Student: Ana Prada
Class: Advance Quant.

Predicting Food insecurity in NYC

Introduction and motivation: Urgency of COVID-19

This project will focus on developing a model to understand how to predict food insecurity in NYC. It develops both a logistic regression and a tree model to address the following question "What best factors that are available to us best predict the places where there is a significant amount of food insecurity individuals?". The Coronavirus Disease of 2019 (COVID-19) pandemic threatened the lives and livelihoods of people and significantly impacted the food system in New York City. Before the COVID-19 crisis began, more than 1.1 million people lived in food-insecure households. According to reports by the Food Bank for NYC, approximately 75% of food pantries and soup kitchens surveyed during the first months of the pandemic reported an increase in the number of visitors and nearly one-third reported the number of visitors at their programs had more than doubled. Black and Latino adults were more than twice as likely as white adults to report that their household did not get enough to eat[1].

The pandemic is not yet over, and the future remains tenuous for people who have experienced uncertain access to enough food for their families. It is likely that it will take time for food insecurity levels to recover. There is an urgent need to reimagine how to distribute resources and shift our focus on resilient, long-term and creative ways to redistribute local food systems. We all deserve equal access to nutritious, culturally appropriate food especially during a pandemic that has left thousands of people unemployed and on the verge of eviction.

Food is central to the health, well-being, cultural heritage, economic and social resilience of low-income communities, which are often communities of color. Healthy, sustainable, accessible foods can mitigate the risks of experiencing diet-related illnesses, food insecurity, social isolation, and environmental degradation. There are two main dimensions to food security that should be considered: the production and supply of an adequate quality and quantity of food, and the ability of people to access food. What households spend for food is determined by both item prices and selections. Household food surveys show[2] low-income families tend to spend their food dollars differently and spend less per pound for nearly all broad food groups than do all households combined. They are able to do this by purchasing lower cost items within the broad food groups. The Bureau of Labor Statistics' Consumer Expenditure Survey data reveals (CES) that poor households devote a greater share of their income to food spending than did wealthier households. The lowest 20% of income households spend on average 20% ($4,850) on Food and 45% on Housing, while the 20% highest income Households spend 11% ($13,200) and 30% respectively. Food spending increases with household income, as wealthier households buy higher-quality food items and more convenience foods. But for lower-income households, who live in communities that have a higher ratio of small stores to supermarkets than high income communities, a higher proportion of spending goes

---

[1] https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-recessions-effects-on-food-housing-and

[2] Coleman-Jensen, Alisha, Matthew P. Rabbitt, Christian A. Gregory, and Anita Singh. 2020. Household Food Security in the United States in 2019, ERR-275, U.S. Department of Agriculture, Economic Research Service.

toward food because, on average, the estimated size of the price difference between small stores and supermarkets is 10 percent.

Most of the studies on the disadvantages of poor urban neighborhoods have focused on the quality of public community facilities. However, the quantity and quality of local private amenities, such as grocery stores and restaurants, can also have important quality of life implications for communities. Some research suggests that a smaller number of retails implies a more limited choice, and the lack of competition leads to higher prices where "the poor pay more" for many basic goods and services. Following Meltzger and Schuetz, we focused on understanding neighborhood stores whose customers represent primarily the immediate vicinity. These retails reflect most likely the composition of neighborhood residents. Literature suggests that the goods most likely to be sold at neighborhood stores include groceries, health and beauty products, and general household items, such as cleaning and household supplies. In addition to retail, some prime services like laundry services, coffee shops, and limited-service restaurants, and beauty salons were considered in this report.

Meltzger and Schuetz's research suggests poor neighborhoods are more disadvantaged in food service than in retail, and within retail, the differences are smallest for basic necessities, such as grocery stores and pharmacies. Also, poor neighborhoods have a much higher proportion of unhealthy chain restaurants. However, predominantly Latino neighborhoods have more diverse food services and greater physical access to retail corridors than predominantly White and Black neighborhoods. Together, these results suggest that residents in relatively low-income neighborhoods have retail activity nearby, but that it is less dense and composed of smaller and less diverse options (both of which could have implications for the quality and cost of the goods and services). Finally, the results showed that low-income neighborhoods have greater access to transit and more retail space per building. This is important because, in spite of possessing some characteristics that would, theoretically, make these neighborhoods more appealing to retail businesses, they still face less retail access overall.

Data preparation

I used the library(tidycensus)

```
library(tidyverse)
options(tigris_use_cache = TRUE)


NYC <- get_acs(state = "NY", county = "New York" | "Richmond" | "Kings" | "Bronx" | "Queens", geography = "tract",
               variables = "B19013_001", geometry = TRUE)
head(NYC)
```

I selected the following variables and divided them in three different categories, Economic, Social and Housing:

Dependent variable: Food Insecure[3] (% of residents), I choose to convert this into a Dummy variable where I recorded as 1 the tracts that were > than the third quantiles. After running the first logistic regression I had to change this and create a new one where 1 were the tracts > =than the median.

Housing and Infrastructure

1. ex_high_cost_h_tract -> Extreme Housing Burden (% of renters)
2. more_than_one_tract-> Overcrowding (% of units with more than one occupant per room)
3. Number of food retails by census tract (Count point per census tract)
4. no_kitchen_tract -> Lack Kitchen (% of housing units)

Demographics

1. black_pop_tract -> Black (% of total population)
2. latino_pop_tract -> Latino (% of total population)
3. disabled_tract -> Disabled (% of total population)
4. lonely_tract -> Living Alone (% of households)
5. sixtyfive_tract

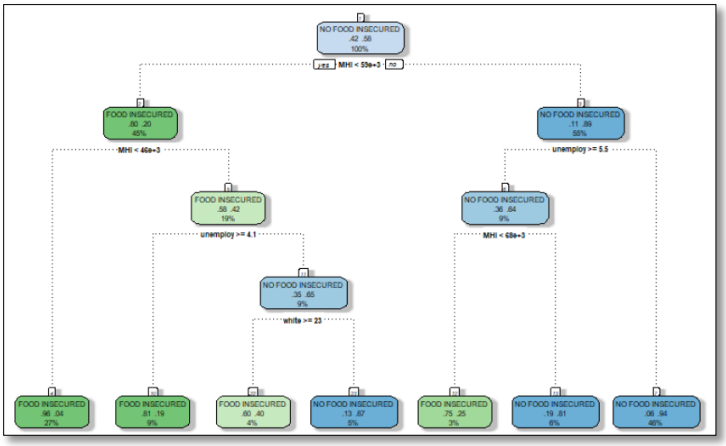
Economic

1. Median Household income
2. % unemployment
3. %poverty (This variable had to be dropped because there were some circularity issues)

The N/A are census tracts where there are no population, like Central Park. Those tracts were erased from the data set.

[3] Measure of America, Social Science Research Council. 2016.

Food insecurity Logistic Regression and Random Tree Model




```
Call:
glm(formula = food2 ~ MHI + unemploy + singlemom + black + latino +
    disabled + familychild + high + more_one, family = binomial,
    data = FOOD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4983  -0.2761  -0.0124   0.1943   3.7484

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.771e+00  9.745e-01   6.949 3.69e-12 ***
MHI         -1.441e-04  9.028e-06 -15.960  < 2e-16 ***
unemploy     7.021e-01  5.352e-02  13.118  < 2e-16 ***
singlemom    8.099e-02  2.380e-02   3.403 0.000667 ***
black       -1.577e-02  4.612e-03  -3.419 0.000628 ***
latino      -3.736e-02  5.486e-03  -6.810 9.76e-12 ***
disabled     1.430e-01  2.454e-02   5.828 5.62e-09 ***
familychild -4.817e-02  1.178e-02  -4.090 4.31e-05 ***
high        -3.551e-02  9.357e-03  -3.795 0.000148 ***
more_one     1.064e-01  1.632e-02   6.520 7.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2858.32  on 2100  degrees of freedom
Residual deviance:  959.05  on 2091  degrees of freedom
AIC: 979.05
```

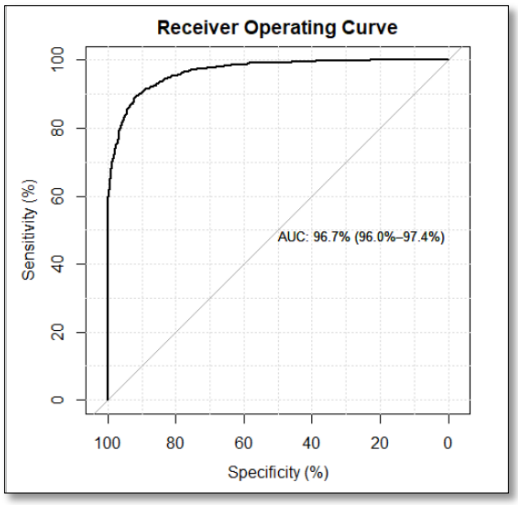
The Final Model (Fig. 1) includes nine statistically significant predictors to Food Insecurity. The main contributor, after controlling for the variation contributed by remainder of the variables in the model and holding all else constant, are weather the census tract has a low median household income ($p < 0.5$) as compared to men, which decreases the chances of Food insecurity. Looking closer at the Random Tree, it is interesting how both MHI and unemployment are the most important predictor of food insecurity, also Census tracts with a White population higher than 23%. Other significant contributors include, Unemploymeny, and Latino. In terms of their predictive power (z scores) and contribution to the model the variables rank as follows: **MHI:** |-15.960|, **Unemployment:** |13.118|, **Latino:** |-6.810|, **Overcrowding(more_one):** |6.520|, **Disability:** |5.828|, **Family with children:** |-4.090|, **Black** |-3.419|, **High rent burden:** |-3.795|, **Singlemom:** |3.403|. This model has a total accuracy of 0. 0.904 with a threshold of 0.5. From the 1219 that were predicted to be Food secured, 1126 we Secured and 93 were Insecured. From the 882 that were predicted to be Food insecured, 775 were Insecured and 107 were secured. Given the sensitivity of the subject matter we want to evaluate the model based on its ability to detect true positives and having low false positives being especially important as these represent the number of people that are predicted to not be food insecured and in fact were.

The Food insecurity calculation considers several variables, to create a method to predict which census tracts in NYC are food insecured and which ones are not. In addition to performing with almost 90% percent accuracy, the algorithm revealed that being not white and living in overcrowding settings and disability are fundamental factors in determining whether a census tract is food insecured or no, those in high income census tracts, especially mostly white census tracts, are more likely to be Food Secured than cesus tracts with similar incomes but less white population. The fact that in both models, Income, ethnicity and race are significant in Food insecurity shows, how Food Insecurity is a product of structural racism and inequality.

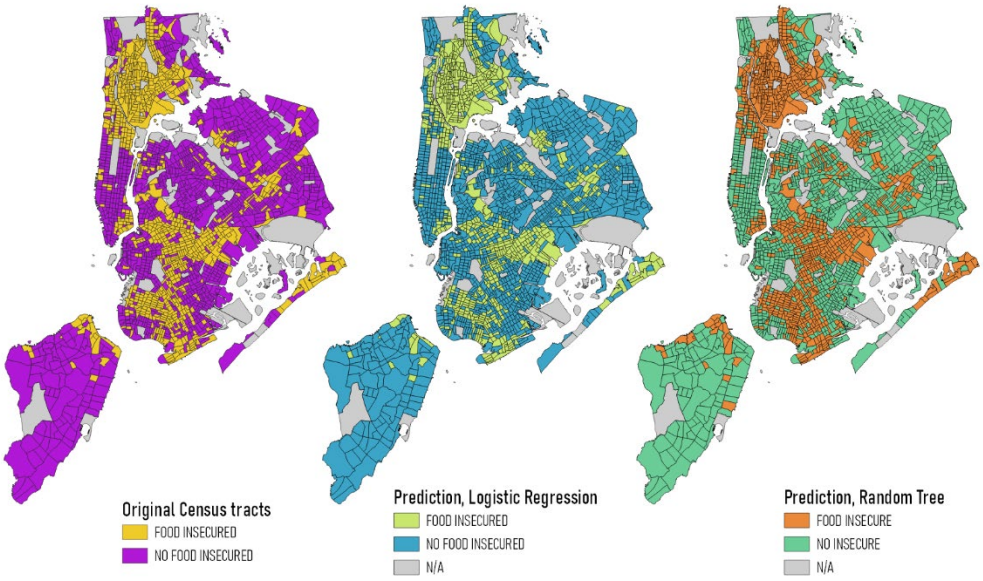
Policy debates around food environments and their impacts on health have been dominated by the notion that low-income neighborhoods of color are food deserts, because they lack large supermarkets and therefore may have a limited access to fresh, affordable, and healthy foods. Other authors argue that this

conceptualization is misleading and potentially detrimental to the health of poor communities because it ignores the contribution of smaller stores, particularly that of so-called ethnic markets. This conceptualization around food deserts reflects classes and racialized biases of foodscapes, by ignoring the day-to-day relationship with food in low-income communities, while favoring private corporate intervention. It is fundamental to look beyond just food availability as a predictor of food insecurity and really understand the reasons why income and race are so important predictors. There is a presumed "emptiness" and "vacancy" embedded in the understanding of food deserts. Advocates and policymakers outside of these neighborhood spaces often overlook or do not see the ways in which residents make their own ways to navigate food insecurity and reflect their hopes and desires for their communities more broadly.

The ROC depicts the relationship between True Positive Rate and False Positive Rate as the threshold is shifted. The graph on the right represents the final model. Here we can see that the final model has a higher AUC 96.7%. On the other hand, what the left graph above shows is that we are able to significantly increase True Positive Rate up to about 0.96 without sacrificing much in False Positive Rate. This indicates that the model can be tuned with the threshold to hit that point. However, depending on the goal of the model, whether optimizing for True Positive Rate or False Positive Rate is more important, the model can be tuned appropriately.



Finally, the following maps show the accuracy of the Logistic Regression and the Random Tree in predicting Food insecurity.



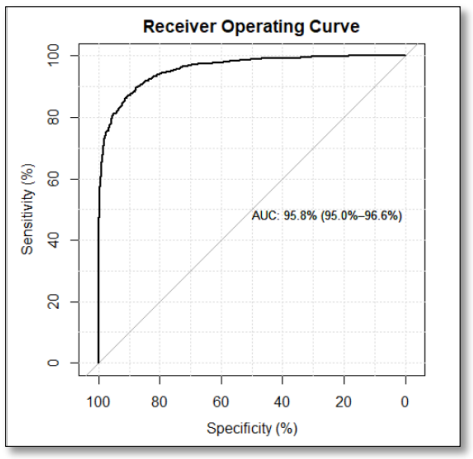
APPENDIX:

ECONOMIC MODEL

The economic model includes data from the 2014-2018 ACS. Five variables were included in this model, % people below poverty, %people with SNAPS, Median Household Income, %Unemployed, % no health insurance

MODEL - ROC CURVE

```
Call:
glm(formula = food2 ~ households + MHI + unemploy + unhealth,
    family = binomial, data = FOOD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7095  -0.3238  -0.0242   0.1706   3.4403

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.258e+00  6.057e-01   3.728 0.000193 ***
households   9.552e-02  1.119e-02   8.533  < 2e-16 ***
MHI         -1.056e-04  7.895e-06 -13.376  < 2e-16 ***
unemploy     5.561e-01  4.706e-02  11.817  < 2e-16 ***
unhealth    -5.641e-02  1.426e-02  -3.957  7.6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2858.32  on 2100  degrees of freedom
Residual deviance:  998.81  on 2096  degrees of freedom
AIC: 1008.8

Number of Fisher Scoring iterations: 7
```



DIAGNOSTIC

Pseudo R^2 for Logistic Regression		
Hotitanicer and Lemeshow R^2	**0.622**	
Cox and Snell R^2	**0.571**	
Nagelkerke R^2	**0.768**	

PREDICTION AND ACCURACY

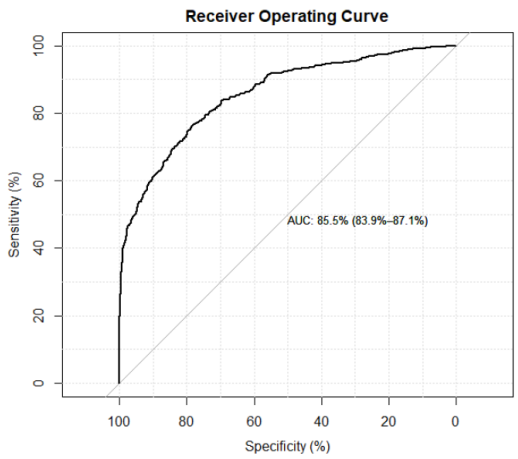
Prediction Economic characteristics		
	NOT INSECURED	INSECURED
INSECURED	104	753
NOT INSECURED	1115	129
Accuracy of the model	0.889100428367444	

DEMOGRAPHIC MODEL

The demographic model includes data from the 2014-2018 ACS. Seven variables were included in this model, % white people, % black people, % people with disability, % people over 65, % family with children, % latino, % people living alone. When I included single mothers, "latino" became significant.

MODEL - ROC CURVE

```
Call:
glm(formula = food2 ~ white + black + singlemom + disabled +
    sixtyfive + familychild + latino + lonely, family = binomial,
    data = FOOD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6007  -0.6981  -0.3886   0.5670   2.7370

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -4.273700   0.542786  -7.874  0.00000000000000344 ***
white       -0.017060   0.003850  -4.431  0.00000938077284964 ***
black       -0.019840   0.004038  -4.914  0.00000089252310997 ***
singlemom    0.097511   0.015856   6.150  0.00000000077611841 ***
disabled     0.308059   0.020781  14.824 < 0.0000000000000002 ***
sixtyfive   -0.138208   0.013791 -10.021 < 0.0000000000000002 ***
familychild  0.042500   0.009401   4.521  0.00000615720241323 ***
latino      -0.016381   0.004669  -3.508              0.000451 ***
lonely       0.063400   0.008110   7.817  0.00000000000000539 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2858.3  on 2100  degrees of freedom
Residual deviance: 1867.6  on 2092  degrees of freedom
AIC: 1885.6

Number of Fisher Scoring iterations: 5
```



DIAGNOSTIC

Pseudo R^2 for Logistic Regression		
Hotitanicer and Lemeshow R^2	**0.347**	
Cox and Snell R^2	**0.376**	
Nagelkerke R^2	**0.506**	

PREDICTION AND ACCURACY

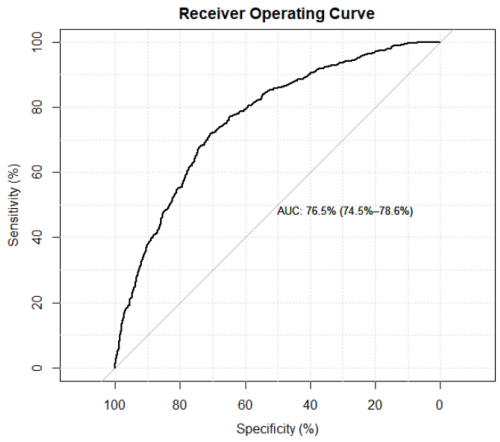
Prediction Economic characteristics		
	NOT INSECURED	INSECURED
INSECURED	160	579
NOT INSECURED	1059	303
Accuracy of the model	0.779628748215136	

HOUSING MODEL

The demographic model includes data from the 2014-2018 ACS. Five variables were included in this model, % high rent burden. % extreme rent burdened, % households without kitchen, % overcrowding, Number of food retails per Census tract.

MODEL - ROC CURVE

```
Call:
glm(formula = food2 ~ high + ex_high + no_kitc + more_one + RFT,
    family = binomial, data = FOOD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4879  -0.8817  -0.5669   1.0386   2.1036

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -3.946308   0.256419 -15.390 < 0.0000000000000002 ***
high         0.042258   0.006127   6.896      0.00000000000533 ***
ex_high      0.002654   0.006336   0.419                0.675
no_kitc     -0.016614   0.037492  -0.443                0.658
more_one     0.088819   0.008262  10.750 < 0.0000000000000002 ***
RFT          0.056737   0.008762   6.476      0.00000000009436 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2858.3  on 2100  degrees of freedom
Residual deviance: 2404.0  on 2095  degrees of freedom
AIC: 2416

Number of Fisher Scoring iterations: 4
```



DIAGNOSTIC

Pseudo R^2 for Logistic Regression		
Hotitanicer and Lemeshow R^2	**0.159**	
Cox and Snell R^2	**0.194**	
Nagelkerke R^2	**0.262**	

PREDICTION AND ACCURACY

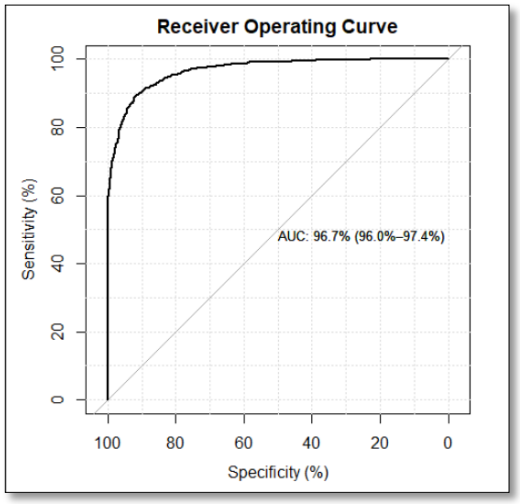
Prediction Economic characteristics		
	NOT INSECURED	INSECURED
INSECURED	236	484
NOT INSECURED	983	398
Accuracy of the model	0.698238933841028	

FINAL MODEL

The Final model was composed out of the significant variables of the Economic, Demographic and Housing models that include data from the 2014-2018 ACS. The demographic model includes data from the 2014-2018 ACS. Thirteen variables were included in the first try and after dropping 4 non-significant variables (% No health insurance, %people with disability, % people over 65, % people living alone). The final model includes nine variables were included in this model, and % people below poverty, %people with SNAPS, Median Household Income, %Unemployed, % no health insurance, % white people, % black people, % latino,% single mom with children% people with disability, % people over 65, % family with children, % high rent burden. % households without kitchen, % overcrowding.

MODEL - ROC CURVE

```
Call:
glm(formula = food2 ~ MHI + unemploy + singlemom + black + latino +
    disabled + familychild + high + more_one, family = binomial,
    data = FOOD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4983  -0.2761  -0.0124   0.1943   3.7484

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.771e+00  9.745e-01   6.949 3.69e-12 ***
MHI         -1.441e-04  9.028e-06 -15.960  < 2e-16 ***
unemploy     7.021e-01  5.352e-02  13.118  < 2e-16 ***
singlemom    8.099e-02  2.380e-02   3.403 0.000667 ***
black       -1.577e-02  4.612e-03  -3.419 0.000628 ***
latino      -3.736e-02  5.486e-03  -6.810 9.76e-12 ***
disabled     1.430e-01  2.454e-02   5.828 5.62e-09 ***
familychild -4.817e-02  1.178e-02  -4.090 4.31e-05 ***
high        -3.551e-02  9.357e-03  -3.795 0.000148 ***
more_one     1.064e-01  1.632e-02   6.520 7.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2858.32  on 2100  degrees of freedom
Residual deviance:  959.05  on 2091  degrees of freedom
AIC: 979.05
```



DIAGNOSTIC

Pseudo R^2 for Logistic Regression		
Hotitanicer and Lemeshow R^2	**0.872**	
Cox and Snell R^2	**0.695**	
Nagelkerke R^2	**0.934**	

PREDICTION AND ACCURACY

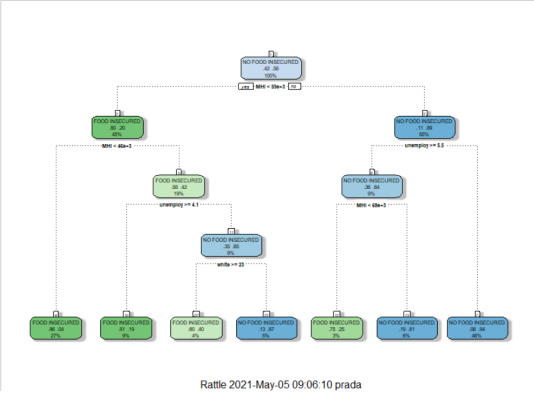
Prediction Economic characteristics (0.3)		
	NOT INSECURED	INSECURED
INSECURED	183	822

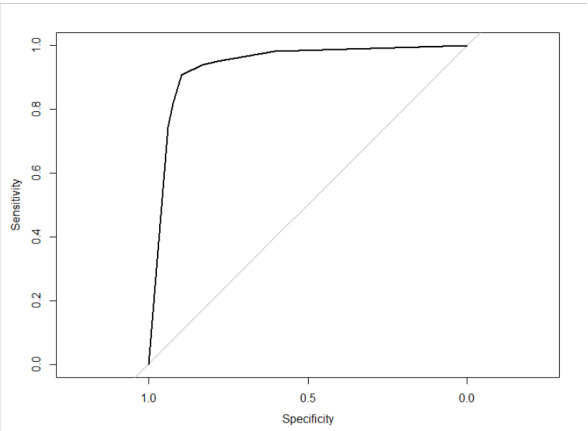
NOT INSECURED	1036	60
Accuracy of the model	0.884340790099952"	
Prediction Economic characteristics (0.5)		
	NOT INSECURED	INSECURED
INSECURED	93	775
NOT INSECURED	1126	107
Accuracy of the model	0.904807234650167	
Prediction Economic characteristics (0.7)		
	NOT INSECURED	INSECURED
INSECURED	45	700
NOT INSECURED	1174	182
Accuracy of the model	0.89195621132793	

RANDOM TREE

The final model includes nine variables were included in this model, and % people below poverty, %people with SNAPS, Median Household Income, %Unemployed, % no health insurance, % white people, % black people, % people with disability, % people over 65, % family with children, % high rent burden. % households without kitchen, % overcrowding.

STARTER TREE – 4 NODES



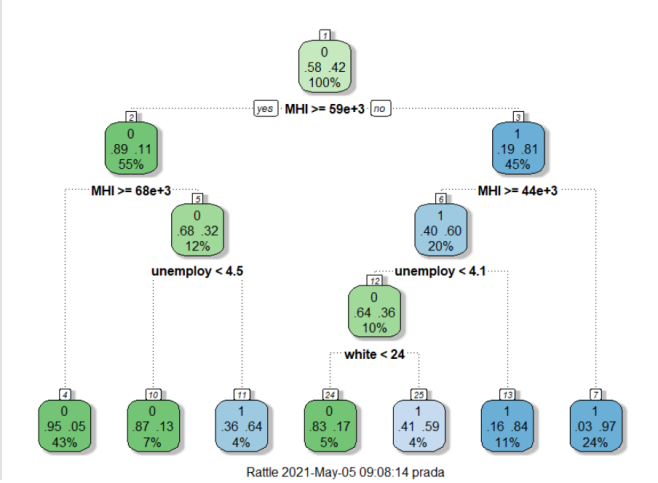


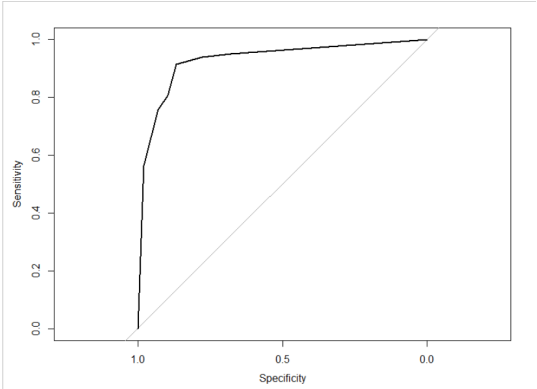
AUC = 0.9349

PREDICTION AND ACCURACY

Prediction Economic characteristics		
	NOT INSECURED	INSECURED
INSECURED	114	790
NOT INSECURED	1105	92
Accuracy of the model	Accuracy: 0.864348162291169	

TREE TRAIN OBSERVATIONS 75%





AUC = 0.9275

Prediction Economic characteristics		
	NOT INSECURED	INSECURED
INSECURED	211	15
SECURED	32	161
Accuracy of the model	Accuracy: 0.887828162291169	

- **For the confusion matrix had to change from numeric to integer food2**

Accuracy : 0.9403
95% CI : (0.9132, 0.961)
No Information Rate : 0.58
P-Value [Acc > NIR] : <0.0000000000000002

Kappa : 0.8769

Mcnemar's Test P-Value : 0.2301

Sensitivity : 0.9091
Specificity : 0.9630
Pos Pred Value : 0.9467
Neg Pred Value : 0.9360
Prevalence : 0.4200
Detection Rate : 0.3819
Detection Prevalence : 0.4033
Balanced Accuracy : 0.9360

'Positive' Class : 1