Data Analysis Assignment 02 - Resume Data

Influence of Gender and Race on job application callbacks

Divya Sharma (ds655)

Model Documentation

Data Source and Dictionary: OpenIntro

This experiment data comes from a study that sought to understand the influence of race and gender on job application callback rates. The study monitored job postings in Boston and Chicago for several months during 2001 and 2002 and used this to build up a set of test cases. Over this time period, the researchers randomly generating resumes to go out to a job posting, such as years of experience and education details, to create a realistic-looking resume. They then randomly assigned a name to the resume that would communicate the applicant's gender and race.

Research Question: How do race and gender influence job application callback rates?

1. Overview

The resume data in the OpenIntro Library is a dataset of Resumes that were used to apply for job profiles, and whether or not they recieved a callback. The resume dataset contains the following fields -

- *Job Details* - These include details such as City, Industry, Job Title, Private/Non Profit, required education, and required skills
- *Applicant Details* - Details about the applicant, such as Gender, Race, years of education, college degree, skills, and years of experience
- *Resume Details* - Details about the resume, such Email available, Resume Quality
- *Callback* - whether the applicant received a call back for this job posting for their resume (1 or 0) - this will be the *dependent variable*

The source data contains 4,870 rows and 30 columns. Out of 4,870 job-resume combinations, 392 received a callback. The dataset will be used to train a logistic regression model to predict the probability of receiving an interview invite, given the gender and socioeconomic class of the applicant.

2. Data Cleaning and EDA

2.1: Missing Values

- The `job_req_min_experience` column contains 2,746 NULL values - this is 56% missing, however, only 156 of these postings have a requirement for education. We can assume that if they are missing this field then they are entry level jobs that do not require experience
- The `job_fed_contractor` column has 1,768 (36%) NAs.
- The `job_ownership` column has 1.992 unknowns

2.1: Data Cleaning

For variables that are stored as numeric 0 and 1 but are actually flags (computer_skills, job_req_any etc) - converting them to factors before feeding this to the model. The variables include -

- `gender` - Gender (male or Female)
- `resume_quality` - Resume Quality (high or low)
- `race` - Race (black or white)
- `job_equal_opp_employer` - Whether the employer is an equal opportunity employer (0 or 1)
- `job_fed_contractor` - Whether employer is a federal contractor (0 or 1)
- `job_req_any` - Whether job has any requirements (0 or 1)
- `job_req_communication` - Whether job requires communication skills (0 or 1)
- `job_req_education` - Whether job requires education (0 or 1)
- `job_req_computer` - Whether job requires computer skills (0 or 1)
- `job_req_organization` - Whether job requires organization skills (0 or 1)
- `honors` - Whether applicant has honors (0 or 1)
- `worked_during_school` - Whether applicant worked during school (0 or 1)
- `computer_skills` - Whether applicant has computer skills (0 or 1)
- `special_skills` - Whether applicant has special skills (0 or 1)
- `volunteer` - Whether applicant is a volunteer (0 or 1)
- `military` - Whether applicant was in the military (0 or 1)
- `employment_holes` - Whether applicant has any gaps in employment (0 or 1)
- `has_email_address` - Whether resume has an email address (0 or 1)

2.2: Missing Values

- `job_req_min_experience` - this column has values like 'some' and blanks. The 'some' have been replaced by 0.5 (minimum experience), and the blanks have been replaced by 0. 56% of the data (2,746 rows) have blanks, and 21% of the data (1,064 rows) has the value 'some'
- `job_fed_contractor` - This column has 0s, 1s and NAs. 1,768 values are NAs which accounts for 36% of the data. The NAs have been replaced by 0s as we can assume majority of the jobs are not federal contractors

2.3: Exploratory Data Analysis (EDA)

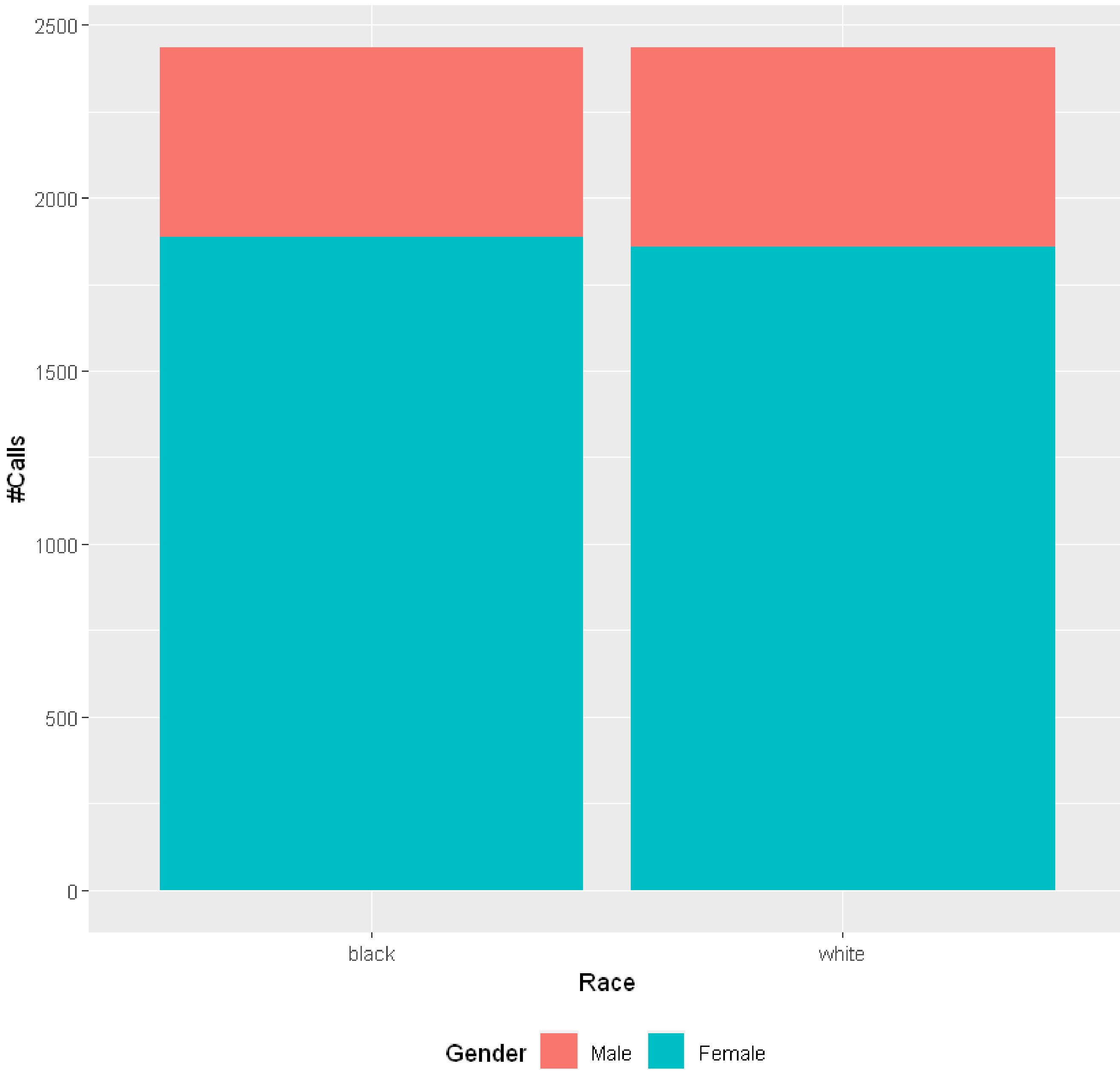
EDA is done to examine the correlations between the predictor variables such as gender, race, resume details etc and the outcome variable which is received callback.

We observe that in gender, females received higher callbacks compared to males(309 vs 83), however there were a lot more applications by females as compared to males (3746 vs 1124). Overall, females received callbacks `8.25%` times which is higher than males (`7.38%`)

While looking at race, we observed that there were equal applications for black and white people, however, black people got callbacks `6.45%` times which is much much lower than that of white people (`9.65%`) times.

```
# Plotting the bar chart
ggplot(resume, aes(x = race, fill = gender_factors)) +
  geom_bar() +
  labs(title = "Fig 2.1: Relationship between Gender and Received Callback", x = "Race", y = "#Calls")
  theme(legend.position = "bottom")+
  scale_fill_discrete(name = "Gender", labels = c("Male", "Female"))
```

Fig 2.1: Relationship between Gender and Received Callback



3. Modeling

We will be using this data to predict whether or not a callback was received, based on the provided data of job details, applicant details, and resume quality. This is an inference problem, so we are more interested in what variables are significant towards receiving a callback, rather than the accuracy of the model.

One major issue that we can face in this model is that of class imbalance, as only 392 out of 4,870 (~8%) job-resume combinations got a callback

Currently, Logistic Regression is a good choice for this problem due to a variety of reasons -

- Logistic Regression is a powerful tool for modeling the probability of a binary outcome
- It can be used to account for the effects of multiple independent variables on the outcome variable
- It is easier to interpret and explain to stakeholders

```
Logistic Regression Results - Coefficient + Confidence Intervals @ 97.5% for variables
================================================================
                                                  Dependent variable:
```

	Callback Received
Job City: Chicago	-0.400*** (-0.684, -0.115)
Job Industry: Finance/Insurance/Real Estate	-0.179 (-0.687, 0.330)
Job Industry: Manufacturing	-0.377 (-0.946, 0.192)
Job Industry: Other Service	0.071 (-0.245, 0.387)
Job Industry: Transportation/Communication	0.609** (-0.023, 1.241)
Job Industry: Wholesale and Retail Trade	-0.102 (-0.498, 0.295)
Job Type: Manager	-0.548** (-1.060, -0.037)
Job Type: Retail Sales	-0.400* (-0.901, 0.102)
Job Type: Sales Rep	-0.635*** (-1.162, -0.108)
Job Type: Secretary	-0.275* (-0.651, 0.102)
Job Type: Supervisor	-0.441* (-1.013, 0.132)
Gender: F	0.003 (-0.359, 0.365)
Race: White	0.442*** (0.190, 0.694)
Has Honors: True	0.655*** (0.211, 1.099)
Has Years of experience: True	0.023** (-0.001, 0.047)
Has Computer Skills: True	-0.212 (-0.554, 0.129)
Has Employment Holes: True	0.363*** (0.087, 0.639)
Constant	-2.351*** (-3.055, -1.647)
Observations	4,870
Log Likelihood	-1,312.902
Akaike Inf. Crit.	2,661.803
Note:	*p<0.1; **p<0.05; ***p<0.01

4. Results

Now that we have built a logistic regression model, we can assess the performance using the following metrics -

4.1 Assessing Model Performance - APR Metrics

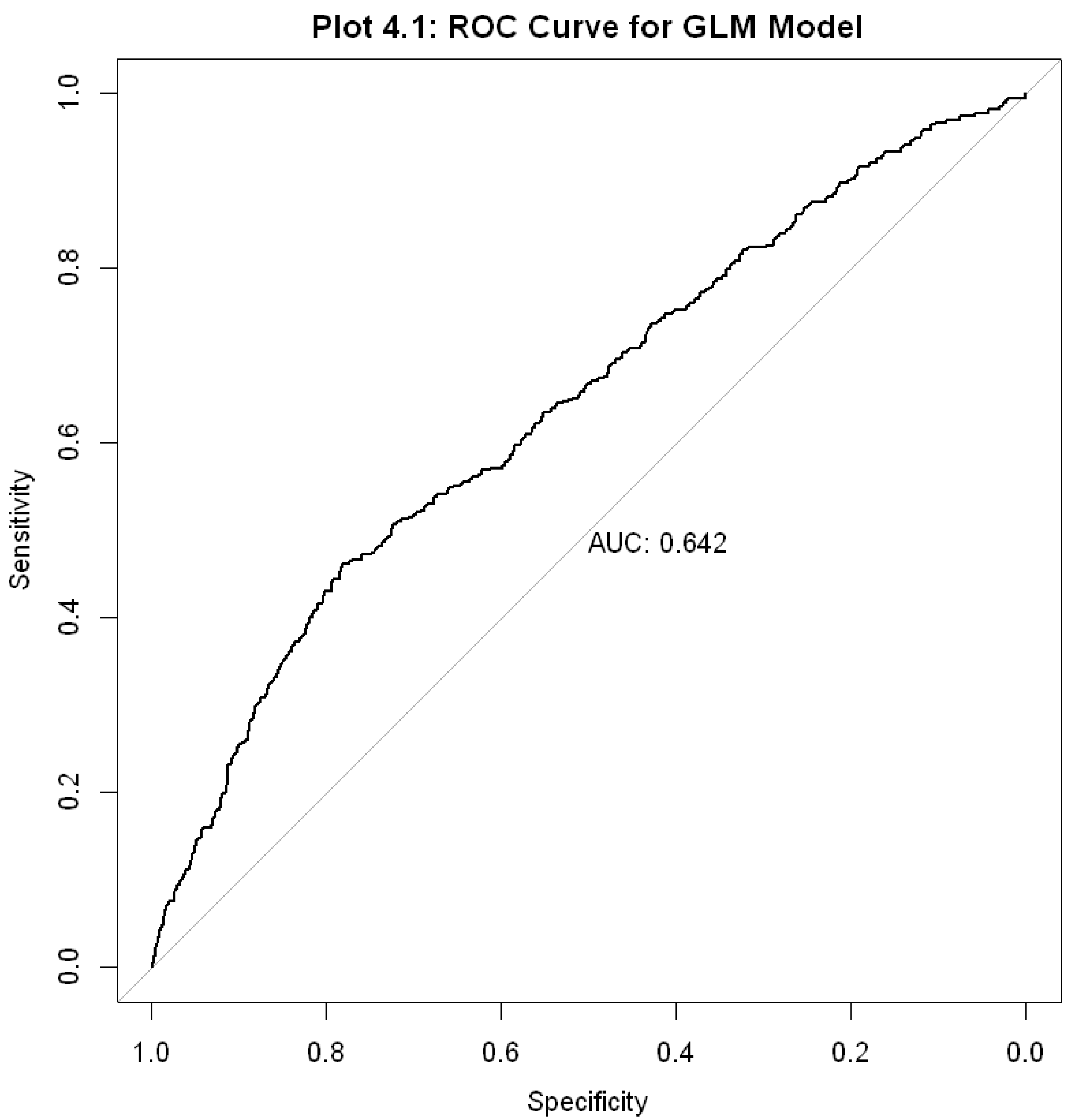
	Metric	Value
1	Sensitivity	0.939
2	Specificity	0.161
3	Pos Pred Value	0.927
4	Neg Pred Value	0.186
5	Precision	0.927
6	Recall	0.939
7	F1	0.933
8	Prevalence	0.920
9	Detection Rate	0.863
10	Detection Prevalence	0.931
11	Balanced Accuracy	0.550
12	Accuracy	0.876
13	Kappa	0.106
14	AccuracyLower	0.866
15	AccuracyUpper	0.885
16	AccuracyNull	0.920
17	AccuracyPValue	1
18	McnemarPValue	0.031

- `Accuracy`: A model with an accuracy of **0.87** predicts the correct outcome **87%** of the time. *Note that Accuracy is not a good measure of model performance due to class imbalance*
- `Precision`: A precision of **0.19** predicts the positive outcome correctly **19%** of the time when it predicts a positive outcome.
- `Recall`: A model with a recall of **0.2** correctly identifies **20%** of the positive cases.
- `Kappa`: A model with a kappa of **0.12** has a fair agreement between the predicted and actual outcomes, after accounting for the possibility of agreement occurring by chance.

4.2 Assessing Model Performance - ROC Curve

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```



An ROC of > 0.5 means that the model is better at predicting than chance. An ROC of 0.658 indicates that the model is able to predict the probability of a callback with reasonable accuracy.

5. Future Work

While the model can infer the most significant factors that resulted in recieving a callback, moving forward we can fix the class imbalance issue by using sampling methods. Once there is a better ratio of callbacks to non-callback applications, we can feed that data to the model.

This will lead to a better model that can predict whether a job-resume combination will get a callback or not.