Anita Blege - Assignment 4(Final Project)

Anita Blege

2022-10-15

The titanic dataset describes the survival status of all passengers on the Titanic which sank in 1912 and killed 1,502 of its passengers and crew members. This dataset does not include information about crew members. The goal of modeling and performing analytics on this data is to gain insights on how important features such as sex, passenger class, and fare can affect the survival rate of passengers in similar situations and disasters especially taking into consideration several several social biases.

The csv dataset was obtained from the Stanford CS109 repository (<https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html>)

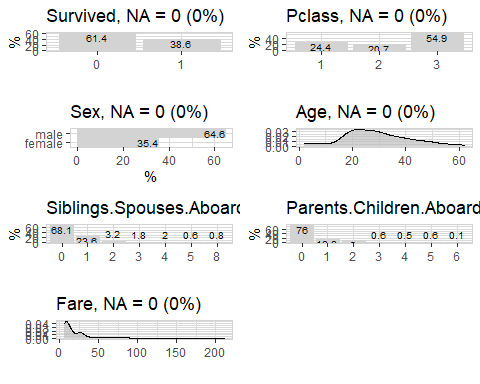
describe(titanic)

## vars n mean sd median trimmed mad min max  
## Survived 1 887 0.39 0.49 0.00 0.36 0.00 0.00 1.00  
## Pclass 2 887 2.31 0.84 3.00 2.38 0.00 1.00 3.00  
## Sex\* 3 887 1.65 0.48 2.00 1.68 0.00 1.00 2.00  
## Age 4 887 29.47 14.12 28.00 28.96 11.86 0.42 80.00  
## Siblings.Spouses.Aboard 5 887 0.53 1.10 0.00 0.27 0.00 0.00 8.00  
## Parents.Children.Aboard 6 887 0.38 0.81 0.00 0.19 0.00 0.00 6.00  
## Fare 7 887 32.31 49.78 14.45 21.50 10.32 0.00 512.33  
## range skew kurtosis se  
## Survived 1.00 0.47 -1.78 0.02  
## Pclass 2.00 -0.62 -1.29 0.03  
## Sex\* 1.00 -0.61 -1.63 0.02  
## Age 79.58 0.45 0.28 0.47  
## Siblings.Spouses.Aboard 8.00 3.67 17.64 0.04  
## Parents.Children.Aboard 6.00 2.73 9.63 0.03  
## Fare 512.33 4.76 32.99 1.67

The data has about 887 rows and 7 columns which are rcolnames(titanic), 545 passengers died while 342 survived. The table below summarizes and describes the dataset rdescribe(titanic):

During data preparation, missing values, wrong column types, and normalization will be performed to prepare the data for the modeling stage. Below is a visual exploration of the raw data to understand the general distribution of the dataset. From the exploration we note that none of the columns have missing or empty values.

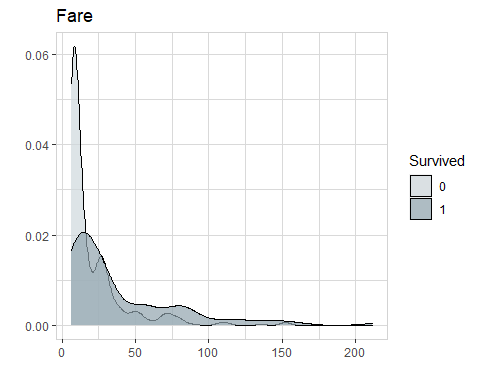
titanic %>% explore\_all



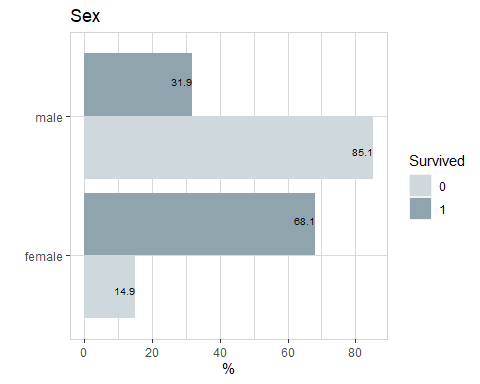
table(titanic$Survived)

##   
## 0 1   
## 545 342

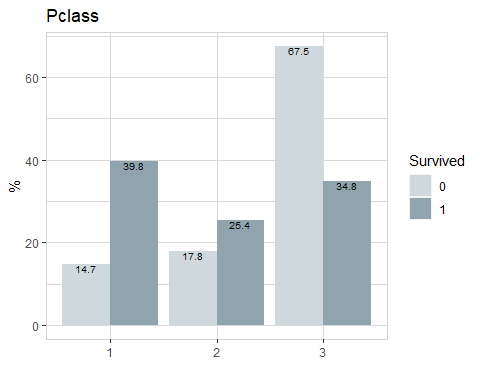
titanic %>% explore(Fare, target = Survived, split = TRUE)



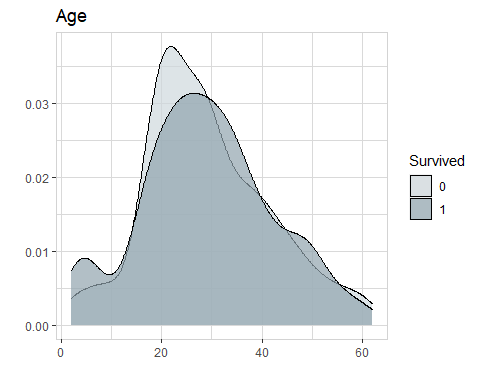
titanic %>% explore(Sex, target = Survived, split = TRUE)



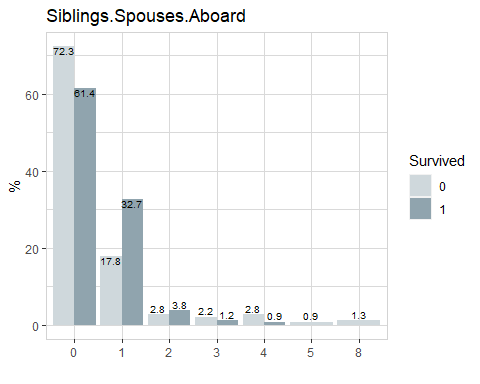
titanic %>% explore(Pclass, target = Survived, split = TRUE)



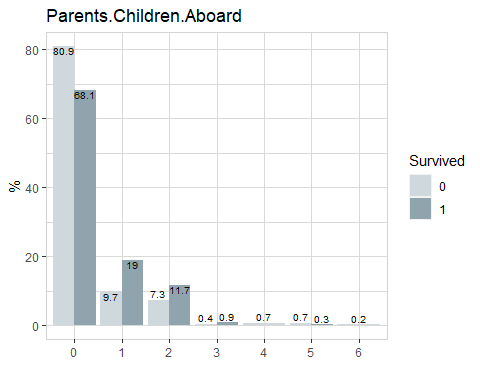
titanic %>% explore(Age, target = Survived, split = TRUE)



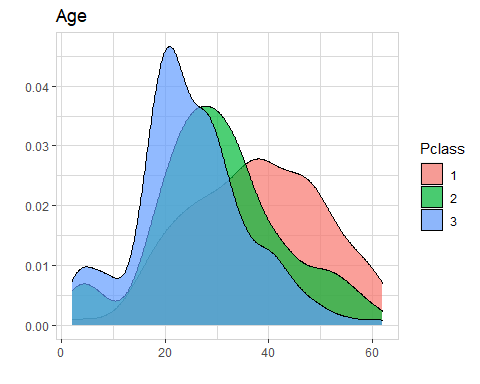
titanic %>% explore(Siblings.Spouses.Aboard, target = Survived, split = TRUE)



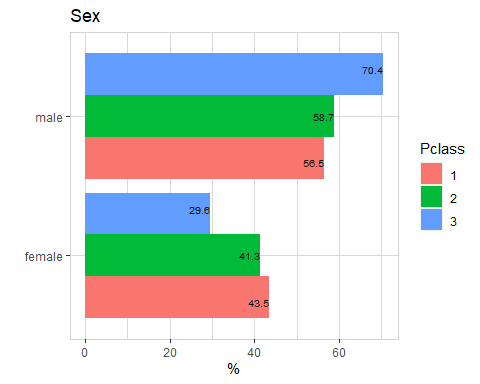
titanic %>% explore(Parents.Children.Aboard, target = Survived, split = TRUE)



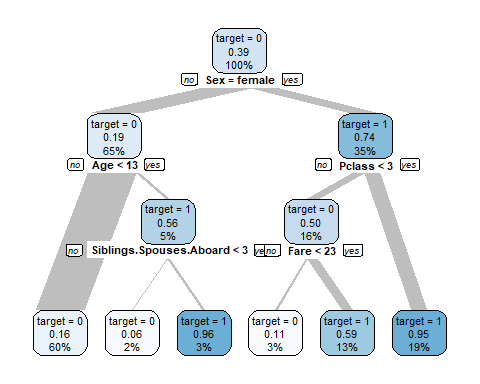
titanic %>% explore(Age, target = Pclass)



titanic %>% explore(Sex, target = Pclass)



titanic %>% explain\_tree(target = Survived)



titanic %>%   
 ggplot(aes(x = Age, fill = Survived)) +  
 geom\_histogram() +  
 facet\_wrap(~Sex + Pclass) +  
 theme\_test() +  
 theme(  
 plot.title = element\_text(family = "Times New Roman", hjust = 0.5),  
 axis.text = element\_text(family = "Times New Roman", face = "bold"),  
 axis.title = element\_text(family = "Times New Roman", face = "bold"),  
 legend.title = element\_blank(),  
 legend.text = element\_text(family = "Times New Roman")  
   
 ) +  
 labs(title = "Survival Rate, Age, Sex, and Passenger Class")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database

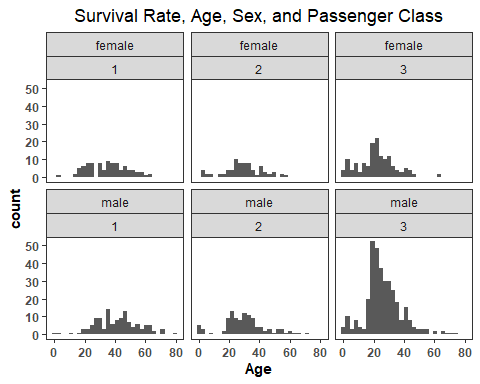
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database

## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database  
  
## Warning in grid.Call(C\_stringMetric, as.graphicsAnnot(x$label)): font family not  
## found in Windows font database

## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database  
  
## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database

## Warning in grid.Call.graphics(C\_text, as.graphicsAnnot(x$label), x$x, x$y, :  
## font family not found in Windows font database

## Warning in grid.Call(C\_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font  
## family not found in Windows font database



We see in the simple decision flow chart above that only 19% of male passengers less than age 13 survived. The decision tree and the distribution above suggests that sex, fare, and age increase the chances of survival of the passengers.

A logistic regression model will be used on data split to 75/25 to predict the survival of passengers in the test data. Accuracy is the metric that will be used for assessing the results of the model.

#val\_split <- createDataPartition(titanic$Species, p=0.70, list=FALSE)  
  
 #split data into 75/25 for training and validating the model.  
 train\_titanic <- titanic[1:667,]  
 test\_titanic <- titanic[668:889,]  
   
 #create the model  
   
 titanic\_log\_reg <- glm(Survived ~.,family=binomial(link='logit'),data=train\_titanic)  
  
 summary(titanic\_log\_reg)

##   
## Call:  
## glm(formula = Survived ~ ., family = binomial(link = "logit"),   
## data = train\_titanic)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5408 -0.6340 -0.4154 0.6312 2.3835   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.1650100 0.6503784 7.942 2.00e-15 \*\*\*  
## Pclass -1.1757050 0.1728042 -6.804 1.02e-11 \*\*\*  
## Sexmale -2.7346107 0.2254995 -12.127 < 2e-16 \*\*\*  
## Age -0.0361857 0.0086594 -4.179 2.93e-05 \*\*\*  
## Siblings.Spouses.Aboard -0.3217242 0.1253562 -2.566 0.0103 \*   
## Parents.Children.Aboard -0.1284693 0.1407896 -0.912 0.3615   
## Fare -0.0009965 0.0030665 -0.325 0.7452   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 892.88 on 666 degrees of freedom  
## Residual deviance: 605.34 on 660 degrees of freedom  
## AIC: 619.34  
##   
## Number of Fisher Scoring iterations: 5

#use anova to analyze deviance  
   
 anova(titanic\_log\_reg, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Survived  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 666 892.88   
## Pclass 1 63.940 665 828.94 1.282e-15 \*\*\*  
## Sex 1 199.950 664 628.99 < 2.2e-16 \*\*\*  
## Age 1 11.611 663 617.38 0.0006557 \*\*\*  
## Siblings.Spouses.Aboard 1 10.904 662 606.48 0.0009594 \*\*\*  
## Parents.Children.Aboard 1 1.039 661 605.44 0.3080034   
## Fare 1 0.103 660 605.34 0.7477738   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#for accuracy  
#confusionMatrix(data=model\_result, reference=test\_titanic$Survived)

#Testing For Accuracy Of The Model  
  
baseline\_accuracy = 545 / (545 + 342)  
  
predict\_train = predict(titanic\_log\_reg, type = "response")  
  
table(train\_titanic$Survived, predict\_train >= 0.5)

##   
## FALSE TRUE  
## 0 349 57  
## 1 80 181

accuracy = (181 + 349) / nrow(train\_titanic)  
sensitivity = 181 / (181 + 80)  
specificity = 349 / (349 + 57)  
  
cat("accuracy: ", accuracy, " > ", "baseline: ", baseline\_accuracy)

## accuracy: 0.7946027 > baseline: 0.6144307

#make predictions on test data  
   
 model\_result <- predict(titanic\_log\_reg,newdata=test\_titanic,type='response')  
 model\_result <- ifelse(model\_result > 0.5,1,0)  
   
 test\_titanic$Survived = as.numeric(model\_result >= 0.5)  
 table(test\_titanic$Survived)

##   
## 0 1   
## 148 72

#save results of predictions to csv  
 Predictions = data.frame(test\_titanic)  
 write.csv(file = "Titanic Predictions", x = Predictions)

The Model is found to be 79.4% accurate with a baseline of 61.6%. For the test data, it predicted that 72 out of 220 test passengers survived.