Final Project - IST 387

Ethan Jose

Table of Contents

Introduction

My goal as a consultant for eSC was to analyze and understand energy usage across an area of South and North Carolina in hopes of determining significant factors in energy usage and consumption. In today's world, we see consistent increases in temperature during the summer so my goal was to examine July 2018's data, the hottest month. Using exploratory analysis and predictive modeling, I hope to provide several key recommendations based on my scientific research of the data.

Business Questions Addressed:

1. What factors impact the demand for energy in the specific Area within South and North Carolina?
2. What factors impact the efficiency of cooling a house during hot summer months?
3. What **climate** related factors affect energy efficiency?
4. What **physical** factors relating to a house affect energy efficiency?
5. What **insulation** and **building materials** impact energy efficiency the most?
6. What are the most important factors to prepare for when the temperature increased by 5 degrees in the month of July?

Data Munging, Filtering, Prepping, and Cleaning

MARKDOWN PAGES 1-8

MERGING

- The main challenge for this dataset is combining the static individual house data, the hourly energy use data, and weather data by county
- To combine these I:
    - Looped over every row in the static data, and for each building id in each row, I read in the parquet energy data- filtering to only get July (saving large amounts of runtime)
    - Additionally, in each row I also read in the weather data from the 'in.county' column to get that house's respective weather data
        - I used fread which is significantly faster than read_csv or read.csv
        - I filtered for only July as well to prevent mismatched binding of data
    - Finally I used rbind to merge the respective static house data rows with the parquet energy data and csv weather data

CLEANING and Prepping

- First I fixed column names that were hard to use in R, specifically those used in the weather data columns
- I also used indexing to determine if any of the energy data was all zeros, meaning they would be useless in our analysis, and removed these
- I also calculated the total energy consumption by adding each energy consumption that was not all zeroes together
- I next decided on my predictor values, focusing on important weather metrics like temperature and wind speed
- In addition to the climate predictors I also made multi-level factors for insulation values, window type, window area, and material
    - I did this in order to analyze these values numerically instead of as strings, which would be difficult to analyze efficiently

- This presented a challenge in the fact that I am not a specialist in build material, insulation, or construction types so my interpretation of ranks could be different than others. I considered pursuing one hot encoding to truly understand which values would be impactful for different ranks but within the timeframe, I decided it would be best to settle for hierarchical integer factor values where for the most part, it is clear what the hierarchy is

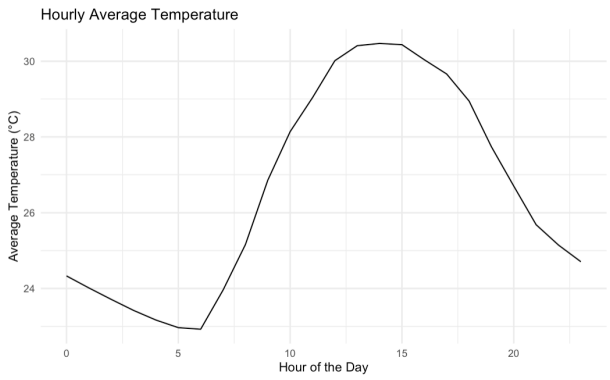
Final Prep and Selection

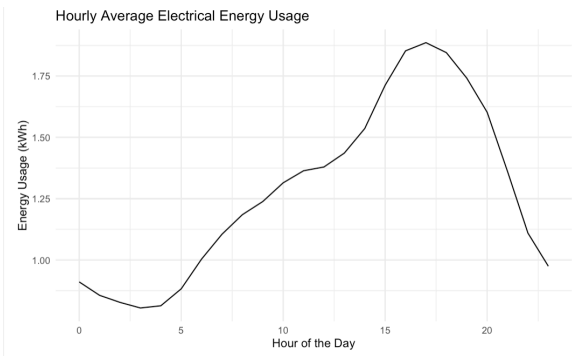
- My last step was simply selecting the predictor values using dplyr, and making the hour and day columns readable so we could see each row clearly
- I also created a column for energy by square feet, to standardize our model results, using total energy consumption would produce useless models as it is essentially the result of what we are predicting
- We need to target a variable that does not exist right now for the model to work.

DESCRIPTIVE STATISTICS AND VISUALIZATIONS

MARKDOWN PDF PAGES 8-12

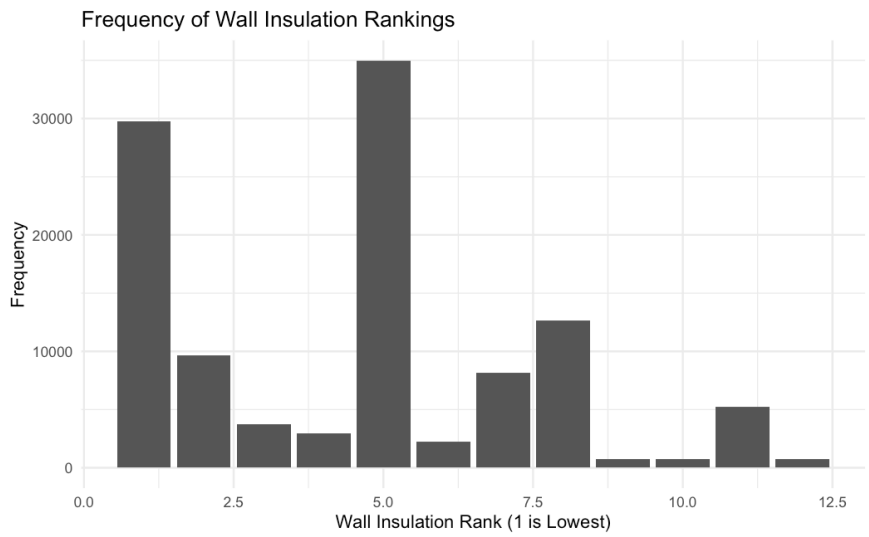
- The overall dataset contains 111,600 rows, where each hour, and day of July per house is a row
- My finalized dataset which was used for modeling was 111,600 rows and had 20 columns used in modeling and 6 columns used for description, such as latitude, longitude, county, building id, and hour.



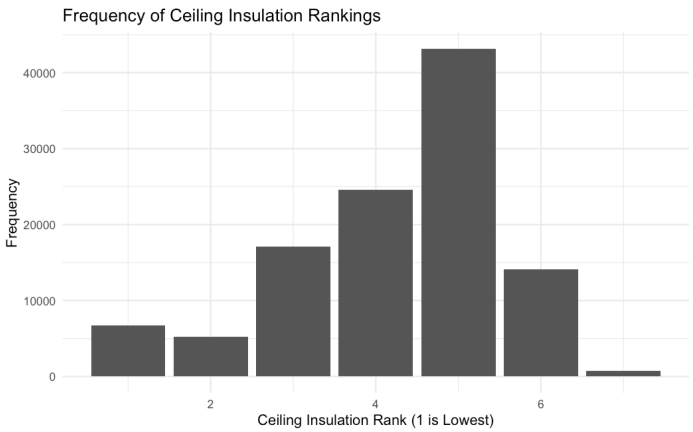


These two images show the average hourly electrical energy usage and average hourly temperature for a given day in July. We can see how there is a large degree of correlation between the two factors in terms of time.

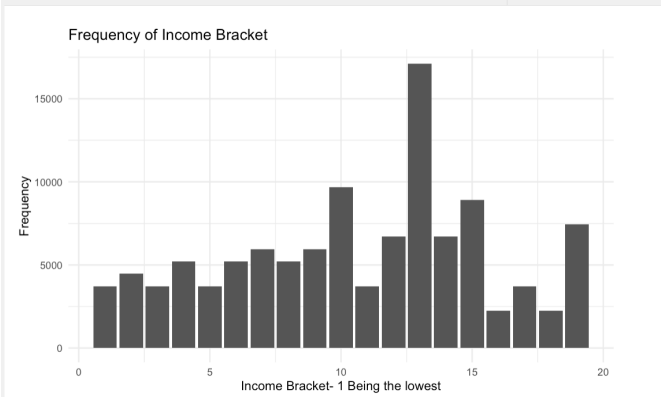
Intuitively this output would be expected as a majority of electrical systems “work harder” during the hot hours of the day, particularly AC units



The two above plots show two different things. The first shows the average temperature during July hourly, which is sure to be correlated with energy consumption. The second plot is the frequency of wall insulation type based on my factored rankings.



The above plot shows the frequency of ceiling insulation types according to my ranking



This final descriptive plot shows the frequency of house's income levels based on my ranking with The lowest being 1 and the highest being 19. The majority of incomes fit in the middle bracket.

Modeling Techniques Used

MARKDOWN PDF PAGES 12-22

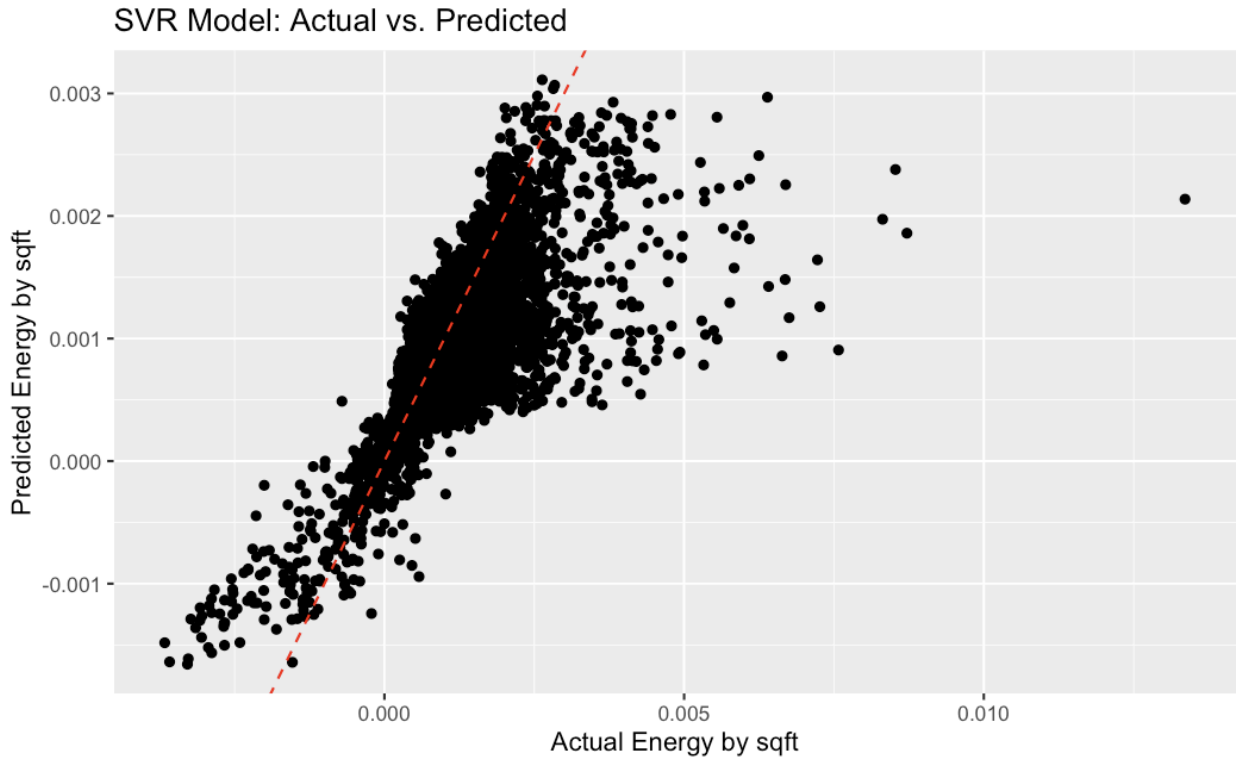
I utilized four different modeling technique for this business case:

1. Linear Regression
2. Treebag Modeling
3. Decision Tree
4. Support Vector Regression

I applied all of these models on our target variable of Energy per Square Footage

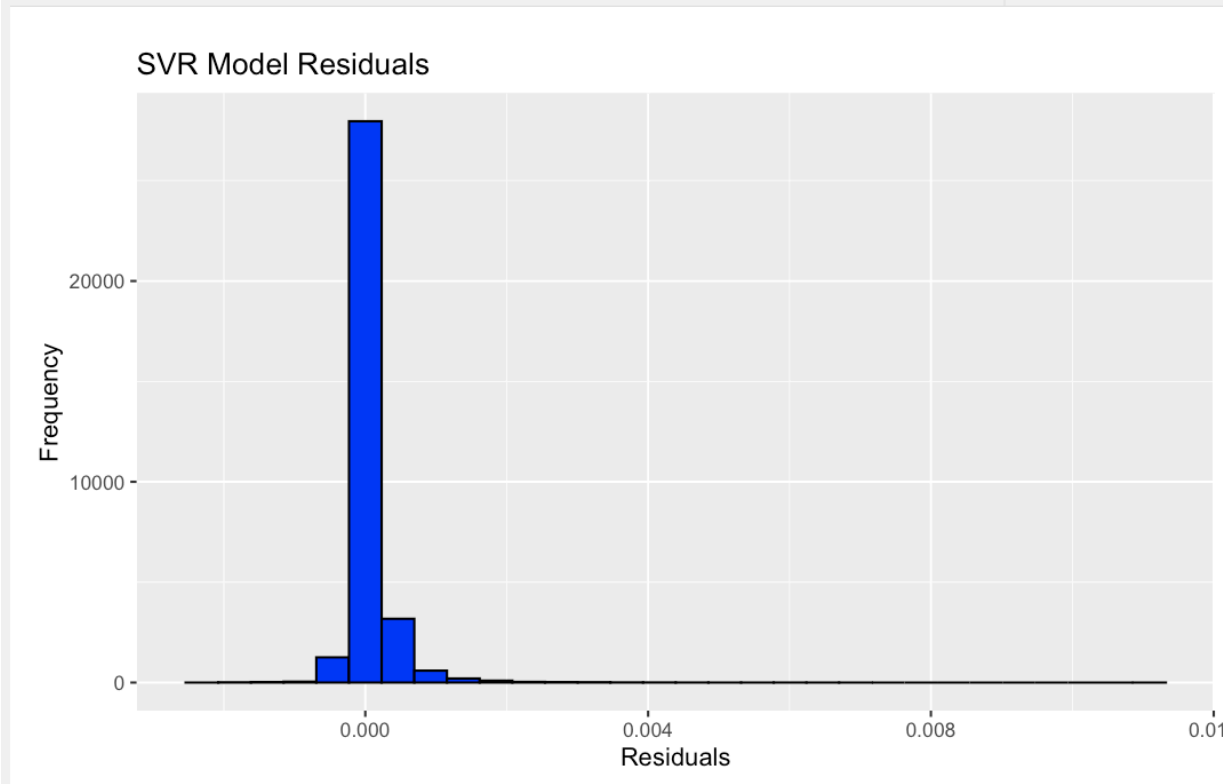
Results: Overall each model performed well when being evaluated on Root Mean Squared Error, all with very small decimal values in the results. This was a good sign overall and showed that there was some degree of significance between my predictor variables and my target variables.

Support Vector Regression (RMSE: 0.0003285604)



In the above SVR plot, we can see that there is a significant concentration of intercept between our predictions and actual values, with some outliers. The concentration above and below zero indicate that there is not a higher concentration above or below zero indicating an evely spread variability. And the outliers show that some predictors do not fit the model or deviation in the data in regards to the model.

Similar to above, this histogram shows a relatively symmetrical distribution around zero. The narowness of the plot indicates the model is making predictions with a low variability.



In this SVR plot, we can see that there is a significant concentration of intercept between our predictions and actual values, with some outliers. Note that my training data was 70% of all of my data. I chose to focus only on all of my predictors at once in this case to show the overall accuracy of the prediction.

Treebag Model (RMSE: 0.0004199181523087 69")

	Overall <dbl>
horizontal_rad	100.000000
in.sqft	98.125507
diffuse_rad	93.312939
rankedIncome	73.387956
temperature	69.873834
normal_rad	63.457668
window_ranking_integer	63.322072
floorInsulRank	60.224803
slabInsulRank	57.177620
rankedDucts	49.099139

	Overall <dbl>
in.bedrooms	22.072982
windowAreaRanking	16.008192
ceilingInsulRank	9.320015
wind_speed	6.797792
wallInsulRank	5.907963
rankedHVAC_Efficiency	0.000000

Here is my tree bag showing the relative importance of each predictor variable making predictions in the model. Our radiation and square footage have the highest increase in Mean Squared Error when omitted. On the lower end of the spectrum bedrooms and window area ratio are significant, but at a much smaller degree than our top predictors. Ranked HVAC efficiency has dud values in the dataset for both this and our linear regression model and is scientifically insignificant.

It will be interesting to compare the temperature's value in the increase in MSE when omitted when we explore the 5-degree increase in values.

Linear Regression Model (RMSE): 0.0004642069)

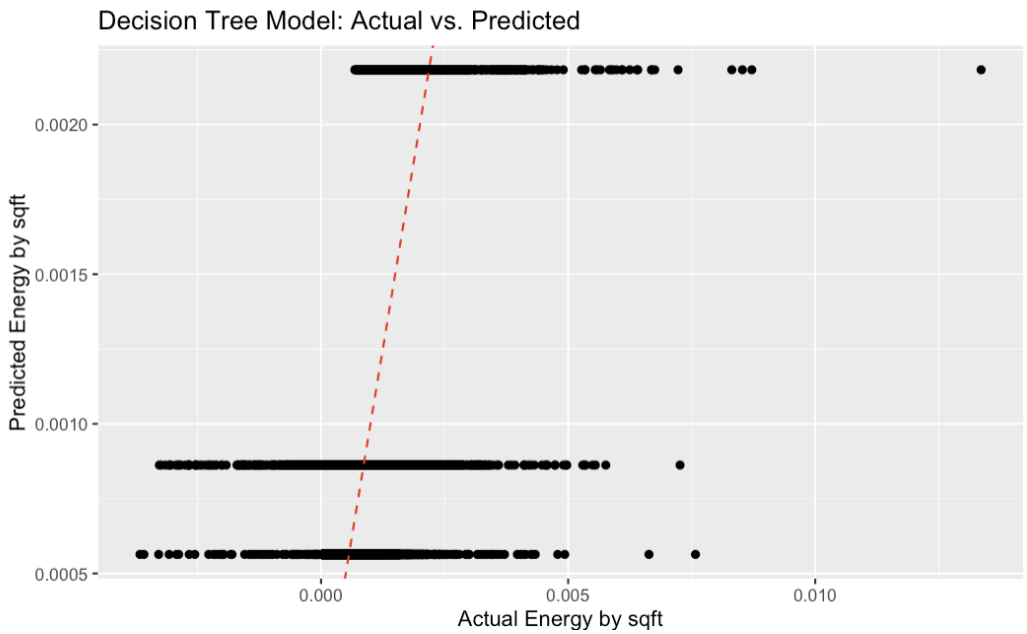
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.907e-04  2.167e-05 -18.033  < 2e-16 ***
temperature              5.268e-05  6.228e-07  84.585  < 2e-16 ***
wind_speed               1.486e-05  9.150e-07  16.236  < 2e-16 ***
horizontal_rad          -7.421e-07  1.993e-08 -37.239  < 2e-16 ***
normal_rad               5.064e-07  1.569e-08  32.275  < 2e-16 ***
diffuse_rad              7.722e-07  2.695e-08  28.653  < 2e-16 ***
window_ranking_integer -4.295e-06  5.451e-07  -7.880 3.31e-15 ***
windowAreaRanking      -5.936e-06  1.133e-06  -5.240 1.61e-07 ***
rankedIncome           -1.769e-06  3.419e-07  -5.175 2.29e-07 ***
in.bedrooms            -7.455e-05  1.924e-06 -38.741  < 2e-16 ***
rankedDucts            -5.877e-06  5.168e-07 -11.372  < 2e-16 ***
rankedHVAC_Efficiency  -7.465e-08  9.036e-07  -0.083  0.93417
usageLevel              1.412e-04  1.994e-06  70.800  < 2e-16 ***
wallInsulRank           1.815e-06  6.454e-07   2.813  0.00492 **
ceilingInsulRank       -4.906e-06  1.294e-06  -3.792  0.00015 ***
floorInsulRank          3.190e-05  2.099e-06  15.197  < 2e-16 ***
slabInsulRank           1.139e-05  1.991e-06   5.724 1.04e-08 ***
wallTypeRank           -1.282e-05  2.747e-06  -4.665 3.09e-06 ***
in.geometry_stories    -5.240e-05  2.826e-06 -18.543  < 2e-16 ***
in.sqft                -9.659e-08  1.322e-09 -73.067  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0004682 on 111580 degrees of freedom
Multiple R-squared:  0.2666,    Adjusted R-squared:  0.2665
F-statistic:  2135 on 19 and 111580 DF,  p-value: < 2.2e-16
```

Here is my linear regression model for energy per square feet given the predictor values. We can see that P values for each predictor coefficient are significant apart from ranked HVAC efficiency.

Intuitively HVAC efficiency should be an important predictor of energy consumption however it seems as though either my ranking and knowledge of heat pumps or data collection could result in this predictor being insignificant.

This model however certainly does provide important predictive knowledge, with 26.6% of the variance in electrical energy usage per foot being explained by our significant predictor variables.

Decision Tree Model (RMSE: 0.0004771165)



This scatterplot of prediction vs actual likely indicates that my datas categorical variables, sush as window type, area, and insulation type, do not have enough levels and distribute horizontally in lines. I stayed away from using this model as RPART decision tree may not fit my modeling dataset in an efficient or useful way. Despite this the RSME value is low so predictions were accurate in some cases, but the model evidently is over or underfit for our needs.

INSIGHTS AND CONCLUSION

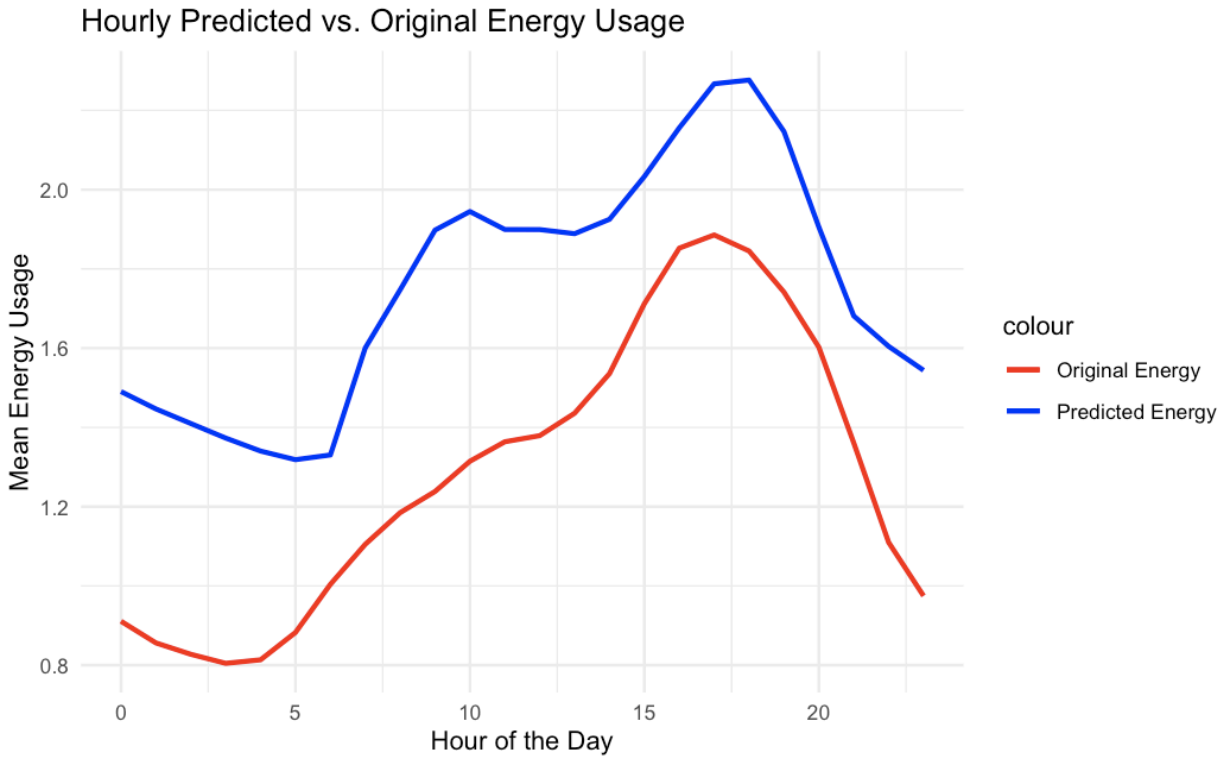
JUSTIFICATION FOR MODEL

- This prediction is based on my decision to use the svr model, which had the lowest RMSE, and an evenly distributed scatter of predictions meaning it was not over or under fitted. It also had low variance when observing the residuals in a histogram, where everything was symmetrically centered around 0 (for the most part)
- In addition plots for the DT indicated a bad fit between my data and the model. In terms of linear regression, the predictions were valid, but the SVM seems to perform better across the board. The treebag model was valid as well but is better used, in my case, for determining the importance of the individual predictors

INSIGHTS BASED ON PREDICTION

- During peak times (3 and 5pm) one can expect total hourly energy consumption to increase from an average of 1.8 kWh to an average of 2.3 kWh in the event of a 5 degree increase in temperature.
- This is based on my predictions being applied to a duplicate dataset where the temperature is increased by 5 degrees, and we recalculate total energy consumption by multiplying our predicted values by the houses square footage.

SEE MARKDOWN PAGE 22



RECOMMENDATIONS/CONCLUSION

- The main recommendation overall to ESC would be to allocate extra bandwidth to electrical energy between 7am and 6pm, where energy usage is above any normal current energy usage on average
- Additionally I would explore sun based radiation in the context of climate as its shifts hourly had strong correlations between energy usage per square feet
    - Our prediction based on a 5 degree increase presents its own inherent issue as sun radiation was not accounted for and temperature increased would likely be correlated to changes in sun based radiation
    - ESC would need to explore what other climate based ramifications a 5 degree temperature increase would have

- I would also pursue more research into income disparities when it comes to energy efficiency, as this problem has a correlation between overall energy consumption and income bracket
- The same goes for window, insulation, and duct quality, which could be prone to worse efficiency when installed in lower income areas at a higher rate
- I would also refine the analysis with construction materials and HVAC unit type as my analysis and ranking could be improved with the knowledge of how each of these is ranked in a hierarchical fashion

ALL code is referenced in the PDF of my markdown Code.