PREDICTING LOW INCOME LOW ACCESS TRACTS IN US USING LOGISTIC REGRESSION, LDA AND RDA

Rohan Shinde

November 11, 2022

SMU, ISI Delhi

Table of contents

Table of Contents

- What is the project about?
- Objective of the project
- About data
- Discussion
- Methods used
- Results
- Strengths and limitations of models

WHAT IS THIS PROJECT ABOUT?

What is this project about?

What is food security?

- Measure of the availability of food and individuals' ability to access it
- Food security is characterised by the availability, accessibility and affordability of food at all times

Indicators of Food Security

- Measure of the availability of food and individuals' ability to access it
- Availability: Enough food is available to all people
- Accessibility: Food is within geographical reach of every person
- Affordability: Each person has the necessary funds to get adequate, safe and nutritious food.

What is this project about?

Availability

- Complicated by the disparate patterns of population density among the state (because of the Urban-Rural divide)
- Business-wise, itmakes much more sense to have stores in an urban area, compared to a rural ones
- Even in Urban areas, it is not necessary that the stores near a particular neighbourhood sell healthy foods

What is this project about?

Affordability

- Major factor in the obesity epidemic that is going on in the country
- People are forced to choose between having enough food on the table or having an inadequate amount of healthy food
- Even though nutritious food might be accessible, it does not mean that it is affordable
- This ties food insecurity with the issue of poverty

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

OBJECTIVE OF PROJECT

Objective of Project

- A paradigm shift has occurred wherein policy makers are switching from a targeted manner of delivering food subsidies to a Universal Basic Income scheme (UBI scheme).
- A UBI scheme: 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 present evidence to show that UBI is not the most efficient way to go about it

Objective of Project

- **Primary objective**: Predict Low Income Low Access tracts based on other socio-economic factors
- **Secondary objective**: Comparison between the efficiency of a proposed solution, UBI and the present method, SNAP.

ABOUT THE DATA

About the data

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
- Neighborhood-level indicators of resources, such as the average income of the neighborhood and the availability of public transportation

- Several indicators are available to measure food access along these dimensions

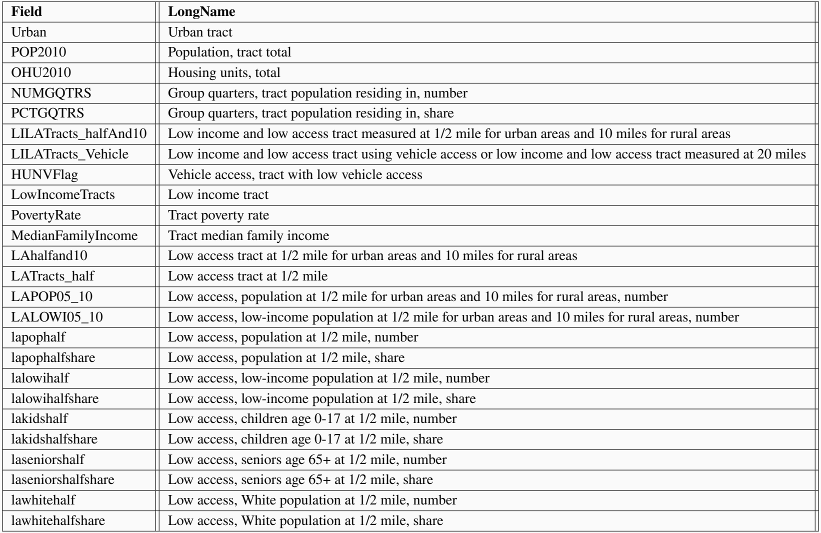
About the data

- 72531 observations and 147 variables
- Accessibility indicators for 3 different distance measures from nearest supermarket/superstore- half mile, 1 mile and 10 miles
- Other Census tract- specific variables like Population, Median Family Income, Poverty Rate, Number of Group Quarters, etc.
- 62% of the data is missing for accessibility indicators for 10 mile distance measure

About the data

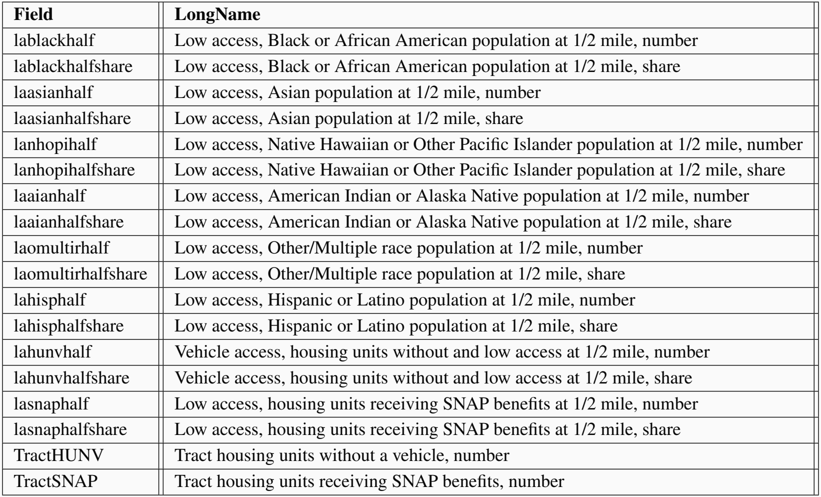
- 35% of it is missing for 1 mile distance
- As a result, we only predict the Low Income Low Access tracts using half mile measure in urban and 10 mile measure in rural tracts measure
- *LILATracts_halfAnd10* response denotes if a specific tract is a Low-income census tracts where a significant number (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

Variables used in the project



Field	LongName
Urban	Urban tract
POP2010	Population, tract total
OHU2010	Housing units, total
NUMGQTRS	Group quarters, tract population residing in, number
PCTGQTRS	Group quarters, tract population residing in, share
LILATracts_halfAnd10	Low income and low access tract measured at 1/2 mile for urban areas and 10 miles for rural areas
LILATracts_Vehicle	Low income and low access tract using vehicle access or low income and low access tract measured at 20 miles
HUNVFlag	Vehicle access, tract with low vehicle access
LowIncomeTracts	Low income tract
PovertyRate	Tract poverty rate
MedianFamilyIncome	Tract median family income
LAhalfand10	Low access tract at 1/2 mile for urban areas and 10 miles for rural areas
LATracts_half	Low access tract at 1/2 mile
LAPOP05_10	Low access, population at 1/2 mile for urban areas and 10 miles for rural areas, number
LALOWI05_10	Low access, low-income population at 1/2 mile for urban areas and 10 miles for rural areas, number
lapophalf	Low access, population at 1/2 mile, number
lapophalfshare	Low access, population at 1/2 mile, share
lalowihalf	Low access, low-income population at 1/2 mile, number
lalowihalfshare	Low access, low-income population at 1/2 mile, share
lakidshalf	Low access, children age 0-17 at 1/2 mile, number
lakidshalfshare	Low access, children age 0-17 at 1/2 mile, share
laseniorshalf	Low access, seniors age 65+ at 1/2 mile, number
laseniorshalfshare	Low access, seniors age 65+ at 1/2 mile, share
lawhitehalf	Low access, White population at 1/2 mile, number
lawhitehalfshare	Low access, White population at 1/2 mile, share

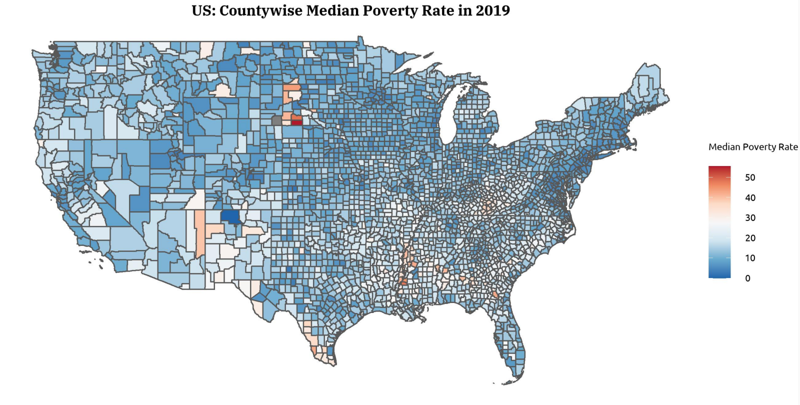
Variables used in the project



Field	LongName
lablackhalf	Low access, Black or African American population at 1/2 mile, number
lablackhalfshare	Low access, Black or African American population at 1/2 mile, share
laasianhalf	Low access, Asian population at 1/2 mile, number
laasianhalfshare	Low access, Asian population at 1/2 mile, share
lanhopihalf	Low access, Native Hawaiian or Other Pacific Islander population at 1/2 mile, number
lanhopihalfshare	Low access, Native Hawaiian or Other Pacific Islander population at 1/2 mile, share
laaianhalf	Low access, American Indian or Alaska Native population at 1/2 mile, number
laaianhalfshare	Low access, American Indian or Alaska Native population at 1/2 mile, share
laomultirhalf	Low access, Other/Multiple race population at 1/2 mile, number
laomultirhalfshare	Low access, Other/Multiple race population at 1/2 mile, share
lahisphalf	Low access, Hispanic or Latino population at 1/2 mile, number
lahisphalfshare	Low access, Hispanic or Latino population at 1/2 mile, share
lahunvhalf	Vehicle access, housing units without and low access at 1/2 mile, number
lahunvhalfshare	Vehicle access, housing units without and low access at 1/2 mile, share
lasnaphalf	Low access, housing units receiving SNAP benefits at 1/2 mile, number
lasnaphalfshare	Low access, housing units receiving SNAP benefits at 1/2 mile, share
TractHUNV	Tract housing units without a vehicle, number
TractSNAP	Tract housing units receiving SNAP benefits, number

EDA

EDA



Source: Calculated from USDA Food Access Reasearch Atlas Data

Figure: Median Poverty Rate

EDA

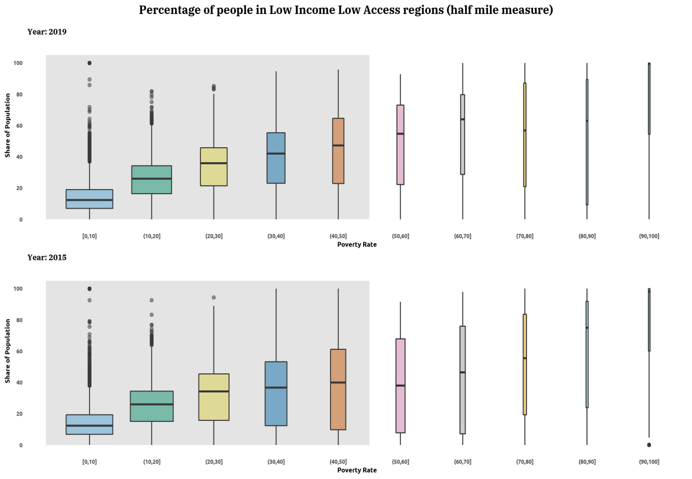
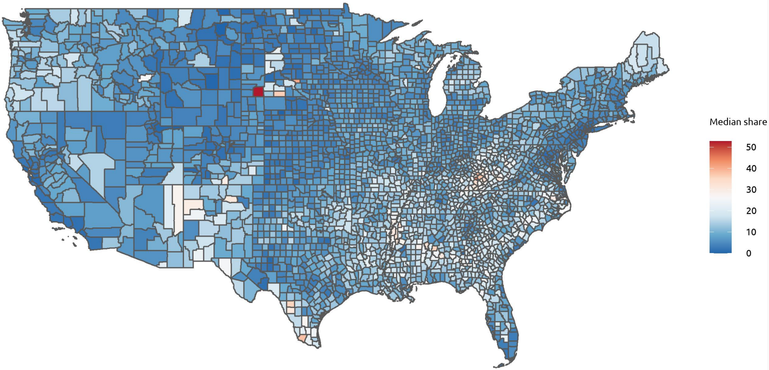


Figure: Poverty Rate in LILA regions (half mile measure)

EDA

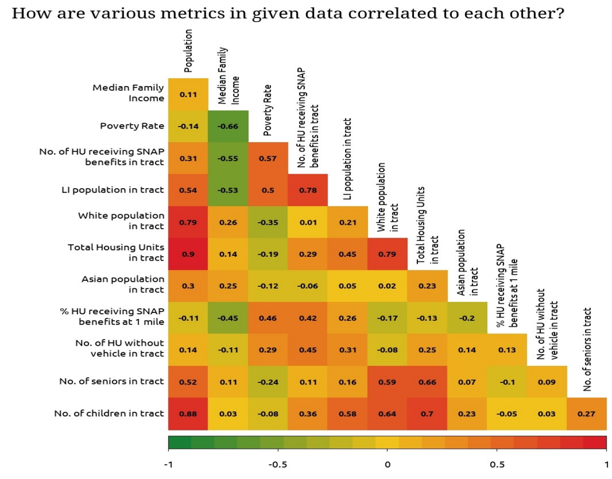
US: Countywise median percentage of population receiving SNAP benefits in 2019



Source: Calculated from USDA Food Access Reasearch Atlas Data

Figure: Median Percentage of people receiving SNAP benefits

EDA



How are various metrics in given data correlated to each other?

	Population	Median Family Income	Poverty Rate	No. of HU receiving SNAP benefits in tract	LI population in tract	White population in tract	Total Housing Units in tract	Asian population in tract	% HU receiving SNAP benefits at 1 mile	No. of HU without vehicle in tract	No. of seniors in tract
Median Family Income	0.11										
Poverty Rate	-0.14	-0.66									
No. of HU receiving SNAP benefits in tract	0.31	-0.55	0.57								
LI population in tract	0.54	-0.53	0.5	0.78							
White population in tract	0.79	0.26	-0.35	0.01	0.21						
Total Housing Units in tract	0.9	0.14	-0.19	0.29	0.45	0.79					
Asian population in tract	0.3	0.25	-0.12	-0.06	0.05	0.02	0.23				
% HU receiving SNAP benefits at 1 mile	-0.11	-0.45	0.46	0.42	0.26	-0.17	-0.13	-0.2			
No. of HU without vehicle in tract	0.14	-0.11	0.29	0.45	0.31	-0.08	0.25	0.14	0.13		
No. of seniors in tract	0.52	0.11	-0.24	0.11	0.16	0.59	0.66	0.07	-0.1	0.09	
No. of children in tract	0.88	0.03	-0.08	0.36	0.58	0.64	0.7	0.23	-0.05	0.03	0.27

Figure: Correlation plot

METHODS USED

Methods Used : Feature Engineering

Elastic Net Regularization

- Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models
- In high dimensional data (with n observations), the LASSO selects at most n variables before it saturates
- 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

Methods Used : Feature Engineering

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

Methods Used : Feature Engineering

Principal Component Analysis

- 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
- Allows analysis of datasets that may contain, for example, multicollinearity, missing values, categorical data, and imprecise measurements
- Goal is to extract the important information from the data and to express this information as a set of summary indices called principal components

Methods Used : Feature Engineering

- ’Preserving as much variability as possible’
- Finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other
- The first principal component is the direction in space (the space of predictor data points) along which projection have the largest variance

MODELS

Methods: Models

Logistic Regression

- 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
- 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
- The parametric model of Logistic Regression can be written as

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

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

Methods: Models

- The parameter $W = \{w_0, w_1, \cdots, w_n\}$ of the Logistic Regression is chosen by maximizing the conditional data likelihood

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

- 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.

Methods: Models

Linear Discriminant Analysis

- Used as a tool for classification, dimension reduction, and data visualization
- Often produces robust, decent, and interpretable classification results
- 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.
- 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

Methods: Models

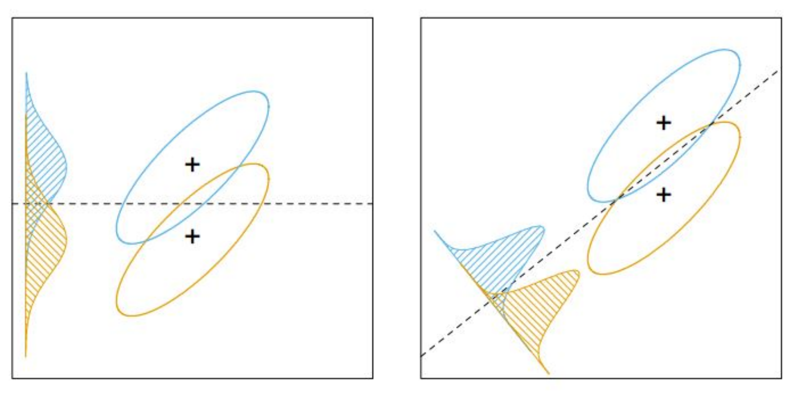


Figure: Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel)

Methods: Models

Regularized Discriminant analysis

- Trade-off between LDA and QDA
- 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
- RDA shrinks the separate covariances of QDA toward a common covariance as in LDA
- 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 (in the rda() function of klaR)

Methods: Models

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

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

Methods: Models

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.

Methods: Assessment Metrics

Assessment metrics 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.

Methods: Assessment Metrics

- **Accuracy and Balanced Accuracy**:

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

$$\text{Balanced accuracy} = \frac{\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}}{2}$$

- **Precision and Recall/ Sensitivity and Specificity**:

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

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

Methods: Assessment Metrics

- **F1-score**:

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

- **Youden's J statistic (also called Youden's index)**:

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

Methods: Assessment Metrics

ROC (Receiver Operating Characteristic) curve

- **True Positive Rate or Sensitivity**:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- **False Positive Rate**:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- 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

Methods: Assessment Metrics

AUC

- Stands for ”Area under the ROC Curve”
- Provides an aggregate measure of performance across all possible classification thresholds
- Classification-threshold-invariant. Measures the quality of the model’s predictions irrespective of what classification threshold is chosen.

RESULTS

Results: Logistic Regression

Initial Pre processing before fitting models

- Train-test split
- 10 fold Cross Validation for hyperparameter tuning

Logistic Regression

- Avoid overfitting and reduce multicollinearity: used elastic net regularisation
- 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

Results: Logistic Regression

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*
- The next step was to tune the cutoff probabilities for the logistic regression

Results: Logistic Regression

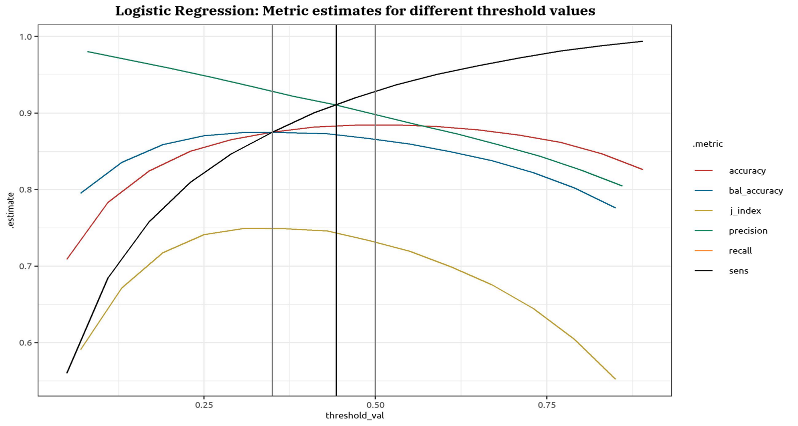


Figure: Values of different metrics for different values of cutoff-probabilities in Logistic Regression

Results: Logistic Regression

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

- Accuracy of this model on the test data : 0.884
- AUC of this model on the test data : 0.949

Results: Logistic Regression

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

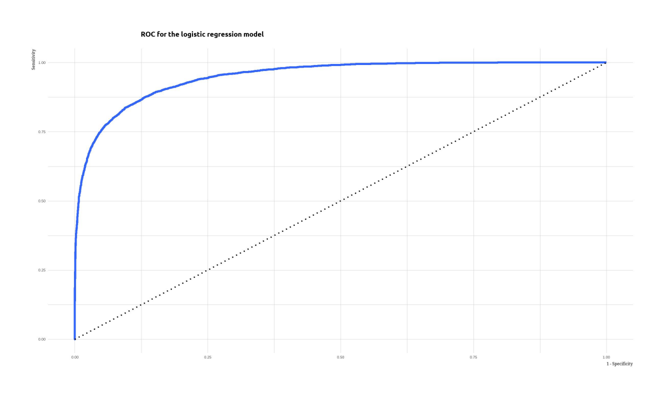


Figure: ROC curve for Logistic Regression Model

Results: Logistic Regression

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:

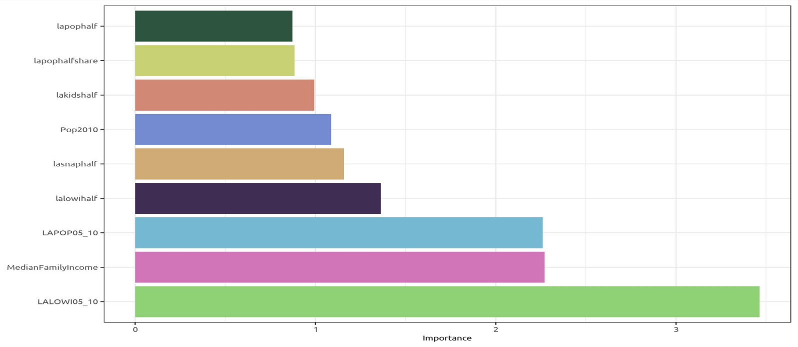


Figure: Variable Importance plot for the Logistic Regression model

Results: Linear Discriminant Analysis

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
- Cutoff Probability for LDA model without PCA: 0.36
- Cutoff Probability for LDA model without PCA: 0.35

Results: Logistic Regression

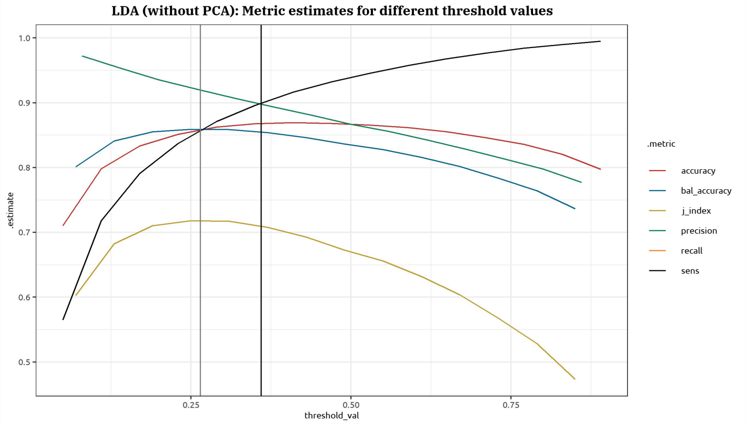


Figure: Values of different metrics for different values of cutoff-probabilities for the first LDA model(one where PCA was not applied)

Results: Logistic Regression

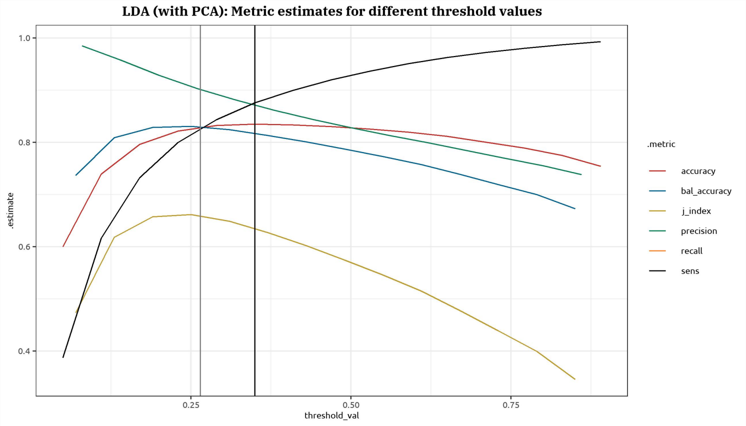


Figure: Values of different metrics for different values of cutoff-probabilities for the first LDA model(one where PCA was applied)

Results: Logistic Regression

Using the above cutoff values the confusion matrix for the first LDA model (Accuracy: 0.853 and an AUC: 0.939) is as follows:

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

while that for the second LDA model with PCA (Accuracy: 0.812 and an AUC: 0.918 is as follows:

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

Results: Linear Discriminant Analysis

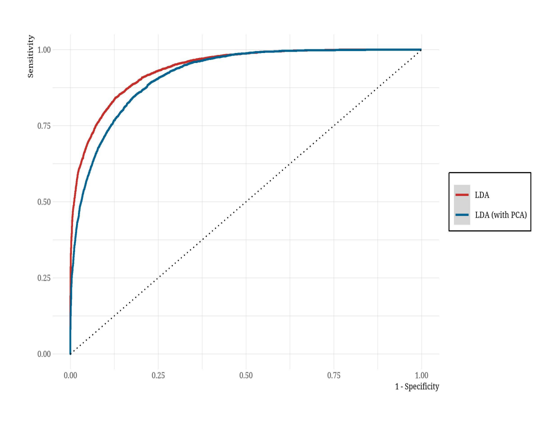


Figure: ROC curves for LDA models

Results: Regularized Discriminant Analysis

Regularized Discriminant Analysis

- We denote the λ seen in Methods section as *frac_common_cov* and γ as *frac_identity* 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

Results: Regularized Discriminant Analysis

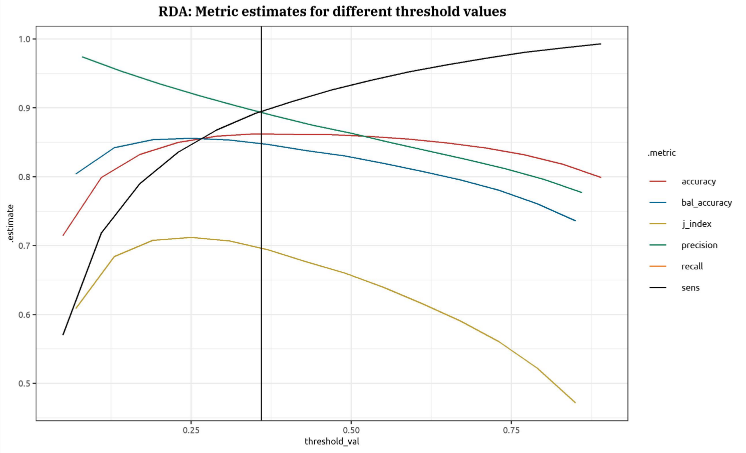


Figure: Values of different metrics for different values of cutoff-probabilities in RDA

Results: Regularized Discriminant Analysis

The ROC curve for RDA is as follows:

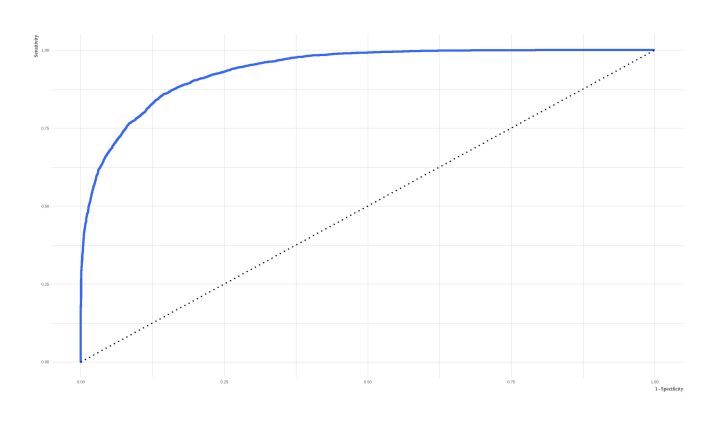


Figure: ROC curve for Regularized Discriminant Analysis

Results: Regularized Discriminant Analysis

- AUC for this model: 0.938
- Accuracy for this model: 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 models, the AUC had dropped quite a bit when calculated on test data as opposed to that calculated on training data

Results: Final Comparison of Models

Final Comparison of Models

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

Results: Final Comparison of Models

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

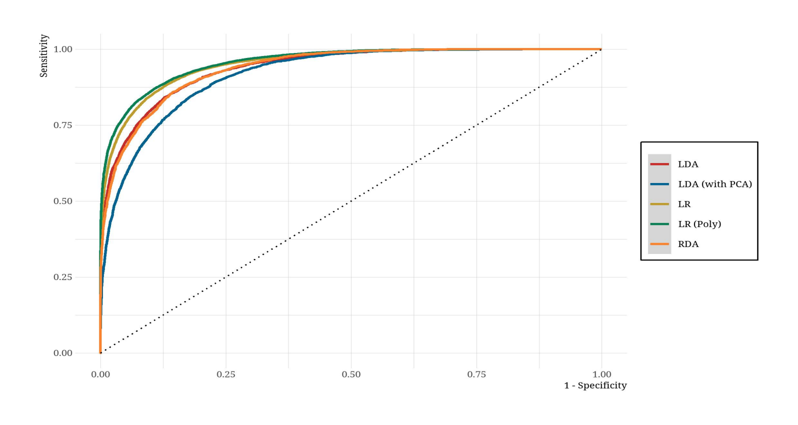


Figure: ROC curves for all models

Results: Final Comparison of Models

- 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.

CONCLUSIONS

Results: Final Comparison of Models

- One of the variables with coefficient of largest magnitude in the Logistic 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
- 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

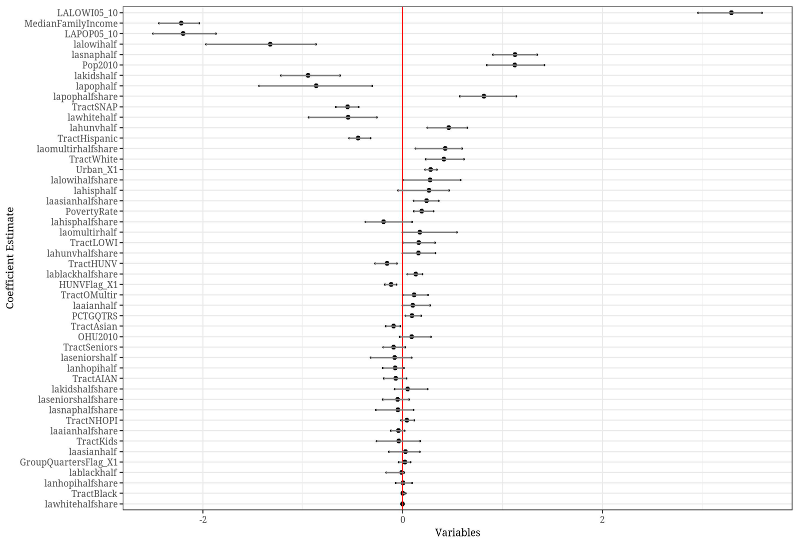
STRENGTHS AND LIMITATIONS

Strengths and Limitations

- Miss out few peculiar geographical region based trends
- Can be tackled using spatial analysis
- Random Forests and KNN algorithms
- Major strength: able to explain most of the variation in the Low Income Low Access designation of a tract

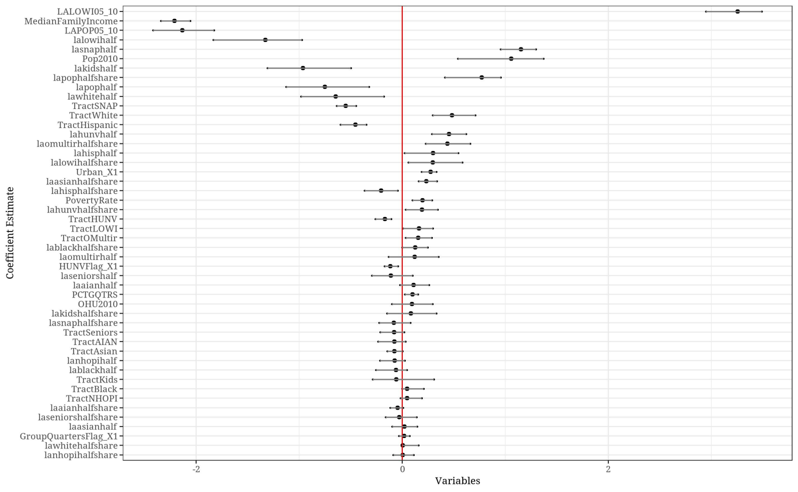
Arigato Gozaimasu (Thank You)

LR: Bootstrap Estimates of Coefficients



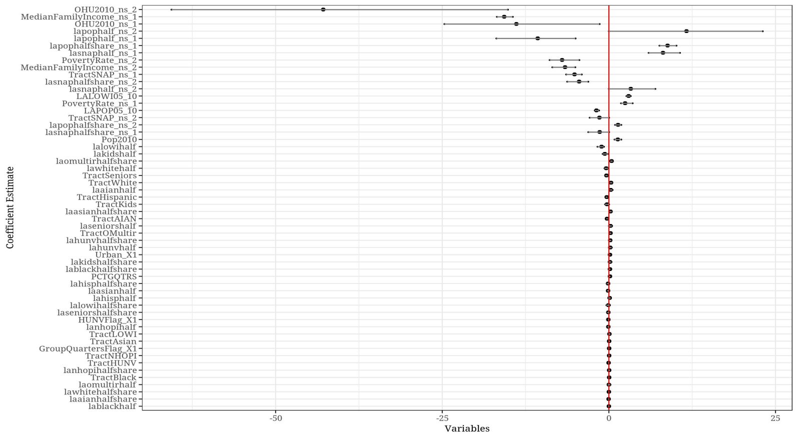
LR: Bootstrap Estimates of Coefficients

- **Mixture = 0.6**



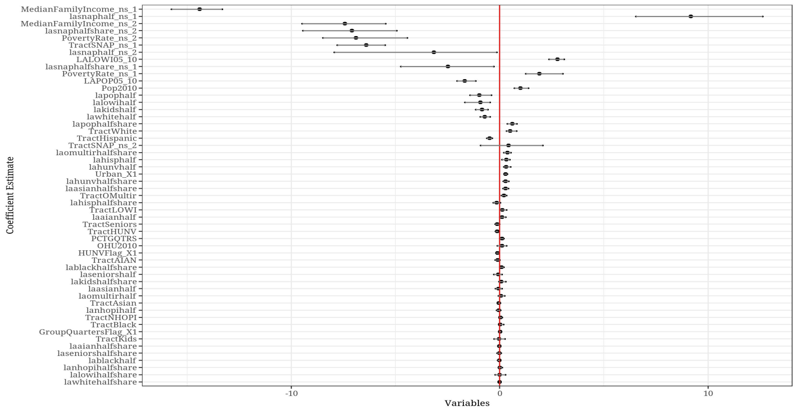
LR: Bootstrap Estimates of Coefficients

- **Logistic Regression model (Elastic Net Regularization applied) with quadratic transformations of appropriate variables**



LR: Bootstrap Estimates of Coefficients

- **Logistic Regression model (Elastic Net Regularization applied) with quadratic transformations of appropriate variables: Variables corresponding to highly sensitive estimates removed**


MedianFamilyIncome_ns_1
lasnaphalf_ns_1
MedianFamilyIncome_ns_2
lasnaphalfshare_ns_2
PovertyRate_ns_2
TractSNAP_ns_1
lasnaphalf_ns_2
LALOWI05_10
lasnaphalfshare_ns_1
PovertyRate_ns_1
LAPOP05_10
Pop2010
lapophalf
lalowihalf
lakidshalf
lawhitehalf
lapophalfshare
TractWhite
TractHispanic
TractSNAP_ns_2
laomultirhalfshare
lahisphalf
lahunvhalf
Urban_X1
lahunvhalfshare
laasianhalfshare
TractOMultir
lahisphalfshare
TractLOWI
laaianhalf
TractSeniors
TractHUNV
PCTGQTRS
OHU2010
HUNVFlag_X1
TractAIAN
lablackhalfshare
laseniorshalf
lakidshalfshare
laasianhalf
laomultirhalf
TractAsian
lanhopihalf
TractNHOPI
TractBlack
GroupQuartersFlag_X1
TractKids
laaianhalfshare
laseniorshalfshare
lablackhalf
lanhopihalfshare
lalowihalfshare
lawhitehalfshare
Coefficient Estimate
Variables