Analysis of NCAA Women's Basketball Data

Josie Peterburs
Stat 410 Spring 2025

Introduction

The game of basketball relies on hundreds of statistics to quantify team and player performance. Identifying the most critical statistics for a team to improve its win rate is a key question for coaches and analysts. To address this, I conducted a logistic regression analysis to predict wins based on common basketball statistics. The dataset for this project was obtained from the March Machine Learning Mania Competition on Kaggle, where participants use historical NCAA basketball data to forecast the outcomes of the men's and women's March Madness tournaments. In this project, I use the 2024-2025 women's regular season detailed results dataset to identify the factors most influential in determining game outcomes. The primary goal is to construct a model that selects the most impactful statistics while maintaining interpretability, allowing the findings to be easily communicated to the coaching staff.

The dataset comprises 34 variables, mainly statistics for winning and losing teams, with each row representing a single game. Key variables of interest include field goals attempted and made, free throw percentage, assists, turnovers, and rebounds. I also incorporated a variable for effective field goal percentage (eFG%) to account for the added value of three-point shots. Effective field goal percentage was used as an evaluation tool for Loyola Women's basketball to judge team performance, and the team with the higher effective field goal percentage won in nearly all of our games this past season. These variables were used to predict a binary response variable, added to the dataset, indicating whether the home team won the game (1 for home team win, 0 for home team loss), with neutral-site games removed.

Methods

To better represent the data and to make interpretation easier, all of the variables were aligned relative to the home team. Given that the response variable is based on whether the home team won, variables such as WFGM and LFGM were relabeled as HomeFGM and AwayFGM, respectively. Additionally, features like field goal difference (FGM_diff), turnover difference (TO_diff), field goal percentage ratio, and effective field goal percentage ratio (eFG_pct_ratio) were added to provide comparisons between the two teams. The total number of rebounds (HomeR and AwayR) for each team was also calculated. Finally, the dataset was subset to include only games from the 2024-2025 season.

For the first model, I made a logistic regression model using backward selection with all variables in the maximum model. I did not include variables that were unique identifiers or non-statistical variables in the maximum model, so season, day number, and team ID were not included in the maximum model. Furthermore, when I first made this model, the only two

predictors left after variable selection were HomeScore and AwayScore. This makes sense, as the score of the game determines the winner, so the most important factor is being able to score more points than the other team. However, I did not want my model to include the score variables since this conclusion is so obvious. Therefore, HomeScore and AwayScore were also not included in the maximum model. After applying backward selection, I checked for multicollinearity. HomeFGM and AwayFGM were highly correlated, so I removed AwayFGM. This left me with a final model including HomeFGM, HomeFGM3, AwayFGM3, HomeFTM and AwayFTM.

The second model I made was a logistic regression model using stepwise selection with a subset of the variables included in the maximum model. The variables I chose to include were Home_eFG, Away_eFG, HomeR, AwayR, HomeAst, AwayAst, TO_diff, FGM_diff, and eFG_pct_ratio. These were the variables I was specifically interested in seeing how they affect the probability of winning. After stepwise selection, all of the variables were kept in the model except the two assist variables. After checking for multicollinearity, the final model included Home_eFG, HomeR, Away_R, TO_diff, and FGM_diff.

Lastly, I made a simple Random Forest model to compare with these logistic regression models. I used the same variables that were included in the model created using stepwise selection with a subset of the variables. This model was created to explore the trade-off between accuracy (random forest) and interpretability (logistic regression).

Results

To assess model performance, I generated predicted probabilities for each game and classified games as home wins or losses using a threshold of 0.5. The confusion matrix for this classification showed that the model correctly predicted the outcome for a substantial majority of games. The accuracy, sensitivity (true positive rate), and specificity (true negative rate) all indicated that both the backward and stepwise selection models perform well at distinguishing between home wins and losses.

Using the 2023 season as a test dataset, the accuracy of the backward selection model was 0.8447 while the accuracy of the stepwise selection model was 0.8824. Because of this, I will use the stepwise selection model as the final model and will focus on it in this analysis. The random forest model had a test accuracy of 0.8697, meaning it performed worse than the stepwise selection model.

The logistic regression model, adjusted for VIF, demonstrated significant relationships between key statistics and the likelihood of a home win. The model included Home_eFG, HomeR, AwayR, TO_diff, and FGM_diff. The coefficients revealed that higher Home_eFG and HomeR significantly increased the odds of a home win ($p < 2e\text{-}16$), with coefficients of 0.179 and 0.220, respectively. Conversely, higher AwayR and TO_diff decreased the odds, with coefficients of -0.087 and -0.239, respectively ($p < 2e\text{-}16$). Additionally, a greater FGM_diff (home field goals minus away field goals) strongly increased the odds of a home win (coefficient = 0.260, $p < 2e\text{-}16$). These relationships make sense, as having a higher effective field goal

percentage and number of rebounds increases the likelihood of the home team winning. Because TO_diff is away turnovers subtracted from home turnovers, if TO_diff is a negative number, it means the home team has more turnovers than the away team. Since more turnovers are generally detrimental, this aligns with the negative coefficient, indicating that, keeping other variables constant, a team with more turnovers is less likely to win. Lastly, for field goal difference, having better shooting than the away team has a significant impact on the home team's ability to win.

I also examined the area under the ROC curve (AUC), which measures the model's ability to discriminate between the two classes across all possible thresholds. The AUC for the stepwise selection model was 0.957, indicating a very good level of discrimination between home wins and home losses. The ROC curve visually confirms this high discriminatory power.

To check for potential issues with the model, I plotted the residuals and checked for any obvious patterns or outliers. The residual plots did not reveal any major problems, suggesting that the logistic regression model was appropriate for the data.

Though the random forest did not perform as well as the stepwise selection model, it provided additional insights into variable importance. The Mean Decrease in Accuracy for HomeR was the highest at 132.32, followed by FGMdiff at 129.12. TO_diff and Home_eFG also played significant roles. The Mean Decrease in Gini scores further supports these findings, with FGM_diff having the highest score (1863.14), highlighting its importance in node splitting within the random forest. Because a random forest model is more difficult to interpret than a logistic regression, the final model is the stepwise selection model using a subset of the variables in the maximum model as shown below.

$$\text{logit}(HomeWin) = -12.2408 + 0.1793 HeFG + 0.2197 HR - 0.0871 AR - 0.2394 ToDiff + 0.2596 FGMdiff$$

Conclusion

This study demonstrates the application of statistical modeling to identify key performance indicators in women's NCAA basketball. Both logistic regression and random forest analyses highlighted the importance of specific, actionable metrics for home team success. Notably, effective field goal percentage, rebounding, turnover differential, and field goal difference emerged as significant predictors of home win probability.

The key takeaways for coaches and analysts are clear: prioritizing improvements in effective field goal percentage, particularly at home, securing rebounds while limiting the opponent's rebounding opportunities, minimizing turnovers (or maximizing turnover differential), and outperforming the opponent in field goal conversions can substantially increase the likelihood of winning home games.

While this analysis provides valuable insights, future research could incorporate additional factors such as player-specific statistics, opponent quality, and game dynamics (e.g., pace of play) to refine the models further. Additionally, developing interactive tools that allow coaches to simulate the impact of improving specific metrics on win probability could enhance

the practical application of these findings. By utilizing both statistical analysis and actionable coaching strategies, teams can use data to gain a competitive edge and enhance their performance on the court.

Appendix

````
---
title: "410 Models"
format:
  html:
    self-contained: true
    embed-resources: true
editor: visual
execute:
  echo: true
  eval: true
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(warning = FALSE)
```

# 410 Project - Models

```{r, include=FALSE}
library(tidyverse)
library(ggpubr)
library(MASS)
library(car)
library(randomForest)
library(caret)
```

```{r}
wbb <- read.csv("C:/Users/gigip/personal/personal projects/march
madness/WRegularSeasonDetailedResults.csv")

# make effective field goal percentage variable
wbb <- wbb %>% mutate(WeFG = 100*(WFGM + 0.5 * WFGM3) / WFGA, LeFG = 100*(LFGM + 0.5
* LFGM3) / LFGA)
```

# Regression Analysis

## Data Prep

### Add Binary Outcome Variable

- add variable that signifies whether the home team won the game
````

- 1 if home team won
- 0 if home team lost
- removed neutral games

```{r}
wbb$HomeWin <- ifelse(
  wbb$WLoc == "H", 1,       # Home team won
  ifelse(wbb$WLoc == "A", 0, NA)  # Home team lost (Away won)
)
wbb <- wbb %>% filter(WLoc != "N") # Remove neutral games
```

Put Variables in Terms of Home/Away Teams

```{r}
# Initialize new columns for home/away stats
home_stats <- c("Score", "FGM", "FGA", "FGM3", "FGA3", "FTM", "FTA", "OR", "DR", "Ast", "TO",
"Stl", "Blk", "eFG")
away_stats <- paste0("Away_", home_stats)  # e.g., "Away_FGM"
home_stats <- paste0("Home_", home_stats)  # e.g., "Home_FGM"

# Loop through each stat and assign home/away values
for (i in seq_along(home_stats)) {
  stat <- gsub("Home_", "", home_stats[i])  # e.g., "FGM"

  # If home team won, W stats = home team, L stats = away team
  wbb[home_stats[i]] <- ifelse(
    wbb$WLoc == "H",
    wbb[[paste0("W", stat)]],  # e.g., WFGM
    wbb[[paste0("L", stat)]]   # e.g., LFGM (home team lost)
  )

  wbb[away_stats[i]] <- ifelse(
    wbb$WLoc == "H",
    wbb[[paste0("L", stat)]],  # e.g., LFGM
    wbb[[paste0("W", stat)]]   # e.g., WFGM
  )
}
```

Create Additional Features

```{r}
# Field goal difference
wbb$FGM_diff <- wbb$Home_FGM - wbb$Away_FGM

# Turnover difference (negative means home team had more TOs)
wbb$TO_diff <- wbb$Home_TO - wbb$Away_TO

# Home/Away FG% ratio
wbb$FG_pct_ratio <- (wbb$Home_FGM / wbb$Home_FGA) /
                    (wbb$Away_FGM / wbb$Away_FGA)

# Home/Away eFG% ratio
wbb$eFG_pct_ratio <- wbb$Home_eFG / wbb$Away_eFG
```

Cleaning

```{r}
wbb_subset <- wbb %>% filter(Season == 2024 | Season == 2025) %>%
dplyr::select(HomeWin:eFG_pct_ratio, NumOT)

# add a few more variables
wbb_subset$Home_R <- wbb_subset$Home_OR + wbb_subset$Home_DR
wbb_subset$Away_R <- wbb_subset$Away_OR + wbb_subset$Away_DR
wbb_subset$HomeWin <- as.factor(wbb_subset$HomeWin)

head(wbb_subset)
```

Create a Test Dataset Using 2023 Season

```{r}
wbb_subset_test <- wbb %>% filter(Season == 2023) %>% dplyr::select(HomeWin:eFG_pct_ratio,
NumOT)

# add a few more variables
wbb_subset_test$Home_R <- wbb_subset_test$Home_OR + wbb_subset_test$Home_DR
wbb_subset_test$Away_R <- wbb_subset_test$Away_OR + wbb_subset_test$Away_DR
wbb_subset_test$HomeWin <- as.factor(wbb_subset_test$HomeWin)

head(wbb_subset_test)
```

Logistic Regression

Max model with all variables

- do not include unique identifiers or non-statistical variables (season, day number, team ID)

- since the location was used to create the response variable, do not include this either

  - more interested in how stats impact win vs lose because teams will have to play at home and away no matter what

Test Model

```{r}
mod <- glm(HomeWin ~ Home_eFG + Away_eFG + TO_diff, family = binomial(), data = wbb_subset)

summary(mod)
vif(mod)
```

Backward Selection

```{r, echo = TRUE, results='hide'}
mod_max <- glm(HomeWin ~ . - Home_Score - Away_Score, data = wbb_subset, family = binomial())
mod_back <- step(mod_max, direction = "backward")
```

```{r}
summary(mod_back)
vif(mod_back)

adj_mod <- glm(HomeWin ~ Home_FGM + Home_FGM3 + Away_FGM3 + Home_FTM + Away_FTM, wbb_subset, family = binomial())
summary(adj_mod)
vif(adj_mod)
```

Accuracy of Backwards Selection

```{r}
# confusion matrix using 24-25 data
actual <- as.factor(wbb_subset$HomeWin)
predict_probs <- predict(adj_mod, type = "response")
predicted <- as.factor(ifelse(predict_probs > 0.5, 1, 0))
```

```
conf_mat <- confusionMatrix(actual, predicted)
conf_mat
```

Accuracy for Backwards Selection on Test Dataset

```{r}
actual <- as.factor(wbb_subset_test$HomeWin)
predict_probs <- predict(adj_mod, newdata = wbb_subset_test, type = "response")
predicted <- as.factor(ifelse(predict_probs > 0.5, 1, 0))

conf_mat <- confusionMatrix(actual, predicted)
conf_mat
```

Variable Selection Using Subset for Max Model

```{r}
mod_max <- glm(HomeWin ~ Home_eFG + Away_eFG + Home_R + Away_R + Home_Ast + Away_Ast
+ TO_diff + FGM_diff + eFG_pct_ratio, family = binomial(), data = wbb_subset)
summary(mod_max)

# stepwise selection
mod <- step(mod_max, direction = "both")
summary(mod)
vif(mod)

# adjust for vif
mod_adj <- glm(HomeWin ~ Home_eFG + Home_R + Away_R + TO_diff + FGM_diff, wbb_subset,
family = binomial)
summary(mod_adj)
vif(mod_adj)
```

Accuracy for Subset Max Model

```{r}
# confusion matrix using 24-25 data
actual <- as.factor(wbb_subset$HomeWin)
predict_probs <- predict(mod_adj, type = "response")
predicted <- as.factor(ifelse(predict_probs > 0.5, 1, 0))

conf_mat <- confusionMatrix(actual, predicted)
```

conf_mat
```

### Accuracy for Stepwise Model on Test Dataset

```{r}
actual <- as.factor(wbb_subset_test$HomeWin)
predict_probs <- predict(mod_adj, newdata = wbb_subset_test, type = "response")
predicted <- as.factor(ifelse(predict_probs > 0.5, 1, 0))

conf_mat <- confusionMatrix(actual, predicted)
conf_mat
```

ROC Curve

```{r}
library(pROC)
# get the predicted probabilities (pi-hats)
pi_hat <- predict(mod_adj, type = "response") # makes predictions off every observation in the data set

# roc() has two inputs: the actual response and the pi-hats
roc_obj <- roc(wbb_subset$HomeWin, pi_hat)
auc_value <- auc(roc_obj)

data.frame(fpr = 1 - roc_obj$specificities, tpr = roc_obj$sensitivities) %>%
  ggplot(aes(x = fpr, y = tpr)) +
  geom_line(color="maroon") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "darkgrey") +
  xlab("False Positive Rate") +
  ylab("True Positive Rate")
auc_value
```

Check Deviance

```{r}
plot(fitted(mod_adj), residuals(mod_adj))
plot(residuals(mod_adj))
```

Compare With Random Forest
```

Compare best logistic regression model with a Random Forest to show trade-offs between simplicity and accuracy.

```{r}
rf_model <- randomForest(
  as.factor(HomeWin) ~ Home_eFG + Home_R + Away_R + TO_diff + FGM_diff,
  data = wbb_subset,
  importance = TRUE  # Shows variable importance
)
print(importance(rf_model))
```

Random Forest Accuracy on Test Dataset

```{r}
predict_probs <- predict(rf_model, newdata = wbb_subset_test)

conf_matrix_rf <- confusionMatrix(predict_probs, wbb_subset_test$HomeWin)
conf_matrix_rf

```

Coefficient Plot

```{r}
library(ggplot2)
coef_df <- data.frame(
  Predictor = names(coef(mod_adj)),
  Effect = coef(mod_adj)
) %>% filter(Predictor != '(Intercept)')

ggplot(coef_df, aes(x = Effect, y = reorder(Predictor, Effect))) +
  geom_col(fill = "maroon") +
  labs(title = "Key Factors Affecting Win Probability",
       y = "Predictor", x = "Effect Size (Log-Odds)") +
  theme_minimal()
```

Bars show which stats most influence win margins (longer = stronger effect).