Predictive Modeling of an Unbalanced Binary Outcome in Food Insecurity Data

Jonathan Fabish[1], **Lauren Davis**[2], **and Seong-Tae Kim**[1]
[1]Department of Mathematics & Statistics, NC A&T State University, Greensboro, NC, U.S.A
[2]Department of Industrial & Systems Engineering, NC A&T State University, Greensboro, NC, U.S.A

Abstract— *Predictive modeling of a rare event using an unbalanced data set leads to poor prediction sensitivity. Although this obstacle is often accompanied by other analytical issues such as a large number of predictors and multicollinearity, little has been done to address these issues simultaneously. The objective of this study is to compare several predictive modeling techniques in this setting. The unbalanced data set is addressed using four resampling methods: undersampling, oversampling, hybrid sampling, and ROSE synthetic data generation. The large number of predictors is addressed using penalized regression methods and ensemble methods. The predictive models are evaluated in terms of sensitivity and F1 score via simulation studies and applied to the prediction of food deserts in North Carolina. Our results show that balancing the data via resampling methods leads to an improved prediction sensitivity for every classifier. The application analysis shows that resampling also leads to an increase in F1 score for every classifier while the simulated data showed that the F1 score tended to decrease slightly in most cases. Our findings may help improve classification performance for unbalanced rare event data in many other applications.*

Keywords: unbalanced data, predictive modeling, penalized regression, variable selection, resampling

1. Introduction

As defined by the USDA, food insecurity is a household-level economic and social condition of limited or uncertain access to adequate food. This is in contrast to hunger, which is an individual level physiological condition which results from food insecurity. According to the Food Research and Action Center in 2016, North Carolina had the 14^{th} highest food insecurity level among all states with 15.1% (603,094) households experiencing food insecurity [1]. Coined in Scotland in the early 1990s, the term “food desert" is used to describe communities which have limited access to affordable and nutritious foods [2]. The U.S. Census Bureau conducts the American Community Survey (ACS) [3] on a yearly basis to provide social, housing, economic, and demographic data. A census block is the smallest geographic area for which they collect and record population data. Census block groups, typically containing between 600 and 300 people, are one level above census blocks in terms of geographical area and are the smallest unit for which the Census Bureau collects and records sample data. They do not cross census tract, county, or state boundaries [4].

The motivation of this study stems from an unsuccessful attempt to build a model to predict the binary food desert status of U.S. census block groups in North Carolina. The objective is to build a predictive model for food insecurity in North Carolina using the predictors statistically selected. Unfortunately, due to a severe unbalance between the classes of the response variable, i.e., only 3.3% of observations are food deserts, prediction sensitivity was low and no trustworthy inferences could be made. It has been shown that resampling methods such as oversampling and undersampling are effective in improving prediction performance in such a situation [8-12].

Much of the recent research on resampling methods in the predictive modeling has involved using decision tree methods and support vector machines. However, the data sets used in these studies did not suffer from the large p problem encountered here. Our data has 2,780 covariates, i.e., there are 2^{2780} possible combinations of predictors. This renders traditional variable selection methods, such as forward and backward selection, too computationally intensive. Furthermore, many of the predictors in the food insecurity data set are known to be linear combinations of others. Thus, multicollinearity is an additional obstacle to overcome. Modern penalized regression methods perform simultaneous parameter estimation and variable selection in a setting with large p and multicollinearity.

We apply oversampling, undersampling, a hybrid of over- and undersampling, and Random Over-Sampling Examples (ROSE) [8] synthetic data generation to an unbalanced binary response variable. These resampling methods are applied to unbalanced data sets, and result in an equal distribution of observations from the majority and minority response classes. It has been shown that these sampling methods lead to improvement in classification sensitivity, while each method has drawbacks [7].

Using the original and resampled data sets we train four penalized logistic regression models, the least absolute shrinkage and selection operator (LASSO) [12], elastic net (ENET) [15], smoothly clipped absolute deviation (SCAD) [17], and minimax concave penalty (MCP) [18], and two ensemble classifiers, random forest [21] and boosting [22].

These classifiers have been successfully applied to big data sets. The penalized regression methods used here are highly interpretable since they shrink many regression coefficients to exactly zero. Ensemble methods tend to improve prediction accuracy, but lose interpretability by combining the results from many classifiers.

Correctly classifying the minority observations is the main purpose of our research, which makes the accuracy itself an unsuitable performance measure. Sensitivity, often called recall, measures the proportion of the minority observations which are correctly classified. Precision measures the proportion of positive predictions which are correct. The F1 score is the harmonic mean of precision and sensitivity. To achieve a high F1 score, a model needs a high precision and high sensitivity which makes it ideal for assessing a predictive model focused on correctly identifying observations from the minority class.

2. Statistical Methods

In this study, we apply novel combinations of well-known methods for dealing with a sparse parameter set and unbalanced binary response variable. Penalized regression methods such as LASSO, ENET, SCAD, and MCP as well as ensemble methods random forest and boosting trees are well known to handle classification problems involving a large number of predictors, p. Oversampling, undersampling, hybrid sampling methods, and synthetic data generation have been used successfully to overcome an unbalanced response variable. We apply combinations of these methods to real and simulated data exhibiting large p and an unbalanced response variable. In this chapter, we briefly summarize the theoretical background of the statistical methods selected.

2.1 Penalized Regression Methods

Let Y be the binary response vector, let $\boldsymbol{X}$ be a $n \times p$ predictor matrix, and let β be the $p \times 1$ vector of regression coefficients. The penalized log-likelihood function is given by

$$\hat{\beta}^{LOG} = \underset{\beta}{\operatorname{argmin}}[L(\beta) + P_{\lambda,\gamma}(\beta)] \tag{1}$$

where,

$$L(\beta) = \sum_{i=1}^{N}[\ln(1 + e^{\beta^T x_i}) - y_i \beta^T x_i] \tag{2}$$

and $P_{\lambda,\gamma}(\beta)$ is one of the penalty functions described in the following.

Penalized regression methods impose a unique penalty function on the coefficients as in equation (1). The methods considered here perform simultaneous variable selection and parameter estimation and are applicable for both continuous and discrete response variables, making them suitable for regression or classification problems with large p. For all the following penalized regression methods, we assume that the covariate matrix has been standardized by subtracting means and dividing by the standard deviation of each respective column.

LASSO. The LASSO, proposed by Tibshirani (1996) [12], imposes the L_1 penalty which is hyperrectangular in nature. The penalty of LASSO is given by

$$P_{\lambda,\gamma}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

which is called the L_1 penalty. The regularization parameter, λ, is data-driven and calculated via cross-validation for LASSO and each of the following penalties as well. The cyclical coordinate descent algorithm (CCDA) along a regularization path is used to compute solutions to the LASSO efficiently. The LASSO solution and other penalized regression solutions are typically computed via CCDA which solves a series of univariate optimization problems until some convergence criteria is met.

ENET. LASSO shrinks many coefficients to exactly zero, which is a highly interpretable form of variable selection, but can be unstable in a setting involving multicollinearity. Ridge regression, which imposes the L_2 penalty, is known to perform well in such a setting. The ENET penalty, proposed by Zou and Hastie (2003) [16], is a convex combination of the L_1 and L_2 penalties given by

$$P_{\lambda,\gamma}(\beta) = \lambda \sum_{j=1}^{p}[(1-\alpha)|\beta_j| + \alpha\beta_j^2]. \tag{4}$$

$0 \leq \alpha \leq 1$ controls the trade-off between the L_1 and L_2 penalties, with $\alpha = 0$ equivalent to the LASSO and $\alpha = 1$ equivalent to ridge regression. We set $\alpha = 0.5$ for the elastic net penalty which yields a strictly convex constraint region which still has non-differentiable corners, enabling it to perform variable selection while remaining more stable than LASSO among highly correlated predictors [13].

SCAD. LASSO and ENET are both biased estimators and employ convex penalties. SCAD and MCP are both continuous piecewise non-convex penalties which start equivalent to the LASSO penalty but weaken as the magnitude of β_j increases. SCAD was introduced by Fan and Li (2001) [17]. The SCAD penalty, defined on $[0, \infty)$, is given by

$$P_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta, & \text{if } \beta \leq \lambda, \\ \frac{\gamma\lambda\beta - 0.5(\beta^2+\lambda^2)}{\gamma-1}, & \text{if } \lambda < \beta \leq \gamma\lambda, \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if } \beta > \gamma\lambda. \end{cases} \tag{5}$$

γ is known as the threshold parameter and determines the point at which the penalty transitions to the subsequent piece of the function for SCAD, and MCP which follows.

MCP. The MCP, proposed by Zhang (2010) [18], was designed to approach to the unbiased estimates faster than

SCAD. The MCP penalty, defined on $[0, \infty)$, is given by

$$P_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2\gamma}, & \text{if } \beta \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \beta > \gamma\lambda. \end{cases} \quad (6)$$

Unlike LASSO or ENET, both SCAD and MCP are known to achieve the oracle property [19], that is, as $n \to \infty$ the model identifies the zero regression coefficients correctly with probability approaching 1 while remaining consistent for non-zero coefficients [20].

2.2 Ensemble Methods

Ensemble methods combine the results of multiple base statistical learning algorithms to construct an improved learning algorithm. In this study we apply two ensemble methods, random forest and boosting for classification.

Random Forest. Random forest employs a resampling method, closely related to bagging, which decreases variance of predictions. In bagging, we draw with replacement M random samples of size n from our training data set of n observations. A different model is trained and the statistic of interest is calculated based on each the M random samples. The final estimate is the mean or mode, for continuous or discrete random variables respectively, of the values computed across all M samples. Random forest differs from bagging in that rather than training on all p predictors at each node, a subset of size $\sqrt{p}$ is drawn at each split in a given tree. The result is a sequence of highly uncorrelated classification trees. Random forest is an effective method of decreasing variance and improving prediction accuracy but variable selection for individual predictors is not achieved. Rather, the out-of-the-bag-sample is used to assess variable importance in terms of mean decrease in accuracy upon permuting the values of each variable in succession, as compared to including all variables in the model [7] [13].

Boosting Trees. Boosting is another powerful resampling method. Unlike random forest, it produces decorrelated samples through an iterative weighting scheme. Boosting with classification trees consists of fitting a sequence of trees in which the first tree is fit to the response variable and each subsequent tree is fit to the residuals of the previous tree. Some benefits of boosting trees are the speed, insensitivity to the scale of the predictors, and relatively high accuracy. However, they have three hyperparameters to tune and are sensitive to overfitting the training data. Cross-validation can help to mitigate these issues [23] [7] [13].

2.3 Resampling Methods for Unbalanced Data

This study applies resampling methods, namely undersampling, oversampling, a hybrid of both, and ROSE synthetic data generation to overcome an unbalanced data set. All of these methods have been shown to improve prediction sensitivity, which is often a primary assessment measure for the type of classification problem considered here [7] [10]. For the purpose of illustrating each resampling methods in the following subsections, we consider a data set of $n = 1000$ observations for a binary response variable Y_i such that $y_i \in \{0, 1\}$ for $i = 1, \cdots, 1000$. Suppose also that $\sum_{i=1}^{1000} Y_i = 100$, so that the minority class label '1' makes up only 10% of the observations.

Oversampling. Oversampling balances the data set by randomly sampling with replacement, from the minority class, the same number observations which make up the majority class [8] and combining the observations from the resampled minority class and entire majority class into a single data set. In the example data set, oversampling would result in a new data set with 1800 observations of which 900 pertain to each class. Potential overfitting of the training data is a concern with this method.

Undersampling. Undersampling consists of randomly sampling without replacement, from the majority class, the same number of observations which make up the minority class and combining the observations from the sampled majority class and entire minority class into a balanced data set [8]. In the example, undersampling generates a data set with 200 observations of which 100 pertain to each class. A downside to this approach is that we eliminate a significant portion of our data set which likely contains useful information.

Hybrid Sampling. A combination of oversampling and undersampling which results in a data set with the same dimensions as the original. First, the minority class is oversampled sequentially until the number of observations from the minority class reaches some proportion p of the desired final sample size n (both of which are required arguments for the function in the ROSE package in R). Next, the majority class is undersampled to yield the balanced data set of size n [8]. In our example, we set $n = 1000$ and $p = 0.5$. The minority set is oversampled until there were 500 observations pertaining to class '1', and the majority set is undersampled to 500 observations, resulting in a data set with $n = 1000$ observations.

ROSE. The ROSE method balances the data set by modeling the joint density of a given observation. The process of data generation for our example data set is as follows. Let the class labels pertain to the set $G = \{0, 1\}$. Let n_g for $g = 0, 1$ be the number of observations pertaining to class g, i.e., $n_0 = 900$ and $n_1 = 100$.

1) For $i = 1, \cdots, 1000$:
    a) Randomly select $g \in G$ with probability 0.5.
    b) From the training set, randomly select with replacement an observation $(y_g, \boldsymbol{x}_g)$ from the n_g observations pertaining to the class g.
    c) Sample $\boldsymbol{x}^*$ from $K_{\boldsymbol{H}_g}(\cdot, \boldsymbol{x}_g)$, where $K_{\boldsymbol{H}_g}$ is the estimated probability distribution centered at $\boldsymbol{x}_g$ with covariance matrix $\boldsymbol{H}_g$.

Upon completing the for loop, the ROSE procedure has generated an independent data set of 1000 observations, each with an equal chance of coming from either class. It is

notable that the synthetic data set does not contain any of the original observations which leaves them available for model validation [8] [9].

2.4 Assessment Measures for Prediction Performance

On an unbalanced testing data set, a classifier that predicts that every observation comes from the majority class will appear to do well in terms of accuracy. This is a common scenario when a classifier is trained on an unbalanced data set. In many such binary classification tasks, correctly identifying observations from the minority class is the primary goal. Therefore, we need an alternative to mean accuracy to assess the quality of predictions made my our model.

Sensitivity, also known as recall, measures the proportion of positive cases correctly predicted. It is insensitive to the unbalanced class distribution. However, a classifier which predicts that every observation comes from the minority class will achieve a high sensitivity,

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}. \quad (7)$$

Precision, also known as positive predictive value, measures the proportion of positive predictions that were correct. This evaluates the validity of attaining a high sensitivity,

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}. \quad (8)$$

The F1 score is the harmonic mean of precision and sensitivity,

$$F1\ Score = \left(\frac{Precision^{-1} + Sensitivity^{-1}}{2}\right)^{-1}. \quad (9)$$

A high sensitivity at the expense of a low precision or a high precision at the expense of a low sensitivity translates to a low F1 score, making it an ideal assessment measure for rare event classification. All of these measures can easily be obtained from the confusion matrix.

3. Simulation Studies

We applied four sampling methods to real and simulated data in order to balance the classes of the response variable prior to prediction using penalized and ensemble classification methods. For comparison, the six classification algorithms selected were also applied on a simulated data set that was balanced from the offset. The mean prediction sensitivity and F1 score computed over 300 iterations are reported to assess the effectiveness of the various classification algorithm/sampling method combinations.

3.1 Design of Simulations

Each iteration generates independent testing and training data using the mvrnorm function in the MASS package in R. We fixed $p = 1000$ and $n = 4000$, and considered two covariance structures, independent and first order autoregressive (AR(1)) with $\rho = 0.5$ and $\sigma^2 = 1$.

In the AR(1) covariance structure, correlation is highest among adjacent predictors and decreases exponentially with distance. The true parameter space is given by $\beta = [3, 3, 0, 3, 2, 0, 2, 0, 0, \cdots, 0]$ where only 5 of the coefficients are non-zero and the remaining parameters consist of zero coefficients. To generate the simulated response variable, the intercept value in the logistic regression was adjusted such that the minority class comprises approximately 10% of all observations.

Next, using the ROSE [8] package in R, oversampling, undersampling, hybrid sampling, and ROSE were applied to the unbalanced training data set. Finally, the six classification algorithms were trained using the now balanced training data sets, as well as the unbalanced training set for reference. Predictions were then made by all models on the testing data. The mean sensitivity and F1 score and their respective standard deviations were calculated over all iterations.

3.2 Results

The results for the Monte Carlo simulations are presented in Tables 1 and 2. In each cell, the top number represents the mean and the bottom number in parenthesis represents the standard deviation.

Table 1 Unbalanced Data: Sensitivity and F1 Score for n = 4000, Independent Covariance

Method	Unbalanced		Under		Over		Hybrid		Rose	
	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1
Lasso	0.41	0.55	0.95	0.39	0.58	0.59	0.70	0.58	0.47	0.37
	(0.08)	(0.07)	(0.03)	(0.05)	(0.08)	(0.04)	(0.07)	(0.04)	(0.08)	(0.05)
ENET	0.20	0.32	0.93	0.34	0.34	0.40	0.46	0.43	0.40	0.329
	(0.07)	(0.09)	(0.04)	(0.05)	(0.07)	(0.06)	(0.08)	(0.05)	(0.08)	(0.05)
SCAD	0.67	0.729	0.95	0.45	0.81	0.639	0.85	0.61	0.72	0.46
	(0.065)	(0.04)	(0.030)	(0.06)	(0.08)	(0.05)	(0.07)	(0.05)	(0.08)	(0.04)
MCP	0.68	0.73	0.96	0.46	0.78	0.64	0.83	0.61	0.73	0.45
	(0.07)	(0.04)	(0.03)	(0.06)	(0.08)	(0.04)	(0.07)	(0.05)	(0.07)	(0.04)
RF	0.00	-	0.84	0.25	0.00	-	0.00	-	0.78	0.39
	(0.00)		(0.06)	(0.03)	(0.00)		(0.00)		(0.08)	(0.06)
Boost	0.06	-	0.89	0.31	0.89	0.40	0.88	0.40	0.94	0.34
	(0.03)		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)

Table 1 represents the simulation results for the six predictive models applied to an initially unbalanced data set with $n = 4000$ and independent covariance, before and after applying four resampling methods. MCP scored highest on the unbalanced data and undersampled data. For the other three resampling methods, boosting attained the highest sensitivity. SCAD, tying with MCP, attained the highest F1 score. There is a pattern of inferior performance of all models on the ROSE data set. Sensitivity improved for all models for all resampling methods. F1 score tended to decrease for penalized regression methods but improved for ensemble methods.

Table 2 Unbalanced Data: Sensitivity and F1 Score for n = 4000, AR(1) Covariance

Method	Unbalanced		Under		Over		Hybrid		Rose	
	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1
Lasso	0.72 (0.04)	0.79 (0.03)	0.96 (0.02)	0.66 (0.03)	0.79 (0.04)	0.76 (0.03)	0.86 (0.03)	0.74 (0.03)	0.86 (0.04)	0.65 (0.03)
ENET	0.63 (0.05)	0.74 (0.03)	0.96 (0.02)	0.62 (0.03)	0.70 (0.04)	0.69 (0.03)	0.79 (0.04)	0.67 (0.03)	0.83 (0.04)	0.63 (0.03)
SCAD	0.81 (0.04)	0.84 (0.02)	0.96 (0.02)	0.71 (0.04)	0.88 (0.03)	0.79 (0.03)	0.90 (0.03)	0.77 (0.03)	0.93 (0.03)	0.67 (0.03)
MCP	0.81 (0.04)	0.84 (0.02)	0.96 (0.02)	0.72 (0.04)	0.88 (0.04)	0.78 (0.03)	0.90 (0.03)	0.77 (0.03)	0.94 (0.03)	0.67 (0.03)
RF	0.03 (0.02)	-	0.95 (0.02)	0.57 (0.03)	0.01 (0.01)	-	0.19 (0.06)	-	0.96 (0.02)	0.55 (0.04)
Boost	0.41 (0.04)	0.56 (0.04)	0.95 (0.02)	0.60 (0.03)	0.94 (0.02)	0.63 (0.03)	0.93 (0.02)	0.64 (0.03)	0.97 (0.02)	0.57 (0.03)

Table 2 represents the simulation results of six predictive models applied to an initially unbalanced data set with $n = 4000$ and AR(1) covariance, before and after applying four resampling methods. SCAD and MCP tied for best performance on the unbalanced data set. All methods attained the peak sensitivity while MCP attained the highest F1 score, on the undersampled data set. The non-convex penalized and ensemble methods performed similarly on the ROSE sampled data set. The distribution of top scores for the other sampling methods is identical to that of Table 1. No combination of resampling and classifier led to an improvement over the F1 score attained by SCAD and MCP on the unbalanced data set.. All models attained higher scores on the ROSE data set with AR(1) covariance than on the ROSE dataset with independent covariance. Once again, sensitivity improved for all models across all resampling methods. F1 score tended to decrease for penalized regression methods across all resampling methods and improve slightly for ensemble methods.

4. Empirical Study

4.1 Data Source

An ideal data set for the study at hand would include a binary response variable representing the food desert status of each block group in North Carolina and a set of mutually uncorrelated predictors which are associated with food desert status. A fixed time data set for the current analysis was compiled from the following sources.

PolicyMap. The binary response variable for this study, limited supermarket access (LSA) status by block group in NC, was attained from Reinvestment Fund (2016) via PolicyMap [6]. LSA status is a measure of whether a block group is well-served by a supermarket or experiences limited access. This aligns nicely with the definition of a food desert, i.e., an area in which residents have limited access to an affordable and healthy diet. For this reason, we chose LSA status as a surrogate variable for food desert status. The data for supermarket location was acquired by Reinvestment Fund from the 2017 Nielsen TDLinx database [24] and includes supermarkets, supercenters, limited assortment stores, and natural food stores, but excludes superettes and dollar stores due to their lack of healthy food options. To account for variability in urban and rural areas, population density of various block groups was considered upon assignment of LSA status. This is achieved by considering how far the distance to the nearest supermarket would need to be reduced to equal the distance of a well-served block group of the same population density class. The LSA status was eventually encoded as a dummy variable having the value of 1 for positive LSA status and 0 for negative LSA status [6], which we refer to here as positive and negative food desert status, respectively.

US Census Bureau. The predictor variables used in the analysis were published by the US Census Bureau in the 2016 American Community Survey [3] and acquired from American Fact Finder [5]. The matrix of covariates was obtained from American Fact Finder [5], a public database on U.S. Census Bureau data, and consists of 2,780 variables representing 5-year estimates of various social, housing, economic, and demographic data from the 2016 ACS. The response variable, limited supermarket access (LSA), which we are using as a surrogate for food desert status, was obtained from the Reinvestment Fund through PolicyMap [6]. Each random variable is discrete numeric and represents the count of a particular characteristic with respect to the given block group.

Data Wrangling. The data from Reinvestment Fund and the US Census Bureau were cleaned using the R statistical programming language, via the tidyverse package. The data were combined using census block group number as a common key. Prior to cleaning, the data set had 6066 observations, each representing a different block group, and 2835 predictor variables associated with each observation. Any predictor that was missing data for 10 or more observations was excluded and then all the remaining incomplete observations were removed. After cleaning, the data set had final dimensions 6062 x 2780 (45769 KB).

4.2 Analysis and Results

Using 5-fold cross-validation, the classification algorithms were applied before and after balancing the data set using the four resampling methods. On each fold, 80% of the food insecurity data set was allocated to the training set. Using the ROSE package for R [8], oversampling, undersampling, hybrid sampling, and ROSE were each applied to the training portion of the data set. Then, the balanced training data were used to train the six selected classification algorithms. LASSO and ENET were applied using the glmnet function from the glmnet [15] package with $\alpha = 1$ for LASSO and $\alpha = 0.5$ for ENET. SCAD and MCP were applied using the ncvreg [19] package. Random forest was applied using the randomForest [21] package. Gradient boosting was applied using the gbm [22] package with $interaction.depth = 1$. 500 trees were grown for both random forest and gradient boosting. Predictions were then made on the testing portion of the data set. The mean prediction sensitivity and F1 score were calculated for each fold and averaged. The results are presented in Table 3.

Since LASSO, SCAD, and MCP all performed well on the undersampled data set, we chose to examine which

predictor variables were selected as significant by these models in this setting. We took the intersection of the significant predictors across all 5-folds of cross-validation. Ideally, knowing which predictors are associated with positive food desert status could provide information to help address the issue of food insecurity throughout the region being studied, North Carolina in this case. The intersection of significant predictors contained about 20 predictors including: 'Geographical Mobility In The Past Year For Current Residence–Micropolitan Statistical Area Level – Moved from principal city', 'Employment Status For The Population $\geq$ 16 Years – Armed Forces', 'Sex By Industry For The Civilian Employed Population $\geq$ 16 Years – Male manufacturing', 'Sex By Industry For The Civilian Employed Population $\geq$ 16 Years – Female Management of companies and enterprises', 'Age By Language Spoken At Home By Ability To Speak English For The Population $\geq$ 5 Years – Speak other Indo-European languages and speak English well', 'YEAR STRUCTURE BUILT – build 1939 or earlier', 'House Heating Fuel – wood', 'Value – 60,000- 69,999$', 'Value – 125,000 – 149999$', 'Value – Not computed', 'Occupancy Status - Vacant', 'Commuting Characteristics By Sex – female; worked outside state of residence', 'Means Of Transportation To Work By Travel Time To Work – Public transport 20 – 24 minutes', 'Household Type By Household Size – non-family households 4-people', 'Race Of Householder - Householder who is Some other race alone', 'Mortgage Status By Selected Monthly Owner Costs As A Percentage Of Household Income In The Past 12 Months – Not computed', 'Rent Asked - 550-599$', 'Bedrooms By Gross Rent – $\geq$ 3 bedrooms; cash rent; less than $300', 'Mortgage Status And Selected Monthly Owner Costs – housing units without a mortgage; 1200-1299$', and 'Sex By Industry For The Civilian Employed Population 16 Years And Over - Information'. Some of these variables seem to be conflicting, such as 'Value – 125,000 – 149999$' and 'Value – 60,000- 69,999$'. It could be informative to separate food deserts based on population density and run an individual analysis for each group since it is likely that urban and rural food deserts have different profiles with respect to these predictors. This may account for the apparent conflict among the covariates selected here.

5. Discussion

The resampling methods were vital to the results of each model in the food insecurity data. Their application led to an increase in sensitivity and F1 score for essentially every model. No usable results were obtained by any model on the unbalanced data set. The ensemble methods failed to obtain usable results for every combination of resampling. Among all the resampling methods, undersampling led to the largest improvement in terms of sensitivity while the other resampling methods improved the F1 score slightly more. SCAD and MCP produced consistent results for every resampling method.

Table 3 Food Insecurity Data: Sensitivity and F1 Score

Method	Unbalanced		Under		Over		Hybrid		Rose	
	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1	Sens.	F1
Lasso	1.00 (0.00)	0.06 (0.00)	0.71 (0.08)	0.13 (0.02)	0.26 (0.03)	0.20 (0.03)	0.30 (0.07)	0.17 (0.04)	0.47 (0.11)	0.18 (0.04)
ENET	0.00 (0.00)	-	0.72 (0.07)	0.13 (0.02)	0.21 (0.03)	0.17 (0.03)	0.30 (0.04)	0.18 (0.03)	0.46 (0.11)	0.18 (0.04)
SCAD	0.07 (0.04)	0.11 (0.06)	0.69 (0.07)	0.12 (0.02)	0.51 (0.10)	0.16 (0.02)	0.48 (0.06)	0.15 (0.02)	0.53 (0.11)	0.17 (0.03)
MCP	0.06 (0.03)	0.10 (0.05)	0.70 (0.06)	0.12 (0.02)	0.54 (0.08)	0.16 (0.02)	0.57 (0.06)	0.16 (0.02)	0.53 (0.08)	0.17 (0.02)
RF	1.00 (0.00)	0.06 (0.00)	1.00 (0.00)	0.05 (0.00)	1.00 (0.00)	0.06 (0.01)	1.00 (0.00)	0.06 (0.01)	1.00 (0.00)	0.03 (0.00)
Boost	1.00 (0.00)	0.06 (0.00)	1.00 (0.00)	0.05 (0.00)	1.00 (0.00)	0.06 (0.00)	1.00 (0.00)	0.06 (0.00)	1.00 (0.00)	0.06 (0.01)

Table 3 represents the simulation results of the six predictive models using unbalanced and balanced food insecurity data sets calculated using 5-fold cross-validation. In each field, the top number represents the mean and the number in parenthesis is the standard deviation. None of the models performed well on the unbalanced data set. The sensitivity and F1 score improved for every model relative to the results from the unbalanced data, except for random forest and boosting on the ROSE data which, for reasons we could not identify, both classified all test observations as class '1'. LASSO achieved the highest sensitivity of 0.929 on the undersampled data set, but performed poorly on the oversampled and hybrid sampled data sets. SCAD and MCP performed well on every resampled data set. Boosting had the highest F1 score on four of the data sets, ranging from 0.261 to 0.324.

With respect to the simulated data with independent covariance, the only resampling method for which every classifier performed well was again the undersampling method. The highest sensitivity was obtained by penalized regression methods applied to the undersampled data set. No model was able to improve over the F1 score attained by SCAD or MCP on the unbalanced data set. Additionally, the sensitivity of SCAD and MCP on the unbalanced data set was superior to that of many of the other models on the balanced data sets. If we knew that the F1 score was a more relevant measure than sensitivity, we would prefer to use SCAD or MCP on the unbalanced data set, rather than another model with a resampling method. SCAD and MCP attained a high sensitivity for every resampling method.

Considering the simulated data with AR(1) covariance, similar to the simulations with independent covariance, no method improved with respect to the F1 score of SCAD or MCP on the unbalanced data. Only if we knew that prediction sensitivity were a more relevant measure than F1 score, would we choose to apply resampling methods. With this covariance structure, ROSE led to prediction results that were competitive with those from undersampling. This is an important fact because sometimes the severity of the unbalanced data can make undersampling impractical. Boosting with the ROSE data set achieved the highest sensitivity. That being said, SCAD and MCP performed consistently across all resampled data sets again and should be considered attractive options in this and the previous settings.

Overall, the combination of penalized regression methods with resampling enabled us to improve the sensitivity and F1 scores such that we could identify predictors associated with positive food desert status in NC. This is promising in that it can help us to gain insight into the social, economic, demographic, and housing data profiles of food deserts in NC or elsewhere in the world. This certainly provides motivation to expand the food desert study to include larger regions. Additionally, this could help us to develop creative ideas to address societal problems such as food insecurity or other similar issues involving rare events.

6. Conclusion

For every simulation except undersampling with penalized regression led to the highest sensitivity. With AR(1) covariance, ROSE with boosting outperformed undersampling for each classifier with respect to sensitivity, but only by a slim margin. In every case, random forest performed poorly on unbalanced data, oversampled data, and hybrid sampled data sets. For random forest, the introduction of duplicate observations due to sampling with replacement crippled the performance. Boosting outperformed SCAD and MCP by a small margin in a few cases but was also unstable when applied to the ROSE food insecurity data. SCAD and MCP were consistent performers across every data set, in particular when undersampling was performed. The results of undersampling are promising but as the number of minority observations m decreases this method becomes less useful since the final number of observations in the resampled data set is $2m$. Interestingly, ROSE tended to perform better on data with AR(1) covariance than on data with independent covariance, making it a highly attractive option when multicollinearity is present.

There are many facets of this investigation that warrant future investigation. One can consider new predictive modeling techniques which can reduce false positive results. Also, there are historical time series data from the ACS that could be incorporated into this work, and the LSA status is available for all of the U.S. on PolicyMap. This would enable a more systematic investigation of the parameters predicted to be associated with food desert status to gain insight into the socioeconomic and demographic profiles of food deserts in various regions of the U.S. The food insecurity data had only 3% minority observations but our simulated data approximately contained 10% minority observations due to the instability in the logistic regression. It is worthwhile to address this issue as well.

Acknowledgment

This project was supported by NSF National Research Traineeship Project: Improving Strategies for Hunger Relief and Food Security using Computational Data Science (Award No. DGE-1735258) and ACE Implementation Project: Data Science and Analytics Advancing STEM Education at NC Carolina A&T State University (Award No. HRD-1719498).

References

[1] Profile of Hunger, Poverty, and Federal Nutrition Programs, Food Research & Action Center. (2017). Available: http://www.frac.org/wp-content/uploads/sos-nc.pdf

[2] National Research Council. *The Public Health Effects of Food Deserts: Workshop Summary,* Washington, DC: National Academies Press, 2009.

[3] American Community Survey, US Census Bureau. (2018). Available: https://www.census.gov/programs-surveys/acs/about.html

[4] Census Blocks and Block Groups, US Census Bureau. Available: https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf

[5] American Fact Finder, US Census Bureau. Available: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

[6] Reinvestment Fund 2016. Limited Supermarket Access (LSA). (2016). PolicyMap. Available: https://www.policymap.com/

[7] B.W. Yap, K.A. Rani, H.A.A. Rahman, S. Fong, Z. Khairudin, and N.N. Abdullah, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets," *DaEng-2013,* Singapore: Springer, 2013.

[8] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: A Package for Binary Imbalanced Learning," *R Journal,* 6(1):82-92, 2014.

[9] G. Menardi, and N. Torelli, "Training and assessing classification rules with unbalanced data," *Data Mining and Knowledge Discovery,* 28(1):92-122, 2012.

[10] G.M. Weiss, K. McCarthy, and B. Zabar, "Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?", *DMIN,* 7:35-41, 2007.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review." *GESTS International Transactions on Computer Science and Engineering.* 30:25-36, 2006.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological),* 58(1):267-288, 1996.

[13] T. Hastie, R. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* New York: Springer, 2009.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R,* New York: Springer, 2013

[15] J. Friedman, T. Hastie, R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software,* 33(1):1-22, 2010.

[16] H. Zou, and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology,* 67(2):301-320, 2005.

[17] J. Fan, and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association,* 96(458):1348–1361, 2001

[18] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics,* 38(2):894–942, 2010.

[19] P. Breheny, and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics,* 5(1):232-253, 2011.

[20] H. Leeb, and B.M. Poetscher, "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," Cowles Foundation Discussion Papers 1500, Cowles Foundation for Research in Economics, Yale University, 2005 (revised Apr 2007).

[21] A. Liaw, and M. Wiener, "Classification and Regression by randomForest," *R News,* 2(3):18-22, 2002.

[22] B. Greenwell, B. Boehmke, B., J. Cunningham, J., and GBM Developers. "gbm: Generalized Boosted Regression Models," R package version 2.1.4.

[23] K. Woodruff, "Introduction to Boosted Decision Trees," Machine Learning Group Meeting, 2017.

[24] Nielson:TDLinx. (2018). Available: https://catalog.data.gov/dataset/nielsen-tdlinx