Applying NFL Statistical Models to CMU Football

By: Eli Cohen, Jordan Gilbert, Marion Haney, Sarah Tandean

Project Advisor: Ron Yurko

Motivation

Statistical models are becoming increasingly important in decision-making in sports, especially in football where the focus is on NFL and Division I football. In this project, we aim to adapt models commonly used in NFL and Division I football analytics to assist Carnegie Mellon's Division III football team in making strategic decisions.

Project Goal: Create an Expected Points model based on play-by-play data that will aid the CMU coaches in gaining a better understanding of their team's success and efficiency.

Data

CMU-Only Data

- CMU Football coaches provided play-by-play data from CMU's 2022 season
- Data contained down, distance, yard line, and play type and result
- The data was missing important information necessary to create an expected points model, such as next scoring event for each play.

Cleaning Process

- Reformatted variables:
  - Turned the yard line variable into yards to opponent's end zone
- Manufactured variables:
  - Broke plays up into games and drives
  - Found the next score event for every play based on possession team

Web Scraping PAC Play-By-Play Data

- Scraped play-by-play data for all teams in President's Athletic Conference (PAC)
- Scraped games played by PAC teams during 2022 from the D3I Football website
- Used text parsing and regex techniques to transform play-by-play information into relevant variables such as down, distance, yard line, and play outcome
- PAC data allows us to create an Expected Points Model for the conference and compare the output to the CMU-only Expected Points Model.

Variable Name	Meaning
Next Score	The next scoring event with respect to the team in possession of the ball
Down	The current down for the team on offense
Distance to First Down	The distance the offense needs to get a first down
Distance to Opponent's Endzone	The distance the offense needs to score a touchdown

Table 1: Variables used in the analysis

Methods

- **Response Variable:** the probability of Next Score Type per play in the dataset.
- For simplicity we omitted safeties and extra points.

$$\text{Next Score Type} \in \{\text{Touchdown } (\mathbf{7}),\ \text{Field Goal } (\mathbf{3}),\ \text{No Score } (\mathbf{0}), -\text{Touchdown } (\mathbf{-7}),\ -\text{Field Goal } (\mathbf{-3})\}$$

- We used **multinomial logistic regression**.
- $\mathbf{X}$ corresponds to the game situation: Down, ln(Distance to First Down), Goal Down, and the interaction between Down and ln(Distance to First Down).
- β_y is the corresponding coefficient vector for Next Score Type.

$$\log\left(\frac{P(\text{Touchdown}|\mathbf{X})}{P(\text{No Score}|\mathbf{X})}\right) = \mathbf{X}\cdot\beta_{\text{Touchdown}},$$
$$\log\left(\frac{P(\text{Field Goal}|\mathbf{X})}{P(\text{No Score}|\mathbf{X})}\right) = \mathbf{X}\cdot\beta_{\text{Field Goal}},$$
$$\log\left(\frac{P(\text{Opponent Field Goal}|\mathbf{X})}{P(\text{No Score}|\mathbf{X})}\right) = \mathbf{X}\cdot\beta_{\text{Opponent Field Goal}},$$
$$\log\left(\frac{P(\text{Opponent Touchdown}|\mathbf{X})}{P(\text{No Score}|\mathbf{X})}\right) = \mathbf{X}\cdot\beta_{\text{Opponent Touchdown}},$$

- **Expected Points:**

$$EP = 7\cdot P(\text{Next Score Type} = \text{Touchdown}|\mathbf{X}) + -7\cdot P(\text{Next Score Type} = \text{Opponent Touchdown}|\mathbf{X}) + 3\cdot P(\text{Next Score Type} = \text{Field Goal}|\mathbf{X}) + -3\cdot P(\text{Next Score Type} = \text{Opponent Field Goal}|\mathbf{X})$$

Analysis & Results

Calibration of the EP Model:

- Predictions for Touchdown and Opponent Touchdown appear close to the observed probabilities.
- Field Goal and Opponent Field Goal less accurate, potentially due to less observed field goals.

Figure 1: EP model predictions vs. observed probabilities for each score type.

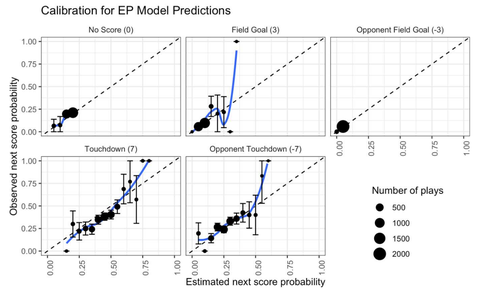


Figure 2: Comparison of scoring event probabilities over the course of the field.
CMU has a higher probability of an opponent scoring between 100 - 75 yds from the end zone.

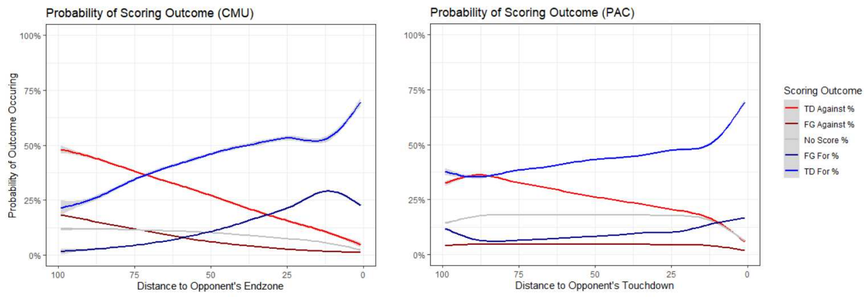
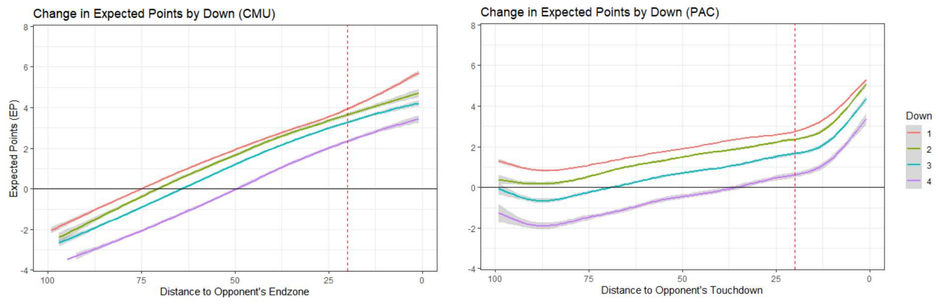


Figure 3: Comparison of expected points change over the course of the field by down.
EP increases as teams move closer to the opponent's end zone.



EPA Efficiency Metric

- **Expected Points Added (EPA):** EP(end of play) - EP(start of play)
- EPA > 0 means the play was efficient.

CMU Efficiency Metric	
Down	Efficient
1	4 Yards or more
2	Gaining half or more distance to 1st Down
3	Converting the 1st Down
4	Converting the 1st Down

Table 1: Two definitions of play efficiency; CMU coach's definition and EPA definition.

Table 2: Percent of offensive efficient plays (for all teams) based on the CMU-defined efficiency metric. **CMU is 9th out of 11 teams for this definition of efficiency.**

PAC Offensive Efficiency Ranking by CMU Defined Metric			
Rank	Team	% of Efficient Plays	# of Plays
1	Grove City	52.60%	812
2	Geneva	50.90%	708
3	St. Vincent	49.10%	598
9	Carnegie Mellon	39.40%	768

Table 3: Percent of offensive efficient plays (for all teams) based on EPA efficiency. **CMU is 8th out of 11 teams for this definition of efficiency. CMU is a bottom-ranked offensive team.**

PAC Offensive Efficiency Ranking by EPA			
Rank	Team	% of Efficient Plays	# of Plays
1	Grove City	44.70%	812
2	St. Vincent	37.90%	598
3	Geneva	37.60%	708
8	Carnegie Mellon	30.10%	768

Table 4: Percent of defensive efficient plays (for all teams) based on the CMU-defined efficiency metric. **CMU is ranked 3rd for this definition of efficiency.**

PAC Defensive Efficiency Ranking by CMU Defined Metric			
Rank	Team	% Efficient Plays Allowed	# of Plays
1	Washington and Jefferson	30.18%	748
2	Westminster	33.92%	794
3	Carnegie Mellon	39.43%	670

Table 5: Percent of defensive efficient plays (for all teams) based on EPA efficiency. **CMU is also ranked 3rd. CMU's defense is top-ranked in terms of efficiency in the PAC.**

PAC Defensive Efficiency Ranking by EPA			
Rank	Team	% Efficient Plays Allowed	# of Plays
1	Westminster	25.57%	794
2	Washington and Jefferson	27.32%	748
3	Carnegie Mellon	28.70%	670

Conclusions

- **Expected points increase per down** and as teams **move closer to first downs and the opponent's end zone**.
- **CMU has a much higher probability of giving up a touchdown between 100 and 75 yards** from the end zone compared to the PAC.
- **CMU doesn't have much of a boost in expected points when entering the red zone**, whereas the PAC at large does.
- CMU was, **offensively, one of the least efficient football teams** in the PAC by either metric in 2022, despite winning the conference.
  - However, they were **one of the most efficient defensively** - defense wins conference championships.

Limitations

- CMU data had rounding error in the distance to first down variable that propagated into the EP model
- Play-by-play data did not include time on the game clock
- Human-input errors of game information

Future Research

- **Incorporate more play information into the model**, such as play call, formation, personnel groupings, and defensive scheme
- **Add in player data to CMU's model** in order to allow for player performance evaluation and attributing parts of EPA to certain players
- Create a database to **continue scraping data during the upcoming 2023 season** and beyond

References

- Ron Yurko - Assistant Professor in Department of Statistics
- Ryan Larsen - Head Coach of CMU Football
- Andy Helms - Offensive Coordinator of CMU Football
- Zach Branson - Assistant Professor in Department of Statistics
- Joel Greenhouse - Professor in Department of Statistics
- Peter Freeman - Associate Teaching Professor in Department of Stat.
- Yurko, R., Ventura, S. & Horowitz, M. (2019). nflWAR: a reproducible method for offensive player evaluation in football. Journal of Quantitative Analysis in Sports, 15(3), 163-183. https://doi.org/10.1515/jqas-2018-0010

Carnegie Mellon University
Statistics & Data Science