Demystifying Food Inequity in the USA

Shyan Koul / Stat-471 / 12.19.2021

https://github.com/shyankoul/shyankoul-final-project

AGENDA

Introduction

Data Description and Exploration

- Sources
- Cleaning
- Description of Variables

Model Building, Evaluation, and Interpretation

- Logistic and Shrinking Classification Models
    - Logistic Classification
    - Ridge Classification
    - Lasso Classification
- kNN Classifcation Model
- Tree-Based Models
    - Random Forest Model
    - Boosted Model

Conclusions

- Takeaways
- Insights

Appendix

EXECUTIVE SUMMARY

The COVID pandemic revealed to many how drastically supply chain issues within the agriculture industry can affect peoples' everyday lives. In 2020, 551 supermarkets closed across the country, particularly in New York, New Jersey, Florida, and California.[1] As a result, many communities were left without access to nearby food, creating food deserts. The reality is, however, that this issue has existed long before COVID, in large part because policymakers and community activists cannot see food inaccessibility coming until it is too late. Here lies the inspiration for this study.

This study uses data from the US Census Bureau to predict which Census Tracts will be flagged by the United States Department of Agriculture to be deemed 'low-access'. This means that a significant portion of the Census Tract does not live within 10 miles of the nearest supermarket or grocery store. The factors considered in making this prediction include, but are not limited to, average household income, racial demographics, percent of residents living in mobile homes, as well as population density.

The analysis was based off the creation of six different predictive models: Logistic Classification, K-Nearest Neighbors Classification, Ridge Classification, Lasso Classification, Random Forests Model, and Generalized Boosted Models The majority of the analyses pointed to the same most significant predictive features, though they varied in predictive accuracy. The most accurate classification model was the Generalized Boosted Model, and the least was the K-Nearest Neighbors Model.

Factors such as the percent of the Census Tract without access to plumbing were large indicators of food inaccessibility across multiple models. The study concludes with the suggestion that to address the predictability of the lack of food options within certain neighborhoods, policymakers should implement a tax on supermarkets and grocery stores that leave Census Tracts classified as low access.

[1] https://www.scrapehero.com/department-and-grocery-store-closures-in-2020/

INTRODUCTION

In 2020, over 38 million Americans were food insecure, including 12 million children.[2] Although many consider the United States to be among the most developed nations in the world, food insecurity—or the lack of available financial resources for food at the household level—is a major social and health concern in the United States. This issue, however, is not solely limited to financial constraints. In many areas of the country, food is inaccessible simply due to geographical limitations. These limitations arise when supermarkets, grocery stores, and other grocers flee certain neighborhoods because they are not as profitable as others. Thus, food is simply not accessible to many communities.

Although it is easy to consider this issue merely an effect of socioeconomic disparities, studies have shown that the issue is actually far more complicated. The aim of this study is to understand the factors that most contribute to food inaccessibility across the country and to predict which communities are most vulnerable to this lack of access. Success will be evaluated through the correct assessment of low-access communities in concurrence with data from the United States Department of Agriculture.

Although these communities have already been identified, this model can be used to predict future food access vulnerability, allowing policymakers and grocers to work together and prevent the introduction of this problem in the first place. For example, if a community's demographics change, it could make them more likely to lose access to food within the next few years. Predicting which communities will be flagged as low access is particularly important because the USDA only conducts its analysis every two years while grocers relocate every day in the United States. Finding which communities are at risk before they are already at risk grants communities time to address the issue before it even becomes an issue.

DATA DESCRIPTION AND EXPLORATION

Data Sources

The first data source used in this analysis was the 2019 Food Access Research Atlas from the United States Department of Agriculture. This atlas provided important metrics for measuring neighborhoods' access to food via supermarkets and grocery stores. The features included were considered primarily as response variables, and each observation represented a Census Tract, neighborhoods that generally have a population around 1,000.

The second relevant data source used was the 2019 Planning Database Block Group Data. This dataset was primarily used as explanatory variables, providing relevant statistics for each Census Tract. Examples include racial demographics, poverty rates, and age distribution.

Data Cleaning Process

Both data sets required rigorous cleaning to be useful for the analysis. Many features within the Food Access Data were categorized as factors rather than numerical values and many presented important statistics as the number of people within a demographic rather than that demographic's percent makeup of the entire population. This data had to undergo a standardizing procedure using the 'mutate' function.

The primary challenge faced with the Census data was that it was organized by block groups, which are smaller neighborhoods within Census Tracts, rather than Census Tracts. Thus, each observation could not immediately be matched with a counterpart in the Food Access Data. After selecting the features most relevant to food access issues, the data was grouped by Census Tract and the weighted averages of the features (using population as a weight) were calculated. In some cases, it was more appropriate to find the sum.

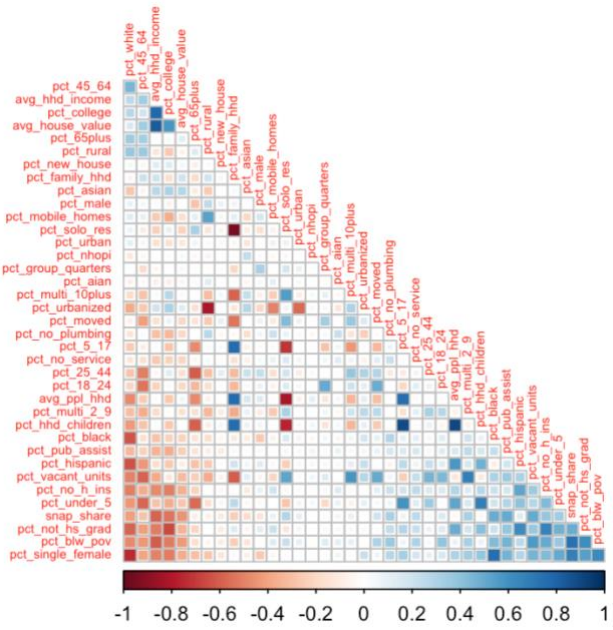
To merge the two data sets, the Geography Identifiers by Block Group and Census Tract numbers of each observation were standardized so that they can serve as the basis of the connection between the two data sets. After merging, there were a total of 70,147 observations and 38 explanatory variables.

The data was divided into training and test data at random using the following method:

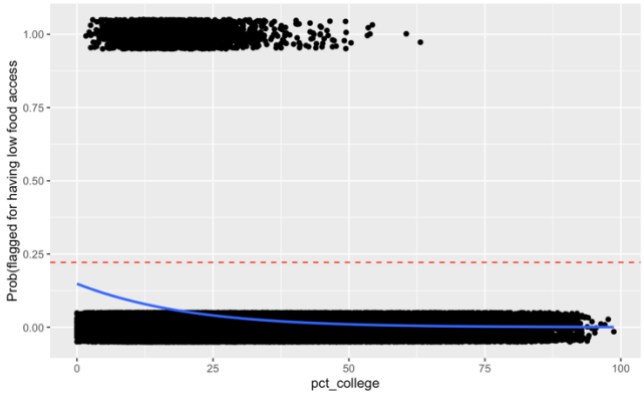
```
set.seed(250)
train_samples <- sample(1:nrow(us_food_access),0.8*nrow(us_food_access))
food_train <- us_food_access[train_samples,]
food_test <- us_food_access[-train_samples,]
```

Description of Variables and Data Exploration

Explanatory Variables: The explanatory variables include demographic descriptors like racial distribution, average household income, public assistance prevalence, age distribution, population density, and marital status.



- There was little multicollinearity found in the explanatory variables. There was a predictable negative correlation among races and a positive correlation among features like "average household income" and "average home price". There were still many instances where these correlations diverged, so they were still deemed to be important to include in analysis.



- Upon initial analysis, it is clear that multiple factors are required for analysis. The strongest predictor for food

Response Variable: Many response variables were considered, and a strenuous process was undergone to determine which of many was most compelling. The first decision was whether to use "% of population not within X miles of a food provider" or the USFDA's "food accessibility problem" flag. Given the purpose of this research is to predict which counties are at risk of food accessibility issues, it was decided to use the USFDA's "food accessibility problem" flag. This way, in the future, the model could quickly identify what Census Tracts would need examination through prediction. The second decision was to determine what mile radius to use to determine food inaccessibility. For this, a logistic regression was run for the 0.5-mile, 1-mile, 10-mile, and 20-mile levels. The 10-mile level output the most statistically significant features. Further, I found 10 miles the most compelling categorization, as it presents an issue to all individuals, regardless of vehicle access, but is not as severe as 20 miles.

- <u>Class imbalance:</u> Only 4.52% of the observations in the training data were flagged to be deemed as in poor conditions for food accessibility (within a 10-mile radius). Given the large sample size, this value is still 2,538 Census Tracts, but it also grants way for later model complications, as models would have to be very well-fit to predict the few instances in which a Census Tract is flagged.
- The response variable is not heavily correlated with any singular explanatory variable, making it a good candidate for prediction models

access flagging in a Census Tract, the percent of residents who attended college, did not predict a single true positive on its own, even with a generous weight for the cost of a false positive.

- There are no categorical features out of the explanatory variables, so having too many levels on categorical features will not be an issue

MODEL BUILDING, EVALUATION, AND INTERPRETATION

LOGISTIC AND SHRINKAGE CLASSIFICATION MODELS

Logistic Classification

The first step of analysis was the creation of a logistic classification model. Since the explanatory variable had a binary output, this method was a good starting point to understand prediction. An ordinary least squares regression would not be an accurate prediction tool, as it would be incredibly underfit. Through exploratory data analysis of the explanatory variables, no assumptions of a logistic classification model were broken (independent observations, large sample size, absence of multicollinearity, linearity of independent variables and log odds, and lack of strongly influential outliers). Unlike linear regressions, logistic regressions do not require a linear relationship between the dependent and independent variables, a Normal residual distribution, or homoscedasticity.

The regression revealed 20 statistically significant relationships with an alpha level of 0.05. The three strongest indicators for food access issues are Census Tracts with large Black populations, many households without plumbing, and large Asian populations. The three strongest indicators against were Census Tracts with large populations using SNAP benefits, large populations living in multi-family housing complexes with 10+ residents, and multiple people living in each household. The misclassification rate was 3.5% and the AUC was 0.9698, indicating a strong model.

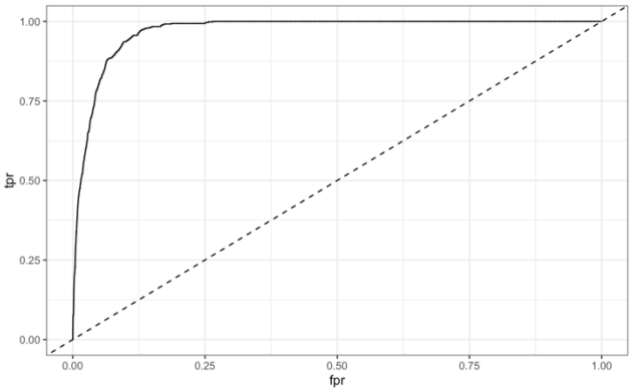
Strongest indicators for:

% Black, % Without Plumbing, % Asian

Strongest indicators against:

% On SNAP, % Multi-Family Housing, Avg People/Household

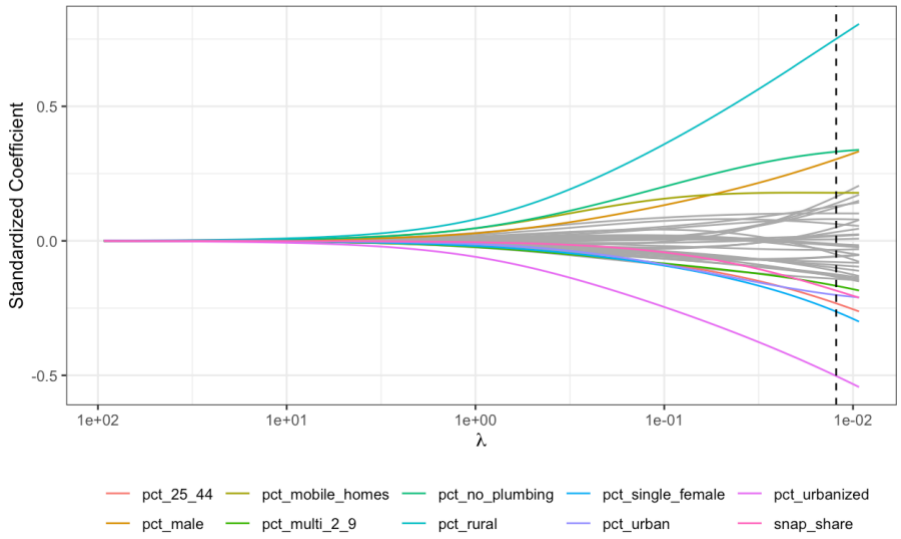
ROC curve for logistic classification



Ridge Classification

For comparison, a ridge classification model was also set onto the dataset. Ridge models have a large advantage over logistic models when there are more predictor variables than there are observations. Although that is not the case in this scenario, ridge regressions are also useful because they do not require unbiased estimators—they in fact add bias to estimators to reduce the standard error of the model. In inducing a penalty for coefficients, ridge regressions disincentivize few variables overhauling much prediction power. The large unidirectional coefficients for pct_urban, pct_urbanized, and pct_rural in the logistic classification model indicates that there may be a need to limit coefficient size.

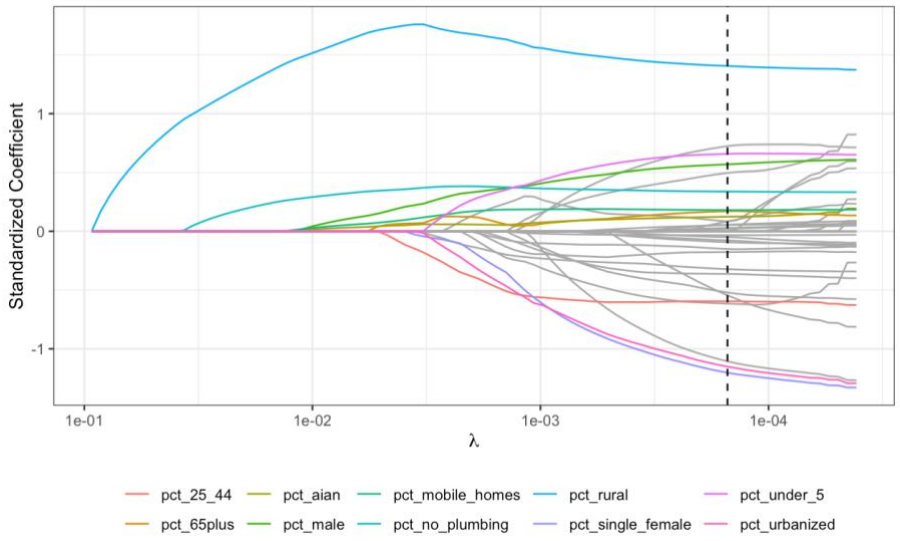
The model was cross-validated to find the optimal lambda value chosen according to the one-standard-error rule. The following shows the top 10 features identified by the ridge classification model:



Lasso Classification

About half of the features included in the logistic classification model were statistically significant at a 0.05 level, and most of the remaining features were not even significant at the 0.1 level. Thus, it is possible that a few features have large effects on the response variable. This structure would incentivize both coefficient shrinkage and feature selection, which lasso classifications do well.

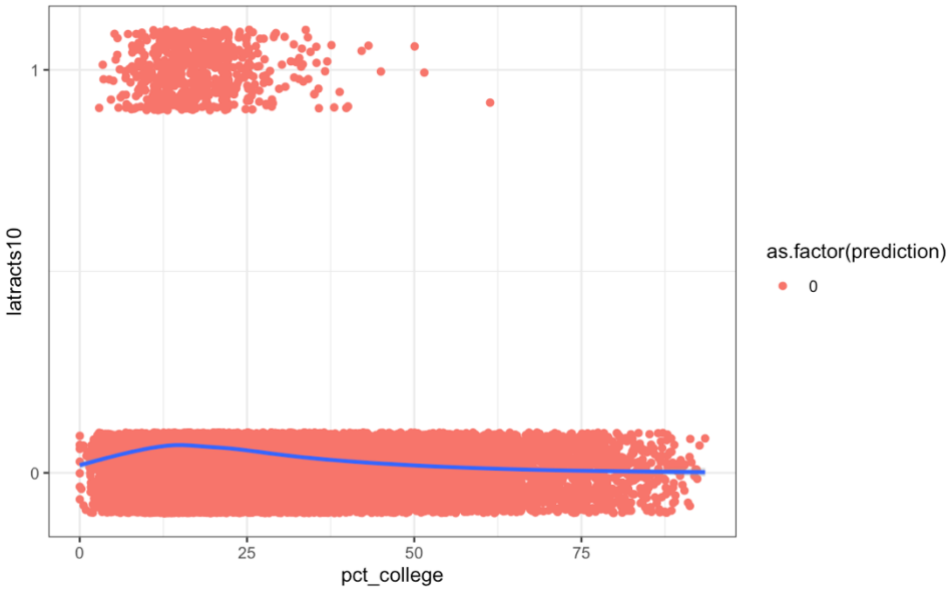
The model was cross-validated to find the optimal lambda value chosen according to the one-standard-error rule. The following shows the top 10 features identified by the lasso classification model:



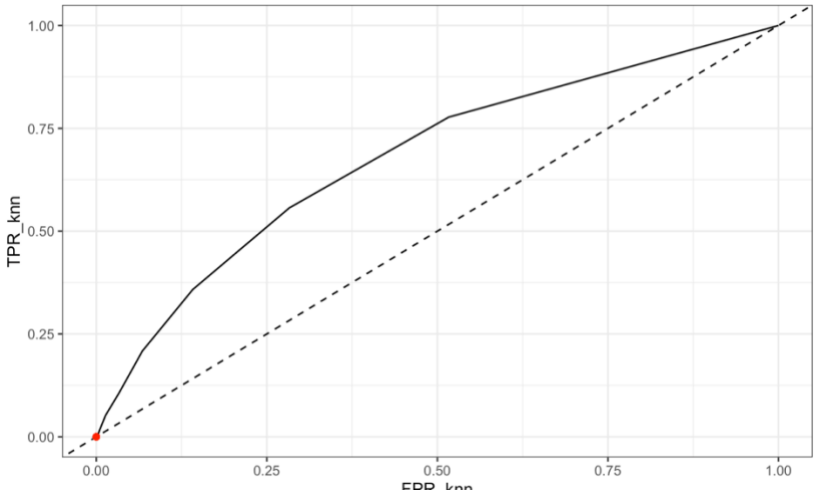
**Note: an elastic net regression model was intended but could not be executed due to run-time errors*

KNN CLASSIFCATION MODEL

kNN classification uses a plurality vote over the k closest data points to test observations. Since it is primarily used for classification, I chose to include it in this comparison of different fitted models.

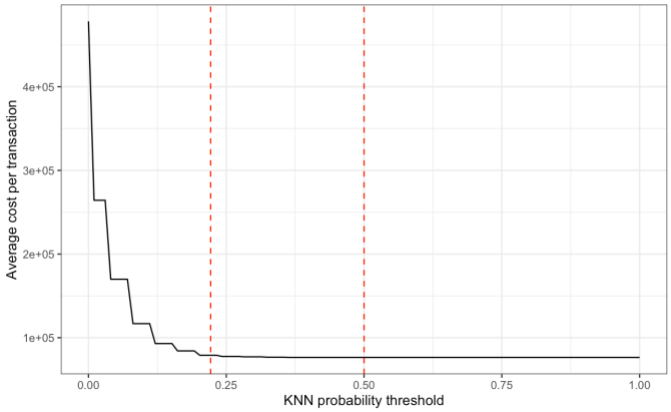


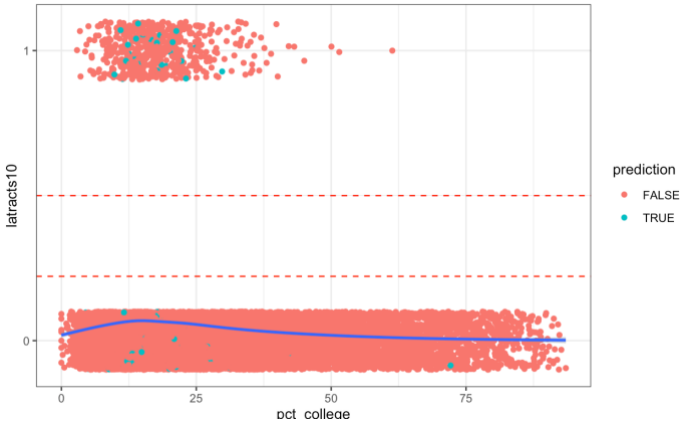
The model was quite unsuccessful in predicting the food accessibility. Shown to the right are the plotted predictions from the kNN model (Note: the blue line represents the probability of a Census Tract being flagged based on all explanatory variables—pct_college is just used in this plot so that the trend is visualizable in 2 dimensions). As is evident, the model never predicts flagging for any observations, which is clearly untrue. This resulted in a low AUC of only 0.6783.



To address this, a weight was introduced to penalize misclassification using the cost of false positives (the

cost of building a supermarket) and the cost of false negatives (the healthcare costs associated with food inequity). Although this weighting did cause some predictions to become 'flagged', it increased the misclassification rate. Whereas it was previously 4.35%, it was now 5.43%.



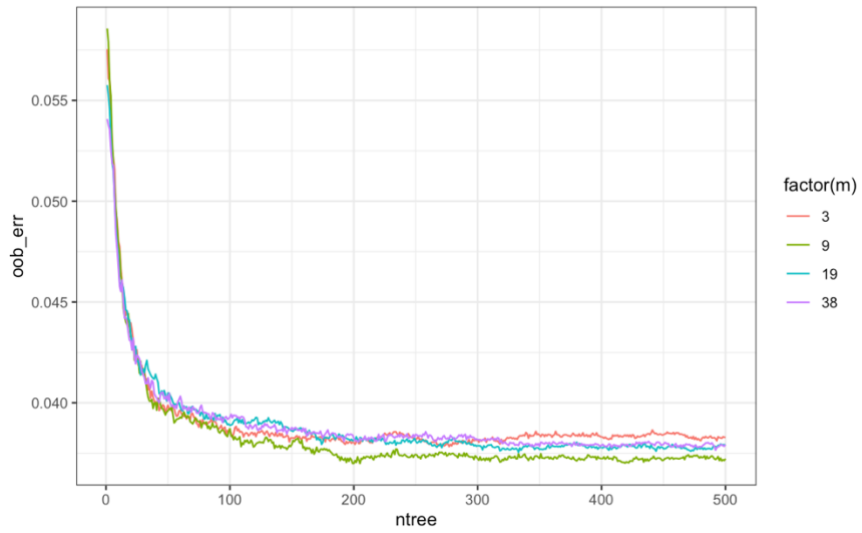


TREE BASED MODELS

Random Forest Model

Random Forest classification models build independent decision trees before combining results at the end of the process. The advantage of their use lies within their ability to grow fully grown decision trees, thus reducing bias. This does, however, come at the cost of a higher variance. As a result, the model creates many uncorrelated trees to maximize the decrease in variance.

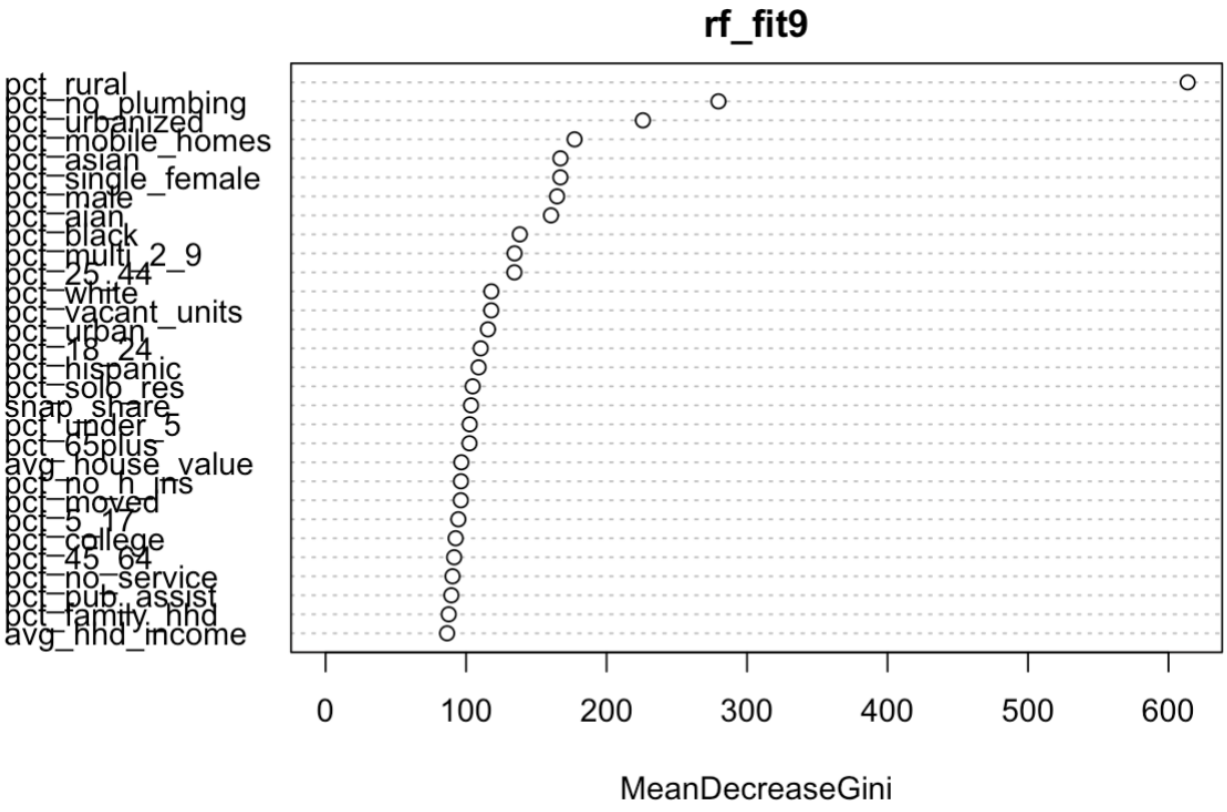
In the creation of this model, first, the optimal m was found through the trial of different values, reducing the possibility of overfitting or underfitting. A trial for many more possible m's was not possible due to process time limitations, but that would have found Out of Bag Error for many more m's selected the m with the lowest OOB error.



From this chart, it is evident that an m of 9 is the OOB-minimizing value. This means that there will be 9 features to consider at each tree split. Next, the ideal number of bootstrap samples, shown by the ntrees axis, was found. Here, it is found that this occurs around B=200, where the OOB error plateaus.

Next, the most Gini-affecting features were found. Gini is a measure of predictive power in explanatory variables. The purity-based importance was not able to found due to processing time limitations.

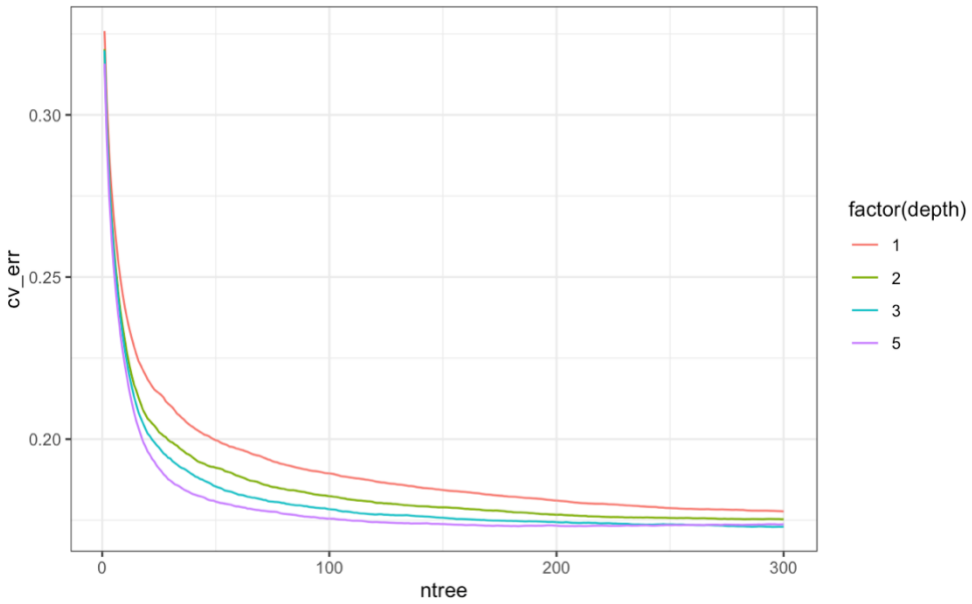
As shown in this plot, rurality was the most important feature in the model, followed by the percent of people who lack access to plumbing and then the percent of individuals living in 'urbanized' areas. Average household income was the least important feature.



Boosted Model

Boosting, in contrast to random forest models, is based on weak learners with high bias and low variance. Boosted models create shallow trees which can even reach as small as decision stumps (trees with two leaves). The algorithm's strategy for reducing error relies on its ability to reduce bias.

In the creation of this model, the first step was to find the optimal tuning interaction depth to minimize CV error.

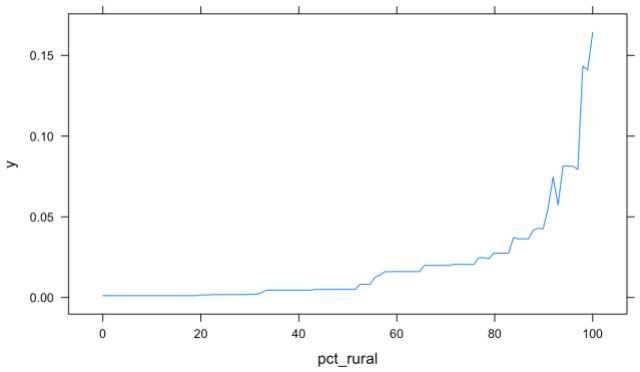


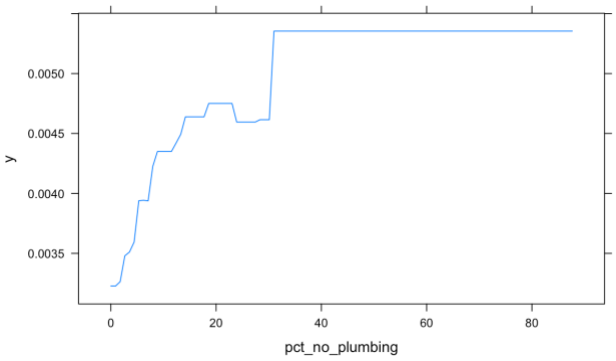
The model building process used the parameters of 300 trees being built at various interaction depths. The corresponding CV error distribution was plotted for each of these. The results showed factor depth 5 being optimal.

The optimal number of trees was later found to be 183. Using the optimal depth and the optimal number of trees grown, we then found the relative imporatnce of each factor. The top 10 variables were as follows:

pct_rural	54.44333331
pct_no_plumbing	7.72699600
pct_aian	4.96136861
pct_single_female	4.14605988
pct_male	4.01041611
pct_mobile_homes	2.98285196
pct_white	2.93830516
pct_black	2.40801039
pct_25_44	2.13547407
pct_solo_res	1.59084171

It is evident that rurality is by far the most important variable at play. The path dependence plots of the two most important features are shown below:





These plots describe the general relationship between the explanatory variable on the x axis and the response variable on the y axis, all other factors held equal. This shows that living in a rural area or an area without plumbing significantly increase a Census Tract's access to food. This makes sense because rural areas do not generally have as many resources made available to them and areas without plumbing are well correlated with low development. What is surprising, however, is that poverty is not an important factor, but lack of plumbing is.

CONCLUSION

Model Comparison

	3 most important variables	Misclassification Error
Logistic Model	Pct_black;Pct_no_plumbing;pct_asian	0.0350
kNN Model	N/A	0.0435
Ridge Model	Pct_25_44;pct_mobile_homes;pct_no_plumbing	0.0409
Lasso Model	Pct_25_44;pct_aian;pct_mobile_homes	0.0363
Random Forest Model	Pct_rural;pct_no_plumbing;pct_urbanized	0.0346
Boosted Model	Pct_rural;pct_no_plumbing_pct_aian	0.0251

The boosted model performed the best, with a misclassification rate of only 2.51%. This makes sense because boosted models and random forest models often produce highly accurate predictions.

A notable comparison is that between the simple logistic model and the ridge and lasso logistic models. Although the ridge and lasso models are a lot more complex and intensive, the logistic model performed better than them both. This is most likely because logistic models perform best with a large ratio of the # of observations / # of features. The Lasso Model performed better than the Ridge model, suggesting that a few coefficients have large influence rather than most coefficients having a small influence.

Unsurprisingly, the kNN model performed the worst. This is because kNN models do much worse than their predictive counterparts when there are many features at play. This is because it becomes difficult for the model to find overarching trends that are not simply based on distance.

There were some pretty surprising differences between the most important variables—each predictive modeling class had its own “most important” feature. Some features, however, such as pct_no_plumbing, pct_moble_homes, and pct_aiain, were found in the top three across model classes.

Overall Conclusions, Recommendations

Based on the findings of this study, I would recommend to policymakers that they subsidize supermarkets or grocery stores in communities with high risk of food inaccessibility. From this analysis, it is evident that there are a few factors that greatly increase the likelihood that a Census Tract has no food shopping options within 10 miles of their home. Based on comparison with previous data, this lack of accessibility has not been permanent for decent amount of the communities affected, meaning that there were previously food options in their area that have since left. Therefore, it is important to identify areas at risk of losing their access to food and taking proactive measures to keep existing grocers in the area. In this way, even the False Positives from the model can be useful, as they show communities share many characteristics with communities flagged at risk. Local governments should institute a “watch period” over Census Tracts deemed at-risk of food inaccessibility, taxing supermarkets and grocery stores if they choose to leave there.

Limitations

The primary technical limitation of the analysis was processing time. This prevented the use of a net elastic regression and the visualization of purity-based importance scores. Another limitation was the lack of food access data by block group. With this data, more granular analysis could have been conducted, and researchers could even find which specific neighborhoods lack access to food the most. This would also increase the number of observations, further increasing the training accuracy of the models.

There are also other features that would be interesting to consider, such as the percent of the population that works in agriculture or crime rates in the area.

Recommended Follow-Up Analyses

This analysis can act as the foundation for many further studies. One such study could understand how certain features can predict lack of food accessibility to children under 18. This can be done already using the data used in this study and would simply require the changing of the response variable.

Another possible follow-up could be seeing how the different “important” features change as one changes the scope of the lack of food in an area. Before deciding on a 10-mile radius, I performed logistic regressions on 0.5-mile, 1-mile, and 20-mile radii as well, and each presented their own “important” features. It would be fruitful to see why these features might change and how each challenge should be approached differently.

Lastly, this analysis can be replicated in other countries to see whether these trends hold true in different cultural, socioeconomic, or political contexts. This understanding would help isolate the root issues that cause food inaccessibility, and different countries could learn from each other’s practices.

APPENDIX

All additional documentation and graphs can be found in the GitHub repository.