Will You Experience Unemployment?

Tania Ventura, Lauren Sedita, Kristine Lister

Abstract

As of 2017, there are 7.1 million unemployed persons in America. The National Longitudinal Surveys (NLS), sponsored by the U.S. Bureau of Labor Statistics (BLS) is a set of longitudinal studies with the goal to collect data on the labor force experiences of adults and young adults over their lifetime. **The purpose of this project is to train a model to predict whether someone will or will not experience unemployment in their lifetime based on the NLSY79 dataset published by the Bureau of Labor Statistics.**

1 Data Analysis

1.1 Data Background

The National Longitudinal Survey of the Youth (NLSY79) began in 1979, with a cohort of 12,686 young men and women who were 14 to 22 years of age when first surveyed. Survey questions delve into areas including labor market behavior, educational experiences, family background, government program participation, family life, health issues, assets and income, and the Armed Services Vocational Aptitude Battery, which measures knowledge and skills including reading and mathematics. The sample members have been surveyed every year from 1979 to 1994 then biennially from 1994 to 2014.

1.2 Dataset Characteristics

Our dataset is a subset of the NLSY79 survey and comprises of a variable matrix with 18 features and 9900 rows each pertaining to a single individual. Our column features are broken down by type: 5 nominal variables, 1 ordinal variable, and 12 variables of which 9 are discrete and 3 are continuous. These variables focus on different attributes of a person's life that can have an impact whether they will experience unemployment or not. We have determined that biennial surveys do not have an impact on our chosen features. The following features are:

CASEID, AGE1979, SEX, RACE, AFQT-3, ASVAB-8, HIGHESTGRADE, HEALTHLIMIT, INCOME, INDUSTRY, NUMJOBS, URBAN-RURAL, REGION, MARSTAT-COL (marital), YR ENT LF (year entered labor force), YR EMP (years employed after entering LF), YR UNEMP (years unemployed after leaving LF), YR OUT (years out of the LF).

1.2 Data Visualization

Several different visualization techniques were used to further understand our data. We started by making simple visuals to help understand the participants in the study based on demographics such as sex and race. We discovered that there are more females than males in the sample (5170 vs 4736), and that the majority race in the sample was Non-Black/Non-Hispanic (5866 people), followed by Black (2375 people), and Hispanic (1665 people). Distributions can be seen in the charts below.

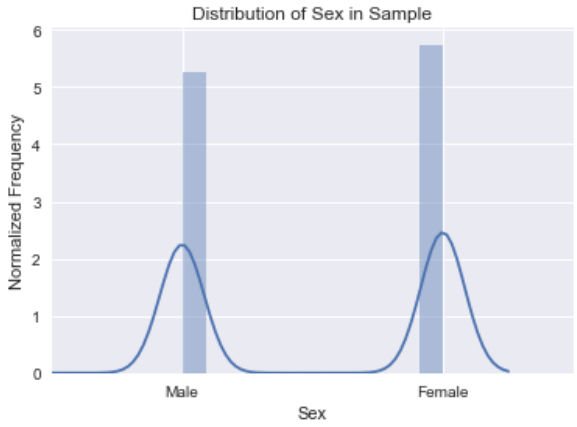


Figure 1. Distribution of Sex in Sample

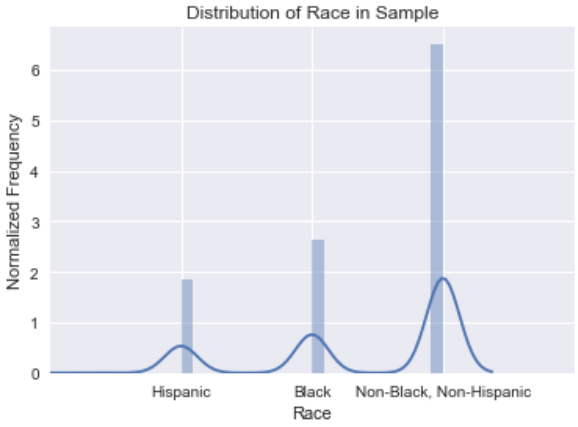


Figure 2. Distribution of Race in Sample

Using a target Boolean variable (0 never unemployed, 1 has experienced unemployment at least once), and intelligence scores AFQT-3 and ASVAB-8, we could try to understand how intelligence factors into unemployment. From the box plots below we can see that these test scores are generally higher for individuals who have not ever experienced unemployment (missing values were deleted before creating the plots).

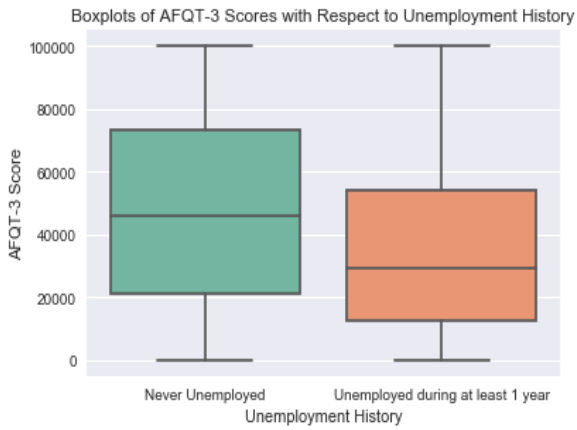
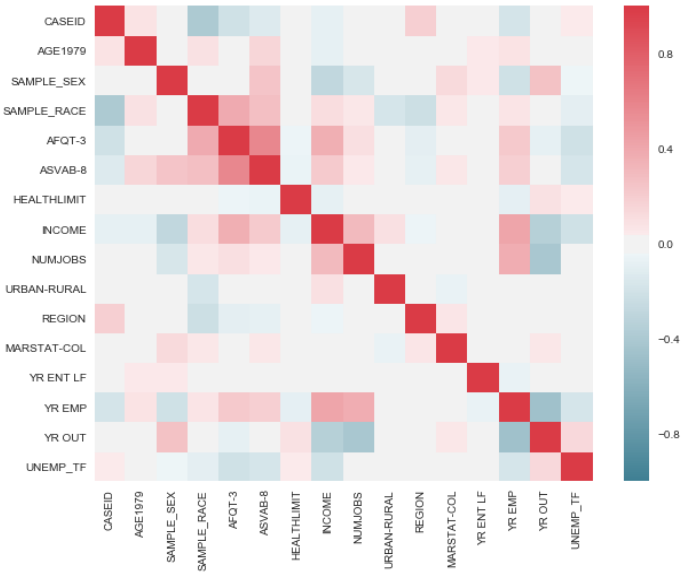


Figure 3. AFQT-3 Scores

Figure 4. AVSAB-8 Scores



We also wanted to find out just how correlated our different features were to one another. To do this we created a Correlation Visualization graph. To the left, we can see that variables in red have a positive correlation, and variables with a blue square have a negative correlation. Looking at our target variable, UNEMP_TF, there are not a lot of strong positive/negative correlations, which will likely cause us to have to do more feature engineering in the future and incorporate this knowledge when creating a model.

Figure 5. Correlation Visualization

1.3 Cleaning the data

As with any survey, there were a lot of areas of the dataset where there were a significant number of missing values. With this dataset, an individual may not have replied to a question because they refused to answer, were not interviewed at that time, or did not know the answer to a particular question. For the purpose of this project, we decided to eliminate any entries that had a missing value for any of the features. We also eliminated the features 'INDUSTRY' and 'HIGHEST GRADE' because there were an overwhelming number of missing values. After deleting this data to help clean up the set, we were left with 7,625 data points from the original 9,900. Our encoding methods and model were used with this cleaner data set.

1.4 Dummy Encoding

This data set includes demographic information such as sex, race, marital status, and region of the country where the individual lives. For variables such as urban-rural, the data was already in a format of 0s and 1s that we could utilize, but sex was labeled with 1s and 2s, race with 1s, 2s, and 3s, and region with 1s, 2s, 3s, and 4s, We decided that using dummy variables would be easier for interpretation, since a variable such as race or region with float values cannot be interpreted in the way that would make sense. Encoding was done on the cleaned data set.

2 Model Selection and Results

2.1 logistic regression

Because we are trying to decide whether one will experience unemployment in their lifetime, we felt that a logistic regression model would be a good place to start with since the outcome is binary. For a logistic regression model, the target variable must be discrete (unemployed at some time, or not). This would provide us with a good baseline model before attempting other classification algorithms to the data in the future. A logistic regression model has an objective of $h(x) = \log\left(1 + e^{(-yw^T x)}\right) + \lambda\|w\|_1$ to predict the probability of an outcome occurring given a set of predictor variables. Utilizing LogisticRegression() in Python's sklearn library, we did a first pass and got the following results on a test set (worth 20 percent of the total data):

	precision	recall	f1-score	support
0	0.69	0.85	0.76	935
1	0.63	0.41	0.49	590
avg / total	0.67	0.68	0.66	1525

The overall accuracy score on the test set was 67 percent.

3 Improvements

Moving forward we are planning on using a random forest next, as well as other models. We believe these would be appropriate models to use because it can help with data sets that have a large feature space as well as non-linear interactions.