IBM Supervised Learning: Regression Project - Energy Efficiency and Heating Load Prediction

This study looks into using linear regression to predict energy efficiency in the form of heating load. Various regression techniques are considered – initially, linear, and polynomial regressions are used. Following this, regularization techniques such as ridge, lasso and elastic net regularisation were used. It was found that polynomial regression was the most accurate on our testing set. However, this may be due to the small number of features provided in the dataset.

1 Introduction

The heating load of the building is defined as the amount of heat energy that must be added to the space in order to maintain the temperature in an acceptable range. This is of particular interest as not only by designing more efficient buildings can heating costs be reduced but additionally, the carbon footprint from heating such buildings is reduced.

The dataset was taken from the UCI machine learning repository, accessible at https://archive.ics.uci.edu/ml/datasets/Energy+efficiency and this report builds on work done by Tsanas and Xifara[1].

Our model will be focussed on prediction of the heating load of various buildings.

2 Data Exploration and Cleaning

The dataset has eight features denoted by X1... X8 and two target variables denoted by Y1 and Y2 and shown in Table 1.

Code	Variable Name
X1	Relative Compactness
X2	Surface Area
X3	Wall Area
X4	Roof Area
X5	Overall Height
X6	Orientation
X7	Glazing Area
X8	Glazing Area Distribution
Y1	Heating Load
Y2	Cooling Load

Table 1: Variables

For the purposes of this study, the cooling load variable was dropped and not used for training or prediction purposes.

In addition to this, rows with N/A or missing values were dropped. As there was no non-numerical data, one-hot encoding was not required.

The data was then visualised in Figure 1.

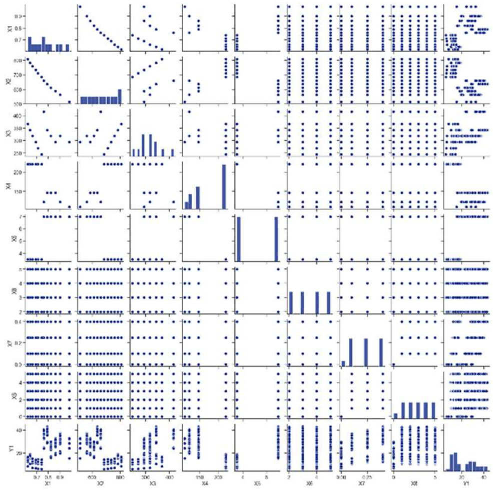

Figure 1: Pairplot of the cleaned dataset

3 Regression Models

The dataset was split into a train set and test set with percentages 30% and 70% respectively.

The regressions trained on the dataset were: standard linear regression, polynomial regression, ridge regression, lasso regression and elastic net regression – the latter 3 being regularisation techniques.

For the ridge, lasso and elastic net regressions, models were trained using theoretically “good” alpha values and then 4-fold cross-validation was used, comparing a range of alpha values in an attempt to optimise the regressions. The reason for the latter producing greater errors lies in the fact that for the first models, polynomial features were created up to a degree of 12 but for the latter they were only created up to a degree of 5 in order to speed up model training.

The chosen error statistic to measure the error and compare the regressions was mean-squared error.

3.1 Linear Regression

The linear regression equation can be described as:

$$y_{pred} = \sum_{i} \beta_i x_i$$

Where β is a weight as determined from fitting the model and x refers to the features of the dataset. Note that x_0 = 1 in order to add a constant term.

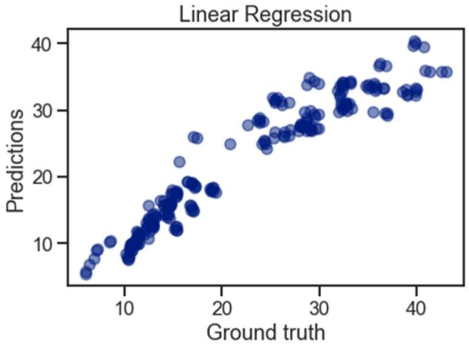

Figure 2: Standard linear regression

Using standard linear regression produced a fairly large error of 8.15 and as a result, other regressions were explored.

3.2 Polynomial Regression

In order to increase the number of features, polynomial regression was used. After testing the various degrees from 2 to 10, it was found that using a degree of 4 minimized the in the test set.

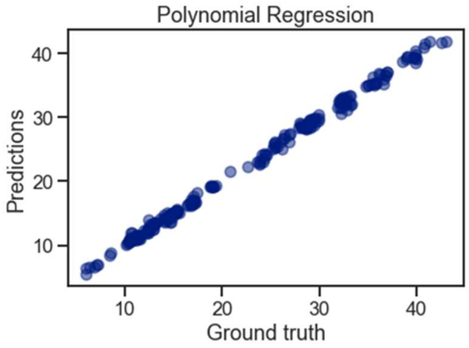

Figure 3: Polynomial regression (degree = 4)

Polynomial regression was very accurate and had a mean-squared error of 0.289.

Ridge Regression

Using ridge regression with an alpha of 0.001, a mean-squared error of 0.363 was produced.

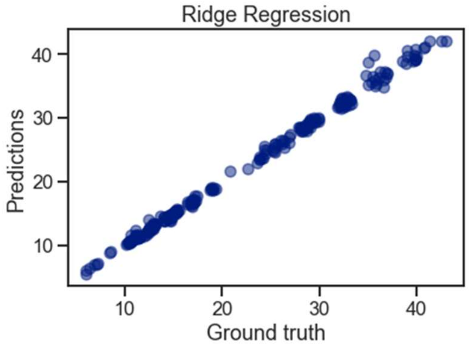

Figure 4: Ridge regression (alpha = 0.001)

Using ridge regression with an alpha of 0.005, a mean-squared error of 0.691 was found.

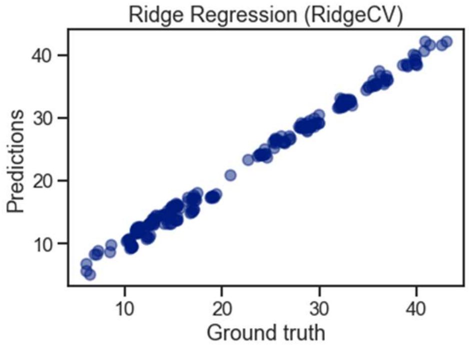

Figure 5: RidgeCV regression (alpha = 0.005)

Lasso Regression

Using lasso regression with an alpha of 0.0001 with polynomial features yup to a degree of 12 produced a mean-squared error of 0.353.

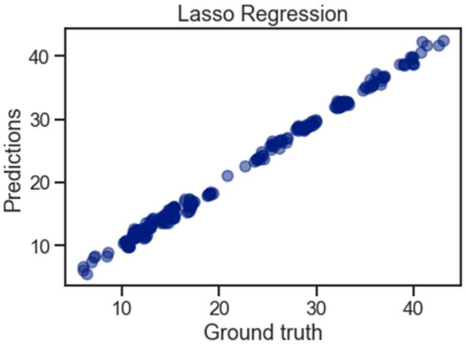

Figure 6: Lasso regression (alpha = 0.0001)

Using lasso regression with polynomial features up to a degree of 5 with cross validation produced a mean-squared error of 2.82.

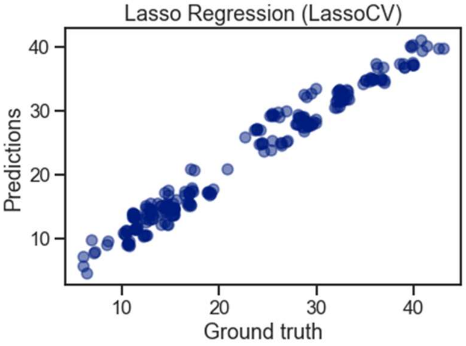

Figure 7: LassoCV regression (alpha = 0.0001)

Elastic Net Regression

Using elastic net regression with an alpha of 0.005 and L1 ratio of 0.9, a mean squared error of 4.49 was found.

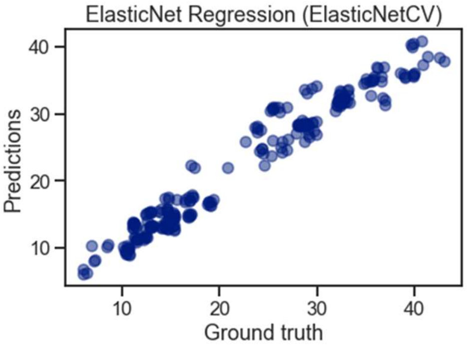

Figure 8: Elastic net regression (alpha = 0.005, L1 ratio = 0.9)

4 Results

Polynomial regression appears to be the most accurate of the various techniques used. This is likely the case due to the small amount of data available, the data only had 8 features. As a result, the use of regularisation techniques was not as helpful.

We recommend polynomial regression as it is the most accurate and still remains fairly explainable.

However, these regularisation techniques were still compared as shown in Figure 9.

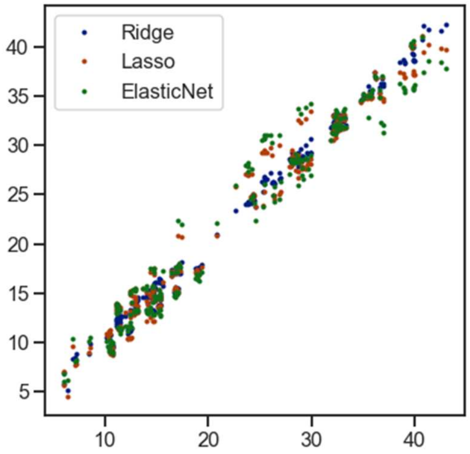

Figure 9: Comparison of regularisation techniques

From this it is clear ridge regression worked best on the dataset out of these regularisation regressions, this provides insights into why these techniques were less successful – lasso regression (and hence elastic net) seeks to reduce the number of features. However, given the already small number of features, this reduces the accuracy of the model.

From the linear regression, we are able to draw insights into the importance of features within the model. As features were scaled with MinMax scaling, comparisons can be drawn.

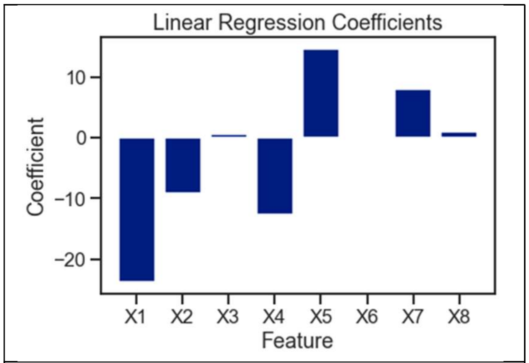

Figure 10: Linear regression coefficients

From Figure 10, it is clear wall area, orientation, and glazing area distribution (X3, X6 and X8) are not significant factors in the heating load of a building.

5 Conclusion and Next Steps

In conclusion, it is clear regression can be used to determine the heating load of a building. In this case, it appears polynomial regression provides the greatest accuracy out of the models that were trained and tested. However, models may be able to perform better with a greater number of features.

References

[1] 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 (the paper can be accessed from [Web Link])