SPORTS
INFO SOLUTIONS

FOOTBALL ANALYTICS CHALLENGE

FOR THE

Elijah Cavan
William Ngana

BOYS & GIRLS CLUBS OF AMERICA

"The only sign we have in the locker room is from 'The Art Of War': Every battle is won before it is fought."

- Bill Belichick.

Part 1

Exploring the Data

Key Assumptions

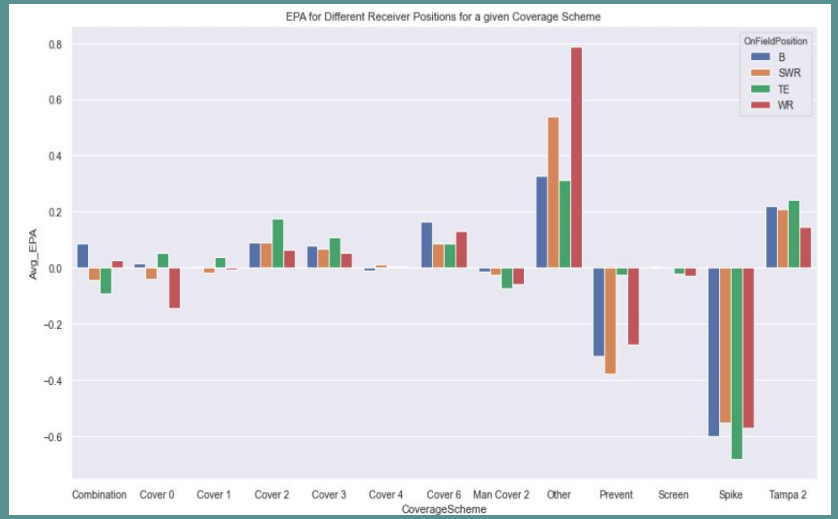
Q. Which route combinations were most popular in the NFL in 2020? Of these route combinations, which perform best against each coverage type?

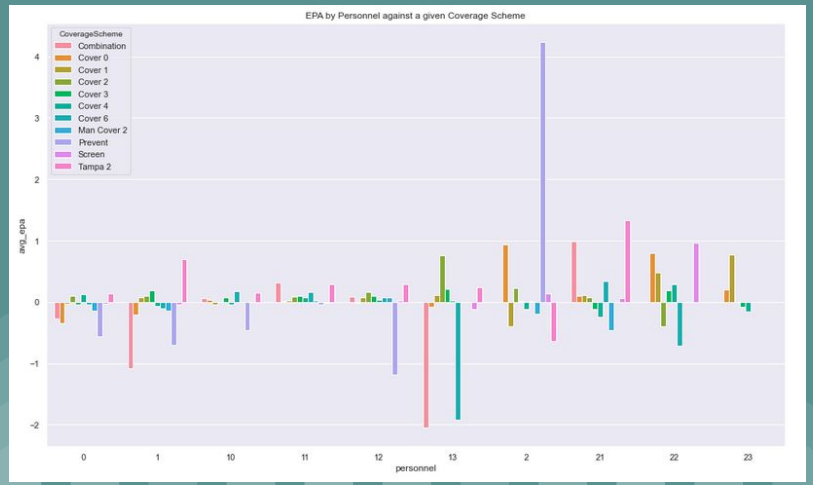
We assumed:

- A route combination was a list of routes run on the play
- Over a large enough sample, weather would play only a small part
- Turnovers, TDs were less sticky then pass break ups and first downs
- Receivers mainly played one position (i.e slot or out wide)
- All plays are equally weighted (i.e a 10 yard gain in the 4th quarter is the same as a 10 yard game in the first - although this is handled by EPA)
- A receiver is anyone who runs a route - including TEs and RBs
- Removed all plays where the QB is not on the field

What position excels against what coverage?

- Slot receivers are generally better than wide outs against Tampa 2. TEs are the best receivers in general against cover 0, cover 1 & cover 2. This makes sense since TEs are probably being covered by LBs in these schemes.
- We also see how the various personnel groupings fare against a given coverage; for example 21 personnel is most efficient at the combination and tampa 2 coverages (and useless for the cover type - i.e cover 0, cover 1 coverages)





Part 2

Cleaning and Engineering

Feature Engineering

We added a variety of features to help train the ML models and draw insights from the data. In particular, we calculated results which depended solely on the Coverage Scheme. Some examples of new features we used:

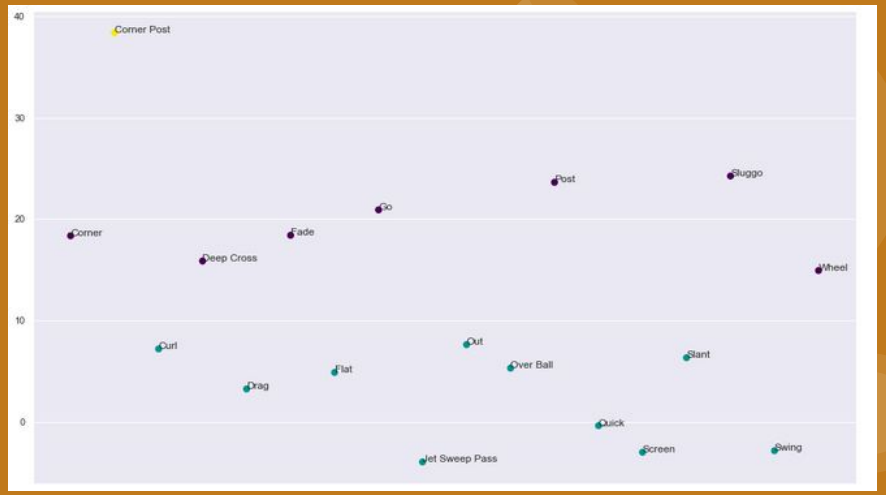
- Air yards
- YAC (yards after the catch)
- Pressure rate per Coverage Type
- Pass breakups per coverage type
- Receiver class (is a receiver especially good at running long routes, short routes, ect)
- Route length (is a given route long or short)
- Personnel groupings

Cleaning

- Consolidated several route types into a route class (for example rb-screen and bubble-screen are credited with the ‘screen’ class)
- Removed plays in which the QB is not on the field
- Removed players who played less than 50 snaps
- Removed the coverages ‘Spike’ and ‘Other

How long is that route?

In addition to defining new features, we also changed some of the pre existing columns. Here we are showing a scatter plot of the KMeans clusters for routes based on air yards (these are classified as Short, Medium, Long routes). As well as a histogram of Points above Average for WRs. This was done to simplify looking at the routes- there are less combinations if we replace the route names with the route length.

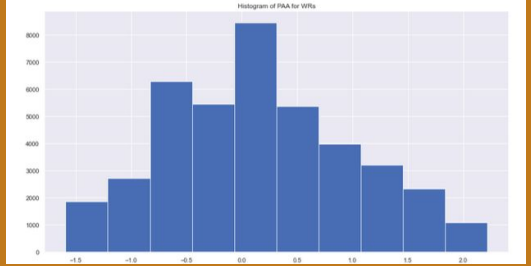


Sample of High Preforming WRs

['Calvin Ridley' 'Julio Jones' 'Stefon Diggs' 'Cole Beasley'
 'Keenan Allen' 'Michael Thomas' 'Tyreek Hill' 'Jarvis Landry'
 'Corey Davis' 'Davante Adams' 'Justin Jefferson' 'Brandon Aiyuk'
 'DeAndre Hopkins' 'Cooper Kupp']

Top WRs in Madden

	Name	RosterPosition	overall_rating
39966	Michael Thomas	WR	99.0
111101	DeAndre Hopkins	WR	98.0
7631	Julio Jones	WR	97.0
49192	Tyreek Hill	WR	96.0
101691	Davante Adams	WR	94.0



Part 3

Q. Part 1: What are the most Popular Routes/ Route Combinations

Most Popular Routes

```
Most popular route combinations 10 personnel:

(('Curl', 'Flat', 'Flat', 'Go', 'Out'), 31.0)
(('Curl', 'Curl', 'Curl', 'Flat', 'Flat'), 30.0)
(('Curl', 'Curl', 'Flat', 'Flat', 'Go'), 29.0)
Most popular route combinations 11 personnel:

(('Run Fake', 'rpo'), 115.0)
(('Curl', 'Curl', 'Curl', 'Flat', 'Flat'), 38.0)
(('Curl', 'Curl', 'Flat', 'Flat', 'Go'), 30.0)
Most popular route combinations 12 personnel:

(('Curl', 'Curl', 'Curl', 'Flat', 'Flat'), 21.0)
(('Run Fake', 'rpo'), 51.0)
(('Corner', 'Deep Cross', 'Flat', 'Quick', 'Run Fake'), 17.0)
Most popular route combinations 0 personnel:

(('Screen', 'Screen'), 54.0)
(('Curl', 'Flat', 'Flat', 'Go', 'Out'), 17.0)
(('Flat', 'Flat', 'Go', 'Out', 'Slant'), 15.0)
Most popular route combinations 1 personnel:

(('Curl', 'Curl', 'Curl', 'Flat', 'Flat'), 8.0)
(('Curl', 'Curl', 'Flat', 'Flat', 'Go'), 6.0)
(('Screen', 'Screen'), 14.0)
```

```
Most popular route combinations by route length:

(('long', 'short', 'short', 'short', 'short'), 1285.0)
(('short', 'short', 'short', 'short', 'short'), 1083.0)
(('medium', 'short', 'short', 'short', 'short'), 1004.0)
(('long', 'medium', 'short', 'short', 'short'), 980.0)
(('long', 'long', 'short', 'short', 'short'), 785.0)
(('short', 'short', 'short', 'short'), 740.0)
```

Note: The combinations for the rest of the coverage types can be found in the notebook

Here we show the most popular route combinations by personnel groupings and by Coverage type. Notice- the most popular route combinations are not always the ones that seek to maximize EPA.

Most Popular combinations for: Combination

	Route_list	personnel	mean	count
9	['Corner', 'Curl', 'Flat', 'Flat', 'Slant']	10	0.294190	10
24	['Corner', 'Flat', 'Flat', 'Out']	10	-1.178749	8
84	['Flat', 'Flat', 'Flat', 'Flat', 'Out', 'Swing']	10	-0.735372	6

Most Popular combinations for: Cover 0

	Route_list	personnel	mean	count
82	['Corner', 'Flat', 'Slant']	11	-0.016449	12
312	['Fade', 'Flat', 'Out']	11	-0.162710	12
294	['Fade', 'Flat', 'Flat', 'Out']	0	-1.204635	10

Most Popular combinations for: Cover 1

	Route_list	personnel	mean	count
2589	['Flat', 'Flat', 'Go', 'Out']	11	-0.284107	52
418	['Corner', 'Flat', 'Flat', 'Out']	10	-0.742640	48
2577	['Flat', 'Flat', 'Go', 'Out', 'Slant']	0	-0.890183	35

Most Popular Routes cont..

Most Popular Route length combinations for: Combination

	route_len_list	personnel	mean	count
44	['medium', 'short', 'short', 'short', 'short']	10	0.578493	45
29	['long', 'short', 'short', 'short', 'short']	10	0.895750	25
20	['long', 'medium', 'short', 'short', 'short']	0	-0.663793	25

Most Popular Route length combinations for: Cover 0

	route_len_list	personnel	mean	count
195	['medium', 'short', 'short']	11	0.157812	75
181	['medium', 'short', 'short', 'short', 'short']	0	-0.272336	65
182	['medium', 'short', 'short', 'short', 'short']	10	-0.042063	50

Most Popular Route length combinations for: Cover 1

	route_len_list	personnel	mean	count
611	['medium', 'short', 'short', 'short', 'short']	10	0.056856	565
682	['short', 'short', 'short', 'short', 'short']	10	-0.147920	515
500	['long', 'short', 'short', 'short', 'short']	10	-0.109546	480

Most Popular for Combination

	Route	route_len	mean	count
5	Flat	Short	0.071175	152
1	Curl	Short	0.387381	67
7	Out	Short	0.166135	52

Most Popular for Cover 0

	Route	route_len	mean	count
5	Flat	Short	-0.178837	616
1	Curl	Short	-0.011489	228
12	Slant	Short	0.026088	214

Most Popular for Cover 1

	Route	route_len	mean	count
6	Flat	Short	-0.011508	5128
2	Curl	Short	-0.084758	2880
7	Go	Long	0.011684	2078

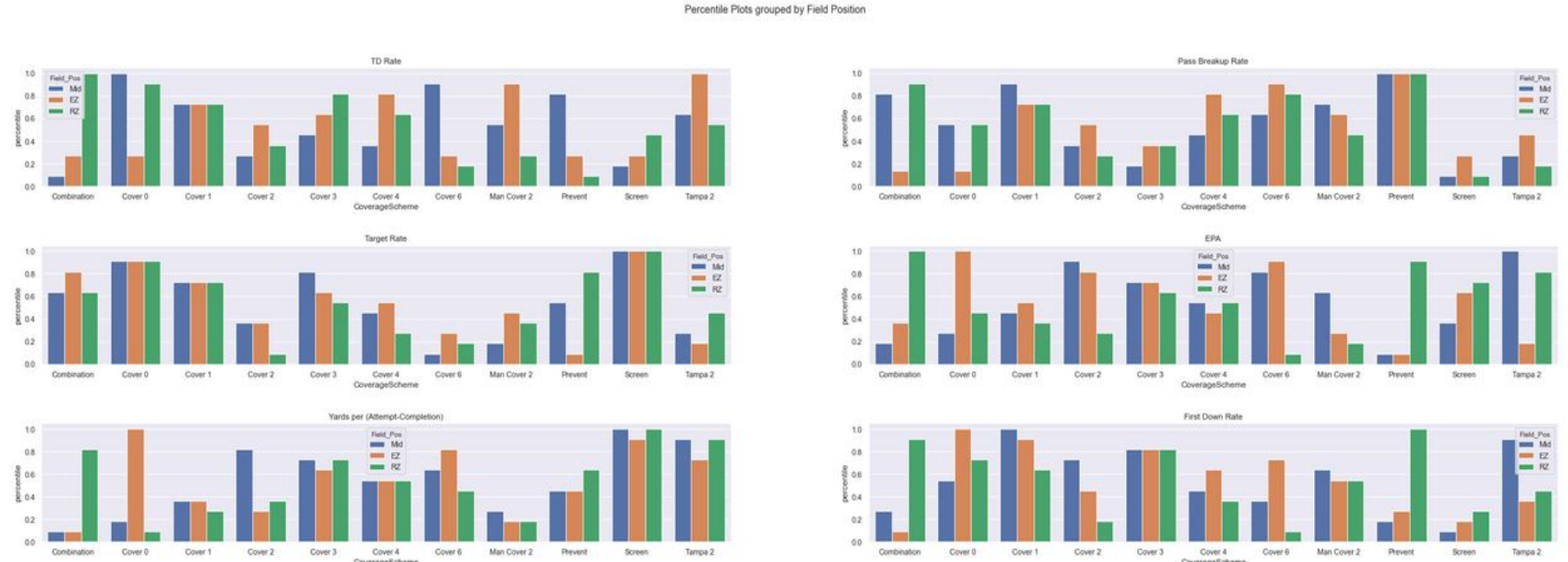
Part 4

Models & Analysis

Q. Part 2: Of these route combinations, which perform best against each coverage type?

Stacking Coverage Schemes Against One-Another

Plotting the most important features per coverage type based on percentiles relative to other coverages. Further split by Field Position (Redzone, Own Endzone, Middle of the Field)



Percentile Plots: Insights

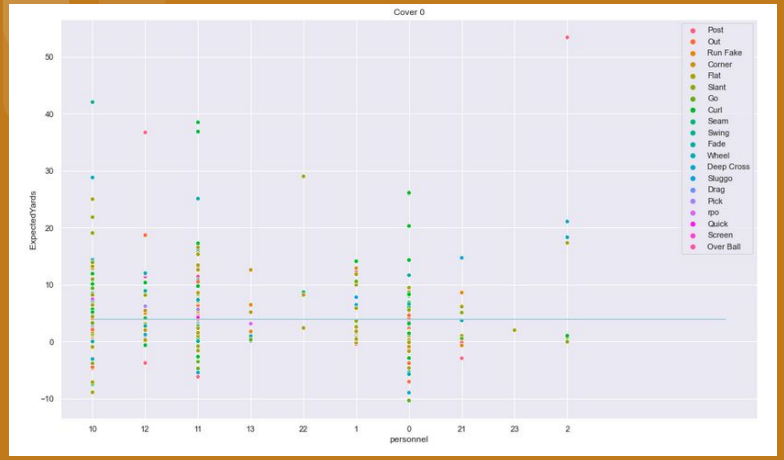
We have determined that:

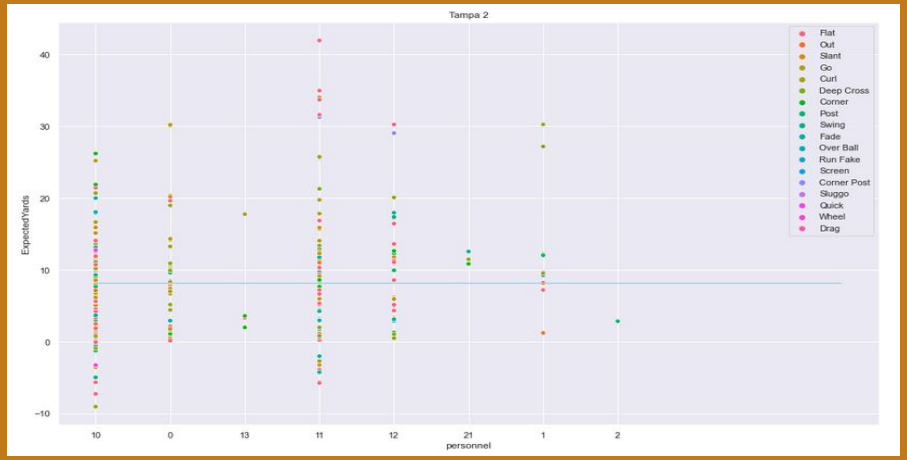
- The cover schemes (cover 0, cover 1 for example) tend to be more steady, while schemes like Tampa 2, Combination can wildly fluctuate in effectiveness based on field position
- The more DBs you have in coverage tends to limit targets; but you end up giving up high EPA plays/ many first downs
- Prevent has a high pass break up rate; but also a high target rate. This tells us that game situation plays a huge role in what play schemes and coverages are run.

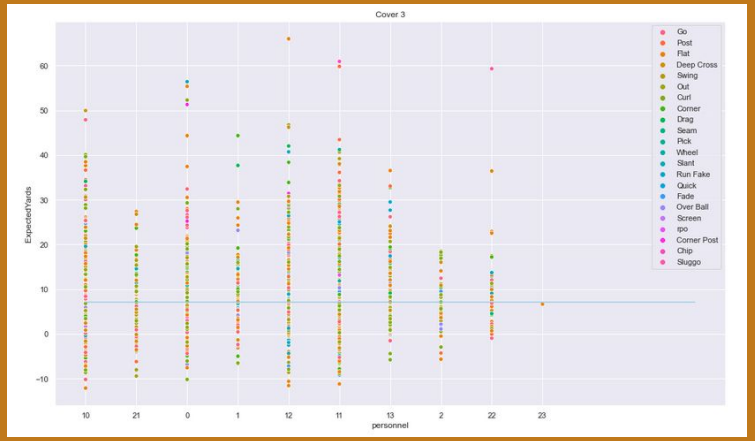
ML Models: General Info

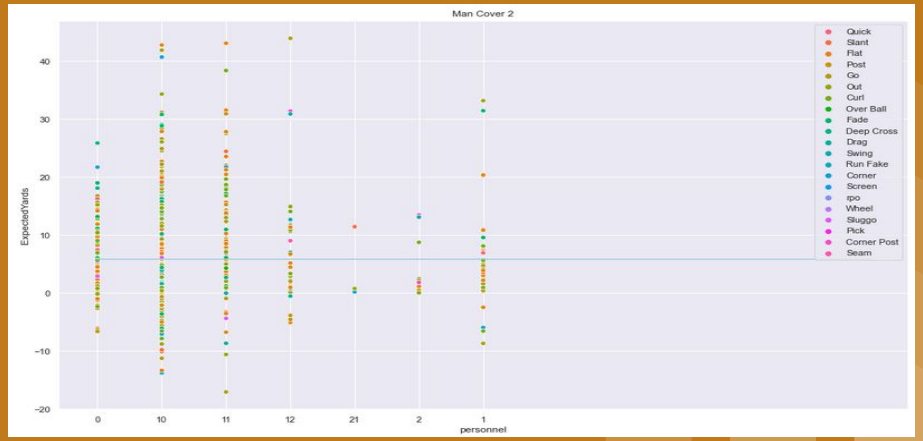
- Logistic regression and random forests regression models used
- Features used:
  - Team stats class: (ex. Point differential) allows the model to determine the difference between the two teams playing (including how good one offense is compared to the opposing defense)
  - Player stats (ex. Points above average) allows the model to distinguish between a good QB throwing to a good receiver v.s a bad QB throwing to a bad receiver
  - Cover Scheme class: coverage scheme statistics such as the average pass break ups or pressure rates for a given coverage scheme
  - Routes class: the route combinations and route lengths used on the play. Including if a receiver is running a ‘specialist’ route (a route the receiver is much better than the average receiver in running). Also the personnel grouping is included.

ML Model: Expected Yards (98% accuracy)







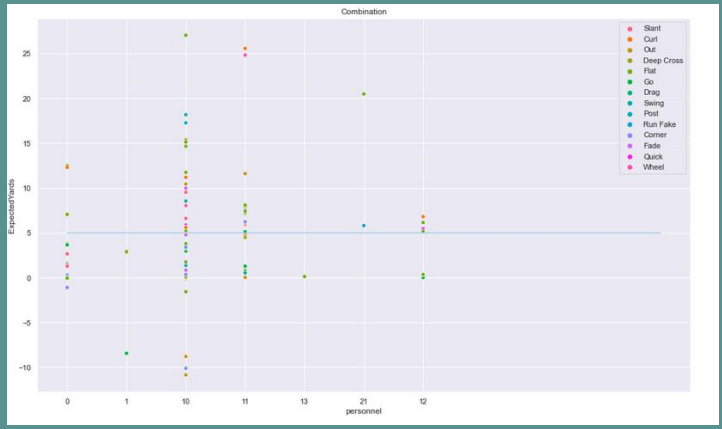


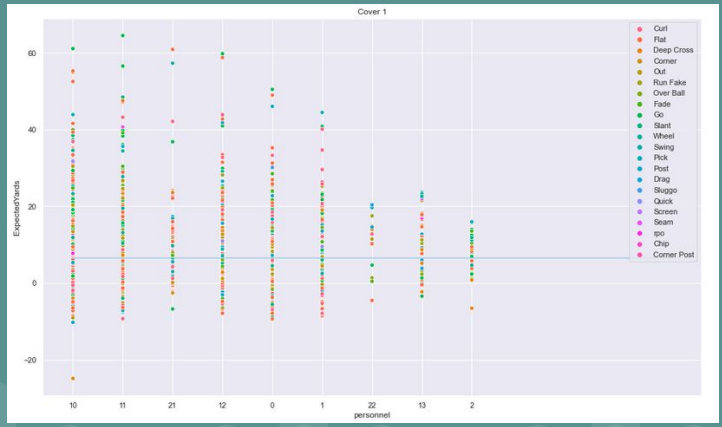
Expected Yards: Insights

The plots in the previous slide show a sample of the expected yards model (98% accuracy) versus personnel groupings and the targeted route on the play.We can draw several insights from these plots. For example: The combination scheme is least successful against 10 personnel, but does a good job versus 0, 12, 23 personnel. Routes that are effective against combination include slants, curls and flats.

Conversely, Cover 1 succeeds versus 22, 13 and 02 personnel; but is burned by 10,11,12 personnel for example. Routes that are effective against Cover 1 include Go routes, slants and flats.

The blue line represents the average offensive yardage against that coverage scheme





Best Route Combos by Expected Yards


```
Combination
mean      10.08
count      3.00
Name: ['Corner', 'Curl', 'Fade', 'Out', 'Slant'], dtype: float64
Cover 0
mean      4.415
count     6.000
Name: ['Corner', 'Flat', 'Flat', 'Slant'], dtype: float64
Cover 1
mean       3.9552
count     25.0000
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Flat'], dtype: float64
Cover 2
mean      10.422222
count      9.000000
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Go'], dtype: float64
Cover 3
mean       5.218485
count     33.000000
Name: ['Curl', 'Curl', 'Flat', 'Flat', 'Go'], dtype: float64
Cover 4
mean       6.62
count     20.00
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Go'], dtype: float64
Cover 6
mean      12.553
count     10.000
Name: ['Curl', 'Flat', 'Flat', 'Go', 'Out'], dtype: float64
Man Cover 2
mean      6.65625
count     8.00000
Name: ['Flat', 'Flat', 'Go', 'Out', 'Slant'], dtype: float64
Prevent
mean      12.076
count      5.000
Name: ['Go', 'Go', 'Go', 'Go', 'Go'], dtype: float64
Tampa 2
mean      16.147273
count     11.000000
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Flat'], dtype: float64
```

ML Model: Probability of First Down (97% accuracy)

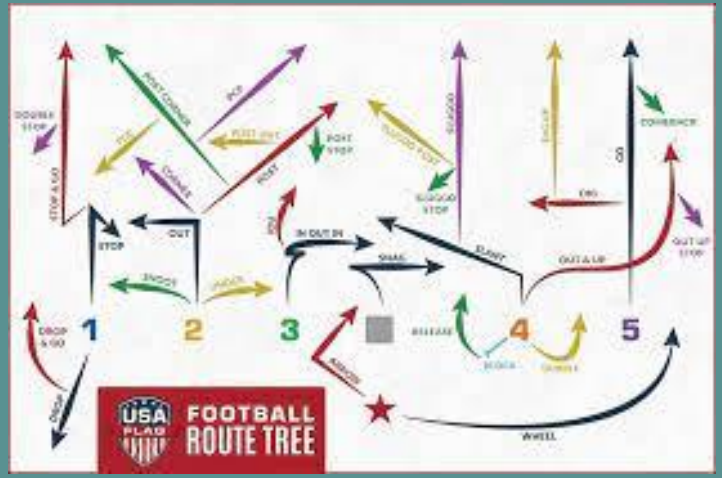
ML model which calculates the probability of a first down based on the route being run on the play (as well as a number of other features). The plots show the probability of a first down given that an optimal route is being run, compared to the average probability for all route combinations. In part 1 we showed the most popular route combinations; the optimal routes are now the routes that maximize EPA against a given coverage. The blue line represents the probability of a first down for the average route combination.

Probability of First Down: Insights

We have determined that:

- For every coverage scheme, running the optimal routes is more effective (obvious since they are chosen based on EPA). But for some coverages these optimal routes are more effective than for others (see EDA plot, receivers are not always the best position player to target for a given coverage).
- The cover schemes (as opposed to combination schemes, Tampa 2, ect) are the most effective schemes at limiting first downs

Best Route Combos by Probability of First Down




```
Combination
mean     1.0
count    3.0
Name: ['Corner', 'Curl', 'Fade', 'Out', 'Slant'], dtype: float64
Cover 0
mean     0.833333
count    6.000000
Name: ['Corner', 'Flat', 'Flat', 'Out'], dtype: float64
Cover 1
mean      0.615385
count    26.000000
Name: ['Flat', 'Flat', 'Flat', 'Go', 'Out'], dtype: float64
Cover 2
mean      0.619048
count    21.000000
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Flat'], dtype: float64
Cover 3
mean      0.545455
count    44.000000
Name: ['Curl', 'Flat', 'Flat', 'Flat', 'Go'], dtype: float64
Cover 4
mean      0.35
count    20.00
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Go'], dtype: float64
Cover 6
mean      0.7
count    10.0
Name: ['Curl', 'Flat', 'Flat', 'Go', 'Out'], dtype: float64
Man Cover 2
mean     0.625
count    8.000
Name: ['Flat', 'Flat', 'Go', 'Out', 'Slant'], dtype: float64
Prevent
mean     1.0
count    4.0
Name: ['Corner', 'Flat', 'Go', 'Out', 'Post'], dtype: float64
Tampa 2
mean      0.636364
count    11.000000
Name: ['Curl', 'Curl', 'Curl', 'Flat', 'Flat'], dtype: float64
```

Limitations and Improvements

There are several ways we could improve this analysis. We were also limited by the data.

- No tracking data (given us alternative ways to classify routes)
- Only 1 year of data
- We did not take into account how often a receiver plays, or how a good receiver might ‘warp’ the defence
- We did not take into account how running a play multiple times decreases its effectiveness
- We didn’t look at different formation types (21 personnel v.s 12 personnel for example)
- We did not account for the strength of a given DB and how that might affect plays run by the offense
- Is a route effective or is the defense conceding the play. For example, if I get an 8 yard gain against a ‘prevent’ coverage; is this an effective play or am I doing what the defense wants me to do.

References: Additional Datasets Used

Madden ratings: https://maddenratings.weebly.com/

Strength indicators:
https://www.pro-football-reference.com/years/2020/index.htm

That’s all Folks!