STAT 447B Group Report

Predicting Site Energy Consumption Using Climate Variables and Building Characteristics

By Anjali Chauhan, Sumit Meghlani, Idris Hedayat, Sameer Shankar

March 31, 2022

1. Summary

In this analysis, we sought to develop a model that predicts a site's energy consumption by studying the effects of climate variables and building characteristics on the Site EUI (Energy Usage Intensit)y to provide crucial information for potential optimized energy retrofitting. This helps in improving a building's asset performance (utilities). Retrofitting offers a potential upside in the overall performance of the building through improved energy efficiency, increased staff productivity, reduced maintenance costs, and better thermal comfort. By training an ensemble of prediction models like XGBoost, Gradient Boosting Machine, Light GBM, and Support Vector Machine with important variables selected through variable importance method using Random Forest Regressor and Correlation Analysis and predicting the Site EUI, we achieved a high prediction performance with the lowest Root Mean Square Error (RMSE) of 0.1878 along with Interval Score of -, Average Interval Length of ___ and a covergae of ___ for the ______ Model.

2. Introduction

Climate change is an urgent, and multi-faceted issue heavily impacted by infrastructure. Addressing climate change involves mitigation of Greenhouse Gas (GHG) emissions via changes to electricity systems, transportation, buildings, industry, and land use. According to a report issued by the International Energy Agency, the life cycle of buildings from construction to demolition were responsible for 37% of global energy-related CO_2 emissions in 2020. Yet it is possible to drastically reduce the energy consumption of buildings. For example, retrofitted buildings can reduce heating and cooling energy requirements by 50-90 percent. Therefore, it is important to optimize energy retrofitting by accurately predicting Site EUI of these buildings. The study aims to investigate and understand the effect of climate variables and building characteristics on the building's EUI (Energy Usage Intensity).

The analysis aims to address the following questions:

- Does the effect of building characteristics outweighs the effect of climate variables on Site EUI and vice-versa?
- Is there a need to build separate models for different Facility types or Building classes?

More specifically, the analysis has the following objectives to answer the above questions:

- To find important climate variable(s) and/or building characteristics variable(s) and determine the effect of said variable(s) on the site EUI
- To translate the relationship between the response and multiple predictors into insightful visualizations
- To model and predict the site EUI values

This report summarizes all of the primary statistical modelling and analysis results associated with the study. The remainder of this report is organised as follows: Section 3 describes the data collection, provides measurement of the variables and summarizes the data. Section 4 presents the data preprocessing and statistical modelling techniques used to answer the aforementioned research questions. Section 5 summarizes and interprets the results of the statistical analysis conducted. Appendices are provided for further exploratory data analysis along with the code used for the statistical modelling. Lastly, Section 7 presents the limitations and challenges in conducting this analysis.

3. Data

3.1 Description

The data was collected in collaboration with Climate Change AI (CCAI) and Lawrence Berkeley National Laboratory (Berkeley Lab). Data contains roughly 100,000 observations of building energy usage records collected over 7 years and from a number of states within the United States. The dataset consists of building characteristics (e.g. floor area, facility type, etc), and weather data for the building location (e.g. annual average temperature, annual total precipitation, etc), as well as the energy usage for the building (Site EUI). Each row in the data corresponds to the a single building observed in a given year. There are 75757 rows and 64 columns with 64448 rows containing null values.

Table 1: Description of Building Characteristics Variables Used for Analysis

	Variable	**Unit**	**Description**
1.	***(Response)*** Site EUI	$kBtu\ ft^{-2}$	The amount of heat and electricity consumed by a building as reflected in utility bills
2.	Energy Star Rating	-	It is a measure with a score between 1-100 where a higher energy rating means that the building performs better
3.	Floor Area	ft^2	Floor area of the building
4.	Year Built	-	Anonymized year in which the weather and energy usage factors were observed
5.	Elevation	ft	Elevation of the building location
6.	Facility Type	-	Building usage type
7.	Building Class	-	Building classification (residential/commercial)
8.	State Factor	-	Anonymized state in which the building is located
9.	Year Factor	-	Anonymized year in which the weather and energy usage factors were observed

Table 2: Description of Climate Variables Used for Analysis

	Variable	**Unit**	**Description**
1.	[month]'s Minimum Temperature	$°F$	Minimum temperature in [month] at the building location where month from January to December **(12 variables)**
2.	[month]'s Average Temperature	$°F$	Average temperature in [month] at the building location where month from January to December **(12 variables)**
3.	[month]'s Maximum Temperature	$°F$	Maximum temperature in [month] at the building location where month from January to December **(12 variables)**
4.	Cooling Days	-	The number of degrees where the daily average temperature exceeds 65 $°F$

	Variable	Unit	Description
5.	Heating Days	-	The number of degrees where the daily average temperature falls under 65 $^{\circ}F$
6.	Precipitation	$inches$	Annual precipitation at the building location
7.	Snowfall	$inches$	Annual snowfall at the building location
8.	Average Temperature	$^{\circ}F$	Average temperature over a year at the building location
9.	Days Below 30F	-	Total number of days below 30 $^{\circ}F$ at the building location
10.	Days Above 80F	-	Total number of days above 80 $^{\circ}F$ at the building location
11.	Direction of Max Wind Speed	$^{\circ}$	Wind direction for maximum wind speed at the building location
12.	Direction of Peak Wind Speed	$^{\circ}$	Wind direction for peak wind gust speed at the building location
13.	Max Wind Speed	ms^{-1}	Maximum wind speed at the building location
14.	Days With Fog	-	Number of days with fog at the building location

Table 3: Summary Statistics of All Climate Variables

Var	site_eui	energy_star_rating	floor_area	year_built	ELEVATION	Year_Factor
Count	75757	49048	75757	73920	75757	75757
Std	58.26	28.66	246875.8	37.05	60.66	1.47
Min	1.00	0.00	943	0.00	-6.40	1.00
25%	54.53	40.00	62379	1927.00	11.90	3.00
50%	75.29	67.00	91367	1951.00	25.00	5.00
Mean	82.58	61.05	165984	1952.31	39.51	4.37
75%	97.28	85.00	166000	1977.00	42.70	6.00
Max	997.87	100.00	6385382	2015.00	1924.50	6.00

Due to a large number climate variables, summary statistics hasn't been provided in the report. Please refer to the code.

3.2 Exploratory Data Analysis

In our Exploratory Data Analysis, we aimed to find how individual explanatory individuals relate and behave alongside the response variable Site EUI as well other explanatory variables. Through wrangling of the data aim to suggest potential transformations to the data in hopes of finding behaviors that will improve performances of future models.

#Investigating Relationships with the Response

#Spearman Correlation

In a preliminary attempt of gathering information on covariate relationships with Site EUI, we used Spearman rank correlation coefficients to investigate potential linear relationships with the response. In particular we choose Spearman correlation which is invariant to monotone increasing effect of transforms. We find that energy star rating has by far the largest Spearman correlation with Site EUI in terms of magnitude with -0.654 [REPLACE WITH VALUE FROM SUBSET ??], suggesting a relatively strong decreasing relationship with the response. We also note relatively higher magnitudes for monthly min and average temperatures, in particular a decreasing relationship with colder months temperature measures, namely January and February, while there are increasing relationships with Summer months auch as June and July. In context this can be expected as there is a lesser requirement of energy for uses such as heating and lighting in building in the summer months when days last longer and are warmer.

#Plots Investigating relationships with Site EUI

In order to enable us to explore the data more effectively there are measures that can be taken such as subsetting the data, given the large size of the data set. Via the summary tables we also see that site EUI is right skewed, so a cube root transformation for the sole purpose of investigating relationship will be more effective.

#building charactersitics vs site eui

building characteristics are of particular interest in context of the data, and so we investigated relationships between Site EUI and building variables. The most noteworthy relationships was with energy star rating, which was as expected from the Spearman correlations, and found a clearer and more linear relationship with the response than almost all other variables.

For elevation we used binning for the classes based on the variables quantiles from summary statistics. We see a notable increase in site EUI initially as buildings add a floors requirement more total energy usage, before this plateaus, which is understandable in context as buildings aren't generally built above a certain height and number of floors.

For state factor we see that the classes are very imbalanced for Stat 10 with only 15 datapoints, while stat 6 has much more than the others with 50840, indicating we could undersample for this specific state factor, while combining stat 10 with another state factor level.

#Investigating Relationships between explanatory variables

It is in our interest to explore potential relationships between covariates in the dataset, as these could potentially lead to issues namely multicollinearity. The monthly minimum, maximum, and avaerge temperatures for colder months display relatively strong relationships with other climatic variables relating to cooler weather including "days below…", "heating degree days" and "snowfall" covariates. We gather that the strong negative relationship in context shows as temperature increases in these cold months, the amount of snow fall and days below 0F and 20F decrease as expected. Tther noteworthy relationship

Fig. 1: Relationship between Climate Variables and Streamflow

Looking more closely at various variables broken out by `gridcode` (*see Fig. 5 as an example*), some interesting results were seen. For many `gridcode`'s, there appears to be a much stronger linear relationship between climate variables and the stream flow and these relationships appear to have different intercept and slope values. This suggests that there may not be a one-size-fits-all approach to fitting a regression model based on annual climate variables alone and different slopes for different `gridcode`'s may need to be considered for further analysis in the future.

4. Methods

4.0. Pipeline

Below in Fig. 6, we have a Proof-of-Concept pipeline that addresses all of the client's research questions. A breakdown of each of the steps shown in the end-to-end workflow diagram is covered below.

Fig. 6: End-to-End Pipeline

4.1 Data Pre-Processing

As part of the data pre-processing pipeline, we start by dropping duplicate rows, and columns with a high percentage of missing values (over 50%). With the remaining missing values, we used the median imputation method (filling missing values of the column with the median). Furthermore, we dropped highly correlated features using pairwise correlation analysis, and binned the facility type column (due to unbalanced classes). Lastly, we removed rows with extreme Site EUI values using the IQR method and we used an ordinal encoding to encode all categorical columns.

4.2 Feature Selection

One of the most important step in our pipeline was Feature Selection as it is one of the main objectives. We have used several different methods in order to achieve this objective:

4.2.1. Pairwise Correlation Analysis Before using any of the traditional feature selection techniques mentioned below, we investigated if any of the features were highly correlated. We see from the Correlation Matrix below that there is high correlation for monthly temperature averages particularly those around the sames seasons suggesting future implementation of new features in the place of seperate months. We also see high correlation for other climatic features such as days above and below variable, as well as high correlation for snowfall and precipitation inches, also indicating potential interactions or new replacement covariates in place of these terms

```
<!--  -->
<!-- knitr::include_graphics('image/corr.png') -->
<!-- # plt.figure(figsize=(16, 6)) -->
<!-- # # define the mask to set the values in the upper triangle to True -->
<!-- # mask = np.triu(np.ones_like(df.corr(), dtype=np.bool)) -->
<!-- # heatmap = sns.heatmap(df.corr(), mask=mask, vmin=-1, vmax=1, annot=False, cmap='BrBG') -->
<!-- # heatmap.set_title('Triangle Correlation Heatmap', fontdict={'fontsize':10}, pad=16) -->

<!-- \begin{center} -->
<!-- \textbf{Fig. 9}: Heatmap Presenting Correlation Between Variables -->
<!-- \end{center} -->
```

4.2.2. Variable Importance using boruta package This method is built on a random forest classifier. It ranks features based on their importance measure i.e. Mean Decrease Accuracy (MDA) where higher means more important. MDA measures how much accuracy the model losses by excluding each variable. The more the accuracy degrades, the more important the variable is. From the variable importance plot we see the 3 most importance variables by far are energy star rating floor area and year built. We proposed the notion of potential interactions with these terms. [NOT SURE EXACTLY ON REASONING FOR THE INTERACTION LATER ??]. We also note particular importance of Facility Type, Elevation, Days below 20F, Ferbruary average temperature, January average temperature, as well as building class, relative to the other features in the dataset.

Fig 12: Variable Importance (using 'boruta' package)

4.2.3. Interactions and Additional features Based on the variable importance and pairwise correlations we decided on adding interaction terms for future models. We decided on including interactions of; - floorarea and year built - floor area and energy star rating - all 3; floor area, year built, and energy star rating - floor area and sum of Cooling Heating degree Days

We also implemented new features containing other features that were otherwise correlated with one another, notably monthyl temperature variabels. Thus we decided to introduce seasonal terms, where Spring took the average of March April May temperatures, Summer took average of Juny July August, Fall took average of September October November, and Winter took December January and February.

We also introduced terms for "days below.." and "days above.." temperature variables; - Freezing days: total days below 0 10 F - Cold days: total days below 30 and 20 F - Warm days: total days above 80 and 90 F - Hot days: days above 100 and 110 F

Along with a feature covering both snowfall and precipitation: - Snow Rain inches: sum total inches of snowfall and precipitation

In adding these features we removed the originals contained within these, in turn this would help leave us with more parsimonious models to interpret.

4.2.4 Feature Transformations

In addition to Z-Score Standardization, which refers to scaling and centering of the distribution to ensure the data is not on a varying scale and is internally consistent, several feature transformations were performed to deal with both left and right skewed feature distributions. Table 3 shows all the transformations performed in order to achieve the desired results which are highlighted by Fig. 7 and Fig. 8.

Table 3: Feature Transformations

Variable	*Description*	*Transformation*	*Skewness*
floor_area	Floor Area	Fifth Root Transform	Right Skewed
year_buily	Year Built	Shifted (1200 - year built) & Squared Transform	Left Skewed
energy_star_Rating	Energy Star Rating	Squared Transform	Left Skewed
ELEVATION	Building Elevation	Sixth Root Transform	Right Skewed
avg_temp	Average Temperature	None	None Obvious
SpringTemp	Average Temperatures of March April May	None	None Obvious
SummerTemp	Average Temperatures of Juny July August	None	None Obvious
FallTemp	Average Temperatures of September October November	None	None Obvious
WinterTemp	Average Temperatures of December January February	None	None Obvious
floorxBuilt	Interaction Term Floor Area and Year Built	Sixth Root Transform	Left Skewed
floorxEnergy	Interaction Term Floor Area and Energy Star Rating	Sixth Root Transform	Left Skewed
floorxBuilt	Interaction Term Floor Area and Year Built	Sixth Root Transform	Right Skewed
floorxHeatCool	Interaction Term Floor Area and Sum of Heating and Cooling Degree Days	Sixth Root Transform	Right Skewed
freezing_days	Total days below 0 and 10 F	Fifth Root Transform	Slight Right Skewed
cold_days	Total days below 20 and 30 F	square Root Transform	Slight Right Skewed
warm_days	Total days above 80 90 F	Square Root Transform	Left Skewed
hot_days	Total days above 100 110 F	Dropped	Too Few Data

4.3. Model Training and Validation

A good variety of models were implemented as a part of our analysis for extensive results. Two sets of models where trained to capture both the effect of individual co-variate terms and the interaction terms on the predictive performance of the model. These models were cross-validated (10-fold, Repeated CV) and their hyper parameters were fine tuned using Random Search. Please refer to Table 3 and Table 4 for results that denote the predictive performance of the models.

4.3.1. Linear Models

After feature selection, we trained a Linear Regression model with a 10-fold cross validation to test the model performance on the training data. We train a similar set of models with interaction terms as features and we see slightly better results than the former method.

4.3.2. Support Vector Machines

To add further complexities to the previous models, we trained Support Vector Machines that expanded our feature space using different kernels. We have a radial kernel to compare performance of models without the complexities relating to Linearity respectively. In the radial kernel, only the neighboring behaviour of data is taken into account which means only those data points influence the modelling compared to the Linear SVM whose performance is similar to a Linear model. From Table 3 and Table 4, we see that the Radial SVM without interaction terms performs slightly better than Radial SVM with the interaction terms.

4.3.3. Tree Models

We used three tree-based models such as Random Forest, XGBoost, Quantile Regression Forest (QRF) and Gradient Boosting Machine to improve the performance compared to the above models. From Table 3 and Table 4, we see that the Random Forest, QRF and XGBoost with interaction terms has better results compared to without interaction terms. However, the GBM without interaction terms gives slightly better results than the GBM with interaction terms. The best tree model is the Quantile Regression Forest Model with the lowest RMSE of 0.227 (with interaction terms).

4.3.4. Ensemble Models

We trained two Ensemble models with/without the interaction terms that combine the above listed 6 models to produce improved results. These models generally produce more accurate predictions than a single model. From Table 3 and Table 4, we see that that the Ensemble Model with interaction terms has significantly better results compared to the Ensemble Model without interaction terms. Therefore, based on the validation evaluation metrics from Table 3 and 4, we chose the Ensemble Model with the interaction terms as our best model.

Fig. 17: Comparing Prediction Performance for Different Models Without and With Interaction Terms

5. Model Results

From Table 3, we see that the ensemble model have the lowest RMSE , meaning it has the best predictive performance out of all the models with no interaction terms.

Table 4: Comparing Prediction Performance of Different Models Without the Interaction Terms

Models	RMSE
Linear Model	0.203
Quantile Regression Forest	0.256
Random Forest	0.255
XGBoost Linear	0.255
Radial Support Vector Machine	0.218
Gradient Boosting Machine	0.262
Ensemble Model	**0.1986**

From Table 4, we see that once again that the ensemble of all the 6 models listed has the lowest RMSE, meaning it has the best predictive performance out of all the individual models with interaction terms.

Table 5: Comparing Prediction Performance of Different Models With the Interaction Terms

Models	RMSE
Linear Model	0.196
Quantile Regression Forest	0.227
Random Forest	0.234
XGBoost Linear	0.242
Radial Support Vector Machine	0.267
Gradient Boosting Machine	0.274
Ensemble Model	**0.1878**

From Table 5, we use our best performed models on test dataset to evaluate whether our models are still valid when applying on test dataset, we can see that the RMSE from both models are quqite consistent with their result in Training dataset in Table 3 and 4.

Table 6: Prediction Performance of Best Models on Testing Dataset

Models	RMSE
Ensemble Model Without Interaction Terms	0.192664
Ensemble Model With Interaction Terms	0.2042341

The two plots below show how much the prediction from our best 2 models deviated from actual value therefore giving us a rough estimation of whether a model is a good fit or not. We clearly see that both models'prediction are quite near the actual values, most points were near to the fitted line indicating a good fit.

Comparing the predictive performance of the best models from both Table 3 and Table 4, we see that the latter (Ensemble Model, RMSE: 0.1795) has a better performance. Therefore, the best model we chose for making predictions was the Ensemble Model with the interaction terms.

6. Outlier Detection

We examined the stream flow values in our dataset by using the interquartile range (IQR) method, classifying any stream flow values more than 1.5 times the IQR below the first quartile or above the third quartile as an outlier. Looking at Fig. 14, there are 16 extreme high points that are outliers but no extreme low points. Grouping stream flow values by gridcodes is looking at each watershed separately. The average stream flow values vary from watershed to watershed. The average stream flow values for watershed at grid code 14 is higher than the rest of the watersheds. On the other hand, the average stream flow values for watershed at grid code 1264 is lower than the rest of the locations. Gridcodes 1212 and 1391 have the most outliers and it is worth looking into and learning more about those areas.

Fig. 20: Streamflow Outliers vs Gridcode

7. Limitations

- After feature selection, there is some potential that the best features for each model type were not selected. When selecting the features that would be included in the hyper-parameter tuning for the models we used the results of both Boruta (random forest based selection method) and Forward Stepwise Regression (regression based selection method). The selected features from each both agreed with each other so they seemed reliable, however we did not do an exhaustive search over all variables for each of the models and it is possible there were better combinations.

- Gridcode being selected as an important feature may lead to poor model performance due to lack of data to properly fit especially in the 'with interaction' case. There are 23 separate gridcodes, so there are only 20-40 observations for each gridcode which is not very much (especially for tree based models). Having access to more data could result in much higher performance.

- Each observation in the dataset is an aggregation of climate data collected over the year (i.e. they are the average values collected over the year). This limits the forecasting power of the predictive models that we have fit as we would need to use the forecasted explanatory variables to predict the stream flow, which will most likely lower the performance of the model.

8. Conclusion

After all the model iterations and improvements, we were able to achieve fairly good results with the Ensemble Model taking into account the interaction effect between variables. These results have large implications when it comes to water resource management economically. We have conducted our primary analysis taking into account the spatial data (e.g. gridcodes) which serves as a good MVP to predict the streamflow.

As a side interest and to expand upon the idea of predicting the streamflow solely using climate variables and not any spatial and temporal data, we conducted an analysis and trained models without this data and the predictive performance dropped significantly. Although this addresses the client's first research question of whether one catchment can be used to extrapolate stream flow in another catchment, the limitation we faced was lack of training data. To be able to model such a complex problem using climatic variables, we need more data to train our model which can help with improving the predictive performance of the model.

To address the second research question of whether or not we can detect the unusual streamflow activity accurately, we built a proof of concept pipeline for the outlier detection system. It will take the features from our prediction model as input and label the observations as either anomalous or regular depending on the anomaly scores which are the measures of deviation from normal behavior. We will face the same challenge here - lack of training data. This will lead to an increase in false positives and false negatives in the outlier detection system which can have detrimental consequences. For example, not being able to detect a subtle increase in the streamflow (false negatives) which could lead to irrigation problems and in severe cases even floods. Or getting a huge pool of outlier values (false positives) that will raise false alarms of anomalous behavior more often than desired.

9. Future Research

There are two areas of further research that would help address the limitations in this study; first is further investigating outliers and anomaly detection and second is training a model on a more granular time frame.

- The outlier detection that we have done is appropriate for identifying outliers in the current dataset but does not have a real-world application. Training a classification model to predict extreme values in streamflow and/or training an anomaly detection model to find unusual patterns in streamflow and the climate variables. Both models could have more real-world application in flood/drought prevention which is useful in fields such as agriculture.

- Additionally, requiring input data for the whole year for the model makes it impractical to accurately predict streamflow values in the future as we would need to use forecasted values for the input variables. Having the data be on a more granular time frame would greatly benefit the analysis and real-world predictive power of the model as we could train the predictive streamflow model only using past variables. The resulting model would not rely on using forecasted input variables, which addresses a critical limitation of this study.

10. References

- Government of Canada / Gouvernement du Canada. (2021, November 25). Government of Canada / gouvernement du Canada. Climate. Retrieved February 5, 2022, from https://climate.weather.gc.ca/glossary_e.html
- US Department of Commerce, N. O. A. A. (2012, March 8). Snow measurement guidelines. Snow Measurement Guidelines. Retrieved February 5, 2022, from https://www.weather.gov/gsp/snow
- Janssen, J., & Ameli, A. A. (2021). A Hydrologic Functional Approach for Improving Large-Sample Hydrology Performance in Poorly Gauged Regions. Water Resources Research, 57(9), e2021WR030263.
- Statistical interaction: More than the sum of its parts. Statistics Solutions. (2021, June 22). Retrieved February 21, 2022, from https://www.statisticssolutions.com/statistical-interaction-more-than- the-sum-of-its-parts/

11. Appendix