Cornell University
Operations Research and
Information Engineering

An Analysis of Male Unemployment

ORIE 4741: Kristine Lister (kl799), Lauren Sedita (les278), Tania Ventura (tv72)

***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.**

.

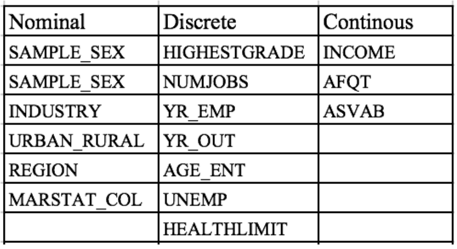
I. Data Analysis

A. Data Background

The National Longitudinal Survey of the Youth[1] (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.

B. Dataset Characteristics

Our dataset is a subset of the NLSY79 survey and comprise of a variable matrix with 16 features and 2,512 rows each pertaining to a single male individual. Our column features are broken down by type: 6 nominal variables, and 10 variables of which 7 are discrete and 3 are continuous (Figure 1). These variables focus on different attributes of a person's life that can have an impact whether they will experience unemployment or not. The following features are: YR_ENT_LF (year entered labor force), YR_EMP (years employed after entering LF), YR_OUT (years out of the LF), AGE_ENT (age when entered the labor force).

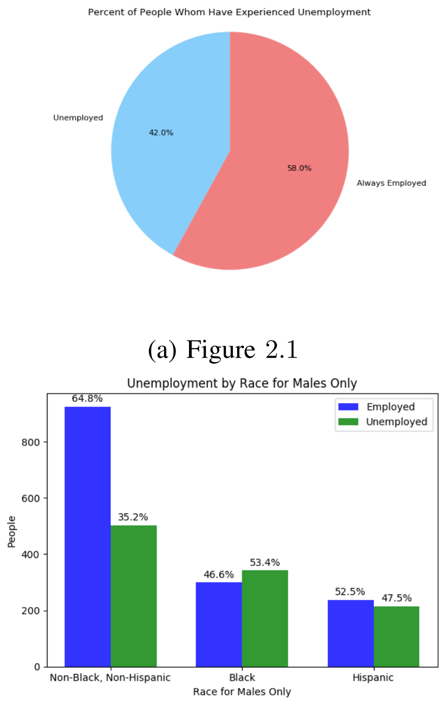


Nominal	Discrete	Continous
SAMPLE_SEX	HIGHESTGRADE	INCOME
SAMPLE_SEX	NUMJOBS	AFQT
INDUSTRY	YR_EMP	ASVAB
URBAN_RURAL	YR_OUT	
REGION	AGE_ENT	
MARSTAT_COL	UNEMP	
	HEALTHLIMIT	

Figure 1: Data Types of All Features.

C. Data Visualization

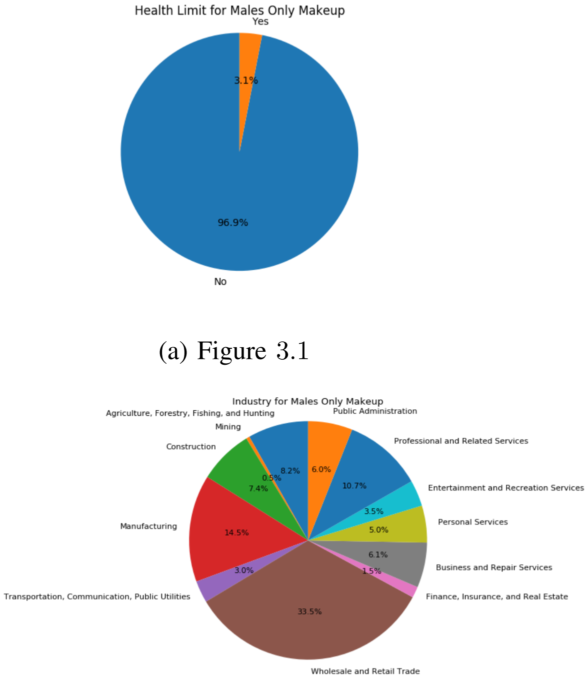
We mainly separated our data between bar graphs with corresponding pie charts. This made it very easy for our group to visualize distinct separations among our participants based on different features. It is useful to start helping us determine which features might be more significant when determining whether or not someone will experience unemployment. In the bar graphs, the height of the bars corresponds to the number of people in each category. Above each bar is a percentage, and that is the percentage of employed vs unemployed people in that category. For example, the unemployment by marriage status above the bars for "Never Married" the percentages are 48.4 and 51.6 which means 48.4 percent of men who never married never experienced unemployment and 51.6 percent of men who were never married have experienced unemployment.



(a) Figure 2.1

(b) Figure 2.2

Figure 2: The largest percentage gap of unemployment vs employment is within non-Black, non-Hispanic males.



(a) Figure 3.1

(b) Figure 3.2

Figure 3: Fig. 3.1 shows that only a small percentage of men have a health condition that limits their type of work. Fig. 3.2 shows industry.

D. *Cleaning the Data*

As with any survey, there were areas of the dataset where there were 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. While this removed a number of respondents, we felt that 2,512 was a large enough representation of the cohort.

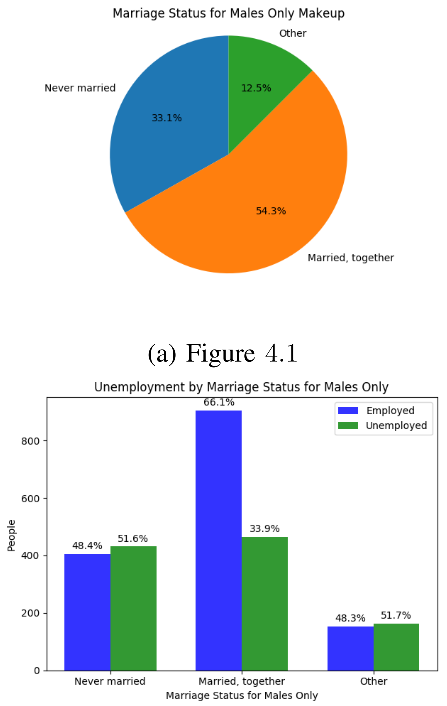
E. *Dummy Encoding*

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.

II. Model Selection and Results

A. *Logistic Regression Model*

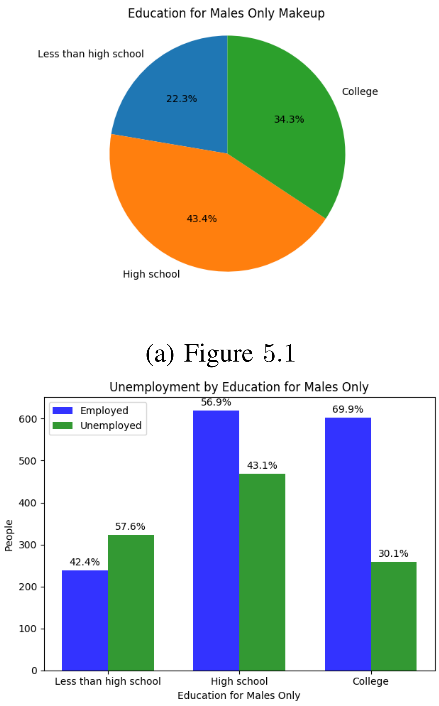
Logistic regression can be used to predict the probability that a person will experience unemployment. Logistic regression fits a lo-



(a) Figure 4.1

(b) Figure 4.2

Figure 4: The largest percentage gap of unemployment vs employment is within married males.



(a) Figure 5.1

(b) Figure 5.2

Figure 5: The largest percentage gap of unemployment vs employment in highest education is college.

gistic function to the data, and the logistic function is defined below:

$$P(Y=1) = \frac{1}{1+e^{-\beta^T X}}$$

where $P(Y=1)$ is the probability that one will experience unemployment, β is the vector of fitted coefficients, and X is the vector of observed predictor variables. If the predictor variables are independent, one can also use the model to get the odds ratio: the increase in probability of experiencing unemployment due to an increase of one unit of a continuous predictor variable or from observing one category over another in a categorical predictor variable. The proof of this is shown below: First define the odds function, the probability of experiencing unemployment divided by the probability of never experiencing unemployment:

$$Odds(X) = \frac{P(Y=1)}{P(Y=0)}$$

$$= \frac{P(Y=1)}{1-P(Y=1)} = e^{\beta^T X}$$

When one predictor variable of X is increased by one unit or one categorical variable is changed to another value, one can find the odds ratio. If there were only two predictor variables then an increase in one unit of the predictor variable, would lead to an odds ratio of:

$$Odds\ ratio = \frac{Odds(x+1)}{Odds(x)}$$

$$= \frac{e^{\beta_0+\beta_1 x_1+\beta_2(x_2+1)}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} = e^{\beta_2}$$

So the increase in the odds of experiencing unemployment due to observing one unit increase is related to the coefficient, and we can use the coefficients to determine how the predictor variables affect the probability of experiencing unemployment. In order to use this interpretation of the model coefficients, the independence of predictor variables must be determined.

1) Testing Correlation: Various statistical tests can be used to test for independence between predictor variables. For categorical predictor variables, one can use Pearson's Chi-Squared Test for Independence. For continuous predictor variables, one can use Pearson's Product Moment Correlation. Both of these tests are available in R packages. Pearson's Chi-Squared Test for Independence tests the null hypothesis that the variables are correlated. Therefore low p-values imply that the null hypothesis can be rejected and the variables are not correlated. The results of the test are shown in Figure 7 (see Tables section).

The only categorical predictor variable with correlations is the variable "Does health limit the amount of work you can do?" However because only 3% of the male respondents answered yes, we believe that the Pearson's Test may not be accurate. Therefore we will still use the question in our predictor set. Pearson's Product Moment Correlation Test tests the null hypothesis that the variables are not correlated. Therefore small p-values imply that one can reject the null hypothesis and assume that the variables are correlated. The results for this test are shown in Figure 8 (see Tables section).

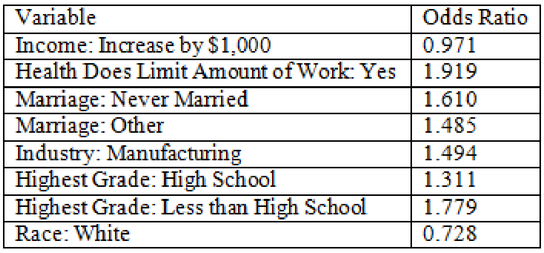
Because all of the p-values are below the significance level of 0.05, all of the continuous variables are correlated. Therefore to maintain the probabilistic interpretation of the model coefficients, only models that have one continuous variable can be used.

2) Model Selection: To narrow down the possible models to be considered, we use R's "bestglm" function with the objective of minimizing the Akaike Information Criterion, whose formula is shown below.

$$AIC = 2k - 2ln(L)$$

Where k is the number of parameters of the model and L is the likelihood function. AIC has both a regularization term, k, and a loss function, $-2ln(L)$. Given the type of model, a list of prediction variables, the outcome variable, and the selection criteria, the "bestglm" function returns the top five models that minimize the selection criteria, which allows us to use those models for validation. We ran the bestglm function on five sets of predictor variables; each subset had all of the categorical variables and only one of the continuous variables, in order to preserve independence between predictor variables. We found that the top five models with the continuous variables income and years out of the labor force had the lowest AIC values for all twenty-five models considered by bestglm. The set of variables included in the models and the corresponding AIC values are shown in Figure 9 (see Tables section).

To find the most accurate model, we use K-fold cross validation with K equal to 251. We evaluate our models using the error rate, the fraction of misclassified observations. This requires a cutoff level, α, for which observations with predicted probabilities over α are classified as having experienced unemployment and observations with predicted probabilities under α are classified as never being employed. In order to find the model with the smallest error rate, we ran K-fold cross validation on each model with α values ranging from 0.36 to 0.60, and found which α value minimized the error for each model. We then compared these smallest errors amongst the models and selected the model with the smallest error as the most accurate model. Our most accurate model had a balanced accuracy rate of 66.1%, corresponding to an α value of 0.48. The fitted coefficients and associated p-values for the best model are shown in Figure 10 (see Tables section). The calculated odds-ratio for each significant predictor variable are shown in Figure 6.



Variable	Odds Ratio
Income: Increase by $1,000	0.971
Health Does Limit Amount of Work: Yes	1.919
Marriage: Never Married	1.610
Marriage: Other	1.485
Industry: Manufacturing	1.494
Highest Grade: High School	1.311
Highest Grade: Less than High School	1.779
Race: White	0.728

Figure 6: Odds-Ratio for Significant Predictor Variables.

The odds ratios show that having a higher income and being non-Black and non-Hispanic both make one less likely to experience unemployment. Meanwhile having health limit the amount of work one is able to do, never having been married or not currently married, working in a manufacturing industry, and having less than a college degree all make one more likely to experience unemployment.

B. Decision Tree Model

A decision tree is either a classification or regression tree that aims to segment the predictor space into different regions. By segmenting this predictor space, observations can be predicted according to a set of rules which formulate the tree structure. Decision trees are simple to create and easy to interpret, and can handle mixed data types with ease. However, they can also be unstable with parameter tuning, which is a challenge that we encountered when doing cross validation for model parameters.

When segmenting the predictor space, we use a greedy approach and divide $Z \epsilon R^p$ with recursive binary splits. For each split, we choose the best possible one among all features without looking ahead that will minimize the variability within each rectangle. To do this, we tried two different loss functions, the Gini index and cross entropy. Each split attempts to minimize a specific criterion.

$$Gini: \; G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$Entropy: \; D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk})$$

Where $\hat{p}_{mk} = \frac{no.\ obs.\ in\ kth\ class}{no.\ obs.\ in\ R_m}$

A decision tree will continue to split the predictor space until each terminal node is small and each resulting box has very low variance. However, if the tree runs too deep and we make too many splits, there is potential for overfitting to the training set, and we will not be able to generalize. To fix this, we used cross validation to determine the optimal depth of the tree, which ended up being 7 branches.

For this project, we utilized a classification tree with two classes. Given a set of training data, which comprised 80 percent of the total data, we trained the tree and tested it with the remaining 20 percent. By utilizing grid search in Python's sklearn, we trained different models with various parameters for maximum depth, loss function, and maximum number of features to consider. Most of the time, the best result was utilizing the Gini index as a loss function (which is the default in sklearn), using all features when making decisions, and having a maximum tree depth of 7. However, as mentioned earlier in this section, trees can be unstable with parameter tuning, and we did have models which performed better with other values for these parameters with a different training set. Utilizing all of the features, on average we got an accuracy score of 75 percent on the test set with a single tree.

In order to dive deeper with the decision tree, we also decided to look at which features in the dataset were the most important for the model. We were able to use the attribute feature_importances_ in sklearn to look at "the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance" (sklearn documentation). Therefore, the higher the value, the more important the feature is for the model. Since the parameter tuning was fairly unstable throughout this process, we ended up with different feature importances for different models, but our best decision tree model utilized the following features: Age, AFQT and ASVAB scores, Highest Grade completed, income, number of jobs held, urban/rural, age when entered the labor force, years out of the labor force, region of country, and marital status. We decided to retrain the model using only these features and discarding the rest. This brought the model accuracy on the test set up to 78 percent on average.

C. Random Forest Model

We were very satisfied with the relative performance of the decision tree, but wanted to see if an ensemble method such as a random forest could perform even better. In many cases, the predictability of a decision tree is not as competitive as other machine learning models, and an ensemble method of using multiple trees can vastly improve accuracy. A random forest model uses a set number of decision trees to train, and outputs the mode of the classes from the individual trees. Different trees in the forest are trained using different parts of the training set, such as different combinations of features, in order to reduce the variance of the overall model. However, this does mean that random forests are not as easy to interpret as a single tree.

We trained a random forest on a training set comprising of 80 percent of the total data, and tested it on the remaining 20 percent. Like the decision tree model, we used grid search as a way of doing cross validation for parameters. We decided to use cross validation for the number of trees in the forest and maximum number of features to consider. In the end, we used a forest of 100 trees and used all features (although individual trees did use subsets). The overall accuracy on the test set from this random forest was about .78 on average, which was not much better in terms of performance than the decision tree. However, the random forest model was much more consistent in terms of results, and we did not have serious fluctuations in accuracy. This leads to a more robust model in general, although it is harder to interpret than the single decision tree.

III. Conclusion

Using techniques we learned in class (k-fold cross validation, logistic regression, decision trees) and techniques learned outside of class (random forests), we were able to find models that can predict whether one will experience unemployment given the dataset we used. With random forests, we achieved our best accuracy of 78%. Emphasizing what was mentioned earlier, some key results we found using logistic regression was that having a higher income and being non-Black and non-Hispanic both make one less likely to experience unemployment. Meanwhile having health limit the amount of work one is able to do, never having been married or not currently married, working in a manufacturing industry, and having less than a college degree all make one more likely to experience unemployment. Our results can be used by government agencies or non-profits to create programs that can start combating significant issues such as racial discrimination in hiring practices or removing barriers to achieve higher education and to recognize that low wage workers and people with health limitations are at a higher risk of experiencing unemployment.

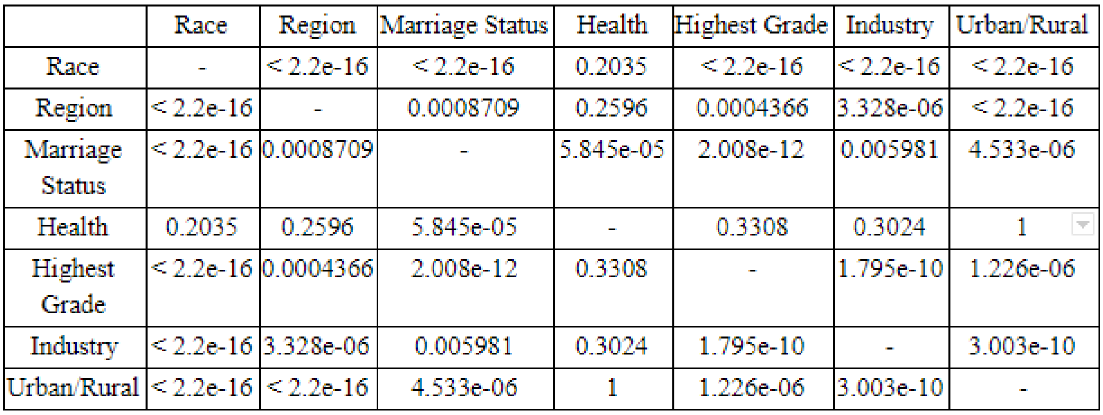
We do feel confident that our results can be used in production to see whether one will experience unemployment. Because we did not use female data, female participants might have had other key features that our models did not pick up as significant. In the future, there is potential to look into female participants and find models to compare to the results we achieved on this project.

IV. References

[1] U.S. Department of Labor by Center for Human Resource Research. (2001). A Guide

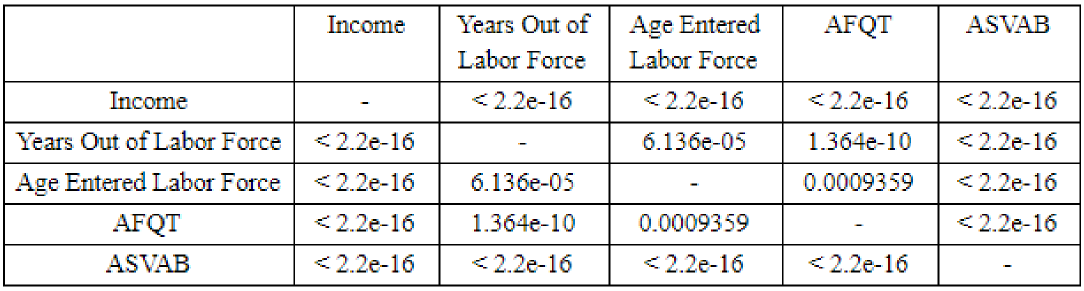
to the 1979–2000 National Longitudinal Survey of Youth Data. Columbus, OH: The Ohio State University. Retrieved from https://www.bls.gov/nls/79guide/2001/nls79g0.pdf.

V. Tables



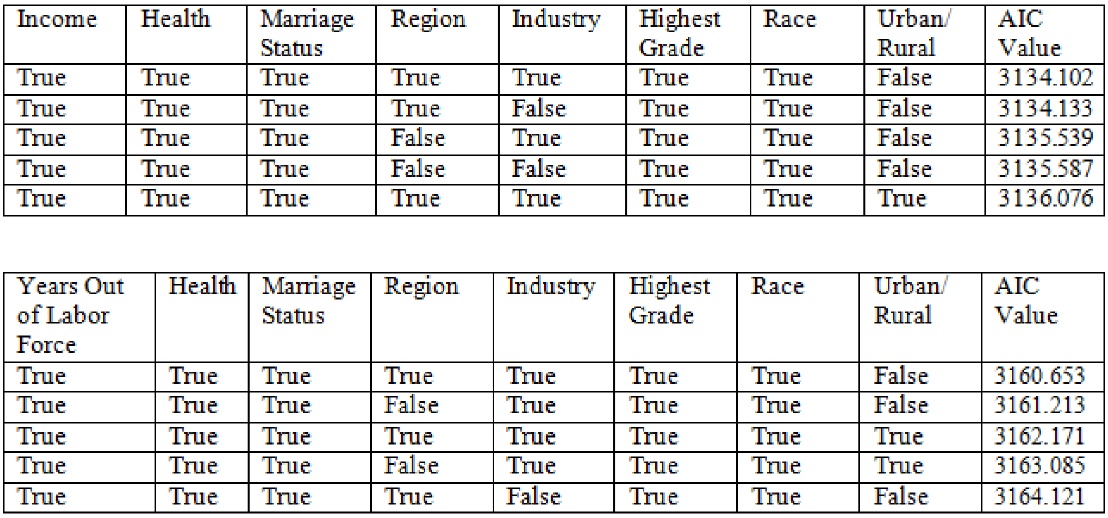
	Race	Region	Marriage Status	Health	Highest Grade	Industry	Urban/Rural
Race	-	< 2.2e-16	< 2.2e-16	0.2035	< 2.2e-16	< 2.2e-16	< 2.2e-16
Region	< 2.2e-16	-	0.0008709	0.2596	0.0004366	3.328e-06	< 2.2e-16
Marriage Status	< 2.2e-16	0.0008709	-	5.845e-05	2.008e-12	0.005981	4.533e-06
Health	0.2035	0.2596	5.845e-05	-	0.3308	0.3024	1
Highest Grade	< 2.2e-16	0.0004366	2.008e-12	0.3308	-	1.795e-10	1.226e-06
Industry	< 2.2e-16	3.328e-06	0.005981	0.3024	1.795e-10	-	3.003e-10
Urban/Rural	< 2.2e-16	< 2.2e-16	4.533e-06	1	1.226e-06	3.003e-10	-

Figure 7: Pearson's Chi-Square Test for Independence P-values for Categorical Variables.



	Income	Years Out of Labor Force	Age Entered Labor Force	AFQT	ASVAB
Income	-	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16
Years Out of Labor Force	< 2.2e-16	-	6.136e-05	1.364e-10	< 2.2e-16
Age Entered Labor Force	< 2.2e-16	6.136e-05	-	0.0009359	< 2.2e-16
AFQT	< 2.2e-16	1.364e-10	0.0009359	-	< 2.2e-16
ASVAB	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	-

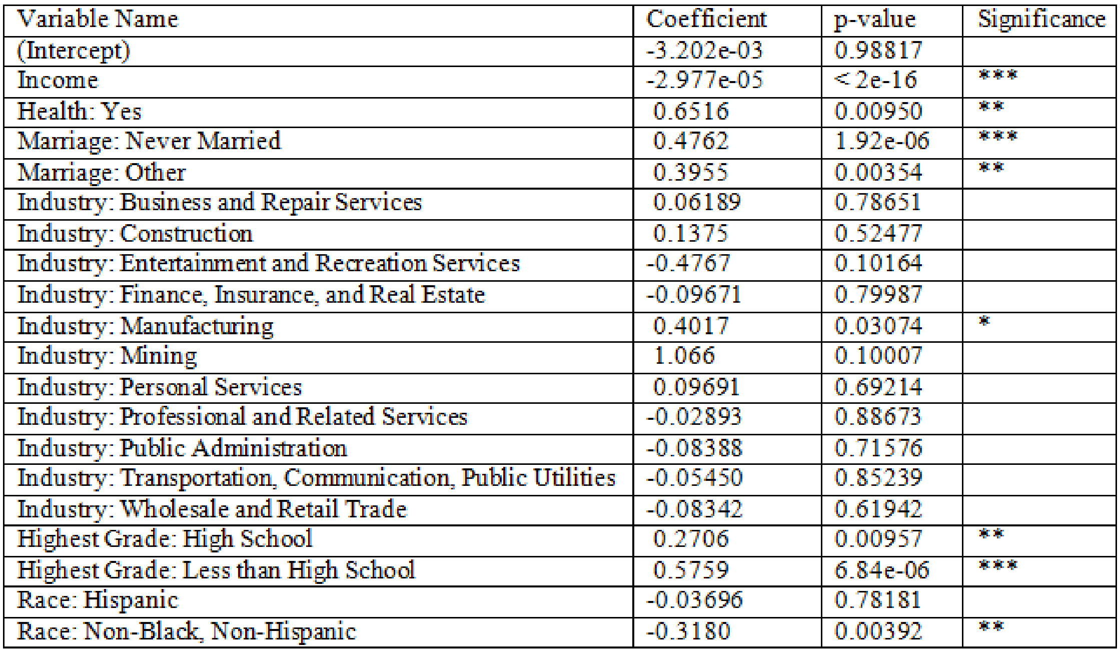
Figure 8: Pearson's Product Moment Correlation P-values for Continuous Variables.



Income	Health	Marriage Status	Region	Industry	Highest Grade	Race	Urban/ Rural	AIC Value
True	True	True	True	True	True	True	False	3134.102
True	True	True	True	False	True	True	False	3134.133
True	True	True	False	True	True	True	False	3135.539
True	True	True	False	False	True	True	False	3135.587
True	True	True	True	True	True	True	True	3136.076

Years Out of Labor Force	Health	Marriage Status	Region	Industry	Highest Grade	Race	Urban/ Rural	AIC Value
True	True	True	True	True	True	True	False	3160.653
True	True	True	False	True	True	True	False	3161.213
True	True	True	True	True	True	True	True	3162.171
True	True	True	False	True	True	True	True	3163.085
True	True	True	True	False	True	True	False	3164.121

Figure 9: Top Ten Models by AIC Values.



Variable Name	Coefficient	p-value	Significance
(Intercept)	-3.202e-03	0.98817	
Income	-2.977e-05	< 2e-16	***
Health: Yes	0.6516	0.00950	**
Marriage: Never Married	0.4762	1.92e-06	***
Marriage: Other	0.3955	0.00354	**
Industry: Business and Repair Services	0.06189	0.78651	
Industry: Construction	0.1375	0.52477	
Industry: Entertainment and Recreation Services	-0.4767	0.10164	
Industry: Finance, Insurance, and Real Estate	-0.09671	0.79987	
Industry: Manufacturing	0.4017	0.03074	*
Industry: Mining	1.066	0.10007	
Industry: Personal Services	0.09691	0.69214	
Industry: Professional and Related Services	-0.02893	0.88673	
Industry: Public Administration	-0.08388	0.71576	
Industry: Transportation, Communication, Public Utilities	-0.05450	0.85239	
Industry: Wholesale and Retail Trade	-0.08342	0.61942	
Highest Grade: High School	0.2706	0.00957	**
Highest Grade: Less than High School	0.5759	6.84e-06	***
Race: Hispanic	-0.03696	0.78181	
Race: Non-Black, Non-Hispanic	-0.3180	0.00392	**

Figure 10: Logistic Regression Results for Best Model. Significance codes: 0: ‘***’, 0.001: ‘**’, 0.01: ‘*’, 0.05: ‘.’