CS 216 Project Proposal

Andrew Weatherman, Cristina Sniffen, Brennan Hurd, Alan Suh, Andrew Gao

Introduction:

College basketball is among the fastest-growing and most-watched sports in the United States. A staggering 18.1 million viewers tuned in to watch the 2022 NCAA (National Collegiate Athletic Association) national championship. The total tournament averaged 10.7 million viewers over 67 games – a rate up 13% compared to a then-record 2021 session.[1] Since 1996, 14 of the past 24 NCAA tournament finals have drawn more viewers than the most-watched game in that year's NBA Finals series. And take away Finals series not featuring Michael Jordan, LeBron James, or Kobe Bryant, and the NCAA takes the viewership cake – every single time.[2]

While the popularity of men's college basketball shows no signs of slowing, neither does its rule book. Following complications caused by the COVID-19 pandemic, the NCAA revised a long-contentious rule: Players that transfered to a new school would be immediately eligible to play; no longer would they be forced to sit out an entire year. As expected, this opened a flood gate of player movement and major roster shake-ups. With this rule expected to cause downstream effects for a number of years to come, there is no better time to apply machine learning to publicly-available transfer data and statistics.

Research Questions:

Our analysis will focus on predicting the performance of high-major transfers using individual recruiting rankings, game-by-game data, and season-long averages for over 60 raw and rate-based statistics. Notably, we will investigate whether **transfer performance can be accurately and reliably predicted using public data**. *Performance* will be evaluated against individual player offensive rating – which is an estimation of points produced per 100 possessions. These data, as will be explained in the *Data Sources* section, are easily accessible using a member-built API. The analysis will leverage supervised machine learning with the gradient-boosted XGBoost model.

While many head coaches prefer to delegated important decisions to their intuition or eyes, analytics still commands a vital role in the meeting rooms of many top college programs. Using open-source data, as to ease resource inequality between teams, to model would-be transfer performance could prove to be an invaluable tool in the analytics shed for coaches and programs.

Data Sources:

Our research will source data from the public-facing `cbbstat` API – which is designed and maintained by group member Andrew Weatherman. Built atop a Fast API framework in Python, `cbbstat` provides extensive and clean men's college basketball data on ten-plus endpoints. The API returns tidy data that has already been cleaned, processed, and merged; removing the need for data scraping and wrangling will provide our team a surplus of time to be allocated towards building, training, and polishing the model.

`cbbstat` compiles completely open-source data from Barttorvik, ESPN, 247Sports, Rivals, and Verbal Commits. While the API framework was built in

[1] NCAA. *D1 Men's Basketbell Championship Game Sets Single Game Viewing Records*

[2] Sportico. *March Madness Championship Game Ratings*

Python, Andrew primarily used R to collect the raw data, process it, and clean it. With a desire for fully reproducible work, this will likely be the only data source used by our team. Other data from Synergy, CBBAnalytics, and KenPom would be useful complements but are locked behind a paywall.

The API also features a convenient R wrapper package – `toRvik`, similarly developed and maintained by Andrew Weatherman. While the language of choice is undecided, `toRvik` would be used to collect the data in R and `pandas` in Python.

Course Modules:

While the research will inevitably incorporate a number of course modules, three will be emphasized:

Probability (Module 3):

The probability module will be used to interpret the findings of our model. While no algorithm has been selected, we will train and develop a number of models that take numeric independent variables (recruiting rank, player game stats, player season averages and aggregates) to predict a numeric dependent one (individual offensive rating). The model with the best fit to the testing data set will be chosen as our model of choice.

Statistical Inference (Module 5):

We will use this module to investigate various groups within our data set to find statistically significant differences, if any exist, using hypothesis testing. We will do this by grouping players by characteristics including teams, conferences, and opponent strength in order to find underlying trends in which to explore further. Finding trends within our data will be beneficial for guiding our model toward better predictions by the use of feature selection. P-values will be used to gauge significance, with Anova testing used to find significance between several groups. This module will be utilized during the data investigation and feature selection stages.

Visualization (Module 8):

Likely, `ggplot2`, `gt`, and `Shiny` (R) will be leveraged to visualize model results – both statically and interactively. Our project members have experience with plotting in R and dockerize and shipping Shiny applications. While post-model visualizations ultimately conclude as public-facing mediums of research findings, the visualization module will be used throughout the life-cycle of the research. Before data is fed to a model for training, visualizing the set to identify any outliers or skews is vital to efficacy. Tufte's data visualization principles, with an added understanding of modern designs, will guide our work.

Collaboration Plan:

Between meetings, we plan to use text as our primary mode of communication. This will allow for quick, convenient interaction. Furthermore, we will use Google Drive for any written work that needs to be shared, such as this proposal. We met as a group to discuss the data set we would be using and brainstorm feasible research questions. Collectively, we decided that working with R rather than Python might work better for us, seeing as though this is the platform the group has the most experience with. Furthermore, the data file is an R script, which will make the conversion easier. Following this meeting, we used WeTransfer.com to share the dataset with each other so that we all have access to it. In general, we will be meeting on Monday evenings for about half an hour to talk through programming issues, questions, or written analysis relating to our project. Ideally, these meetings will be held in-person in a common, West Campus study area. However, if a group member(s) cannot attend in-person, we can set up a Zoom discussion for them instead. Due to travel plans, the meeting on Monday, October 21st will be entirely on Zoom. In addition to group meetings, we plan to work on the project individually on an as-needed basis (approximately an hour or two per week), but understand that this may need to be adjusted as the deadline approaches. Project responsibilities will be

decided and agreed upon on a weekly basis.

To share project data, code, and writing, we are using a public repository on GitHub. This allows us to track work and progress collaboratively, and provides strong version control mechanisms. Everyone in our group has used GitHub for other classes/personal projects, and feel comfortable with the software. Our project will be housed at the following GitHub repository: https://github.com/andreweatherman/CS216