MODELING FOOD AND HOUSING INSECURITY

by

Francis Biney, Ebenezer Nkum and Tolulope Adeyina

Contents

1 Abstract

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.

2 Introduction

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.

3 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.

3.1 Data Preprocessing and Feature Engineering

The original data contain multiple columns for one question. We collapse them into one column with their respective coding. The original data contain 7087 respondents who received an ID but responded to no question. We deleted all of them. We also deleted all respondents who dropped after the 5th question. We removed irrelevant variables with no predictive variable from the data. We corrected inconsistency values in the **Age** variable. The final data before the impute had 5175 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.
- 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.

3.2 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.

4 Analysis plan

4.1 Model Development

We explored logistic regression (both multinomial and binomial), logistic regression with LASSO regularization and random forest to assess the risk factors housing insecurity. We will obtain 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 will examine 3 of them to determine which factors are associated with food insecurity.
The responses we explored were:

- Q26. "The food that I bought just didn’t last, and I didn’t have money to get more." Was that often, sometimes, or never true for you in the last 12 months ?
- 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 will report the best model with highest accuracy. We will 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.

5 Analysis and Results

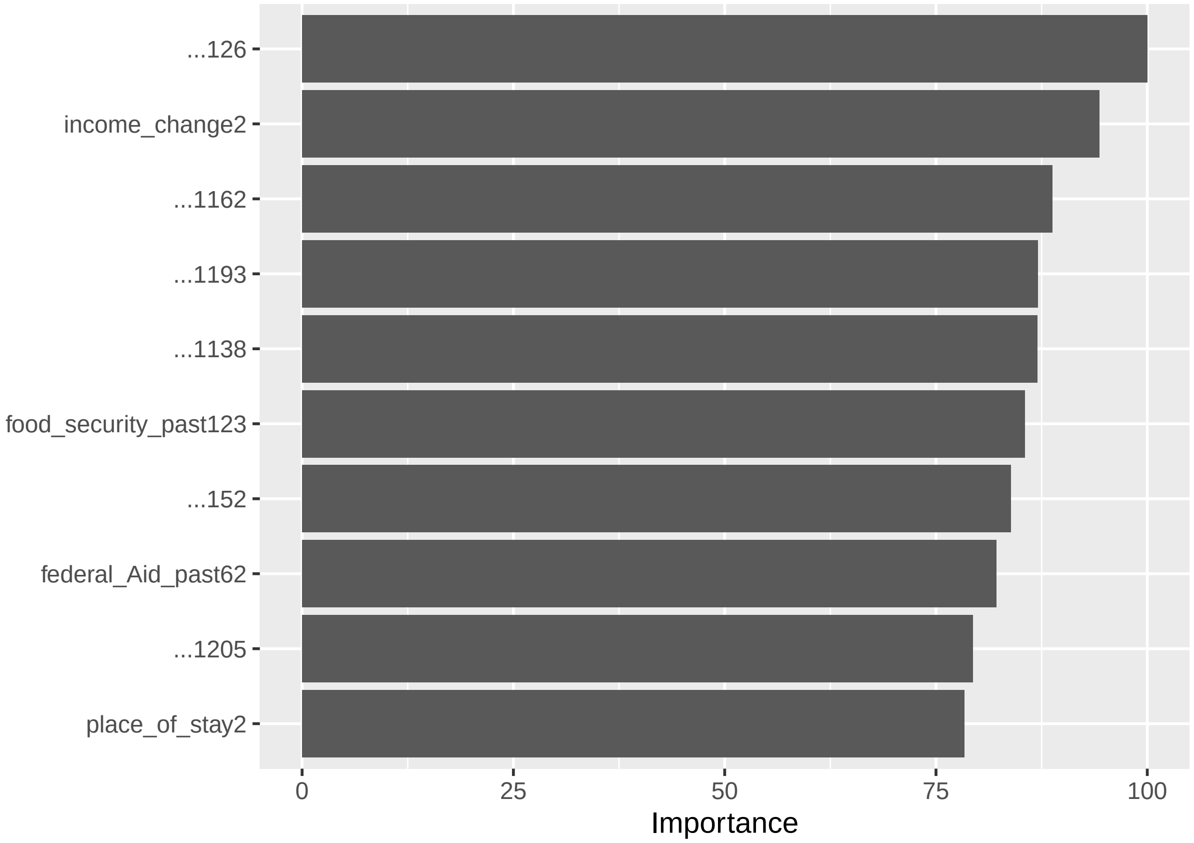
5.1 Housing Insecurity model(1)

```
## New names:
## * `` -> ...1

## Rows: 254 Columns: 31

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (31): ...1, Enrollment, Employ_Status, Place_of_work, Hours_work_per_wee...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

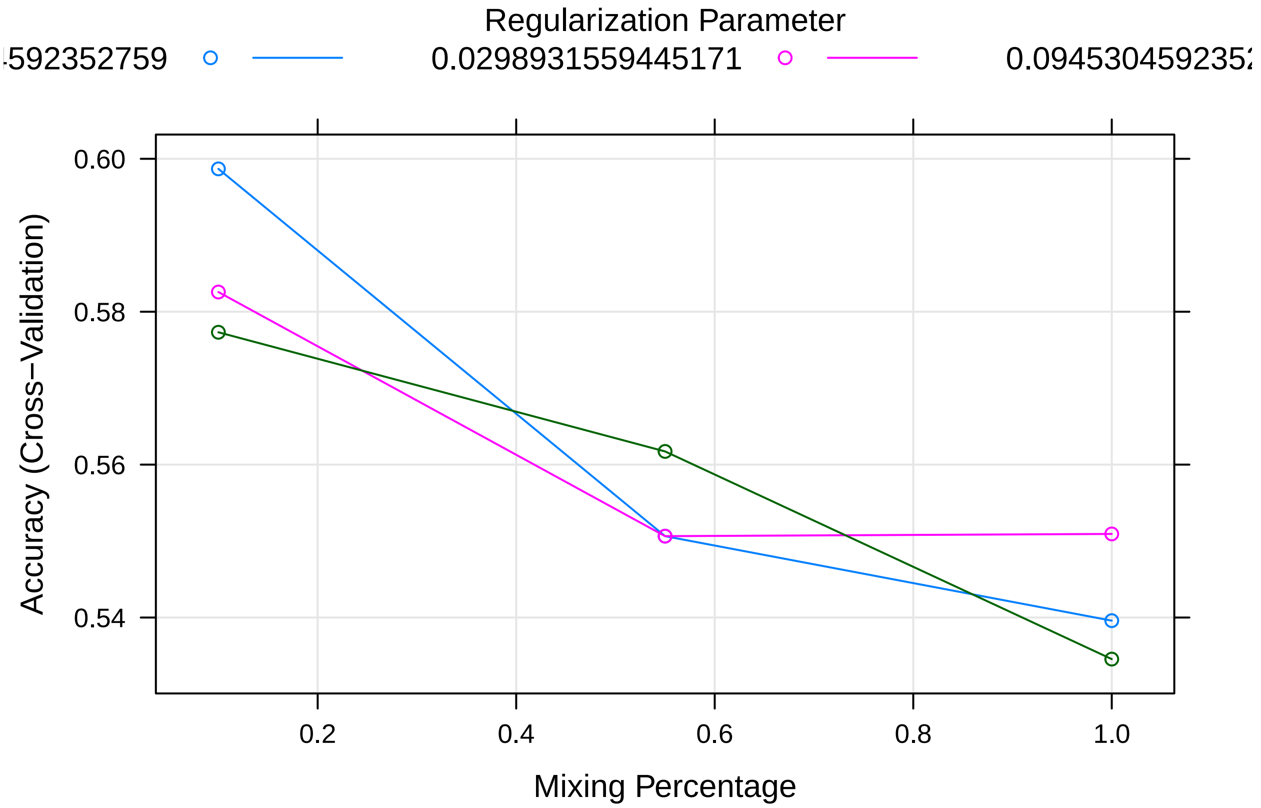


```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction  1  2  3
##          1 23 15  5
##          2  9  6  3
##          3  3  0  1
##
## Overall Statistics
##
##                Accuracy : 0.4615
##                  95% CI : (0.337, 0.5897)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 0.9142
##
##                   Kappa : 0.0134
##
##  Mcnemar's Test P-Value : 0.1718
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.6571  0.28571  0.11111
## Specificity            0.3333  0.72727  0.94643
## Pos Pred Value         0.5349  0.33333  0.25000
## Neg Pred Value         0.4545  0.68085  0.86885
## Precision              0.5349  0.33333  0.25000
## Recall                 0.6571  0.28571  0.11111
## F1                     0.5897  0.30769  0.15385
## Prevalence             0.5385  0.32308  0.13846
## Detection Rate         0.3538  0.09231  0.01538
## Detection Prevalence   0.6615  0.27692  0.06154
## Balanced Accuracy      0.4952  0.50649  0.52877
```

Here, we fit multinomial logistic regression. 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 1.1. Also their plots are provided

5.2 Housing Insecurity model(2)



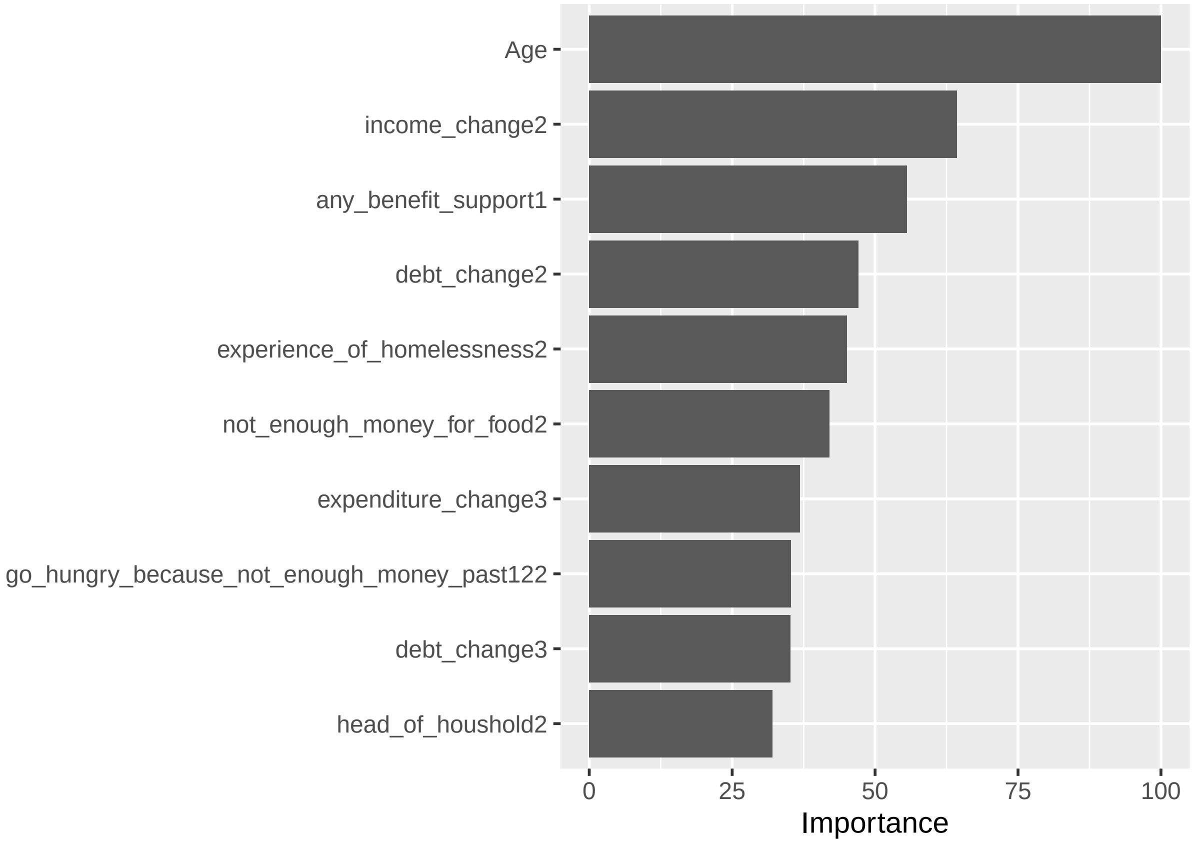
```
## glmnet variable importance
##
##   variables are sorted by maximum importance across the classes
##   only 20 most important variables shown (out of 322)
##
##                1       2      3
## ...1162 2.4652 0.63195 4.6287
## ...152  2.9166 0.09998 4.5482
## ...1205 2.2637 0.64047 4.4358
## ...140  2.8036 0.06377 4.3990
## ...126  1.5278 1.16245 4.2218
## ...170  1.5815 0.98282 4.0959
## ...179  1.5960 0.89974 4.0273
## ...151  1.5717 0.88125 3.9846
## ...1207 1.3553 1.09068 3.9776
## ...1242 0.9797 1.40570 3.9170
## ...156  1.2369 1.05470 3.8232
## ...1164 1.4610 0.68781 3.6804
## ...195  0.8621 1.25690 3.6506
## ...1175 0.4091 1.67918 3.6199
## ...1193 2.0876 3.42156 0.0000
## ...1146 2.2557 3.19065 0.0000
## ...1138 1.7765 3.18797 0.0000
## ...185  1.2636 3.09660 0.3014
## ...1251 0.6240 0.88127 3.0368
## ...198  2.4164 3.02717 0.0000
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3
##          1 32 18  7
##          2  3  3  1
##          3  0  0  1
##
## Overall Statistics
##
##                Accuracy : 0.5538
##                  95% CI : (0.4253, 0.6773)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 0.4518143
##
##                   Kappa : 0.0911
##
##  Mcnemar's Test P-Value : 0.0003132
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.9143  0.14286  0.11111
## Specificity            0.1667  0.90909  1.00000
## Pos Pred Value         0.5614  0.42857  1.00000
## Neg Pred Value         0.6250  0.68966  0.87500
## Precision             0.5614  0.42857  1.00000
## Recall                0.9143  0.14286  0.11111
## F1                    0.6957  0.21429  0.20000
## Prevalence            0.5385  0.32308  0.13846
## Detection Rate        0.4923  0.04615  0.01538
## Detection Prevalence  0.8769  0.10769  0.01538
## Balanced Accuracy     0.5405  0.52597  0.55556
```

Here, we fi multinomial logistic regression with LASSO regularization. We use 10-fold cross validation to estimate the model parameters. The results are reported below:

5.3 Housing Insecurity model(3)

```
## Random Forest
##
## 189 samples
##  29 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  0.0000000000
##    2    0.5452924  0.0000000000
##    3    0.5452924  0.0000000000
##    4    0.5452924  0.0000000000
##    5    0.5452924  0.0009009009
##    6    0.5452924  0.0033799245
##    7    0.5612963  0.0523824512
##    8    0.5577680  0.0581387530
##    9    0.5772125  0.1085636689
##   10    0.5645224  0.0903382048
##   11    0.5802437  0.1329084960
##   12    0.5913548  0.1647789427
##   13    0.5862671  0.1573193717
##   14    0.5933138  0.1750936399
##   15    0.5895224  0.1741799535
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 14.
```



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3
##          1 34 19  8
##          2  1  2  1
##          3  0  0  0
##
## Overall Statistics
##
##                Accuracy : 0.5538
##                  95% CI : (0.4253, 0.6773)
##     No Information Rate : 0.5385
##     P-Value [Acc > NIR] : 0.4518
##
##                   Kappa : 0.0603
##
##  Mcnemar's Test P-Value : 1.402e-05
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3
## Sensitivity            0.9714  0.09524   0.0000
## Specificity            0.1000  0.95455   1.0000
## Pos Pred Value         0.5574  0.50000      NaN
## Neg Pred Value         0.7500  0.68852   0.8615
## Precision              0.5574  0.50000       NA
## Recall                 0.9714  0.09524   0.0000
## F1                     0.7083  0.16000       NA
## Prevalence             0.5385  0.32308   0.1385
## Detection Rate         0.5231  0.03077   0.0000
## Detection Prevalence   0.9385  0.06154   0.0000
## Balanced Accuracy      0.5357  0.52489   0.5000
```

Here, we fit random forest. We report the confusion matrix and variable of importance for the model.

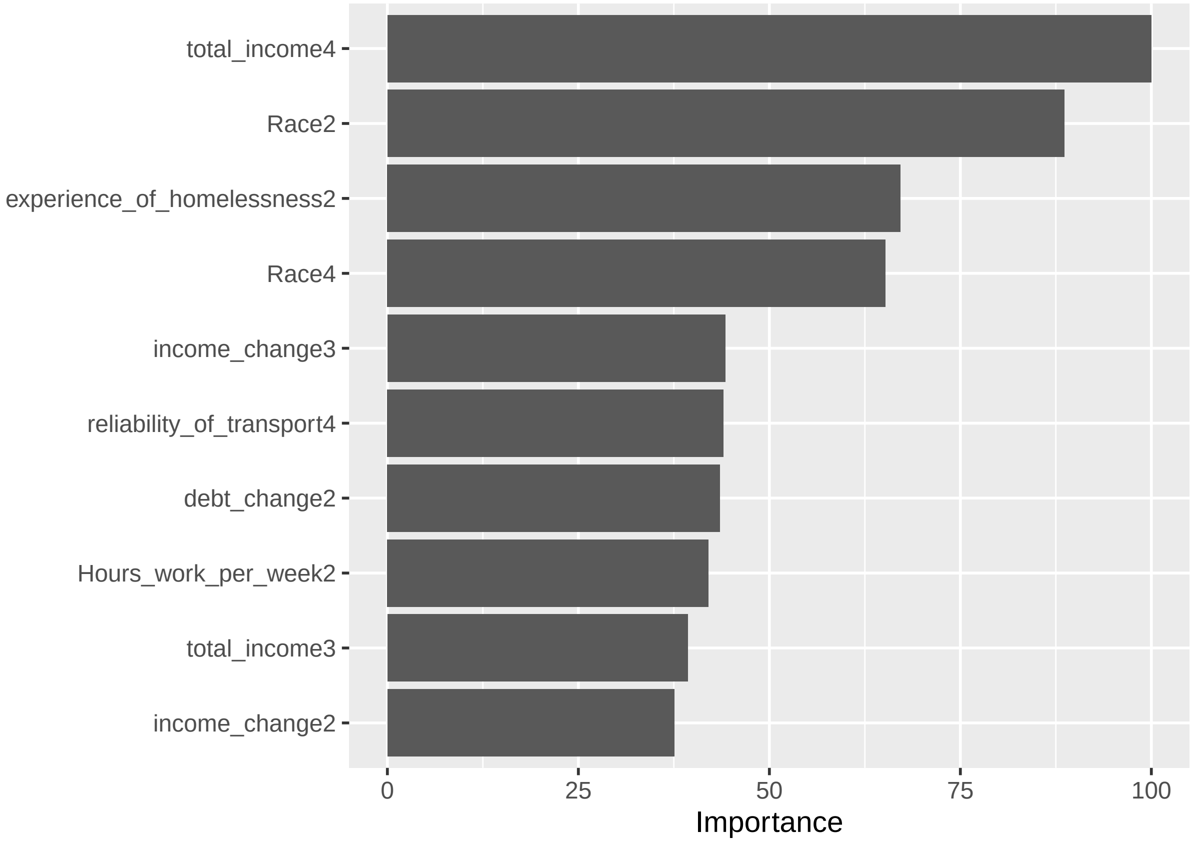
6 Food Insecurity model

We fitted 4 models but we select the best one with the highest accuracy

The response that was selected for the food insecurity model is:
Q31. In the past 12 months, were you ever hungry but didn't eat because there wasn't enough money for food?

6.1 Food Insecurity model(1)

```
## Rows: 5175 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (30): Enrollment, Employ_Status, Place_of_work, Hours_work_per_week, Rac...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

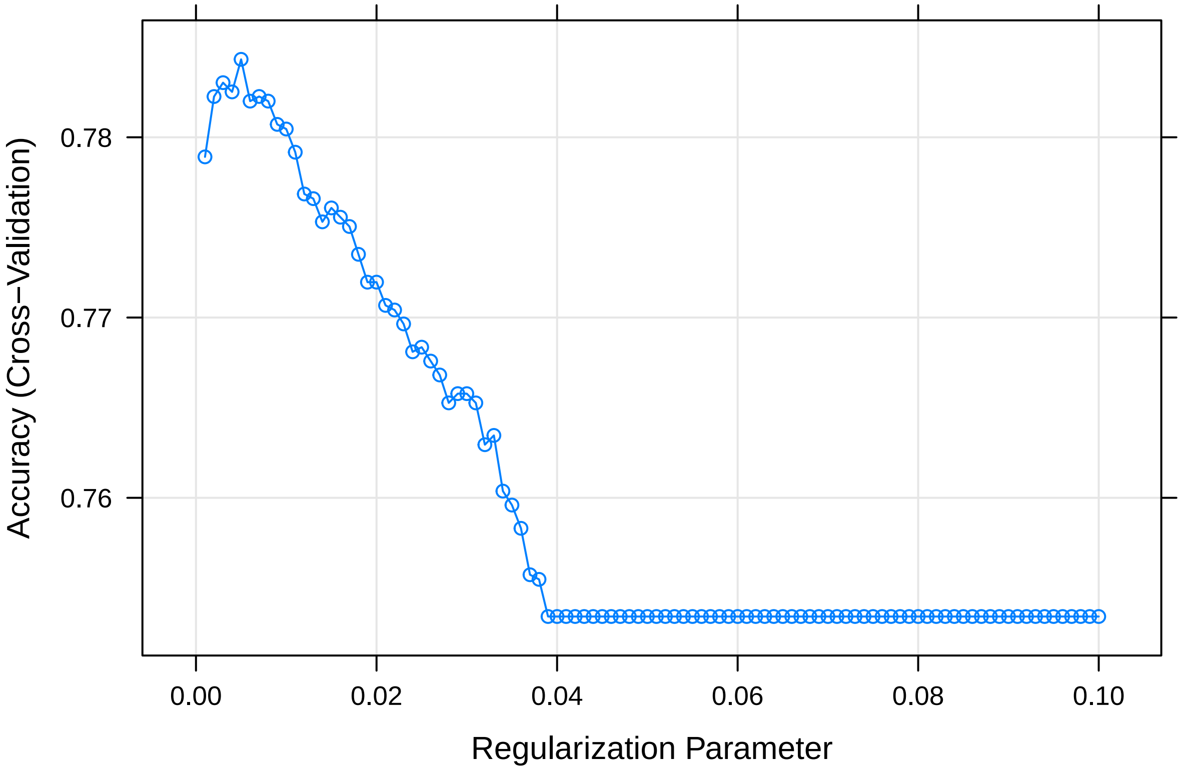


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   2
##          1 109  80
##          2 210 895
##
##                Accuracy : 0.7759
##                  95% CI : (0.7522, 0.7983)
##     No Information Rate : 0.7535
##     P-Value [Acc > NIR] : 0.03201
```

```
##
##                   Kappa : 0.3009
##
##  Mcnemar's Test P-Value : 3.587e-14
##
##             Sensitivity : 0.34169
##             Specificity : 0.91795
##          Pos Pred Value : 0.57672
##          Neg Pred Value : 0.80995
##               Precision : 0.57672
##                  Recall : 0.34169
##                      F1 : 0.42913
##              Prevalence : 0.24652
##          Detection Rate : 0.08423
##    Detection Prevalence : 0.14606
##       Balanced Accuracy : 0.62982
##
##        'Positive' Class : 1
##
```

Here, we fit multinomial logistic regression. 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 1.1. Also their plots are provided

6.2 Food Insecurity model(2)



```
## glmnet variable importance
##
##   only 20 most important variables shown (out of 66)
##
##                              Overall
## total_income4                 1.1482
## experience_of_homelessness2   0.9062
## Race4                         0.6711
## Race2                         0.6584
## debt_change2                  0.5805
## income_change2                0.5130
## permanent_address_past122     0.4617
## total_income3                 0.4333
## reliability_of_transport4     0.4324
## income_change3                0.4013
## Hours_work_per_week2          0.3962
## academic_level5              0.3440
## head_of_houshold2            0.3399
## federal_Aid_past64           0.3192
## reliability_of_transport2    0.3078
## expenditure_change2          0.2731
## College5                     0.2552
## academic_level3              0.2267
## Race7                        0.2244
## total_income2                0.2209
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   1   2
##          1  93  58
##          2 226 917
## 
##                Accuracy : 0.7805
##                  95% CI : (0.757, 0.8028)
##     No Information Rate : 0.7535
##     P-Value [Acc > NIR] : 0.01228
## 
##                   Kappa : 0.282
## 
##  Mcnemar's Test P-Value : < 2e-16
## 
##             Sensitivity : 0.29154
##             Specificity : 0.94051
##          Pos Pred Value : 0.61589
##          Neg Pred Value : 0.80227
##               Precision : 0.61589
##                  Recall : 0.29154
##                      F1 : 0.39574
##              Prevalence : 0.24652
##          Detection Rate : 0.07187
##    Detection Prevalence : 0.11669
##       Balanced Accuracy : 0.61602
```

```
##
##        'Positive' Class : 1
##
```

Here, we fi multinomial logistic regression with LASSO regularization. We use 10-fold cross validation to estimate the model parameters. The results are reported below:

6.3 Food Insecurity model(3)

Here, we fit random forest. We report the confusion matrix and variable of importance for the model.

7 Discussion

7.1 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: really, 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 0.523 with 95% CI : (0.395, 0.646). 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)

7.2 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 Q26: "The food that I bought just didn't last, and I didn't have money to get more", 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), Multinomial logistic regression with LASSO penalty (MLR-LASSO) and Random Forest (RF). For the fist response Q26 the model that gave th best prediction was MLR withe prediction accuracy of

0.648 and 95% confidence interval of (0.621, 0.674). The best model for Q2 was RF with prediction accuracy of (0.7203, 0.769) and confidence interval of (0.720, 0.769). For Q3 the best model was MLR-LASSO with prediction accuracy of 0.780 and confidence interval of (0.757, 0.8028).

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.