MGT Madness - Team 15 Progress Report

Project Background:

Sports, especially the NCAA March Madness tournament, are popular for entertainment and analysis. The tournament attracts millions of viewers and spectators, generating significant revenue for the NCAA. In 2022, the tournament had 10.7 million TV viewers, 685,000 attendees, and generated an estimated $1.15 billion in revenue [1].

With millions watching and even more following on social media, March Madness is a great opportunity for advertising. With a huge captive audience, companies can convert consumers into loyal brand users, create brand exposure, and introduce new products and services. However, the same number of viewers do not watch all 67 games. So advertisers need to determine which games to focus advertising efforts on. Although this is a complex task, in-game attendance can be used to approximate overall “demand” for a game. For example, a First Four game with approximately 4,000 spectators will more than likely not have the same number of viewers as a Final Four game with around 70,000 spectators.

In addition, March Madness is a gold mine for sports betting. Firms such as DraftKings and FanDuel earn a commission with each bet made, in addition to revenue from their websites and app ads. Reliable game prediction models are essential to maximizing this revenue through improving lines and parlays, increasing the monetary value of bets made, and encouraging more bettors to use their website and app.

Predicting the Men’s Division I March Madness Tournament outcome is a challenging task for analytics enthusiasts. It involves forecasting 67 games played over three weeks, including possible upsets, player injuries, and changing data. Achieving certainty in this prediction process is difficult.

To predict the outcomes of matchups in March Madness, various analyses are conducted. Live win probabilities, calculated on a play-by-play basis, consider factors like remaining game time, score difference, possession, and pre-game win probabilities. The excitement index measures the rate of change of a team’s chance of winning during a game, impacting viewership. Elo ratings, which rate a team’s chances of winning based on factors like location, conference, and game type, are also used.

The goal of MGT Madness is to create machine learning models that reliably predict the attendance at NCAA tournament games and the outcome of games given different match ups between teams, other pre-game characteristics and player statistics.

Planned Approach and Models

In this project, we want to answer the following questions:

1. What factors are the biggest influencers on NCAA tournament game attendance?
2. What predictors can be used to accurately predict the outcome of NCAA tournament games?

To answer these questions, we plan to use a generalized linear model (GLM) to predict game attendance, and a probit regression model to predict game winners. Models will be trained, validated, and tested, and various comparison methods will be used to select the final models (accuracy, summary statistics, principal component analysis, decision trees, etc.)

Modeling Attendance:

Since attendance is a numeric value, a generalized linear or multiple linear regression model can be used to predict game attendance. Depending on how our data cleaning goes, we may use predicted % arena capacity in lieu of predicted raw attendance. This would make the model more realistic, and ensure that predictions do not go well over the capacity of a game’s arena.

In any case, we plan to test multiple combinations of factors to find the best model, using the root mean square error (RMSE) and adjusted R-squared as points of comparison.

Potential factors to investigate include:

- Team rankings (such as Massey Ordinal, Elo, AP Polls, Net Ranking, and Coaches Poll)
- Relative strength of each team’s conference
- “Excitement factor” of each team
- Game location relative to each team’s campus
- University enrollment of each team
- Game’s round in tournament

Of these factors, we hypothesize that game location, tournament round, and relative conference strength will be the most influential on game attendance.

Modeling Game Results:

Since basketball game results are binary (either a win or loss), a probit or logistic regression model is great fit. We will compare potential models using accuracy, precision, and specificity, among other statistics.

Potential factors to investigate include:

- Team ranking differential
- Team's quality regular season wins
- Game location relative to each team's campus
- Game's round in tournament
- Team efficiency (eg. offensive points per 100 possessions and points allowed per 100 possessions)
- Statistics over the last 5 games

We anticipate that each team's ranking differential, quality wins, location, and efficiency will be the biggest factors determining the win/loss outcome of a game.

Data Preparation and Cleaning

The project involved gathering data on game outcomes, betting odds, attendance, and AP poll rankings for college basketball games between 2012 and 2022. The data was collected using a combination of web scraping and querying APIs.

The data collection process began with using datasets from the `ncaahoopR` library, which included box score, play-by-play, rosters, and schedule files. The resulting data frame served as the basis for amending and creating other features.

Similarly, we used the `hoopR` package to retrieve betting data for each game using the `espn_mbb_betting(<game_id>)` function. Due to the size of the data files, this data had to be stored within three different files.

To collect data on AP poll rankings, the we wrote a web-scraping function that collected data from sports-reference.com using the `rvest` package. The data was then processed using the `tidyverse` package and `lapply()` function to combine scraped data frames into a single table.

For attendance data, we used the `game_id` identifier to query the ESPN API and retrieve data on attendance. We also attempted to use the `hoopR` package, which had data in its repository files, but ultimately used the ESPN API.

Finally, we processed and merged all of the collected data into one data frame using the `data_cleanup.R` script. The script included processing steps such as filtering and renaming variables, geocoding college locations, and merging data frames.

Data Source Overview:

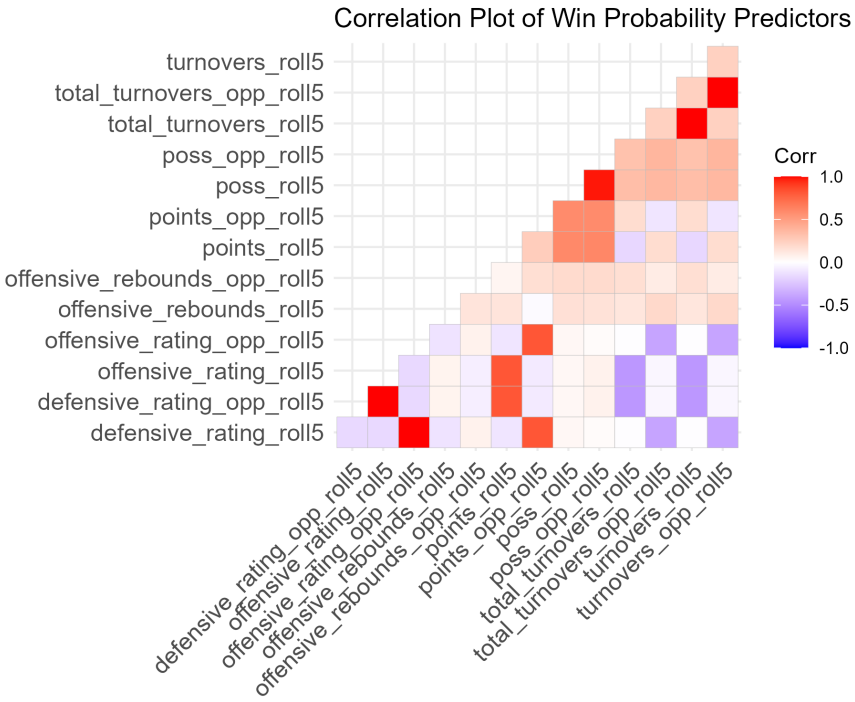
- ESPN API has a wealth of information ranging from team information (location, logo, colors, jerseys, etc.) to box-score and game-time information
- sports-reference.com contains AP poll ranking data in addition to other ranking sources
- The ncaahoopR data contains information around box scores, play-by-play and team information. The box scores and play-by-play data will be instrumental for our project. This will be one of the main sources for key statistics such as points scored, rebounds, steals, win probabilities, etc. Below is an example of one of the box scores
    - https://github.com/lbenz730/ncaahoopR (https://github.com/lbenz730/ncaahoopR)
    - https://github.com/lbenz730/ncaahoopR_data (https://github.com/lbenz730/ncaahoopR_data)
- NCAA Men's Basketball Data: - Records from 2000, including game attendance, team records per season, week-by-week Associated Press Poll Records - These records are mostly in .pdf format
    - https://www.ncaa.org/sports/2013/11/27/ncaa-men-s-basketball-records-books.aspx (https://www.ncaa.org/sports/2013/11/27/ncaa-men-s-basketball-records-books.aspx)
- The Kaggle datasets include: - Basic information: Team ID's and Names; Tournament seeds since 1984; final scores of all regular season, tournament games; etc. - Team Box Scores: game-by-game stats at a team level (free throws, rebounds, turnovers, etc.) - Geography: the city locations of all games since 2009 - Public Rankings: Weekly team rankings from multiple metrics - Supplemental Information: Coaches, conference affiliations, bracket structure, etc.
    - https://www.kaggle.com/competitions/mens-march-mania-2022/data (https://www.kaggle.com/competitions/mens-march-mania-2022/data)
    - https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset (https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset)

player_id	position	MIN	FGM	FGA	3PTM	3PTA	FTM	FTA	OREB	DREB	REB	AST	STL	BLK	TO	PF	PTS
4592187	F	30	5	8	0	1	7	10	6	4	10	2	1	1	2	2	17
4065653	F	12	0	1	0	1	0	0	0	2	2	0	0	1	0	1	0
4431669	C	28	6	12	0	1	6	7	2	4	6	1	0	0	2	5	18

- Because we were scraping data from ESPN and other sources, there were times when we were throttled and even blocked from pulling data. This was an unforeseen issue that ultimately required our team to take a step back and rethink our approach. We found a publicly available ESPN API that allowed us to pull the necessary data in a more efficient manner than web scraping.
    - There are many different ways for us to clean our data and one of the biggest challenges is determining what is right for our project. Do we want to use stats based on the last 5 games, 10 games? Do we want to focus solely on how a single team performs or do we want to compare how the team performs vs how the opponent performs? These are the kinds of questions that we've asked ourselves to determine how the data will be formatted for analysis.

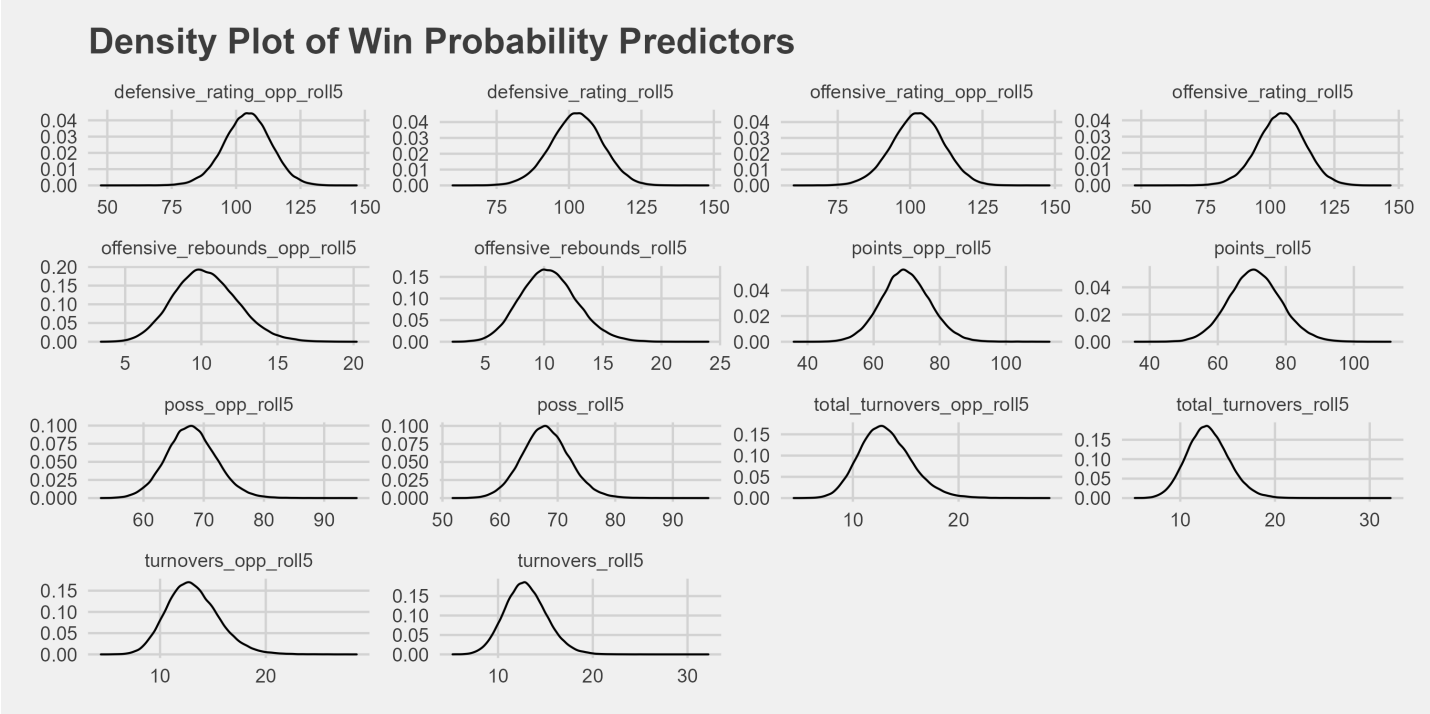
Data Exploration:

We created a correlation matrix of our factors. Due to how we calculated some of our features, many of the relationships were not surprising such as the highly correlated rating factors. However, things like points scored over the last 5 games didn't correlate with total turnovers in the last 5 games as strongly as expected. Meaning that points off turnovers isn't as large of a factor to points scored as anticipated.



Data Exploration - Correlation Plot of Win Probability Predictors

In addition to our correlations, we examined the distributions of factors to determine how well we hit normality assumptions and for scaling purposes in later analyses. Due to the large size of the data, our predictors were approximately normal.



Data Exploration - Density Plot of Win Probability Predictors

Initial Modeling

Initially, we aimed to predict game outcomes by creating models based on select features. We engineered new columns to transform data into a time-series format, focusing on recent games rather than future ones. Our exploratory model analyzed stats like the point differential, offensive/defensive efficiency ratings, turnover differential, and home/away/neutral status of teams in the last five games.

To establish a performance baseline for our own win probability predictive models, we used ESPN's contingency table predictions. We compared the actual outcomes of games with ESPN's predictions and found their accuracy to be approximately 73.26% (precision = 73.81%, specificity = 72.64%). This serves as a useful benchmark as we continue to refine our own predictions.

ESPN Confusion Matrix (omitting the first 5 games of each team of each season)

	Predicted Win	**Predicted Loss**
Actual Win	36,564	12,946
Actual Loss	12,976	34,445

Our initial win probability model was a logistic regression model through the `glm(, family='binomial')` function in R. We considered all factors from our correlation matrix. From this, we found that there was no real difference in terms of AIC between the logit and probit. The confusion matrix below shows that our Logit model's accuracy = 65.35%, precision = 65.78%, specificity = 63.59%.

Logit Model Confusion Matrix (omitting the first 5 games of each team of each season)

	Predicted Win	**Predicted Loss**
Actual Win	33,193	16,317
Actual Loss	17,267	30,154

As expected, ESPN does outperform our initial logit model. This will provide us with a good baseline to reference for our future modeling attempts.

Of note, the coefficient of the factor of home / away from our logit model initially confirmed the common assumption that there is a home field advantage. These results may change as we continue to develop our models.

Next Steps

The next steps involve building different models and performing validation for a dataset. The models that will be built include correlation analysis, principal component analysis, linear regression, and decision trees. These models will be used to predict attendance and win/loss outcomes in the dataset.

For **attendance prediction**, the Mean Squared Error (MSE) metric will be used for model validation. MSE measures the average squared difference between the predicted attendance values and the actual attendance values in the dataset.

For **win/loss prediction**, the dataset will be split into training, testing, and validation sets. The training set will be used to train the model, the testing set will be used to evaluate the model's performance on unseen data, and the validation set will be used to evaluate the model's final performance.

In addition to the train/test/validate splits, cross-validation will also be performed. Cross-validation is a technique used to evaluate the performance of a model by splitting the dataset into multiple parts, training the model on one part, and testing it on another. This process is repeated multiple times, with different parts of the dataset used for training and testing each time.

Finally, accuracy and recall metrics will be used to evaluate the performance of models predicting win/loss outcomes. Accuracy measures the proportion of correctly predicted outcomes, while recall measures the proportion of actual outcomes that were correctly predicted.

Works Cited

- [1] Bubel, Jennifer. "How Much Money Do Universities Get for Going to the NCAA March Madness Tournament?" *Diario AS*, 28 Feb. 2023, https://en.as.com/ncaa/how-much-money-do-universities-get-for-going-to-the-ncaa-march-madness-tournament-n/. (https://en.as.com/ncaa/how-much-money-do-universities-get-for-going-to-the-ncaa-march-madness-tournament-n/.)
- [2] Parker, Tim. "How Much Does the NCAA Make off March Madness?" Edited by Jefreda R Brown, *Investopedia*, Investopedia, 9 Mar. 2023, https://www.investopedia.com/articles/investing/031516/how-much-does-ncaa-make-march-madness.asp#:~:text=In%202022%2C%2045%20million%20Americans,see%20the%20heftiest%20cash%2Dout. (https://www.investopedia.com/articles/investing/031516/how-much-does-ncaa-make-march-madness.asp#:~:text=In%202022%2C%2045%20million%20Americans,see%20the%20heftiest%20cash%2Dout.)
- [3] Pomeroy, Ken. "The Possession", *Kenpom*, 19 Mar. 2004, https://kenpom.com/blog/the-possession/#:\~:text=The%20most%20common%20formula%20for,and%20FTA%20%3D%20free%20throw%20attempts (https://kenpom.com/blog/the-possession/#:~:text=The%20most%20common%20formula%20for,and%20FTA%20%3D%20free%20throw%20attempts)
- [4] Korpar, Lora. "March Madness Betting Expected to Exceed $3 Billion, Set All-Time High", *Newsweek*, 14 Mar. 2022, https://www.newsweek.com/march-madness-betting-expected-exceed-3-billion-set-all-time-high-1687917 (https://www.newsweek.com/march-madness-betting-expected-exceed-3-billion-set-all-time-high-1687917)
- [5] Coleman, Madeline. "March Madness: How a fan used a machine to nail his bracket", *Sports Illustrated*, 31 Mar. 2021, https://www.si.com/college/2021/03/31/march-madness-fan-trained-machine-predict-bracket-will-geoghegan (https://www.si.com/college/2021/03/31/march-madness-fan-trained-machine-predict-bracket-will-geoghegan)
- [6] Consoli, John. "Advertisers Go Mad For March Madness", *TV News Check,* 16 Mar 2022, https://tvnewscheck.com/business/article/advertisers-go-mad-for-march-madness/ (https://tvnewscheck.com/business/article/advertisers-go-mad-for-march-madness/)