UNITED STATES MILITARY ACADEMY

SCORE MODULE PRESENTATION

MA388: SABERMETRICS

SECTION D1

LTC MICHAEL POWELL

BY

CADET KEVIN BURKE `26, CO H2
CADET JOHN CLAUSING ’27, CO H2

WEST POINT, NEW YORK

5 MAY 2024

KBJC I CERTIFY THAT I HAVE COMPLETELY DOCUMENTED ALL SOURCES THAT I USED TO COMPLETE THIS ASSIGNMENT AND THAT I ACKNOWLEDGED ALL ASSISTANCE I RECEIVED IN THE COMPLETION OF THIS ASSIGNMENT.

_____ I CERTIFY THAT I DID NOT USE ANY SOURCES OR RECEIVE ANY ASSISTANCE REQUIRING DOCUMENTATION WHILE COMPLETING THIS ASSIGNMENT.

SIGNATURE:_____

_____

MA388 Sabermetrics: SCORE

NFL Touchdown vs. Field Goal Expectancy Matrices

CDT Kevin Burke and CDT John Clausing

1. Learning Goals:

The list of the learning outcomes for those who complete the module: (1) Understand how to build expectancy matrices for touchdowns and field goals in terms of distance and yards-to-go with NFL (American Football) data (2) Use a General Additive Model (GAM) to fit the non-linear scoring data for the touchdown matrix (3) Be able to read the matrices and explain the functions used to build them (4) Understand how to compare the matrices to determine the higher-scoring play-call in each situation

2. Introduction: This module seeks to answer this question: What are the expected points for a touchdown and field goal in each 4th down situation? Touchdown and field goal expectancy are the expected number of points that will be scored by the respective plays. The plays are subdivided into different 4th down situations by (1) yards until 1st down (anywhere from 1-10), also known as yards-to-go, and (2) the distance in yards from the goal line, which marks the beginning and front of the end zone (where touchdowns are scored). The expected points in these different 4th down situations are the generalized mean of the points scored in each situation. This is an important concept because it allows coaches and players to determine whether or not a field goal should be called instead of going for a 1st down when currently on 4th down. The data used is restricted to plays within 40 yards of the goal line (double the Red Zone distance). The Red Zone is defined as the last 20 yards heading into an opponent's end zone. The Red Zone distance was doubled in this module to provide more data points but mainly to make the decision between the two play options harder, since calling a traditional play is advantageous in the red zone. This module's results give an insight on the situations that are easiest to score a touchdown in and when it is advantageous to pursue a field goal instead. This matters significantly because in close games, going for and scoring a field goal can prevent a loss or overtime, but also if a team is playing the strategic long game, going for an early field goal can set them up to win in the end.

(see the section *4. Methods* for an in-depth explanation of GAMs and Matrices and how they will be used in this module.)

3. Data: https://www.kaggle.com/datasets/maxhorowitz/nflplaybyplay2009to2016

This data set was found on Kaggle through Google searching. It is detailed NFL play-by-play data from 2009-2018. It was compiled by Ron Yurko, Sam Ventura, and Mark Horowitz in order to makeup for the lack of publicly available NFL data sources. This is basically the equivalent of PitchF/x for baseball. It can be downloaded from the link above as an Excel (.csv) file. It can be imported to RStudio by placing it in your working directory and read-in using the notation found below in the example. The sample size of the file is 449,371. Variables (255 total) included are gameID, playID, side of field, teams, game seconds, drive number, downs, yardage line, yards to go, play type, yards gained, pass length, air yards, run yards, and more. The variables ranges' are dependent on their categories: air yards or total yards on a play can be negative all the way up to 100, but yards to go varies from 0 to around 20 at most. Other variables are binary and represented by a 0 or 1, like if the QB dropped back or not on the play. Below is a sample of the data and some of the variables included:

```
library(tidyverse)
library(mgcv)
library(broom)
```

```r
library(knitr)

#Read in the data from the csv file

footballData = read.csv("NFLPlayByPlay.csv")

#Columns are selected to provide a sample of the dataset

sample_example <- footballData %>%
  select(GameID, HomeTeam, AwayTeam, Drive, down,
         yrdln, ydstogo, PlayType) %>%
  filter(down == 4) %>%
  head(10)

sample_example %>%
  kable()
```

GameID	HomeTeam	AwayTeam	Drive	down	yrdln	ydstogo	PlayType
2009091000	PIT	TEN	1	4	44	8	Punt
2009091000	PIT	TEN	2	4	4	8	Punt
2009091000	PIT	TEN	3	4	41	21	Punt
2009091000	PIT	TEN	4	4	19	7	Field Goal
2009091000	PIT	TEN	5	4	21	16	Punt
2009091000	PIT	TEN	8	4	44	22	Punt
2009091000	PIT	TEN	9	4	38	5	Punt
2009091000	PIT	TEN	10	4	13	6	Field Goal
2009091000	PIT	TEN	15	4	45	1	Punt
2009091000	PIT	TEN	16	4	4	11	Punt

In the table above, the columns are variables selected for each data point (each play). They are filtered so only those plays that occurred on 4th down are available (notice there are either punts or field goals). Then, only the first 10 plays were taken, as this is just a sample to show what the data looks like.

4. Methods / Instructional Content:

https://www.espn.com/nfl/story/_/id/33059528/nfl-game-management-cheat-sheet-punt-go-kick-field-goal-fourth-downs-plus-2-point-conversion-recommendations

The article above from ESPN provides a graph based on a model that determines whether a coach should go for a 1st down, kick a field goal, or punt. The graph's variables include yards to go and yards to end zone. The model is considered more aggressive than the average coach (as stated by the author), as it goes for 4th downs most of the time, especially when within 3 yards of a 1st down. "ESPN's model considers the win probability expected given a fourth-down success and fourth-down failure, and weighs those by the expected conversion rate of that fourth down." Our model is simpler than this, but we still focus on the idea of when a FG is more beneficial than to go for a conversion. From this article we streamlined our module idea to create two expectancy matrices, one for touchdowns and a second for field goals, which can then be contrasted against one another every play to see which will score more points.

https://www.sportingnews.com/us/nfl/news/nfl-fourth-down-conversion-chart-rate-by-distance/vofkeub6xwms6imajxqkfipp

The article above provided the evidence that showed how significant our analysis could be. "Based on data going back to 2013, teams have had a success rate of 39.4 percent going for it on fourth-and-8. If teams wanted to try for it on fourth-and-9, that success rate dips a whopping 9.8 percent down to 29.6 percent." This stark contrast in conversion rate, especially when factoring in field position, can greatly affect what

play is called. This effected the content and structure of our model because it made us specifically have each yard to go be its own state in the matrix, due to how strong of an effect it has historically had on the league in regards to conversion rate.

For this module, we will use a Generalised Additive Model (GAM) to fit the data to produce our touchdown points expectancy matrix. In a similar manner to a linear regression, a GAM produces an output based off a set of input parameters. However, a GAM can be used to fit nonlinear data.

In practice, the GAM equation can be seen below where β_0 is an intercept, and each variable is the input to a function eg. $f_n(x_n)$.

$$g\left(E[y]\right) = \beta_0 + f_1(x_1) + f_2(x_2) + \ \ldots \ f_n(x_n)$$

An example of a GAM in practice is the expected points matrix for field goals. In this case, the expected amount of points is $g\left(E[y]\right)$ which is some function of the distance from the end zone $f_{YardLine}(x_{YardLine})$. See how to apply a GAM to the data in R below.

```
#Field Goal GAM

#Step 1 Isolate field goal data from the play by play data. The variables of interest are the FieldGoal

footballFG <-footballData %>%
  filter(PlayType == "Field Goal") %>%
  select(FieldGoalResult, FieldGoalDistance)

#Step 2 Create a GAM for Field Goal probability as a function of yardline.

gamFGMatrix <- footballFG %>%
   gam((FieldGoalResult == "Good") ~ s(FieldGoalDistance), family = binomial, data = .)

#Step 3 Create a dataframe for field goal probabilities.

FGMatrix <- data.frame(FieldGoalDistance = 1:40)

#Step 4 Apply the model to the field goal dataframe and add a column for the point value of a field goa

PredictedFG <- gamFGMatrix %>%
  augment(type.predict = "response",
          newdata = FGMatrix) %>%
  rename(FGProbability = .fitted) %>%
  mutate(ExpectedPoints = 3 * FGProbability) %>%
  select(FieldGoalDistance, ExpectedPoints, FGProbability)

#Step 5 Output the data to a table.

PredictedFG %>%
  kable(caption = "Table 2: Point Expectancy and Probability Matrix for a Field Goal
        given Yard Line", digits = 3)
```

Table 2: Table 2: Point Expectancy and Probability Matrix for a Field Goal given Yard Line

FieldGoalDistance	ExpectedPoints	FGProbability
1	2.997	0.999
2	2.997	0.999
3	2.996	0.999
4	2.996	0.999
5	2.995	0.998
6	2.994	0.998
7	2.993	0.998
8	2.992	0.997
9	2.991	0.997
10	2.989	0.996
11	2.988	0.996
12	2.986	0.995
13	2.983	0.994
14	2.981	0.994
15	2.978	0.993
16	2.974	0.991
17	2.970	0.990
18	2.965	0.988
19	2.960	0.987
20	2.953	0.984
21	2.946	0.982
22	2.938	0.979
23	2.928	0.976
24	2.918	0.973
25	2.906	0.969
26	2.893	0.964
27	2.878	0.959
28	2.862	0.954
29	2.843	0.948
30	2.822	0.941
31	2.798	0.933
32	2.772	0.924
33	2.744	0.915
34	2.713	0.904
35	2.679	0.893
36	2.645	0.882
37	2.609	0.870
38	2.572	0.857
39	2.534	0.845
40	2.497	0.832

5. Exercises/Activities:

Now that you have seen the fundamentals of how a GAM works and how it can be applied to map a nonlinear relationship with one parameter, we will do the same but now with two parameters to model touch down point expectancy based off of yardline and how many yards until the next first down.

```
# Gather data to find the success of touchdowns in the fourth down. (Isolate the desired dataset)
```

```
footballDataTD <- footballData %>%
  group_by(GameID, Drive) %>%
  mutate(pts_scored = ifelse(PlayType == "Field Goal", (FieldGoalResult == "Good")*3,
                             max(Touchdown)*6)) %>%
  mutate(pts_scored = max(pts_scored)) %>%
  select(down, yrdln, ydstogo, PlayType, pts_scored) %>%
  ungroup() %>%
  filter(down == 4, ydstogo <= 10) %>%
  mutate(ydstogo = ifelse(ydstogo > yrdln, yrdln, ydstogo)) %>%
  group_by(yrdln, ydstogo) %>%
  summarize(mean_points = mean(pts_scored),n = n()) %>%
  filter(!is.na(mean_points))
```

With the desired data isolated, create a dataframe for yardline and yards to go until the next first down. This will serve as the skeleton to which you will apply the GAM you create.

```
TDMatrix <- data.frame(yrdln = rep(1:100, 10)) %>%
  group_by(yrdln) %>%
  mutate(ydstogo = 1:10)
```

Now use the data isolated to produce a GAM for the expected touchdown values using the desired parameters. The code for the model should follow a similar format to the GAM created for the field goal matrix.

```
gamTDMatrix <- footballDataTD %>%
   gam((mean_points / 6) ~ s(yrdln, ydstogo), family = binomial, data = .)
```

Use the GAM to fill in the blank dataframe and multiply each value by the points of a touchdown to create the expectancy matrix.

```
PredictedPoints <- gamTDMatrix %>%
  augment(type.predict = "response",
          newdata = TDMatrix) %>%
  rename(mean_points = .fitted)

PredictedPoints <- PredictedPoints %>%
  mutate(Values = 6 * mean_points)

# Turn the dataframe into a matrix

TouchDownMatrix <- PredictedPoints %>%
  group_by(yrdln, ydstogo) %>%
  select(-c(mean_points, .se.fit)) %>%
  pivot_wider(
    names_from = ydstogo,
    values_from = Values,
    names_prefix = "YdsToGo"
  ) %>%
  ungroup()
```

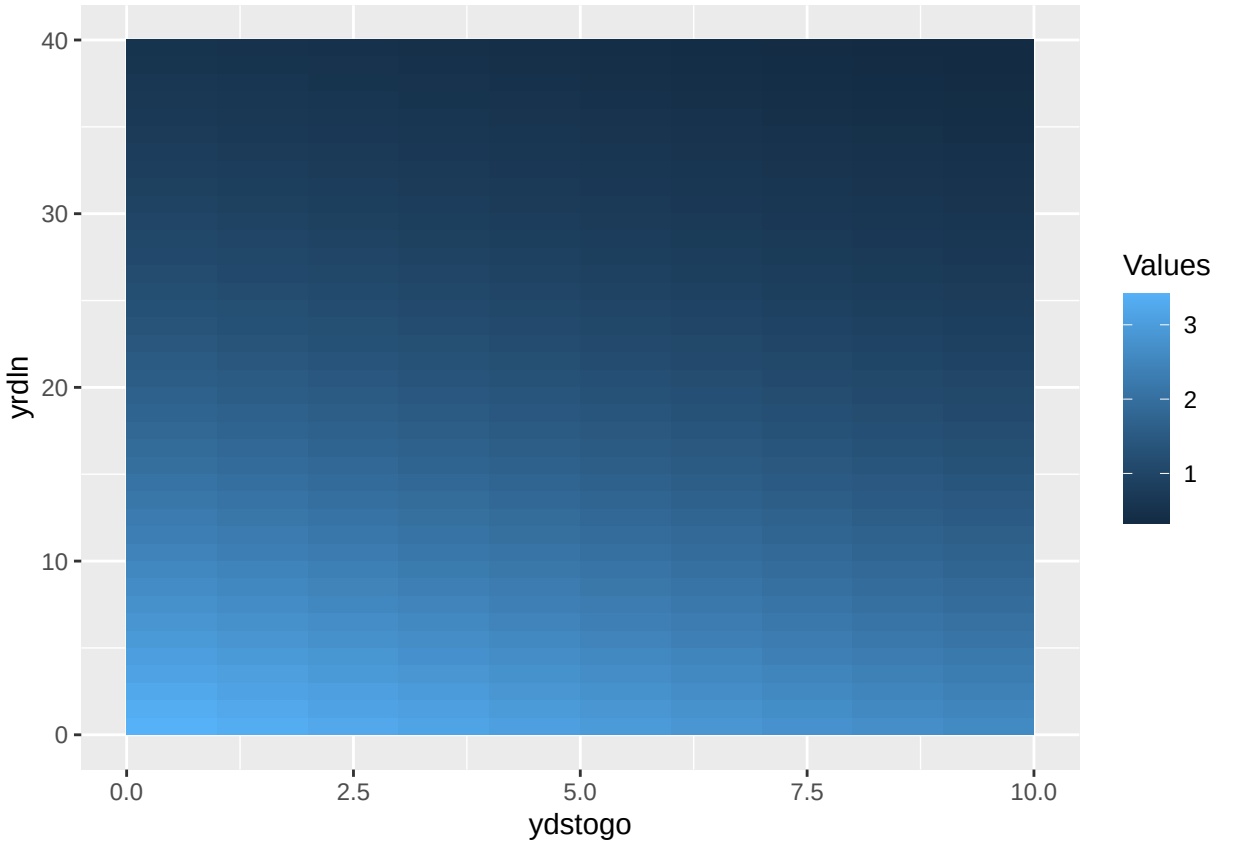
With this newly created expectancy matrix, you can now display the data using a table or heatmap.

```
# Output the matrix as a table

TouchDownMatrix %>%
  head(40) %>%
  kable(caption = "Table 1: Point Expectancy Matrix on Fourth Down given Yard Line and Yards
        until First Down", digits = 3)
```

Table 3: Table 1: Point Expectancy Matrix on Fourth Down given Yard Line and Yards until First Down

yrdln	YdsToGo1	YdsToGo2	YdsToGo3	YdsToGo4	YdsToGo5	YdsToGo6	YdsToGo7	YdsToGo8	YdsToGo9	YdsToGo10
1	3.419	3.327	3.233	3.139	3.045	2.951	2.857	2.763	2.670	2.577
2	3.322	3.229	3.135	3.041	2.947	2.853	2.759	2.666	2.573	2.481
3	3.224	3.131	3.037	2.942	2.848	2.755	2.661	2.569	2.477	2.386
4	3.126	3.032	2.938	2.844	2.750	2.657	2.564	2.473	2.382	2.292
5	3.028	2.934	2.840	2.746	2.653	2.560	2.468	2.378	2.288	2.200
6	2.929	2.835	2.742	2.648	2.556	2.464	2.374	2.284	2.196	2.109
7	2.831	2.737	2.644	2.552	2.460	2.369	2.280	2.192	2.105	2.020
8	2.733	2.640	2.547	2.456	2.365	2.276	2.188	2.101	2.016	1.933
9	2.635	2.543	2.451	2.361	2.272	2.184	2.097	2.013	1.929	1.848
10	2.539	2.447	2.357	2.268	2.180	2.093	2.009	1.926	1.845	1.765
11	2.443	2.353	2.263	2.176	2.089	2.005	1.922	1.841	1.762	1.685
12	2.348	2.259	2.172	2.085	2.001	1.918	1.837	1.758	1.681	1.606
13	2.255	2.168	2.081	1.997	1.914	1.833	1.754	1.677	1.603	1.530
14	2.164	2.078	1.993	1.910	1.830	1.751	1.674	1.599	1.527	1.456
15	2.074	1.989	1.907	1.826	1.747	1.670	1.596	1.523	1.453	1.385
16	1.985	1.903	1.822	1.744	1.667	1.592	1.520	1.450	1.382	1.316
17	1.899	1.819	1.740	1.663	1.589	1.517	1.447	1.379	1.313	1.250
18	1.815	1.736	1.660	1.586	1.513	1.444	1.376	1.310	1.247	1.186
19	1.733	1.656	1.582	1.510	1.440	1.373	1.307	1.244	1.184	1.125
20	1.653	1.579	1.507	1.437	1.370	1.304	1.241	1.181	1.122	1.066
21	1.575	1.504	1.434	1.367	1.301	1.239	1.178	1.120	1.064	1.010
22	1.500	1.431	1.363	1.298	1.236	1.175	1.117	1.061	1.007	0.956
23	1.428	1.360	1.295	1.233	1.172	1.114	1.059	1.005	0.953	0.904
24	1.357	1.293	1.230	1.170	1.112	1.056	1.002	0.951	0.902	0.855
25	1.290	1.227	1.167	1.109	1.053	1.000	0.949	0.900	0.853	0.808
26	1.224	1.164	1.106	1.051	0.998	0.946	0.898	0.851	0.806	0.763
27	1.162	1.104	1.048	0.995	0.944	0.895	0.849	0.804	0.761	0.720
28	1.101	1.046	0.993	0.942	0.893	0.846	0.802	0.759	0.718	0.680
29	1.043	0.990	0.940	0.891	0.844	0.800	0.757	0.717	0.678	0.641
30	0.988	0.937	0.889	0.842	0.798	0.755	0.715	0.676	0.639	0.604
31	0.935	0.886	0.840	0.796	0.753	0.713	0.674	0.638	0.603	0.570
32	0.884	0.838	0.794	0.751	0.711	0.673	0.636	0.601	0.568	0.537
33	0.836	0.792	0.750	0.709	0.671	0.634	0.600	0.567	0.535	0.505
34	0.790	0.748	0.707	0.669	0.633	0.598	0.565	0.534	0.504	0.476
35	0.746	0.706	0.667	0.631	0.597	0.564	0.532	0.503	0.475	0.448
36	0.704	0.666	0.630	0.595	0.562	0.531	0.501	0.473	0.447	0.421
37	0.664	0.628	0.593	0.561	0.530	0.500	0.472	0.445	0.420	0.396
38	0.626	0.592	0.559	0.528	0.499	0.471	0.444	0.419	0.395	0.373
39	0.590	0.558	0.527	0.497	0.469	0.443	0.418	0.394	0.372	0.350
40	0.556	0.525	0.496	0.468	0.442	0.417	0.393	0.371	0.349	0.329

```r
# Create a visual of the matrix

PredictedPoints %>%
  filter(yrdln <= 40) %>%
  ggplot(aes(x = ydstogo, y = yrdln, fill = Values)) +
  geom_raster(hjust = 0, vjust = 0)
```



Congratulations, you have now successfully produced touchdown expectancy and field goal expectancy matrices. With these, you can compare the expected point values to make a decision backed by the data. For example, at the 5 yard line with 1 yard to go, the expected value is 3.028 points which is higher than the value of a fieldgoal at the 5 yard line, 2.995 points. Therefore, there is a higher return for attempting the touchdown. However, at the 12 yard line with 9 yards to go, the touchdown expectancy is 1.681 points where as a field goal at the 12 is expected to generate 2.986 points. In this situation, it would be advantageous to take the field goal as opposed to attempting a touchdown.

6. Wrap-Up/Conclusions:

After completing the exercise in Section 5, one should be able to see two tables, produced by the expectancy matrices, and a heat map, based on the touchdown expectancy matrix. In the Touchdown expectancy matrix table, each value within the table is the expected points of 4th down plays in different field positions. As one moves from left to right on the table, the values almost always go down. As one moves from the top to the bottom of the table, the values almost always go down. Each value is the expected number of points that will be scored, which are the generalized mean of the points scored in each situation in actuality. The heat map (the gradient chart with a blue scale) depicts these results in color with the brighter blue being more points than the darker blue. The Field Goal expectancy matrix is simpler than the touchdown one as it only has one value for expected points per row (each row being the distance from the goal line). As one

moves down the table, the expected points slowly goes down, with the difference between 1 and 40 yards away being much closer compared to the the touchdown matrix's difference between 1 and 40.

An example of using these charts would be determining which play to call when 5 yards away. For a field goal, that has a set expected points value of 2.995 with a 0.998 probability of the field goal being made. On the other hand, the touchdown matrix presents the expected points for 5 yards in 10 different states, based on yard-to-go until a first down. With 1 yard-to-go, the expected points value is higher at 3.028, however, from 2 to 10 yards-to-go, it is lower, ranging from 2.934 down to 2.200 expected points. In conclusion, if you want to score more points on that specific play, or score in general, if you don't have 1 yard-to-go, you should kick a field goal because the expected points for a field goal are higher. But it isn't that simple. Although in that moment the field goal is better in terms of scoring in that position, that doesn't meant that a touchdown won't be scored after getting a littler closer and getting a few more first downs, but, you run the risk of turning over the ball if you don't convert the first down. Essentially, play calling comes down to long-term game planning, the current situation, and a team's preference based on their offensive options, which isn't accounted for in this module, but this module provides the data to inform decision-making.

By completing this module, the learning outcomes below were hopefully accomplished: (1) Understand how to build expectancy matrices for touchdowns and field goals in terms of distance and yards-to-go with NFL (American Football) data (2) Use a General Additive Model (GAM) to fit the non-linear scoring data for the touchdown matrix (3) Be able to read the matrices and explain the functions used to build them (4) Understand how to compare the matrices to determine the higher-scoring play-call in each situation

The first goal was accomplished in Sections 4 and 5, where in Section 4 matrices were explained and in Section 5 they were built with football in mind. In Section 5, you grouped 4th down plays initially by drives and assigned scoring outcomes so that one could filter the plays by yards-to-go and distance. Then you built a less complex matrix for field goal probability to compare the first to. The second goal was accomplished in Sections 4 and 5 as Section 4 explained them and in Section 5 you used one. In Section 5, you used a GAM to fit the touchdown data because the data itself isn't linear i.e. being closer and closer to the end zone will not increase the expected points by a constant amount every yard you get closer. Goal 3 was accomplished in the above part of Section 6 and 5 as you should be able to explain how matrices are built after Section 5 and we discussed the tables and heat map results within this section. Finally, goal 4 was also accomplished in the above part of this section.

Through accomplishing these lesson objectives, one can answer what the touchdown and field goal expectancy is in each 4th down situation and thus, which play is the best call.

Another way the exercise and information above can be applied in a different sport is observing the value difference between a 3-pointer and a 2-pointer in the NBA, specifically looking at an in-the-paint shot vs. a 3-pointer to see which is more effective. Today, the 3-pointer is a more efficient shot, but it wasn't always seen that way. A future skill that could be used to build on what's presented here is adding pace or number of possessions as variables and then looking at them through a "+" statistic to see how eras in football have affected them. This would show how different eras affected play calling and use of time, which in turn effected plays on 4th downs, displaying thoughts on the importance of general field position (specifically in regards to turnovers) vs. the importance of scoring position.

Acknowledgement of Assistance

Powell, Mike LTC. Assistance given to the authors, oral and written discussion. We met with LTC Powell to help us with our touchdown matrix because our result didn't make sense. Some scoring situations like 38 yards away and 10 to go had a higher probability to score then 15 yards away with 2 to go (a fabricated but logically correct example). LTC Powell helped us by correcting our code and instead of looking at individual plays, we grouped by drives and found the expectancy using that approach. Due to this, we assigned points to play types, allowing us to accurately track which plays (ones where scoring occurred) were being observed. He then recommended that we make a heat map of the data as well which we implemented in the Exercises section of our SCORE module. West Point, NY. 23APR2024.

Works Cited

"Detailed NFL Play-By-Play Data 2009-2018." *Www.kaggle.com*, www.kaggle.com/datasets/maxhorowitz/nflplaybyplay2009to2016.

"How to Format Number of Digits with Kable." *Stack Overflow*, stackoverflow.com/questions/66630454/how-to-format-number-of-digits-with-kable.

Kirenz, Jan. "Generalized Additive Models (GAM) — Introduction to Regression Models." *Kirenz.github.io*, kirenz.github.io/regression/docs/gam.html.

"NFL 4th down Conversion Chart, Explained: Breaking down the NFL's Success Rates by Distance & More to Know." *Www.sportingnews.com*, www.sportingnews.com/us/nfl/news/nfl-fourth-down-conversion-chart-rate-by-distance/vofkeub6xwms6imajxqkfipp.

Walder, Seth. "NFL Game Management Cheat Sheet: Guide to Fourth Downs and 2-Point Conversions." *ESPN.com*, 14 Jan. 2023, www.espn.com/nfl/story/_/id/33059528/nfl-game-management-cheat-sheet-punt-go-kick-field-goal-fourth-downs-plus-2-point-conversion-recommendations.