National Longitudinal Survey of Youth | 1997

Predicting Income and Mental Health Status

Atticus, Owen, Tucker

Data Set and Predictors

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. We used Data from 2019 only.

General predictor variable categories:

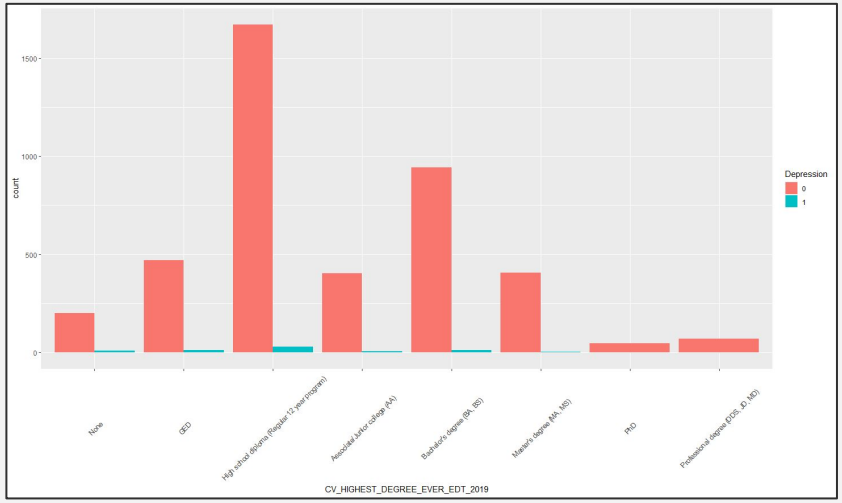
- Education level
- Demographic Information
- Health status
- Socioeconomic status

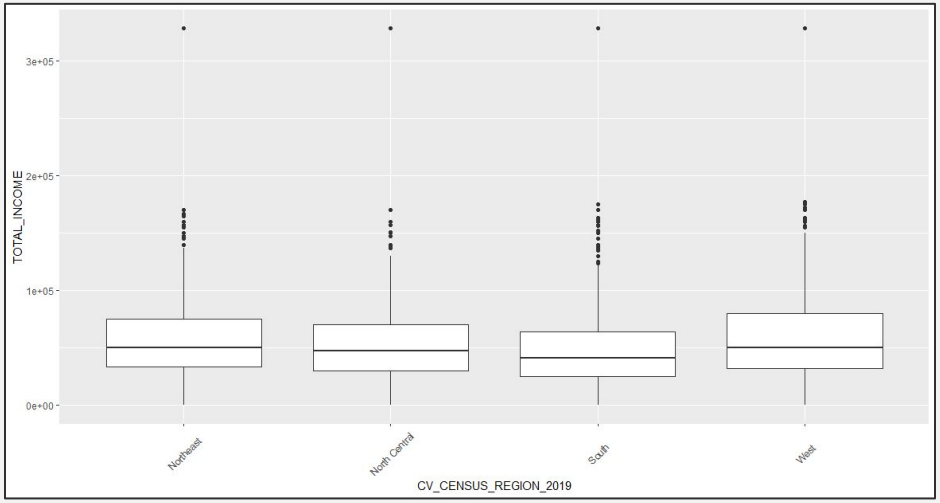
Research Questions

1.) What is the relationship between a number of predictors, such as education level, household status, income level, demographics, and symptoms of depression (categorical, two levels) and can symptoms of depression be predicted by these factors?

2.) What is the relationship between respondent related predictors such as education, household, and demographics related factors and the income (continuous) of the respondents and can income level be predicted by these factors?

Exploratory Data Analysis (EDA)





- Found major underrepresentation of respondents categorized as having symptoms of depression

Classification

Predicting Depression in Respondents

Response variable: Depression

RESPONSE CHOICE: "I felt depressed."

- 0 Rarely/None of the time/1 Day
- 1 Some/A little of the time/1-2 Days
- 2 Occasionally/Moderate amount of the time/3-4 Days
- 3 Most/All of the time/ 5-7 Days

- Engineered into 0 (not depressed) for response choices 0,1,2 and a 1 (depressed) for response choice 3

- Response counts:

0	1
4209	71

One Hot Encoding

CV_CENSUS_REGION_2019	CV_INCOME_FAMILY_2019	CV_HH_POV_RATIO_2019	CV_HH_SIZE_2019
Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)	188857	1123	2
South (AL, AR, DE, DC, FL, GA, KY, LA, MD, MS, NC, OK, SC, T...	67000	263	4
Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)	150000	892	2
Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)	263000	1301	3

CV_CENSUS_REGION_2019.South..AL..AR..DE..DC..FL..GA..KY..LA..MD..MS..NC..OK..SC..TN...TX..VA..WV.	CV_CENSUS_REGION_2019.West..AK..AZ..CA..CO..HI..ID..MT..NV..NM..OR..UT..WA..WY.
0	0
1	0
0	0
0	0

- With one hot encoded categorical predictor variables: 64 variables total
- Had to remove one level from each one hot encoded variable due to dummy variable trap
- Left with 50 variables total

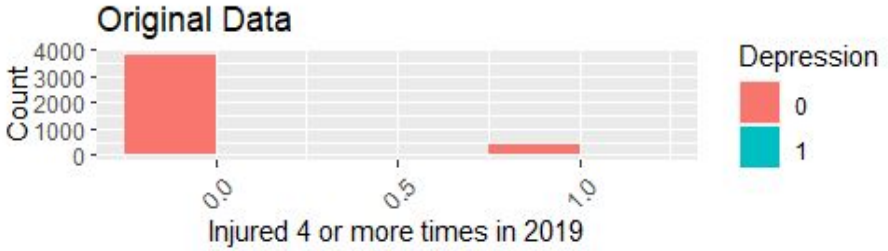
Using SMOTE (Systematic Minority Oversampling Technique)

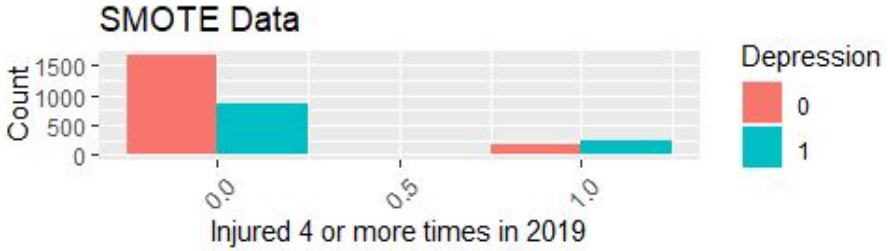
- Generates Synthetic minority class observations based on k nearest neighbors
- Used k of 5 (default k)
- Created even class distribution per current literature[1]
- Depression response distribution after SMOTE:

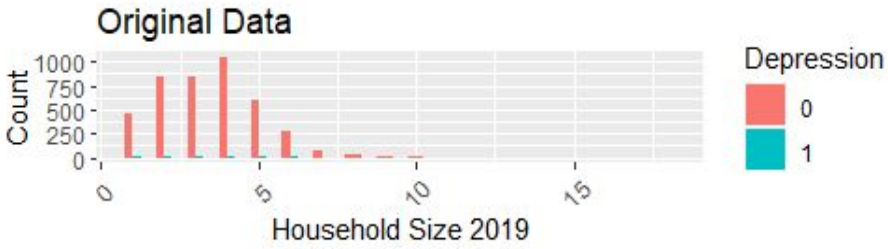
0	1
1863	1846

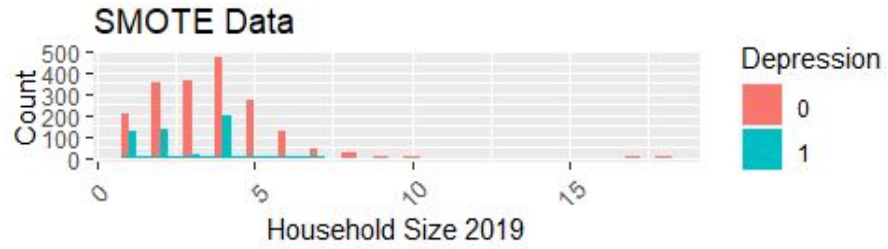
A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification

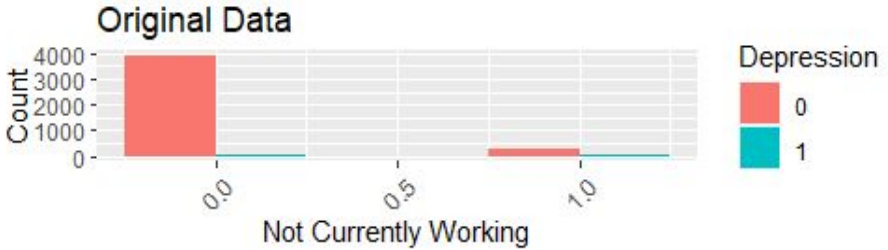
Distribution of Data with SMOTE

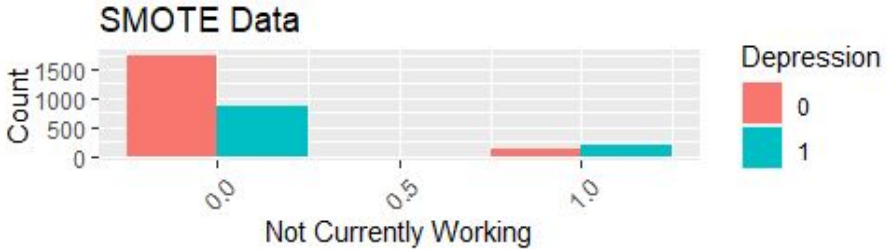












LASSO Regression w/ SMOTE data - Depression

- LASSO selected 18 variables, all other coefficients set to zero

- Accuracy: 78.77%
- Recall: 72.73%

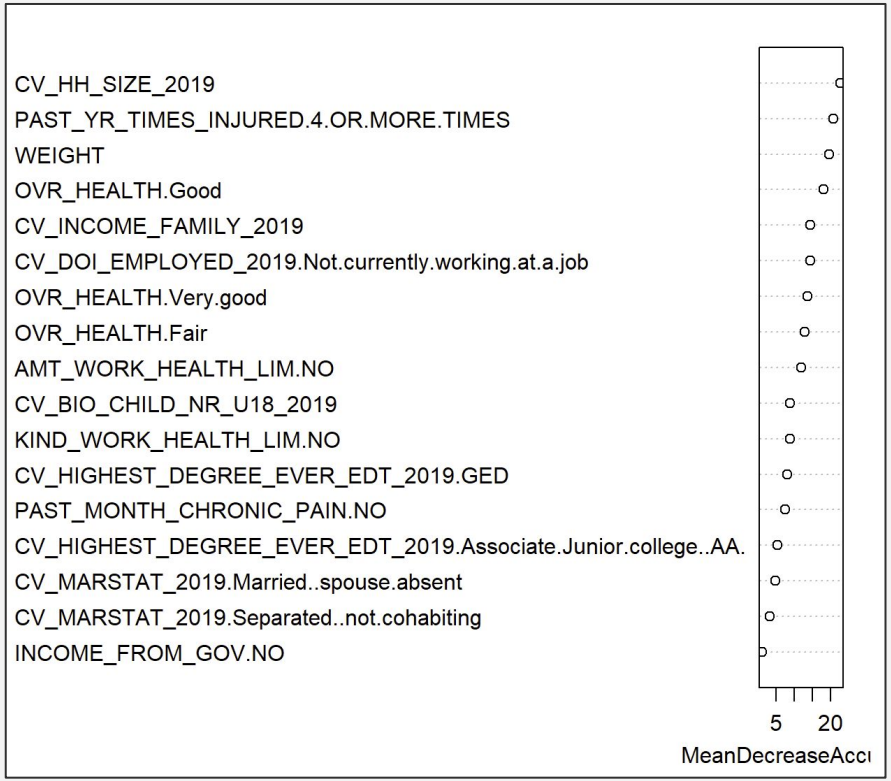
- McFadden's Pseudo R^2: 0.37

pred.dep	0	1
0	377	127
1	56	368

- Most important Variables:
  - Not currently employed in 2019
  - Household Size
  - 4 or more injuries in past year

Random Forest Classifier - Depression

- Accuracy: 96.34%
- Recall: 93.33%
- Mtry = 5
- trees = 100



Regression

Predicting Income in Respondents

- ➔ Income: total annual income in USD
- ➔ One-hot encoded predictors were used

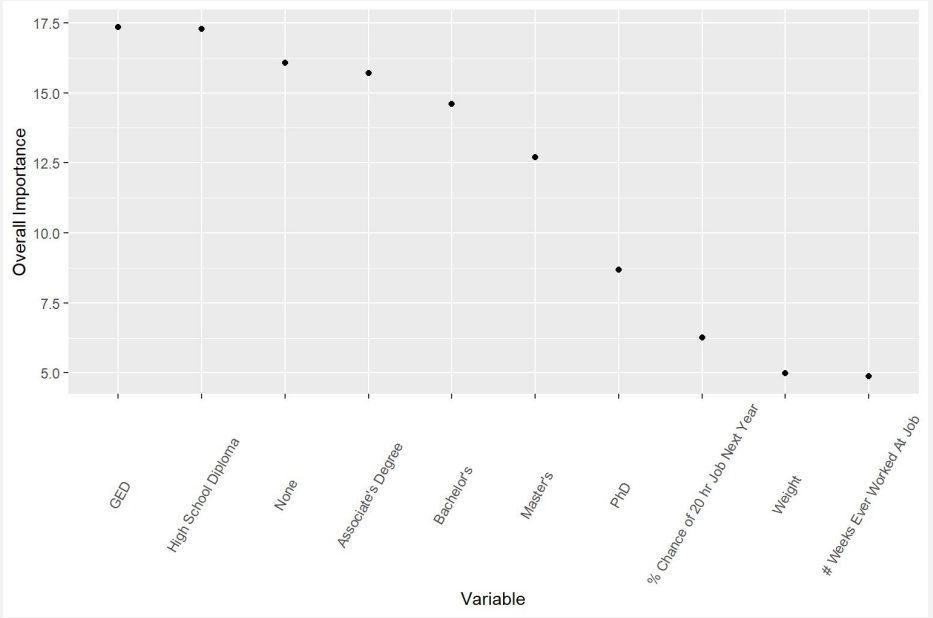
Linear Regression & K-Fold Cross Validation

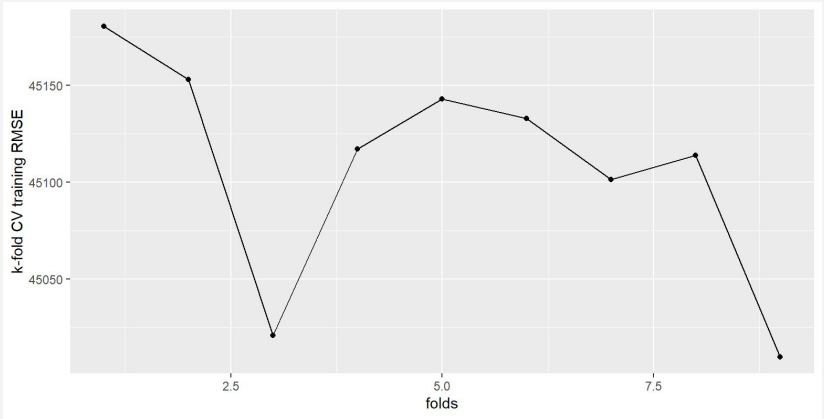
Linear Regression

- All predictor variables were used for both models
- Test RMSE: $44,881.39
- Most important predictors related to income & employment
- R^2 of 0.21

K-Fold CV

- Used K = 10
- CV RMSE: $45,009.73





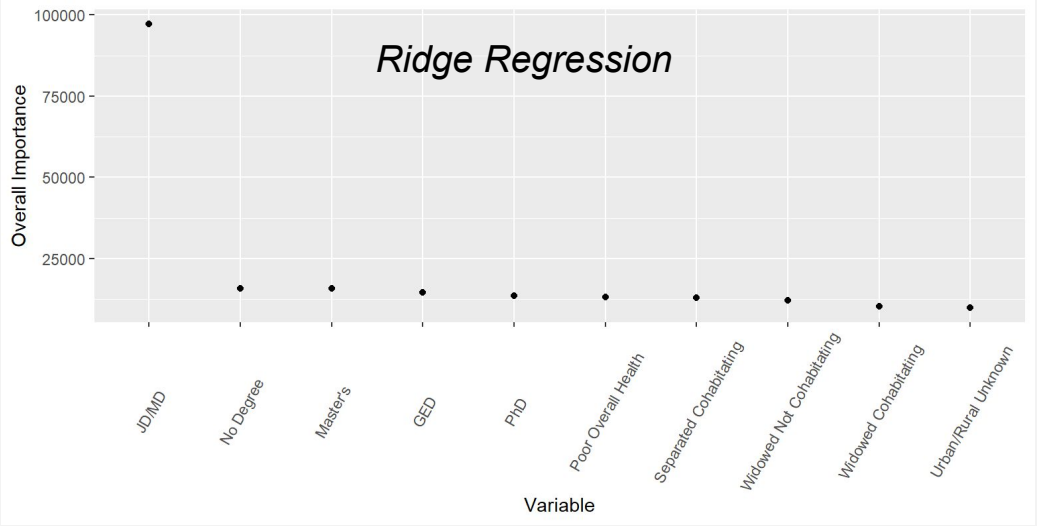
Lasso & Ridge Regression

Lasso regression test RMSE: $44,815.46

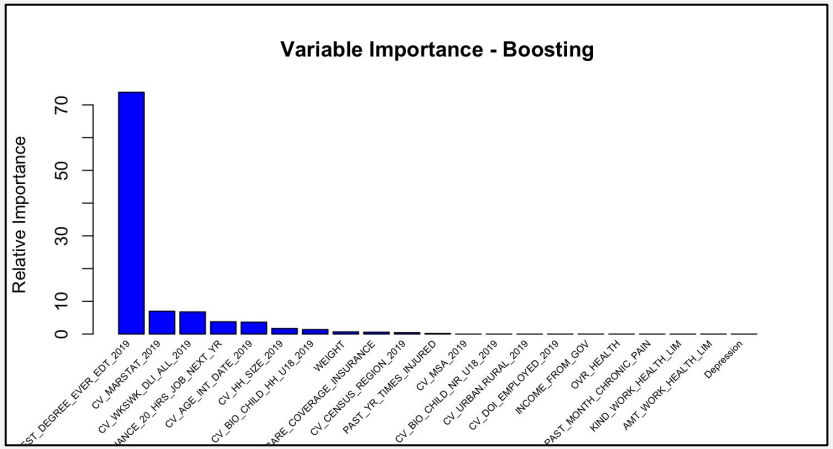
- Most important predictors related to education, health, and marital status

Ridge Regression test RMSE: $44,932.47

- Most important predictors related to education, health, and marital status with more emphasis on marital status



Boosting Trees

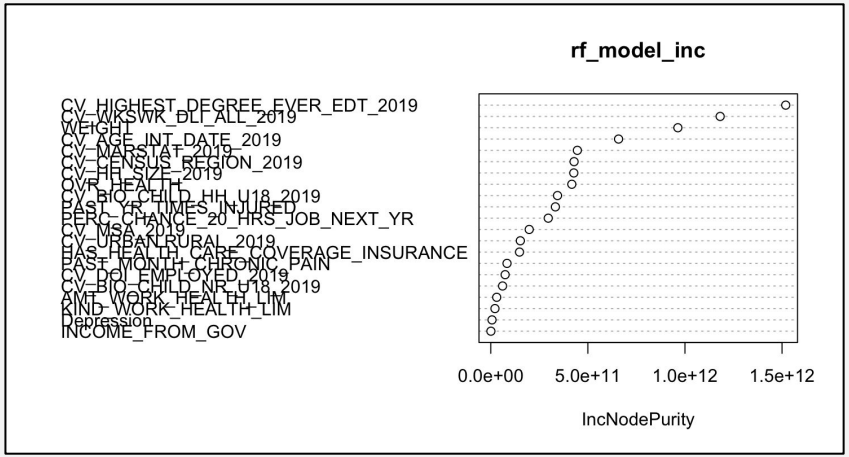


Variable importance for income using reduced boosting model.

- Full model RMSE of 47288.3
- Reduced model RMSE of 47470.01

➔ Little to no explanation of variance

Random Forest Regression Trees



Variable importance for income using reduced RF model

- Full model RMSE of 44782.52
- Reduced model RMSE of 46661.93

➔ Little to no explanation of variance

Summary

Classification (Symptoms of Depression)

- Success with SMOTE - but must be wary of overfitting, overlapping classes, loss of majority class information
- Random Forest was our best model
- Most important predictors for Depression:
  - Not being employed in 2019
  - being injured 4 or more times in 2019

Regression (Income)

- Consistently inaccurate prediction results
  - Poor test RMSE across all regression based models
- Difficult to predict income based on given predictors
  - Would be worthwhile to look into reassessing and choosing new predictors for future work
- Level of education was found to have the overall highest association with income level

Thank you!

Any questions?