STAT 539, Data Analysis Proposal

Justin Gomez. Paul Harmon

3/28/2017

1 Background

NCAA College Basketball, including the regular season and the NCAA tournament, is a major cornerstone of college sports. Each year, from mid-November to mid-March, the best Men's and Women's college basketball teams face off against each other. Each teams' season culminates in a bid to enter the either the NCAA tournament or its subsidiary, the NIT. Sixty four teams start the tournament and after several weeks of play, the final two teams play for their respective championships.

While the games themselves have tremendous economic impact, the NCAA tournament's effects reach beyond the court. Broadcasting the tournament cost CBS nearly 10 billion dollars in 2016 (Forbes). The effect on the schools' academics is palpable, too; Smith (2008) notes that "a college's profile will increase with big-time athletics, as will the perception of the school itself". Indeed, some have even contended the increased visibility has impacts on academic metrics at those schools including increased student quality and retention. During the tournament, millions of fans fill out brackets to try and pick the winners and losers of each contest.

Naturally, methods for accurately picking winners and losers of college basketball games are

of interest for several reasons. For fans, predictive models can aid in filling out brackets in early March. For coaches, better understanding which factors influence the outcome of a game can help them understand how better to develop game plans and coach student athletes to victory. Finally, for campus administrators, developing models to assess and predict how a team might perform in an upcoming year may be helpful for management or marketing purposes. This analysis focuses mainly on the development of a binary regression model in order determine which factors may influence the probability of a home team winning a given basketball game.

2 Data Set Description

The data were obtained from a repository on Kaggle.com related to the March Machine Learning Mania competition in 2016. A great deal of information was originally pulled down, much more than we could ever use in any meaningful way. In the end, we have settled on fourteen predictor variables to include in our study of our single response: whether the home team won or not. Our data set contains more than sixty-four thousand games spanning fourteen seasons of basketball. Team rosters are continuously changing as students enter and leave programs for a myriad of reasons. This leads to teams being relatively unchanged within a single year, but year to year we could see very different teams, so it seems to make sense to account for this behind the scenes shift by adding a season variable to our model. Whether or not a game goes into overtime is also a variable of interest as it may provide some insight into the level of competition between the teams; games that go into overtime likely involve more evenly matched teams than those that do not, and thus the probability of the home team winning may be close to 50%. To get an idea for how many points winning teams might need to put up in a typical game, we will also include the winning team's score.From there,

our data consists of performance statistics for both teams summarized as either percentages or differentials. For percentage predictors, we have field goal, three-point, and free-throw percentages for both the winning and the losing teams. All percentages were calculated as the number of made shots divided by the number of atempted shots. Next, are our differentials. We will be considering total rebounds (offensive rebounds plus defenseive rebounds), assists, turnovers, steals, blocks, and personal fouls when calculating these differentials, and we will always subtract the losing team's game statistics from the winning team's (winner-loser). Altogther in this data set, we

3 Goals and Questions of Interest

We are interested in determining the factors that influence the outcomes of college basketball games in order to predict which teams are more likely to win a matchup. The key questions of interest are as follows.

1. Does home court advantage influence the outcome of a game, after controlling for on-court performance factors?

2. Do winning teams turn the ball over less than losing teams?

3. Do any of the variables measured modify the effect of other covariates of interest?

The key goal of the analysis is to create a good predictive model that can, with some reasonable degree of accuracy, generate predictions as to whether a team will win or lose a game. We plan to break the dataset into training and testing sets in order to assess the quality of predictions. Since the data pertain to multiple seasons, the training and testing sets are going to be chosen using a stratified random sampling method that evenly samples games form each season.

4 Preliminary Data Exploartion

5 Analysis Plan and Modeling