Playoff Classification

Travis Ho

2023-10-27

Setup

```
#Load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Data

```
#Load NFL stats dataset
nfl <- read.csv("./data/NFL_data.csv")
```

This dataset was obtained from the Kaggle dataset called “NFL 2010 to 2022 Stats to predict 2023 Winner” This is a dataset containing team statistics from all NFL teams from 2010 to 2022. Stats include but are not limited to point differential, win percentage, and score percentage. It contains 384 teams with 19 different statistical variables.

Research Question

In the NFL, playoff teams all look a little different. Some teams dominate the season to make the playoffs and some teams barely squeak in. However there is a chance that it is possible to classify a playoff team based on some statistics. Statistics such as point differential, red zone percentage, score percentage, and turnover percentage could all be telling if a team made the playoffs or not. Based on the different characteristics of teams, can we correctly classify whether the team is a playoff team or is not a playoff team?

- H0: The classification of a team as a playoff team or not a playoff team is not possible with these variables.
- HA: The classification of a team as a playoff team or not a playoff team is possible with at least one of these variables.

Variables of Interest

The dependent variable in this study is playoff which is whether a team made the playoffs or not. This is a string and binary variable that is either yes (Y) the team made the playoffs or no (N) they did not make the playoffs.

The independent variables are point differential, red zone percentage, score percentage, and turnover percentage. Point differential is how many more points total a team scored over the season than they gave up. This is a continuous variable that is all integers. Red zone percentage is the percent of times a team scored when they got into the redzone (20 yards or closer to goal). This is a continuous variable on the range 0-100 and is numeric. Score percentage is the percent of times a team scored when they got the ball. This is a continuous variable on the range 0-100 and is numeric. Turnover percentage is the percent of times a team turned over the ball when they had possession. This is a continuous variable on the range 0-100 and is numeric.

Data Wrangling

```r
#Create a new dataframe with only the variables we need
stats <- nfl %>%
         select(Playoff,
                PD,
                RedZone_perc,
                Score_perc,
                Turnover_perc)
```

```r
#Change Yes playoffs to 1 and No playoffs to 0
stats$Playoff <- factor(ifelse(stats$Playoff == "Y", 1, 0))
```

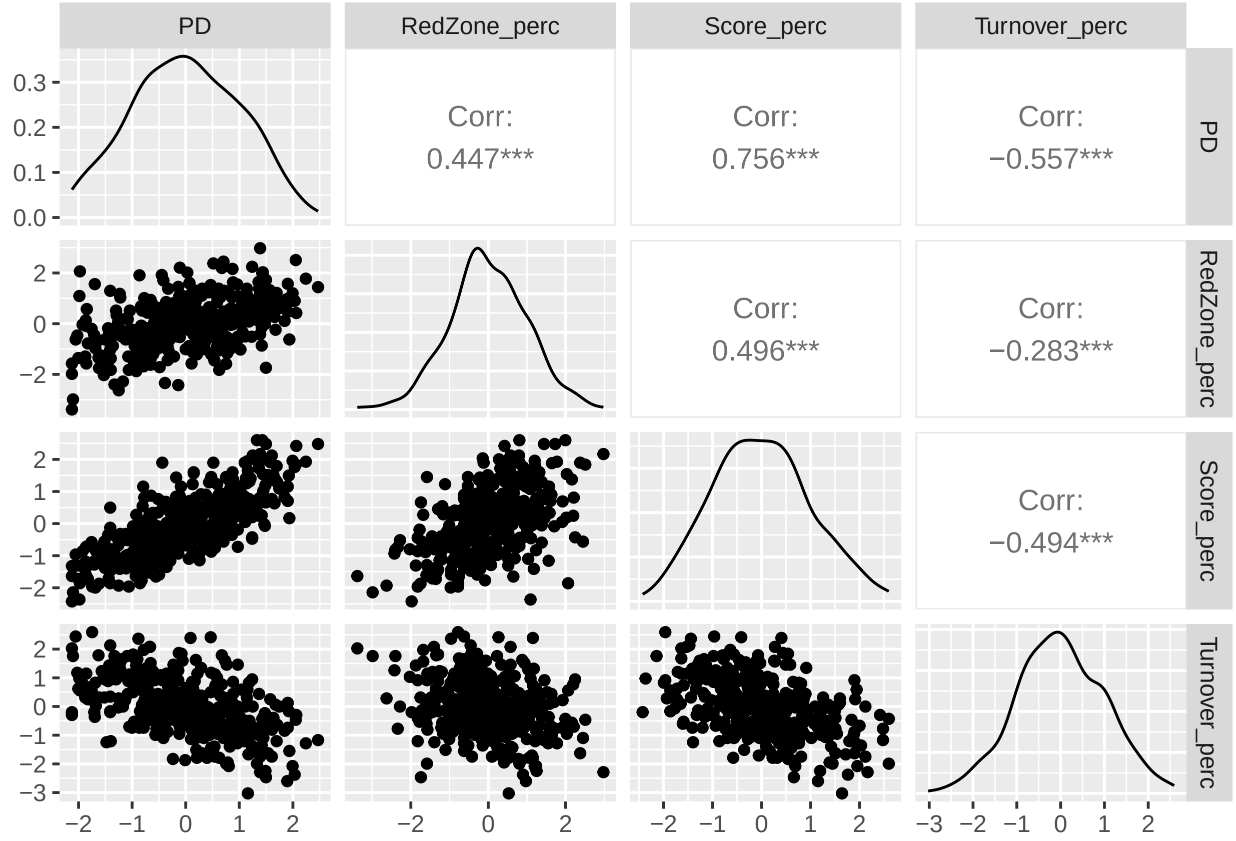
```r
#Check frequencies of Playoffs
table(stats$Playoff)
```

```
## 
##   0   1 
## 236 148
```

```r
#Normalize the continuous numerical variables
library(bestNormalize)
set.seed(123)
selected_columns <- c("PD","RedZone_perc","Score_perc","Turnover_perc")
stats_normal <- lapply(
  stats[selected_columns],
  function(x){
    bestNormalize(x)$x.t
  }
)
#Convert it to a dataframe and add playoff column to it
stats_normal <- as.data.frame(stats_normal)
stats_normal$Playoff <- stats$Playoff
```

#Analysis and Data Visualization

```r
stats_normal %>%
  select(-Playoff) %>%
  ggpairs()
```



There is a pretty strong positive correlation between point differential and scoring percentage (0.756. There is a somewhat strong negative correlation between point differential and turnover percentage (-0.557).

```r
#Fit to logistic model
log_nfl <- glm(Playoff ~ .,data = stats_normal, family = "binomial")
```

```
summary(log_nfl)
```

```
##
## Call:
## glm(formula = Playoff ~ ., family = "binomial", data = stats_normal)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.0797     0.1840  -5.868 4.42e-09 ***
## PD              2.6077     0.3369   7.739 9.99e-15 ***
## RedZone_perc   -0.2027     0.1925  -1.053   0.2924
## Score_perc      0.2871     0.2605   1.102   0.2704
## Turnover_perc  -0.5954     0.1984  -3.000   0.0027 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.99  on 383  degrees of freedom
## Residual deviance: 243.03  on 379  degrees of freedom
## AIC: 253.03
##
## Number of Fisher Scoring iterations: 6
```

Both Redzone percentage and Score percentage are not significant so we will need to remove them. First lets check multicollinearity though.

```
#Check VIF
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(log_nfl)
```

```
##            PD  RedZone_perc    Score_perc Turnover_perc
##      1.215920      1.169383      1.324646      1.038544
```

All VIF values look like they are all pretty low so multicollinearity should not be an issue.

```
#Remove Non-significant Predictors, Remove RedZone_perc first
log_nfl2 <- glm(Playoff ~ PD + Score_perc + Turnover_perc,data = stats_normal, family = "binomial")

summary(log_nfl2)
```

```
##
## Call:
## glm(formula = Playoff ~ PD + Score_perc + Turnover_perc, family = "binomial",
##     data = stats_normal)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.0839     0.1839  -5.894 3.77e-09 ***
## PD              2.5798     0.3372   7.651 1.99e-14 ***
## Score_perc      0.2073     0.2505   0.827  0.40797
## Turnover_perc  -0.5889     0.1984  -2.968  0.00299 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.99  on 383  degrees of freedom
## Residual deviance: 244.14  on 380  degrees of freedom
## AIC: 252.14
##
## Number of Fisher Scoring iterations: 6
```

Score_perc is still not significant so we need to remove it.

```
#Remove Non-significant Predictors, Remove Score_perc
log_nfl3 <- glm(Playoff ~ PD + Turnover_perc,data = stats_normal, family = "binomial")

summary(log_nfl3)
```

```
##
## Call:
## glm(formula = Playoff ~ PD + Turnover_perc, family = "binomial",
##     data = stats_normal)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.0856     0.1840  -5.900 3.64e-09 ***
## PD              2.6992     0.3080   8.764  < 2e-16 ***
## Turnover_perc  -0.6125     0.1967  -3.114  0.00185 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 511.99  on 383  degrees of freedom
## Residual deviance: 244.82  on 381  degrees of freedom
## AIC: 250.82
```

```
##
## Number of Fisher Scoring iterations: 6
```

All predictors are significant now.

```
#Check VIF of variables again
vif(log_nfl3)
```

```
##            PD Turnover_perc
##      1.020141      1.020141
```

Looks good all values are very low so multicollinearity is not a problem.

```
coefficients <- coef(log_nfl3)
odds_ratios <- exp(coefficients)

odds_ratios
```

```
##   (Intercept)            PD Turnover_perc
##     0.3377091    14.8683320     0.5419777
```

```
#Obtain the probabilities
probs <- predict(log_nfl3,
                 type = "response")

classes <- factor(
  ifelse(
    probs > 0.50,1,0
  )
)
```

```
library(caret)
```

```
## Loading required package: lattice
```

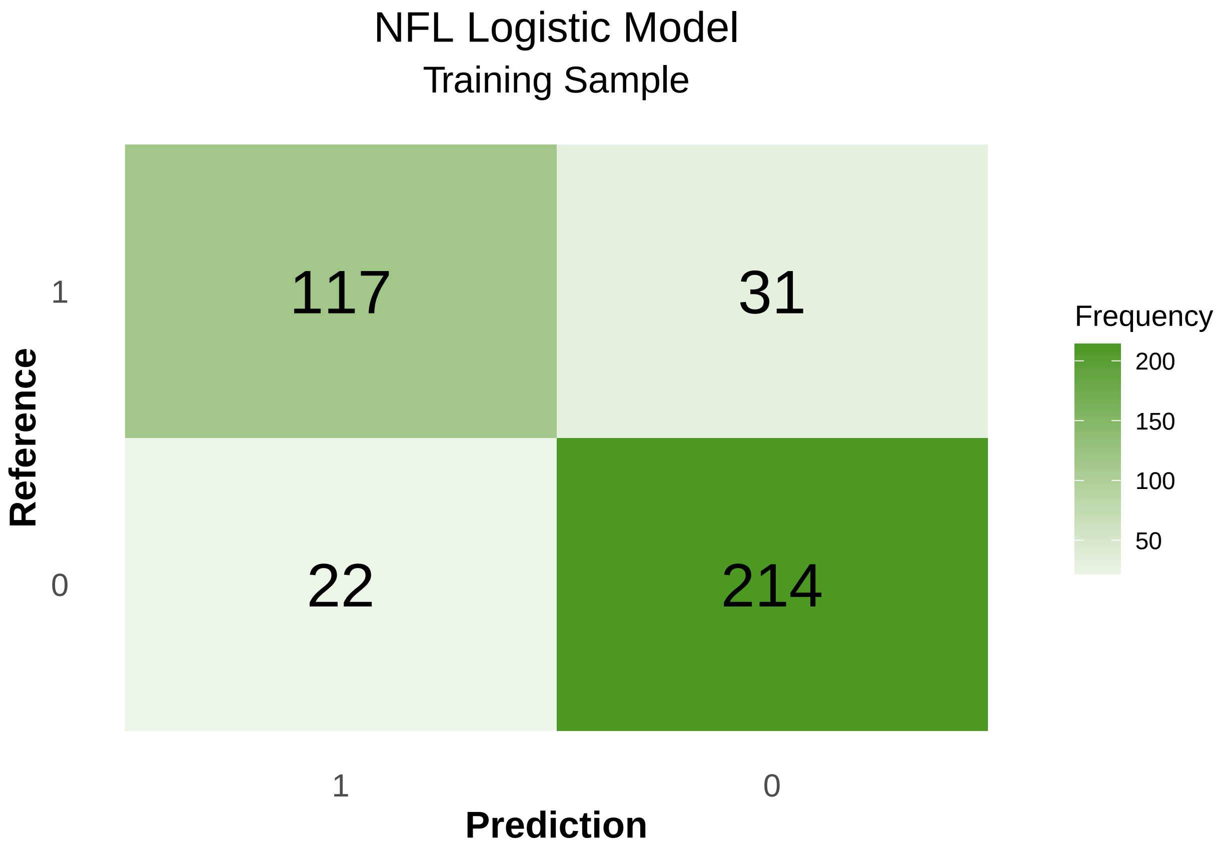
```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

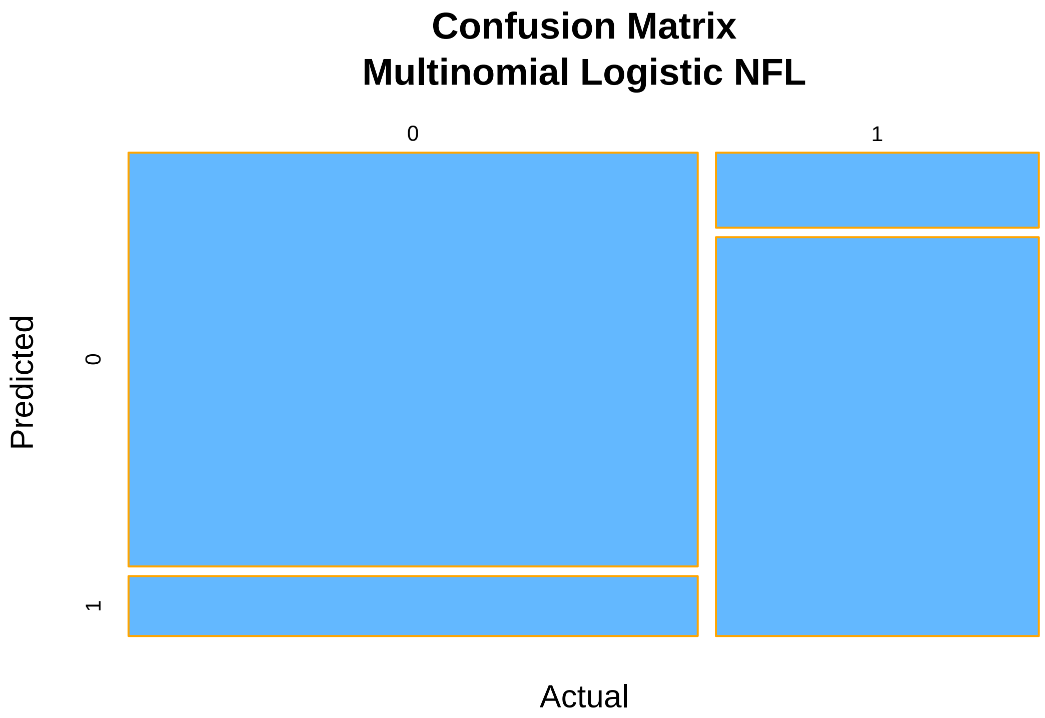
```
train_confusion <- confusionMatrix(
  data = classes,
  reference = stats_normal$Playoff,
  positive = "1"
);train_confusion
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 214  31
##          1  22 117
##
##                Accuracy : 0.862
##                  95% CI : (0.8234, 0.8949)
##     No Information Rate : 0.6146
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7053
##
##  Mcnemar's Test P-Value : 0.2718
##
##             Sensitivity : 0.7905
##             Specificity : 0.9068
##          Pos Pred Value : 0.8417
##          Neg Pred Value : 0.8735
##              Prevalence : 0.3854
##          Detection Rate : 0.3047
##    Detection Prevalence : 0.3620
##       Balanced Accuracy : 0.8487
##
##        'Positive' Class : 1
##
```

```
# Set up for visualization
train_df <- as.data.frame(train_confusion$table)
train_df$Prediction <- factor(
  train_df$Prediction, levels = rev(levels(train_df$Prediction))
)
# Visualize confusion matrix
train_df %>%
  ggplot(
    aes(x = Prediction, y = Reference, fill = Freq)
) +
  geom_tile() +
  geom_text(aes(label = Freq), size = 8) +
  scale_fill_gradient2(
    low = "white", high = "#4B9823",
    name = "Frequency"
) +
  labs(
    title = "NFL Logistic Model",
    subtitle = "Training Sample"
) +
theme(
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  axis.text = element_text(size = 12),
  axis.title = element_text(size = 14, face = "bold"),
  plot.title = element_text(size = 16, hjust = 0.5),
  plot.subtitle = element_text(size = 14, hjust = 0.5)
)
```



```
mosaicplot(
  table(
    classes,
    stats_normal$Playoff
  ),
  main ="Confusion Matrix\nMultinomial Logistic NFL",
  xlab = "Actual",
  ylab = "Predicted",
  color = "steelblue1",
  border = "orange"
)
```



```
library(rms)
```

```
## Loading required package: Hmisc
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'rms'
```

```
## The following objects are masked from 'package:car':
##
##     Predict, vif
```

```
lrm_nfl <- lrm(
  formula = Playoff ~ PD + Turnover_perc,
  data = stats_normal
)
lrm_nfl
```

```
## Logistic Regression Model
##
## lrm(formula = Playoff ~ PD + Turnover_perc, data = stats_normal)
##
##                        Model Likelihood      Discrimination    Rank Discrim.
##                              Ratio Test             Indexes          Indexes
## Obs            384    LR chi2     267.17      R2       0.681    C       0.932
##  0             236    d.f.             2      R2(2,384)0.499    Dxy     0.863
##  1             148    Pr(> chi2) <0.0001    R2(2,272.9)0.622    gamma   0.863
## max |deriv| 2e-08                             Brier    0.100    tau-a   0.410
##
##               Coef    S.E.   Wald Z Pr(>|Z|)
## Intercept     -1.0856 0.1840 -5.90  <0.0001
## PD             2.6992 0.3080  8.76  <0.0001
## Turnover_perc -0.6125 0.1967 -3.11  0.0018
```

Discussion

The initial research question that we initially asked was whether the characteristics of point differential, redzone score percentage, score percentage, or turnover percentage could correctly classify whether a team was a playoff team or not. In the initial logistic model that was created (lrm_nfl) it was displayed that two of the predictors (redzone score percentage and score percentage) were not significant variables. As a result, a new logistic model (lrm_nfl2) was created which displayed that score percentage was still not significant. Consequently, another new model (lrm_nfl3) was generated which displayed point differential and turnover percentage as significant predictors. After creating this model, multicollinearity had to be checked if it was an issue. The vif of point differential and turnover_perc were both 1.02 which indicates that multicollinearity is not an issue.

After all assumptions were met, the odds ratios were obtained. When controlling for turnover percentage, each point differential increase leads to 14.87 times higher odds that the team is a playoff team. When controlling for point differential, each point increase in turnover percentage leads to 1.46 times lower odds (OR=0.54) that the team is a playoff team. It should be noted however that because the data was normalized, the point increases are based on the normalized scale rather than the true scale of the predictors.

Finally, in evaluating the classification of our model, a confusion matrix was produced. The confusion matrix indicated an 0.862 accuracy rate which means that the model correctly classified 86.2% of the data points based on the two predictors. The sensitivity was 0.7905 which means that 79.05 of the positives (playoff teams) were correctly classified. The Specificity is 0.9068 which indicates that 90.68% of the negatives (non-playoff teams) were correctly classified. The kappa value was 0.7053 which indicates almost perfect agreement. Beyond the calculations, the visual of the confusion matrix as well as the mosaic plot both display the overwhelming amount of true playoff teams and true non-playoff teams as opposed to false playoff teams and false non-playoff teams Lastly, in the lrm model, chi squared being significant indicates that the prediction is better than chance. Also the r^2 value is 0.681 indicating that 68.1% of the variance in the classification can be explained by point differential and turnover percentage. The C/AUC is 0.932 which means that the model is a very strong model.

As a result of all these factors, we can conclude that of the four predictors, point differential and turnover percentage can correctly classify whether a team was a playoff team or not.