PREDICTING UNEMPLOYMENT

WHICH VARIABLES PLAY THE MOST SIGNIFICANT ROLE IN PREDICTING EMPLOYMENT STATUS?

Austin Cicale, Theodore Popkin, Josiah Jones

I. Introduction

Unemployment rate is one of the most talked about economic indicators and is often used as a measure of how well an economy is performing. Unemployment rate refers to the number of people actively looking for employment divided by the total number of people in the labor force. On a societal level, high unemployment rate means that an economy is underperforming, since output and consumption are both lower than they should be. On an individual level, high levels of unemployment means that more people are not able to afford necessary goods and provide for themselves or their families. Extremely low unemployment is also not ideal because it could point to an overheating economy, where businesses can't hire additional workers and employees have difficulty finding more suitable jobs.[1] Despite its importance to individuals and societies, predicting unemployment is a particularly difficult task. That is what we seek to address with our research. In this paper, we will address which factors about an individual play the most significant role in determining employment status. Specifically, we focus on unemployment in 2021.

Accurately predicting unemployment could be of significant importance at an individual, organizational, and societal level. At an individual level, it would allow us to determine which factors are more likely to lead to unemployment. That is, predicting unemployment would allow individuals more information on the consequences of their potential choices and would therefore allow people to better manage risk. For example, if we know that during business cycles, people

[1] Hayes, A. (2023, August 9). *What Is Unemployment? Understanding Causes, Types, Measurement*. Investopedia. https://www.investopedia.com/terms/u/unemployment.asp

with Trait X are more likely to be employed, it could be an incentive for people to move towards Trait X or gain skills that relate with that trait. At an organizational level, it would also be helpful for businesses. If companies were able to know what unemployment will look like in the future, they would be better equipped to match candidates with positions and adjust resource allocation to focus on hiring plans, training programs, etc.

Finally, at a societal level, predicting unemployment would be helpful in implementing policies, especially with monetary and fiscal policy. The Federal Reserve could use predictions to create economic stability and minimize the impacts of economic downturns. At the same time, having a clear understanding of unemployment will allow the public to have more confidence in the economy, which would allow for more long-term planning (investing, purchasing houses, etc). It could also inform decisions in other sectors of policies such as inequality; specifically, predicting unemployment could be a measure of society's biases. For example, if we determine that people of a certain race are more likely to be unemployed than others, it may point to more systemic issues that should be addressed, or at least require further research into why. This would allow politicians to address economic inequality and other systemic biases.

To create our model, we will be using data from the Integrated Public Use Microdata Series (IPUMS) and American Community Survey (ACS) to obtain information regarding unemployment rates of each month in 2021 and characteristics of people who were unemployed during this time.[2] Thus, we will be contributing to this question by focusing on if we are able to predict if someone will be unemployed based on specific characteristics they have. Before we began modeling the data, we assumed the biggest factors on one's unemployment status are education level, race, and age. We determined these as the biggest factors that play a role into whether one is hired or not and therefore would play the biggest part in unemployment. After

[2] IPUMS USA. https://usa.ipums.org/usa/

doing our models, we found that there are actually a number of different factors that play a role in employment status. This makes sense, but we were intrigued that there were a number of factors that go beyond what goes on the job application, such as marriage status or number of years spent in the United States.

II. Literature Review

There have been a number of ways researchers have attempted to predict unemployment. Alternative attempts to predict unemployment use variables like unemployment insurance claims[3], future expectations of the economy[4], and the comovement between aggregate economic activity and unemployment flows (i.e. Tasci model)[5]. Furthermore, there are also autoregressive models and professional forecasters to predict unemployment. The Federal Reserve publishes some of these in the Greenbook/Tealbook and the Survey of Professional Forecasters.

The results from these models differ from ours because other models tend to focus on factors that influence the unemployment rate, but our model seeks to answer what factors influence *who* becomes unemployed. That is, other models look at a macro-level, but we want to zoom in to see if there is any way to predict which individuals are more likely to become unemployed. This seems like a more interesting question to us, because it makes our conclusions more applicable to the average citizen who reads them. While the conclusions of other models focus on how many people will be unemployed, our data will tell people which factors or attributes they should try to attain in order to stay employed.

---

[3] Zheng, P. (Claire). (2020, Summer). *Predicting unemployment from unemployment insurance claims*. Predicting Unemployment from Unemployment Insurance Claims. https://www.ibrc.indiana.edu/ibr/2020/summer/article1.html

[4] Blanchflower, D. G., & Bryson, A. (2021, August 23). *The economics of walking about and predicting unemployment*. NBER. https://www.nber.org/papers/w29172

[5] Meyer, B., & Tasci, M. (2015, February). *Lessons for Forecasting Unemployment in the United States: Use Flow Rates, Mind the Trend*. Federal Reserve Bank of Atlanta: Working Paper Series. https://www.atlantafed.org/-/media/documents/research/publications/wp/2015/01.pdf

The results from a study like this could possibly be very divisive, as we could conclude that people of certain races, genders, or other inherent attributes that people can't choose are more likely to be employed than people who don't meet that standard. This could be problematic and quickly become a political issue. Furthermore, our conclusions could be quite varied depending on the circumstances. For example, if we had more time and resources, it might be more helpful (and accurate) to look at different markets across time and space to see which variables prove themselves to be most important. For example, variables that affect who is unemployed would probably differ dramatically between China in 2000 and the United States in 2023 (or even within just one country in that timeframe). Therefore, our conclusions are probably not able to tell a fully reliable casual story. Still, we think that the model/strategy is beneficial and should be further studied because of its applicability if done right.

III. The Model

Our exploration into predicting unemployment for individuals, within the labor force, in 2021 is underpinned by a pragmatic approach, focusing on fundamental assumptions and an empirical model designed for clarity and relevance.

Behavioral Assumptions

At the core of our model are behavioral assumptions that individuals make employment decisions based on discernible factors. Education, age, and race are identified as pivotal determinants, and we posit that these decisions fluctuate in response to broader economic cycles. For instance, certain traits may gain prominence during economic upswings, influencing hiring patterns.

Informational Assumptions

Our model assumes that individuals possess information about their own characteristics and a general awareness of prevailing economic conditions. However, we acknowledge the inherent imprecision in this information. Perfect foresight is a rarity, and individuals may make decisions based on imperfect but accessible knowledge.

Institutional Assumptions

Considering the institutional backdrop, our model takes into account the broader economic environment, labor market policies, and societal norms of the United States in 2021. While the primary focus is on individual characteristics, we recognize the external influences shaping employment opportunities, including governmental policies, industry practices, and unexpected events such as COVID-19.

Empirical Model

The empirical model is rooted in logistic regression, chosen for its simplicity and effectiveness in analyzing relationships between employment status and key variables. Education levels, race, and age emerge as focal points in our analysis. The empirical equation takes the form: $Employment\ Status = \beta 0 + \beta 1 \times Education + \beta 2 \times Race + \beta 3 \times Age + \epsilon$

- Here, β0 is the intercept,
- β1, β2, β3 are coefficients,
- and ϵ represents the error term.

This simplicity aligns with our objective of providing transparent insights without unnecessary complexity.

While this model captures essential relationships, we acknowledge its simplicity and the need for potential refinements based on empirical observations. Flexibility is inherent, and we

remain open to adjustments as we delve deeper into the intricacies of unemployment dynamics. In essence, our model is a practical tool that bridges theoretical foundations with real-world observations, offering a clear lens into the factors influencing employment status. As our research unfolds, we anticipate further nuances that may necessitate fine-tuning, ensuring the model's continued relevance and effectiveness.

IV. Data Description

Data Source and Collection Period

The dataset utilized for this study originates from IPUMS, obtained through their comprehensive online database. Specifically, the data derives from the ACS and encompasses information collected during the year 2021.

Nature of Data and Observation Units

The nature of the data is cross-sectional, focusing on individuals residing within the United States throughout the calendar year of 2021. Each observation within the dataset represents an individual respondent. Initially, the dataset comprised a total of 3,252,599 observations. Following thorough data cleaning procedures, the sample was refined to a dataset containing 261,202 observations, ensuring data accuracy and reliability for subsequent analyses.

Limitation and Imbalance in the Dataset

An inherent limitation within the dataset surfaced due to an imbalance in the distribution of observations between employed and unemployed individuals. A significant majority, approximately 94% of the observations, pertained to individuals classified as employed. This skewness in class distribution posed a challenge in effectively scrutinizing the nuanced relationships distinguishing the employed from the unemployed cohort.

This imbalance could significantly impact the validity of classification modeling techniques, as they tend to favor the majority class (employed individuals) in their predictions, thereby hindering the exploration of underlying relationships among the diverse groups.

To address this imbalance, oversampling techniques were employed, specifically leveraging the ROSE package. This approach facilitated the creation of a more balanced training dataset by artificially enhancing the representation of the minority class (unemployed individuals). This strategic oversampling aimed to rectify the disproportionate class distribution, enabling a more comprehensive analysis of the underlying relationships between the employed and unemployed segments of the dataset.

Descriptive Data Statistics

Table 1 presents comprehensive statistical information for all fifteen variables encompassed within our dataset, delineated by variable abbreviations alongside descriptive explanations. The statistical summaries are stratified based on the employment status, our dependent variable. The initial row in the table represents the employment status variable (EMPSTAT), depicting the count of individuals classified as employed or unemployed within our dataset. For numeric variables such as FAMSIZE, NCHILD, and AGE, the statistical summary section provides the mean and standard deviation for each variable, distinguished by employment status. Conversely, categorical variables in our dataset exhibit the levels within each category alongside observation counts for every level, further stratified based on employment status.

Table 1. Descriptive Statistics of Data

Variable	Description	Statistical Summary	
EMPSTAT	Employment status	**Employed** Count: 245134	**Unemployed** Count: 16068
FAMSIZE	Number of own family members in household	**Employed** Mean: 3.35 SD: 1.79	**Unemployed** Mean: 3.41 SD: 1.89
NCHILD	Number of own children in household	**Employed** Mean: 1.02 SD: 1.18	**Unemployed** Mean: 0.91 SD: 1.15
RELATE	Relationship to household head	**Employed** Head (1): 121877 Spouse (2): 68851 Child (3): 16237 Child-in-law (4): 1819 Parent (5): 4938 Parent-in-law (6): 764 Sibling (7): 4994 Sibling-in-law (8): 1425 Grandchild (9): 419 Other relatives (10): 3506 Partner, friend, visitor (11): 13896 Other non-relatives (12): 6408	**Unemployed** Head (1): 6485 Spouse (2): 4216 Child (3): 2013 Child-in-law (4): 154 Parent (5): 570 Parent-in-law (6): 80 Sibling (7): 478 Sibling-in-law (8): 130 Grandchild (9): 62 Other relatives (10): 325 Partner, friend, visitor (11): 962 Other non-relatives (12): 593
SEX	Sex	**Employed** Male: 133050 Female: 112084	**Unemployed** Male: 7730 Female: 8338
AGE	Age	**Employed** Mean: 45.28 SD: 13.10	**Unemployed** Mean: 44.63 SD: 14.54
MARST	Marital status	**Employed** Married, spouse present (1): 148067 Married, spouse absent (2): 9199 Separated (3): 5939 Divorced (4): 21406 Widowed (5): 4559 Never married/single (6): 55964	**Unemployed** Married, spouse present (1): 8116 Married, spouse absent (2): 705 Separated (3): 475 Divorced (4): 1650 Widowed (5): 358 Never married/single (6): 4764
RACE	Race	**Employed** White (1): 5699 Black/African American (2): 17844 American Indian (3): 3120 Chinese (4): 18222 Japanese (5): 1706 Other Asian (6): 55736 Other race, nec (7): 43857	**Unemployed** White (1): 3527 Black/African American (2): 1590 American Indian (3): 205 Chinese (4): 1251 Japanese (5): 86 Other Asian (6): 3265 Other race, nec (7): 2929

		Two major races (8): 45611 Three + major races (9): 2039	Two major races (8): 3055 Three + major races (9): 163
HISPAN	Hispanic origin	**Employed** Not Hispanic (0): 149543 Mexican (1): 52075 Puerto Rican (2): 848 Cuban (3): 5810 Other (4): 36858	**Unemployed** Not Hispanic (0): 9715 Mexican (1): 3164 Puerto Rican (2): 61 Cuban (3): 353 Other (4): 2776
CITIZEN	Citizenship status	**Employed** US citizen born abroad (1): 16480 Naturalized citizen (2): 129376 Not a citizen (3): 99278	**Unemployed** US citizen born abroad (1): 1108 Naturalized citizen (2): 8284 Not a citizen (3): 6676
YRSUSA2	Years in the United States, intervalled	**Employed** 0 to 5 years (1): 26954 6 to 10 years (2): 25480 11 to 15 years (3): 26160 16 to 20 years (4): 32440 21+ years (5): 134100	**Unemployed** 0 to 5 years (1): 2066 6 to 10 years (2): 1824 11 to 15 years (3): 1681 16 to 20 years (4): 2108 21+ years (5): 8389
SPEAKENG	Speaks English	**Employed** Does not speak English (1): 12042 Yes, speaks only English (3): 52476 Yes, speaks very well (4): 98619 Yes, speaks well (5): 50426 Yes, but not well (6): 31571	**Unemployed** Does not speak English (1): 1056 Yes, speaks only English (3): 3257 Yes, speaks very well (4): 5737 Yes, speaks well (5): 3386 Yes, but not well (6): 2632
SCHOOL	School attendance	**Employed** No, not in school (1): 227274 Yes, in school (2): 17860	**Unemployed** No, not in school (1): 14402 Yes, in school (2): 1666
EDUCD	Educational attainment	**Employed** No schooling completed (2): 11158 Kindergarten or preschool (3): 283 Grade 1,2,3,4, or 5 (4): 4414 Grade 6,7, or 8 (5): 10765 Grade 9,10, or 11 (6): 10639 Some college, no degree (7): 32993 12th grade, no diploma (61): 7913 High school diploma (63): 41814 GED or equivalent (64): 6388 Associate's degree (81): 16777 Bachelor's degree (101): 53495 Master's degree (114): 32182 Degree beyond bachelor's (115): 7486 Doctoral degree (116): 8827	**Unemployed** No schooling completed (2): 943 Kindergarten or preschool (3): 18 Grade 1,2,3,4, or 5 (4): 287 Grade 6,7, or 8 (5): 693 Grade 9,10, or 11 (6): 869 Some college, no degree (7): 2688 12th grade, no diploma (61): 694 High school diploma (63): 3402 GED or equivalent (64): 535 Associate's degree (81): 1194 Bachelor's degree (101): 3015 Master's degree (114): 1330 Degree beyond bachelor's (115): 217 Doctoral degree (116): 183
VETSTAT	Veteran status	**Employed** Veteran: 4780 Not a veteran: 240354	**Unemployed** Veteran: 277 Not a veteran: 15791

V. Econometric Model

Basic Specification

To comprehend the relationship between variables and employment status, a logistic regression model was formulated to predict employment status using all available observations and variables within the dataset.

Table 2 presents a summary of predictor variables utilized in the basic model. Several predictor variables displayed varying degrees of significance in predicting employment status. Notably, FAMSIZE ($p \approx 0.52$), NCHILD ($p \approx 0.22$), and VETSTAT ($p \approx 0.93$) exhibited notably high p-values, signifying a lack of substantial association with employment status. Additionally, none of the CITIZEN factor variables showed statistical significance, except for one CITIZEN factor variable, demonstrating proximity with a p-value of approximately 0.10.

Conversely, a majority of the variables showcased statistical significance as predictors of EMPSTAT. Particularly, EDUCD116, representing individuals with a Doctoral degree, emerged with the most substantial coefficient magnitude in the model, recording a coefficient of -1.315595. Interpreting this coefficient, amidst constant variable conditions, indicates that possessing a doctoral degree decreases the log odds of unemployment by approximately 1.32 compared to individuals with no schooling (EDUCD2). This underscores a robust negative association between higher education levels and the likelihood of unemployment.

However, a critical issue surfaced concerning the model's predictive performance. Employing a classification threshold of 0.5 led the model to predict all observations as employed, resulting in an error rate slightly exceeding 6%. While seemingly a low error rate, this predictive behavior signifies a bias towards the majority class (employed individuals) due to the imbalanced distribution of data classes.

To mitigate this bias and enhance the identification of potential unemployment indicators, subsequent analysis will be conducted utilizing a dataset with more balanced classes.

Methodology and Model Selection

Logistic regression serves as the foundational baseline model in this study due to its widespread application in binary classification tasks and its interpretability. It provides a probabilistic framework to predict categorical outcomes, making it an ideal starting point to assess the relationship between predictors and employment status. Its linear nature enables the estimation of coefficients, aiding in the identification of significant predictors influencing employment status. The complementary modeling methods that will be used throughout the data analysis include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification trees and random forests, boosting, support vector machines (SVM), and k-nearest neighbor (kNN).

LDA and QDA were selected due to their ability to handle multi-class classification problems efficiently. LDA seeks to maximize class separation by projecting data onto a lower-dimensional space, assuming normally distributed classes with equal covariances. QDA differs as it can accommodate potentially nonlinear decision boundaries, thus extending beyond logistic regression's linear boundaries.

The decision tree approach, including ensemble methods like random forests, was chosen for its proficiency in identifying complex, non-linear relationships. Unlike logistic regression, these methods create hierarchical decision structures, allowing interactions between predictors to be captured more comprehensively. Random forests, by aggregating multiple trees, mitigate overfitting tendencies and often yield superior predictive accuracy.

Boosting algorithms were incorporated for their iterative nature, repeatedly adjusting model weights to minimize prediction errors. This iterative improvement mechanism enables boosting to uncover subtle relationships among predictors and offers enhanced predictive performance by correcting misclassifications from preceding models, supplementing logistic regression's predictive capabilities.

SVMs were chosen for their robustness in handling high-dimensional data and identifying non-linear decision boundaries. By transforming data into higher dimensions, SVMs can efficiently separate classes with complex patterns, thereby extending the modeling scope beyond logistic regression's linear boundaries.

kNN was employed to leverage local patterns within the dataset. By classifying an observation based on its nearest neighbors, kNN supplements logistic regression by capturing local structures in the data that might not be evident in linear models. Its simplicity and ability to identify non-linear relationships provide additional insights.

Identification and Assumptions

In logistic regression, identification stems from estimating coefficients that represent the relationship between predictor variables and the log odds of the binary outcome. The model assumes a linear relationship between the log-odds of the outcome variable and the predictor variables. It identifies predictors' effects on the probability of the outcome through the logistic function, enabling interpretation in terms of odds ratios. Logistic regression assumes a linear relationship between predictors and the log-odds of the outcome, observations are independent of each other, and the absence of multicollinearity among predictor variables.

LDA assumes normally distributed classes with equal covariances, while QDA relaxes the equal covariance assumption. Classification trees and random forests do not assume linearity

in predictor relationships, thereby accommodating non-linear interactions among predictors. Boosting assumes the additive nature of improving model performance iteratively by adjusting weights. SVM assumes separability by a margin or uses kernel methods to map data into higher dimensions, thereby identifying non-linear decision boundaries. kNN presumes that observations close to each other in the predictor space share similar outcomes.

Each method utilized in this study is underpinned by distinct assumptions and mechanisms for identification. Logistic regression and the alternative methods encompass varying assumptions, enabling diverse approaches to capture relationships between predictors and employment status. Acknowledging these assumptions is crucial for interpreting the results and understanding the strengths and limitations of each method in predicting employment status.

Table 2. Basic Logistic Regression Model

Variable	Coefficient	Std. Error	Prob.	Variable	Coefficient	Std. Error	Prob.
(Intercept)	-2.681706	0.082427	< 2e-16	HISPAN1	-0.327367	0.036957	< 2e-16
FAMSIZE	-0.004622	0.007220	0.522049	HISPAN2	-0.171884	0.135869	0.205847
NCHILD	-0.013536	0.011015	0.219131	HISPAN3	-0.371576	0.062718	3.13e-09
RELATE2	0.197788	0.023748	< 2e-16	HISPAN4	-0.100875	0.036127	0.005234
RELATE3	0.690329	0.037395	< 2e-16	CITIZEN2	0.059526	0.036457	0.102517
RELATE4	0.475984	0.089944	1.21e-07	CITIZEN3	0.018920	0.038570	0.623759
RELATE5	0.489225	0.048605	< 2e-16	YRSUSA22	-0.043473	0.034288	0.204842
RELATE6	0.342881	0.121882	0.004905	YRSUSA23	-0.170809	0.036017	2.11e-06
RELATE7	0.362422	0.055389	6.02e-11	YRSUSA24	-0.169781	0.034738	1.02e-06
RELATE8	0.336732	0.097936	0.000585	YRSUSA25	-0.137146	0.032219	2.08e-05
RELATE9	0.832120	0.141064	3.66e-09	SPEAKENG3	-0.351717	0.043525	6.43e-16
RELATE10	0.310178	0.064853	1.73e-06	SPEAKENG4	-0.274279	0.040045	7.43e-12
RELATE11	0.080635	0.038434	0.035905	SPEAKENG5	-0.206966	0.039605	1.73e-07
RELATE12	0.263797	0.048538	5.48e-08	SPEAKENG6	-0.029576	0.038962	0.447798
SEXFemale	0.213020	0.017162	< 2e-16	SCHOOL2	0.197507	0.030714	1.27e-10
AGE	0.005797	0.000877	3.85e-11	EDUCD3	-0.306188	0.246325	0.213858
MARST2	0.123460	0.045128	0.006223	EDUCD4	-0.233217	0.070420	0.000927
MARST3	0.221203	0.052145	2.21e-05	EDUCD5	-0.212343	0.052466	5.18e-05
MARST4	0.241069	0.032720	1.74e-13	EDUCD6	-0.029335	0.049657	0.554681
MARST5	0.098248	0.059510	0.098752	EDUCD7	-0.090168	0.041769	0.030873
MARST6	0.253849	0.029905	< 2e-16	EDUCD61	0.043329	0.052732	0.411260
RACE2	0.168286	0.032758	2.79e-07	EDUCD63	-0.070332	0.039450	0.074615
RACE3	-0.020801	0.078691	0.791522	EDUCD64	0.017240	0.057460	0.764158

RACE4	0.040513	0.036130	0.262156	EDUCD81	-0.212054	0.047522	8.11e-06
RACE5	-0.213218	0.112884	0.058916	EDUCD101	-0.397793	0.041782	< 2e-16
RACE6	-0.140154	0.027021	2.14e-07	EDUCD114	-0.651229	0.047418	< 2e-16
RACE7	-0.044641	0.036504	0.221366	EDUCD115	-1.000229	0.078691	< 2e-16
RACE8	0.026699	0.033643	0.427438	EDUCD116	-1.315595	0.084236	< 2e-16
RACE9	0.210735	0.085686	0.013918	VETSTATVeteran	-0.005583	0.063604	0.930052

Note: logistic regression model including all observations and variables

VI. Results

In pursuit of identifying the most influential variables affecting employment status, diverse modeling methods were employed. Leveraging a random sampling approach, 70% of our dataset was allocated as a training set, while the remaining 30% constituted the test or validation set. Addressing the imbalance in class observations within our data, we applied the Random Over Sampling Examples (ROSE) package in R, employing sampling techniques and a smoothed bootstrap approach. This methodology facilitated the generation of artificial data within the training set, mirroring its original size while mitigating the disparity in class distribution. Notably, the original training set exhibited over 93% “employed” observations, whereas the augmented training set achieved a more balanced representation, with approximately 50% “employed” observations. By training the models on data with improved class balance, we aim to enhance the discernment of variables impacting employment. Subsequently, the models constructed using the refined training set will be utilized to predict observations within the test set, allowing assessment of prediction accuracy.

Logistic Regression

In our endeavor to predict employment status using logistic regression, the final model incorporated all available predictor variables, excluding NCHILD, FAMSIZE, and VETSTAT, which showed no statistically significant impact on employment status prediction within our model. The predictor variable summary statistics are detailed in the model summary (Table 3).

Notably, all variables exhibited statistical significance, with EDUCD116 retaining the largest coefficient of approximately -1.28, akin to the basic logistic model discussed earlier. AGE, the sole numeric variable in the model, displayed the smallest coefficient at approximately 0.006. This coefficient implies that a one-year increase in age elevates the log odds of unemployment by 0.006 units. The model's performance was evaluated on the testing dataset, with the prediction outcomes summarized in the confusion matrix (Table 4). The logistic regression model accurately predicted the employment status for approximately 59.62% of the observations in the test set. Specifically, it correctly identified 59.62% of employed individuals and 59.59% of unemployed individuals. These results highlight the model's relatively balanced predictive ability concerning both employment and unemployment within the test set.

Table 3. Logistic Regression Model

Variable	Coefficient	Std. Error	Prob.	Variable	Coefficient	Std. Error	Prob.
(Intercept)	-0.0333166	0.0465389	0.474061	HISPAN2	-0.3318839	0.0834161	6.93e-05
RELATE2	0.1806104	0.0133611	< 2e-16	HISPAN3	-0.4361329	0.0361662	< 2e-16
RELATE3	0.7200161	0.0210369	< 2e-16	HISPAN4	-0.1533622	0.0213746	7.23e-13
RELATE4	0.6136927	0.0526795	< 2e-16	CITIZEN2	0.1140348	0.0213486	9.22e-08
RELATE5	0.4990924	0.0318179	< 2e-16	CITIZEN3	0.0592565	0.0225700	0.008653
RELATE6	0.0843820	0.0789816	0.285352	YRSUSA22	-0.0330105	0.0205865	0.108823
RELATE7	0.3657460	0.0330771	< 2e-16	YRSUSA23	-0.2049455	0.0214202	< 2e-16
RELATE8	0.3232148	0.0582793	2.92e-08	YRSUSA24	-0.1911892	0.0206533	< 2e-16
RELATE9	0.9192224	0.0977157	< 2e-16	YRSUSA25	-0.1528950	0.0190742	1.09e-15
RELATE10	0.2877776	0.0382020	4.96e-14	SPEAKENG3	-0.3544759	0.0262549	< 2e-16
RELATE11	0.0960018	0.0224633	1.92e-05	SPEAKENG4	-0.3190179	0.0244483	< 2e-16
RELATE12	0.2537354	0.0298342	< 2e-16	SPEAKENG5	-0.2245489	0.0242655	< 2e-16
SEXFemale	0.2512560	0.0100210	< 2e-16	SPEAKENG6	-0.0372867	0.0242166	0.123629
AGE	0.0060840	0.0004774	< 2e-16	SCHOOL2	0.2368296	0.0188050	< 2e-16
MARST2	0.1261479	0.0264718	1.89e-06	EDUCD3	-0.0471854	0.1359430	0.728518
MARST3	0.1963666	0.0313956	3.99e-10	EDUCD4	-0.2545186	0.0417544	1.09e-09
MARST4	0.2716250	0.0187346	< 2e-16	EDUCD5	-0.2179130	0.0312261	2.98e-12
MARST5	0.1841383	0.0353783	1.94e-07	EDUCD6	-0.0840359	0.0306310	0.006079
MARST6	0.2516642	0.0166544	< 2e-16	EDUCD7	-0.0694317	0.0255637	0.006607
RACE2	0.1269965	0.0199442	1.92e-10	EDUCD61	0.1001690	0.0329254	0.002348
RACE3	0.0255460	0.0469252	0.586167	EDUCD63	-0.0561654	0.0241827	0.020204
RACE4	-0.0106500	0.0213716	0.618256	EDUCD64	0.0689683	0.0352329	0.050289
RACE5	-0.2095300	0.0620970	0.000740	EDUCD81	-0.1932764	0.0286278	1.46e-11

RACE6	-0.1816197	0.0154406	< 2e-16	EDUCD101	-0.3600270	0.0250675	< 2e-16
RACE7	-0.0178540	0.0215038	0.406384	EDUCD114	-0.6069604	0.0274439	< 2e-16
RACE8	0.0706098	0.0198924	0.000386	EDUCD115	-0.8981350	0.0406269	< 2e-16
RACE9	0.0858298	0.0525352	0.102309	EDUCD116	-1.2766792	0.0426614	< 2e-16
HISPAN1	-0.3719410	0.0216706	< 2e-16				

Table 4. Logistic Regression Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	43833	1958
Predicted: Unemployed	29683	2887

Prediction Accuracy: *Employed: 59.62%* *Unemployed: 59.59%* *Total: 59.62%*

LDA and QDA

LDA and QDA models were constructed to build upon insight from the logistic regression model. Notably, similar to the logistic regression model, the optimal performance for both LDA and QDA models was observed upon excluding NCHILD, FAMSIZE, and VETSTAT from the predictive framework. The evaluation of our LDA model, showcased through the confusion matrix (Table 5), revealed that it accurately predicted the employment status for approximately 59.69% of all observations in the test set. Specifically, it achieved a correct prediction rate of 59.7% for employed individuals and 59.59% for unemployed individuals within the validation set. While both LDA and logistic regression rendered identical predictions for unemployed individuals, the LDA exhibited a slight enhancement in correctly predicting employed individuals. Conversely, our QDA model presented slightly different predictions. Its predictive accuracy improved notably to 66.94%, with an increased correct prediction rate of 68.13% for employed individuals. However, the QDA model demonstrated a reduction in predictive accuracy for unemployed individuals, declining to 48.79%. Despite slight disparities in predictive results among these modeling techniques, the most accurate predictions emerged

consistently when employing the same set of predictor variables, notably excluding NCHILD, FAMSIZE, and VETSTAT. This consistent pattern underscores the reinforcement of our previous findings regarding certain predictor variables that do not significantly contribute to predicting employment status.

Table 5. LDA Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	43889	1958
Predicted: Unemployed	29627	2887

Prediction Accuracy: *Employed: 59.70%* *Unemployed: 59.59%* *Total: 59.69%*

Table 6. QDA Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	50090	2481
Predicted: Unemployed	23426	2364

Prediction Accuracy: *Employed: 68.13%* *Unemployed: 48.79%* *Total: 66.94%*

Classification Trees

Following the construction of a classification tree to predict employment status, the resulting model comprised merely four out of the original fourteen variables. The variables included in the tree construction were AGE, EDUCD, RELATE, and SEX. The tree's structure and branch characteristics are visualized in Figure 1, delineating the five internal nodes and six terminal nodes (leaves). At each internal node, branch splits to the left indicate individuals satisfying that node's condition, while splits to the right denote individuals not meeting the node's condition. The values attributed to each leaf denote the count of observations classified as employed or unemployed from the entire training set falling within that branch path. The evaluation of the classification tree model's predictive performance on the test set is summarized

in Table 7, presenting the confusion matrix. The model achieved an overall predictive accuracy of 59.11%, precisely predicting 59.06% of employed individuals and 59.86% of unemployed individuals. Remarkably, these predictive outcomes closely align with those attained by logistic regression, LDA, and QDA models, despite utilizing only four variables as predictors. The model's simplicity, incorporating a concise set of predictors, might render it preferable, offering vital insights into the most influential predictors of employment status.

Figure 1. Classification Tree Predicting Employment Status

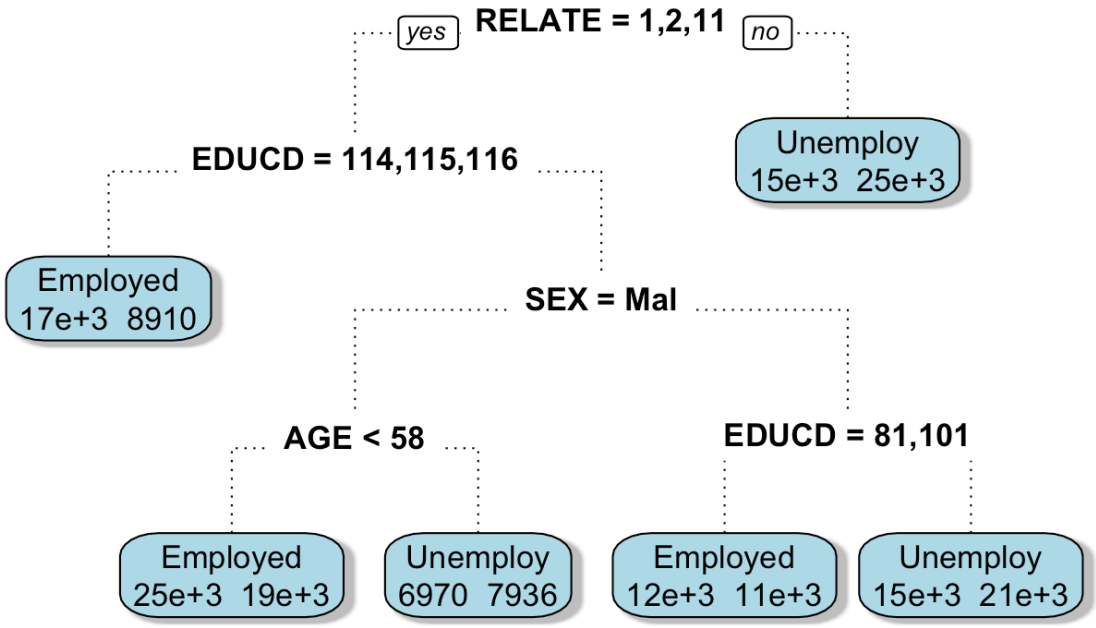

Table 7. Classification Tree Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	43417	1945
Predicted: Unemployed	30099	2900

Prediction Accuracy: Employed: 59.06% Unemployed: 59.86% Total: 59.11%

Random Forest

After training the Random Forest model and employing it to predict observations within our test set, the model revealed valuable insights through its variable importance plot (Figure 2). This plot assesses variable importance by considering both their contribution to accuracy and the degree of misclassification, as measured by MeanDecreaseGini. Notably, the variable importance plot highlights AGE, EDUCD, and RACE as the three most crucial predictors. The inclusion of these variables in the Random Forest model aligns with their selection as predictors in the classification model, reinforcing their significance in predicting employment status. However, a notable finding arises regarding FAMSIZE, which ranks as the fourth most important predictor in the Random Forest model, despite not exhibiting substantial contribution in previous modeling techniques. Conversely, VETSTAT emerges as the least influential variable, reaffirming previous findings of its minimal impact on predicting employment status. According to its confusion matrix in Table 8, the Random Forests model achieved a higher total predictive accuracy of 79.58%. It notably excelled in predicting employed individuals, with a success rate of 82.89%. However, contrasting with its performance in predicting employed individuals, the model demonstrated lower efficacy in predicting unemployment, yielding a success rate of 27.97%. This asymmetry in predictive outcomes might explain why the Random Forest model attributed increased importance to variables like FAMSIZE. While this variable might be more important in accurately classifying employed individuals, it reflects potential challenges in discerning the intricate dynamics associated with unemployment prediction.

Figure 2. Random Forest Variable Importance Plot

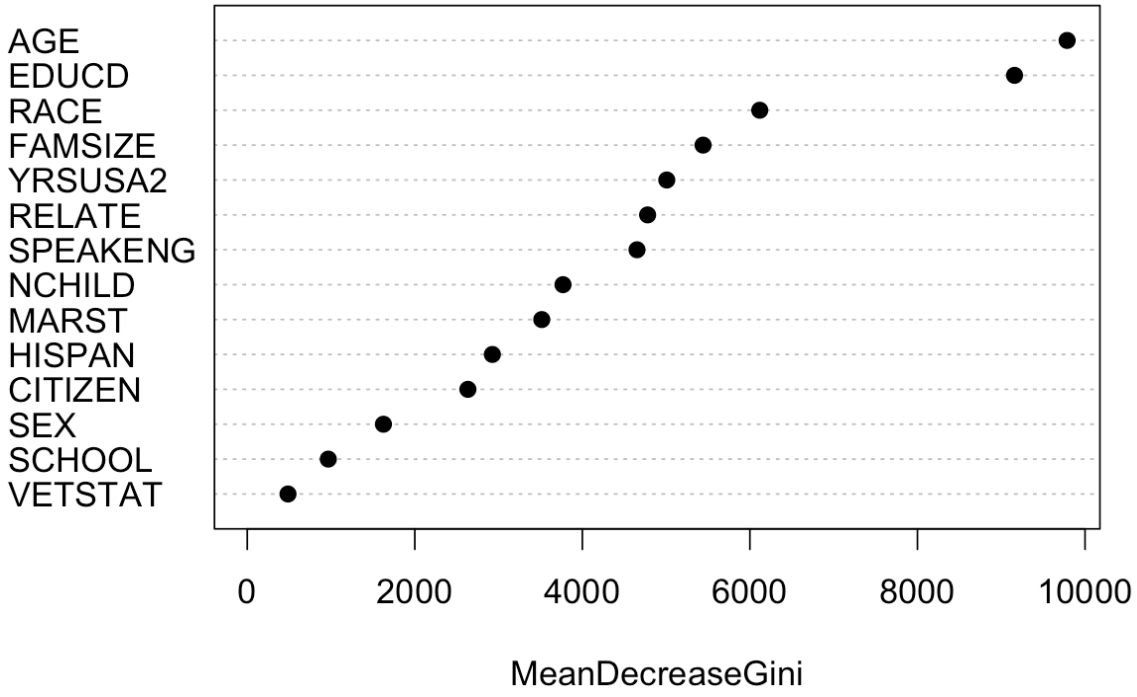

Table 8. Random Forest Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	61007	3490
Predicted: Unemployed	12509	1355

Prediction Accuracy: *Employed: 82.98%* *Unemployed: 27.97%* *Total: 79.58%*

Boosting

A Boosting model, employing 100 trees with a shrinkage value of 0.1, was constructed to predict employment status. The resulting relative influence plot (Figure 3) delineates EDUCD, RELATE, SEX, and AGE as the four most influential variables within the model, mirroring the variables selected in the classification tree model. This alignment further substantiates the assertion that these four variables stand as pivotal predictors among our variable set. Notably, VETSTAT's relative influence value of zero concurs with earlier findings, reinforcing its

insignificance in predicting employment status. The relative influence of FAMSIZE, however, undergoes a notable shift, ranking eighth in importance compared to its fourth-place ranking in the Random Forest model. This shift aligns with the more balanced prediction accuracy achieved for both employed and unemployed individuals within this model. The Boosting model's confusion matrix (Table 9) showcases improved balance in prediction accuracy for employed and unemployed individuals, accurately classifying 61.54% of employed individuals and 59.59% of unemployed individuals, culminating in a total prediction accuracy of 61.42%. This balanced predictive accuracy substantiates the refined importance attributed to FAMSIZE within the Boosting model, aligning with the model's overall performance in rendering a more equitable prediction accuracy for both employment statuses.

Figure 3. Boosting Relative Influence Plot

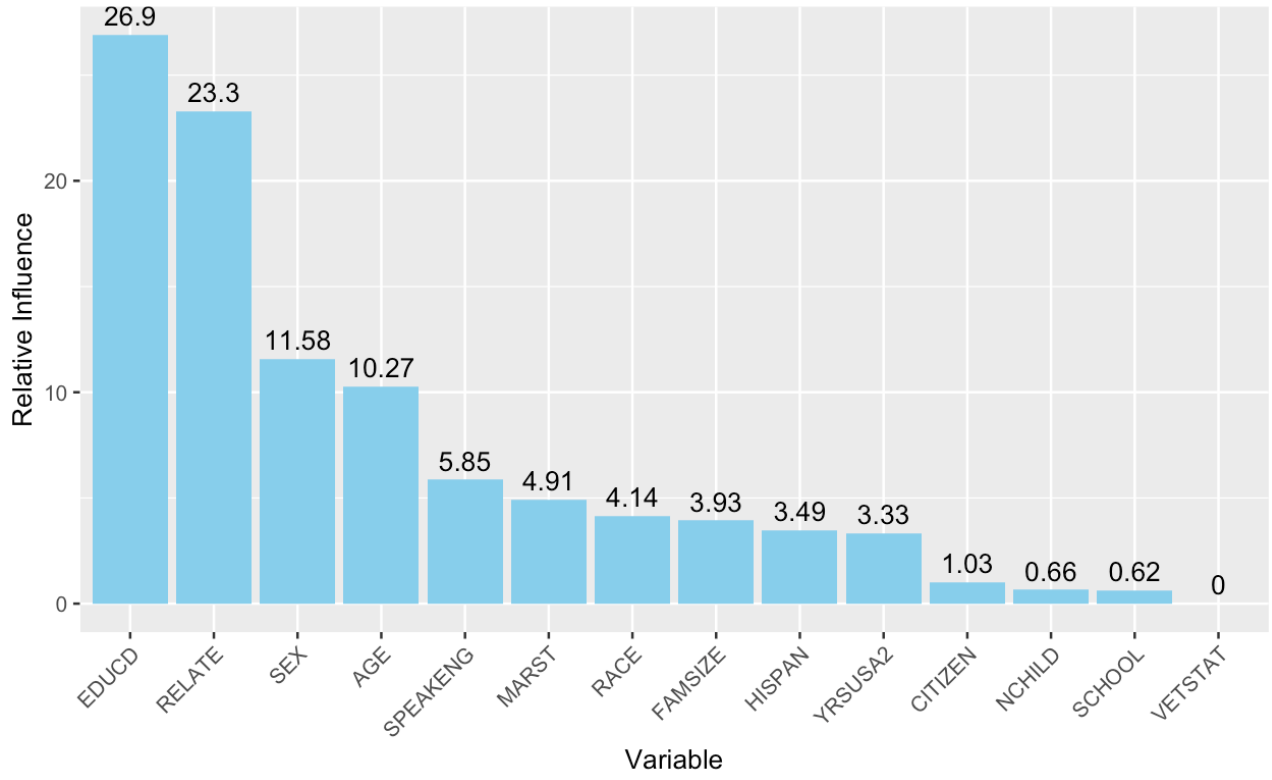

Table 9. Boosting Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	45242	1958
Predicted: Unemployed	28274	2887

Prediction Accuracy: *Employed: 61.54%* *Unemployed: 59.59%* *Total: 61.42%*

SVM

To delve deeper into the predictive potential of key variables consistently identified as influential predictors of employment status, a linear SVM was constructed using EDUCD, RELATE, SEX, and AGE variables. Due to computational constraints, the model was trained on a random 25% sample of our training observations. Subsequently, this SVM model underwent testing using our designated test set. The outcomes, detailed in the SVM confusion matrix (Table 10), revealed an overall prediction accuracy of 49.03%. Notably, the model accurately classified 47.92% of employed individuals and 65.88% of unemployed individuals. While there was a decline in our ability to predict employment status for the employed individuals, this SVM model demonstrated the highest accuracy in predicting unemployment compared to all prior models. It's important to note that the model's performance might have been impacted by training it on a smaller subset rather than the entire training dataset. Nevertheless, the inclusion of the four variables - EDUCD, RELATE, SEX, and AGE - showcased significant importance in accurately predicting the employment status, particularly for unemployed individuals.

Table 10. Support Vector Machine Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	35229	1653
Predicted: Unemployed	38287	3192

Prediction Accuracy: *Employed: 47.92%* *Unemployed: 65.88%* *Total: 49.03%*

KNN

The final model employed in predicting employment status was the k-nearest neighbor (kNN) method, trained on a randomly sampled 60% subset of our complete training set. All predictors were utilized except for VETSTAT and SCHOOL, as determined by their insignificance in both random forest and boosting analyses. The exploration of various k-values revealed that setting k to 5, representing the five nearest neighbors, yielded the optimal results. Employing this kNN model on our test set produced a prediction accuracy of 60.33%, as depicted in the kNN confusion matrix (Table 11). Notably, the model correctly identified employed individuals 61.06% of the time, while accurately predicting unemployed individuals 49.29% of the time. The lower accuracy in predicting the unemployed class may be attributed to the constraints of our smaller training sample or the possibility of an alternative k-value that might have improved predictions. Despite not achieving as advantageous results as previous methods, this model demonstrated a higher success rate in predicting employed individuals compared to other methodologies.

Table 11. K-Nearest Neighbors Confusion Matrix

	Actual: Employed	**Actual: Unemployed**
Predicted: Employed	44888	2457
Predicted: Unemployed	28628	2388

Prediction Accuracy: *Employed: 61.06%* *Unemployed: 49.29%* *Total: 60.33%*

VII. Conclusion

In conclusion, our study delved into predicting unemployment among individuals within the labor force aged in 2021, emphasizing crucial factors such as education, sex, age, and relationship status. Utilizing various models — logistic regression, discriminant analysis, classification trees, random forest, boosting, support vector machine, and k-nearest neighbors —

we unearthed insights that shed light on influential predictors. Our findings highlight that education attainment, gender, age, and relationship status emerged as the most influential predictors. This prominence is particularly evident in the classification tree and boosting models, underlining their strong predictive power. Specifically, these variables play a significant role in shaping an individual's likelihood of unemployment, providing valuable insights for both personal decision-making and organizational strategies. Additionally, our analysis revealed a noteworthy finding concerning veteran status.. Across all modeling methods, veteran status did not exhibit a significant effect on predicting unemployment. This consistent result underscores the limited impact of veteran status on unemployment status within our studied demographic. Race emerged as a crucial factor, as indicated by the Random Forest model. The model suggests that race plays a significant role in influencing unemployment rates. Notably, our logistic regression model further emphasized this point, indicating that Black individuals were more likely to be unemployed than white individuals. In navigating the complex landscape of unemployment, our study provides practical insights for individuals, organizations, and policymakers. The identified predictors offer nuanced guidance for decision-making, and the overarching conclusions guide a more focused approach to addressing unemployment challenges. Our research stands as a comprehensive contribution, emphasizing the significance of continued inquiry to decipher the multifaceted nature of unemployment and chart a course toward more resilient and equitable economic landscapes. The insights gleaned invite future researchers to explore specific aspects illuminated by our study.