Submission by Team Inferno for This is Statistics Fall Data Challenge 2021

Pranav Krishna* 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.

- 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.

*B. Math. Third year, SMU, ISI Bagalore

†B. Math. Third year, SMU, ISI Bagalore

[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 but 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

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.

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.

4 Methods

4.1 Cluster Analysis

The regression model is 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 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.

Following this, we used k-means clustering to classify the states into distinct clusters with a similar distribution. In this method, states were grouped such that the sum of euclidean distance from the centroid of the group is minimised.This quantity is called the Within-Cluster-Sum of Squares(WSS). This sum was calculates for the number of clusters ranging from 1 to 10, and the states were classified into 4 clusters based on the optimum value. This means that beyond 4 clusters, there was no significant decrease in WSS.

4.2 Feature Selection

Feature selection was done using the Boruta for the default random forest model. The way it works is that the whole matrix of the available data is shuffled around to create "Shadow

Attributes" and the matrix is expanded by adding these as the columns. Then, the Random Forest model is fitted and an attribute is deemed a significant predictor if it has a feature importance than the best performing shadow attribute.

After this, we neglected the flagged variables because they depend more upon the definition used by USDA ERS and thus, are secondary to the socio-economic indicators.

4.3 Regression Model

Logistic regression is a type of a Generalised Linear Model used when we want to model the probability of a binary outcome. It is the model of choice for instances when we want to predict if any give photograph contains a dog. It can also be "Multinomial" when we care predicting 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. In our case, the predictor variables were selected using the output of the Boruta random forest classifier. 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, i.e. $\sum_i a_i x_i$. We can solve the equation for log odds as

$$\frac{p}{1-p} = b^{\sum_i a_i x_i}$$

.

```
df <- farad %>%
  dplyr::select(PovertyRate_2019,MedianFamilyIncome_2019,
                TractSNAP_2019,
                TractLOWI_2019,
                TractWhite_2019,
                OHU2010,
                TractAsian_2019,
                lasnap1share_2019,
                TractHUNV_2019,
                TractSeniors_2019,
                TractKids_2019,
                Pop2010,
                LILATracts_1And10_2019)



df2 <- df[complete.cases(df),]

model1 <- glm(LILATracts_1And10_2019~. ,
              data=df,family=binomial)
model2 <- MASS::stepAIC(model1)
```

FIGURE 1: R code to create the Logistic regression model **model1** in R

5 Results

5.1 Cluster Analysis

In each case (Figure 2, Figure 3, and Figure 4), there was a clear elbow at 4 clusters in the WSS. Thus, we selected 4 as the number of clusters in our analysis.

From the chloropleths it is evident that there are a few states like Arkansas, Mississippi, New Mexico, and Georgia which had high percentage of populations living in Low Access tracts and even had a high state median Poverty Rate. States like Louisiana, South Carolina and Alabama belonged to clusters with high percentage of people in Low Income Low Access tracts. From the distribution of clusters, we notice that the percentage of people in LILA tracts is relatively higher in the Southern and South-Eastern regions of US.

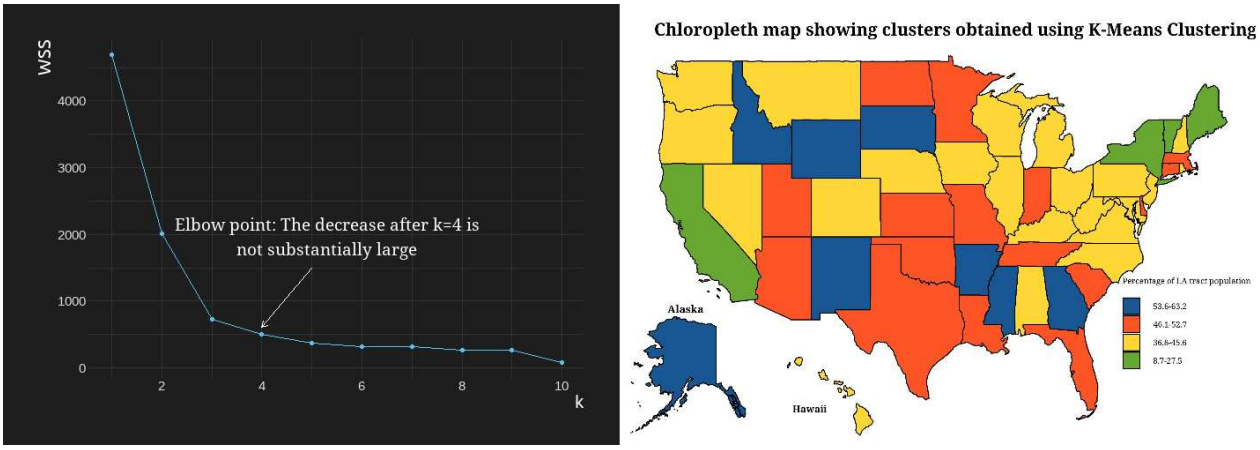

FIGURE 2: The "Within Sum of Squares" plot (left) and chloropleth map (right) showing the clusters of States based on percentage of people living in Low Access Tracts.

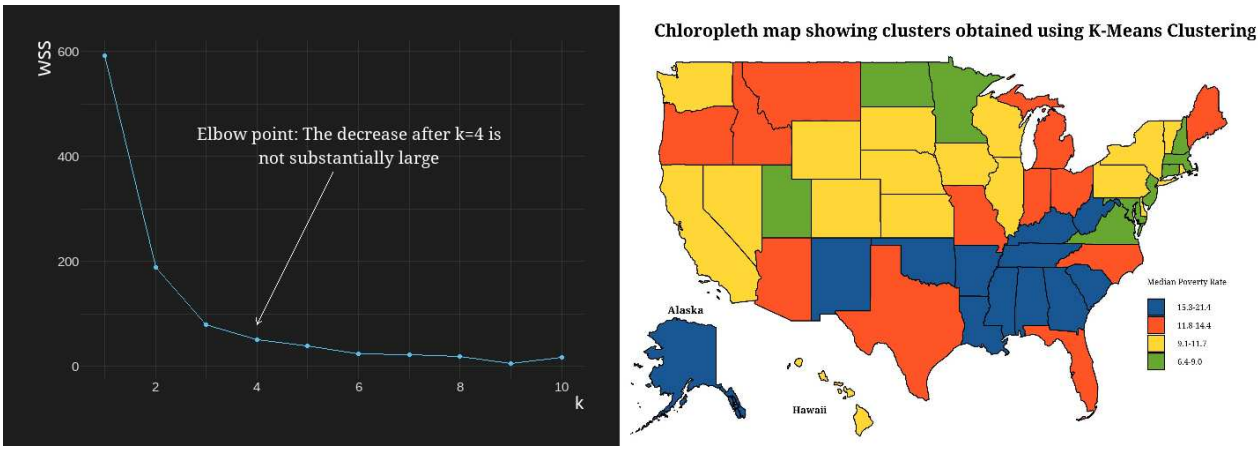

FIGURE 3: The "Within Sum of Squares" plot (left) and chloropleth map (right) showing the clusters of States based on Median Poverty Rate of the States.

5.2 Feature Selection

For each column of the relative importance, we calculated the average and sorted from most important to least important. As expected the flag for Low Income Tracts and the flag for Low access tracts(1 mile limit for Urban and 10 mile limit for Rural tracts) were the most important factors. This is because the Low Income Low Access designation is the product of the two indicator random variables. Other than that, we go on to find that socio-economic indicators like Poverty rate and median family income are important predictor variables.

Compared to the relative importance of these socio-economic indicators, 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. Broadly speaking, the number of White, Asian, Black and Hispanic people figure in the top 22 relatively important predictors, but compared to the socio-economic indicators of a tract, they are not as important.

The top 22 relatively important indicators are summarised in the table as follows-

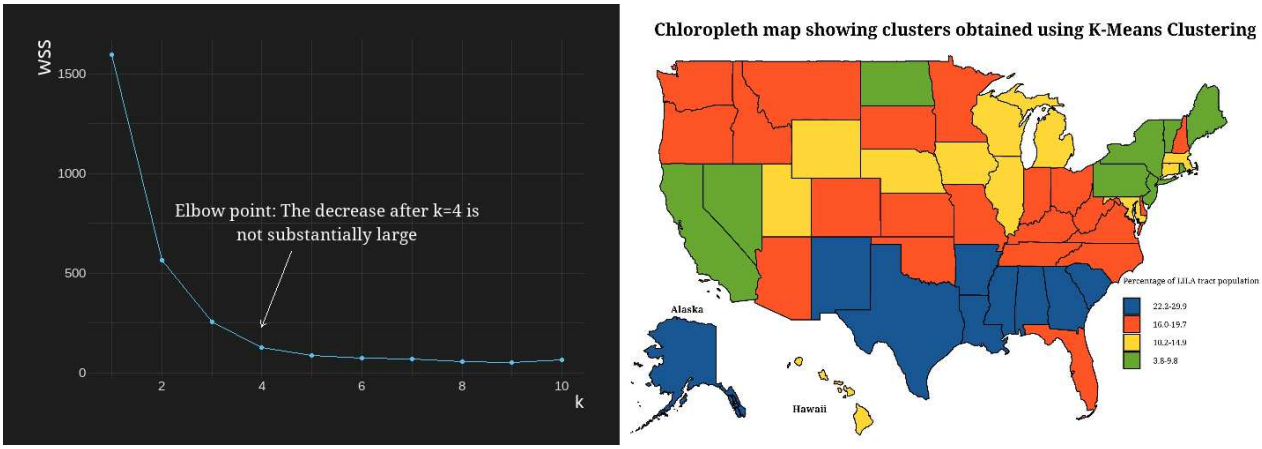

FIGURE 4: The "Within Sum of Squares" plot (left) and chloropleth map (right) showing the clusters of States based on percentage of people living in Low Income Low Access Tracts.

	Relative Importance
LowIncomeTracts	71.28
LA1and10	27.42
PovertyRate	26.23
LATracts1	22.56
MedianFamilyIncome	22.24
TractSNAP	19.14
TractWhite	18.56
TractLOWI	18.40
LATracts10	17.46
lasnap1share	15.62
OHU2010	15.01
TractAsian	14.90
TractSeniors	14.71
TractKids	14.67
TractHUNV	14.54
Pop2010	14.23
lalowi1share	13.84
LAPOP1_10	13.29
LALOWI1_10	13.12
TractOMultir	12.73
TractBlack	12.38
TractHispanic	12.27

5.3 Regression

The regression model based on the results of feature selection is given in Figure 5

```
summary(model1)
```

```
##
## Call:
## glm(formula = LILATracts_1And10_2019 ~ ., family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9708  -0.5252  -0.2703  -0.0665   3.6648
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.954e-01  1.128e-01   5.278 1.31e-07 ***
## PovertyRate_2019         -3.617e-03  2.181e-03  -1.658  0.09728 .
## MedianFamilyIncome_2019 -4.813e-05  1.438e-06 -33.474  < 2e-16 ***
## TractSNAP_2019           -1.706e-03  1.494e-04 -11.422  < 2e-16 ***
## TractLOWI_2019            6.425e-04  3.436e-05  18.697  < 2e-16 ***
## TractWhite_2019          -2.611e-04  1.589e-05 -16.429  < 2e-16 ***
## OHU2010                   1.604e-04  6.442e-05   2.490  0.01278 *
## TractAsian_2019           1.971e-04  5.992e-05   3.289  0.00101 **
## lasnap1share_2019         7.836e-02  2.048e-03  38.264  < 2e-16 ***
## TractHUNV_2019            1.326e-03  1.755e-04   7.551 4.30e-14 ***
## TractSeniors_2019        -2.708e-04  6.341e-05  -4.270 1.95e-05 ***
## TractKids_2019           -8.156e-04  6.867e-05 -11.877  < 2e-16 ***
## Pop2010                   2.538e-04  3.234e-05   7.850 4.16e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 48585  on 52045  degrees of freedom
## Residual deviance: 33072  on 52033  degrees of freedom
##   (20485 observations deleted due to missingness)
## AIC: 33098
##
## Number of Fisher Scoring iterations: 7
```

FIGURE 5: The result of the summary call on the model in R

We also wish to see measures of how well our model fits. The output produced by summary(model1) included indices of fit (shown below the coefficients), including the null

and deviance residuals and the AIC. One measure of model fit is to check how well the model is able to predict the LILA status of the tract given the important metrics derived from the random forest classifier Boruta.

```
glm_probs<- data.frame(probs=predict(model1,type="response")) %>% as_tibble()

glm_pred<- glm_probs %>%
  mutate(pred=ifelse(probs>.5, 1, 0))

glm_pred = cbind(df2, glm_pred)

glm_pred %>%
  count(pred, LILATracts_1And10_2019) %>%
  spread(LILATracts_1And10_2019, n, fill = 0)
```

```
##   pred     0    1
## 1    0 40909 5826
## 2    1  1927 3384
```

```
glm_pred %>%
  summarize(score = mean(pred == LILATracts_1And10_2019))
```

```
##       score
## 1 0.8510356
```

FIGURE 6: Summarizing score for how well the aforementioned model fits the actual data

Another measure of model fit is the significance of the overall model. This test asks whether the model with predictors fits significantly better than a model with just an intercept (i.e., a null model). The test statistic is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-squared with degrees of freedom equal to the differences in degrees of freedom between the current and the null model (i.e., the number of predictor variables in the model).

```
with(model1,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail=F))
```

```
## [1] 0
```

The chi-square of 15512.71 with 12 degrees of freedom and an associated p-value of less than 0.001 (R approximates values very close to 0 with 0) tells us that our model as a whole fits significantly better than an empty model.

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

Here TP = True Positives, FN = False Negatives, FP = False Positives and FN = False Negatives. 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.

An AUC value of 0.8578 signifies that the model indeed provides a quite good fit to the data pertaining LILA status of the tract and the relevant metrics for the same.

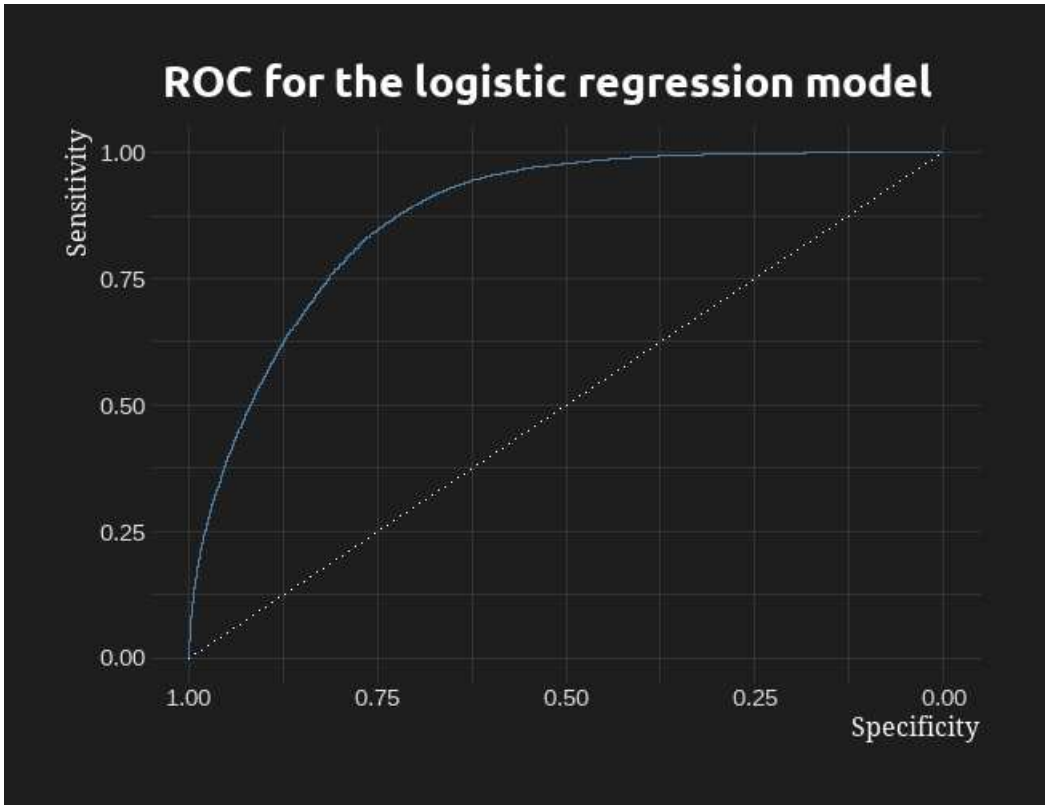

FIGURE 7: The Area under ROC in our logistic model is about 0.8578

6 Recommendations

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.

7 Strengths and limitations

The main limitation for our regression model is validation, or more precisely, the lack of it. Because we have predicted the results on training data itself, we cannot just assume that the model is good for predicting LILA status from the important factors derived from Boruta. Moreover we cannot split the data into training and test data lest we might miss out few peculiar geographical regions which would affect the overall model had they been included in the training data. So instead of splitting the data into train and test data we try to find the statistical validity of this logistic model using other means.

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.