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[1]**
  - Add score difference and away indicator
  - Scraped from athletics.cmu.edu

DATE	OPPONENT	SCORE	FG	PCT	3PT	PCT	FT	PCT	OFF	DEF	REB	AST	TO	STL	BLK	PF	PTS
Nov 12	Allegheny	L, 86-81	32-68	47.1	5-18	27.8	12-17	70.6	12	23	35	23	10	10	3	16	81
Nov 15	at Mount Union	L, 94-72	25-61	41.0	8-22	36.4	14-18	77.8	12	23	35	19	14	9	6	20	72
Nov 20	Bethany (W.V.)	W, 87-84	33-61	54.1	7-19	36.8	14-23	60.9	10	24	34	18	15	9	1	23	87

- **Opponent Rankings[2]**
  - Merge opponent rankings onto season statistics based on the opponent of each game

Rank	Team	W	L	Pct
1	Yeshiva (Styline)	29	1	96.7
2	Swarthmore (Centennial)	28	1	96.6
3	Randolph-Macon (ODAC)	28	2	93.3
3	Wittenberg (NCAC)	28	2	93.3
5	Saint John's (MN) (MIAC)	27	2	93.1
6	Mount Union (OAC)	27	3	90.0
7	St. Joseph (CT) (Great Northeast)	26	3	89.7
7	Brockport (SUNYAC)	26	3	89.7
7	St. Thomas (MN) (MIAC)	26	3	89.7
10	Neb. Wesleyan (American Rivers)	29	4	86.2

- **School Schedule[3]**
  - Scrape each column of the school calendar schedule data from a pdf to csv
  - Extract dates of exams and breaks by text parsing

2019-2020 Official Academic Calendar — **Carnegie Mellon University**

Spring 2020 Semester & Mini-4 - amended

Semester: *(M-14, T-15, W-15, Th-14, F-13) Total= 71*
Mini-4: *(M-7, T-7, W-7, Th-6, F-6) Total= 33*

Date	Day		Previous Date
March 18	W	Mini-4 Classes Begin	March 16
March 18	W	Summer 2020 Registration Begins	
March 27	F	Mini-4 Course Add Deadline	March 20
March 27	F	Mini-4 Course Audit Grade Option Deadline	March 20
March 27	F	Mini-4 Course Drop Deadline to Receive Tuition Adjustment	March 20
~~March 30~~	M	~~Semester Pass/Fail Grade Option Deadline~~ CANCELED	March 30
April 6	M	Semester Course Withdrawal Grade Deadline; No Course Withdrawal after this date	March 30
~~April 10~~	F	~~Mini-4 Course Drop Deadline; Assign Withdrawal Grade after this date~~	April 3

- **Game Schedule**
  - Scrape each column of the game schedule data from a Word doc to csv
  - Merge game schedule onto season statistics using the date column
  - Data provided by Coach Tony

Men's Basketball
2019-20 Schedule

DATE	OPPONENT	SITE	TIME
Nov. 12	**Allegheny College**	**Home**	**8:00PM**
Nov. 15	@University of Mount Union	Alliance, OH	8:00PM
Nov. 20	**Bethany**	**Home**	**7:30PM**
Nov. 23	@Marymount University	Alexandria, VA	3:00PM
Nov. 30	***Double Tree/Carnegie Mellon Invitational***	**Home**	
	Salisbury		**1:00PM**
Dec. 4	**LaRoche College**	**Home**	**7:30PM**
Dec. 7	**Penn State-Behrend**	**Home**	**2:00PM**
Dec. 19	**Washington & Jefferson**	**Home**	**4:00PM**
Dec. 21	**Chatham University**	**Home**	**2:00PM**
Dec. 31	@Walsh	North Canton, OH	1:00PM
Jan. 6	@Drew University	Madison, NJ	4:00PM

Modeling

Model

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.

Game Logistics	
OPP_RANK	Ranking of the opponent
TIME_NUM	Game time, recorded as military time
TRAVEL	How the team traveled to the game. It is a categorical variable (bus, plane or none)

Performance Indicators	
FG.PCT	Field goal percentage
X3PT.PCT	Three-point shot percentage
FT.PCT	Free throw percentage
OFF	Offensive rebounds
DEF	Defensive rebounds
AST	Assists
TO	Turnovers
STL	Steals
BLK	Blocks
PF	Personal fouls

Academic Schedule	
diffToNearestExam	Days away from the closest exam day
checkExam	If the game happens during exam week
diffToNearestBreak	Days away from the closest break

Coefficients and significances

Logistics Variables

	Estimate	Pr(>|t|)
(Intercept)	-100.9374	0.0000
OPP_RANK	0.0349	0.0000
TIME_NUM	0.2851	0.0881
TRAVELbus	-4.1907	0.0060
TRAVELplane	-2.5573	0.0722

Academic Schedule Variables

	Estimate	Pr(>|t|)
diffToNearestExam	0.0490	0.1600
checkExamTRUE	2.2836	0.2720
diffToNearestBreak	0.0506	0.0079
TRAVELbus	-4.1907	0.0060

Game Performance Variables

	Estimate	Pr(>|t|)
FG.PCT	0.7119	0.0000
X3PT.PCT	0.3230	0.0000
FT.PCT	0.1773	0.0000
OFF	0.6380	0.0000
DEF	1.0180	0.0000
AST	0.0403	0.8151
TO	-0.7933	0.0000
STL	1.1249	0.0000
BLK	0.2377	0.2737
PF	-0.0133	0.9223

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 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 a 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

- [1]https://athletics.cmu.edu/sports/mbkb/archive
- [2]http://stats.ncaa.org/rankings/change_sport_year_div
- [3]https://www.cmu.edu/hub/calendar/