

NFL Expected Completion Probability

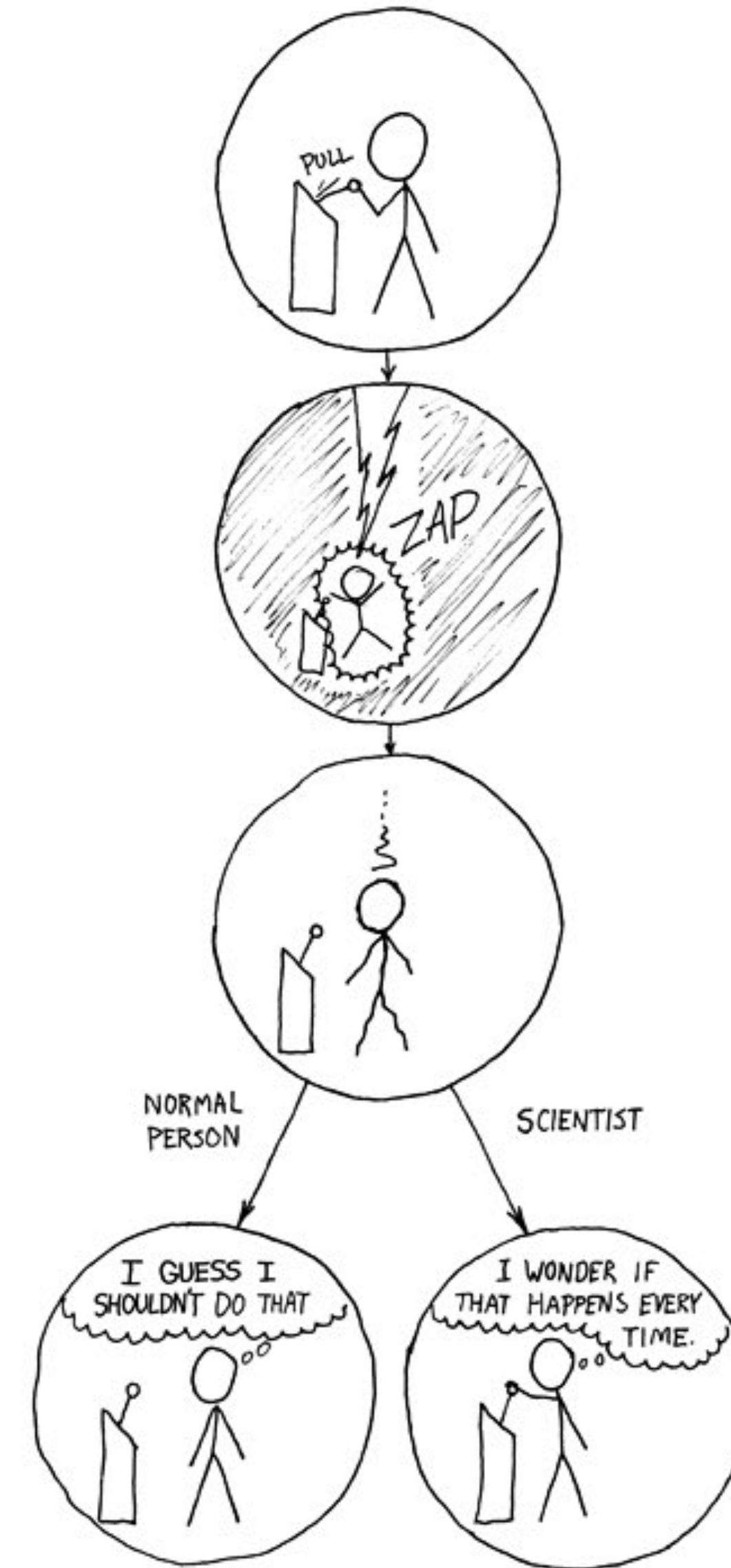
**Modeling passes in the NFL with publicly available data and
reproducible research**

By: Zach Feldman
Advised By: Dr. John Draper

Societal Impact

Reproducible, Public Data Pushes Innovation

- Reproducible research is vital for new ideas, transparency, and learning
- Public NFL research has skyrocketed since the release of nflscrapR (Horowitz, Ventura, Yurko)
- Completion Percentage Over Expected has become a popular metric for quarterback evaluation
- No consensus public model



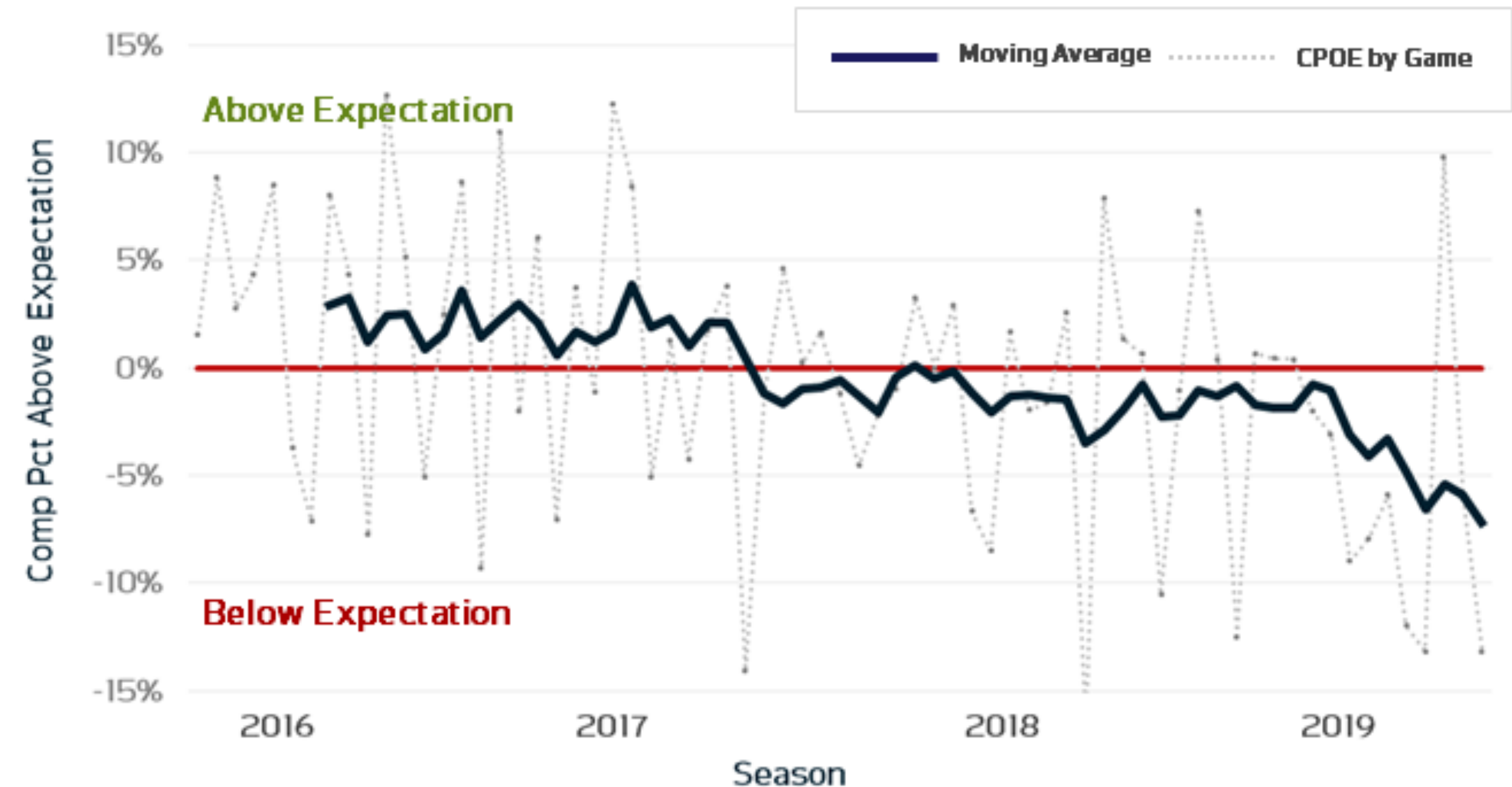
source: xkcd

Analytical Uses

Completion Percentage Over Expected

- CPOE, popularized by Josh Hermsmeyer, of fivethirtyeight.com, in 2019, is a valuable metric for quarterback evaluation
- Relatively stable year-over-year
- Performance against expected is easily comparable through time
- Provides context that raw completion percentage does not
- NFL Next Gen Stats has an expected completion model utilizing proprietary tracking data

Tom Brady Comp Pct Over Expectation (Moving Average)
Since 2016

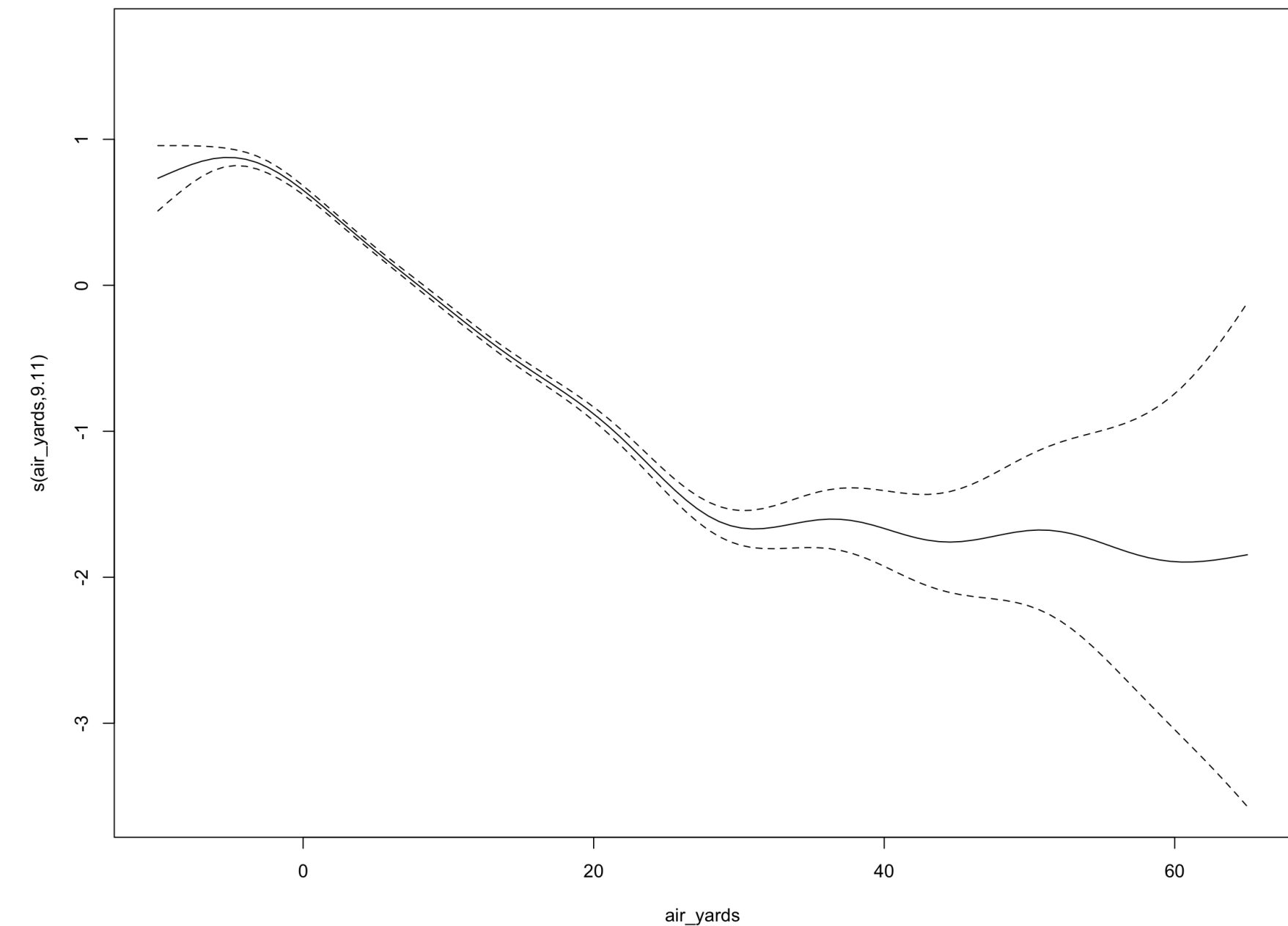


Model Development

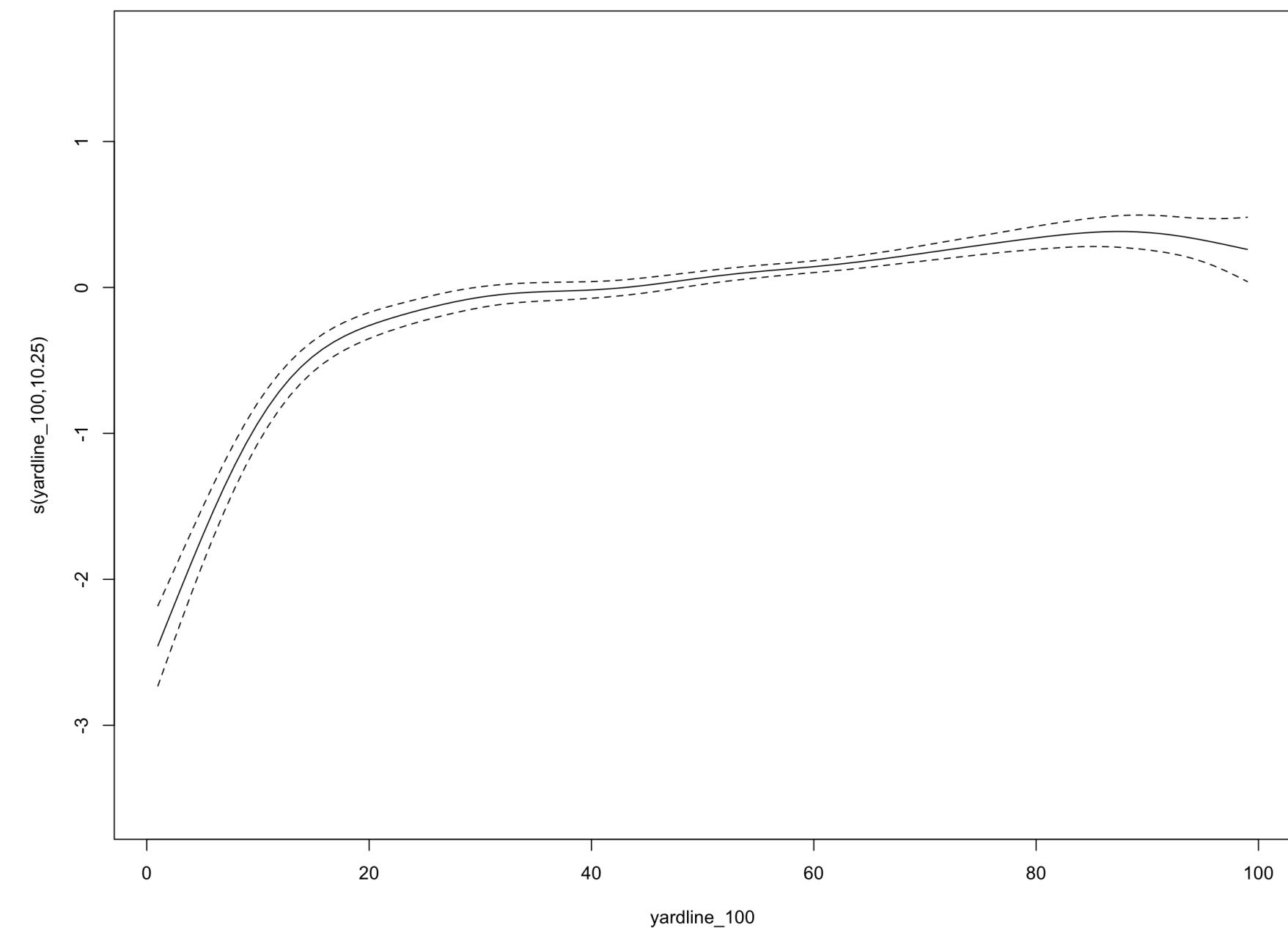
GAMs - Generalized Additive Models

- GAMs provide many benefits over linear models
- Predictors are fit with smoothed splines
- Allows for a more non-parametric approach
- Most public models use GAMs, which was my reason for starting with them
- After a lot of time using them, I wondered about other possible models

Partial Effect of Smoothed Air Yards



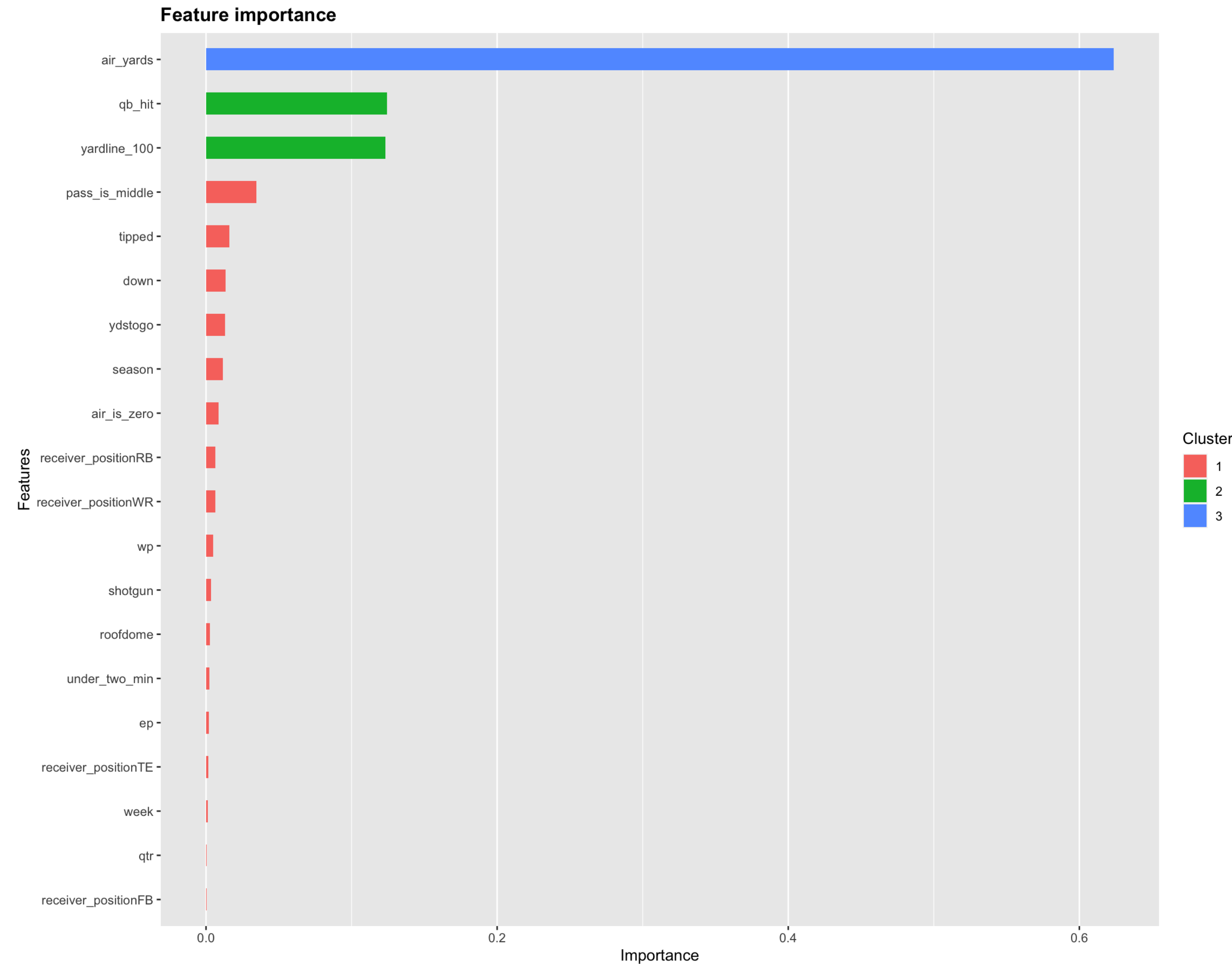
Partial Effect of Smoothed Yard Line



Model Development

XGBoost - eXtreme Gradient Boosting

- Provides a very fast and portable implementation of Gradient Boosting
- Also provides regularization parameters to prevent overfitting - especially useful when we know one variable is going to drive most of the prediction
- I utilize boosted decision trees for logistic probabilities
- Not much tuning needed to be done given how much one variable dominates the model
- Better performance, but harder to interpret effects of variables
- Many variables only add a little extra, but collectively improve upon just air yards



Results

	Full In Sample MSE	30% Out of Sample MSE	QB CPOE YoY R^2	Test 2019 MSE
XGBoost	0.201	0.202	0.119	0.194
GAM	0.204	0.203	0.101	0.196
Air Yards Only - GAM	0.214	0.212	0.115	0.206

- XGBoost is the best performer in every category, but not demonstrably by any means
- While XGBoost was the best performing, the GAM still gives us value from variable interpretation
- While Air Yards alone can get us good results, it's clear we can still improve on it
- Plan to implement the XGBoost model into the new R package nflfastR created by Sebastian Carl and Ben Baldwin