Harsh Giria, Sofia Fiorillo, Yash Kothari, Anusha Hegde, Joshua Dierker
Arya Farahi
12/10/2021

Final Report:
Modeling Energy Efficiency

I. Introduction

Per the World Economic Forum, buildings are responsible for 40% of global energy consumption and 33% of greenhouse gas emissions**[B]**. Tsanas and Xifara's paper "suggests that building energy consumption has steadily increased over the past decades worldwide, and heating, ventilation and air conditioning (HVAC), which have a catalytic role in regulating the indoor climate, account for most of the energy use in the buildings"**[A]**. Clearly, we should strive for energy-efficient building design in order to combat climate change. By determining which building parameters have the most influence on heating and cooling load, we can design buildings to be more energy-efficient. In this project, our goal was to assess the heating and cooling load of buildings as a function of various building parameters. To accomplish this, we conducted a regression analysis to predict the heating and cooling of buildings in order to identify which building parameters influence them the most.

II. Data

The dataset was generated from simulations run using the Ecotect software. The simulations used 12 different Ecotect building shapes under various parameter combinations to generate 768 buildings. The dataset has 8 features, which are the building parameters that we will use to predict the response variables, which are the heating and cooling loads of the buildings. These features include Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distribution. Relative Compactness is derived by comparing the volume to surface ratio of the shape to that of the most compact shape with the same volume. Orientation is the cardinal direction that the front of the building is facing. Glazing area is the proportion of the building's Surface Area occupied by glass windows; and Glazing Area Distribution is the Portion of the building that has glass windows (glazing). In terms of cleaning the dataset, our work was cut out for us since it was generated from a software simulation. There was already one observation per cell and one column per variable. We dropped the empty columns from the dataset and renamed the remaining columns from their original names to their respective building parameters for better

readability. Lastly, we encoded the nominal categorical variables such as Orientation as indicator variables. We did not take any steps to normalize the data because no models that rely on distance were used, so normalization would have no impact on the regression analysis.

III. Exploratory Analysis

To visualize our data we created violin plots to show the distribution of the explanatory variables when compared to the heating (See Fig1.) and cooling load (Fig2.) since it not only gives us a boxplot, but also shows us the density of the different data points.

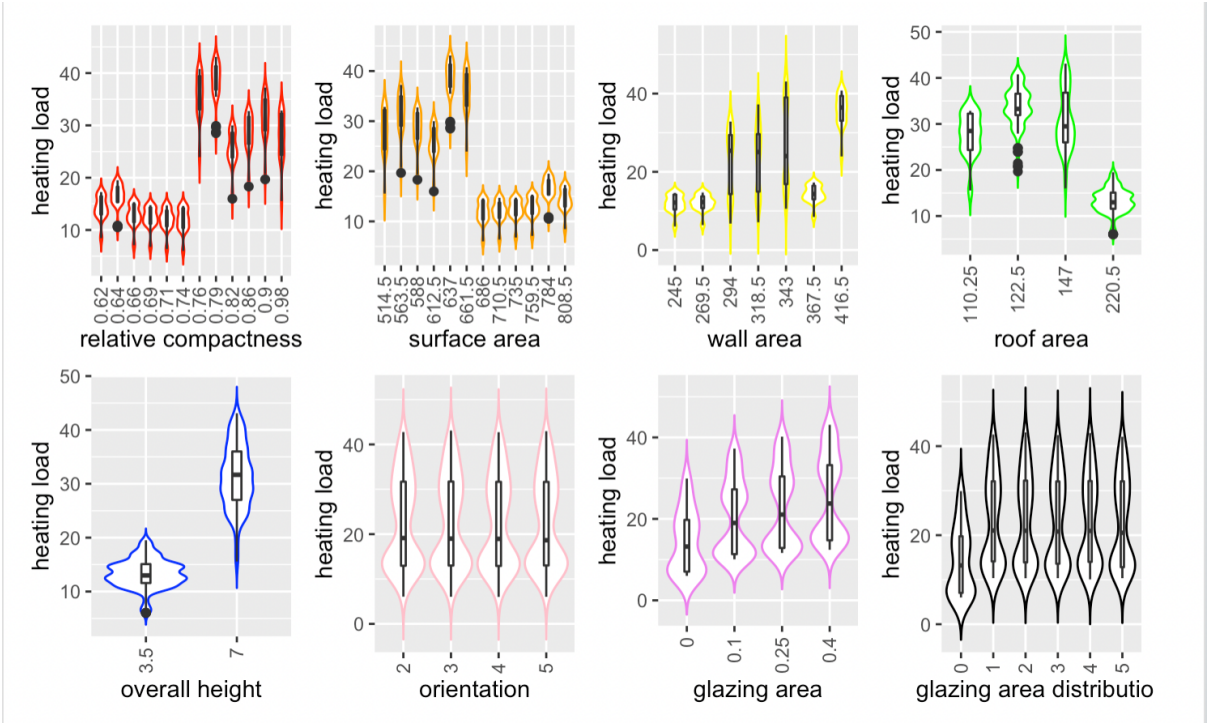

Fig. 1: Heating Load per X variable.

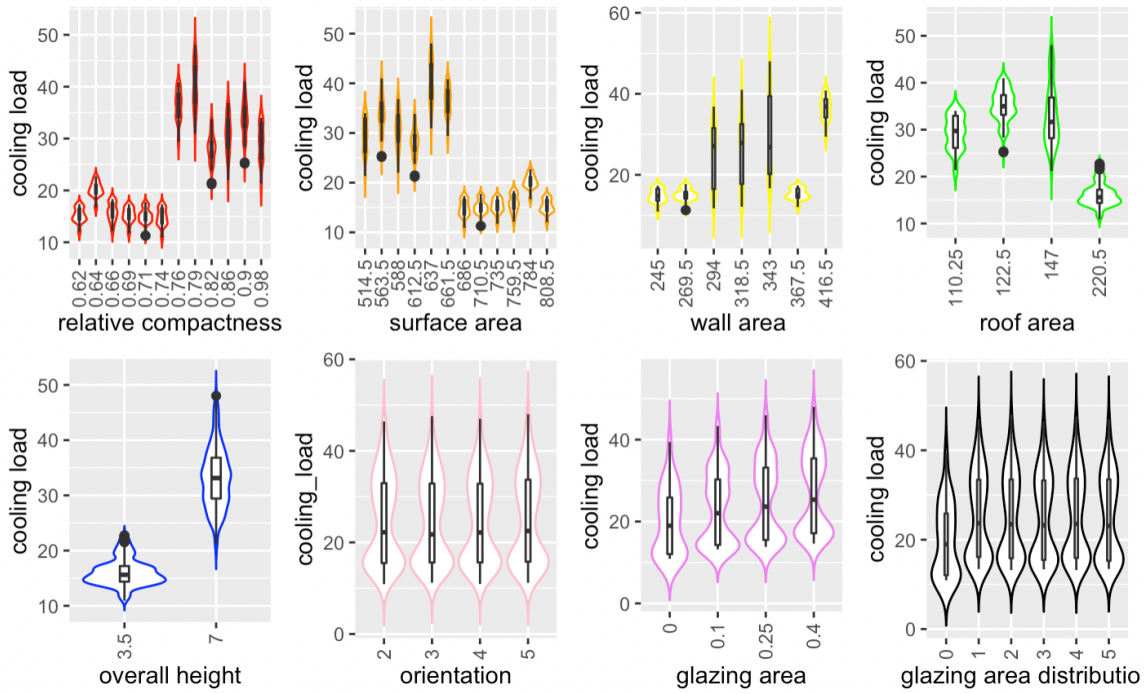

Fig. 2: Cooling Load per X variable.

From the graphs we can observe that the heating load seems to follow a step function when analyzed against relative compactness and surface area. For relative compactness there seems to be a threshold (of around 0.74), that when surpassed, triggers the heating load to increase. For the surface area, we can infer that when a threshold of around 662 is surpassed, the heating load diminishes.

We also observe that overall height seems to have a great impact on the heating load. A higher height seems to be related to a high heating load. However, since there are only two levels for the overall height, it is difficult to interpret any real correlation. Orientation, glazing area and glazing area distribution seem to have little or no effect on the heating load. Finally, we infer that the same conclusions can be applied to the cooling load as its graphs are the same as the heating load.

To explore this relationship further, we remove the categorical variables- the orientation and the glazing area distribution. Then, we compute the correlation matrix for the variables (See Fig3.).

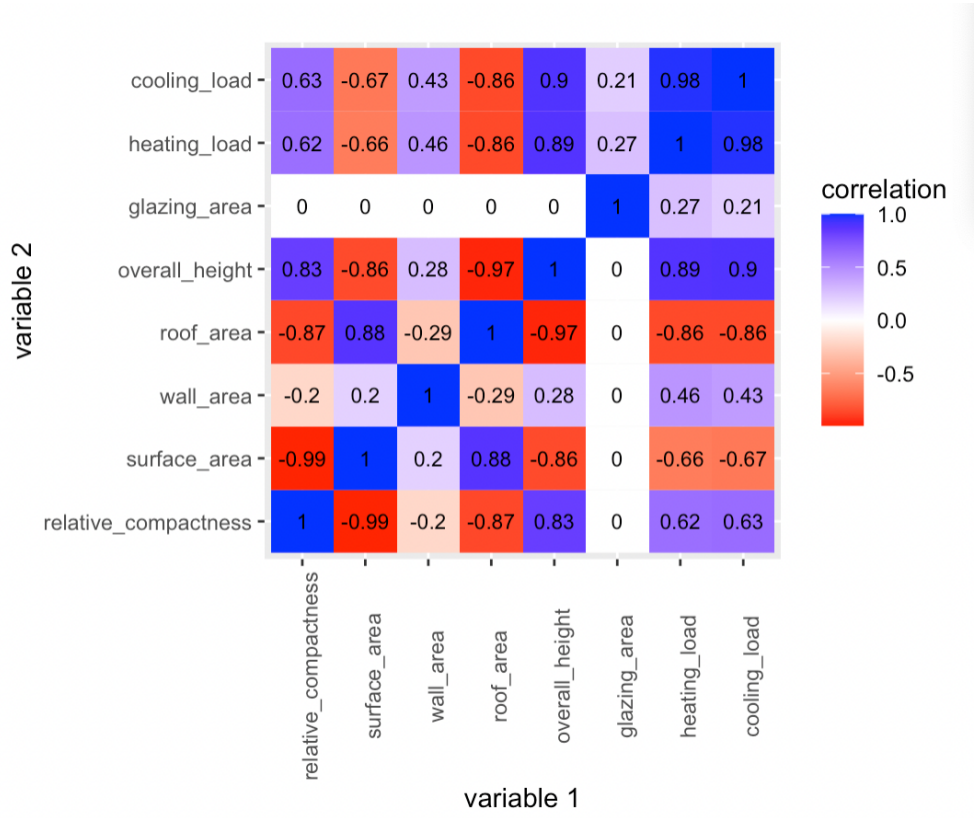

Fig. 3: Correlation Matrix

We observe that the variables that seem to be highly positively correlated to heating and cooling load are overall height, and relative compactness. On the other hand, roof area and surface area seem to be highly negatively correlated to heating and cooling load. This leads us to our first hypothesis- **Relative compactness, surface area, roof area, and overall height appear to be important for the prediction of heating and cooling loads.**

Upon further inspecting the graphs we come up with our second hypothesis, **The underlying relationship between the explanatory and the response variables does not appear to be linear.** This indicates that any functional relationship of the explanatory variables and the response variables is not trivial. This also implies that classical models such as linear regression may fail to find an accurate mapping of the explanatory variables to the response variables. To find further evidence for our hypothesis, we create a predictive model.

IV. Modeling

After the data was cleaned and exported to a tidy file, it was imported into a python environment. 3 data frames were created: one containing all the explanatory variables, and one with each response variable (heating load and cooling load).

Using the train_test_split library from sklearn the data was divided into training and testing samples, while the training data was used to fit the different models, the testing data was used to understand the accuracy of the predictions obtained with the models and ensure there was no overfitting.

First we fitted the linear model (one for each response variable), calculated the testing MSE and obtained the regression coefficients. The same was done afterwards for the SVR and Random Forests regression. After selecting the best model out of three analyzed, a SHAP analysis was performed to be able to interpret the results and understand the relationship between the explanatory and the heating and cooling load.

V. Discussion

The testing MSE obtained with the linear model was of 9.3 for the heating load and 9.8 for the cooling load. The values seemed relatively low, and when plotting the predicted vs actual values (see Fig. 4) a diagonal line could be visualized, which could initially indicate that the linear regression is a good fit to our data. However, the coefficients of the model were also studied and it was discovered that they had unreasonably high orders of magnitude (Fig. 5), making it obvious that the linear regression model does not capture the relationship between the explanatory and response variables.

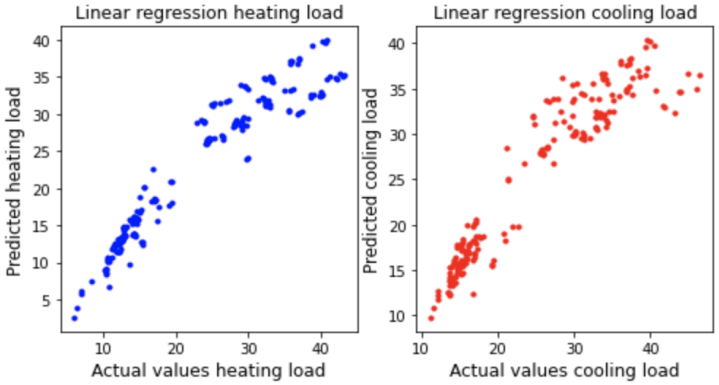

Fig. 4: Actual vs. predicted values using the linear regression model

	Variables	Heating load coefficients	Cooling load coefficients
0	relative compactness	-5.884656e+01	-6.496461e+01
1	surface area	-1.379962e+09	-4.794439e+11
2	wall area	1.379962e+09	4.794439e+11
3	roof area	2.759924e+09	9.588879e+11
4	overall height	3.864760e+00	4.212142e+00
5	glazing area	1.621128e+01	1.323144e+01
6	glazing area distribution1	4.960494e+00	2.164747e+00
7	glazing area distribution2	4.656717e+00	1.775645e+00
8	glazing area distribution3	4.558893e+00	1.590791e+00
9	glazing area distribution4	4.699235e+00	2.047108e+00
10	glazing area distribution5	4.314578e+00	1.441225e+00
11	orientation2	-3.422383e+05	-1.189048e+08
12	orientation3	-3.422382e+05	-1.189048e+08
13	orientation4	-3.422383e+05	-1.189048e+08
14	orientation5	-3.422385e+05	-1.189048e+08

Fig. 5: Coefficients for each of the variables

Later on an SVR regression model was fitted to the training data. However, this resulted in a testing MSE of 40.2 and 32.7 for the heating and cooling load respectively, and there is an obvious pattern when plotting the actual vs predicted values for the response variables (Fig. 6). We can therefore conclude that the SVR regression model is not a good fit to our data.

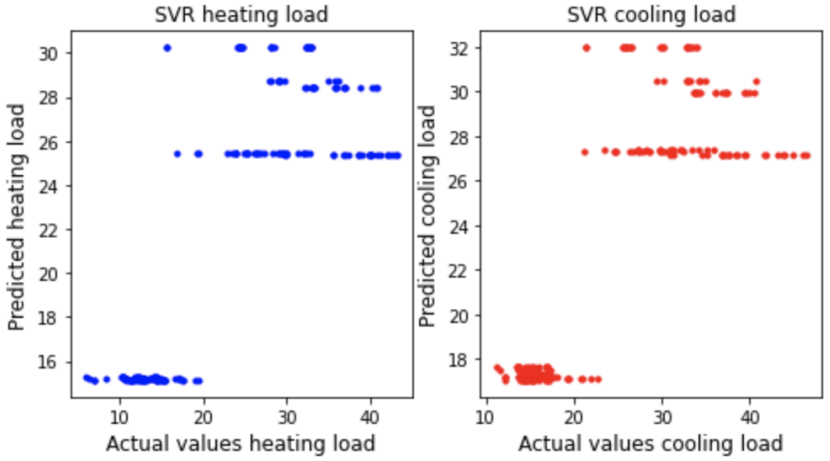

Fig. 6: Actual vs. predicted values using the SVR regression model

Finally a Random Forest regression model was fitted and the testing MSE decreased significantly to 0.3 and 4.0 for the heating and cooling load, and the plot of actual vs predicted values is very close to a diagonal line (Fig. 7). We can see that although both are very accurate, the heating load one seems to be more, which is something that might be worth further exploring.

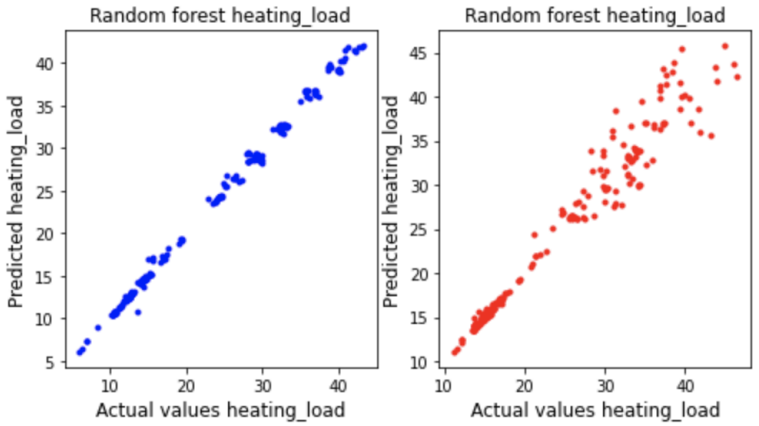

Fig. 7: Actual vs. predicted values using the Random Forest regression model

Not unexpectedly, out of the three models analyzed, this is the one that best fits the data and shows consistently superior performance, which confirms our initial hypothesis that the relationships between the explanatory and response variables are quite complicated to be adequately captured by a linear model. Now after having selected the best one, a SHAP analysis was carried out and the results are shown in Fig. 8.

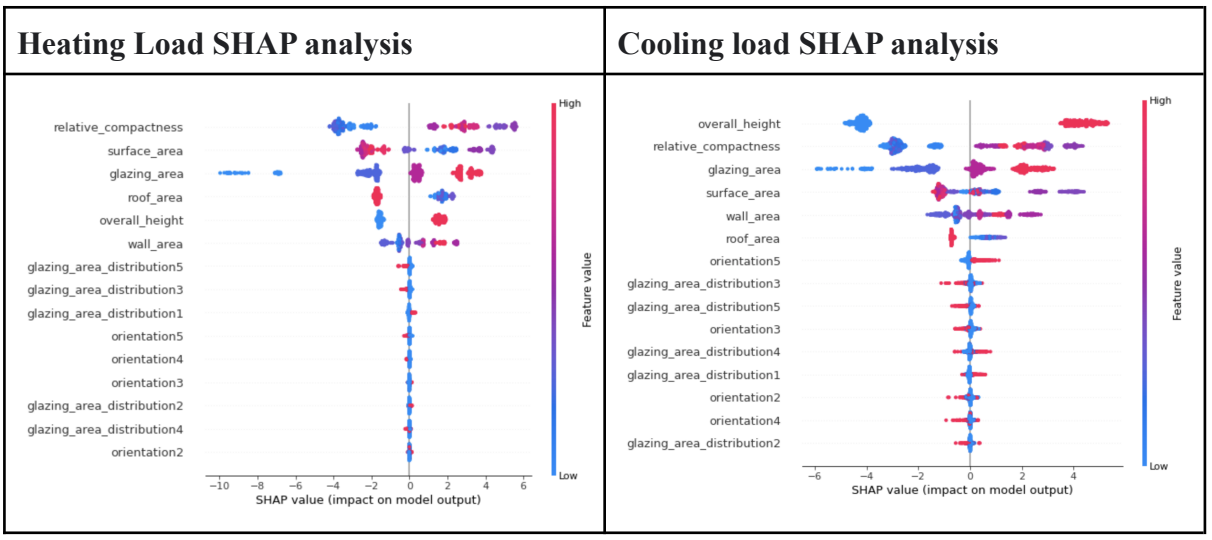

Fig. 8: SHAP analysis obtained from the Random Forest model

For the heating load, we can observe that the most important factors are relative compactness, surface area and glazing area. Heating load seems to have a linear behavior according to glazing area (lower values of glazing area are associated with lower SHAP values), but a non-linear behavior according to surface area and relative compactness. For instance, for instance for the surface area we can see that average values generate the highest SHAP values. The analysis also suggests that neither roof nor overall height are important factors for the estimation of the heating load, variables we initially thought would be, following the

exploratory analysis. Their apparent effect on the response variables might be explained by the high correlations that exist between the predictive variables.

In the case of cooling load, the most important factors seem to be overall height, relative compactness, glazing area and surface area. It would be interesting to analyze why overall height seems to have a significantly higher effect on cooling load than on heating load. Roof area doesn't seem to be important, which is explained by the same reason outlined above. Similarly, the cooling load seems to behave linearly according to the glazing area. However, the behavior is more complex according to relative compactness and surface area. Finally, higher overall height values are related to a higher cooling load; however, with only two levels contemplated in the analysis, it is difficult to make any more inferences about the effect of this variable.

Potential Limitations

- One of the limitations of the study includes the lack of data levels for the overall height of the building- since there were only two levels available, determining the actual effect height has on the heating or cooling load can be inaccurate.
- Moreover, there are many confounding variables present in the real world which the data does not seem to factor in since it is a simulation with these factors kept constant- such as the outdoor temperature and humidity and how much it differs from the desired temperature indoors, other internal appliances' effect on heating and cooling load, occupancy and its effect on heating and cooling loads, and so on.
- The type of glazing affects the amount of heat a building loses to leaks and how much heat energy a building absorbs through sunlight **[C]**, however, there is no variable present to account for glazing type. A future study could incorporate the effect of the type of glazing on heating and cooling loads.

VI. Conclusion

From the analysis we conducted on energy efficiency, we were able to draw some of the following conclusions. The first conclusion that we drew was that the relationship between the input variables and the output variables was not linear based on the analyses that we conducted (except for glazing area) . Therefore, the best predictive model for our data was the Random Forest regression model as it had the lowest testing MSE and accurately captured the non-linear relations that the data presented. Based on a further SHAP analysis, the most

important variables for predicting the heating or cooling load were relative compactness, glazing area, and surface area.

In terms of next steps for a future study, more variables could be added to the experiment to obtain a more accurate prediction for heating and cooling loads, such as humidity or temperature which were given fixed values in this experiment. Due to interactions between variables, some of the effects that we currently found might change if we contemplate some of these factors. In addition, some of the variables studied currently have only a few levels; for example, there are only two levels for height of the building. In the future, next steps could involve capturing more levels of each variable to better understand the data.

The analysis conducted in this study is crucial as determining which variables best predict heating and cooling load can help to design buildings in a more energy efficient manner which can help with global warming issues and improve our planet's overall health.

VII. Acknowledgment

Together, all team members had regular meetings where we chose a dataset, looked over the dataset, and discussed the desired direction of the project. Below are individual contributions of each team member.

- Sofia did part of the data visualization and its analysis, and carried out the modelling and its corresponding discussion.
- Harsh assisted in visualizing the data, added limitations to this report, improved upon the exploratory analysis and its interpretation, and found references that relate to improvements for future studies.
- Yash assisted in cleaning the dataset to make it tidy, visualized the data, and interpreted the results for the exploratory analysis. He also assisted in analyzing the data further to generate the hypothesis.
- Joshua introduced the project, its motivations, and its relevance to the world which we all inhabit. He also explained the dataset and facilitated collaboration by configuring the GitHub repositories.

- Anusha consolidated information from all team members and put together the slide deck presentation. She also developed conclusions from the analyses conducted and wrote about these in the paper.

(The link to our website is https://sharkman424.github.io/sds322_team4.github.io/)

VIII. Bibliography

A. A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012

B. Gayle Markovitz Partnerships Editor, and Gayle Markovitz. “COP26: Here Are 7 of the World's Greenest Buildings and Best Solutions to Build Sustainably.” *World Economic Forum*, World Economic Forum, 2 Nov. 2021, https://www.weforum.org/agenda/2021/11/cop26-buildings-green-architecture-build-better-now-climate-change/.

C. Bulow-Hube, H. The effect of glazing type and size on annual heating and cooling demand for Swedish offices. Canada: N. p., 1998. Web.