NFL Big Data Bowl 2021 Project Paper

Ty Hammond, Blake Hammarstrom, Jaret Stickrod, Luke Spencer

Colorado State University

STAT 445: Statistical Machine Learning

Dr. Andee Kaplan

November 29, 2024

Motivation

The topic we chose to study is how the defender position in the NFL impacts the result of a play. The motivation for this project comes from the importance of defensive strategies in the outcome of NFL plays. The number of defenders in the box is a gametime decision that significantly impacts the result of a play. Specifically, the focus of this project is to see whether an increase in the number of defenders in the box causes a decrease in yards gained. The significance of this study is the insights the findings can provide to coaches on balancing their defensive playbook between rushing and coverage plays. For example, having more pass-rushers on the play could put pressure on the quarterback but leave downfield open for receivers. On the other hand, not having enough pass-rushers gives the quarterback more time to get the pass off, potentially leaving options open downfield, or leaving the middle vulnerable to a run play. Being able to understand the relationship between defender position and result of play would create a quantifiable measurement that could assist coaches and coordinators in game.

Our motivation for this project comes from a passion for analytics and football. We share an understanding of the game but being able to dive deeper and find measurable insights from game data creates a different tier of knowledge. Sports analytics is in the middle of an uprising and our project aims to be a part of it.

Methodology

The first part of this study investigates the relationship between the number of defenders in the box, the number of pass rushers, and the outcome of a play. This analysis focused exclusively on passing plays, which were filtered from the dataset to ensure consistency in play type and to help minimize the size of the data. The play result response variable is the yards gained by the offense, excluding penalty yardage. Variables in interest were used in their original form and in binned formats to investigate potential categorical effects. The dataset was further cleaned by removing missing observations for the selected variables, resulting in a final dataset with 17,330 observations. Exploratory data analysis (EDA) was conducted using visualizations, including boxplots, heatmaps, and a correlation matrix to assess the relationships between the predictors and play results. These analyses provided insights into the associations between play results and the predictors, with the number of pass rushers showing a slightly stronger correlation (0.032) compared to defenders in the box (0.006). Both linear regression and generalized additive models (GAMs) were used to model the relationship between the predictors and the response variable. The linear regression model assumed a direct linear relationship, while the GAM accounted for potential non-linearities by incorporating smooth terms for the predictors. The dataset was split into training (80%) and test (20%) sets to evaluate model performance and five-fold cross-validation was

used to assess the generalizability of the models on the training data. Performance metrics, including root mean square error (RMSE), mean absolute error (MAE), and R-squared, were calculated for both cross-validation and test set evaluations.

Conclusions & Results

The linear regression model showed small explanatory power, with an adjusted R-squared of 0.0005, indicating the predictors explained only 0.05% of the variance in the play result. Among the predictors, the number of pass rushers was statistically significant (estimate = 0.30, p = 0.002), suggesting a modest positive effect on play outcomes, while the number of defenders in the box was not significant (estimate = -0.044, p = 0.606). The residual standard error was 9.893, and the test set RMSE was 9.892, consistent with cross-validation results. The GAM model provided a slight improvement, with an adjusted R-squared of 0.0014. The smooth terms for the number of pass rushers and defenders in the box were statistically significant (p = <0.001 and p = 0.001, respectively), suggesting non-linear relationships. Cross-validation selected an optimal model with three degrees of freedom, yielding a test set RMSE of 9.879, slightly better than the linear model. Both models indicated weak relationships between the predictors and play result, highlighting the complexity of predicting play outcomes and suggesting the need for new predictors or different modeling approaches.

References

The National Football League. 2021. NFL Big Data Bowl 2021. Retrieved Oct 29, 2024, from https://www.kaggle.com/competitions/nfl-big-data-bowl-2021/data

Kaplan, Andee. Statistical Learning, Class Notes. PDF File. 2024. https://dsci445-csu.github.io/notes/2_stat_learning/20240829_2_stat_learning.pdf

Kaplan, Andee. Regression, Class Notes. PDF File. 2024. https://dsci445-csu.github.io/notes/3_regression/20240910_3_regression.pdf

Kaplan, Andee. Classification, Class Notes. PDF File. 2024. https://dsci445-csu.github.io/notes/4_classification/20240919_4_classification.pdf

Kaplan, Andee. Regularization, Class Notes. PDF File. 2024. https://dsci445-csu.github.io/notes/6_regularization/20241015_6_regularization.pdf

Kaplan, Andee. Nonlinear, Class Notes. PDF File. 2024. https://dsci445-csu.github.io/notes/7_nonlinear/20241029_7_nonlinear.pdf