Slaads - Project Proposal

Prepared by:	Kevin Gasiorowski Piotr Senkow Srikanth Nanduri Vineet Dcunha ~~Riley Edwards~~ (*dropped class*)
Date (mm/dd/yyyy):	03/08/2020

Goals

Betting on NCAA men's basketball is extremely popular. According to Reuters (1), about 47 million people bet a combined $8.5 billion on the NCAA tournament tournament in 2018. But to get into the tournament requires a successful regular season that is gauged by overall winning percentage. In this study, we performed an exhaustive analysis on how well *basic* and *advanced* regular season statistics from 2015 through 2019 predict overall regular season winning percentage for NCAA Division I Men's Basketball teams. This document summarizes the major findings of this analysis including which variables were most important to predicting winning percentage along with a comparison of the basic vs. advanced statistics in predicting winning percentage.

Methods

The models were created using the Win Percentage as the response variable. Our analysis was divided into multiple sections for the following criterias.

- Advanced statistics: Advanced statistics that are based on formulas that combine traditional, "basic" statistics into an overall rating.
- Offensive statistics: Includes stats that are collected while the team is attacking.
- Defensive statistics: Includes stats that are collected while the team is defending.
- Neutral statistics: Attributes within the data which are neither considered defensive or offensive fall into this category
- Basic statistics: Including all the attributes in the dataset

The statistics from the above criterias were used as the independent variables in our models.
The models were built using the first order, second order and interaction terms. The final models were validated against the forward selection and backward elimination methods.
The independent variables were checked for multicollinearity using the VIF function.
The histogram graph was plotted for the residuals to find any unconventional data which cannot be determined using the regression models.
The QQ plot and PP plot determined 95% of the data was plotted on the line.

The stabilization transformation was applied based on the residuals plots.

Overview
Our results suggest several key findings, including:

- Scoring-Margin is the best single basic statistic at predicting winning percentage
- Offensive Rating is the best single advanced statistic at predicting winning percentage
- Models built with only offensive statistics explained only a 54% variability in the overall win percentage.
- Models built with defensive statistics explained 58% of the variability in win percentage of a team.
- Models built with neutral statistics explained 89% of the variability in win percentage of a team.
- Of the 19 independent basic variables (game-play statistics) evaluated, 4 variables are to describe 88.8% of the variability in a team's winning percentage. The variables are as follows:
    - Scoring Margin (Stat Category: Neutral)
    - Opponents Three point Field Goal percentage (Stat Category: Defense)
    - Rebounding Margin (Stat Category: Neutral)
    - Personal Fouls Per Game (Neutral)
- Models built with the advanced statistics explained 88.7 percent of the variability in the win percentage which is .07% lower than primarily using basic variables.

Further details on each of these findings are provided below.

Best Single Basic Statistic
For the data analysis, the basic statistics measures were extracted and broken up into several categories: offensive, defensive and neutral. In order to identify the appropriate independent variables that explain the highest degree of variability in a team's winning percentage. We utilized the Akaike Information Criteria measure calculated with the forward selection linear modeling process to determine the best and most valuable single statistic for the linear model.

From our analysis, Scoring Margin was the independent variable without which the linear model would lose significant information and thus loses effective predictive value. The basic statistic scoring margin is calculated as the average difference between a team's score and the opponent's score. The statement is additionally supported by measuring the correlation value between scoring margin and win percentage. The data shows that scoring margin has 0.94 correlation value with win percentage which shows a strong, positive linear dependence of win percentage with scoring margin.

Additionally, a linear model built with only the Scoring Margin as the independent variable explains 88.6% of the variability in win percentage. After breaking up the dataset into training and test data, the linear model with only scoring margin was able to predict a team's win percentage with 94% accuracy.

Additional Models Evaluated

While scoring margin is a strong statistical measure, in an effort to dig deeper and conduct additional analysis, we reviewed offensive, defensive and neutral variables independently. To begin with the offensive variables, the four significant variables for predicted win percentage were Points per Game, FG percentage, Three pointers made, Free throw percentage and rebounds per game. These four variables were able to explain 53% of the variability in win percentage. The model was accurate 75% of the time in predicting the values in the test data.

Similarly, for the defensive variables, the important variables for predicting win percentage were Opponents Field Goal Percentage, Opponent 3 Point Field Goal Percentage, Opponent Average 3 Pointers Made/Game, Opponents Points per Game, Steals Per Game and Blocked Shots Per Game. ~ 58% of variability in dependent variables can be predicted by these variables.

Best single advanced statistic

Utilizing the advanced statistics, the independent variable ***Offensive Rating*** (ORtg) performed better than any other single advanced statistic in predicting winning percentage. Offensive rating is one of the most popular and commonly referenced advanced statistics, as we found it highlighted on every single site that we researched. This advanced statistic is calculated by normalizing the average number of points scored per 100 possessions. When performing forward selection, it was found to result in the lowest amount of information lost for a given model. On its own, it is able to explain 61.21% of variance in Winning Percentage.

However, what we found interesting, is that by combining other advanced statistics that considered turnovers, rebounds, and field goal efficiency, we were able to produce a model that was much more accurate and robust, explaining 88.73% of the variance in winning percentage and correctly predicting 95% of our test data. This robust model **did not** include this Offensive Rating statistic (when we tried to add it, the p-value for its test statistic was insignificant). Therefore, one might conclude that while scoring a lot of points does somewhat predict winning, strong defense, rebounding, and efficient scoring is a much stronger predictor for winning.

Are the "Four Factors" to winning basketball still relevant?

According to statistician Dean Oliver's 2004 book "Basketball on Paper: Rules and Tools for Performance Analysis" (2), the "Four Factors of Basketball Success" are:

1. Shooting (40%)
2. Turnovers (25%)
3. Rebounding (20%)
4. Free Throws (15%)

The number in parentheses is the approximate weight Mr. Oliver assigned each factor.

Using our data from NCAA men's regular season games from 2015 through 2019, we engineered variables corresponding to Mr. Oliver's "four factors" and did find we were able to build a model that explained 88.19% of variance in winning percentage.

However, we were also able to build a model, again using shooting, turnovers, and rebounds, but replacing free throws with TrueShootingPercentage which gives additional consideration for 3-pointers and less consideration for free throw percentage. This model performed slightly better, explaining 88.78% of the variance in winning percentage.

Therefore, one may wish to further analyze if free throws remain as critical an aspect for NCAA basketball winning success as Mr. Oliver proposed in 2004, or if perhaps the game has evolved and more importance should now be given to 3-pointers over free throws.

Conclusion

Splitting up the data into 4 logical partitions of offensive, defensive, neutral, and advanced statistics brought a lot of insights to our team. The first one that became quickly apparent to us was that offensive and defensive statistics alone were not enough to predict a team's winning percentage. Predicting a team's winning percentage using those stats alone produced models that only explained 58% of the data's variability for the defensive statistics and 54% for the offensive. Neutral and advanced statistics, however, produced models with adjusted R-squared scores of 88%.

With that in mind we identified the best basic and advanced statistics were scoring margin and offensive rating respectively. After keeping those statistical definitions in consideration, it made sense in hindsight as to why those statistics were the main driving forces in predicting a team's winning percentage. For example, scoring margin was defined as the average difference between a team's score and the opponent's score. It so happened to be the case that teams with higher positive scoring margins tended to have higher winning percentages; after all, teams that score more points, win. The same could be said for offensive rating which was defined as the average number of points scored per 100 possessions.

Equipped with this knowledge, we sought to test statistician Dean Oliver's hypothesis on his idea of the "Four Factors of Basketball Success". He claimed that shooting attributed to a team's success at 40%, turnovers at 25%, rebounds at 20%, and free throws at 15%. Using the best statistics which were previously identified to test this hypothesis, we produced a model that explained 88.1% of the variance in the independent variable which was able to correctly predict up to 95% of our test data. We also slightly improved the model by switching out the free throw statistic for 'TrueShootingPercentage' which gave additional consideration for 3-pointers and less consideration for free throw percentage. This slight improvement in model performance made us wonder if free throws are as important as Dean Oliver made them out to be or if 3-pointers are more relevant in college basketball nowadays.

References

1. https://www.reuters.com/article/us-basketball-ncaa-gambling/americans-to-bet-85-billion-on-ncaas-march-madness-basketball-tournament-report-idUSKCN1QZ0YH
2. https://www.basketball-reference.com/about/factors.html