Case Study: Energy Consumption in London Households

STAT 471 Final Project

Kamran Elahi, Awad Irfan, M. Abdullah Khalid

May 2, 2021

Contents

1 Project Background **2**

2 Executive Summary **2**

3 Data Cleaning and Exploratory Data Analysis **3**

4 Model Building **10**

4.1 Linear Regression 10

4.2 Lasso Regression 13

4.3 Decision Trees 14

4.4 Random Forests 16

4.5 Boosting 18

5 Evaluation and Interpretation **20**

6 Conclusions **21**

1 Project Background

Energy consumption is critical for livelihood, and recently a lot of attention has been placed on how energy consumption can be effectively managed. There are more advocates than ever before for a shift toward renewable energy sources, and at the same time some governments cannot acquire enough traditional energy to keep their cities powered. When it comes to making these decisions and solving these problems, it is critical to first understand how energy is used.

Predicting energy consumption is important for many different stakeholders, including energy companies, consumers, alternative energy providers, and even policy makers. Understanding what factors influence energy consumption can better help these individuals make informed decisions when it comes to building, budgeting, marketing, and implementing laws. It is thus important for these stakeholders to better understand what factors influence energy consumption across different factors.

The use of this information will vary for different stakeholders, meaning that the types of factors they are interested in using to predict energy consumption will be different as well. Energy companies, both traditional and alternative, will want to understand which types of consumers use the most energy; traditional providers want to build infrastructure to be able to better serve their customers with high demand, and renewable energy providers want to target the same individuals to both to let them know the benefit of saving costs as well as make the most impact by convincing them to switch. Thus, these players would be interested in demographic factors, such as household income.

Individual consumers are aware of their demographic information, and are likely not interested in making predictions about others. They may, however, be interested in budgeting for electricity bills throughout the year, or may want to make the shift to renewable energy at the best possible time. Residents in hotter cities may also realize that they could have significant savings by switching to alternative sources of energy while their counterparts in more moderate climates may realize that the cost of implementing such alternatives would outweigh the savings. Thus, these individuals would likely care less about demographic factors, and more about climate and temporal data.

Finally, policy makers that have to decide how to allocate energy resources, or which individuals should receive support to implement alternative energy first would likely be interested in all of these predictors. We thus attempt to predict energy consumption using a variety of factors, including both demographic predictors as well as ones related to time, and weather.

2 Executive Summary

This project seeks to predict energy consumption behaviors based on a combination of different demographic, temporal, and weather-based predictors. We analyze 3.5 million observations across different consumers and days. Because the temporal and weather predictors are constant for the same day, and the demographic predictors are the same for the same ACORN class, we group the data by these predictors, looking at the average consumption usage across groups, to get a condensed data set of about 15,000 observations. This condensed data set is used for the final data analysis and model building.

We test a total of seven models, including three classical regression models and four tree based models. Our final model set included a basic regression, a modified regression (in which factors are removed to control for collinearity), a lasso regression, a basic decision tree, a decision tree tuned for splits, a random forest model, and a tree boosting model. We tune some of the more sophisticated models on their underlying attributes. For example, we tune our random forest model and our boosting model for number of trees, random forest for number of attributes per tree, and boosting for the depth of each tree.

We find that advanced tree-based models work the best. Specifically, the random forest and boosting models had test MSEs of 1.90 and 2.19 respectively, compared to the next highest test MSE of 3.13 from the basic linear regression. We are thus confident in the predictive capability of the random forest model for the population on which the data was collected. We found through the tree based models that the most important factors were ACORN class and the maximum temperature of the day.

3 Data Cleaning and Exploratory Data Analysis

The initial dataset (retrieved from https://www.kaggle.com/jeanmidev/smart-meters-in-london) consisted of four different tables, which are as follows:

- daily_dataset.csv: This dataset consists of data related to energy consumption for each day and each household from December 2011 to January 2014.
- informations_households.csv: This dataset links each household to their Acorn group (defined later).
- uk_bank_holidays.csv: This dataset lists the days which are bank holidays in the United Kingdom from December 2011 to January 2014.
- weather_daily_darksky.csv: This dataset details the weather metrics for each day in London from December 2011 to January 2014

Since the goal of the study is to predict the energy consumption based on weather and household metrics, we need to combine these different tables into a single table before we proceed to explore and analyze the dataset. The nulls were dropped, integrated variables were created for holidays and weekends, and all of these tables were finally joined (on day) into a single table as shown below.

```
##        LCLid        day energy_std energy_sum is_holiday temperatureMin
## 1  MAC000131 2011-12-15      0.239       9.51          0           4.08
## 2  MAC000131 2011-12-16      0.281      14.22          0           1.80
## 3  MAC000131 2011-12-17      0.188       9.11          0           0.24
## 4  MAC000131 2011-12-18      0.203      10.51          0          -0.56
## 5  MAC000131 2011-12-19      0.259      15.65          0          -0.84
## 6  MAC000131 2011-12-20      0.288      17.16          0           5.66
## 7  MAC000131 2011-12-21      0.222      11.28          0           5.93
## 8  MAC000131 2011-12-22      0.267      10.62          0           8.08
## 9  MAC000131 2011-12-23      0.249      13.97          0           5.28
## 10 MAC000131 2011-12-24      0.151       7.94          0           3.17
##    temperatureMax humidity cloudCover windSpeed dewPoint pressure uvIndex
## 1            7.97     0.77       0.42      4.71     2.41      997       1
## 2            4.68     0.88       0.70      3.71     1.60      988       1
## 3            5.35     0.86       0.37      3.99     0.96     1008       1
## 4            5.49     0.84       0.22      3.60    -0.31     1016       1
## 5            6.64     0.94       0.47      2.70     2.45     1014       1
## 6            8.26     0.81       0.48      4.25     3.64     1015       1
## 7           12.14     0.94       0.67      2.90     8.60     1018       0
## 8           12.14     0.87       0.38      4.31     8.07     1025       1
## 9           11.44     0.85       0.74      4.85     7.08     1018       0
## 10           8.22     0.80       0.37      4.46     2.79     1028       1
##    is_weekend   Acorn month
## 1           0 ACORN-E    12
## 2           0 ACORN-E    12
## 3           1 ACORN-E    12
## 4           1 ACORN-E    12
## 5           0 ACORN-E    12
## 6           0 ACORN-E    12
## 7           0 ACORN-E    12
## 8           0 ACORN-E    12
## 9           0 ACORN-E    12
## 10          1 ACORN-E    12
```

The following are the variables in our dataset:

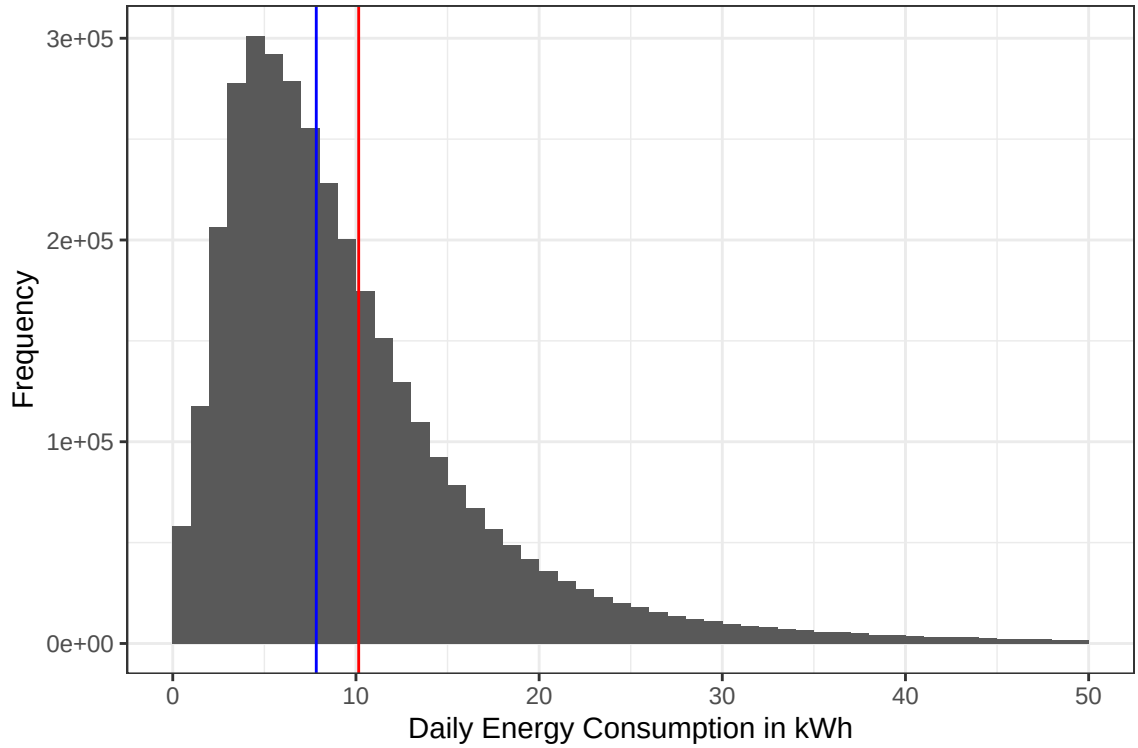
- LCLid: Unique identifier for a household
- day: The day of the measurement
- energy_std: Half Hourly Standard Deviation of the energy measurement
- energy_sum: Total energy consumption for that day in kWh (kiloWatt-hours)
- is_holiday: Binary variable that is 1 if the day of the measurement is a holiday and 0 otherwise
- temperatureMin: Minimum Temperature recorded on the day in Celsius
- temperatureMax: Maximum Temperature recorded on the day in Celsius
- humidity: Percentage level of Humidity recorded on the day
- cloudCover: Percentage of Cloud Cover recorded on the day
- windSpeed: Speed of the wind recorded on the day in metres per second
- dewPoint: Dew point recorded on the day in Celsius
- pressure: Pressure recorded on the day in mbar
- uvIndex: The ultraviolet index recorded on the day
- is_weekend: Binary variable that is 1 if the day of the measurement is a weekend and 0 otherwise
- month: Month of the day of the measurement
- Acorn: Classification Acorn group that the household belongs to. The UK government classifies households into 18 different Acorn groups based on their lifestyle and income. The Acorn group ranges from A to R. The higher the alphabet, the better the standard of living for the household. (https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf)

Each row in the table above represents a household and a day. Since the temperature and acorn data will be the same for each day and acorn, we need to group by day and Acorn when training models on this data. We will use both the original data and the grouped data for our analysis, but only the grouped data for our models. The grouped data is shown below:

```
##           day   Acorn energy_std energy_sum is_holiday temperatureMin
## 1  2011-11-23 ACORN-D     0.2408       9.28          0           3.81
## 2  2011-11-23 ACORN-E     0.3666      12.31          0           3.81
## 3  2011-11-23 ACORN-F     0.0814       4.42          0           3.81
## 4  2011-11-23 ACORN-G     0.0812       5.34          0           3.81
## 5  2011-11-23 ACORN-L     0.1350       5.62          0           3.81
## 6  2011-11-23 ACORN-Q     0.2192       5.80          0           3.81
## 7  2011-11-24 ACORN-D     0.2348      14.48          0           8.56
## 8  2011-11-24 ACORN-E     0.1893      10.23          0           8.56
## 9  2011-11-24 ACORN-F     0.0988       6.48          0           8.56
## 10 2011-11-24 ACORN-G     0.0905       9.44          0           8.56
##    temperatureMax humidity cloudCover windSpeed dewPoint pressure uvIndex
## 1            10.4     0.93       0.36      2.04     6.29     1027       1
## 2            10.4     0.93       0.36      2.04     6.29     1027       1
## 3            10.4     0.93       0.36      2.04     6.29     1027       1
## 4            10.4     0.93       0.36      2.04     6.29     1027       1
## 5            10.4     0.93       0.36      2.04     6.29     1027       1
```

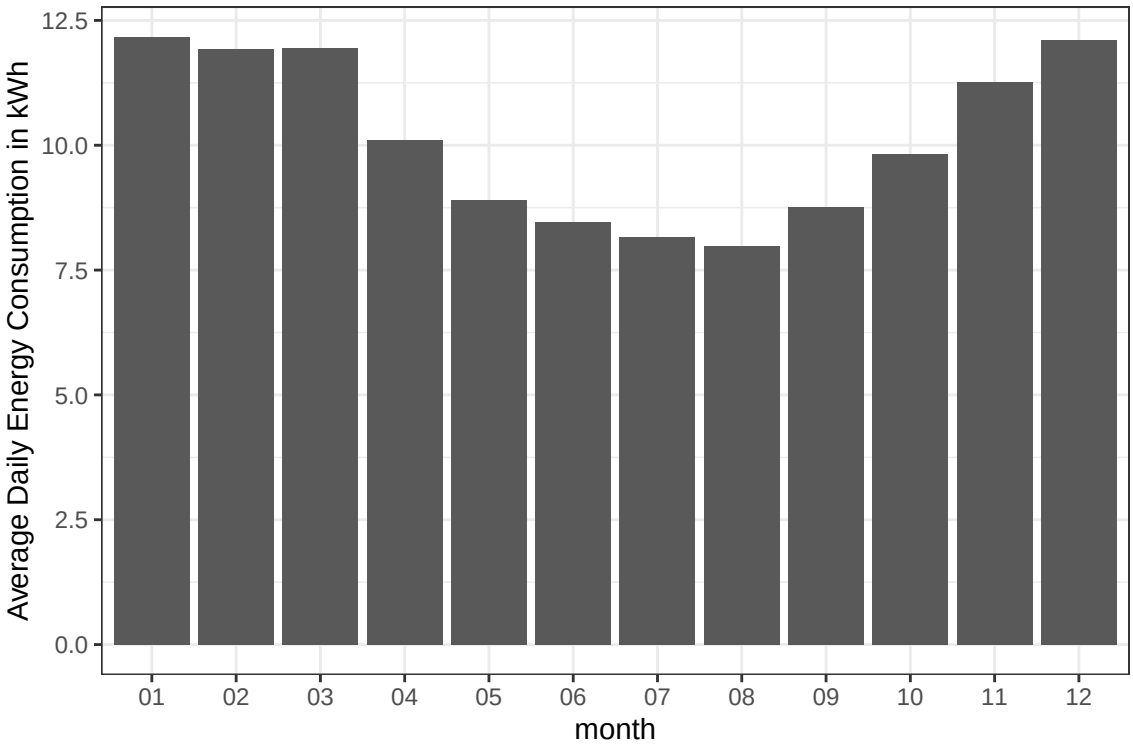
```
## 6            10.4     0.93       0.36      2.04     6.29     1027        1
## 7            12.9     0.89       0.41      4.04     8.56     1027        1
## 8            12.9     0.89       0.41      4.04     8.56     1027        1
## 9            12.9     0.89       0.41      4.04     8.56     1027        1
## 10           12.9     0.89       0.41      4.04     8.56     1027        1
##    is_weekend month
## 1           0    11
## 2           0    11
## 3           0    11
## 4           0    11
## 5           0    11
## 6           0    11
## 7           0    11
## 8           0    11
## 9           0    11
## 10          0    11
```

Let's visualize the distribution of Daily Energy Consumption.

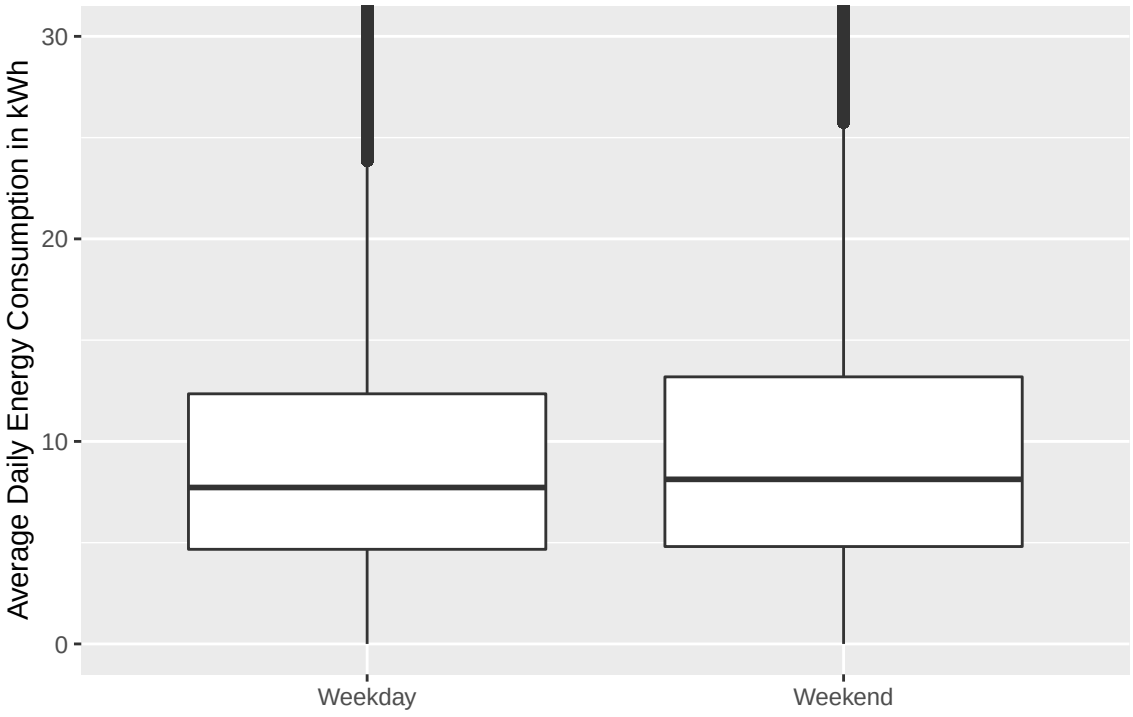


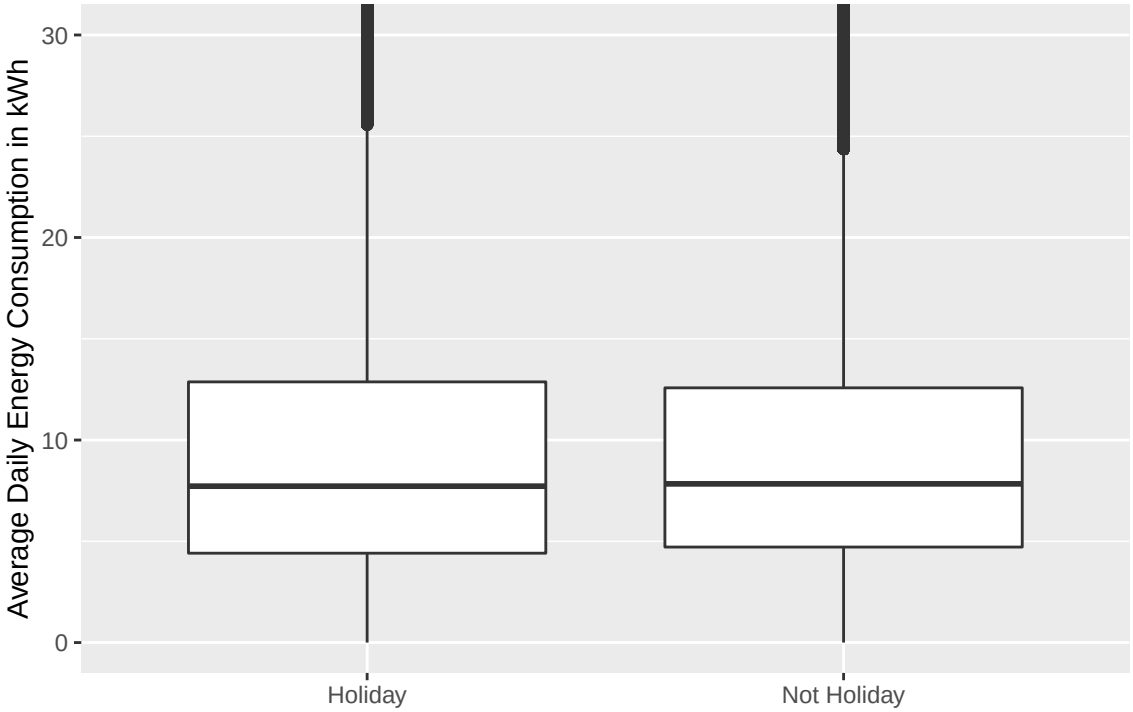
As we can see above, the distribution is skewed to the right. The mean (red line) is around 10 which is greater than the median (blue line), which is around 8.

Let's compare energy consumption with month and specific days (weekends and holidays)



More energy is consumed during the winter months, which makes sense because temperatures are very cold in London in the winter, prompting the need for heaters, but not that hot in the summer to have the need for air conditioning.

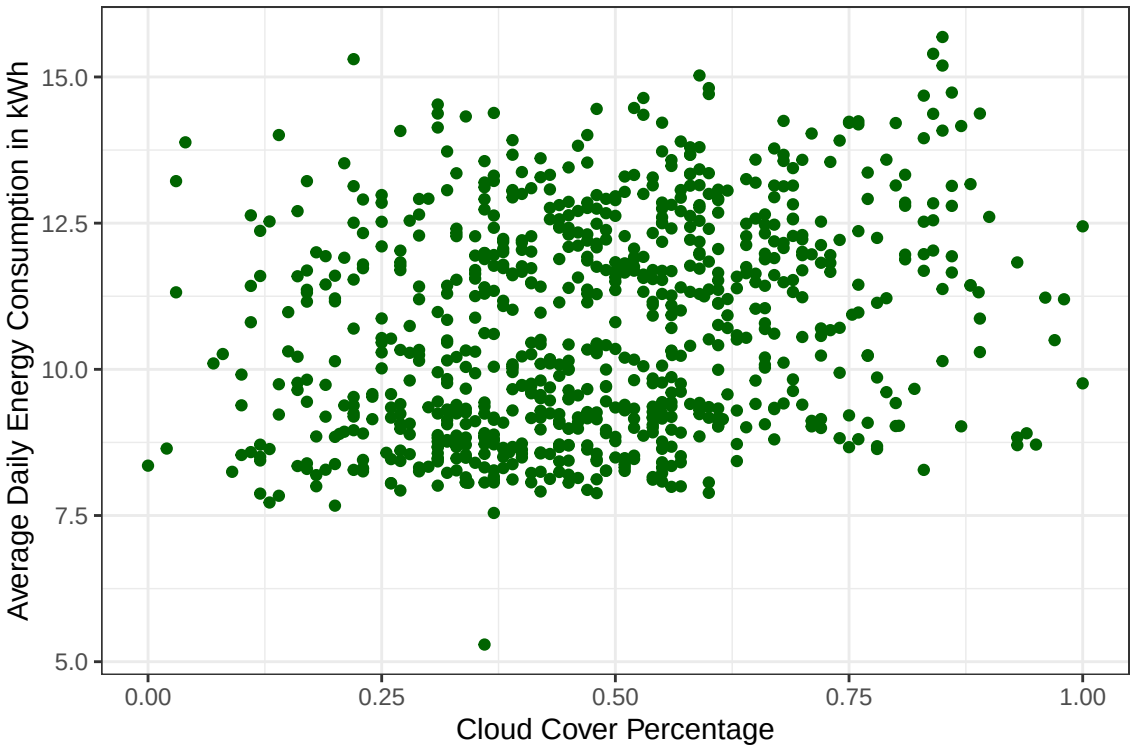




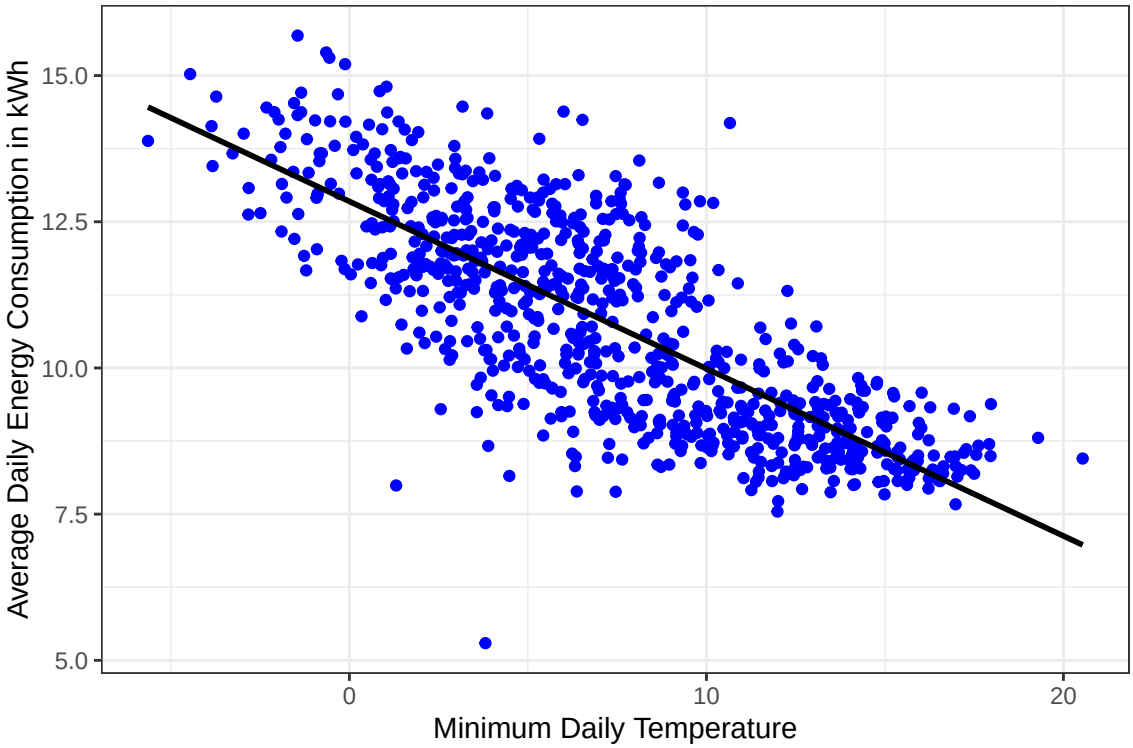
We observe that the energy consumption for weekends and holidays is slightly higher than normal days but there is not much of a noticeable difference.

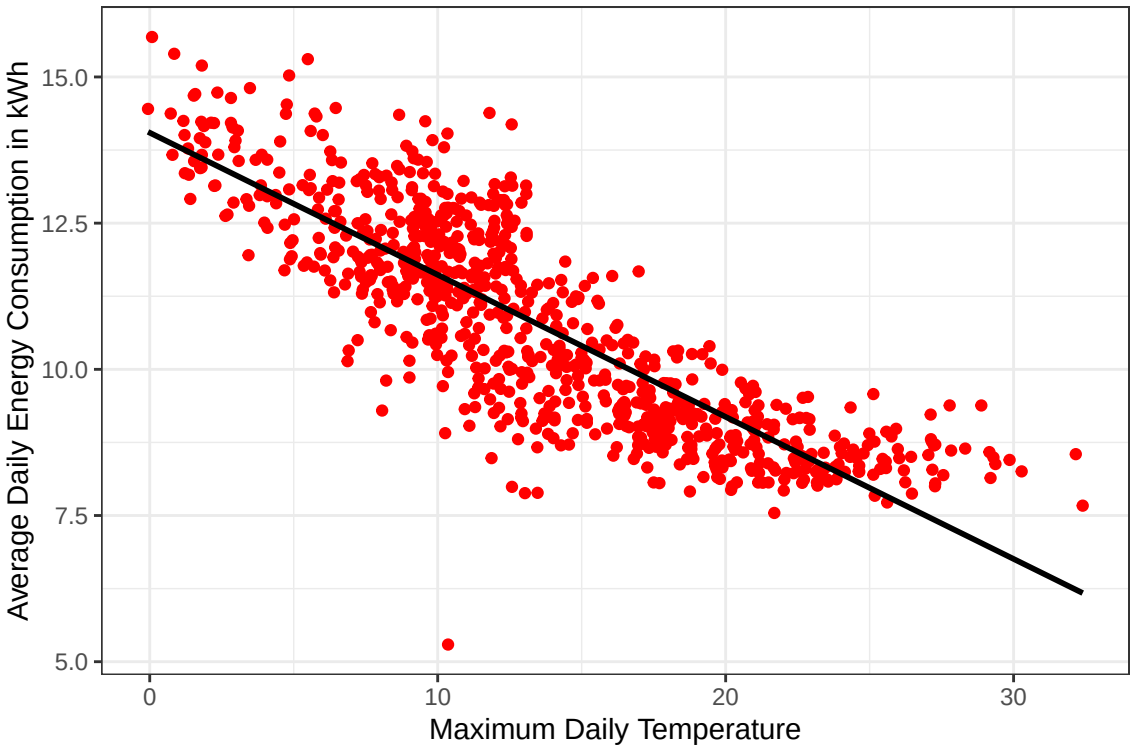
Now, since we will be looking at plots on the grouped dataset (which we will use on our model), we must split the training and test dataset. Our EDA from here will be done on the training dataset.

Let's look at the relationship between cloud cover, temperature and energy consumption.

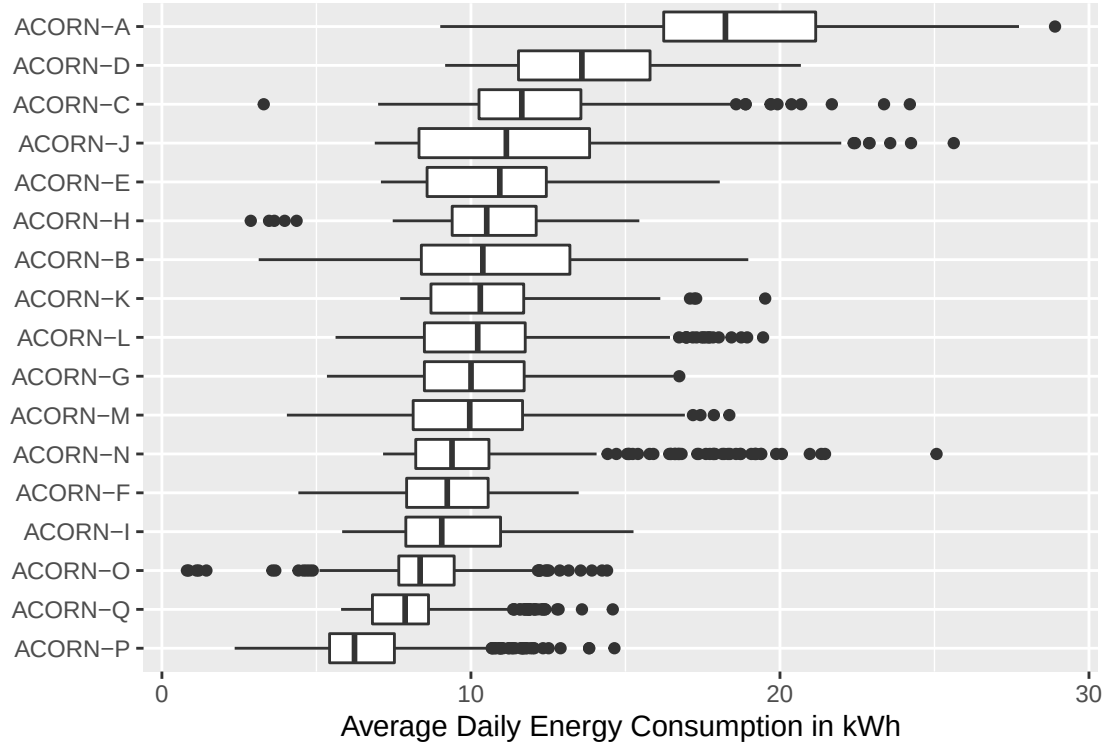


Surprisingly, there does not seem to be a lot of correlation between cloud cover and energy consumption.

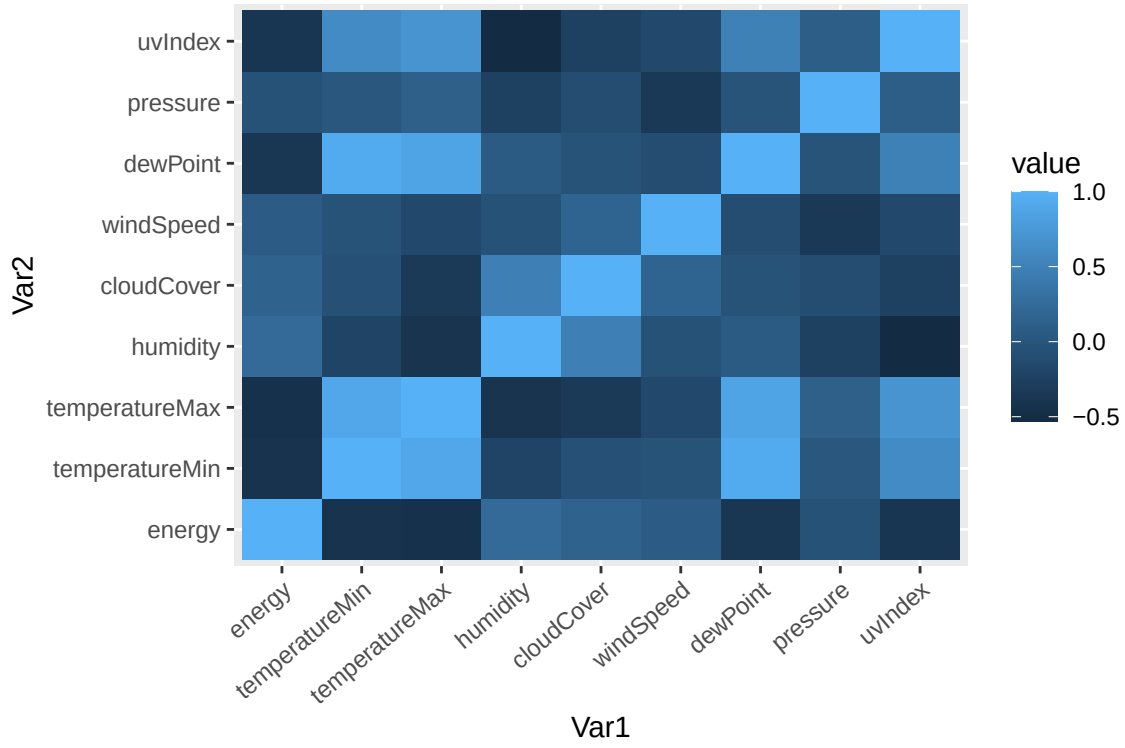




There is a lot of strong negative correlation with temperature and Daily Energy Consumption.



Since the higher alphabet ACORNS are wealthier, it makes sense that these ACORNS consume more energy, which is visible in the boxplot above. ACORN-A (Lavish Lifestyles) consume the most energy while ACORN-P (Struggling Estates) consume the least energy.



There are some obvious correlations between temperatures, dewpoints and uvindex. But apart from that, there are no standout or interesting correlation values. The most correlated variables with energy are temperature and dew point.

4 Model Building

4.1 Linear Regression

Let us first try a basic linear regression model. We will only use this model to make predictions so we do not need to deal with multicollinearity right now.

We get a test mean squared error of **3.14**. Now, to analyze the co-efficients, we will drop and transform our features so that there is little to no multicollinearity. We will first consider the average temeperature instead of the maximum and minimum temperature (which are correlated with each other). After that, we will drop the other predictors which have a Variance Inflation Factor (VIF) greater than 10.

```
##              is_holiday              is_weekend      as.factor(month)02
##                    1.03                    1.03                    1.85
##      as.factor(month)03      as.factor(month)04      as.factor(month)05
##                    3.99                    6.58                    8.56
##      as.factor(month)06      as.factor(month)07      as.factor(month)08
##                    9.08                    9.46                    7.98
##      as.factor(month)09      as.factor(month)10      as.factor(month)11
##                    4.92                    2.09                    1.66
##      as.factor(month)12 as.factor(Acorn)ACORN-A as.factor(Acorn)ACORN-B
##                    2.11                    2.29                    2.34
## as.factor(Acorn)ACORN-C as.factor(Acorn)ACORN-D as.factor(Acorn)ACORN-E
##                    2.31                    2.32                    2.38
```

```
## as.factor(Acorn)ACORN-F as.factor(Acorn)ACORN-G as.factor(Acorn)ACORN-H
##                    2.40                    2.39                    2.40
## as.factor(Acorn)ACORN-I as.factor(Acorn)ACORN-J as.factor(Acorn)ACORN-K
##                    2.25                    2.39                    2.34
## as.factor(Acorn)ACORN-L as.factor(Acorn)ACORN-M as.factor(Acorn)ACORN-N
##                    2.38                    2.36                    2.31
## as.factor(Acorn)ACORN-O as.factor(Acorn)ACORN-P as.factor(Acorn)ACORN-Q
##                    2.41                    2.33                    2.38
## as.factor(Acorn)ACORN-U          temperatureAvg                humidity
##                    2.34                  137.19                   20.87
##              cloudCover               windSpeed               dewPoint
##                    1.82                    1.51                  122.96
##                pressure     as.factor(uvIndex)1     as.factor(uvIndex)2
##                    1.36                   11.91                    9.36
##     as.factor(uvIndex)3     as.factor(uvIndex)4     as.factor(uvIndex)5
##                    9.56                   17.42                   16.61
##     as.factor(uvIndex)6     as.factor(uvIndex)7
##                   16.23                    3.17
```

Since temperatureAvg, dewPoint and uvIndex are heavily correlated and have a VIF greater than 10, we will remove dewPoint and uvIndex, keep temperatureAvg, and then run a linear regression on this dataset.

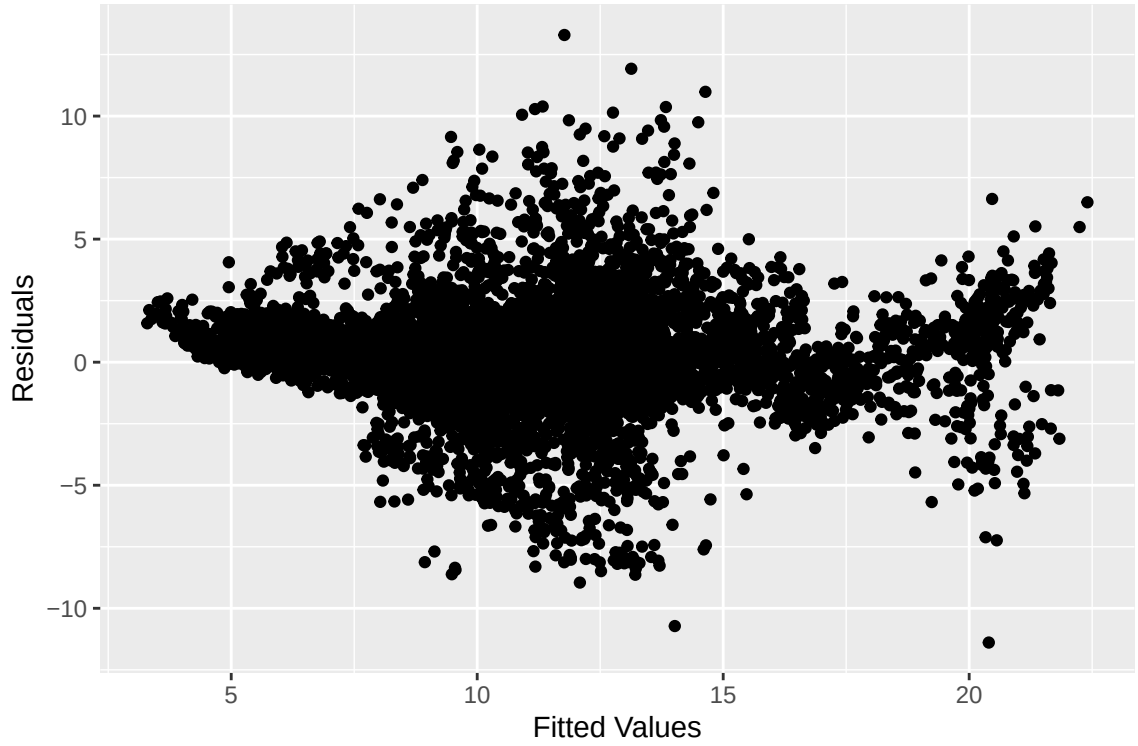
```
##
## Call:
## lm(formula = energy_sum ~ is_holiday + is_weekend + as.factor(month) +
##     as.factor(Acorn) + temperatureAvg + humidity + cloudCover +
##     windSpeed + pressure, data = train_data_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.392  -0.746  -0.076   0.641  13.295
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              13.407011   1.785980    7.51  6.5e-14 ***
## is_holiday                0.270690   0.113225    2.39    0.017 *
## is_weekend                0.512455   0.035976   14.24  < 2e-16 ***
## as.factor(month)02       -0.337303   0.071873   -4.69  2.7e-06 ***
## as.factor(month)03       -0.494928   0.078771   -6.28  3.4e-10 ***
## as.factor(month)04       -1.699534   0.082923  -20.50  < 2e-16 ***
## as.factor(month)05       -2.227173   0.088286  -25.23  < 2e-16 ***
## as.factor(month)06       -2.199982   0.094370  -23.31  < 2e-16 ***
## as.factor(month)07       -1.725821   0.106830  -16.15  < 2e-16 ***
## as.factor(month)08       -1.751297   0.106530  -16.44  < 2e-16 ***
## as.factor(month)09       -1.782487   0.092508  -19.27  < 2e-16 ***
## as.factor(month)10       -1.171516   0.082938  -14.13  < 2e-16 ***
## as.factor(month)11       -0.633374   0.076058   -8.33  < 2e-16 ***
## as.factor(month)12        0.001045   0.069737    0.01    0.988
## as.factor(Acorn)ACORN-A   7.305730   0.110159   66.32  < 2e-16 ***
## as.factor(Acorn)ACORN-B  -0.820655   0.109441   -7.50  6.9e-14 ***
## as.factor(Acorn)ACORN-C   0.761432   0.109831    6.93  4.3e-12 ***
## as.factor(Acorn)ACORN-D   2.320210   0.109633   21.16  < 2e-16 ***
## as.factor(Acorn)ACORN-E  -0.704886   0.108584   -6.49  8.8e-11 ***
## as.factor(Acorn)ACORN-F  -2.138146   0.108388  -19.73  < 2e-16 ***
```

```
## as.factor(Acorn)ACORN-G -1.261351   0.108415  -11.63  < 2e-16 ***
## as.factor(Acorn)ACORN-H -0.664119   0.108389   -6.13  9.2e-10 ***
## as.factor(Acorn)ACORN-I -1.928094   0.110826  -17.40  < 2e-16 ***
## as.factor(Acorn)ACORN-J  0.135721   0.108453    1.25    0.211
## as.factor(Acorn)ACORN-K -1.158753   0.109316  -10.60  < 2e-16 ***
## as.factor(Acorn)ACORN-L -0.958502   0.108597   -8.83  < 2e-16 ***
## as.factor(Acorn)ACORN-M -1.322043   0.109046  -12.12  < 2e-16 ***
## as.factor(Acorn)ACORN-N -1.523866   0.109903  -13.87  < 2e-16 ***
## as.factor(Acorn)ACORN-O -2.873342   0.108169  -26.56  < 2e-16 ***
## as.factor(Acorn)ACORN-P -4.701203   0.109397  -42.97  < 2e-16 ***
## as.factor(Acorn)ACORN-Q -3.471901   0.108627  -31.96  < 2e-16 ***
## as.factor(Acorn)ACORN-U -0.042108   0.109366   -0.39    0.700
## temperatureAvg          -0.163125   0.005568  -29.30  < 2e-16 ***
## humidity                 0.748513   0.251990    2.97    0.003 **
## cloudCover               0.781622   0.104835    7.46  9.5e-14 ***
## windSpeed                0.015395   0.011146    1.38    0.167
## pressure                -0.000329   0.001678   -0.20    0.844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.77 on 12153 degrees of freedom
## Multiple R-squared:  0.733,  Adjusted R-squared:  0.732
## F-statistic:  928 on 36 and 12153 DF,  p-value: <2e-16
```

As we can see above, most of the variables are statistically significant at the 5% level. The variables that are not significant are windSpeed and pressure. ACORN-J (Starting Out), month 12 (December), and Holiday are also not statiscally significant at the 5% level.

For every 1 degree increase in Celsius temperature, there is a decrease in energy consumption of 0.17 kWh. For a household that belongs to ACORN A (Lavish Lifestyles), there is a 7.2 kWh increase in energy consumption (which is pretty close to the average energy consumption). The summer months also have negative coefficients, along with the low income ACORN groups.

We get a similar test mean squared error of **3.15**.

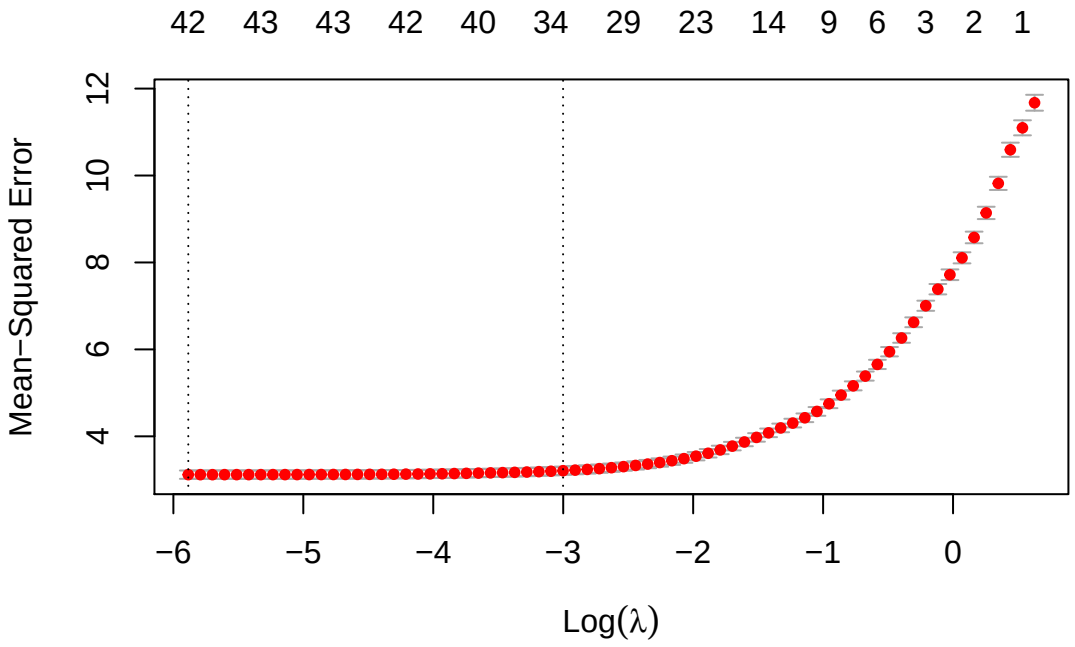


However, looking at the residual plot, we see evidence of a trend in the error terms vs the fitted values. This violates the assumption of linear regression of the error terms being independent and so we must try other models.

4.2 Lasso Regression

We will now try penalized linear regression models i.e Lasso Regression. Since Lasso Regression automatically performs shrinkage and selection of the features, we do not need to modify the features in any way.

We will first use cross validation to choose the best alpha value for both.



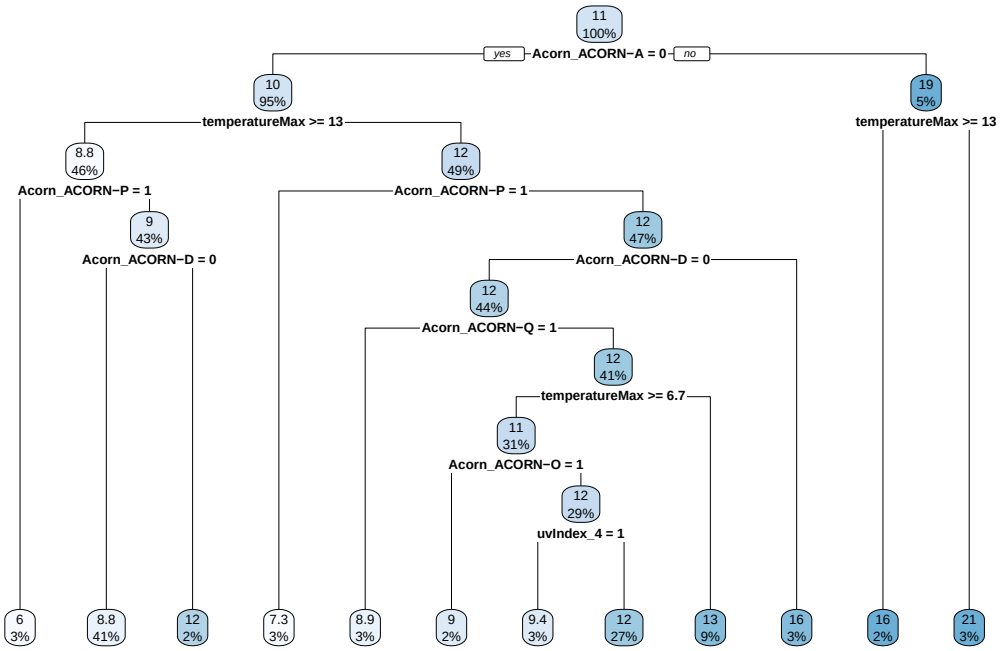
The best value for lambda according to the 1 standard error rule is 0.0468 at which 34 of the features are selected.

We will now a train a lasso model with this lambda value.

We get a test mean squared error of **3.23**, which is similar to the error we got with unpenalized Linear Regression. Thus, penalizing does not offer us any benefit in terms of better predictions.

4.3 Decision Trees

We will now train a decision tree regressor on our training dataset.



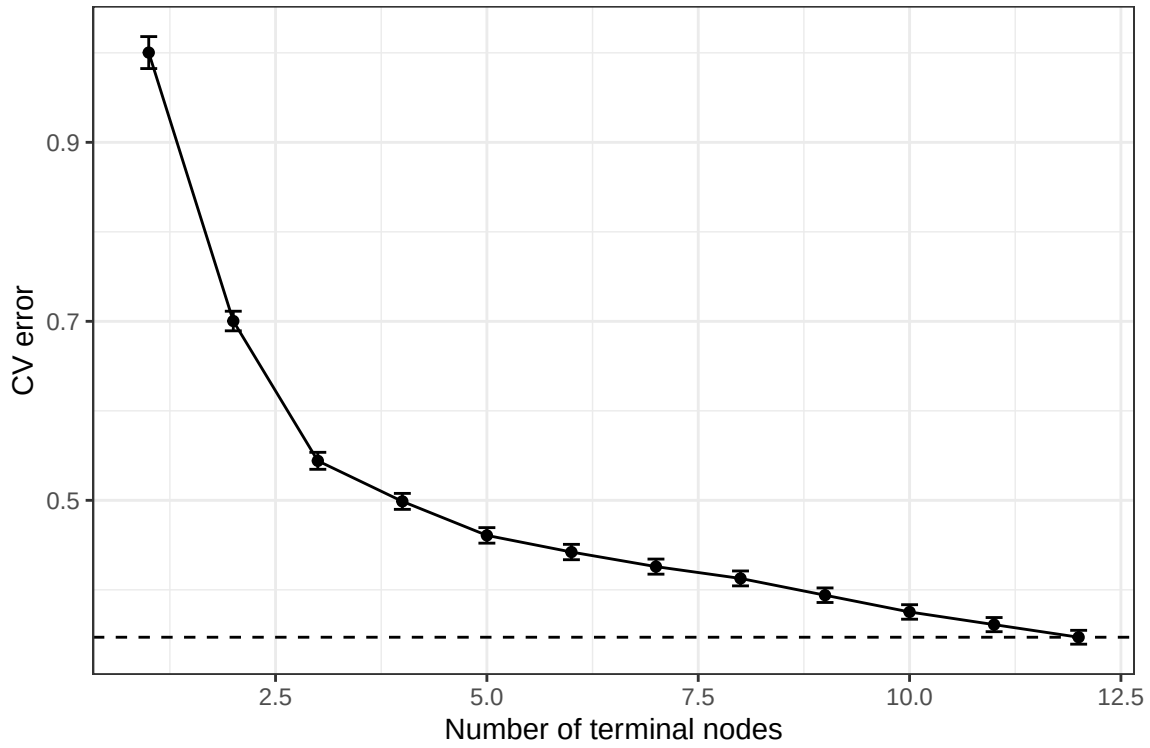
As you can see above, the primary splitting feature is whether a household belongs to ACORN-A (Lavish Lifestyles) or not.

We get a test mean squared error of **4.06** using an untuned decision tree

We will now use cross validation to tune this decision tree and get the best value for minsplit

```
##
## Regression tree:
## rpart(formula = energy_sum ~ ., data = train_data_las)
##
## Variables actually used in tree construction:
## [1] Acorn_ACORN-A  Acorn_ACORN-D  Acorn_ACORN-O  Acorn_ACORN-P  Acorn_ACORN-Q
## [6] temperatureMax uvIndex_4
##
## Root node error: 1e+05/12190 = 12
##
## n= 12190
##
##       CP nsplit rel error xerror  xstd
## 1  0.30      0       1.0    1.0 0.018
## 2  0.16      1       0.7    0.7 0.011
## 3  0.05      2       0.5    0.5 0.009
## 4  0.04      3       0.5    0.5 0.009
## 5  0.02      4       0.5    0.5 0.009
## 6  0.02      5       0.4    0.4 0.009
## 7  0.02      6       0.4    0.4 0.008
## 8  0.02      7       0.4    0.4 0.008
## 9  0.02      8       0.4    0.4 0.008
## 10 0.01      9       0.4    0.4 0.008
```

```
## 11 0.01     10       0.4    0.4 0.008
## 12 0.01     11       0.3    0.3 0.008
```

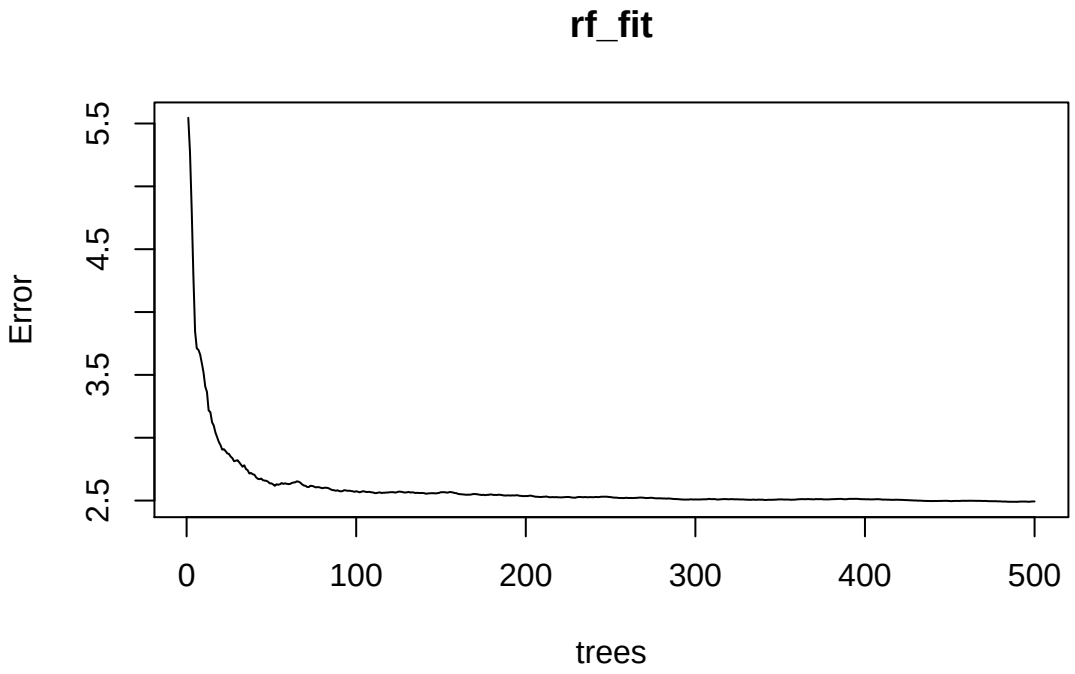


The optimal decision tree has an nsplit of 10 and CP of 0.0101.

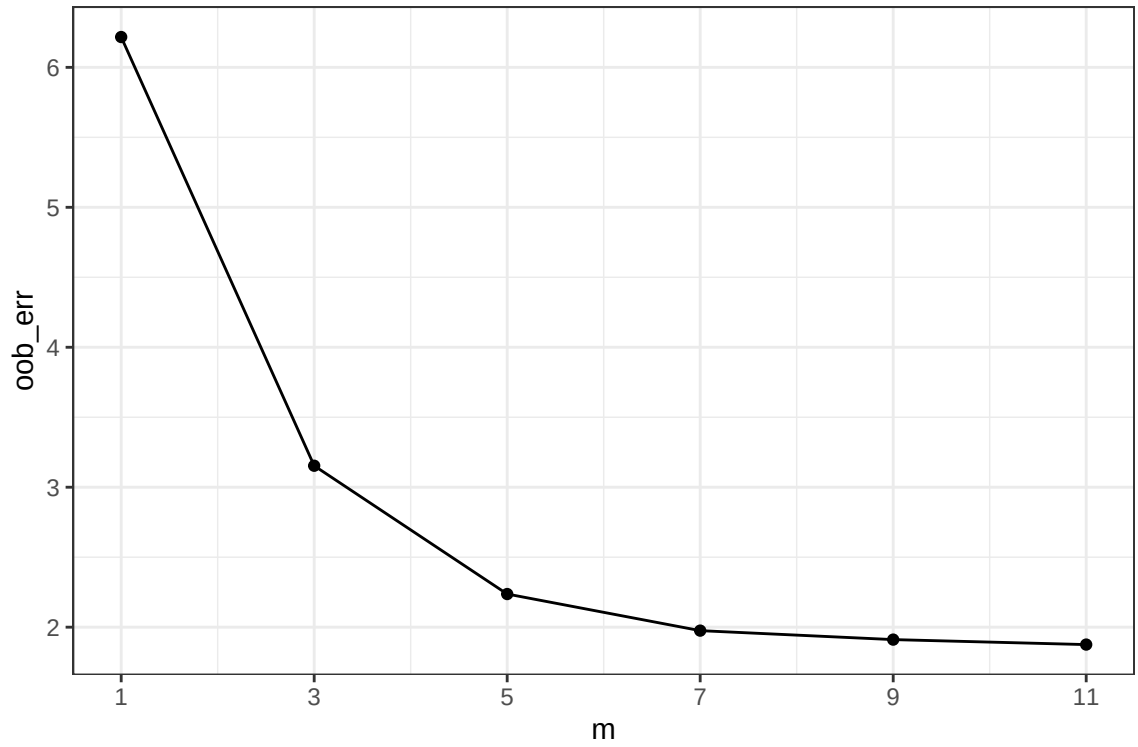
The test mean squared error actually increased to **4.21** with the optimal decision tree. This suggests that it is not useful to fit a single decision tree, even when pruned.

4.4 Random Forests

Let's try fitting a random forest to our dataset!



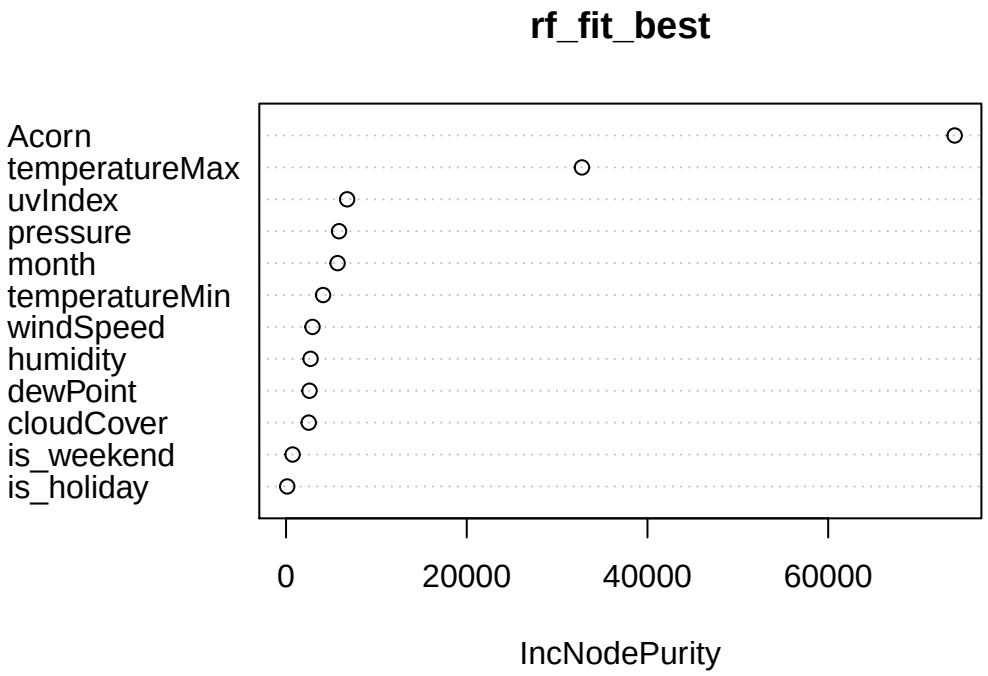
Beyond 200 trees, the error stabilizes so we will use a default of 200 trees from now on.



The best value for mtry is 11 at which the OOB error is about 2. We will now train a final random forest with 200 trees and 11 mtry.

We get a test mean squared error of **1.90**, which is much better than the decision tree and linear regression

models.

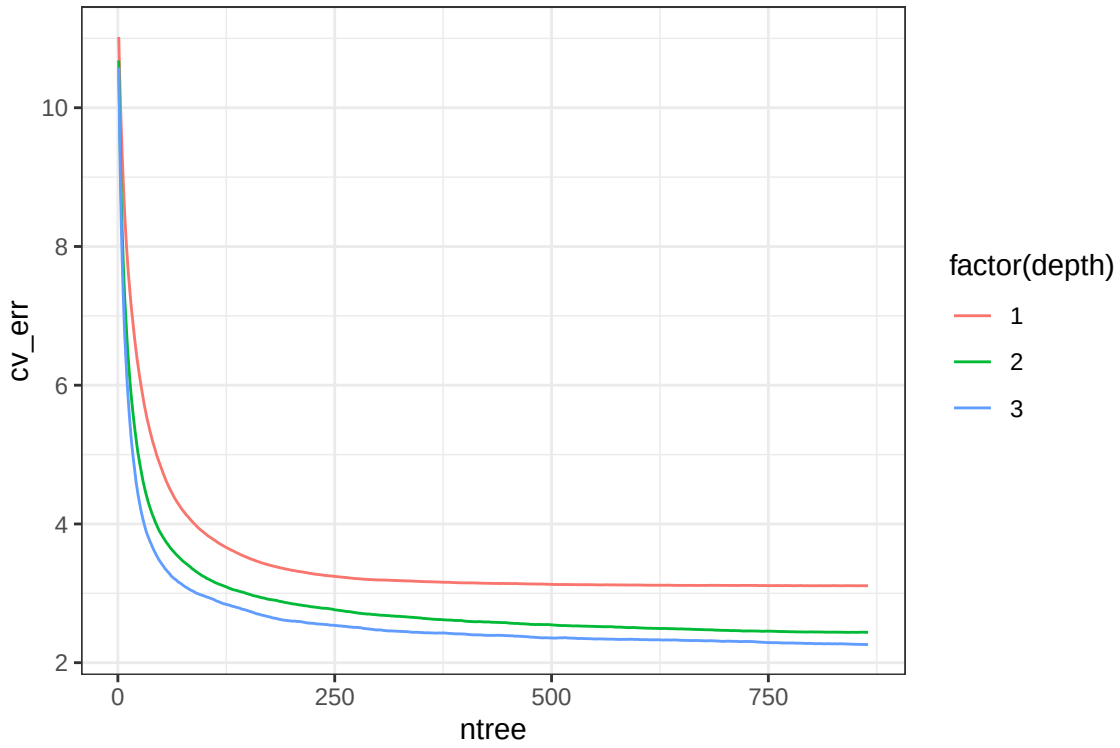


The most important features are Acorn and temperatureMax. All the other features have a similar node purity.

4.5 Boosting

We find the optimal number of trees to be 865.

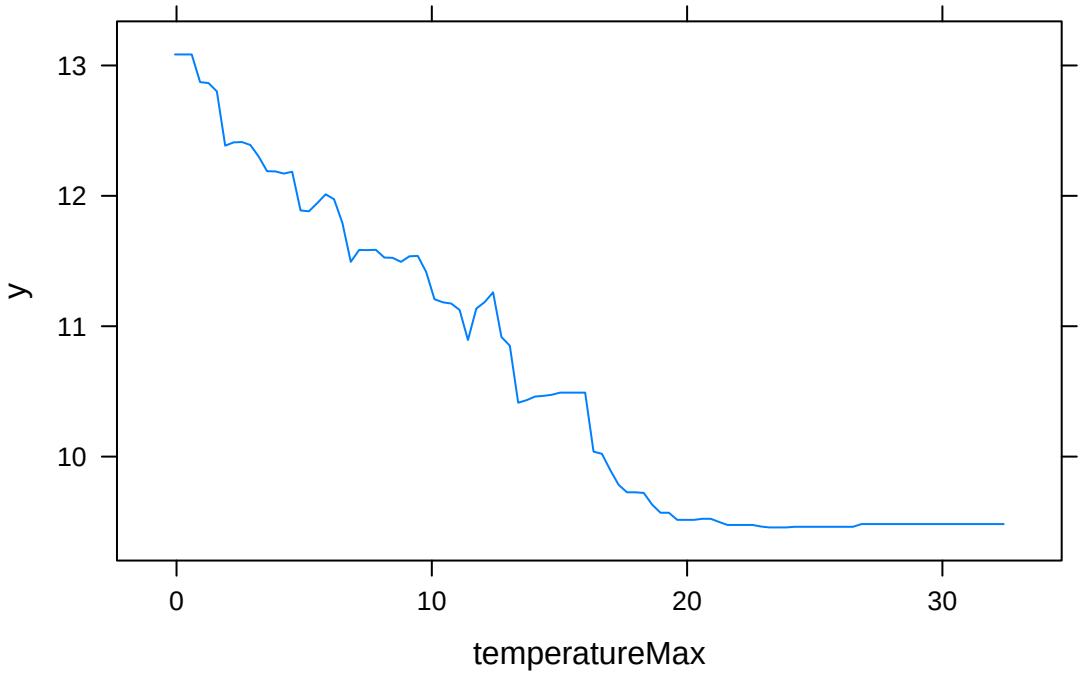
Now we tune the interaction depth.



We find that the most optimal interaction depth is 3, at which the cross-validation error is about 2, and the optimal number of trees is 864.

```
##                              var rel.inf
## ‘Acorn_ACORN-A‘ ‘Acorn_ACORN-A‘   30.94
## temperatureMax   temperatureMax   21.69
## ‘Acorn_ACORN-P‘ ‘Acorn_ACORN-P‘    6.73
## ‘Acorn_ACORN-D‘ ‘Acorn_ACORN-D‘    5.88
## temperatureMin   temperatureMin    3.33
## pressure               pressure    3.30
```

We see that ACORN A and temperatureMax are the two most important variables, by far, in predicting energy consumption.



The partial dependence plot shows that as temperatureMax increases, the predicted value for energy consumption decreases. But after a temperature of around 22 Celsius, energy consumption stabilizes indicating that increases in temperature beyond 22 Celsius don't contribute to lower energy consumption.

The test mean squared error for the Boosting model is **2.19**. This is better than the linear regression and decision tree models, but slightly worse than the Random Forest Regressor.

5 Evaluation and Interpretation

We tested a total of seven models: three linear models and four tree-based models.

Our first linear model ran simple regression on all predictors. Our modified simple regression controlled for collinearity by removing some factors. Finally, we used a lasso model to shrink and select less important predictors. We then tuned the LASSO model to find the ideal lambda value.

Our first tree based model was a simple decision tree built with the default decision tree settings. Then, we tuned the decision tree to find the optimal number of terminal nodes based on the test data. Surprisingly, this yielded a worse test MSE than the vanilla decision tree.

We also trained a random forest model with 200 trees and 11 factors to consider when splitting the tree (m). We chose the number of trees based on the level at which more trees had a zero or trivial impact on the test error. We tuned the number of factors per tree based on cross validation.

For boosting, we tuned the optimal number of trees and the optimal interaction depth. We found that an interaction depth of 3 performed better than 1 or 2, and in order to keep trees small, we kept the interaction depth at 3 and did not test values more than this. We found that, at an interaction depth of 3, the ideal number of trees is 864.

Here are our results summarized:

Model	Test Mean Squared Error
Simple Linear Regression	3.13
Modified Simple Regression	3.15
Lasso Penalized Regression	3.23
Untuned Decision Tree	4.06
Tuned Decision Tree	4.21
Random Forest	1.90
Boosting	2.19

Figure 1: results-mse

6 Conclusions

We found that simple linear regression worked decently well, and that removing predictors with collinearity and penalized regression did not help improve the model performance. This is likely due to the fact that the data did not comply with the assumptions of linear regression, specifically violating independent error terms, making it difficult to predict accurately even when modifications were made.

Tree-based models do not require as rigorous assumptions as linear regression does, and should thus perform overall better on the data. We find that, although the performance of a single tree is very poor, the performance of enhanced tree methods (random forest and boosting) return incredible results. Based on the test MSE, we recommend using the Random Forest model to predict energy consumption on any given day.

The most important features, according to our random Forest model, are income/high-level demographic data (ACORN score), and the temperature. Thus, commercial entities such as electricity companies and alternative energy providers should first target high net worth individuals when it comes to increasing revenue. Further, policy makers or activists trying to limit excessive energy consumption should consider providing education on energy efficiency and energy saving methods to high net worth individuals first.

The results also show that high temperature is very important when it comes to predicting energy consumption. Thus, residents in hotter climates should sincerely consider renewable energy sources. Further, budget-conscious consumers should be aware that their electricity bills will rise with hotter weather. They may wish to budget appropriately, or may choose to take their holidays on days that are predicted to be hotter.

The two largest limitations of this data are the lack of diversity in geographic data and the definition of Acorn data.

All data was collected in London. We recognize that in traditionally hotter climates, low temperatures may be more relevant in predicting energy consumption than higher temperatures; the variation from the norm is what is most likely to cause consumers to change their behavior. Additionally, it may be the case that some factors which do not vary within London (i.e. urban vs. rural status, altitude, geographical location,

etc.) are important in predicting energy consumption. Thus, the models may be limited in their ability to predict data outside of London. We would like to perform a follow-up study that looks at observations from other cities as well.

The ACORN status was a convenient way to put together several demographic attributes in a way that defines similar groups of consumers. This is useful for model building in that some correlated features are combined and helps interpretation for commercial players to better understand the precise market they should target. However, the ACORN variable ignores some variability within these groups, and the cutoffs are somewhat arbitrary. Although we find a lot of predictive power from the ACORN features, it may be more insightful to build a model on the underlying features that compose the model. This may help explain outlier individuals who barely meet the definition of their ACORN status.

Regardless, we found that the more sophisticated tree based methods were very effective in their predictive power, and performed very well on test data. We believe that, although there are ways to improve the model, it provides robust insight into consumer's energy usage behaviors. We believe the use of our random forest model to predict usage will be effective for many stakeholders, even outside of London.