NFL Fourth Down Analytics

Data-driven insights into fourth down strategy in professional football

Vivian Liu
Mentored by Dr. Anton Dahbura and Tad Berkery
JHU Sports Analytics Research Group (https://sports-analytics.cs.jhu.edu/)

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Introduction

Analytics are driving sports. More and more sports teams are turning to analytics teams to drive their decision making. This is particularly true in the sport of football and the National Football League (NFL).

A large shift in NFL strategy has come in the area of fourth downs. Over the last decade, the number of attempts to "go for it" on fourth down has increased by over 75%. This is in part due to analytic efforts showing payoffs for risk taking. These models developed within each team or by large companies such as ESPN, can often be seen as graphics during broadcasts, but are not accessible to the general viewer.

With a more accessible version of the model which can be run locally or used through a website, everyone can engage with fourth down strategy.

Objectives

The main objective of this project was to develop an accessible fourth down conversion model for all users, from coaches to the average viewer.

The project was developed under the following goals:

1. Develop a fourth down conversion probability model
2. Develop a framework for fourth down decision recommendations
3. Determine an optimal fourth down strategy
4. Build an accessible Python package for model training and game simulations which includes an ETL pipeline
5. Create a website with pre-trained models which provides simple-to-use features for people to use

Project Setup

This is a Python-based project.

Data
Data for model training is publicly available play-by-play and player-level data from the NFL. The public Python package nflreadpy was used to import data. Git Large File Storage and AWS S3 are used to store data for streamlined input into the website.

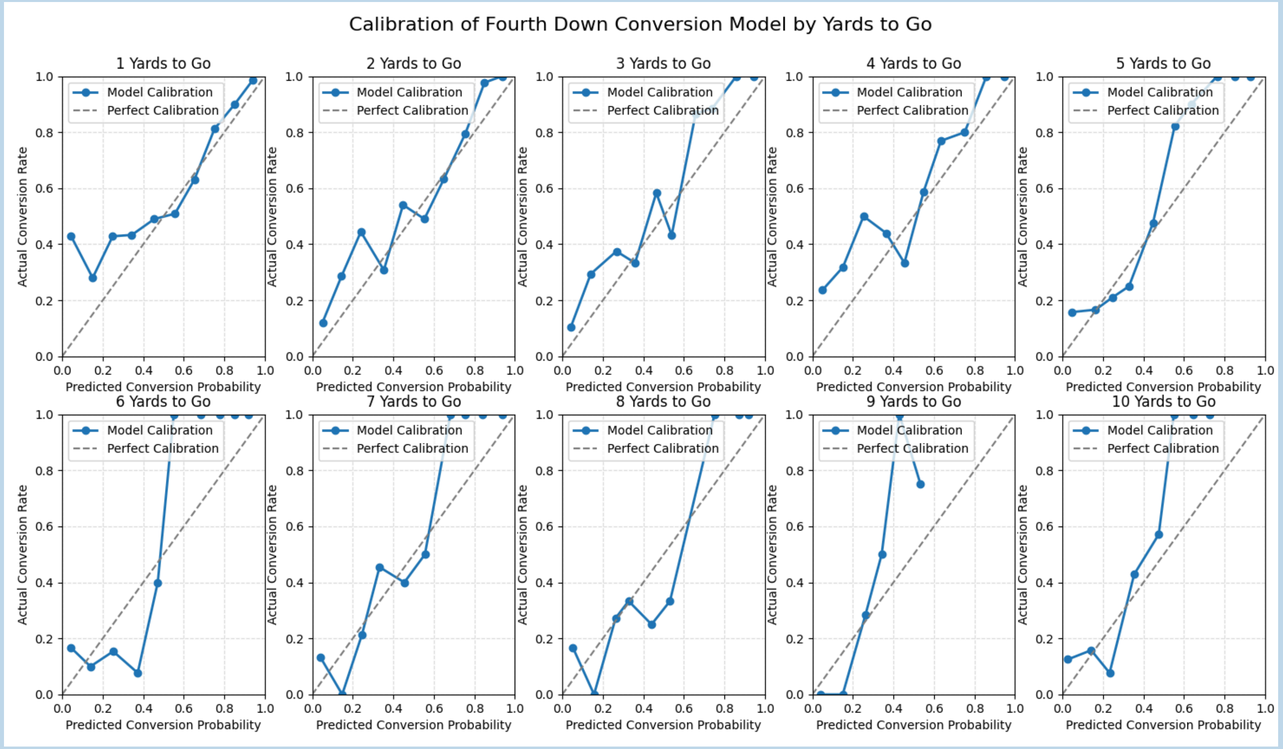
Project Features

1. Two-step model
   a. Primary conversion probability model
   b. Secondary EPA estimation model
2. Strategy analysis
   a. Markov Decision Process-based game simulation outcomes with conservative, neutral, and aggressive fourth down strategies
3. Website with three features
   a. Manual fourth down decision analysis
   b. Live and historical fourth down analysis
   c. Fourth down decision game for users

Deliverables and Results

1 Primary Conversion Model

- Primary input variables include yardline, yards to go, score, and spread line
- Examples of engineered features include fourth down offense and defense rates, quarterback average EPA, and yards traveled in previous drives
- A Generalized Additive Model (GAM) was used for interpretability and control over input variables
- The model was trained on data from 2017-2024 due to the dynamic nature of football strategy



2 Fourth Down Decision Output

The final output accounts for different play options including the possibility of punting, kicking a field goal, or going for it. Combined with the conversion model, the secondary stages account for additional aspects of the game.

- The secondary EPA model was an XGBoost model taking a simpler set of inputs. The output includes EPA for punts, field goals, and going for it
  - The field goal outcome is player-dependent, accounting for their accuracy
  - The EPA for going for it is conditional on a successful conversion
- The EPA output and conversion probability are combined for an overall score for the play

SCORE = EPA * CONVERSION PROBABILITY

- The final decision output accounts for "must go" situations. For example, losing with seconds left at the end of a game means teams should definitely go for it. A "must go" threshold was calculated empirically using Win Probability Added and determined to be a 7% win probability

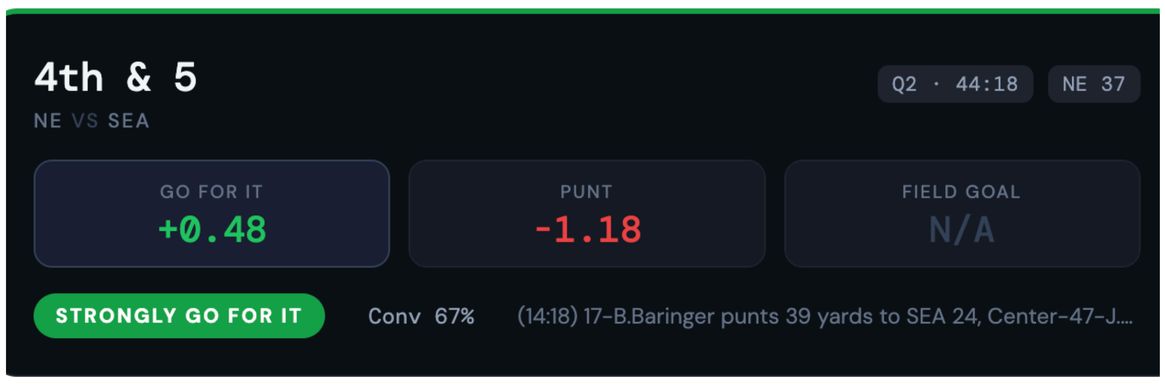
Scan to explore the website yourself!

3 Website

The website is written using Streamlit for simple UI and straightforward design. The website uses stored data instead of loading it from a package at each startup. It uses the pretrained models to give decision outputs for historical, live, and hypothetical fourth down situations.

Live Data Features

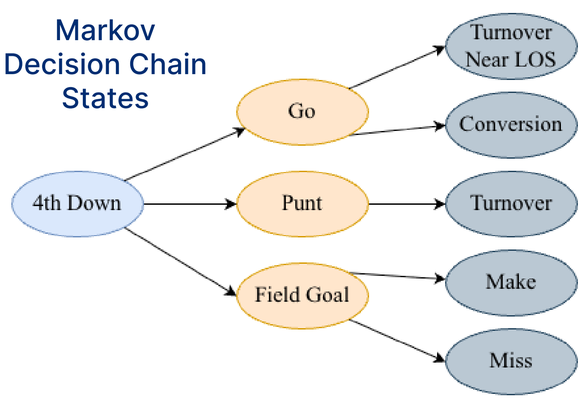
- The website uses a script calling the ESPN API to integrate live data into the Live Games tab
- This data comes from ESPN Scoreboard and Gamecast



4 ETL Pipeline, Model Training, and Strategy Simulations

The project contains an ETL pipeline which can be downloaded and run to load and clean the data used to train the models.

It also contains game simulations under conservative, neutral, and aggressive fourth down strategies to compare the outcomes. Compared to generally conservative coaching decisions made in the real world, the simulation outcomes demonstrate that a neutral strategy (i.e., slightly more aggressive than baseline) gives the highest potential win rate of 0.54.



Fourth Down Policy Comparison for 1000 Simulated Games	
Policy	**Average Win Rate**
Conservative	0.492
Neutral	0.547
Aggressive	0.514

Conclusion

This project aims to make football analytics more accessible.

I propose the outlined GAM conversion probability model as a tested solution to provide interpretability and transparency. The website provides a landing page for football fans.

The models also show that a less conservative fourth down strategy can increase win rates.