Predicting Low Income Low Access Tracts in US using Logistic Regression, LDA and RDA

Rohan Shinde*

Abstract

Food insecurity is an important policy challenge in USA. In 2008, the US Farm bill requested the USDA to measure the extent of "Food Deserts" and find out its consequences and causes. In 2013, the ERS switched the term "Food Deserts" with "Low Income and Low Access" to reflect the reality more clearly. A logistic regression model is developed for predicting the Low Income Low access designation of a tract using socio-economic factors. The results show that a relaxation in the eligibility for SNAP benefits is much more effective at decreasing food insecurity than a UBI scheme. While the submission only talks about simple changes in these factors, the model can be applied to a more complex set of scenarios.

1 Introduction

United States of America is the third most populous country in the world, only behind India and China. Around 10.5% of the households were food insecure in the year 2020, unchanged from the year 2019. This translates to 38.3 million persons being food insecure.[1]

When the households with children are considered, the condition becomes much worse. Around 14.8% of households with children experienced food insecurity. This translates to 6.1 million children experiencing food insecurity.

These facts raise an important question: What constitutes Food Security?

Food security is characterised by the availability, accessibility and affordability of food at all times. Limited access to supermarkets, supercenters, grocery stores, or other sources of healthy and affordable food may make it harder for some people to eat a healthy diet.

- Availability means that enough food is available to all people, which includes domestic production, stocks from previous years and net imports
- Accessibility means that food is within geographical reach of every person.

*MStat (NB Stream) 1st Year, SMU, ISI Delhi

[1]https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/key-statistics-graphics/

- Affordability means that each person has the necessary funds to get adequate, safe and nutritious food.

Based on data from the ERS, food availability is not an issue. This mirrors the global scenarios where even though there is adequate production of food, too much is wasted and there are pockets of extreme food insecurity and even acute hunger.The availability and affordability dimensions of food security warrant further probing.

Availability is further complicated by the disparate patterns of population density among the states and even within a state. This is because of the Urban-Rural divide. Because of high population density in Urban areas, a single store can serve many more citizens than a similar store in a rural area. Thus, business-wise, it makes much more sense to have stores in an urban area, compared to a rural one. Thus, a large section of the rural population is bound to have lower access. Even in Urban areas, the issue is far from being simple. It is not necessary that the stores near a particular neighbourhood sell healthy foods. People are forced to choose between consuming enough calories and paying other expenses, or having adequate nutritious food and lacking funds for other expenses.

Affordability is a major factor in the obesity epidemic that is going on in the country. Because people are forced to choose between having enough food on the table or having an inadequate amount of healthy food. The food that is cheaper generally is more processed, has more salt and has more calories. Thus, even though nutritious food might be accessible, it does not mean that it is affordable. This ties food insecurity with the issue of poverty.

A paradigm shift has occurred wherein policy makers are switching from a targeted manner of delivering food subsidies to a Universal Basic Income scheme. A UBI scheme is a financial scheme wherein the citizens receive a legally-fixed and equal amount of money without any means testing or restriction on how it can be spent. We use a binary logistic regression model to show that increasing eligibility for SNAP benefits is much more effective.[2] [3]

In summary, food insecurity is a multi dimensional problem. Even though accessibility is not a problem, availability and affordability present challenges that must be solved to address the issue. We present evidence to show that UBI is not the most efficient way to go about it.

2 Objective

The objective of this project is to make a logistic regression model and make recommendations using it. The main comparison we make is between the efficiency of a proposed solution, UBI and the present method, SNAP. A secondary objective is to use the regression

[2]https://news.yahoo.com/los-angeles-launching-us-biggest-130450934.html

[3]https://www.leoweekly.com/2021/11/universal-basic-income-like-pre-filed-bill-would-provide-1k-to-some-kentuckians/

model as an analytical framework, and evaluate the main factors that lead to a tract becoming Low Income and Low Access.

3 Data

There are many ways to measure food store access for individuals and for neighborhoods, and many ways to define which areas are low-income and low access—neighborhoods that lack healthy food sources. Most measures and definitions consider at least some of the following indicators of access:

- Accessibility to sources of healthy food, as measured by distance to a store or by the number of stores in an area;
- Individual-level resources that may affect accessibility, such as family income or vehicle availability; and
- Neighborhood-level indicators of resources, such as the average income of the neighborhood and the availability of public transportation.

In the Food Access Research Atlas, several indicators are available to measure food access along these dimensions. For example, users can choose alternative distance markers to measure low access in a neighborhood, such as the number and share of people more than one-half mile to a supermarket or 1 mile to a supermarket. Users can also view other census-tract-level characteristics that provide context on food access in neighborhoods, such as whether the tract has a high percentage of households far from supermarkets and without vehicles, individuals with low income, or people residing in group quarters

The model mentioned in the previous section needs two main data types. The first being the Low Income Low Access designation and the second being a host of socio-economic predictors that go into the regression model. Both of these were sourced from the original data of Food Access Research Atlas available at website of USDA ERS.

In 2015, USA has 50 states and District of Columbia, which will be expressed as 51 states henceforth. District of Columbia is a city-state and is the capital of the country.

In the Food Access Research Atlas, several indicators are available to measure food access along these dimensions. For example, users can choose alternative distance markers to measure low access in a neighborhood, such as the number and share of people more than one-half mile to a supermarket or 1 mile to a supermarket. Users can also view other census-tract-level characteristics that provide context on food access in neighborhoods, such as whether the tract has a high percentage of households far from supermarkets and without vehicles, individuals with low income, or people residing in group quarters. One significant hurdle for the analysis was the lack of data for many variables, for example, the Low Income and Low Access population count was missing for a significant fraction of the tracts. Preliminary analysis suggested that we should fill these in with 0, but we have

not done that. The main reason behind this is that the Food Access Research Atlas, made available by USDA did not replace the missing data with zeroes. Similarly, the data on racial composition was also missing for a large number of tracts. Considering this lack of data, we used what was available and made the best use of it.

The original data has about 72531 observations and 147 variables. The variables include a lot of accessibility indicators for 3 different distance measures from nearest supermarket/superstore- half mile, 1 mile and 10 miles. There are also other Census tract-specific variables like Population, Median Family Income, Poverty Rate, Number of Group Quarters, etc. included in the data. About 62% of the data is missing for accessibility indicators for 10 mile distance measure and about 35% of it is missing for 1 mile distance measure. Such high missing values cannot be just removed from the data to generalise well nor can they be imputed since some counties have missing value for some variables in all it's census tracts. As a result, we only predict the Low Income Low Access tracts using half mile measure in urban and 10 mile measure in rural tracts. (The *LILATracts_halfAnd10* response denotes if a specific tract is a Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than one-half mile from the nearest supermarket, supercenter, or large grocery store for an urban area or greater than 10 miles for a rural area)

The models in this report are aimed towards classifying, predicting and understanding the effects of socioeconomic explanatory variables on the Low Income and Low Access designation of a tract. However, it is also worth understanding how availability and accessibility varies across the states. For accessibility, we expressed the total number of Low access persons at 1/2 mile for urban areas and 10 miles for rural areas as a proportion of the total population of the state. For affordability, we used the average poverty rate.

4 Methods

4.1 Feature Engineering

4.1.1 Elastic Net Regularization

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. In high dimensional data (small number of observations n as compared to the number of variables p), the LASSO selects at most n variables before it saturates. Also if there is multicollinearity in the data, then the LASSO tends to select one variable from a group of correlated variables and ignore the others. To overcome these limitations, the elastic net adds a quadratic part ($\|\beta\|^2$) to the penalty, which when used alone is ridge regression. The estimates from the elastic net method are defined by

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1)$$

The quadratic penalty term makes the loss function strongly convex, and it therefore has a unique minimum. The Elastic Net Regularization has been used for Logistic Regression models in this project.

4.1.2 Principal Component Analysis

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. PCA is a very flexible tool and allows analysis of datasets that may contain, for example, multicollinearity, missing values, categorical data, and imprecise measurements.

The goal is to extract the important information from the data and to express this information as a set of summary indices called principal components. This means that 'preserving as much variability as possible' translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Finding such new variables, the principal components (PCs), reduces to solving an eigenvalue/eigenvector problem.

We can find the principal components mathematically by finding the projections of the data which maximize the variance. The first principal component is the direction in space (the space of predictor data points) along which projection have the largest variance. The second principal component is the direction which maximizes variance among all directions orthogonal to the first. The kth component is the variance-maximizing direction orthogonal to the previous $k1$ components.Continuing so on, there are in total, p principal components. The complete set of principal components can be worked out by solving the eigenvalue problem. [4]

4.2 Models

4.2.1 Logistic Regression Model

In statistics, the logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination)[5].

Logistic regression is a type of a Generalised Linear Model used when we want to model the probability of a binary outcome. It can also be "Multinomial" when we care predicting

[4]https://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch18.pdf

[5]https://en.wikipedia.org/wiki/Logistic$_regression$

a categorical variable with multiple possibilities. The main reason for using this is that in case of binary outcomes, the assumption of heteroskedasticity is violated. Because the definitions of the flags in the data were decided by the USDA ERS, we decided to use the unflagged socioeconomic indicators in our analysis.

The basic idea of Logistic regression is that we assume a linear relationship between the log odds, i.e, $\log\frac{p}{1-p}$ and the predictor variables. It is a parametric form for the distribution $P(Y|X)$ where Y is a discrete value and $X = \{x_1, x_2, \cdots, x_n\}$ is a vector containing discrete or continuous values. The parametric model of Logistic Regression can be written as

$$P(Y = 1|X) = \frac{1}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}$$

and

$$P(Y = 0|X) = \frac{\exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}$$

The parameter $W = \{w_0, w_1, \cdots, w_n\}$ of the Logistic Regression is chosen by maximizing the conditional data likelihood. It is the probability of the observed Y values in the training data. The constraint can be written as

$$W \leftarrow \underset{W}{\operatorname{argmin}} \sum_{l} \ln(P(Y^l \,|\, X^l, W)$$

The maximum-likelihood method is computationally intensive and, although it can be performed in desktop spreadsheet software, it is best suited for statistical software packages. The output of logistical regression is reported in terms of odds ratios, which is the numerical odds (bounded by 0 and infinity) of the binary, dependent variable being true, given a one-unit increase in the independent variable.

Logistic regression can be modified to handle categorical explanatory variables through definition of dummy variables, but this becomes impractical if there are many categories. Similarly, one can extend the approach to cases in which the response variable is polytomous (i.e., takes more than two categorical values). Also, logistic regression can incorporate product interactions by defining new explanatory variables from the original set, but this, too, becomes impractical if there are many potential interactions.

4.2.2 Linear Discriminant Analysis

LDA is used as a tool for classification, dimension reduction, and data visualization. It has been around for quite some time now. Despite its simplicity, LDA often produces robust, decent, and interpretable classification results. Linear discriminant analysis (LDA) finds a linear combination of features that separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid over-fitting ("curse of dimensionality") and also reduce computational costs. In particular an LDA model projects a feature space (a dataset of n-dimensional samples) onto a smaller subspace of dimension k (where $k \leq n-1$) while maintaining the class-discriminatory information.

The classical LDA approach assumes that the data in each binary class comes from a multivariate Gaussian distribution (possibly with different mean vectors but with same covariance matrix Σ). The linear decision boundary can then be obtained by maximising the log likelihood of the conditional class probability (given the data) using Bayes' rule (assuming some prior marginal distributions of the classes).

The process of LDA using Fisher Discriminant method can be summarized as follows[6]:

- We first compute the d-dimensional mean vectors $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathrm{D}_i}^{n} \mathbf{x}_k$ for the different classes from the dataset.
- We then compute the scatter matrices (in-between-class and within-class scatter matrix). The within-class scatter matrix $\mathrm{S_W}$ is computed by the following equation:

$$\mathrm{S_W} = \sum_{i=1}^{c} \mathrm{S}_i$$

  where $\mathrm{S}_i = \sum_{\mathbf{X} \in \mathrm{D}_i}^{n} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^{\mathrm{T}}$ (scatter matrix for every class) and $\mathbf{m}_i$ is the mean vector $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathrm{D}_i}^{n} \mathbf{x}_k$. We then calculate the between-class scatter matrix $\mathrm{S_B}$ as follows

$$\mathrm{S_B} = \sum_{i=1}^{c} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^{\mathrm{T}}$$

  where m is the overall mean, and $\mathbf{m}_i$ and N_i are the sample mean and sizes of the respective classes.
- Next, we solve the generalized eigenvalue problem for the matrix $\mathrm{S_W^{-1}S_B}$ to obtain the linear discriminants.
- We are not interested in projecting the data into a subspace that improves the class separability, and also reduces the dimensionality of our feature space. In order to do that, we need to decide which eigenvector(s) we want to drop for our lower-dimensional subspace, and hence take a look at the corresponding eigenvalues of the eigenvectors. The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data, and those are the ones we want to drop.

[6][https://sebastianraschka.com/Articles/2014_python_lda.html

- In LDA, the number of linear discriminants is at most $c - 1$ where c is the number of class labels, since the in-between scatter matrix S_B is the sum of c matrices with rank 1 or less. Note that in the rare case of perfect collinearity (all aligned sample points fall on a straight line), the covariance matrix would have rank one, which would result in only one eigenvector with a nonzero eigenvalue.
- After sorting the eigenpairs by decreasing eigenvalues, it is now time to construct our $d \times k$-dimensional eigenvector matrix W. In the last step, we just transform our samples onto the new subspace via the equation Y = XW. (where X is a $n \times d$-dimensional matrix representing the n samples, and Y are the transformed $n \times k$-dimensional samples in the new subspace

4.2.3 Regularized Discriminant Analysis

Regularized discriminant analysis is sort of a trade-off between LDA and QDA. Remember, in LDA we assume there is a common covariance matrix for all of the classes. QDA assumes different covariance matrices for all the classes. Regularized discriminant analysis is an intermediate between LDA and QDA, developed by J. Friedman. RDA shrinks the separate covariances of QDA toward a common covariance as in LDA.

Let's look at the covariance matrix estimation. Here below, we have a parameter, α , preselected to control which end you want to favor. Then, the , is a convex combination of the common covariance matrix in LDA and a separate covariance matrix estimated as in QDA. The parameter controls the complexity of the model.

RDA estimates the covariance matrix controlling the amount of tuning towards the common covariance from LDA, different covariance matrices from QDA as well as the Identity matrix. The *rda()* function from the *klaR* package, which *caret* utilizes, makes an additional modification to the covariance matrix, which also has a tuning parameter γ. In *rda()* function, the covariance matrix of the Gaussian distribution for each class k can be given as

$$\hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_k(\lambda) + \gamma \frac{1}{p} \text{tr}(\hat{\Sigma}_k(\lambda))I$$

where $\hat{\Sigma}_k(\lambda) = (1 - \lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma}$ (here, $\hat{\Sigma}_k$ is the covariance matrix of the kth class used in QDA while, Σ is the common covariance matrix of all classes used in LDA) . Both γ and λ can be thought of as mixing parameters, as they both take values between 0 and 1. For the four extremes of γ and λ, the covariance structure reduces to special cases:

- ($\gamma = 0$ and $\lambda = 0$): QDA - individual covariance for each group.
- ($\gamma = 0$ and $\lambda = 1$): LDA - a common covariance matrix.
- ($\gamma = 1$ and $\lambda = 0$): Conditional independent variables - similar to Naive Bayes, but variable variances within group (main diagonal elements) are all equal.

- ($\gamma = 1$ and $\lambda = 1$): Classification using euclidean distance - as in previous case, but variances are the same for all groups. Objects are assigned to group with nearest mean.

4.3 Evaluation Metrics

[7]

The metrics discussed below are calculated from the confusion matrix obtained after predictions on the test data. The confusion matrix is an organized way of mapping the predictions to the original classes to which the data belong. We familiarize ourselves with the following terms that appear in the confusion matrix:

- **True Positive (TP)** refers to a sample belonging to the positive class being classified correctly.
- **True Negative (TN)** refers to a sample belonging to the negative class being classified correctly.
- **False Positive (FP)** refers to a sample belonging to the negative class but being classified wrongly as belonging to the positive class.
- **False Negative (FN)** refers to a sample belonging to the positive class but being classified wrongly as belonging to the negative class.

4.3.1 Accuracy and Balanced Accuracy

In binary classification, *Accuracy* is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. Mathematically this can be expressed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The Balanced Accuracy on the other hand, is defined as follows:

$$\text{Balanced accuracy} = \frac{\text{TPR} + \text{TNR}}{2}$$

Balanced accuracy normalizes the true positive and true negative predictions by the number of positive and negative samples. It can serve as an overall performance metric for a model, whether or not the true labels are imbalanced in the data, assuming the cost of FN is the same as FP.

[7] https://en.wikipedia.org/wiki/Precision_and_recall

4.3.2 Precision and Recall/ Sensitivity and Specificity

In a classification task, the *precision* for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class). Mathematically,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

In a classification task, a precision score of 1.0 for a class C means that every item labelled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labelled correctly)

Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been). Recall is also sometimes referred to as *True Positive Rate* or *Specificity*. Mathematically,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A recall of 1.0 means that every item from class C was labelled as belonging to class C (but says nothing about how many items from other classes were incorrectly also labelled as belonging to class C.

4.3.3 F1-score

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$\text{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F_1 score is approximately the average of the two when they are close, and for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean. Since in this measure, precision and recall are evenly weighted, it is sometimes criticized as a evaluation metric due to this bias.

4.3.4 Youden's J statistic (also called Youden's index)

[8]

The Youden's J statistic is defined as follows:

$$\text{J} = \text{specificity} + \text{sensitivity} - 1$$

[8] https://en.wikipedia.org/wiki/Youden%27s_J_statistic

Intuitively, the Youden's J statistic can be thought of as the probability of an informed decision (as opposed to a random guess). The Youden's index is often used in conjunction with receiver operating characteristic (ROC) (defined below) analysis. The index is defined for all points of an ROC curve, and the maximum value of the index may be used as a criterion for selecting the optimum cut-off point.

4.3.5 ROC and AUC

Another method of seeing the fit is the ROC curve. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate or Sensitivity : It is a synonym for recall and is therefore defined as follows:
$$TPR = \frac{TP}{TP + FN}$$
- False Positive Rate (FPR) is defined as follows:
$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC is desirable for the following two reasons:

- AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

5 Results

5.1 Logistic Regression

The initial model that was fit on the *model_data* was a logistic regression model. Since *model_data* has 49 explanatory variables and 1 binary response variable (namely *LILA-Tracts_halfAnd10*, we fit a elastic net regularized logistic regression model on the complete

model_data. While creating the workflow process, all categorical predictors were converted to dummy variables and the numeric variables were normalized (scaled and shifted accordingly) to save computation time.

To avoid overfitting and to reduce the multicollinearity in the data we used elastic net regularisation and tuned the *penalty* of the regularisation as well as the *fraction of LASSO regularization* (in our code *mixture = 1* signifies LASSSO regularization and *mixture = 0* signifies the ridge regularization) in the model. The tuning was done after the data was firstly split into training and test sets and then the parameters were tuned on the basis of high average AUC values obtained on using **10 fold Cross Validation**.The best tuning parameter values for the *penalty* and *mixture* parameters stated above were obtained as :

penalty	mixture	Average AUC
0.000001	1	0.952
0.00000422	1	0.952
0.0000178	1	0.952
0.000001	0.6	0.952
0.00000422	0.6	0.952

The parameters were calculated using a grid search on the tuning parameter values for *penalty* and *mixture*. We had also considered other hyperparameter values taht are within one standard deviation of the average AUC but since our main goal is to reduce the number of explanatory variables, a *mixture* value of 1 was itself chosen for further predictions.

The next step was to tune the cutoff probabilities for the logistic regression. What this means is that we need to find a cutoff - probability above which we would classify the predicted probability obtained from the logistic regression for a specific tract, as either being a LILA tract or not. For this we considered different set of metrics discussed earlier to find the optimum cutoff probability. Note that since we wish to minimize the number of tracts we classify as belonging to non-LILA status but are actually LILA tracts, what we basically wish is to maximise recall/sensitivity (using the coding convention that in the binary encoding of LILA tract, we set non-LILA tract as a success rather than a LILA tract). From the figure below, we set our cutoff-probability value as 0.44

Using these hyperparamter values and the chosen cutoff-probability we obtain the following confusion matrix for the test data we created during our initial modeling phase:

	Truth = 0	Truth = 1
Predicted = 0	8810	1142
Predicted = 1	528	3868

The accuracy of this model on the test data was found out to be 0.884 while the AUC for the model on the test data turned out to be 0.949 suggesting quite a well fit.

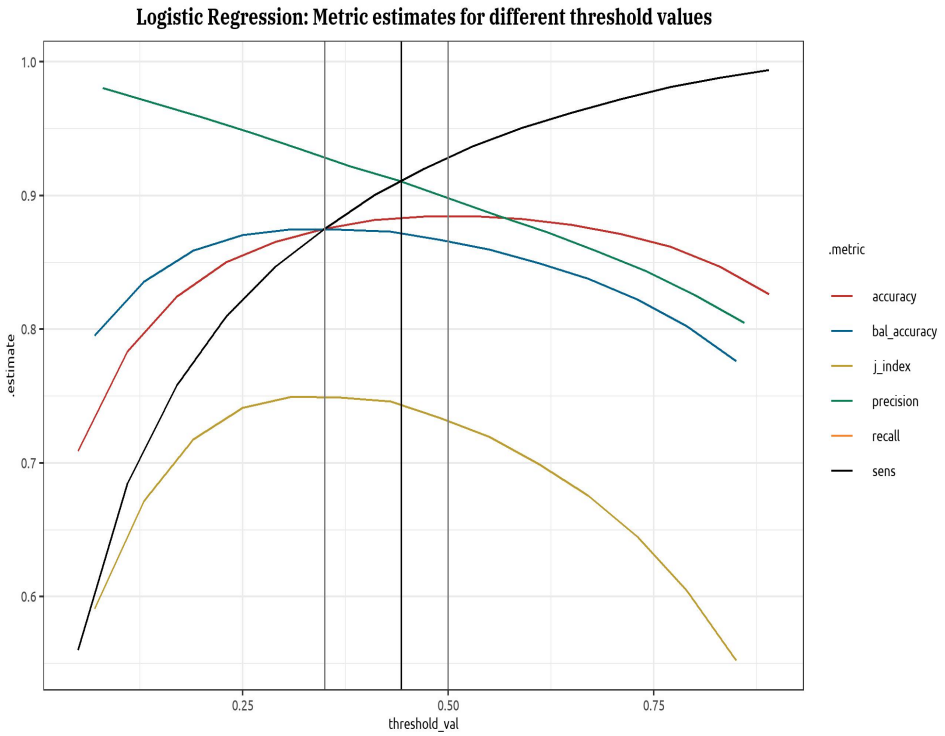


FIGURE 1: Values of different metrics for different values of cutoff-probabilities in Logistic Regression

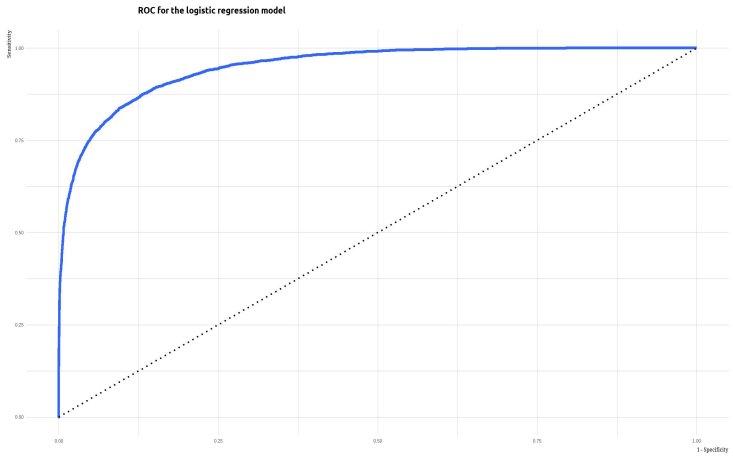


FIGURE 2: ROC curve for Logistic Regression Model

The ROC curve for this model is shown in the figure below

We can also find the relative importance of each variable in this logistic regression model. The 9 variables with highest relative importance (in absolute values) are shown in the figure below:

Compared to the relative importance of socio-economic indicators like Median Family

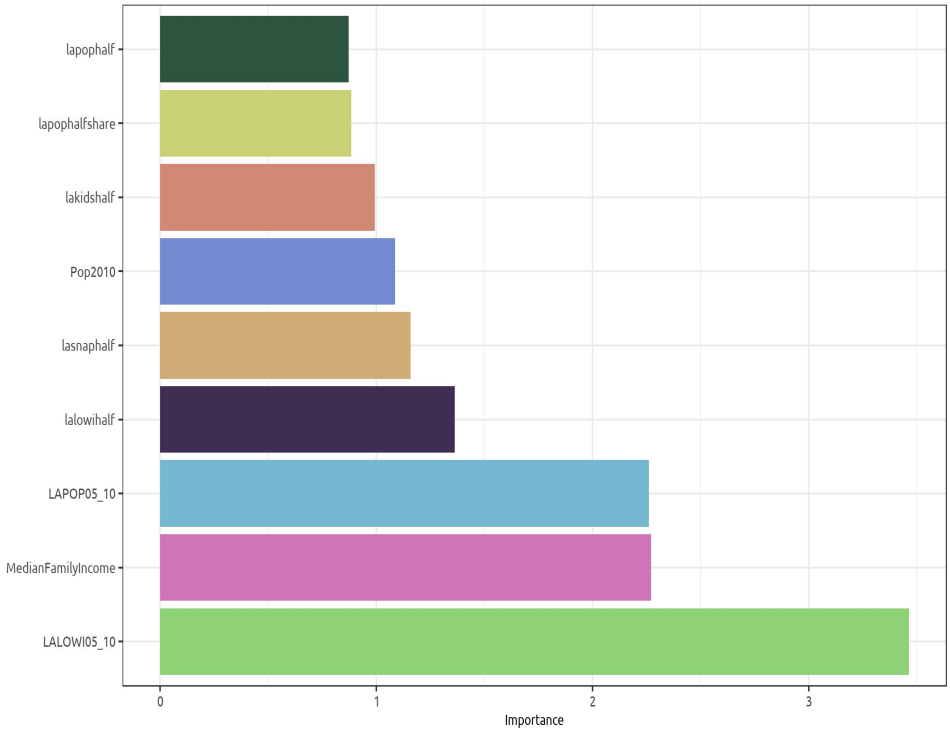


FIGURE 3: Variable Importance plot for the Logistic Regression model

Income, Poverty Rate, Low Access tract population,etc. the racial composition of the population in a tract does not seem to be a very important factor for predicting the Low Income Low Access status of a tract.

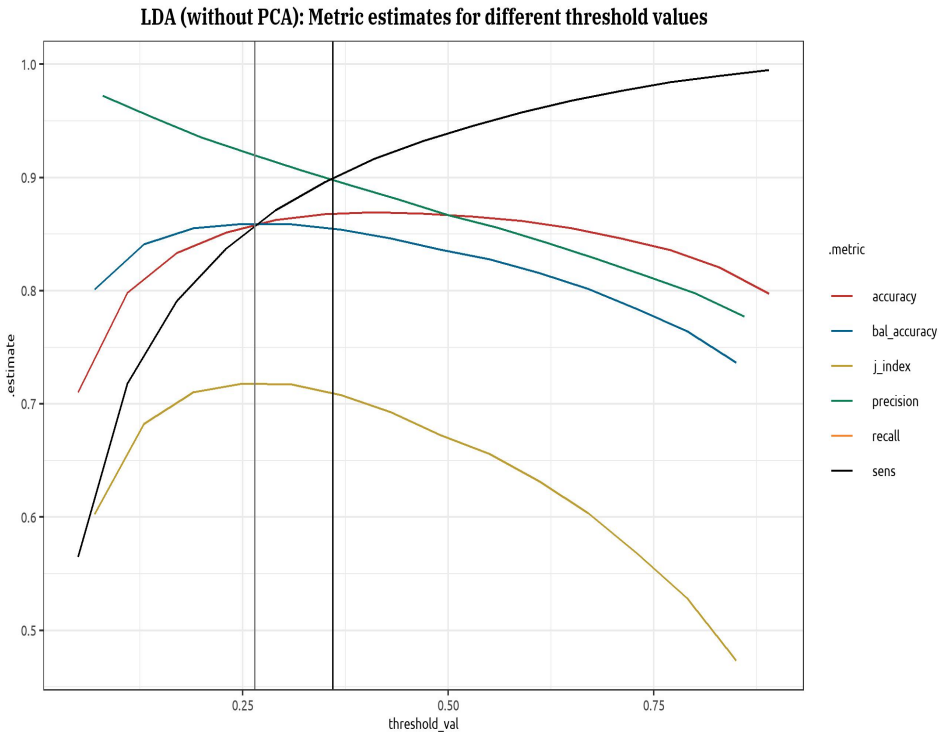


FIGURE 4: Values of different metrics for different values of cutoff-probabilities for the first LDA model(one where PCA was not applied)

5.2 Linear Discriminant Analysis

Since there are no regularization methods in the *MASS* package for the Linear Discriminant Analysis, we have fit two LDA models;

- The first model with all 49 predictors wherein all categorical predictors were converted to dummy variables and the numeric variables were normalized (scaled and shifted accordingly) but no other feature engineering process was applied
- The second model being same as the first one except that PCA was applied as a feature engineering step before the model was fit

In both the models, we have again tuned the cutoff probability to increase recall/sensitivity, accuracy and AUC. The cutoff probability for the first LDA model was found out to be 0.36 while that for that for the second LDA model was found out to be 0.35.

Using the above cutoff values the confusion matrix for the first LDA model is as follows:

	Truth = 0	Truth = 1
Predicted = 0	8997	1765
Predicted = 1	341	3245

while that for the second LDA model with PCA is as follows:

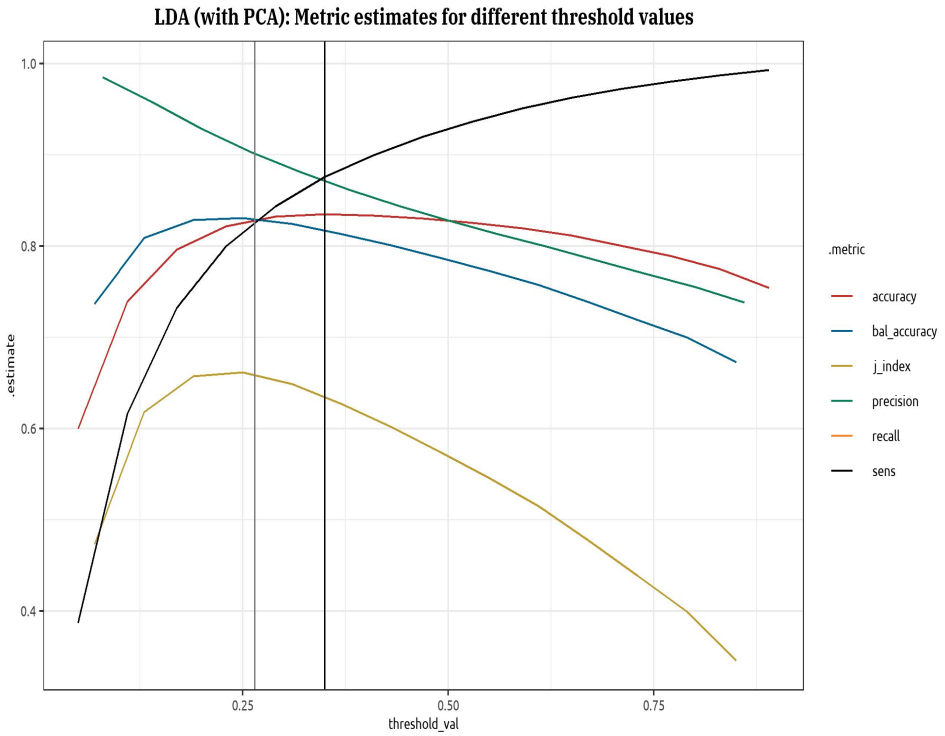


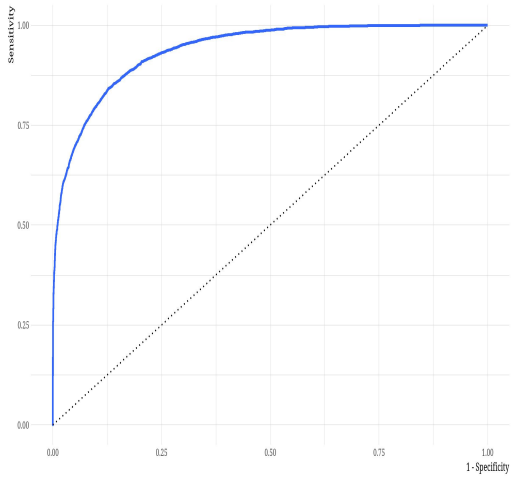
FIGURE 5: Values of different metrics for different values of cutoff-probabilities for the first LDA model(one where PCA was applied)

	Truth = 0	Truth = 1
Predicted = 0	9002	2366
Predicted = 1	336	2644

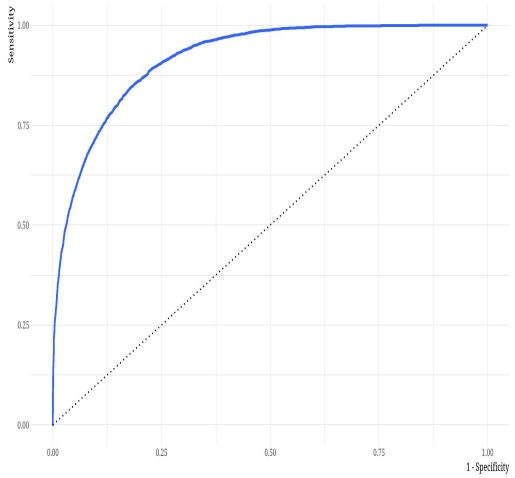
- LDA model without PCA had an accuracy of 0.853 and an AUC of 0.939 on test data
- LDA model with PCA had an accuracy of 0.812 and an AUC of 0.918 on test data

The ROC curves for the two models are shown in figure 6:

Thus, the LDA model without PCA seems to outperform the LDA model with PCA by just a small margin. But it is critical to note here that the variables used in the LDA model with PCA are inherently uncorrelated whereas that in the former had few variables that had a bit of collinearity among themselves. Moreover the model complexity also is reduced quite a bit after using PCA as opposed to that of without using PCA.



(A) ROC curve for LDA model without PCA



(B) ROC curve for LDA model with PCA

FIGURE 6: ROC curves for Linear Discriminant Analysis models

5.3 Regularized Discriminant Analysis

As seen in the Methods section earlier we have 2 parameters that we can tune in RDA. We denote the λ seen there as *frac_common_cov* and γ as *frac_identity*. To tune these hyperparameters we again turn towards **10 fold Cross Validation** as solution. We perform an extensive crossing grid search for these hyperparameter values and choose the combination that maximises the AUC. The first few combinations of these parameters that result in the highest AUC are as follows

frac_common_cov	*frac_identity*	Average AUC
1	0.2	0.939
1	0.4	0.935
1	0.6	0.931
0.9	0.2	0.930
0.9	0.4	0.928

From these we choose *frac_common_cov* = 1 and *frac_identity* = 0.2 which achieves the highest AUC of 0.938. Similar to LDA we again need to tune the cutoff probabilities too. We choose a near-optimum cutoff probability value from the plot below

The ROC curve for RDA is as follows:

As noted earlier, the AUC for this model is about 0.938 while the accuracy for this model is 0.851. An interesting thing to note here is that the AUC of the training data was almost equal to that of the testing data. Whereas in the case of Logistic regression and LDA

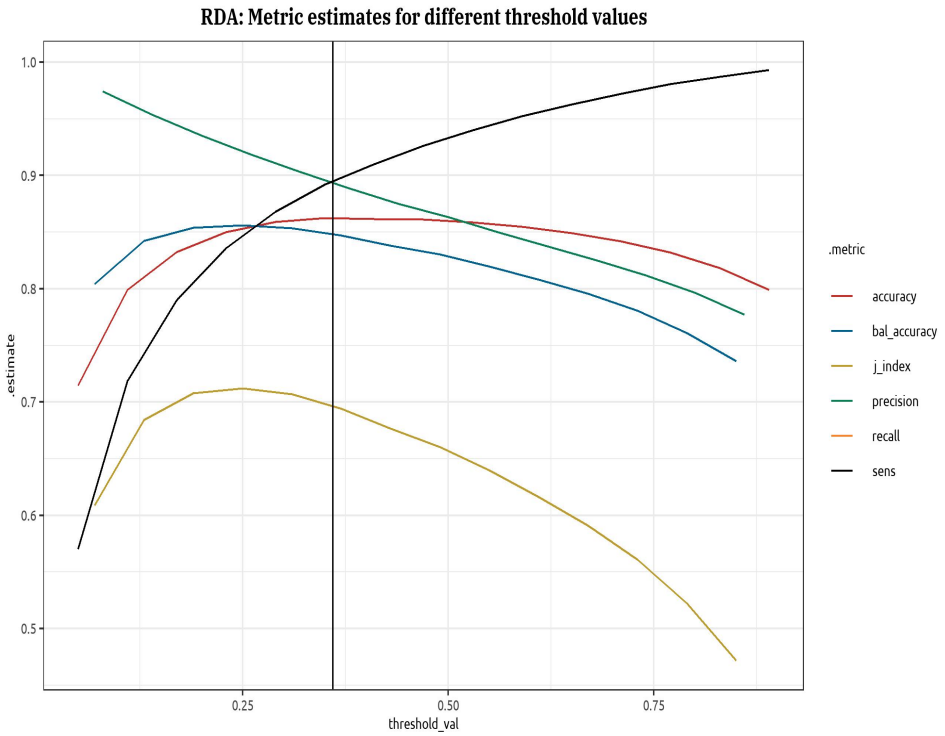


FIGURE 7: Values of different metrics for different values of cutoff-probabilities in RDA

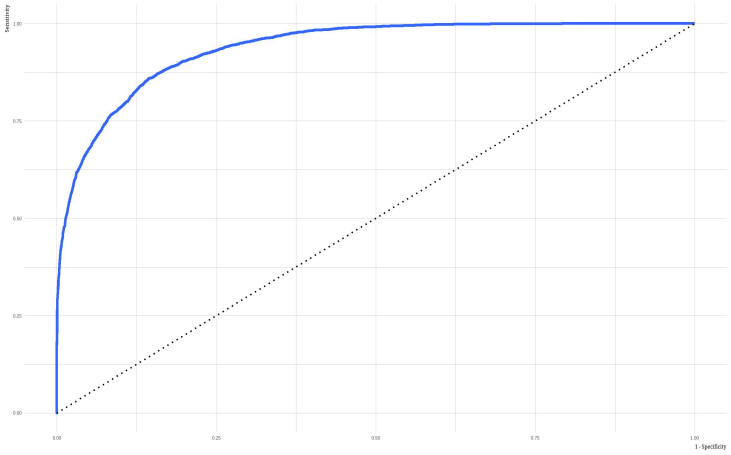


FIGURE 8: ROC curve for Regularized Discriminant Analysis

models, the AUC had dropped quite a bit when calculated on test data as opposed to that calculated on training data.

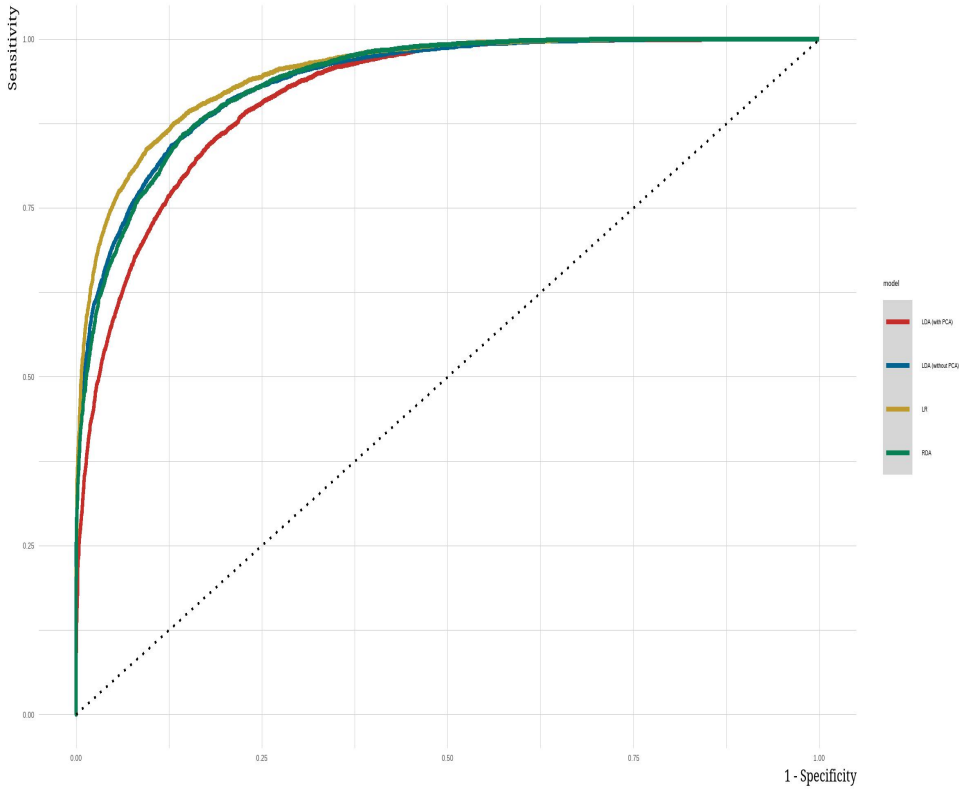


FIGURE 9: ROC curves for all models

6 Final Comparison of Models

We compare the above discussed models for various set of metrics discussed earlier as well as compare their ROC curves.

Metric	LR	LDA (without PCA)	LDA (with PCA)	RDA
Accuracy	0.884	0.853.2	0.812	0.851
AUC	0.949	0.939	0.918	0.938
J-Index	0.716	0.611	0.492	0.606
Precision	0.880	0.905	0.887	0.897
Recall	0.772	0.648	0.528	0.646
Specificity	0.943	0.963	0.964	0.960
Balanced Accuracy	0.858	0.806	0.746	0.803

The ROC curves can be compared as shown in the figure below

From the table and the plot above it is very clear that Logistic Regression outperforms LDA and RDA. Moreover due to distributional assumptions, Logistic Regression is more easy to interpret than LDA and RDA.

7 Conclusions

The largest magnitude in the regression model is of the variable measuring the share of the tract population that has low access and is availing SNAP benefits. This suggests efficient targeting by the SNAP program. One way this is useful to policymakers is that it suggests limited impact of a UBI scheme. For example, the impact of a 1000per year scheme would be the same as increasing the applicant base for SNAP participation by one percentage point in every tract.

$$\left|\begin{array}{l}\text{logOdds(UBI TO 1000 USD}\\ \text{to every family)}\end{array}\right| - \left|\begin{array}{l}\text{logOdds(Increase in eligibility for}\\ \text{SNAP by 1\% in every tract)}\end{array}\right| = \frac{4.813 - 7.836}{1000} \tag{1}$$

The latter is significantly cheaper to implement and does not involve any new legislation Thus, we recommend that un-targeted UBI schemes not be implemented and instead, the Federal Poverty line or the amount of SNAP benefits allowed to the beneficiaries be increased to improve food security.

Secondly, the coefficient for the number of kids and Seniors in a tract have negative coefficients. This suggests two alternative possibilities. Either the issue of food security is less prominent among kids, or that the Low Income Low access tract definition needs overhauling to accommodate the status of children and senior citizens. Because a larger proportion of households with kids are plagued by food insecurity than the general population, we conclude it is the latter possibility. Thus, we conclude that ERS needs to develop another way of measuring food insecurity that is more targeted towards children and seniors. Especially because these sections of the population have more expansive and varied and nutritional needs than the general proportion.

8 Strengths and limitations

The main limitation for our models is that we cannot split the data into training and test data lest we might miss out few peculiar geographical region based trends which would affect the overall model had they been included in the training data. This could be tackled using spatial analysis alongwith the analyses we provided. Moreover, Random Forests and KNN algorithms have also shown quite significant predictions in literature for such classification problems in general.

The main strength of the model is that it is able to explain most of the variation in the Low Income Low Access designation of a tract. It can be used to model more complicated scenarios.