Homework 3

PSTAT 131 John Wei

Contents

Classification

For this assignment, we will be working with part of a Kaggle data set that was the subject of a machine learning competition and is often used for practicing ML models. The goal is classification; specifically, to predict which passengers would survive the Titanic shipwreck.

Load the data from `data/titanic.csv` into *R* and familiarize yourself with the variables it contains using the codebook (`data/titanic_codebook.txt`).

Notice that `survived` and `pclass` should be changed to factors. When changing `survived` to a factor, you may want to reorder the factor so that *"Yes"* is the first level.

Make sure you load the `tidyverse` and `tidymodels`!

```
library(tidymodels)
library(tidyverse)
library(ISLR)
library(ISLR2)
library(discrim)
library(poissonreg)
library(corrr)
library(corrplot)
library(klaR)
library(pROC)
library(tinytex)
set.seed(4167)
```

```
titanic <- read_csv("titanic.csv")
```

```
titanic
```

```
## # A tibble: 891 x 12
##    passenger_id survived pclass name       sex     age sib_sp parch ticket  fare
##           <dbl> <chr>     <dbl> <chr>      <chr> <dbl>  <dbl> <dbl> <chr>  <dbl>
##  1            1 No            3 Braund, M~ male     22      1     0 A/5 2~  7.25
##  2            2 Yes           1 Cumings, ~ fema~    38      1     0 PC 17~ 71.3
##  3            3 Yes           3 Heikkinen~ fema~    26      0     0 STON/~  7.92
##  4            4 Yes           1 Futrelle,~ fema~    35      1     0 113803 53.1
##  5            5 No            3 Allen, Mr~ male     35      0     0 373450  8.05
```

```
## 6             6 No             3 Moran, Mr~ male     NA      0      0 330877  8.46
## 7             7 No             1 McCarthy,~ male     54      0      0 17463  51.9
## 8             8 No             3 Palsson, ~ male      2      3      1 349909 21.1
## 9             9 Yes            3 Johnson, ~ fema~    27      0      2 347742 11.1
## 10           10 Yes            2 Nasser, M~ fema~    14      1      0 237736 30.1
## # ... with 881 more rows, and 2 more variables: cabin <chr>, embarked <chr>
```

```r
titanic$survived <- factor(titanic$survived)
titanic$pclass <- factor(titanic$pclass)
```

```r
titan <- titanic %>% arrange(desc(survived))
```

```r
titan
```

```
## # A tibble: 891 x 12
##    passenger_id survived pclass name       sex     age sib_sp parch ticket  fare
##           <dbl> <fct>    <fct>  <chr>      <chr> <dbl>  <dbl> <dbl> <chr>  <dbl>
##  1            2 Yes      1      Cumings, ~ fema~    38      1     0 PC 17~ 71.3
##  2            3 Yes      3      Heikkinen~ fema~    26      0     0 STON/~  7.92
##  3            4 Yes      1      Futrelle,~ fema~    35      1     0 113803 53.1
##  4            9 Yes      3      Johnson, ~ fema~    27      0     2 347742 11.1
##  5           10 Yes      2      Nasser, M~ fema~    14      1     0 237736 30.1
##  6           11 Yes      3      Sandstrom~ fema~     4      1     1 PP 95~ 16.7
##  7           12 Yes      1      Bonnell, ~ fema~    58      0     0 113783 26.6
##  8           16 Yes      2      Hewlett, ~ fema~    55      0     0 248706 16
##  9           18 Yes      2      Williams,~ male     NA      0     0 244373 13
## 10           20 Yes      3      Masselman~ fema~    NA      0     0 2649    7.22
## # ... with 881 more rows, and 2 more variables: cabin <chr>, embarked <chr>
```

Remember that you'll need to set a seed at the beginning of the document to reproduce your results.

Question 1

Split the data, stratifying on the outcome variable, `survived`. You should choose the proportions to split the data into. Verify that the training and testing data sets have the appropriate number of observations. Take a look at the training data and note any potential issues, such as missing data.

```r
titan_split <- initial_split(titan, prop = 0.80,
                             strata = survived)
titan_train <- training(titan_split)
titan_test <- testing(titan_split)
```

```r
titan_train %>% print(n = 100)
```

```
## # A tibble: 712 x 12
##     passenger_id survived pclass name      sex     age sib_sp parch ticket   fare
##            <dbl> <fct>    <fct>  <chr>     <chr> <dbl>  <dbl> <dbl> <chr>   <dbl>
##   1            1 No       3      "Braund~ male    22        1     0 A/5 2~   7.25
##   2            5 No       3      "Allen,~ male    35        0     0 373450   8.05
##   3            6 No       3      "Moran,~ male    NA        0     0 330877   8.46
##   4            7 No       1      "McCart~ male    54        0     0 17463   51.9
```

```
##   5           8 No        3      "Palsso~ male    2         3     1 349909  21.1
##   6          15 No        3      "Vestro~ fema~  14         0     0 350406   7.85
##   7          17 No        3      "Rice, ~ male    2         4     1 382652  29.1
##   8          19 No        3      "Vander~ fema~  31         1     0 345763  18
##   9          21 No        2      "Fynney~ male   35         0     0 239865  26
##  10          25 No        3      "Palsso~ fema~   8         3     1 349909  21.1
##  11          28 No        1      "Fortun~ male   19         3     2 19950  263
##  12          30 No        3      "Todoro~ male   NA         0     0 349216   7.90
##  13          36 No        1      "Holver~ male   42         1     0 113789  52
##  14          41 No        3      "Ahlin,~ fema~  40         1     0 7546      9.48
##  15          43 No        3      "Kraeff~ male   NA         0     0 349253   7.90
##  16          46 No        3      "Rogers~ male   NA         0     0 S.C./~   8.05
##  17          47 No        3      "Lennon~ male   NA         1     0 370371  15.5
##  18          49 No        3      "Samaan~ male   NA         2     0 2662    21.7
##  19          50 No        3      "Arnold~ fema~  18         1     0 349237  17.8
##  20          52 No        3      "Noswor~ male   21         0     0 A/4. ~   7.8
##  21          55 No        1      "Ostby,~ male   65         0     1 113509  62.0
##  22          58 No        3      "Novel,~ male   28.5       0     0 2697     7.23
##  23          60 No        3      "Goodwi~ male   11         5     2 CA 21~  46.9
##  24          63 No        1      "Harris~ male   45         1     0 36973   83.5
##  25          65 No        1      "Stewar~ male   NA         0     0 PC 17~  27.7
##  26          70 No        3      "Kink, ~ male   26         2     0 315151   8.66
##  27          72 No        3      "Goodwi~ fema~  16         5     2 CA 21~  46.9
##  28          73 No        2      "Hood, ~ male   21         0     0 S.O.C~  73.5
##  29          74 No        3      "Chrono~ male   26         1     0 2680    14.5
##  30          77 No        3      "Stanef~ male   NA         0     0 349208   7.90
##  31          78 No        3      "Moutal~ male   NA         0     0 374746   8.05
##  32          84 No        1      "Carrau~ male   28         0     0 113059  47.1
##  33          87 No        3      "Ford, ~ male   16         1     3 W./C.~  34.4
##  34          88 No        3      "Slocov~ male   NA         0     0 SOTON~   8.05
##  35          90 No        3      "Celott~ male   24         0     0 343275   8.05
##  36          91 No        3      "Christ~ male   29         0     0 343276   8.05
##  37          92 No        3      "Andrea~ male   20         0     0 347466   7.85
##  38          93 No        1      "Chaffe~ male   46         1     0 W.E.P~  61.2
##  39          94 No        3      "Dean, ~ male   26         1     2 C.A. ~  20.6
##  40          95 No        3      "Coxon,~ male   59         0     0 364500   7.25
##  41          96 No        3      "Shorne~ male   NA         0     0 374910   8.05
##  42          97 No        1      "Goldsc~ male   71         0     0 PC 17~  34.7
##  43         100 No        2      "Kantor~ male   34         1     0 244367  26
##  44         101 No        3      "Petran~ fema~  28         0     0 349245   7.90
##  45         102 No        3      "Petrof~ male   NA         0     0 349215   7.90
##  46         103 No        1      "White,~ male   21         0     1 35281   77.3
##  47         104 No        3      "Johans~ male   33         0     0 7540     8.65
##  48         105 No        3      "Gustaf~ male   37         2     0 31012~   7.92
##  49         106 No        3      "Mionof~ male   28         0     0 349207   7.90
##  50         109 No        3      "Rekic,~ male   38         0     0 349249   7.90
##  51         111 No        1      "Porter~ male   47         0     0 110465  52
##  52         112 No        3      "Zabour~ fema~  14.5       1     0 2665    14.5
##  53         113 No        3      "Barton~ male   22         0     0 324669   8.05
##  54         114 No        3      "Jussil~ fema~  20         1     0 4136     9.82
##  55         115 No        3      "Attala~ fema~  17         0     0 2627    14.5
##  56         116 No        3      "Pekoni~ male   21         0     0 STON/~   7.92
##  57         117 No        3      "Connor~ male   70.5       0     0 370369   7.75
##  58         118 No        2      "Turpin~ male   29         1     0 11668   21
```

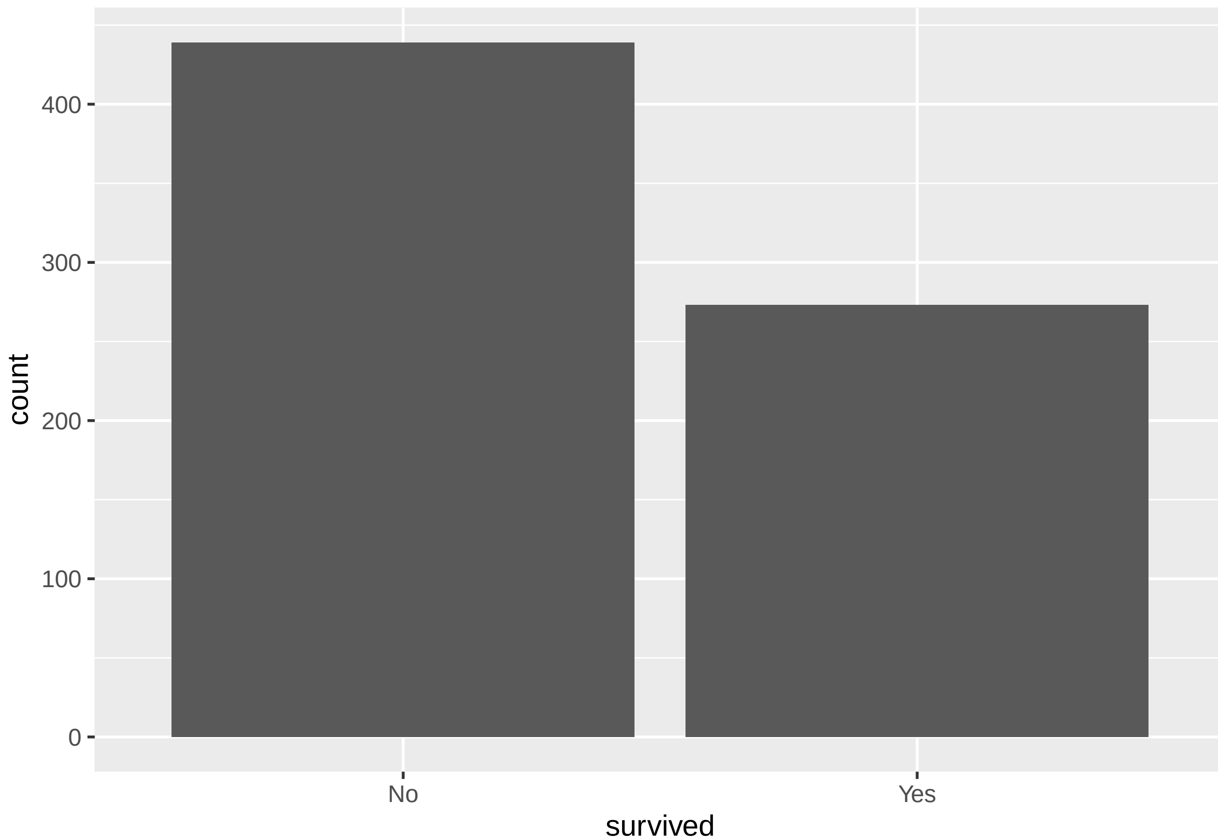
```
## 59        119 No      1       "Baxter~ male   24         0      1 PC 17~ 248.
## 60        121 No      2       "Hickma~ male   21         2      0 S.O.C~  73.5
## 61        122 No      3       "Moore,~ male   NA         0      0 A4. 5~   8.05
## 62        123 No      2       "Nasser~ male   32.5       1      0 237736  30.1
## 63        125 No      1       "White,~ male   54         0      1 35281   77.3
## 64        127 No      3       "McMaho~ male   NA         0      0 370372   7.75
## 65        130 No      3       "Ekstro~ male   45         0      0 347061   6.98
## 66        131 No      3       "Drazen~ male   33         0      0 349241   7.90
## 67        133 No      3       "Robins~ fema~  47         1      0 A/5. ~  14.5
## 68        135 No      2       "Sobey,~ male   25         0      0 C.A. ~  13
## 69        136 No      2       "Richar~ male   23         0      0 SC/PA~  15.0
## 70        138 No      1       "Futrel~ male   37         1      0 113803  53.1
## 71        139 No      3       "Osen, ~ male   16         0      0 7534     9.22
## 72        140 No      1       "Giglio~ male   24         0      0 PC 17~  79.2
## 73        141 No      3       "Boulos~ fema~  NA         0      2 2678    15.2
## 74        144 No      3       "Burke,~ male   19         0      0 365222   6.75
## 75        145 No      2       "Andrew~ male   18         0      0 231945  11.5
## 76        148 No      3       "Ford, ~ fema~   9         2      2 W./C.~  34.4
## 77        150 No      2       "Byles,~ male   42         0      0 244310  13
## 78        151 No      2       "Batema~ male   51         0      0 S.O.P~  12.5
## 79        153 No      3       "Meo, M~ male   55.5       0      0 A.5. ~   8.05
## 80        155 No      3       "Olsen,~ male   NA         0      0 Fa 26~   7.31
## 81        156 No      1       "Willia~ male   51         0      1 PC 17~  61.4
## 82        159 No      3       "Smilja~ male   NA         0      0 315037   8.66
## 83        161 No      3       "Cribb,~ male   44         0      1 371362  16.1
## 84        163 No      3       "Bengts~ male   26         0      0 347068   7.78
## 85        164 No      3       "Calic,~ male   17         0      0 315093   8.66
## 86        168 No      3       "Skoog,~ fema~  45         1      4 347088  27.9
## 87        169 No      1       "Bauman~ male   NA         0      0 PC 17~  25.9
## 88        170 No      3       "Ling, ~ male   28         0      0 1601    56.5
## 89        172 No      3       "Rice, ~ male    4         4      1 382652  29.1
## 90        174 No      3       "Sivola~ male   21         0      0 STON/~   7.92
## 91        175 No      1       "Smith,~ male   56         0      0 17764   30.7
## 92        178 No      1       "Isham,~ fema~  50         0      0 PC 17~  28.7
## 93        179 No      2       "Hale, ~ male   30         0      0 250653  13
## 94        180 No      3       "Leonar~ male   36         0      0 LINE     0
## 95        181 No      3       "Sage, ~ fema~  NA         8      2 CA. 2~  69.6
## 96        182 No      2       "Pernot~ male   NA         0      0 SC/PA~  15.0
## 97        183 No      3       "Asplun~ male    9         4      2 347077  31.4
## 98        186 No      1       "Rood, ~ male   NA         0      0 113767  50
## 99        189 No      3       "Bourke~ male   40         1      1 364849  15.5
## 100       190 No      3       "Turcin~ male   36         0      0 349247   7.90
## # ... with 612 more rows, and 2 more variables: cabin <chr>, embarked <chr>
```

There exist missing data in the observations age and cabin. The missing ages would probably change our data a little bit. Some of the variables may be correlated to each other. It is a good idea to use stratified sampling as we want to focus on and group by the people who either survived (or didn't survive). There may be differences in those populations - for example, where they were staying or their age.

Question 2

Using the **training** data set, explore/describe the distribution of the outcome variable `survived`.

```r
titan_train %>%
  ggplot(aes(x = survived)) +
  geom_bar(group = 1)
```

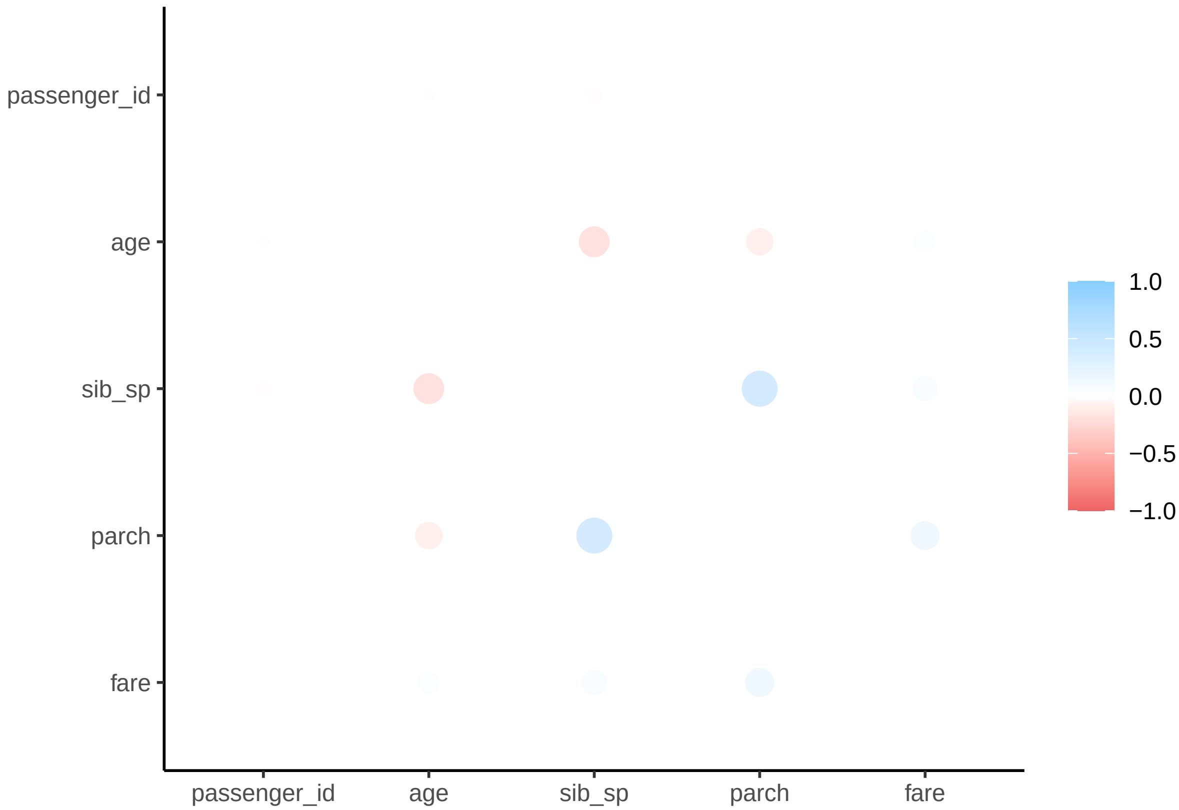


Using a barplot, we see more people in the training set did not survived - approximately 60% people did not survive.

Question 3

Using the **training** data set, create a correlation matrix of all continuous variables. Create a visualization of the matrix, and describe any patterns you see. Are any predictors correlated with each other? Which ones, and in which direction?

```r
cor_titan_train <- titan_train %>%
  dplyr::select(-c(survived, pclass, name, sex, ticket, cabin, embarked)) %>%
  correlate(use = "pairwise.complete.obs", method = "pearson")
rplot(cor_titan_train)
```



Most of the variables do not have correlation with each other. sib_sp and parch have strong positive correlation, parch and fare have a slightly positive correlation, sib_sp and fare have a very slight positive correlation, age and parch have a slightly negative correlation, and sib_sp and age have a decently negative correlation.

Question 4

Using the **training** data, create a recipe predicting the outcome variable `survived`. Include the following predictors: ticket class, sex, age, number of siblings or spouses aboard, number of parents or children aboard, and passenger fare.

```
titan_recipe <- recipe(survived ~ pclass + sex + age + sib_sp +
                           parch + fare, data = titan_train) %>%
  step_impute_linear(age) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("sex"):fare) %>%
  step_interact( ~ age:fare)
```

Question 5

Specify a **logistic regression** model for classification using the `"glm"` engine. Then create a workflow. Add your model and the appropriate recipe. Finally, use `fit()` to apply your workflow to the **training** data.

```
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
```

```r
log_wkflow <- workflow() %>%
  add_model(log_reg) %>%
  add_recipe(titan_recipe)
log_fit <- fit(log_wkflow, titan_train)
log_fit %>%
  tidy()
```

```
## # A tibble: 10 x 5
##    term            estimate std.error statistic  p.value
##    <chr>              <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)      4.31      0.624        6.91  4.87e-12
##  2 age             -0.0521    0.0120      -4.36  1.33e- 5
##  3 sib_sp          -0.428     0.124       -3.44  5.92e- 4
##  4 parch           -0.100     0.129       -0.777 4.37e- 1
##  5 fare            -0.00314   0.00840     -0.374 7.08e- 1
##  6 pclass_X2       -1.13      0.340       -3.31  9.21e- 4
##  7 pclass_X3       -2.40      0.356       -6.73  1.71e-11
##  8 sex_male        -2.58      0.280       -9.19  3.80e-20
##  9 sex_male_x_fare -0.00654   0.00624     -1.05  2.95e- 1
## 10 age_x_fare       0.000314  0.000179     1.75  8.02e- 2
```

Question 6

Repeat Question 5, but this time specify a linear discriminant analysis model for classification using the `"MASS"` engine.

```r
lda_mod <- discrim_linear() %>%
  set_mode("classification") %>%
  set_engine("MASS")
lda_wkflow <- workflow() %>%
  add_model(lda_mod) %>%
  add_recipe(titan_recipe)
lda_fit <- fit(lda_wkflow, titan_train)
```

Question 7

Repeat Question 5, but this time specify a quadratic discriminant analysis model for classification using the `"MASS"` engine.

```r
qda_mod <- discrim_quad() %>%
  set_mode("classification") %>%
  set_engine("MASS")
qda_wkflow <- workflow() %>%
  add_model(qda_mod) %>%
  add_recipe(titan_recipe)
qda_fit <- fit(qda_wkflow, titan_train)
```

Question 8

Repeat Question 5, but this time specify a naive Bayes model for classification using the `"klaR"` engine. Set the `usekernel` argument to `FALSE`.

```r
nb_mod <- naive_Bayes() %>%
  set_mode("classification") %>%
  set_engine("klaR") %>%
  set_args(usekernel = FALSE)
nb_wkflow <- workflow() %>%
  add_model(nb_mod) %>%
  add_recipe(titan_recipe)
nb_fit <- fit(nb_wkflow, titan_train)
```

Question 9

Now you've fit four different models to your training data.

Use `predict()` and `bind_cols()` to generate predictions using each of these 4 models and your **training** data. Then use the *accuracy* metric to assess the performance of each of the four models.

Which model achieved the highest accuracy on the training data?

```r
titan_log<-predict(log_fit, new_data = titan_train, type = "prob")
titan_log_col<-bind_cols(titan_log, titan_train)
log_reg_acc<- augment(log_fit, new_data= titan_train) %>% accuracy(truth=survived, estimate=.pred_class
log_reg_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.808
```

```r
titan_lda<-predict(lda_fit, new_data = titan_train, type = "prob")
titan_lda_col<-bind_cols(titan_lda, titan_train)
lda_reg_acc<- augment(lda_fit, new_data= titan_train) %>% accuracy(truth=survived, estimate=.pred_class
lda_reg_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.799
```

```r
titan_qda<-predict(qda_fit, new_data = titan_train, type = "prob")
titan_qda_col<-bind_cols(titan_qda, titan_train)
qda_reg_acc<- augment(qda_fit, new_data= titan_train) %>% accuracy(truth=survived, estimate=.pred_class
qda_reg_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.787
```

```r
titan_nb<-predict(nb_fit, new_data = titan_train, type = "prob")
titan_nb_col<-bind_cols(titan_nb, titan_train)
nb_reg_acc<- augment(nb_fit, new_data= titan_train) %>% accuracy(truth=survived, estimate=.pred_class)
nb_reg_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.775
```

```
accuracies <- c(log_reg_acc$.estimate, lda_reg_acc$.estimate,
                qda_reg_acc$.estimate, nb_reg_acc$.estimate)
models <- c("Logistic Regression", "LDA", "Naive Bayes", "QDA")
results <- tibble(accuracies = accuracies, models = models)
```

```
results
```

```
## # A tibble: 4 x 2
##   accuracies models
##        <dbl> <chr>
## 1      0.808 Logistic Regression
## 2      0.799 LDA
## 3      0.787 Naive Bayes
## 4      0.775 QDA
```

The Logistic Regression model has the highest accuracy of all the models at 80.76%.

Question 10

Fit the model with the highest training accuracy to the **testing** data. Report the accuracy of the model on the **testing** data.

Again using the **testing** data, create a confusion matrix and visualize it. Plot an ROC curve and calculate the area under it (AUC).

How did the model perform? Compare its training and testing accuracies. If the values differ, why do you think this is so?

```
bind_cols(predict(log_fit, new_data=titan_test), titan_test %>% dplyr::select(survived))
```

```
## # A tibble: 179 x 2
##    .pred_class survived
##    <fct>       <fct>
##  1 Yes         Yes
##  2 Yes         Yes
##  3 Yes         Yes
##  4 No          Yes
##  5 Yes         Yes
##  6 Yes         Yes
##  7 Yes         Yes
##  8 Yes         Yes
##  9 Yes         Yes
## 10 Yes         Yes
## # ... with 169 more rows
```

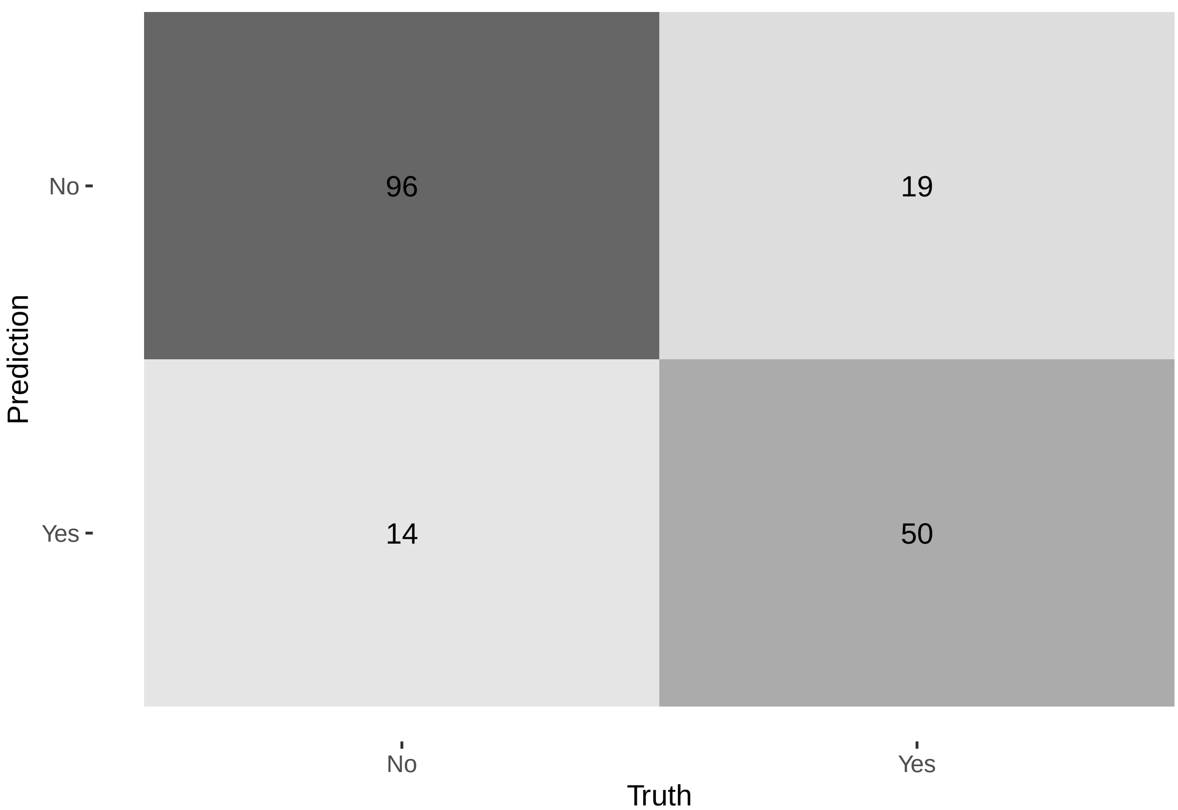
```
bind_cols(predict(log_fit, new_data=titan_test), titan_test %>% dplyr::select(survived)) %>% accuracy(t
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.816
```

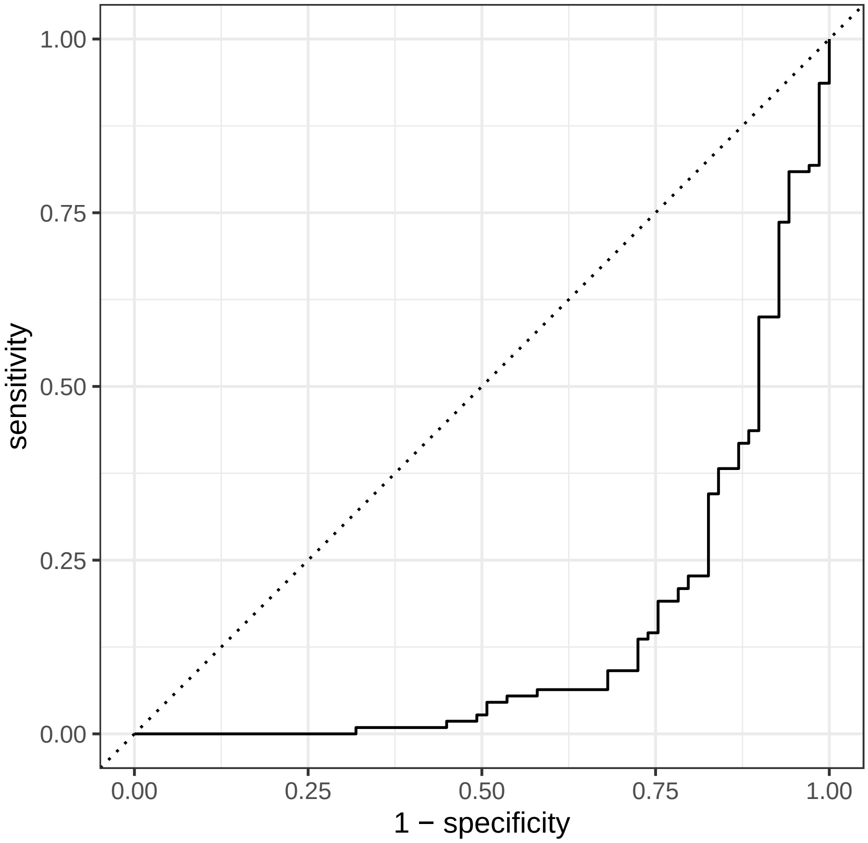
```
augment(log_fit, new_data = titan_test) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

```
##           Truth
## Prediction No Yes
##        No  96  19
##        Yes 14  50
```

```
augment(log_fit, new_data = titan_test) %>%
  conf_mat(truth = survived, estimate = .pred_class) %>% autoplot(type = "heatmap")
```




```
augment(log_fit, new_data = titan_test) %>%
  roc_curve(survived, .pred_Yes) %>%
  autoplot()
```




```
pROC::auc(augment(log_fit, new_data = titan_test)$survived, augment(log_fit, new_data = titan_test)$.pr
```

```
## Area under the curve: 0.8589
```

The training and test accuracies are similar (81.56 vs 80.76%). The values differ slightly, perhaps because of overfitting in the training set, the method of measuring accuracies, and/or correlation differences between training/test data. The confusion matrix looks like it predicts the right outcome most of the time.