Rene Guerra - Homework 3

Sunday, October 30th 2022

```
Titanic$survived <- factor(Titanic$survived, levels= c('Yes', 'No'))
Titanic$pclass <- factor(Titanic$pclass)

levels(Titanic$survived)
```

```
## [1] "Yes" "No"
```

1

```
set.seed(2022)

Titanic_split <- initial_split(Titanic, prop= 0.80, strata= survived)

Titanic_train <- training(Titanic_split)
Titanic_train
```

```
## # 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, M~ male     22      1     0 A/5 2~  7.25
##  2            6 No       3      Moran, Mr~ male     NA      0     0 330877  8.46
##  3            7 No       1      McCarthy,~ male     54      0     0 17463  51.9
##  4            8 No       3      Palsson, ~ male      2      3     1 349909 21.1
##  5           13 No       3      Saunderco~ male     20      0     0 A/5. ~  8.05
##  6           14 No       3      Andersson~ male     39      1     5 347082 31.3
##  7           15 No       3      Vestrom, ~ fema~    14      0     0 350406  7.85
##  8           17 No       3      Rice, Mas~ male      2      4     1 382652 29.1
##  9           19 No       3      Vander Pl~ fema~    31      1     0 345763 18
## 10           21 No       2      Fynney, M~ male     35      0     0 239865 26
## # ... with 702 more rows, and 2 more variables: cabin <chr>, embarked <chr>
```

```
Titanic_test <- testing(Titanic_split)
Titanic_test
```

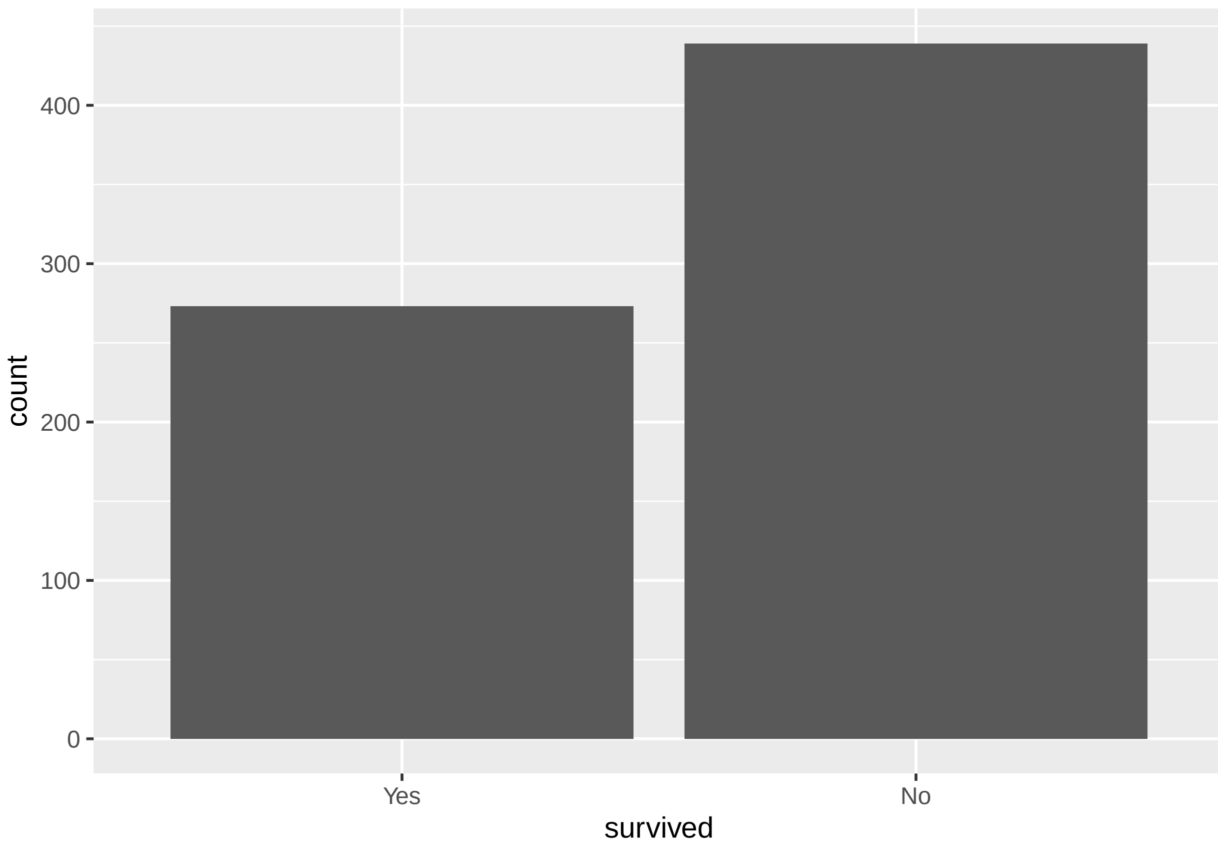
```
## # A tibble: 179 x 12
##    passenger_id survived pclass name      sex     age sib_sp parch ticket   fare
##           <dbl> <fct>    <fct>  <chr>     <chr> <dbl>  <dbl> <dbl> <chr>   <dbl>
##  1            5 No       3      Allen, M~ male     35      0     0 373450   8.05
##  2            9 Yes      3      Johnson,~ fema~    27      0     2 347742  11.1
##  3           28 No       1      Fortune,~ male     19      3     2 19950  263
##  4           39 No       3      Vander P~ fema~    18      2     0 345764  18
##  5           49 No       3      Samaan, ~ male     NA      2     0 2662    21.7
##  6           50 No       3      Arnold-F~ fema~    18      1     0 349237  17.8
##  7           54 Yes      2      Faunthor~ fema~    29      1     0 2926    26
##  8           69 Yes      3      Andersso~ fema~    17      4     2 31012~   7.92
##  9           74 No       3      Chronopo~ male     26      1     0 2680    14.5
## 10           75 Yes      3      Bing, Mr~ male     32      0     0 1601    56.5
```

```
## # ... with 169 more rows, and 2 more variables: cabin <chr>, embarked <chr>
```

It is a good idea to use stratified sampling for this data because the sample comes from all categories and is divided into subcategories that will potentially derive different results for the outcome variable we are evaluating.

2

```
Titanic_train %>%
  ggplot(aes(x= survived)) + geom_bar()
```



Most people did not survive based on the training data set.

3

```
is.numeric(Titanic$passenger_id)
```

```
## [1] TRUE
```

```
is.numeric(Titanic$survived)
```

```
## [1] FALSE
```

```
is.numeric(Titanic$pclass)
```

```
## [1] FALSE
```

```
is.numeric(Titanic$name)
```

```
## [1] FALSE
```

```r
is.numeric(Titanic$sex)
```

```
## [1] FALSE
```

```r
is.numeric(Titanic$age)
```

```
## [1] TRUE
```

```r
is.numeric(Titanic$sib_sp)
```

```
## [1] TRUE
```

```r
is.numeric(Titanic$parch)
```

```
## [1] TRUE
```

```r
is.numeric(Titanic$ticket)
```

```
## [1] FALSE
```

```r
is.numeric(Titanic$fare)
```

```
## [1] TRUE
```

```r
is.numeric(Titanic$cabin)
```

```
## [1] FALSE
```

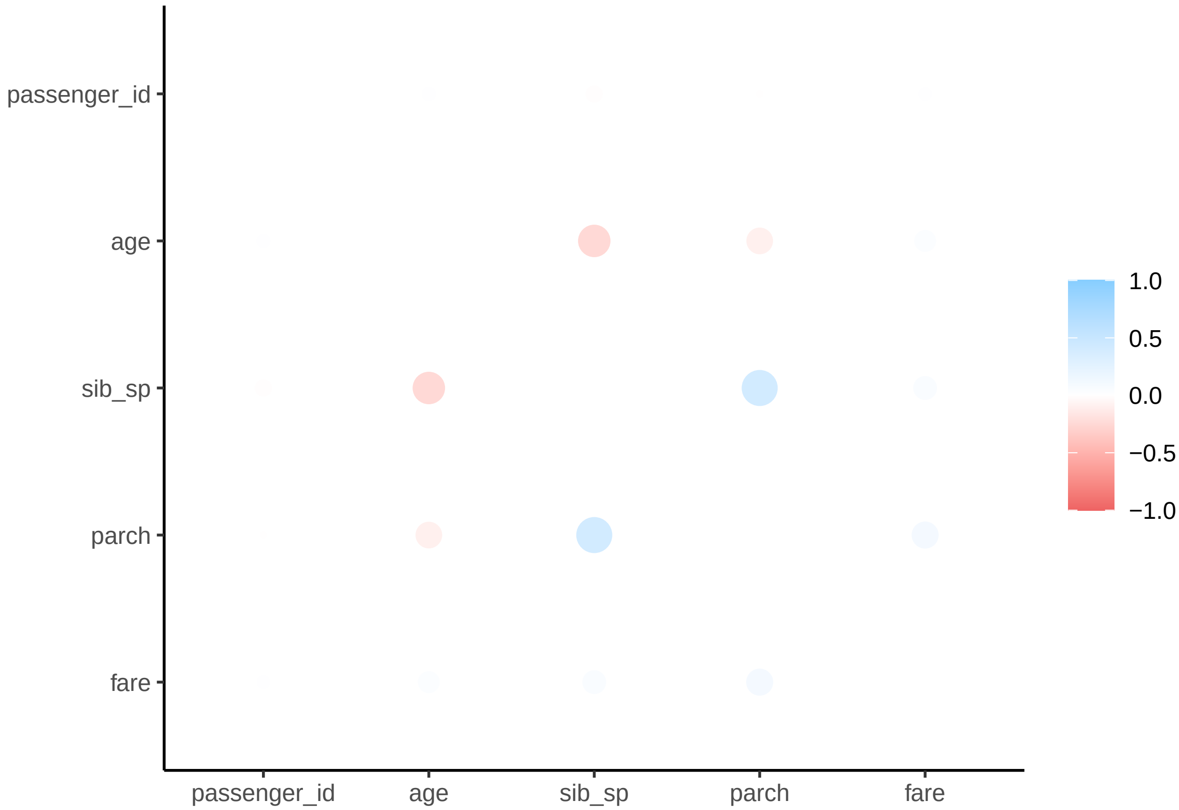
```r
is.numeric(Titanic$embarked)
```

```
## [1] FALSE
```

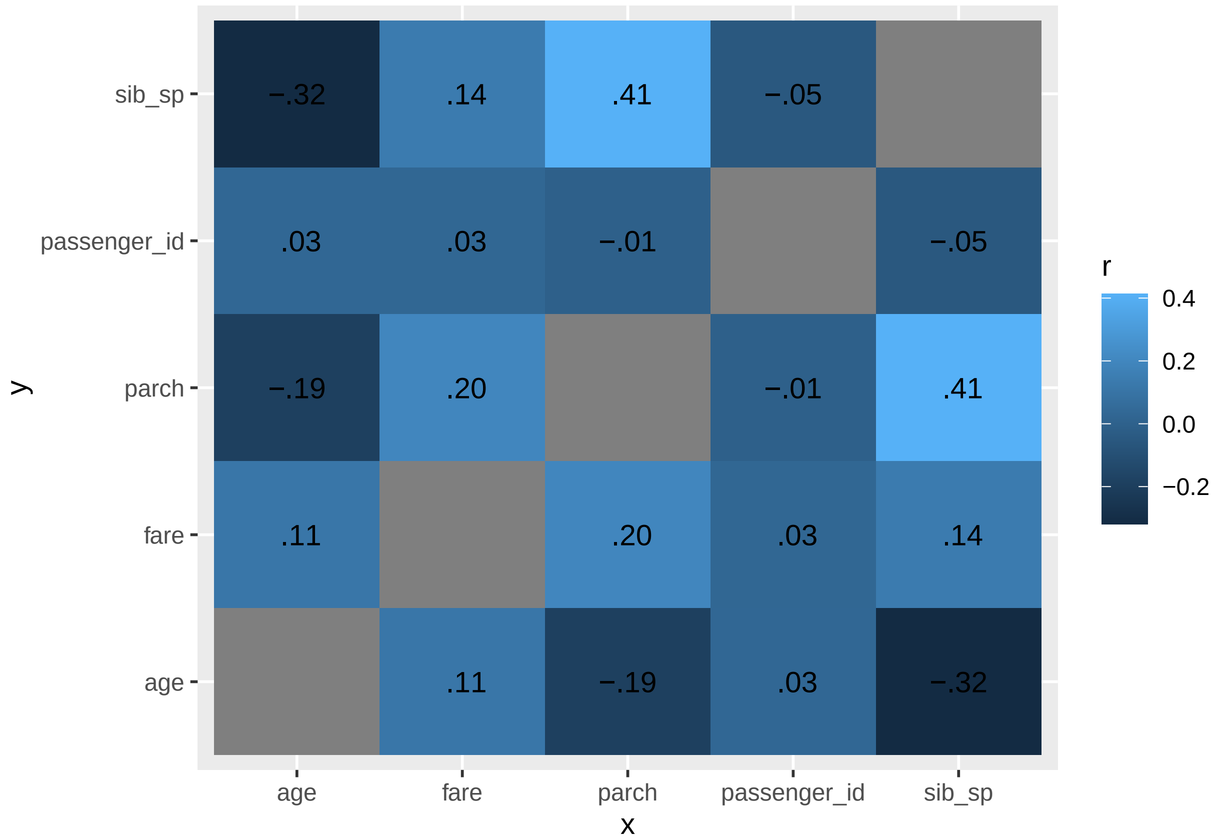
```r
Cor_Titanic <- Titanic_train %>%
  select(-c(survived, pclass, name, sex, ticket, cabin, embarked)) %>%
  correlate()
```

```
## Correlation computed with
## * Method: 'pearson'
## * Missing treated using: 'pairwise.complete.obs'
```

```r
rplot(Cor_Titanic)
```




```
Cor_Titanic %>%
  stretch() %>%
  ggplot(aes(x, y, fill= r)) +
  geom_tile() +
  geom_text(aes(label= as.character(fashion(r))))
```



The main pattern I see is the negative correlation between sib_sp and age. The correlation matrix shows that older passengers were not accompanied by siblings and younger passengers did not have spouses. Another significant negative correlation is parch and age. Older passengers did not have parents on board and younger passengers were not accompanied by children. On the other hand, sib_sp and parch have a significantly positive correlation. This means that most children and siblings had parents aboard, thus there were families.

4

```
Titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp + parch + fare, Titanic_train) %>%
  step_impute_linear(age, impute_with = imp_vars(sib_sp)) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(~ starts_with("sex"):fare + age:fare)

Titanic_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          6
##
## Operations:
##
## Linear regression imputation for age
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("sex"):fare + age:fare
```

```r
Interaction1 <- lm(survived ~ sex + fare, data= Titanic_train)
Interaction1
```

```
##
## Call:
## lm(formula = survived ~ sex + fare, data = Titanic_train)
##
## Coefficients:
## (Intercept)      sexmale         fare
##    1.344099     0.510056    -0.001788
```

```r
Interaction2 <- lm(survived ~ age + fare, data= Titanic_train)
Interaction2
```

```
##
## Call:
## lm(formula = survived ~ age + fare, data = Titanic_train)
##
## Coefficients:
## (Intercept)          age         fare
##    1.566710     0.004009    -0.002702
```

5

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

glm_Titanic
```

```
## Logistic Regression Model Specification (classification)
##
## Computational engine: glm
```

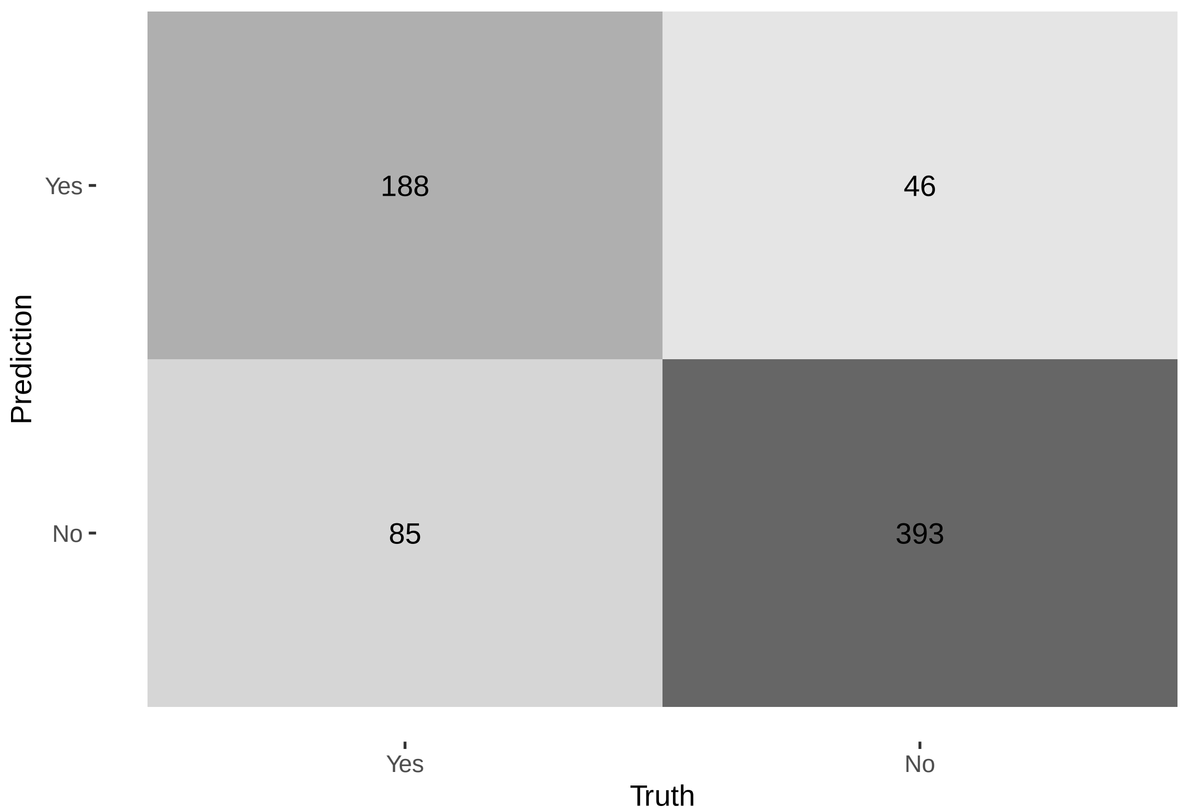
```r
glm_Titanicflow<- workflow() %>%
  add_model(glm_Titanic) %>%
  add_recipe(Titanic_recipe)

TitanicFit1 <- fit(glm_Titanicflow, Titanic_train)

TitanicFit1 %>%
  tidy()
```

```
## # A tibble: 10 x 5
##    term             estimate std.error statistic  p.value
##    <chr>               <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)     -3.88      0.624       -6.22  5.02e-10
##  2 age              0.0539    0.0121       4.47  7.94e- 6
##  3 sib_sp           0.520     0.130        4.02  5.92e- 5
##  4 parch            0.207     0.138        1.50  1.34e- 1
##  5 fare            -0.0106    0.0117      -0.906 3.65e- 1
##  6 pclass_X2        0.878     0.338        2.60  9.33e- 3
##  7 pclass_X3        1.95      0.343        5.70  1.18e- 8
##  8 sex_male         2.44      0.303        8.07  7.24e-16
##  9 sex_male_x_fare  0.0139    0.00930      1.49  1.35e- 1
## 10 fare_x_age      -0.000235  0.000203    -1.16  2.47e- 1
```

```r
predict(TitanicFit1, new_data = Titanic_train, type= "prob")
```

```
## # A tibble: 712 x 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.0989    0.901
##  2    0.0986    0.901
##  3    0.272     0.728
##  4    0.0792    0.921
##  5    0.170     0.830
##  6    0.0181    0.982
##  7    0.783     0.217
##  8    0.0476    0.952
##  9    0.515     0.485
## 10    0.232     0.768
## # ... with 702 more rows
```

```r
augment(TitanicFit1, new_data= Titanic_train) %>%
  conf_mat(truth= survived, estimate= .pred_class) %>%
  autoplot(type= "heatmap")
```




```r
glm_accuracy <- augment(TitanicFit1, new_data= Titanic_train) %>%
  accuracy(truth= survived, estimate= .pred_class)

glm_accuracy
```

```
## # A tibble: 1 x 3
```

```
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.816
```

6

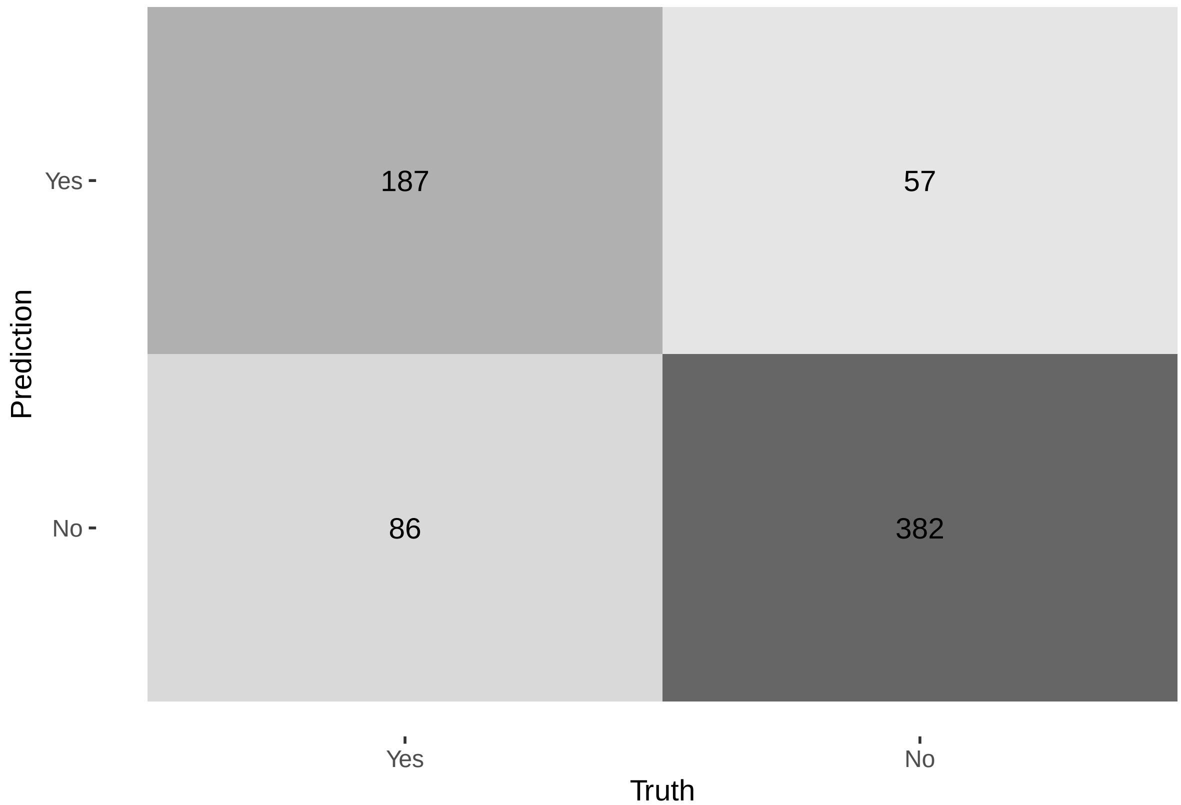
```r
lda_Titanic <- discrim_linear() %>%
  set_engine("MASS") %>%
  set_mode("classification")

lda_Titanic
```

```
## Linear Discriminant Model Specification (classification)
##
## Computational engine: MASS
```

```r
lda_TitanicFlow <- workflow() %>%
  add_model(lda_Titanic) %>%
  add_recipe(Titanic_recipe)

TitanicFit2 <- fit(lda_TitanicFlow, Titanic_train)

predict(TitanicFit2, new_data = Titanic_train, type= "prob")
```

```
## # A tibble: 712 x 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.0603    0.940
##  2    0.0566    0.943
##  3    0.224     0.776
##  4    0.0576    0.942
##  5    0.100     0.900
##  6    0.0118    0.988
##  7    0.844     0.156
##  8    0.0380    0.962
##  9    0.602     0.398
## 10    0.164     0.836
## # ... with 702 more rows
```

```r
augment(TitanicFit2, new_data= Titanic_train) %>%
  conf_mat(truth= survived, estimate= .pred_class) %>%
  autoplot(type= "heatmap")
```




```r
lda_accuracy <- augment(TitanicFit2, new_data= Titanic_train) %>%
  accuracy(truth= survived, estimate= .pred_class)

lda_accuracy
```

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

7

```r
qda_Titanic <- discrim_quad() %>%
  set_engine("MASS") %>%
  set_mode("classification")

qda_Titanic
```

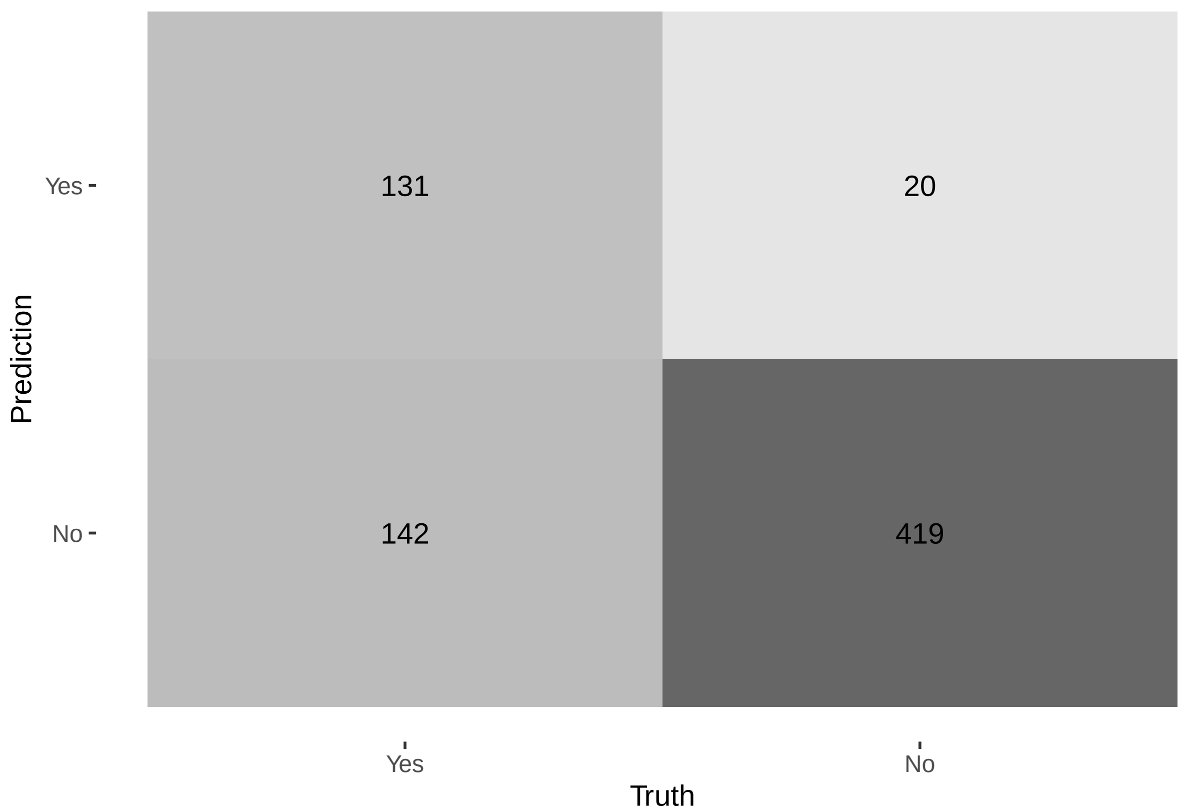
```
## Quadratic Discriminant Model Specification (classification)
##
## Computational engine: MASS
```

```r
qda_TitanicFlow <- workflow() %>%
  add_model(qda_Titanic) %>%
  add_recipe(Titanic_recipe)

TitanicFit3 <- fit(qda_TitanicFlow, Titanic_train)
```

```r
predict(TitanicFit3, new_data = Titanic_train, type= "prob")
```

```
## # A tibble: 712 x 2
##       .pred_Yes .pred_No
##           <dbl>    <dbl>
##  1 0.00456         0.995
##  2 0.00387         0.996
##  3 0.0434          0.957
##  4 0.0000286       1.00
##  5 0.00744         0.993
##  6 0.00269         0.997
##  7 0.436           0.564
##  8 0.0000000712    1.00
##  9 0.207           0.793
## 10 0.00754         0.992
## # ... with 702 more rows
```

```r
augment(TitanicFit3, new_data= Titanic_train) %>%
  conf_mat(truth= survived, estimate= .pred_class) %>%
  autoplot(type= "heatmap")
```




```r
qda_accuracy <- augment(TitanicFit3, new_data= Titanic_train) %>%
  accuracy(truth= survived, estimate= .pred_class)

glm_accuracy
```

```
## # A tibble: 1 x 3
```

```
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.816
```

8

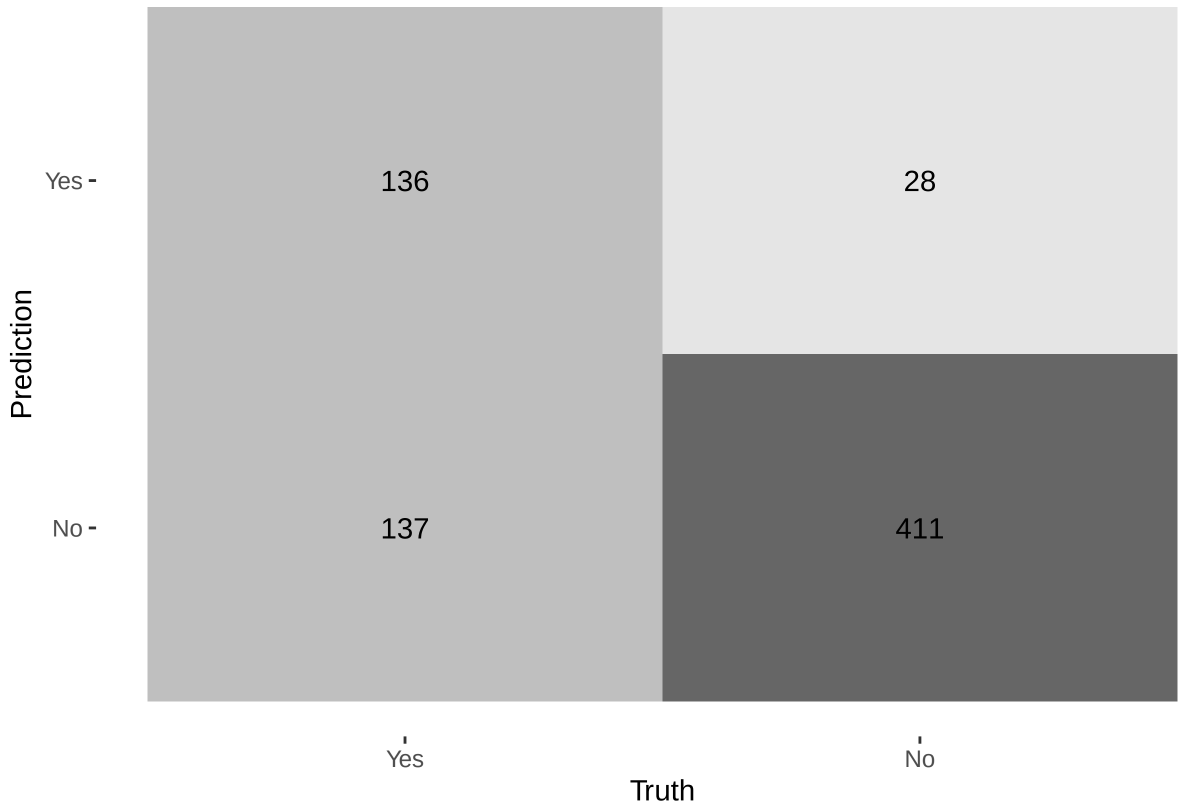
```
nB_Titanic <- naive_Bayes() %>%
  set_engine("klaR") %>%
  set_mode("classification") %>%
  set_args(usekernel= FALSE)

nB_Titanic
```

```
## Naive Bayes Model Specification (classification)
##
## Engine-Specific Arguments:
##   usekernel = FALSE
##
## Computational engine: klaR
```

```
nB_TitanicFlow <- workflow() %>%
  add_model(nB_Titanic) %>%
  add_recipe(Titanic_recipe)

TitanicFit4 <- fit(nB_TitanicFlow, Titanic_train)

predict(TitanicFit4, new_data= Titanic_train, type= "prob")
```

```
## # A tibble: 712 x 2
##      .pred_Yes .pred_No
##          <dbl>    <dbl>
##  1 0.0120         0.988
##  2 0.0122         0.988
##  3 0.415          0.585
##  4 0.0000633      1.00
##  5 0.0142         0.986
##  6 0.000744       0.999
##  7 0.370          0.630
##  8 0.000000292    1.00
##  9 0.277          0.723
## 10 0.118          0.882
## # ... with 702 more rows
```

```
augment(TitanicFit4, new_data= Titanic_train) %>%
  conf_mat(truth= survived, estimate= .pred_class) %>%
  autoplot(type= "heatmap")
```




```
nB_accuracy <- augment(TitanicFit4, new_data= Titanic_train) %>%
  accuracy(truth= survived, estimate= .pred_class)

nB_accuracy
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.768
```

9

```
TotalAccurate <- c(glm_accuracy$.estimate, lda_accuracy$.estimate, qda_accuracy$.estimate, nB_accuracy$

ModType <- c("Logistic Regression", "LDA", "QDA", "Naive Bayes")

Performance <- tibble(TotalAccurate= TotalAccurate, ModType= ModType)

Performance %>%
  arrange(-TotalAccurate)
```

```
## # A tibble: 4 x 2
##   TotalAccurate ModType
##           <dbl> <chr>
## 1         0.816 Logistic Regression
## 2         0.799 LDA
## 3         0.772 QDA
## 4         0.768 Naive Bayes
```

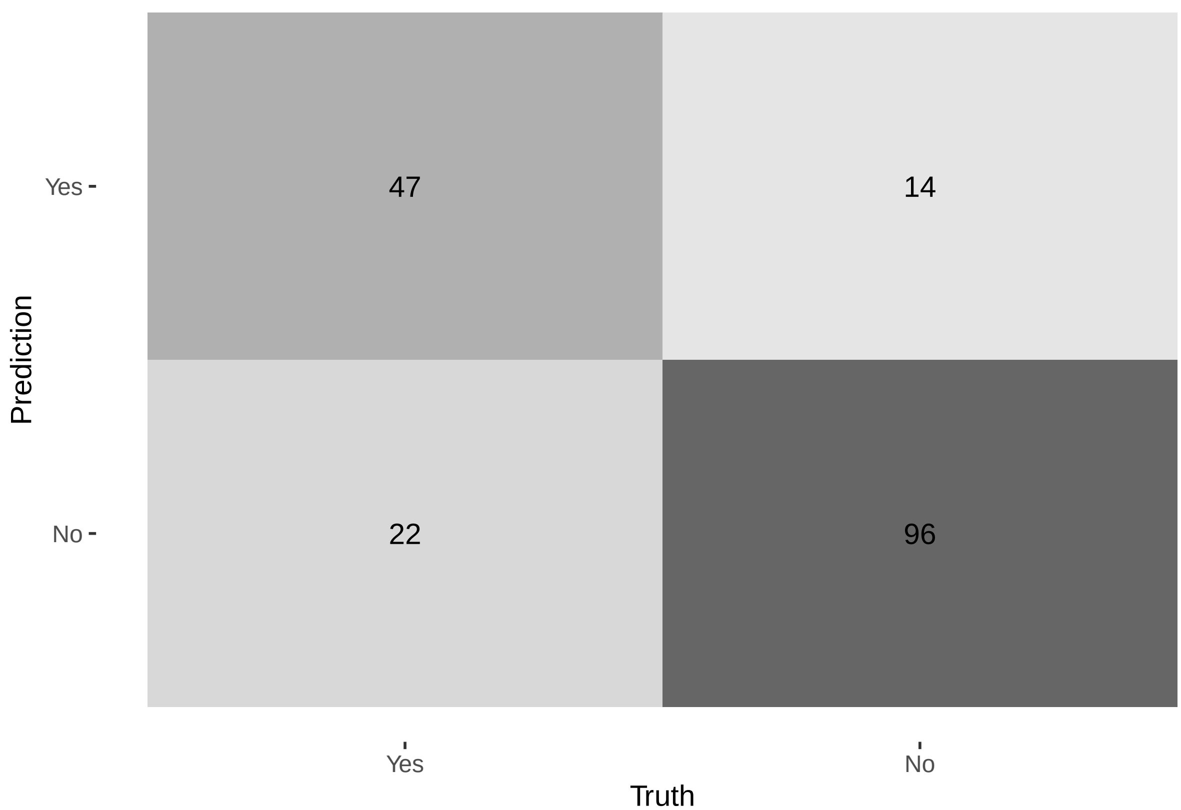
The model that achieved the highest accuracy on the training data is Logistic Regression with 0.8160112 accuracy.

10

```
predict(TitanicFit1, new_data= Titanic_test, type= "prob")
```

```
## # A tibble: 179 x 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.0860    0.914
##  2    0.561     0.439
##  3    0.223     0.777
##  4    0.546     0.454
##  5    0.0627    0.937
##  6    0.669     0.331
##  7    0.798     0.202
##  8    0.203     0.797
##  9    0.0834    0.917
## 10    0.119     0.881
## # ... with 169 more rows
```

```
augment(TitanicFit1, new_data= Titanic_test) %>%
  conf_mat(truth= survived, estimate= .pred_class) %>%
  autoplot(type= "heatmap")
```

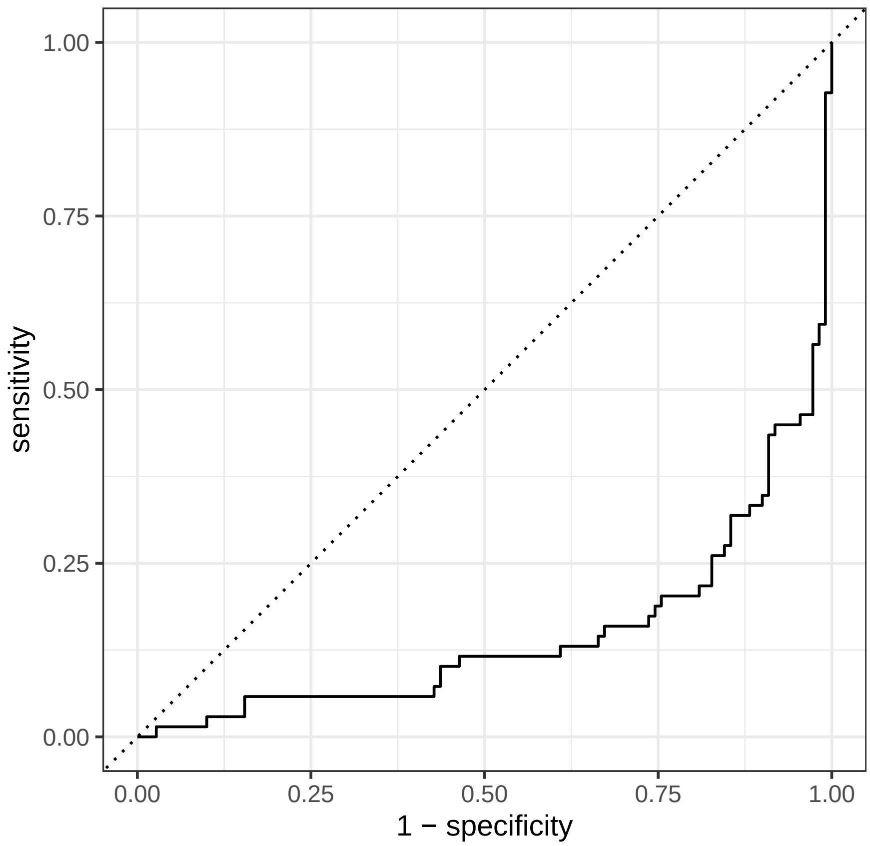



```
Add_metric <- metric_set(accuracy, sensitivity, specificity)
```

```r
augment(TitanicFit1, new_data= Titanic_test) %>%
  Add_metric(truth= survived, estimate= .pred_class)
```

```
## # A tibble: 3 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.799
## 2 sensitivity binary         0.681
## 3 specificity binary         0.873
```

```r
augment(TitanicFit1, new_data= Titanic_test) %>%
  roc_curve(survived, .pred_No) %>%
  autoplot()
```



The ROC curve is below the random classifier. Bad model for classification of variable for no survival. The area under the curve is 0.87 The training accuracy is 0.8160112. The testing accuracy is 0.7988827. Accuracy results are not significantly different and it is normal to have a higher training accuracy since the model is optimized to train data.