Predicting Food Deserts in the United States

Natalie Cygan, Grace Kim, Priscilla Lui

Motivation

- In 2010, the USDA reported that over 23 million Americans lived in areas with limited access to affordable and nutritious food, known as food deserts.
- This limited access to affordable and nutritious food has significant health impacts, particularly for children, as processed sugar- and fat-laden foods contribute to health epidemics such as obesity.[1]
- **Goals**:

  1) Build a predictor to classify areas likely to become food deserts.

  2) Identify which features are the best predictors of food deserts.

Figure 1: Food deserts in the US (USDA Food Access Research Atlas)[2]

Feature Engineering

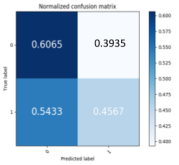
- **Prediction labels:** Food desert or Not food desert. *(Source: 2015 USDA Food Access Research Atlas)*
- **Features:** Ten features were socioeconomic features of 12,060 zip codes in 2015. Ten additional features were the percent changes for those same socioeconomic features from 2012 to 2015. *(Source: US Census)*
- **Final dataset**: 12,060 labeled data points with 20 features each (2337 of which were food deserts).

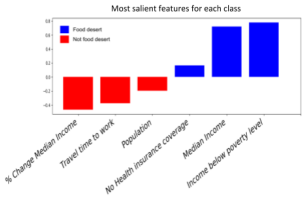
Sample Census Features	
Unemployment rate	% Change in unemployment rate
Median rent	% Change in Median rent
College Education Rates	% Change in College Education Rates
Geographic mobility	% Change in Geographic mobility
...	...

Model Rankings

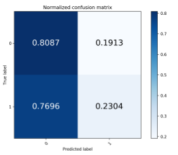
We tried 4 different models and evaluated them with respect to our goals:

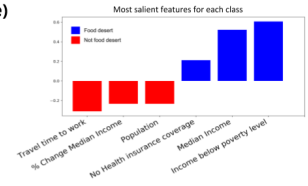
Support Vector Machine



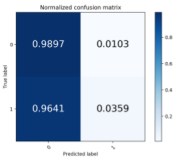


Logistic Regression (baseline)





Neural Network



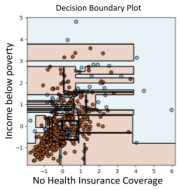
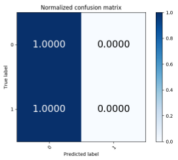
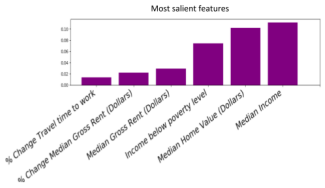


Figure 2: Plot of decision boundaries using a Decision Tree Classifier. The axes are the two most salient features, which were chosen from our baseline model.

Random Forest Classifier





Results & Analysis

F1-Score Comparison

Classifier	'Not Food desert'	'Food desert'
SVM	0.70	0.30
Logistic Regression	0.81	0.23
Neural Network	0.89	0.07
Random Forest	0.89	0.0

- For all models, the F1-Scores for 'Food desert' were much lower than those of 'Non Food desert' because the dataset was unbalanced.
- SVM and Logistic Regression performed well, perhaps due to balancing of class weights.

Most Important Features:

- Income below poverty level
- Median income, % Change in median income
- Number of people w/o health insurance coverage

Next Steps to Address Challenges:

- Experimenting with more ways to balance data
- Training NN to have higher precision
  Attempts thus far:
  - Determining the # of NN Layers and Nodes
    - Decision Tree Classifier to find ~# of hidden layers (see fig. 2)
    - Grid Search to iterate through permutations of # of layers and nodes per layers
  - Unbalanced dataset:
    - 20% was labeled as 'Food desert'
    - Tried over-sampling, and balancing class weights
- If time permits, capturing more historical trend information.

References

[1] "Living in a Food Desert: How Lack of Access to Healthy Foods Can Affect Public Health | Notes From NAP". notes.nap.edu. 2011-01-25.
[2] United States Department of Agriculture-Economic Research Service. Food Access Research Atlas (formerly known as the Food desert locator). ers.usda.gov/data-products/food-access-research-atlas.aspx.