Summary of Credit Scoring Assignment

University of Connecticut FNCE 5352

Matt McDonald

April 20, 2021

Summary

Fifteen submissions were collected (including one submission by Professor McDonald). Of these, 8 were scoreable since they included a prediction for every record (37,500) in the "ConsumerCred-train.csv" file, and 7 still need remediation. This document summarizes the results of scoring these submissions, and includes an analysis of the distribution of the results.

Processing

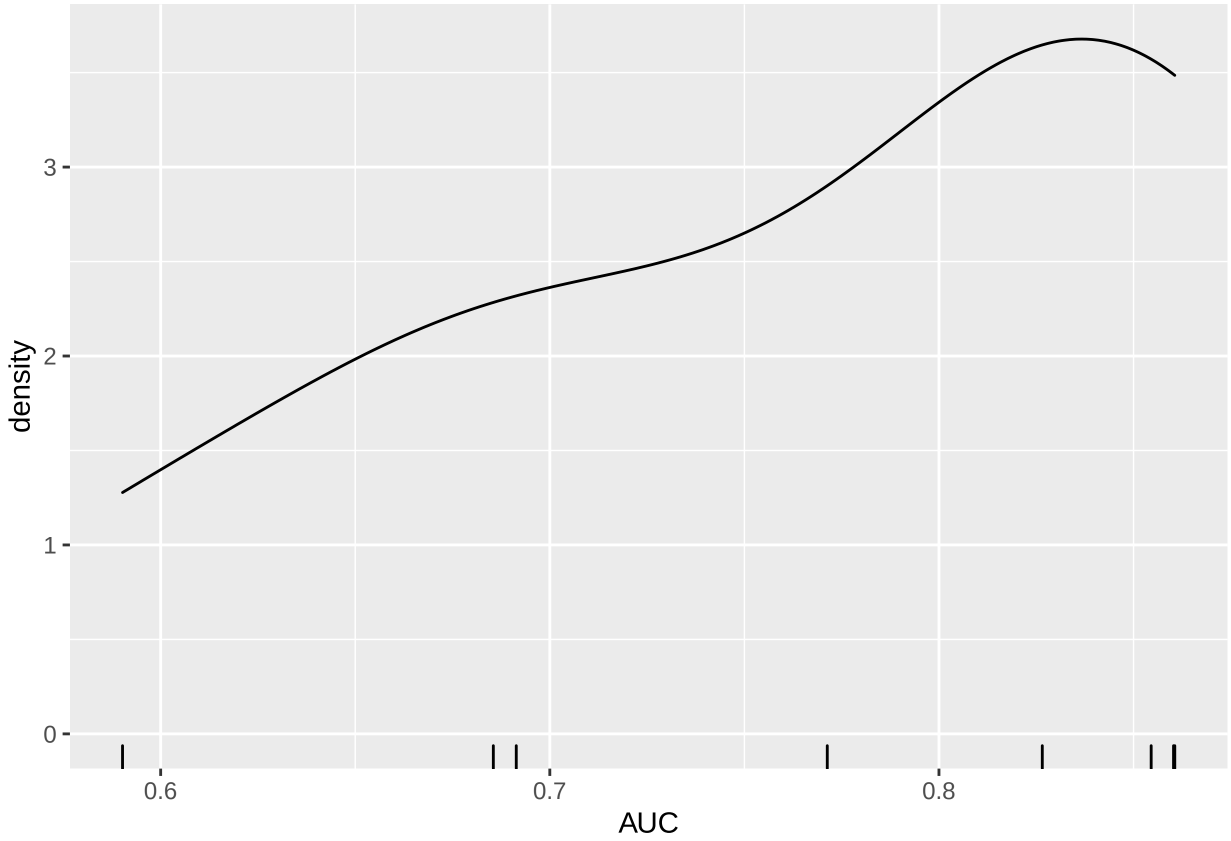
The following code evaluates the submissions against the *solution* file, which is coded with a Default indicator for each or the records in the test file.

```r
library(tidyverse)
library(pROC)

#Load in the solution file
solutionfile <- here::here('Assignments', 'ConsumerCredit', 'solution.csv')
solution <- read_csv(solutionfile)

#All submissions are found in the "grading" folder
submissionsfolder <- here::here('Assignments', 'ConsumerCredit', 'submissions')

#create the "submissions" tibble
#file is the file name
submissions <- tibble(file=dir(submissionsfolder))
#csv is a list column containing a tibble with the submission from the team
submissions <- submissions %>%
  mutate(csv=map(file, ~ read_csv(paste(submissionsfolder, .x, sep='//'))))

#7 submissions have the wrong number of rows
submissions %>% mutate(nrow=map_dbl(csv,nrow)) %>% filter(nrow != 37500) %>% select(-file)

#this function scores the submission
getAUC <- function(csv, colnum=2){
  out <- 0

  if (nrow(csv) == nrow(solution)) {
    rocobj <- roc(
      response = solution$SeriousDlqin2yrs,
      predictor = pull(csv[,colnum])
    )
    out <- auc(rocobj)
  }
  out
}

predcol=2
#use my getAUC function to score the data contained in column 'csv'
submissions <- submissions %>%
  mutate(AUC = map2_dbl(csv, predcol, getAUC))
```

Distribution of scores

The following graph show the distribution of scores. All submissions have AUC values greater than 50%, which indicates that every conforming submission had some predictive power. An AUC value of 50% indicates a completely random model.

```
ggplot(submissions %>% filter(AUC > 0), aes(x=AUC)) + geom_density() + geom_rug()
```



Notes

The data was taken from the Kaggle competition “Give Me Some Credit”, which can be found at the following link: https://www.kaggle.com/c/GiveMeSomeCredit

This Kaggle competition contains a lot of discussion about approaches that can be used to process the data and improve performance. As it stands, the best score obtained by the class is close to the winning score of 0.8695. However, our highest score would not have cracked the top 100 scores in this competition.

The code Professor McDonald used to generate his results can be found at https://github.com/mattmcd71/fnce5352_spring2021/tree/main/Assignments/ConsumerCredit under the file names *modelingexample.R*.

This summary was writted using RMarkdown, which is a useful tool in RStudio for communicating results. A helpful CheatSheet can be found at https://rmarkdown.rstudio.com/