Analytics Assignment 2 - Resume Data

Revanth Chowdary Ganga (rg361)

1. Overview

The aim of this analysis is to study **how race and gender influence job application callback rates.**

The Resume Dataset from OpenIntro library in R is used for this analysis. The outcome variable is binary in nature (received or didn't receive callback) and hence a **Logistic Regression Model** has been used to make an **Inference** on the factors (esp. Race and Gender) which may impact the callback rate.

This document walks-through the following steps used for the project:

1. Data: Overview, Cleaning and Processing
2. Modelling: Model Selection, Variable Selection, Model Output, Model Assessment
3. Results
4. Future Work

2. Data Overview and Cleaning

2.1 Data Overview

The Resume Dataset was created by sending artificially generated resumes to different employers in Chicago and Boston during 2001 and 2002 and checking if the resumes got picked for further steps.

The Dataset has **4,870** Observations of **30** variables, The variables can be broadly classified into the following categories (with selected examples):

- employer related: location, industry, contractor info, equal opportunity employer status
- job requirement related: educational qualifications or if computer skills etc. are required for the job
- Applicant Information: Gender, Race, educational qualifications, prior work experience
- resume: overall classification of the resumes quality (low / high)
- Callback: a binary `outcome variable` which represents if the resume got picked for further steps

Out of the 4,870 observations, only **392** observations (~8% of total) have a positive outcome (resume selected) so the data is unbalanced. However, since we are doing Inference no balancing procedures have been done as it is not required.

2.2 Data Cleaning and Processing

Before modelling, the Data was cleaned and processed to prepare it for modelling.

The following general cleaning steps were performed:

1. Checking for NAs: out of all the columns, only the `federal contractor status` column had NAs
2. Converting to suitable types: wherever applicable, columns were converted to the relevant type (mostly Factor) e.g. job type, city, columns with requirement information for skills etc.
3. Combining levels: in some columns, some specific values had very few observations, in these cases they were combined with other suitable categories to reduce standard errors e.g. experience required
4. Imputation: some columns which had blanks (intentional, not NAs) for few observations were imputed with suitable assumed values e.g. experience required

The following assumptions or decisions were made during the cleaning process:

1. federal contractor status column was excluded for further analysis due to the very high % of NAs (1,768 | ~36%)
2. for the job experience required column, "some" and "0.5" were combined with 1 and wherever blanks were present it was assumed that the jobs had no requirements so they were imputed with 0
3. Derived Variables were created (start with "satisfy_") to have boolean values to represent if the applicant met certain conditions, this helps in removing the influence of outlier points and multicollinearity. e.g. minimum job experience, computer skills

No Observations were dropped from the dataset during the cleaning process.

3. Modelling

3.1 Model Selection

For this project, Logistic Regression method was used to infer the factors affecting the callback outcomes for the following reasons:

1. Binary outcome variable makes logistic regression ideal for this analysis
2. It is interpretable and can be used to study how gender and race affect the chances of receiving a callback
3. The model is robust (to noise and outliers) and simplistic (easy and fast implementation and processing).

3.2 Variable Selection

Post the Data Cleaning and Processing, the variables were analysed and some of the variables were not included in the modelling process for the following resons:

1. Identifier Variables: These aren't relevant to the research question and would cause over-fitting if used e.g. Job Ad Id, First Name
2. Missing Values: variables which had very high missing values or NAs which couldn't be assumed/imputed e.g. Job ownership, Fed contractor
3. Non Matching: fields which couldn't be quantified due to no matching field in applicant information e.g. requirement fields for communication, organization
4. Redundant: Once the Derived Variables were created (Explained in Sec 2.2), the source variables are no longer required e.g. computer skills, job experience
5. Other Fields: These fields were dropped base on priori variable selection e.g. volunteering and worked during school

While there was potential confounding possibilities with job type, role, experience required, computer skills since we are using the derived variables this issue is eliminated along with multicollinearity

The Final Model uses the following Input variables:

1. Employer and Job Related Variables: Location (City), Industry, Role, Equal Opportunity Employer, Education Requirement
2. Applicant Related Variables: Race, Gender, Honors, Special Skills, Employment Holes, college degree
3. Derived Variables: Job Experience, Computer Skills
4. Other: Resume Quality

3.3 Model Outputs

The Generated Model has the following parameters:

Logistic Regression Results

	Dependent variable:
	Callback Received
Job City: Chicago	-0.441*** (-0.712, -0.171)
Job Industry: Finance/Insurance/Real Estate	-0.134 (-0.646, 0.377)
Job Industry: Manufacturing	-0.309 (-0.885, 0.266)
Job Industry: Other Service	0.119 (-0.200, 0.438)
Job Industry: Transportation/Communication	0.636** (0.001, 1.271)
Job Industry: Wholesale and Retail Trade	-0.163 (-0.563, 0.238)
Job Role: Manager	-0.357 (-0.869, 0.156)
Job Role: Retail Sales	-0.169 (-0.672, 0.334)
Job Role: Sales Rep	-0.362 (-0.898, 0.173)
Job Role: Secretary	-0.261 (-0.642, 0.120)
Job Role: Supervisor	-0.363 (-0.945, 0.218)
Equal Opportunity Employer: True	0.274** (-0.012, 0.561)
Job Requires Education: True	-0.457** (-0.951, 0.037)
Applicant Race: White	0.451*** (0.198, 0.705)
Applicant Gender: Male	-0.007 (-0.368, 0.353)
Applicant Has College Degree: True	0.191 (-0.113, 0.495)
Applicant Has Honors: True	0.625*** (0.182, 1.068)
Applicant Has Special Skills: True	0.735*** (0.460, 1.011)
Applicant Has Employment Holes: True	0.370*** (0.079, 0.661)
Resume Quality: High	0.199* (-0.066, 0.463)
Applicant Satisfies Experience Requiremnt: True	0.548 (-1.840, 2.936)
Applicant Satisfies Computer Skills Requiremnt: True	-0.003 (-0.806, 0.800)
Constant	-3.564*** (-6.131, -0.996)
Observations	4,870
Log Likelihood	-1,290.688
Akaike Inf. Crit.	2,627.377
Note:	*p<0.1; **p<0.05; ***p<0.01

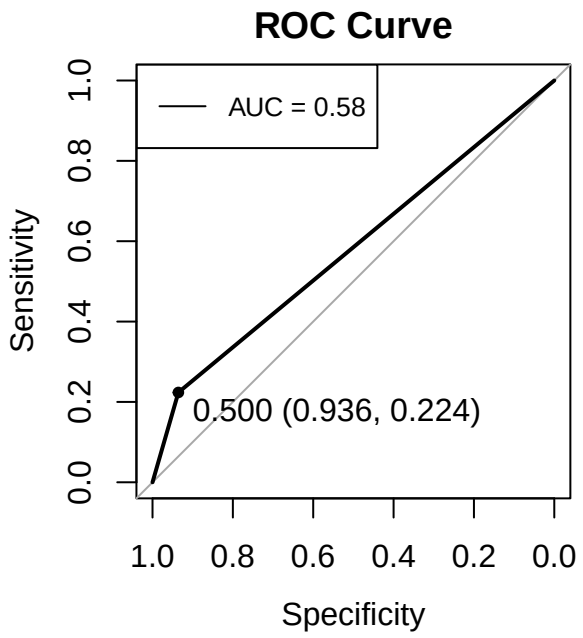
3.4 Model Assessment

The following metrics were used for evaluating the model:

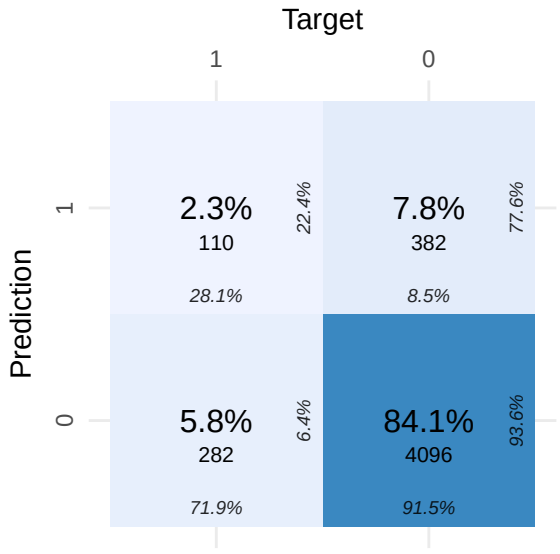
1. Kappa Score - It calculates the degree of agreement between the model's predictions and the actual class labels while accounting for the possibility of chance agreement. Kappa Score of the model is ~**0.18**
2. Sensitivity Score - measures the proportion of actual positives which are correctly identified as such, this is a good metric to use since our positive case (callback received) has very few observations in the dataset. The sensitivity of the model is **0.28**
3. ROC-AUC measures the model's ability to discriminate between the positive and negative classes. The ROC-AUC for model is **0.58**

Note: Accuracy is not a good metric to use for this analysis since there is a very high imbalance in our dataset, but for reference purposes, the accuracy of the model is **86.37%**

The ROC-AUC plot of the model is as follows:



The Confusion Matrix for the model is as follows:

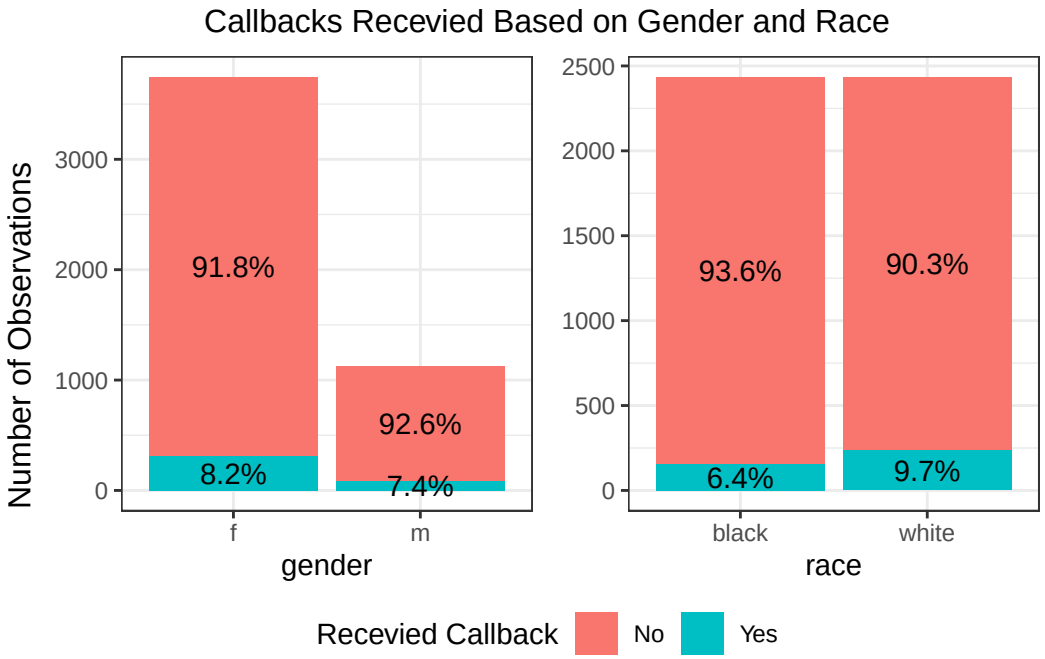


4. Results

The coefficient for the the gender variable at level male is -0.07 This means that keeping all other variables constant, if an applicant is a male, he has a lower chance of getting a call back than a female applicant.

The coefficient for the race variable at level white is 0.45 This means that keeping all other variables constant, if an applicant is a white person, they have a higher chance of getting a call back than a black applicant.

These Inferences can be confirmed by analyzing the below graphs, The % of women getting a callback is higher than that of men and similarly white people get more callbacks as compared to a black person.



5. Future Work

While the model was able to Infer the influence of race and gender on receiving a callback, the model can be improved by having more balanced data. some fields in the dataset did not have a matching field for applicant such as communication skills, organizational skills etc. which if available could also be used in the model.

The Data is also old (2001-2002) so the results from it may not be be as applicable today since there are more variables such as mode of application, if the resume was scanned using any AI etc. So having more recent data with new fields can make the model better and more applicable.