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 Introduction

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.

1.1 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.

1.2 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.

1.3 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.

2 Data Preprocessing

2.1 Data Description

2.2 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 1: 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)

2.3 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).

3 Analysis Plan

This document serves to provide a road map to the approaches our group intend to adopt and execute in terms of our modeling approaches and general outline of our final report. We, however, remark that the various plans set forth are subject to changes according to recommendations and suggestions received from our domain experts (Dr. Amy Wagler and her research colleagues on this same project) upon consultations, and any other future modifications that we shall consider appropriate.

3.1 Exploratory Data Analysis

In this section, we will provide 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 will primarily be looking at the distributions of our target variable (death event) and all the predictor variables.

In terms of the types of graphs or visualizations, I would like my group to consider creating comparative boxplots/histograms, segmented or comparative bar graphs and, possibly, density plots. We will critically assess these graphs and settle on a few which will help us to effectively tell a compelling story about the data. We will build animation effects into some of the graphs where possible or appropriate.

- **Cross Tabulations (Contingency Tables)**: Provide numerical summary of how responses differ by some categories.
- **Comparative Boxplots/Histograms/Density plots**: These graphs will be used to describe the distribution of the continuous or numeric variables across the levels of the target variable.
- **Bar graphs (Segmented/Comparative bar graphs) and Mossaic Plots**: To explore how the available categorical predictors are distributed across the levels of the target or response variable.
- **Correlation plots** will be utilized to detect the presence of multicolinearity among the continuous predictors.

3.1.1 Missing Data/Outliers and their treatments

The available data will be inspected for missing and unusual values. Any missing data and/or outlying data identified from our data preprocessing and exploratory analysis will be treated appropriately depending on the nature of the missingness and the information we are able to gather about the possible reasons for the missingness or unusual values. Possible methods to likely to be considered include imputation, listwise deletion, among others.

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, a division of the entire into say training data and testing data will be carried out. 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. This partitioning will be done by using the strategy of V -fold cross-validation on the misclassification errors, with V = n, where n would specified base on the sample size of the entire data set. We will randomly divide the data into V folds with stratification on the target variable category. This assures that similar proportions of 0s and 1s are preserved in each fold, and set.seed() will be used for easily reproducible results.

3.2.2 Predictive Models

In this part of the analysis, since the nature of the tasks present a classification problem, we will fit several classification models to the target variable as a function of all the possible predictors in the dataset that we will deem fit. All the models considered are *supervised* learning algorithms.

Candidate models include Logistic regression, Linear Discriminant Analysis (LDA), K-nearest neighbors (KNN), Support Vector Machines/Classifiers, Naive Bayes' algorithm, Multivariate Adaptive Regression Splines, Artificial Neural Network, and Tree-based models such as Random Forest, Bagging, or Boosting. Due to the sensitivity of the KNN model to different scales, the continuous variables will be normalized and kept across the other models for consistency, that is, if we end up fiting a KNN model.

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.

Binary Logistic Regression, we will fit a regularized logistic regression model such as LASSO as one of the baseline classifiers for comparison. LASSO does not require the target variable to be normally distributed, no homogeneity of variance assumption is required, and with the aid of graph and output interpretation is effective under LASSO. LASSO, however, requires more data to achieve stability and is effective mostly on linearly separable.

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 Classification Trees According to (James et al., 2013), for a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. In interpreting the results of a classification tree, we are often interested not only in the class prediction corresponding to a particular terminal node region, but also in the class proportions among the training observations that fall into that region. A recursive binary splitting is used to grow a classification tree. Classification trees are often preferred to other classifiers for the following reasons. One thing, they are non-linear classifiers which do not require the data to be linearly separable. Two, they are easy to read and very easy to explain to people. After a model is generated, it's easy to report back to others regarding how the tree works. Moreover, trees can be displayed graphically, and are easily interpreted even bya non-expert (especially if they are small). Also, decision trees can easily handle qualitative predictors. However, trees can be very non-robust. In other words, a small change in the data can cause a large change in the final estimated tree. For this reason, we will use tree models that help circumvent this challenge and greatly improves the predictive performance by aggregating many decision trees. Such tree models include Random Forest, Boosting, and Bagging.

3.2.2.4 Random forest The random forest is a great predictive model for high dimensional data, and we will use RF as one of the baseline classifiers model. RF has high performance and accuracy with regards to modeling, provides feature importance of estimates, and can automatically handle missing values with no scaling required. However, given that RF has some advantages over other classifiers, its prediction time is high, can overfit the data, and also require more computational time and resources.

3.2.2.5 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.

3.2.2.6 Naive Bayes Classifier It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It works on Bayes theorem of probability to predict the class of unknown data sets. Naive Bayes algorithm is particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. When the assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. It should be however noted that this assumption of independent predictors may be unrealistic since, in real life, it is almost impossible that we get a set of predictors which are completely independent. Another disadvantage of the Naive Bayes classifier which might discourage us from using it is that it works only with categorical variables, so one has to transform continuous features to discrete, which might lead to the loss of a lot of information.

3.2.2.7 Multivariate Adaptive Regression Splines We will employ at least 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.

3.2.2.8 Artificial Neural Network We will fit at least two different artificial neural networks (ANN) models, specifically with different numbers of layers and different numbers of units. ANN learning algorithms are quite robust o noise in the training data set, they are often used where the fast evaluation of the learned target function is required and its ability to learn and model non-linear and complex relationships. However, the ANN classifier has a disadvantage of unexplained functioning of the network and also no specific rule for determining the structure of artificial neural networks, the appropriate network structure is achieved through experience and trial and error.

3.3 Model Validation

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

3.4 Performance/Evaluation Metrics

The following metrics will be 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 Area Under the Precision Recall Curve (AUC-PR)

The area under the precision-recall curve (AUC-PR) is a model performance metric for binary responses that is appropriate for rare events and not dependent on model specificity. Precision-recall curves have also been recognized as useful for classification performance assessment for unbalanced binary responses in bioinformatics. Like AUC-ROC, AUC-PR is a threshold-independent metric that calculates the area under a curve, where the curve is defined by a trade-off between different aspects of performance as the threshold applied to the model's predictions varies. Both ROC and PR curves are functions of the confusion matrix (Sofaer, Hoeting, and Jarnevich, 2019).

3.4.4 Model Selection

Base on the calculated metrics for each classifier algorithm, the best model will be selected, for instance, the model with the highest AUROC will be best among the others, however, AUROC alone is not the measure for selection, hence a contingent table will be obtained with classifier models on the row and metric on the column and base on all the metric estimates the one with the best performance across all metrics will be our final predictive model for analysis.

3.4.5 Model Assessment

Our final obtained classifier model will be validated using the test data provide. That is we will use our model to predict on test data given, specifically looking at some important measures such as specificity and sensitivity of the predicted outcome by the model. Also, we seek to have a prediction accuracy of approximatley 98% and above for our model.

3.5 Model Deployment

To aid easy deployment and implementation of our proposed model(s), we plan to develop a web application with the R Shiny App package.This dashboard will provide users, among other features, the flexibility to interact with the visualizations and models to assess performance based on changing parameters. We are aiming at an application similar to the online prediction tool developed by Dominguez-Rodriguez et al. (2021).

4 Analysis and Results

4.1 Exploratory Data Analysis

4.1.1 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 2: 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%
	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%

Table 2: A summary of variables of interest *(continued)*

Variable Name	Levels	Coded As	Obervations	Percentage
	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%
	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%

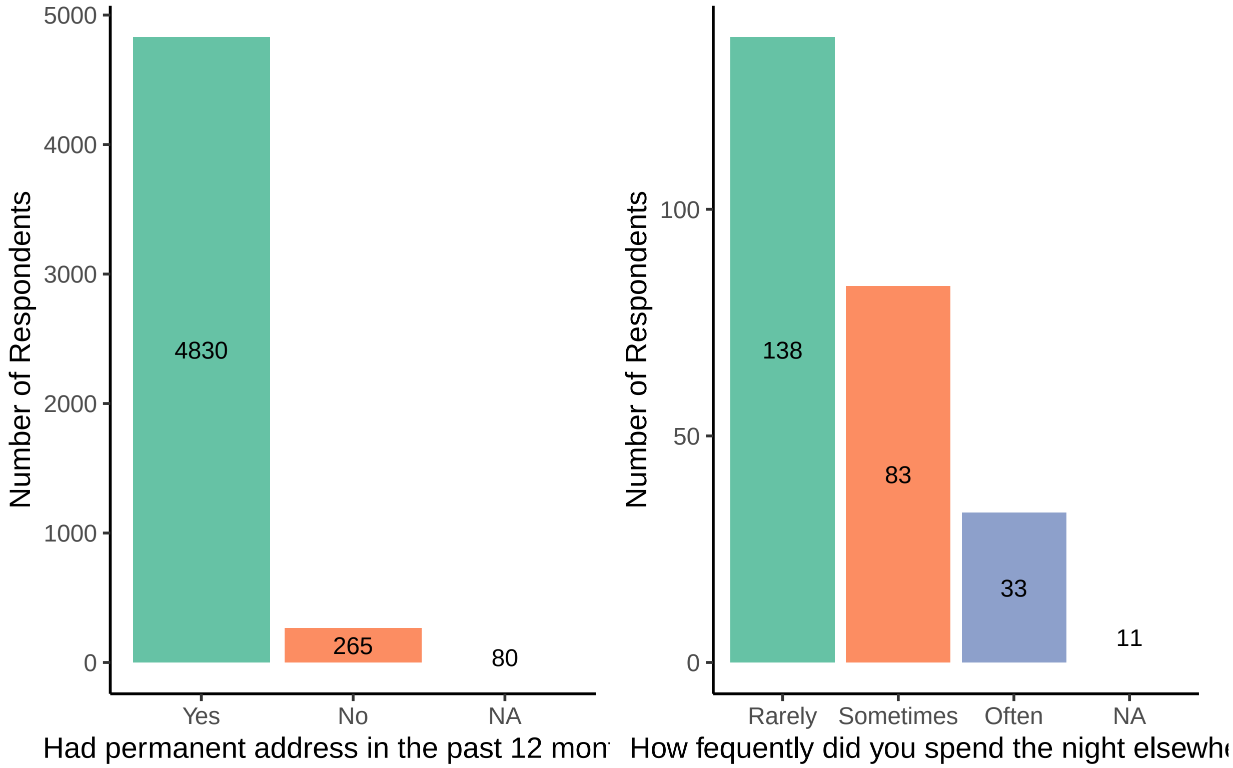
Table 2: A summary of variables of interest *(continued)*

Variable Name	Levels	Coded As	Obervations	Percentage
	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%
	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%

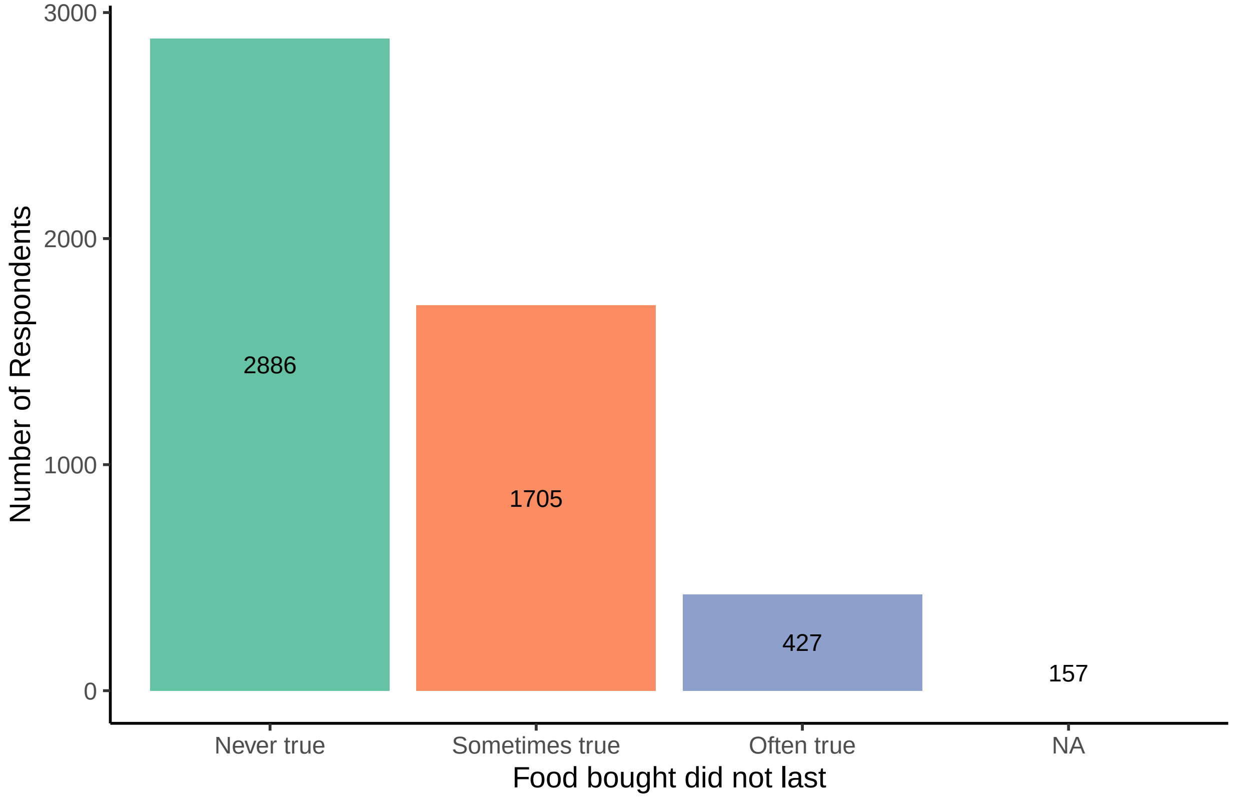
Table 2: A summary of variables of interest *(continued)*

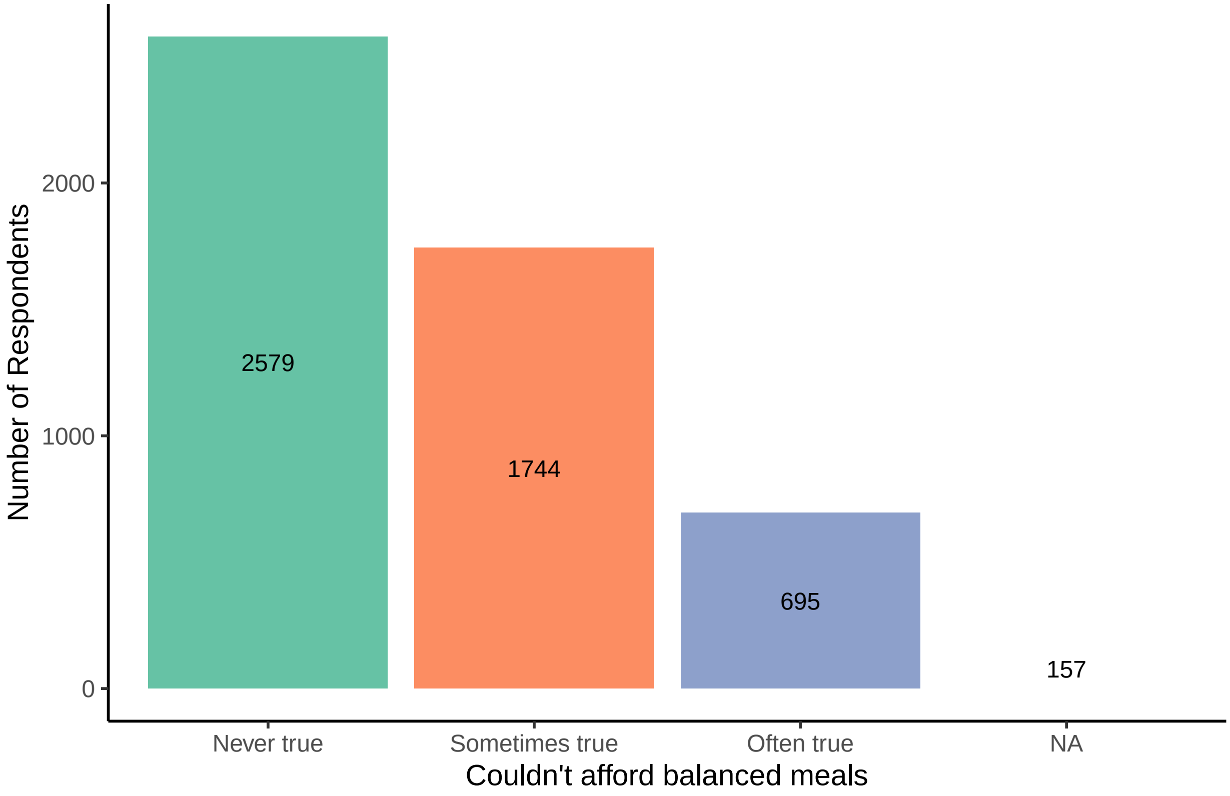
Variable Name	Levels	Coded As	Obervations	Percentage
	NA		157	3.03%
FI_q28	Yes	1	1684	32.54%
	No	2	3334	64.43%
	NA		157	3.03%
FI_q30	Yes	1	1676	32.39%
	No	2	3322	64.19%
	NA		177	3.42%
FI_q31	Yes	1	1268	24.50%
	No	2	3730	72.08%
	NA		177	3.42%
expenditures_changed	Decreased	1	1323	25.57%
	About the same	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%
	About the same	2	2441	47.17%
	Increased	3	2207	42.65%
	NA		238	4.60%

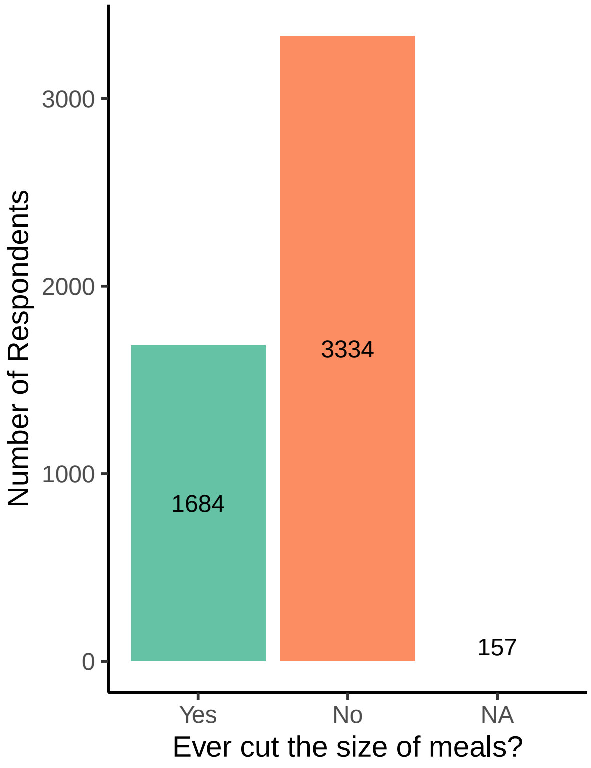
4.1.2 Distribution of Housing Insecurity Responses

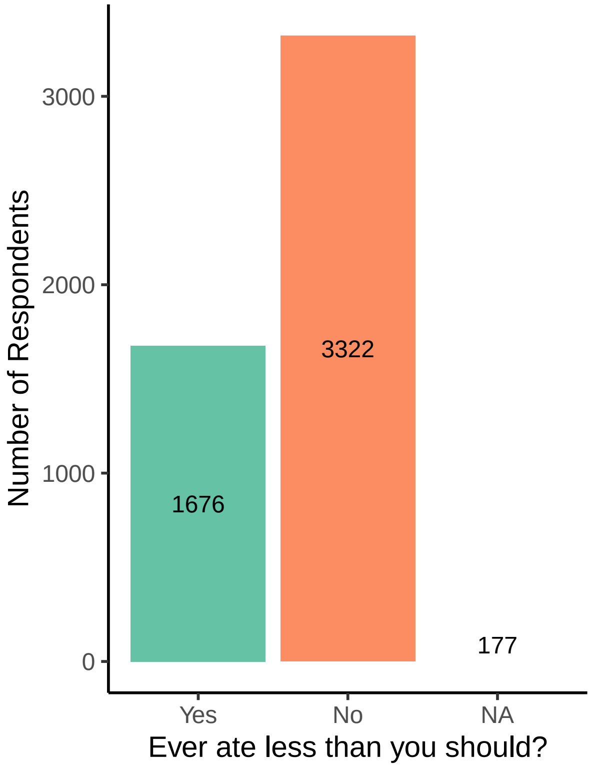


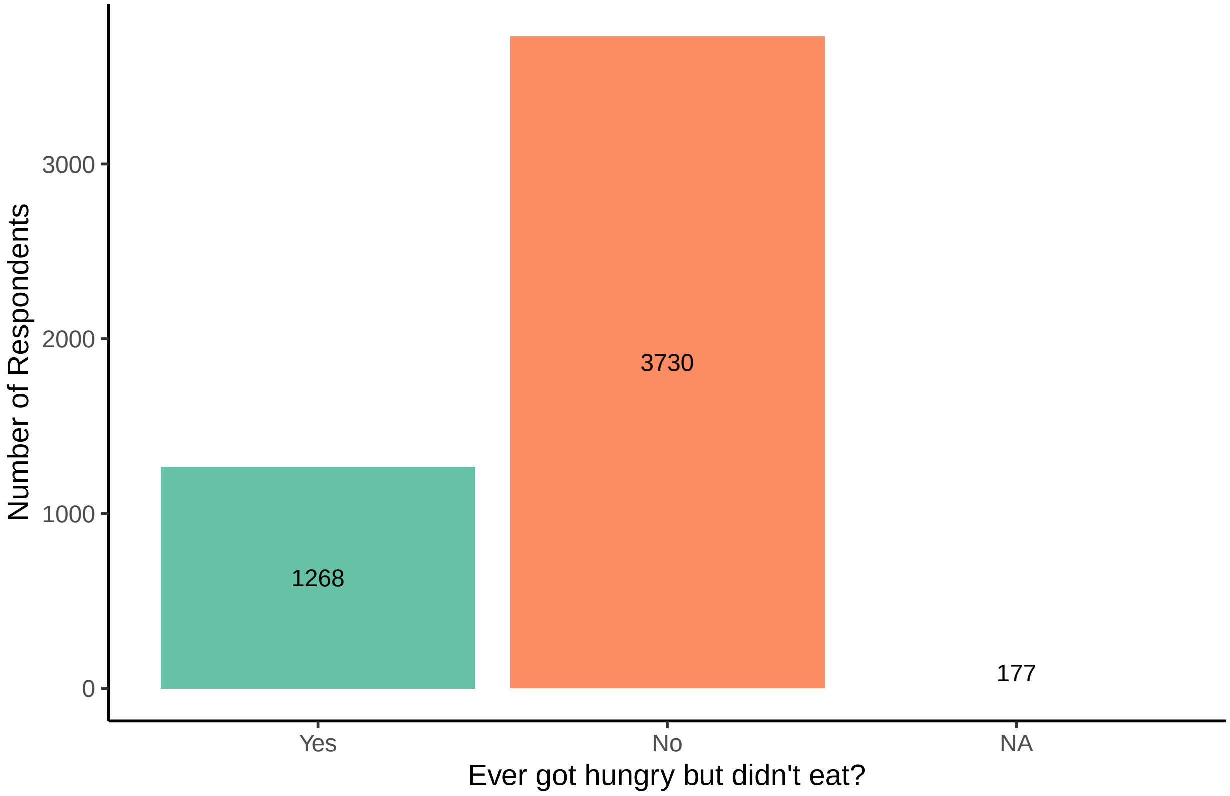
4.1.3 Distribution of Food Insecurity Responses




Number of Respondents
2000
1000
0
2579
1744
695
157
Never true
Sometimes true
Often true
NA
Couldn't afford balanced meals


Number of Respondents
3000
2000
1000
0
1684
3334
157
Yes
No
NA
Ever cut the size of meals?


Number of Respondents
3000
2000
1000
0
1676
3322
177
Yes
No
NA
Ever ate less than you should?



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.

4.2 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)

4.2.1 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.

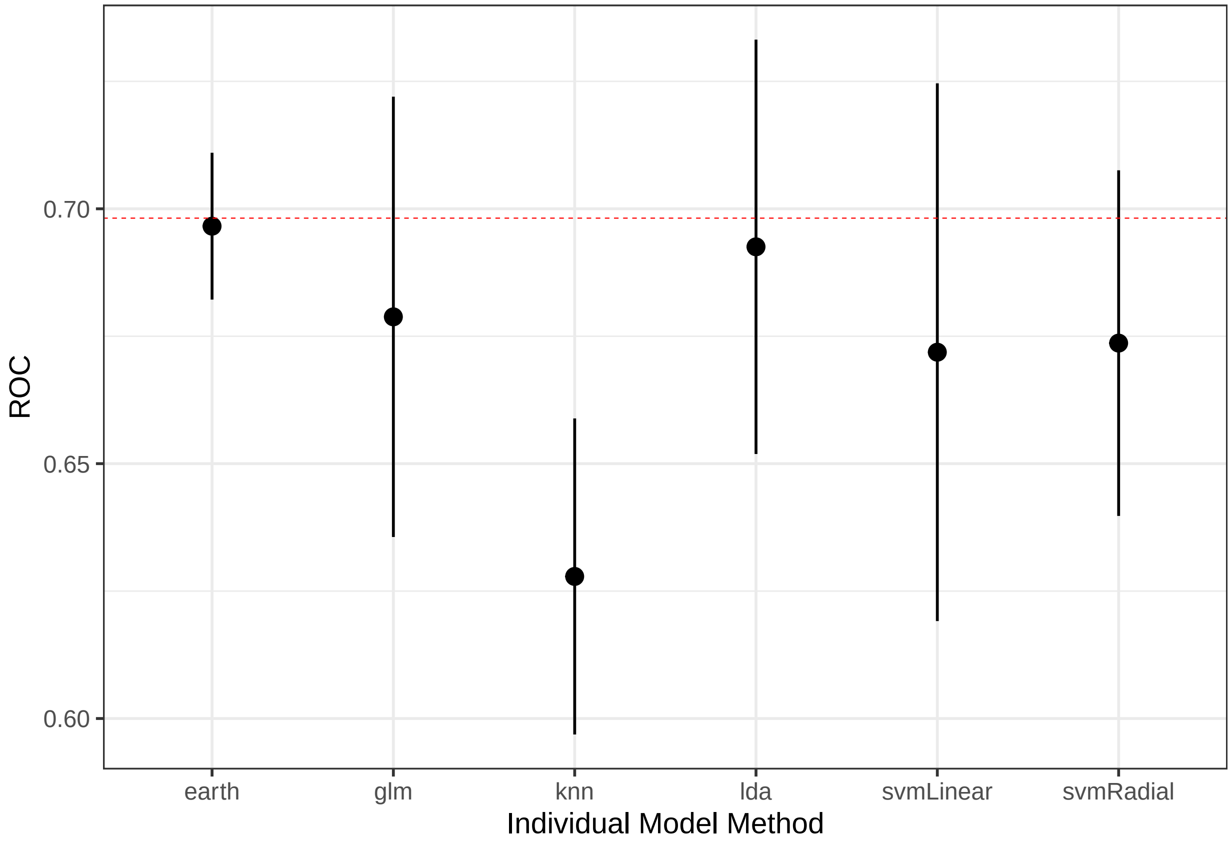
4.2.2 Modeling Housing Insecurity

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).

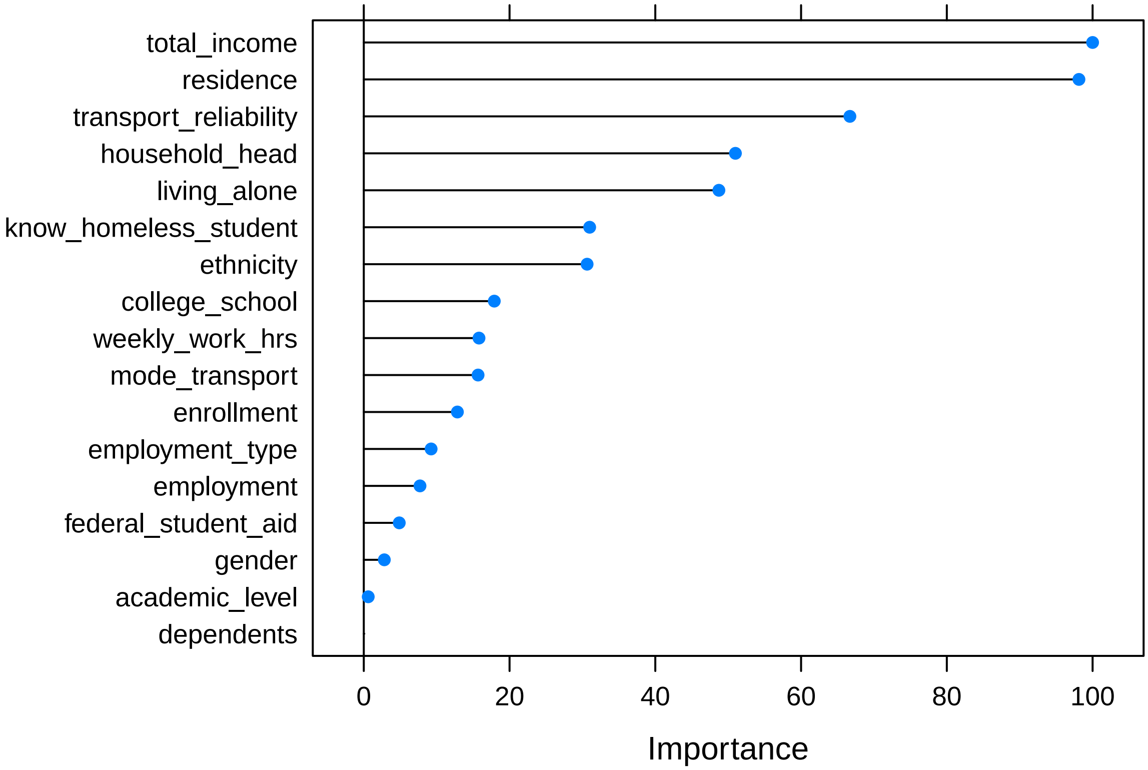
Table 3: Evaluation metrics for Permanent House Address as a response variable

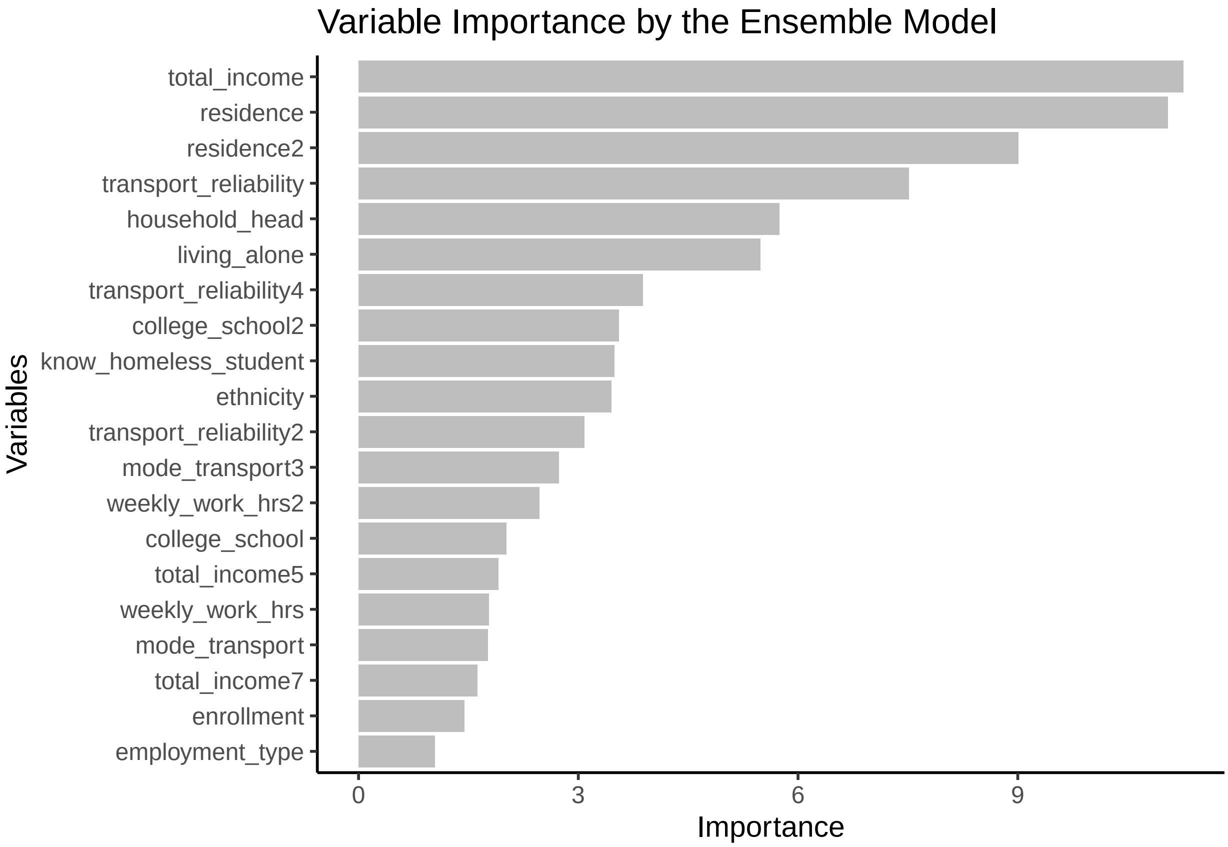
Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.24	0.76	0.77	0.53	0.86	0.7058
LDA	0.22	0.78	0.79	0.55	0.87	0.7149
KNN	0.37	0.63	0.64	0.58	0.77	0.637
MARS	0.2	0.8	0.82	0.48	0.89	0.7206
SVM Linear	0.22	0.78	0.79	0.51	0.87	0.7151
SVM Radial	0.17	0.83	0.85	0.45	0.91	0.7011

4.2.2.1 Results for Permanent Address 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.



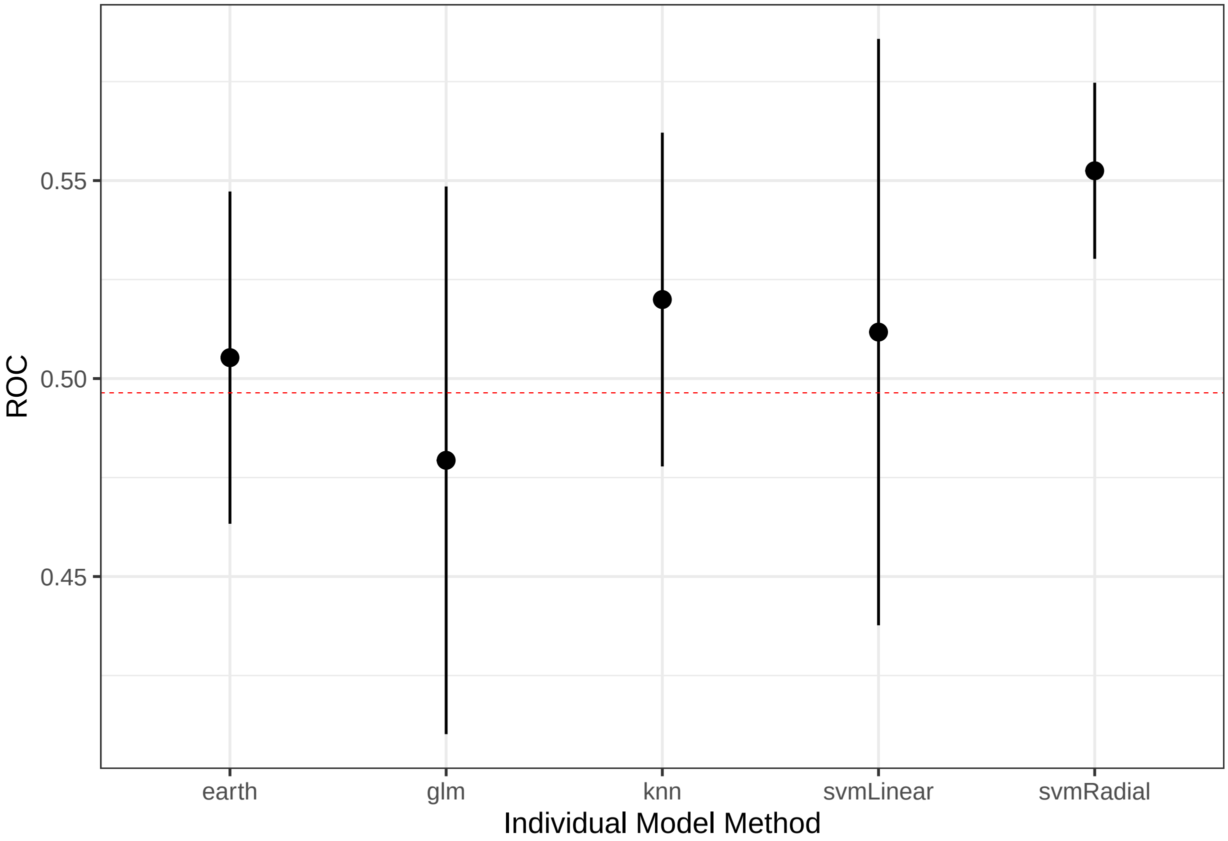


Our initial preferred model based on the evaluation metrics been SVM Radial would have total income, residence, transport reliability, household head, and living alone are the top five variables that accurately predicts the housing insecurity, with employment, federal aid, gender, and dependents been 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.

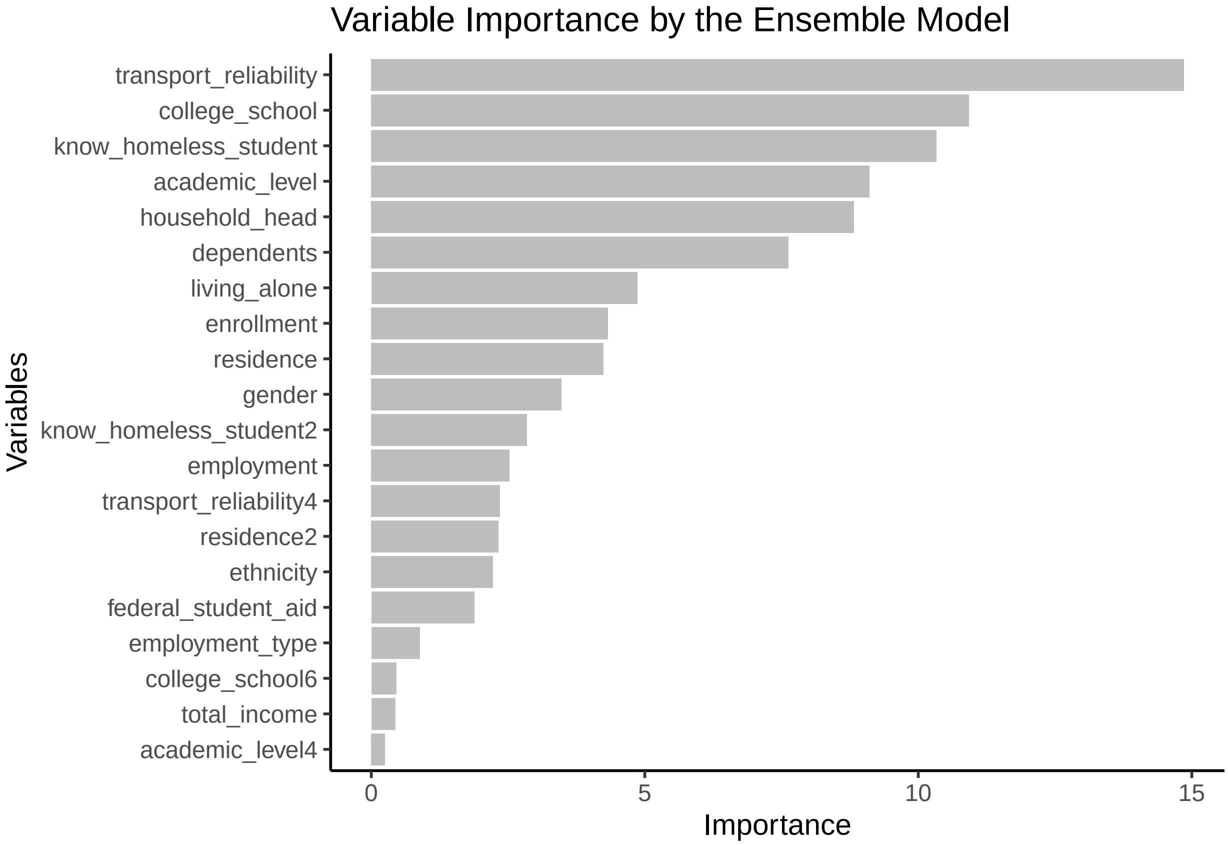
Table 4: Evaluation metrics for Spent Night Elsewhere as a response variable

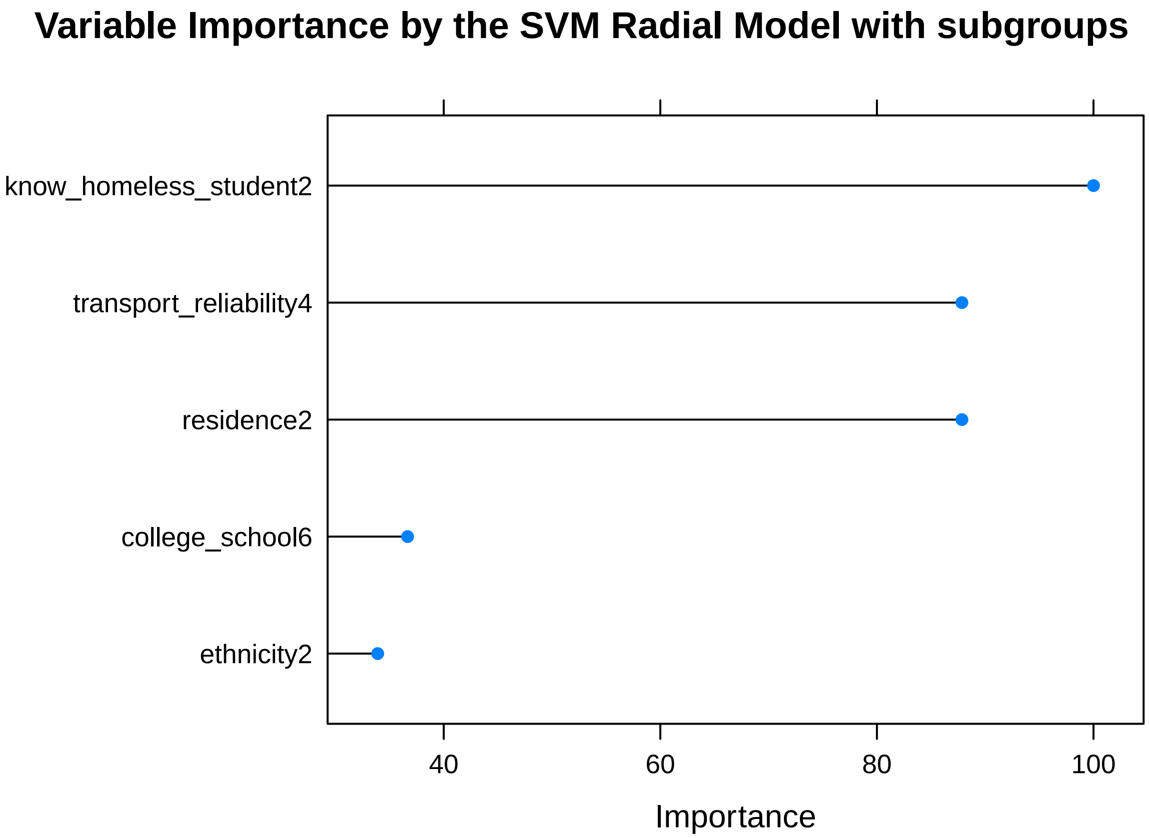
Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.54	0.46	0.35	0.56	0.37	0.5216
KNN	0.48	0.52	0.49	0.56	0.48	0.5294
MARS	0.45	0.55	0.65	0.47	0.56	0.5952
SVM Linear	0.54	0.46	0.51	0.42	0.46	0.5012
SVM Radial	0.51	0.49	0.54	0.44	0.49	0.5498

4.2.2.2 Results for Spending the night elsewhere 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.





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.

4.2.3 Modeling Food Insecurity

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

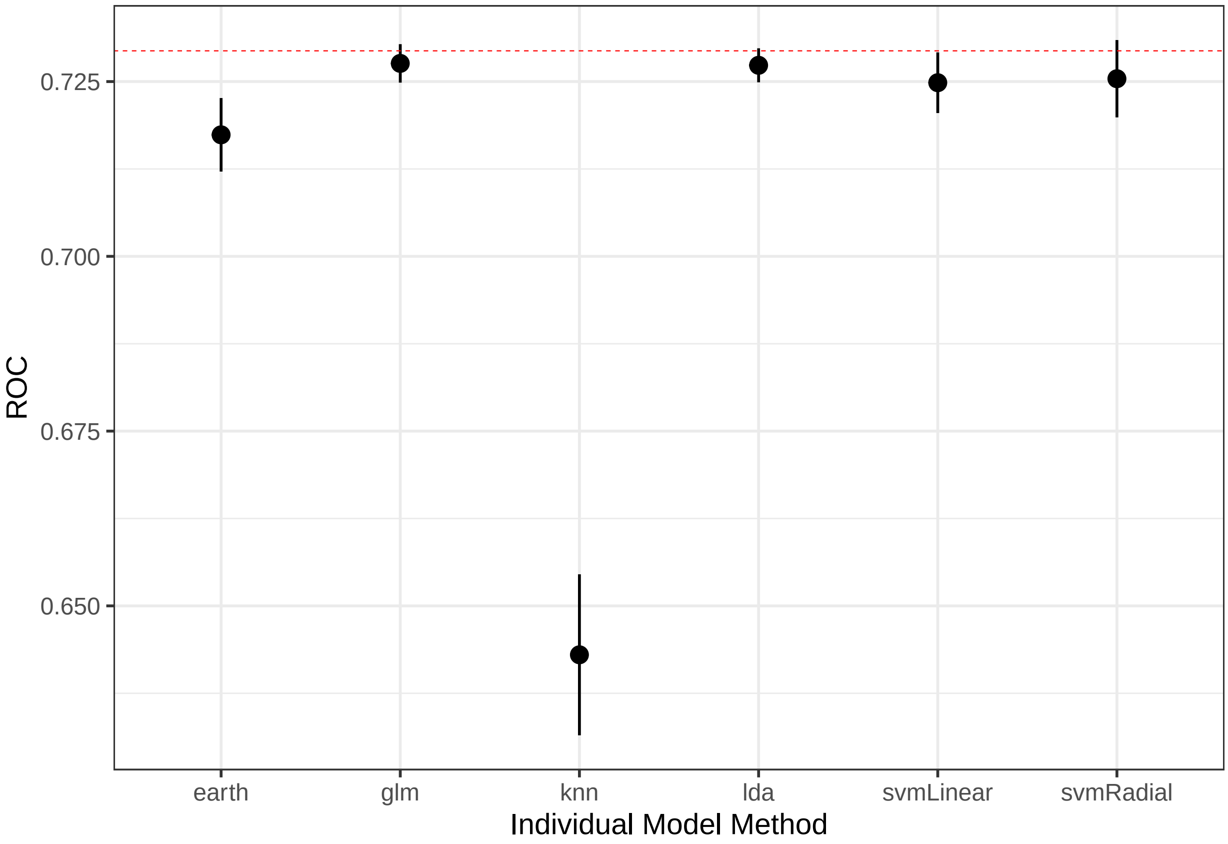
enough money for food? (FI_q31).

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.

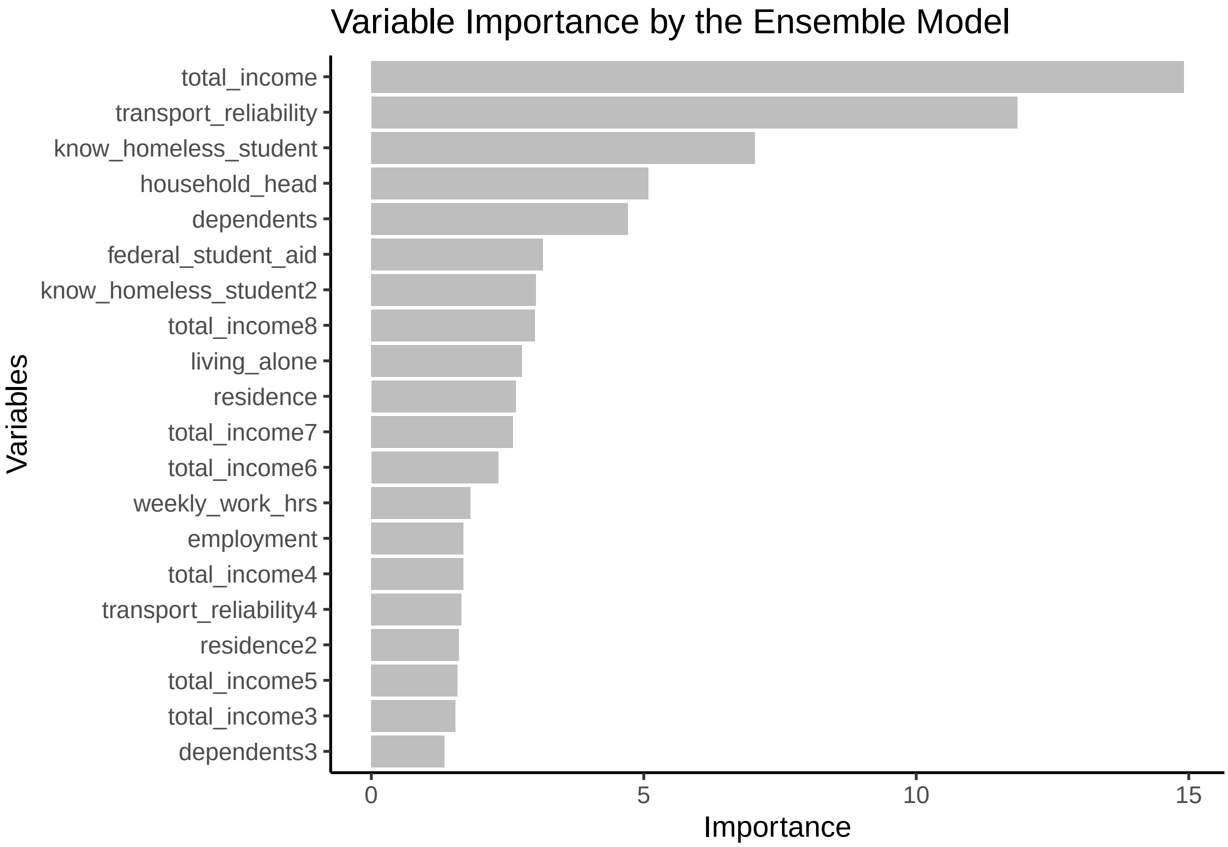
Table 5: Evaluation metrics for Food Insecurity I as a response variable

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	0.58	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

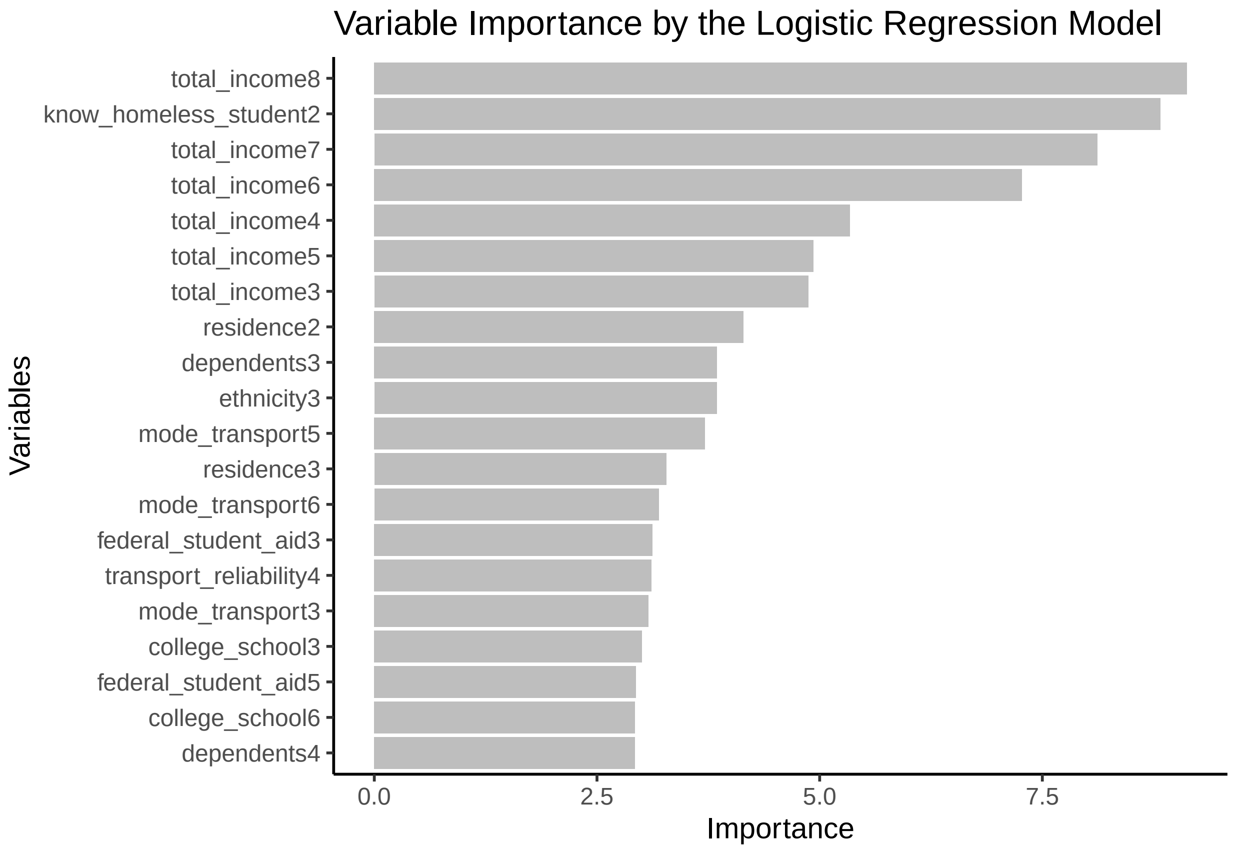
4.2.3.1 Results for Food Insecurity Response I Based on the table values and the evaluation metric criteria, with the exception of knn model , all the others are preferred as our initial best models.



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.



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.

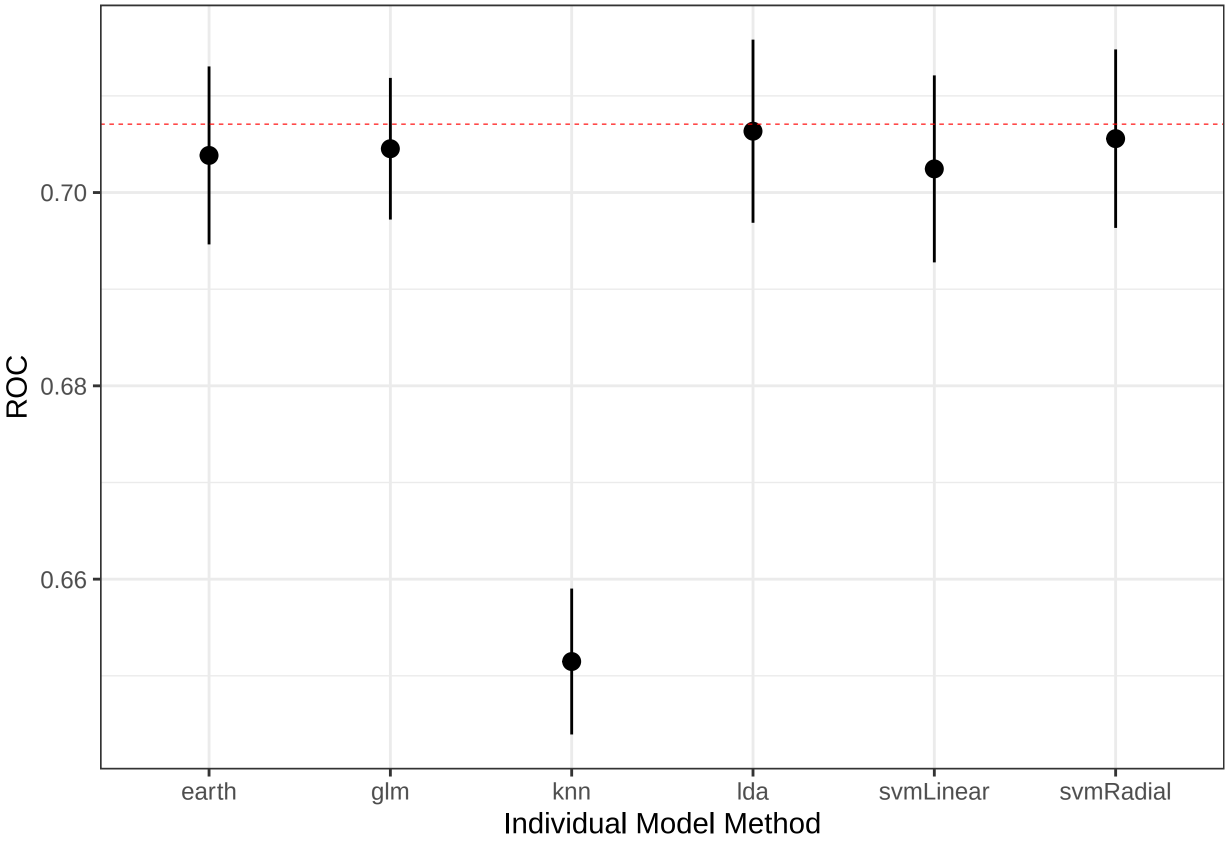


Given the plot above, the top 5 subgroups are total income8, know homeless student 2, total income 7, residence2, and ethnicity3.

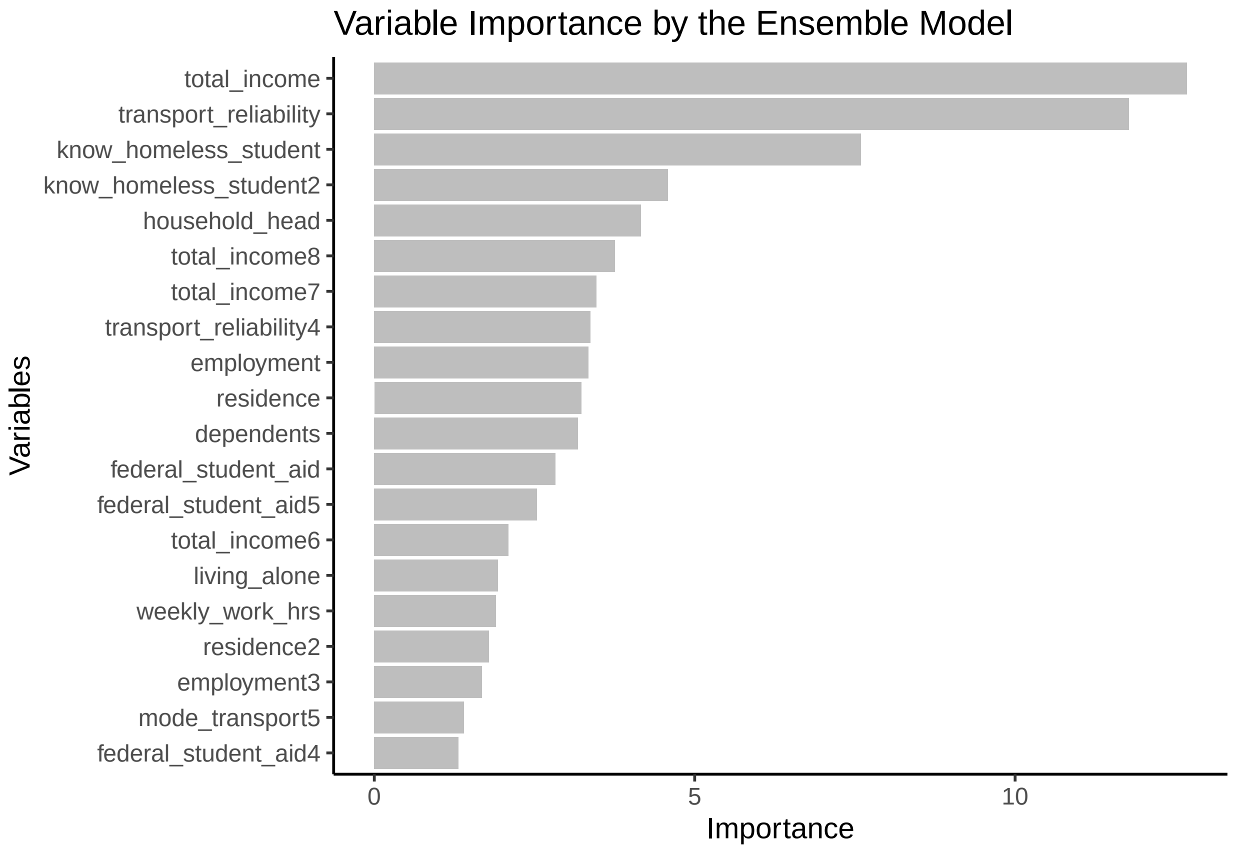
Table 6: Evaluation metrics for Food Insecurity II as a response variable

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	0.72	0.61	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

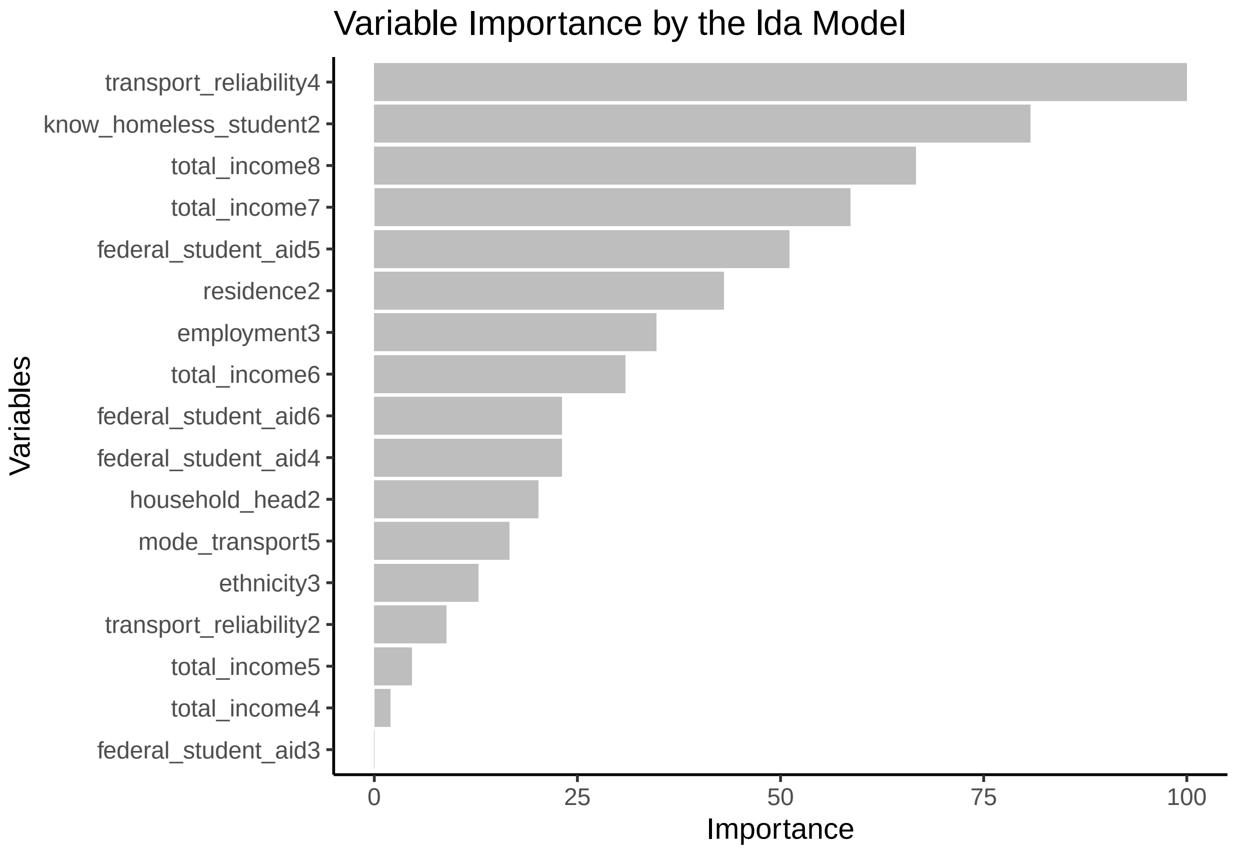
4.2.3.2 Results for Food Insecurity Response II Modeling food insecurity 2 as response, the best initial models are all models apart from the knn model.



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.



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.

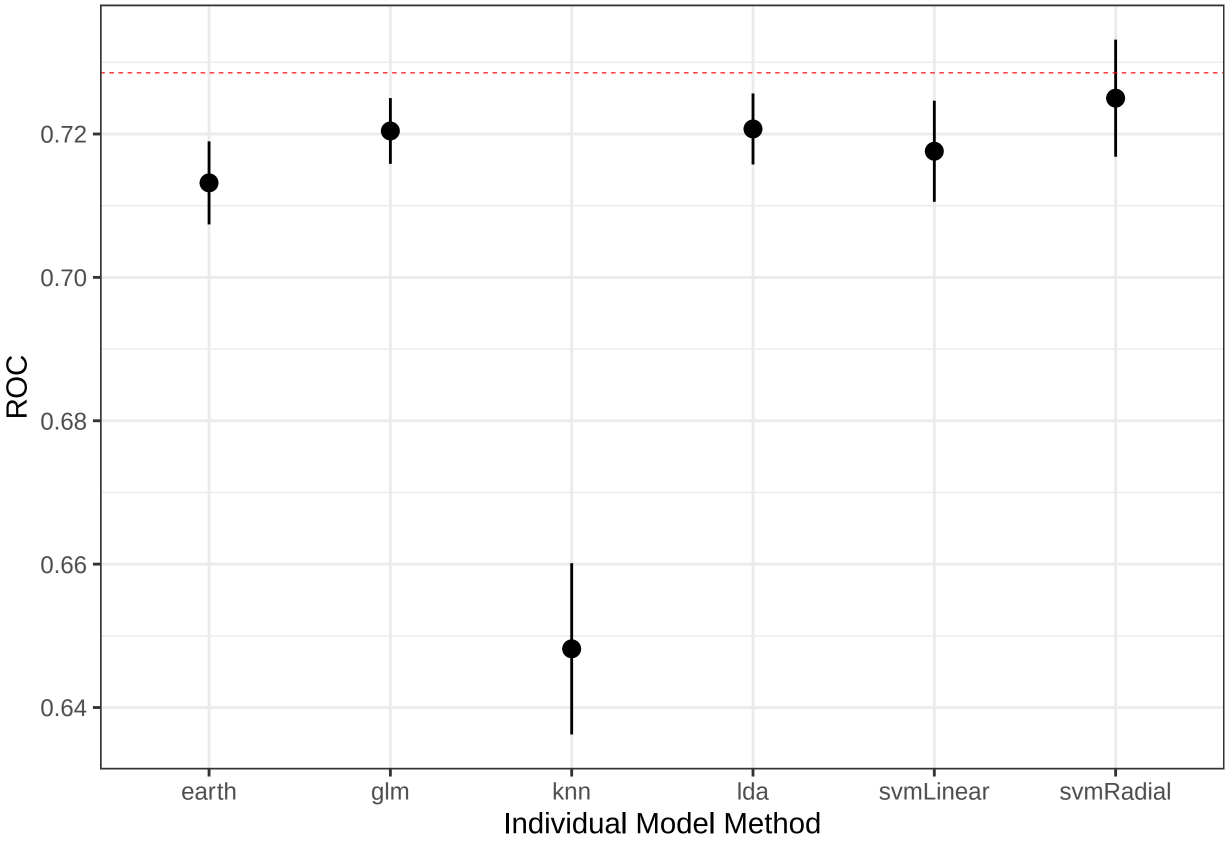


Given the plot above, the top 5 subgroups are transport reliability4, know homeless student2, total income8, federal student aid5, and residence2.

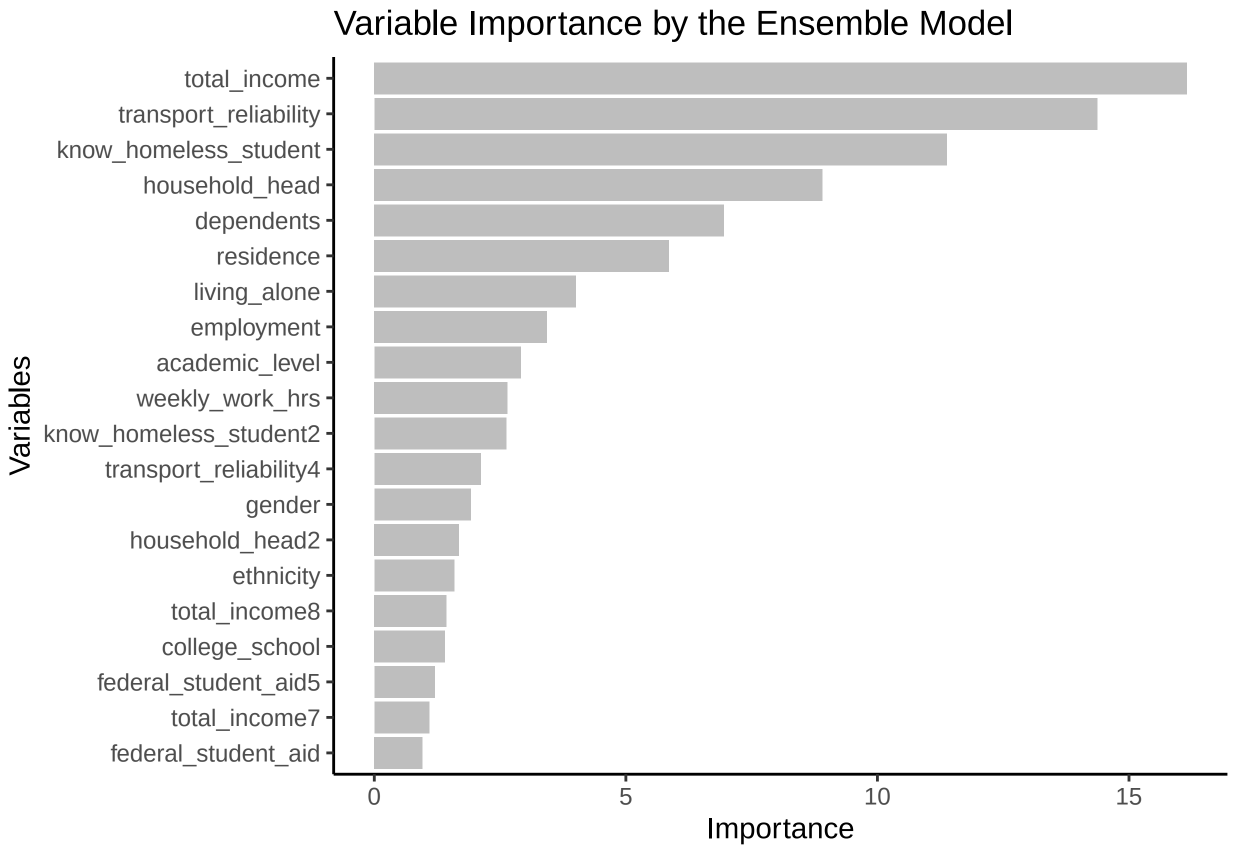
Table 7: Evaluation metrics for Food Insecurity III as a response variable

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	0.67	0.5	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

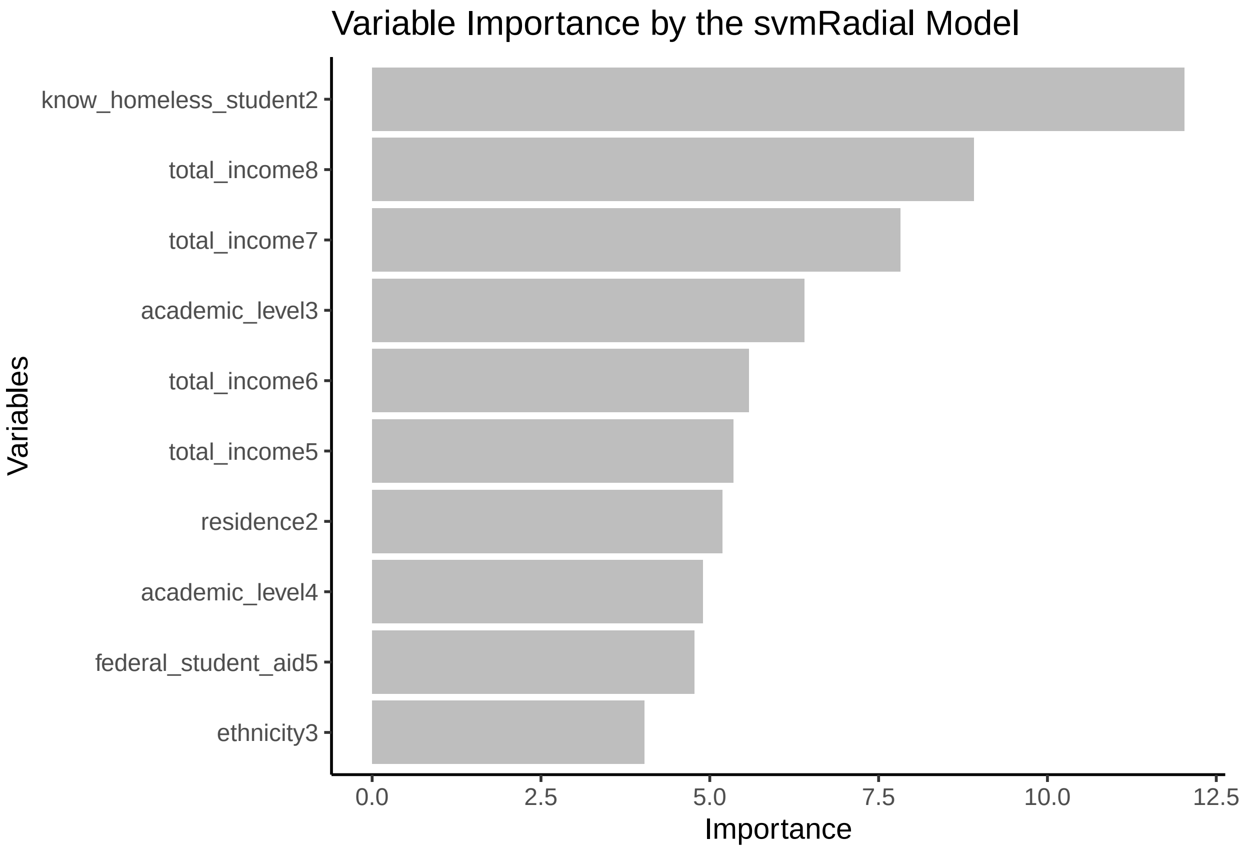
4.2.3.3 Results for Food Insecurity Response III Given the table values of evaluation metric criteria, our initial best models are all but knn and MARS.



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



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.

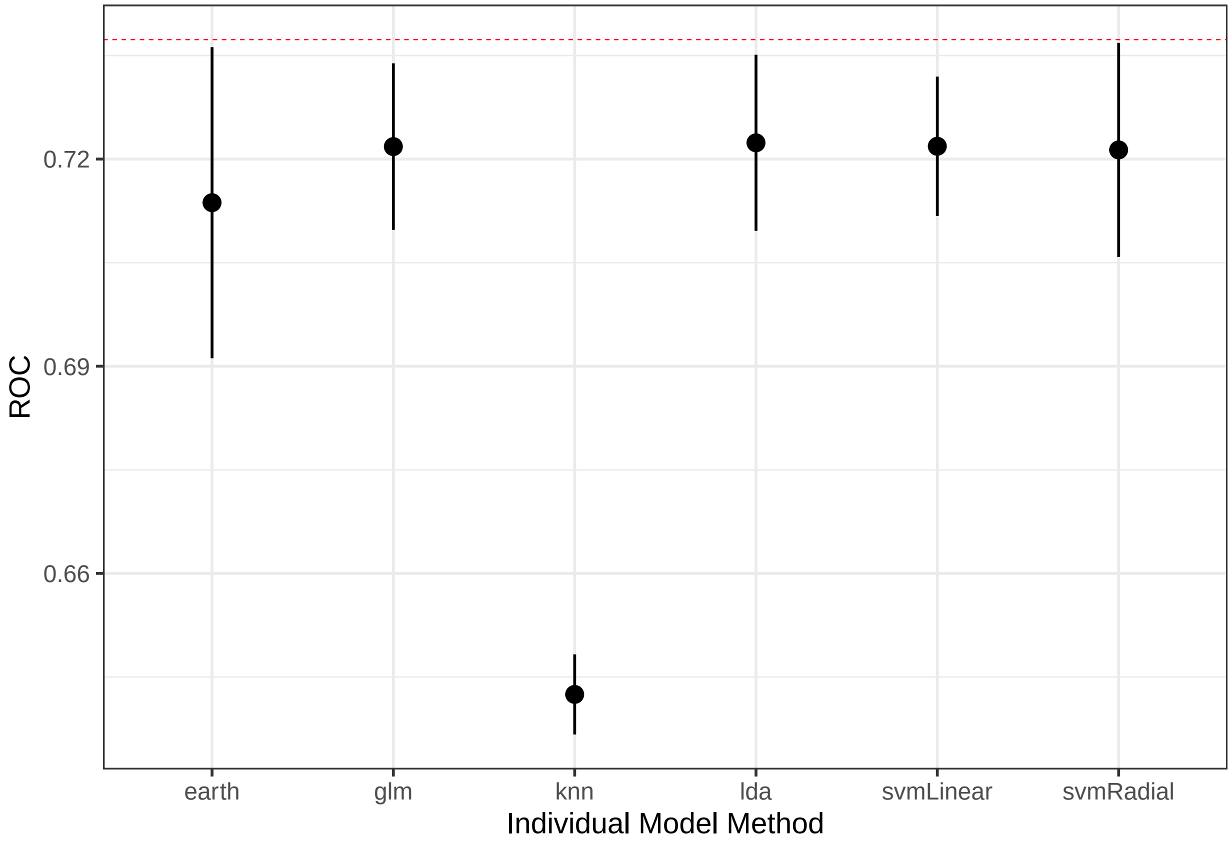


Given the plot above for food insecurity 3, the top 5 subgroups are know homeless student2, total income8, total income7, academic level3, and residence2.

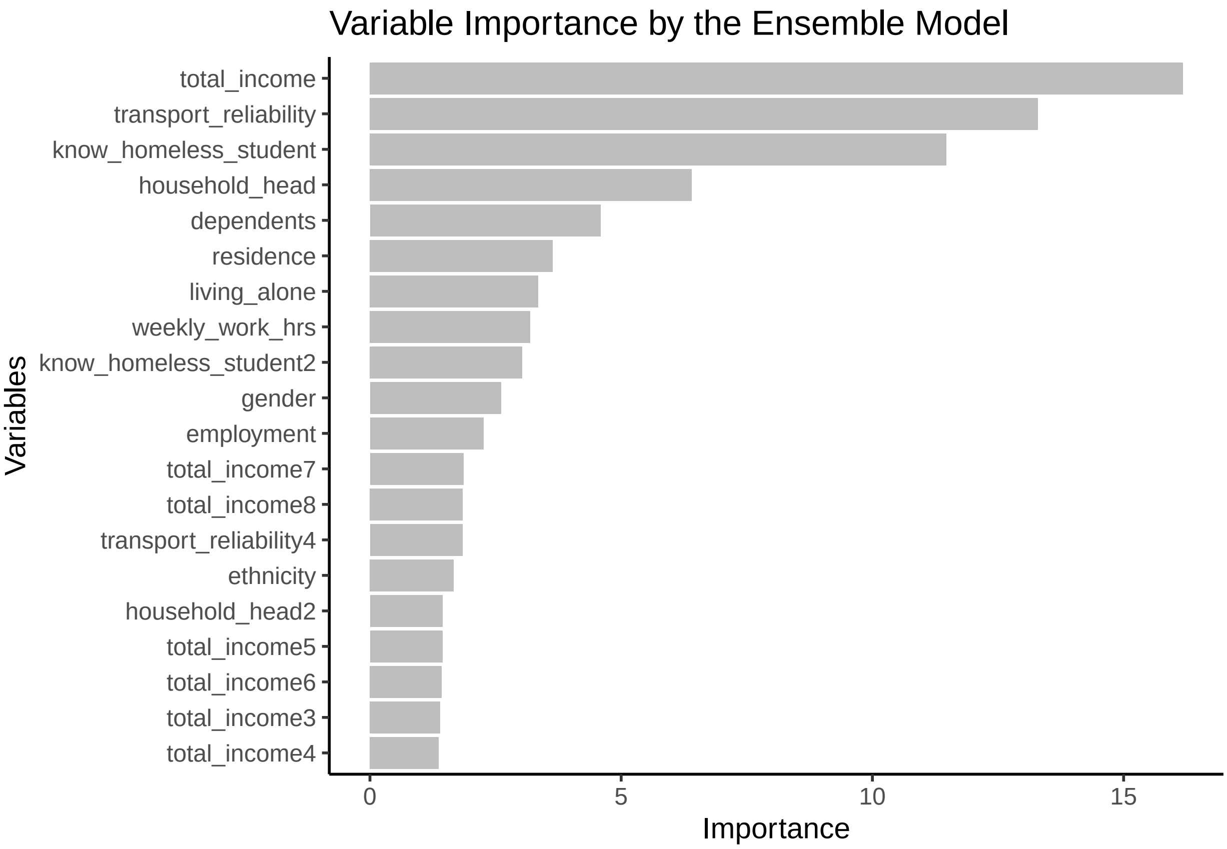
Table 8: Evaluation metrics for Food Insecurity V as a response variable

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	0.37	0.63	0.56	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

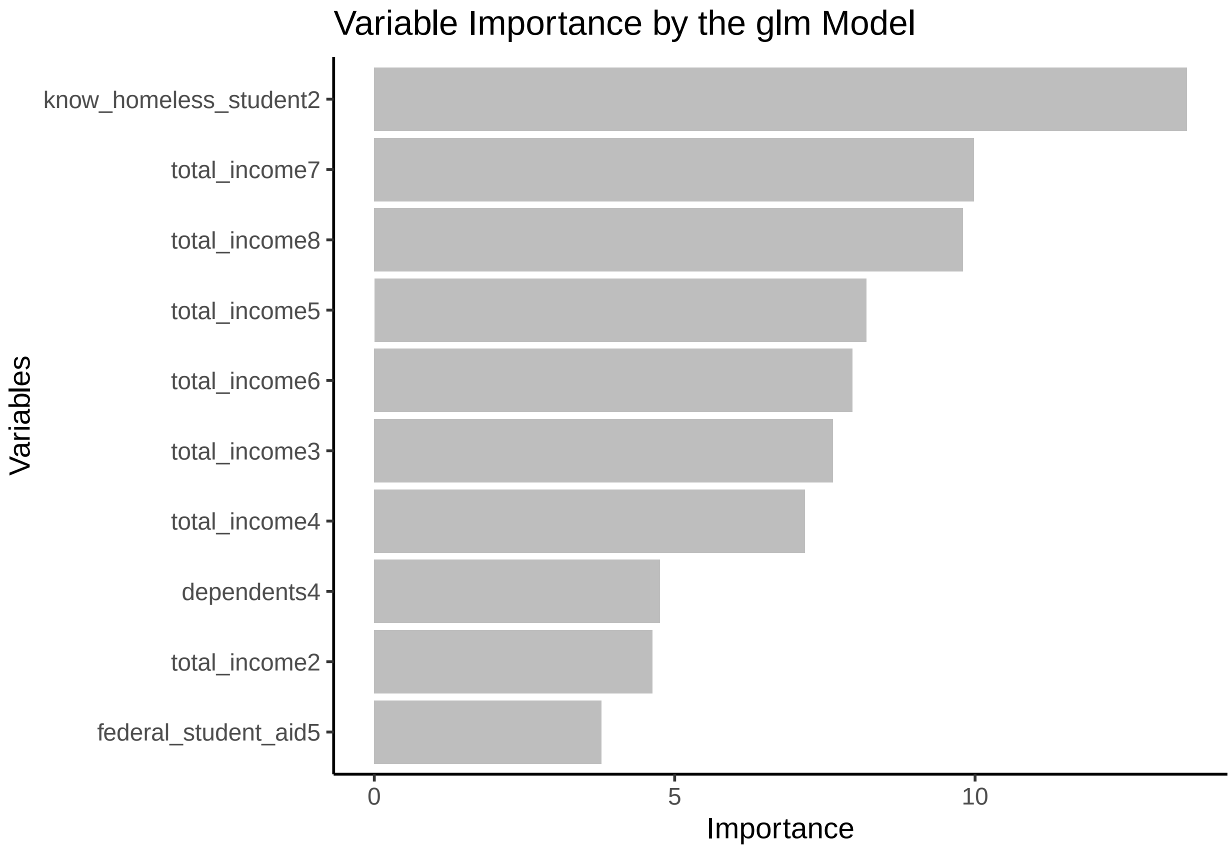
4.2.3.4 Results for Food Insecurity Response IV Our best models according the evaluation metrics criteria are glm and SVM Radial for food insecurity response 4.



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.



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.



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.

5 Discussion

Our initial preferred model or models are selected based on the evaluation metrics accuracy, misclassification, and 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 are 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 modelings are chosen a plot of variables of importance are obtained to determine which variables are of the essence 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.

In a 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. Also, we observed issues of multiple responses with some of the variables, especially, for the gender variable. We did not find it reasonable for respondents to be given the chance to select more than one options for their gender.

6 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.

7 Future Work (Phase II)

7.1 Rule for defining response variable for Food and Housing Insecurities

In the phase of the analysis, we will provide a rule for classifying a student as food insecure or housing insecure. These rules were used in selecting the appropriate response variables for both food and housing insecurities.

We hope to construct 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”.

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/

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

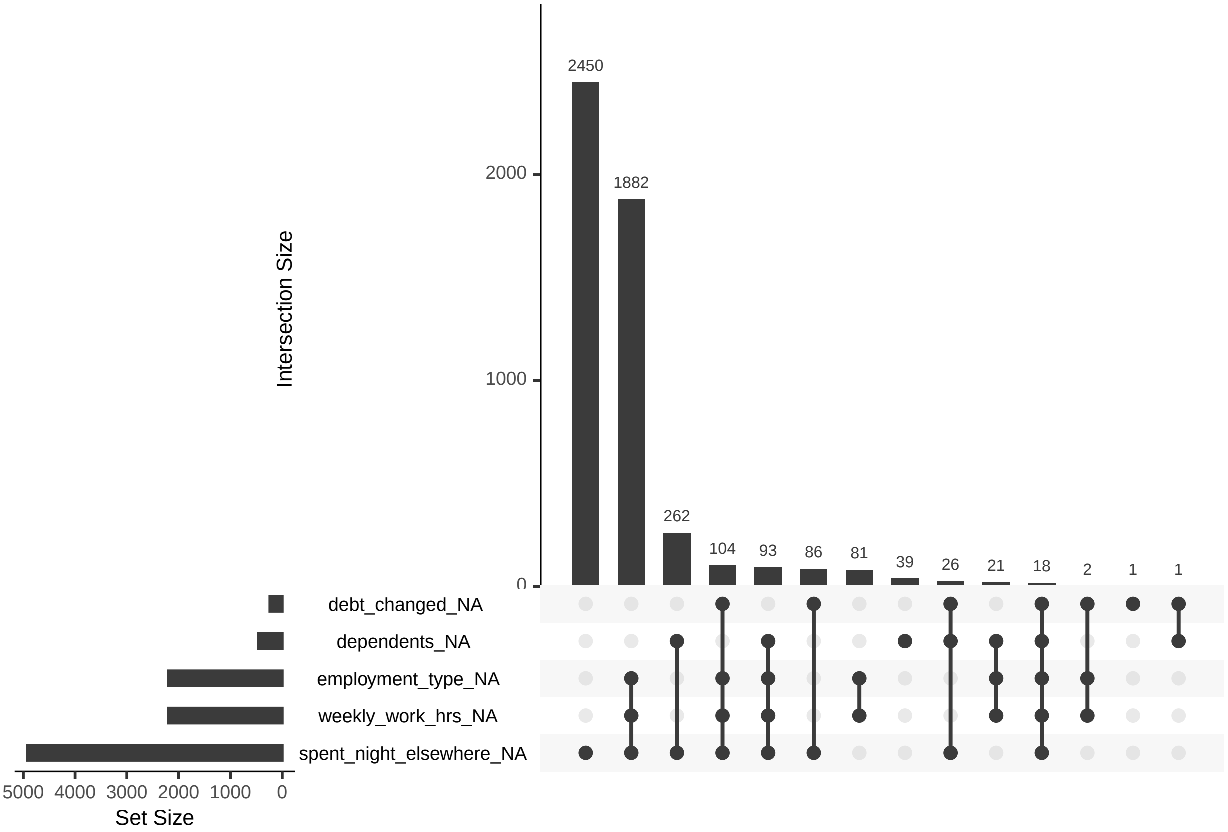
Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577.

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.

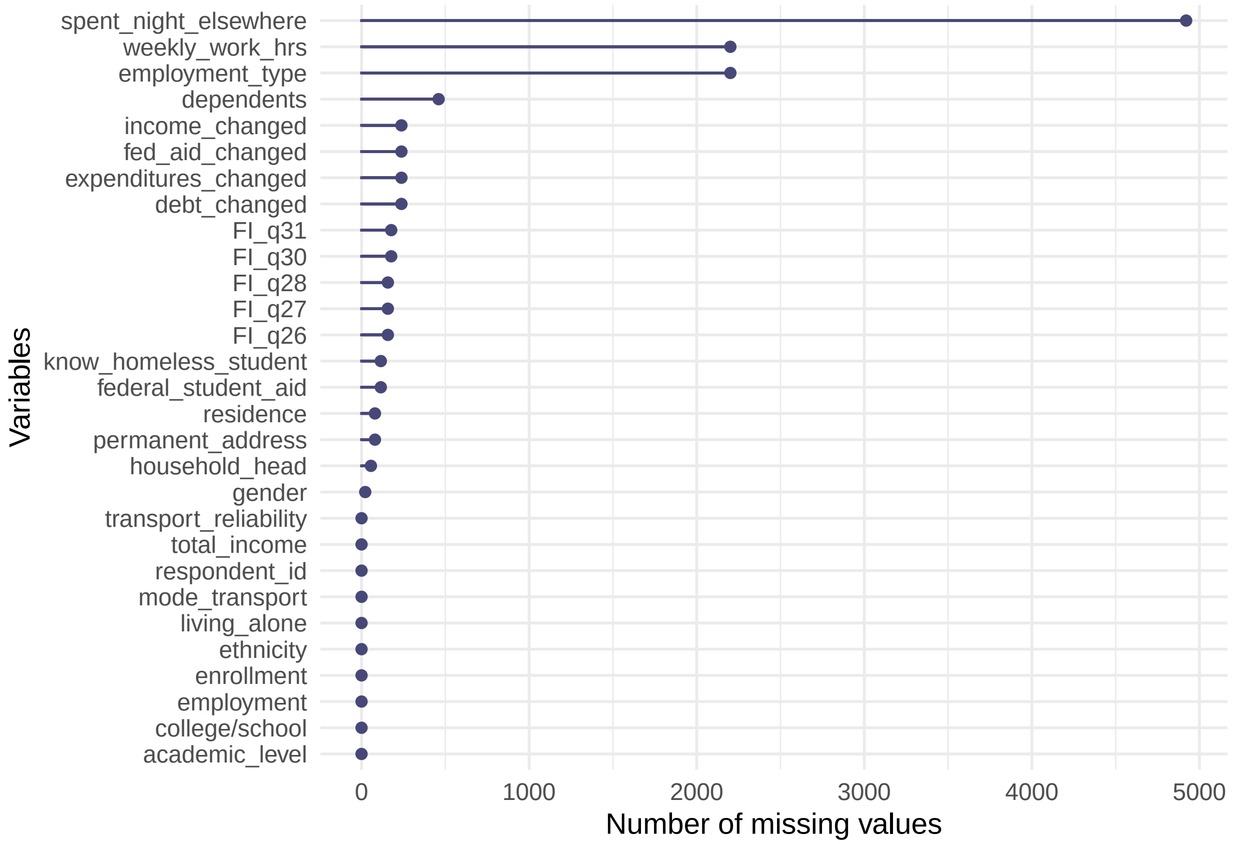
Wikipedia. (2021). Support-vector machine. Retrieved October 15, 2021, from https://en.wikipedia.org/wiki/Support-vector_machine

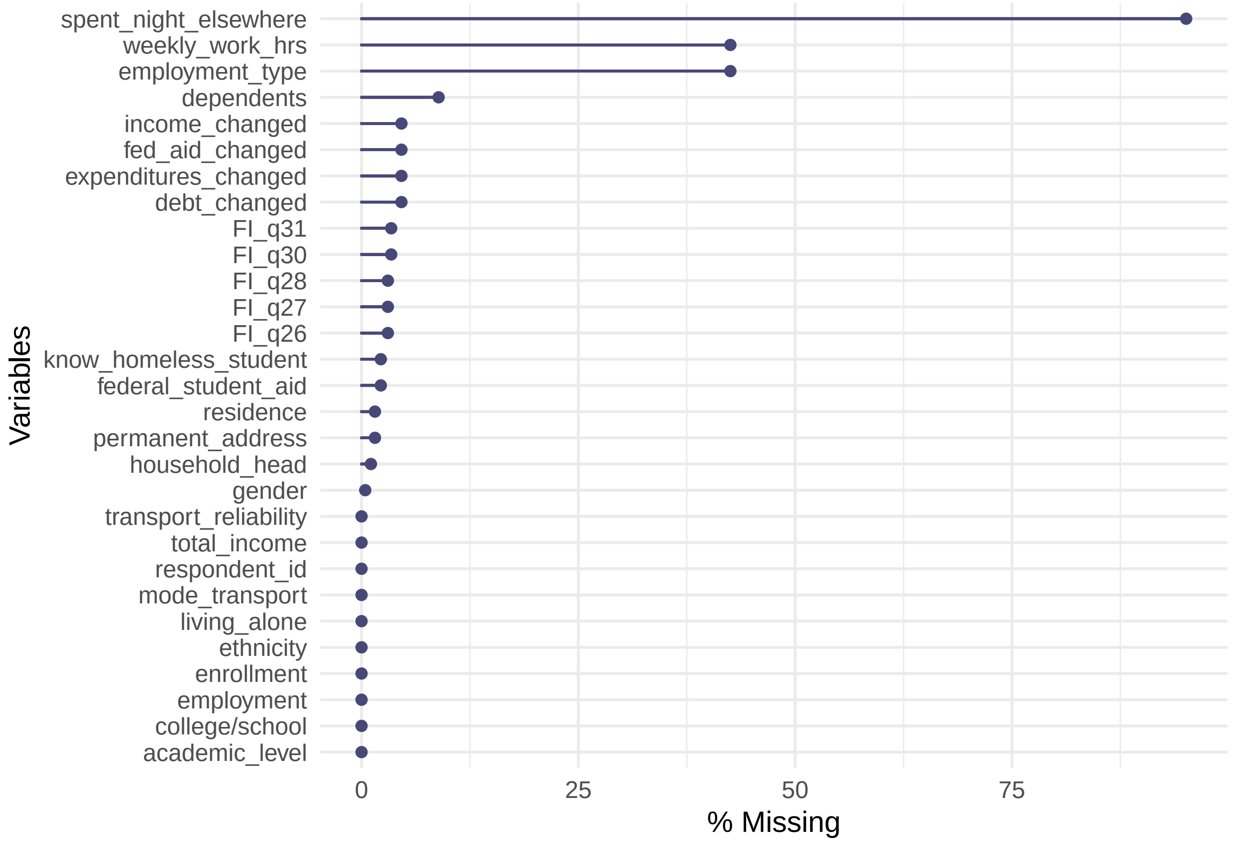
Apendix

7.2 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


spent_night_elsewhere
weekly_work_hrs
employment_type
dependents
income_changed
fed_aid_changed
expenditures_changed
debt_changed
FI_q31
FI_q30
FI_q28
FI_q27
FI_q26
know_homeless_student
federal_student_aid
residence
permanent_address
household_head
gender
transport_reliability
total_income
respondent_id
mode_transport
living_alone
ethnicity
enrollment
employment
college/school
academic_level
Variables
0
1000
2000
3000
4000
5000
Number of missing values


spent_night_elsewhere
weekly_work_hrs
employment_type
dependents
income_changed
fed_aid_changed
expenditures_changed
debt_changed
FI_q31
FI_q30
FI_q28
FI_q27
FI_q26
know_homeless_student
federal_student_aid
residence
permanent_address
household_head
gender
transport_reliability
total_income
respondent_id
mode_transport
living_alone
ethnicity
enrollment
employment
college/school
academic_level
Variables
0
25
50
75
% Missing

7.3 Codes

Codes for the analysis is available upon request.