Modeling Food and Housing Insecurity at UTEP

Phase II Report

George E. Quaye   John Koomson   Willliam O. Agyapong

University of Texas, El Paso (UTEP)

Department of Mathematical Sciences

Contents

1   Discussion

It was clearly noticed from our analysis that given each best model, predicting whether one is food insecure or housing insecure for the six responses, the factors or variables such as total income, transportation reliability, type of residence, specifically residence off-campus with family, whether or not one is the head of the household and the number of dependents on the person were part of the major contributing factors. However, we made a weird observation, which is among the subgroups at risk for food insecurity factors, total income for people earning more than $100,000 were occurring in all four responses, which analytical and socially does not make sense. In general, total income being an underlying factor for both food insecurity and housing insecurity, clearly suggests that the income on which students live for their livelihood appears not to be enough.

Considering the factors affecting housing insecurity responses, permanent house address and spending the night elsewhere, both were found to be associated with transport reliability, household head, the college one comes from, and whether the person is living alone, and total income as the most important variables among others. That is to say, the chances of one being housing insecure rely on one's total income, one being a household head with more dependents, and how reliable their transportation is with regards to one's income and expenditure. Also considering which subgroups or populations are at risk of being housing insecure, factors including having a residence off-campus with family, coming from either a college of education or college of science, and having a very reliable transportation mode. Specifically, it appears that if a student belongs to the college of science or college of education and stays off-campus with the family coupled with having a reliable common mode of transportation is at risk at a point in time of housing insecurity should that person is being deprived of such benefits.

Furthermore, the Logistics model and support vector machine dominated as the best model for modeling and predicting food insecurity response variables. The topmost factor predicting food insecurity from all four modeled response was the total income of the person, mode of transport, and its reliability, whether one know of any homeless student, household head, the number of dependents, and finally whether one is employed or not. Suggesting that one becoming food insecure depends on one income, numbers of dependents, and more as depicted by the model graphs. Identifying which person or persons are at risk of food being insecure was characterized by people living off-campus with family, someone been black or African American, and enrolled as a junior and having more than four dependents. Specifically, a Black American enrolled as a junior living with a family off campus with more than four dependents is a high risk of being food insecure.

Our initial preferred model or models were selected based on the evaluation metrics accuracy, misclassification, fbeta, and the area under the curve (AUC) from each response variable table obtained. That is, a model with the least misclassification, higher accuracy, an fbeta score approaching one, and a higher AUC. However, our initial preferred model or models may or may not be sufficient enough to make a predictive decision because they may involve irregularities. As an improvement to phase one of our modelings, we employed post hoc predictive modeling

to ascertain the initial preferred models. Specifically, we employed the ensemble analysis model to train our obtained several related but different analytical models (SVM, LDA, etc..) and then synthesized the results into a single score or spread in order to improve the accuracy of prediction. We made use of this idea of the ensemble method because a single model based on a data sample can have biases, high variability, or outright inaccuracies that affect the reliability of its analytical findings. By combining these different models we can reduce the effects of those limitations and provide better predictive information from the learning model. After the best predictive models are chosen a plot of variables of importance were obtained to determine which variables are most significant in predicting a given response variable, and also to answer the research question of which subgroup are at risk of a given food and house insecurity response.

2   Analysis and Results

2.1   Exploratory Data Analysis

2.1.1   Summary Statistics

Table 2 shows how the various variables used throughout the analysis are distributed across their respective levels, after data cleaning, accounting for double counting in multiple response instances, and collapsing levels where we considered appropriate. Missing values are denoted by **NA**, so variables with the additional **NA** level are those with missing observations.

Table 1: A summary of variables of interest

Variable Name	Levels	Coded As	Obervations	Percentage
enrollment	Full Time	1	4325	83.57%
	Part Time	2	850	16.43%
employment	Full Time	1	968	18.71%
	Part Time	2	2006	38.76%
	No	3	2201	42.53%
employment_type	On-Campus	1	491	16.51%
	Off-Campus	2	2483	83.49%
weekly_work_hrs	19 hours or less	1	1337	44.96%
	More than 19 hours	2	1637	55.04%
ethnicity	Hispanic	1	4462	86.22%
	Asian	2	115	2.22%
	Black/African American	3	135	2.61%
	White/Caucasian	4	291	5.62%

Table 1: A summary of variables of interest *(continued)*

Variable Name	Levels	Coded As	Obervations	Percentage
	Other	5	172	3.32%
gender	Female	1	3508	67.79%
	Male	2	1583	30.59%
	Non-binary/Non-conforming	3	62	1.20%
	NA		22	0.43%
total_income	< 10,000	1	851	16.44%
	10,000-19,999	2	862	16.66%
	20,000-29,999	3	806	15.57%
	30,000-39,000	4	649	12.54%
	40,000-49,999	5	538	10.40%
	50,000-69,999	6	651	12.58%
	70,000-99,999	7	494	9.55%
	>=100,000	8	324	6.26%
academic_level	Freshman	1	709	13.70%
	Sophomore	2	739	14.28%
	Junior	3	1275	24.64%
	Senior	4	1628	31.46%
	Masters	5	559	10.80%
	Doctoral/Professional	6	265	5.12%
college/school	Business	1	573	11.07%
	Education	2	484	9.35%
	Engineering	3	838	16.19%
	Health Sciences	4	577	11.15%
	Liberal Arts	5	1274	24.62%
	Science	6	822	15.88%
	Nursing	7	423	8.17%
	Pharmacy	8	41	0.79%
	Other	9	55	1.06%
	More than one	10	88	1.70%
mode_transport	Car (alone)	1	3359	64.91%
	Car (someone drives) & Carpool	2	805	15.56%
	Bus/Public transport	3	398	7.69%
	Bike, Trolley, or Walk	4	179	3.46%
	Other	5	125	2.42%

Table 1: A summary of variables of interest *(continued)*

Variable Name	Levels	Coded As	Obervations	Percentage
	Not applicable	6	309	5.97%
transport_reliability	Not reliable at all	1	92	1.89%
	Somewhat reliable	2	429	8.82%
	Fairly reliable	3	1628	33.46%
	Very reliable	4	2717	55.84%
living_alone	Yes	1	441	8.52%
	No	2	4734	91.48%
dependents	0	1	3802	73.47%
	1	2	367	7.09%
	2 - 3	3	448	8.66%
	4 or more	4	98	1.89%
	NA		460	8.89%
household_head	Yes	1	1155	22.32%
	No	2	3964	76.60%
	NA		56	1.08%
residence	On-Campus	1	133	2.57%
	Off-Campus with family	2	4093	79.09%
	Off-Campus not with family	3	821	15.86%
	Other	4	48	0.93%
	NA		80	1.55%
permanent_address	Yes	1	4830	93.33%
	No	2	265	5.12%
	NA		80	1.55%
spent_night_elsewhere	Rarely	1	138	52.08%
	Sometimes	2	83	31.32%
	Often	3	33	12.45%
	NA		11	4.15%
know_homeless_student	Yes	1	965	18.65%
	No	2	4095	79.13%
	NA		115	2.22%
federal_student_aid	Grants	1	1293	24.99%
	Work Study	2	256	4.95%
	Loans	3	581	11.23%
	Scholarship	4	237	4.58%

Table 1: A summary of variables of interest *(continued)*

Variable Name	Levels	Coded As	Obervations	Percentage
	Multiple Aids	5	1652	31.92%
	Other	6	992	19.17%
	None	7	49	0.95%
	NA		115	2.22%
FI_q26	Never true	1	2886	55.77%
	Sometimes true	2	1705	32.95%
	Often true	3	427	8.25%
	NA		157	3.03%
FI_q27	Never true	1	2579	49.84%
	Sometimes true	2	1744	33.70%
	Often true	3	695	13.43%
	NA		157	3.03%
FI_q28	Yes	1	1684	32.54%
	No	2	3334	64.43%
	NA		157	3.03%
FI_q30	Yes	1	1676	32.39%
	No	2	3322	64.19%
	NA		177	3.42%
FI_q31	Yes	1	1268	24.50%
	No	2	3730	72.08%
	NA		177	3.42%
expenditures_changed	Decreased	1	1323	25.57%
	About the same	2	2201	42.53%
	Increased	3	1413	27.30%
	NA		238	4.60%
income_changed	Decreased	1	2702	52.21%
	About the same	2	2046	39.54%
	Increased	3	189	3.65%
	NA		238	4.60%
fed_aid_changed	Decreased	1	1175	22.71%
	About the same	2	3388	65.47%
	Increased	3	374	7.23%
	NA		238	4.60%
debt_changed	Decreased	1	289	5.58%

Table 1: A summary of variables of interest *(continued)*

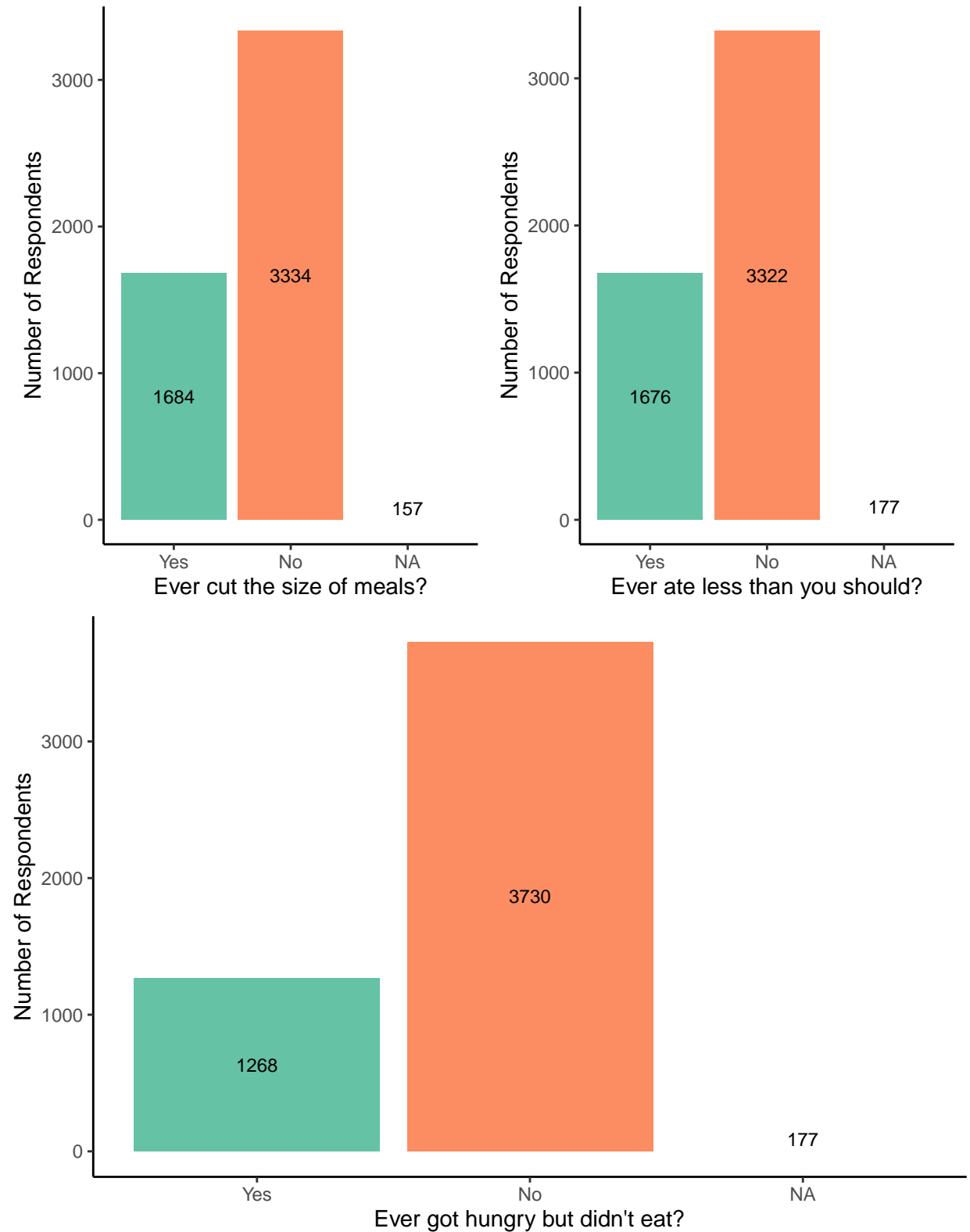| Variable Name	Levels	Coded As	Obervations	Percentage
	About the same	2	2441	47.17%
	Increased	3	2207	42.65%
	NA		238	4.60%

2.1.2   Distribution of Housing Insecurity Responses
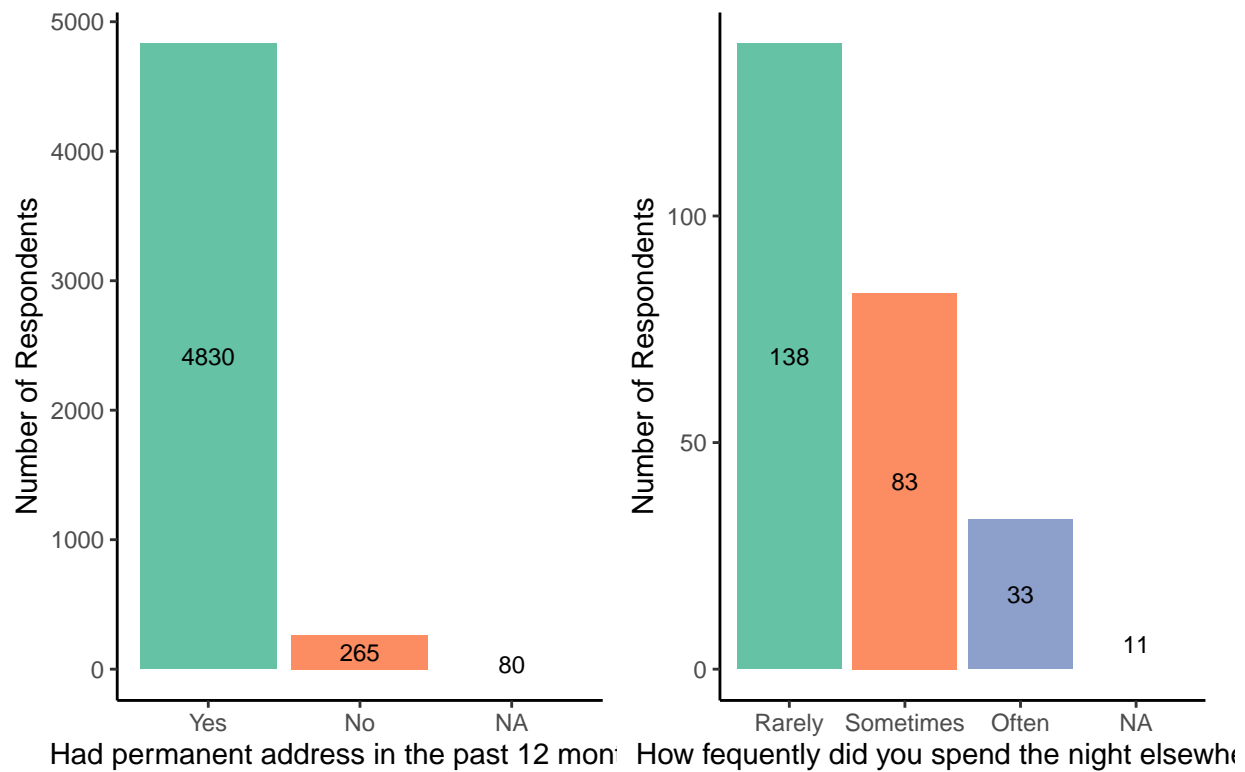
2.1.3 Distribution of Food Insecurity Responses
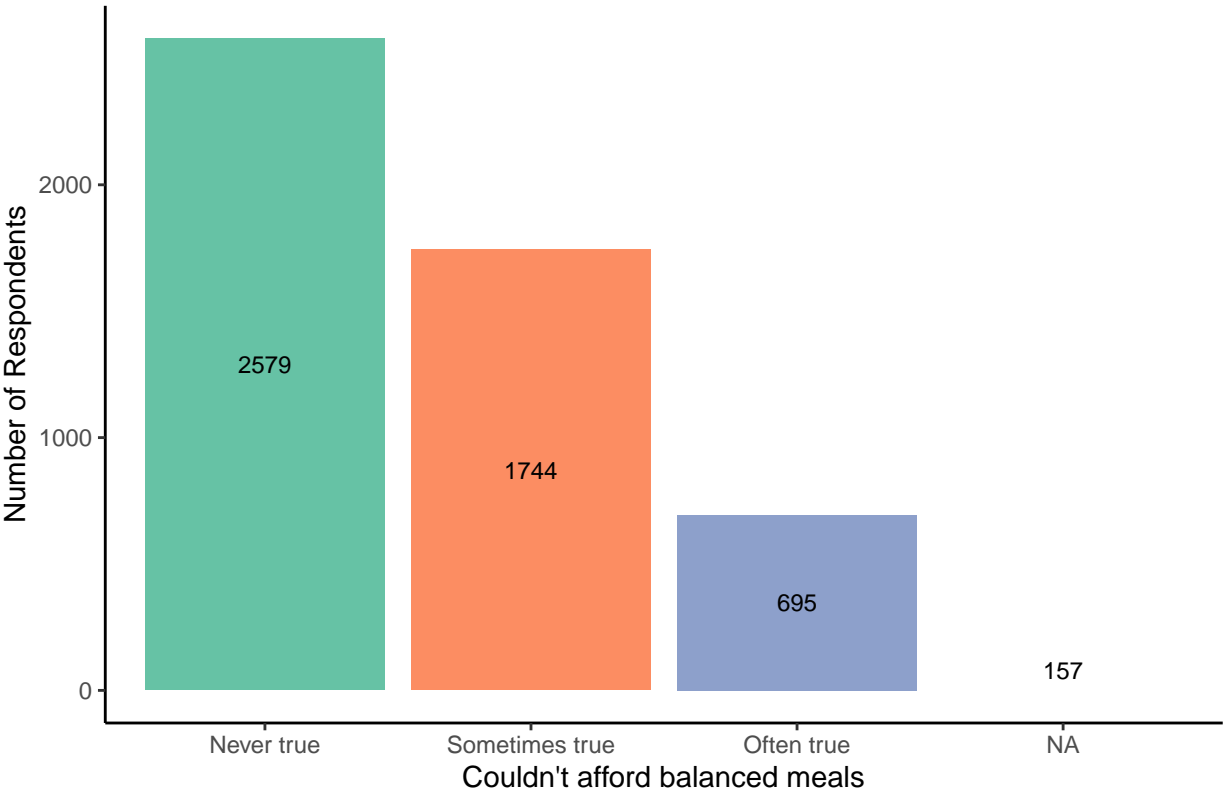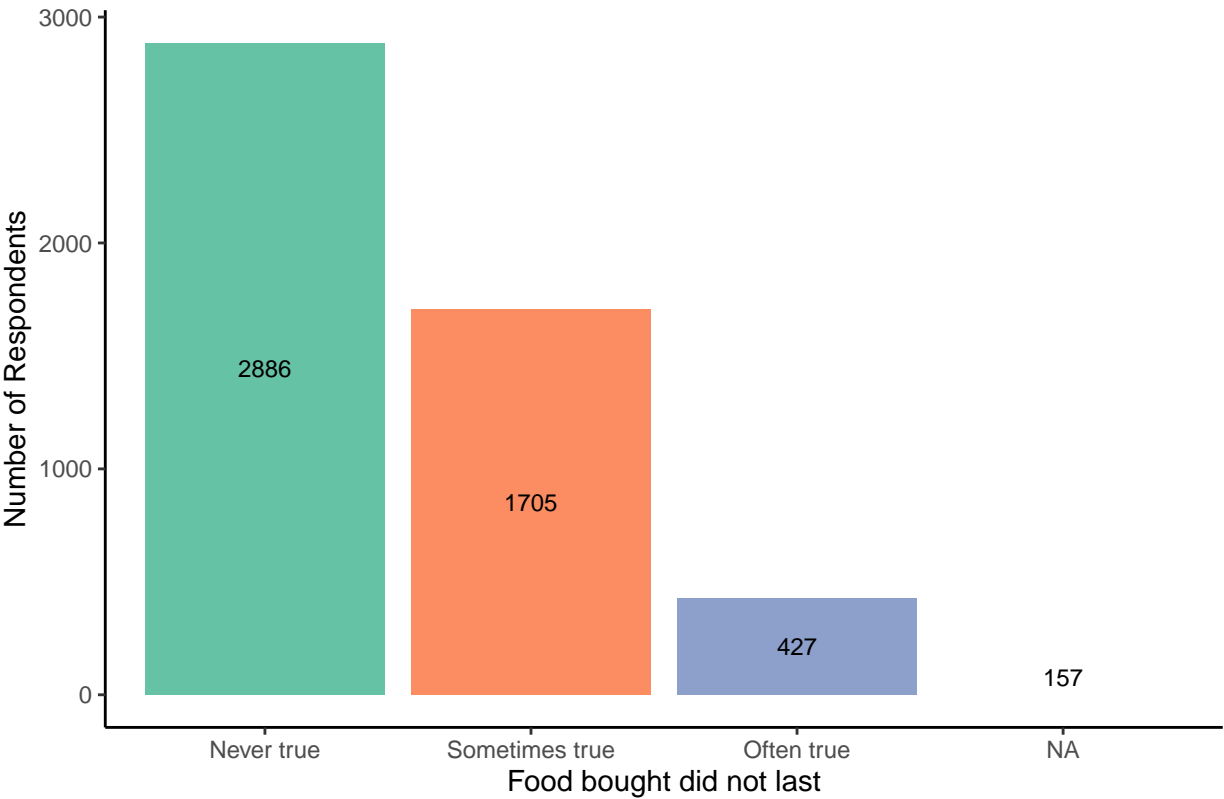
The table and graphs above display the number of missing values with missing percentages as well as the distribution of the potential response variables. From the results, the following observations were made:

- It is clear that all variables have some amount of missing observation given they are all above 50%, hence a

necessary missing data treatment is required.

- The response variables also contains missing values, hence we filter all missing observation with respect to each response variable out and used the remaining data set for further imputation and analysis analysis.

- The mice package aided in imputation the missing values in the predictor variables, specifically by using the median, mice was chosen because it is robust to data and its imputation style.

- The selected response variables exhibited highly imbalanced classification. The imbalanced classification was treated with both sampling methods under the ROSE package.

2.2   Model Building

Given our data set and its structure, the following supervised classification machine learning algorithms were employed to obtained a predictive model for each given response variable;

- Logistic Regression

- Linear Discriminant Analysis (LDA)

- K-Nearest Neighbors (KNN)

- Multivariate Adaptive Regression Splines (MARS)

- Support Vector Machines with Linear Kernel (SVM Linear)

- Support Vector Machines with Radial Basis Kernel (SVM Radial)

2.2.1   Partitioning data set

For the purpose of training and validation of each derived model ,and the estimation of the performance metrics, the entire data set was partitioned into training and testing in a ratio of 2/3 and 1/3 respectively.

### 2.2.2   Results for Housing Insecurity Responses

Here, we modeled housing insecurity as a response based on whether or not a respondent had a permanent address (permanent_address) as well as whether or not a respondent spent the night elsewhere given that they did not have a permanent address in the past six months to the survey period (spent_night_elsewhere).

Results for Permanent Address

Table 2: Evaluation metrics for Permanent House Address as a response variable

Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.24	0.76	0.77	0.53	0.86	0.7058
LDA	0.22	0.78	0.79	0.55	0.87	0.7149
KNN	0.37	0.63	0.64	0.58	0.77	0.637
MARS	0.2	0.8	0.82	0.48	0.89	0.7206
SVM Linear	0.22	0.78	0.79	0.51	0.87	0.7151
SVM Radial	0.17	0.83	0.85	0.45	0.91	0.7011

Based on the table values our best initial models are SVM Radial and the MARS (earth) model from the permanent housing address response.
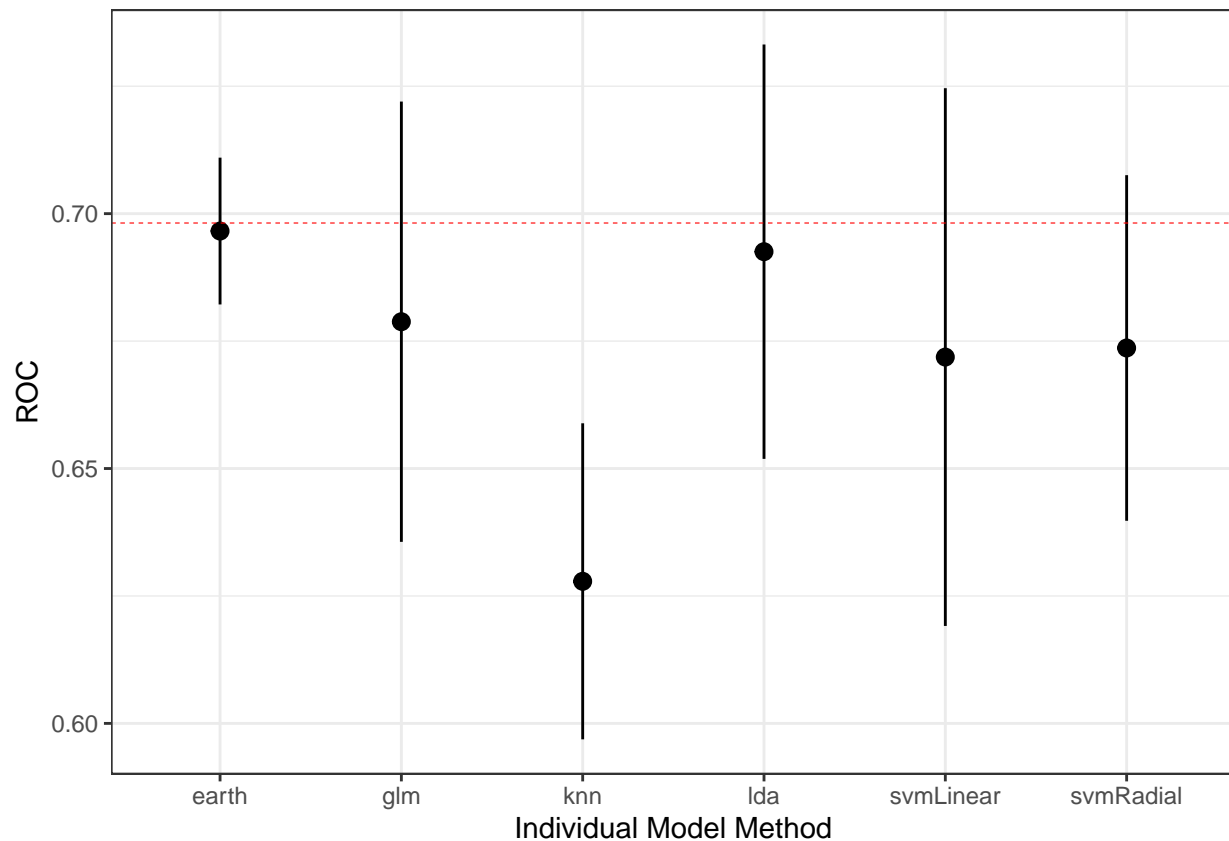
The ensemble model came up with the MARS (earth) model as the best performing model given its line of reference. Therefore, we confidently conclude that the MARS model which appeared at both our necessary and sufficient measure for selection criteria is the model predictive model for modeling the housing insecurity permanent address.

Variable Importance by the Ensemble Model
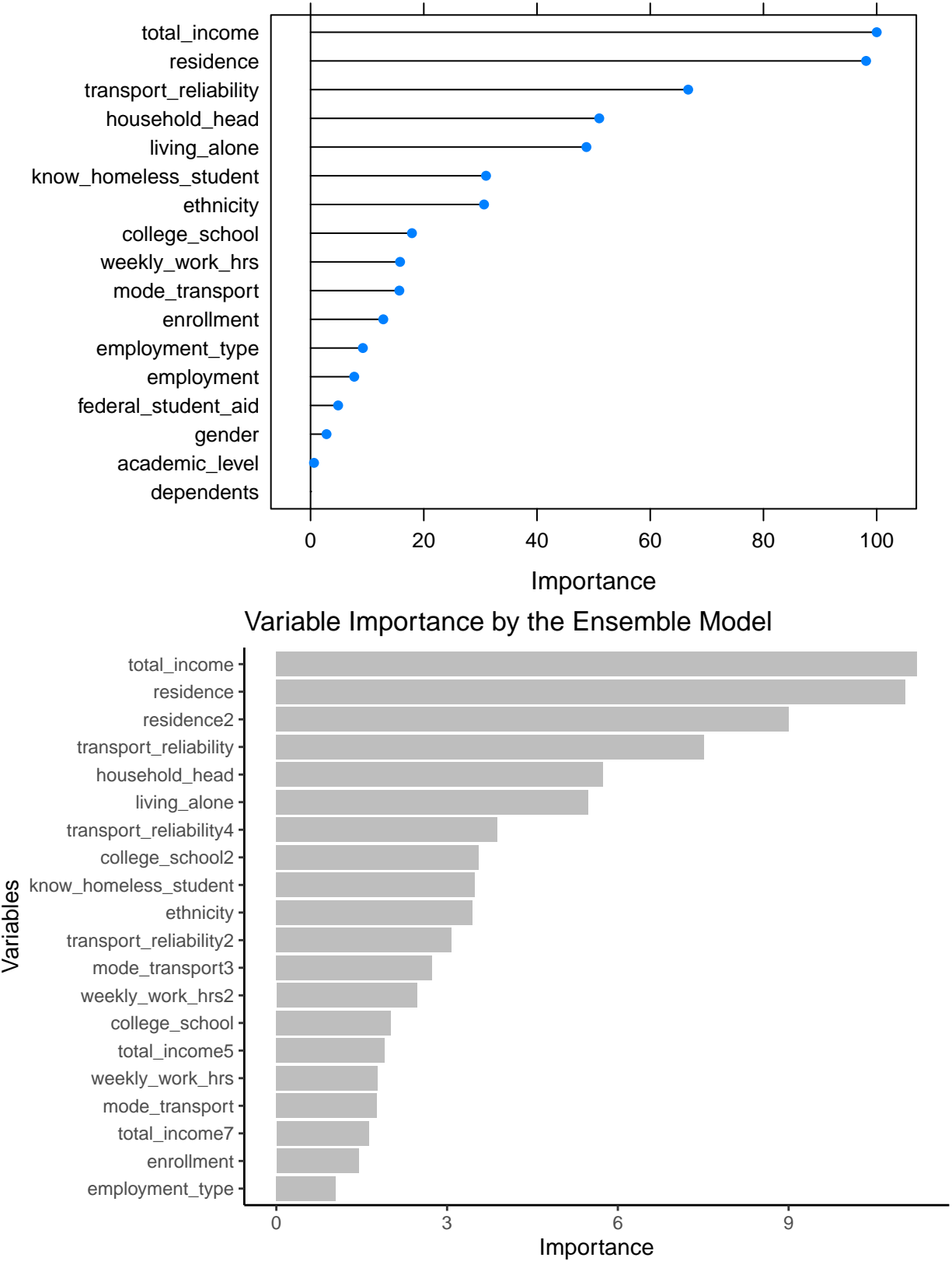
Our initial preferred model based on the evaluation metrics been SVM Radial would have total income, residence, transport reliability, household head, and living alone are the top five variables that accurately predicts the housing insecurity, with employment, federal aid, gender, and dependents been least of importance. However our final model

by the ensemble method which is the MARS model also had same variables of most importance as that of the SVM radial. However this MARS model plot further classifies the important variable residence2 , transport reliability and college_school2 as subgroups of importance.

Results for Spending the night elsewhere

Table 3: Evaluation metrics for Spent Night Elsewhere as a response variable

| Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.54	0.46	0.35	0.56	0.37	0.5216
KNN	0.48	0.52	0.49	0.56	0.48	0.5294
MARS	0.45	0.55	0.65	0.47	0.56	0.5952
SVM Linear	0.54	0.46	0.51	0.42	0.46	0.5012
SVM Radial	0.51	0.49	0.54	0.44	0.49	0.5498

Taking spending the night elsewhere, our initial best model is the MARS (earth) model given the table values, that is MARS model has the least misclassification, higher accuracy, higher fbeta, and higher AUC. However we add SVM radial to our best model since its comes next to the MARS model for example that AUC and fbeta.
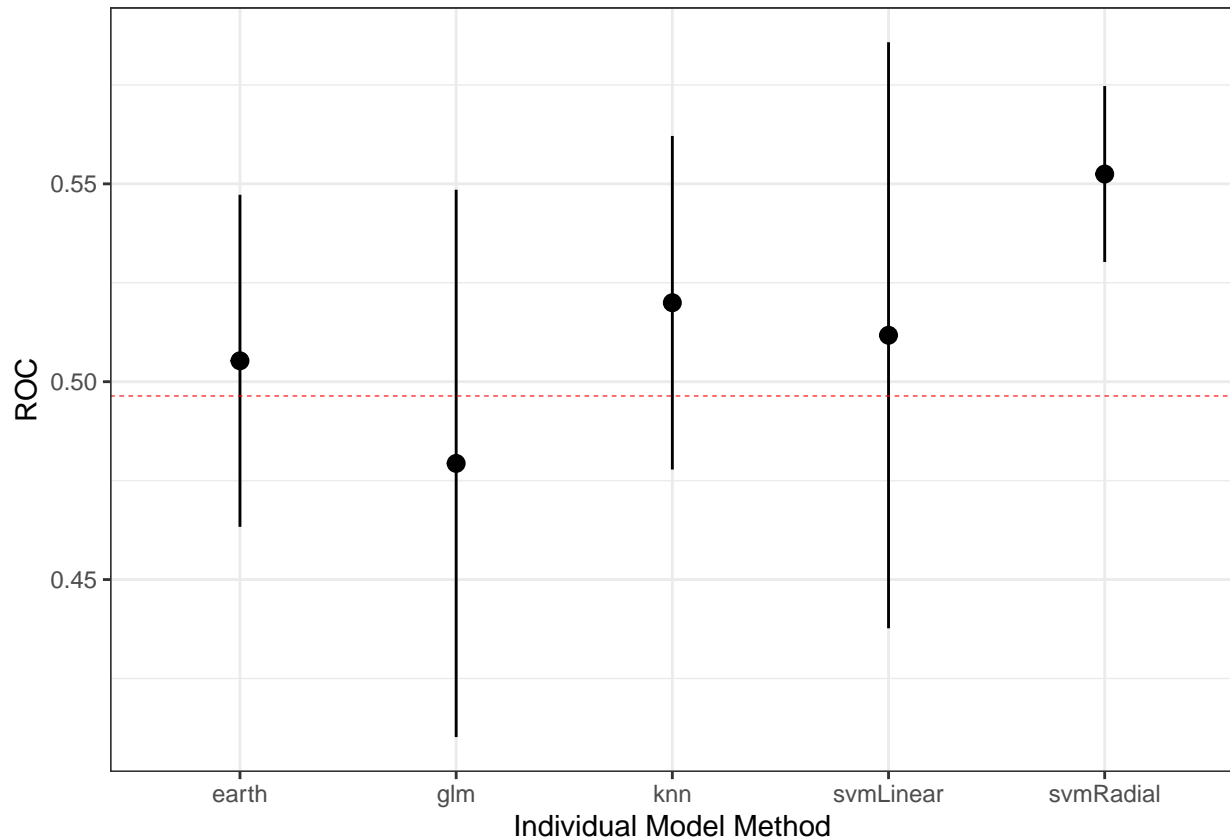
According to the ensemble modeling, the SVM Radial is the best given the reference line. Hence the SVM Radial becomes our best model given spending the night elsewhere as the response variable.

Variable Importance by the Ensemble Model
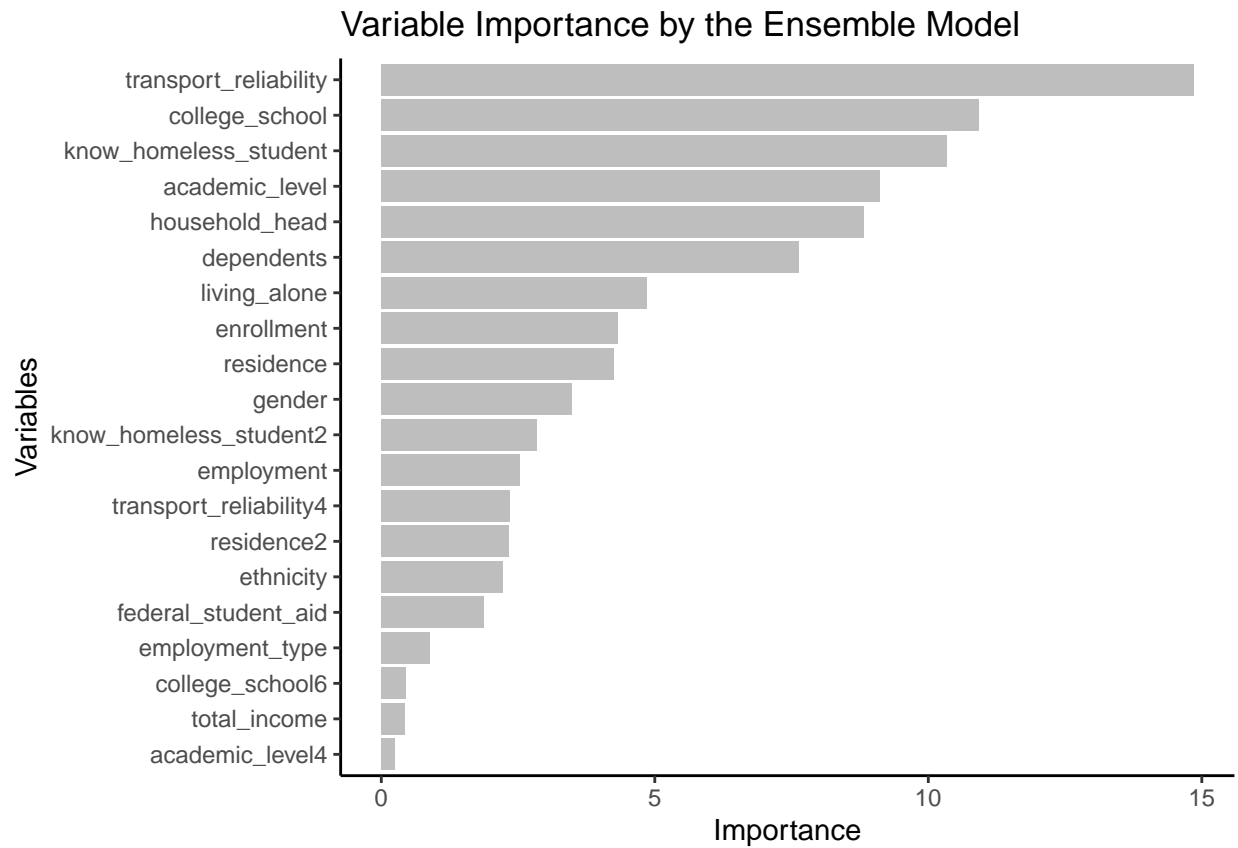
Variable Importance by the SVM Radial Model with subgroups
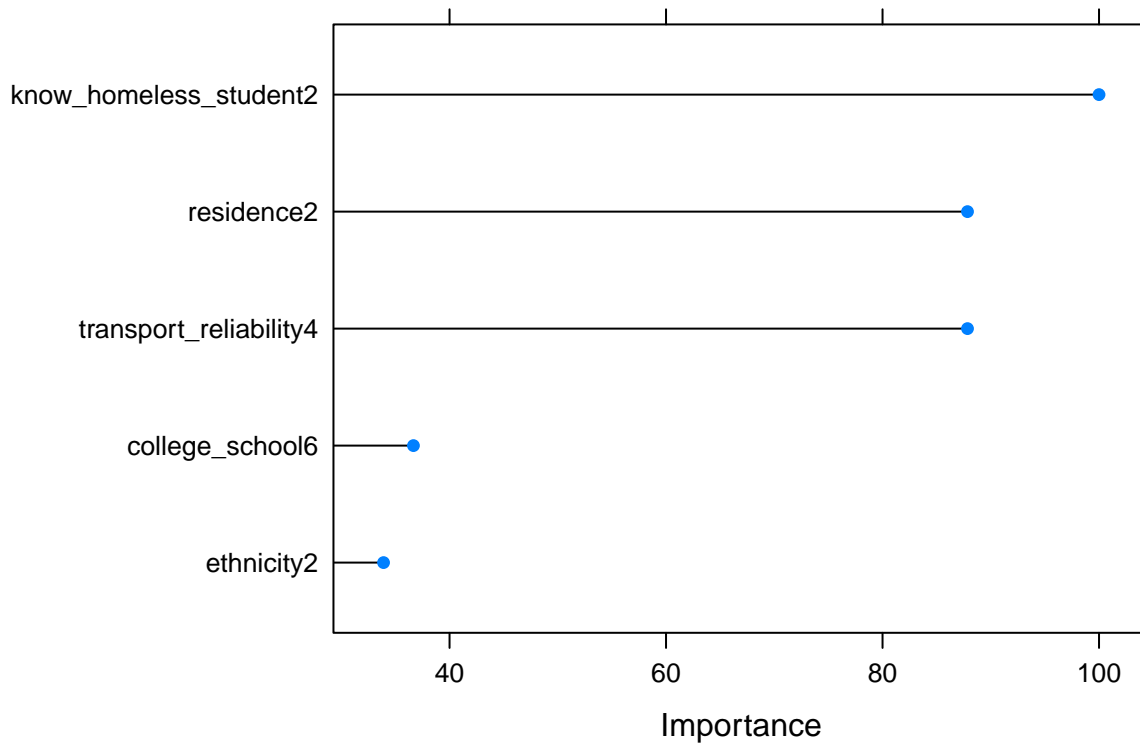
According to our variable of an important plot from the best model (SVM radial), the top 5 variables that determine housing insecurity are transport reliability, college, known homeless students, academic level, and household head. Also know homeless student2, residence2, transport reliability 4, college school6 and ethnicity2 are the subgroups of importance.

2.2.3 Results for Food Insecurity Responses

In this section, we report results for modeling food insecurity in terms response variables based on survey questions as described below:

- **Food Insecurity I**: "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? (FI_q26).

- **Food Insecurity II**: "I couldn't afford to eat balanced meals." Was that often, sometimes, or never true for you in the last 12 months? (FI_q27).

- **Food Insecurity III**: 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? (FI_q28).

- **Food Insecurity IV**: In the past 12 months, were you ever hungry but didn't eat because there wasn't

enough money for food? (FI_q31).

Though Food Insecurity I and II originally had three levels (Often true, Sometimes true, and Never true), we modeled them as dichotomous responses where the "Often true" and "Sometimes true" levels were put in a "Yes" group as a sign of food insecurity, while the "Never true" level was classified as a "No" group for respondents not at the risk of food insecurity.

Results for Food Insecurity Response I

Table 4: Evaluation metrics for Food Insecurity I as a response variable

Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.33	0.67	0.66	0.67	0.64	0.7342
LDA	0.34	0.66	0.65	0.67	0.63	0.7335
KNN	0.37	0.63	0.57	0.68	0.58	0.6747
MARS	0.34	0.66	0.66	0.66	0.64	0.7319
SVM Linear	0.34	0.66	0.66	0.66	0.63	0.7314
SVM Radial	0.33	0.67	0.66	0.67	0.64	0.7322

Based on the table values and the evaluation metric criteria, with the exception of knn model , all the others are preferred as our initial best models.
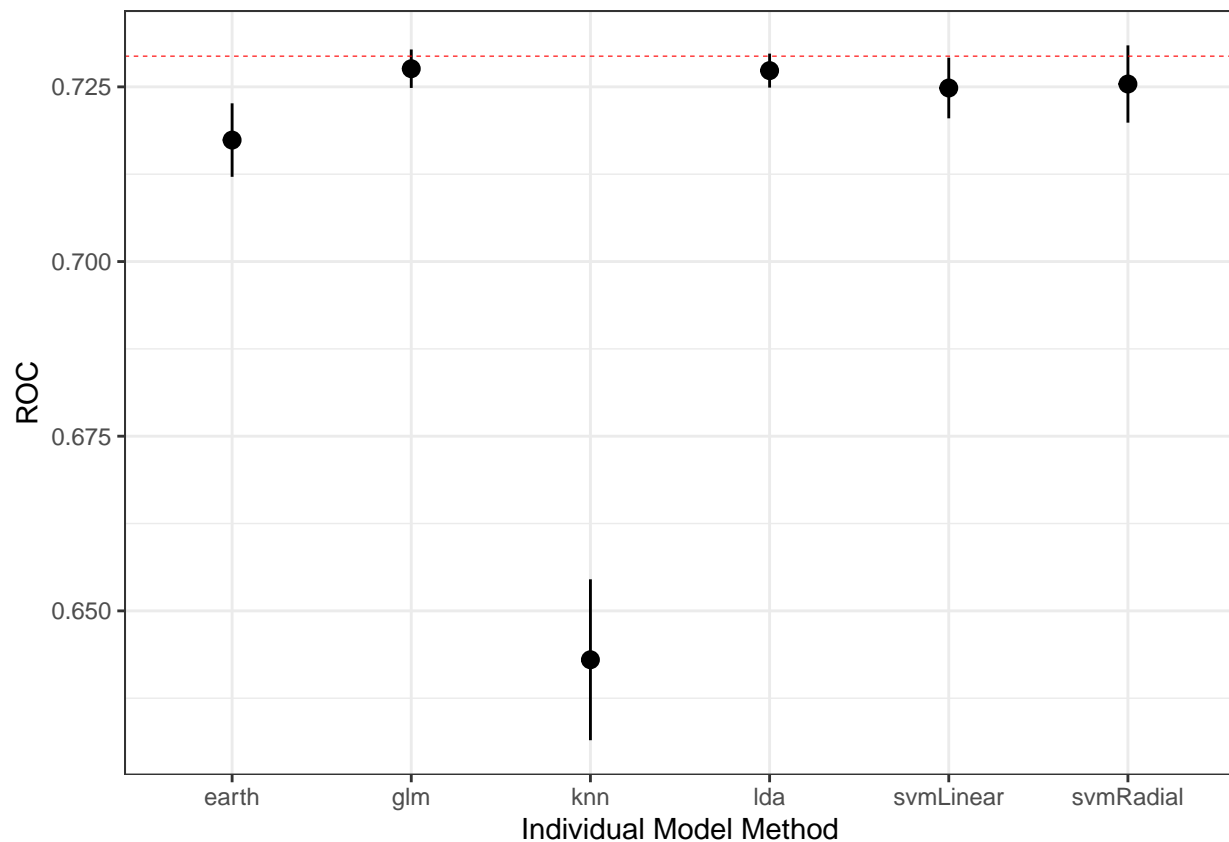
From the ensemble model, both glm and LDA are considered the best model for predicting food insecurity response 1, however, we decide on logistics regression as our best model.

Variable Importance by the Ensemble Model

Using glm as our best model the variable of importance or factor that predict food insecurity 1 as response are total income, transport reliability, know homeless student, household head and dependents.

Variable Importance by the Logistic Regression Model
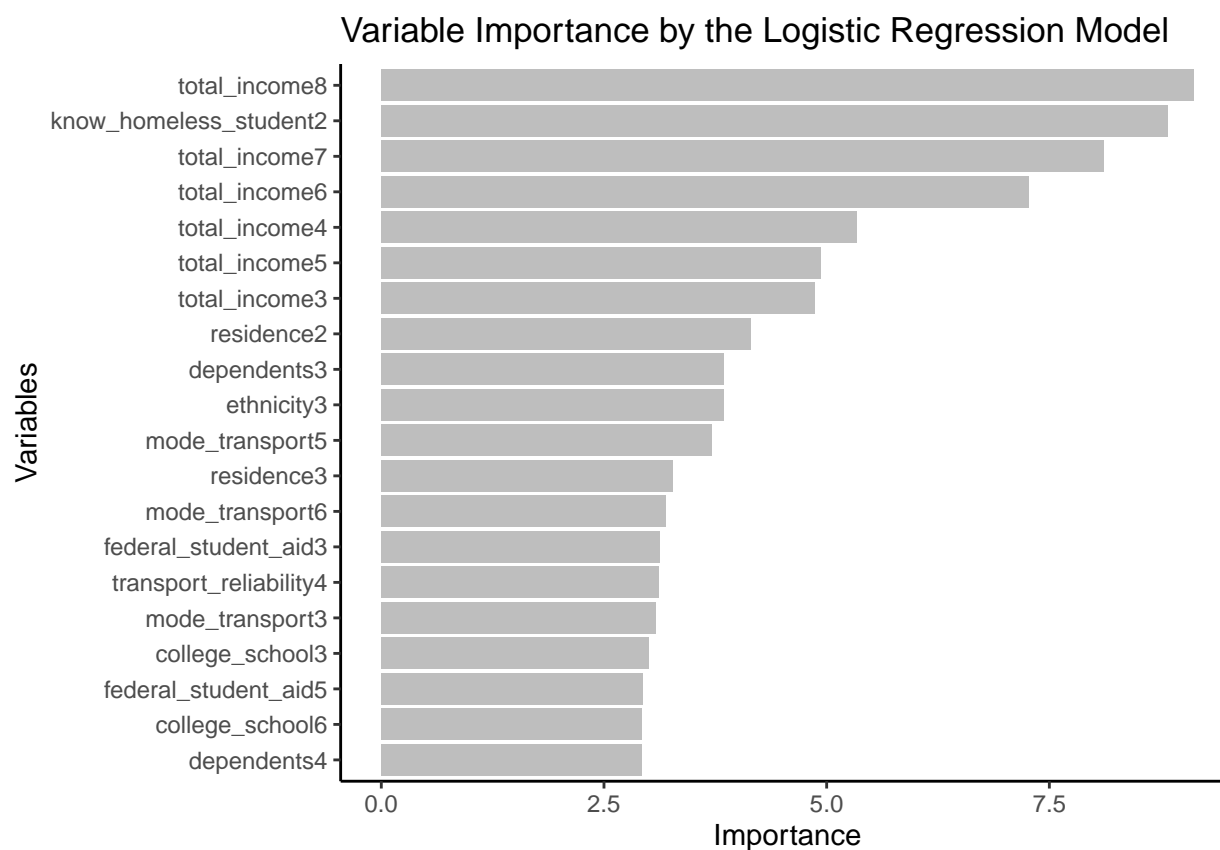
Given the plot above, the top 5 subgroups are total income8, know homeless student 2, total income 7, residence2, and ethnicity3.

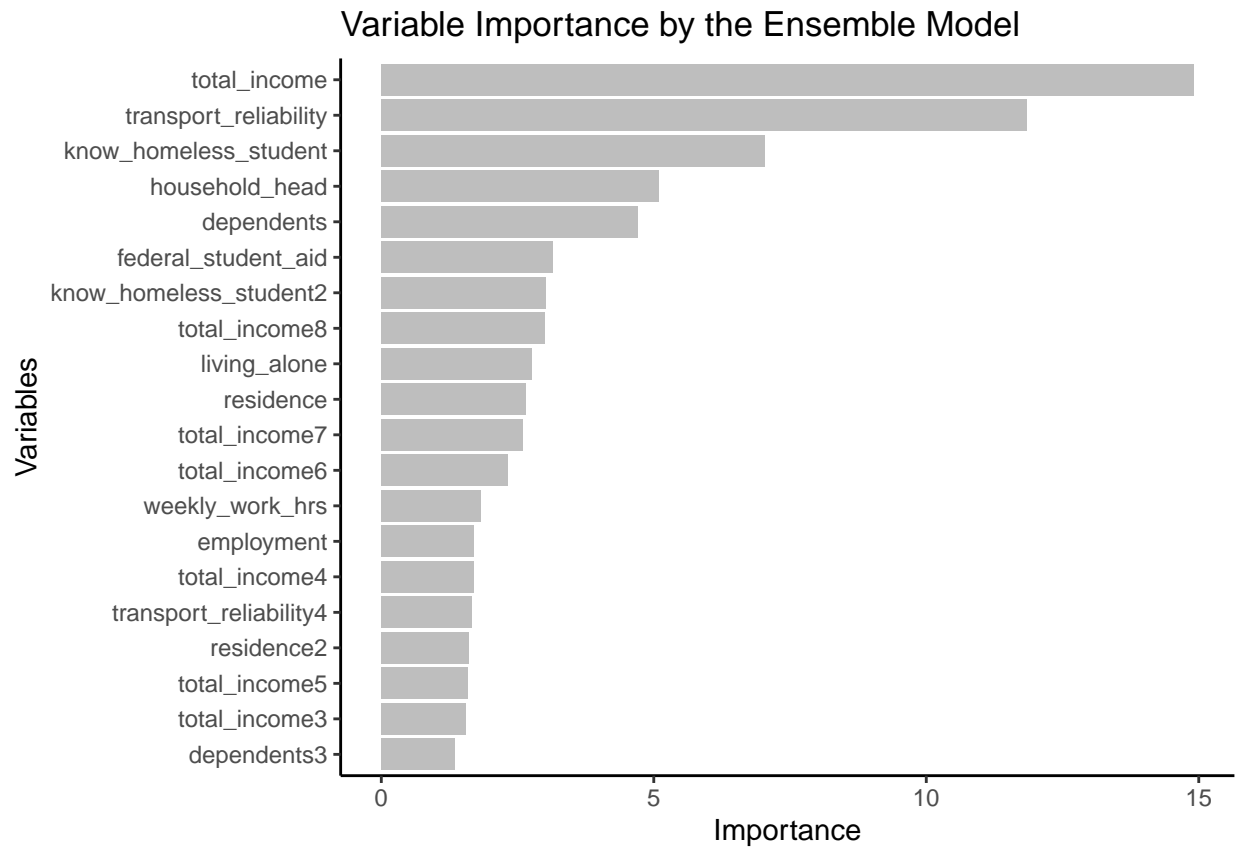**Results for Food Insecurity Response II**

Table 5: Evaluation metrics for Food Insecurity II as a response variable

Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.34	0.66	0.64	0.67	0.66	0.7123
LDA	0.34	0.66	0.65	0.67	0.66	0.7132
KNN	0.37	0.63	0.56	0.7	0.6	0.6689
MARS	0.35	0.65	0.64	0.65	0.65	0.7078
SVM Linear	0.35	0.65	0.64	0.67	0.65	0.7116
SVM Radial	0.35	0.65	0.63	0.67	0.65	0.7113

Modeling food insecurity 2 as response, the best initial models are all models apart from the knn model.
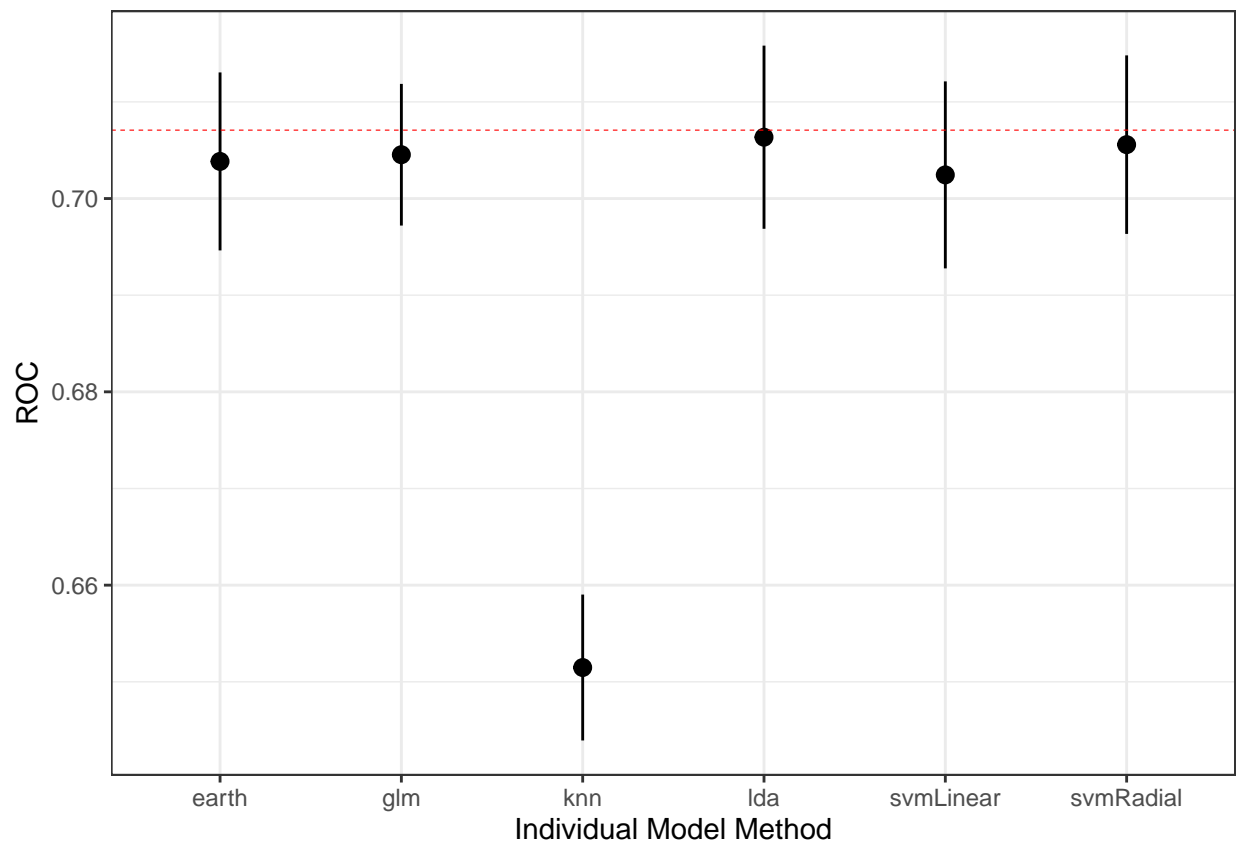
The best model from the ensemble model is the LDA, however, the SVM Radial can as well the best model given that it is closer to the line of reference as indicated from the plot above.

Variable Importance by the Ensemble Model

The plot above indicates that treating food insecurity 2 as a response through the LDA model by ensemble model, the 5 most important factors are total income, transport reliability, know homeless student, household head, and employment.

Variable Importance by the lda Model
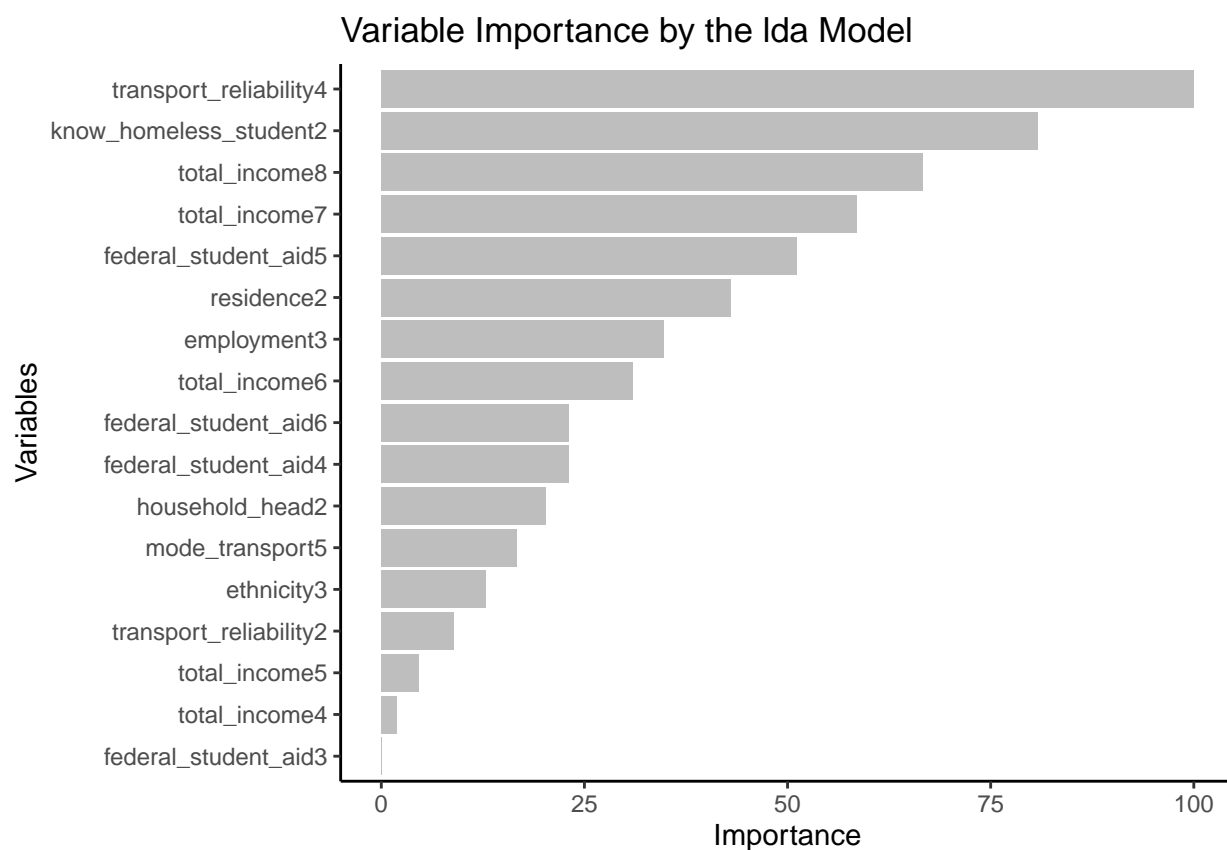
Given the plot above, the top 5 subgroups are transport reliability4, know homeless student2, total income8, federal student aid5, and residence2.
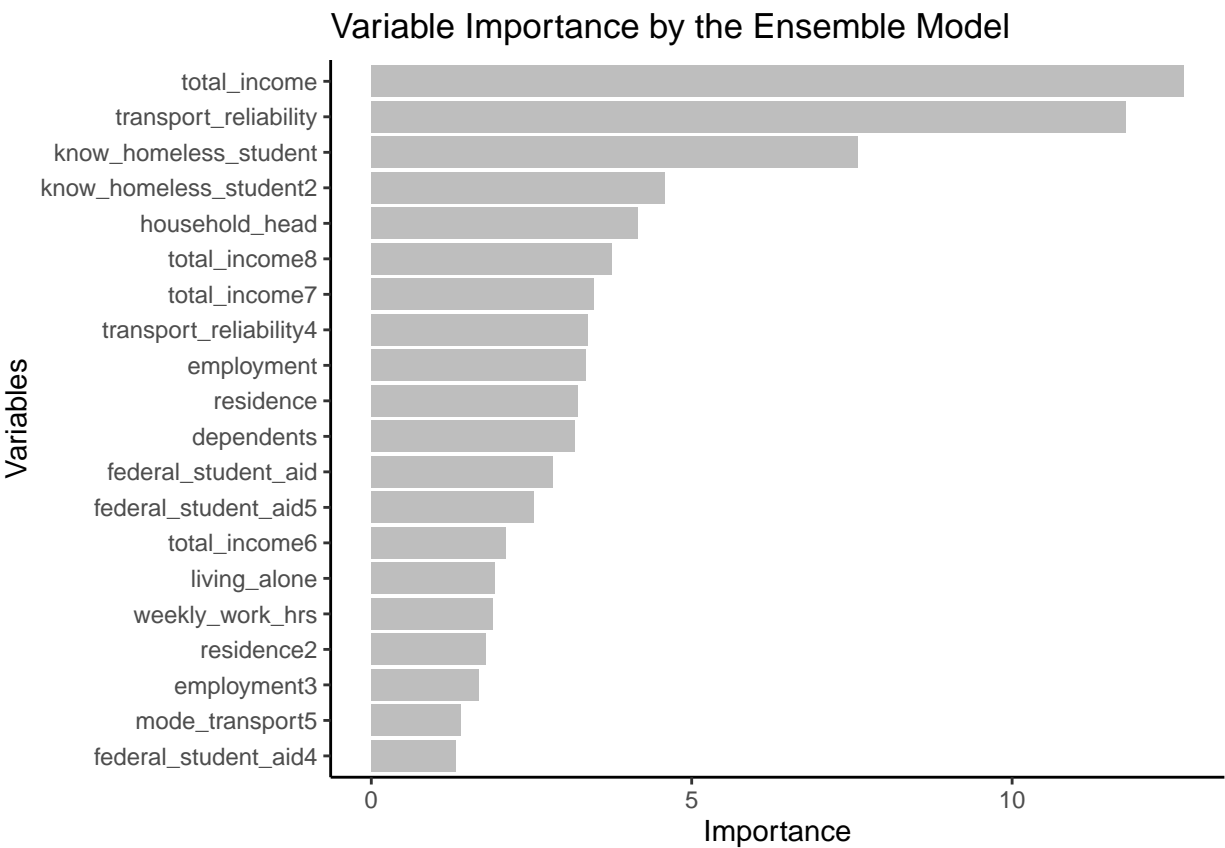
Results for Food Insecurity Response III

Table 6: Evaluation metrics for Food Insecurity III as a response variable

Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.35	0.65	0.59	0.68	0.55	0.7004
LDA	0.34	0.66	0.62	0.69	0.57	0.7017
KNN	0.38	0.62	0.54	0.66	0.51	0.6515
MARS	0.35	0.65	0.59	0.68	0.55	0.6975
SVM Linear	0.34	0.66	0.6	0.69	0.56	0.7006
SVM Radial	0.33	0.67	0.59	0.71	0.56	0.7033

Given the table values of evaluation metric criteria, our initial best models are all but knn and MARS.

The final best model through the ensemble method is the SVM Radial for the food insecurity 3.

Variable Importance by the Ensemble Model

The plot above indicates that treating food insecurity 3 as a response through the SVM Radial model by ensemble model, the 5 most important factors are total income, transport reliability, know homeless student, household head, and depedents.
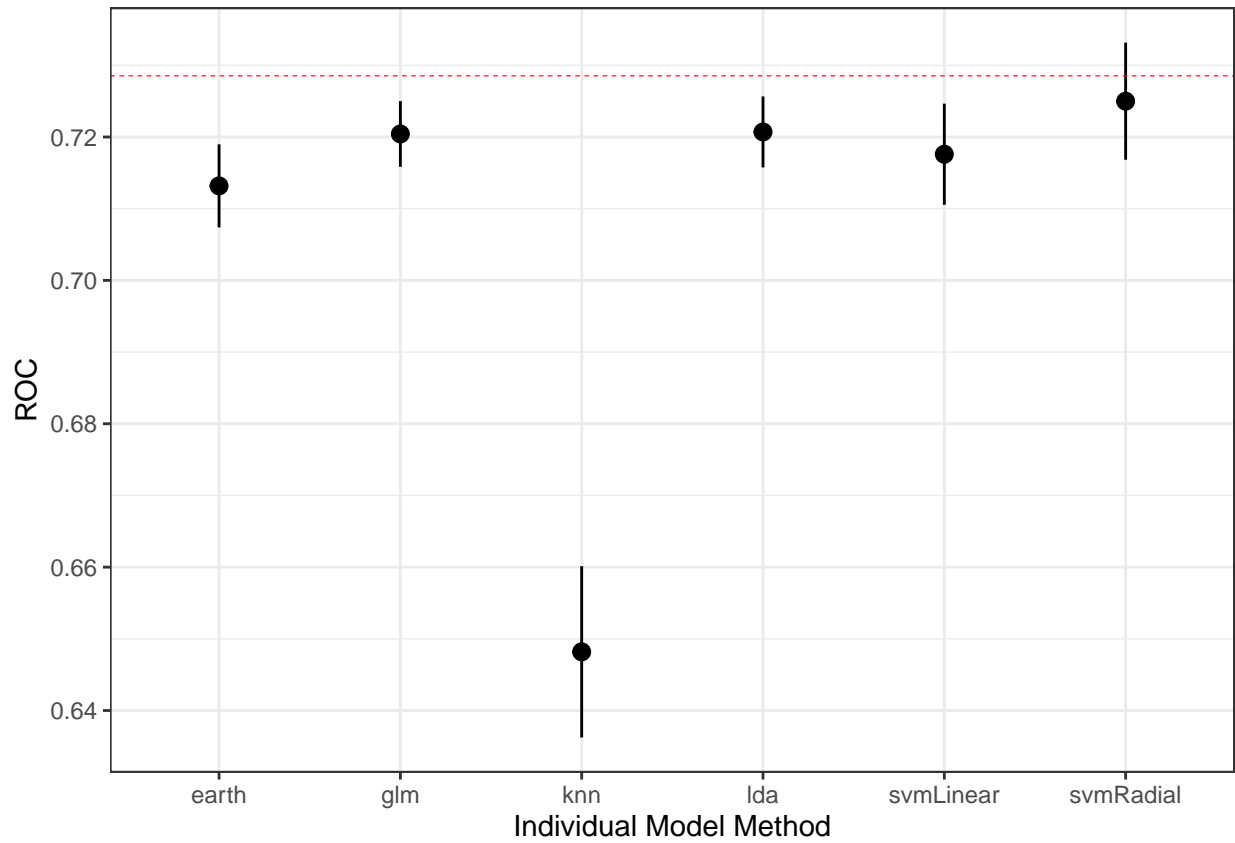
Variable Importance by the svmRadial Model
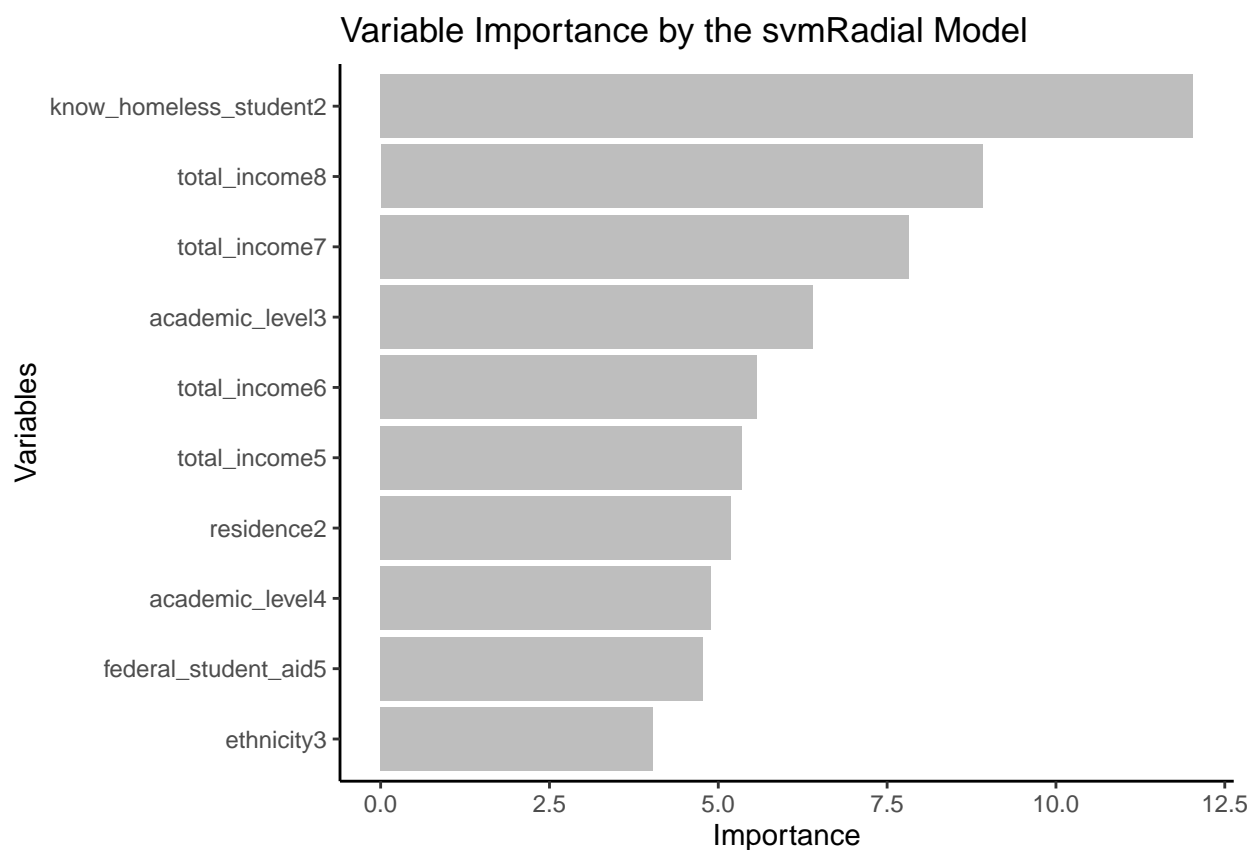
Given the plot above for food insecurity 3, the top 5 subgroups are know homeless student2, total income8, total income7, academic level3, and residence2.
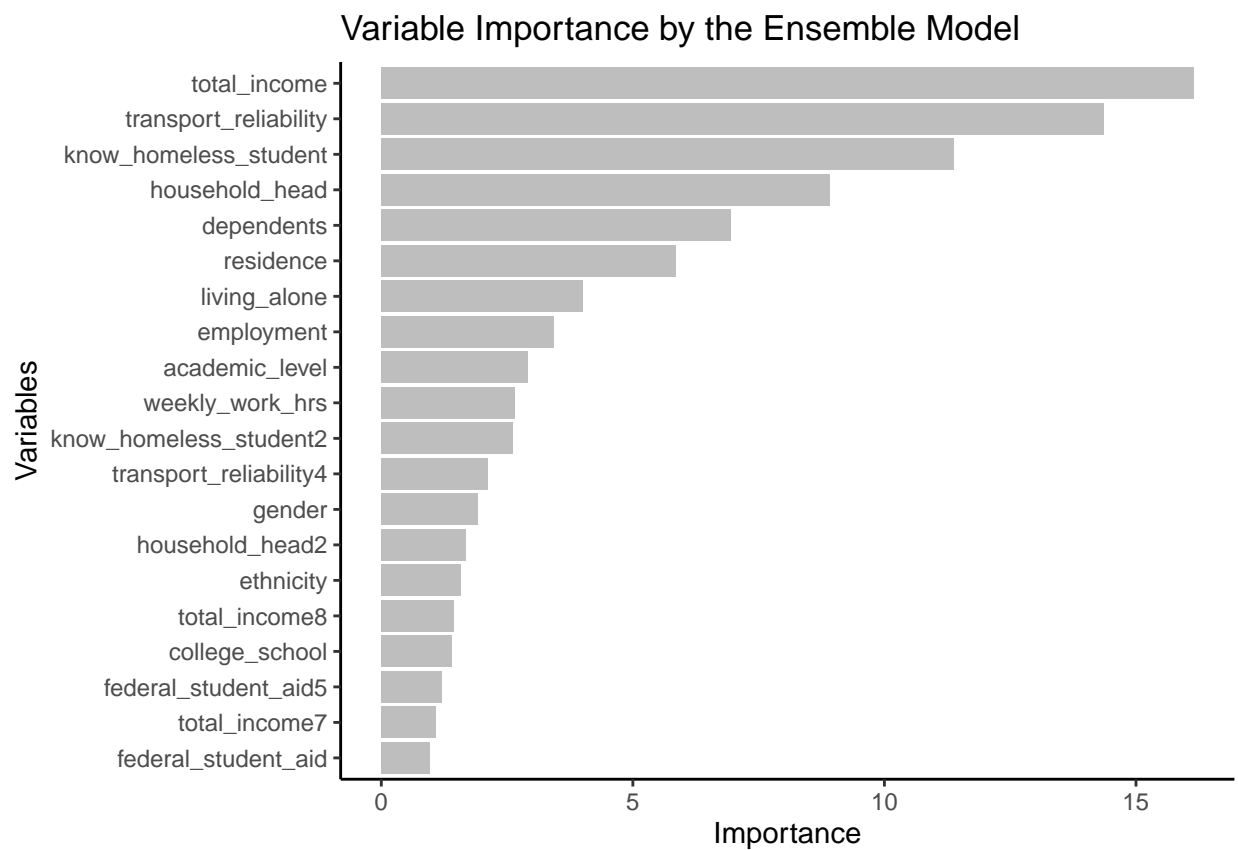
Results for Food Insecurity Response IV

Table 7: Evaluation metrics for Food Insecurity V as a response variable

Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.31	0.69	0.65	0.7	0.52	0.7325
LDA	0.32	0.68	0.65	0.69	0.51	0.7312
KNN	0.38	0.62	0.55	0.65	0.43	0.6433
MARS	0.32	0.68	0.63	0.69	0.5	0.7262
SVM Linear	0.32	0.68	0.65	0.7	0.51	0.731
SVM Radial	0.3	0.7	0.62	0.73	0.52	0.7375

Our best models according the evaluation metrics criteria are glm and SVM Radial for food insecurity response 4.

Now according to the ensemble method, both initial best models are best for predicting the response variable, however, we decided on the glm model as per the ensemble model.

Variable Importance by the Ensemble Model
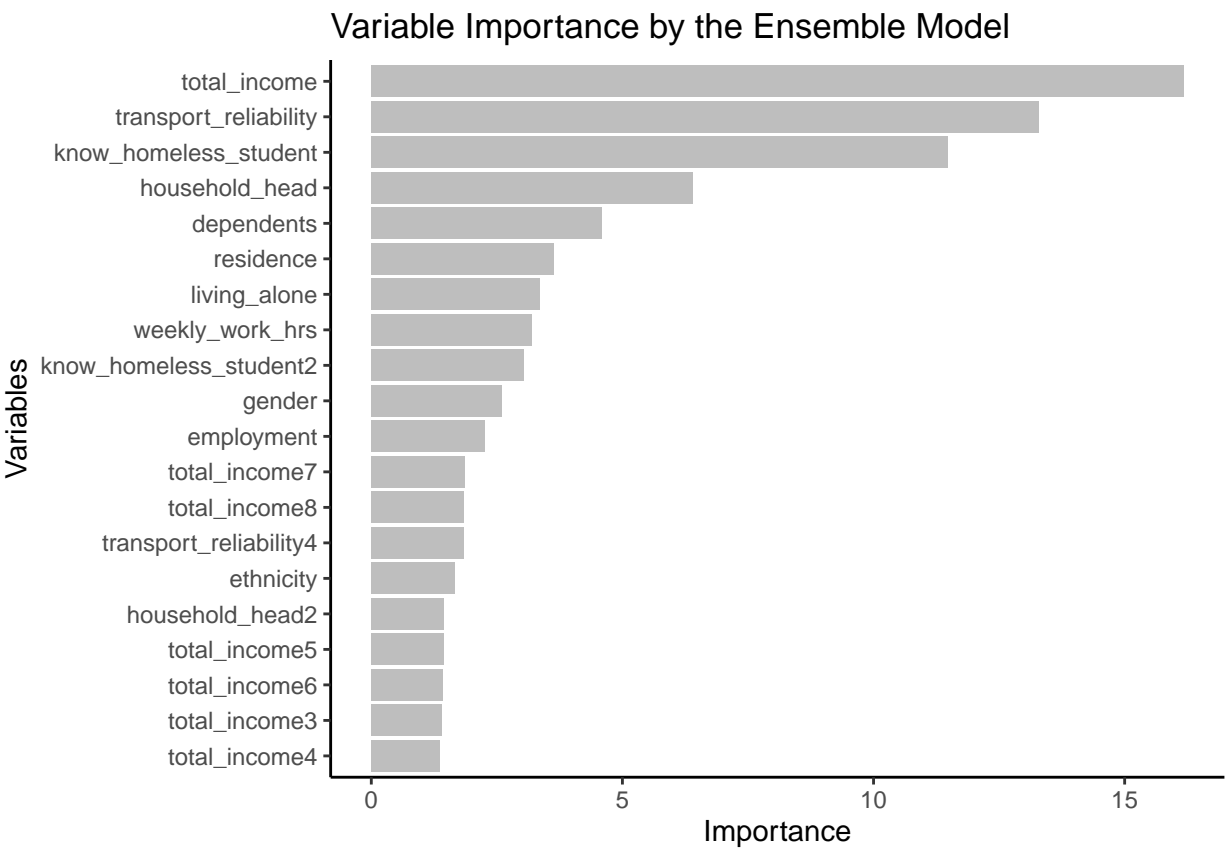
The plot above indicates that treating food insecurity 4 as a response through the glm model by ensemble model, the 5 most important factors are total income, transport reliability, know homeless student, household head, and dependents.
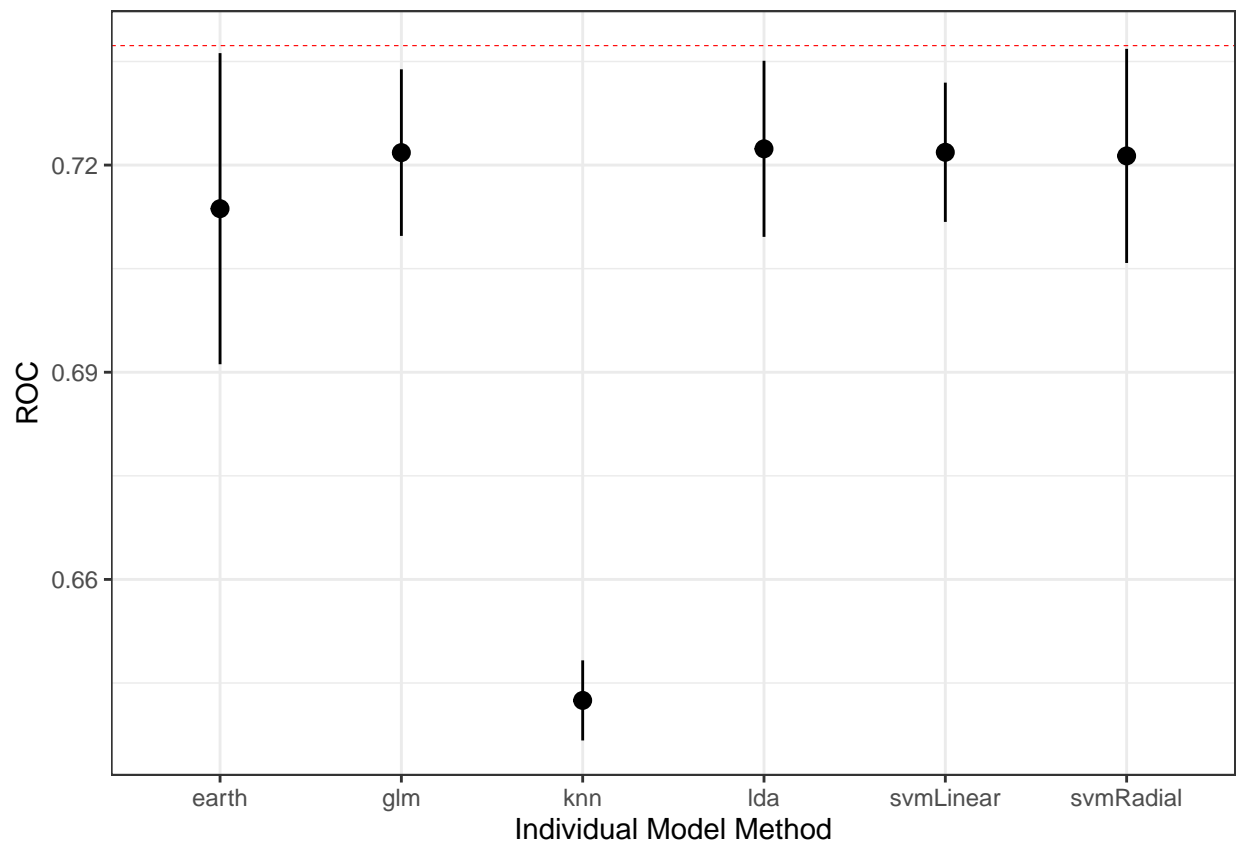
Variable Importance by the glm Model
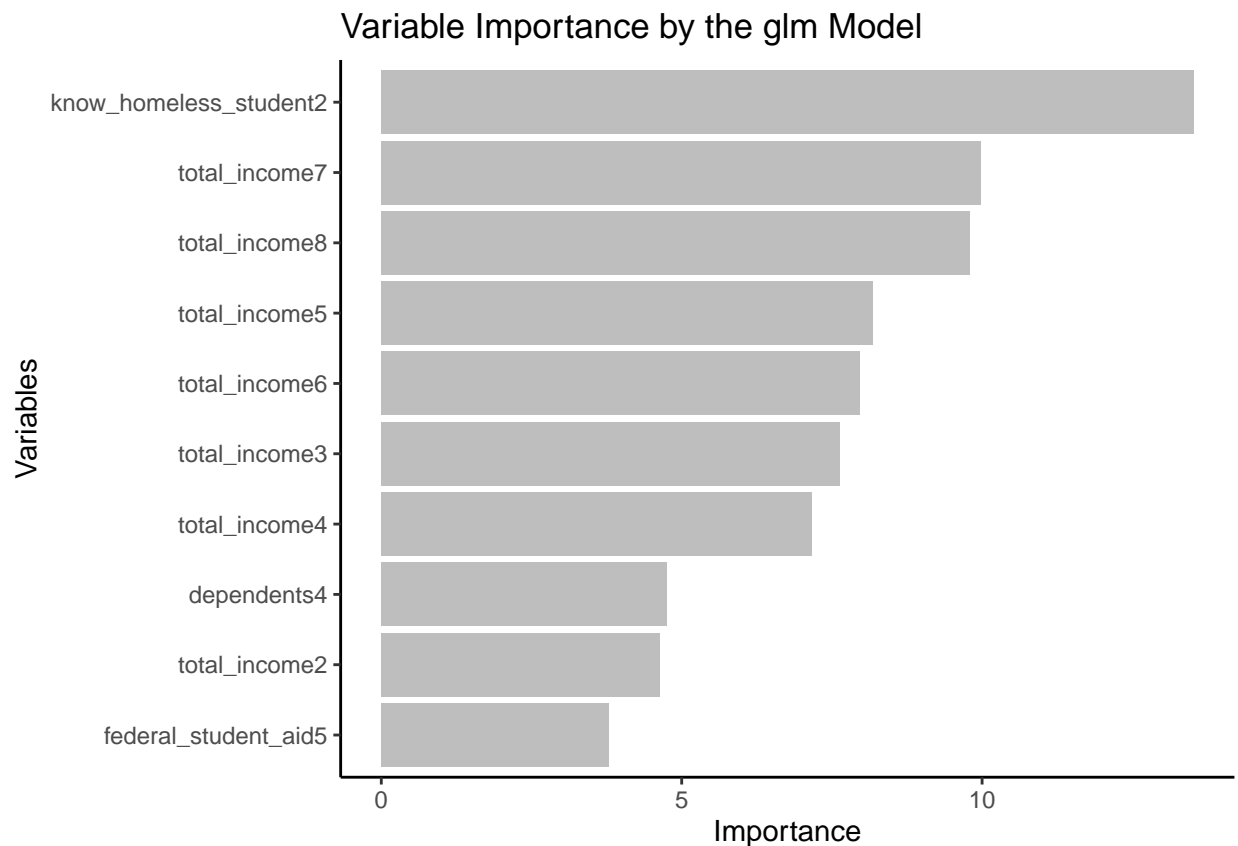
Given the plot above for food insecurity 3, the top 10 subgroups involves known homeless student2, total income (2,3 - 8), dependents4 and federal student aid5.

3   Methodology

This section serves to provide a road map to the approaches our group adopted and executed in terms of our analysis.

3.1   Exploratory Data Analysis

In this section, we provided descriptive analyses of the data including numerical summaries and graphical displays as a first-hand insight into the data to help ourselves and our audiences to better understand the underlying dataset. We primarily looked at the distributions of our target variables and all the predictor variables.

In terms of the types of graphs or visualizations, we considered segmented or comparative bar graphs. We critically assessed these graphs and settled on a few which help us to effectively tell a compelling story about the data.

- **Cross Tabulations (Contingency Tables)**: Provide numerical summary of how responses differ by some categories.

- **Bar graphs (Segmented/Comparative bar graphs)**: To explore how the available categorical predictors

are distributed across the levels of the target or response variable.

3.1.1 Missing Data/Outliers and their treatments

The available data was inspected for missing and unusual values. The missing data and/or outlying data identified from our data preprocessing and exploratory analysis was treated appropriately depending on the nature of the missingness and the information we were able to gather about the possible reasons for the missingness or unusual values. In such instnaces, the `mice` package was used for data imputation.

3.2 Modeling Approaches

3.2.1 Data Partitioning

For the purpose of predictions or testing on the machine learning algorithm models or estimation of the evaluation metrics for the best model selection, the entire dataset were partitioned into training data and testing data. The training data would be used to fit the models and the testing data was used to derive the metrics for model selection or testing.

3.2.2 Predictive Models

In this part of the analysis, since the nature of the task present a classification problem, we fit several classification models to the target variable as a function of all the possible predictors we selected from the original dataset. All the models considered are *supervised* learning algorithms. Candidate models included Logistic regression, Linear Discriminant Analysis (LDA), K-nearest neighbors (KNN), Support Vector Machines/Classifiers, Multivariate Adaptive Regression Splines.

3.2.2.1 Logistic Regression   Logistic Regression is very simple and one of the mostly used traditional machine learning algorithms. It is a statistical model for predicting binary classes. However, as remarked by James, Witten, Hastie, and Tibshirani ([2013](#)), multiple-class extension of Logistic Regression exists but *Discriminant Analysis* models are commonly used in place of it. The dependent variable in here follows Bernoulli distribution. Logistic Regression model uses the logistic function or sigmoid function for predictive modeling of the given problem, which takes value between 0 and 1. 1 will be predicted if the curve goes to positive infinity and 0 if it goes to negative infinity. The logistic Regression model performs the predictive analysis based on the relationship between the binary dependent variable and the other one or more independent variables from the given dataset.

3.2.2.2 K-Nearest Neighbors Classifier   K-nearest neighbors regression (KNN regression) is one of the simplest and best-known non-parametric methods. Unlike the Logistic Regression, no assumptions are made about the shape of the decision boundary. Therefore, we can expect this approach to dominate Logistic Regression when

the decision boundary is highly non-linear. It uses a distance metric such as the Euclidean distance for separation between points in the feature space. In the KNN classification model, the prediction is purely based on neighbor data values without any assumption on the dataset. The K in the name of the model represents the number of nearest neighbor data values. This parameter can be tuned to find an optimal value. Based on K, the decision is made by the KNN algorithm on classifying the given dataset.

3.2.2.3   Support Vector Machines (SVM)   Support Vector Machines; SVM is a great classification machine learning model for classifying data of clear classes. Also does not require a predetermined cutoff point for prediction. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall (Wikipedia, 2021). We will consider using this classification model because it is versatile and effective in high dimensional spaces. They are versatile in the sense that different Kernel functions can be specified for the decision function. Common kernels include polynomial kernels and radial kernels. Support vector classifiers are also able to address the problem of possibly non-linear boundaries between classes. We will try at least three kernel SVM for this analysis noticeably, the Linear SVM and two Non-Linear kernels SVM such as the Gaussian radial basis function (RBF) and Polynomial kernel SVM through packages such as caret and kernlab.

3.2.2.4   Multivariate Adaptive Regression Splines   We employed one MARS model with an appropriate set-up of parameters as another baseline classifier. The MARS algorithm is also suitable for a large number of predictor variables, robust to outliers, it automatically detects interactions between variables, and despite its complexity, it is an efficient fast algorithm. However, it is susceptible to overfitting, more difficult to understand and interpret as against others, and not good with missing data.

3.3   Model Validation

The k-fold cross-validation resampling technique was used to validate our prediction models.

3.4   Performance/Evaluation Metrics

The following metrics were employed to examine the performance of the models for best predictive power.

3.4.1   F-beta Score (F_β)

The F-beta score or F-beta measure is the weighted harmonic mean of precision and recall, reaching its optimal value at 1 and its worst value at 0. It turns the F-measure into a configurable single-score metric for evaluating a binary classification model based on the predictions made for the positive class. The beta parameter determines the weight of recall and precision in the combined score. $\beta < 1$ lends more weight to precision, while $\beta > 1$ favors recall. That is, a smaller beta value, such as 0.5, gives more weight to precision and less to recall, whereas a larger beta value, such as 2.0, gives less weight to precision and more weight to recall in the calculation of the score. It is a useful metric to use when both precision and recall are important but slightly more attention is needed on one or the other, such as when false negatives are more important than false positives, or the reverse (Machine Learning Mastery (2021), Scikit Learn (2021)).

We can compute the F-beta score by

$$F_\beta = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall},$$

where **precision** and **recall** are defined below.

The choice of the beta parameter will be used in the name of the F-beta score. For example, a β value of 1 is referred to as the F_1-measure or the F_1 score. A β value of 2 is referred to as F_2-measure or F_2-score.

Precision

Precision is a metric that quantifies the number of correct positive predictions made. It is computed as the ratio of correctly predicted positive classes divided by the total number of positive classes that were predicted as in the formula below.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{True Positives}}{\text{Total Predicted Positives}}$$

Recall

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. The intuition for recall is that it is not concerned with false positives and it minimizes false negatives (Machine Learning Mastery, 2021). It is calculated as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{True Negatives}} = \frac{TP}{TP + NP}$$

3.4.2 Area Under the Receiver Operator Curve (AUROC)

The AUROC curve is a graphical illustration of the performance of the prediction model. The ROC curve is the relationship between the recall and precision over varying threshold values. The threshold is the positive predictions of the model. The AUROC curve is plotted by keeping the x-axis a false positive rate and the y-axis as a true positive rate. Its value ranges from 0 to 1 (Theerthagiri, Jeena Jacob, Usha Ruby, and Yendapalli, 2021).

3.4.3 Model Selection

Base on the metrics calculated using the trained models and testing data for each classifier algorithm, the best model was selected, for instance, the model with the highest AUROC be best among the others, however, AUROC alone is not the measure for selection, hence a table other performance measures together with the AUC were used to estimate the one with the best performance across all metrics.

3.5 Model Deployment

To aid easy deployment and implementation of our proposed model(s), we planned to develop a web application with the R Shiny App package.This dashboard was to serve as an effective form of communicating our results to our client and other audiences. This was to provide among other features the flexibility to interact with the visualizations and models to assess performance based on changing parameters. We were aiming at an application similar to the online prediction tool developed by Dominguez-Rodriguez et al. (2021). Unfortunately, due to time constraint we were unable to develop such a communication tool.

4 Background

In this section, we provide an overview of Food and Housing Insecurity in Colleges in the U.S. in addition to the challenges faced by households in America. Also, we provide a description of the surveyed data obtained by The University of Texas at El Paso, processing of data cleaning and variable description.

This section primarily provides a background to the study by means of an overview of food insecurity and housing insecurity in the US in general and at El Paso, the satellite region of our target population, in particular.

4.1 Overview of the study

This project is aimed at analyzing data based on a survey constructed in summer 2019 and 2020 by a team of researchers at UTEP (Moya, Crouse, Schober, Wagler) to assess the state of food and housing insecurity among UTEP students. A slightly modified version of the survey was administered summer 2020. In this work, we seek to assess whether there are any changes to food and housing insecurities since 2019. This is particularly important

since the worldwide pandemic hit during this period. The primary and secondary research questions are listed below.

1. **Primary Research Question**: What factors are associated with food insecurity (FI) and housing insecurity (HI) among UTEP students?

2. **Secondary Research Question**: Which subpopulations of students are most at risk for FI and HI at UTEP?

There are two phases to the project, phase I is concerned with the primary research question while phase II will be based on the secondary research question.

4.2   Overview of Food Insecurity

Food insecurity is defined by the United States Department of Agriculture (USDA) as the lack of consistent access to enough food for active healthy life. It can also be defined as a physical discomfort, thus the definition can relate to an individual or a family's lack of resources to obtain enough food for a healthy living. With the increasing spread of COVID-19, and many Americans losing their jobs, one of America's challenge worsened: hunger.During this period, communities and colleges in America in particular faced the longest queue's for food compared to past years.

According to the USDA, 13.7 million or 10.5% of all households in the U.s. experienced food insecurity at some point during the COVID-19 pandemic. In addition, it was observed that, nearly 40% of college students were food insecure. That means,24 of every 60 students are food insecure. After several years of research on food insecurity, researchers discovered food insecurity as a likelihood to deter college students from completing their degree.

Further research from USDAS stated over the past 20 years and a study completed in 2017 showed that, racial minorities, first-generation students, low income students, student with children and LGBTQ+ students are highly affected by COVID-19. These variables were included in analyzing the factors affecting food insecurities at UTEP.

4.3   Overview of Housing Insecurity

Housing insecurities includes difficulty in paying for living expense, overcrowding, substandard living and frequently moving out. Though housing insecurities can affect anyone, lower income families are known to be highly impacted since they pay high proportion of their lower income on high cost rent.

According to a 2019 report by The Hope Center for college, Communities and Justice, nearly 3 out of 5 students experience having insecurities in 2018. Also, 14% of 4 year student reported experiencing homelessness. Despite the many challenges of housing insecurities, the impact of COVID-19 have increased the likelihood of homelessness among college students. For example, many students lost their jobs and financial due to COVID-19 pandemic.

4.4   Data Description

4.4.1   Data cleaning and naming of variables

In this section, we provide a brief description of the data for prediction and a method for renaming and collapsing the variables. The initial data from the survey comprises 12,536 observations with 104 columns. From the data, we observed enrollment and employment as the variables with the minimum number of missing values. In preparation of our data for modeling, we excluded the following questions from the our analysis: respondent ID, the age of the respondent, what pronouns do you use to describe yourself, where do you live, and variables associated with changes due to COVID-19. Table 1 shows the recoded names of the variables we considered in our analysis and their corresponding descriptions based on the survey questions.

Table 8:   Table of recoded variable names from survey data

Variable Name	Description
enrollment	Your enrollment at UTEP (Q1)
employment	Are you employed (Q2)
employment_type	You are consistenly working at which employment type (Q2)
weekly_work_hrs	How many hours a week are you consistently working (Q3)
ethnicity	What is your ethnicity (Q4)
gender	Which gender do you identify with
total_income	The The estimated income for your household in 2019 (Q9)
academic_level	What is your academic level (Q10)
college/school	Which College/School are you a student of (Q11)
mode_transport	The common mode of commute (Q12)
transport_reliability	How reliable is it getting you to/from college (Q13)
living_alone	Do you live alone (Q14)
dependents	Number of dependents (Q15 and Q16 combined)
household_head	Are you the head of the household (Q17)
residence	Where do you live (Q19)
permanent_address	In the past 12 months, have you had a prmament adress (Q20)
spent_night_elsewhere	How frequently did you spent the night somewhere due to lack of housing (Q22)
know_homeless_student	Do you know UTE students who have experienced homelessness (Q23)
federal_student_aid	Have you received financial aid in the past 10 months (Q25)
FI_q26	The food that I bought just didn't last. and I didn't have enough money to get more (Q26)
FI_q27	I couldn't afford to eat balanced meal, how often (Q27)
FI_q28	Did you ever cut the size of your meal (Q28)
FI_q30	In the pst 12 months, have you eat less than you felt you should because there was not enough money (Q30)
FI_q31	In the past 12 months, were you hungry but didn't eat because there wasn't enough money for food (Q31)
expenditures_changed	Since COVID-19, has there been any personal or household change in expenditures (Q32)
income_changed	Since COVID-19, has there been any personal or household change in financial Income (Q33)
fed_aid_changed	Since COVID-19, has there been any personal or household change in financial Aid (Q34)
debt_changed	Since COVID-19, has there been any personal or household change in debt (Q35)

4.4.2   Collapsing the variables

In other to obtain a feasible result, the following variables were collapsed due to insignificant proportion of respondent answering those questions. With regards to **Ethinicity**, we maintained Hispanic , Black/African American, White/Caucassian, Asian and collapsed American Indian, Native Hawaiian, mixed race and other into **other** . We decided not to keep a mixed race level because the total reduced to 88 after we assigned (Hispanic, white) to Hispanic, and (Hispanic, Black) to Black. **Gender** was reduced to 3 levels: female, male, and non-binary/non-conforming. With respect to **Income** we kept first 5 levels and 100000 or more (11)->8, collapse

(6,7)->6, (8,9,10)->7. For **Academic level**, we keep first 5, and collapse last 2 -> 6. For **College**, we maintain first 6 levels and colapse the remaining ones into other (level 7). For **Commute**, we keep car, Bus/public and other, and collapse car & carpool (2,3) -> 9, collapse bike, trolley, walk (5,6,7) -> 10. Keep Not applicable (0). For, **Federal Aid**, Emergency loan was reduced from 316 to 50, so we added those to loans(3).

5 Recommendations

From the results and discussions, we make the following recommendations:

- There should be proper advertisement of food pantries and shelters if there exist some for low income students to survive on in the absence of adequate financial aid.

- With regards to multiple responses, we are of the opinion that this should not be allowed for the gender variable in subsequent surveys.

- We also recommend in later part on this work constructing a measure of food insecurity as a dichotomous response variable that combines what we consider to be four main dimensions of food insecurity embodied in five of the survey questions, namely, a stable source of food , lack of healthy/balanced meals, inadequate size of food – eating less and going hungry . Using this definition as proposed by the USDA, we selected $Q26, Q27, Q28$ and $Q30$ respectively as our response variable for food insecurity. Therefore, an individual who participated in the survey is classified as being at risk of food insecurity if they answered yes to all of questions $Q26, Q27, Q31$, and answered yes to either $Q28$ or $Q30$ since these two questions are similar in terms of measuring a single dimension of food insecurity. Thus, an individual who reports four conditions that indicate food insecurity are classified as "food insecure".

6 Data Science Best Practices Adopted

We sought to improve upon our initial work from the first phase of the project by incorporating three key practices in terms of data science ethics and processes that we have learned from the Data Science Collaboration course and beyond. The three practices we focused on include programming most aspects of the project, ensuring reproducibility of our entire work, and improvements to the communication of our results. It is interesting to note that due to the shared conviction of all members of the group to always uphold best practices right from the beginning of any data science project they set forth to emback on, you would find out that most of the practices we talk about here, especially the programming and reproducibility, carry over directly from phase I of the project. The next subsections discuss each one of these practices in details.

6.1   Programming or Process Automation

Taking our inspiration from the talk by Peng (2014) on Reproducible Research Checklist, we programmed and automated various aspects of our project in order to aid reproducibility, allow for convenient extension of the project to future survey data , and to help avoid common errors associated with manual data processing. Here, we highlight what we did and how we went about with their implementations.

6.1.1   Data Cleaning

- From the very beginning of our data processing, instead of manually specifying the file or directory paths, we used R codes to dynamically determine the paths to prevent our files reading codes from breaking down on other people's computer platforms. This helped with our collaboration on github by giving each member the convenience to clone the github repository of the project and start running the codes without having to change any file paths. To achieve this, we used the `rstudioapi::getSourceEditorContext` function to get the file path and extracted the directory name with the help of the `dirname` function.

- **A program to remove unwanted parts of variable names** One will notice that the original variable names from the excel data file included the actual survey questions which does not make it easy for reading and processing. Instead of manually editing the names in the excel file, we used regular expressions with the help of the R pattern matching and replacement function `gsub` to automatically strip off the redundant character strings in the original variable names to facilitate the initial stages of our data cleaning. At this stage we needed just the question number part of the variable names so we took advantage of the fact that each question number ended with two dots followed by the unwanted characters after importing the data into R. The regular expression pattern used is `"\\..*"`, which helped us match all characters after the two dots and replacing them with an empty string to get rid of the unwanted parts of the variable names.

- We also created a program to track variables that were part multiple response survey questions. For example, because multiple responses were enabled for ethnicity on the survey, the dataset provided to us had separate columns for the each level of ethnicity. The goal then was to consolidate the individual parts belonging to the same survey question into one variable. We first created a count variable to record the total number of **non-na** responses for each question. We then used two helper functions we created called `na_to_zero` and `zero_to_na` to convert all the **NA's** to zeros, and to restore the **NA's** in the end, respectively.

6.1.2   EDA

The exploratory data analysis section also received a lot of programming. Functions were written to create graphs and summary statistics tables. One area we want to talk about is where we wrote R codes to dynamically alternate the background colors of table rows to make it easy to read all the records pertaining to a particular variable in

Table 2 of the report. Instead of manually specifying the row indexes in the *stripe_index* argument of `kable` function to show where the records for a variable begins and ends, we did this programmatically by creating a `get_stripe_index` function.

6.1.3   Modeling

Codes for training models and generating reports for the responses were manually duplicated for each response in our phase I report. We decided to improve upon this area because of the potential dangers of having to make changes at many different places when there comes the need to modify aspects of the modeling processes. With some programming, we were able to abstract the modeling and results reporting codes into functions. These functions are available in `modeling_functions.R` file in the github repository. First, there is a function called `modeling` for training the models for different responses. This function takes in the name of the response variable (a character string), model list, and a k-fold (an integer value for the number of cross-validations) as argument. Other functions created include evaluation metrics and results table functions, as well as a variable importance function.

6.2   Reproducibility

Results or project reproducibility has become a major concern for data scientists, and the scientific and research community. To this end, several authors have taken it upon themselves to solve this issue of reproducibility. For instance, Yu and Kumbier ([2020](#)) proposed a framework known as PCS which discusses reproducibility through the stability component.

6.2.1   Version control

In this work, we used Github as the version control to ensure reproducibility and collaboration by team members. Version or source control is the practice of tracking and managing changes to projects. These version controls help teams to keep track and modify their code in a special kind of database. This ensures effective collaboration of team members since teams can track back and compare earlier versions of code. By using version control, we were able to document a complete history of every file and code. These changes include the creation and deleting of files and editing of codes. In addition, team members can branch and merge. For example, to ensure faster completion of the project, we worked independently on "branches" and merged our files together after completion.

6.2.2   R Rackage Management

Being able to reproduce a work in the R Statistical Software requires one to have the required packages and attach them to the R environment. However, it can become a problem keeping track of the many packages used in a project, so we organized all packages that our project depended on and employed a package manager called `pacman` (Rinker and Kurkiewicz, [2018](#)) for managing add-on packages to load them. The benefit is that this package manager will

automatically install packages required but not available on the user's version of R, and then load the packages.

6.3   Communication

Our findings and procedure were communicated through a written document in Rmarkdown. Also, our analysis results and report was presented through the concept of flipping the paradigm of communication by Mannshardt (2021) as follows;

- Discussion

- Results comprising EDA and Model results

- Methodology and Data Preprocessing

- Background

7   Reflection

We employed several data science rules and guidelines that ensure a veridical data science project. One way for data scientists to maximize the impact of their research or project is to improve their collaborations with domain experts. Throughout the data cleaning, data manipulation, modeling, and communication phase, we ensured collaboration among team members and domain experts and enforced good data science practices and guidelines in our work. In this section, we provide ways in which these concepts we learned in class helped us work on this project. Thus, we emphasize the impact of ASCCR by Vance and Smith (2019) to our work. Those guidelines include but are not limited to ensuring reproducibility, communication, good modeling practices, and this project's adequate documentation.

For the duration of this project, we had the opportunity to grow as young data science professionals by developing a solid relationship with the domain expert. During the variable screening, the domain expert provided advice on extracting some of the variables from the survey questions and guided the team in ways to achieve efficient results. This advice and guidance inspired us to produce an impactful research result by our team. We achieved this feat by employing the ASCCR framework, which ensures a good relationship between the domain expert and data scientist. This was since we had good relationship with the domain expert.

In addition, content in ASCCR ensures that the data scientist includes Qualitative, Quantitative, and Qualitative components in data science projects. For example, while working on this project, we selected the appropriate quantitative or modeling approaches to use since the domain experts provide a refined formulated domain problem. Having a clear understanding of the domain problem is imperative to select the proper data cleaning, imputation, and modeling approaches.

Furthermore, communication under ASCCR ensured that data scientist apply practical skills in their collaboration with domain experts to understand the results of the project. In general, we ensured effective communication of our model results to expected audience. For example, the paradigm for presenting academic work and project is literature review, proof and formulas, new work, results, and the decision. But we observed that, the appropriate way to present our project report is to flip the paradigm (Mannshardt, 2021): results, new work, proof and formulas, background, and history. We observed that the audience (decision makers, policy makers, among others) plays important role on how to present results.

Moreover, we used version control and GitHub to improve on collaboration and reproducibility of our codes. We observed that GitHub enabled us to work independently on the project and keep track of any code changes. Also, it provided a secure way to store the codes for the project.

Finally, we see other changes that we made including the development of functions for flexibly training models and generating reports of the results as a great addition to the project. This is because there could be the need for

people who gain access to our project through the GitHub repository to want to try other models or choose to model other responses. With the help of these functions, it should be easy or possible for anyone to effortlessly accomplish this. All that need to be done will be to specify a single model name or a list of models supported by the caret package in R, and with a little tweak to the modeling functions one can also model any other variable from the survey data as a response variable.

References

Dominguez-Rodriguez, S., Villaverde, S., Sanz-Santaeufemia, F. J., Grasa, C., Soriano-Arandes, A., Saavedra-Lozano, J., . . . others. (2021). A bayesian model to predict covid-19 severity in children. *The Pediatric Infectious Disease Journal*, *40*(8), e287–e293.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Machine Learning Mastery. (2021). A gentle introduction to the fbeta-measure for machine learning. Retrieved October 15, 2021, from https://machinelearningmastery.com/fbeta-measure-for-machine-learning/

Mannshardt, E. (2021). What's our point? Flipping the paradigm for communication in statistics and data science. *Amstat*.

Peng, R. D. (2014). *Youtube video: Rproducible Research Checklist*. John Hopkins. Retrieved from https://www.youtube.com/watch?v=VOaN33aAcEw&ab_channel=RogerPeng

Rinker, T. W., and Kurkiewicz, D. (2018). *pacman: Package management for R*. Buffalo, New York. Retrieved from http://github.com/trinker/pacman

Scikit Learn. (2021). Retrieved October 15, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html

Theerthagiri, P., Jeena Jacob, I., Usha Ruby, A., and Yendapalli, V. (2021). Prediction of covid-19 possibilities using k-nearest neighbour classification algorithm. *Int J Cur Res Rev| Vol*, *13*(06), 156.

Vance, E. A., and Smith, H. S. (2019). The asccr frame for learning essential collaboration skills. *Journal of Statistics Education*, *27*(3), 265–274.

Wikipedia. (2021). Support-vector machine. Retrieved October 15, 2021, from https://en.wikipedia.org/wiki/Support-vector_machine

Yu, B., and Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, *117*(8), 3920–3929.

Apendix

7.1 Visualizing Missing Data
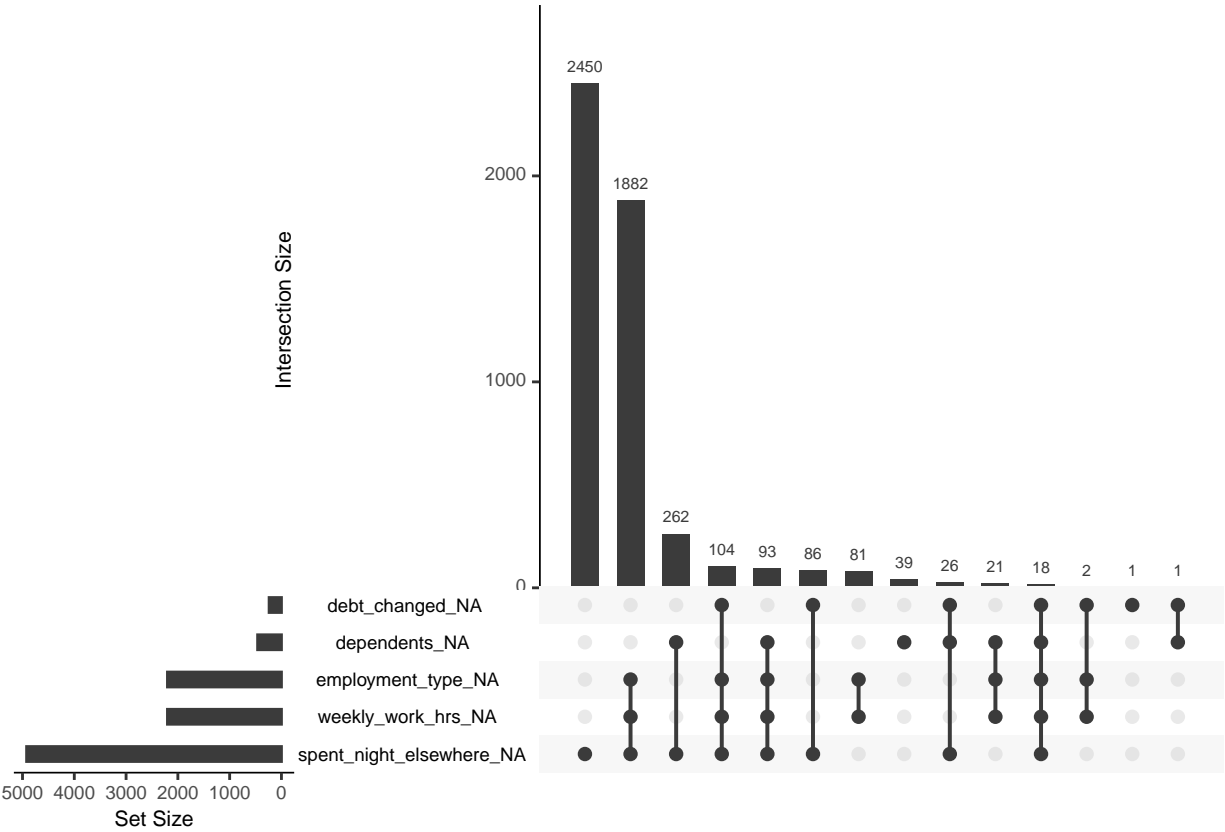
variable	n_miss	pct_miss
spent_night_elsewhere	4921	95.0917874
employment_type	2201	42.5314010
weekly_work_hrs	2201	42.5314010
dependents	460	8.8888889
expenditures_changed	238	4.5990338
income_changed	238	4.5990338
fed_aid_changed	238	4.5990338
debt_changed	238	4.5990338
FI_q30	177	3.4202899
FI_q31	177	3.4202899
FI_q26	157	3.0338164
FI_q27	157	3.0338164
FI_q28	157	3.0338164
know_homeless_student	115	2.2222222
federal_student_aid	115	2.2222222
residence	80	1.5458937
permanent_address	80	1.5458937
household_head	56	1.0821256
gender	22	0.4251208
respondent_id	0	0.0000000
enrollment	0	0.0000000
employment	0	0.0000000
ethnicity	0	0.0000000
total_income	0	0.0000000
academic_level	0	0.0000000
college/school	0	0.0000000
mode_transport	0	0.0000000
transport_reliability	0	0.0000000
living_alone	0	0.0000000
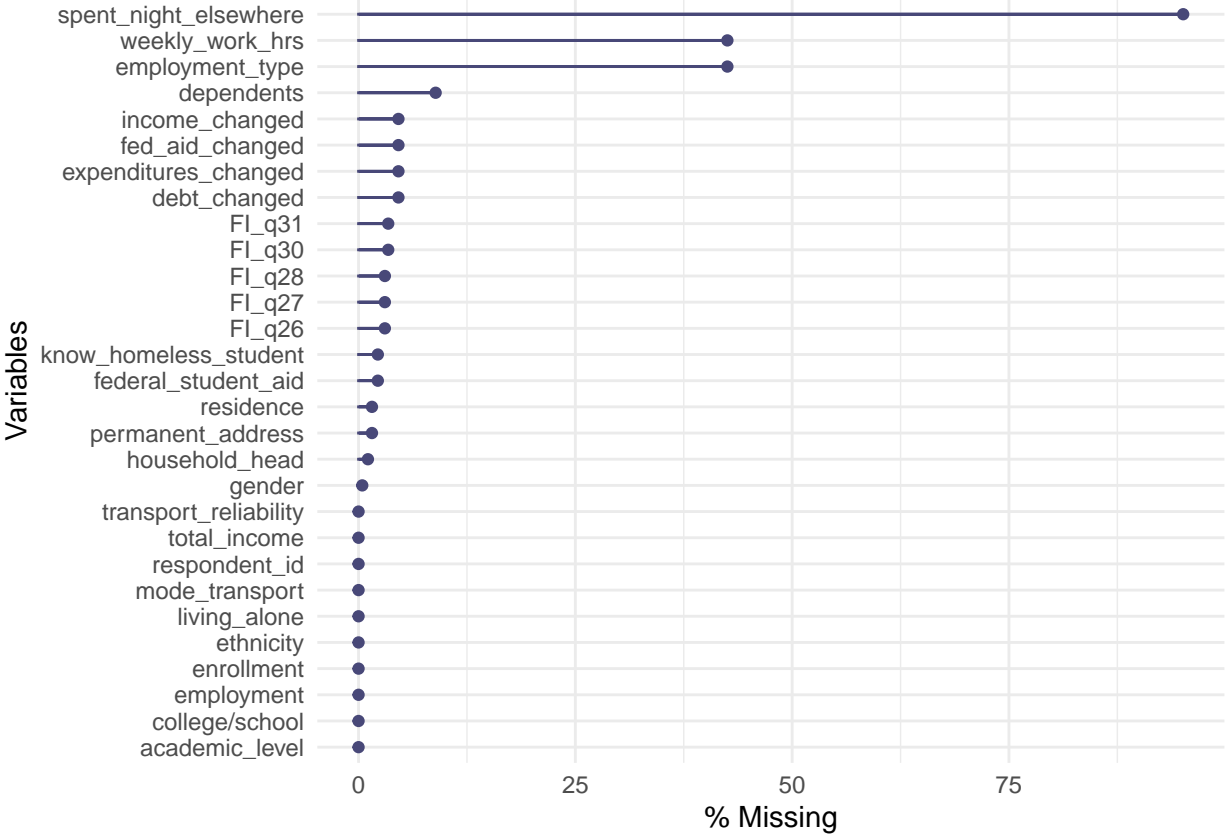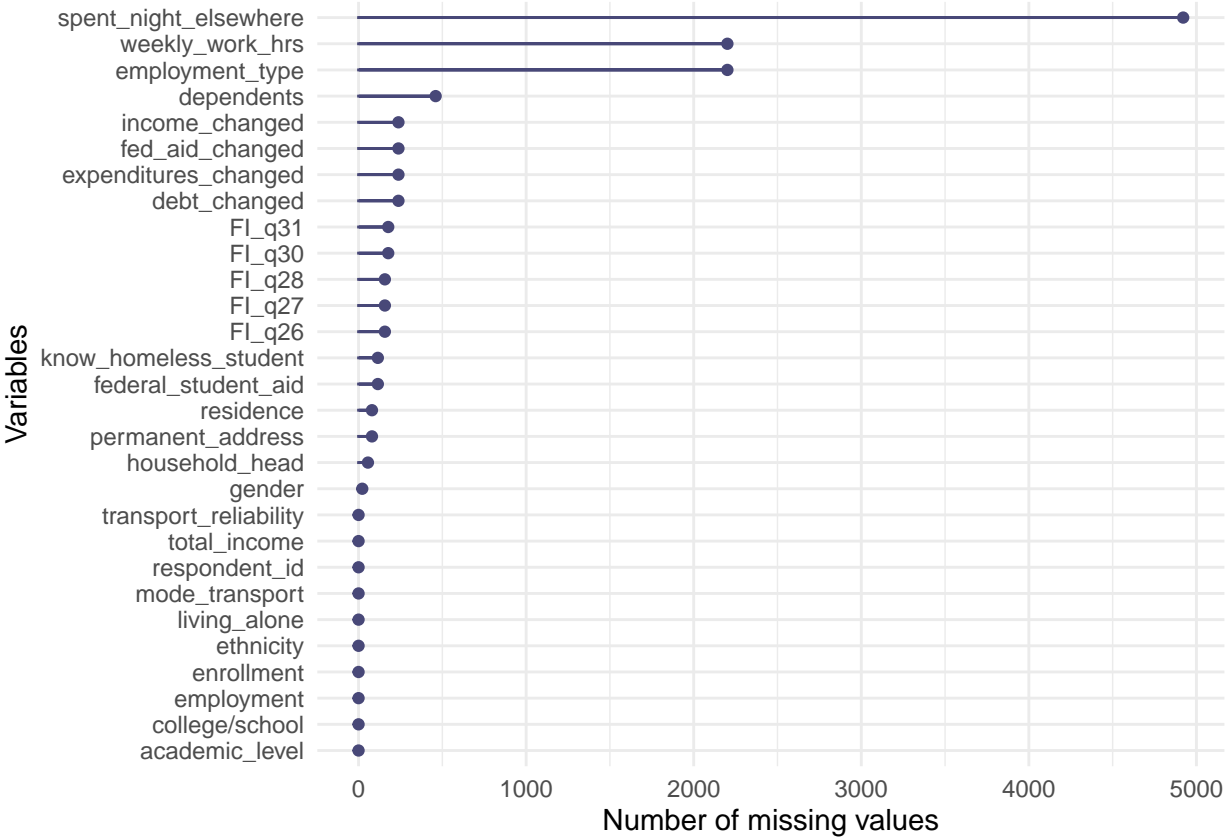
7.2 Codes

Codes for the analysis is available via a github repository upon request.