Final

Jimmy Peterson

2025-07-29

#install Rmarkedown insatall.packages(tinytex")

#if you wish to create PDFs or WOrk files from Latex install the following tinytex::install_tinytex()

Predictive Modeling of Student Employment Using R

Jimmy Peterson

GA Tech, Summer 2025

Abstract

These data were obtained from Toxic Release Inventory (TRI) data from 1987-2022 that examined the effect of pollution on on infant birth outcomes (specifically, *birth weights*) in Louisiana's infamous **Cancer Alley**. The effects of pollutants (e.g., CS_2) were analyzed per $mile^2$.
These data are a part of an on-going project with my Professor.

Table of Contents

Background

Use this section to talk about what motivates your project. Is this furthering your learning in data science? Or are you using this data in an ongoing capacity? Also, what has been written if anything about your topic? Give 1-2 citations.

Dataet Utilized

Description of Dataset The College Student Placement Factors Dataset (the "Data") to be utilized in the exercise is comprised of data associated with 10,000 college students from 100 colleges that applied for work post-graduation. The Data includes nine independent variables related to a student's academic performance and preparedness for the workplace and whether the student successfully achieved job placement (the dependent variable).

Source of the Data The Data was sourced from Kaggle.com. Kaggle.com is an online platform known for hosting data science competitions and providing a repository of free datasets for analysis and research. The Data is licensed by MIT.edu and is described as "a realistic, large-scale synthetic database". The Data was downloaded in CSV format.

View of Data

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
##   College_ID              IQ         Prev_Sem_Result       CGPA
##  Length:10000       Min.   : 41.00   Min.   : 5.000   Min.   : 4.540
##  Class :character   1st Qu.: 89.00   1st Qu.: 6.290   1st Qu.: 6.290
##  Mode  :character   Median : 99.00   Median : 7.560   Median : 7.550
##                     Mean   : 99.47   Mean   : 7.536   Mean   : 7.532
##                     3rd Qu.:110.00   3rd Qu.: 8.790   3rd Qu.: 8.770
##                     Max.   :158.00   Max.   :10.000   Max.   :10.460
##  Academic_Performance Internship_Experience Extra_Curricular_Score
##  Min.   : 1.000       Length:10000          Min.   : 0.000
##  1st Qu.: 3.000       Class :character      1st Qu.: 2.000
##  Median : 6.000       Mode  :character      Median : 5.000
##  Mean   : 5.546                             Mean   : 4.971
##  3rd Qu.: 8.000                             3rd Qu.: 8.000
##  Max.   :10.000                             Max.   :10.000
##  Communication_Skills Projects_Completed  Placement
##  Min.   : 1.000       Min.   :0.000      Length:10000
##  1st Qu.: 3.000       1st Qu.:1.000      Class :character
##  Median : 6.000       Median :3.000      Mode  :character
##  Mean   : 5.562       Mean   :2.513
##  3rd Qu.: 8.000       3rd Qu.:4.000
##  Max.   :10.000       Max.   :5.000
```

```r
# Remove College_IDs from CLG0050 to CLG0100
data <- data %>%  filter(!(College_ID >= "CLG0050" & College_ID <= "CLG0100"))

#remove Prev_Semester_Results, Academic_Performance, Extra_Curricular_Score, Communication_Skills, Proj

data <- data %>%
  select(-Prev_Sem_Result, -Academic_Performance, -Extra_Curricular_Score, -Communication_Skills, -Proj

#change column and row names

data <- data %>%
  rename(
    College = College_ID,
    Internship = Internship_Experience,
    GPA = CGPA
      )
# Remove College from CLG0050 to CLG0100
data <- data %>%  filter(!(College >= "CLG0050" & College <= "CLG0100"))

#change character objects into numbers in columns Internship and Placement

data <- data %>%
  mutate(
    Internship = case_when(
      Internship == "Yes" ~ 1,
      Internship == "No" ~ 2,
      TRUE ~ NA_real_
    ),
    Placement = case_when(
      Placement == "Yes" ~ 1,
      Placement == "No" ~ 2,
      TRUE ~ NA_real_
    )
  )

#change the College ID to simple number groupings

data$College <- sub("^CLG00", "", data$College)
summary(data)
```

Cleaning of the Data

```
##    College                IQ              GPA           Internship
##  Length:4881        Min.   : 41.00   Min.   : 4.540   Min.   :1.00
##  Class :character   1st Qu.: 90.00   1st Qu.: 6.270   1st Qu.:1.00
##  Mode  :character   Median :100.00   Median : 7.580   Median :2.00
##                     Mean   : 99.54   Mean   : 7.535   Mean   :1.61
##                     3rd Qu.:109.00   3rd Qu.: 8.790   3rd Qu.:2.00
##                     Max.   :150.00   Max.   :10.460   Max.   :2.00
##    Placement
##  Min.   :1.000
##  1st Qu.:2.000
```

```
##  Median :2.000
##  Mean   :1.832
##  3rd Qu.:2.000
##  Max.   :2.000
```

```
view(data)
```

A Multiple Regression Model

dfaddfad

Now let's take a look at the regression summary.

Next Steps

These are the Midterm goals that I was not (yet) able to accomplish and why.