MODELING FOOD AND HOUSING INSECURITY

by

Francis Biney, Ebenezer Nkum and Tolulope Adeyina

Contents

1 Executive Summary

This project set out to find which subpopulations of University of Texas at El Paso (UTEP) students are most at risk for food and housing insecurity. The data set used is a direct survey constructed to assess thet state of food and housing insecurity among UTEP students by a team of researchers (Moya, Crouse, Schober and Wagler). Each student participants record consists of 39 variables, containing sociodemographic data (variables 1-19) and data about food and housing status (variable 20-37). There were $5,175$ total participants, used for this project. Among the variables collected, variables; **In the past 12 months, were you ever hungry but didn't eat because there wasn't enough money for food?** and **In the last 12 months, since (today's date),did you ever cut the size of your meals or skip meals because there was not enough money for food?** were used to assess those at risk for food insecurity, and variables; **In the past 12 months, have you had a permanent address?** and **Due to lack of permanent address or housing options, how frequently did you spend the night elsewhere in the past six months due to lack of permanent housing?** were used to assess those at risk for housing insecurity. We would like to mention that, the variables that were used to assess the housing insecurity the former suffered from imbalance proportion of representation and the latter suffered from lack of representation. However models used have some tendency to correct these imbalances to a certain degree.

2 Main Report

Our analysis reveals that debt increase is the strongest single predictor of food insecurity among UTEP students. The other predictors that are most statistically significant are:

- Age
- Total income (less $10,000$)
- Federal Aid in the past 12 month (about the same)
- Income change (about the same)
- Any benefit support (Supplemental Nutrition Assistance Program)
- College (Liberal Arts)
- Hours work per week (19 hrs or less)
- Expenditure change (about the same)

These factors have 76.91% accuracy of identifying students who are at risk for food insecurity with sensitivity of 27.5% and specificity of 93.33%. Intuitively, these factors classify subpopulation of students who have their debt increased, with total annual income below $10,000$, having their expenditure decreased and head of household is 30% at risk of food insecurity.
Also, the analysis reveals that students staying on-campus or off-campus not with family is the strongest predictor of housing insecurity among UTEP students. The other predictors that are most statistically significant are:

- Age
- Mode of transport (Bike and walk)
- Total Income (below $10,000$ anually)
- Income change (About the same)
- College (of Education)
- Hours work per week (19 hrs or less)
- federal Aid in the past six month (Work study)

These factors have 95.44% accuracy of identifying students who are at risk for housing insecurity with sensitivity of 3.3% and specificity of 99.99%. Intuitively, these factors classify subpopulation of students who have stay on-campus or off-campus not with family with age less 27, with mode of transport (carpool, Bus, Trolley, Bike, Walk) and federal aid about the same to at risk for housing insecurity.
These finding is actionable and it was not obvious in advance. A student with increased debt presumably will be at risk for food insecurity.

3 Recommendation

- The major challenge is correct the data representation. Correct wording of questions to get the right response should be considered.
- We recommend that this subpopulation identified here should be closely studied to validate the findings

4 Detail Analysis

4.1 Introduction

In this project, we seek to find factors that are associated with food insecurity and housing insecurity among University of Texas at El Paso (UTEP) students. We further examined subpopulations of students who are at most risk for food and housing Insecurity. We employed the well-know models of logistic regression, logistic regression with LASSO regularization, random forest and classification tree and choose the best model with highest accuracy. In recent years, there has been an increasing awareness among educators about the prevalence of food and housing insecurities in college and university settings. Largely, this is due to the efforts of scholars such as (Gupton 2014) and (Goldrick-Rab, Broton, and Eisenberg 2015) (among others), who have succeeded in raising awareness about extreme cases of insecurities, such as hunger and homelessness. As the United States simultaneously endures a historic pandemic and an economic recession, many college students are having trouble accessing basic needs. A recent survey from the Hope Center for College, Community, and Justice,(Goldrick-Rab et al. 2020) found that more than half of students are experiencing food insecurity, housing insecurity, or homelessness. In addition, more than two-thirds of students lost a job or suffered cuts to pay or hours, and many have been unable to get financial assistance from their campus or the federal government. One of the survey's most troubling findings is that students of color—especially Black students, Pacific Islander or Native Hawaiian students, and Indigenous students—are being disproportionately affected. We seek to find what may be specific among UTEP students.

4.2 Data

The data come from an electronic survey completed by 5449 students attending UTEP at all levels. There are 5449 rows each representing responses of the participants. There are 37 distinct variables (Questions) for the survey. There are 36 categorcal variables and only one numerical variable (Age). Out of the 37 distinct variables, there two variables that can be used as a response variable for housing insecurity model and four variables that can be used a response variable for the food insecurity model. The main challenge in processing this data set was the large number of missing values in the response variables.

4.3 Data Preprocessing and Feature Engineering

The original data contain multiple columns for some specific questions. We collapse them into one column with their respective coding. The original data contain $7,087$ respondents who received an ID but responded to no question. We deleted all of them. We also deleted all respondents who dropped after answering their 5th question. We removed irrelevant variables

with no predictive power from the data. We corrected inconsistency values in the **Age** variable. The final data before the impute had $5,175$ rows and 30 columns.

Missing Values Treatment

- We imputed two variables with missing rate less 50% with modal class.
- We imputed the rest of missing values with random forest algorithm.
- Even though one response variable for the housing insecurity model has missing rate of 95.09%, we decided to fit a sub-model with the data present (Down sampling).

4.4 Data Exploration

With data distribution, approximately 84% of those who participated in the survey were full students, with 16% being part time students. 42.53% of the participants are working and the 57.47% are not working. Hispanics/Latins make about 76% of the entire respondents.

5 Analysis plan

5.1 Model Development

We fitted logistic regression (both multinomial and binomial), logistic regression with LASSO regularization and random forest to assess the risk factors housing insecurity. We obtained the variable important and determine the highest predictor of the response variable.

The logistic regression model is widely used in the social and biological sciences. The model is especially useful is demographic research in the assessment of the effects of the explanatory factors on the relative risk of outcomes. In this case, the logistic model will provide the probability of a particular outcome occurring. It supports categorizing data into discrete classes by studying the relationship from a given set of labeled data. It is easier to implement and makes no assumption about the distributions of the classes un feature space.

The second model we explore is the LASSO regularization of the logistic regression. We need regularization to introduce bias to the model and to decrease the variance. This method will set regression coefficients for irrelevant variables to zero. This provides a system for selecting important variables but it does not necessarily provide a way to rank them.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way.(Breiman 2001)

In the project we have 3 questions to answer:

- Which factors are associated with housing Insecurity ?
- Which factors are associated with food insecurity?
- Which subpopulations are most at risk for Food and housing Insecurity?

Housing Insecurity With the housing insecurity, we have two response variable to help us identify the category of the students sample that are at the risk of housing insecurity. The responses we explored were:

- Due to lack of permanent address or housing options, how frequently did you spend the night elsewhere in the past six months due to lack of permanent housing?
- In the past 12 months, have you had a permanent address?

The first response variable has three (3) categorical variable (often, sometimes and rarely) and the second response has 2 categorical variable (Yes, No). However we had 95.09% missing rate so we decided to fit a sub-model with the data present.

Food Insecurity

Here we have 4 possible response variables to explore. We examined 2 of them to determine which factors are associated with food insecurity.

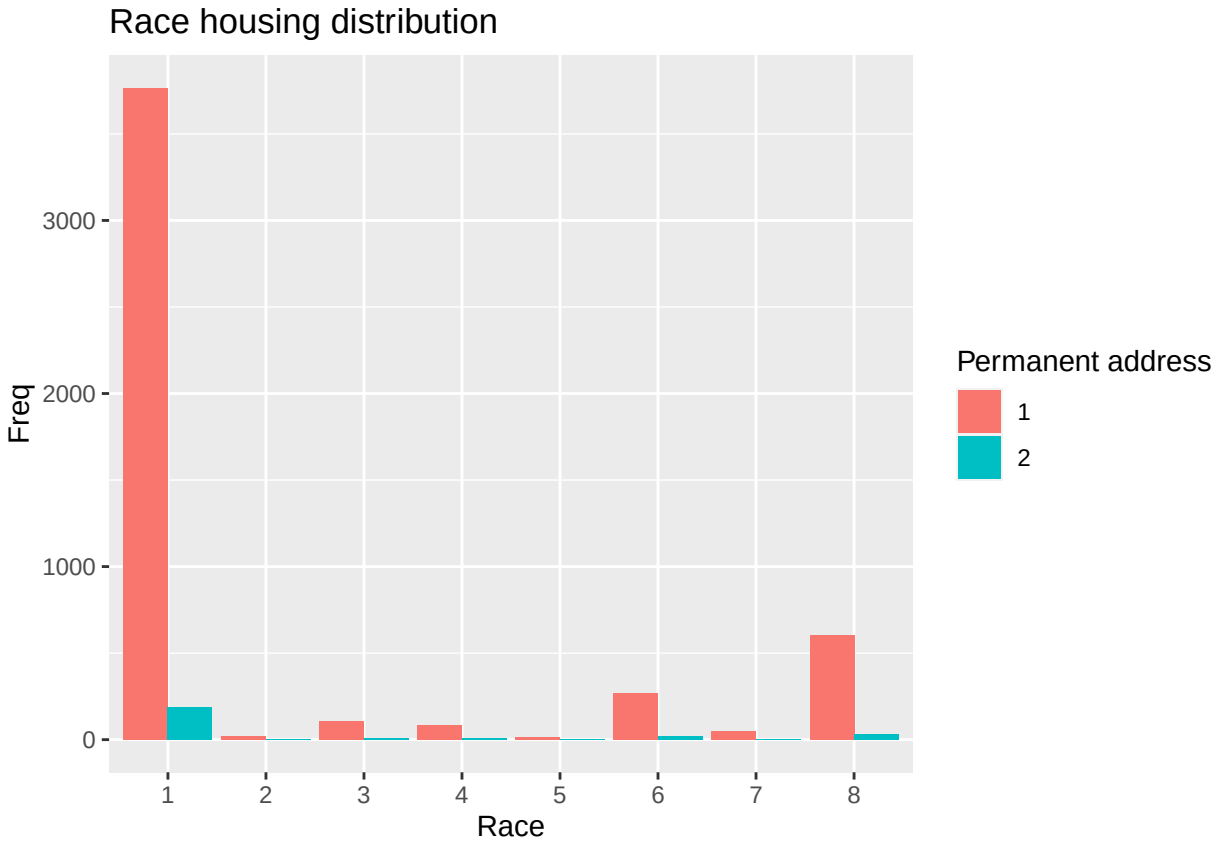
The responses we explored were:

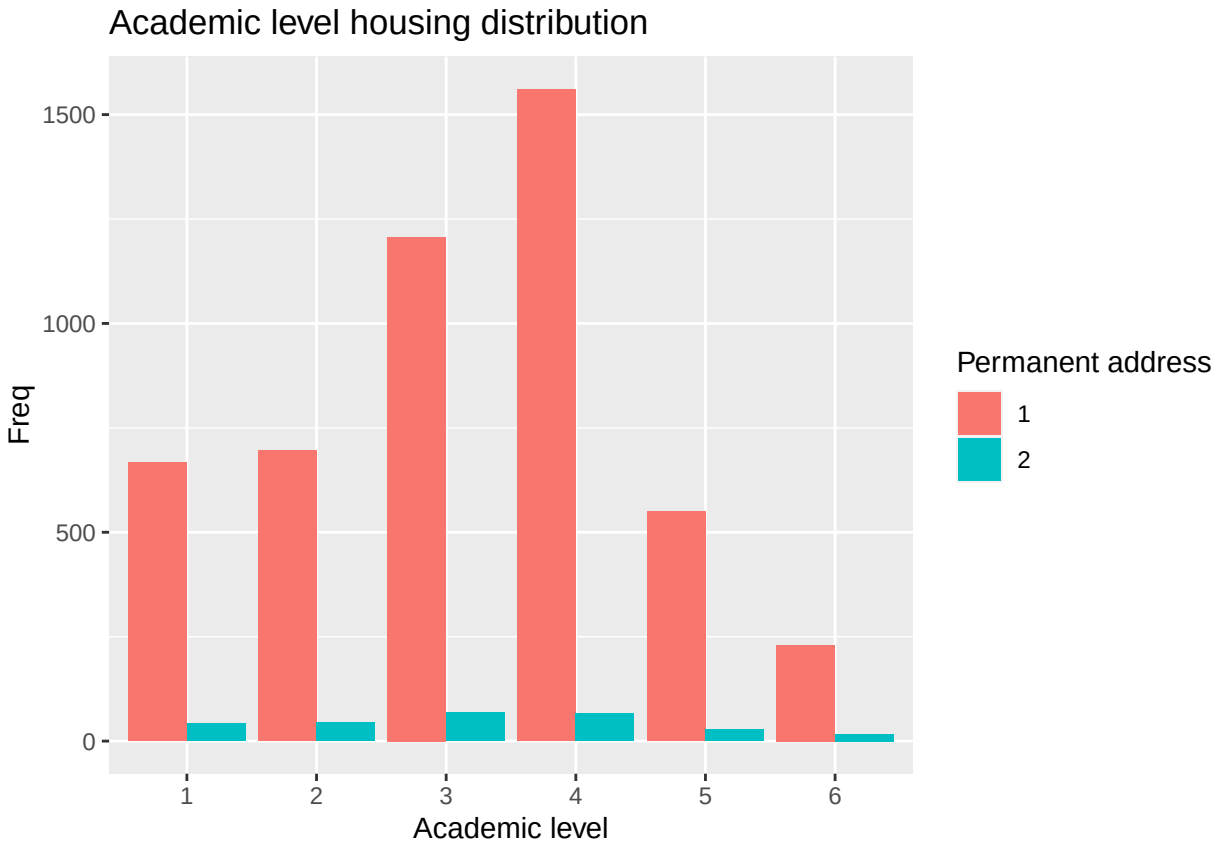
- Q28. In the last 12 months, since (today's date), did you ever cut the size of your meals or skip meals because there was not enough money for food?
- Q31. In the past 12 months, were you ever hungry but didn't eat because there wasn't enough money for food?

Each of these have two categorical variable. We report the best model with highest accuracy. We also check the consistency of the variable importance with the different response. We performed some exploration on the first 3 variable of importance and provided recommendation.

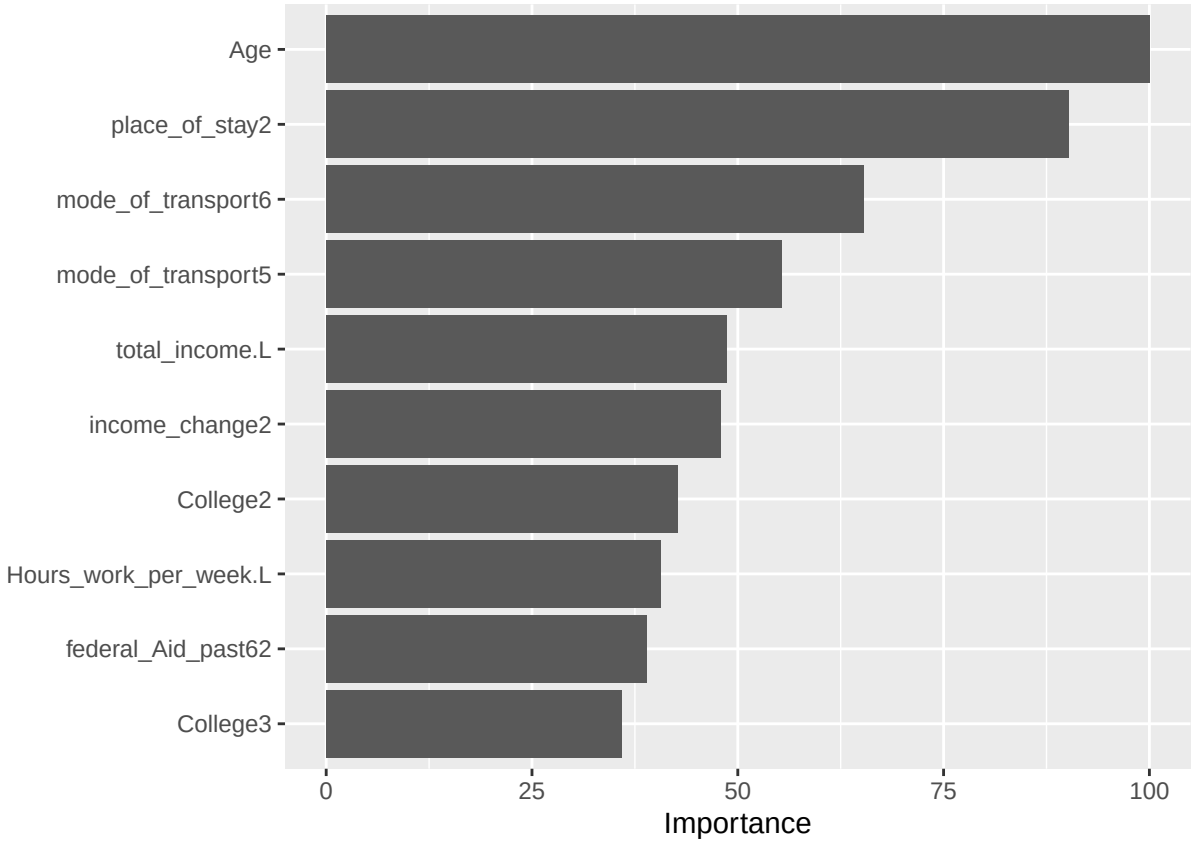
6 Analysis and Results

6.1 Housing Insecurity model




Academic level housing distribution
Freq
1500
1000
500
0
1
2
3
4
5
6
Academic level
Permanent address
1
2

6.2 Housing Insecurity mode (l) : First response variable (Permanent address past 12 month)




```
## Generalized Linear Model
##
## 5175 samples
##   23 predictor
##    2 classes: '1', '0'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4657, 4657, 4658, 4658, 4658, 4658, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9472473  0.01087074

## Confusion Matrix and Statistics
```

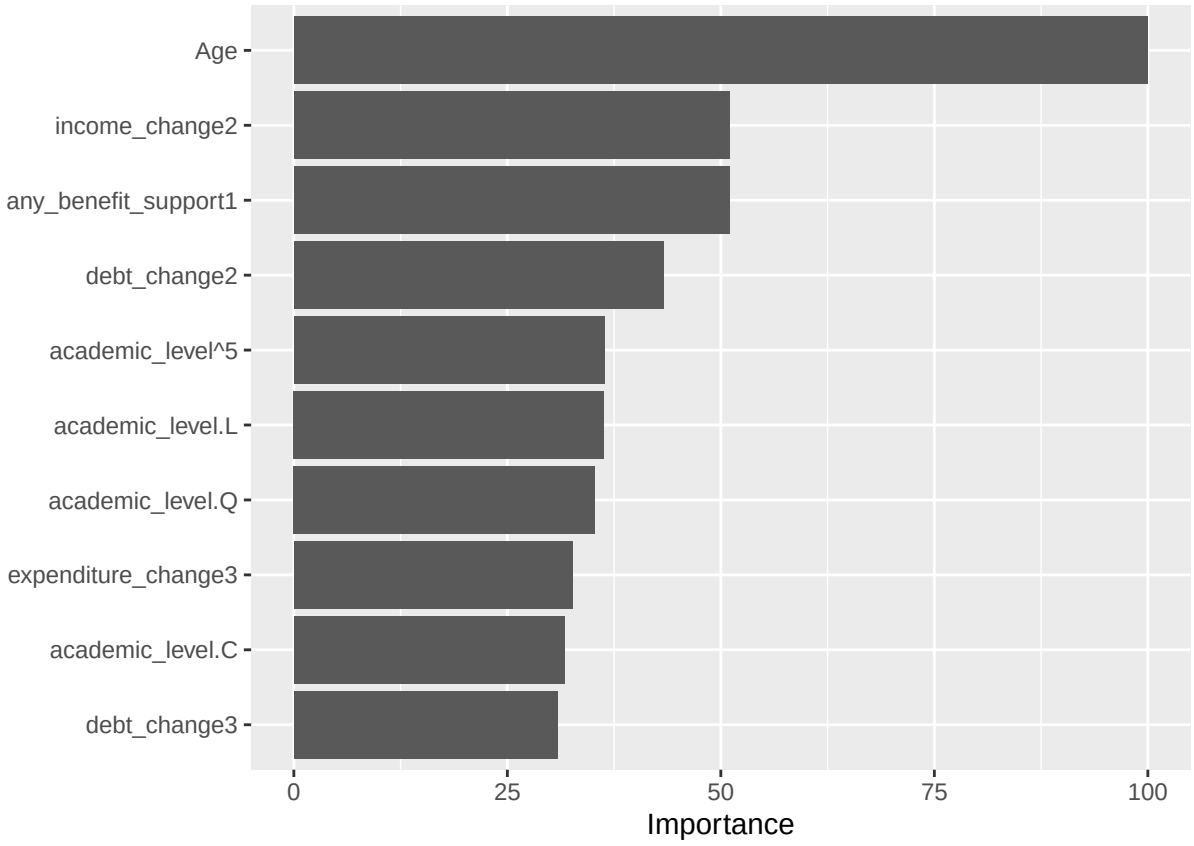
```
##
##           Reference
## Prediction    1    0
##          1 1233   58
##          0    1    2
##
##                Accuracy : 0.9544
##                  95% CI : (0.9416, 0.9651)
##     No Information Rate : 0.9536
##     P-Value [Acc > NIR] : 0.4816
##
##                   Kappa : 0.0593
##
##  Mcnemar's Test P-Value : 3.086e-13
##
##             Sensitivity : 0.99919
##             Specificity : 0.03333
##          Pos Pred Value : 0.95507
##          Neg Pred Value : 0.66667
##               Precision : 0.95507
##                  Recall : 0.99919
##                      F1 : 0.97663
##              Prevalence : 0.95363
##          Detection Rate : 0.95286
##    Detection Prevalence : 0.99768
##       Balanced Accuracy : 0.51626
##
##        'Positive' Class : 1
##
```

Binomial logistic regression gave the highest accuracy. We use 10-fold cross validation to estimate the model parameters. The results are reported below: We report the confusion matrix and variable of importance for the model. At an accuracy of 95.44%; Age, place of stay, mode of transport and total income were found to be the first 4 important variables.

6.3 Housing Insecurity model : Second response variable (Spent the night elsewhere)

6.4 Housing Insecurity model (2)

```
## Random Forest
##
## 189 samples
##  23 predictor
##   3 classes: '1', '2', '3'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 170, 171, 170, 170, 171, 170, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    1    0.5452924  0.0000000
##    2    0.5667446  0.0799735
##    3    0.5704483  0.1196426
##    4    0.5734795  0.1403227
##    5    0.5653314  0.1306708
##    6    0.5647173  0.1391030
##    7    0.5666374  0.1493743
##    8    0.5737622  0.1652209
##    9    0.5681287  0.1564291
##   10    0.5633138  0.1497457
##   11    0.5664620  0.1617438
##   12    0.5638791  0.1552982
##   13    0.5784893  0.1816277
##   14    0.5670175  0.1606598
##   15    0.5769201  0.1889712
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 13.
```




```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3
##          1 30 16  8
##          2  5  5  0
##          3  0  0  1
##
## Overall Statistics
##
##                Accuracy : 0.5538
##                  95% CI : (0.4253, 0.6773)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 0.4518
##
##                   Kappa : 0.1092
##
```

```
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.8571  0.23810  0.11111
## Specificity            0.2000  0.88636  1.00000
## Pos Pred Value         0.5556  0.50000  1.00000
## Neg Pred Value         0.5455  0.70909  0.87500
## Precision              0.5556  0.50000  1.00000
## Recall                 0.8571  0.23810  0.11111
## F1                     0.6742  0.32258  0.20000
## Prevalence             0.5385  0.32308  0.13846
## Detection Rate         0.4615  0.07692  0.01538
## Detection Prevalence   0.8308  0.15385  0.01538
## Balanced Accuracy      0.5286  0.56223  0.55556
```

Random forest gave the highest accuracy for prediction. At an accuracy of 55.38%; Age, income change, benefit support and debt change were found to be first 4 variable of importance.

7 Food Insecurity model

7.1 Using the response variable : In the past 12 months, were you ever hungry but didn't eat because

there wasn't enough money for food?

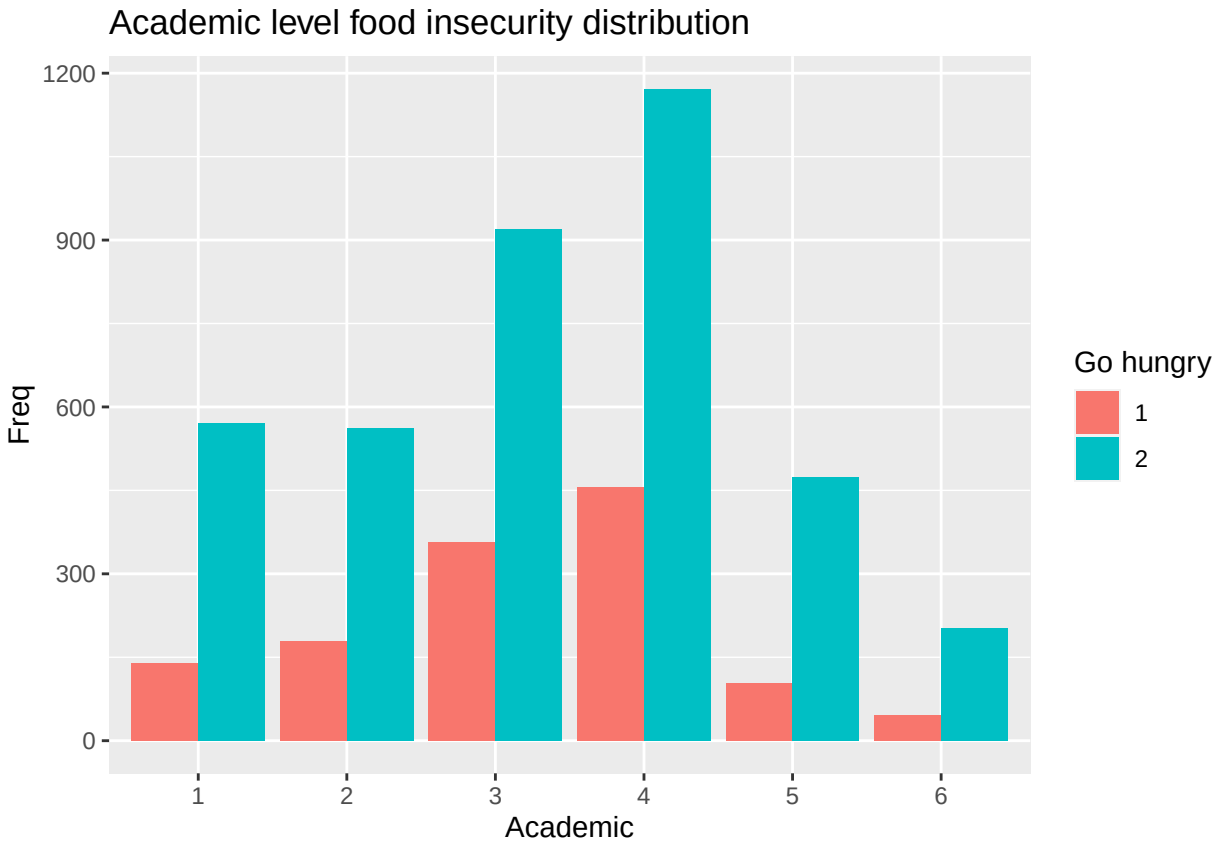
```
# Academic level food insecurity distribution



c = as.data.frame(table(df_food_new_1$go_hungry,df_food_new_1$academic_level))

p <- ggplot(data = c, aes(x=Var2, y= Freq, fill=Var1)) +
     geom_bar(stat="identity", position=position_dodge())
```

```
print(p  + labs(title= "Academic level food insecurity distribution",
                      x="Academic", fill = "Go hungry"))
```

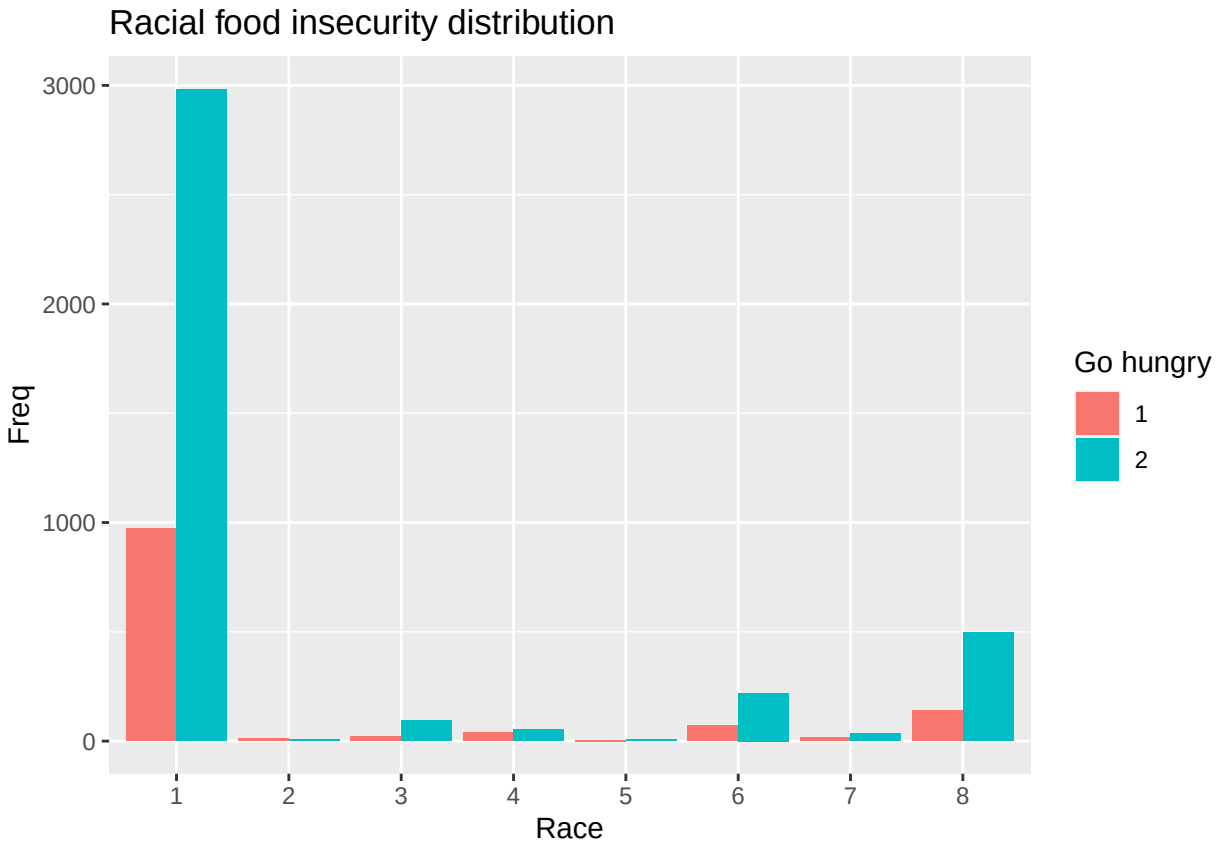


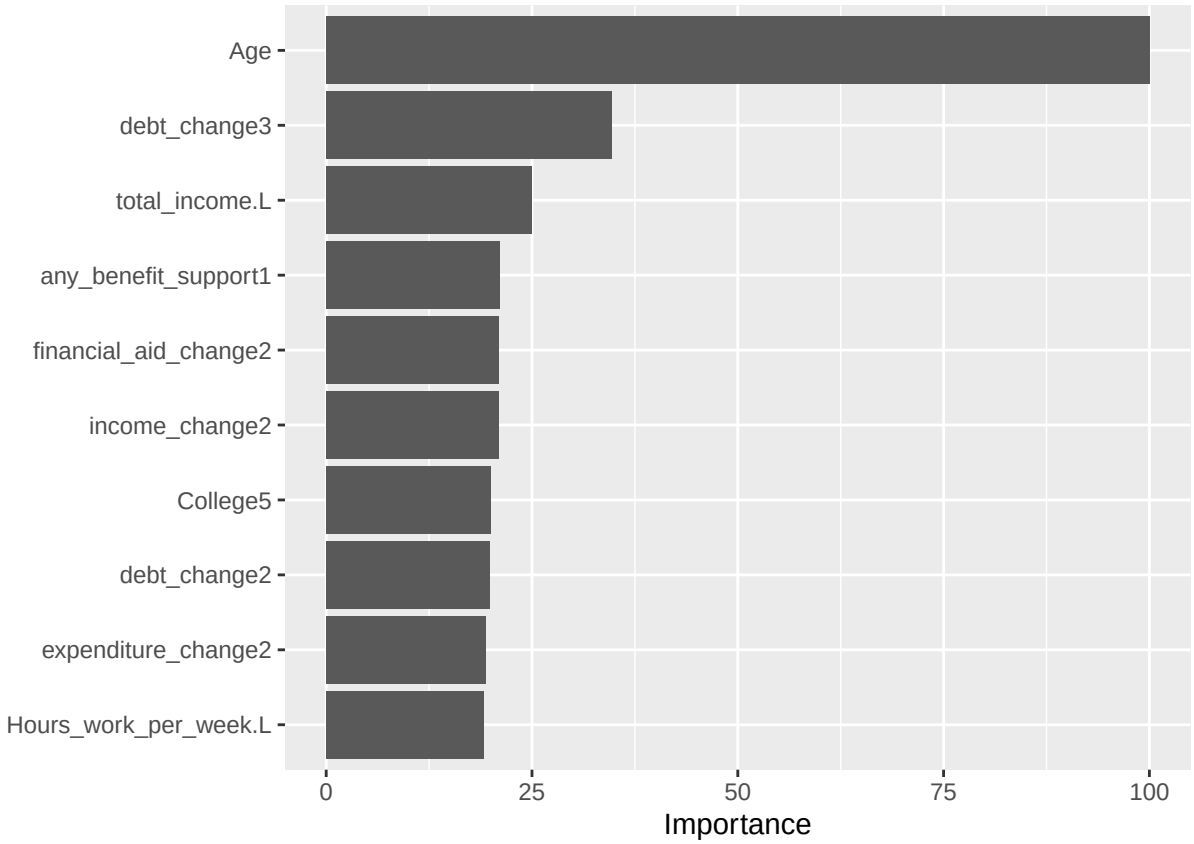

```
# Racial housing distribution

c = as.data.frame(table(df_food_new_1$go_hungry,df_food_new_1$Race))

p <- ggplot(data = c, aes(x=Var2, y= Freq, fill=Var1)) +
     geom_bar(stat="identity", position=position_dodge())


print(p  + labs(title= "Racial food insecurity distribution",
                      x="Race", fill = "Go hungry"))
```


Racial food insecurity distribution
3000
2000
1000
0
Freq
1
2
3
4
5
6
7
8
Race
Go hungry
1
2

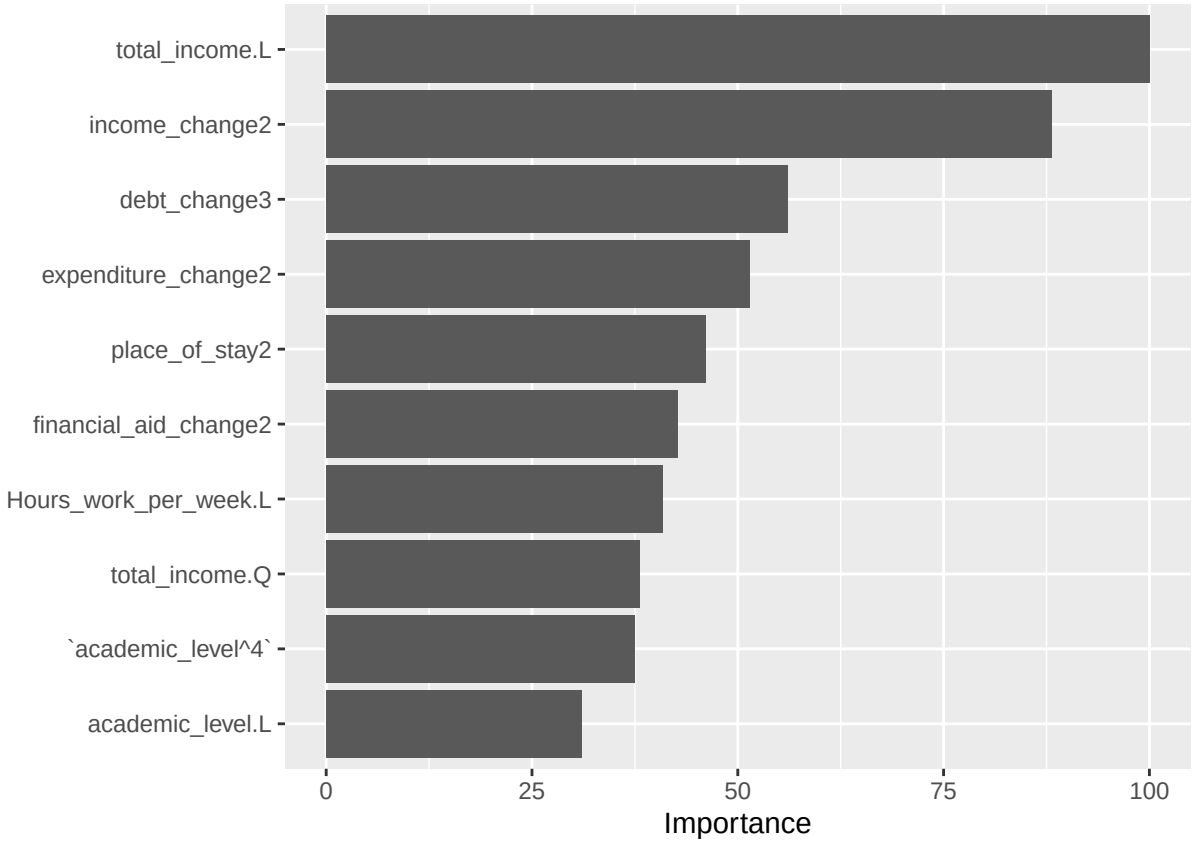



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1  81  60
##          0 239 915
##
##                Accuracy : 0.7691
##                  95% CI : (0.7452, 0.7918)
##     No Information Rate : 0.7529
##     P-Value [Acc > NIR] : 0.09259
##
##                   Kappa : 0.2359
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.25312
```

```
##             Specificity : 0.93846
##          Pos Pred Value : 0.57447
##          Neg Pred Value : 0.79289
##               Precision : 0.57447
##                  Recall : 0.25312
##                      F1 : 0.35141
##              Prevalence : 0.24710
##          Detection Rate : 0.06255
##    Detection Prevalence : 0.10888
##       Balanced Accuracy : 0.59579
##
##        'Positive' Class : 1
##
```

Random forest gave the highest accuracy for prediction. At an accuracy of 76.91%; Age, debt change, total income and benefit support were found to be first 4 variable of importance.

7.2 Using the response variable : eat_less



```
## Generalized Linear Model
##
## 3880 samples
##   21 predictor
##    2 classes: '1', '0'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3492, 3491, 3492, 3492, 3492, 3493, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7103231  0.2849311
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1 170 117
##          0 252 756
##
##                Accuracy : 0.7151
##                  95% CI : (0.6896, 0.7395)
##     No Information Rate : 0.6741
##     P-Value [Acc > NIR] : 0.0008295
##
##                   Kappa : 0.293
##
##  Mcnemar's Test P-Value : 3.042e-12
##
##             Sensitivity : 0.4028
##             Specificity : 0.8660
##          Pos Pred Value : 0.5923
##          Neg Pred Value : 0.7500
##               Precision : 0.5923
##                  Recall : 0.4028
##                      F1 : 0.4795
```

```
##              Prevalence : 0.3259
##          Detection Rate : 0.1313
##    Detection Prevalence : 0.2216
##       Balanced Accuracy : 0.6344
##
##        'Positive' Class : 1
##
```

Binomial logistic regression gave the highest accuracy. We use 10-fold cross validation to estimate the model parameters. The results are reported below: We report the confusion matrix and variable of importance for the model. At an accuracy of 71.51%; total income , income change, debt change , and expenditure change were found to be the first 4 important variables.

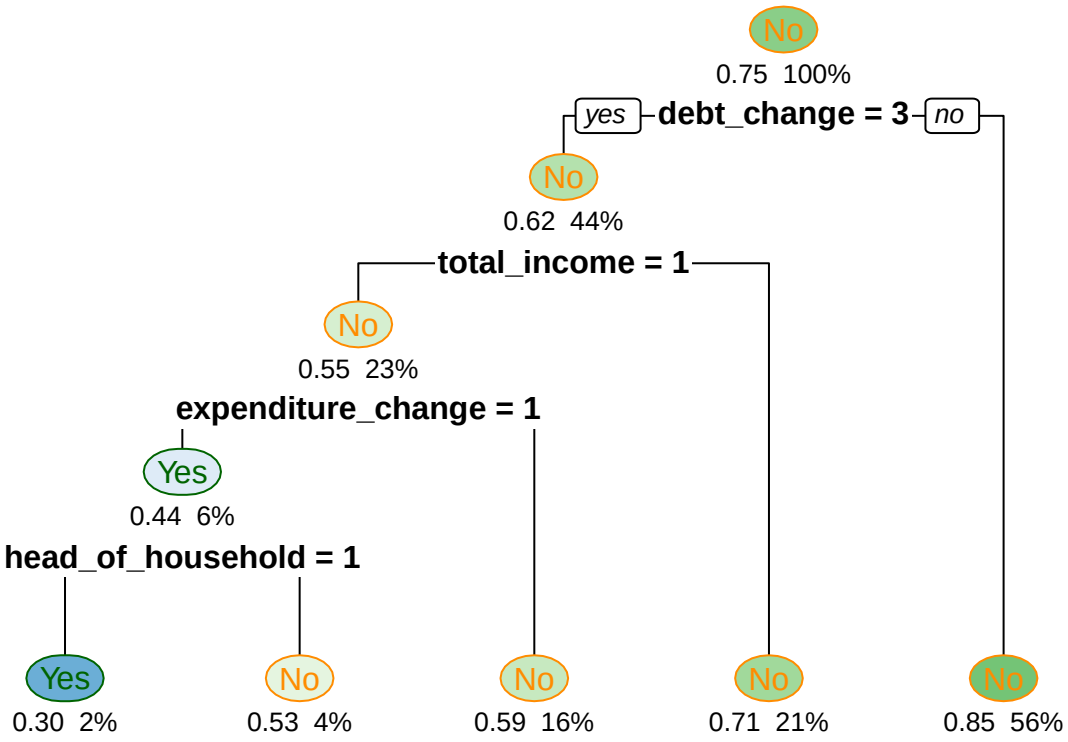
8 Subpopulation at risk for food and housing insecurity.

8.1 Food Insecurity

The most important predictors with response, go hungry because there was no money for by the highest performing model with accuracy of 76.91% are

- Age
- Debt Change
- Total income
- Federal Aid in the past 12 month
- Income change
- Any benefit support
- College
- Hours work per week
- Expenditure change

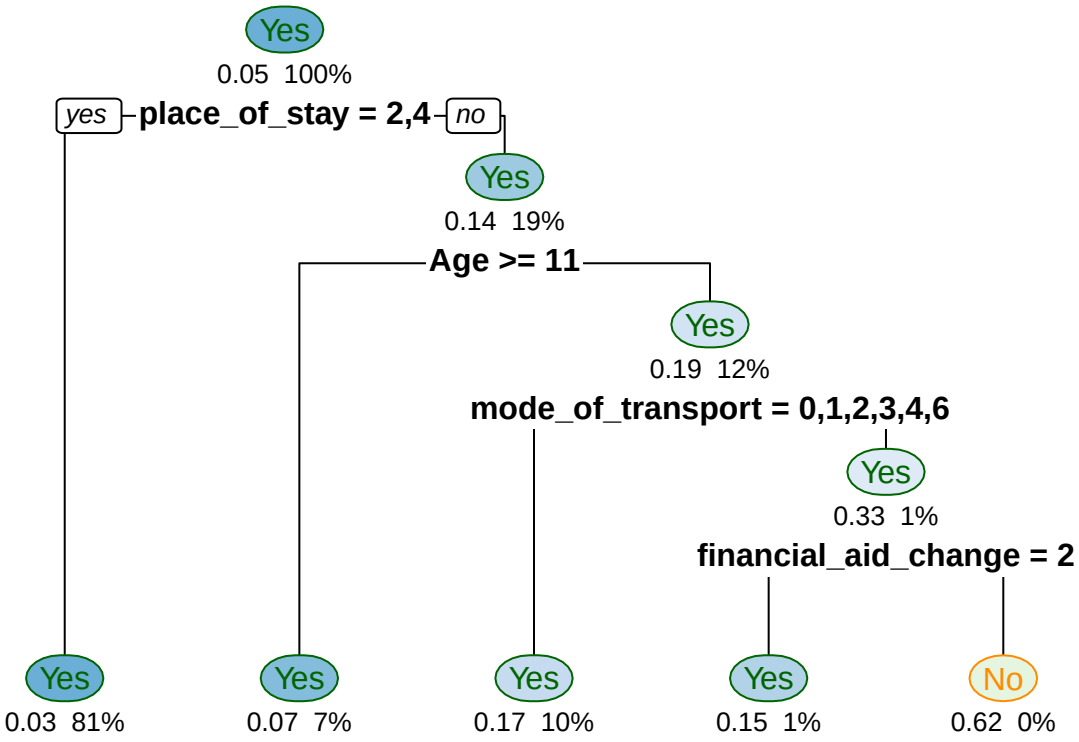
```
df_food_tree <- df_food2      #df_food2[, c(5,22,9,20,17,16,11,4,19,18)]
```



8.2 Housing Insecurity

The most important predictors with response variable, **permanent address** by the correct predicting model with accuracy of 95.44% are:

- Age
- Place of stay
- mode of transportation
- Total income
- Income change
- College
- Hours work per week
- federal Aid past 12 month



9 Discussion

9.1 Housing Insecurity model

Three models were built for the housing insecurity response: Q22:"frequency of night elsewhere in the past six months due to lack of permanent". Q22 has three categories: rarely, sometimes, and often.

The models consider are Multinomial logistic regression(MLR), Multinomial logistic regression with LASSO penalty (MLR-LASSO) and Random Forest (RF). 10-fold cross validation was used to select the optimal hyper-parameter. Random forest gave the best prediction accuracy of 0.585 with confidence interval(CI) of (0.456, 0.706) followed by MLR-LASSO with prediction accuracy of 55.38% with 95% CI : $(0.4253, 0.6773)$. The top five risk factor for each response from RF's variable important plots are Age(Q5), any_benefit_support(Q24),income_change(Q33), depth_change(Q35) and experience_homelessness(Q13)

9.2 Food Insecurity model

We built three models for three food insecurity response and determined the top 5 feature or risk factors that explains these response. The response considered are Q28: "In the last 12 months, since (today's date), did you ever cut the size of your meals or skip meals because there was not enough money for food?" and Q31:"Were you ever hungry but didn't eat because the wasn't enough money for food".
The models consider in each case are Multinomial logistic regression(MLR), binomial logistic regression with LASSO penalty (MLR-LASSO) and Random Forest (RF).
For the first response (Q26) the model that gave the best prediction was binomial logistic regression with prediction accuracy of 71.51% and 95% confidence interval of $(0.6896, 0.7395)$. The best model for Q31 was Random Forest with prediction accuracy of 76.91% and confidence interval of $(0.7452, 0.7918)$. The top five risk factor for each response form the estimated coefficients and variable important plots for response Q26 are Total_income (Q9), Place_of_stay (Q19), Race (Q6),Reliability_of_transportation (Q13) and experience_homelessness(Q23). For Q28 the risk factor are depth_change(Q35), experience homelessness(Q23), income_change(Q33) and reliability_of_transportation(Q13). Risk factors for Q31 are Total income (Q9), experience_homelessness(Q13), Race (Q6), depth_change(Q35) and income_change(Q33).

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Goldrick-Rab, Sara, Katherine Broton, and Daniel Eisenberg. 2015. "Hungry to Learn: Addressing Food and Housing Insecurity Among Undergraduates." *Wisconsin Hope Lab*, 1–25.

Goldrick-Rab, Sara, Vanessa Coca, Gregory Kienzl, Carrie R Welton, Sonja Dahl, and Sarah Magnelia. 2020. "# RealCollege During the Pandemic: New Evidence on Basic Needs Insecurity and Student Well-Being."

Gupton, Jarrett T. 2014. "Engaging Homeless Students in College." In *Student Engagement in Higher Education*, 237–52. Routledge.