NLSY 2019 Depression & Income Analysis

Group Members: Owen Patrick, Atticus Patrick, Tucker Paron

Introduction

Our dataset, The National Longitudinal Survey of Youth (NLSY97), is a large dataset of 8,984 participants born between 1980 and 1984. The survey first took place in 1997 where all participants were living in the United States, then interviews were held annually from 1997 until 2020. Survey participants were allowed to "skip" certain questions, and this is denoted for each question. There is data on mental health, employment, substance use, familial status, and education history. These answers are both numeric and categorical.

Our two primary response variables that we are trying to predict with our model are depression and income. There are quite a few questions in the survey relating to depression but we decided to focus on one in particular. It asks the respondent to respond to the statement "I felt depressed" with four response options. We decided to engineer this variable into a binary response variable. We encoded no as a 0 and yes as a 1 for depression. Also, we decided to predict the incomes of respondents for our continuous variable. The question in the survey asks "TOTAL INCOME FROM WAGES AND SALARY IN PAST CALENDAR YEAR."

Due to the fact that respondents were able to skip certain questions we had to impute or remove rows with these skips. We decided to set most skips to NA. For the "Valid Skips" we decided to either impute these to a zero or NA depending on the variable where we thought it made sense. After doing these imputations we then kept only complete cases of data and it reduced the number of observations from 8,984 to 4,280. Additionally, because we had so many categorical variables we decided to one hot encode them. In order to avoid the dummy variable trap due to the multicollinearity of the categories within each categorical variable, we included m-1 categories for our models (given m categories.) The left out category can then be considered the reference value. Predictor Variables are listed in the code appendix.

Research Questions

In this project we decided to address two main questions. First we examined the relationship between a number of predictors, such as employment, education, household & demographics and symptoms of depression (with classification with levels 0 and 1). Second, we addressed the relationship between respondent related predictors such as employment, education, household & demographics related factors and the income (continuous) of the respondents.

Methods

Classification

1. Logistic Regression (lasso/ridge)

Initially a model predicting depression using all 49 predictor variables (once one hot encoded) was trained and tested but this model only returned one true positive prediction correctly. This is due to there being only 71 people labeled as depressed, which is what we want to predict, versus 4209 people labeled as not depressed. Two problems needed to be addressed: the extreme minority class of respondents labeled as depressed (5-7 days in the past week) and the abundance of unnecessary predictor variables. To address the minority class, SMOTE (Systematic Minority Oversampling Technique) was used. This function created synthetic observations based on k nearest neighbors within the minority class. Based on current literature, the optimal choice is to balance the minority and majority classes evenly so this was the goal. To deal with the abundance of predictors, LASSO regression was implemented. This returned coefficients for the most important variables and set the other predictor variable's coefficients to zero (Figure 1). Combining these two techniques, a final logistic regression model was built with the subset of important predictors and a balanced sampling of depressed and non-depressed individuals (1846 and 1863, respectively.) This model performed quite well as will be discussed later on.

2. Naive Bayes Classifier

Naive Bayes is a popular method involving the use of Bayes' rule and Bayesian inference to determine the conditional probability of an event. In this case, Naive Bayes was used to determine: P(Exhibiting symptoms of Depression |

Predictors B0…Bn) based on P(Predictors B0…Bn | Exhibiting symptoms of Depression). In Naive Bayes, it is assumed all predictors are independent of each other, and while this was not true in our study and is not generally the case, Naive Bayes is known to yield effective prediction power. For the Naive Bayes model, the subset of important features determined by Lasso regression was used for both training and testing. The resulting accuracy was assessed using a confusion matrix.

3. Random Forest Classifier

A random forest classifier was trained and tested using the selected important variables and the upsampled data set from SMOTE. This tree was built with a mtry of 5 and number of trees = 100. This model was very accurate but there are major concerns of overfitting.

Regression

1. Linear regression (lasso/ridge)

Linear Regression was the first method utilized when addressing respondent income. We used 67 predictors for this (with multiple predictors for each original variable created by one-hot encoding as mentioned in the introduction). Four approaches were used to predict income here – linear regression, linear regression with k-fold cross validation, lasso regression, and ridge regression. K-fold cross validation was performed with 2 through 10 folds to determine the ideal number of folds. Similarly, lasso and ridge regression were iteratively run to determine the ideal penalty term (lambda) for the given training data. Lasso regression was also used to perform feature selection, as the penalty term in lasso regression generally forces predictor variables with little predictive power to zero.

2. Boosting

A boosting model was used to predict TOTAL_INCOME. There were 22 predictor variables used after removing all related income variables (ie. variables that were highly collinear). The boosting model used 100 trees, an interaction depth of 3, a shrinkage parameter of .01 and a bagging fraction of .5. These parameters were close to the default values and were moderate selections given the small sample size (relative to the number of variables). A secondary boosting model was used with only 10 of the most highly influential variables; however, as is noted in the results section, this model performed worse.

3. Random Forest

A random forest was also used to attempt to predict TOTAL_INCOME. Again, 22 predictors were used after removing the collinear variables. A forest size of 100 (ntree = 100) was used due to the aforementioned small relative data size. Like the boosting model, a secondary RF model was established also with only the 10 most highly influential variables. This too performed worse than the original.

Results

Depression

1. Logistic Regression

The logistic regression model using SMOTE had a final accuracy of 78.77% and a recall of 72.73% which were computed from the decision matrix in Figure 2. We care about recall because this is the proportion of respondents truly labeled as depressed that we successfully predicted to be depressed. Further evaluating this model, McFadden's Pseudo R-Squared was calculated. This value is 0.37 which shows that the model is an excellent fit (lower values are expected for McFadden's R-squared.) The most important features in this model were not being employed in 2019, being injured 4 or more times in the past year, and household size were the most important variables in this model. Despite this performance, it is vital to realize that overfitting could be present since there were so many synthetic minority observations generated (about 1800.)

2. Naive Bayes Classifier

Naive Bayes classifier was relatively accurate with an accuracy score of 72.31%. However, as shown by the confusion matrix in Figure 4, there was a high number of false negatives. This means that out of all actual people with symptoms of depression, we are not predicting enough people to have symptoms of depression.

3. Random Forest Classifier

The final random forest classification model had an extremely high accuracy of 96.34% and a recall of 93.33%. This was trained and tested on the SMOTE data and was our best performing model, however with such a high accuracy this model may be very prone to overfitting. Household size, being injured 4 times in the past year, and weight were the most important variables in this model for predicting depression (figure 3.)

Income

1. Linear Regression (lasso/ridge)

The four models used were linear regression, linear regression with cross validation, lasso regression, and ridge regression. All models were trained on full sets of predictor variables (this full set can be found in the code appendix). K-fold cross validation was carried out with k = 2, 3, …, 10, and k = 10, shown in Figure 5, was determined to be the best performing version with a test root mean squared error (RMSE) of $45,009.73. Linear regression yielded a test RMSE of $44,881.39. The most important variables (Figure 6) were a GED as a highest degree, a high school diploma as a highest degree, and no degree as highest degree. Lasso regression yielded a test RMSE of $44,815.46. The most important predictor variables (Figure 7) were a DDS/JD/MD as a highest degree, a Masters as a highest degree, and a PhD as a highest degree. Ridge regression resulted in a test RMSE of $44,955.3 with the most important variables (Figure 8) being a DDS/JD/MD as a highest degree, no degree, and a Masters as a highest degree.

2. Boosting

The boosting model yielded poor results with a RMSE of 47288.3. This means that on average our model's predictions were incorrect by over $47000. Given that the standard deviation of TOTAL_INCOME is 51065.24, our model is explaining little to no variance. When tuning the model to only include the ten most influential variables (see **Figure 9**), the result was an even worse performance with an RMSE of 47470.01. Some examples of the most important variables were highest degree received, marital status, and weeks worked in the past year.

3. Random Forest

The Random Forest implementation yielded similar results to that of the Boosting model. The initial implementation with all 22 (non-collinear) variables had a RMSE of 44782.52. This is slightly better than the Boosting model, but still explains little variation in regards to the single standard deviation of 51065.24. The updated model using only the ten most influential variables (see **Figure 10**) possessed an RMSE of 46661.93. This is significantly worse than the original. Some examples of the most important variables were highest degree received, weeks worked, weight and age.

Conclusions

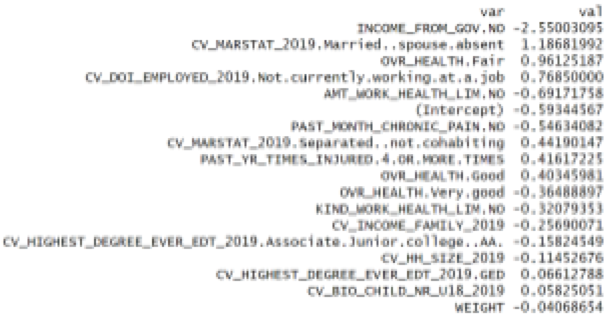
Depression

The naive bayes model performed fairly well in terms of accuracy, but did not perform well in terms of recall as it yielded a high number of false negatives. This suggests the Naive Bayes model is not ideal for interpretation of feature importances and predictions. The best performing model was the random forest model with above 90% accuracy. The baseline random forest with no smote did not correctly predict any people with depression correctly so this jump up in accuracy must be taken with a grain of salt (due to overfitting.) The logistic regression model also performed well but not to the same standard. Overall, it appears that household size, not being employed, being injured frequently, and weight seem to be the most significant variables for predicting depression.

Income

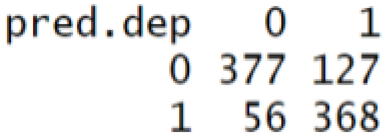
All models predicting income were assessed in terms of root mean squared error (RMSE). With regards to the four linear models performed, the best performing model was lasso regression with a RMSE of $44,815.46. The most important predictors in all four models were heavily related to highest degree received. With regards to the boosting and random forest models, there were unfortunately no well performing models. The Random Forest model was the best performing model overall with a RMSE of $44782.52. However, there were still valuable findings such as what variables are most relevant to the responses. Given our models' poor performance, one should be cautious gleaning any solid insights from variable importance plots; however it may be telling that both the boosting and random forest models selected marital status, highest degree received, and weeks worked as the most influential variables. These are not surprises, as one would expect income to rise with education level and typically the more one works the more they earn. Additionally, marriage status could potentially be explained by people having to earn more to have an established family and marriage.

Figures



```
                                                          var         val
                                           INCOME_FROM_GOV.NO -2.55003095
                          CV_MARSTAT_2019.Married..spouse.absent  1.18681992
                                              OVR_HEALTH.Fair  0.96125187
              CV_DOI_EMPLOYED_2019.Not.currently.working.at.a.job  0.76850000
                                       AMT_WORK_HEALTH_LIM.NO -0.69171758
                                                  (Intercept) -0.59344567
                                   PAST_MONTH_CHRONIC_PAIN.NO -0.54634082
                      CV_MARSTAT_2019.Separated..not.cohabiting  0.44190147
                          PAST_YR_TIMES_INJURED.4.OR.MORE.TIMES  0.41617225
                                              OVR_HEALTH.Good  0.40345981
                                         OVR_HEALTH.Very.good -0.36488897
                                      KIND_WORK_HEALTH_LIM.NO -0.32079353
                                      CV_INCOME_FAMILY_2019 -0.25690071
CV_HIGHEST_DEGREE_EVER_EDT_2019.Associate.Junior.college..AA. -0.15824549
                                              CV_HH_SIZE_2019 -0.11452676
                           CV_HIGHEST_DEGREE_EVER_EDT_2019.GED  0.06612788
                                    CV_BIO_CHILD_NR_U18_2019  0.05825051
                                                       WEIGHT -0.04068654
```

Figure 1: Coefficients of LASSO Regression Model



```
pred.dep   0   1
       0 377 127
       1  56 368
```

Figure 2. Confusion Matrix of Final Logistic Regression Model

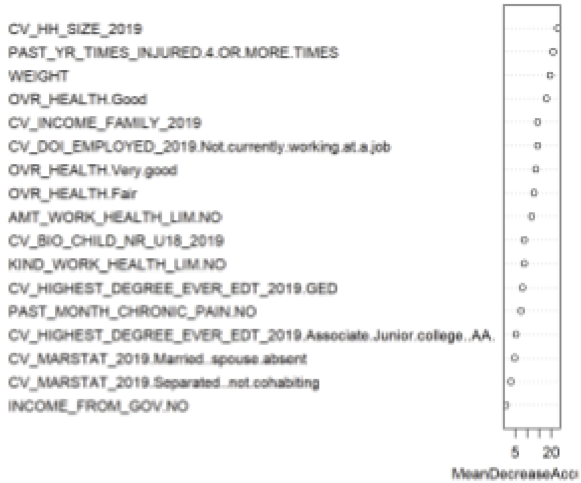
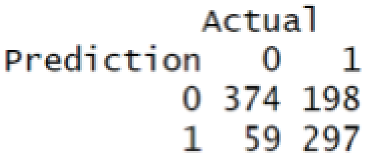


Figure 3. Variable Importance for Random Forest (Depression classification)



```
          Actual
Prediction   0   1
         0 374 198
         1  59 297
```

Figure 4. Naive Bayes Confusion Matrix

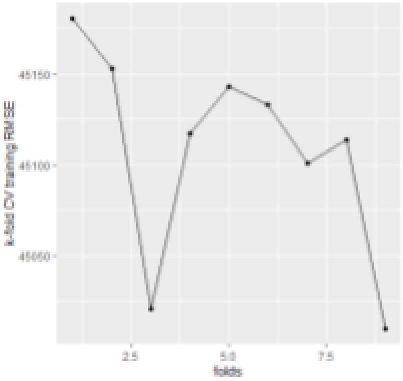


Figure 5. K-fold cross validation for linear regression

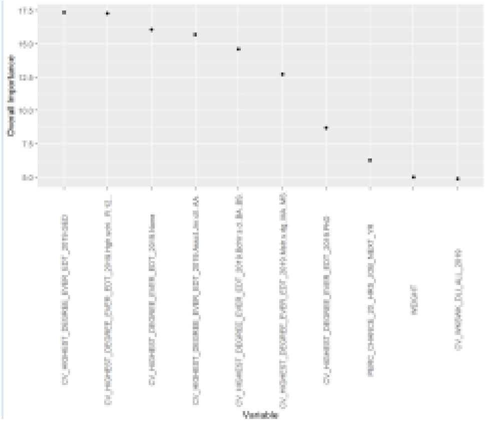


Figure 6. Linear Regression Variable Importances for Income

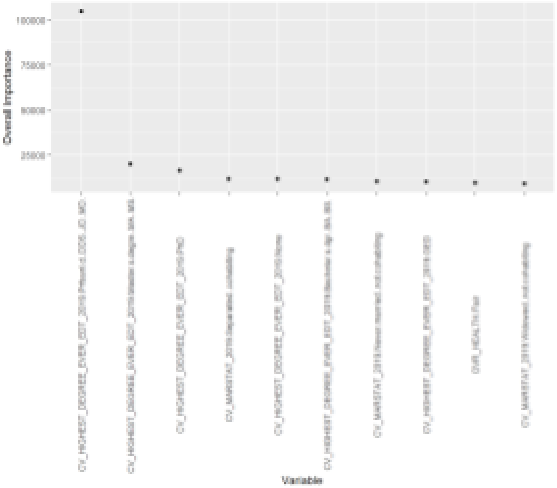
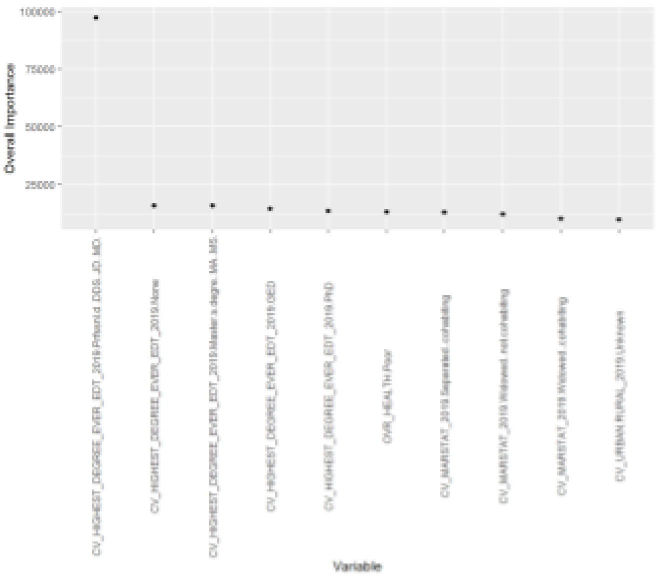


Figure 7. Variable Importance for Lasso Regression



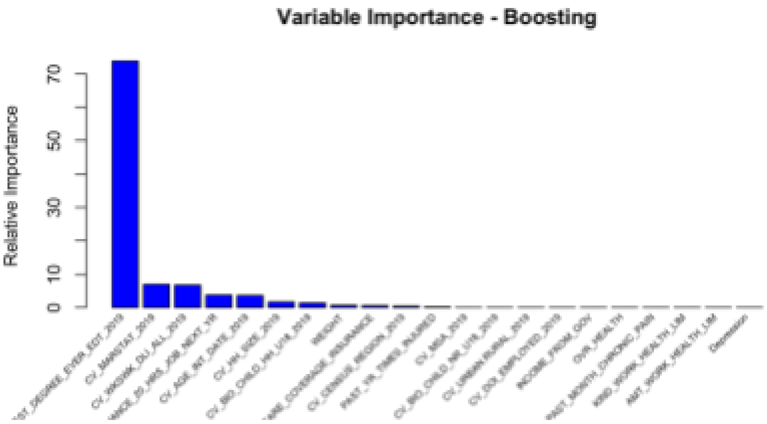


Figure 9. Variable importance plot for Boosting model predicting TOTAL_INCOME.

Figure 8. Variable Importance for Ridge Regression

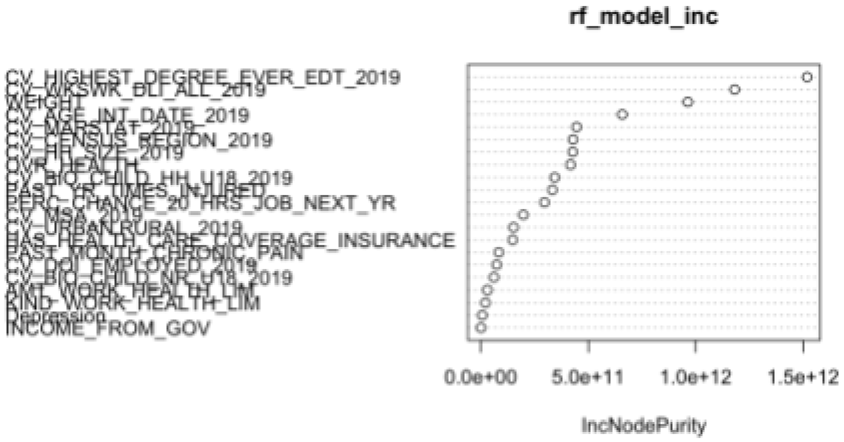


Figure 10. Variable importance plot for Random Forest predicting TOTAL_INCOME.

Code Appendix:

Depression Predictor Variables

```
"INCOME_FROM_GOV.NO", "CV_MARSTAT_2019.Married..spouse.absent", "OVR_HEALTH.Fair",
"CV_DOI_EMPLOYED_2019.Not.currently.working.at.a.job", "AMT_WORK_HEALTH_LIM.NO",
"PAST_MONTH_CHRONIC_PAIN.NO", "CV_MARSTAT_2019.Separated..not.cohabiting",
"PAST_YR_TIMES_INJURED.4.OR.MORE.TIMES", "OVR_HEALTH.Good", "OVR_HEALTH.Very.good",
"KIND_WORK_HEALTH_LIM.NO", "CV_INCOME_FAMILY_2019",
"CV_HIGHEST_DEGREE_EVER_EDT_2019.Associate.Junior.college..AA.", "CV_HH_SIZE_2019",
"CV_HIGHEST_DEGREE_EVER_EDT_2019.GED", "CV_BIO_CHILD_NR_U18_2019", "WEIGHT"
```

Income Predictor Variables

```
[1] "CV_AGE_INT_DATE_2019"
 [2] "CV_CENSUS_REGION_2019.Northeast..CT..ME..MA..NH..NJ..NY..PA..RI..VT."
 [3] "CV_CENSUS_REGION_2019.North.Central..IL..IN..IA..KS..MI..MN..MO..NE..OH..ND..SD..WI."
 [4]
"CV_CENSUS_REGION_2019.South..AL..AR..DE..DC..FL..GA..KY..LA..MD..MS..NC..OK..SC..TN...TX..VA..WV
."
 [5] "CV_CENSUS_REGION_2019.West..AK..AZ..CA..CO..HI..ID..MT..NV..NM..OR..UT..WA..WY."
 [6] "CV_HH_SIZE_2019"
 [7] "CV_HIGHEST_DEGREE_EVER_EDT_2019.None"
 [8] "CV_HIGHEST_DEGREE_EVER_EDT_2019.GED"
 [9] "CV_HIGHEST_DEGREE_EVER_EDT_2019.High.school.diploma..Regular.12.year.program."
[10] "CV_HIGHEST_DEGREE_EVER_EDT_2019.Associate.Junior.college..AA."
[11] "CV_HIGHEST_DEGREE_EVER_EDT_2019.Bachelor.s.degree..BA..BS."
[12] "CV_HIGHEST_DEGREE_EVER_EDT_2019.Master.s.degree..MA..MS."
[13] "CV_HIGHEST_DEGREE_EVER_EDT_2019.PhD"
[14] "CV_HIGHEST_DEGREE_EVER_EDT_2019.Professional.degree..DDS..JD..MD."
[15] "CV_MARSTAT_2019.Never.married..cohabiting"
[16] "CV_MARSTAT_2019.Never.married..not.cohabiting"
[17] "CV_MARSTAT_2019.Married..spouse.present"
[18] "CV_MARSTAT_2019.Married..spouse.absent"
[19] "CV_MARSTAT_2019.Separated..cohabiting"
[20] "CV_MARSTAT_2019.Separated..not.cohabiting"
[21] "CV_MARSTAT_2019.Divorced..cohabiting"
[22] "CV_MARSTAT_2019.Divorced..not.cohabiting"
[23] "CV_MARSTAT_2019.Widowed..cohabiting"
[24] "CV_MARSTAT_2019.Widowed..not.cohabiting"
[25] "CV_MSA_2019.Not.in.CBSA"
[26] "CV_MSA_2019.In.CBSA..not.in.central.city"
[27] "CV_MSA_2019.In.CBSA..in.central.city"
[28] "CV_MSA_2019.In.CBSA..not.known"
[29] "CV_MSA_2019.Not.in.country"
[30] "CV_BIO_CHILD_HH_U18_2019"
[31] "CV_BIO_CHILD_NR_U18_2019"
[32] "CV_URBAN.RURAL_2019.Rural"
[33] "CV_URBAN.RURAL_2019.Urban"
[34] "CV_URBAN.RURAL_2019.Unknown"
[35] "CV_WKSWK_DLI_ALL_2019"
[36] "CV_DOI_EMPLOYED_2019.Not.currently.working.at.a.job"
```

```
[37] "CV_DOI_EMPLOYED_2019.Current.working.at.a.job"
[38] "CV_DOI_EMPLOYED_2019.Military.service..but.no.job..reported"
[39] "PERC_CHANCE_20._HRS_JOB_NEXT_YR"
[40] "TOTAL_INCOME"
[41] "INCOME_FROM_GOV.NO"
[42] "INCOME_FROM_GOV.YES"
[43] "OVR_HEALTH.Excellent"
[44] "OVR_HEALTH.Very.good"
[45] "OVR_HEALTH.Good"
[46] "OVR_HEALTH.Fair"
[47] "OVR_HEALTH.Poor"
[48] "WEIGHT"
[49] "PAST_MONTH_CHRONIC_PAIN.NO"
[50] "PAST_MONTH_CHRONIC_PAIN.YES"
[51] "KIND_WORK_HEALTH_LIM.NO"
[52] "KIND_WORK_HEALTH_LIM.YES"
[53] "AMT_WORK_HEALTH_LIM.NO"
[54] "AMT_WORK_HEALTH_LIM.YES"
[55] "PAST_YR_TIMES_INJURED.NONE"
[56] "PAST_YR_TIMES_INJURED.1.TIME"
[57] "PAST_YR_TIMES_INJURED.2.TIMES"
[58] "PAST_YR_TIMES_INJURED.3.TIMES"
[59] "PAST_YR_TIMES_INJURED.4.OR.MORE.TIMES"
[60] "EMOTIONAL_HEALTH_ISSUE.NONE"
[61] "EMOTIONAL_HEALTH_ISSUE.1.TIME"
[62] "EMOTIONAL_HEALTH_ISSUE.2.TIMES"
[63] "EMOTIONAL_HEALTH_ISSUE.3.TIMES"
[64] "EMOTIONAL_HEALTH_ISSUE.4.OR.MORE.TIMES"
[65] "HAS_HEALTH_CARE_COVERAGE.INSURANCE.NO"
[66] "HAS_HEALTH_CARE_COVERAGE.INSURANCE.YES"
[67] "Depression"
```