Leeds School of Business
UNIVERSITY OF COLORADO BOULDER

Advanced Data Analytics Project: Analysis of Energy Consumption in Sceaux, France

Final Write-up

Team 2: Jacob Pulitzer, Portia Gifford, MacKenzie Murphy, Sam Al Qarzi, Ross McDonald

Table of Contents

Business Understanding

Business Problem

Today, approximately 65% of the global energy generation comes from fossil fuels. As dire as the ramifications are on this dependency of toxic sources to generate power, a significant amount of this energy goes to waste for the most part due to consumers' misuse. As environmental researchers, our primary goal is to direct energy suppliers and consumers alike to move towards renewable, clean, and safe energy sources. With the data and analysis, we are hoping to shine a light on where and when the highest amount of energy is being consumed in order to better target and educate homeowners' on power waste. Ultimately, we believe that our results will help consumers become more energy conscious and aware leading to increased efficient energy consumption, which will, in turn, have a positive impact on the environment.

Addressing the Problem with Analytics

Through the data, our primary goal was for our descriptive analysis to help us uncover an understanding of all factors that result in power mismanagement. The use of data analytics will allow the team to hone in on when the greatest amount of power use is taking place. For example, calculating the correlations and averages of all variables in the dataset will aid in determining the relations and influence of those features on power usage. With this knowledge, we can confidently output solutions and advise consumers on how to be more efficient in their usage.

Similarly, time variables will map the effect of time on the trend of energy consumption. Testing this effect will educate consumers on how to optimize their consumption in the future. We believe that our analysis will provide consumers with the information that they need to become aware of when their power misuse is taking place, as granular as the day of the week, holidays, etc. and as broad as to which quarters of the year. As a consequence, as consumers efficiently manage their energy consumption, more power will be saved leading to fewer operating power plants and less environmental contamination.

Data Understanding

Data Description

For the project we analyzed a multivariate data set on household power consumption. The data was collected in Sceaux (7km of Paris, France) between December 2006 and November 2010 (47 months). The data set consists of seven variables: day, time, average household global_active_power by minute, average household global_reactive_ power by minute, voltage, Global_intensity, Sub_metering_1, Sub_metering_2, and Sub_metering_3. Sub_metering_1, Sub_metering_2, and Sub_metering_3 relate to active energy consumption in the Kitchen, Laundry room, and Climate control systems respectively. Our data was sourced by Georges Hebrail (georges.hebrail '@' edf.fr), Senior Researcher, EDF R&D, Clamart, France Alice Berard, TELECOM ParisTech Master of Engineering Internship at EDF R&D, Clamart, France.

Analytics problem

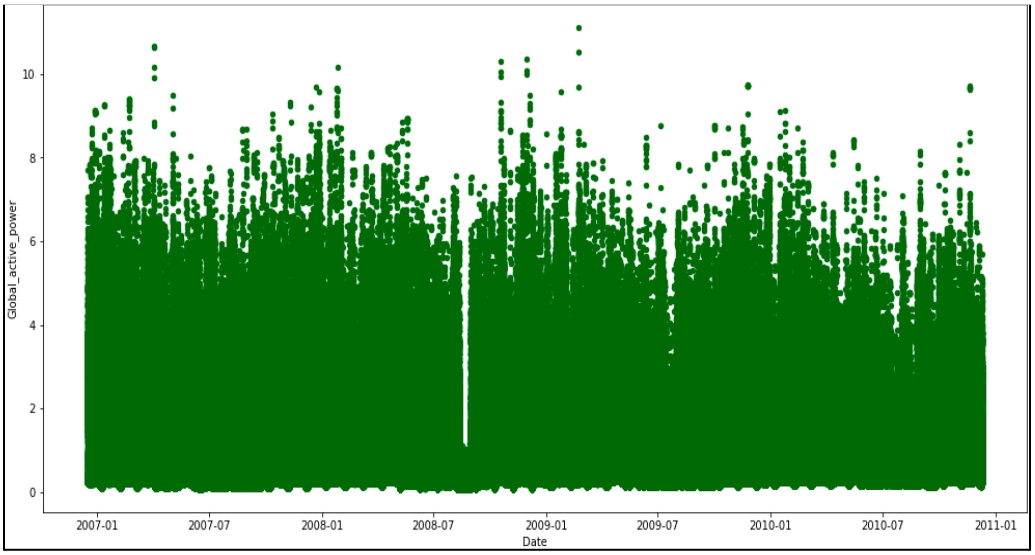
Our analytics problem is supervised due to the type of data we have acquired. All independent variables in the data lead to the output of household energy consumption. Our intention as stated above is to utilize all the independent features in the data to better predict the continuous variable of power usage given the most important predictors we find. We understand that both date and time variables in the dataset are essential to predict change in power usage over time. Similarly, we have some information in the dataset about global averages and voltage data that could play a part in leading total energy consumption to fluctuate.

Data Preparation

Data Integration & Formatting

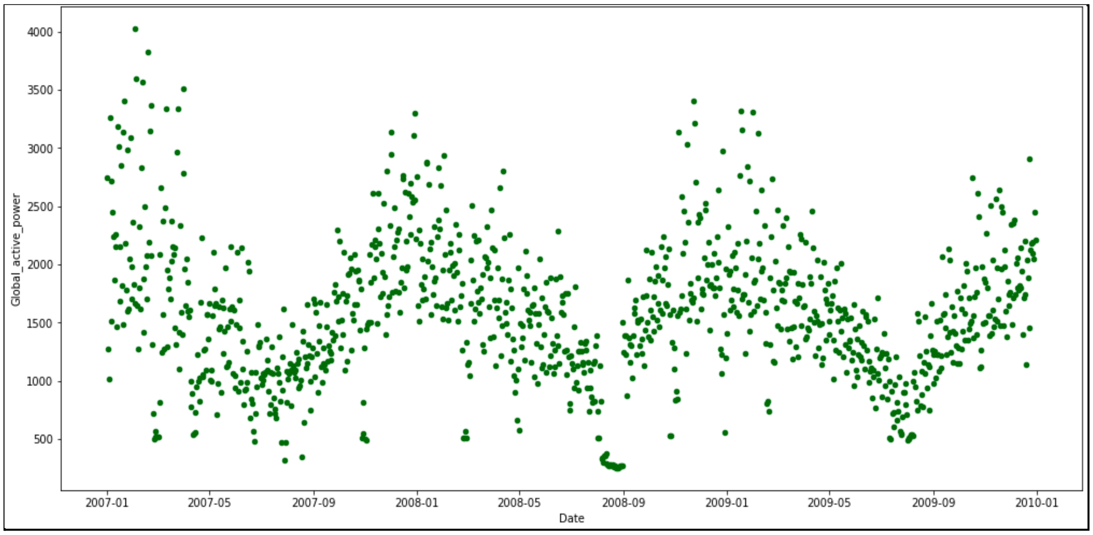
Originally, the data was a text file that we had to convert to a csv and then launch our data preparation process. The dataset has about 8 columns and around 2.1 million records, each record in the data mirrors a minute data which consist of date, time, power consumption, voltage, power intensity, 3 columns for submetering for different parts of the house.After reading the data in Rstudio, we started the preparation process by checking the data for nulls and duplicates as well as formatting the columns into their right formats. Fortunately, the dataset had a few missing values in one of the columns and no duplicates. With his huge frame of data, we faced many issues in RStudio including plotting the data as well as performing simple analytical calculations. *Figure 1,* found below, is a simple visualization of power consumption against date variable, as shown, the density of data is troubling which led us to narrow the dataset to a manageable number of records before we proceed with further analytics.

Figure 1: Global Active Power per Minute by Date



We narrowed the dataset by taking the hourly average which left us with around 35000 records. With this amount of data, we were able to get a sufficient understanding of the data by visualizing the variables and conducting some descriptive analysis of the target variable. Because this data is a time series data, our main goal was to fit the data on a time series model (e.g ARIMA, SARIMA). However, we encountered some challenges in Rstudio converting the dataset to a time series data. After extensive research, we reached the decision of totaling the original data by day to obtain meaningful results from our models, and we excluded years with incomplete cycles in the dataset. *Figure 2* found below shows the daily total power consumption for the three cycled years in the dataset (2007, 2008, 2009).

Figure 2: Total Daily Global Active Power by Date



Modeling

To perform our analysis of energy usage we decided to use 5 different models. Each model can be used to better understand how different time factors relate to energy usage, and further the models can be used to predict future energy usage. The 5 different models we built are as follows: Time-series analysis with Facebook Prophet, regression tree, classification tree to predict high energy days, XGBoost for best predictability, and a multivariate regression for feature selection.

Facebook Prophet

Description

As we encountered some issues to convert the data to a time series data in Rstudio, we decided to look to another alternative. Facebook Prophet is a python package widely implemented in similar situations where time seasonality holds an influence over the predictive variable. The package comes packed with various helpful features including all time seasonalities (e.g yearly, monthly, weekly, and daily seasonality) as well as holiday effects. Prophet models only take two inputs which are the timestamp variable as *ds* and the target variable noted as *y*. Prophet models can accept as many predictors as needed that can be fed into the model as external regressors. However, as we understood from our correlation analysis that all variables are in huge part factored in into our target, we proceeded with fitting our target on time variable only. Prior to fitting the prophet model, we split the daily data into train and test sets using 2007 and 2008 data as train set (amounting to 731 records) and 2009 data as validation set (amounting to 365 records). We set up the model's parameters to include all seasonality features and holiday effects of French national holidays since the data was collected in France.

Results/Findings

Unfortunately, the model did not perform as well as we had hoped. It produced an R_Squared value of 0.37 and mean absolute error of 311.42 which is not that bad given that the data points range between 200 and 4000. *Figure 3* shows the fitted train set of two years power consumption data continued with Prophet predicted power consumption of 2009.

Figure 3: Time Series Forecast for 2009

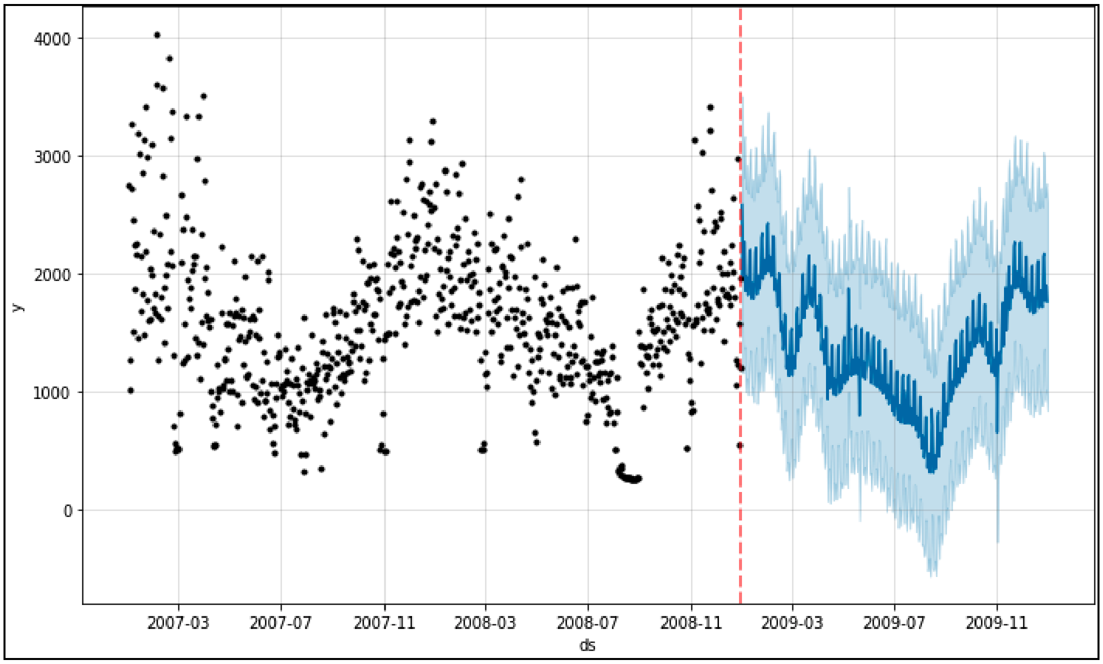
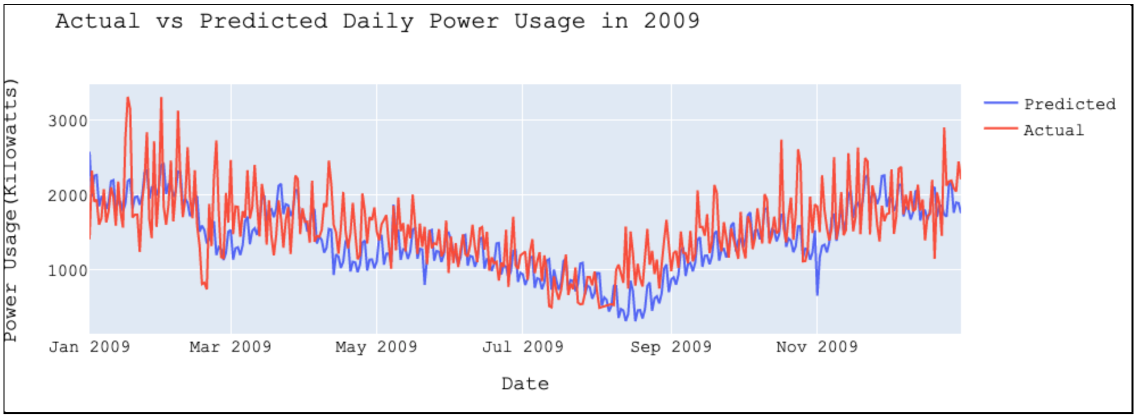


Figure 4 below shows Prophet daily predictions in 2009 compared the actual daily power consumptions in the same year. As trained, the model partially succeeded to capture the trend of summer power consumption from mid May until mid August. However, it failed to learn the effect of daily seasonality in the beginning as well as the end of the year where power consumption was higher due to extreme weather conditions.

Figure 4: Time Series Actual vs. Predicted Daily Power Usage



Conclusion

Based on our findings from this model, we conclude that the model had insufficient training data to accurately capture the trend of power usage. Two years worth of data apparently was not adequate for the model to train on. We believe that if the model was fed more training data, it would have increased its predictive accuracy.

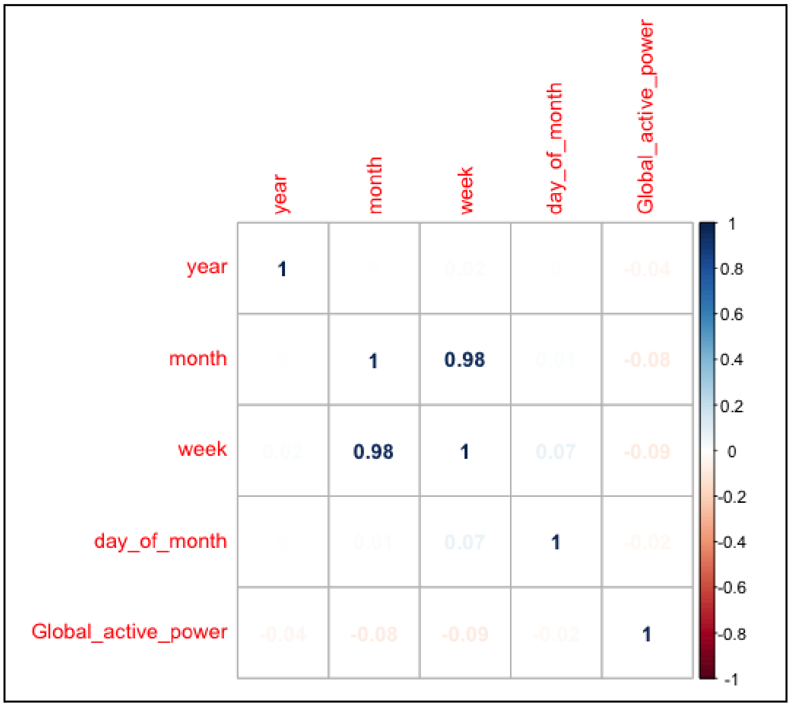
Classification Tree

Description

Classification trees can be very useful when trying to predict a binary outcome. Because we are interested in determining which times in the year are more likely to see high energy usage, we built a classification tree. The goal of this tree model was to predict days in the upcoming year that will have a high energy usage. We defined high energy usage as a day where the total energy usage is greater than the upper quartile, or over 1909.5 kilowatts/day.

Before building the tree, it was important to remove any variables that are highly correlated with each other. The month and week variables are almost directly correlated with each other at 95% (*see figure 5*). We decided to remove the week variable because month had a higher correlation to our target variable, Global_active_power. When building the tree model, all observations up to 2009 were used as the training set, and observations in 2009 were used as a test set. By splitting the observations into a train and test set we are able to determine how much predictive power the model has. The time variables used to predict energy usage are as follows: quarter, month, day_of_week, day_of_month, iswkend, is_holiday, and year.

Figure 5: Correlation Plot

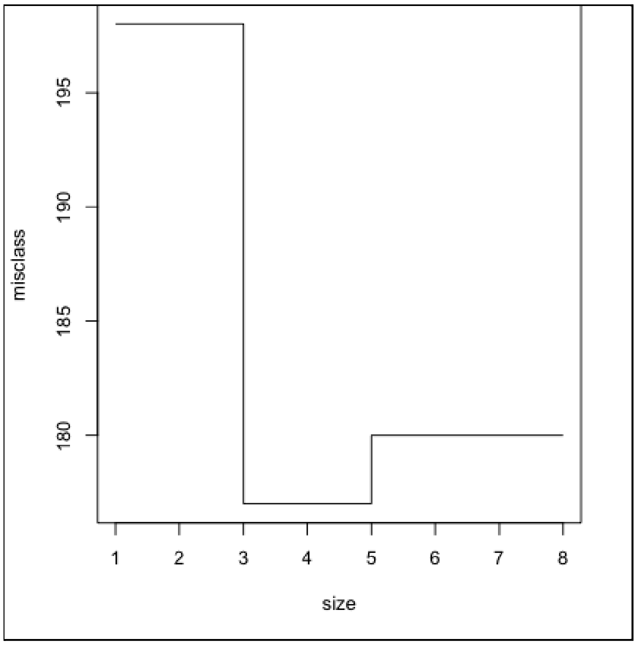


After fitting the data, the variables used in the construction of the model are quarter, month, iswkend, day_of_month, and day_of_week. Once running the initial tree, we ran a 10-fold cross validation on the training set, and then pruned the tree to the number of leaves with the lowest misclassification rate. Lastly, we ran predictions on the training set with the pruned tree to determine the predictive power of the model.

Results/Findings

Before pruning our tree, the model achieved an accuracy of 79% on the training set and 78% on the test set. However, to find the optimal number of leaf's and to validate the results, it is important to run a cross validation. After running a 10-fold cross validation and pruning the tree to the optimal number of leaf's (*see figure 6)*, which was three, the model achieves an accuracy of 75% on the training data and 80% on the testing data. The variables dropped from the model during the pruning are month, iswkend, and day_of_month. The most significant factor in predicting a high energy day is if the observation falls on a Saturday, Sunday, or a Tuesday (for some odd reason). This model puts a lot of weight on quarter two and three, making it nearly impossible for an observation during these time periods to be classified as a high power use day.

Figure 6: Misclassification error rate by size of tree



Looking at the pruned classification tree (*see figure 7*) we can see that an observation in quarter two and three will almost guarantee low or average power usage. Next, if it is a Friday, Monday, Thursday, or Wednesday, we predict low to average power usage. The only time we predict high power usage is when the observation is not in the second or third quarter, and not on a Friday, Monday, Thursday, or Wednesday. When comparing the pruned tree (*figure 7)* with the unpruned tree (*figure 8)* we can see that day_of_month, month, and iswkend all dropped out of the model.

Figure 7: Pruned Classification Tree

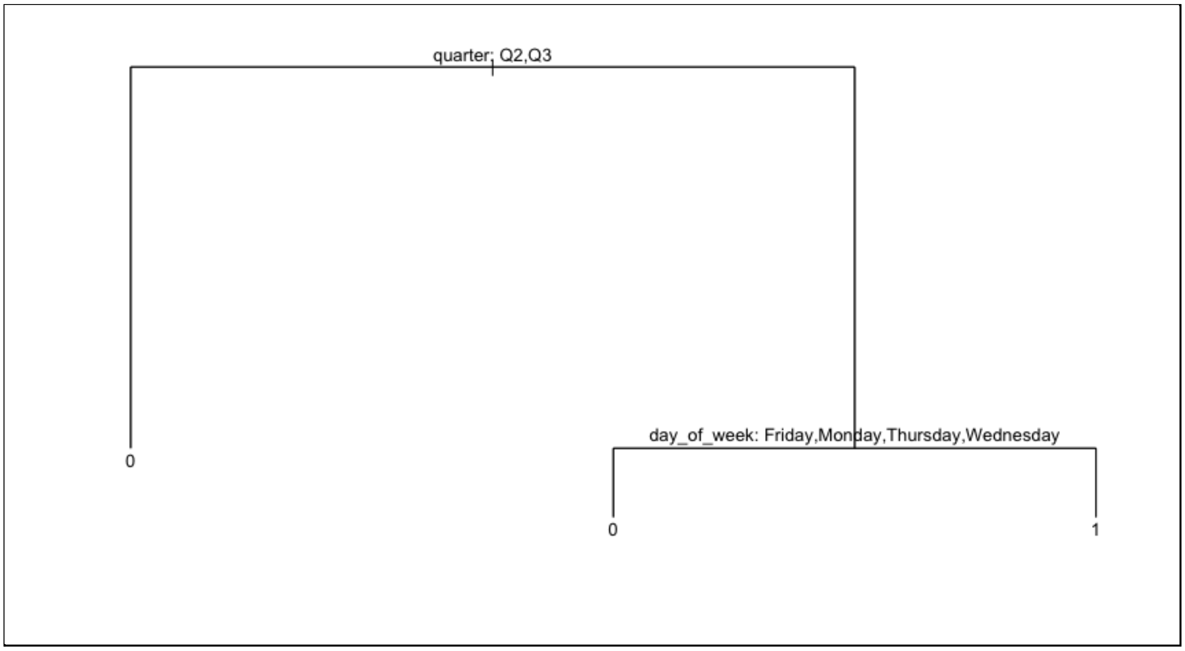
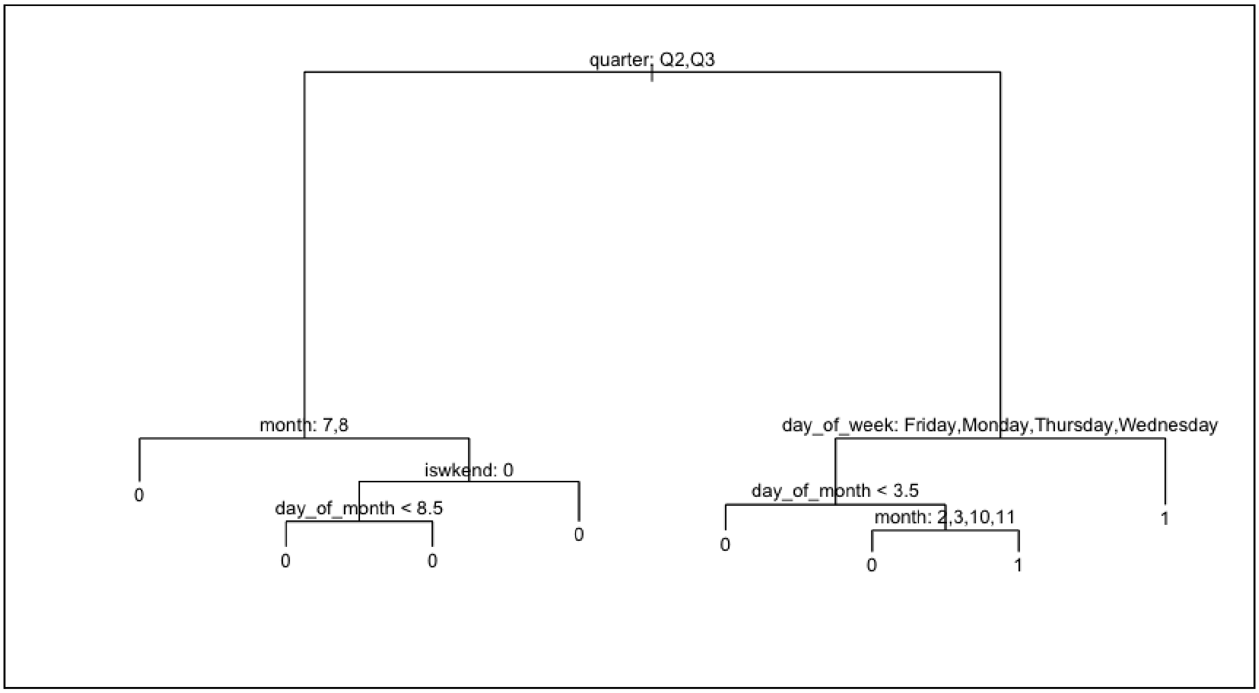


Figure 8: Unpruned Classification Tree



Regression Tree

Description

To aid in our understanding of which time factors result in the highest and lowest global active energy, we built a simple regression tree. The same preprocessing steps used in the decision tree were used to build this tree, but our target variable was total active energy use rather than high active energy. To build the tree, we first fit the tree to the training set. Next, we ran a cross validation and pruned the tree to the optimal number of terminal nodes, determined by the lowest deviance of residuals. Lastly, we ran predictions on our test set, and calculated the RMSE to determine the model with the highest predictive power..

Results/Findings

Because we are trying to determine how much predictive power our model has, we used RMSE (root mean squared error) to find the optimal fit. The unpruned base regression tree (*see figure 9*) resulted in predictions with a RMSE of 382.03, meaning on average the model can predict the global active power within 382.03 kilowatts on any given day. After pruning the tree to a size of 4, determined with a cross validation (*see figure 10*), we were able to achieve a RMSE of 370.68, improving our model's predictive power. The variables used in the base tree construction, prior to pruning, were month, if it is a weekend or not, the day of week, day of month, and the year, consisting of 10 terminal nodes. After pruning, the variables used to construct the tree were month and day of week. The model predicts the highest energy levels, at 2215 kilowatts, will occur on Saturday and Sunday in the months of January, February, March, October, November, and December (*see figure 11)*. The lowest predicted energy levels will fall in June and July, with a predicted active energy of only 901.5 kilowatts.

Figure 9: Unpruned Regression Tree

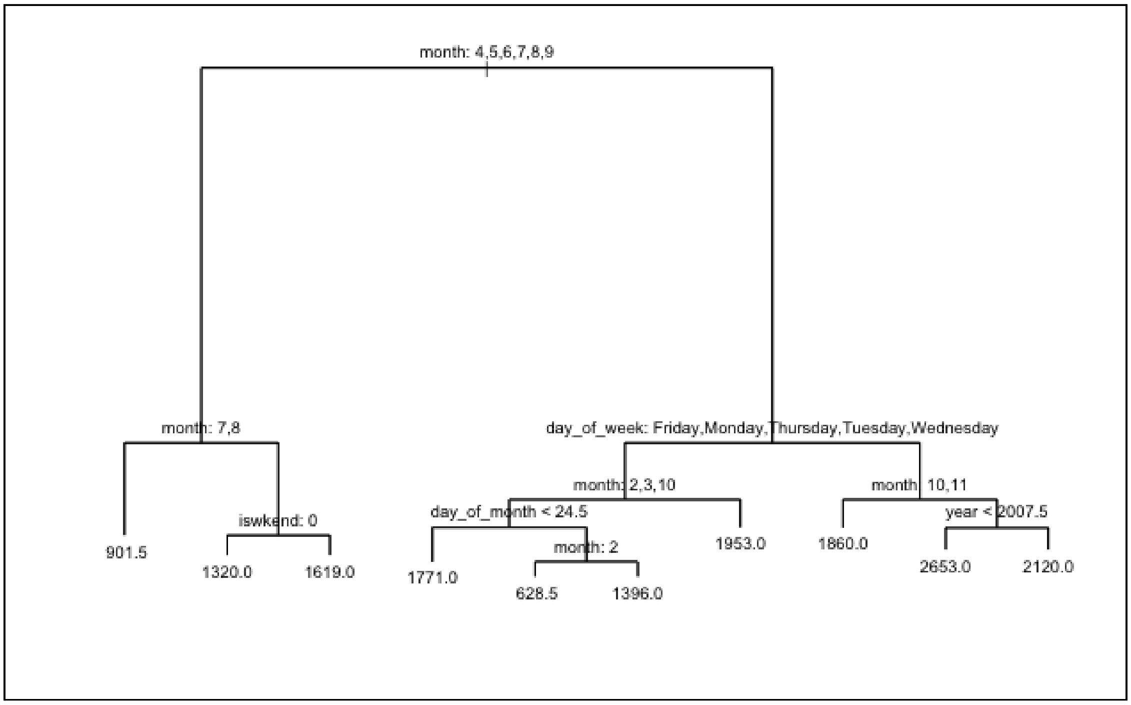


Figure 10: Cross Validation for Regression Tree

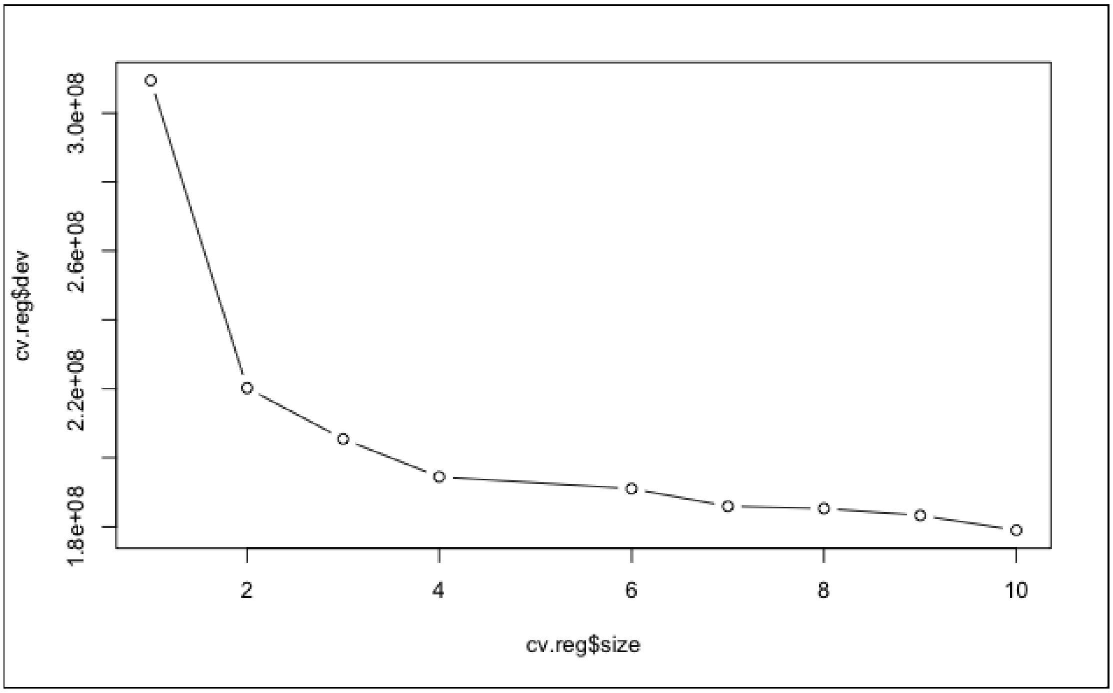
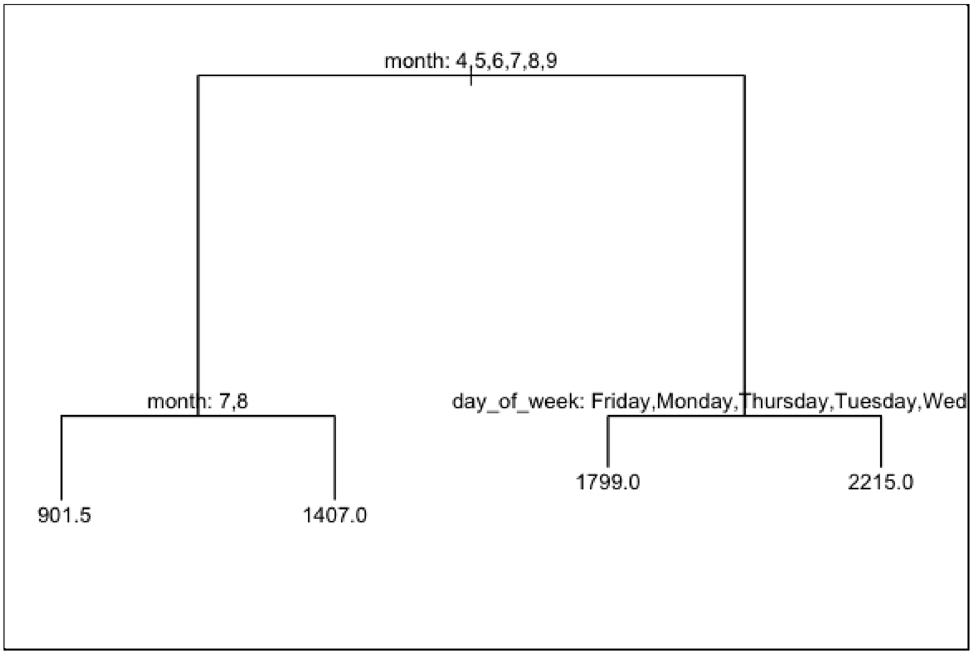


Figure 11: Pruned Regression Tree



XG boost

Description

XGboost also called Extreme Gradient Boosting, was an algorithm that was originally created during a machine learning competition on kaggle and now has been adapted as a package on R. The XGboost algorithm is extremely useful and is much faster than other approaches to tree modeling. The XGboost algorithm often outperforms other tree based models because of its efficiency, ease of use and highly accurate results. The model improves upon the basic gradient boosting model, it does so well because it uses multiple trees at the same time; on tries to predict a target and the other tries to predict the residuals from the first tree, the third uses the second tree to predict the residuals and so on and so on. The model also eliminates the possibility of overfitting by using LASSO (L1) and Ridge (L2) both of which are regularization techniques.

Lasso shrinks the less important features coefficients to zero, removing the feature. Ridge(L2) regression helps to reduce multicollinearity by adding a degree of bias to the regression estimates. In combination with each other we get a model that is extremely accurate and usually outperforms other models.

In answering our business question we wanted to understand which features contributed the most to Global active power. XGboosts can be used for both predicting binary classification probability or it can be used with regression when the outcomes are none binary. For our purposes we want to look at regressions since our dependent variable is not binary and is a continuous variable.

To begin I wanted to reduce the amount of variables within the data set that were directly correlated to our outcome variable, as well as use the same features that were used in other models for the project so we can cross reference our results. I removed columns; Voltage, Sub_metering_1, Sub_metering_2, Sub_metering_3, Global_intensity, Date, Global_reactive_power, X). Next I wanted to convert the columns Quarter, and days of week to their own column, to do this I used a process called one hot encoding. This way each Quarter and Day of the week had their own column that the algorithm could use to evaluate their global active power

Results/Findings

The output of our Xgboost model showed that the most important features within the data were the weekdays of Thursday, Tuesday, Friday and that it is a weekend (*observed in figure 12*). This however is somewhat surprising, based on the other models performed on the data we would expect Saturday to be an important variable as well as quarters, we would expect to see higher usage during the weekends when people are home as well as increased usage in the winter seasons. The Mean Squared Error is .079, and RMSE is .281 which are both very low which would indicate that these predictions are very accurate. Some issues however, are that we had to remove most of the data and the remaining data is mostly dummy coded. I would not recommend using this model as an indicator of which days result in the most power usage however with more data using an xgboost model could be extremely helpful for a deeper understanding on how certain days or features can impact power usage.

Figure 12: Feature importance

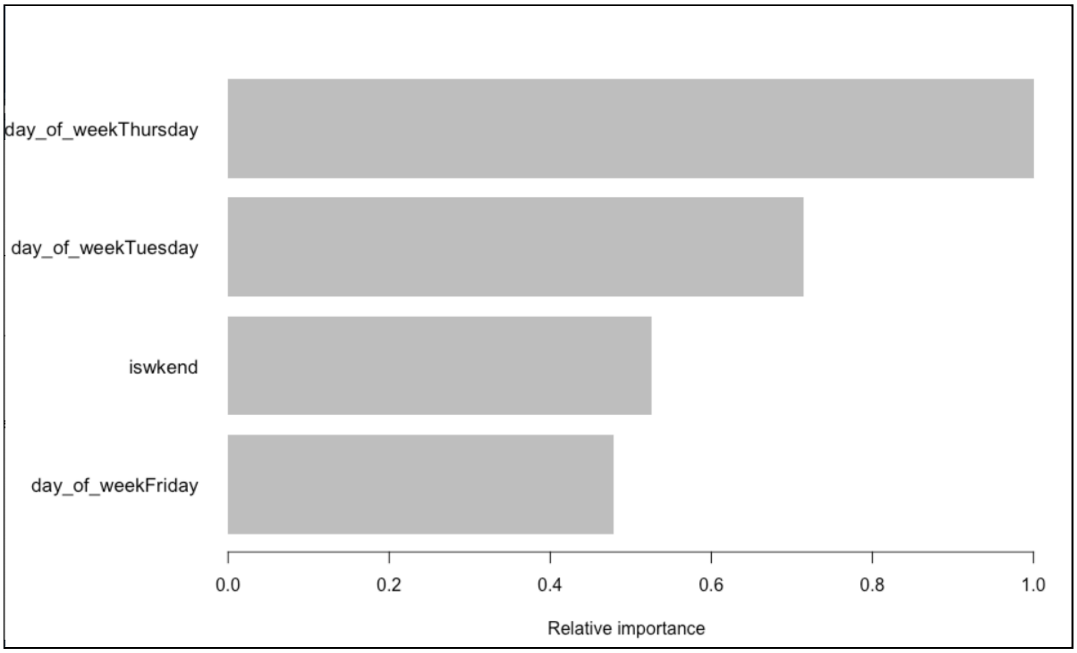
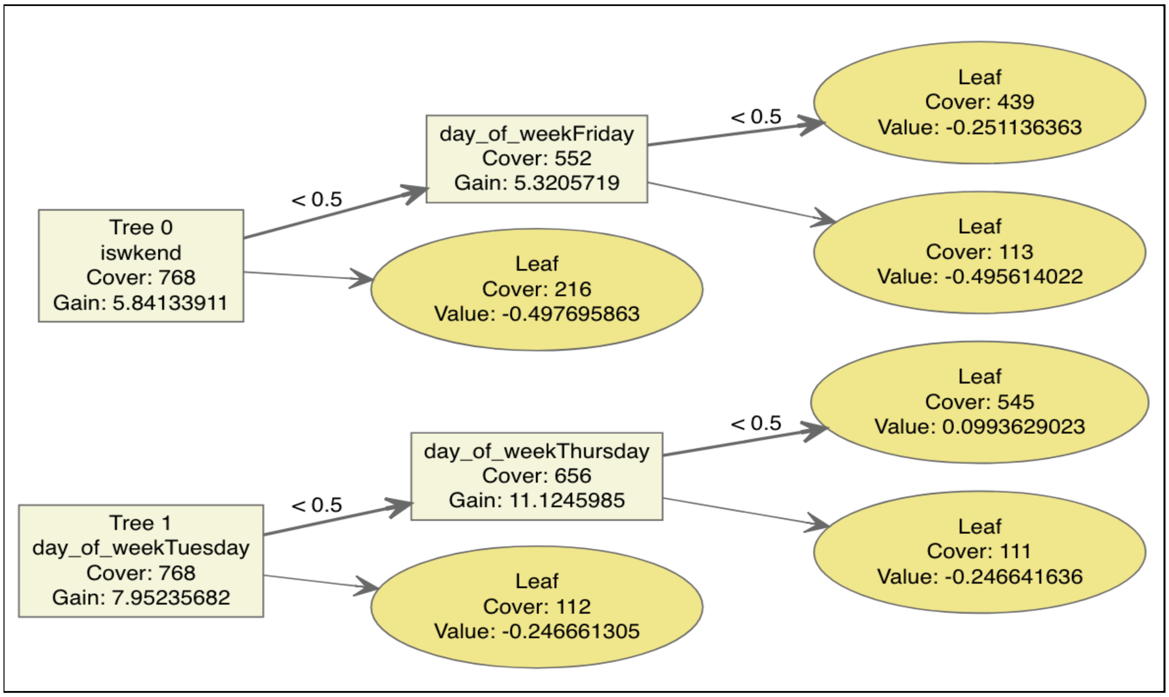


Figure 13 : Xgboost tree model



Linear Regression

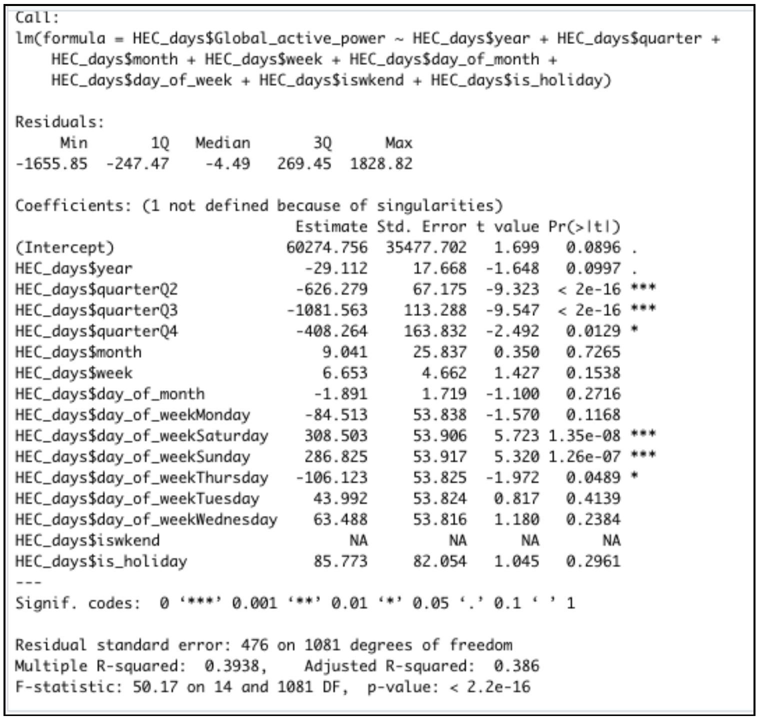
Description

It was clear for the team from the beginning that a linear regression would be necessary and useful in understanding which specific time period variables we created were correlated with the highest and lowest amount of energy consumption. Although the initial regression was performed using all of the variables given in the dataset, the results made it clear that focusing solely on the time variable would be the most impactful. This decision can be attributed to the fact that variables such as "voltage" are essentially a direct factor of energy consumption, so was clearly displayed as highly significant. Instead, the variable "Global_active_power" was regressed on the years, quarters, months, weekends, days of week and holidays.

Results/Findings

The results from the regression output show the variables 'isWkend" and "quarter3" have the largest significance, the weekend variable is the most positively significant at 312.77 and quarter 3 as the most significant in terms of having a negative estimate of -1069.55. In addition, variable “quarter2” is found to be somewhat significant at -621.26 and as well as “month” at an estimate of 35.97.

Figure 14: Linear Regression Model 1



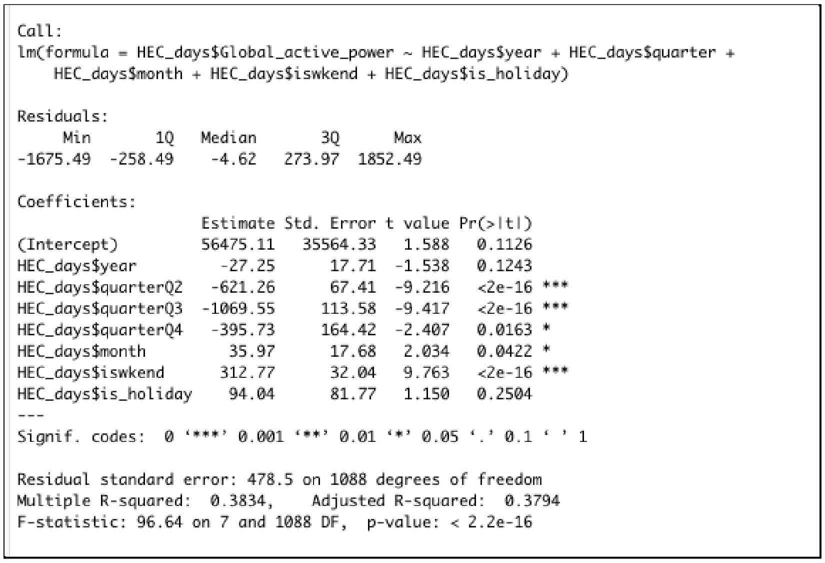
```
Call:
lm(formula = HEC_days$Global_active_power ~ HEC_days$year + HEC_days$quarter +
    HEC_days$month + HEC_days$week + HEC_days$day_of_month +
    HEC_days$day_of_week + HEC_days$iswkend + HEC_days$is_holiday)

Residuals:
     Min       1Q   Median       3Q      Max
-1655.85  -247.47    -4.49   269.45  1828.82

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                60274.756  35477.702   1.699   0.0896 .
HEC_days$year                -29.112     17.668  -1.648   0.0997 .
HEC_days$quarterQ2          -626.279     67.175  -9.323  < 2e-16 ***
HEC_days$quarterQ3         -1081.563    113.288  -9.547  < 2e-16 ***
HEC_days$quarterQ4          -408.264    163.832  -2.492   0.0129 *
HEC_days$month                 9.041     25.837   0.350   0.7265
HEC_days$week                  6.653      4.662   1.427   0.1538
HEC_days$day_of_month         -1.891      1.719  -1.100   0.2716
HEC_days$day_of_weekMonday   -84.513     53.838  -1.570   0.1168
HEC_days$day_of_weekSaturday  308.503     53.906   5.723 1.35e-08 ***
HEC_days$day_of_weekSunday    286.825     53.917   5.320 1.26e-07 ***
HEC_days$day_of_weekThursday -106.123     53.825  -1.972   0.0489 *
HEC_days$day_of_weekTuesday    43.992     53.824   0.817   0.4139
HEC_days$day_of_weekWednesday  63.488     53.816   1.180   0.2384
HEC_days$iswkend                   NA         NA      NA       NA
HEC_days$is_holiday            85.773     82.054   1.045   0.2961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 476 on 1081 degrees of freedom
Multiple R-squared:  0.3938,	Adjusted R-squared:  0.386
F-statistic: 50.17 on 14 and 1081 DF,  p-value: < 2.2e-16
```

Figure 15: Linear Regression Model 2



```
Call:
lm(formula = HEC_days$Global_active_power ~ HEC_days$year + HEC_days$quarter +
    HEC_days$month + HEC_days$iswkend + HEC_days$is_holiday)

Residuals:
     Min       1Q   Median       3Q      Max
-1675.49  -258.49    -4.62   273.97  1852.49

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         56475.11   35564.33   1.588   0.1126
HEC_days$year         -27.25      17.71  -1.538   0.1243
HEC_days$quarterQ2   -621.26      67.41  -9.216   <2e-16 ***
HEC_days$quarterQ3  -1069.55     113.58  -9.417   <2e-16 ***
HEC_days$quarterQ4   -395.73     164.42  -2.407   0.0163 *
HEC_days$month         35.97      17.68   2.034   0.0422 *
HEC_days$iswkend      312.77      32.04   9.763   <2e-16 ***
HEC_days$is_holiday    94.04      81.77   1.150   0.2504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 478.5 on 1088 degrees of freedom
Multiple R-squared:  0.3834,	Adjusted R-squared:  0.3794
F-statistic: 96.64 on 7 and 1088 DF,  p-value: < 2.2e-16
```

Evaluation

Overall Evaluation of all Models

All of our models had a low accuracy level averaging around 38% except the classification tree which had a 80% accuracy rate. We believe that all of our linear models had an inadequate learning curve with the low amount of data they were fed.

The decision tree model performed the best, with an accuracy of 80%. Because we are predicting unbalanced classes (77% of the observations are not high power days, 22.6% of them are), the model clearly holds some predictive power. If we were to simply predict all observations as not high power days, we would achieve a base accuracy of 75% (only misclassifying 25% of the days as not high energy use when they were). Considering this model has an 80% accuracy, we gain appx 4% prediction accuracy when using this model.

Deployment

Our Plan

As environmental researchers, it is our responsibility to take our findings and urge the population towards a change in behavior. From our analysis, we were able to find the factors that cause the most energy consumption, with this in mind, raising awareness and spreading the word of the harmful effects these factors have is paramount to the health of the planet. In the analysis, it was evident that weekends were a significant factor as well as the month being an indicator to the total Global Active Power. By spreading information about the harmful effects of the high levels of energy consumption during these days, we hope that we can get homeowners to switch to more efficient appliances within their homes and ultimately lower the total Global Active Power value.

Issues, Risks, and Concerns

Throughout the course of this analysis, we were able to identify the times of year that contributed to the highest levels of energy consumption. However, as discussed in the Facebook Prophet model, we failed to get the accuracy we originally were hoping for. This would be a concern our group has moving forward. Another aspect that requires further research before deployment would be to find why there are slight differences in the results of our models. They all are pointing towards the same conclusion, but have minor differences that we would like to focus on.

Ethical Considerations

The ethical considerations that should be considered and addressed in the future, pertain to the accuracy of the results being low and the fact that the data only takes into account one specific location. The low accuracy of the models could have implications on the insufficiency of the data after averaging it by day rate. As for the data set soley encompassing the energy consumption for the region of Sceaux, France, we cannot make the same recommendations to other regions or countries due to the different variables that would be present for each.

The models we produced do not have the best accuracy scores, meaning we could be pushing for change during the wrong times. The process of mitigating these risks include gathering more data to train our models on and being able to spend more time on the building and analysis aspects of this report. With that being said, if we are able to make change happen in the population even at the wrong times, it is still a benefit towards lowering the total Global Active Power.