1 Introduction

The energy efficiency Data Set is collected from UCI Machine Learning Repository through kaggle.
In this data set we study the energy efficiency of 768 buildings by measuring the heating load and cooling load. those loads are (or can be) functions of 8 building parameters as the features of the data set:

- **Relative compactness**, which is the volume to surface ratio of a building (or a closure in general).
- **Surface area**, as the surface area of a building.
- **Orientation**, characterized by four integers, where each is direction of a building like north (a north facing building), south, etc.
- **Glazing area**, is the total area of glass on the wall.
- **Glazing area distribution**, that shows how glass area is distributed across the building.
- **Wall area**, **Roof area** and **Overall height** are obvious.
- **Heating load** and **cooling load**, as the amount of load required for heating and cooling the building, respectively.

This data set can be used either for predicting a new building energy efficiency or interpretation, to see which features have the most impact on the energy efficiency of the buildings. I used models that are focused on interpretation.

2 Cleaning and Feature Engineering

The data set doesn't have any null value. All features are either float or integers, therefore we don't need to transform them to numeric variables via one hot coding.

3 Regression models

I applied three regression models. First I chose K-fold train-split to 3 splits and set the shuffle mode. Linear Regression gives the lowest R^2 error for around 0.90. After that I imported GridSearchCV for both Lasso and Ridge to find the best degree for the polynomial features and the best alpha value for the regularization. I found out that the best degree for the polynomial features for both lasso and ridge is 5 and the best alpha is 0.01. Ridge has slightly higher R^2 than lasso but for both were around 0.99. Therefore I choose ridge as my final model for interpretation of my data set.

4 Interpretation

For interpretation of my data set I choose ridge which had the highest R^2 error. As I mentioned the best degree of polynomial features for our data set was 5. Since such high degree of interactions of features needs more complicated interpretation of interactions, for more straightforward interpretation, I limited the degree of the polynomial features to 2 and

keep alpha 0.01. By setting the degree of the polynomial features to 2, the R^2 error has been decreased a bit to 0.98, which shows it is still very much reliable. By looking into the derived coefficients and sorting them, I have the following interpretations:

- At the first order, **relative compactness**, **wall area** and **overall height** have the highest impact on the energy load of the building. On the other hand, **orientation** and **glazing area distribution** have nearly no impact.
- At the second order, *relative compactness* $\times$ *overall height* and *wall area* $\times$ *overall height* are the most important features of a building for its energy efficiency.

For a deeper interpretation of this data set one should consider higher degree of polynomial features to consider higher ordered interactions and also study the relation of each featured with heating and cooling loads. For example, although orientation has not much impact on the energy efficiency of the building but still there is differences between different directions.