Food Desert Data Mining Project

IST 707

Winter 2022

Miranda Braman

Kaitlyn Keebaugh

Matthew Pergolski

Victor Yamaykin

Table of Contents

[Introduction 3](#_Toc97583348)

[The Data 3](#_Toc97583349)

[Exploratory Data Analysis 4](#_Toc97583350)

[Models 4](#_Toc97583351)

[Clustering 4](#_Toc97583352)

[Association Rule Mining (ARM) 4](#_Toc97583353)

[Decision Trees 5](#_Toc97583354)

[Results 6](#_Toc97583355)

[Conclusions 6](#_Toc97583356)

# Introduction

The Healthy Food Financing Initiative (HFFI) defines food deserts as low-income and low-access census tract areas. Low-income means that there is a poverty rate of more than 20% or below 80% of the statewide median family income. Low-access means at least 500 people and/or at least 33% of the population lives more than 1 mile from a large grocery store (10 miles, in the case of rural areas).1 The United States Department of Agriculture uses data from the census every 5 years to build their Food Access Research Atlas. There is labeled data for 2006, but not for 2010, 2015, or 2019.

This analysis sets out to answer two prominent questions in terms of racial justice:

* Are food deserts growing or shrinking and where?
* Can classification help to identify food deserts and learn about the surrounding communities?

# The Data

With the data sets for 2006 and 2019 combined, there are over 160 attributes. In exploratory data analysis, there were 10 attributes chosen, including the target for food desert. In the final analysis, additional attributes for ethnicity will be added.

* Census Tract ID
* State
* County
* Population in 2010
* Occupied Housing Units (OHU) in 2010
* If people live in an urban area
* If they live in a low-income area
* Poverty rate
* The median family income in the community
* If it is a food desert or not

There were several methods of cleaning the data for each classification algorithm.

* 1. Subsection for each variable

# Exploratory Data Analysis

EDA for each variable

# Models

## Clustering

Following the exploratory analysis section, the unsupervised learning method of ‘Clustering’ was chosen to determine potential groupings / subsets within the ‘Food Desert’ data set. Ideally, various states would be grouped together to determine potential similarities, meaningful structure, generative features, etc.

To begin the clustering process, the data set needed to be transformed in terms of aggregating observations by the ‘State’ variable. State data was then converted from a basic column/variable to a row name of the data frame. Non-numeric variables were then omitted from the analysis (i.e., ‘State,’ ‘CensusTract’, and ‘County’).

An additional analysis step was conducted to determine the optimal amount of clusters to use for the data. All in all, the ‘Average Silhouette Method’ and its respective R/RStudio functions determined the ‘best’ number of clusters to choose was 3 via the ‘fviz\_nbclust’ function from the ‘factoextra’ package.

Chart, line chart

Description automatically generated

Following this munging/preliminary step in the process, various models and functions were prepped for use within the analysis section of the overall ‘Clustering’ topic; for the hierarchical aspect, ‘agnes’ and ‘hclust’ algorithms were chosen. Before running the respective algorithms, an overall method was determined before use. Available options for these methods in determining distance are listed as follows:

* Average
* Single
* Complete
* Ward

In order to determine which method would serve best for the analysis, a custom function was developed to generate, compute, and compare agglomerative coefficients (which conveys overall quality and fit for the clustering algorithm). In ranking from best to worst, the ‘ward’ method indicated ~0.88, followed by ‘complete,’ ‘average,’ and ‘single which generated results of ~0.85, ~0.73, and ~0.60, respectively.

Since ‘ward’ yielded the highest agglomerative coefficient value, this method was chosen for the ‘agnes’ and ‘hclust’ algorithms. With keeping 3 clusters in mind, the following dendrogram was generated (with state information shown near the bottom of the figure):

A picture containing diagram

Description automatically generated

Lastly, the ‘K-Means’ algorithm was also chosen as a viable algorithm for the ‘Clustering’ topic. The three (3) overall clusters consisted of 16, 3, and 31 observations (which equate to 50 – the total number of US states). The ‘Sum of Squares’ values by cluster equate to the following:

* ~75.7
* ~11.5
* ~108.9

The algorithm is also accompanied by a visualization via the ‘fviz\_cluster’ function of the same ‘factoextra’ package that was introduced earlier in this section.

Chart

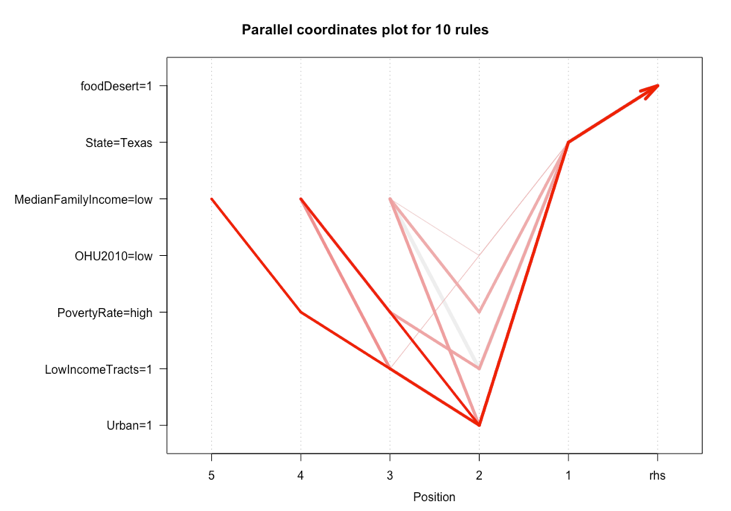
Description automatically generated

As shown in the Cluster Plot, three (3) distinctive clusters are formed, containing US state information. The larger the geometric point, the higher the mean poverty rate for the state. These subsets of states may lead to valuable classification in latter stages of the report when supervised learning methods are introduced to the audience. We may be able to preliminarily state that clusters with higher poverty rates may be correlated with food deserts.

## Association Rule Mining (ARM)

To use the apriori algorithm, some of the attributes had to be discretized; in particular, the data for population, occupied housing units, poverty rate, and median family income were put in categories such as “low,” “medium,” and “high.” Another issue was the imbalanced data set. The method of “over” sampling helped to get a similar number of observations for food desert or not. The rules were sorted by support, lift, and confidence. The best results were achieved with support = 0.25, confidence = 0.7, and minimum length = 3. Here is an example of the rules found:

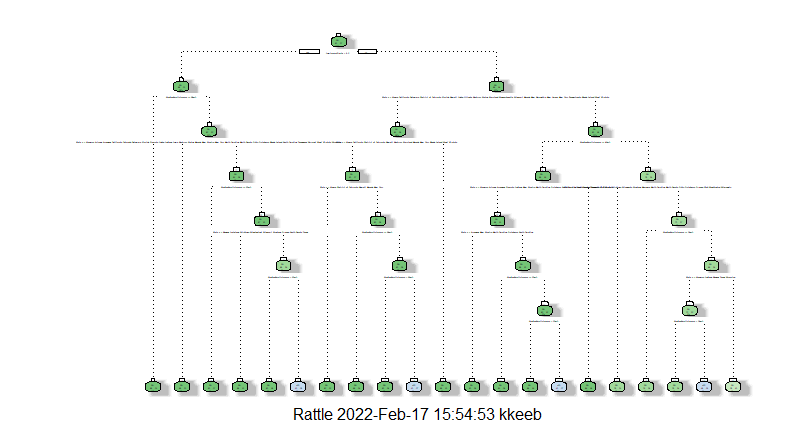
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Left-hand side | Right-hand side | Support | Confidence | Coverage | Lift | Count |
| State=Texas, Urban=1,  LowIncomeTracts=1,  PovertyRate=high,  MedianFamilyIncome=low | foodDesert=1 | 0.039 | 0.82 | 0.047 | 1.6 | 5203 |



## Decision Trees

Decision trees are supervised learning because they can build a model using a training data set and then predict outcomes based on a testing data set. These decision trees are the classification type because they are deciding whether an area is a food desert or not. Since this data set is so big, to split it into a training and testing set, only some variables were chosen. They include Median Family Income, Low Income Tract, and States. The data was split 80/20 training and testing.

The function prop.table was used to make sure each data set had the same amount of yes and no food desert outcomes. Without scaling down the variables and the function, this tree could get too complex and could start overfitting. To try and prevent this, a set.seed of 341 was used and a control function of max depth was used to make sure the longest tree branch was only 7 nodes long. This produced great accuracy, precision, and recall. The next decision trees will be smaller and more specific, so there is a better model for predicting where the food deserts really are.



# Results

discuss results, issues, and limitations for all the analysis. Note which ones worked well, and why, which ones did not, and why, etc.

# Conclusions

What was the outcome - what did you find, discover, predict, classify? Why does it matter to humans?

So far, the results of the classification algorithms confirm the low-income and high poverty rate are strong indicators of a food desert. In the next iteration of the research, there will be additional attributes related to low access and ethnicity.