Logistic Regression Analysis of Resume

Jiechen Li

Data Overview

Sourced from OpenIntro, my dataset contains 4,870 entries with 30 variables, centered on the query: "How do race and gender affect callback rates from job applications?" The analysis evaluates the link between demographics and the "received_callback" outcome.

Data Cleaning

1. **Check for Data Structure**

The target variable, "received_callback", is binary: 0 for "no callback" and 1 for "received callback". Key predictors are "race" ("white" or "black") and "gender" ("f" for female, "m" for male), which will be binary encoded. While the analysis incorporates various predictors, over-reliance on categorical ones may introduce bias. The continuous "years_experience" is categorized as:

Entry Level	Mid Level	Senior Level
1 - 5 years	6 - 9 years	10 years +

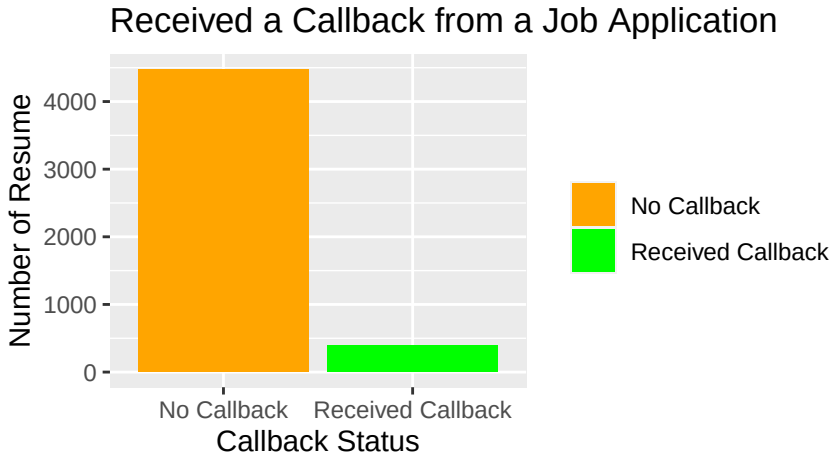
Lastly, all relevant variables will be converted to factors.

2. **Check for Missing Values**

The missing value analysis reveals 1,768 absent entries for "job_fed_contractor". While the dataset is mostly complete, addressing these gaps is vital for the accuracy of further analyses.

3. **Check for Distributions**

The data shows a marked imbalance in the outcome variable: 392 received callbacks compared to 4,478 that didn't. Such skewness can bias machine learning models, such as logistic regression, causing them to lean towards the majority class.



Based on the data cleaning and preprocessing insights, several assumptions and potential challenges emerge that warrant careful consideration:

- The severe imbalance in the "received_callback" variable may lead models to bias predictions towards the majority class (no callback).
- Many categorical predictors can result in potential multicollinearity and overfitting risks.
- The "years_experience" variable may confound the relationship between predictors and outcomes if not handled properly.

Modeling

Justification for Logistic Regression

To study the binary "received_callback" outcome, logistic regression is ideal. Suited for dichotomous outcomes like callbacks (1) or none (0), it offers interpretable odds ratios, clarifying the impact of predictors like race or gender on callback chances.

Variable Selection (a priori)

Using a priori variable selection, based on the resume dataset with "received_callback" as the outcome and race and gender as key predictors, I have selected variables potentially influencing callbacks and correlating with race or gender as follows:

- **Outcome Variable**:
  received_callback: Indicator for if there was a callback from the job posting for the person listed on this resume.

- **Predictor Variables**:
  race: Inferred race associated with the first name on the resume.
  gender: Inferred gender associated with the first name on the resume.
  years_experience: Years of experience listed on the resume.
  job_city: City where the job was located.
  volunteer: Indicator for if volunteering was listed on the resume.
  employment_holes: Indicator for if there were holes in the person's employment history.
  worked_during_school: Indicator for if the resume listed working while in school.

Logistic Regression Modeling

Model 1:

I assessed how years of experience impacted callback rates. Although the model prediction revealed an outlier, its removal didn't significantly alter the results, so we kept the original model.

Model 2:

I constructed a model using seven predictor variables. "years_experience" has been categorized into three levels and renamed "experience_category", while the other six variables are binary.

Summary output table

1). **Race** stands out prominently: White applicants are about 56% more likely to receive a callback than black applicants, a finding that's statistically significant ($p < 0.001$) and supported by the 95% CI (1.26, 1.93).

2). **Location** is crucial: Boston-based applications have significantly higher odds of a callback than those in Chicago, with the difference being statistically significant ($p < 0.001$) and the 95% CI ranging from 0.53 to 0.82.

3). **Experience** carries weight: Senior level applicants have a distinct advantage, with a 51% greater likelihood of callbacks compared to entry level, validated by a p-value of 0.004 and a 95% CI (1.14, 2.00).

Surprisingly, candidates with employment gaps have a statistically significant ($p < 0.001$) 81% increased chance of callbacks, a finding that's both unexpected and supported by the 95% CI (1.42, 2.32).

Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
Race			
Black	—	—	
White	1.56	1.26, 1.93	**<0.001**
Gender			
Female	—	—	
Male	0.93	0.70, 1.21	0.6
Job Location			
Boston	—	—	
Chicago	0.66	0.53, 0.82	**<0.001**
Working Experience			
Entry Level	—	—	
Mid Level	0.98	0.76, 1.27	0.9

Senior Level	1.51	1.14, 2.00	**0.004**
Volunteer History			
No	—	—	
Yes	1.09	0.87, 1.36	0.4
Employment Gaps			
No	—	—	
Yes	1.81	1.42, 2.32	**<0.001**
Worked During School			
No	—	—	
Yes	1.12	0.88, 1.43	0.4

[1]OR = Odds Ratio, CI = Confidence Interval

Predict Probability Upon evaluating the model's predictions, I observed that the chances of receiving a callback are evenly distributed.

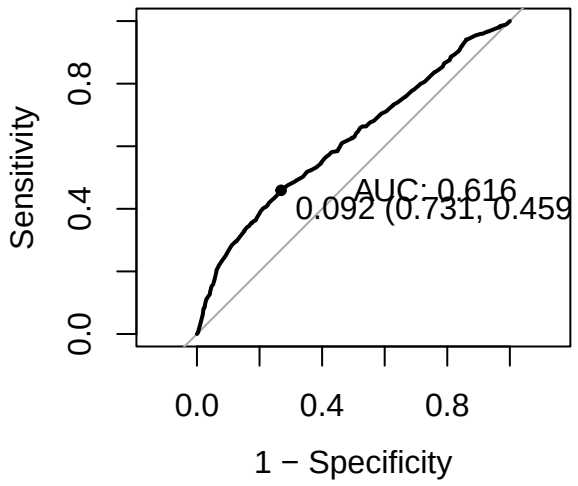
Cook's Distance Through Cook's distance, I pinpointed three potential outliers. After excluding them in "model_3", p-values revealed no notable enhancement over Model 2. Given their non-influence, Model 2 will be retained.

VIF The "experience_category" variable, having three levels, may introduce multicollinearity, especially with sparse observations in some levels. Hence, I assessed multicollinearity using VIF on Model 2.
1). **race**: An adjusted GVIF of 1.000479 indicates negligible multicollinearity, so race's effect on callbacks is clear from other predictors' influence.

2). **gender**: An adjusted GVIF of 1.069440 suggests minor multicollinearity, warranting slight caution when interpreting gender's influence.

Regarding the other variables: their adjusted GVIF values are slightly above 1, pointing to mild multicollinearity.

ROC Curve I will generate an ROC curve for Model 2 and identify the optimal threshold for the confusion matrix. The ROC suggests Model 2 can distinguish between classes, with its effectiveness gauged by the AUC and the balance between sensitivity and specificity.



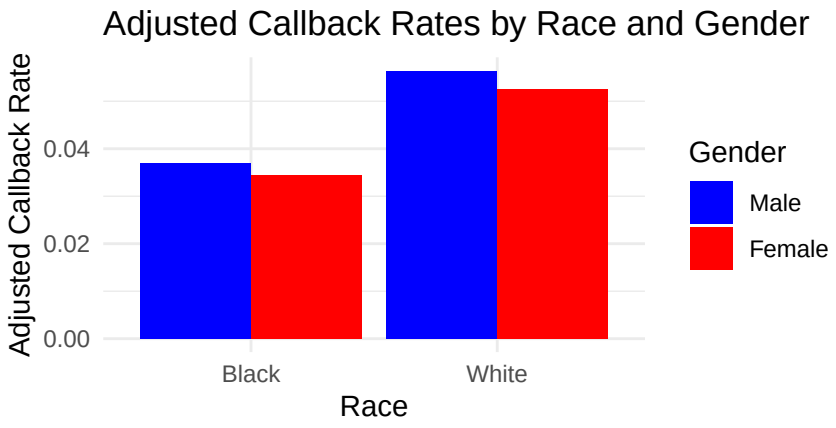
Confusion Matrix With a threshold set at 0.092 to address dataset imbalance and emphasize recall, the model achieves about 72% accuracy. Though specificity is decent, its low sensitivity and precision suggest a conservative stance in predicting "yes" for callbacks, resulting in potential false negatives.

Results

The bar chart visualizes the predicted callback rates for job applicants, adjusted for various factors including job city, experience category, volunteering history, employment gaps, and whether the applicant worked during school. The rates are segregated based on race and gender, offering insights into potential disparities.

Race & Gender Dynamics: White males have the highest callback rates, followed by white females, black males, and black females.

Despite adjustments, racial disparities in callback rates remain pronounced, with gender adding another layer of complexity.



Future Work

By emphasizing both the strengths of the current work and the need to address its limitations, it sets a constructive direction for future research efforts.

Strengths:

- Holistic inclusion of diverse variables provides a comprehensive view of callback factors.
- Rigorous statistical techniques validate the findings, with adjustments made for potential confounders.

Limitations:

- Data imbalance poses a potential bias risk.
- Possible multicollinearity and unobserved confounders might affect the results.
- The findings' generalizability beyond the specific dataset is uncertain.