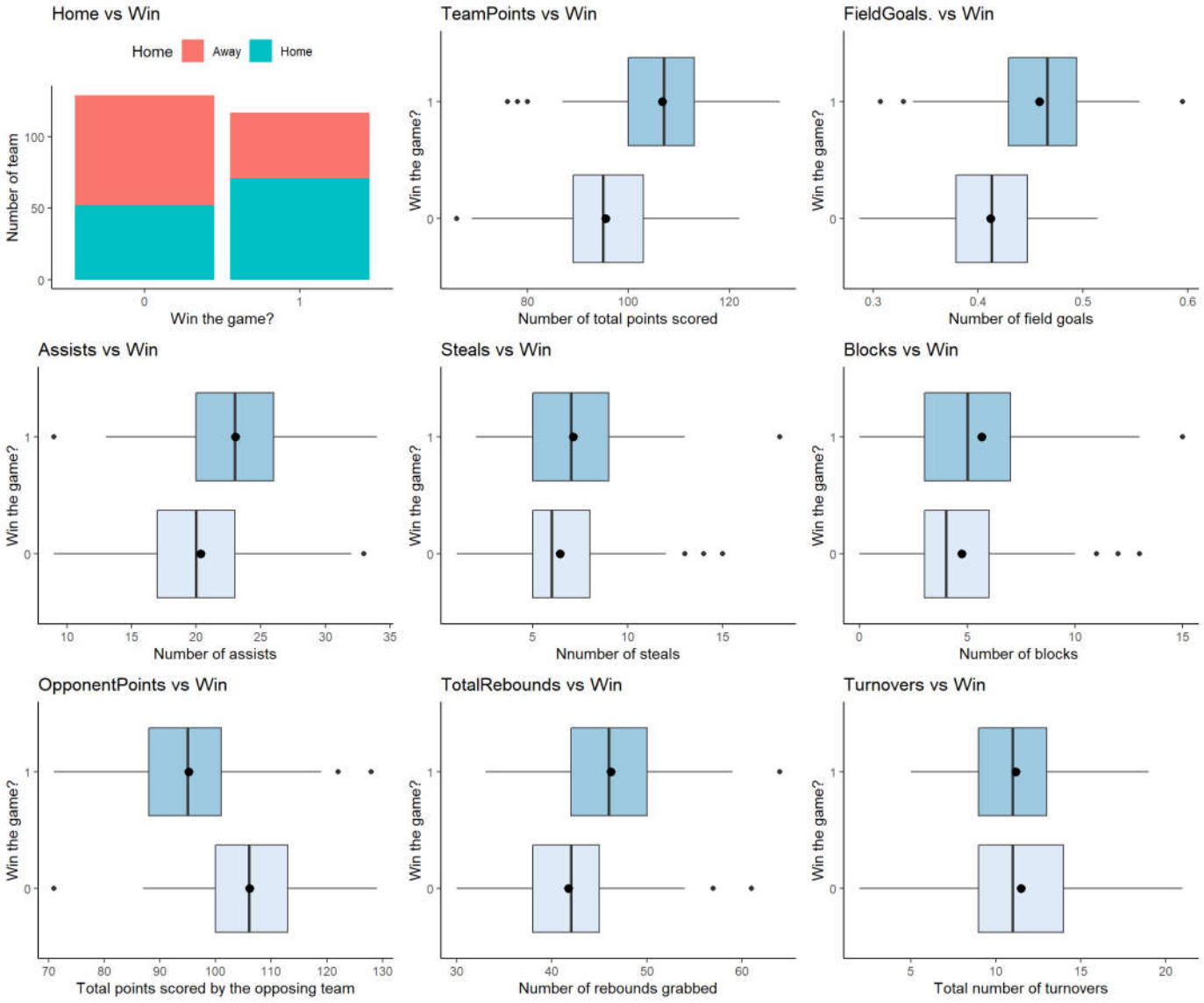
Data Analysis Assignment 3

Yuanjing Zhu

10/27/2022

Q1: Create plots to explore the relationships between Win and the following variables: Home, TeamPoints, FieldGoals. (with a period!), Assists, Steals, Blocks, OpponentPoints, TotalRebounds, and Turnovers.



For nba games by Charlotte Hornets teams before 2017-10-01:
From the bar plot of Home vs Win, we can see that the number of home team winning the game is larger than a visiting team and the number of away team losing is larger than a home team, which means that an nba team is more likely to win the game when it is the home team. From the boxplot of TeamPoints vs Win, we can see that a team scoring more points has higher chance to win. If a team scores more than 105 points, it has higher chance to win but if it scores less than 95 points, it is more likely to lose. Similar for FieldGoals.(with period), when a team has more field goals made in the game versus field goals attempted in the game, the frequency of winning is larger. The box plot of Assists and Win tells us that more assists made in the game leads to higher chance of winning the game, which maks sense since by definition assist means passing leading to a successful field goal. For steals and blocks, if a team has larger number of steals and blocks, it is more likely to win, but the difference is not very large. For OppoentPoints, it shows an opposite trend from TeamPoints where more points opponent team made,

more likely the team would lose. For total number of rebounds vs win, a team is more likely to win it it has larger number of rebounds. Total number of turnovers does not seem to have an impact on the odds of winning the game from the last box plot.

Q2: Identify at least two pairs and briefly explain why we should not include them in the model at the same time.

1. TeamPoints and FieldGoals
2. FieldGoals and FieldGoals.(with period)
3. OffRebounds and TotalRebounds

By looking at the code book, the first pair I identified is TeamPoints and FieldGoals. This is because total points equal the sum of number of goals multiply points of each goal. The more field goals made in the game will result in higher total points scored, so they should not be included in the model simultaneously to avoid multicolinearity.

By definition, FieldGoals, FieldGoals.(with period) and FieldGoalsAttempted should not be included together because there is inherent connection between these three: $FieldGoals. = FieldGoals/FieldGoalsAttempted$. Variation in FieldGoals will affect the value of FieldGoals.(with period).

Similar as OffRebounds and TotalRebounds since TotalRebounds includes OffRebounds, i.e. the increasing in OffRebounds will result in an increase in TotalRebounds.

Q3: Fit a logistic regression model for Win using Home, TeamPoints, FieldGoals., Assists, Steals, Blocks, TotalRebounds, and Turnovers as predictors. Present the output of the fitted model and interpret the significant coefficients in terms of the odds of your team winning an NBA game.

Logistic regression Model Regressing Win

Predictor	Estimate	SE	z value	p-value
(Intercept)	-29.80	3.94	-7.56	0.00
HomeHome	1.11	0.40	2.75	0.01
TeamPoints	-0.01	0.03	-0.34	0.74
FieldGoals._percent	0.44	0.08	5.78	0.00
Assists	-0.11	0.05	-2.23	0.03
Steals	0.39	0.08	4.74	0.00
Blocks	0.04	0.07	0.60	0.55
TotalRebounds	0.28	0.04	6.22	0.00

[a] Null deviance: 340.44
[b] Residual deviance: 195.38
[c] AIC: 213.38

Predictor	Estimate	SE	z value	p-value
Turnovers	-0.17	0.06	-3.03	0.00

[a] Null deviance: 340.44
[b] Residual deviance: 195.38
[c] AIC: 213.38

95% Confidence Interval

Predictor	2.5%	97.5%
Intercept	0.00	0.00
Home	1.39	6.81
TeamPoints	0.94	1.04
FieldGoals._percent	1.35	1.82
Assists	0.81	0.99
Steals	1.27	1.76
Blocks	0.91	1.19
TotalRebounds	1.22	1.45
Turnovers	0.75	0.94

Note: the FieldGoals.is coded in the data as a decimal, I converted it to percent prior to fitting the model.

According to p-value, Home, FieldGoals.(%), Assists, Steals, TotalRebounds and Turnovers are considered significant predictors in regressing winning the game at the 0.05 level. Here is the interpretation of significant coefficients:

Odds of winning the game are 3 ($e^{1.11}$) times higher for home team compared to away team and we are 95% confident that the true odds ratio comparing home and away team lies between 1.39 and 6.81.

For every one percent increase in the ratio of number of field goals made in the game and the number of field goals attempted in the game, the odds of winning increase by a factor of 1.55 ($e^{0.44}$), and we are 95% confident that the true odds ratio lies between 1.35 and 1.82.

For every one unit increase in the number of assists, the odds of winning will decrease 10% ($e^{-0.11} = 0.9$) and we are 95% confident that the true odds ratio lies between 0.81 and 0.99.

For every one unit increase in the number of steals, the odds of winning will increase by a factor of 1.48 ($e^{0.39}$) and we are 95% confident that true odds ratio lies between 1.27 and 1.76.

For every one unit increase in total number of rebound, the odds of winning will increase by a factor of 1.32 ($e^{0.28}$) and we are 95% confident that true odds ratio lies between 1.22 and 1.45.

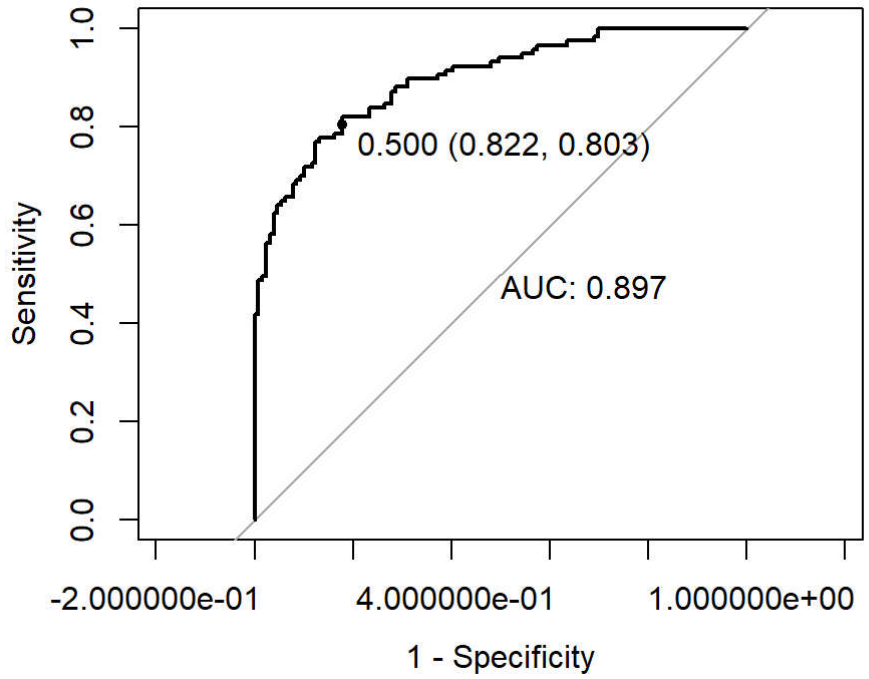
For every one unit increase in turnover, the odds of winning will decrease 16% ($e^{-0.17} = 0.84$) and we are 95% confident that true odds ratio lies between 0.75 and 0.94.

Q4: Are there are any concerns regarding multicollinearity in this model?

Home	TeamPoints	FieldGoals._percent	Assists	Steals	Blocks	TotalRebounds	Turnovers
1.274	2.033	3.440	1.528	1.527	1.164	2.268	1.287

From the VIF table, we can see that vif values for all predictors are less than 5, so multicollinearity is not a big concern in this model.

Q5: Using 0.5 as your cutoff for predicting wins or losses (1 vs 0) from the predicted probabilities, what is the accuracy of this model? Show the ROC curve and give the AUC.



Using 0.5 as cutoff predicting wins and losses, the accuracy of this model is 81.3%. From the ROC curve, we can see that the auc score is 0.897. It's close to 1, indicating our model has good performance.

Q6: Now add Opp.FieldGoals., Opp.TotalRebounds, Opp.TotalFouls, and Opp.Turnovers as predictors to the previous model. Interpret coefficients for significant terms, if any.

The result tables including point estimate, standard error, z-value, p-value and confidence intervals are in the appendix for reference.

The p-value for each predictor indicates that Home, TeamPoints, FieldGoals.(%), Turnovers, Opp.FieldGoals., Opp.TotalRebounds, and Opp.Turnovers are statistically significant at the level of 0.05. Here are the interpretations.

Odds of winning the game are 4.6 ($e^{1.53}$) times higher for home team compared to away team and we are 95% confident that the true odds ratio comparing home and away team lies between 1.3 and 19.3.

For every one unit increase in the number of total points scored in the game, the odds of winning will increase by a factor of 1.15 ($e^{0.14}$) and we are 95% confident that true odds ratio lies between 1.03 and 1.32.

For every one percent increase in the ratio of number of field goals made in the game and the number of field goals attempted in the game, the odds of winning increase by a factor of 1.49 ($e^{0.4}$),and we are 95% confident that the true odds ratio lies between 1.13 and 2.1.

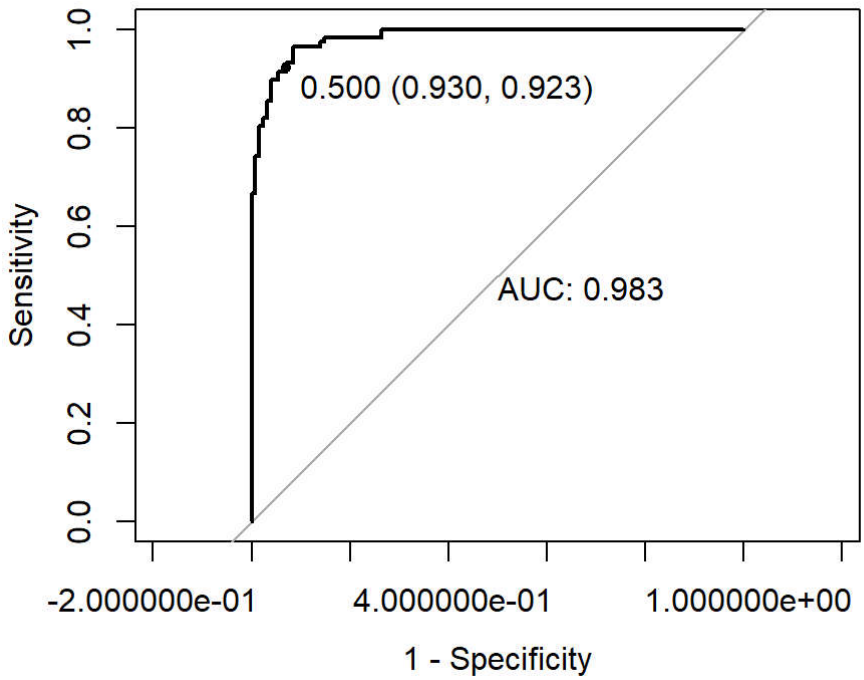
For every one unit increase in the total number of turnover, the odds of winning will decrease by 42% ($e^{-0.55} = 0.58$) and we are 95% confident that the true odds ratio lies between 0.44 and 0.72.

For every one percent increase in the ratio of number of field goals made by the opposing team in the game and the number of field goals attempted by the opposing team in the game, the odds of winning decrease by 58% ($e^{-0.86} = 0.42$), and we are 95% confident that the true odds ratio of is between 0.29 and 0.56.

For every one unit increase in the number of rebounds grabbed by the opposing team, the odds of winning will decrease by 30% ($e^{-0.36} = 0.7$) and we are 95% confident that the true odds ratio lies between 0.57 and 0.82.

For every one unit increase in the number of opponent's turnovers, the odds of winning will increase by a factor of 1.84 ($e^{0.61}$) and we are 95% confident that true odds ratio lies between 1.38 and 2.64.

Q7: What is the accuracy of this new model? Show the ROC curve and give the AUC. Which model better predicts the odds of winning?



The accuracy of this model was increased to 92.68% and the roc-auc score was also improved to 0.983. In terms of sensitivity, the new mode can predict 92.31% of teams who won the game, while the prior model can only detect 80.31%. Specificity also increased from 82.17% to 93.02%, which means the new model has better chance to correctly predict the team who lose the game. Based on roc-auc socre, sensitivity, specificity, the new model better predicts the odds of winning.

Q8: Use the model that you selected in question 7 to predict out-of-sample probabilities for the nba_reduced_test data. Using 0.5 as your cutoff for predicting wins or losses (1 vs 0) from the out-of-sample predicted probabilities, what is the out-of-sample accuracy? How well does your model do in predicting data for the 2017/2018 season?

Using the improved model to predict whether teams in the test set will win the game for the 2017/2018 season and comparing with the true results, our model achieved 86.59% out-of-sample accuracy. From the ROC curve (in appendix), the roc-auc score is 0.984 (very close to 1), which means that our model can predict winning very well. The sensitivity is 0.9444 indicating our model can identify 94.44% of those teams who won the game. The specificity is 0.8043, meaning 80.43% of teams losing the game can be predicted correctly by this model.

Q9: Using the change in deviance test, test whether including Opp.Assists and Opp.Blocks in the model at the same time would improve the model. Is there any other variable in this dataset which we did not consider that you think might improve our model? Which one and why?

Here I performed a chi-squared test between two models with and without Opp.Assists and Opp.Blocks. The p-value is 0.46, which means the added predictors are not statistically significant. However, after I calculated the confusion matrix, the accuracy of the model including Opp.Assists and Opp.Blocks is increased from 86.59% to 89.02%. Therefore, including these two predictors did improve the model.

Then I added another variable: Opp.X3PointShots.(with period). I chose this variable because it logically makes sense that if a team's opposing team has higher chance shooting 3-point goal, the smaller chance this team will win. To test whether it is a significant variable, first I fitted a logistic regression model with Opp.X3PointShots. added. The p-value is extremely small, indicating the variable is statistically significant. From the result table in the appendix, the coefficient of Opp.X3PointShots. indicates that one percent increase in the ratio of number of 3 point shots made by the opposing team and number of 3 point shots attempted by the opposing team, the odds of winning will decrease by 16% ($e^{-0.17} = 0.84$). Then I conducted chi-sqaure test between the 2 models with and without Opp.X3PointShots. and the p-value indicates that there is significant difference between the two models. Finally, I plotted the roc curve (in appendix) and obtained the auc score. The auc score of the model with Opp.X3PointShot. is 0.989, which is higher than the previous one. So adding Opp.X3PointShot.(with period) could improve our logistic regression model.

Q10: What do you conclude from this analysis?

1. According to the coefficients of our logistic regression model, predictor "Home" also has the greatest impact on whether a team will win. If a team is a guest team, it may suffer from real-life disadvantages such as time zone changes, tough travel, unfamiliarity of field as well as psychological factors, which will much lower the chance of winning the game.
2. FieldGoals., Opp.FieldGoals., and turnovers are also significant in affecting team winning or losing, so coach should work on these fields to increase the odds of winning. For example, coaches can focus on training methods for improving basketball shooting proficiency and tell players to put more effort in attempting to stop the opposition from scoring. Coaches should also try to figure out ways to reduce turnovers, such as practicing footwork, using passing drills with fast moving targets, etc.

Appendix

Q6 Result table after adding Opp.FieldGoals., Opp.TotalRebounds, Opp.TotalFouls, and Opp.Turnovers

Logistic regression Model2 Regressing Win

Predictor	Estimate	SE	z value	p-value
(Intercept)	9.19	11.61	0.79	0.43
HomeHome	1.53	0.68	2.25	0.02
TeamPoints	0.14	0.06	2.26	0.02
FieldGoals._percent	0.40	0.16	2.56	0.01
Assists	-0.01	0.08	-0.12	0.91
Steals	0.35	0.18	1.96	0.05
Blocks	-0.09	0.10	-0.91	0.36
TotalRebounds	0.17	0.09	1.83	0.07
Turnovers	-0.55	0.13	-4.27	0.00
Opp.FieldGoals._percent	-0.86	0.17	-5.08	0.00
Opp.TotalRebounds	-0.36	0.09	-3.94	0.00
Opp.TotalFouls	0.10	0.10	0.96	0.34
Opp.Turnovers	0.61	0.16	3.73	0.00

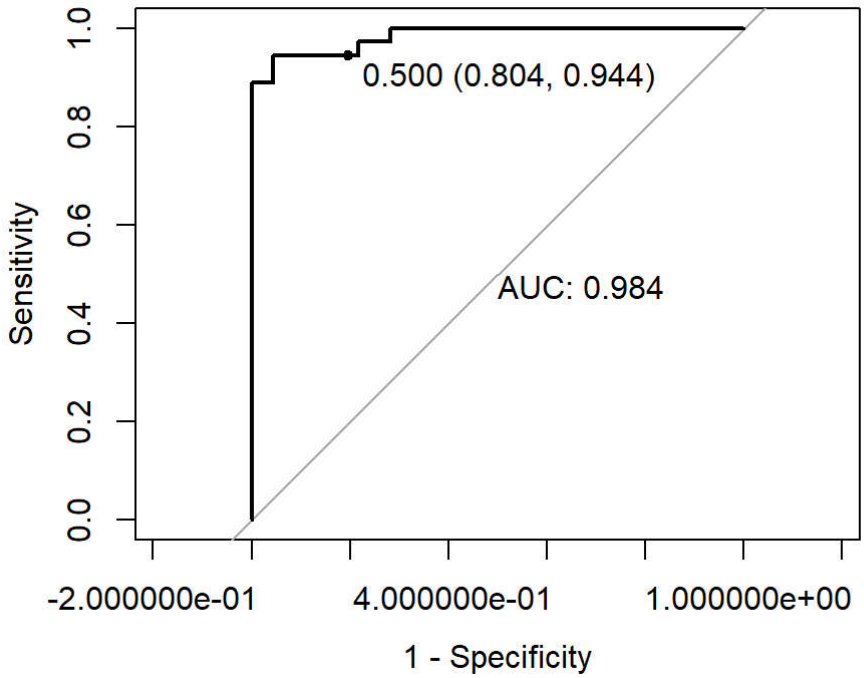
[a] Null deviance: 340.44
[b] Residual deviance: 66.263
[c] AIC: 94.263

95% Confidence Interval

Predictor	2.5%	97.5%
Intercept	0.00	1.727262e+14
Home	1.30	1.931000e+01
TeamPoints	1.03	1.320000e+00
FieldGoals.	1.13	2.100000e+00
Assists	0.84	1.160000e+00
Steals	1.02	2.060000e+00

Predictor	2.5%	97.5%
Blocks	0.74	1.120000e+00
TotalRebounds	1.00	1.420000e+00
Turnovers	0.44	7.200000e-01
Opp.FieldGoals.	0.29	5.600000e-01
Opp.TotalRebounds	0.57	8.200000e-01
Opp.TotalFouls	0.90	1.360000e+00
Opp.Turnovers	1.38	2.640000e+00

Q8 ROC curve in predicting data for the 2017/2018 season



Q9 Result table after adding Opp.3PointShots.

Logistic regression Model4 Regressing Win

Predictor	Estimate	SE	z value	p-value
(Intercept)	31.64	17.16	1.84	0.07
HomeHome	1.07	0.82	1.30	0.19
TeamPoints	0.26	0.09	2.82	0.00

[a] Null deviance: 340.44
[b] Residual deviance: 82.323
[c] AIC: 108.32

Predictor	Estimate	SE	z value	p-value
FieldGoals._percent	0.29	0.17	1.72	0.09
Assists	-0.02	0.09	-0.27	0.79
Steals	0.31	0.19	1.68	0.09
Blocks	0.02	0.11	0.15	0.88
TotalRebounds	0.06	0.11	0.58	0.56
Turnovers	-0.65	0.15	-4.25	0.00
Opp.FieldGoals._percent	-1.07	0.24	-4.42	0.00
Opp.TotalRebounds	-0.55	0.15	-3.75	0.00
Opp.TotalFouls	0.03	0.12	0.22	0.83
Opp.Turnovers	0.75	0.19	3.90	0.00
Opp.3PointShots._percent	-0.17	0.05	-3.36	0.00

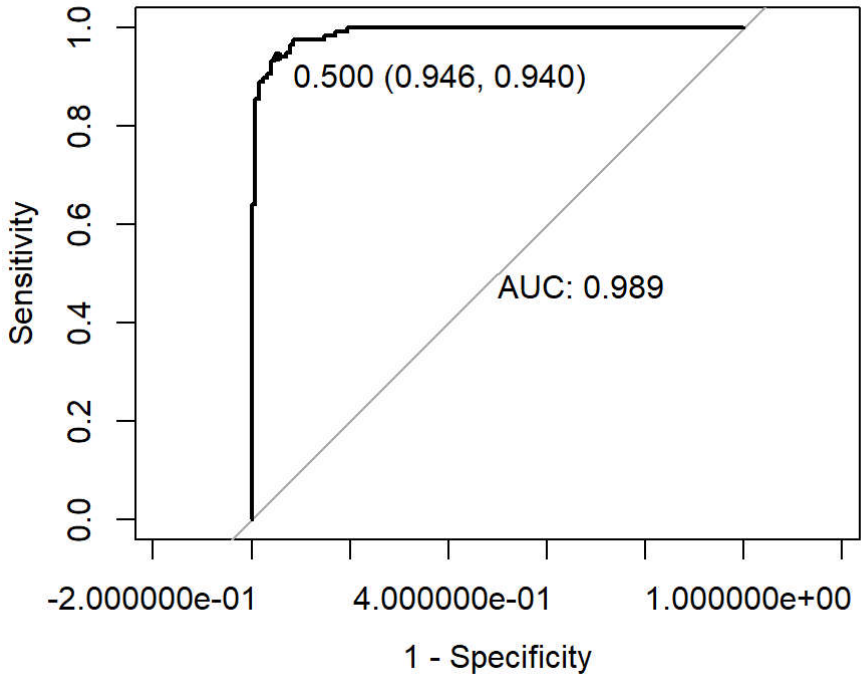
[a] Null deviance: 340.44
[b] Residual deviance: 82.323
[c] AIC: 108.32

95% Confidence Interval

Predictor	2.5%	97.5%
Intercept	2.06	1.036742e+30
Home	0.61	1.599000e+01
TeamPoints	1.11	1.610000e+00
FieldGoals.	0.99	1.940000e+00
Assists	0.81	1.170000e+00
Steals	0.96	2.020000e+00
Blocks	0.81	1.270000e+00
TotalRebounds	0.86	1.320000e+00
Turnovers	0.37	6.800000e-01
Opp.FieldGoals.	0.20	5.100000e-01
Opp.TotalRebounds	0.41	7.300000e-01

Predictor	2.5%	97.5%
Opp.TotalFouls	0.81	1.320000e+00
Opp.Turnovers	1.51	3.240000e+00
Opp.3PointShots.	0.75	9.200000e-01

Q9 ROC curve after adding new variable: Opp.3PointShots.



```
h1, h4 {
  text-align: center;
}

knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyr)
library(dplyr)
library(ggplot2)
library(brew)
library(stargazer)
library(patchwork)
library(corrplot)
library(caret)
library(kableExtra)
library(broom)
library(car)
library(leaps)
library(MASS)
library(gridExtra)
library(pROC)
library(PerformanceAnalytics)
# Data
nba <- read.csv('nba_games_stats.csv', header = 1, stringsAsFactors = TRUE)

## clean and subset the data
# Set factor variables
nba$Home <- factor(nba$Home)
nba$Team <- factor(nba$Team)
nba$WINorLOSS <- factor(nba$WINorLOSS)
# Convert date to the right format
nba$Date <- as.Date(nba$Date, "%Y-%m-%d")
# Also create a binary variable from WINorLOSS.
# This is not always necessary but can be useful
#particularly for R functions that prefer numeric binary variables
#to the original factor variables
nba$Win <- rep(0,nrow(nba))
nba$Win[nba$WINorLOSS=="W"] <- 1
nba$Win <- as.factor(nba$Win)

# Charlotte hornets subset
nba_reduced <- nba[nba$Team == "CHO", ]
# Set aside the 2017/2018 season as your test data
nba_reduced_train <- nba_reduced[nba_reduced$Date < "2017-10-01",]
nba_reduced_test <- nba_reduced[nba_reduced$Date >= "2017-10-01",]

dim(nba_reduced)
head(nba_reduced)
summary(nba_reduced)
str(nba_reduced)
# EDA
## y: Win
## x: Home, TeamPoints,FieldGoals. (with a period!), Assists, Steals, Blocks, OpponentPoints,
TotalRebounds, and Turnovers

# Win - Home
```

```
gg1 <- ggplot(nba_reduced_train, aes(x=Win,fill=Home)) + geom_histogram(stat="count") +
  labs(title="Home vs Win",
       x="Win the game?",y="Number of team") +
  theme_classic() +
  theme(legend.position="top")

 # ggplot(nba_reduced_train, aes(Win, ..count..)) + geom_bar(aes(fill = Home), position = "do
dge") + coord_flip() +
  #labs(title="Home or away game vs Win the game",
  #     x="Win the game?",y="Number of team") +
#  theme_classic() +
 # scale_fill_brewer(palette="Blues")
  #theme(legend.position="none")

# Win - TeamPoints
gg2 <- ggplot(nba_reduced_train,aes(x=Win, y=TeamPoints, fill=Win)) +
  geom_boxplot() + coord_flip() +
  labs(title="TeamPoints vs Win ",
       x="Win the game?",y="Number of total points scored") +
  stat_summary(fun.y="mean")+
  theme_classic() +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")

# Win - FieldGoals.
gg3 <- ggplot(nba_reduced_train,aes(x=Win, y=FieldGoals., fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  scale_fill_brewer(palette="Blues") +
  labs(title="FieldGoals. vs Win",
       x="Win the game?",y="Number of field goals") +
  theme_classic() + theme(legend.position="none")

# Win - Assists
gg4 <- ggplot(nba_reduced_train,aes(x=Win, y=Assists, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  labs(title="Assists vs Win",
       x="Win the game?",y="Number of assists") +
  theme_classic() +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")

# Win - Steals
gg5 <- ggplot(nba_reduced_train,aes(x=Win, y=Steals, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  scale_fill_brewer(palette="Blues") +
  labs(title="Steals vs Win",
       x="Win the game?",y="Nnumber of steals") +
  theme_classic() + theme(legend.position="none")

# Win - Blocks
gg6 <- ggplot(nba_reduced_train,aes(x=Win, y=Blocks, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
```

```
  labs(title="Blocks vs Win",
       x="Win the game?",y="Number of blocks ") +
  theme_classic() +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")

# Win - OpponentPoints
gg7 <- ggplot(nba_reduced_train,aes(x=Win, y=OpponentPoints, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  scale_fill_brewer(palette="Blues") +
  labs(title="OpponentPoints vs Win",
       x="Win the game?",y="Total points scored by the opposing team") +
  theme_classic() + theme(legend.position="none")

# Win - TotalRebounds
gg8 <- ggplot(nba_reduced_train,aes(x=Win, y=TotalRebounds, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  labs(title="TotalRebounds vs Win  ",
       x="Win the game?",y="Number of rebounds grabbed") +
  theme_classic() +
  scale_fill_brewer(palette="Blues") +
  theme(legend.position="none")

# Win - Turnovers
gg9 <- ggplot(nba_reduced_train,aes(x=Win, y=Turnovers, fill=Win)) +
  geom_boxplot() + coord_flip() +
  stat_summary(fun.y="mean") +
  scale_fill_brewer(palette="Blues") +
  labs(title="Turnovers vs Win  ",
       x="Win the game?",y="Total number of turnovers") +
  theme_classic() + theme(legend.position="none")

grid.arrange(gg1,gg2,gg3,gg4,gg5,gg6,gg7,gg8,gg9, nrow = 3)


#potential_corr_cols <- c("TeamPoints","FieldGoals", "FieldGoalsAttempted", "FieldGoals.", "O
ffRebounds", "TotalRebounds")

#chart.Correlation(nba_reduced_train[, potential_corr_cols], histogram=TRUE)

nba_reduced_train$FieldGoals._percent <- nba_reduced_train$FieldGoals. * 100
nbareg1 <- glm(Win ~ Home + TeamPoints + FieldGoals._percent + Assists + Steals + Blocks + To
talRebounds + Turnovers, data = nba_reduced_train, family = binomial)
summary(nbareg1)

nbareg1 %>%
  tidy() %>%
  kable(caption = "Logistic regression Model Regressing Win",
        col.names = c("Predictor", "Estimate", "SE", "z value", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Null deviance: 340.44","Residual deviance: 195.38","AIC: 213.38"))%>%
  kable_styling(position="center", full_width = T)
conf_nba <- exp(confint(nbareg1))
```

```
conf_nba_df <- data.frame(conf_nba)
rownames(conf_nba_df) <- NULL
Predictor <- c("Intercept", "Home","TeamPoints", "FieldGoals._percent", "Assists", "Steals",
"Blocks", "TotalRebounds", "Turnovers")
cbind(data.frame(Predictor), conf_nba_df) %>%
  kable( caption="<center>95% Confidence Interval</center>",
         bookstabs = T,
         col.names = c("Predictor", "2.5%", "97.5%"),
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center") %>%
  kable_styling(full_width = T)

vif_nba <- vif(nbareg1)
stargazer(vif_nba, type = 'html', out = 'vif_nba.html')
confusionMatrix(as.factor(ifelse(fitted(nbareg1) >= 0.5, "1","0")), nba_reduced_train$Win,pos
itive = "1")
roc(nba_reduced_train$Win, fitted(nbareg1), print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)

nba_reduced_train$Opp.FieldGoals._percent <- nba_reduced_train$Opp.FieldGoals. * 100
nbareg2 <- glm(Win ~ Home + TeamPoints + FieldGoals._percent + Assists + Steals + Blocks + To
talRebounds + Turnovers + Opp.FieldGoals._percent + Opp.TotalRebounds + Opp.TotalFouls + Opp.
Turnovers, data = nba_reduced_train, family = binomial)
summary(nbareg2)
confusionMatrix(as.factor(ifelse(fitted(nbareg2) >= 0.5, "1","0")), nba_reduced_train$Win,pos
itive = "1")

roc(nba_reduced_train$Win, fitted(nbareg2), print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)

nba_reduced_test$FieldGoals._percent <- nba_reduced_test$FieldGoals. * 100
nba_reduced_test$Opp.FieldGoals._percent <- nba_reduced_test$Opp.FieldGoals. * 100

nba_test_predict <- predict(nbareg2, nba_reduced_test, type = "response")

confusionMatrix(as.factor(ifelse(nba_test_predict >= 0.5, "1","0")), nba_reduced_test$Win,pos
itive = "1")

roc(nba_reduced_test$Win, nba_test_predict, print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)
nbareg3 <- glm(Win ~ Home + TeamPoints + FieldGoals._percent + Assists + Steals + Blocks + To
talRebounds + Turnovers + Opp.FieldGoals._percent + Opp.TotalRebounds + Opp.TotalFouls + Opp.
Turnovers + Opp.Assists + Opp.Blocks, data = nba_reduced_train, family = binomial)

anova(nbareg2, nbareg3, test = "Chisq")

nba_test_predict <- predict(nbareg3, nba_reduced_test, type = "response")

confusionMatrix(as.factor(ifelse(nba_test_predict >= 0.5, "1","0")), nba_reduced_test$Win,pos
itive = "1")

# add Opp.3PointShots.
nba_reduced_train$Opp.3PointShots._percent <- nba_reduced_train$Opp.3PointShots. * 100
nbareg4 <- glm(Win ~ Home + TeamPoints + FieldGoals._percent + Assists + Steals + Blocks + To
talRebounds + Turnovers + Opp.FieldGoals._percent + Opp.TotalRebounds + Opp.TotalFouls + Opp.
```

```
Turnovers + Opp.3PointShots._percent, data = nba_reduced_train, family = binomial)
summary(nbareg4)
# Deviance
anova(nbareg2, nbareg4, test = "Chisq")
# roc-auc
roc(nba_reduced_train$Win, fitted(nbareg4), print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)
nbareg2 %>%
  tidy() %>%
  kable(caption = "Logistic regression Model2 Regressing Win",
        col.names = c("Predictor", "Estimate", "SE", "z value", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Null deviance: 340.44","Residual deviance: 66.263","AIC: 94.263"))%>%
  kable_styling(position="center", full_width = T)
conf_nba2 <- exp(confint(nbareg2))
conf_nba_df2 <- data.frame(conf_nba2)
rownames(conf_nba_df2) <- NULL
Predictor <- c("Intercept", "Home","TeamPoints", "FieldGoals.", "Assists", "Steals", "Blocks"
, "TotalRebounds", "Turnovers", "Opp.FieldGoals.", "Opp.TotalRebounds", "Opp.TotalFouls", "Op
p.Turnovers")
cbind(data.frame(Predictor), conf_nba_df2) %>%
  kable( caption="<center>95% Confidence Interval</center>",
         bookstabs = T,
         col.names = c("Predictor", "2.5%", "97.5%"),
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center") %>%
  kable_styling(full_width = T)

nba_test_predict <- predict(nbareg2, nba_reduced_test, type = "response")
confusionMatrix(as.factor(ifelse(nba_test_predict >= 0.5, "1","0")), nba_reduced_test$Win,pos
itive = "1")
roc(nba_reduced_test$Win, nba_test_predict, print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)
nbareg4 %>%
  tidy() %>%
  kable(caption = "Logistic regression Model4 Regressing Win",
        col.names = c("Predictor", "Estimate", "SE", "z value", "p-value"),
        digits = c(2, 2, 2, 2, 2),
        align = "l") %>%
  add_footnote(c("Null deviance: 340.44","Residual deviance: 82.323","AIC: 108.32"))%>%
  kable_styling(position="center", full_width = T)
conf_nba4 <- exp(confint(nbareg4))
conf_nba_df4 <- data.frame(conf_nba4)
rownames(conf_nba_df4) <- NULL
Predictor <- c("Intercept", "Home","TeamPoints", "FieldGoals.", "Assists", "Steals", "Blocks"
, "TotalRebounds", "Turnovers", "Opp.FieldGoals.", "Opp.TotalRebounds", "Opp.TotalFouls", "Op
p.Turnovers", "Opp.3PointShots.")
cbind(data.frame(Predictor), conf_nba_df4) %>%
  kable( caption="<center>95% Confidence Interval</center>",
         bookstabs = T,
         col.names = c("Predictor", "2.5%", "97.5%"),
         digits = c(2, 2),
         align = "l") %>%
  kable_styling(position="center") %>%
```

```
  kable_styling(full_width = T)
# add Opp.3PointShots.
nbareg4 <- glm(Win ~ Home + TeamPoints + FieldGoals. + Assists + Steals + Blocks + TotalRebou
nds + Turnovers + Opp.FieldGoals. + Opp.TotalRebounds + Opp.TotalFouls + Opp.Turnovers + Opp.
3PointShots., data = nba_reduced_train, family = binomial)
summary(nbareg4)
# Deviance
anova(nbareg2, nbareg4, test = "Chisq")
# roc-auc
roc(nba_reduced_train$Win, fitted(nbareg4), print.thres=0.5, print.auc=T, plot=T,legacy.axes=
T)
```