Studying Data from the Food Environment Atlas - Group 1

Erdos Institute Data Science Bootcamp, May 2024

Executive Summary

The purpose of this project is to explore relationships in the Food Environment Atlas' dataset. We have trained models to predict obesity, food insecurity, and persistent poverty. Many communities are still struggling with food insecurity, and much of the population does not even have convenient access to a grocery store. We hope that our analyses can give some insight into the complexities surrounding food access.

Our Team

Craig Corsi, Omeiza Olumoye, Tatum Rask, Sayantan Sarkar

Introduction

The Food Environment Atlas is an aggregated dataset and visualization web app consisting of U.S. county, state, and regional data on over 280 indicators pertaining to food consumption, health, and community. The atlas comprises nine categories: Access and Proximity to Grocery Stores, Store Availability, Restaurant Availability and Expenditures, Food Assistance, Food Insecurity, Food Taxes, Local Foods, Health and Physical Activity, and Socioeconomic Characteristics. Data from as early as 2007 and as late as 2018 is included, and many indicators have data from multiple years. The atlas' visualizer allows users to observe data visually and develop an intuition for the complexities surrounding endemic hunger and illness.

Problem

Our goal is to predict multiple outcomes: adult obesity rate in 2017 (variable code 'PCT_OBESE_ADULTS17' in the dataset), household very low food security from 2015-2017 ('VLFOODSEC_15_17'), and the persistent-poverty category ('PERPOV10').

Data preprocessing and cleaning

In our exploratory data analysis, we considered indicators that were collected up to two years prior to a given response. That is, any data collected between 2015 and 2017 were considered as features for adult obesity rate and household food insecurity. As an exception to this, we also considered indicators which typically persist over multiple decades, such as whether a given county is considered to be a metro area ('METRO13').

Missing data values for county data were imputed with either the national or state mean value, while state data was extrapolated from county data by taking a weighted average. We also included external data consisting of county population estimates and latitude/longitude, both from the U.S. Census Bureau. We updated the data for Kusilvak Census Area, AK and Oglala Lakota County, SD, with the current names and FIPS identifiers for these counties. Moreover, we combined the entries for Bedford County, VA and the former independent city, Bedford, VA, and recalculated data values to represent Bedford's reversion to being a town within Bedford County in 2013.

For county data, we also implemented a custom train-test split and k-fold cross-validation routine that stratifies the data geographically. This is because we wanted to stratify our splits by multiple categorical variables including state, but stratifying by state introduced imbalances in other variables. Our custom splits ensure that the closest neighboring counties to a county in the test set are in the training set, while still maintaining the same relative frequency of non-geographic categorical features.

Models

Obesity: Five input features were chosen to train these models based on lasso regression. We trained a multiple linear regression model and a k-nearest-neighbors model (k = 5), with average MSEs 5.975 and 8.763, respectively, on the holdout sets. These performed better than the baseline model, which predicts the mean obesity rate and has an average MSE of 16.43 on the holdout sets. Since multiple linear regression performed the best, we focused on that model.

Very Low Food Security: We have chosen a few features based on their correlation score with the target variable. Multiple linear regression, decision tree and random forest regressor have been used to find out which features are responsible for very low food security. The best model, a random forest model trained on a set of ten features, had an MSE of 0.006384 and an R2-score of 0.994251.

Persistent-Poverty: We trained logistic regression, LDA, QDA, and random forest models and used a random classifier as a baseline model. Three instances of each model were trained on sets of 7, 18, and 20 features. We only considered models which labeled 11% or more of the counties as persistent-poverty counties, based on the actual persistent-poverty rate of 11.1% in 2010. The optimal classifier was a random forest classifier trained on 18 features, with a prediction probability threshold of 0.27. The average accuracy score was 91.68%, but on the test set it only labeled 9.84% of counties as persistent-poverty counties, so the accuracy score increased to 92.06%.