Classification

Joshua Durana

2022-09-25

Logistic Regression

Logistic regression calculates the probability of an instance being a certain classification. It uses the log odds from the parameters and calculates whether it’s a positive or negative class. The algorithm is not intensive to run and gives you a probabilistic output. But, similar to linear regression it’s prone to underfit.

Load Data and Set Factors

```
#Load Data
airplaneData <- read.csv("Data/airplaneData.csv", header = TRUE)

#Convert Columns in to Factors
cols <- c("Inflight.wifi.service", "Departure.Arrival.time.convenient", "Ease.of.Online.booki
ng", "Food.and.drink", "Online.boarding", "Seat.comfort", "Inflight.entertainment", "On.boar
d.service", "Leg.room.service", "Baggage.handling", "Checkin.service", "Inflight.service", "C
leanliness", "satisfaction")
airplaneData[cols] <- lapply(airplaneData[cols], as.factor)

#Drop X and ID Column
airplaneData <- subset(airplaneData, select = -c(X, id, Gate.location))
```

Train and Test Sets

```
set.seed(2022)
i <- sample(1:nrow(airplaneData), .80*nrow(airplaneData), replace = FALSE)
train <- airplaneData[i,]
test <- airplaneData[-i,]
```

####Data Exploration

```
#Show the first 6 rows of the data frame
head(train)
```

```
##        Gender     Customer.Type Age  Type.of.Travel    Class Flight.Distance
## 101175 Female disloyal Customer  35 Business travel Eco Plus             590
## 41668  Female    Loyal Customer  71 Business travel Business             918
## 68287    Male disloyal Customer  20 Business travel Business             312
## 10473  Female disloyal Customer  41 Business travel Business            1310
## 99576    Male    Loyal Customer  47 Business travel Eco Plus             199
## 8029     Male    Loyal Customer  41 Business travel Business            3224
##        Inflight.wifi.service Departure.Arrival.time.convenient
## 101175                     2                                 2
## 41668                      4                                 5
## 68287                      4                                 4
## 10473                      3                                 3
## 99576                      4                                 2
## 8029                       5                                 5
##        Ease.of.Online.booking Food.and.drink Online.boarding Seat.comfort
## 101175                      2              3               2            1
## 41668                       5              4               4            3
## 68287                       4              3               4            3
## 10473                       3              2               3            2
## 99576                       2              4               4            4
## 8029                        3              5               4            4
##        Inflight.entertainment On.board.service Leg.room.service
## 101175                      3                1                1
## 41668                       4                4                4
## 68287                       3                5                5
## 10473                       2                3                3
## 99576                       4                1                3
## 8029                        5                5                5
##        Baggage.handling Checkin.service Inflight.service Cleanliness
## 101175                4               4                4           3
## 41668                 4               4                4           1
## 68287                 5               4                5           3
## 10473                 4               3                4           2
## 99576                 2               2                5           4
## 8029                  5               2                5           5
##        Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
## 101175                         19                       19
## 41668                           0                        0
## 68287                          37                       39
## 10473                           0                        0
## 99576                          10                       12
## 8029                            0                        0
##                   satisfaction
## 101175 neutral or dissatisfied
## 41668  neutral or dissatisfied
## 68287                satisfied
## 10473                satisfied
## 99576                satisfied
## 8029                 satisfied
```

```
#Output the name of all the columns
names(train)
```

```
##  [1] "Gender"                            "Customer.Type"                    
##  [3] "Age"                               "Type.of.Travel"                   
##  [5] "Class"                             "Flight.Distance"                  
##  [7] "Inflight.wifi.service"             "Departure.Arrival.time.convenient"
##  [9] "Ease.of.Online.booking"            "Food.and.drink"                   
## [11] "Online.boarding"                   "Seat.comfort"                     
## [13] "Inflight.entertainment"            "On.board.service"                 
## [15] "Leg.room.service"                  "Baggage.handling"                 
## [17] "Checkin.service"                   "Inflight.service"                 
## [19] "Cleanliness"                       "Departure.Delay.in.Minutes"       
## [21] "Arrival.Delay.in.Minutes"          "satisfaction"
```

```
#Get information on each row
str(train)
```

```
## 'data.frame':    83123 obs. of  22 variables:
##  $ Gender                           : chr  "Female" "Female" "Male" "Female" ...
##  $ Customer.Type                    : chr  "disloyal Customer" "Loyal Customer" "disloyal
Customer" "disloyal Customer" ...
##  $ Age                              : int  35 71 20 41 47 41 58 58 29 41 ...
##  $ Type.of.Travel                   : chr  "Business travel" "Business travel" "Business t
ravel" "Business travel" ...
##  $ Class                            : chr  "Eco Plus" "Business" "Business" "Business" ...
##  $ Flight.Distance                  : int  590 918 312 1310 199 3224 577 239 328 919 ...
##  $ Inflight.wifi.service            : Factor w/ 6 levels "0","1","2","3",..: 3 5 5 4 5 6 3
6 3 5 ...
##  $ Departure.Arrival.time.convenient: Factor w/ 6 levels "0","1","2","3",..: 3 6 5 4 3 6 3
6 6 5 ...
##  $ Ease.of.Online.booking           : Factor w/ 6 levels "0","1","2","3",..: 3 6 5 4 3 4 3
6 3 4 ...
##  $ Food.and.drink                   : Factor w/ 6 levels "0","1","2","3",..: 4 5 4 3 5 6 3
6 6 4 ...
##  $ Online.boarding                  : Factor w/ 6 levels "0","1","2","3",..: 3 5 5 4 5 5 5
6 3 6 ...
##  $ Seat.comfort                     : Factor w/ 6 levels "0","1","2","3",..: 2 4 4 3 5 5 6
6 6 6 ...
##  $ Inflight.entertainment           : Factor w/ 6 levels "0","1","2","3",..: 4 5 4 3 5 6 5
6 6 5 ...
##  $ On.board.service                 : Factor w/ 6 levels "0","1","2","3",..: 2 5 6 4 2 6 5
3 6 5 ...
##  $ Leg.room.service                 : Factor w/ 6 levels "0","1","2","3",..: 2 5 6 4 4 6 5
4 6 5 ...
##  $ Baggage.handling                 : Factor w/ 5 levels "1","2","3","4",..: 4 4 5 4 2 5 4
1 5 4 ...
##  $ Checkin.service                  : Factor w/ 6 levels "0","1","2","3",..: 5 5 5 4 3 3 5
3 6 5 ...
##  $ Inflight.service                 : Factor w/ 6 levels "0","1","2","3",..: 5 5 6 5 6 6 5
5 5 5 ...
##  $ Cleanliness                      : Factor w/ 6 levels "0","1","2","3",..: 4 2 4 3 5 6 6
6 6 6 ...
##  $ Departure.Delay.in.Minutes       : int  19 0 37 0 10 0 0 15 0 0 ...
##  $ Arrival.Delay.in.Minutes         : num  19 0 39 0 12 0 0 24 0 19 ...
##  $ satisfaction                     : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1
2 2 2 2 2 2 1 2 ...
```

```
#Get the dimensions of the data frame
dim(train)
```

```
## [1] 83123    22
```

```
#Get the summary of each column
summary(train)
```

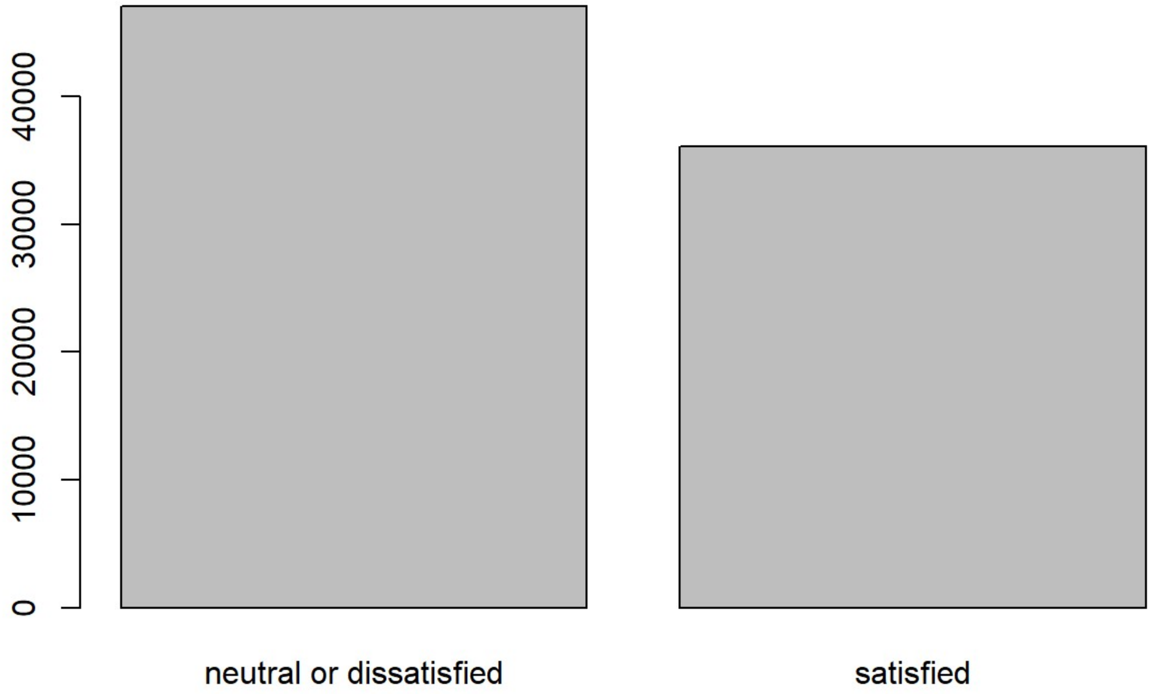
```
##     Gender          Customer.Type           Age        Type.of.Travel
##  Length:83123       Length:83123       Min.   : 7.00   Length:83123
##  Class :character   Class :character   1st Qu.:27.00   Class :character
##  Mode  :character   Mode  :character   Median :40.00   Mode  :character
##                                        Mean   :39.41
##                                        3rd Qu.:51.00
##                                        Max.   :85.00
##
##     Class           Flight.Distance Inflight.wifi.service
##  Length:83123       Min.   :  31    0: 2488
##  Class :character   1st Qu.: 414    1:14239
##  Mode  :character   Median : 842    2:20675
##                     Mean   :1188    3:20692
##                     3rd Qu.:1739    4:15872
##                     Max.   :4983    5: 9157
##
##  Departure.Arrival.time.convenient Ease.of.Online.booking Food.and.drink
##  0: 4262                           0: 3605                0:   85
##  1:12346                           1:14043                1:10204
##  2:13898                           2:19242                2:17620
##  3:14360                           3:19565                3:17830
##  4:20384                           4:15655                4:19514
##  5:17873                           5:11013                5:17870
##
##  Online.boarding Seat.comfort Inflight.entertainment On.board.service
##  0: 1939         0:    1      0:   12                0:    3
##  1: 8550         1: 9603      1: 9911                1: 9477
##  2:14103         2:11988      2:14111                2:11853
##  3:17326         3:14953      3:15282                3:18191
##  4:24667         4:25437      4:23583                4:24659
##  5:16538         5:21141      5:20224                5:18940
##
##  Leg.room.service Baggage.handling Checkin.service Inflight.service Cleanliness
##  0:  380          1: 5801          0:    1         0:    3          0:   10
##  1: 8254          2: 9192          1:10323         1: 5661          1:10610
##  2:15647          3:16495          2:10265         2: 9166          2:12949
##  3:16066          4:29958          3:22788         3:16236          3:19607
##  4:23032          5:21677          4:23273         4:30403          4:21764
##  5:19744                           5:16473         5:21654          5:18183
##
##  Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
##  Min.   :   0.0             Min.   :   0.00
##  1st Qu.:   0.0             1st Qu.:   0.00
##  Median :   0.0             Median :   0.00
##  Mean   :  14.8             Mean   :  15.19
##  3rd Qu.:  12.0             3rd Qu.:  13.00
##  Max.   :1592.0             Max.   :1584.00
##                             NA's   :251
##                    satisfaction
##  neutral or dissatisfied:47065
##  satisfied              :36058
```

```
##
##
##
##
##
```

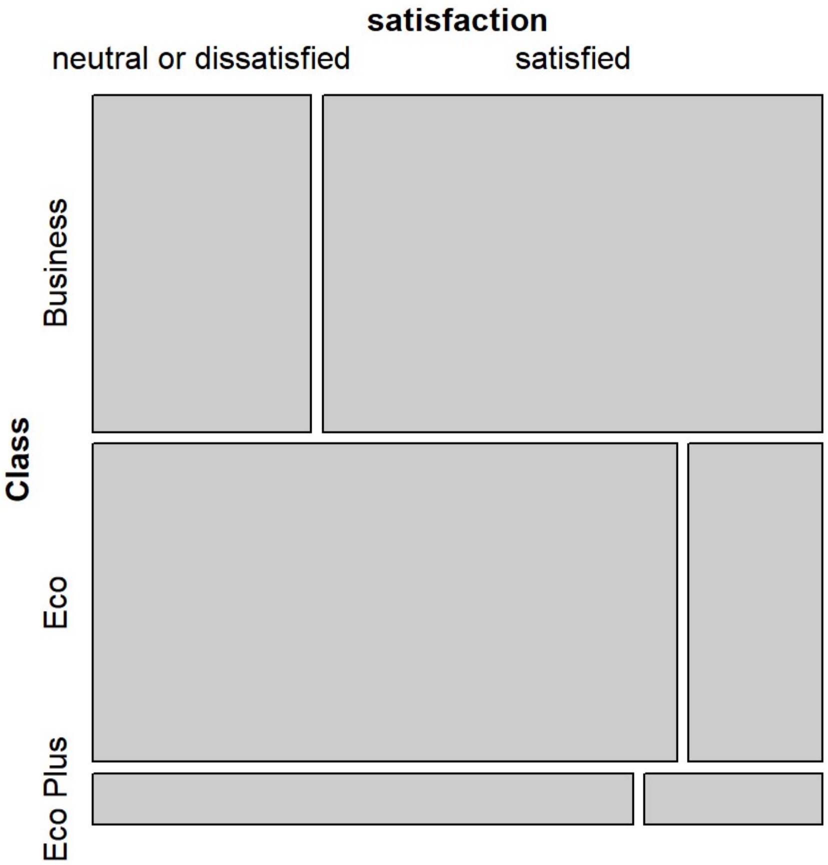
```
#Histogram of Satisfaction
barplot(table(train$satisfaction))

#Mosaic of Class and Satisfaction
library(vcd)
```

```
## Loading required package: grid
```



```
mosaic(table(train[,c(5,22)]))
```



Logistic Regression

```
airplaneLog <- glm(satisfaction~Customer.Type, data=train, family=binomial)
summary(airplaneLog)
```

```
##
## Call:
## glm(formula = satisfaction ~ Customer.Type, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1403  -1.1403  -0.7334   1.2150   1.6999
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.17590    0.01912  -61.51   <2e-16 ***
## Customer.TypeLoyal Customer  1.08800    0.02060   52.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113771  on 83122  degrees of freedom
## Residual deviance: 110639  on 83121  degrees of freedom
## AIC: 110643
##
## Number of Fisher Scoring iterations: 4
```

The deviance residuals quantifies a given point’s contribution to the overall likelihood. It seems good since the quartiles are symmetric and the median is close to 0. The null deviance measures the lack of fit of the model with only the intercept. The residual measures the lack of fit of the model of the entire model. We want the residual deviance to be much smaller than the null deviance, which is the case with our model. The Akaike Information Criterion (AIC) are used to compare between models and lower is the better. The Fisher Scoring iterations tells us how many times the glm function iterated to the maximum likelihood estimates for the coefficients.

Naive Bays

```
library(e1071)
airplaneNB <- naiveBayes(satisfaction~., data=train)
airplaneNB
```

```
## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
## 
## A-priori probabilities:
## Y
## neutral or dissatisfied               satisfied 
##               0.5662091               0.4337909 
## 
## Conditional probabilities:
##                          Gender
## Y                            Female      Male
##   neutral or dissatisfied 0.5112929 0.4887071
##   satisfied               0.5008320 0.4991680
## 
##                          Customer.Type
## Y                         disloyal Customer Loyal Customer
##   neutral or dissatisfied        0.24653139     0.75346861
##   satisfied                      0.09928449     0.90071551
## 
##                          Age
## Y                             [,1]     [,2]
##   neutral or dissatisfied 37.58317 16.46885
##   satisfied               41.79963 12.77713
## 
##                          Type.of.Travel
## Y                         Business travel Personal Travel
##   neutral or dissatisfied       0.5073834       0.4926166
##   satisfied                     0.9262577       0.0737423
## 
##                          Class
## Y                           Business        Eco   Eco Plus
##   neutral or dissatisfied 0.25630511 0.64702008 0.09667481
##   satisfied               0.76482334 0.19354928 0.04162738
## 
##                          Flight.Distance
## Y                              [,1]      [,2]
##   neutral or dissatisfied  928.3673  789.1561
##   satisfied               1526.5854 1127.6785
## 
##                          Inflight.wifi.service
## Y                                    0            1            2            3
##   neutral or dissatisfied 0.0001487305 0.2037607564 0.3297779666 0.3297142250
##   satisfied               0.0688058129 0.1289311665 0.1429363803 0.1434910422
##                          Inflight.wifi.service
## Y                                    4            5
##   neutral or dissatisfied 0.1349410390 0.0016572825
##   satisfied               0.2640468135 0.2517887847
## 
```

```
##                          Departure.Arrival.time.convenient
## Y                                  0          1          2          3
##   neutral or dissatisfied 0.04738128 0.13415489 0.16360353 0.17140125
##   satisfied               0.05635365 0.16728604 0.17188973 0.17452438
##                          Departure.Arrival.time.convenient
## Y                                  4          5
##   neutral or dissatisfied 0.26508021 0.21837884
##   satisfied               0.21931333 0.21063287
## 
##                          Ease.of.Online.booking
## Y                                  0          1          2          3
##   neutral or dissatisfied 0.02592160 0.18570063 0.28477637 0.28724105
##   satisfied               0.06614344 0.14706861 0.16193355 0.16767430
##                          Ease.of.Online.booking
## Y                                  4          5
##   neutral or dissatisfied 0.15559333 0.06076702
##   satisfied               0.23107216 0.22610794
## 
##                          Food.and.drink
## Y                                    0            1            2            3
##   neutral or dissatisfied 0.0010411134 0.1737596940 0.2282800382 0.2292149155
##   satisfied               0.0009983915 0.0561872539 0.1906927728 0.1952964668
##                          Food.and.drink
## Y                                    4            5
##   neutral or dissatisfied 0.1964304685 0.1712737703
##   satisfied               0.2847911698 0.2720339453
## 
##                          Online.boarding
## Y                                  0          1          2          3
##   neutral or dissatisfied 0.01835759 0.15610326 0.26454903 0.31872942
##   satisfied               0.02981308 0.03336292 0.04581508 0.06447945
##                          Online.boarding
## Y                                  4          5
##   neutral or dissatisfied 0.19721662 0.04504409
##   satisfied               0.42667369 0.39985579
## 
##                          Seat.comfort
## Y                                    0            1            2            3
##   neutral or dissatisfied 2.124721e-05 1.586317e-01 1.974078e-01 2.495910e-01
##   satisfied               0.000000e+00 5.926563e-02 7.479616e-02 8.891231e-02
##                          Seat.comfort
## Y                                    4            5
##   neutral or dissatisfied 2.367789e-01 1.575693e-01
##   satisfied               3.963892e-01 3.806368e-01
## 
##                          Inflight.entertainment
## Y                                    0            1            2            3
##   neutral or dissatisfied 0.0002549665 0.1813874429 0.2355465845 0.2371826198
##   satisfied               0.0000000000 0.0381052748 0.0838926174 0.1142326252
##                          Inflight.entertainment
## Y                                    4            5
##   neutral or dissatisfied 0.1936045894 0.1520237969
##   satisfied               0.4013256420 0.3624438405
## 
##                           On.board.service
## Y                                     0            1            2            3
##   neutral or dissatisfied 6.374163e-05 1.620737e-01 1.869755e-01 2.645278e-01
##   satisfied               0.000000e+00 5.127850e-02 8.466914e-02 1.592157e-01
##                           On.board.service
## Y                                     4            5
##   neutral or dissatisfied 2.420482e-01 1.443111e-01
##   satisfied               3.679350e-01 3.369017e-01
## 
##                           Leg.room.service
## Y                                    0           1           2           3
##   neutral or dissatisfied 0.005205567 0.139827898 0.240582174 0.248932328
##   satisfied               0.003743968 0.046397471 0.119917910 0.120638971
##                           Leg.room.service
## Y                                    4           5
##   neutral or dissatisfied 0.204015723 0.161436311
##   satisfied               0.372455488 0.336846192
## 
##                           Baggage.handling
## Y                                   1          2          3          4
##   neutral or dissatisfied 0.08588123 0.13746946 0.26845851 0.32977797
##   satisfied               0.04878252 0.07548949 0.10704975 0.40038272
##                           Baggage.handling
## Y                                   5
##   neutral or dissatisfied 0.17841283
##   satisfied               0.36829552
## 
##                           Checkin.service
## Y                                     0            1            2            3
##   neutral or dissatisfied 2.124721e-05 1.664082e-01 1.630086e-01 2.672687e-01
##   satisfied               0.000000e+00 6.908314e-02 7.191192e-02 2.831272e-01
##                           Checkin.service
## Y                                     4            5
##   neutral or dissatisfied 2.668437e-01 1.364496e-01
##   satisfied               2.971324e-01 2.787454e-01
## 
##                           Inflight.service
## Y                                     0            1            2            3
##   neutral or dissatisfied 6.374163e-05 8.507383e-02 1.357059e-01 2.625730e-01
##   satisfied               0.000000e+00 4.595374e-02 7.707028e-02 1.075489e-01
##                           Inflight.service
## Y                                     4            5
##   neutral or dissatisfied 3.372357e-01 1.793477e-01
##   satisfied               4.029896e-01 3.664374e-01
## 
##                           Cleanliness
## Y                                     0            1            2            3
##   neutral or dissatisfied 0.0002124721 0.1816424094 0.2159991501 0.2359927759
##   satisfied               0.0000000000 0.0571579123 0.0771812081 0.2357313218
##                          Cleanliness
## Y                                    4            5
##   neutral or dissatisfied 0.2142568788 0.1518963136
##   satisfied               0.3239225692 0.3060069887
## 
##                          Departure.Delay.in.Minutes
## Y                             [,1]     [,2]
##   neutral or dissatisfied 16.52908 40.43984
##   satisfied               12.53242 34.53602
## 
##                          Arrival.Delay.in.Minutes
## Y                             [,1]     [,2]
##   neutral or dissatisfied 17.17240 40.74798
##   satisfied               12.60165 35.22496
```

For continuous data such as Age, it outputs the means and standard deviation for each satisfaction levels. For discrete variables, it'll output the probabilities of a certain factor being satisfied or not.

Testing

```
#Logical Regression
prob <- predict(airplaneLog, newdata=test, type="response")
pred <- ifelse(prob>.5, 2, 1)
acc <- mean(pred==as.integer(test$satisfaction))
acc
```

```
## [1] 0.5685001
```

```
#Naive Bayes
pred1 <- predict(airplaneNB, newdata=test, type="class")
table(pred1, test$satisfaction)
```

```
## 
## pred1                     neutral or dissatisfied satisfied
##   neutral or dissatisfied                   10619      1070
##   satisfied                                  1195      7897
```

```
mean(pred1==test$satisfaction)
```

```
## [1] 0.8910062
```

The accuracy on logistic regression isn't bad, but I feel it could be better if I chose different predictors. The accuracy on the Naive Bayes model is much better than logistic regression, most likely due to

For logistic regression works well larger data sets and runs faster than other algorithms. But, logistic regression has a high bias that causes it to underfit Naive Bayes works well with smaller data sets and can work with multiple dimensions better than logistic regression. But, if the predictors are not independent it hurts the algorithm’s performance.

Accuracy is the most common metric to evaluate results in classification, it gives you the percentage of correct predictions to the number of observations. But it doesn’t take to account false or true positives.

Sensitivity measures the ratio of accurate classifications from all of the true predictions. This means that if the model predicts something to be true, the sensitivity measures if the model is correct. Specificity is similar to sensitivity, but for false predictions. Both, don’t give you a full picture of the accuracy of the model.

Kappa is similar to accuracy, but adjusts to account for the chance of a correct prediction. One drawback is there’s not a universally agreed way to interpret Kappa.

The ROC curve shows us the how the false and true positive rates interact with each other. The AUC is the area under the ROC curve and helps us compare other ROC curves and ranges from 0 to 1.