Basketball Logistics and Performance Indicator Analysis

Violet Dong, Ryan Mahtab, and Shurui Zeng
Carnegie Mellon University

Motivation
- Exploring the important factors that influence game outcome for Carnegie Mellon’s DIII men’s basketball team
- Logistical factors
- Performance Indicators
- Special emphasis placed on analyzing the effect of game time on game outcome

Background
- Carnegie Mellon is in NCAA Division III and belongs to the UAA conference in DIII
- A single basketball season is typically 20-25 games long, divided into non-UAA games and UAA conference games
- Non-UAA games happen in the fall semester and UAA conference games are in the spring
- The coach sets the schedules for non-UAA games a year in advance while the conference sets the schedules for UAA games seven years in advance
- The team typically commutes to non-UAA away games by bus and UAA away games by airplane

Data Sources
- All of our data sources span from the 2012-2013 season up to and including the 2019-2020 season
- Three main categories of data
  - Game Logistic Data
    - Schedules of each game detailing start times
  - Performance Indicators
    - In-game statistics
  - NCAA Division III end-of-season school rankings based on overall season record
  - Academic Schedule Data
    - Academic calendar PDFs with dates, events, and descriptions

Data Integration
- Season Statistics
  - Add score difference and away indicator
  - Scraped from athletics.cmu.edu
- Opponent Rankings
  - Merge opponent rankings onto season statistics based on the opponent of each game
- School Schedule
  - Scrap each column of the school calendar schedule data from a pdf to csv
  - Extract dates of exams and breaks by text parsing

Modeling
- SCOREDIFF ~ OPP_RANK + TIME_NUM + FG.PCT + X3PT.PCT + FT.PCT + OFF + DEF + AST + TO + STL + BLK + PF + diffToNearestExam + checkExam + diffToNearestBreak + TRAVEL
- The above model is a generalized linear model with a Gaussian response variable
- The response variable, SCOREDIFF is normally distributed
- It is assumed that the variables a linearly correlated with the response variable

Variables
- Combined all the variables from games logistics, performance indicators and academic schedule

Results
- Explain & Interpret Coefficients & significances
  - Game performance variables like FG.PCT, X3PT.PCT have strong positive correlations
  - Academic Schedule Variables also have influence on the score difference. Difference to nearest school break has a positive correlation with the score. One day further away from break will result in a 0.05 increase in score difference.
  - Game Logistics Data has a slightly significant influence on score difference. Travel by bus shows a 4.19 decrease in the score comparing to home games and one hour later game start would result in a 0.2851 increase in score difference.
- Accuracy
  - Average out-of-sample MSE of our model using 5-fold cross validation is 56.9605
  - Estimated Standard Error of this MSE is 8.5963
  - The reduced model SCOREDIFF ~ OPP_RANK + TRAVEL + diffToNearestBreak has an average out-of-sample MSE of 140.1486 and a corresponding Standard Error of 24.0486
  - Compared to the reduced model, our full model performs better at predicting Score Difference after taking into account game time and the performance indicators

Conclusion
- Although the time of the game is not the most statistically significant, the later the game, the better the team performance
- The further the game dates away from break dates the better the performance, controlling for game performance and logistics of game.

Acknowledgements and Sources
- We would like to thank Professor Nugent, Dr. Centor, Coach Tony and Stefanie Santo for their help and guidance throughout the project