Food Desert Prediction | Practicum I

Sam Blumer

July 1, 2019

Causes, Effects, and Predictions of Food Desert Occurrence Relative to Urbanization in the United States.

According to the American Nutrition Association (2011), “Food deserts are defined as parts of the country vapid of fresh fruit, vegetables, and other healthful whole foods, usually found in impoverished areas. This is largely due to a lack of grocery stores, farmers’ markets, and healthy food providers.” Over the past several decades, as urbanization has increased, food deserts in rural areas have also increased.

Part of the reason that rural areas risk becoming food deserts is because young families move away, and market pressures continue to squeeze small grocers and retailers. Food deserts are defined as counties in which all residents must drive more than 10 miles to the nearest supermarket chain or supercenter. The Great Plains are especially lacking in easy-access grocers. Research has found that, and we have confirmed in our study, that residents of food deserts tend to be older, poorer, and less educated. A big reason to be concerned about food deserts is that health can be compromised by lack of food access, as many residents do not consume adequate amounts of fresh fruits or vegetables, and they often lack adequate dairy and protein in their diet.

This has become a big problem because while food deserts are often short on whole food providers, especially fresh fruits and vegetables, instead, they are heavy on local quickie marts that provide a wealth of processed, sugar, and fat laden foods that are known contributors to our nation’s obesity epidemic.

With so many indicators available for analysis, we are going to attempt to use these indicators to predict counties in which a food desert may exist in the future.

We are going to be using a variety of packages that will allow us to process, explore, analyze, and ultimately offer tools for feature engineering and machine learning. Note that we will also be using a local Spark instance for processing and storing large amounts of our data.

For our local Spark instance, we are going to use Spark 2.1.0 and set-up particular configuration specifications based off the computer we are using for our analysis. This analysis had a basic configuration using 2 cores and 8GB of local RAM to store and process our data via our Spark cluster.

DATA SOURCES

We have multiple data sources that we are going to inspect and aggregate to make a more complete data set for our final evaluation. In total, we used 11 different data sources. Each data source provided some information related to the demographics, socioeconomic, geographic, cultural, and economic factors for each county in the United States. Included in the contents of this reports includes the data dictionary and raw data for each table.

In total, we ended up with 3,258 rows, one row representing one US county, and 63 variables. 62 of our variables are predictors, with the "Desert" variable being the response. A "1" in the Desert column represents a food desert, leaving the "0" to represent all non-food deserts. 3 of our columns are descriptive, and offer the FIPS, State, and County as unique identifiers. This leaves a total of 59 predictors.

DATA PRE-PROCESSING

All data was collected from a variety of sources found in the Resources document, and are stored in separate CSV files, each of which was loaded individually to the local Spark cluster. Therefore, we needed to merge all files together into a flat file, where each row represents one US county, and each column is a descriptor of that county.

Additionally, since we intend on running the data through machine learning algorithms, we need to specify and correct individual columns for processing. This includes updating data types to numeric and factor where appropriate.

Note that the complete files contain counties from Puerto Rico, which will not be used in our analysis. We quickly identified those and removed them from the analysis as data was not complete for all Puerto Rican counties.

All these processes are found in the raw code.

MISSING DATA

During the merge process, turning our individual CSVs into one flat file, we essentially used a left outer join on our first file and the food desert file. The latter only contains values for counties that included food deserts, therefore, any county from our flattened file that did not have a value in the food desert file will have an NA. Since we simply want to use this as a bivariate response column, we can change the NA values to 0s, then change the data type.

Finally, with the data cleaned and pre-processed, we can scan the table for any missing values. As Fig. 1 illustrates, < .1% of our data was missing. This ended up being just one row that was missing several data points, and for simplicity's sake, we removed the row from the data.

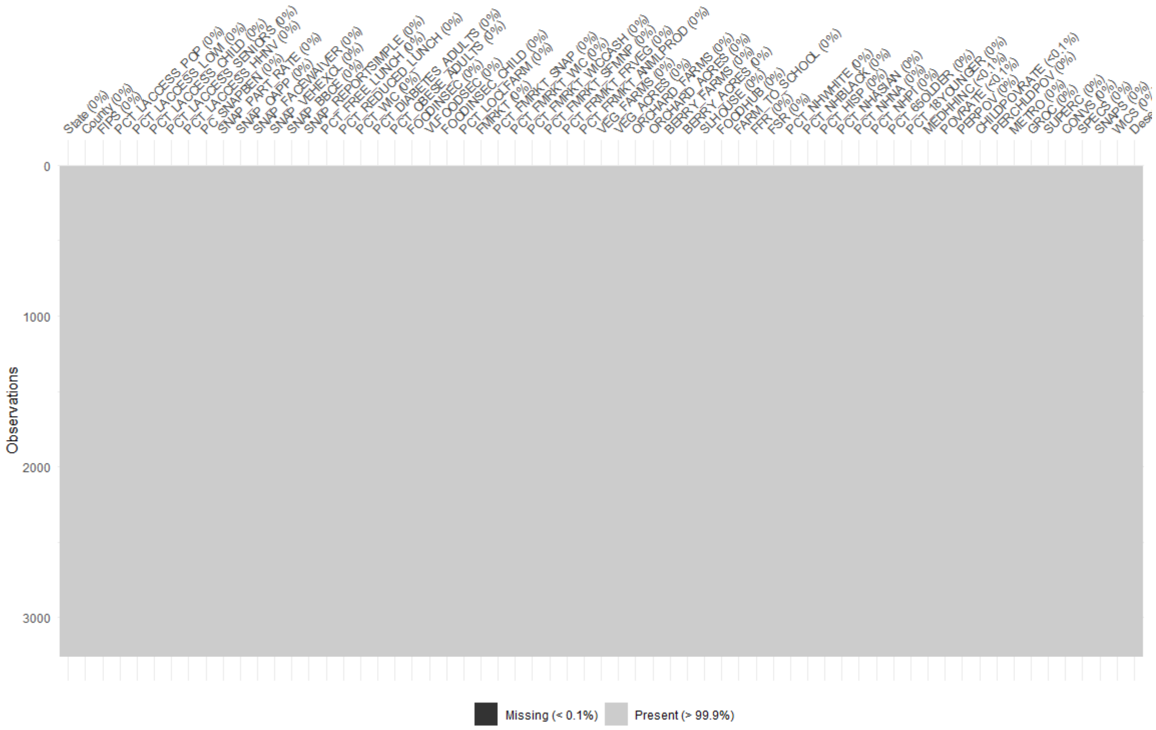


Figure 1

DATA EXPLORATION

With the data imported and cleaned, we can begin our data exploration process. Ultimately, we are looking at two different relationships. The first is identifying and exploring data points that are associated with food deserts as a whole. Second is exploring how the characteristics of rural food deserts differ from urban food deserts. Our data contains a large number of demographic data points relative to each county. These data points include income levels, poverty rates, educational attainment, ethnic makeup, industry employment and more. Since we expect to identify principal components in our next step, we are going to aggregate our demographic data points around our Rural and Urban counties and explore the differences between the two.

As noted earlier, there has been an extended period of time in which urbanization has increased. As we expect, this urbanization means that individuals are leaving rural areas for urban ones. To confirm this, we can plot the total populations of both urban and rural areas.

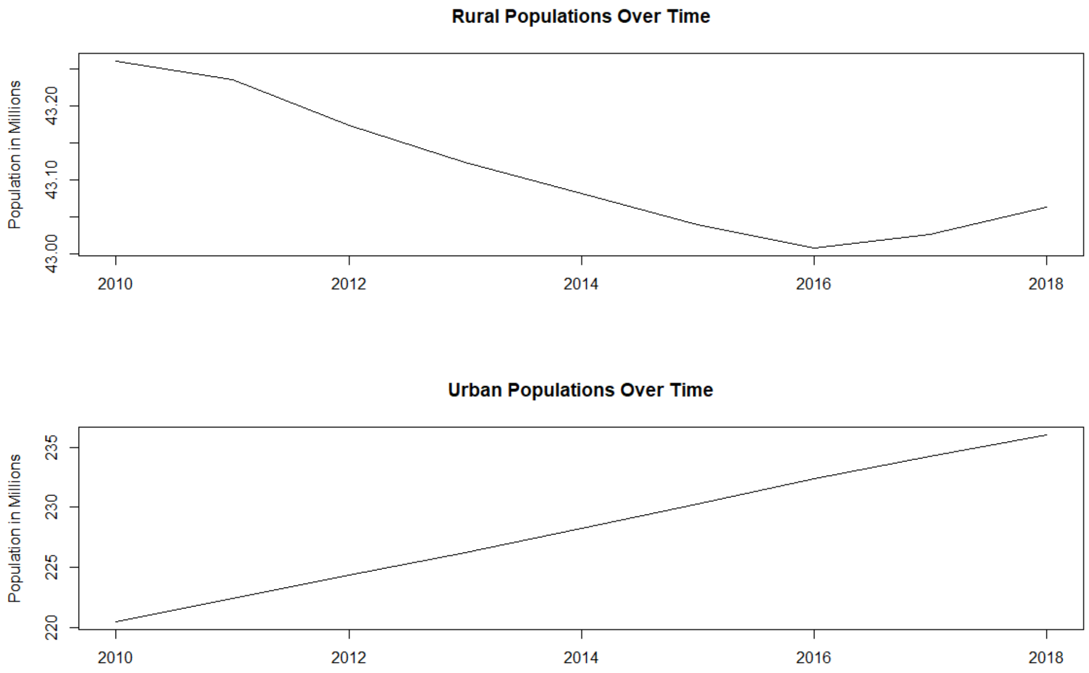
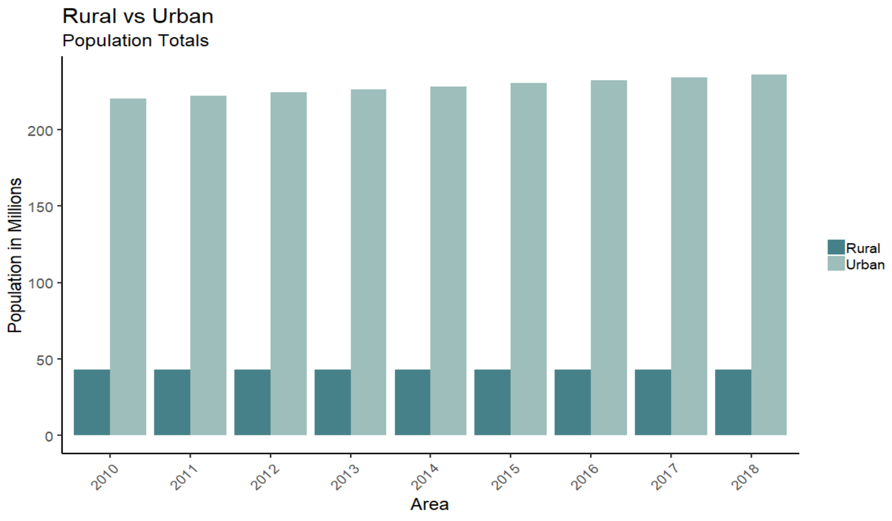


Figure 2

Figure 2 demonstrates that from 2010 – 2016 there is an inverse linear relationship between Rural Populations and Urban Populations, giving credit to the earlier claim that urban areas have grown while rural areas have shrunk. We can see this change year-by-year in Figure 3 below.



Source: USDA Economic Research Service

Figure 3

Because our focus is on urban and rural areas, we need to understand the differences between the two groups, including how populations and sparsity of residents may play a part in how food deserts form.

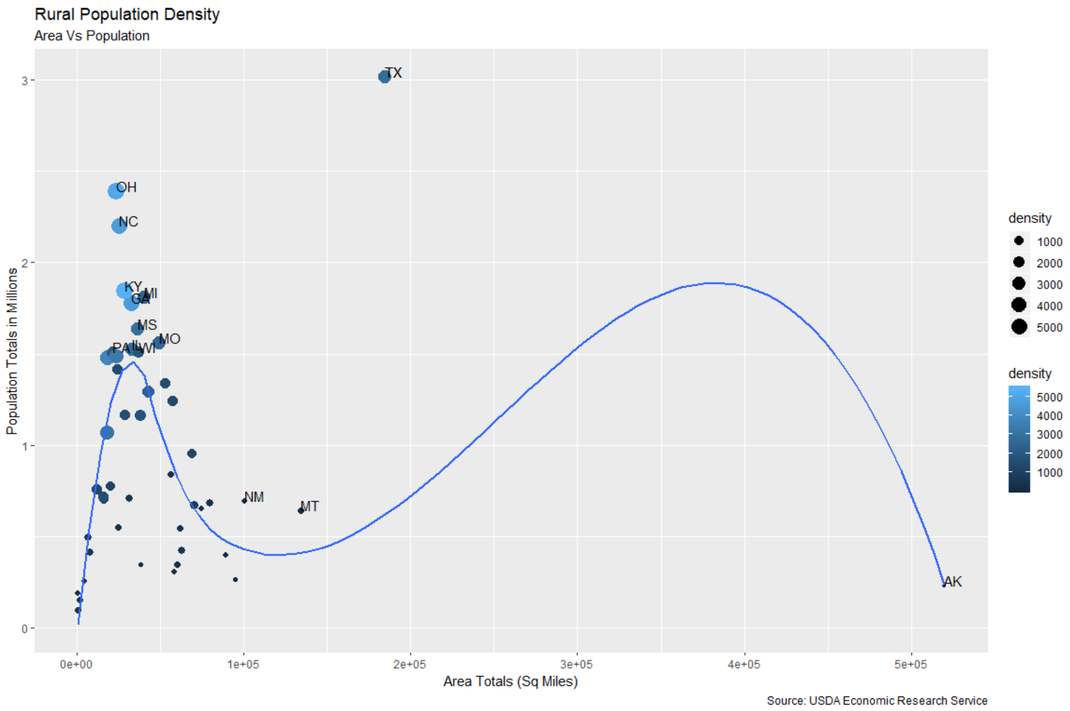


Figure 4

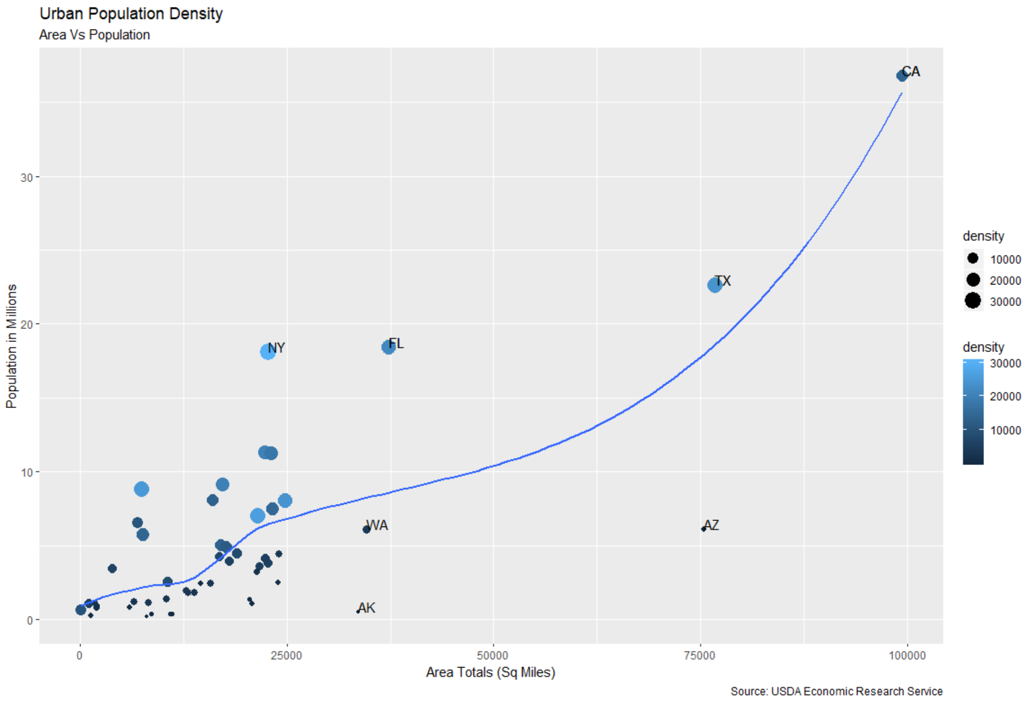
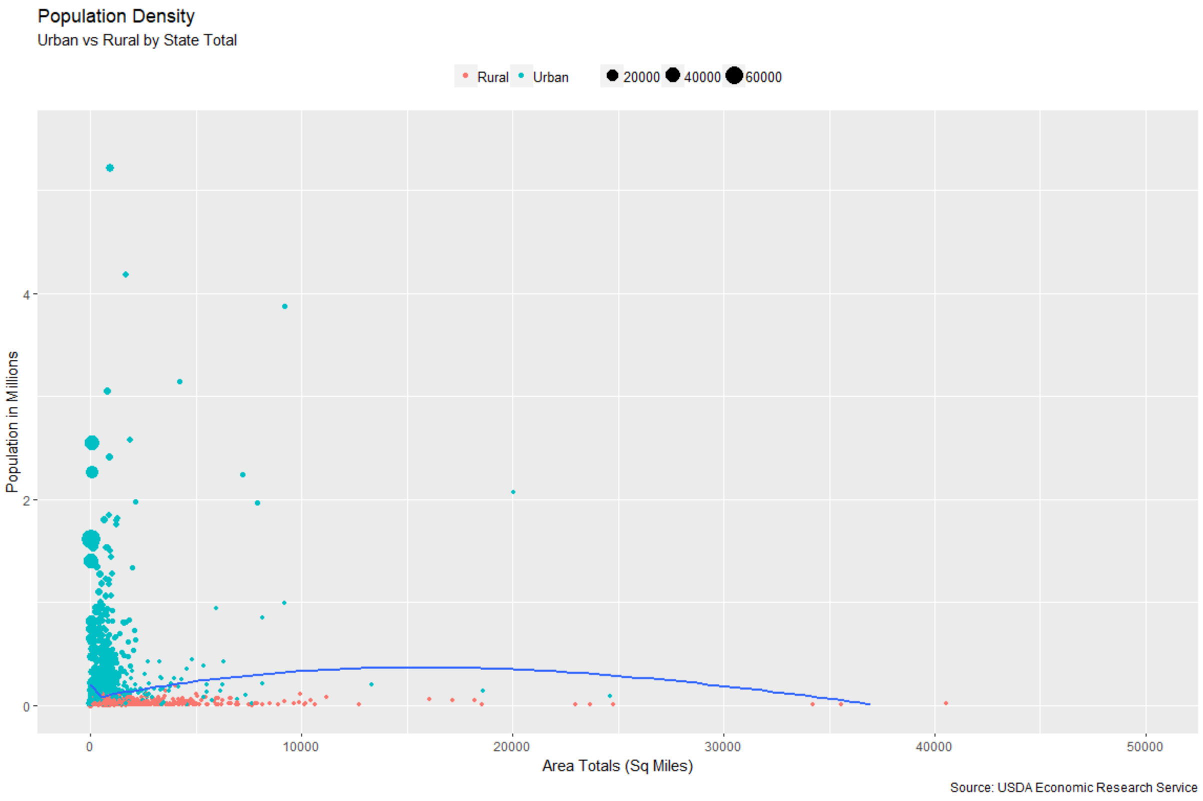


Figure 5

Figures 4 and 5 illustrate populations by land area for each state's rural and urban areas, respectively. From these visuals we can extrapolate that rural areas and populations are much more clustered than urban areas. This is opposition to urban areas in which there is a more linear relationship between population and area. For example, California has the largest amount of urban areas of any state, and that correlates with having the highest population of urban residents.

When comparing the urban and rural areas in terms of land mass, population, and density, we can see that urban areas in total are much smaller in overall landmass, while boasting a much higher total population. This again is in contrast to rural areas where populations are much smaller despite the fact that they, in total, have a greater amount of land. This gives credit to the idea that grocers are not investing in smaller communities simply due to the fact that there are not enough people to justify it, leading to an increase in rural food deserts.



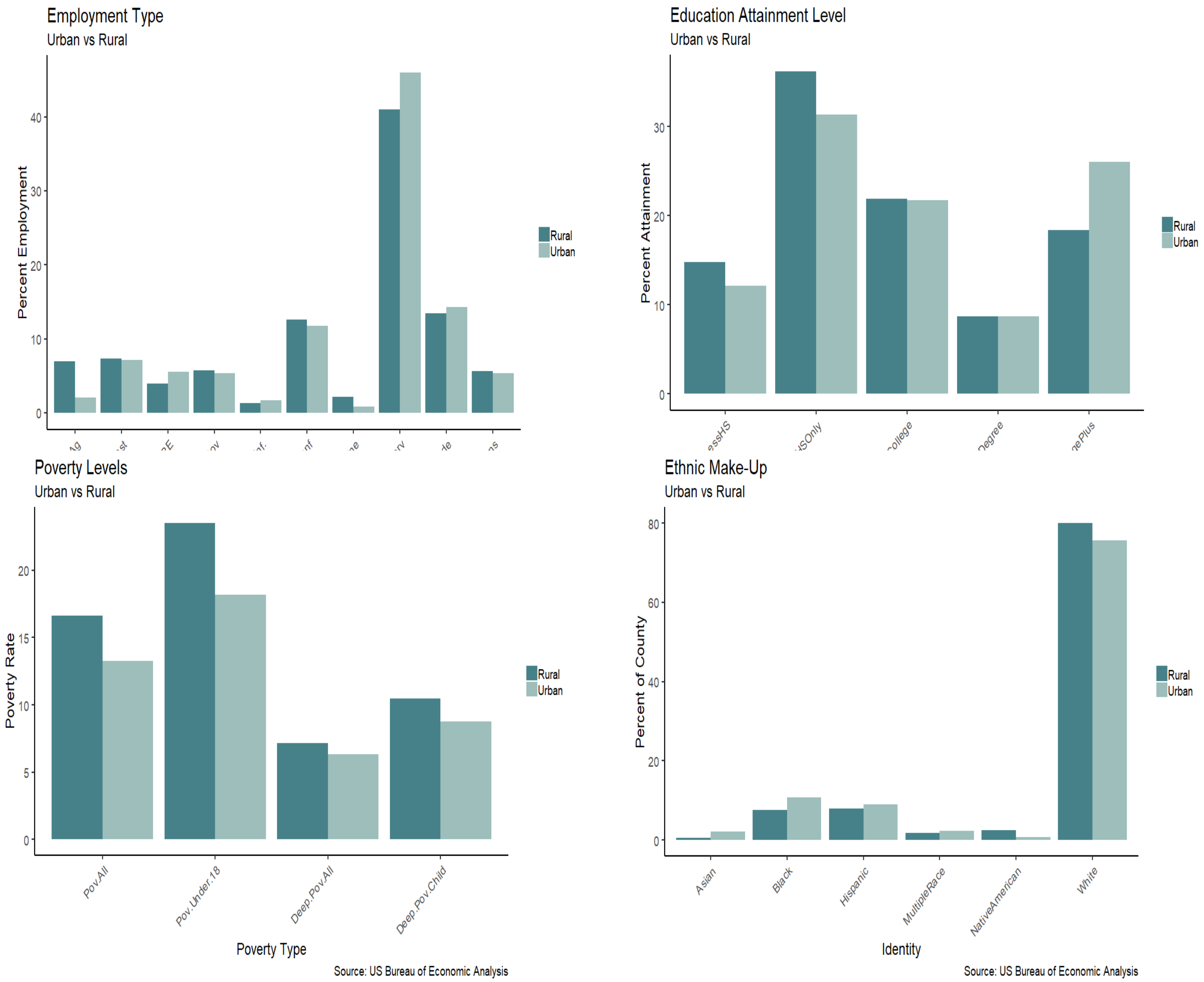
Population Density

Urban vs Rural by State Total

Source: USDA Economic Research Service

Figure 6

Of course, it is not simply population density differences that exist between urban and rural areas; therefore, we can explore our demographic characteristics of both areas and compare them.

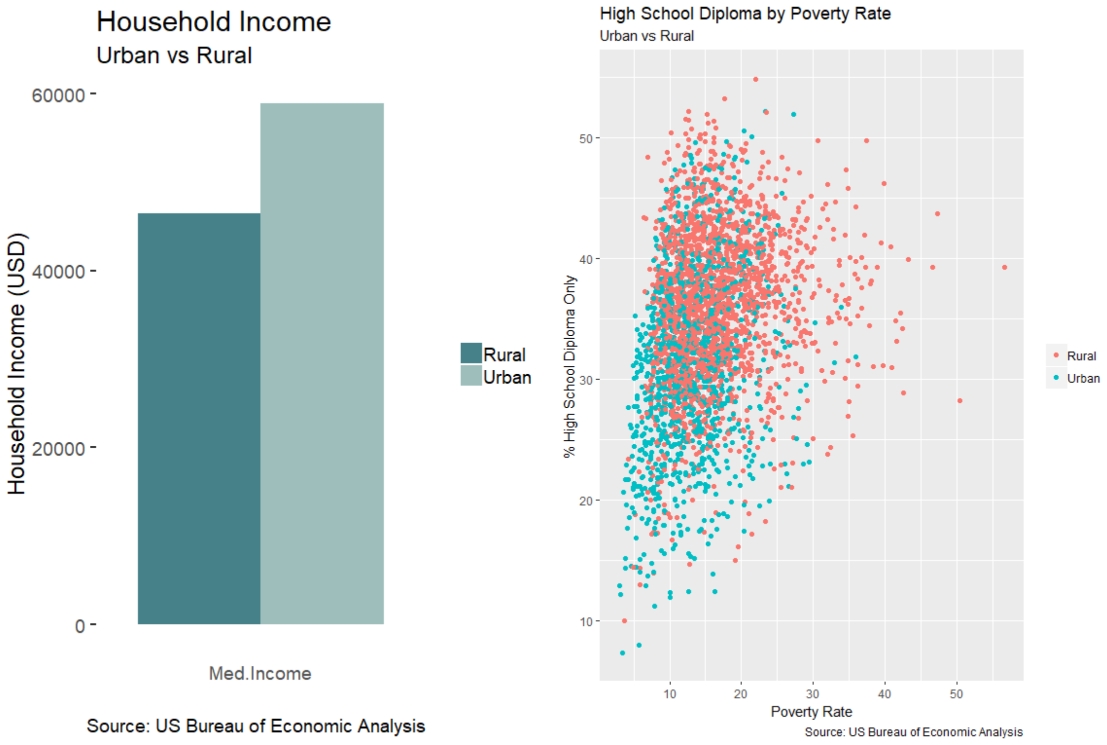


Source: US Bureau of Economic Analysis

Source: US Bureau of Economic Analysis

Figure 7

From the graphs above, we start to get a picture of what rural areas look like: They tend to be heavily employed in agriculture, construction, manufacturing, and transportation. The majority of residents have only a high school diploma with less than 20% of residents having completed at least a bachelors degree. Rural area residents are predominantly white and experience poverty at higher rates than urban residents across the board including child poverty and deep, persistent poverty. As we might expect, this lack of education and predominant employment industries lead to lower incomes overall. Which we can correlated to education levels overall in Figure 8.


Household Income
Urban vs Rural

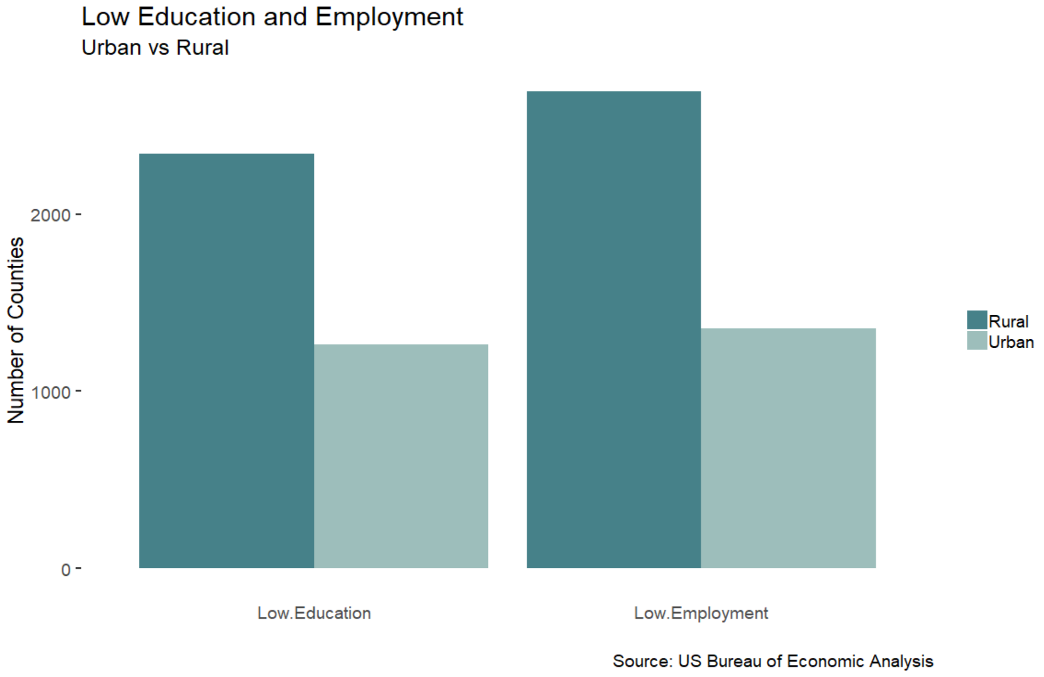
Source: US Bureau of Economic Analysis

High School Diploma by Poverty Rate
Urban vs Rural

Source: US Bureau of Economic Analysis

Figure 8

It makes sense, then that there are nearly double the number of rural counties that are considered to have low education and low employment levels when compared to urban counties.


Low Education and Employment
Urban vs Rural

Source: US Bureau of Economic Analysis

Figure 9

MODEL PREPARATION

Currently, we have 63 total columns in our dataset. Our final column (Desert) is the response, and the first 3 columns (FIPS, State, and County) are identification attributes. Therefore, we need to make sure that they are removed from our model. With such a large number of predictors, we can create a quick formula for our models, as opposed to typing each variable into the glm formula(*x1, x2, x3….xn*). Our first model will use all the predictors in order to set a baseline accuracy rate that we can try to build on with feature engineering.

Using all 59 remaining predictors, our formula is as follows:

```
Desert ~ PCT_LACCESS_POP + PCT_LACCESS_LOWI + PCT_LACCESS_CHILD +
    PCT_LACCESS_SENIORS + PCT_LACCESS_HHNV + PC_SNAPBEN + SNAP_PART_RATE +
    SNAP_OAPP + SNAP_FACEWAIVER + SNAP_VEHEXCL + SNAP_BBCE +
    SNAP_REPORTSIMPLE + PCT_FREE_LUNCH + PCT_REDUCED_LUNCH +
    PCT_WIC + PCT_DIABETES_ADULTS + PCT_OBESE_ADULTS + FOODINSEC +
    VLFOODSEC + FOODINSEC_CHILD + PCT_LOCLFARM + FMRKT + PCT_FMRKT_SNAP +
    PCT_FMRKT_WIC + PCT_FMRKT_WICCASH + PCT_FMRKT_SFMNP + PCT_FRMKT_FRVEG +
    PCT_FRMKT_ANMLPROD + VEG_FARMS + VEG_ACRES + ORCHARD_FARMS +
    ORCHARD_ACRES + BERRY_FARMS + BERRY_ACRES + SLHOUSE + FOODHUB +
    FARM_TO_SCHOOL + FFR + FSR + PCT_NHWHITE + PCT_NHBLACK +
    PCT_HISP + PCT_NHASIAN + PCT_NHNA + PCT_NHPI + PCT_65OLDER +
    PCT_18YOUNGER + MEDHHINC + POVRATE + PERPOV + CHILDPOVRATE +
    PERCHLDPOV + METRO + GROC + SUPERC + CONVS + SPECS + SNAPS +
    WICS
```

Our model includes a variety of factors of different scales. For example, we have percentages of total population, as well as sums of land area and population totals, which, if left untreated, would inaccurately skew the model results. Therefore, we need to scale these values. A simple step to do this includes scaling all numeric factors in one subset, and then rejoining them to the factor variables.

FEATURE ENGINEERING

RANDOM FOREST FEAUTRE EXPLORATION

As mentioned previously, the dataset, at this point, contains 59 predictors, and it is likely that some of these impact our response more than others, and that some may not impact the response at all. We can therefore use the Random Forest Variable Importance function to identify the variance of each variable on our model and adjust it accordingly.

The first test used for variable importance is the Random Forest Variable Importance (RFVI) algorithm, the results for which are seen in Figure 10.

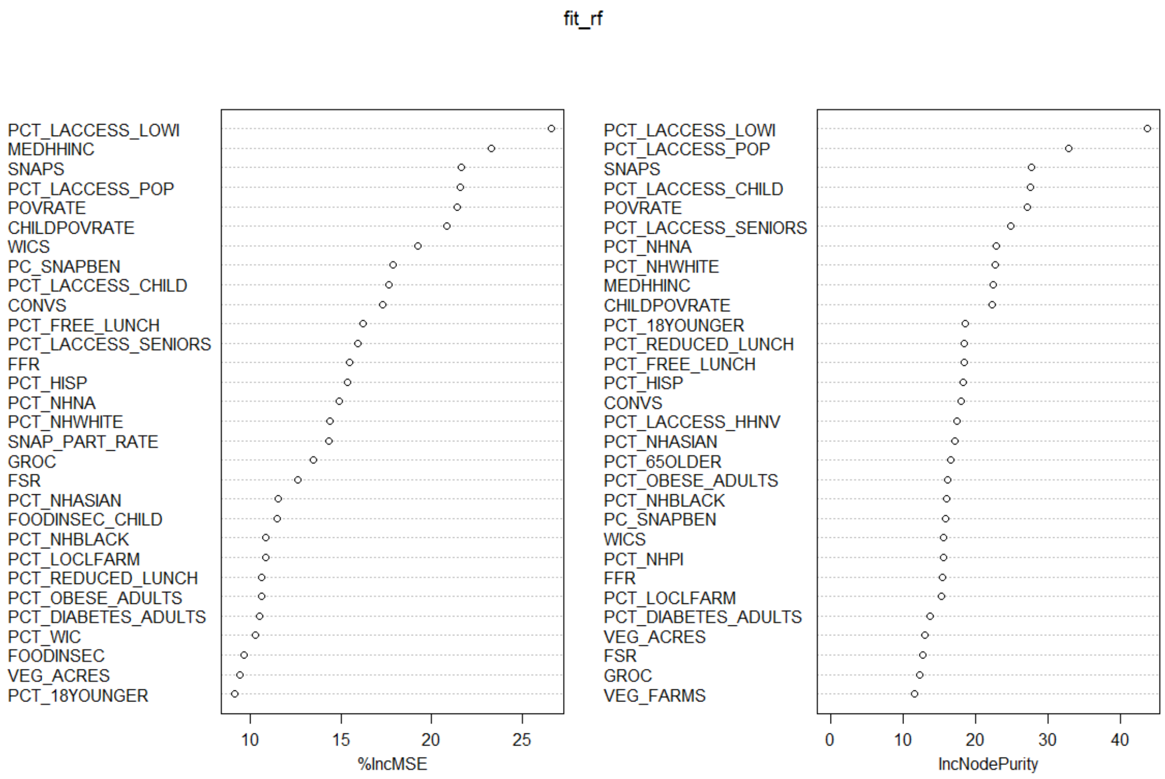


Figure 10

The RFVI plot sorts our variables in order of decreasing importance to the model based of the mean Gini score. This allows us to see that our top five most important variables in terms of predicting a food desert include the percent of residents in a county with low access to food and low income, the median household income, the number of SNAP authorized stores, the percent of the population that is not low income, but does have low access to food, and the overall poverty rate.

The RFVI model also uses a basic random forest to predict food deserts, the results of which will be used as a baseline for which we can compare other models.

```
## Call:
##  randomForest(formula = as.formula(fla), data = scaled_final, importance = 
TRUE, na.action = na.roughfix) 
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 19
## 
##           Mean of squared residuals: 0.1837564
##                     % Var explained: 21.61
```

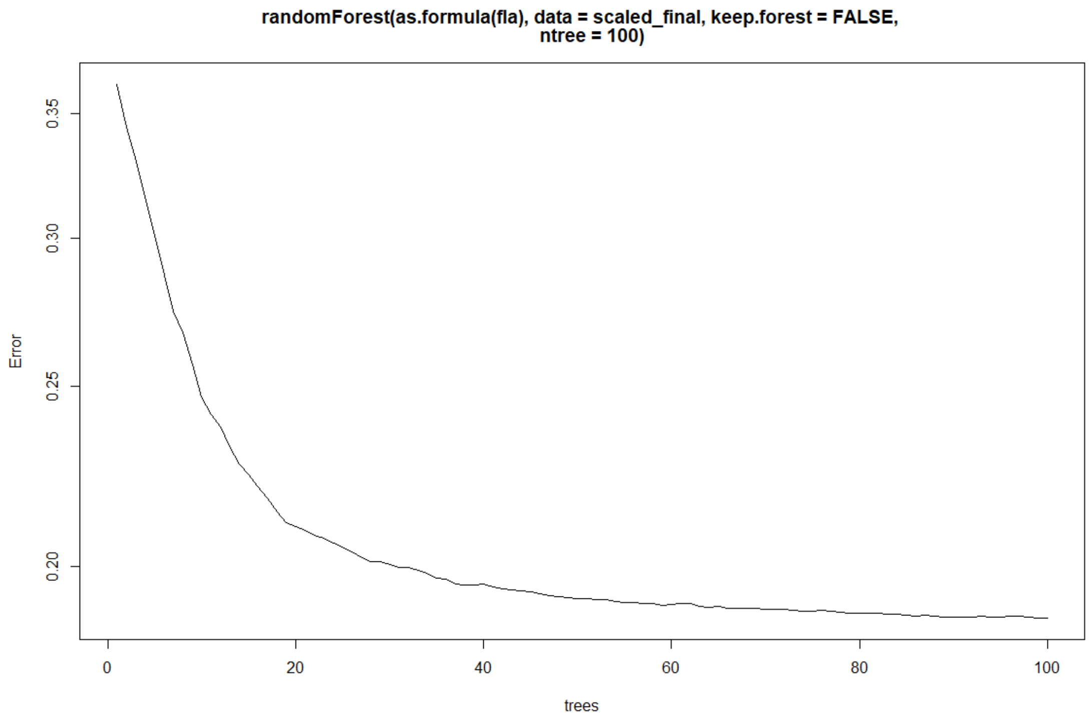


Figure 11

The RFVI used 500 trees with 19 variables split at each, however, we can see that the accuracy was low, and that the error rate remained mostly consistent after just 20 trees.

Since the purpose of our study is to identify *likelihood* or *probability* of becoming a food desert, we want to build a baseline probit regression using all predictors. For this process, we are going to use the sparklyr and dplyr packages.

We split the data into a training and testing split of 75/25 and set the seed to 123 for reproducibility.

```
set.seed(123)
partitions <-  scaled_final %>%
  sdf_random_split(training = 0.75, test = 0.25, seed = 123)

scaled_final_training <- partitions$training
scaled_final_test <- partitions$test
```

We are using the ml_logistic_regression algorithm from spark, where we can use a binomial setting to return probabilities.

```
#fit the first probit model
mysparkprobit <- scaled_final_training %>%
  ml_logistic_regression(fla, data = scaled_final, family = "binomial",  maxi
t = 100)
```

```
pred <- ml_predict(mysparkprobit, scaled_final_test,  type="response")
```

With our baseline model created, we can get an accuracy measure as well as plot some of our results for better interpretation and comparison for future models.

Because we are using all the predictors in our model it is likely that it is overfit. We will want to tweak our model; however, we also want to set baselines for performance and accuracy against some other common algorithms.

As a check on our first Feature Importance algorithm (RFVI in Fig. 11), we are going to run a separate model for validation. This second model uses a spark decision tree that will identify, and rank features based off impact on the response.

```
                  feature  importance
1         PCT_LACCESS_LOWI 0.376882552
2              PCT_NHWHITE 0.145775668
3                    SNAPS 0.084185074
4                    CONVS 0.054166533
5                  POVRATE 0.045278869
6             PCT_18YOUNGER 0.034212061
7                  PCT_WIC 0.030390810
8             CHILDPOVRATE 0.025727681
9                VEG_ACRES 0.025619780
10                PCT_HISP 0.021957443
11              PC_SNAPBEN 0.020810756
12          PCT_FREE_LUNCH 0.015293579
13                    GROC 0.014274198
14               VEG_FARMS 0.014040021
15               VLFOODSEC 0.013789922
16         FOODINSEC_CHILD 0.012921691
17             PCT_65OLDER 0.012177218
18        PCT_LACCESS_HHNV 0.010395186
19                    WICS 0.010036215
20       PCT_REDUCED_LUNCH 0.009512735
21             BERRY_FARMS 0.008168365
22             SNAP_OAPP_0 0.005743964
23       PCT_LACCESS_CHILD 0.005554081
24         PCT_LACCESS_POP 0.003085600
25 PCT_LACCESS_SENIORS 0.000000000
26          SNAP_PART_RATE 0.000000000
27             SNAP_OAPP_1 0.000000000
28       SNAP_FACEWAIVER_1 0.000000000
29       SNAP_FACEWAIVER_0 0.000000000
30          SNAP_VEHEXCL_1 0.000000000
```

```
31            SNAP_BBCE_1 0.000000000
32 SNAP_REPORTSIMPLE_1 0.000000000
33 PCT_DIABETES_ADULTS 0.000000000
34      PCT_OBESE_ADULTS 0.000000000
35             FOODINSEC 0.000000000
36          PCT_LOCLFARM 0.000000000
37                 FMRKT 0.000000000
38         PCT_FMRKT_SNAP 0.000000000
39          PCT_FMRKT_WIC 0.000000000
40     PCT_FMRKT_WICCASH 0.000000000
41       PCT_FMRKT_SFMNP 0.000000000
42       PCT_FRMKT_FRVEG 0.000000000
43   PCT_FRMKT_ANMLPROD 0.000000000
44         ORCHARD_FARMS 0.000000000
45         ORCHARD_ACRES 0.000000000
46           BERRY_ACRES 0.000000000
47               SLHOUSE 0.000000000
48             FOODHUB_0 0.000000000
49        FARM_TO_SCHOOL_0 0.000000000
50                   FFR 0.000000000
51                   FSR 0.000000000
52           PCT_NHBLACK 0.000000000
53           PCT_NHASIAN 0.000000000
54              PCT_NHNA 0.000000000
55              PCT_NHPI 0.000000000
56              MEDHHINC 0.000000000
57              PERPOV_0 0.000000000
58          PERCHLDPOV_0 0.000000000
59               METRO_0 0.000000000
60                SUPERC 0.000000000
61                 SPECS 0.000000000
```

This spark decision tree reports some of the same variable importance rankings as our RFVI model, including poverty rate, residents with low access and low income, and number of SNAP authorized stores. However, it also indicates that the percentage of the populations that is white is a key factor. From our initial exploration of the data, it is likely that there are some of these variables that are confounding, but again, in our first run through, we are going to include all variables and set a baseline.

MODELS

We built four different models for comparisons sake. These included a Logistic Regression (our baseline from above), Decision Tree, Random Forest, and a Gradient Boosted Tree. In order to determine the accuracy of each, we compared them against each other, as well as a control that represents a 50/50 chance of making a correct choice.

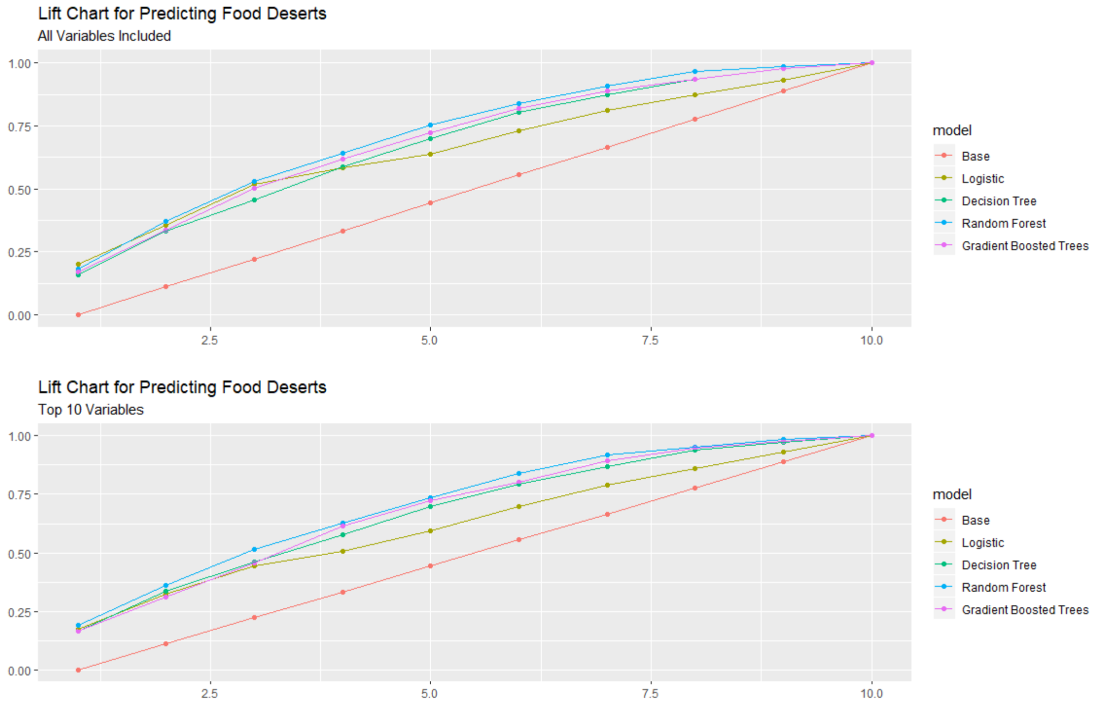
Once we ran through the models using all 59 predictors, we changed our formula to only include the top 10 most important variables per our spark decision tree. These included:

```
1     PCT_LACCESS_LOWI 0.376882552
2          PCT_NHWHITE 0.145775668
3                SNAPS 0.084185074
4                CONVS 0.054166533
5              POVRATE 0.045278869
6        PCT_18YOUNGER 0.034212061
7              PCT_WIC 0.030390810
8         CHILDPOVRATE 0.025727681
9            VEG_ACRES 0.025619780
10            PCT_HISP 0.021957443
```

The variables in total, according to the spark decision tree, account for 84% of the variance in the dataset.

The results for both tests (all variables vs top 10) are seen in Figure 12 and Figure 13 below.

When using all 59 predictors for our models, our logistic regression (the initial baseline) performed the worst, however the decision tree, gradient boosted tree, and random forest performed almost equally, with the random forest model representing the model with the best performance. Not surprisingly, when only using our top ten predictors, which account for 84% of the variance in our data, we see almost identical results; logistic regression has the poorest performance and all other models are consistent. Perhaps most important to note is the fact that all models perform well above the 50/50 baseline. This is specifically illustrated by our lift chart in Fig. 12.



Lift Chart for Predicting Food Deserts

All Variables Included

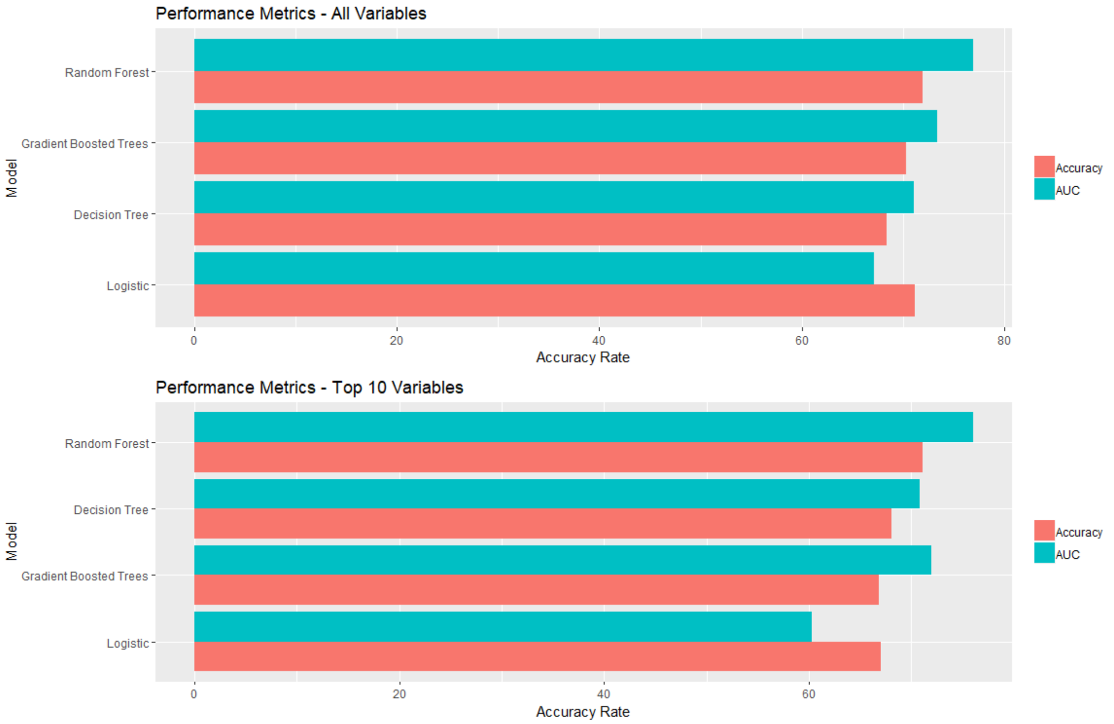
Lift Chart for Predicting Food Deserts

Top 10 Variables

Source: https://cfss.uchicago.edu/notes/sparklyr/

Figure 12

Our second set of visuals in Figure 13 represents the measured accuracy – that is the predictions measured against the actual known values. This is perhaps where we start to see a deviation in performance relative to our models that use all the predictors, and those that use only the top 10. In comparing the accuracy performance between the logistic regression with all variables and the model with only the top 10, we see an accuracy rate of about 62% vs 68%, respectively. This means that the logistic regression model that used all predictors was *less* accurate than the model that used only the top ten. This contrasts with the other models that either had equivalent, or better, performance when using all predictors, rather than just the top ten. However, we continue to see similar performance from our decision tree, random forest, and gradient boosted trees. Overall, the best performing model is the random forest using just the top ten variables based on decision tree importance, with an accuracy rate and AUC of 75% and 71%, respectively.


Performance Metrics - All Variables

Performance Metrics - Top 10 Variables

Source: https://cfss.uchicago.edu/notes/sparklyr/

Figure 13

For our final model, we are going to predict food deserts from 2015 census data.

MAKE PREDICTIONS ON 2015 DATA USING TEST RESULTS FROM 2011

Like our initial dataset, we need to import, pre-process and clean the final dataset. Because the columns are the same, we can simply repeat the processes from above. Once we finished that process, we verify that we are not missing any data.

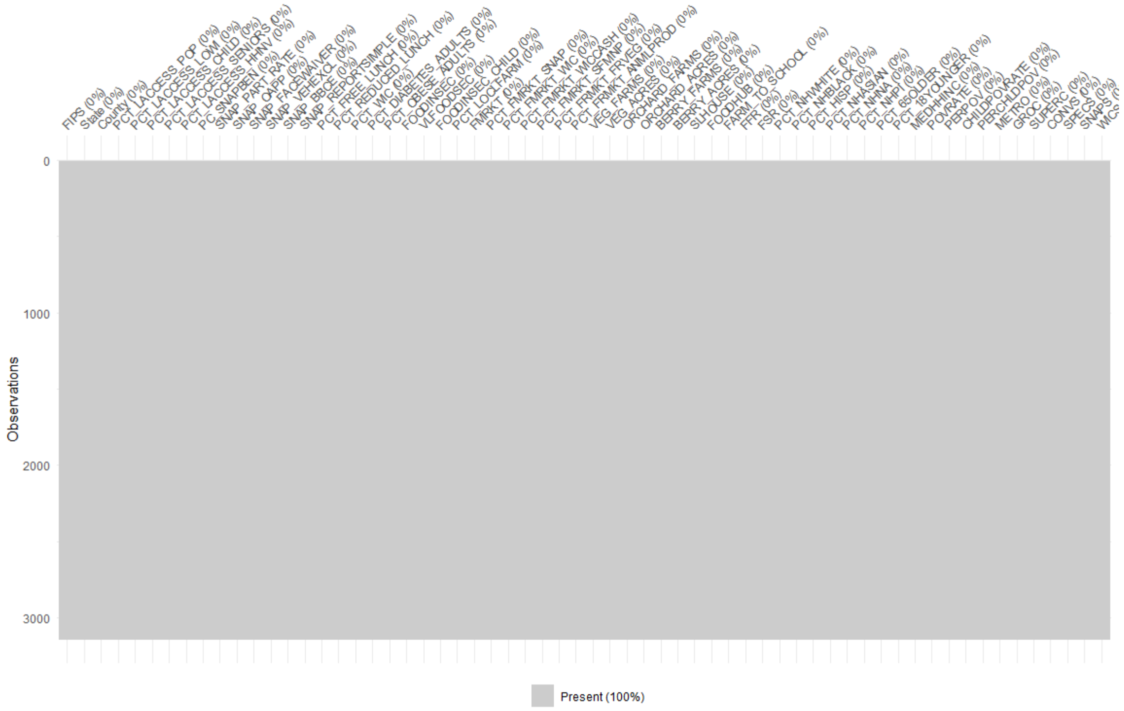


Figure 14

Our ultimate goal is to predict the likelihood, or probability, that a county will contain a food desert, so despite the testing above demonstrating that the Random Forest model is the most accurate, we are going to use a probit regression.

```
final_probit <- scaled_final %>% ml_logistic_regression(fla1, data = scaled_f
inal, family = "binomial",  maxit = 100)
```

```
final_probit_pred <- ml_predict(final_probit, finaloutput2015)
```

```
#separate the prediction information from the demographic info
```

```
final_vis <- final_probit_pred %>%
  select(FIPS, State, County, prediction, probability_1, probability_0)
```

RESULTS

The results of our final model are going to offer predictions and probabilities that a given county will contain a food desert, keeping in mind that food deserts are defined as counties in which all residents must drive more than 10 miles to the nearest supermarket chain or supercenter.

The model used the data from 2011 to make predictions for all 3000+ counties in 2015. A sample of the final results is below. The first row can be interpreted as:

Kent, DE has a 62% chance of *not* containing a food desert based off 2015 information. However, there is a 38% chance that it may.

```
     FIPS  State County     prediction probability_1 probability_0
     <chr> <chr> <chr>           <dbl>         <dbl>         <dbl>
   1 10001 DE    Kent                0         0.380         0.620
   2 10003 DE    New Castle          0         0.472         0.528
   3 10005 DE    Sussex              0         0.430         0.570
   4 1001  AL    Autauga             0         0.308         0.692
   5 1003  AL    Baldwin             0         0.404         0.596
   6 1005  AL    Barbour             1         0.564         0.436
   7 1007  AL    Bibb                0         0.310         0.690
   8 1009  AL    Blount              0         0.263         0.737
   9 1011  AL    Bullock             1         0.736         0.264
  10 1013  AL    Butler              0         0.371         0.629
```

Result Visualizations

All visualizations below are based off the predictions from our final model.

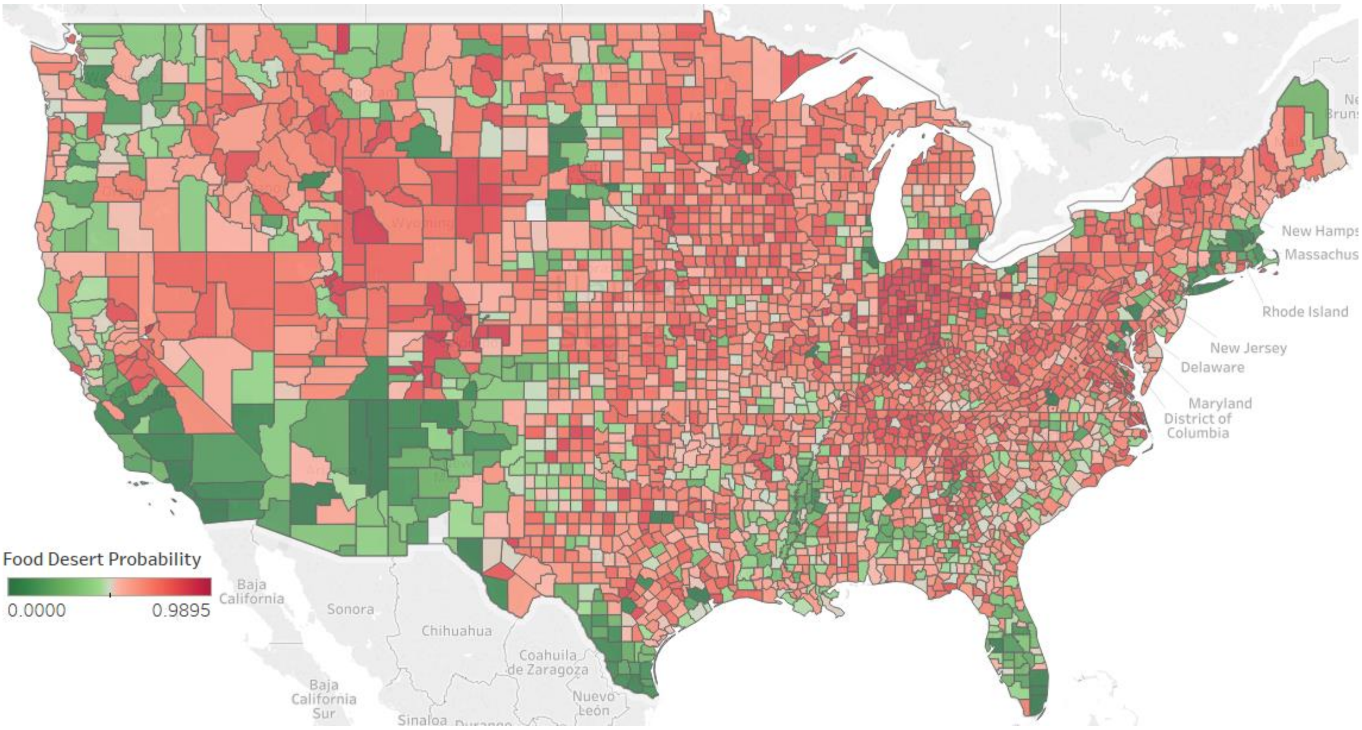
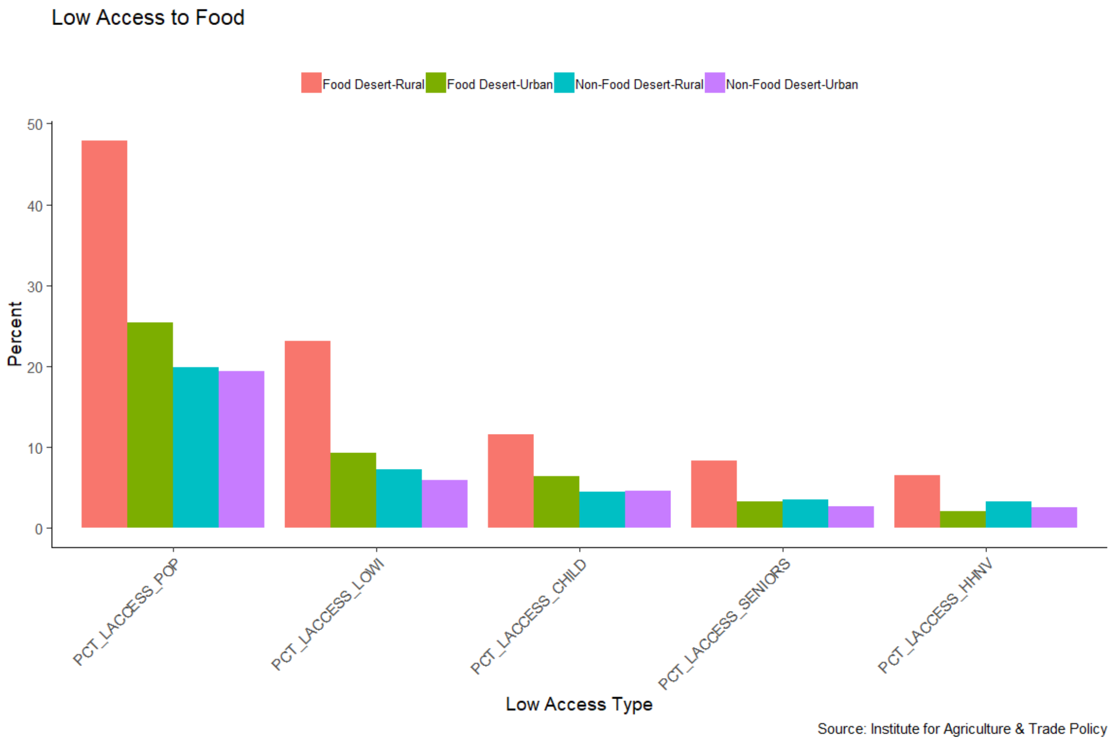


Figure 15

Using a simply interactive Tableau dashboard, we can visualize counties most at risk based off our probit regression. As expected, we see that areas in the mid-west and great plains are the most at risk for containing food deserts, while urban areas (see Denver, Chicago, New York, and Miami as examples) are not likely to contain food deserts.

Additionally, we can derive other descriptive pieces of information by visualizing properties specific to Rural/Urban Food Deserts and Rural/Urban Non-Food Deserts.

For example, figure 16 plots our number one most important feature (low income and low access population percentage) relative to how our model predicts a county is defined in 2015. Rural food deserts have nearly 50% of their population with low access, and ~25% of the population being low-income and having low access to food.

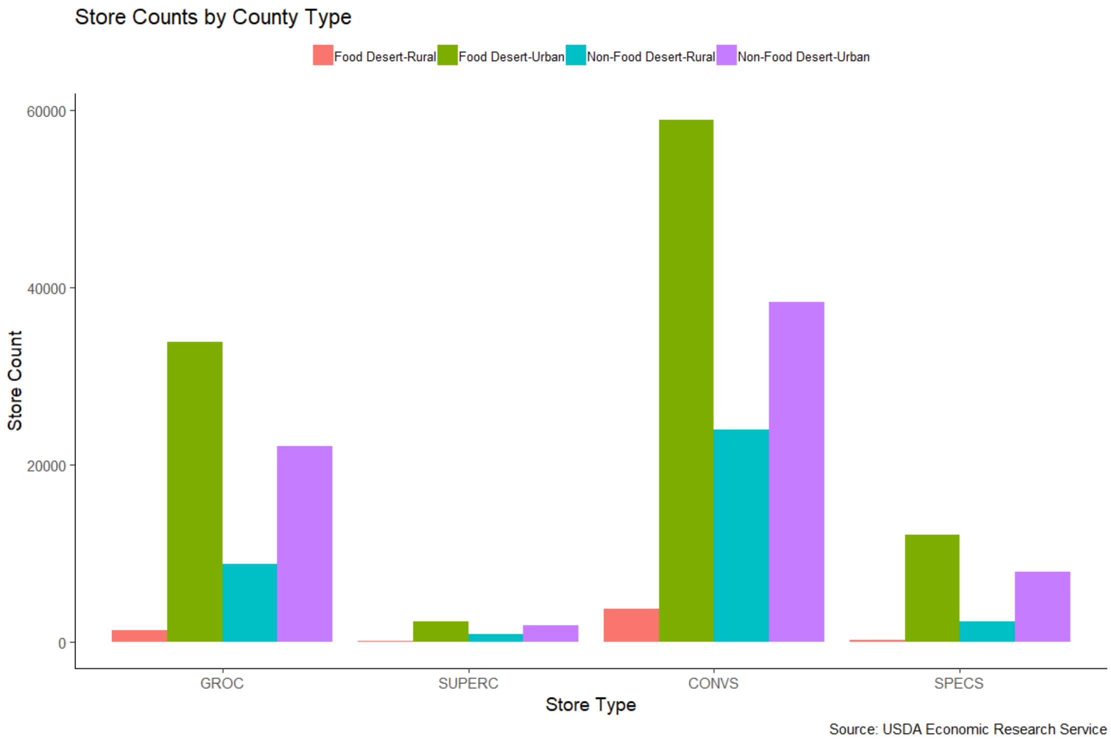

Low Access to Food

Source: Institute for Agriculture & Trade Policy

Figure 16

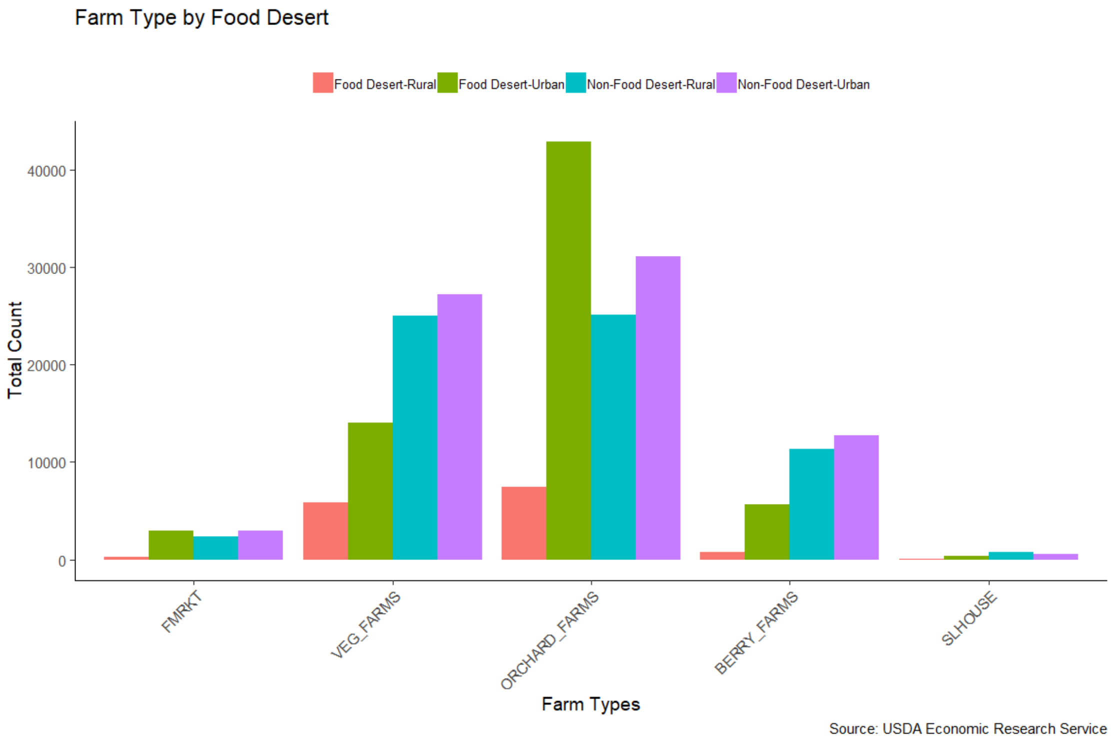
Other plots below visualize properties of urban/rural food deserts and urban/rural non-food deserts, including:

- The total number of groceries, superstores, convenience, and specialty stores
- Number of farmers markets, vegetable farms, orchards, berry farms, and slaughterhouses
- Poverty rates
- Ethnicity
- Industry employment
- Education levels


Store Counts by County Type

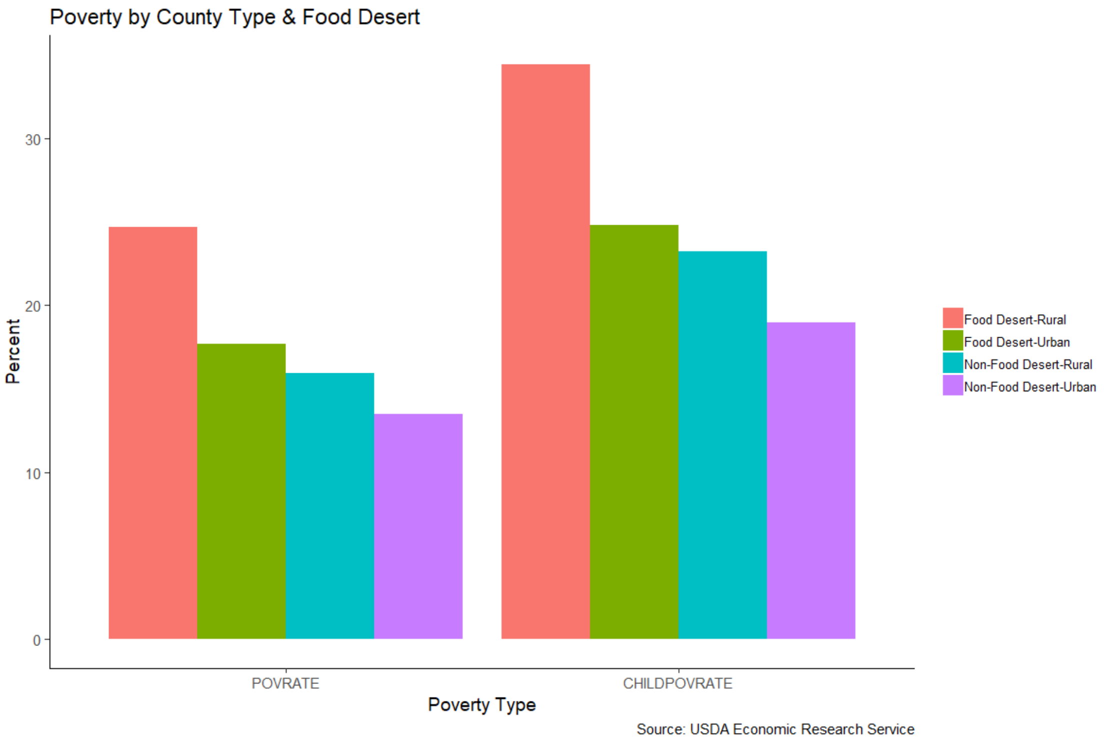
Source: USDA Economic Research Service

Figure 17


Farm Type by Food Desert

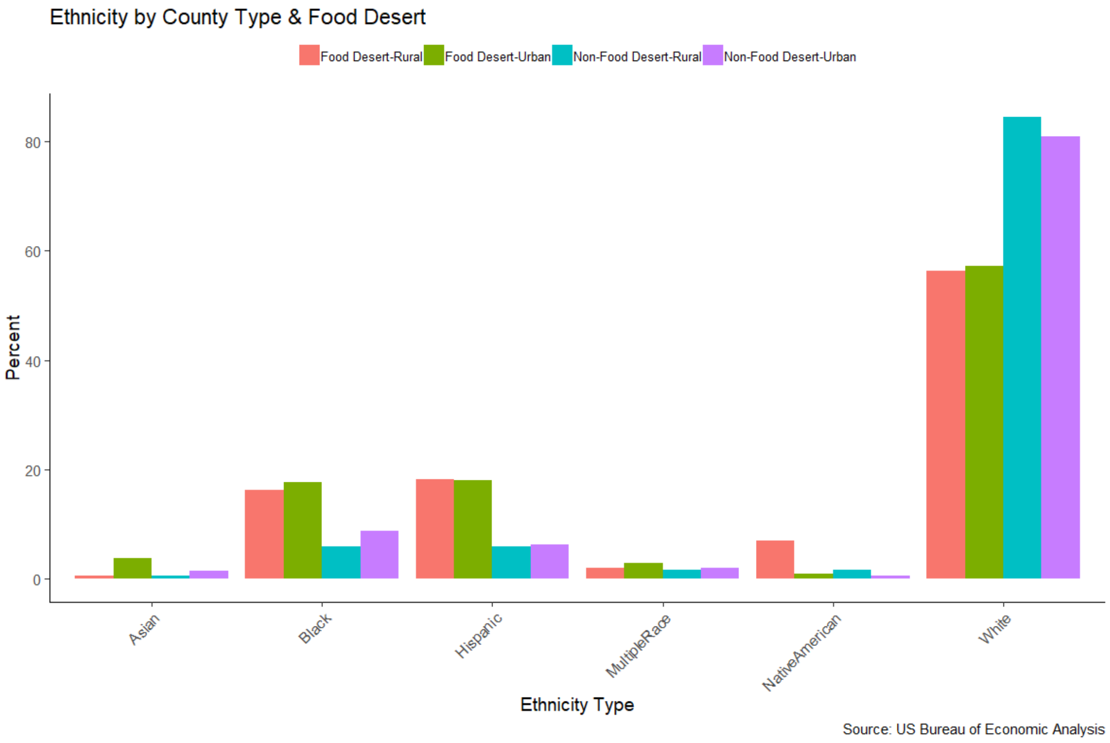
Source: USDA Economic Research Service

Figure 18


Poverty by County Type & Food Desert

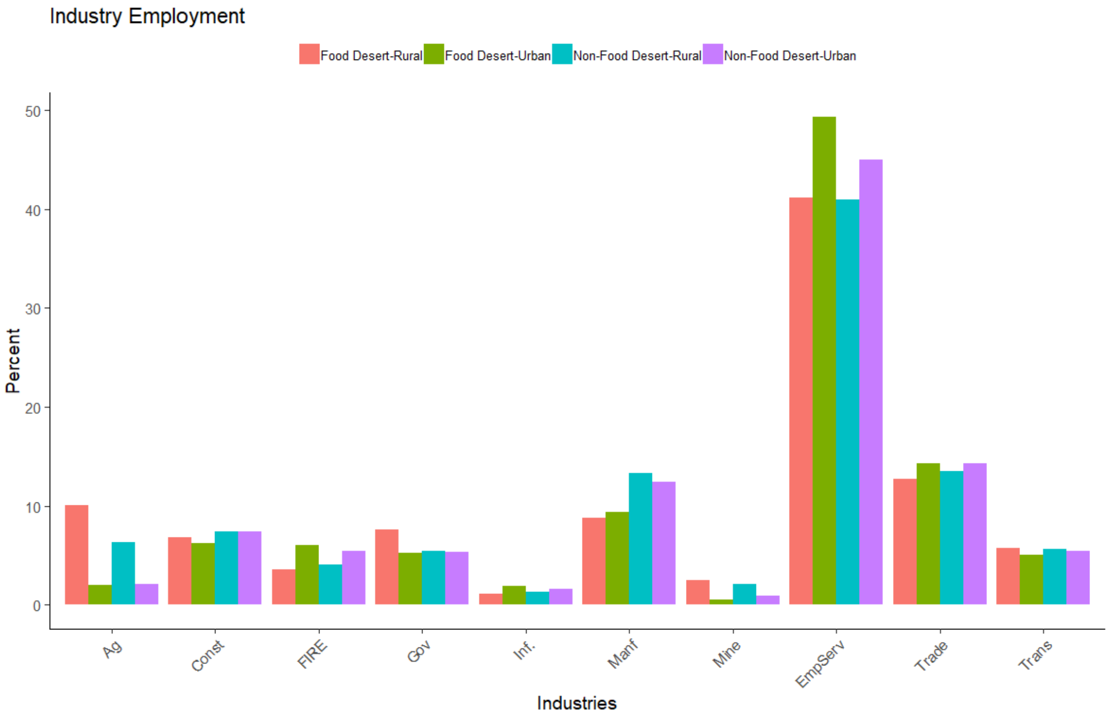
Source: USDA Economic Research Service

Figure 19


Ethnicity by County Type & Food Desert

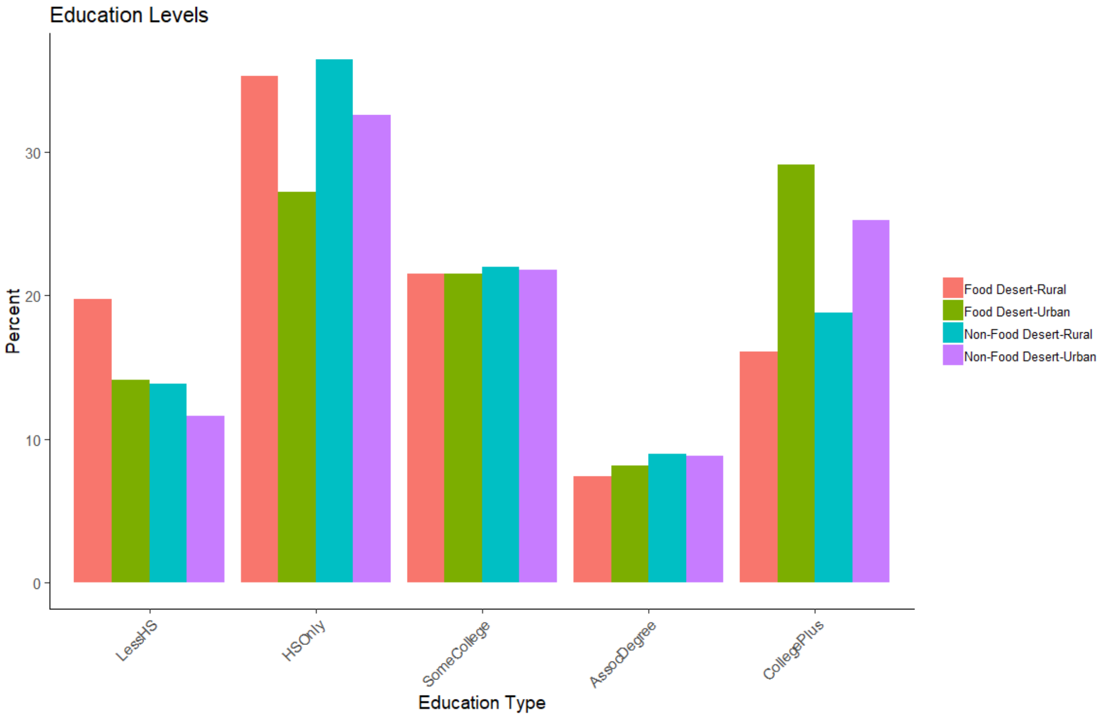
Source: US Bureau of Economic Analysis

Figure 20



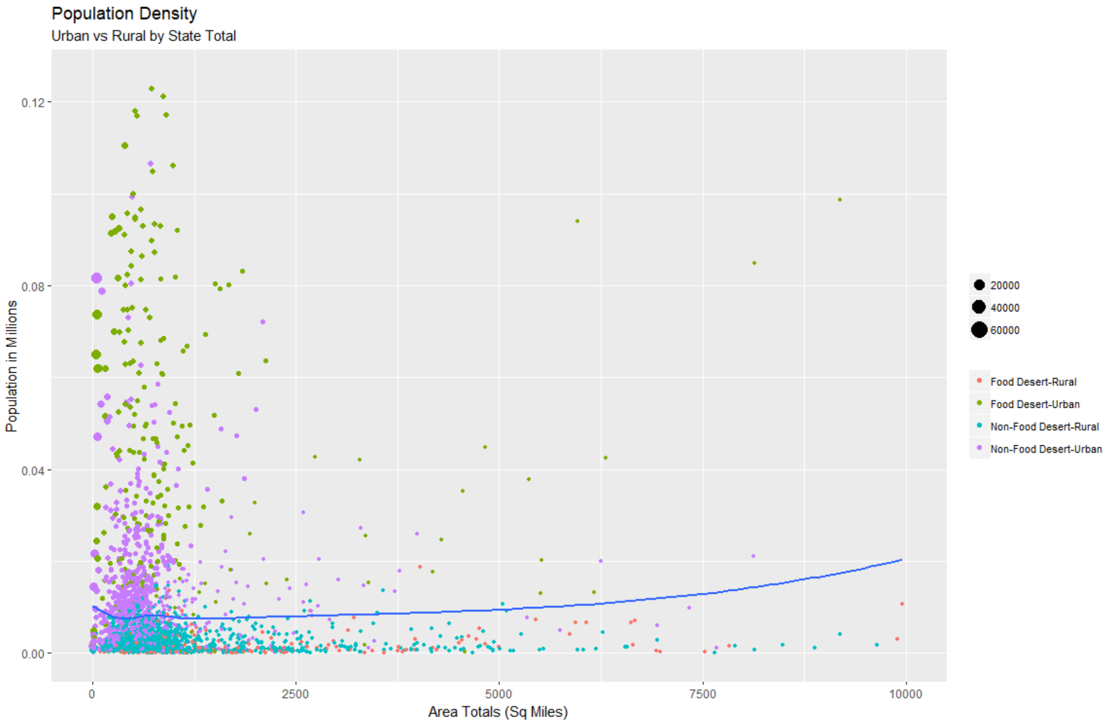
Source: US Bureau of Economic Analysis

Figure 21



Source: US Bureau of Economic Analysis

Figure 22



Source: US Bureau of Economic Analysis

Figure 22

Conclusion

Overall, we were able to identify that with 59 predictor variables, we were able to most accurately predict food deserts with a 75% accuracy rate. When using this same random forest model to predict potential food deserts based off 2015 data, we found that our results were consistent with initial exploratory findings, including:

- Food Deserts tend to consist of people of:
- Low Income
- Low Education
- Employment in Manufacturing and Agriculture
- White

In addition, we found that Rural areas, including those that lost population between 2011 and 2015, tend to have a higher likelihood for containing a food desert. We can see the specific breakdown in figure 23, below.

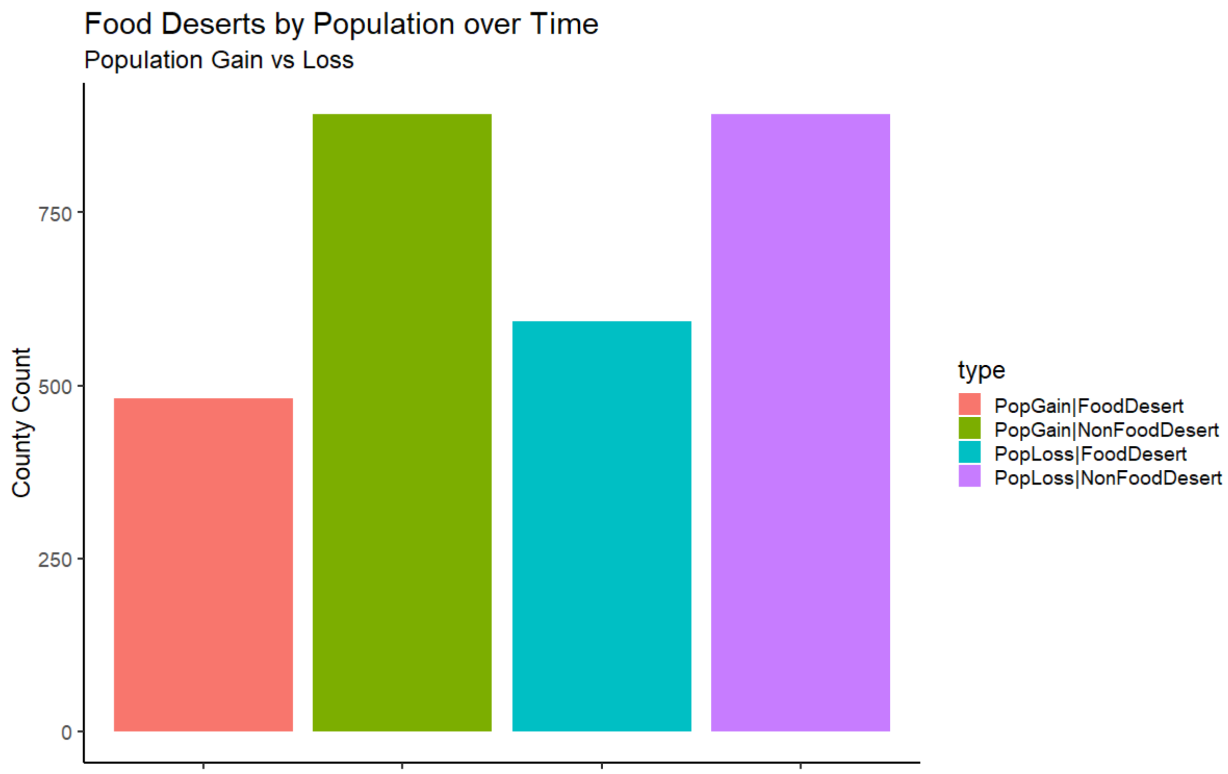

Food Deserts by Population over Time
Population Gain vs Loss
County Count
0
250
500
750
type
PopGain|FoodDesert
PopGain|NonFoodDesert
PopLoss|FoodDesert
PopLoss|NonFoodDesert

Figure 23