DATA 606 Data Project Proposal - Jawaid Hakim

Jawaid Hakim

2022-10-30

Contents

1 Load library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(predictrace)
```

1.0.1 Data Preparation

```
df <- read_csv(file = "https://raw.githubusercontent.com/himalayahall/CUNY-DATA606/master/Project/resume
               col_types = cols(

                    job_ad_id = col_character(),
                    firstname = col_character(),

                    job_city = col_factor(),
                    job_industry = col_factor(),
                    job_type = col_factor(),
                    job_ownership = col_factor(),
                    job_req_school = col_factor(),
                    resume_quality = col_factor(),
                    race = col_factor(),
                    gender = col_factor(),
                    received_callback = col_factor(),

                    years_experience = col_integer(),
                    years_college = col_integer(),

                    .default = col_logical(),

               ))

df$received_callback <- recode_factor(df$received_callback, `0` = "No",  `1` = "Yes")

# I was curious to see how easily one can guess sex from the first name. We make use the predictrace
# library to guess gender from first name alone.
# Looks like predictrace guessed all genders correctly. It suggests that, based on the first name on
# the resume, employers are likely to guess the gender consistently.
gender_guess <- predict_gender(df$firstname)
gender_guess <- gender_guess %>%
                    mutate(likely_gender = ifelse(likely_gender == 'male', 'm', 'f'))
df['gender_guessed'] = as.factor(gender_guess$likely_gender)

# Our central hypothesis is that perceived race of applicants might influence the callback rate.
# Again, I would like to check whether employers would be able to consistently guess the same race
# based on first name.

race_guess <- predict_race(df$firstname)
df['race_guessed'] <- race_guess$likely_race

# We will use predictrace once more. In this case, it fails to predict the race for 11 unique
# first names. All 11 names appear to be *African American* names, which is an interesting
# observation on possible bias in the `predictrace' library!

print (unique(df$firstname[is.na(df$race_guessed)]))

# For our purposes, we can categorize all 11 black race. So it seems plausible that employers
# would be able to make a consistent guess regarding race based on first name.
df$race_guessed <- ifelse(is.na(df$race_guessed), 'black', df$race_guessed)
```

1.0.2 Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

1. All other factors being equal does the perceived **race** of applicants have a meaningful impact on the callback rate?

1.0.3 Cases

What are the cases, and how many are there?

Each case represents a randomly generated resumes There are 4870 cases.

1.0.4 Data collection

Researchers randomly generated realistic resumes to send to job postings in Boston and Chicago. They then randomly assigned a *first name* to the resume that would communicate the race and gender of the application.
First names for the study were selected so that the names would predominantly be recognized as belonging to black or white individuals.

The dataset is available on OpenIntro.

1.0.5 Type of study

What type of study is this (observational/experiment)?

This was an **experiment** conducted over several months during 2001 and 2002 in Boston and Chicago.

1.0.6 Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

OpenIntro.

1.0.7 Response Variable

What is the response variable, and what type is it (numerical/categorical)?

The response variable is the **received_callback**, it is categorical

1.0.8 Explanatory Variable(s)

The explanatory variable is **race** of the applicant inferred by the first name.

1.0.9 Relevant summary statistics

Provide summary statistics for each of the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```r
table(df$race)
```

```
##
## white black
##  2435  2435
```

```r
table(df$received_callback)
```

```
##
##   No  Yes
## 4478  392
```

```r
tbl <- table(df$race, df$received_callback)
tbl
```

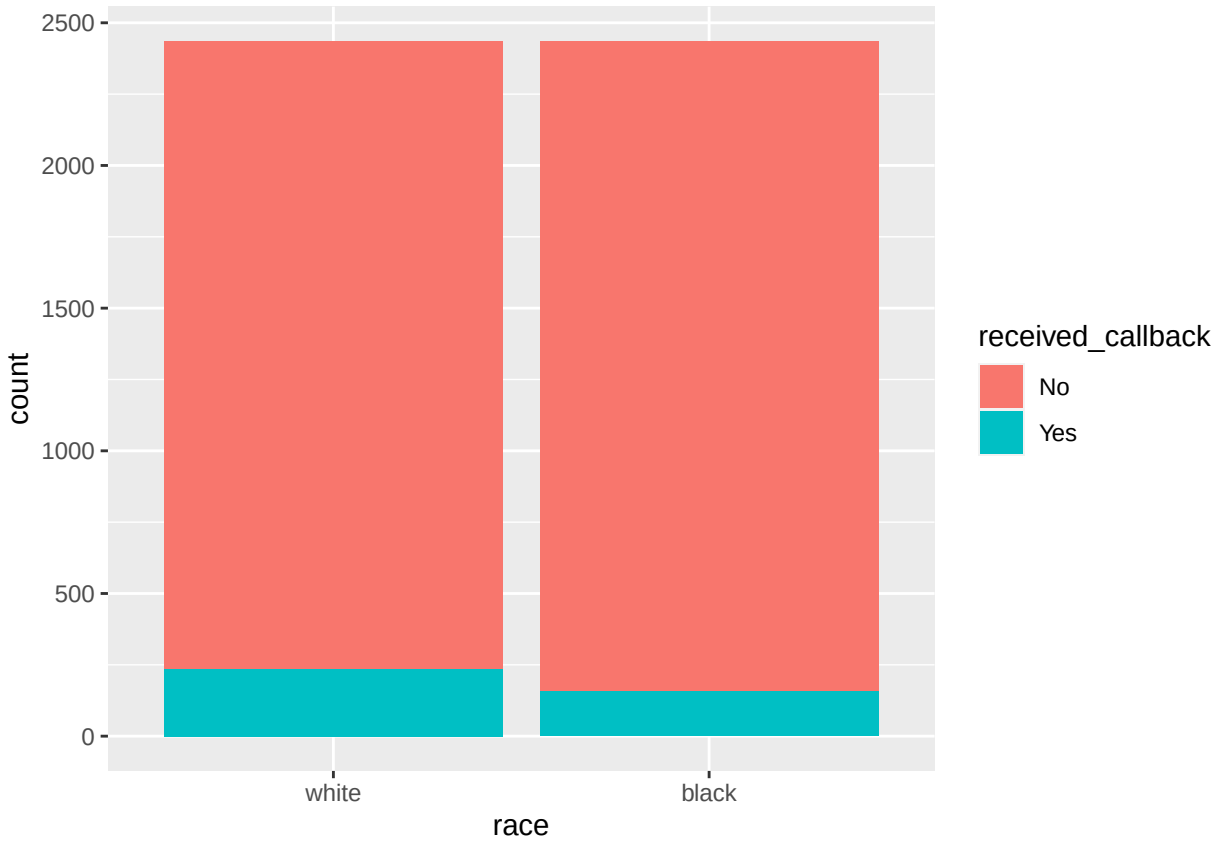
```
##
##           No  Yes
##   white 2200  235
##   black 2278  157
```

```r
prop.table(tbl, 2)
```

```
##
##                No       Yes
##   white 0.4912908 0.5994898
##   black 0.5087092 0.4005102
```

```r
#ggplot(data=df, aes(x=race, y=received_callback, fill=race)) + geom_bar(stat="identity")
plt <- ggplot(df,
       aes(x = race,
           fill = received_callback)) +
  geom_bar(position = "stack")

plt
```


2500
2000
1500
1000
500
0
count
white
black
race
received_callback
No
Yes