KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Identifying the Discouraged Workforce: A Dual Model Analysis with Logistic Regression and Random Forest

Desyne Martinez & Edwin Trejo-Rivera
Advisors: Dr. Sudhashree Sayenju, Dr. Thomas Hodges

INTRODUCTION

- The Bureau of Labor Statistics (BLS) releases a monthly job report, including the national unemployment rate.
- Different measures of unemployment exist, one of which accounts for discouraged workers—those who have stopped looking for work due to job-market-related reasons.
- As of **September 2024**, the unemployment rate was approximately **3.9%**, which rose to **4.1%** when including discouraged workers.
- Unemployment leads to anxiety, depression, low self-esteem, marital dissatisfaction, and increased mortality risk.
- While nonprofits and government programs offer resources, identifying discouraged workers who have lost hope remains a critical challenge.
- This project will develop predictive models using **2017** data collected from almost **4,000** mental health programs of over **187,000** New York residents analyzing variables related to demographics, health, and assistance programs to identify this hidden population.
- Given the mental health issues that influence unemployment, examining this population makes them a better candidate for identifying discouraged workers.

METHODS

- ***Logistic Regression*** was used to determine whether our model could successfully classify discouraged workers from encouraged workers by including variables like (1) Race, (2) Preferred Language, (3) Household Composition, (4) Residency, and (5) Veteran's Status.
- ***Random Forest classification model*** was implemented to more effectively capture the complex relationships among the **53** variables, enabling improved classification of discouraged versus encouraged workers. Random forests achieve this by combining predictions from multiple decision trees and averaging their predictions to improve accuracy and reduce overfitting.
- ***Odds-Ratio Table*** was created to calculate the relationship between predictor variables and the likelihood of an individual being discouraged to work, including the risk ratio, 95% confidence interval (CI) of the risk ratio, and odds ratio.
- ***The ROC Curve*** illustrates each model's ability to differentiate between classes by plotting the trade-off between sensitivity and specificity. In this analysis, three curves represent the performance of each distinct model, providing a visual comparison of their classification capabilities.

DESYNE'S LINKEDIN **EDWIN'S LINKEDIN**

Anticipated Graduation: May 2025

Table 1: Odds Ratios Estimates for Variables in the Logistic Regression (Census Model)

Variables	Odds Ratio	95% Confidence Limits for Odds Ratio		Percentage Change in Odds (%)
Race: Black	0.78	0.74	0.82	-22.00%
Preferred Language: Indo-European	3.1	2.65	3.64	210.01%
Household Composition: Cohabitating with others	0.57	0.54	0.60	-42.83%
Residency: Staten Island	1.3	1.16	1.46	30.43%
Veteran's status	1.24	1.10	1.40	24.1%

Figure 1. Bar Chart of Most Important Variables by Mean Decrease Accuracy (Random Forest)

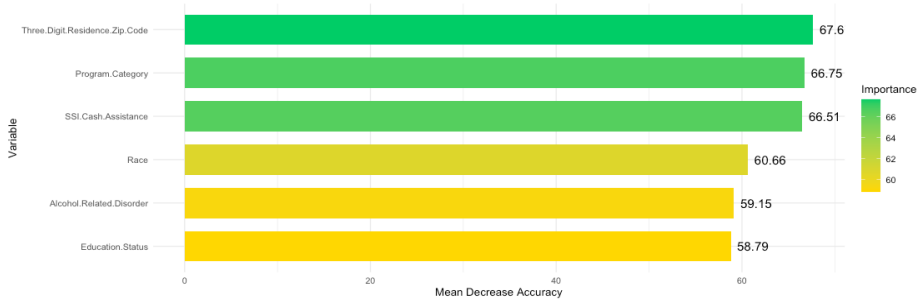


Figure 2: ROC Curve for Random Forest and Logistic Regression Models

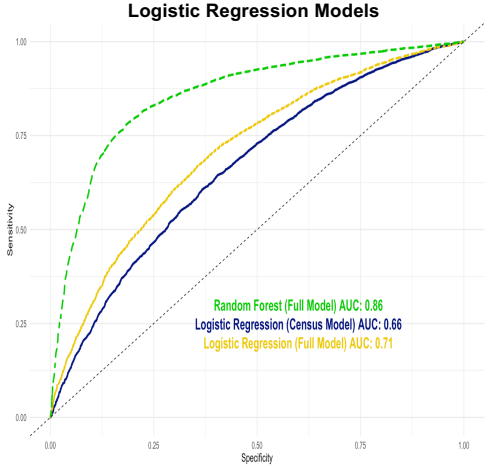
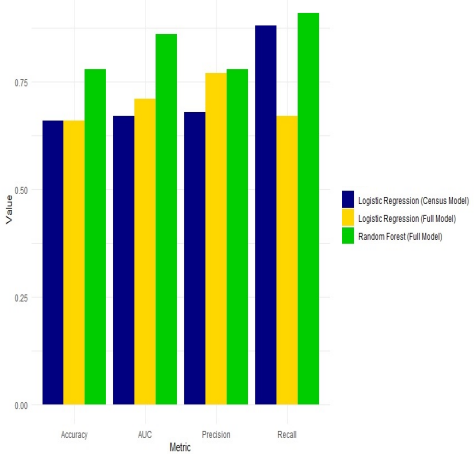


Figure 3: Clustered Bar Charts for the Performance Metrics of the Models



RESULTS

- ***LOGISTIC REGRESSION (FULL MODEL):*** The results highlighted several key predictors and yielded an AUC of 0.71 as displayed in **Figure 2**. However, due to complex relationships among variables, we applied a Random Forest model to capture these nuances more effectively.
- ***RANDOM FOREST (FULL MODEL):*** This model achieved a test classification accuracy of 78.65%, outperforming the original logistic model. Additionally, the model demonstrated a sensitivity of 79% and a specificity of 78%. **Figure 1** illustrates the six most influential variables in the Random Forest model, ranked by their ability to decrease model accuracy, highlighting their significance in predicting the target outcome.
- ***LOGISTIC REGRESSION (CENSUS MODEL):*** To enhance the practicality of our project, we streamlined the model by selecting only those variables available in the US census. While the resulting model's accuracy was lower (66%), it remained sufficiently high to effectively classify the discouraged population.
- ***ROC CURVES*: Figure 2** illustrates multiple ROC curves. For the random forest (Full Model) it shows a concordance index of 0.86, indicating that the model correctly distinguishes between a discouraged worker and an encouraged worker 86% of the time for residents of New York City.
- ***LOGISTIC BETA COEFFICIENTS*: Table 1** presents the most significant predictors of discouraged workers. By exponentiating the coefficients, the odds ratios reveal the relationships between each predictor and the likelihood of being a discouraged worker, accounting for the other variables in the model. These odds ratios were derived from the Logistic Regression (Census Model), which utilized a limited number of variables, making the results more interpretable. Below we have interpreted the odds ratio displayed in **Table 1**.

***Race*:**

- *Black :* The odds ratio of 0.78 means that Black individuals have 22% lower odds of being a discouraged worker compared to white individuals.

***Preferred Language*:**

- *Indo-European*: The odds ratio of 3.1 signifies that preferably speaking an Indo-European language in New York City have 210% of higher odds of being a discouraged worker compared to individuals who prefer to speak English.

***Household Composition*:**

- *Cohabitates With Others:* The odds ratio of .57 indicates that cohabitating with others is associated with a 42.83% lower odds of being a discouraged worker when compared to those who live alone.

***Three Digit Zip Code Residence*:**

- *Staten Island*: The odds ratio of 1.3 means that New York City residents living in Staten Island have a 30.43% of higher odds of being a discouraged worker compared to those residing in the Bronx.

***Veteran Status*:**

- The odds ratio of 1.24 signifies that a New York City Veteran resident has a 24.1% higher odds of being a discouraged worker compared to civilians.

- ***Performance Metrics For The Three Models*:** A detailed comparison of the random forest and logistic regression models, including accuracy, AUC, precision, and recall, is presented in **Figure 3**. The Random Forest model although intricate showcases the best accuracy and precision out of the three.

DISCUSSION

- The census model offers practical advantages by focusing on a limited set of easily accessible variables, increasing the likelihood of identifying discouraged workers, and provides a framework for social agencies to identify discouraged workers to implement necessary interventions.
- Recommended actions include targeted skill training, diversity programs, and counseling services to help reduce work discouragement and create a more inclusive workforce.
- The Random Forest model enhances prediction accuracy by effectively capturing complex, nonlinear relationships among a wide range of variables, providing a more robust identification of discouraged workers.