BUILDING A MORTALITY PREDICTOR

AJ STRAUMAN-SCOTT

DATA 606 FINAL PROJECT

MAY 9TH, 2024

ABSTRACT

The paper investigates the impact of demographic and socioeconomic factors on mortality outcomes in the United States, utilizing the Public Use Microdata Sample of the National Longitudinal Mortality Study (NLMS) spanning from 1973 to 2011. By examining variables such as occupation, industry, income, education, race, and ethnicity, the research seeks to uncover the underlying relationships between socioeconomic status and identity demographics and mortality risk.

OVERVIEW

DATA

Single data set from Public Use Microdata Sample (PUMS) of National Longitudinal Mortality Study (NLMS) that tracks the 6 year mortality rates of 745162 individuals

DEPENDENT VARIABLES

Two outcome variables, for two models: death indicator binary variable, and numeric count of days survived after initial interview

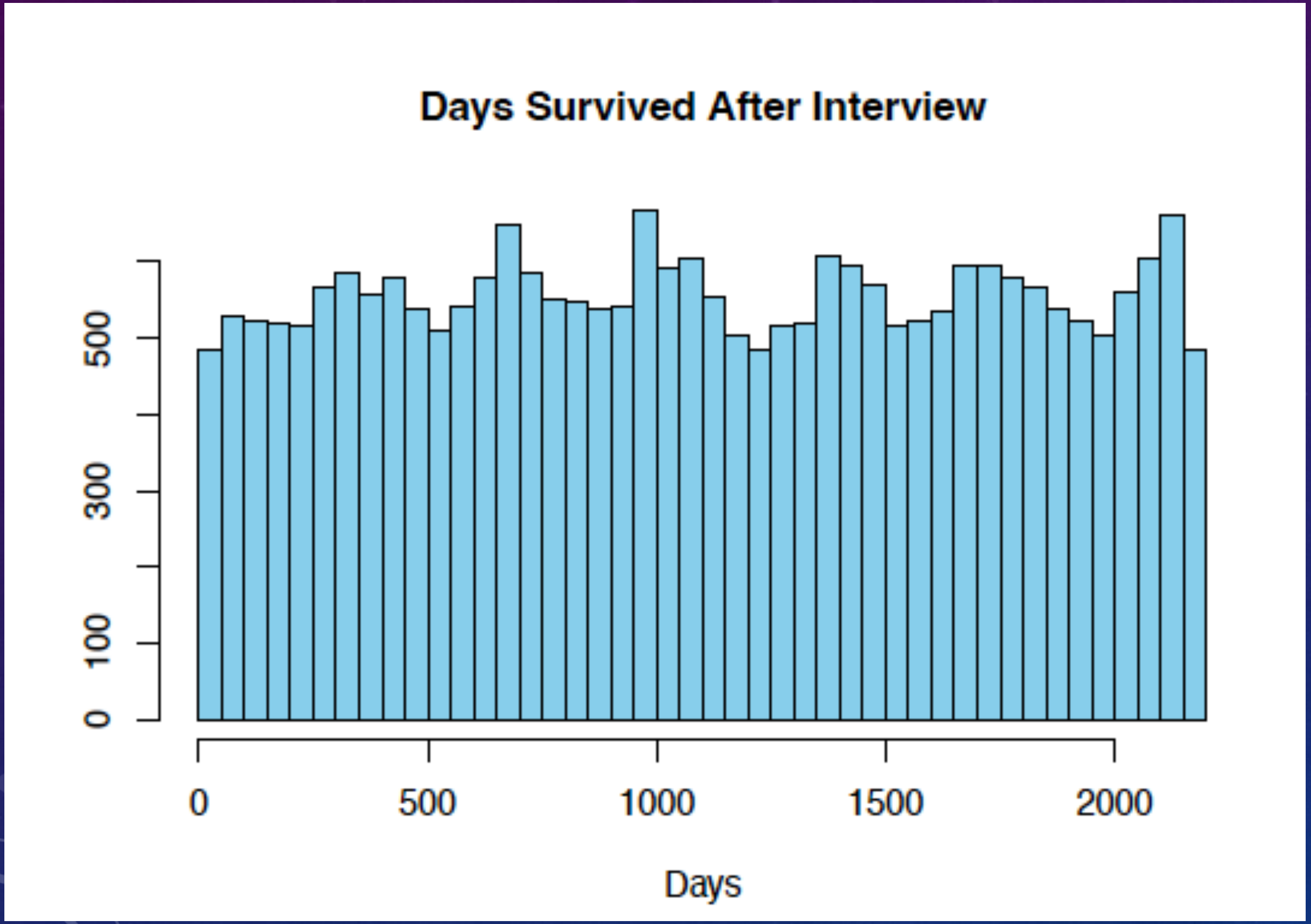
INDEPENDENT VARIABLES

Seventeen possible independent variables, all relating to the demographics and circumstances of the individual

What demographic variables are significant predictors of an individual's death within the next six years?

If that individual does pass away, what variables are significant predictors of the length of days that person survives from the initial interview?

SUMMARY STATISTICS



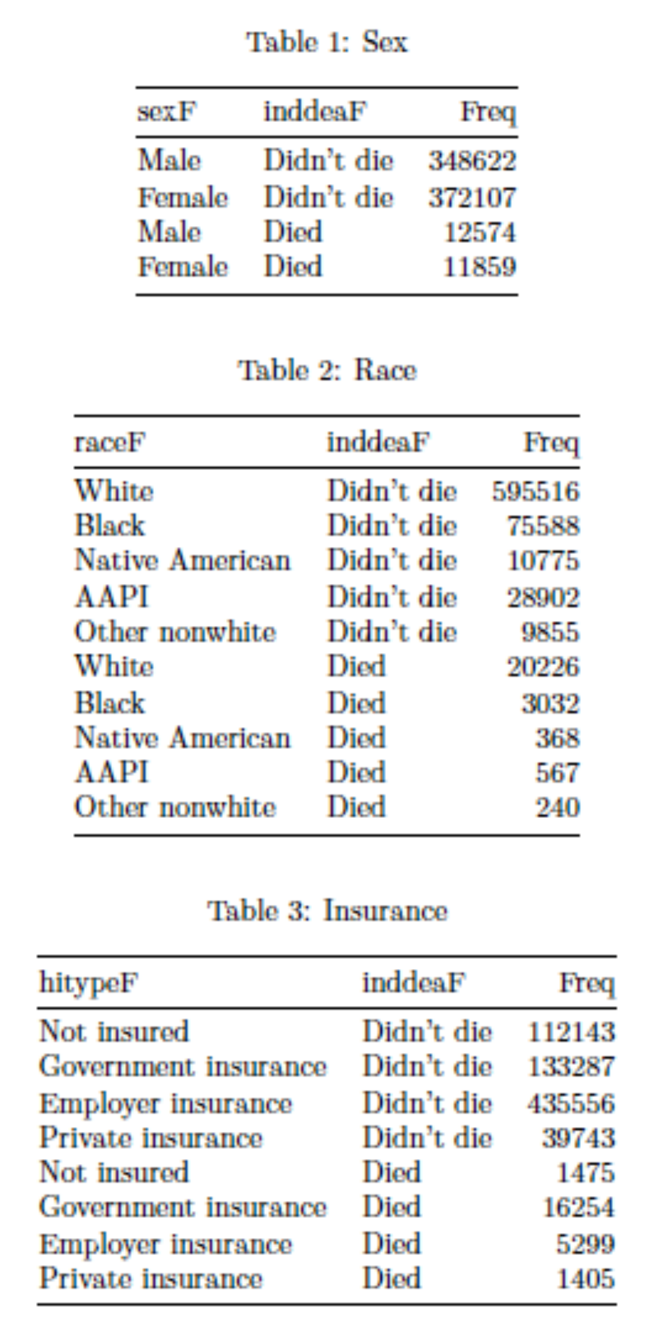


Table 1: Sex

sexF	inddeaF	Freq
Male	Didn't die	348622
Female	Didn't die	372107
Male	Died	12574
Female	Died	11859

Table 2: Race

raceF	inddeaF	Freq
White	Didn't die	595516
Black	Didn't die	75588
Native American	Didn't die	10775
AAPI	Didn't die	28902
Other nonwhite	Didn't die	9855
White	Died	20226
Black	Died	3032
Native American	Died	368
AAPI	Died	567
Other nonwhite	Died	240

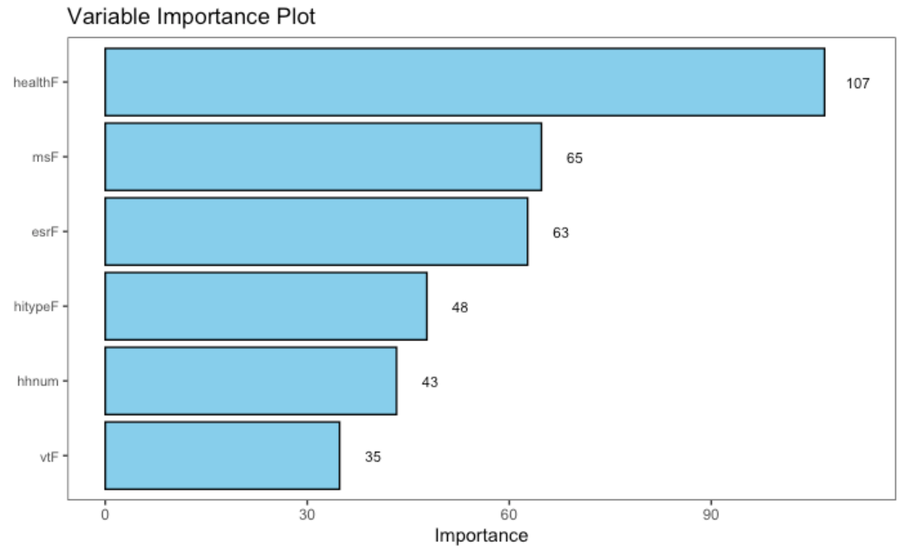
Table 3: Insurance

hitypeF	inddeaF	Freq
Not insured	Didn't die	112143
Government insurance	Didn't die	133287
Employer insurance	Didn't die	435556
Private insurance	Didn't die	39743
Not insured	Died	1475
Government insurance	Died	16254
Employer insurance	Died	5299
Private insurance	Died	1405

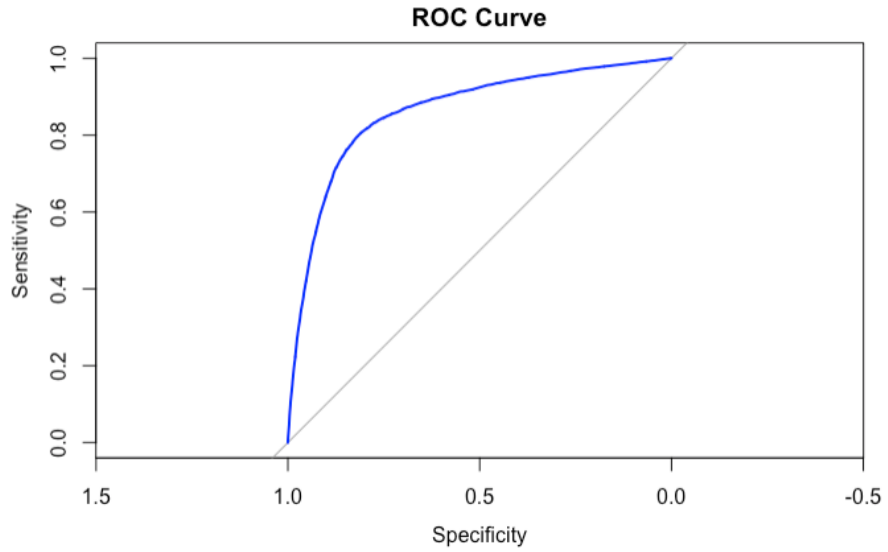
METHODOLOGY

Discover the independent demographic variables' relationship to each of the dependent variables, this analysis will construct two models:

- a logistic regression model of the independent variables' relationship to the death indicator variable, using the full dataset, titled CHANCE OF DEATH model
- a regression model of the independent variables' relationship to the count of days survived variable, using a subset of the data including only those who did die during the study period, titled DAYS LEFT model
    - Regression model dictated by variable’s residual distribution – uniform dependent variable presents challenges!



Variable Importance Plot



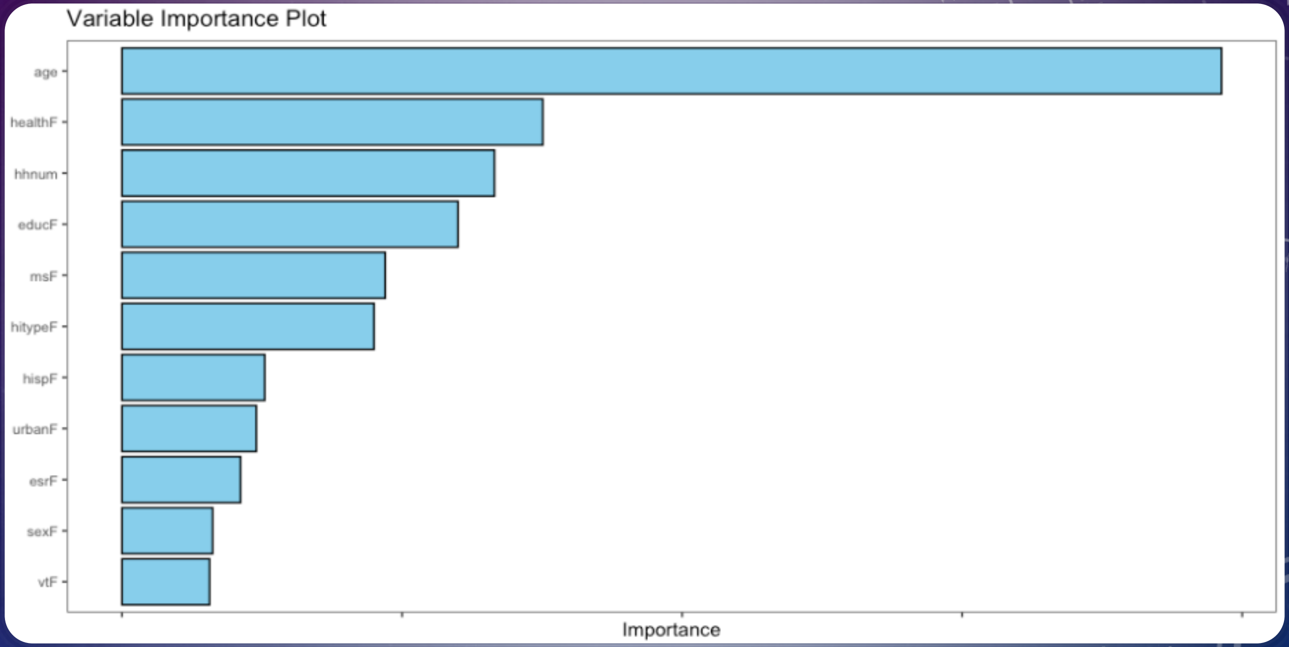
ROC Curve

CHANCE OF DEATH MODEL

- With six independent variables, it's unsurprising the most influential variable is the only variable related to health in the dataset. Marital status and employment type (white collar or blue collar) are nearly identically influential. Health insurance type, number of people in your household and veteran status are the least influential.
- A best-case ROC would look like a 90 degree angle. The ROC curve for the logistic model shows that the model balances specificity and sensitivity well

DAYS LEFT MODEL

- Age is the most important variable in predicting how many days the individual will survive after the interview.
- Quality of health, number of individuals in the household, and level of education are the next most influential.



CONCLUSION

- Encountered challenges such as non-normal residuals distributions and a dependent variable with a uniform distribution
- Fit both a logistic regression Chance of Death model and a randomForest regression Days Left model with moderate-high success
- The non-normal residuals distributions and the uniform distribution of the dependent variable necessitate cautious interpretation of the results.
- Further analysis can include training machine learning models on two unused datsets in the NLMS PUMS six year dataset that begins in 1990, and test them on a subset of the eleven year dataset (also begins in 1990), filtered to only include the deaths in the first six years of the eleven year study.