Food Environment Atlas Predictive Modeling

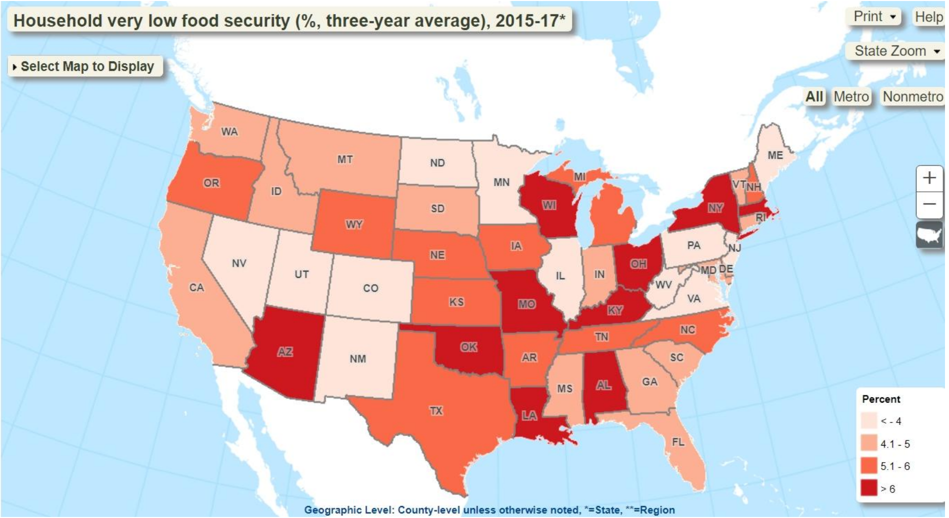
Craig Corsi, Omeiza Olumoye, Tatum Rask, Sayantan Sarkar

The Data

Food Environment Atlas [1]

Over 280 variables within 9 categories:

- Access and Proximity to Grocery Stores
- Store Availability
- Restaurant Availability and Expenditures
- Food Assistance
- State Food Insecurity
- Food Prices and Taxes
- Local Food
- Health and Physical Activity
- Socioeconomic Characteristics



[1] Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Environment Atlas. https://www.ers.usda.gov/data-products/food-environment-atlas/

Guiding Questions

- What factors are most closely connected to poverty, food insecurity, and nutrition-related illnesses?
- How can we understand the complexities surrounding community access to healthy food?
- What communities are in need of food assistance, and how can we implement healthy changes toward long-term improvement?

Data preprocessing and cleaning

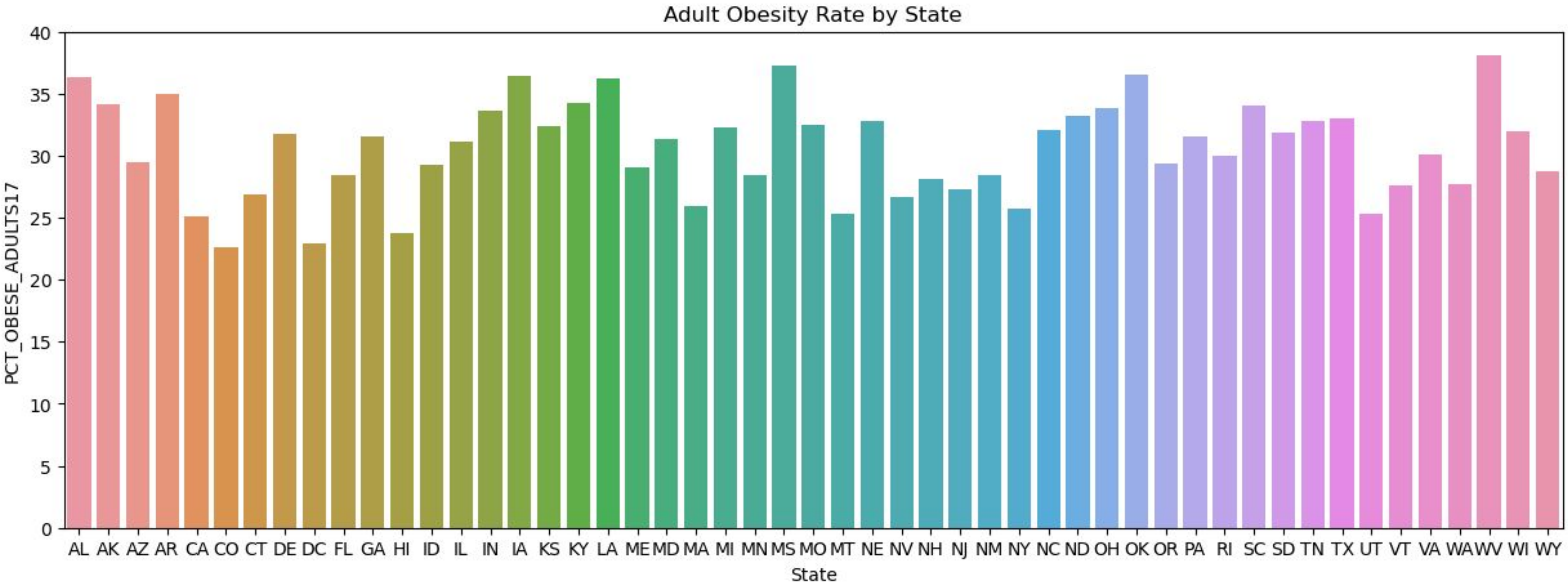
- Null values for county data imputed using national or state average values
- State data created from county data by taking weighted average over all counties
- Population data and latitude/longitude of county centroids included from U.S. Census Bureau estimates
- Updated data with county name changes
- Combined Bedford County, VA with the former independent city of Bedford, VA, and recalculated data accordingly

Stratifying county data geographically

- Used census geographic data to determine the 20 counties closest to each county
- Custom train-test split moves one county to the test set, then moves its 4 closest unsorted neighboring counties to the training set, whenever possible
- Split is also stratified by multiple categorical variables
- Split is reiterated to allow for k-fold cross-validation

Modeling Obesity Rates

Can we predict the adult obesity rate of a state given data about store availability and food assistance?

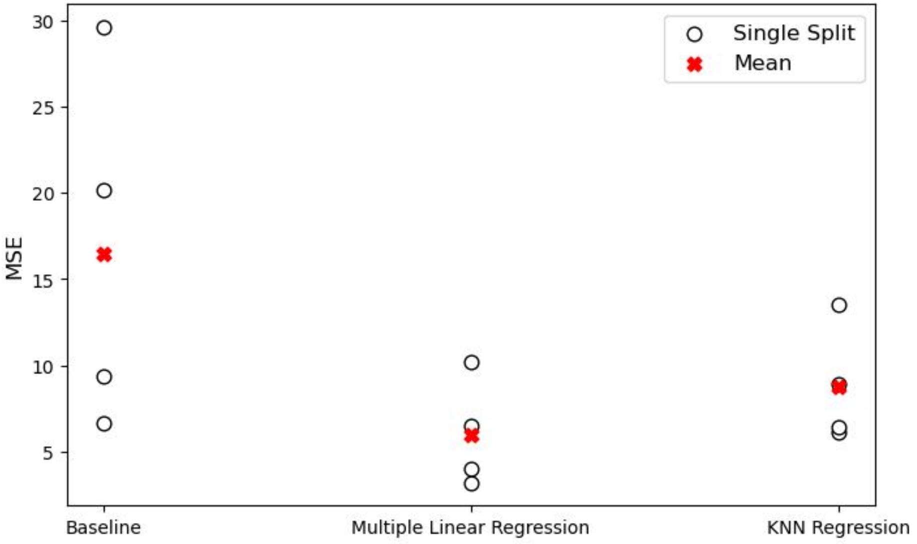


Modeling Obesity Rates - The Process

- Dependent Variable: Percent of Adults Obese in 2017
- 5 Independent Variables:
  - Chosen from data on stores access and food assistance programs
  - Chosen using lasso regression

Variable Code	Variable Name
SUPERCPTH16	Supercenters & club stores/1,000 pop, 2016
CONVSPTH16	Convenience stores/1,000 pop, 2016
SPECSPTH16	Specialized food stores/1,000 pop, 2016
WICSPTH16	WIC-authorized stores/1,000 pop, 2016
FSRPTH16	Full-service restaurants/1,000 pop, 2016

- Compared 3 different models:
  - Training sets: 30 states
  - Validation sets: 10 states



Modeling Obesity Rates - Results

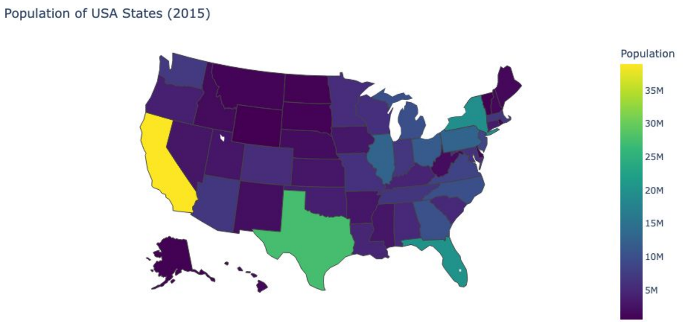
- Multiple Linear Regression Model

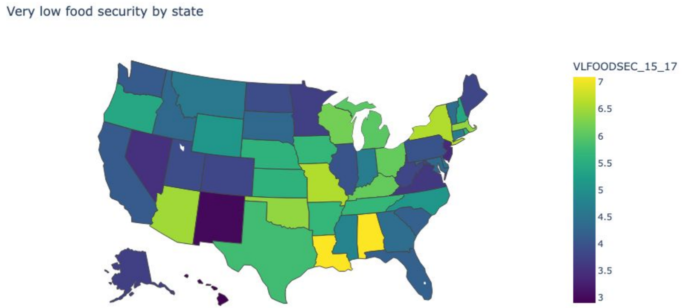
Variable Name	Coefficient
Supercenters & club stores/1,000 pop, 2016	140.023399
Convenience stores/1,000 pop, 2016	12.569778
Specialized food stores/1,000 pop, 2016	-16.744763
WIC-authorized stores/1,000 pop, 2016	16.214871
Full-service restaurants/1,000 pop, 2016	-8.981100

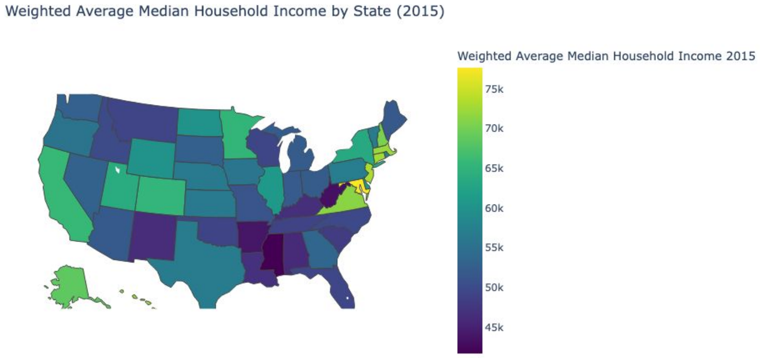
- MSE on training set: 4.611
- MSE on testing set: 6.011

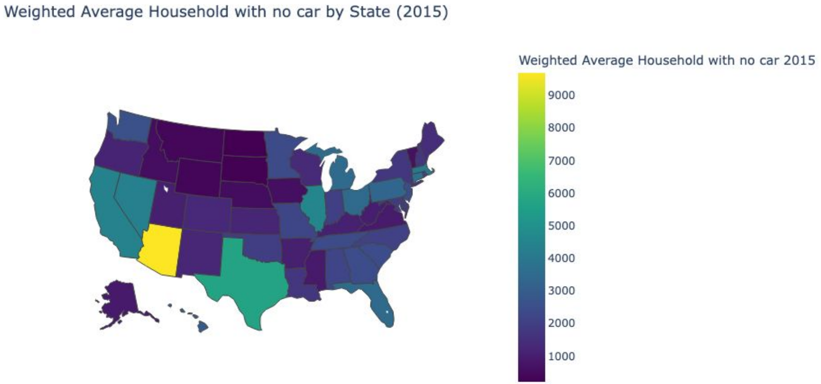
	State	Predicted Obesity Rate	True Obesity Rate
0	IA	33.218550	36.4
1	ME	30.231381	29.1
2	MS	37.216259	37.3
3	TX	31.464512	33.0
4	PA	28.724669	31.6
5	AR	33.925234	35.0
6	MO	31.013048	32.5
7	ID	30.665324	29.3
8	NH	29.257105	28.1
9	SD	36.071725	31.9
10	RI	25.514364	30.0

Very low food security: initial look at the data of interest









Very low food security: actual vs predicted

VLFOODSEC_15_17: Very low food security average between 2015 and 2017.

LR: Linear regression

DT: Decision tree

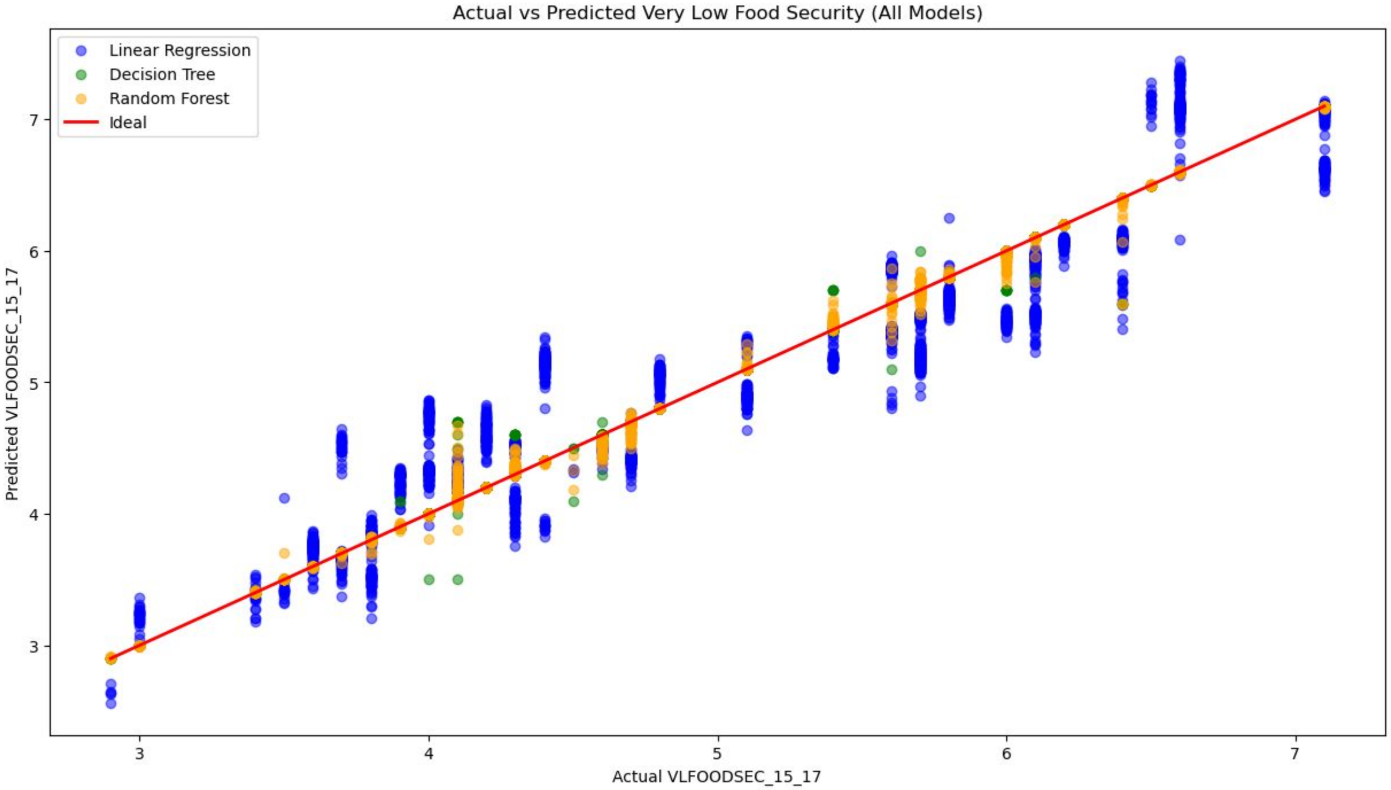
RF: Random forest

State	County	VLFOODSEC_15_17	Predicted_VLFOODSEC_15_17_LR	Predicted_VLFOODSEC_15_17_DT	Predicted_VLFOODSEC_15_17_RF
AK	Anchorage	3.7	4.512340438062400	3.7000000000000000	3.699999999999990
AL	Chilton	7.1	6.60382613823215	7.100000000000000	7.100000000000010
AR	Polk	5.7	5.219999457719100	5.700000000000000	5.7000000000000000
AZ	Navajo	6.5	7.017805703186520	6.5	6.5
CA	San Diego	4.1	4.167243712514540	4.1000000000000000	4.122000000000010
CO	Adams	3.8	3.5171225999642800	3.8000000000000000	3.7830000000000100
CT	Tolland	4.7	4.715581541512040	4.7	4.771999999999990
DC	District of Columbia	3.5	4.126193689764500	3.5	3.704000000000000
DE	New Castle	4.5	4.343309675447490	4.5	4.336000000000000

Features used: Poverty rate, low access to SNAP stores, households without cars and low access, median household income and food insecurity average between 2015 and 2017, number of grocery stores, superstores, convenience stores, fast food store and SNAP stores in 2015.

Basis of selection of a feature: Correlation score with the target variable.

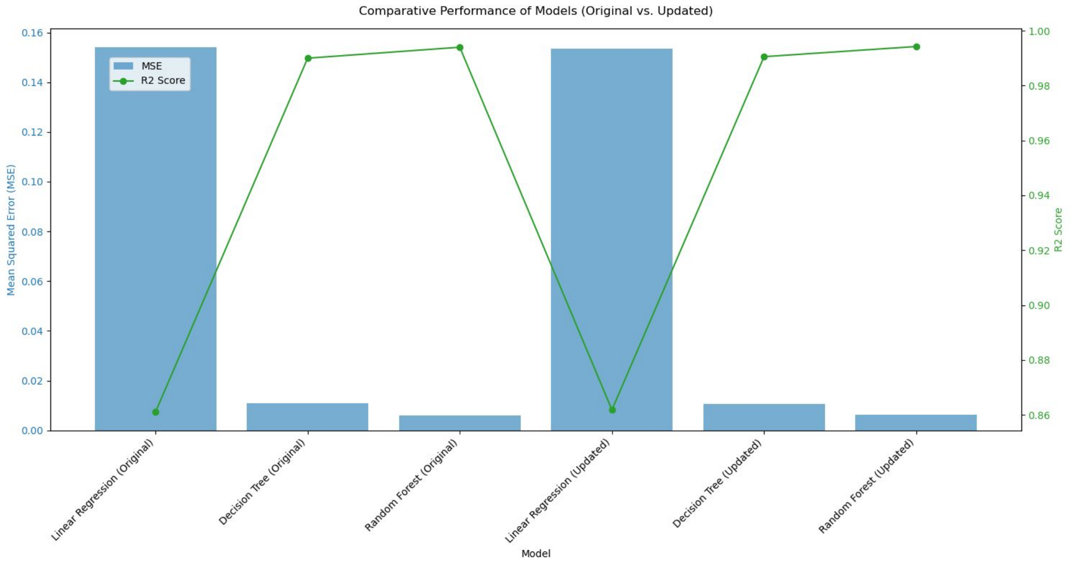
Very low food security: several predictive models



Very low food security: effect of several features

{ML Model} (Original): Features used were- Poverty rate, Low access to SNAP stores, Households without cars and low access, median household income and food insecurity average between 2015 and 2017.

{ML Model} (Updated): Along with the original ones the new features used- number of grocery stores, superstores, convenience stores, fast food store and SNAP stores in 2015.



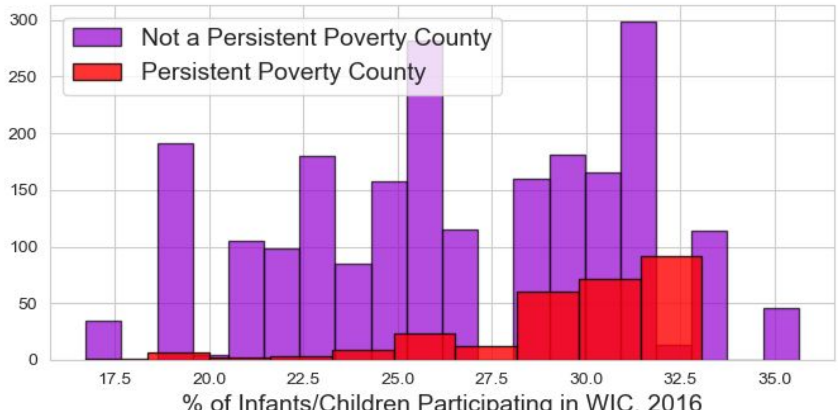
Observation: 1. Random forest model gives the best result.

2. The additional features here do not make much change in the R^2 score.

3. Poverty rate, low access to SNAP stores, median household income and food insecurity average and household without cars are the main features to determine the very low food security.

Classifying Persistent-Poverty Counties

- Counties whose poverty rate exceeded 20% consistently in the past 30 years
- 11.1% of counties nationwide (compare to current poverty rate of 11.6%)
- EDA showed most promising indicators were from food assistance data, as well as convenience stores/full-service restaurants per capita



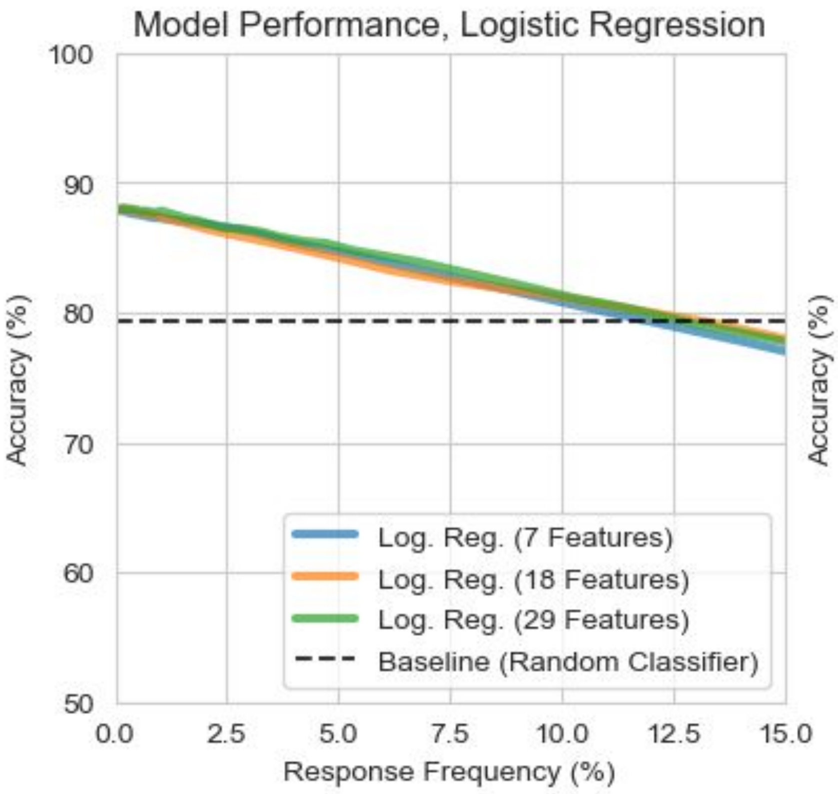
Classifying Persistent-Poverty Counties - Models

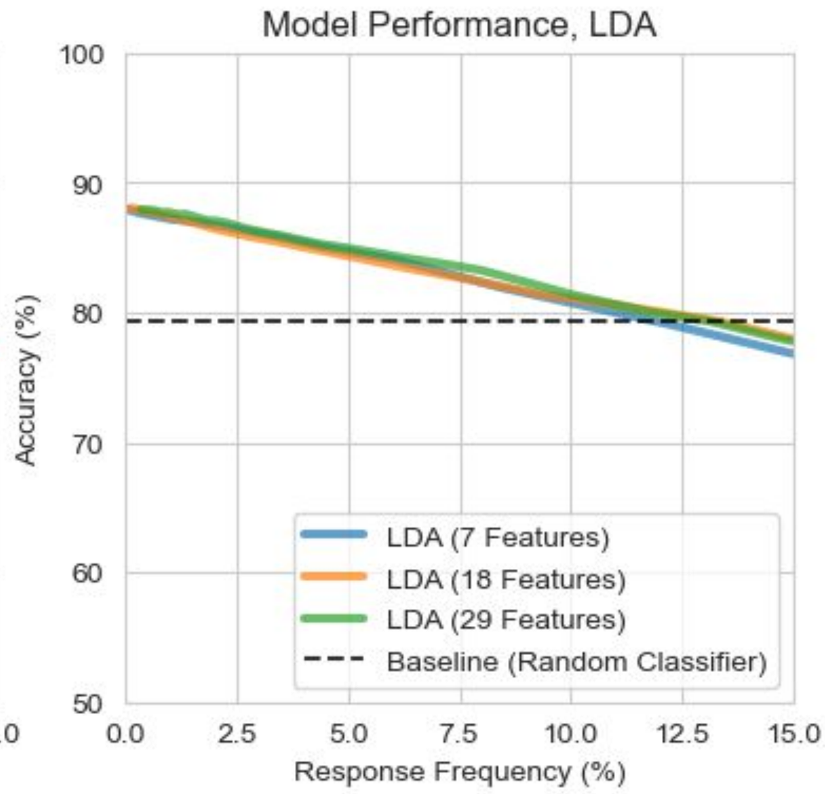
- Baseline model: Random classifier which labels 11.1% of the data as persistent-poverty counties
- Models: Logistic Regression, LDA, QDA, Random Forest
- Each model has 3 instances trained on 7, 18, or 29 features
- 5-fold cross-validation stratified by persistent poverty, metro/nonmetro, and geographic location

Performance metrics:

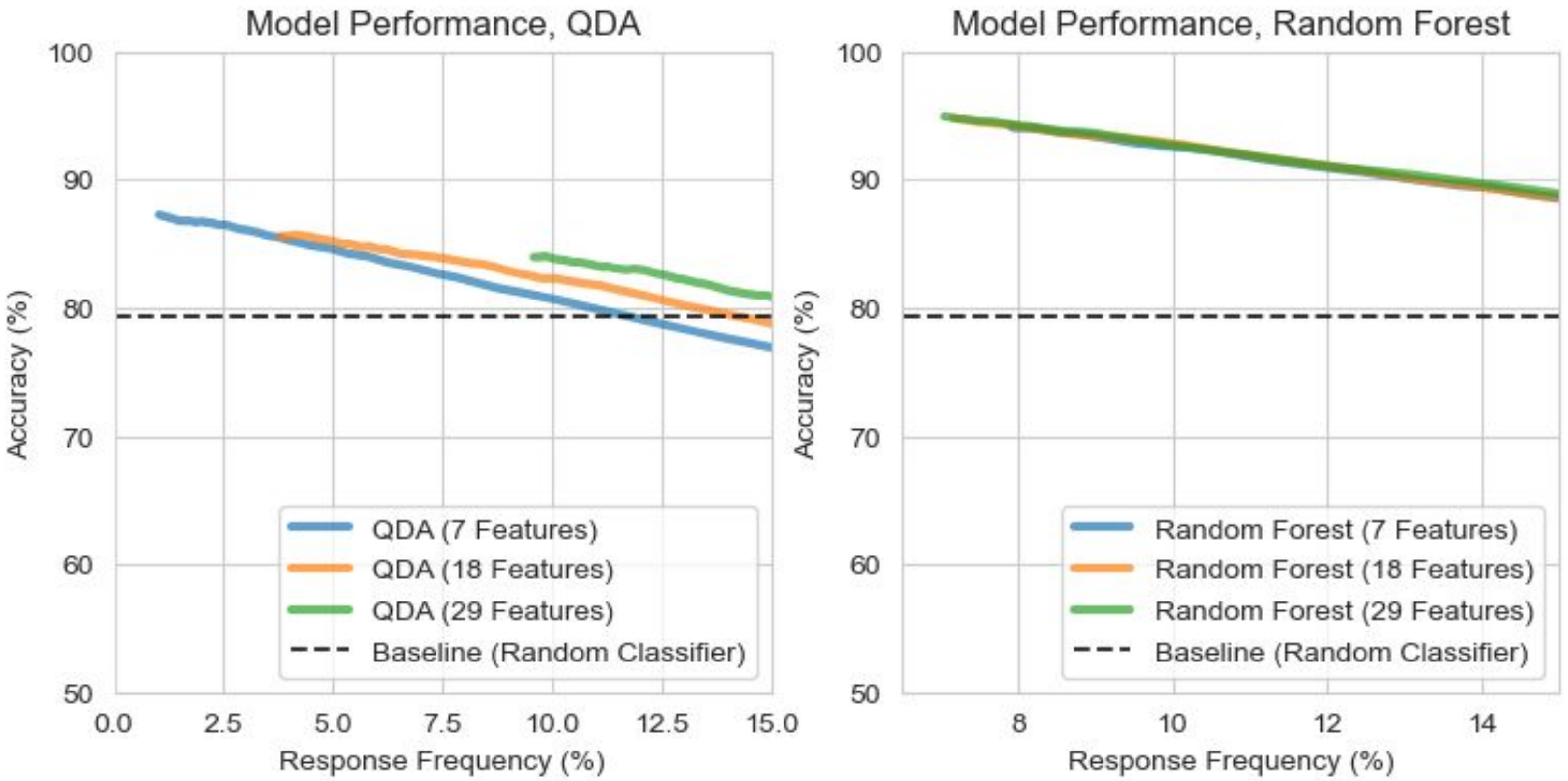
- Accuracy Score
- Frequency of counties predicted as persistent-poverty (we only considered models with prediction frequency above 11%)

Classifying Persistent-Poverty Counties - Performance





Classifying Persistent-Poverty Counties - Performance



Classifying Persistent-Poverty Counties - Results

Best model: Random Forest, 18 features

Mean accuracy on holdout sets: 91.68%
Prediction threshold: 0.27

Accuracy of final model on test set: 92.06%
Frequency of persistent-poverty predictions: 9.84%