variable is FH-Wood. FH_WOOD - Wood used as fuel for main space heating. This variable only appears in colder climates indicating that region-division plays a significant role.

2. How do common household characteristics like the type of housing unit or the number of rooms in the household affect residential consumption?

As shown in the feature importance and shap value visualization, common household characteristics play a key role in the formulation of KWH. On average, most independent variables increase KWH.

3. Do factors such as race, income, education, or any other population classifications need to be addressed for residential consumption impact?

Population classification variables were dropped during the VIF process. Due to the multicollinearity these variables were removed.

## Improvements:

The above findings and methods are good for a basic understanding of finding the significance of electricity consumption. However, it does not answer the question, do smart meters make a difference in electricity consumption? Additionally, due to the recent traumatic experiences these past few years the 2015 data alone is not recommended to use for a future comparison outlook in 2022.

Some other recommendations:

1. Increase the data sources to have 2015-2021 surveys. By doing this option, it will result in a better realistic finding that will be useful to apply to 2022. Enough data to balance out the “abnormal” trends and address the imbalance dataset regarding smart meters.
2. Combining machine learning techniques to create a hybrid model. As stated above, only serval methods worked. If given more time to learn, better models could have been used to get for an accuracy greater than 80%.
3. Technically due to more than 90% of the variables being object type the transformation of dummy variables should have occurred in phase 1 and then proceed to phase 2. Due to time constraints, the dummy transformation occurred in phase 2 after VIF. As a result, there could have been some important variables lost. Additionally, outlines could have been easily spotted. Another area of loss of important variables is during the VIF phase, with some variables combined method could have been used to reduce multicollinearity

and provide important insights. In the future, redesigning of the data cleaning process is needed.

Overall, if the recommendations above were taken it would lead to a potential increase in accuracy and a more widely accepted business algorithm to use for production.

## References:

*Climate Change Indicators: Residential Energy Use*. (2021, July 21). US EPA. Retrieved February 24, 2022, from https://www.epa.gov/climate-indicators/climate-change-indicators-residential-energy-use

Ingrams, S. (2021, July 1). *Smart Meter Roll-Out*. Which? Retrieved February 24, 2022, from https://www.which.co.uk/reviews/smart-meters/article/smart-meters-explained/smart-meter-roll-out-aQJZI0B2Jose

*Residential energy use - Energy Education* (2021, September 21). Energy Education. Retrieved February 24, 2022, from https://energyeducation.ca/encyclopedia/Residential_energy_use

*Use of energy in homes - U.S. Energy Information Administration (EIA)*. (n.d.). U.S Energy Information Administration. Retrieved February 24, 2022, from https://www.eia.gov/energyexplained/use-of-energy/homes.php

New:

https://www.nytimes.com/2022/03/18/climate/global-energy-crisis-conserve.html