Energy Usage Write-Up

After getting the house and energy data, we looked at the variables and found which ones had either zero variance or near zero variance. These variables were then removed to avoid overfitting in the models. Using the variable, in.usage_level, we separated the data frame into 3 different data frames, low, medium, and high, filtered by the energy usage level mean. The 3 data frames had a total of about 8.5 million observations. After we separated into the 3 data frames, we randomly selected 20% of observations from each data frame. So each level was properly represented in the dataset we used with our models. Our final dataset had roughly 850,000 observations. We did random sampling because the original dataset given had about 4 million observations which means our models would've taken a lot of time to run. Since we used random selection, it is essentially the same data because our sample was large enough. So the sample we took is considered an unbiased representation of the entire population of data.

We were tasked with predicting future energy demand which requires specific variables that best represent this. Some variables had only one unique value which doesn't help our project in predicting energy usage if all the observations are the same, so we omitted those. Since some of the variables were very correlated, we made sure to omit the ones that were too similar. We decided that if the variables' correlation coefficient was greater than 0.8, they were omitted. We chose this because 0.8 is a general rule of thumb in data science regarding correlation. We also tested the variables for Near-Zero Variance to see if there was variation within each respective variable. If this value was true, we omitted the variable because it doesn't help us predict if all the data values are the same. Because our goal was predicting future energy usage and to be as accurate as possible, we thought it was best to omit energy output variables because they aren't directly related to energy usage and could alter our results.
We found that variables such as income, house occupants, cooling and heating setpoint best represent energy consumption and would help the model most accurately predict future levels. Our dependent variable was called in.usage_level which is categorical.

Our next task was deciding what models would help us best predict future energy levels. We decided on an ordered logit, an ordered probit, and a random forest.

We decided on an ordered logit model because it handles categorical variables and our dependent variable was categorical. This regression model estimates the relationship between an ordinal dependent variable and a set of independent variables that best describes our dataset. The logit distribution is based on a cumulative standard logistic distribution and assigns probabilities that values will fall below a certain threshold which was the mean in our project. Below is the confusion matrix for the ordered logit model predictions on the test data.

		Actual		
		Low	Medium	High
Predicted	Low	32387	5401	1223
	Medium	9996	78823	3688
	High	0	1312	37100

We then decided to run an ordered probit because it is essentially the same as an ordered logit except with a different distribution. An ordered probit distribution relies on the standard normal

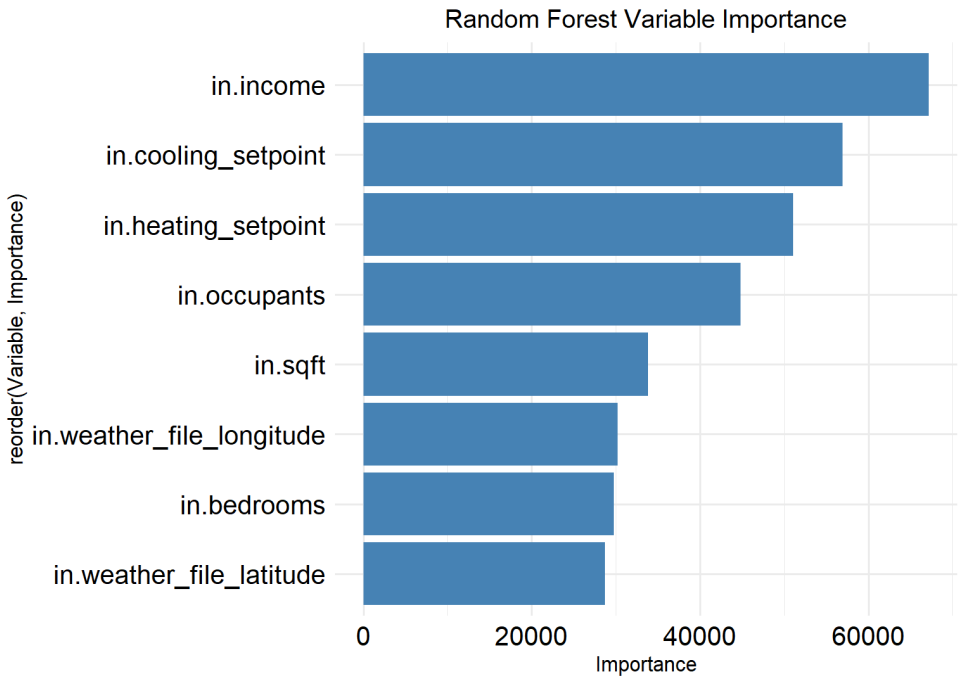
distribution and essentially, predicts the likelihood of an observation falling within a particular ordered category based on its characteristics, assuming that the underlying latent variable follows a normal distribution. Below is the confusion matrix for the ordered probit model predictions on the test data.

		Actual		
		Low	Medium	High
	Low	30933	5274	617
Predicted	Medium	11450	78165	4513
	High	0	2097	36881

Our final model was a random forest which is a supervised machine learning method that combines the output of multiple decision trees to reach a single result. We decided on a random forest because it can handle categorical dependent variables without having to transform them first. Below is a confusion matrix that describes the results of the random forest. Compared to the ordered logit and ordered probit, the random forest was more accurate in predicting energy usage.

		Actual		
		Low	Medium	High
	Low	41805	57	15
Predicted	Medium	108	85609	67
	High	33	67	42169

We wanted to see which variables from the random forest had the most importance within the model. Below is a graph of variable importance from the random forest. The variable with the most importance is income which could mean that companies can target households based on their income to have the biggest impact on energy usage. The plot measured how much a variable contributed to the final prediction which is important if we want to run a more efficient model in the future.



To further confirm that the random forest model produced the most accurate predictions, we have included each of the model accuracies. We were suspicious of the random forest being too accurate and overfitting the data, so we said earlier, we removed all variables with near zero variance and all the energy output variables.

Accuracy	
Ordered Logit	87.28%
Ordered Probit	85.91%
Random Forest	99.80%

Overall, our models predict overall usage will not change, but the time of peak usage changes from 7 AM to 6 PM. This could be because most people get home from school and work around that time so their household energy consumption would naturally increase. We would recommend to customers to adjust setpoints when away from home and should adjust setpoints when away from home to save energy and money. We recommend that higher-income households should be targeted to change consumption as income is seen to be the biggest driver of energy usage from the random forest variable importance plot.