Portfolio CIS 631

Michael Bordeaux

7/26/2023

this is my final project portfolio for CIS 631

-Describe probability as a foundation of statistical modeling, including inference and maximum likelihood estimation

#Course Objective:

#Determine and apply the appropriate generalized linear model for a specific data context

for this learning objective activity 6 is demonstrating using the appropriate glm for a specfic data context.

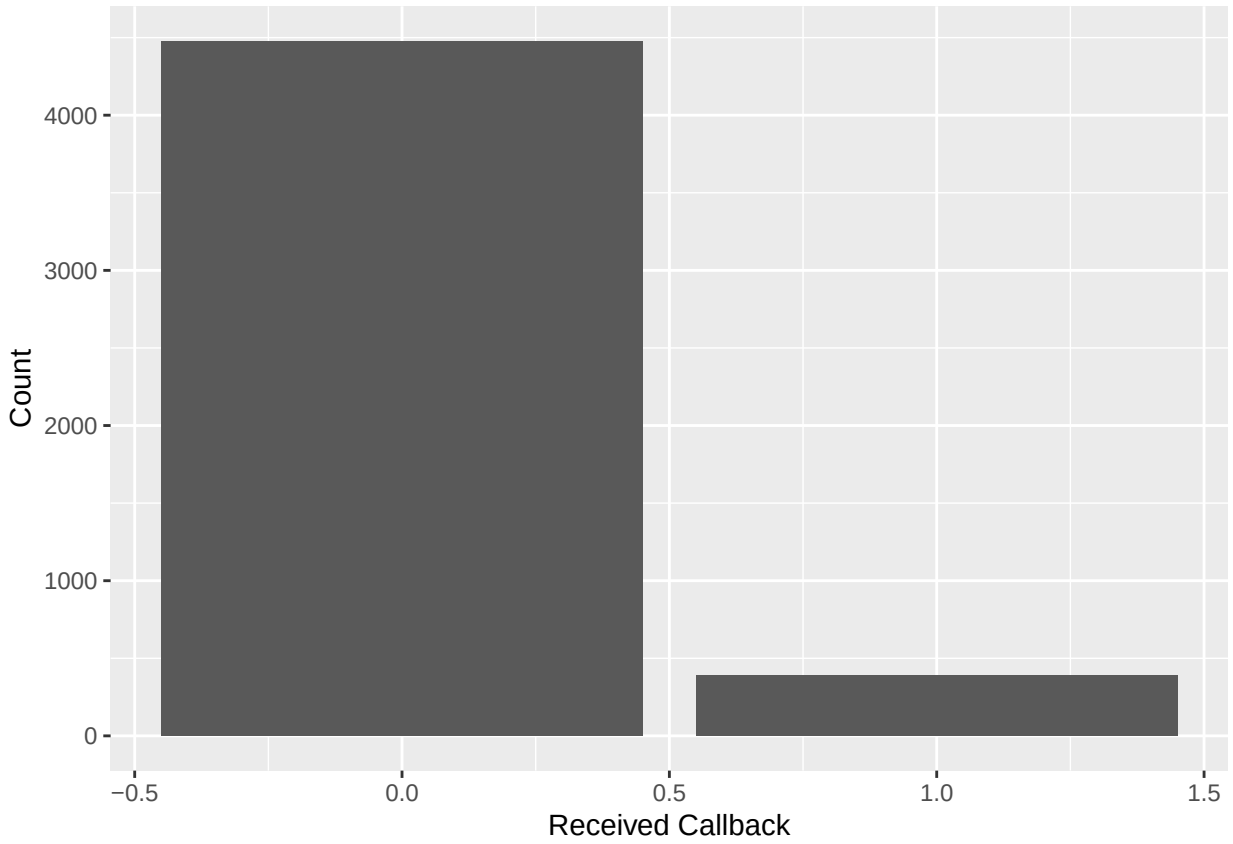
```
url <- "https://www.openintro.org/data/csv/resume.csv"

resume <- read_csv(url)
```

```
## Rows: 4870 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): job_city, job_industry, job_type, job_ownership, job_req_min_exper...
## dbl (20): job_ad_id, job_fed_contractor, job_equal_opp_employer, job_req_any...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

exploratory analysis on received_callback variable:

```
ggplot(resume, aes(x = received_callback)) +
  geom_bar() +
  labs(x = "Received Callback", y = "Count")
```



By looking at the above graph we can see that a majority of these resumes did not receive callbacks.

```
resume$received_callback <- factor(resume$received_callback, labels = c("No", "Yes"))

table_data <- table(resume$received_callback)
total <- sum(table_data)
percent <- prop.table(table_data) * 100

table_df <- data.frame(
  received_callback = levels(resume$received_callback),
  n = table_data,
  percent = percent
)

print(table_df)
```

```
##   received_callback n.Var1 n.Freq percent.Var1 percent.Freq
## 1                No     No   4478           No    91.950719
## 2               Yes    Yes    392          Yes     8.049281
```

looking at the table above our probablity of a "Yes" is only 8% with an odds .08/(1-.08) of roughly 8% also.

we can further explore this data by adding race into it:

Calculating the probability of a randomly selected person percieved as black it would be ~6% and the odds of a randomly selected resume of a person percieved as black being called back is .06/(1-.06) roughly also 6%

```
# The {tidymodels} method for logistic regression requires that the response be a factor variable
resume <- resume %>%
  mutate(received_callback = as.factor(received_callback))
```

```
resume_mod <- logistic_reg() %>%
  set_engine("glm") %>%
  fit(received_callback ~ race, data = resume, family = "binomial")

tidy(resume_mod) %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.675	0.083	-32.417	0
racewhite	0.438	0.107	4.083	0

regression equation :

y= -2.675 + .438*X + E

to simplfy this and look at the equiation for corresponding to resumes/persons perceived as black we'd right it as: y= -2.675 + E

the logg-odds would be: -2.675

and the odds they would be called back is roughly .069 or exp(-2.675)

and the porbablity is .064 of getting called back

linear, trying to fit some sort of a line for some link function explore the data and then say the data means this so thats why I chose this model

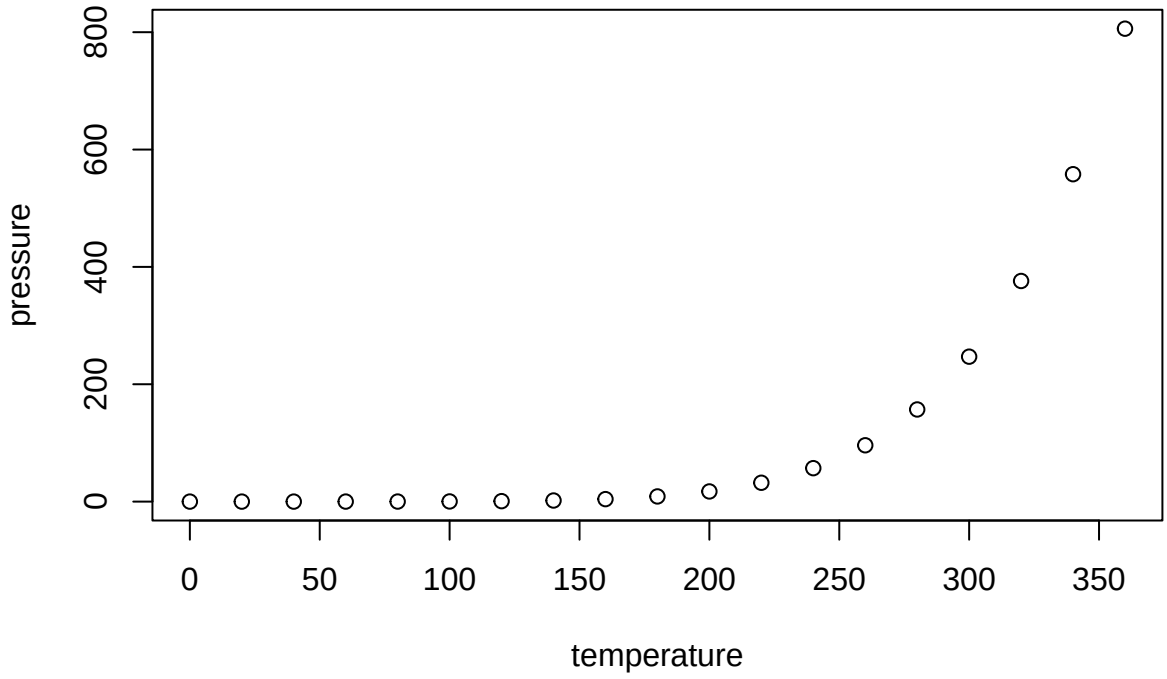
-Conduct model selection for a set of candidate models

-Communicate the results of statistical models to a general audience

-Use programming software (i.e., R) to fit and assess statistical models ** will demonstrate with this portfolio project being coded in R

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```


pressure
0
200
400
600
800
0
50
100
150
200
250
300
350
temperature