Basketball: The Statistics Behind the Statistics

Colton Gearhart

Department of Mathematics and Statistics

Faculty Mentors: Dr. Carl Miller and Dr. Joseph Nolan

Introduction

Substantial research into predictive modelling for the NCAA Men's Basketball Tournament exists; less is readily available when it comes to predicting the outcomes of in-season games. The primary goal of this research was to estimate the probability that a given team will win a game based on how they and their opponent have performed thus far in a season. Results may be useful when it comes to game by game betting models.

Data

- Data from Division I college basketball games from 2008 – 2013 were used.
- Offensive and defensive statistics considered are shown in Figure 01.
- Data were collected at the player level; for each player who entered a game, statistics for that player recorded.
- Individual player-game data were aggregated to obtain cumulative statistics for teams as of the beginning of each game throughout the season (for use in predicting that game).

Methodology

Two simplifying assumptions were made to streamline model development:

- Seasons are independent.
- Successive observations are independent. To avoid dependence of two teams playing the same game, each game was considered from the perspective of the home team; games played at neutral sites were excluded.

Stepwise logistic regression was used to estimate winning probability. Five models, one trained on each season of data, were developed and then used for further analysis.

Analyses

The chart to the right shows variables significant in the models for each year ("+" indicates that a variable was significant). The following variables were significant in every model:

- Wins
- TwoPercent
- OffReb
- Turnovers
- Opp.Wins
- Opp.TwoPercent
- Opp.Blocks
- Opp.Turnovers

This suggests that they are the most important in predicting the probability that a team will win. The next step was to fine tune and assess these models.

Figure 01

		Model				
		809	910	1011	1112	1213
Variable	Wins	+	+	+	+	+
	GamesPlayed		+			+
	Points			+		
	TwoPercent	+	+	+	+	+
	ThreePercent					
	FTPercent		+	+		+
	OffReb	+	+	+	+	+
	DefReb		+	+	+	
	Assists					
	Blocks			+	+	+
	Turnovers	+	+	+	+	+
	Fouls					
	Opp.Wins	+	+	+	+	+
	Opp.GamesPlayed		+			+
	Opp.Points	+		+	+	
	Opp.TwoPercent	+	+	+	+	+
	Opp.ThreePercent	+	+			
	Opp.FTPercent			+	+	+
	Opp.OffReb	+	+	+	+	
	Opp.DefReb				+	
	Opp.Assists		+	+	+	
	Opp.Blocks	+	+	+	+	+
	Opp.Turnovers	+	+	+	+	+
	Opp.Fouls					
Number of Variables		11	15	16	15	13

Figure 02

	Actual Loss	Actual Win
Predicted Loss	A	B
Predicted Win	C	D

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}$$

Model performance was assessed based on the ability of each year's model to predict other years. Based on predicted probability of a win for the home team, each result was classified as either a win or loss.

Probabilities above a specified cutoff were classified as a win and those below were classified as a loss. Predictions were compared to actual results. The calculation for model accuracy based on cutoff is illustrated in Figure 02. This study examined two logical cutoff values in particular: 50% representing each game as essentially a coin toss and 65% which is the overall estimate for the probability that a generic home team will win, based on our data. Figure 03 illustrates the accuracy of each model relative to a specified cutoff.

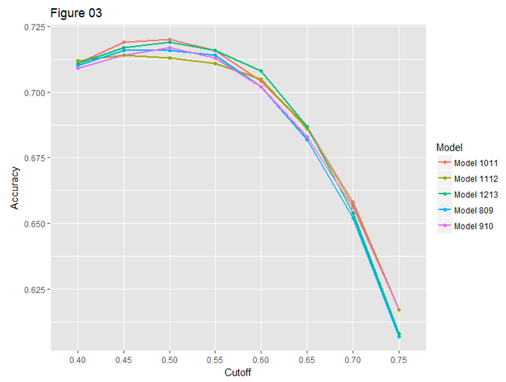
Results suggest that from a model perspective, each game is best viewed as a coin toss. This is somewhat to be expected as the models have already taken into account the perspective of the home team. Therefore, if the home team has a greater than 50% chance to win, it makes sense to classify that as a win.

Analyses were also conducted to examine accuracy and consistency of results across seasons. Additionally, models were examined to see if they do better in the extremes (for this part, prediction probabilities between 40% and 60% would be considered toss-ups and not be included in prediction. Results are shown in Figure 04.

Figure 04

	Model 809	Model 910	Model 1011	Model 1112	Model 1213
testData 809		0.754	0.758	0.752	0.756
testData 910	0.776		0.773	0.773	0.771
testData 1011	0.771	0.766		0.768	0.773
testData 1112	0.776	0.769	0.777		0.777
testData 1213	0.757	0.752	0.754	0.759	

Figure 03



For each model, the resulting accuracies when using different testing datasets are relatively similar. Variability across seasons is small, indicating that our models have strong precision and also seemingly validating our assumption that the seasons may be viewed as independent.

Results/Conclusion

A final model, which included the eight variables found to be significant in all five individual models, was trained using 70% of the data across all seasons with 30% withheld for validation.

- Using the 0.5 cutoff value, this model had an accuracy of 71.8%.
- Considering only the extremes (below 0.4 or above 0.6), this model had an accuracy of 77.0%.

Both models produce better results than the naive approach of picking the home team (resulting in 65% accuracy). Additionally, the year to year consistency appears to be quite strong.

Future Work

- While assumptions were made to simplify the development of the models, it is recognized that they probably are not fully satisfied. Future models might attempt to account for known dependencies.
- Models developed here do not account for spreads, money lines, etc. Future analyses might pull additional data related to betting and incorporate that into models of betting strategy.
- These models do not incorporate game-to-game variability, instead looking only at simple summary statistics for each point in the season. Future work might investigate whether including game-to-game variability could improve upon these results.
- Furthermore, one would expect these models to perform better as more data is available later in a season. Additional assessment is needed to evaluate this sort of "timing" effect.

Acknowledgements

- *Thanks to Dr. Miller for collecting/scraping the data which led to this project.*
- *This research was supported in part by the NKU Burkhardt Consulting Center.*