Assignment: Predicting household energy consumption

Task 1 – Conduct Data Cleaning

After initially loading up the data as a Table, the first step was to clear out any variables that I believed, using domain-specific knowledge, didn't have a significant effect on the Energy requested from the grid. Below is Table 1, where I briefly explain why I deleted each variable.

Variables	Reasoning for removal
Radon Level	Radon is simply a naturally occurring gas that is present in most places, has no impact on energy consumption.
Wind Bearing and Wind speed	Whilst these factors would change the amount of Energy that would be able to be produced by the wind farms, these factors have a relatively small effect on the usage of energy within the household.
Dew point and Pressure	Whilst not completely insignificant, these are two factors that would be affected by other factors (in this case humidity and temperature) so they themselves aren't necessary to predict the energy requested from the grid, and removing these variables helps to remove any possible collinearity.

Table 1: Reasons for removal of variables

I did decide however, to keep many of the weather variables, as extreme weather would mean less people leaving their house as well as more heating/cooling required for the house, so they are important to predict the energy requested from the grid.

Next was changing the 'Weather-Icon' data into numerical data. Integer encoding was used for this process therefore I had to decide on an order to rank the different weathers. I ranked the weather on my theory that the probability that the home inhabitants would leave the house (meaning less energy used) and how much heating they would have to use to keep the house warm. The order I decided on is shown below in Figure 1.

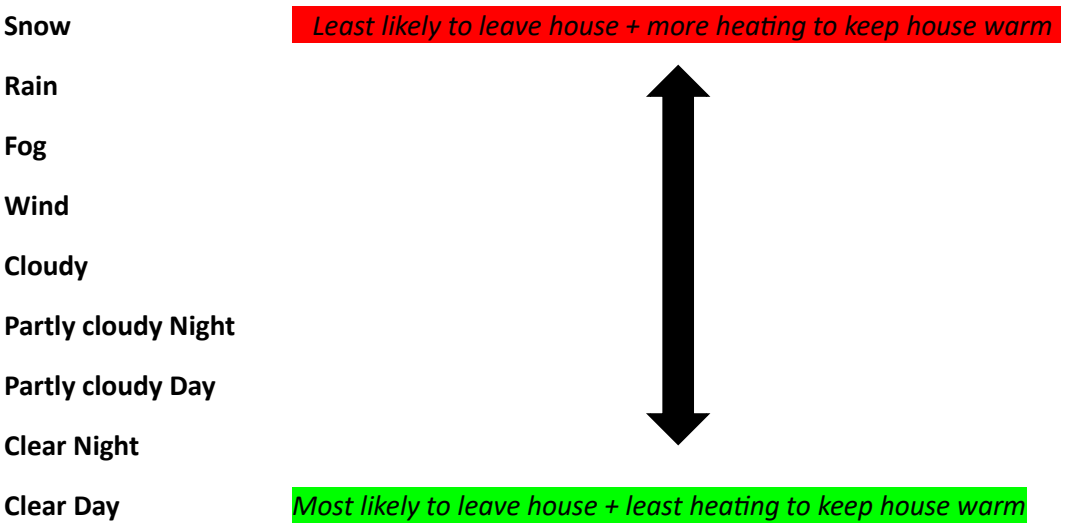

Figure 1: Integer Encoding Order

I then noticed that the usage of appliances in the house will all be directly related to the energy requested from the grid and would most likely be the most important variable, so therefore they have all been left in and have also been totaled. However, I assumed that the 'Solar_kW_' variable was a house's power it receives from its own solar energy, so didn't total this with the other powers,

and will be assuming it will be inversely proportional as if it can provide some of its own power, it won't need as much from the grid.

Next was to remove any missing data. This was done simply by using the *rmmissing* [1] function which removes all rows with any NaN values. However, some values presented themselves as 'inf' therefore I had to replace any 'inf' with 'NaN' (as shown in lines 38-39) then delete the rows using the *rmmissing*[1] function. I decided to do this step before centering and scaling as it made the normalization process more accurate without having infinite or NaN values to alter the data.

Next on the pipeline, I had to center and scale the data. This was simply done by using the inbuilt MATLAB function *normalize*[2] which returns the data with center 0 and standard deviation of 1. Initially I did this bit of code using *for* loops that calculated the means and standard deviations of the columns and then transformed the data, however running this code took a long time so decided to use the *normalize*[2] function to reduce computation time.

Similarly, when it came to removing outliers, I used the inbuilt *rmoutliers*[3] function to do this. I used the moving means detection with a window size of 90 as I believed it provided a smooth representation of the underlying trend of the data, whilst still removing the significant outliers. After removing the outliers, I then removed the individual appliance power draws, as I wanted to get rid of the outliers from those rows before I removed them.

Finally, I ran my table through a bit of code which retrieved the F-stats and P values of each Variable against the Energy requested from grid and removed any variables that didn't satisfy the required values for F-stat and P values given in the week 3 lecture, being F stat needs to be larger than 1 and P-value has to be less than 0.05. This removed temperature, visibility, apparent temperature, precipitation intensity and precipitation probability, suggesting these variables had no significant correlation with the Energy Requested from the grid.

Task 2 and Task 3

In order to create a Linear regressed model as well as higher order polynomial Models, I used the curve fitter app to help generate a piece of code that I could use to plot models for each of my remaining variables after cleaning. I plotted each variable against the Energy Requested from the grid output, using polynomials ranging from power 1 to power 9. When we observe the results of the corresponding R squared values and RMSE values, we can see that as we increase the polynomial power, the R squared Value increases closer to 1(the desired value to show an accurate model [4]). However, despite this, when looking at all the variables, only TotalkW has a close enough value of R squared to 1 (shown in figure 2), suggesting it is the only real good predictor when used on its own. I have included one more example (Figure 3) to show the other variables having far worse accuracy, but please do look at the generated figures to see all variables.

It also has the lowest Value of RMSE suggesting the line of best fit has very minimal error. When comparing the Linear Model against the polynomial models, it is clear that increasing the polynomials makes the model more accurate. However, after we go above about the 3rd or 4th polynomial, the model's accuracy increase very minimally. This information could be useful to help prevent overfitting as the more polynomials we use, the greater the chance of overfitting is, therefore in order to reduce overfitting it would be best to choose the most minimal polynomial that still has high accuracy.

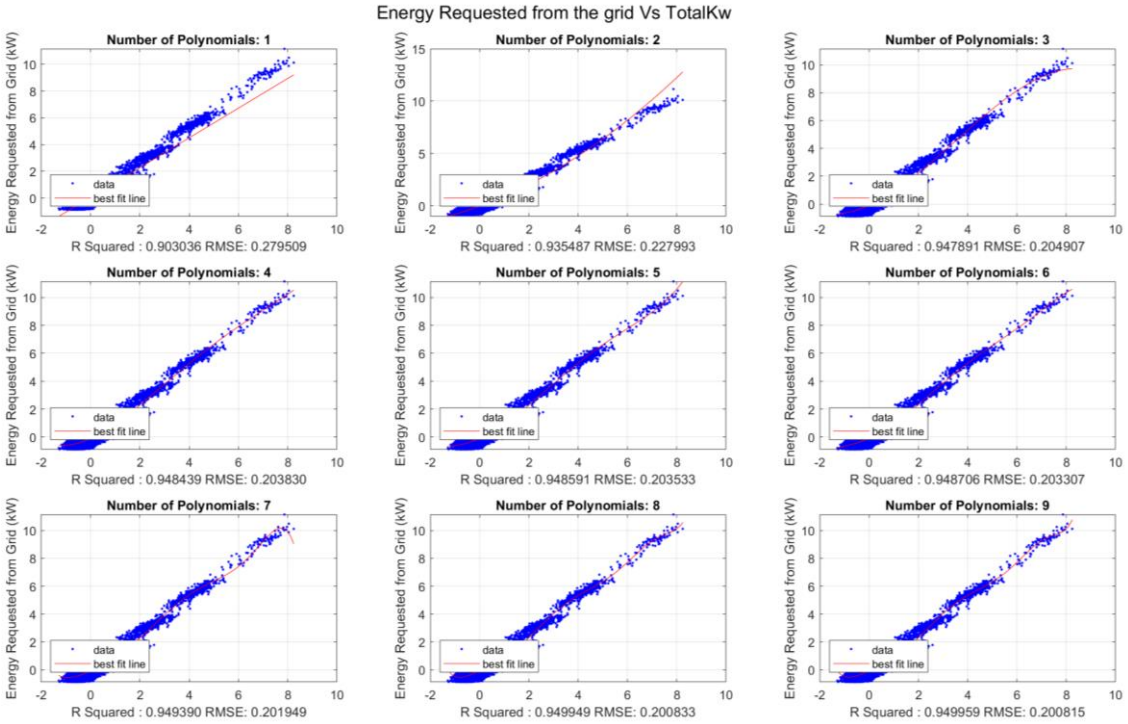

Figure 2: Output v TotalkW

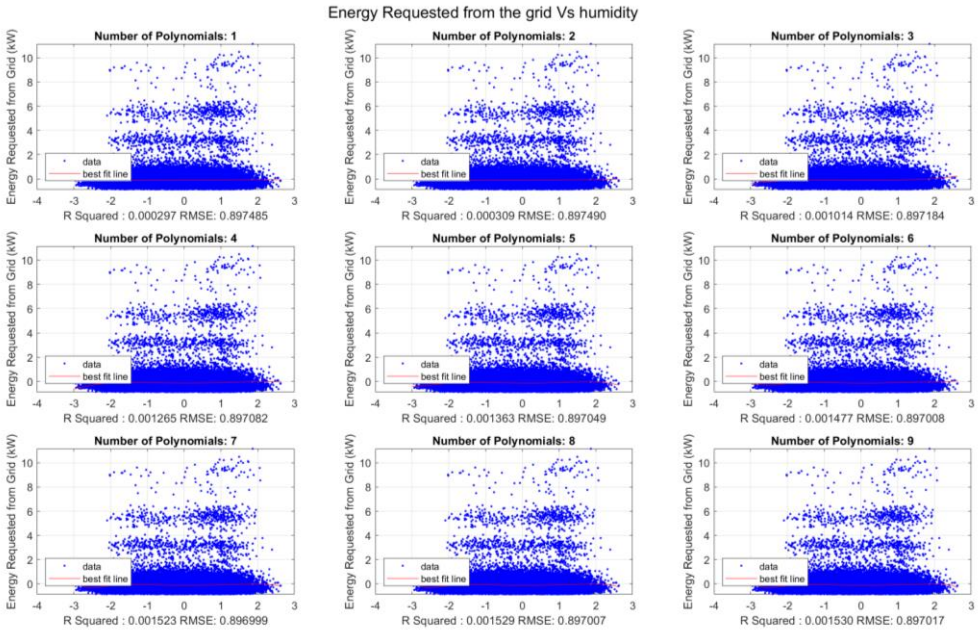

Figure 3: Output v Humidity

In order to reduce the dimensions of the data, and fit it all onto one graph, I used PCA. PCA is a useful tool used in machine learning to reduce dimensionality. It does this by creating a smaller set of predictors/features each of which is a linear combination of the original predictors/features. I used code from Lab 5 to achieve this. Looking at figure 4 we can see the relative contribution of each eigenvector in capturing the variance present in the data.

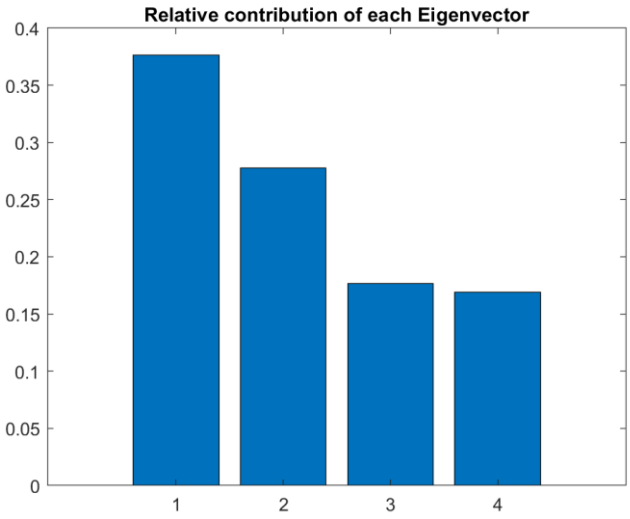

Figure 4: Relative contribution of each eigenvector in capturing the variance present in the data.

Therefore after applying PCA we get the results shown below in figure 5.

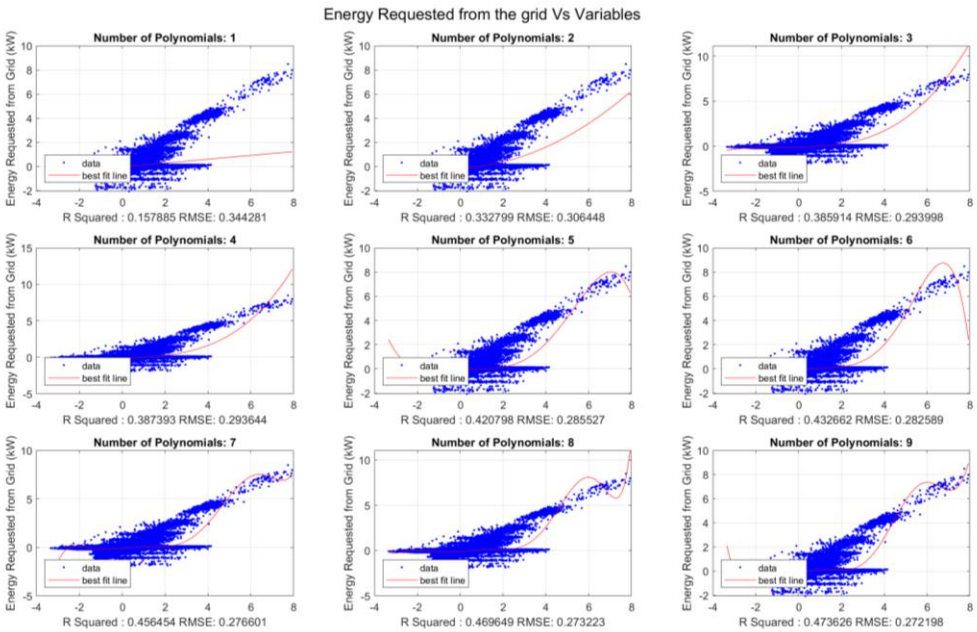

Figure 5: Output vs All variables using PCA

This final Graph reinforces the ideas about higher polynomial being more accurate than linear regression, shown by the increase in R squared and decrease in RMSE. Despite this, the R squared value isn't as optimal as we would like but the RMSE of 0.27 is acceptable in this scope. Overall, we can see there is a significant relation between the final variables and the energy requested from the

grid. This information will be able to be used to help predict when alternative energy production facilities need to be ramped up to meet household energy demands.

In summary, our motivation for this project was to use machine learning to enhance accuracy and efficiency in energy management using past data to create a model. Using data cleaning, I assessed various variables for their significance in predicting energy usage, with domain-specific knowledge and accuracy measures helping me to reduce the data set by removing some insignificant factors such as radon levels. By creating my predictive models, I would be able to assist the national grid in optimizing energy production and distribution, helping the integration of wind power energy into the grid. However, as useful as machine learning can be it still has its current drawbacks. Firstly, we are training the models on potentially personal data based on people's energy uses. This is a bit of an ethical gray area as some may not be happy with their information being used. Additionally, AI technologies can't always take in extreme events that could significantly change the data. An example given the context of our project could be in an exceptionally sunny summer, a house may generate more energy from solar panels, suggesting it would request less energy from the grid. In the future, it will be important to refine predictive accuracy, enhance model interpretability and adapt to dynamic energy market conditions, additionally, more complex problems may require more data and data analysis in order to help create an acceptable and accurate model.

References:

[1] MathWorks, "rmmissing" MathWorks. https://uk.mathworks.com/help/matlab/ref/rmmissing.html (Accessed: April 29, 2024).

[2] MathWorks, "normalize" MathWorks. https://uk.mathworks.com/help/matlab/ref/double.normalize.html?s_tid=doc_ta (Accessed: April 29, 2024).

[3] MathWorks, "rmoutliers" MathWorks. https://uk.mathworks.com/help/matlab/ref/rmoutliers.html?s_tid=doc_ta (Accessed: April 29, 2024).

[4] NCL," Coefficient of Determination, R-squared" NCL. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html (Accessed: April 29, 2024).