Regression Project Report

This report details the methods and outcomes for CS5420 Project 1 - Regression. This project was completed using the NumPy, pandas, ucimlrepo, matplotlib, seaborn, and scikit-learn libraries in Python.

Problem

The goal of this project was to implement linear and ridge regression to predict values related to energy efficiency, in this case heating and cooling load, given 8 features of the house.

Data

The dataset for this project was the Energy Efficiency dataset from UC Irvine Machine Learning Repository. The data has eight features and two target values. The features and targets are as follows.

Name	Feature or Target	Description
Relative Compactness (X1)	Feature	Measure of how compact the building is.
Surface Area (X2)	Feature	Total external surface area of the building.
Wall Area (X3)	Feature	The area of the building's walls
Roof Area (X4)	Feature	The area of the roof of the building.
Overall Height (X5)	Feature	Height of the building.
Orientation (X6)	Feature	Orientation of the building (A numerical representation of north, east, south, or west from 2-5)
Glazing Area (X7)	Feature	The total area of glazing (exterior windows).
Glazing Area Distribution (X8)	Feature	Represents the distribution of window area.
Heating Load (Y1)	Target	The amount of energy required to maintain the baseline indoor temperature

		while heating.
Cooling Load (Y2)	Target	Amount of energy required for cooling under equivalent conditions to Y1.

Methods

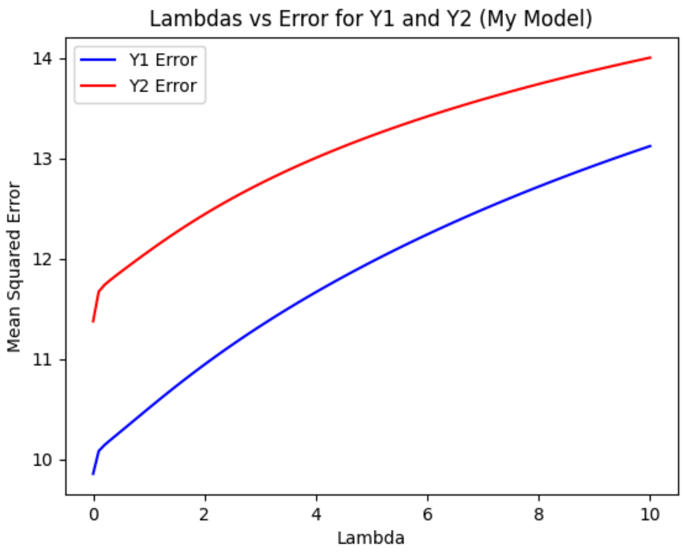
I started the project by splitting the data into the standard 7:1:2 sections using the sample() function from pandas to randomly sample the data. This was in an effort to use good sampling techniques. I then trained my model using the vector formula for linear regression weights. Importantly, when performing my calculations I had to use the NumPy pseudo-inverse function because the matrix proved to not be invertible during my calculations. When attempting to optimize ridge regression I did this by searching every possible lambda value from 0.0 to 10.0 by 0.1 increments.

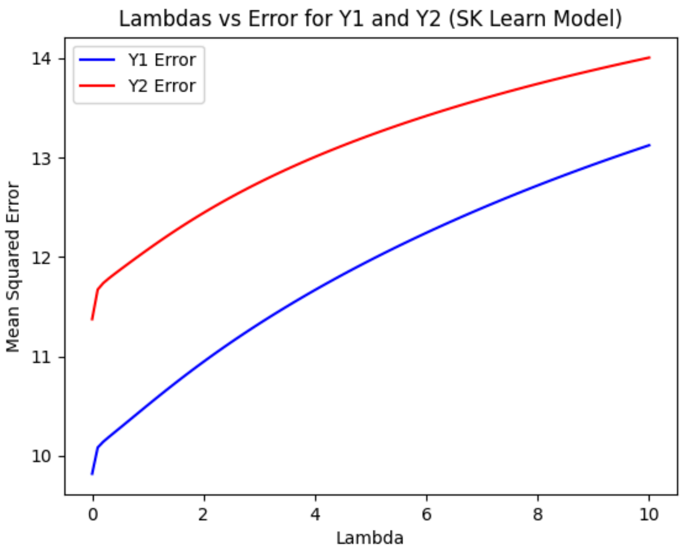
The metric I used to evaluate my models was the mean squared error. Additionally, I compared the output of my models to that of the proven scikit-learn library.

Results

The mean squared error for my implementation of linear regression was 7.565 for Heating Load and 8.123 for Cooling Load. For heating load, my mean squared error differed from that of scikit by 3.5x10^{-9}, so a negligible difference. For the Cooling Load target, my models mean squared error differed by roughly 2.09x10^{-10} from that of scikit, resulting in another negligible difference.

Surprisingly, the best lambda of those searched for this project was 0. This can be seen in the following plots based on the results from both my model and scikit learn.





The two charts are virtually identical, indicating that my implementation of ridge regression is very similar to that of scikit-learn. Further backing up this point, the difference between my model and scikit's model was 0.014 for Y1 and 0.027 for Y2.