Final Report

Gloria Li

12/18/2020

#Introduction

The energy grid as we know it is changing. As the costs of renewable electricity generation continue to fall, and more of these technologies are integrated onto the grid, the age-old challenges of grid reliability and resilience will gain new dimensions. My goals for this project are to examine the extent to which trends on the energy grid are correlated with, and predictive of, grid reliability as measured through interruptions to customer service.

In the first section of this report, I provide background and a problem statement to contextualize the goals of this project. In the next two sections, I describe the steps taken to wrangle the relevant data, as well as the visualization and machine learning methodologies employed to learn about the variables. In the last two sections, I explain the insights gained from the analysis and the steps that could be taken to expand the analysis in the future.

#Background and Problem Statement As anthropologist Gretchen Bakke wrote, "The grid... is the largest machine in the world." For centuries we have seen the one-way flow of electrons, from generation to transmission to distribution, and the top-down concentration of power – both literally and figuratively – in the hands of the nation's electric utilities. However, in the last decade or so, the forces changing this dominant energy paradigm have begun to accelerate.

A primary driver of change is the rise of distributed energy resources, or DERs, such as rooftop photovoltaics, wind turbines, electric vehicles, and battery storage. These are electric generation assets deployed across the grid, supporting both behind-the-meter individual usage and aggregate usage. Alongside DERs, an interesting new entity has emerged: the "prosumer" – a traditional consumer who now has the capacity to produce electricity (usually through rooftop photovoltaics) and sell their excess back to the grid. This exchange, called net metering, has generated considerable policy debate in recent years as rooftop solar capacity in the U.S. has more than tripled between 2015-2020.

Some potential benefits of DERs include reduced greenhouse gas emissions, optimized distribution operations, grid flexibility (i.e., demand response capabilities), and greater customer choice. Some challenges include grid planning for bi-directional power flow, variable or intermittent power generation, and cross-subsidization (often cited in the context of net metering as an energy equity issue). Additionally, utilities are oftentimes concerned about the lost revenue potential that accompanies a growth in DERs. DERs reduce a utility's need to invest in more infrastructural and generation assets, which is how utilities generally earn money, and this cycle creates what many have called the 'utility death spiral'.

It is clear that the future of the energy grid involves DERs, but it is also clear that we do not yet understand the implications of this widescale integration. The North American Electric Reliability Corporation (NERC), which maintains the reliability and adequacy of bulk power generation in four interconnections and six regional entities across North America, has warned that higher DER penetration on the grid may present new challenges to grid reliability. These challenges may be due in part to the necessity for distribution system upgrades, wherein the cost-shifting problem arising from the "utility death spiral" becomes a complicating factor. The high-level research question I explore in this project is: what is the relationship between net-metered distributed generation and grid reliability? In order to contextualize this question, I first looked into the literature that is already publicly available. Particularly instrumental in expanding my understanding of grid reliability was the Public Utility Research Center's (PURC) report on valuing municipal utilities. In the section on utility benchmarking, the report employs an ordinary least squares (OLS) model to evaluate the correlation between variables such as customer density and ownership structure, and a utility's grid reliability scores. The ownership structure of utilities can range from private investor-owned utilities (IOUs) to public

municipality-owned utilities to electric cooperatives, or co-ops, that are common in more rural areas. Average reliability also varies significantly between these different types of utilities, a key insight from the report that later informed my choice of control variables.

Data

The data I used for this project was sourced from the U.S. Energy Information Administration's Form 861, also known as the Annual Electric Power Industry Report. It comprises data collected from approximately 3,300 U.S. respondents, including electric utilities, wholesale power marketers and electric power producers; these respondents, which I refer to generally as utilities, are the units of analysis in my project. The data is divided into annual spreadsheets covering various portions of the survey, such as net metering, utility sales data, and energy efficiency programs. Because 2013 was the first year that my grid reliability metrics of choice were introduced, 2013-2019 formed the theoretical time range I could choose from to conduct my analysis.

When it comes to grid reliability, there are multiple ways in which it is measured. Two of the most common scoring systems include the System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI), which were both introduced in the 2013 EIA-861 survey. I ended up choosing to use SAIDI as my dependent variable because it had the most entries in the data. SAIDI is calculated using the following equation:

$$\frac{\sum Minutes\ of\ Interruption}{\sum Number\ of\ Customers}$$

EIA-861 includes SAIDI reported without Major Event Days (MED), which are interruptions to the electric power system that exceed a normal range, such as hurricanes or wildfire-induced outages. As was done in the PURC report, I used SAIDI without MED to remove the effects that could not have been correlated with my independent variables of interest.

My independent variables included both predictors and controls. While net-metered distributed generation is my predictor of interest, I also included other predictors that are correlated with grid reliability and distributed generation, such as energy efficiency and demand response programs. This served to reduce endogeneity and increase the predictive accuracy of my model. These predictor variables were reported in megawatthours (MWh) sold back to the utility in the case of generation and either MWh saved or number of enrolled customers for energy efficiency and demand response.

Beyond these variables, there remained a significant amount of variation between municipal utilities, IOUs, and electric co-ops leading to SAIDI scores ranging from less than a minute to over 1,200 minutes. Thus, I included control variables from the data such as the state, NERC region, ownership structure, utility sales (MWh), utility customers, number of distribution circuits and voltage optimization counts in those circuits.

```
Data wrangling was an iterative process since the later stages of my modelling sometimes produced findi

I cleaned the data and dealt with missing values by either imputing them with zeroes or dropping them a
```

#Analysis

My analysis consisted of three main methods: feature engineering, visualization, and modelling. After the initial data wrangling, I performed feature engineering to alter certain variables to be more appropriate for my analysis. This included dividing my predictor variables by the total megawatthours sold by a utility or the total customer count in order to make them proportional to the size of a utility. I also divided the number of voltage optimized circuits by total circuits to get the percentage.

Beyond the aforementioned missingness challenge, another problem with my data was that "some utilities reclassified net metering as non-net metering distributed, or vice versa." EIA-861 also reports non-net metering distributed generation; I only used the net metering data in order to avoid double counting, but this could have caused under-counting, depending on how individual utilities changed their reporting practices over time.

At the end of the wrangling, I had 753 total entries for the years ranging from 2013 to 2019 when I dropped NAs and 3,509 entries for the same period when I imputed with zeroes. This difference demonstrates

the tradeoff between a skewed distribution with many zeroes and having more data to improve the predictive power of the model.

My second method included visualization, which was employed throughout the project to both glean insights into the processed data through exploratory data analysis and to communicate my modelling results in a convenient way. For my initial exploration, I created the following plots to better understand the composition of the data.

Include plots here

I also used the Pandas Profiling package to generate a report that summarized key characteristics of my data frame. This report revealed a large number of zeroes (47.9%) in my net metering variable as well as its composite variables, photovoltaic generation percentage (48.4%) and wind generation percentage (73.6%). This was my first hint that there would be very right-skewed distributions even after I dropped rows with missing values instead of imputing, which would have increased the proportion of zeroes further. Therefore, I decided to just use the total net metering variable, which encompassed photovoltaic, wind, and all other types of distributed generation sold back to the grid.

Additionally, when I looked at the distribution of my dependent variable, I saw there were some outliers – keeping this would diminish my model's accuracy so I removed the entries with a SAIDI above 500.

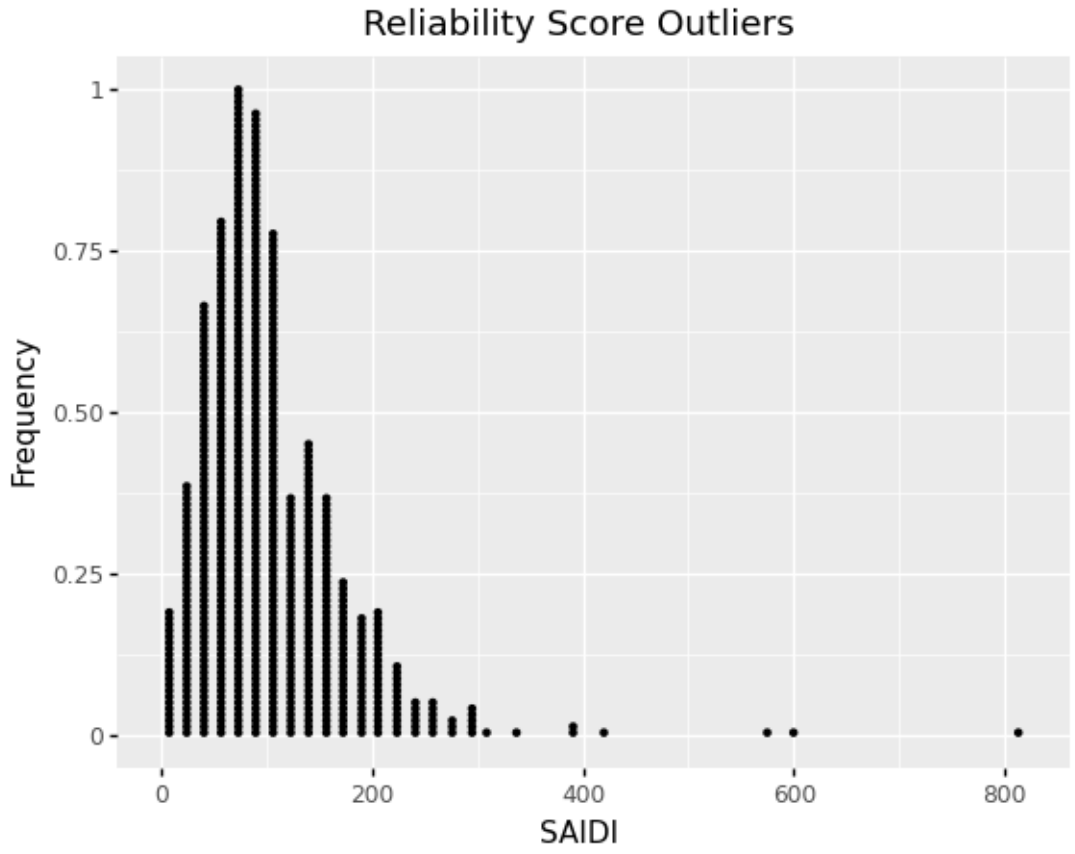

Figure 1: This shows

`I then investigated the correlation between my predictor variables using the Seaborn package. Through t`

Lastly, I modelled my data using various machine learning techniques applicable to regression problems; these included OLS regression, K-Nearest Neighbors (KNN), and Decision Trees variants including Bagged

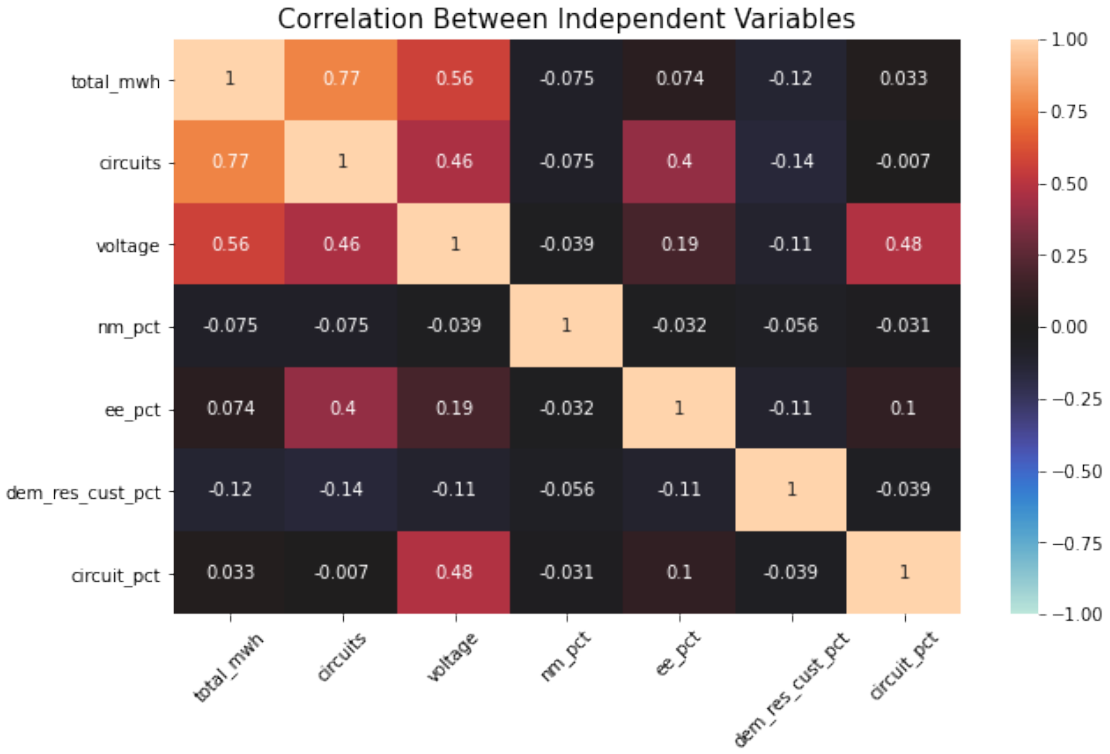

Figure 2: This is an example of how log transformation can reduce the skewedness in a variable's distribution. It shows the variable that represents the number of circuits for a given utility.

regressors and Random Forests. Prior to running my data through the machine learning pipeline, I performed high-level preprocessing on the entire dataset to prepare the data for the pipeline. This pre-processing consisted of dealing with categorical data types and highly skewed distributions through variable transformation.

`The first transformation was straightforward: I turned the categorical control variables such as NERC r`

`Dealing with the distributions of my predictor variables, however, was more complicated; the missingnes`

The second half of modeling involved splitting my data into a test and training set, tuning the hyper-parameters, and then running the training data through my pipeline. I used GridSearchCV to find the estimators that led to the highest predictive accuracy, as measured by Mean Squared Error (MSE) and the R-squared (R2) score. For the KNN, this consisted of finding the ideal number of neighboring data points to evaluate, and for the decision tree models, this consisted of identifying a maximum branch depth and number of predictors to use as inputs.

`Once I had specified my estimators in each model, I generated the pipeline, which first normalized my v`

#Results

I ran several models with different combinations of the data years used (just 2019 or 2013-2019), methods of dealing with missingness (imputing with zeroes or dropping the rows), and variables (including/excluding controls, transformations to reduce skewedness). My initial experiments yielded very low R2 scores, sometimes even in the negatives, which told me that the configurations I used were not very predictive when employed in the pipeline.

My most successful model was a Random Forest regressor with 1250 estimators, a maximum of 40 variables used, and a maximum depth of 40 consecutive choices in each decision tree. When I fitted this model to my test data, it yielded the highest R2 score of 0.61 and a MSE of 1569.63. The MSE is difficult to interpret

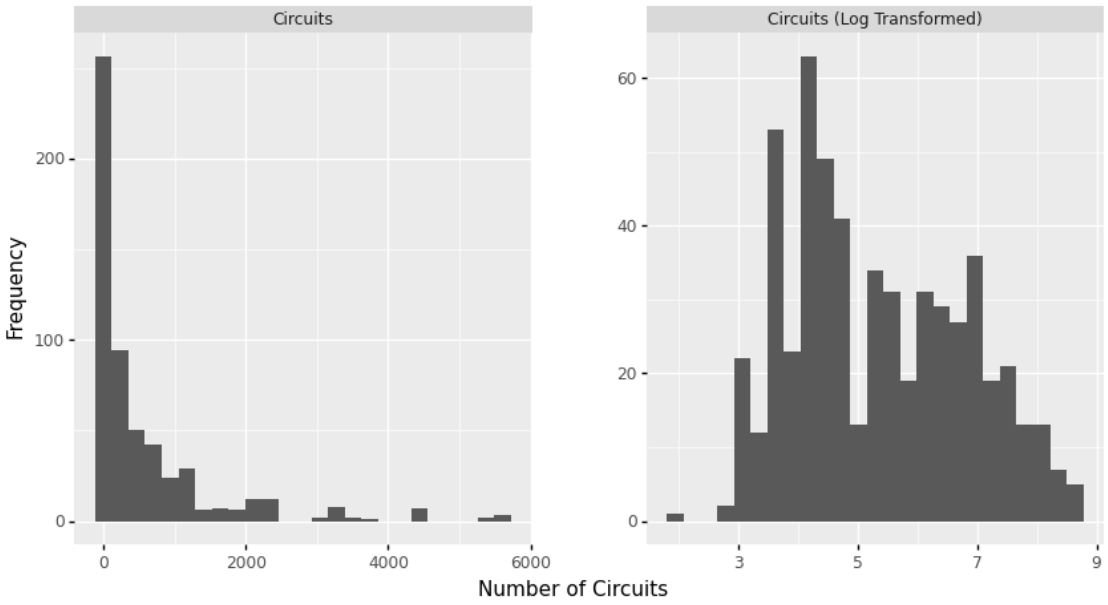

Figure 3: This shows

without context, so I derived the Root MSE (RMSE) to get a number in the units of my dependent variable. The RMSE was 39.62 minutes, meaning that the model's prediction of SAIDI had an average error of ~40 minutes. This may seem high, but it is below the standard deviation of SAIDI in the test dataset, which was 63.66 minutes. Because the highest R2 score is 1, which would entail perfect prediction, the R2 score indicates a moderate fit of the model predictions to the actual data.

Ultimately, this Random Forest model would not be particularly useful for predicting grid reliability using data on net-metered generation, energy efficiency, and demand response. Its prediction will likely fall within one standard deviation of reality, but depending on what your application is, that may still leave too much room for error (~40 minutes of electrical outage). The more interesting insights from this project can be found through the exercise of permuting my variables to see which ones are the most predictive within the model.

In this chart, we see that the most important variable we used was the number of circuits, followed closely by the total MWh sold by a utility; earlier, we saw that these two variables have the highest collinearity, but I kept them both because removing one caused a substantial decrease in predictive power and this chart reinforces why that was the case. My independent variable of interest, the percentage of a utility's generation that was net-metered, is 10th on the list, which seems high compared to the 77 variables that were available but also trails some control variables such as utility ownership (Municipal), state (NC, MI), and NERC regions (SERC, SPP). These controls are high in the list because there is a significant enough SAIDI difference between the utilities that fit their definition and utilities that do not.

#Discussion

My original definition of success in my project proposal was to run an OLS analysis on the correlation between net-metered generation and grid reliability, but this changed because 1) the project was supposed to be oriented more towards prediction than inference and 2) the fact that my other models outperformed the linear regressor shows that the relationships in my data may not be very linear. The project was successful in demonstrating the relative strength of the relationship between net-metered generation, energy efficiency, and demand response, and SAIDI, but the final model was only mildly successful in predicting grid reliability.

`There are several unanswered questions I would explore given more time. One of these is the difference b`

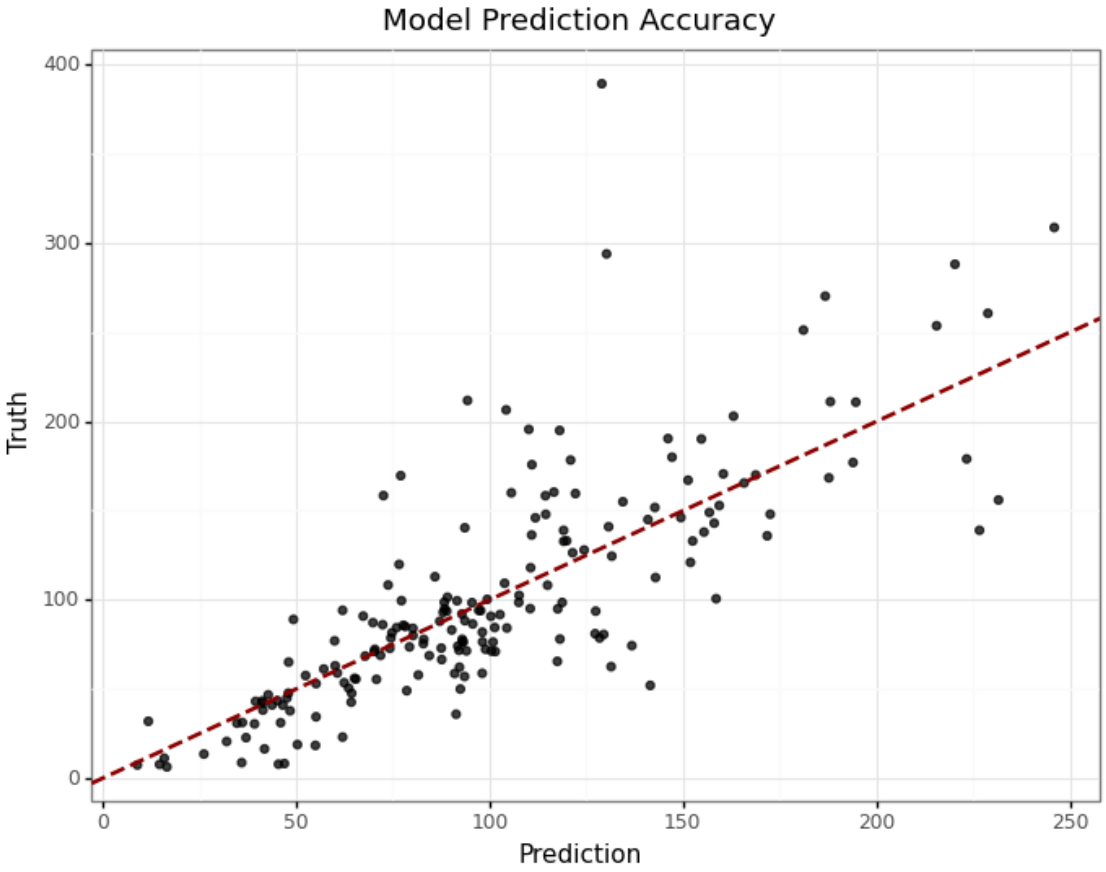

Figure 4: This shows

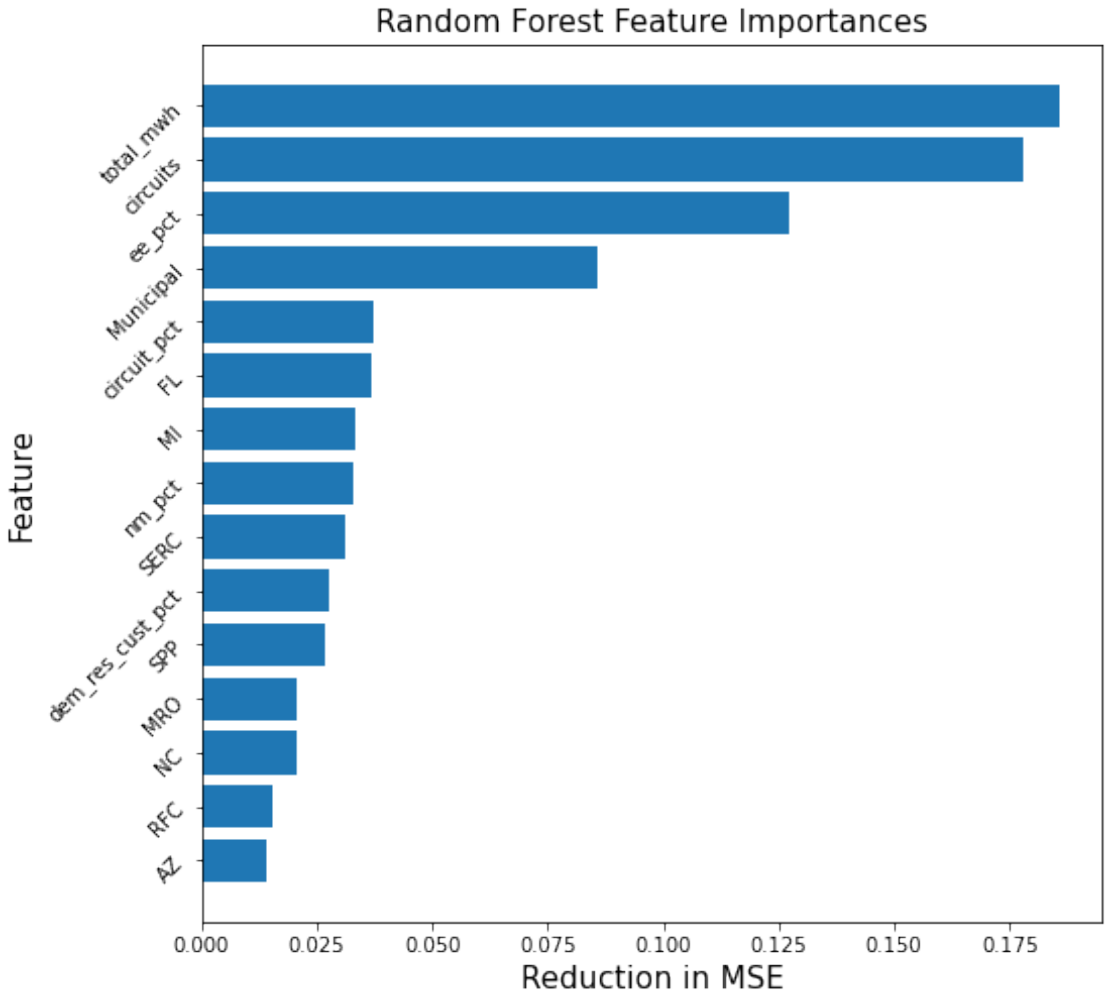

Figure 5: This shows