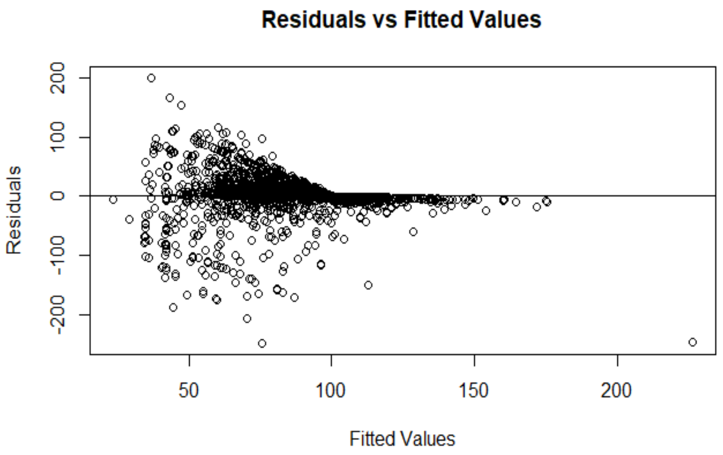
Prediction of building heated from building characteristics

Github code: https://github.com/andrln/stat506finalproject

In this report, I will be answering the question of whether or not one can predict the percentage of a building that is heated from various characteristics about the building, such as its construction material type and the amount of energy used on heating. In order to analyze this issue, I used data from the 2018 Commercial Buildings Energy Consumption survey [Appendix 1]. This survey asked various questions about a sample of buildings across the U.S. and is representative of all buildings nationally.

To utilize the data from this survey, I first needed to take into account fact that it is sample survey data, and adjust for the sampling error. I did this by utilizing example R code given by the CBECs survey user guide [Appendix 2, p13]. Next, I selected the variables I use for my analysis. The response variable is the percentage of the square footage of the building that was heated to at least 50 degrees Fahrenheit, and the predictors variables are the roof and wall construction materials, energy expenditures on heating of all types (electrical, natural gas, fuel, etc), the climate the building is in, the building type, and the total square footage of the building. I also made additional variables, such as more predictor variables such as the total energy expenditures on heating (calculated as the sum of each individual type of expenditure), and the total energy expenditures on heating adjusted for square footage (calculated by dividing by the sqft of the building). I also dropped any building types which were not meant to regularly hold people, like storage buildings.

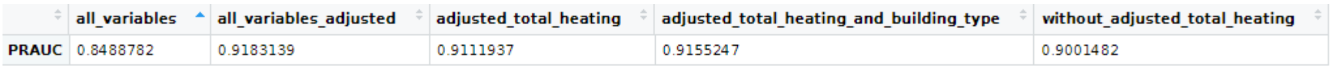
My initial approach was to use multivariate regression to predict the heated percentage of buildings, and use stepwise model selection based on AIC to select predictor variables. However, this type of analysis was quickly found to be not valid, as the assumptions for linear regression were not met as the models had extremely high heteroscedasticity.



This was being caused by the fact that the response variable was in the format of a percentage, meaning it had a minimum value of 0 and maximum of 100. The survey package used for analysis of survey data had no option for an alternative regression model, like beta regression which could handle this data format, like beta regression, so I attempted to improve the issues by limiting the predicted values to the range of the response variable. While this improved the heteroskedasticity issue [Appendix 3], it still was too large for linear regression to be appropriate.

In order to work around these issues, I created a new predictor variable which is a binary version of the initial response variable which is true when the area heated percentage is above 80. and chose to translate the problem into a binary classification one, using logistic regression models instead. This new problem had a severe class imbalance, with only ~20% of the data being in the false category [Appendix 4]. Because of this, I chose to use PRauc to evaluate my models, as it is more reliable when it comes to assessing model performance with unbalanced datasets compared to other metrics like AUC or accuracy [Appendix 5].

Results of some of the models I tested can be found in the table below. Note that a model predicting randomly would have a prauc of the proportion of positives in the data, or about .8, so all values shown should be compared to that. In the all_variables model, I included all predictor variables mentioned initially, but not calculated predictor of total energy expenditures adjusted for square footage.



	all_variables	all_variables_adjusted	adjusted_total_heating	adjusted_total_heating_and_building_type	without_adjusted_total_heating
PRAUC	0.8488782	0.9183139	0.9111937	0.9155247	0.9001482

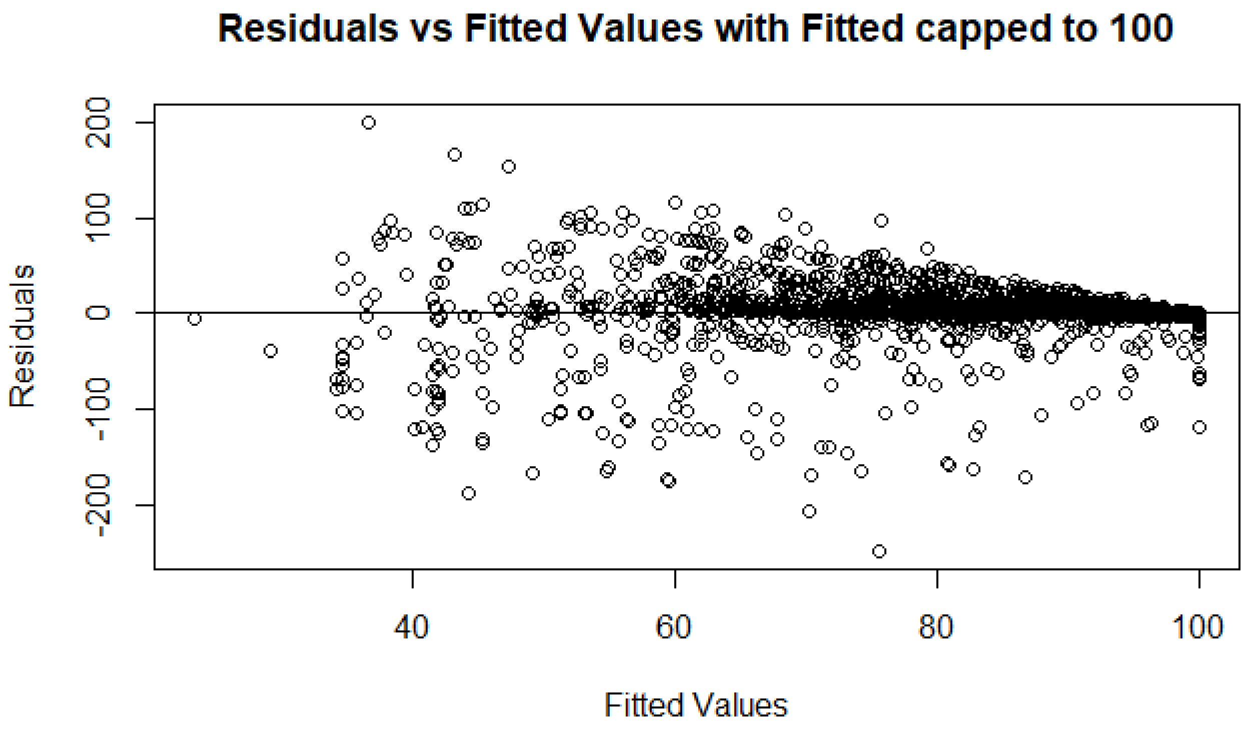
As can be seen in the table, adding the adjusted total expenditures significantly improved the model’s prauc. Other predictors did help aid in prediction, although there was not much difference between them. From this, I would state that my initial hypothesis that the percentage of a building heated above 50F can be predicted through the variables I used was correct, though the actual predictive power was underwhelming. Future analyses exploring the same question should consider using alternative models, such as beta regression, and implementing more in-depth feature selection.

Appendix

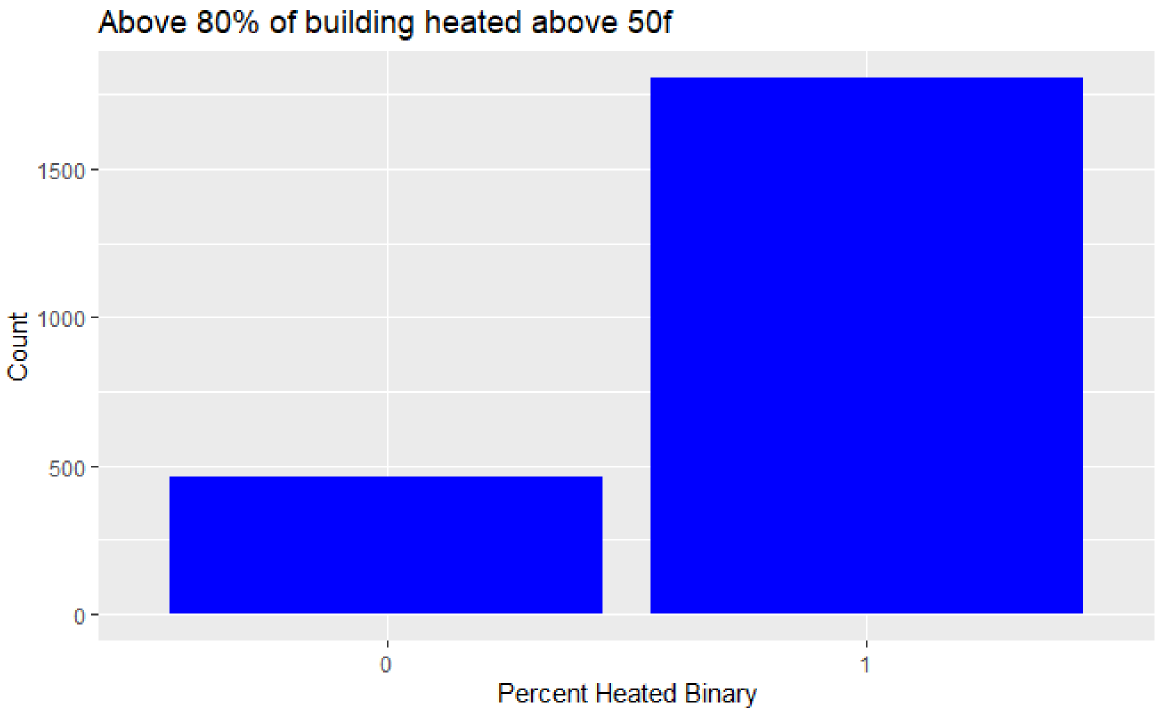
[1]: https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata

[2]: https://www.eia.gov/consumption/commercial/data/2018/pdf/Users%20Guide%20to%20the%202018%20CBECS%20Public%20Use%20Microdata%20File.pdf

[3]



[4]



[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/