Predicting Buildings' Greenhouse Gas Emissions Using Building Characteristics, Ultility Usage, and LASSO Regression[*]

Joseph Burks

December 10, 2025

Climate change is a significant issue today. Several sources influence climate change. One is greenhouse gases like carbon dioxide. Knowing the substantial sources of greenhouse gases can help individuals develop strategies to combat climate change. In this report, we fit a predictive equation using data from 999 observations of buildings subject to the Building Performance Ordinance (BPO) in San Jose, California. We tested several regression models. We tested several models. The best performing is a LASSO regression model: Greenhouse gas emissions = 125.342 + 0.0005 (FLOOR AREA) - 0.903 (ENERGY STAR SCORE) + 0.0003 (TOTAL ELECTRICTY USE). The model had a RMSE of 3778.5594 on our testing dataset, indicating that the model has strong preditive power.

1 Introduction

Climate Change is a critical issue and has the potential to escalate to a cataclysmic disaster. Greenhouse gases such as carbon dioxide trap heat in the atmosphere, contributing to climate change. There are numerous sources of greenhouse gases. In this report, we examine the emissions from a building's energy usage. Knowing which buildings contribute the most greenhouse gases can help city planners develop more useful building strategies or better renewable energy initiatives.

To accurately calculate a building's greenhouse gas emissions, one would need to know the exact sources of the electricity. Several factors can influence the sources of energy, from the location of the building to the time of day. As a result, many places rely on Energy Star, a part of the EPA, to do these calculations. This can be expensive and requires giving data to

[*] Project repository available at: https://github.com/GrumioEstCoquus/MATH261A-final-project.

a third party. This report seeks to develop a predictive model that building owners can use instead.

The report aims to accurately predict a building's CO2 emissions with utility usage and building characteristics. We fit several regression models, including multiple linear regression, LASSO regression, and Ridge regression, on a training subset of the data. Then we tested those models on a validation set. Finally, we calculated the Root Mean Squared Error for the best-performing model on a testing set. We obtained the best model with LASSO regression with a lambda of 43.2615. The fitted LASSO model is Greenhouse gas emissions = 125.342 + 0.0005 (FLOOR AREA) - 0.903 (ENERGY STAR SCORE) + 0.0003 (TOTAL ELECTRICTY USE)

The remainder of this report is structured as follows: Section 2 discusses the data, Section 3 discusses the regression models we used, Section 4 discusses the results, and Section 5 discusses the findings, weaknesses, and further questions.

2 Data

The data we used in this report comes from the San Jose, CA Open Data Portal (Team 2025). In 2018, the city of San Jose adopted the Energy and Water Building Performance Ordinance (BPO). The BPO requires nonresidential and multifamily buildings larger than 20000 ft^2 to track their energy and water use with the EPA ENERGY STAR Portfolio Manager. The data contains 999 observations for buildings subject to the BPO in 2023.

The variables of interest are total greenhouse gas emissions, ENERGY STAR score, year built, surface area, total electricity use, and total water use. Total greenhouse gas emissions is measured in metric tons (C02). ENERGY STAR reports the value based on a building's natural gas use and the electricity coming from nonrenewable sources. It has a mean of 190 metric tons of CO2 and a standard deviation of 677 metric tons of CO2. ENERGY STAR score ranges from 1 to 100. ENERGY STAR also reports the score based on the building's energy performance. It is a comparison to similar buildings nationwide; a score of 50 represents median performance. It has a mean 84 and standard deviation 26. Year built is measured in years. It has a mean 1986 and standard deviation 20. Area, total electricity use, and total water use are all reported by the building owners or managers. Area is measured in square feet, with mean $6.9992 \times 10^4 \text{ft}^2$ and standard deviation $2.25158 \times 10^5 \text{ft}^2$. Total electricity is measured in kilowatt-hours, with mean 4.74586×10^5 kWh and standard deviation 1.81337×10^6 kWh. Total water use is measured in kilo-gallons, with mean 2173 kilo-gallons and standard deviation 1.5907×10^4 kilo-gallons.

3 Methods

We used several regression models in this report. One is multiple linear regression. The general model is as follows: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$ Y_i are the observed responses. In this case, the total greenhouse gas emissions. β_0 is the intercept. It represents the average response value when all predictors are zero. In this case, it contains little meaning, since all predictor variables are strictly positive. The other βs are the slopes corresponding to each predictor variable. The βs represent the average change in total greenhouse gas emissions for a one-unit change in the corresponding predictor variable while keeping the other predictor variables constant. The X_i are the observation predictor values. In this case, the ENERGY STAR score, year built, surface area, total electricity use, and total water use. The ε_i are the error terms and represent the variability in Y not explained by a function of X_i.

There are several assumptions for multiple linear regression. One is that the relationship between the response and predictor must be linear. Meaning, $E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1}$ is correct. The others relate to the ε_i. The ε_i must be independently and identically distributed normal with constant variance ($\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

One model used stepwise regression. Stepwise regression involves fitting simple linear regression models for each potential predictor. For every model, we calculate a test statistic and add the predictor corresponding to the best statistics that exceeds a threshold to the model. This process repeats by adding and removing potential predictors until corresponding to a threshold until no more predictors can be added or removed.

The two models used are regularized regression models. First is Ridge regression. Ridge regression is similar to multiple linear regression but we estimate the parameters by minimizing $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \dots - \beta_{p-1} X_{i,p-1})^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2$ Where $\lambda \geq 0$ controls the level of regularization. We chose the optimal λ by minimizing the AIC on the training data set. LASSO regression is very similar to Ridge regression. We estimate the parameters by minimizing $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \dots - \beta_{p-1} X_{i,p-1})^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|$ Where $\lambda \geq 0$ controls the level of regularization. We also chose the optimal λ by minimizing the AIC on the training data set.

To avoid overfitting, we split the dataset into training, validation, and test sets. We used the following distribution to split the data: 70% training, 15% validation, and 15% test. We trained the models on the training set. We then calculated the Root Mean Square Error (RMSE) for each model with the validation set. We considered the model with the smallest RMSE on the validation set to be the optimal model. Finally, we computed the RMSE of the optimal model with the test set to measure its predictive power. We fit the simple and multiple linear regression models using the lm() function in R (R Core Team 2024). We fit the LASSO and Ridge models in R with the glmnet package (Friedman, Hastie, and Tibshirani 2010).

4 Results

In total we fit seven models. Three simple linear regression models, with predictors total water usage, total electricity usage, and floor area. One model is a multiple linear regression model that used all predictors. One model was obtained with stepwise regression. One model is a LASSO regression model. The last model is a Ridge regression model. Figure 1 shows the models corresponding RMSE on the validation dataset. The LASSO regression had the smallest RMSE (102.8022). The Ridge regression, multiple linear regression, stepwise regression models all had similar RMSE and close to the LASSO regression model. The simple linear regression models with water only and floor area only had much larger RMSE, over 3 times as large as the LASSO regression RMSE. Based on the RMSE the LASSO model is the optimal model for our data.

Table 1: Models and Corresponding Validation RMSE

Model	RMSE
LASSO	102.8022
Ridge	113.0888
All Predictors	117.3212
Stepwise	118.3343
Electricty Only	127.1455
Water Only	311.6665
Floor Area Only	317.2115

Figure 1: Table of all the models tested with their corresponding RMSE on the validation dataset

The fitted LASSO model is: $\hat{Y}_i$ = 125.342 + 5 × 10^{-4} (FLOOR AREA) - 0.903 (ENERGY STAR SCORE) + 3 × 10^{-4} (TOTAL ELECTRICTY USE)

The estimated intercept parameter is $b_0 = 125.342$. This represents the average greenhouse emissions for a building whose floor area is 0, ENERGY STAR score is 0, and electricity use is 0. Since all of these predictors are strictly positive, the intercept does not contain much meaning.

The estimated slope parameter for floor area is $b_1 = 5 \times 10^{-4}$. This means that for each square-foot increase in floor area while keeping the other predictors constant, the building's greenhouse gas emissions will increase on average by 5×10^{-4} metric tons of CO2.

The estimated slope parameter for ENERGY STAR score is b_2 = -0.9034. This means that for each one point increase in ENERGY STAR score while keeping the other predictors constant, the building's greenhouse gas emissions will decrease on average by 0.9034 metric tons of CO2.

The estimated slope parameter for total electricity use is $b_3 = 3 \times 10^{-4}$. This means that for each kWh increase in electricity while keeping the other predictors constant, the building's greenhouse gas emissions will increase on average by 3×10^{-4} metric tons of CO2.

This LASSO regression model has an RMSE on the testing dataset of 378.559. This is relatively small in comparison to the observed greenhouse gas emissions in this. This means that the fitted LASSO model has strong predictive power.

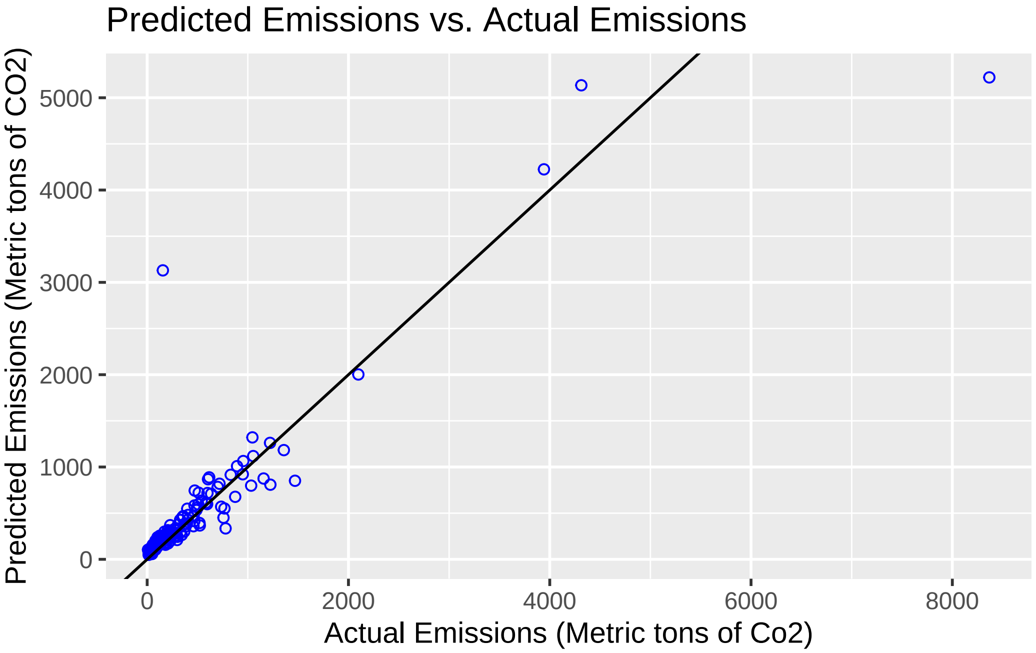

Figure 2: Plot of the predicted buildings' greenhouse gas emissions vs actual buildings' greenhouse gas emissions from the test data set

Figure 2 Shows the LASSO regressions predictions for buildings' greenhouse gas emission vs the actual greenhouse gas emissions based on the testing dataset. If the LASSO model were perfectly accurate, all points would lie along the $y = x$. Besides a couple extreme points, most of the points lie close to the line. This indicates that the LASSO model does a good job predicting buildings' greenhouse gas emissions.

5 Discussion

The LASSO regression model has decent predictive power. The RMSE on the test dataset is less than the standard deviation and relatively small in comparison to the range of total greenhouse gas emissions. Hence, the model can be used to make accurate predictions of buildings' greenhouse gas emissions with total electricity use, ENERGY Star score, and area.

While the LASSO model may have good predictive power, the data's scope is limited. The data are from buildings in San Jose subject to the BPO. The model may not be accurate for buildings not subject to the BPO and in different cities. Moreover, there is no guarantee that the trends identified with the LASSO model will continue in the future.

The scope of this report is limited to developing a predictive model. We did not perform any inference on the estimated parameters. A valuable next step would be running an experiment on another dataset to see if the predictors floor area, ENERGY STAR score, and electricity use are statistically significant.

References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Team, Open Data. 2025. "Building Performance Ordinance." San Jose CA Open Data Portal. https://data.sanjoseca.gov/dataset/building-performance-ordinance.