Assignment 2 ABA

Cam Holecek

2025-02-23

```r
setwd("C:/Users/chole/OneDrive/Advanced BA/QMBE-3730/QMBE-3730")
admit <- read.csv("admit.csv")
```

```r
# Check the structure and summary of the data
str(admit)
```

```
## 'data.frame':    400 obs. of  4 variables:
##  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...
##  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...
##  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ rank : int  3 3 1 4 4 2 1 2 3 2 ...
```

```r
summary(admit)
```

```
##      admit             gre             gpa             rank      
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000  
##  Median :0.0000   Median :580.0   Median :3.395   Median :2.000  
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485  
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000  
```

```r
head(admit)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

```r
#Checking the balance
table(admit$admit)
```

```
## 
##   0   1 
## 273 127
```

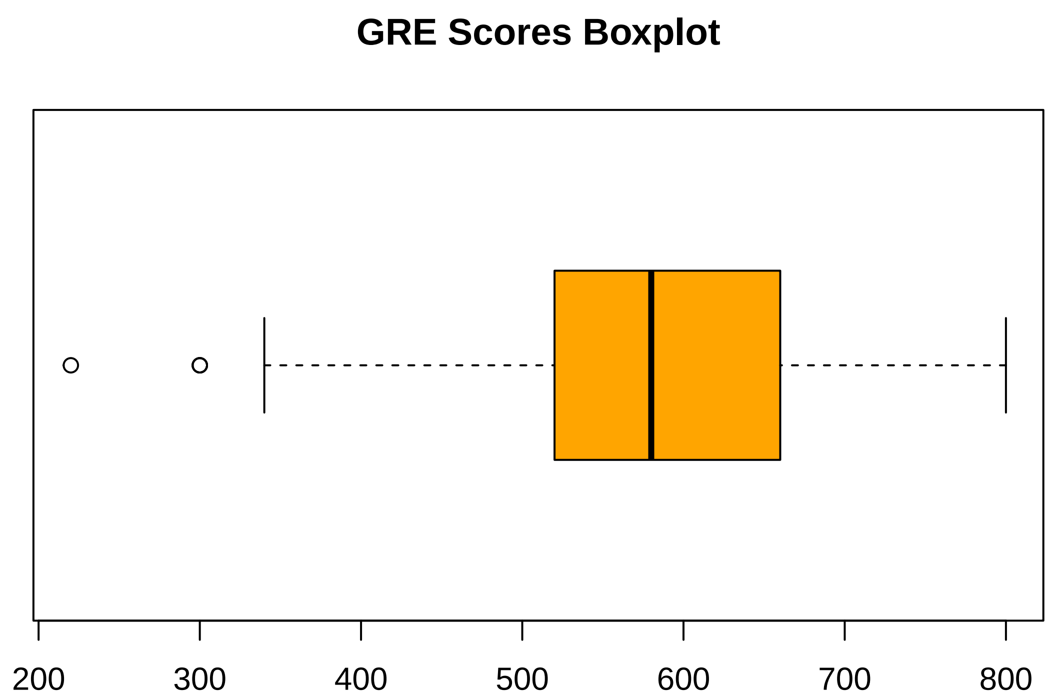
```r
prop.table(table(admit$admit))
```

```
##
##      0      1
## 0.6825 0.3175
```

The majority class is more than twice the size of the minority class, the dataset is not balanced. This imbalance might impact model performance, especially for accuracy.

```r
hist(admit$gre,
     main = "GRE Scores Distribution",
     xlab = "GRE Score",
     col = "skyblue",
     breaks = 20)
```


```r
boxplot(admit$gre,
        main = "GRE Scores Boxplot",
        horizontal = TRUE,
        col = "orange")
```



The GRE scored are skewed since the Histogram is skewed left

```r
#Testing and Training Sets
library(caTools)
set.seed(123)
split <- sample.split(admit$admit, SplitRatio = 0.7)
train_data <- subset(admit, split == TRUE)
test_data <- subset(admit, split == FALSE)
```

```r
#Fit Model
model <- glm(admit ~ gre + gpa + rank, data = train_data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.972151   1.390237  -2.138   0.0325 *
## gre          0.002824   0.001309   2.157   0.0310 *
## gpa          0.608084   0.387848   1.568   0.1169
## rank        -0.650108   0.160419  -4.053 5.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 350.14  on 279  degrees of freedom
## Residual deviance: 318.54  on 276  degrees of freedom
## AIC: 326.54
##
## Number of Fisher Scoring iterations: 4
```

The most important variable for predicting admission status is RANK