Predicting the Prevalence of Obesity in U.S. Counties

Joseph Mintz

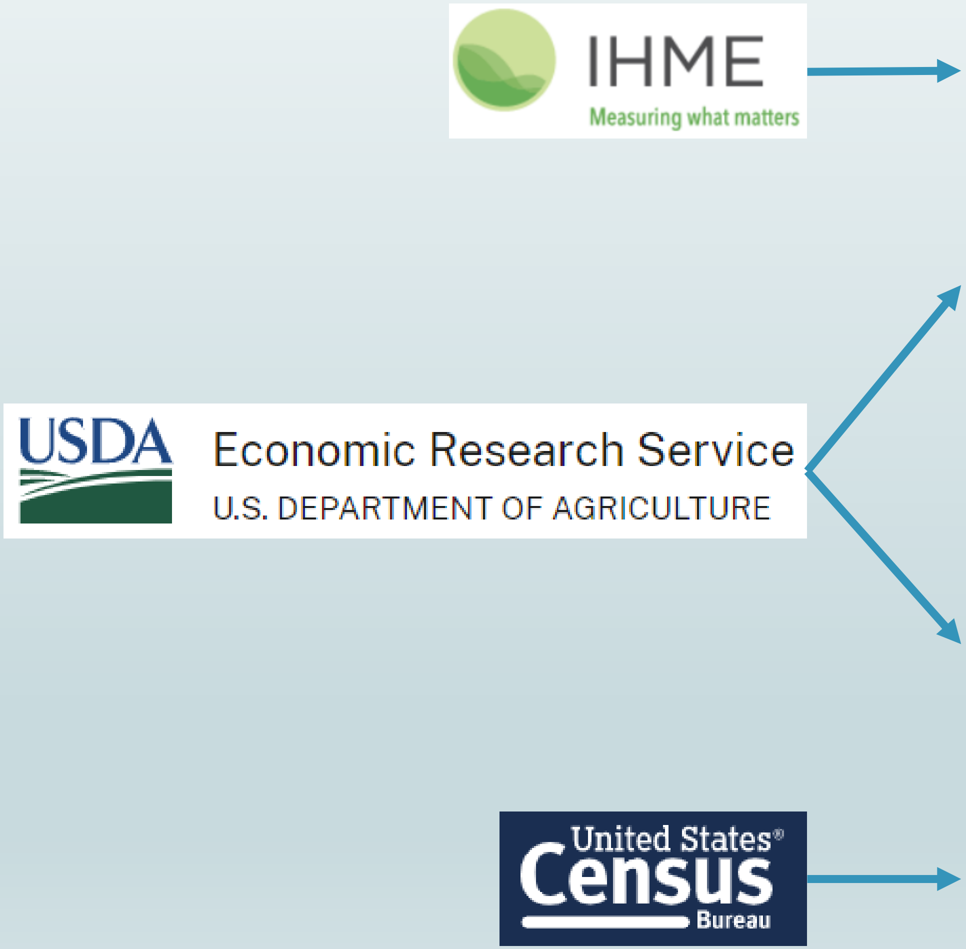
February 14, 2022

Overview and Problem Statement

- America has an obesity epidemic.
- Obesity prevalence has been linked to socioeconomic factors as well as food insecurity, e.g. access to affordable healthy foods.
- What are the most important socioeconomic or food insecurity factors in predicting the prevalence of obesity in U.S. counties?
- And how reliable are these factors in accurately predicting obesity prevalence?

Data

United States county-level data was collected from the following:



1. **Obesity Prevalence**

2. **Food Environment Atlas**: Contains data on factors such as store/restaurant proximity, food prices, food and nutrition assistance programs, and community characteristics, which interact to influence food choices and diet quality.

3. **Educational Attainment**

4. **Poverty and Median Income Estimates**

Exploratory Data Analysis Iterations

Data Cleaning

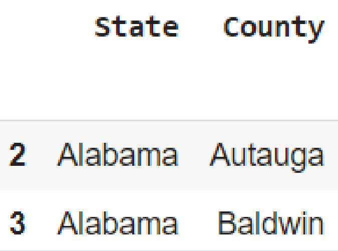
- Check/remove nulls.
- Remove duplicates.
- Remove commas.
- Spaces to underscores.
- Objects to numeric.
- Make State/County/FIPS fields consistent for merging datasets.

Data Exploration

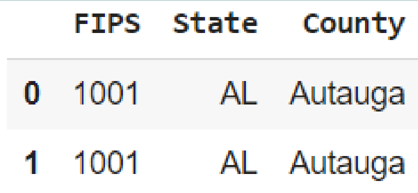
- Correlations
- Distributions
- Descriptive statistics

Feature Engineering

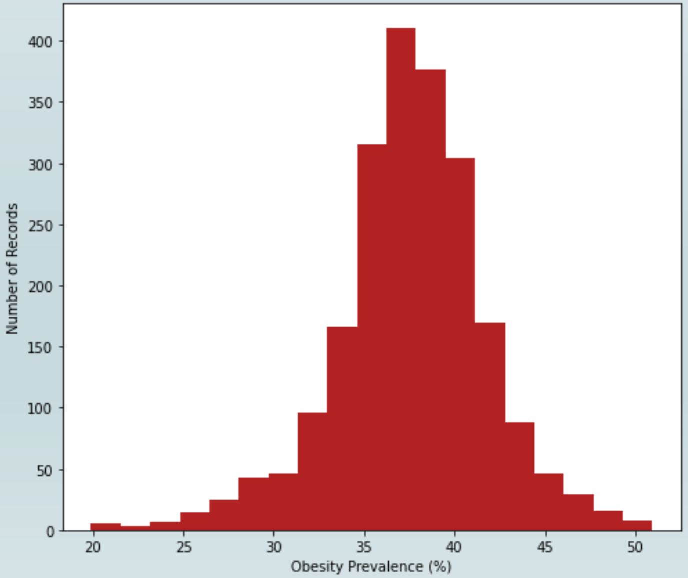
- Find/remove one half of highly correlating, redundant pairs, mostly on Food Environment Atlas.
- Remove years not of interest.
- Combine fields, e.g. create “overall” field from Male/Female data.



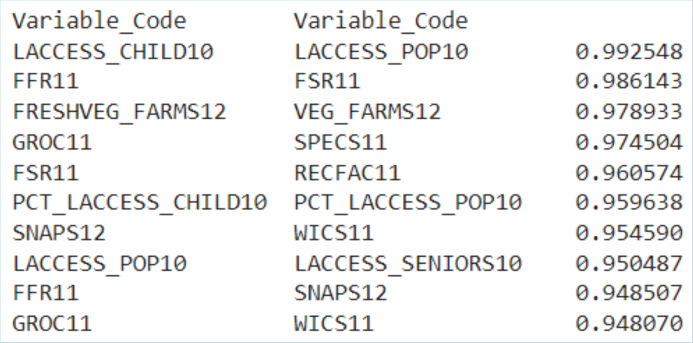
	State	County
2	Alabama	Autauga
3	Alabama	Baldwin



	FIPS	State	County
0	1001	AL	Autauga
1	1001	AL	Autauga



High Correlations

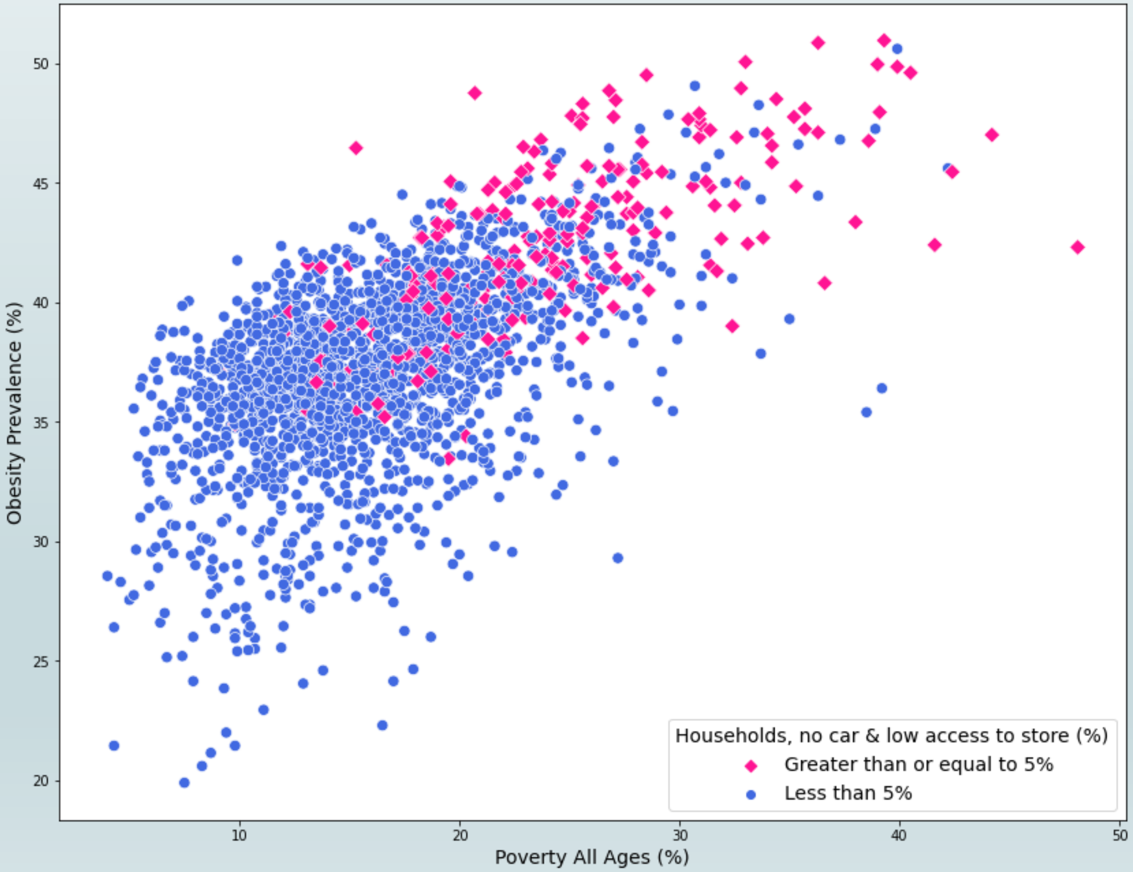


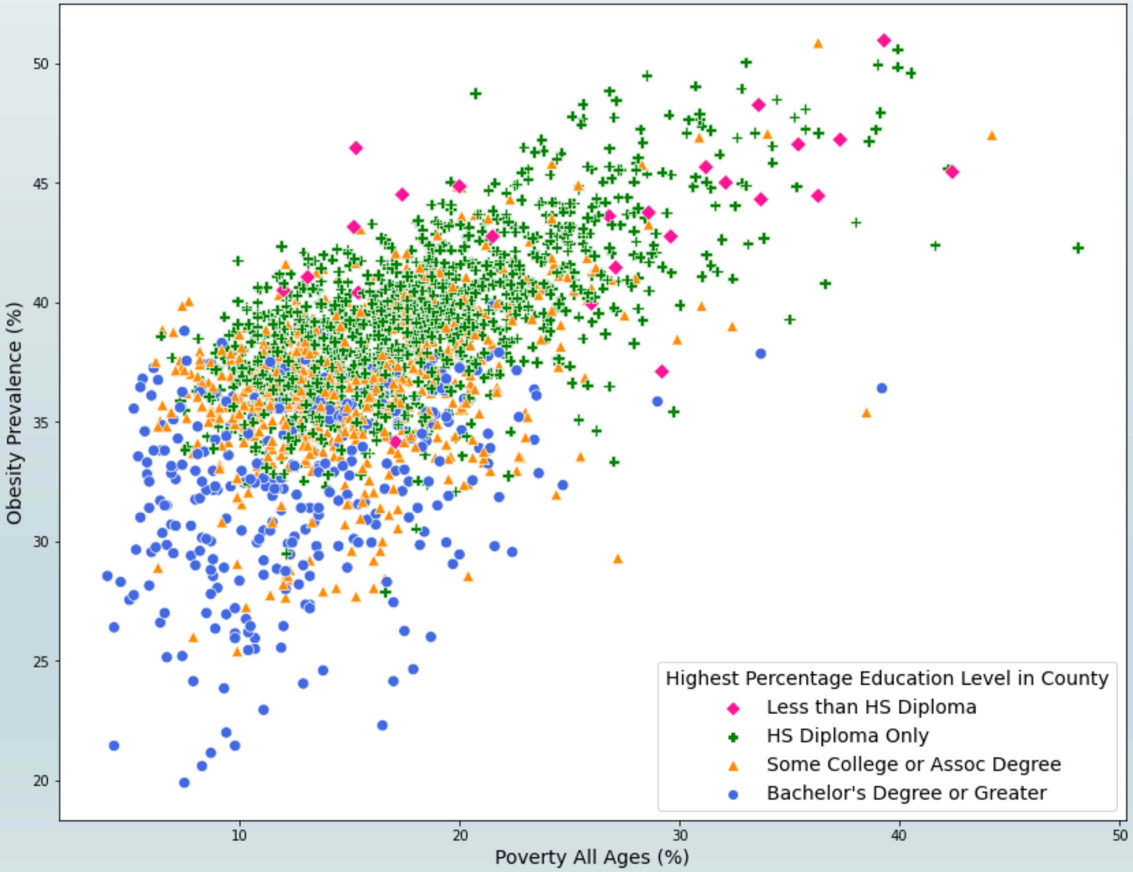
```
Variable_Code         Variable_Code
LACCESS_CHILD10       LACCESS_POP10         0.992548
FFR11                 FSR11                 0.986143
FRESHVEG_FARMS12      VEG_FARMS12           0.978933
GROC11                SPECS11               0.974504
FSR11                 RECFAC11              0.960574
PCT_LACCESS_CHILD10   PCT_LACCESS_POP10     0.959638
SNAPS12               WICS11                0.954590
LACCESS_POP10         LACCESS_SENIORS10     0.950487
FFR11                 SNAPS12               0.948507
GROC11                WICS11                0.948070
```

Data Exploration

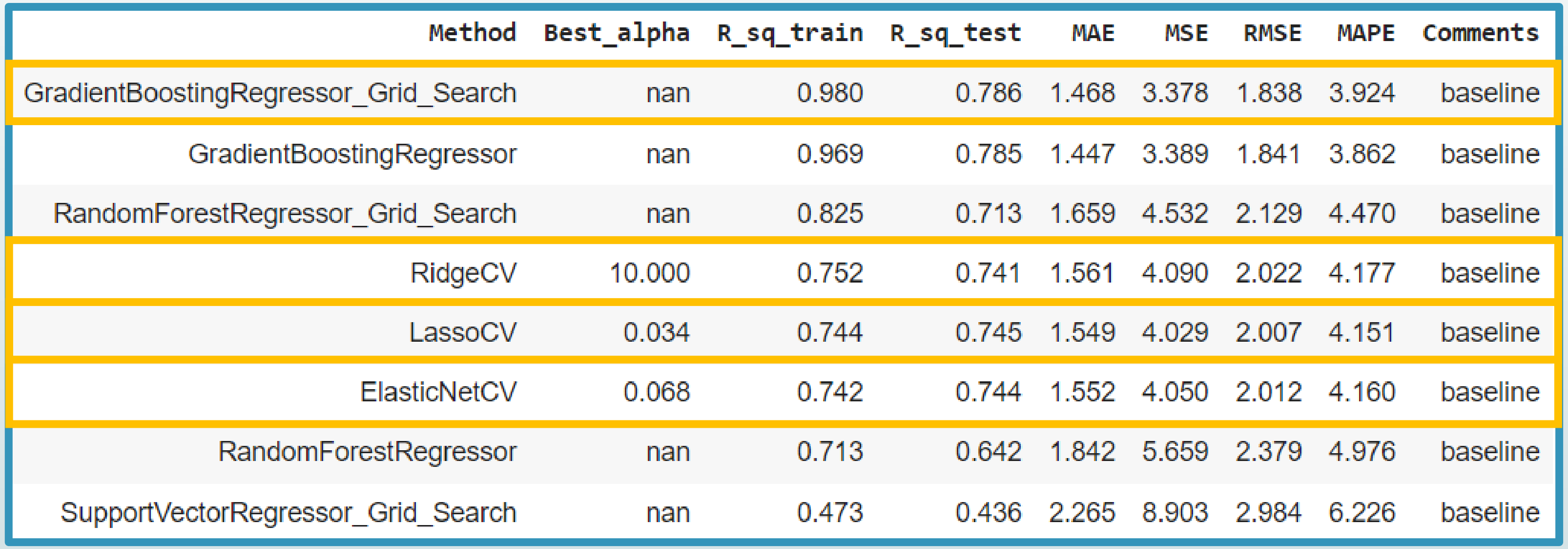
After iterations of data cleaning, exploration, and feature engineering:

- 48 variables
  - 1 Target: Obesity Prevalence by County from 2011
  - 47 Features
- 2166 counties





Modeling

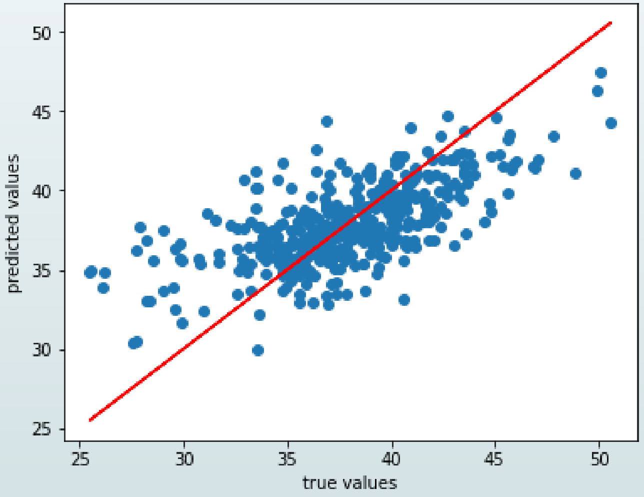


Method	Best_alpha	R_sq_train	R_sq_test	MAE	MSE	RMSE	MAPE	Comments
GradientBoostingRegressor_Grid_Search	nan	0.980	0.786	1.468	3.378	1.838	3.924	baseline
GradientBoostingRegressor	nan	0.969	0.785	1.447	3.389	1.841	3.862	baseline
RandomForestRegressor_Grid_Search	nan	0.825	0.713	1.659	4.532	2.129	4.470	baseline
RidgeCV	10.000	0.752	0.741	1.561	4.090	2.022	4.177	baseline
LassoCV	0.034	0.744	0.745	1.549	4.029	2.007	4.151	baseline
ElasticNetCV	0.068	0.742	0.744	1.552	4.050	2.012	4.160	baseline
RandomForestRegressor	nan	0.713	0.642	1.842	5.659	2.379	4.976	baseline
SupportVectorRegressor_Grid_Search	nan	0.473	0.436	2.265	8.903	2.984	6.226	baseline

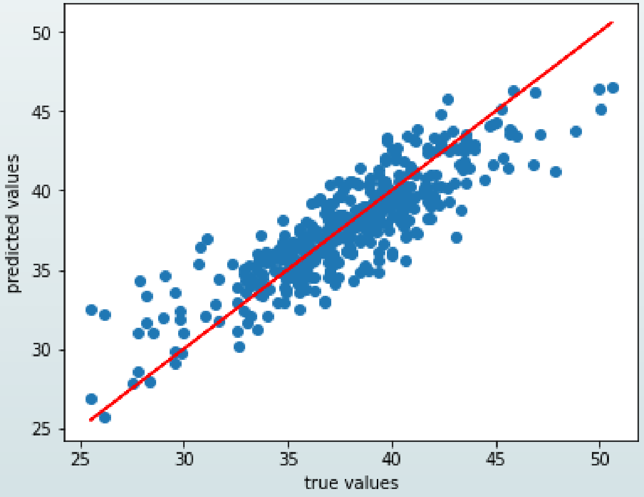
- The Gradient Boosting Regressor (GBR) clearly has the best explanatory power (R^2_train = 0.98)
- However, the GBR R^2_train - R^2_test difference indicates the model is not generalizable, i.e. the model is overfitting.
- Here, generalizability is more important than just a high explanatory power.
- Even though the Ridge, Lasso, and Elastic Net have lower R^2 values, the R^2_train - R^2_test indicate these models are very generalizable. Furthermore, 0.74 can still be considered a satisfactory explanatory and predictive power.
- Among the Ridge, Lasso, and Elastic Net, the Lasso is slightly more generalizable. Lasso is therefore chosen as the best model to move forward.

Predictive Power

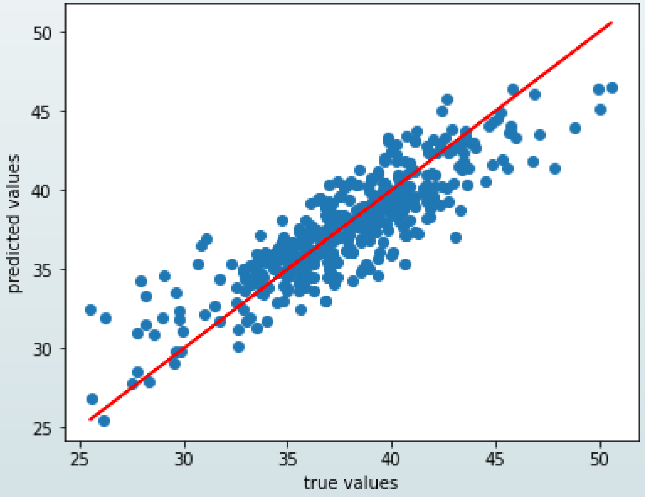
Obesity Prevalence
true and predicted values (SupportVectorRegressor_Grid_Search)



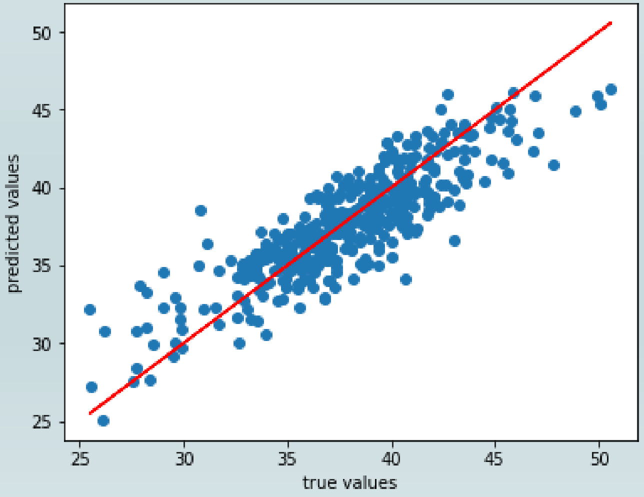
Obesity Prevalence
true and predicted values (ElasticNetCV)



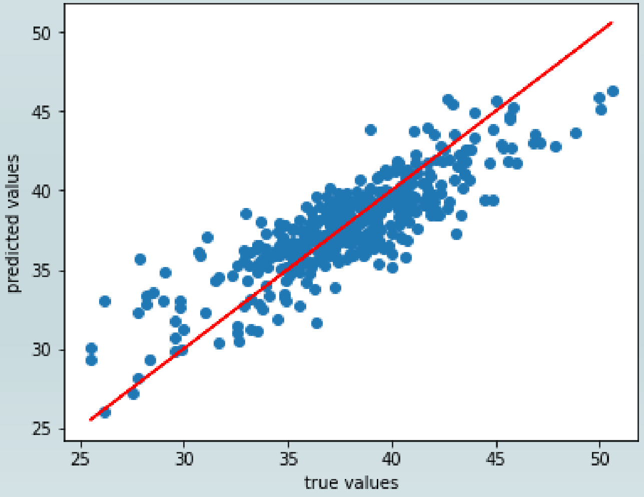
Obesity Prevalence
true and predicted values (LassoCV)



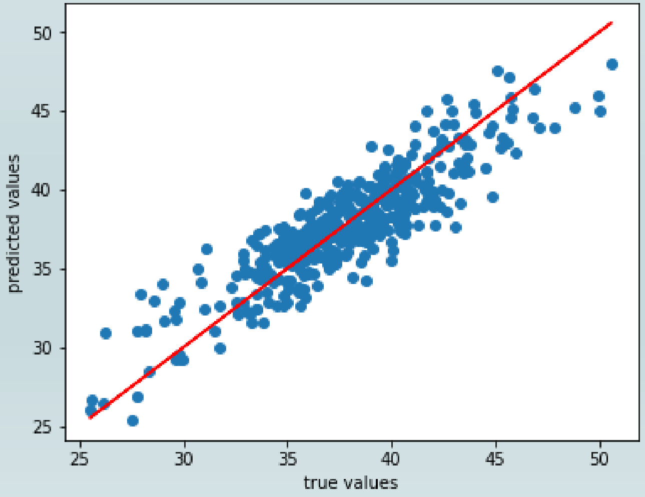
Obesity Prevalence
true and predicted values (RidgeCV)



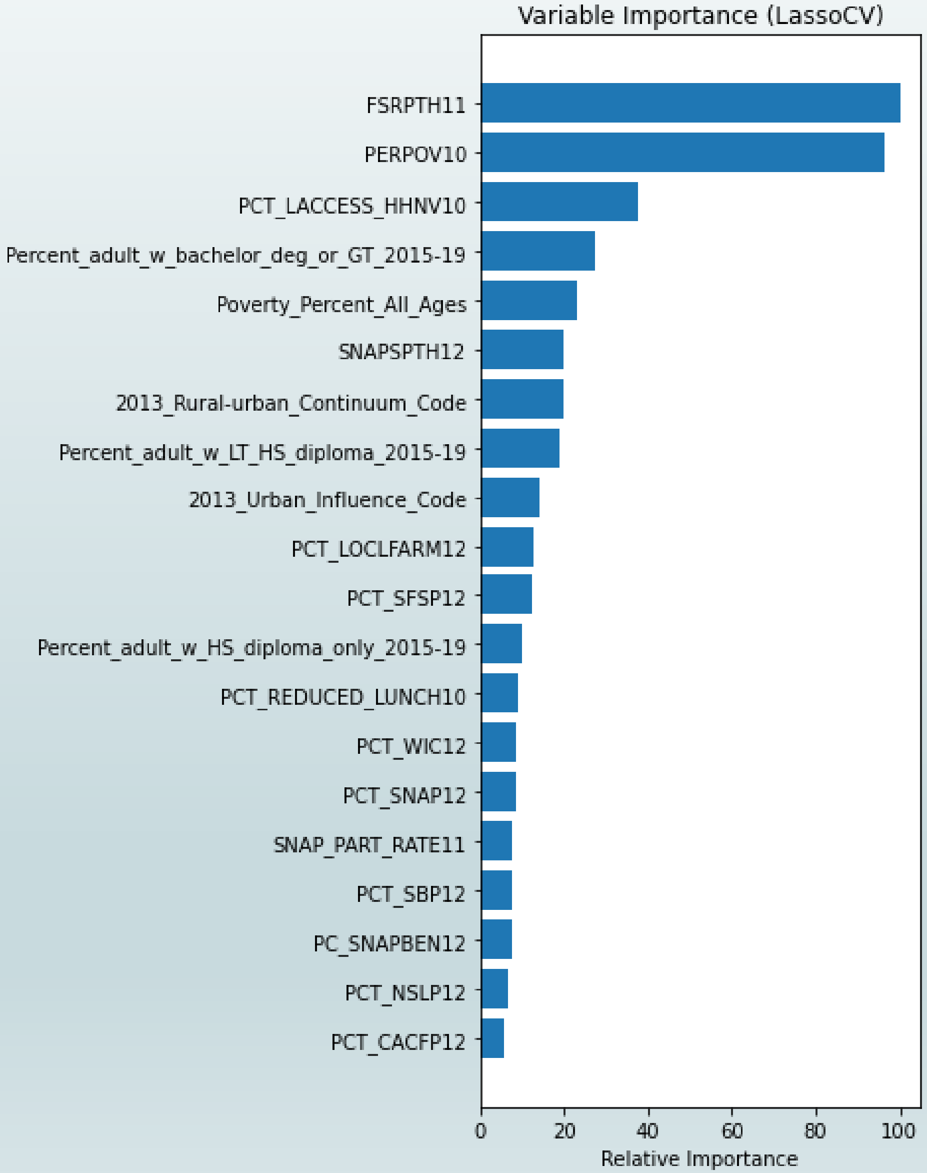
Obesity Prevalence
true and predicted values (RandomForestRegressor_Grid_Search)



Obesity Prevalence
true and predicted values (GradientBoostingRegressor_Grid_Search)



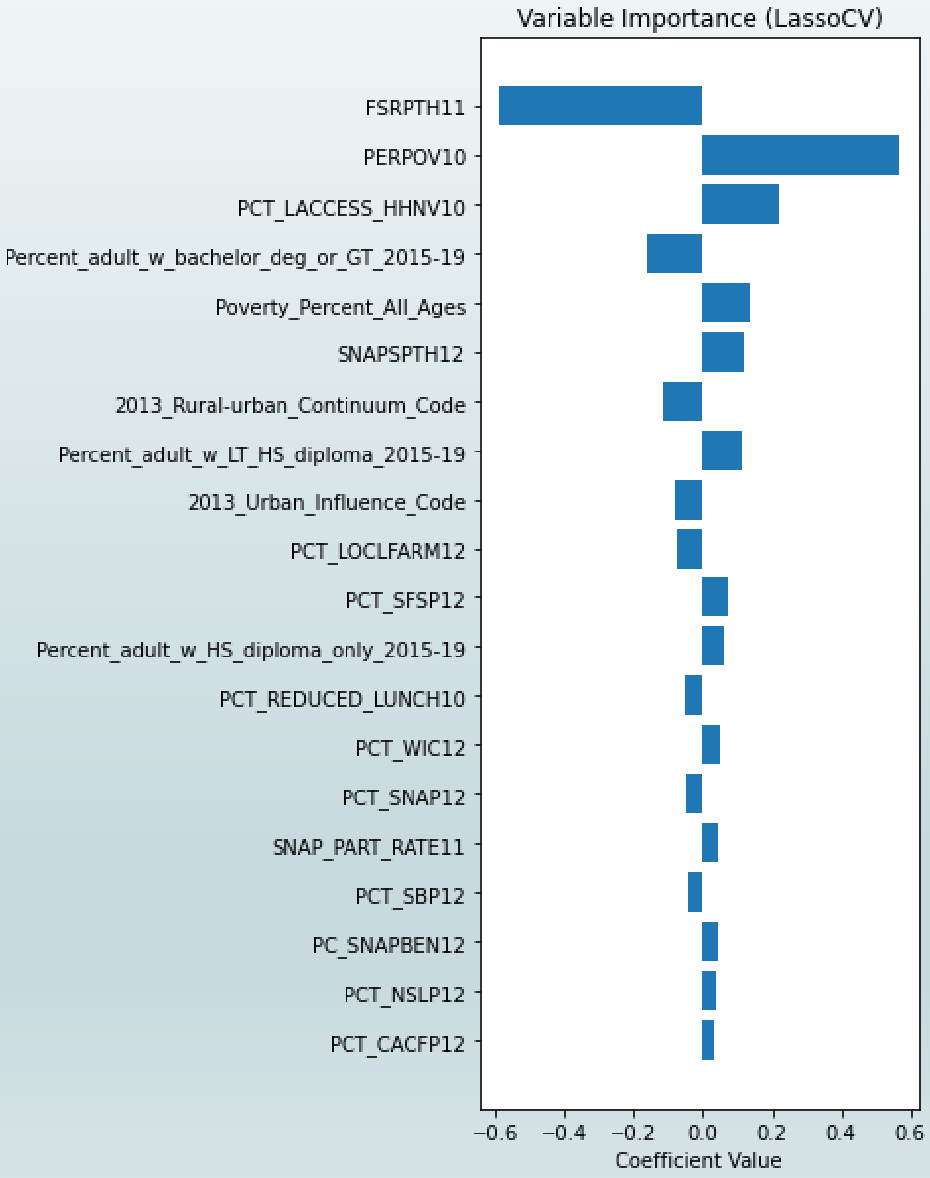
Feature Importance

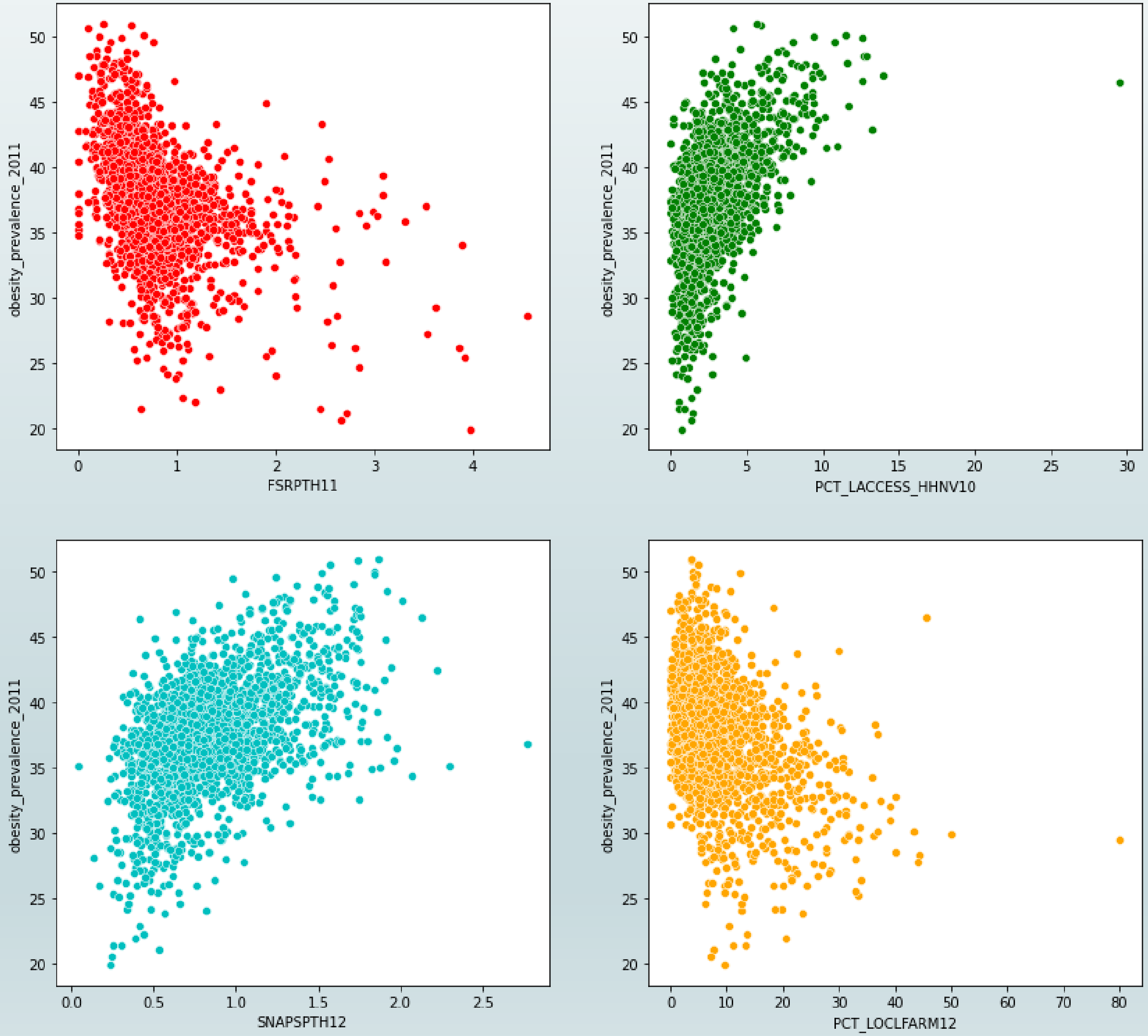


Aside from the fields on education and percentage in poverty, variables from the **Food Environment Atlas** with high feature importance for predicting obesity prevalence include:

- FSRPTH11: Full-service restaurants/1,000 pop, 2011
- PERPOV10: Persistent-poverty counties, 2010
- PCT_LACCESS_HHNV10: Households, no car & low access to store (%), 2010
- SNAPSPTH12: SNAP-authorized stores/1,000 pop, 2012
- PCT_LOCLFARM12: Farms with direct sales (%), 2012
- PCT_SFSP12: Summer Food Service Program participants (% children), 2012
- PCT_REDUCED_LUNCH10: Students eligible for reduced-price lunch (%), 2010

Feature Importance





Conclusions and Next Steps

- Reliable indicators of obesity prevalence in a U.S. county:
  - Socioeconomic factors such as poverty and education level
  - Food insecurity factors, such as low access to affordable healthy food

- Lasso regression is chosen as the best model.
  - Satisfactory explanatory power
  - Ability to generalize well with new data

- Recommended next steps for potential model improvement:
  - Use other subsets of features after further eliminating high correlations / redundancies.
  - Test methods of filling missing data to retain more counties, protecting against sampling bias.

- Use Case: Help local governments implement programs to target communities or groups that lack affordable healthy food options.

Questions?