Predicting Minnesota Timberwolves Total Points and Win Percentages

Henry Smith and Matthew Labuda

Introduction

As the NBA postseason nears the end and the Timberwolves hopes of winning a championship get smaller and smaller with each postseason loss, we took it upon ourselves to figure out what the Timberwolves need to focus on in order to score more points and win games, using data. Thus, we want to answer a few questions.

- What in game statistics are relevant for the Timberwolves when predicting the total number of points in a game?
- What in game statistics are relevant for the Timberwolves when predicting wins and losses?

The goal of our project then also has two parts. For the first question regarding total points, we want to find the model with the highest adjusted r-squared value. Specifically, we are hoping to attain a model with an r-squared value greater than .9 as this would indicate that our model explains a vast majority of the variability in our data of the total points that the Timberwolves score in a game. Additionally, for the second question, we want to figure out what variables are most associated with wins and losses and how those variables compare to what we found in the first question. Previously, research has shown that statistics such as total rebounds and field goal percentage are highly significant in terms of predicting points and the outcomes of NBA games. We believe that these variables will be important variables in our project, as well as a few others. In order to test this, we will use a linear regression to predict total points in a game, and a logistic model to find statistics relevant for Timberwolves wins and losses.

Materials and Methods

In order to complete our dataset, we needed to download three different Timberwolves game log years from basketballreference.com, rename and select the important columns in each year such as field goal percentages (the number of shots made divided by the number of shots taken by the Timberwolves in a game), assists (The number of times a player passed the ball to another player who then scored in one game), total rebounds (the number of times the ball bounced off of the rim after a shot, and someone from the Timberwolves grabbed the ball) and combine them all together. After this transformation was done, a season column was added in order to compare in game stats with these categorical variables. The only removal of data that needed to occur was the first line of each season, as they were only the headers of the columns, and not actual statistics. Next, we combined the seasons on top of each other using rbind. First by combining the 2021 and 2022 seasons, and then combine that dataset with the 2023 season to get our final dataset. Finally, we added a conference column which indicates which conference the Timberwolves opponent is based on a given game, and changed the types of each column from characters to integers and numeric for appropriate statistical analysis (with the exception of the win loss column which was changed to be a factor). For our variables which were percentages such as free throw percentage and three point percentage, we multiplied the values by 100 so they would essentially be on the same numbered scale as other statistics and would not impact the effectiveness of the model.

Since we are investigating two main questions, we needed to look at two main modes for statistical inference. For our first question, we used multiple linear regression to evaluate which variables are linked to the total

points the Timberwolves will score in a game. This allowed us to specifically quantify the impact of single statistics on total points as we were able to see how one unit (shot, percentage point, rebound for example) has an impact on our response variable, points per game. Furthermore, we also were able to see whether our variables are significant or not based on their t-test for individual statistics z statistic and p-value. This let us quickly identify what was useful for predicting and what was not. In our model building, it was crucial to use variables which showed relationships with our response from the eda we carried out in order to understand where our starting point was. Additionally, we did not include every variable such as free throws made and field goals attempted (among others) because there is a strong relationship between those variables and FT%/FG%. From there, we used best subsets to go through every combination of our chosen variables to identify which model would yield the best r-squared value and p-values.

For our second question, we used LASSO regression to identify which variables are important for predicting Timberwolves wins and losses. LASSO allowed us to identify significant predictors by using a penalty (lambda) for adding a variable to the model. This makes non effective predictors coefficients trend towards zero as the penalty increases. Not only does this method point out important variables for predicting Timberwolves wins and losses, it shrinks our model and makes it simpler. Furthermore, by utilizing the one SE rule, we can maintain an effective model, but also simplify it even more by getting rid of some coefficients.

Results

The Timberwolves dataset that we have with one game per row and in game statistics for each column was analyzed through a few main plots when answering our first question related to the total points the Timberwolves will score in a game. The most informative plot we have is Figure 1 which shows that there is a positive relationship between three point percentage, and a small interaction between three point percentage and season. This, along with Figure 2 which is a correlation matrix that compares the different percentage statistics to total points, are the most significant descriptors of relationships within our data. The two strongest correlations were field goal percentage, and three-point percentage, both against points per game. Unfortunately, there were minimal findings of interactions between in game statistics and the conference of the opponent.

Our modeling boiled down to two main categories of models when predicting total points. Our first model was obtained by using the best subsets on variables we found important from eda which include field goal percentage, three-point percentage, steals, season, and season interacted with three-point percentage. This model gives us the highest adjusted r-squared value at .748. Additionally, the variables that are not statistically significant (including the intercept since it represents the 2021 season) are all variables related to the season variable. All season indicators, including the interactions between seasons and 3P% are well above the .05 significance line. The most significant predictor here is FG%. As field goal percentage increases by one percentage point, we expect the total points the Timberwolves score to increase by around 1.4 points (with other variables accounted for). Additionally, our confidence interval shows that we are 95% confident that the true impact that an increase in 1 FG% point has on the total points is an increase of between 1.19 and 1.64, and the only confidence intervals that contained 0 were the ones related to the season indicator variables.

Our second model for predicting total Timberwolves points presents different results. In this model, there is no interaction variable, and all of the predictors are significant. However, the adjusted r-squared is lower at .681. This model includes all of our percentage statistics, as well as steals, turnovers, and assists per game. Turnovers is the only predictor that has a negative relationship with total points. As turnovers increase by 1 unit, the total number of points decreases by around .528 points. The highest and most impactful predictor was FG%. An increase in one percentage point in field goal percentage will increase the total points scored by about 1.1 (with all other variables accounted for). Additionally, our confidence interval shows that we are 95% confident that the true impact that an increase in 1 FG% point has on the total points is an increase of between .88 and 1.32. There were no confidence intervals which contained 0.

When analyzing the data based on wins and losses, there are much more significant indicators from eda than in analyzing the dataset for our first question. Figure 3 is a great representation of what multiple

density plots look like. This specific density plot shows that there is a significant impact on wins and losses based on the total points that the Timberwolves score. The more points scored, the greater probability that they will win. Although this is expected since the team with the higher number of points in basketball wins the game, it is notable that Figure 3's shape is similar for assists, 3P%, FG%, and steals. With these values, there is great promise of finding a model with important predictors of Timberwolves wins and losses. Additionally, we did find some small interactions between shooting percentages and the conference of the opponent. However the interactions were never strong enough to be considered for a model.

Similar to linear regression, our logisitic modeling allowed us to predict Timberwolves wins and losses and provided us with two models. The first model contains the minimum deviance. This first model contains all of the inputted variables we wanted to test (and the ones that were not represented by other in-game statistics such as free throws attempted) except assists. Similarly to linear regression, turnovers have a negative relationship with wins. The more turnovers, the less likely the Timberwolves are to win a game. Each extra turnover is associated with a .818 times increase in the odds of winning a game (negative). Similarly, there is also a negative coefficient for the total points variable. Separately, the largest coefficient in this model is the FG%. This indicates that the higher the Timberwolves' free throw percentage is, the greater chance they have at winning that game. Specifically, each extra field goal percentage point is associated with a 39% increase in the odds of winning a game after controlling for other variables. After manually putting these variables into a glm, it is the case actually that the total rebounds TRB variable is the most significant out of all of them. After conducting a confidence interval, we are 95% confident that the true impact of an increase in one offensive or defensive rebound changes the odds of winning a game by between a factor of 1.187 to 1.399.

Finally, our LASSO model that is one se away maintains a similar deviance to the first model, but contains three less variables and therefore is simpler than our previous model. The variables that were dropped were total points, blocks, and personal fouls. FG% remains the highest valued coefficient, but at almost half the value as it was before the three variables were omitted. Specifically, each extra field goal percentage point is associated with a 22.1% increase in the odds of winning a game after controlling for other variables. Similar to the above model, turnovers remain a negative influence on winning. An extra turnover is associated with a .905 times increase (overall decrease) in the odds of winning a game (negative). After plugging these variables into a glm, FG% remains the most significant predictor with the highest test statistic and lowest p-value but very close to TRB. After conducting a confidence interval, we are 95% confident that the true impact of an increase in one FG% point changes the odds of winning a game by between a factor of 1.21 and 1.47. Finally, we conducted some win percentage probabilities from this model since it was the very last model. We found that having poor defense in our model can be made up for with good offense, and having bad offense can be made up for with good defense. Additionally, we found that if the Timberwolves has 0 steals, 0 rebounds, and 0 turnovers, they would need to shoot 70%+ from both the field and the three point line in order to just get a win probability of 50%.

Discussion

When first diving into this Timberwolves data set, our original question was: Can we predict the amount of points the Timberwolves score in a game based off of some of their Basketball statistics? Our results from all of our models showed us that this is possible due to the several variables we concluded to be significant which explain a majority of our data when examining points per game. The variables examined were Field Goal Percentage, Three Point Percentage, Free Throw Percentage, Season Year, Conference, Assists per Game, Steals per Game, Turnovers per Game, and Rebounds per Game. While we were unable to attain our goal of a model with a .9 r-squared value, we found that Field Goal Percentage, Three Point Percentage, Free Throw Percentage, Steals, Turnovers, and Assists to have the strongest relationships with Timberwolves points per game. Additionally, our second question of interest was what in game statistics are relevant for predicting Timberwolves wins and losses? After eliminating unimportant variables from eda and a variable selection process, we can conclude that steals, three point percentage, field goal percentage, total rebounds, and turnovers are the most important predictors for the Timverwolves from our data. Comparing these results to our literature references, we can see that we have found corroborating evidence to what Fadi

Thabtah found when identifying important predictors for NBA games. In both our analysis and Thabtah's, free throw percentage and total rebounds were very significant indicators of wins and losses.

This makes a lot of sense as all of these predictor variables are values taken from on-court Basketball statistics, not off-court Basketball statistics such as Season and Conference. This implies that players' performances control the main impact on the total points the Timberwolves score each game. This means the Timberwolves are more likely to score low points and therefore lose the game if they have a lot of turnovers, or, vice versa, they are more likely to score a lot of points and win the game if their players are hitting their shots whether it's free throws, three pointers, etc. A confounding variable which could be present in our study is the mental and physical health of the players on the roster. A poor mental or physical health would affect the predictor variables like shooting percentage as well as the response variable of total points per game. Unfortunately, there was no data on this in the site we found so this could be something to examine more in depth in the future.

In terms of generalizability, we believe our study can be applied to other NBA teams rather than just the Timberwolves as every team has similar data with similar trends regarding points per game and the explanatory variables we chose above. However, these trends will most likely vary a little bit per team as there are many star players which can completely alter the outcome of the game. For example, if we looked at the Lakers' statistics, there would be a much bigger emphasis on how many points Lebron got in correlation to the team's total points in that game. Overall, while the trends will be similar for each NBA team, there will definitely be some variation from team to team regarding some confounding variables only found in that particular team like star players or specific player issues.

For limitations, we have to always be careful and remember our study proves an association between variables, not causation. While we found many strong correlations, it is always important to remember this study proves how likely the Timberwolves are to score a low or high amount of points per game. Our study is strong in the sense that we were able to successfuly run a multiple linear regression model as well as a logistic regression model. We also chose very official data, and handpicked the very best predictors to find the very best of the best models. I think one could consider our study to be weak since we only looked at Timberwolves data rather than the NBA data as a whole, or even other leagues like the International or G-league. This would make the generalizability of our study much more valid. Additionally, we only looked at regular statistics that were available, and not advanced statistics that could potentially give us greater detail and answers as to what variables are important in predicting points and wins in the NBA.

Our next steps for future research would be to compare this data to similar data of other NBA teams to see how it meshes together as well as include more advanced in game statistics. This would help us weed out confounding variables and would help with the generalization issue like we stated above. Future research could also include a similar study but with another sport such as Ice Hockey or Football. It would be very interesting to see the main predictors in goals/points/etc for other popular sports and if those are comparable or similar to the ones highlighted above.

Annotated Appendix

Although it was not mentioned, the table of in game stats vs wins and losses was informative. These tell a similar story as the density plots used in eda for our LASSO regression. Every stat had a significant difference between wins and losses. Surprisingly, other than that table, the only plots and tables that were not referenced were figures that told us there was no relationship between the variables we were looking at such as the bar-plot of total rebounds colored by conference, and the interaction plots we made. The bar plot was a normal distributed plot with equal variance between conferences, and all of the interaction plots had marginal differences which were not significant.

Fortunately, we had no missing data, and the only time we needed to really alter our data was deleting a few rows which were redundant but not informative since they only contained the labels of the columns. Additionally, checking assumptions for us was essentially the same story across all of our models. They all passed LINE conditions for the same reasons from the residual plots as well as the conditions (elogits are linear, random, independence) for logistic regression. After finding the model with the highest r-squared for

question 1, we were torn that a handful of the variables were not significant from our t-test for individual predictors. These variables had coefficients that were not equal to 0, but p-values far greater than .05. For these reasons, we got rid of the season variables and its interactions to sacrifice on r-squared, but obtained our final model with all significant predictors. For question 2, we chose the one standard error model since it contained the smallest number of variables but still explained a large portion of the data.

For linear regression, the model output values were the same as they were in our interpretation. For our LASSO model, coefficients needed to be exponentiated in order to get the odds of a variable. Probabilities were calculated through specific inputs into the predict() function.

For our CIR visit, we visited Cheng Vang on Thursday May 4th at the CIR room. Most of our questions were guided towards getting feedback on if a part of our project looks good while the other questions were focused on some small syntax issues. For example, we asked questions such as: "How can we make a good looking table for our variables" and "Does our logistic model look good in regards to the project overview?" We received feedback and answers for every question we had, such as the useful library(kableExtra) Cheng gave us in regard to our question on how to make a clean-looking table. Overall, going to CIR was very beneficial and we are very grateful for Cheng spending some time to help us out.

References:

Jones, E. S. (2016). Predicting outcomes of NBA basketball games (Order No. 10150581). Available from ProQuest Dissertations & Theses Global; SciTech Premium Collection. (1815028507). Retrieved from https://www.proquest.com/dissertations-theses/predicting-outcomes-nba-basketball-games/docview/1815028507/se-2

Fadi, T., Zhang, L., & Neda, A. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 6(1), 103-116. doi:https://doi.org/10.1007/s40745-018-00189-x

Variables	Description
E/W	Conference
season	Season(from 2021, 2022, or 2023
AST	Assists per Game
STL	Steals per Game
TOV	Turnovers per Game
TRB	Total Rebounds per Game
FG%	Field Goal Percentage per Game
3P%	3 Point Percentage per Game
FT%	Free Throws Percentage per Game
Tm	Points per Game

```
twenty_one <- read_csv("~/Stats 272 S23/Project/Henry_Matthew/Wolves 2021.csv")
twenty_two <- read_csv("~/Stats 272 S23/Project/Henry_Matthew/Wolves 2022.csv") #read in data from each
twenty_three <- read_csv("~/Stats 272 S23/Project/Henry_Matthew/Wolves 2023.csv")
twenty_one <- twenty_one %>%
  rename(Date = 3, Opp = 5, WL = 6, Tm = 7, FGA = 10, "FG%" = 11, "3P" = 12, "3PA" = 13, "3P%" = 14, FT
twenty_one <- twenty_one[-c(1,2,4,8,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41)] #select import
twenty_one <- twenty_one %>%
  add_column(season = "21") #add the season as a column for each year
twenty_two <- twenty_two %>%
  rename(Date = 3, Opp = 5, WL = 6, Tm = 7, FGA = 10, "FG%" = 11, "3P" = 12, "3PA" = 13, "3P%" = 14, FT
twenty_two <- twenty_two[-c(1,2,4,8,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41)]
twenty_two <- twenty_two %>%
  add_column(season = "22")
twenty_three <- twenty_three %>%
  rename(Date = 3, Opp = 5, WL = 6, Tm = 7, FGA = 10, "FG%" = 11, "3P" = 12, "3PA" = 13, "3P%" = 14, FT
```

```
twenty_three <- twenty_three[-c(1,2,4,8,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41)]
twenty_three <- twenty_three %>%
  add_column(season = "23")
```

```
test <- rbind(twenty_one, twenty_two) #combine the first two seasons
wolves <- rbind(test, twenty_three) #combine the first two with the third
wolves <- wolves[-c(1,74,157),] #get rid of the labeled rows that are not data
wolves <- wolves %>%
  mutate("E/W" = ifelse(Opp == "DET" | Opp == "ATL" | Opp == "BKN" | Opp == "BOS" | Opp == "CHA" | Opp =
wolves <- wolves %>%
  mutate(Tm = as.integer(Tm),
         TRB = as.integer(TRB),
         AST = as.integer(AST),
         STL = as.integer(STL),
         BLK = as.integer(BLK),
         ORB = as.integer(ORB),
         FT = as.integer(FT),
         PF = as.integer(PF),
         TOV = as.integer(TOV)) #set all the numbered variables as integers
wolves <- wolves %>%
  mutate(`FG%` = as.numeric(`FG%`),
         `3P%` = as.numeric(`3P%`),
         `FT%` = as.numeric(`FT%`)) #set all the continuous variables as numeric
wolves <- wolves %>%
  mutate("FG%" = `FG%`*100,
         "3P%" = `3P%`*100,
         "FT%" = `FT%`*100)
wolves <- wolves %>%
  mutate(WL = as.factor(WL))
```

Comparative Plots

```
## `geom_smooth()` using formula 'y ~ x'
```

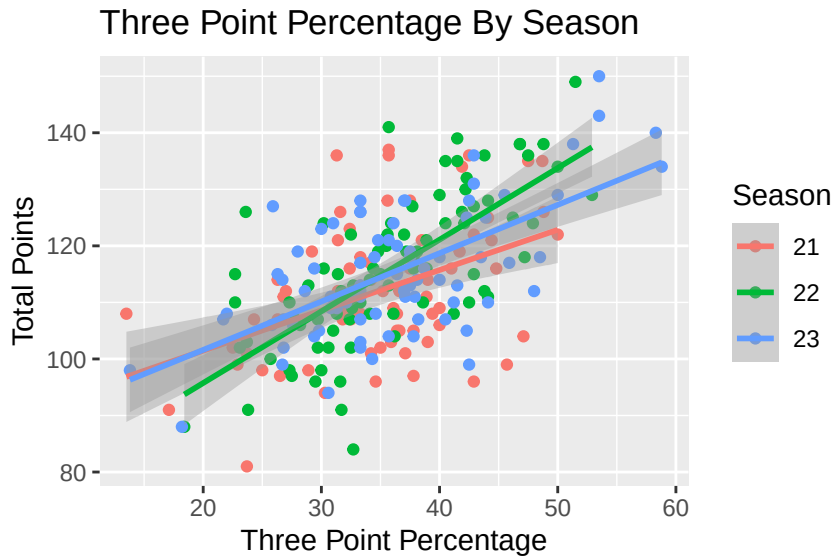

Figure 1

```
##             Tm        FG%        3P%        FT%
## Tm  1.0000000 0.73381250 0.59964904 0.18952094
## FG% 0.7338125 1.00000000 0.57442871 0.02634077
## 3P% 0.5996490 0.57442871 1.00000000 0.04990421
## FT% 0.1895209 0.02634077 0.04990421 1.00000000
```

Figure 2

Logistic Regression EDA

```
ggplot(data = wolves, aes(x = Tm, fill = WL)) +
  geom_density(position = 'fill', alpha = 0.5) +
  labs(x = "Total Points", y = "Density")
```

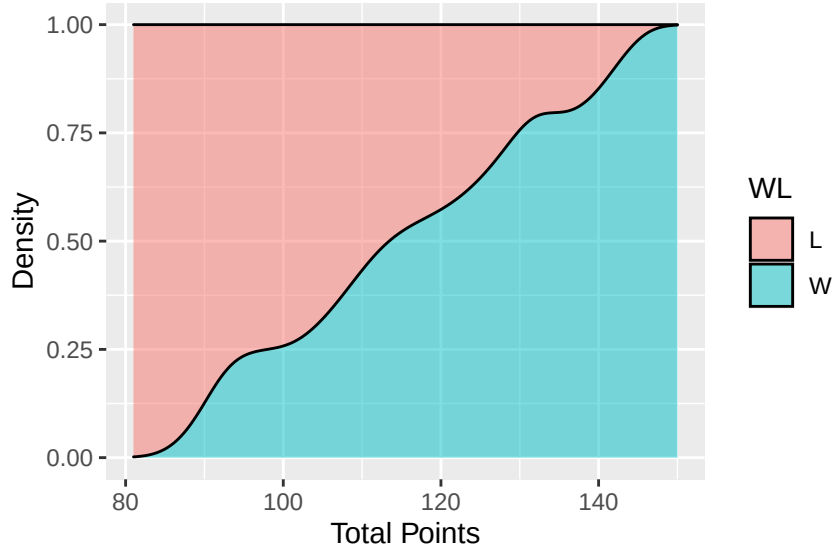

Figure 3

Final Model Question 1

```
model2 <- lm(Tm ~ `FG%` + `3P%` + `FT%` + STL + TOV + AST, wolves)
```

Intermediate Model Question 1

```
m1 <- lm(Tm~`FG%`+`3P%`+STL+season+ season:`3P%`, wolves)
summary(m1)
```

Confidence Interval Question 1

```
confint(model2)
```

```
##                   2.5 %     97.5 %
## (Intercept) 12.7329563 35.4293207
## `FG%`        0.8855834  1.3275748
## `3P%`        0.1839631  0.4737323
## `FT%`        0.1169018  0.3134281
## STL          0.3357835  0.9209362
## TOV         -0.7874748 -0.2692151
## AST          0.2562134  0.7332358
```

Final Model Question 2

```
model3b <- glm(WL ~ STL+`3P%`+`FG%`+TRB+TOV,
               family = binomial, data = wolves)
```

Confidence Intervals Question 2

```
exp(confint(model3a))
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %       97.5 %
## (Intercept) 2.780450e-15 6.611137e-09
## Tm          8.830762e-01 9.812934e-01
## STL         1.273040e+00 1.704436e+00
## `3P%`       1.025423e+00 1.157432e+00
## `FG%`       1.317079e+00 1.731402e+00
## TRB         1.187213e+00 1.399718e+00
## BLK         9.196237e-01 1.222090e+00
## TOV         6.851674e-01 8.729602e-01
## PF          9.536423e-01 1.144776e+00
```

```
exp(confint(model3b))
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %       97.5 %
## (Intercept) 4.320352e-14 1.722327e-08
## STL         1.204021e+00 1.559845e+00
## `3P%`       1.001818e+00 1.119436e+00
## `FG%`       1.213062e+00 1.465928e+00
## TRB         1.143402e+00 1.307065e+00
## TOV         7.460562e-01 9.202848e-01
```