HW 4

Sam Reid

12/29/2023

This homework is designed to give you practice fitting a logistic regression and working with statistical/philosophical measures of fairness. We will work with the `titanic` dataset which we have previously seen in class in connection to decision trees.

Below I will preprocess the data precisely as we did in class. You can simply refer to `data_train` as your training data and `data_test` as your testing data.

```
#this is all of the preprocessing done for the decision trees lecture.

path <- 'https://raw.githubusercontent.com/guru99-edu/R-Programming/master/titanic_data.csv'
titanic <-read.csv(path)
head(titanic)
```

```
##   x pclass survived                                            name    sex
## 1 1      1        1                   Allen, Miss. Elisabeth Walton female
## 2 2      1        1                  Allison, Master. Hudson Trevor   male
## 3 3      1        0                    Allison, Miss. Helen Loraine female
## 4 4      1        0            Allison, Mr. Hudson Joshua Creighton   male
## 5 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6 6      1        1                             Anderson, Mr. Harry   male
##      age sibsp parch ticket     fare   cabin embarked
## 1     29     0     0  24160 211.3375      B5        S
## 2 0.9167     1     2 113781   151.55 C22 C26        S
## 3      2     1     2 113781   151.55 C22 C26        S
## 4     30     1     2 113781   151.55 C22 C26        S
## 5     25     1     2 113781   151.55 C22 C26        S
## 6     48     0     0  19952    26.55     E12        S
##                          home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#replace ? with NA
replace_question_mark <- function(x) {
  if (is.character(x)) {
    x <- na_if(x, "?")
  }
  return(x)
}

titanic <- titanic %>%
  mutate_all(replace_question_mark)

set.seed(678)
shuffle_index <- sample(1:nrow(titanic))
head(shuffle_index)
```

```
## [1]   57  774  796 1044  681  920
```

```r
titanic <- titanic[shuffle_index, ]
head(titanic)
```

```
##         x pclass survived                                             name
## 57     57      1        1                       Carter, Mr. William Ernest
## 774   774      3        0                                 Dimic, Mr. Jovan
## 796   796      3        0                          Emir, Mr. Farred Chehab
## 1044 1044      3        1                      Murphy, Miss. Margaret Jane
## 681   681      3        0                                Boulos, Mr. Hanna
## 920   920      3        0 Katavelas, Mr. Vassilios ('Catavelas Vassilios')
##         sex  age sibsp parch ticket   fare   cabin embarked     home.dest
## 57     male   36     1     2 113760    120 B96 B98        S Bryn Mawr, PA
## 774    male   42     0     0 315088 8.6625    <NA>        S          <NA>
## 796    male <NA>     0     0   2631  7.225    <NA>        C          <NA>
## 1044 female <NA>     1     0 367230   15.5    <NA>        Q          <NA>
## 681    male <NA>     0     0   2664  7.225    <NA>        C         Syria
## 920    male 18.5     0     0   2682 7.2292    <NA>        C          <NA>
```

```r
library(dplyr)
# Drop variables
clean_titanic <- titanic %>%
select(-c(home.dest, cabin, name, x, ticket)) %>%
#Convert to factor level
    mutate(pclass = factor(pclass, levels = c(1, 2, 3), labels = c('Upper', 'Middle', 'Lower')),
    survived = factor(survived, levels = c(0, 1), labels = c('No', 'Yes'))) %>%
na.omit()
#previously were characters
clean_titanic$age <- as.numeric(clean_titanic$age)
clean_titanic$fare <- as.numeric(clean_titanic$fare)
glimpse(clean_titanic)
```

```
## Rows: 1,043
## Columns: 8
## $ pclass   <fct> Upper, Lower, Lower, Middle, Lower, Middle, Lower, Lower, Upp~
## $ survived <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No, No, Yes, N~
## $ sex      <chr> "male", "male", "male", "male", "female", "female", "male", "~
## $ age      <dbl> 36.0, 42.0, 18.5, 44.0, 19.0, 26.0, 23.0, 28.5, 64.0, 36.5, 4~
## $ sibsp    <int> 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0~
## $ parch    <int> 2, 0, 0, 0, 0, 1, 0, 0, 2, 2, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ fare     <dbl> 120.0000, 8.6625, 7.2292, 13.0000, 16.1000, 26.0000, 7.8542, ~
## $ embarked <chr> "S", "S", "C", "S", "S", "S", "S", "S", "C", "S", "S", "S", "~
```

```
create_train_test <- function(data, size = 0.8, train = TRUE) {
    n_row = nrow(data)
    total_row = size * n_row
    train_sample <- 1: total_row
    if (train == TRUE) {
        return (data[train_sample, ])
    } else {
        return (data[-train_sample, ])
    }
}
data_train <- create_train_test(clean_titanic, 0.8, train = TRUE)
data_test <- create_train_test(clean_titanic, 0.8, train = FALSE)
```

Create a table reporting the proportion of people in the training set surviving the Titanic. Do the same for the testing set. Comment on whether the current training-testing partition looks suitable.

```
paste("Proportion survived in training set:" , count(data_train, survived)[2,2]/nrow(data_train))
```

```
## [1] "Proportion survived in training set: 0.398081534772182"
```

```
paste("Proportion survived in testing set:" , count(data_test, survived)[2,2]/nrow(data_test))
```

```
## [1] "Proportion survived in testing set: 0.444976076555024"
```

While not exactly equal, I am satisfied that this split is close enough in terms of survivorhood representation.

Use the `glm` command to build a logistic regression on the training partition. `survived` should be your response variable and `pclass`, `sex`, `age`, `sibsp`, and `parch` should be your response variables.

```
logit_model <- glm(survived~pclass+sex+age+sibsp+parch, data=data_train, family=binomial(link=logit))
```

We would now like to test whether this classifier is *fair* across the sex subgroups. It was reported that women and children were prioritized on the life-boats and as a result survived the incident at a much higher rate. Let us see if our model is able to capture this fact.

Subset your test data into a male group and a female group. Then, use the `predict` function on the male testing group to come up with predicted probabilities of surviving the Titanic for each male in the testing set. Do the same for the female testing group.

```r
test_male <- data_test[data_test$sex == "male",]
test_female <- data_test[data_test$sex == "female",]

fitted_male_results <- predict(logit_model,newdata=test_male, type='response')
fitted_female_results <- predict(logit_model,newdata=test_female, type='response')
```

Now recall that for this logistic *regression* to be a true classifier, we need to pair it with a decision boundary. Use an `if-else` statement to translate any predicted probability in the male group greater than 0.5 into `Yes` (as in Yes this individual is predicted to have survived). Likewise an predicted probability less than 0.5 should be translated into a `No`.

Do this for the female testing group as well, and then create a confusion matrix for each of the male and female test set predictions. You can use the `confusionMatrix` command as seen in class to expidite this process as well as provide you necessary metrics for the following questions.

```r
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
fitted_male_results <- ifelse(fitted_male_results > 0.5,"Yes","No")
fitted_female_results <- ifelse(fitted_female_results > 0.5,"Yes","No")

conf_matrix_male <- confusionMatrix(as.factor(fitted_male_results), test_male$survived, positive = "Yes
conf_matrix_female <- confusionMatrix(as.factor(fitted_female_results), test_female$survived, positive =
```

We can see that indeed, at least within the testing groups, women did seem to survive at a higher proportion than men (24.8% to 76.3% in the testing set). Print a summary of your trained model and interpret one of the fitted coefficients in light of the above disparity.

```r
summary(logit_model)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age + sibsp + parch,
##     family = binomial(link = logit), data = data_train)
##
## Coefficients:
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.903165   0.409280   9.537  < 2e-16 ***
## pclassMiddle -1.291506   0.257421  -5.017 5.25e-07 ***
## pclassLower  -2.404084   0.262022  -9.175  < 2e-16 ***
## sexmale      -2.684206   0.200130 -13.412  < 2e-16 ***
## age          -0.036776   0.007494  -4.907 9.24e-07 ***
## sibsp        -0.395584   0.118587  -3.336  0.00085 ***
## parch         0.032494   0.111916   0.290  0.77155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1121.27  on 833  degrees of freedom
## Residual deviance:  757.87  on 827  degrees of freedom
## AIC: 771.87
##
## Number of Fisher Scoring iterations: 5
```

Student Input According to the model, being male decreases the log-odds of survival by 2.68 compared to being female.

Now let's see if our model is *fair* across this explanatory variable. Calculate five measures (as defined in class) in this question: the Overall accuracy rate ratio between females and males, the disparate impact between females and males, the statistical parity between females and males, and the predictive equality as well as equal opportunity between females and males (collectively these last two comprise equalized odds). Set a reasonable ϵ each time and then comment on which (if any) of these five criteria are met.

```
conf_matrix_male
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  93  28
##        Yes  4   4
##
##                Accuracy : 0.7519
##                  95% CI : (0.6682, 0.8237)
##     No Information Rate : 0.7519
##     P-Value [Acc > NIR] : 0.5473
##
##                   Kappa : 0.1119
##
##  Mcnemar's Test P-Value : 4.785e-05
##
##             Sensitivity : 0.12500
##             Specificity : 0.95876
##          Pos Pred Value : 0.50000
##          Neg Pred Value : 0.76860
##               Prevalence : 0.24806
##           Detection Rate : 0.03101
##     Detection Prevalence : 0.06202
##        Balanced Accuracy : 0.54188
##
##         'Positive' Class : Yes
##
```

```
conf_matrix_female
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No   4   2
##        Yes 15  59
##
##                Accuracy : 0.7875
##                  95% CI : (0.6817, 0.8711)
##     No Information Rate : 0.7625
##     P-Value [Acc > NIR] : 0.354209
##
##                   Kappa : 0.2325
##
##  Mcnemar's Test P-Value : 0.003609
##
##             Sensitivity : 0.9672
##             Specificity : 0.2105
##          Pos Pred Value : 0.7973
##          Neg Pred Value : 0.6667
##              Prevalence : 0.7625
##          Detection Rate : 0.7375
##    Detection Prevalence : 0.9250
##       Balanced Accuracy : 0.5889
##
##        'Positive' Class : Yes
##
```

```
paste("Overall accuracy ratio:", conf_matrix_male[["overall"]]["Accuracy"]/conf_matrix_female[["overall
```

```
## [1] "Overall accuracy ratio: 0.954841885074443"
```

```
paste("Disparate impact:", ((4+4)/(4+4+28+93))/((15+59)/(15+59+4+2)))
```

```
## [1] "Disparate impact: 0.0670437879740205"
```

```
paste("Statistical Parity:",(15+59)/(15+59+4+2)-(4+4)/(4+4+28+93))
```

```
## [1] "Statistical Parity: 0.862984496124031"
```

```r
paste("Predictive Equality:", (15/(15+4))-(4/(4+93)))
```

```
## [1] "Predictive Equality: 0.748236570808465"
```

```r
paste("Equal Opportunity", (59/(59+2))-(4/(4+28)) )
```

```
## [1] "Equal Opportunity 0.842213114754098"
```

Student Input.
Using an epsilon of .2, we would expect Overall accuracy and Disparate Impact to be >.8, and for the other three to be <.2. Only overall accuracy falls within these margins, with the other metrics not being close, meaning our model is pretty accurate, but not at all fair by any metric.

It is always important for us to interpret our results in light of the original data and the context of the analysis. In this case, it is relevant that we are analyzing a historical event post-facto and any disparities across demographics identified are unlikely to be replicated. So even though our model fails numerous of the statistical fairness criteria, I would argue we need not worry that our model could be misused to perpetuate discrimination in the future. After all, this model is likely not being used to prescribe a preferred method of treatment in the future.

Even so, provide a *philosophical* notion of justice or fairness that may have motivated the Titanic survivors to act as they did. Spell out what this philosophical notion or principle entails?

Student Input

The overall concept of "women and children first" is idealized, and can only easily explained by appealing to a chivalric virtue ethic, as opposed to more rational means. It cannot be consequentialist or utilitarian: in the case that overall sum of life matters most, filling the boats first-come-first-serve style should be paramount (more pessimistically, the rule is counterproductive, as in 1912 men would have a better chance of building new families and contributing to the wider society should they survive). The rule itself is deontological in that the people in charge of managing the evacuation were surely going with what their gut told them, rather than worrying about long-term consequences, however this does not tell us why they though it was right. The underlying virtue at the core of this decision was the spiritual sanctity of women and children, and the moral duty of men to protect those who can't protect themselves. A charitable interpretation would be that this is a selfless act by those with greater physical capabilities and thus more societal responsibility, laying down their lives with the knowledge that their sacrifice will save the lives of others. An uncharitable interpretation would be that this is a condescending and archaic principle that infantalizes women, and that either parent would make an equally good caregiver in the event of the death of the other. Either way, there is also the second intention that there should be *some* rule, as it minimizes arguing and time-wasting in the event of a pressing emergency, and having a rule that everyone can accept on honor is helpful no matter the underlying justification.