MGT Madness - Team 15 Project Proposal

1. Michael Munson (GT Id: mmunson34). I currently work as a product development engineer for a company in Minnesota. My educational background includes: Physics (BS), Mathematics (BA), Teaching (MAT); I am working on my Certificate in Data Science for the Chemical Industries.
2. Matthew Rosenthal (GT Id: mrosenthal36). I'm a Technology Analyst for Southwest Airlines in Dallas, TX. In my current role I work to standardize and digitize workforce management processes for our airport employees. I have a BS in Industrial & Systems Engineering from the University of Florida with a Minor in Statistics (2019).
3. Jeffrey Sumner (GT Id: jsumner32). I'm a Data Engineer for Southern Star Central Gas Pipeline. My background is in Econ, Mathematics and Statistics. At Southern Star, some of the things we have work on are forecasting revenue, predict whether we are operating outside of recommended bounds, performing text analytics on survey data, etc.
4. Matthew Royalty (GT Id: mroyalty3). I'm a Data Engineer for Southern Star Central Gas Pipeline. I've been in this role for about 5 years developing data models and building ETL pipelines for our data warehouse initiative.

Background Information on chosen project topic:

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, including from sports betting. In 2022, the tournament had 10.7 million TV viewers, 685,000 attendees, and generated an estimated **$1.15 billion in revenue**, with **$3.1 billion** in sports wagers.

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. Similar ratings, such as **Dr. Joel Sokol's LRMC ratings**, can also be applied.

Problem Statement:

The goal of our project 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.

Primary Research Question:

Can we predict game attendance and outcome based on pre-game statistics? At what level of accuracy can we predict the outcome of games with the data that is currently available?

Supporting Research Questions:

1. What is the relationship between team rankings (Massey Ordinal, Elo, AP Polls, Net Ranking, Coaches Poll, etc.) and attendance?
2. What is the relationship between relative team rankings and the likeliness to win?

Business Justification:

Attendance: Companies with advertising and merchandising budgets strive to convert consumers into loyal brand users efficiently. The initial step of this conversion process is creating brand exposure, where people become aware of the brand's products. March Madness provides numerous high-exposure events, including tournament games and related content. However, not all games offer the same level of exposure, and predicting game attendance in advance will help companies optimize their ad placement for maximum potential conversion rates.

Win / Loss Predictions: March Madness is a significant opportunity for betting companies, as highlighted by approximately $3.1 billion wagered during the 2022 NCAA tournament, according to Fortune.com. These firms, such as DraftKings and FanDuel, earn a commission with each bet made, in addition to revenue from their websites and app ads. While attendance predictions have limited value, reliable game prediction models are essential. Accurate game predictions enable betting companies to improve their lines and parlays, increase the monetary value of bets made, and maximize their revenue.

Data Source Overview:

- 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

Key Variables

To model our research question(s), team ranking (Elo, Massey Ordinal, AP Polls, etc.), fan base size and game location (proximity to home) are likely to be independent variables to consider. Regarding the proximity to home, this data would have to be roughly estimated.

APPROACH/METHODOLOGY

Data Exploration:

We will explore the datasets listed above and determine which would be useful to answer our research question(s).

Data Cleaning:

- Eliminate 2020, 2021 Attendance - For our attendance analysis, we intend to remove 2021 NCAA tournament game attendance data due to the capacity limits due to COVID restrictions.
- Proximity to Home - We plan to calculate the distance between city of game and the participating schools' college campus location coordinates (latitude / longitude differential)
- Momentum - We plan to create a predictor for winning / losing streak (or last 5 or 10 games)
- Win Quality - If available in our data sets, we will utilize NCAA based quality of win statistic of quadrants ("Quad I, II, III, IV win / loss"). If not available, we will create a version of this with our own formula utilizing $PointDifferential \times (HomeOrAway)$
- Team Stat Differential - Team game results (Total Rebounds, total assists, etc.)

Types of Models:

Attendance:

- GLM
- Decision Tree

Win / Loss Predictions:

- Logistic/Probit regression
- Decision tree

How to compare models:

- RMSE
- Specificity/Sensitivity
- Hypothesis testing
    - ANOVA test

Known Hyper-parameters:

- Train-test split ratio; i.e. 70/30
- # of folds in cross validation
- # of trees in decision tree/random forest models
- Probability threshold/cutoff to confirm a win or loss
- # of components in PCA/PCR

Anticipated Conclusions/Hypothesis

Attendance: Multiple Linear Regression or equivalent(s) - We expect to figure out the most effective factors in modeling attendance. Then replicate attendance records to within an acceptable margin and compare to the current season's results. Initially, we anticipate the following to be important factors:

- proximity to home (college campus)
- conference strength
- round of tournament

Win / Loss Predictions: Utilizing Logistic Regression or equivalent, we are aiming for accuracy equal to or better than ESPN predictions. Initially, we anticipate the following to be important factors:

- team ranking differential
- win quality
- proximity of game to hometown
- team efficiency (eg. offensive points per 100 possessions and points allowed per 100 possessions)

Potential Business Impact and Benefits

Attendance: The analysis aims to optimize advertisement placements by predicting game attendance, enabling businesses to maximize their advertising budget. Accurate attendance predictions could also benefit the initial and secondary ticket markets by capturing additional revenue through demand-based pricing.

Win / Loss Predictions: As highlighted in the Business Justification, precise tournament game prediction models can improve betting companies' over/unders, opening lines, and other types of bets. Identifying factors that affect game outcomes can also enhance their in-game betting strategies. Accurately predicting the remaining game results (win/loss, specific stats, etc.) allows them to offer bets that gamblers want to place. Ultimately, providing more profitable bets attracts users to the betting site and increases revenue.

PROJECT TIMELINE/PLANNING:

- March 24: Data Prep procedures completed
- April 2: Group Project: Presentation video and progress report (as stated in syllabus)
- April 7: Models fully functional on datasets
- April 14: Model validation and comparisons completed
- April 16: Group Project: final paper, code and slides (as stated in syllabus)
- April 19: Group Project: final video (as stated in syllabus)

Appendix/Citation:

https://www2.isye.gatech.edu/~jsokol/lrmc/ (https://www2.isye.gatech.edu/~jsokol/lrmc/)

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/)