## Executive Summary

The goal of this project is to provide insight into the different food environments that exist in the United States. It is not secret that the food we consume has a strong impact on our wellbeing. Ensuring that people have access to healthy and affordable foods is a key part of preventative public health measures. To better understand the social, demographic, and economic characteristics that define food accessibility and food environments in the United States, this project replicates an existing study that uses machine learning techniques to predict access to healthful food retailers.

## 1. Background

At the onset of the COVID-19 pandemic, many called the virus 'the great equalizer' to emphasize that all individuals are at risk of getting infected. However, research has detected disparities in who contracts the disease; who gets treated for it; and how severe the infection will be (Williams, 2020). Michelle Ann Williams, Dean of the Harvard T.H. Chan School of Public Health, explained these disparities through the framework of social determinants of health: "social determinants, or the conditions in which we are born, live, work, and play, are the key drivers of these health outcomes and inequities" (Williams, 2020). This is relevant to the access of healthy foods, which is a vital to ensuring preventative public health measures. Preexisting conditions like diabetes and heart disease, which are lethal in combination with COVID-19, are linked to the lack of healthful food options in rural and urban communities (Williams 2020). For instance, in a study examining the relationship between obesity and neighborhood food environments in southern regions of the United States, Morland (2009) found obesity to be lower in areas that had supermarkets. Additionally, Phillips (2009) found a significant relationship between the presence of food swamps and higher hospitalization rates for adults with diabetes.

To further examine the characteristics of food deserts and food swamps, this project will replicate an existing paper that predicts access to healthful food retailers. Amin, Badruddoza, and McCluskey apply statistical learning models to publicly available data sets from the CDC and the Census Bureau; they achieved an out of sample prediction accuracy of 75% (Amin, 2020). Furthermore, an assessment of variable importance revealed that the best performing model relied on population density; presence of Black population; property values; and income to make its prediction (Amin, 2020). These results have important implications for public health professionals and urban planners. Stakeholders can use these results in concert with domain knowledge to direct resources and funds towards creating tailored solutions for food swamps and food deserts.

Amin et. al's methodology primarily relied on non-parametric machine learning methods (i.e, random forest and extreme gradient boosting). The authors decided to avoid parametric techniques because they felt that the way in which demographic, economic, and social factors determine food environments would not be accurately depicted by using a parametric approach (Amin, 2020). They hypothesized that the flexibility and the lack of functional form assumptions that characterize non-parametric approaches would result in a more accurate model. While Amin et al's hypothesis may hold true (as discussed in detail in this report), the use of non-parametric approaches, particularly ensemble methods, has its own challenges. These methods are

notoriously complex and difficult to interpret. While their complexity often leads to better accuracy, the goal of this research question has pressing policy implications. Simply classifying geographic regions as food deserts and food swamps prevents the valuable insights derived from these models to contribute to the policy-making space in this field. As such, in replicating this study, parametric techniques which are optimal for interpretation/inference are used. Additionally, Amin et. al. chose to use the 1-year estimates from the American Community Survey. This project uses the 5-year estimates because these estimates provide a more accurate representation of small geographic area such as census tracts and block groups.

## 2. Data

Prior studies examining food accessibility often rely on distance measures. Researchers compute how far residents have to travel to access a food retailer, often supermarkets (Amin, 2020). While these studies have enriched our understanding of food deserts (i.e, geographic areas where residents have to go out of their way to access a supermarket), they neglect to identify a different food environment that is often prevalent in cities (Amin 2020). This food environment is characterized as having healthful food retailers, often in close proximity. However, healthful food retailers are either too expensive for residents to afford or are outnumbered by non-healthful food retailers such as corner stores and fast-food restaurants. It would be difficult to characterize this food environment by using a distance-based measured. As an alternative, Amin et al. use a density-based approach which identifies the share of healthful food retailers out of all food retailers in a census tract. By using a density-based approach, both food deserts and food swamps are able to be identified and characterized. Ultimately, the characteristics associated with both food environments will result in customized policy solutions instead of a one-size-fits-all strategy (Amin, 2020).

There are two data sources for this project. The first is the Modified Retail Food Environment Index (mRFEI). mRFEI was developed in 2011 by the CDC; it is a continuous density-based measure that represents the percentage of healthy food retailers in a census tract (CDC, 2011). The second source is the 2010 5-year American Community Survey which provides the most accurate tract-level estimates of community and individual characteristics such as age, race, income, poverty, educational attainment, and vehicle availability (United States Census Bureau, n.d.).

### 2.1 Characterization of selected variables

1. *Dimensionality*: each of the attributes obtained from the ACS survey were merged onto the mRFEI index. The merged data frame has 65,345 observations (i.e, census tracts) and 20 attributes (Table 1). Since the number of observations significantly exceeds the number of predictors, this provides more flexibility during preprocessing the data: as will be mentioned below, there are several observations that are missing values across many attributes. These observations were dropped without having to worry about ‘the curse of dimensionality’ creating complexities. In fact, after removing these observations, the dataset was comprised of 49,561 observations.

2. *Attribute type and data model*: out of the 20 attributes four will be highlighted: the mRFEI Index, population density, median household income, and percent of the

population that utilizes public transportation (Table 2). mRFEI index is the source of the outcome variables; the remaining three are among the 20 predictors Amin et al. found to be the most important.

mRFEI, population density, median household income, and public transportation are all continuous attributes: mRFEI has values ranging from 0 to 100 , while public transport has values ranging from 0 to approximately 96 (Table 2). Conversely median household income ranges from 5,000 to 249, 194 while population density ranges from 0.03 to 224,716. These attributes (along with all other attributes in the dataset) are also quantitative and ratio, as they all have a true zero where both multiplication and division are meaningful (Table 1). In looking at these ranges, it is immediately clear that scale may be an issue: unless they are standardized, measures like household income and population density will overpower prediction results.

With regards to the data model, all of the attributes in the data set were pulled from a record type data model and have the census tract as the unit of observation. Since the geographic scope of this study is the census tract level, the data frame already has an appropriate resolution.

Table 1: Characterization of attributes and datasets

| VARIABLES | DATA MODEL | ATTRIBUTE TYPE | UNIT OF OBSERVATION | NUMBER OF MISSING VALUES |
|---|---|---|---|---|
| mRFEI index | record | continuous; quantitative; ratio | census tract | 1,221 |
| Median household income | record | continuous; quantitative; ratio | census tract | 14,671 |
| Population density per square mile | record | continuous; quantitative; ratio | census tract | 14,537 |
| Land area | record | continuous; quantitative; ratio | census tract | 14,537 |
| Population using public transportation (%) | record | continuous; quantitative; ratio | census tract | 14,591 |
| Share of housing units without a vehicle | record | continuous; quantitative; ratio | census tract | 14,622 |
| White population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| Black population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| Asian population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| American Indian and Alaska Native population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| Native Hawaiian and other Pacific Islander population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| Hispanic population (%) | record | continuous; quantitative; ratio | census tract | 14,534 |
| Poverty rate | record | continuous; quantitative; ratio | census tract | 14,607 |
| Households with SNAP | record | continuous; quantitative; ratio | census tract | 14,622 |
| Inequality | record | continuous; quantitative; ratio | census tract | 14,648 |
| Unemployment | record | continuous; quantitative; ratio | census tract | 14,591 |
| Below high school | record | continuous; quantitative; ratio | census tract | 14,539 |
| College no degree | record | continuous; quantitative; ratio | census tract | 14,539 |
| Bachelor's or more | record | continuous; quantitative; ratio | census tract | 14,539 |
| Property value | record | continuous; quantitative; ratio | census tract | 15,060 |
| Rural population | record | continuous; quantitative; ratio | census tract | 14,503 |

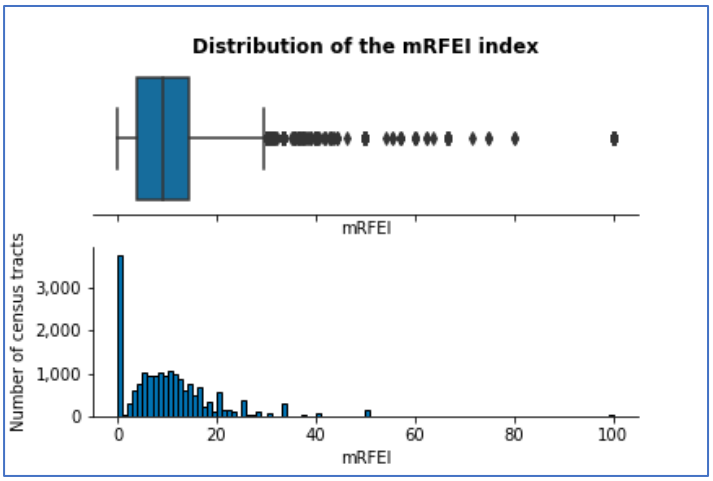
3. *Missingness*: the mRFEI attribute has the 1,221 missing values (Table 1). Removing observations is not ideal as it can mean losing patterns that can aid in the modeling process; however, since the data set has 65,345 census tracts, that 1,221 tracts are missing mRFEI values is not a serious concern. They constitute just 1.9% of all observations. These census tracts will be dropped since the mRFEI is the target variable for this project.
   Out of the four attributes being highlighted, median household income has the highest number of missing values at 14,671 (Table 1). After further exploration of the other

variables, a pattern of missingness was revealed: there are several census tracts that are missing values for groups of variables from the ACS. For instance, the same 14,534 census tracts are missing values for all the race and ethnicity attributes (Table 1). The same applies for the educational attainment attributes. Because of this systematic missingness, imputation is not the ideal method. As such, these census tracts will be dropped from the analysis. Note that the data frame will still have 49, 561 observations.

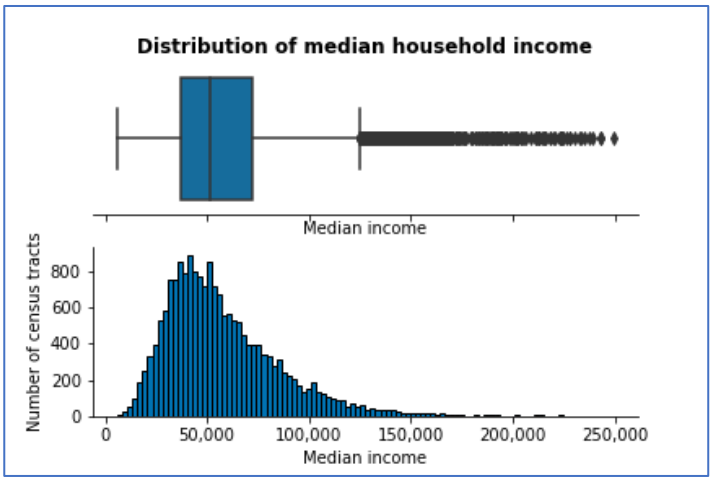
Table 2. Descriptive statistics of selected variables after removing observations with missing values

| | MREFI | MEDIAN HOUSEHOLD INCOME | POPULATION DENSITY | PUBLIC TRANSPORT |
|---|---|---|---|---|
| Count | 49,561 | 49,561 | 49,561 | 49,561 |
| Mean | 11.28 | 5,4646.44 | 5,627.28 | 6.08 |
| Standard deviation | 11.98 | 26,789.84 | 11,842.86 | 12.36 |
| Minimum | 0.00 | 5,000.00 | 0.03 | 0.00 |
| 25% | 3.13 | 36,611.00 | 450.46 | 0.00 |
| 50% | 9.09 | 48,946.00 | 2,526.72 | 1.26 |
| 75% | 15.38 | 66, 217.00 | 5,630.36 | 5.50 |
| Maximum | 100.00 | 249,194.00 | 224,716.67 | 95.73 |

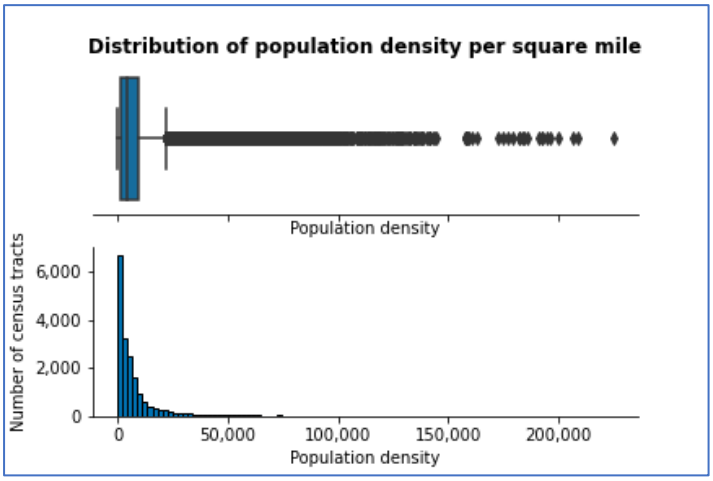
4. *Distributions and Outliers:* the graphical summaries below clearly show a pronounced skew in the distribution of all four attributes (Figure 1). Looking closely at the mRFEI Index, there is a bimodal distribution where the first peak occurs at 0 and the second peak occurs at about 11 (Figure 1: A). As shown in Table 3 and the figures below, the median of this outcome variable is 9.09 while the mean is 11.28—it is comforting to know that the median is close to the mean because, given the presence of outliers, it would not be surprising to see a mean that is much higher. Conversely, the descriptive statistics of public transportation and population density show that the median is significantly different than the mean (Table 2). This is supported by the graphical summaries below where both distributions are significantly skewed and have a large set of outliers that must have impacted the estimation of the mean (Figure 1: C & D). Additionally, the distribution for population density shows that there are quite a lot of census tracts with a population density of zero. Even after removing these observations, given that they do not have any residents, the distribution remained skewed since majority of the census tracts have low population densities, with the lowest being 0.03 (Table 2). Note that these census tracts were retained because they have inhabitants and, thus, are eligible to be considered as a desert, swamp, or a healthful tract.



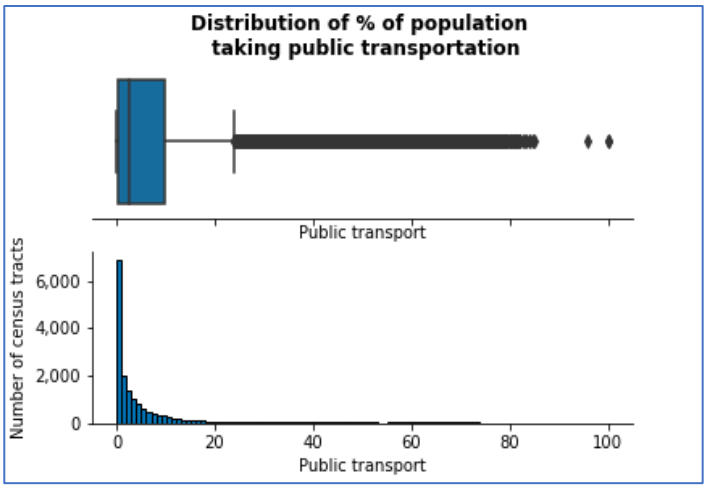

A




B




C




D

*Figure 1 Graphical summary of the distribution of selected attributes*

5. *Correlations/Relationships*: the correlation matrix depicted below shows that the mRFEI attribute does not have a prominent linear relationship with any of the three predictors being highlighted (Figure 2). Given that mRFEI is the source of our target variables, this may mean that these attributes have a non-linear relationship with the outcome. As such, non-parametric approaches will be considered to accurately predict this target. Additionally, population density and public transport are high correlated

with one another (Figure 2). This has implications for our modeling decision since regression methods will be used for all outcome variables. Normally, this would mean that one of the highly correlated attributes would need to be removed to avoid redundancy. However, since Amin et al. selected the 20 attributes by using content expertise (aided by a boosting algorithm), this replication will refrain from dropping any attributes. This will undoubtedly have serious implications on the accuracy of the regression models. However, a non-parametric model which is less sensitive to redundancy will be used to get high accuracy.

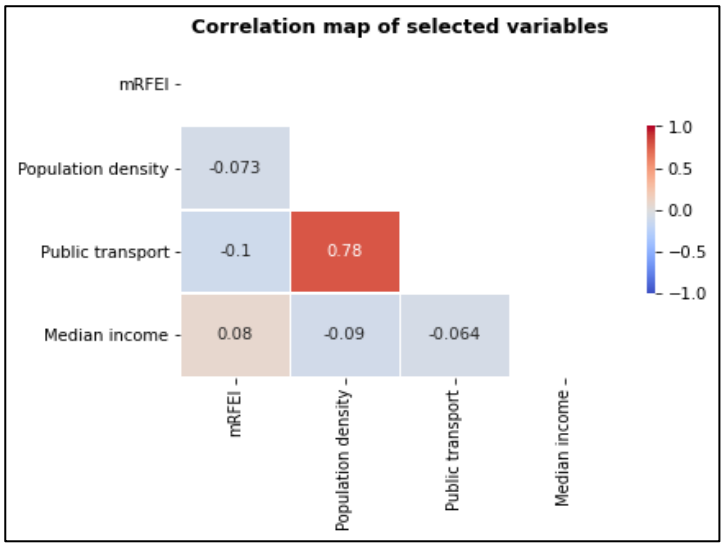



*Figure 2: Correlation Matrix of the mRFEI Index, population density, public transportation, and household income*

## 2.2 Limitations of datasets

A key limitation of the datasets has to do with timeliness/relevance. The CDC has not released an updated version of the mRFEI index; as such, all the figures used to generate the predictive models are 10 years old. This has implications for the performance of the model on unseen data. Although Amin et al.'s research design does contribute to our understanding of food environments and food accessibility, the age of the data sets may prevent the models from being relevant and effective at detecting current food deserts and swamps. There is no guarantee that the characteristics that defined these food environments 10 years ago are the same today.

## 3. Methodology

Following Amin et al's research design, there will be three separate outcome variables derived from the mRFEI index. Thus, there will be three modeling attempts to predict each outcome variable. All three will be binary outcomes.

As noted above, mRFEI values range from 0 to 100: tracts with a value of 0 do not have healthful food retailers and are deemed as food deserts. Tracts with lower scores that are greater

than 0 contain more unhealthy retailers (ex: fast food restaurants) than healthful ones. These tracts are called food swamps. Tracts with high scores contain more healthful retailers than unhealthful ones. These tracts will be considered healthful areas. To create three mutually exclusive outcomes, the median of the mRFEI attribute is used as a cutoff: tracts with mRFEI values less than 9.09 and greater than 0 are food swamps; tracts with values greater than 9.09 are healthful; and tract with a value of 0 are food deserts (Figure 3). Using these mutually exclusive attributes, three binary variables were created. The first outcome, titled *Desert vs. Healthful,* takes on a value of 1 if the census tract is food desert and a value of 0 if the tract is healthful. All food swamps are dropped from the data set. The second outcome, titled *Swamp vs. Healthful,* takes on a value of 1 if the census tract is a food swamp and a value of 0 if the tract is healthful. All food deserts are dropped from the data set. The last outcome, titled *Desert vs. Swamp,* takes on a value of 1 if the census tract is a food desert and a value of 0 if the tract is a food swamp. All healthful tracts are dropped from the data set. This process results in three data sets, each with the same number of attributes but a different outcome.

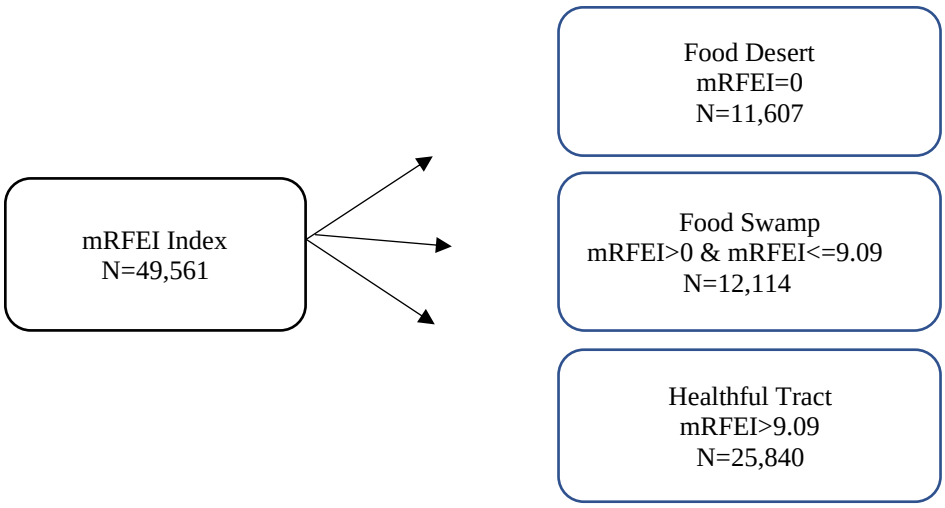



*Figure 3: Derivation of three outcome variables from the mRFEI Index*

The feature set consists of 20 tract-level demographic and economic characteristics obtained from the American Community Survey (Table 1).

### 3.1 Parametric Techniques

A logistic LASSO regression will be applied to the numerical outcome and the three categorical outcomes, respectively. Regression approaches were selected because their simplicity allows for ease of interpretation. Logistic regression was selected for the categorical outcomes because linear regression produces estimates for the probability of class belonging that surpass the acceptable bound between 0 and 1. The LASSO was selected because of its ability to shrink coefficients to zero without having to eliminate what could possibly be important attributes as in subset selection methods. This is especially important since one goal of this project is to infer about the characteristics that differentiate food deserts, food swamps, and healthful areas. Unlike

regular logistic regression, the LASSO not only maximizes the log-likelihood, but it also applies a penalty to the coefficients. Because of the nature of the penalty, LASSO shrinks some of the coefficients to exactly zero, resulting in sparse models with a subset of the predictors having coefficients greater than 0. This technique will make it easier to identify a subset of variables that can be used to characterize and differentiate food deserts, food swamps, and healthful areas. In a similar application, Kassambara (2018), used logistic LASSO regression on a data set containing eight clinical features used to predict the probability of being diabetic. After using LASSO, three out of the eight predictors were shrunk to zero, leading to a sparse and more interpretable model (Kassambara, 2018). While the parametric techniques above are helpful for inference, they are sensitive to outliers. As shown in the above section (and discussed in the implementation appendix), the distribution of most the attributes is heavily skewed, even after log transforming. This will probably affect the accuracy of the classification results from the logistic regression.

### 3.2 Non-parametric Technique

A random forest will be applied to the categorial and numerical outcomes. The random forest is a technique that feeds-off the statistical method of bootstrapping; it's comprised of a group of decorrelated decision trees. Random forest was selected because of its ability to scan through the predictor space and randomly select a subset of attributes when growing the decision trees. This technique is especially important for this project because it actually samples subsets of demographic and economic characteristics in the feature space with replacement and without imposing a restrictive functional form that could cause bias. It will be able to detect patterns that will probably be missed by the parametric methods above. Interpretability will be a challenge; however, the predictive accuracy that is associated with this technique is an appropriate tradeoff, given that its results can be considered alongside that of the parametric techniques. It will be especially interesting if the comparison of variable importance and LASSO coefficients will line-up with what Amin et al. found. Previous application of random forest to predict food access has had positive results: Patel (2019) used population health measures and twitter sentiment analysis to predict food deserts. Random forest had an AUC score of 0.87—second only to Gradient Boosting which had an AUC of 0.88 (Patel, 2019).

## 4. Findings

### 4.1 Deserts vs. Healthful Tracts

Hyper parameter tuning was conducted for the LASSO model (see Implementation Appendix for details). A validation curve was constructed by using 10-fold cross validation to generate average train and test scores for the different parameter values (Figure 4). When the C parameter is equal to 0.021, train and test scores are close to each other and the test score is the highest (Figure 4). After this point, the test score remains constant. While the train and test curves overlap significantly between 0.001 and 0.006, the test score is comparatively lower than at 0.021 (Figure 4). Choosing 0.021 over 0.001 or 0.006 will introduce some variance, but the increase in accuracy is worth this tradeoff. Since LASSO already increases bias by shrinking coefficients to zero (and thus decreases variance), introducing a little more variance should not lead to overfitting. Therefore, a C parameter value of 0.021 was used when fitting the model to the training set.

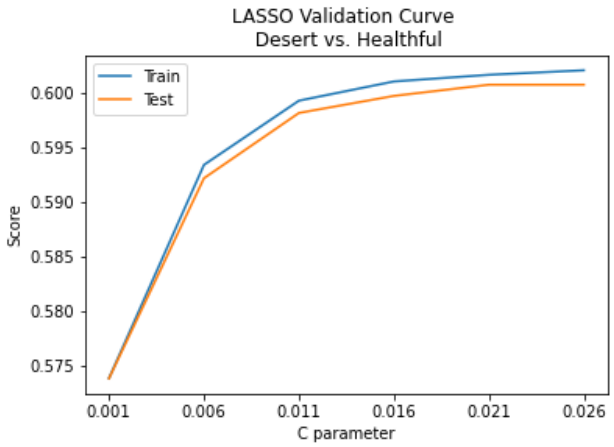



*Figure 4: LASSO validation curve for classifying the outcome, Desert vs. Healthful*

Accuracy and recall were used were used to evaluate and compare model performance. Since the goal of this classification task is to accurately identify and characterize deserts, misclassifying a healthful tract as a desert does not have the same implications as misclassifying a desert. Given the policy goal of identifying and understanding areas without access to healthful foods, it is far more ideal to ensure that as many of the deserts in the data set are being classified as deserts. As shown in Table 3, the random forest had the highest overall accuracy; this is not surprising given the lack of restrictions imposed on the functional form. However, the random forest had the worst recall. It is correctly classifying the deserts in the data set only 47% of the time (Table 3). Conversely, the LASSO and logistic regression both had a comparatively higher recall score of 0.68--both were able to correctly classify 68% of food deserts (Table 3). The accuracy of these models, on the other hand, is quite low: LASSO had an accuracy of 0.59 and logistic regression had an accuracy of 0.58 (Table 3). These are only slightly better than a coin-flip. This is not surprising given the skewedness of some of the attributes.

Table 3: Model evaluation for all target variables

| | Accuracy Logistic regression | Recall Logistic Regression | Accuracy LASSO | Recall LASSO | Accuracy Random Forest | Recall Random Forest |
|---|---|---|---|---|---|---|
| Desert vs. Healthful | 0.58 | 0.68 | 0.59 | 0.68 | 0.69 | 0.47 |
| Swamp vs. Healthful | 0.62 | 0.73 | 0.62 | 0.75 | 0.67 | 0.51 |
| Desert vs. Swamp | 0.74 | 0.65 | 0.71 | 0.57 | 0.74 | 0.64 |

Variable importance from the random forest and coefficient estimates from the logistic and LASSO regressions were generated for inference purposes (Figure 5). For the logistic regression, it appears that land area, population density, and property value have the highest coefficient magnitude. The sign on the estimates is negative, indicating that tracts with high land area, high

population density, and high property values are less likely to be food deserts. Conversely, having a high rural population; a high number of residents who rely on public transportation; and a high median household income makes a census tract more likely to be a food desert. It is surprising that land area is negatively correlated with food deserts, since food deserts are normally associated with rural areas that sit on large land areas. However, it could be very likely that, when compared to deserts, healthful tracts are more likely to be located on large land areas.

The LASSO was able to shrink some of the coefficient estimates to zero, leaving behind a sparse model. Despite the shrinking, the attributes identified by the logistic regression as having the highest magnitudes, regardless of the direction of correlation, were retained. The estimates from the LASSO mirror the logistic regression estimates.

The variable importance from the random forest does not provide any information about the correlation between the attributes and the outcome; it indicates which attributes were most important to split-on in terms of having node purity. Population density, land area, share of the population without a vehicle, and share of the population who is Black were the attributes the random forest relied on most to make its classifications. It is important to note that population density and land area had the highest magnitudes in the LASSO and logistic regression estimates.

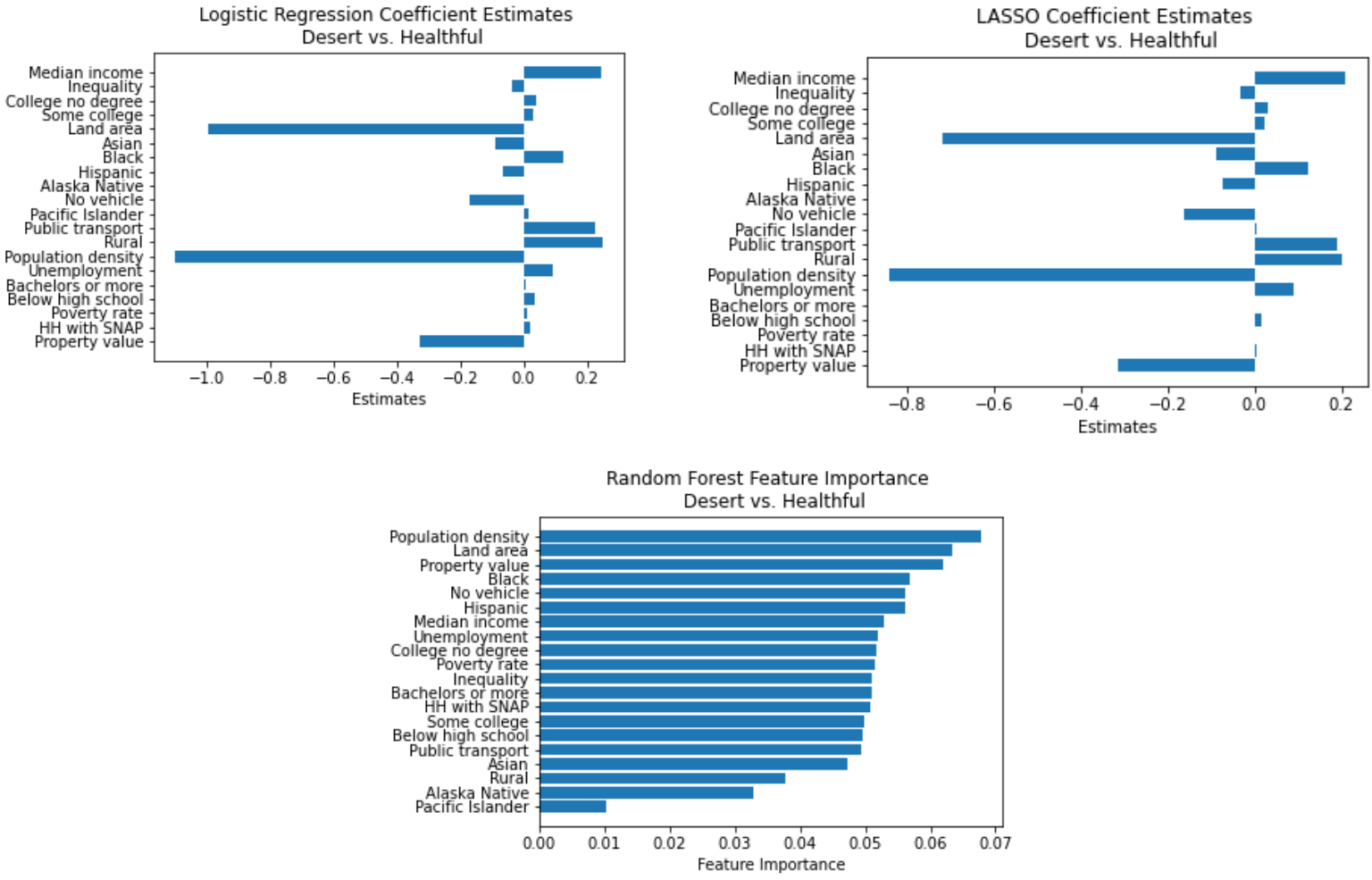



*Figure 5: LASSO and logistic regression coefficient estimates and feature importance from the random forest model for the Desert vs. Healthful outcome.*

### 4.2 Swamps vs. Healthful Tracts

The validation curve for the C parameter showed that 0.026 was the point at which the test score is the highest and also where the distance between the train and test score is the smallest (Figure 6). This means that there will be less overfitting since the score will not change dramatically. For these reasons, 0.026 was selected to be the value of the C parameter when fitting the model to the training set.

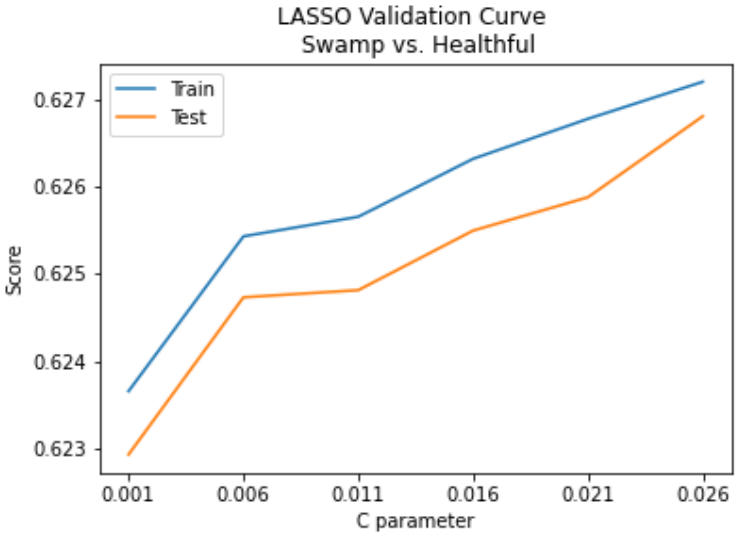



*Figure 6: LASSO validation curve for classifying the outcome, Swamp vs. Healthful*

The classification results when using swamps vs. healthful tracts as the outcome variable were comparatively better than the previous section. The overall accuracy of the LASSO and the logistic regression increased to about 0.62 (Table 3). The recall also increased substantially: the LASSO was able to correctly classify 75% of swamps in the data set; logistic regression correctly classified 73% of swamps (Table 3). This increase in overall accuracy and recall could be attributed to the fact that the imbalance in this data set was not as pronounced as in the desert vs. healthful data set. This means comparatively fewer of the observations in the minority class of the training set were synthetically generated. Having more examples of actual food swamps probably provided more information for both models to make better classifications. Unlike the other two models, the overall accuracy of the random forest actually decreased while the recall increased. Nevertheless, the random forest continues to outperform both models with regard to overall accuracy and to underperform with regard to recall.

The inference portion of this modeling attempt depicted an opposite story from the previous section, confirming Amin et al.'s hypothesis that food deserts and food swamps are different food environments with difference economic and demographic characteristics. The logistic regression estimates show that population density, land area, and the share of the Black population are positively correlated with swamps (Figure 7). Recall that these were negatively correlated with food deserts. Interestingly, the LASSO actually shrinks land area significantly (Figure 7). As a result, the share of the Hispanic population and the share of the population with a bachelor's degree or more have a higher magnitude. This indicates that tracts with a high population density, with a high number of Black and Hispanic residents, and with residents that have high educational attainment are more likely to be food swamps. Conversely, both models found that the share of the

population that lives in the rural part of the tract and tracts with high property values are less likely to be swamps.

The feature importance from the random forest reveals that population density, share of the Black population, and land area were especially important (Figure 7). Although not in the same order, these variables were among the most important attributes in the previous section as well. While the relationship of these attributes to the outcome cannot be definitively determined using just the feature importance, it still provides important considerations for characterizing food environments—population density, race and ethnicity, land area, and mobility could have policy implications as they continue to show up in the coefficient estimates and the feature importance tables.

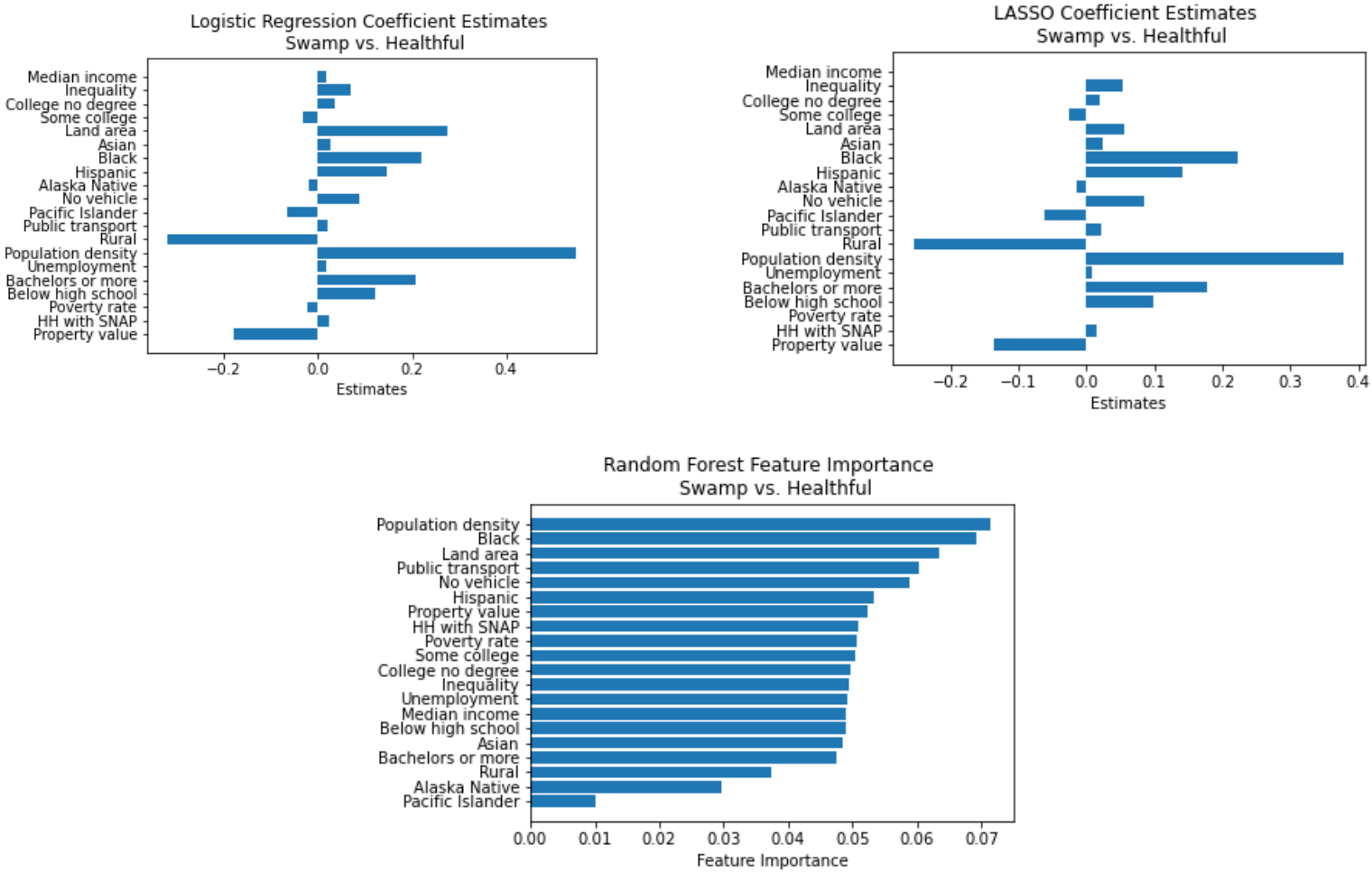



*Figure 7: LASSO and logistic regression coefficient estimates and feature importance from the random forest model for the Swamps vs. Healthful outcome.*

### 4.3 Deserts vs. Swamps

The validation curve for the LASSO hyper parameter is depicted below (Figure 8). Since the train and test curves are closest to one another at 0.001, this value was selected for the C parameter when fitting the model to the training data. Although the test score is highest at 0.006,

the variance is also high at this point as evidenced by the gap between the train and test score (Figure 6). To minimize the possibility of overfitting, 0.001 was selected. Additionally, the increase in the test score associated with 0.006 is only higher by about 0.01 units (Figure 6). This increase in score is not high enough to warrant an increase in variance.

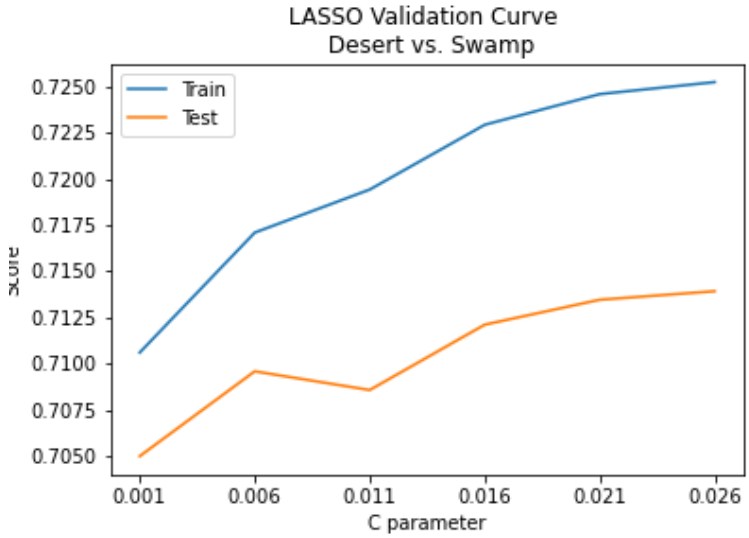



*Figure 8: LASSO validation curve for classifying the outcome, Desert vs. Swamp*

As shown in Table 3, the logistic regression and random forest had the same accuracy score and very similar recall scores. The logistic regression actually overperformed in terms of recall. The LASSO had the lowest accuracy and recall scores--the LASSO recall score from this modeling attempt was actually the lowest than from the previous two sections. LASSO was able to correctly classify 57% of deserts and incorrectly classified the remaining deserts as swamps (Table 3). Overall, it seems like the logistic regression actually performs better than the random forest when employed on a data set that was fairly balanced and didn't require oversampling techniques. Conversely, the attributes could have been more helpful in distinguishing between deserts and swamps rather than distinguishing between deserts and healthful tracts or swamps and healthful tracts.

The most prominent coefficient estimates from the logistic regression are very similar to the previous sections (Figure 9). As before, a higher number of rural residents in a census tract is associated with food deserts. However, most of the attributes actually have very low coefficient estimate magnitudes (Figure 9). This fact is further pronounced in the LASSO estimates, where a majority of the attributes have been shrunk to zero, leaving behind an extremely sparse model. The attribute 'Rural' continues to be positively correlated with deserts, whereas population density, share of the population who is Hispanic, and share of population without a vehicle are negatively correlated with deserts (Figure, 9). These attributes seem to be important in distinguishing between deserts and swamps. One thing to note is that in the logistic regression, land area continues to be negatively associated with deserts. Perhaps the desert tracts retained the

data set are sitting on small land areas. Note that this attribute was actually shrunk to zero by the LASSO model. It seems the LASSO did not find this attribute to be important in distinguishing between deserts and swamps.

The random forest continues to rank population density, the share of the population that lives in the rural part of the census tract, and land area as its top three most important features (Figure 9).

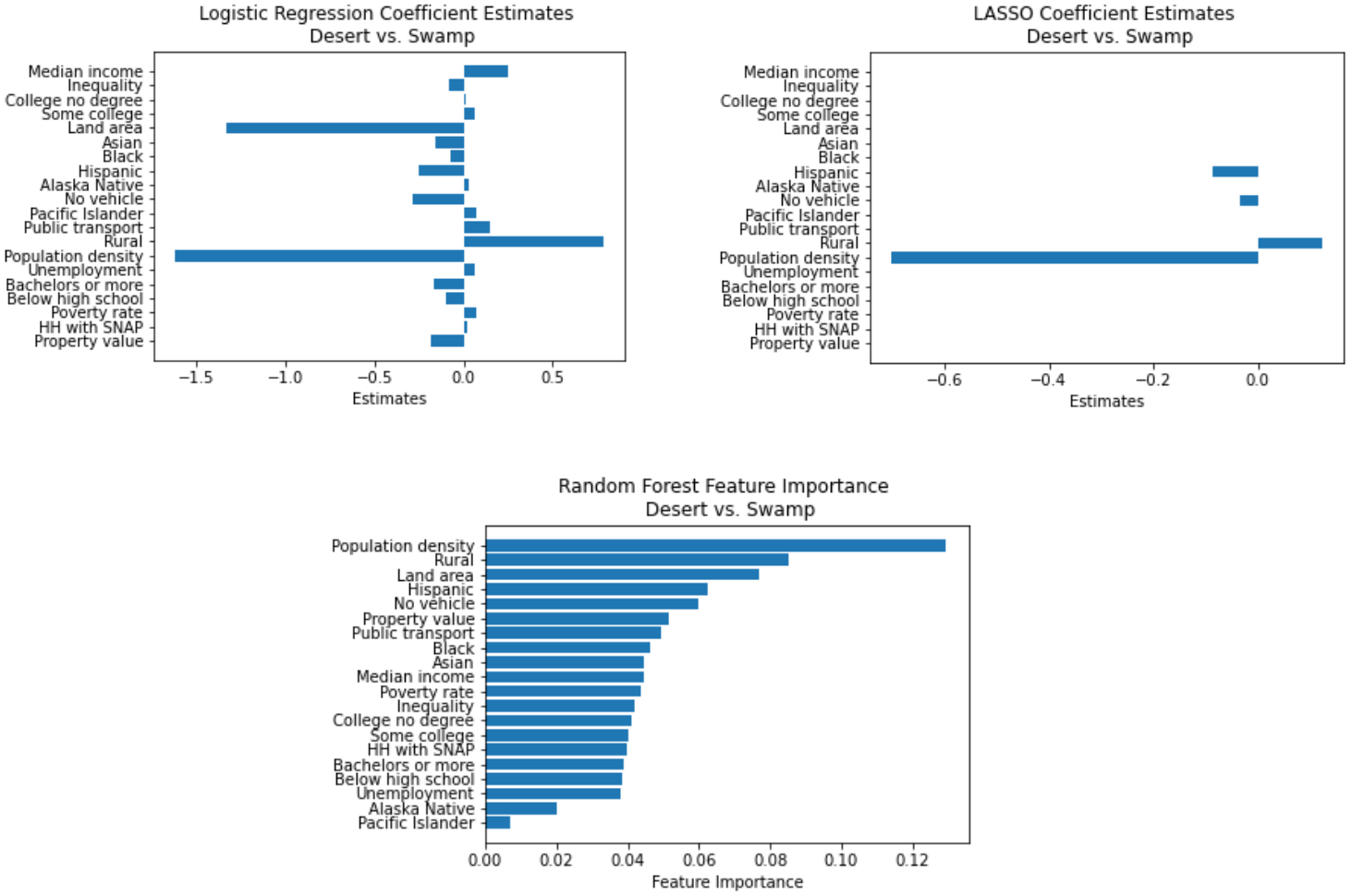



*Figure 9: LASSO and logistic regression coefficient estimates and feature importance from the random forest model for the Desert vs. Swamp outcome.*

## 5. Conclusions

Although the accuracy and recall scores from the models depicted above are not significantly high, they have produced some inferential insights that widen our understanding of food environments and food accessibility in the United States. In particular, the feature importance

from the random forest and the coefficient estimates from the LASSO and logistic regression have shown that food deserts and food swamps have different characteristics. The parametric models consistently found that food deserts were associated with tracts that have a high rural population and tracts where residents tend to rely on public transportation. In some instances, the parametric model also found that median income is positively correlated with deserts. Conversely, food swamps were positively correlated with population density, educational attainment (particularly having a bachelor's degree or more), the share of residents who are Black, and the share of residents who are Hispanic. Interestingly, food deserts were negatively correlated with the attributes that characterize food swamps. This confirms Amin et. al's hypothesis that these two food environments are different from one another. As such, we can tentatively conclude that food deserts tend to be located in rural census tracts where residents rely on public transportation. Food swamps tend to be located in densely populated tracts where residents are Black and Hispanic and have high educational attainment.

With regards to policy-implications, food deserts may benefit from community development programs that seek to encourage healthful food retailers to open branch locations in these areas. However, it is imperative that residents are able to afford the offering of these food retailers. Otherwise, the food desert may become a food swamp. Alternatively, policymakers could also identify if they can provide small-business loans to incentivize residents to open healthful food retail stores. Conversely, food swamps are slightly more difficult to solve. These are areas that have healthful food retailers but are outnumber by non-healthful food retailers. Policymakers can work with existing food retailers to make their offerings more affordable to residents. Additionally, policymakers could look into the business and development decisions that results in these census tracts having more unhealthy food retailers than healthful ones.

These conclusions are not definitive, but great considerations for future work in this area. Analysis like these could have better accuracy and better inferential results if the scope was limited to specific states/cities/census tracts. By limiting the scope, future work will be able to obtain data that is specific to these geographic areas and will perhaps provide a more accurate representation of the wider economic and social factors that are closely tied with food environments.

**Bibliography**

Amin, M.D., Badruddoza, S. & McCluskey, J.J. (2020). Predicting access to healthful food retailers with machine learning. *Food Policy*, 99, 101985. https://www.sciencedirect.com/science/article/pii/S0306919220301895?via%3Dihub

Center for Disease Control. (2011). Modified Food Retail Environment Index [Data file]. Retrieved from https://www.cdc.gov/obesity/downloads/2_16_mrfei_data_table.xls

Census for Diseases Control. (2011). *Census-tract level state maps of the modified food environment index (mREFI)*. https://www.cdc.gov/obesity/downloads/census-tract-level-state-maps-mrfei_TAG508.pdf

Guillaume, L., Nogueira, F., & Aridas, C.K. (2017). ADASYN. *imbalanced-learn.* https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.ADASYN.html#rf9172e970ca5-1

Imbalanced Learn. (n.d.). Compare over-sampling samplers. https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_comparison_over_sampling.html#more-advanced-over-sampling-using-adasyn-and-smote

Kassambara, A. (2018). Penalized Logistic Regression Essentials in R: Ridge, Lasso, and Elastic Net. *Statistical tools for high-throughput data analysis.*http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

Morland, K. & Evenson, K. (2009). Obesity prevalence and the local food environment. *Health & Place*, 15(2), 491-495. https://www.sciencedirect.com/science/article/pii/S1353829208000981?casa_token=da8xeGrMmjMAAAAA:jqfxYhjq2Rx_F3qwpy8ZItMopA-YvwCWmADwA9TXr2KOVD3BzsRxJpb_r5QvklPdPPV6C9PPMbI

Patel, K. (2019). Predicting Food Deserts Via Population Health and Twitter Sentiment Analysis. *Future Vision*. https://medium.com/future-vision/predicting-food-deserts-across-the-united-states-using-population-health-data-and-twitter-9920f4a0e4ea

Phillips, A.Z., & Rodriguez, H.P. (2019). Adults with diabetes residing in "food swamps" have higher hospitalization rates. *Health Services Research*, 54(1), 217-225. https://pubmed.ncbi.nlm.nih.gov/30613953/

Scikit-learn. (n.d.). sklearn.linear_model.LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

United States Census Bureau. (n.d.). American Community Survey (ACS). https://www.census.gov/programs-surveys/acs

United States Census Bureau. (2019). American Community Survey: Data Profiles. [Data file]. Retrieved from https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2019/

United States Census Bureau. (n.d.). When to Use 1-year, 3-year, or 5-year Estimates. https://www.census.gov/programs-surveys/acs/guidance/estimates.html

Williams, M. (2020, July 28). COVID-19 and the social determinants of health. *Harvard Medical School: Center for Primary Care*.http://info.primarycare.hms.harvard.edu/blog/covid-social-determinants-health

## Implementation Appendix

### Preprocessing

A few challenges were apparent in preprocessing the data. The characterization process had revealed some interesting factors. For one, a substantial number of attributes were heavily skewed. A few of them were highlighted in the 'Data' section of this report. To remedy this, log transformations were employed. While most of the skewedness was remedied, there were still a few variables that remained skewed. Ideally, it would be better to convert these continuous attributes into categorical attributes by creating cutoffs within the distribution or converting them to binary attributes. This would perhaps improve the accuracy of the models used, particularly for logistic regression by essentially eliminating the effects of outliers or skewedness. However, doing so would have implications for interpretability. For instance, converting land area (which was heavily skewed after log transformation) into a binary attribute would require determining what classifies as a small land area and what classifies as a large land area. Determining the appropriate cutoff that would apply for every census tract would be difficult. For these reasons, and to align with Amin et al's design process (which retained all variables in their continuous state), the skewed distributions were only log transformed.

### Oversampling technique

In creating the three mutually exclusive food environments, it was immediately clear that class imbalance would be an issue for the *Desert vs. Healthful* and *Swamp vs. Healthful* outcomes. As shown in Figure 3, the data set consists of more healthful tracts than swamps and deserts. To remedy this, Amin et al. used an adaptive synthetic sampling method, which oversamples the minority class in the training set, prior to fitting models (Amin, 2020). This technique is similar to the SMOTE technique but varies because it "focuses on the samples [of the minority class] which are difficult to classify." (Imbalanced Learn, n.d.).

The *Desert vs. Swamp* outcome was balanced; therefore, oversampling techniques were not applied when modeling this outcome.

The adaptive synthetic sampling method was implemented by using the ADASYN algorithm from the imbalanced learn package. (Guillaume, 2020).

### Hyper parameter tuning

Hyper parameter tuning was only conducted for the LASSO model. Logistic regression model does not have any hyper parameters that can be tuned. Random Forest was simply fit using the default. This was done to avoid imposing any restrictions on the random forest that would result in a decline in accuracy.

The LASSO model was tuned by constructing a validation curve for different values for the C parameter. This parameter is the inverse of lambda which determines the impact of the L1 shrinkage penalty. Increasing lambda results in more sparse models since coefficient estimates will be shrunk to zero. Since the C parameter is the inverse of lambda in the scikit-learn implementation, lower values of C are associated with more shrinkage of estimates (Scikit-learn, n.d.).

Train-test split

A learning curve was not constructed to determine the train-test split. Rather, the train-test split was determined based on Amin et. al's methodology which allocated 70% of the data to the training set and 30% of the data to the test set (Amin, 2020).