BSA Football Research

Fall 2024

Research Question

What are the best strategies defensive play callers can use to maximize defensive EPA during “clutch time”?

- *How can we predict EPA in a single play based on a given scenario?*
- *How can we predict what the offense will do?*
- *How can we account for the past 58 minutes of the game when making our predictions?*
- *How can we account for variability in the quality and playstyles of opposing QBs?*

What is “clutch time”?

We are defining “clutch time” as situations where:

- Two minute drill (exactly 120 seconds remaining in the game)
- The point differential is less than or equal to 8 points

What is EPA (Expected Points Added)?

- *Expected points (EP)* is the average number of points yielded at the end of the drive based on field position, down, and distance,
- After the play, *EPA = new EP - old EPA*
- Eg: On their own 35, the EP from drive roughly 0.5 points, but on the opponent’s 35 EP is roughly 2.5 points. If a play starts at the own 35 and gains 30 yards, *EPA = 2.5 - 0.5 = 2*

Data

1. Dataset #1: nflverse play by play
   - This contains formation data, pressure, coverage-type
   - Focus on 2018-2023 → rule changes created different play styles
1. Dataset #2: nflverse comprehensive data
   - This contains data for each game, 1999-2024
   - Contains required EPA data, timestamps within games (372 columns)

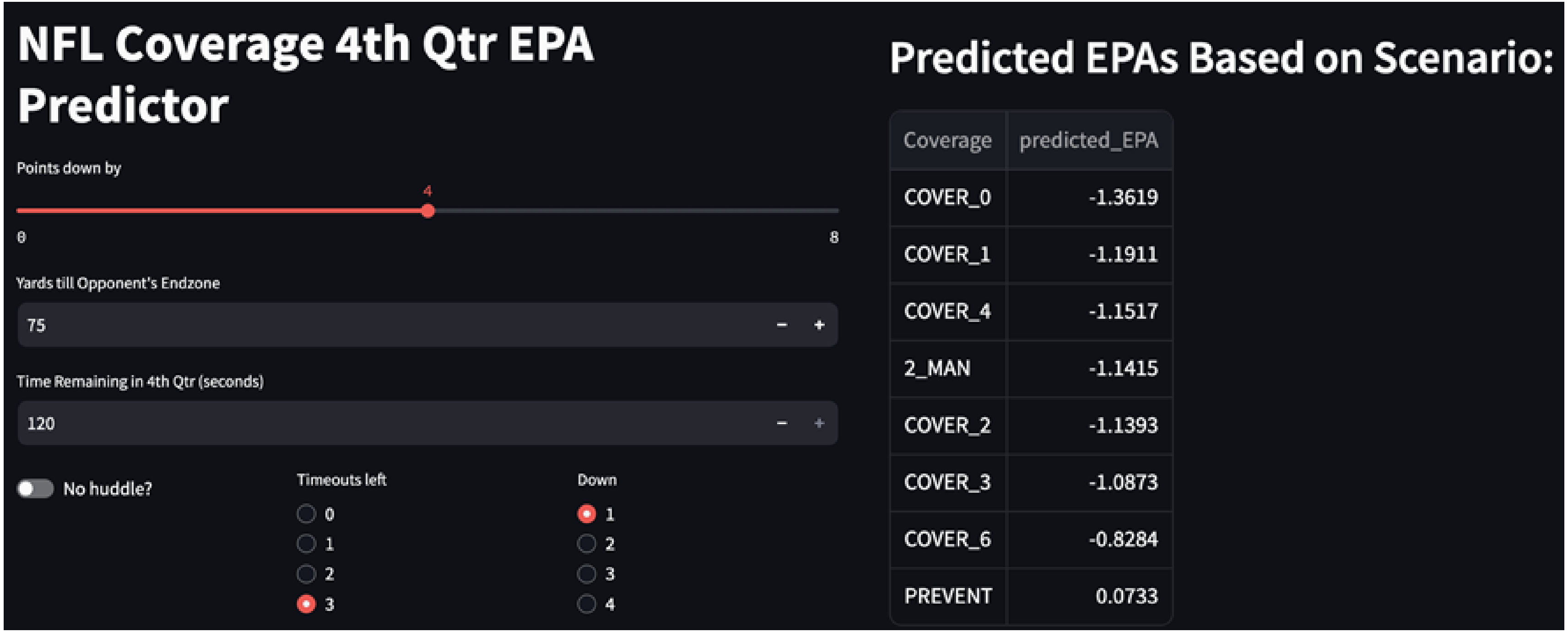
XGBoost Model Background

- This regression model predicts **passing** EPA against all coverages given a situation and route targeted.
- Used XGBoost because it efficiently handles complex data and reduces potential complications.
- Selected several features to represent a specific potential scenario in a play.
- Uses datasets 1 and 2, mostly 1 due to coverage data
  - Only uses plays during clutch time
- It evaluates a given play in a vacuum and does *not* account for the previous 58 minutes of the game.

XGBoost Model Features

Situational	**Defensive Properties**	**Offensive Properties**
Points down by (0-8)	Number of defenders in box	Pass length (short/long)
Down	Number of pass rushers	Pass location (left, middle, right)
Distance to 1st down	Time to throw	Did play go out of bounds?
Yardline	Was QB pressured?	No huddle?
Timeouts left	Coverage played	Route targeted
Seconds remaining (0-120)		

XGBoost Model App



One can input a scenario to the model using this web app constructed with Streamlit

XGBoost Model Evaluation

Model Metric Evaluation:

- *Mean Absolute Error (MAE):* **0.782**
- *Mean Squared Error (MSE):* **1.472**
- *Root Mean Squared Error (RMSE):* **1.213**
- *R-squared (R^2):* **0.207**

Baseline Metric Evaluation:

	Mean Baseline	Median Baseline
Mean Absolute Error (MAE)	**1.885**	**1.908**
Mean Squared Error (MSE)	**0.904**	**0.892**
R-Squared (R^2)	**0.0**	**-0.012**

Bayesian Route Classifier

Goal

- Use play by play data (via Next Gen Stats) to classify the route concept of individual plays prior to the snap
- Trained on a random subset of the data from 2018-2023

Final Method

- Used a naive Bayes classifier for multiclass prediction
- Chosen over more complex and tree-based models due to model simplicity and interpretability
    - Accuracy gains were marginal in comparison to changes in these factors (despite assumption of feature independence)

Bayesian Route Classifier

Model Results

- Accuracy of 0.3817 on testing data
- No Information Rate: 0.1613
- P value (Accuracy > NIR): < 2.2e-16

Most Important Predictors (in order of importance)

- Team of possession
- Defensive personnel
- Defensive coverage
- Offensive formation
- Offensive personnel

Bayesian Network (Joint Probability)

- **Advantages**
  - Computationally resourceful
  - Versatile
  - Reliable

- **Goal**
  1. Split by scenarios
  2. Find weights of regressors
  3. Run predictions

Bayesian Network (Joint Probability)

- **Results**
  - MLE
  - MAE
    - Normalization
  - MSE
  - R^2 (Unreliable)
  - STDEV and MEAN

- **Features**
  - Play type
  - Time (seconds)
  - WP
  - Down and distance

QB Clustering

- Goals of clustering
  - reduce complexity
  - wanted to account for QB quality, playing style, mobility
- Features used
  - Style: Attempts/game, Aggressiveness%, Time to throw, intended air yards/attempt
  - Mobility: Rushing Y/A, Y/G, TD/G, A/G
  - Quality: QBR
- Methodology
  - Scaled Data, removed outliers
  - kMeans clustering w/ elbow technique to select # of clusters

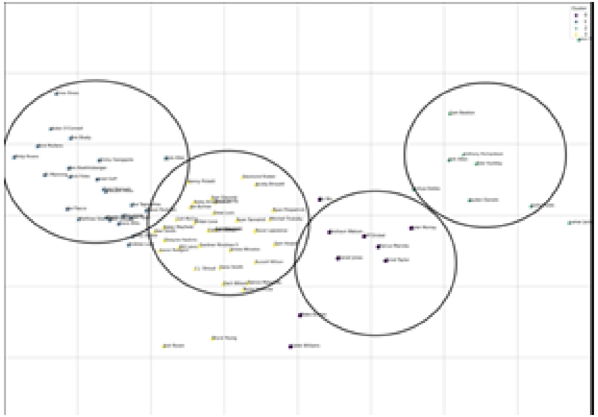
QB Clustering (cont.)

- Add visualizations and examples of classes
- PCA for visualization

- Ended w/ 6 style clusters (could be reduced) (ex: Gunslingers, Quick short passers, dual-threat playmakers)
- 4 mobility clusters (could be reduced) (ex: scramblers, not mobile, low-usage vs high-usage dual-threats)

Style

Mobility



Challenges

- Data limitations → the play by play data lacked sufficient route concept logs
- Toughness of predicting certain routes due to their lack of usage during crunch time
    - This anyways makes its unpredictability not a significant challenge
- Issues with using time series data

Future Plans

- Merging all our solutions to these subproblems to solve the overarching question
- Create singular deliverable based on merged models where users can input a scenario to get predictions
- Accounting for QB's passing location tendencies, time to throw, difficulty to pressure, and quality via QBR
- Using QB Clustering to improve accuracy of route classification and EPA predictions
- Consider clustering defensive tendencies