Springboard Data Science Capstone Project

Correlations between poverty, food environment and diabetes in the United States

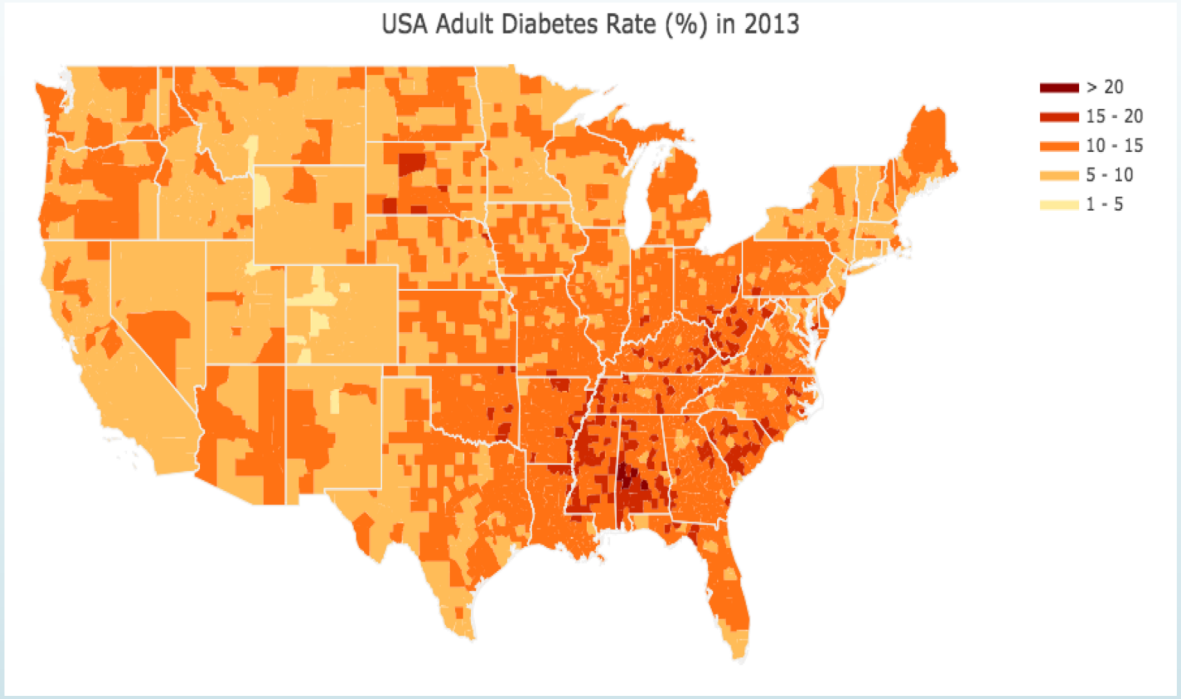
Shuangshuang Liu

January 2020

1. Background

- Globally, by 2014 there are 422 million people with diabetes, up from 108 million in 1980;
- In the United States, more than 30 million people (all ages, 2015) have diabetes;
- In the last 20 years the number of adults diagnosed with diabetes has more than doubled as the American population has aged and become more overweight or obese.

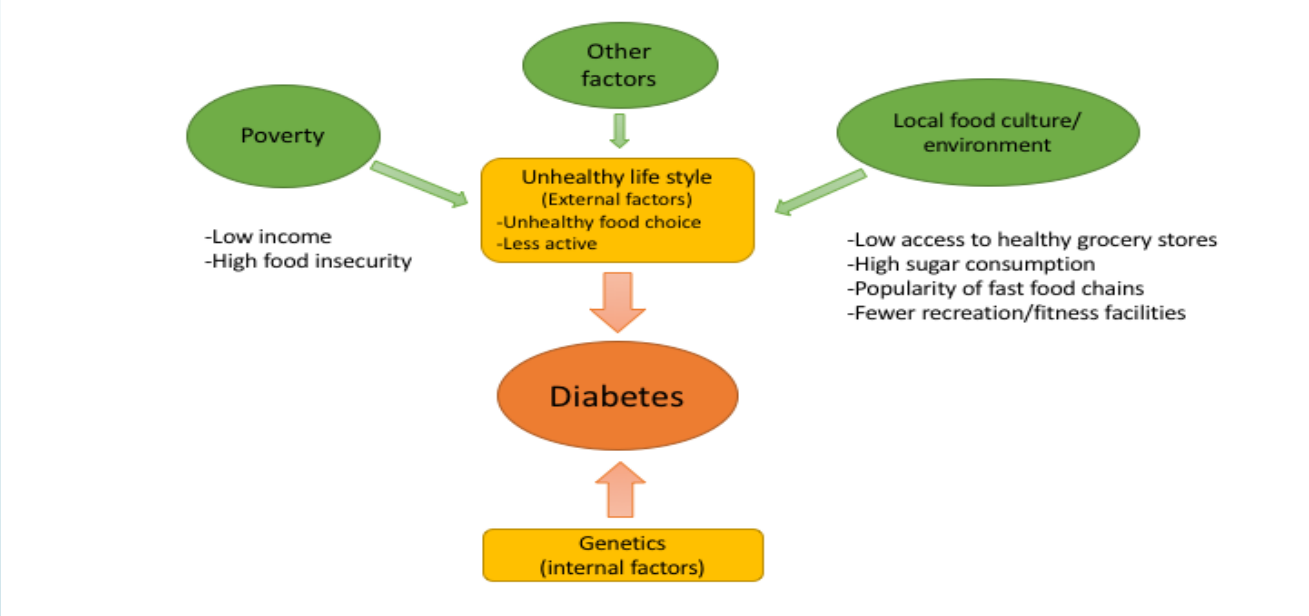
(Diabetes quick facts, Centers for Disease Control and Prevention)



Data source: Food Environment Atlas (https://www.ers.usda.gov/data-products/food-environment-atlas/)

- Measures to prevent or delay the onset of type 2 diabetes include healthy diet, regular physical activity, maintaining a normal body weight and avoiding tobacco use (International Diabetes Federation).

- We hypothesized that **poverty level** and **local food culture/environment**, as measured by factors such as access to grocery stores, among other possible factors, may contribute to **unhealthy life style** and thus be associated with diabetes prevalence. The hypothesis can be summarized in the following diagram.



Specific questions addressed:

1) Are there significant differences in adult diabetes rate between metro and non-metro counties (i.e., urbanity, which is related to access to different types of stores), as well as poverty-persistent and non-poverty counties?

2) What are the most important features that may contribute to diabetes rate? Among those features which ones are more significantly correlated with diabetes rate, indicators of poverty level or food environment?

3) Are there any patterns of food environment across U.S. counties?

2. Methods

Dataset

The "Food Environment Atlas" database will be used for analysis: (https://www.ers.usda.gov/data-products/food-environment-atlas/)

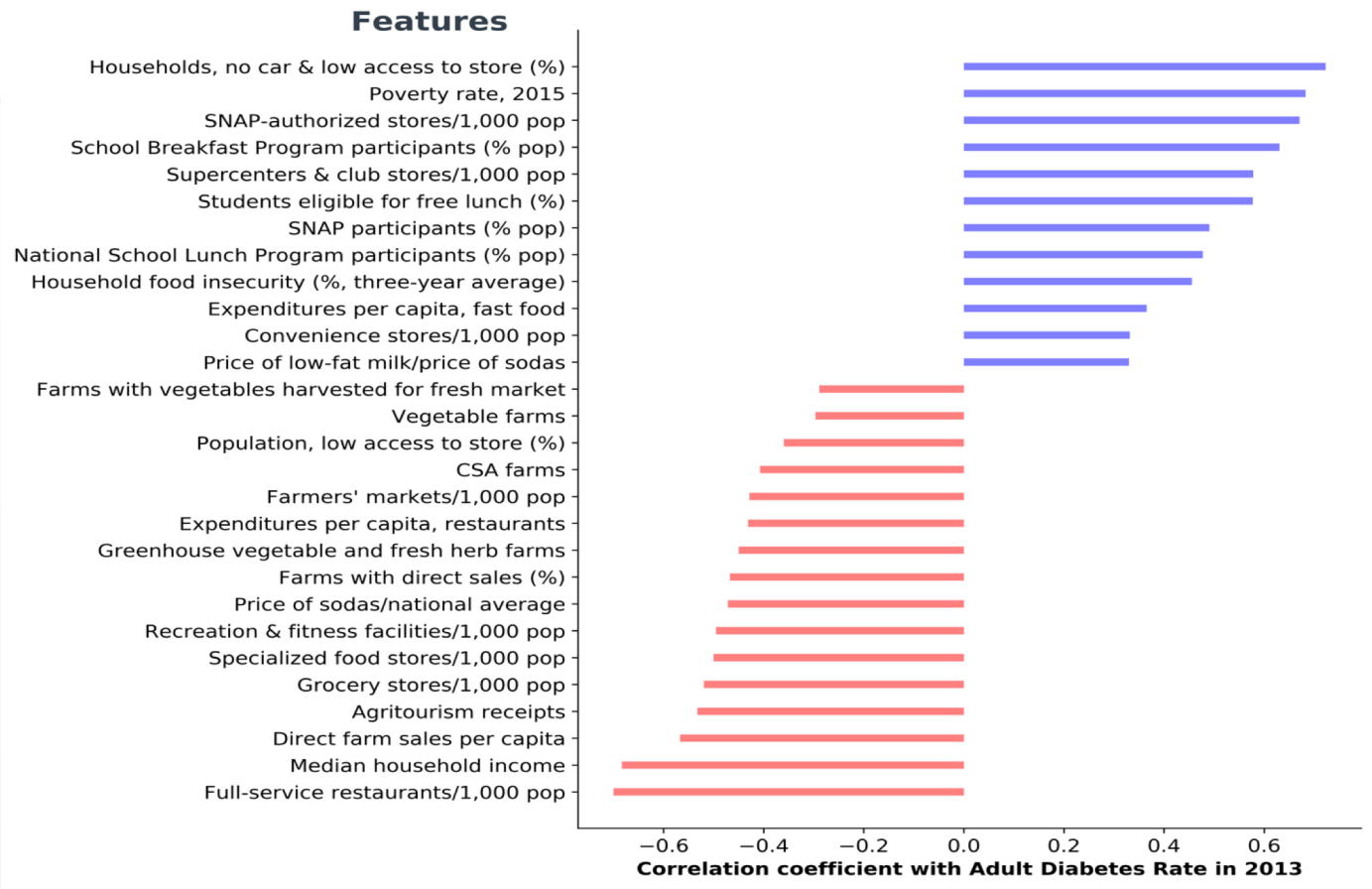
Statistical Analysis and Modeling

1) Mean comparisons evaluating the effects of urbanity and poverty

2) Predicting diabetes with food environment features (supervised learning)
   - Multiple Linear Regression
   - Decision Trees
   - Random Forest
   - XGBoost

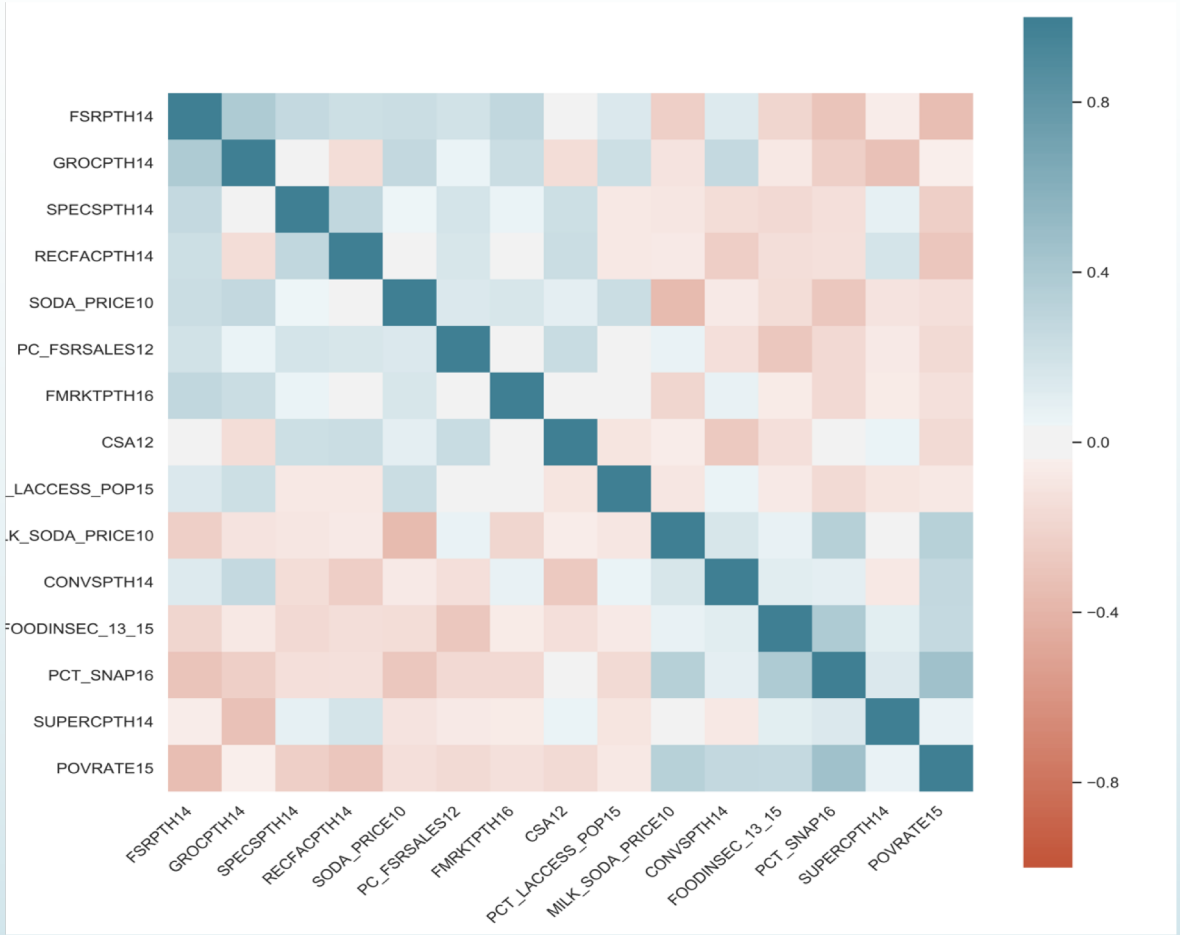
3) Finding patterns of food-economic environment among counties (unsupervised learning)
   - Principal component analysis (PCA)
   - K-means clustering

Feature selection: 28 initial features were selected



Simple linear regression correlation coefficients between 28 Selected features against adult diabetes rates in 2013 based on state average values ($p < 0.05$, unadjusted for multiple testing). When data of a feature for multiple years are available, only the most recent year is considered.

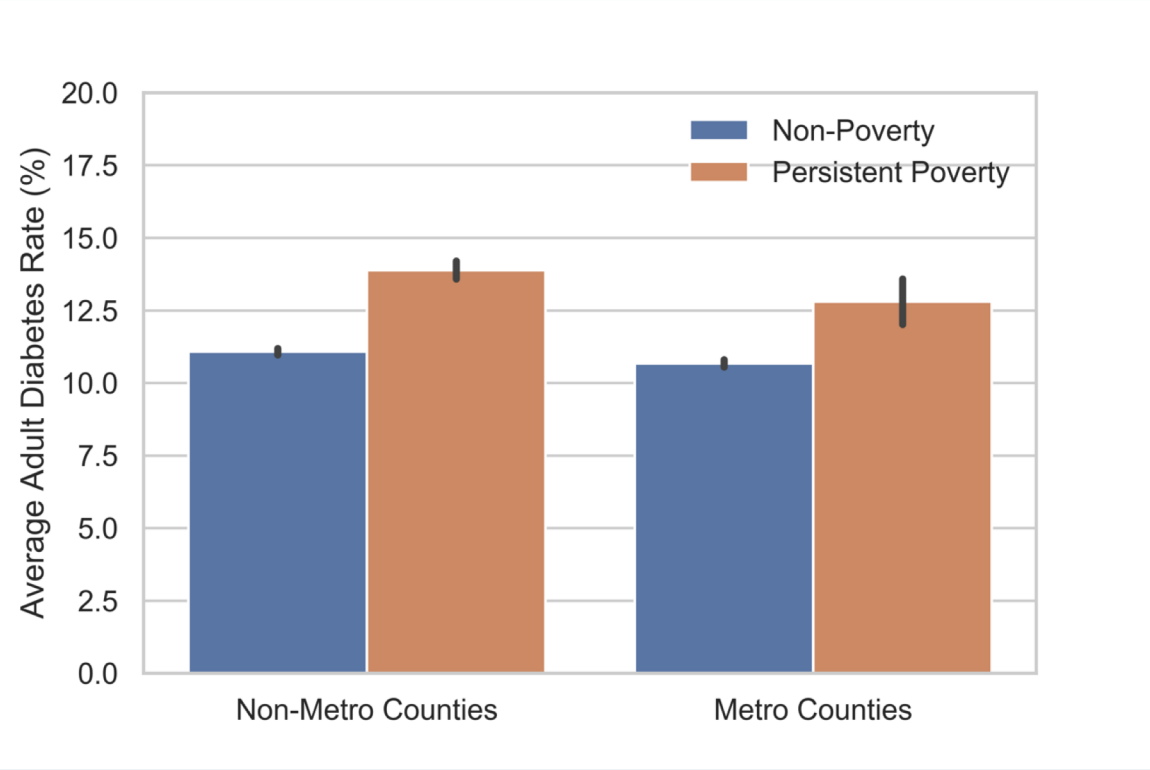
After removing highly correlated features, 15 final features were selected for statistical analysis



Correlation matrix among the 15 selected features from the food environment atlas.

3. Results

3.1 Both urbanity and poverty had significant effects on adult diabetes rates among the counties ($p<0.05$ for both cases).

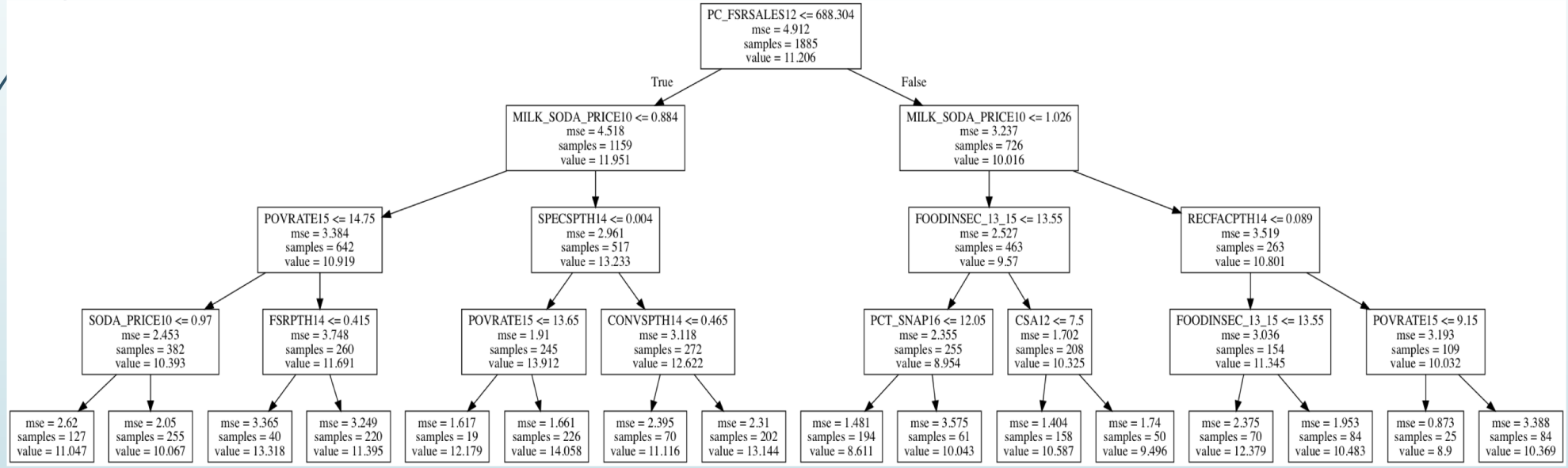


3.2 Random forest model is choosing as the final model after comparing **several models** for predicting diabetes rates.

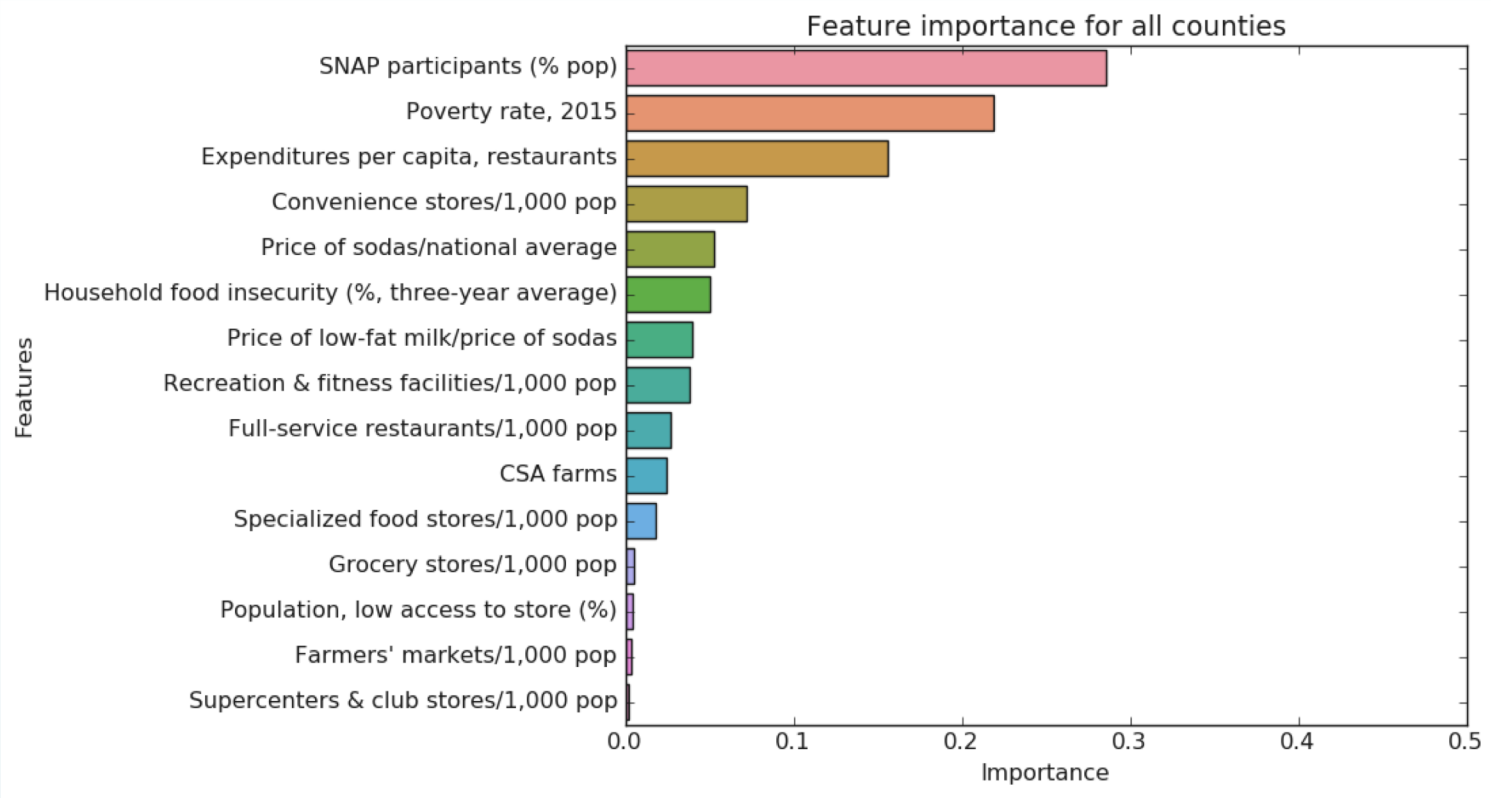
Model (All counties)	Training dataset R^2	Testing dataset R^2
Linear regression	0.51	0.56
Decision tree	0.55	0.50
Random forest	**0.68**	**0.63**
XGBoost	0.83	0.69

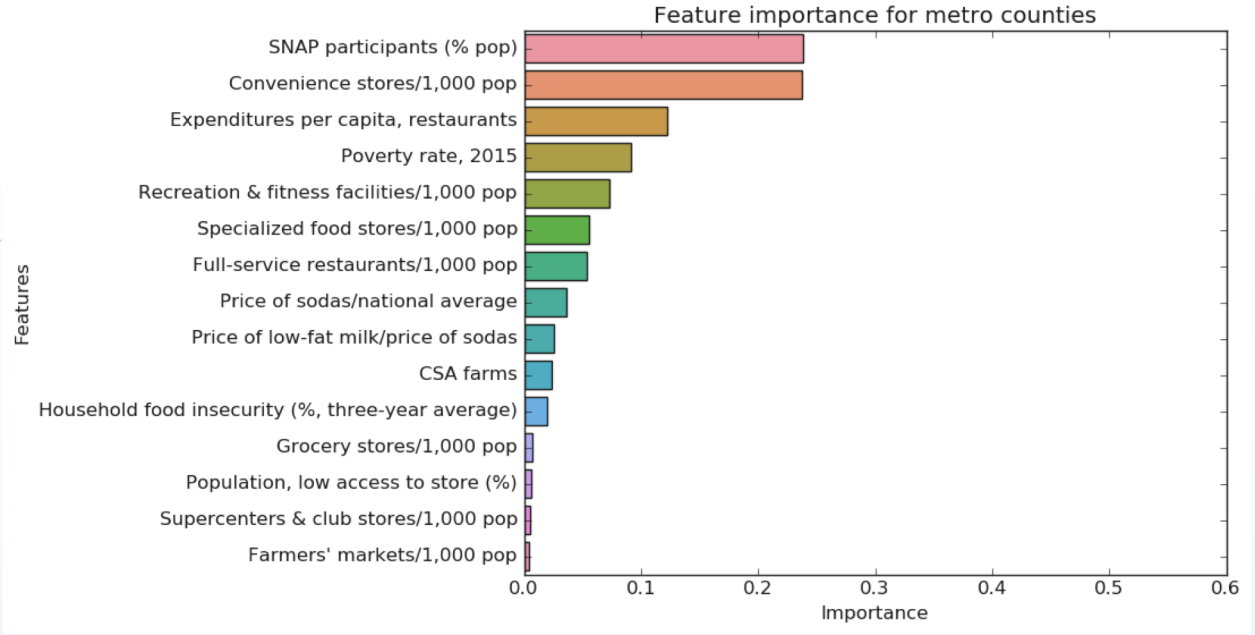
Decision Tree results:

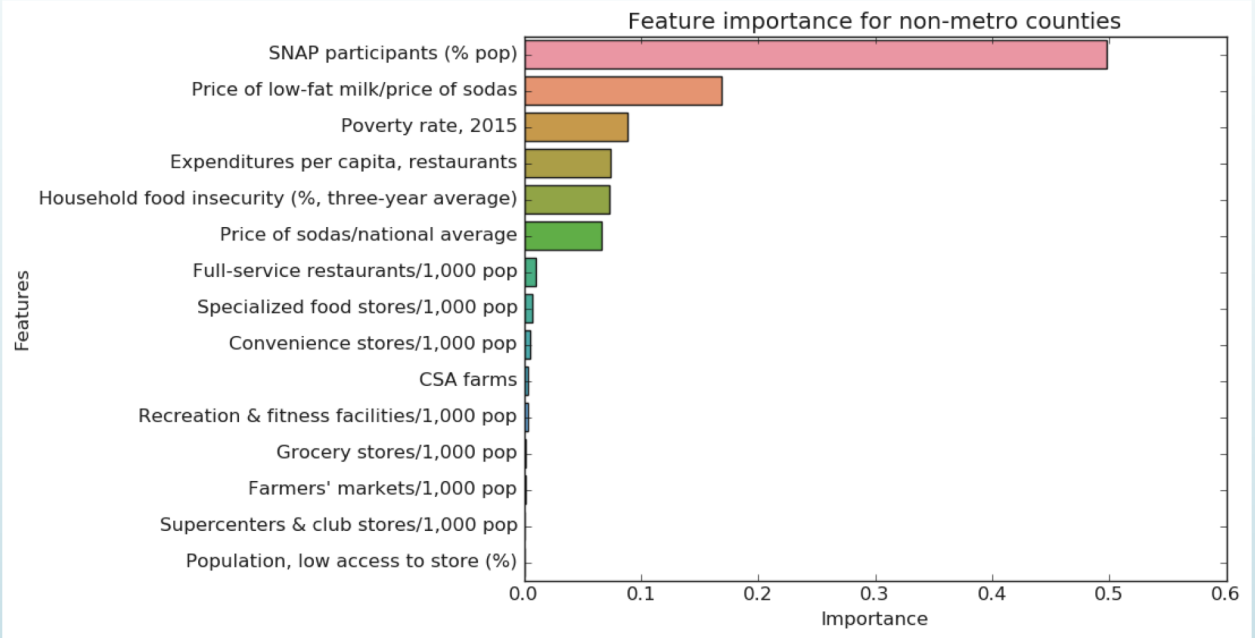
- The most important feature (root node) is '**restaurant expenditures per capita**' (PC_FSRSALES12).
- 2nd split:
  Price of low-fat milk/price of sodas (MILK_SODA_PRICE10)
- 3rd split:
  Poverty rate (POVRATE15)
  Specialized food stores/1,000 pop (SPECSPTH)
  Food insecurity, three year average (FOODINSEC)
  Recreation & fitness facilities/1,000 pop (RECFACPTH14)

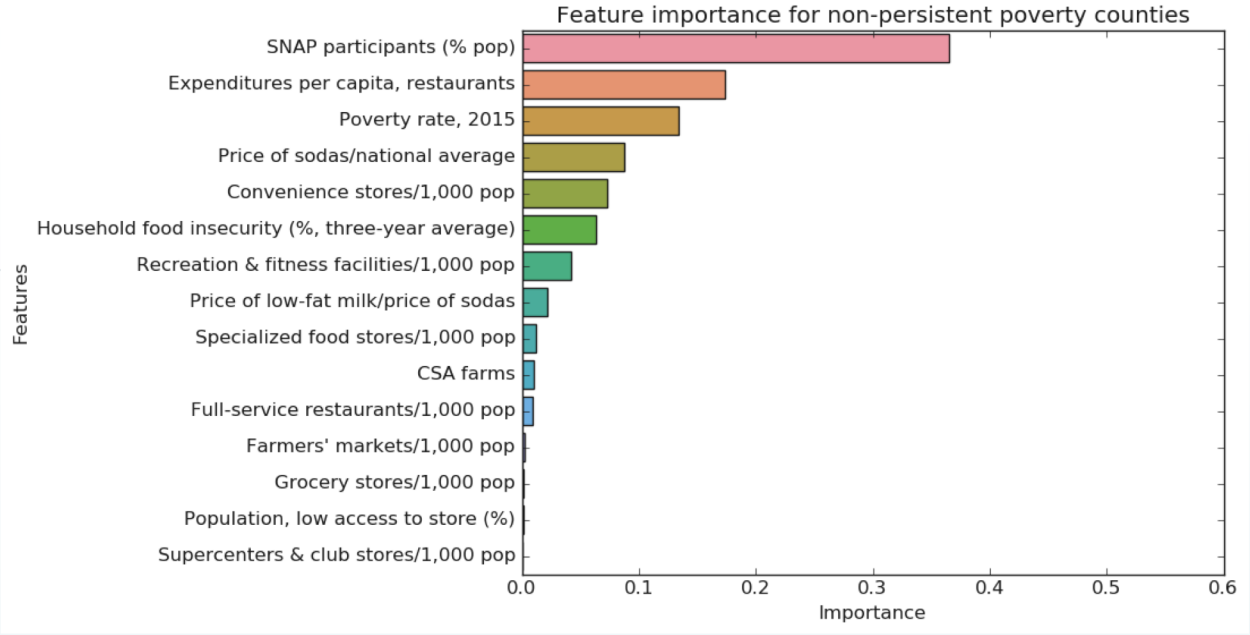


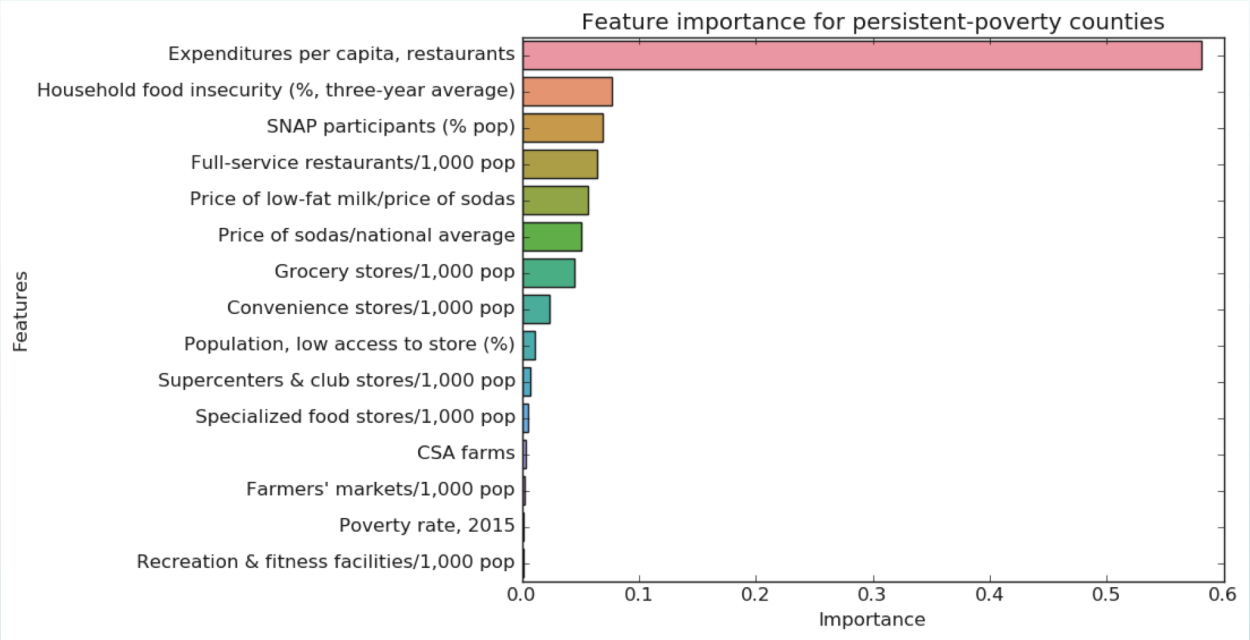
Random forest model results: ranking of feature importance

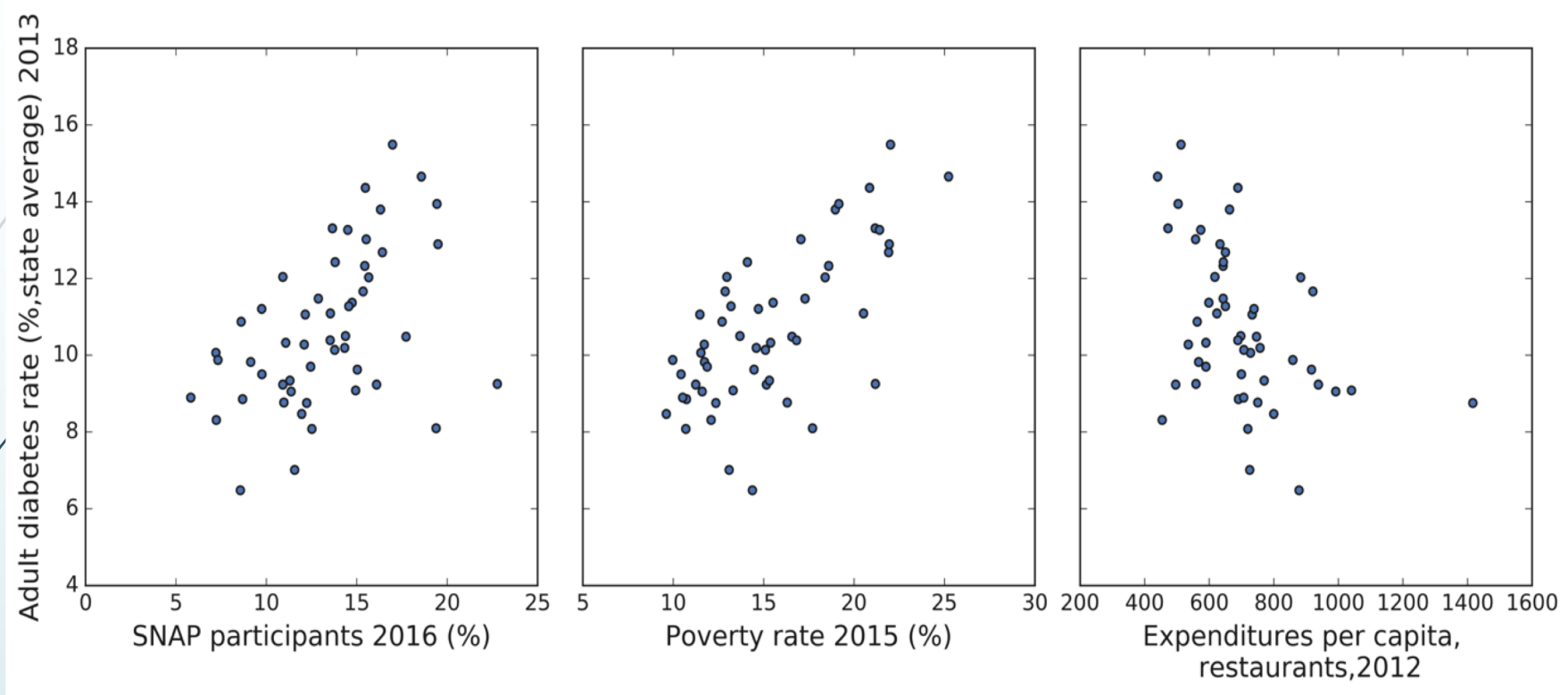



Feature importance for metro counties
SNAP participants (% pop)
Convenience stores/1,000 pop
Expenditures per capita, restaurants
Poverty rate, 2015
Recreation & fitness facilities/1,000 pop
Specialized food stores/1,000 pop
Full-service restaurants/1,000 pop
Price of sodas/national average
Price of low-fat milk/price of sodas
CSA farms
Household food insecurity (%, three-year average)
Grocery stores/1,000 pop
Population, low access to store (%)
Supercenters & club stores/1,000 pop
Farmers' markets/1,000 pop
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance


Feature importance for non-metro counties
SNAP participants (% pop)
Price of low-fat milk/price of sodas
Poverty rate, 2015
Expenditures per capita, restaurants
Household food insecurity (%, three-year average)
Price of sodas/national average
Full-service restaurants/1,000 pop
Specialized food stores/1,000 pop
Convenience stores/1,000 pop
CSA farms
Recreation & fitness facilities/1,000 pop
Grocery stores/1,000 pop
Farmers' markets/1,000 pop
Supercenters & club stores/1,000 pop
Population, low access to store (%)
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance


Feature importance for non-persistent poverty counties
SNAP participants (% pop)
Expenditures per capita, restaurants
Poverty rate, 2015
Price of sodas/national average
Convenience stores/1,000 pop
Household food insecurity (%, three-year average)
Recreation & fitness facilities/1,000 pop
Price of low-fat milk/price of sodas
Specialized food stores/1,000 pop
CSA farms
Full-service restaurants/1,000 pop
Farmers' markets/1,000 pop
Grocery stores/1,000 pop
Population, low access to store (%)
Supercenters & club stores/1,000 pop
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance


Feature importance for persistent-poverty counties
Expenditures per capita, restaurants
Household food insecurity (%, three-year average)
SNAP participants (% pop)
Full-service restaurants/1,000 pop
Price of low-fat milk/price of sodas
Price of sodas/national average
Grocery stores/1,000 pop
Convenience stores/1,000 pop
Population, low access to store (%)
Supercenters & club stores/1,000 pop
Specialized food stores/1,000 pop
CSA farms
Farmers' markets/1,000 pop
Poverty rate, 2015
Recreation & fitness facilities/1,000 pop
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance



Correlations between the top three features and adult diabetes rate (state averages), based on random forest model.

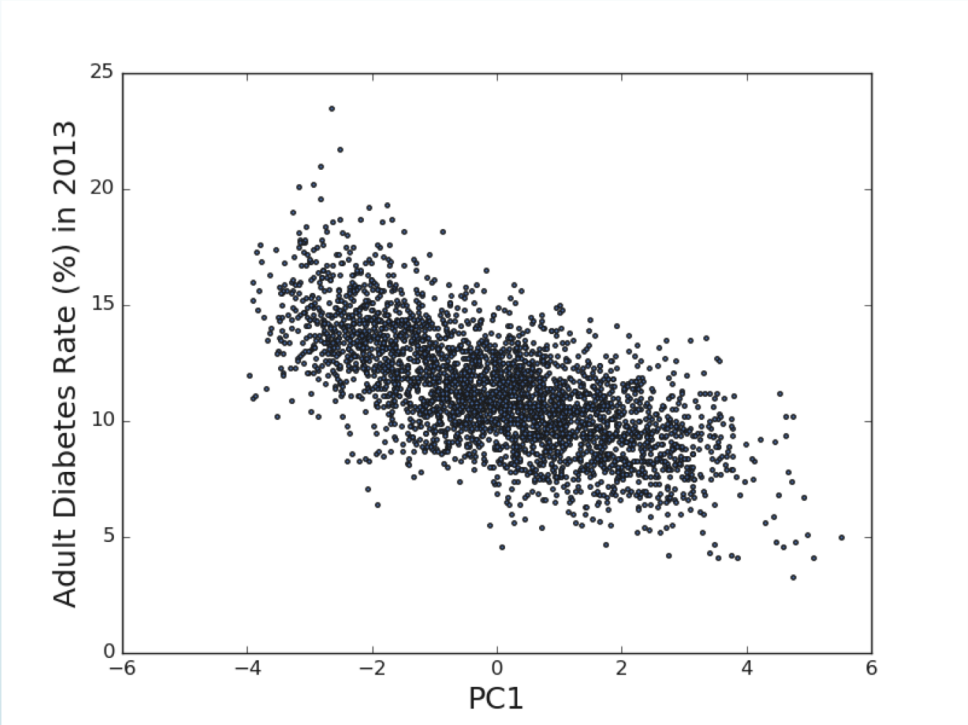
Rankings of top features by random forest model when considering all counties, urbanity or poverty level.

Feature name	All counties	Metro counties	Non-Metro counties	Non-persistent poverty counties	Persistent poverty counties
SNAP participants (% pop)	1	1	1	1	3
Poverty rate	2	4	3	3	14
Expenditures per capita, restaurants	3	3	4	2	1
Convenience stores/1000 pop	4	2	9	5	8
Price of low-fat milk/price of sodas	7	9	2	8	5
Household food insecurity (%, three-year average)	6	11	5	6	2

3.3 The landscape of food environment across the United States

PCA results

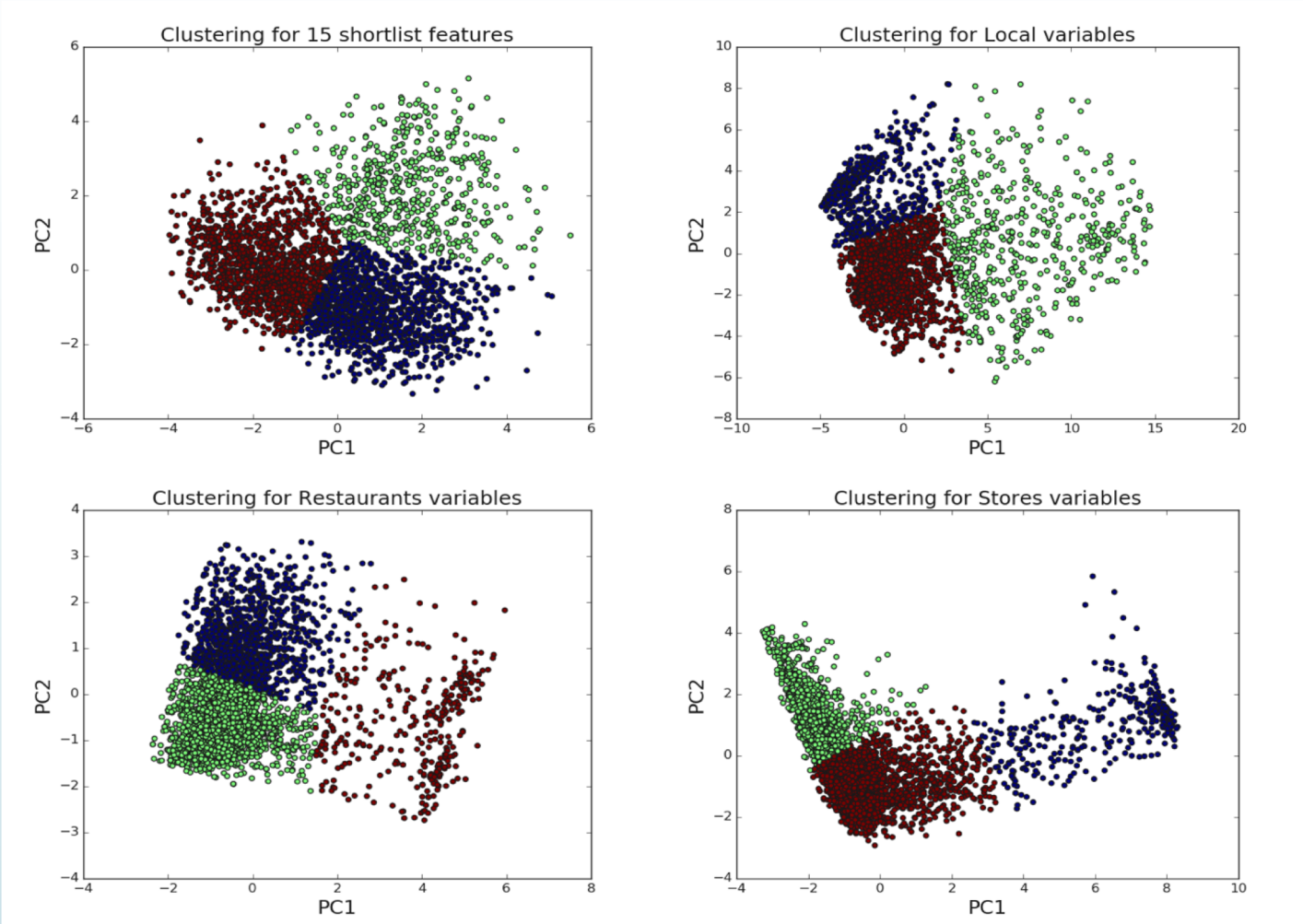
For the group of 15 shortlist features, the first four PCs explained 51% of variance.
For the 'Local' group, first three PCs explained 51% variance; for the 'Restaurants' and 'Stores' groups, the first two PCs explained 61% and 66% variances, respectively.



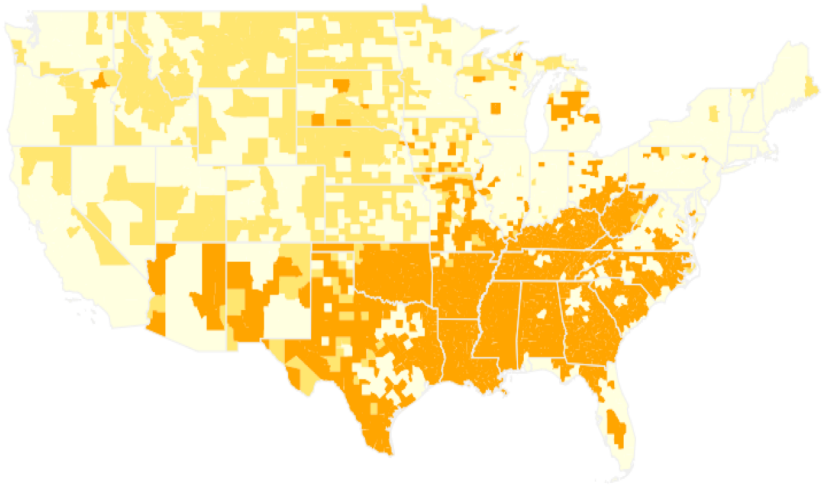
Correlations between PC1 of 15 shortlist features and adult diabetes rate.

3.3 (Cont.)

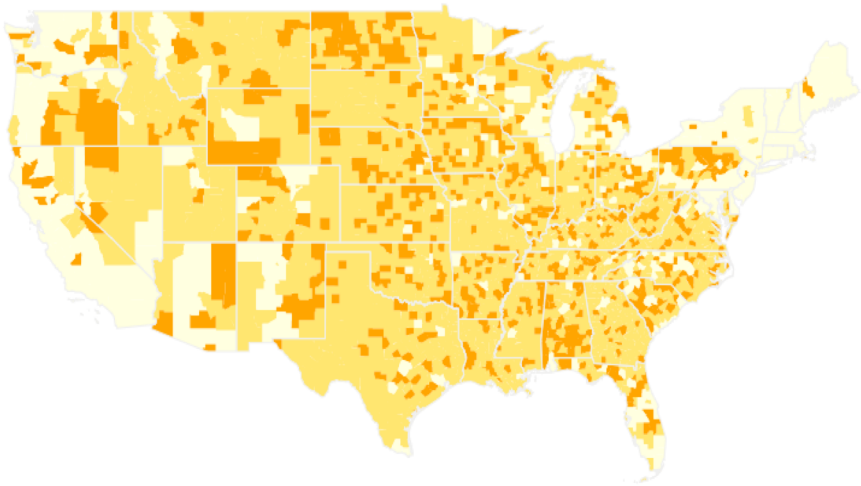
K-means clustering results (with k=3, on data of the first two PC components from PCA on the 15 shortlist features, and groups 'Local', 'Restaurants', and 'Stores').



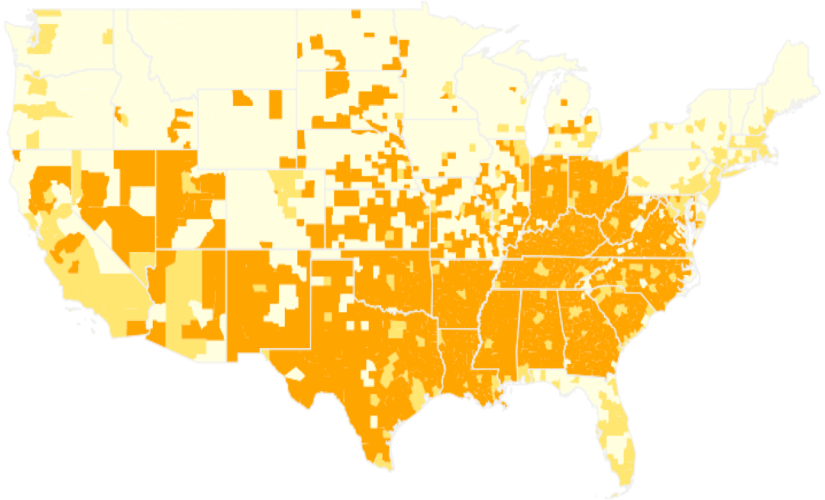
Three clusters among U.S. counties based on 15 shortlist features



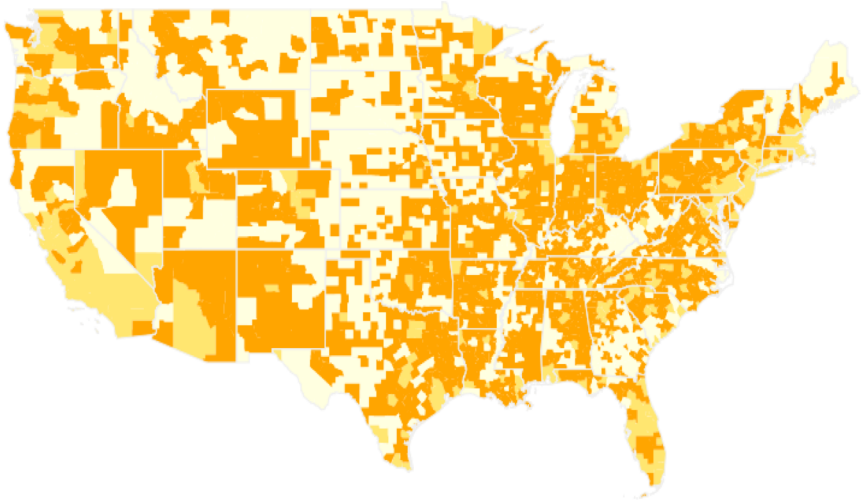
Three clusters among U.S. counties based on Local features



Three clusters among U.S. counties based on Restaurants features



Three clusters among U.S. counties based on Stores features



4. Conclusions

- Features related to **economic status** generally played the most important role in predicting diabetes rate. The top ranked features included **SNAP participants**, **poverty rate** and **expenditures at restaurants.**

- Among persistent-poverty counties, the expenditures at restaurants became the most predictive feature.

- PCA on 15 diabetes-predicting features revealed three major clusters of U.S. counties: the **central south and southeast** (higher poverty and diabetes rate), **west and east coast and the great lakes area** (lower poverty rate and diabetes rate), and the rest counties.

- Counties at central south and southeast are high-risk areas and should be allocating more resources to improve the prevention and treatment of diabetes.