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

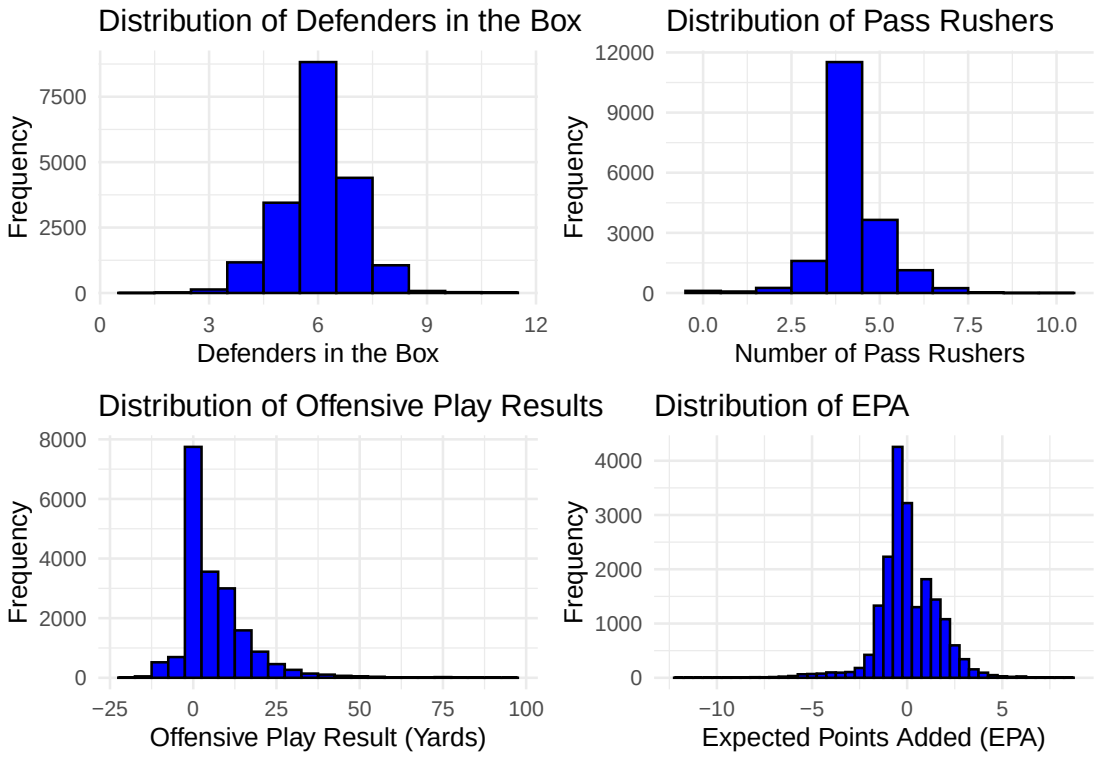
The topic we chose to study is how the defender position in the NFL impacts the result of a play. The motivation for this project comes from the importance of defensive strategies in the outcome of NFL plays. The number of defenders in the box is a gametime decision that significantly impacts the result 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. We recieved our data from the 2021 NFL Big Data Bowl. This was a past Kaggle competition focusing on defensive strategies.

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

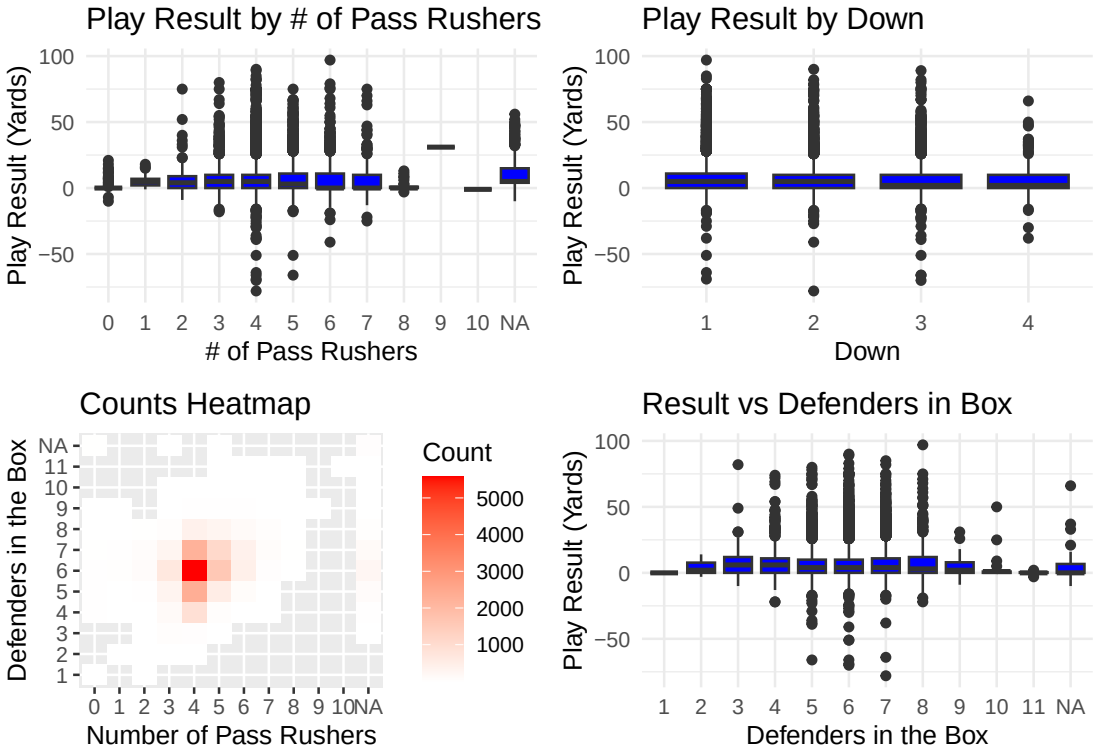
The first part of this study was doing our exploratory data analysis (EDA). Within this we wanted to focus on three different main aspects. First we want to assess the overall data quality of the "plays.csv" file that we retrieved from Kaggle. We wanted to find the unique value counts, counts and proportions of NAs, and data types for all variables. After assessing all of these aspects we determined that there was no imputation necessary as even the variables with the highest proportion of NAs had below a 0.05% rate. Instead we would just drop the NAs rows. The other aspects of our data quality exploration all showed that we had a quality data set and minimal data cleaning was necessary. The next step of EDA was to look at individual variables that could be important to our analysis. We did this with many variables but four important ones cans be seen visually here:



The second set of models used were a Random Forest and an XGBoost model. The second set of models focused specifically on the number of defenders in the box and the number of pass rushers. The data set was first cleaned by omitting missing values and focusing on the two predictors. The data set was split into a training set (80%) and a test set (20%) to make sure the model was evaluated based on unseen data. The Random Forest model used 100 decision trees. The trained model made predictions based on the test data which were then compared with the actual values using a confusion matrix. Two predictors were used for the second set of models to specifically focus on the number of defenders in the box and the number of pass rushers. This decision was made to further expand on the

research question and for the predictors' relevance in determining play outcomes in football. The goal behind including only these two predictors was to improve interpretability of the data while still providing some meaningful insights. While the Random Forest model did not perform well in predicting play result (4.59%), we were able to see that the number of pass rushers had a more significant impact on play result than the number of defenders in the box did. A similar approach was taken with the XGBoost model. The two same predictors were used with the play result as the response. The RMSE was calculated to compare test results to the other models. The focus on the same two features simplified the model significantly which made the analysis more interpretable, while directly addressing the research question.

The last and final aspect of our EDA was looking at combinations of variables. This was primarily focused on our research question and again we made many different visualizations such as the four below.



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. We then decided to attempt to predict EPA or expected points added from presnap indicators. We chose EPA as this was seen as the best measure of offensive success. This meant only using the variables quarter, down, yardsToGo, offenseFormation, defendersInTheBox, and absoluteYardlineNumber as our predictors. Using a 70-30 train and test split we attempted to use KNN, radial SVM, random forest, ridge, and lasso modeling techniques. This provided a good all around base to attack the problems from multiple angles. We also slighly shorted our dataset by taking a sample of 10,000 observations. This was done to help the model run faster. We also compared our results against a baseline MSE where we just predicted the mean. Within our models we utilized caret for some train control. Our K-Nearest Neighbors (KNN) model predicts EPA by identifying the k closest observations in the feature space and averaging their EPA values. The Support Vector Machine (SVM) model uses a radial basis function kernel to create a decision boundary, predicting EPA by optimizing the separation of data points. The Random Forest (RF) model aggregates predictions from multiple decision trees, reducing variance and improving prediction accuracy for EPA. The Ridge Regression model applies regularization, penalizing large coefficients to prevent overfitting while predicting EPA. The Lasso Regression model applies regularization as well, encouraging sparsity by setting some coefficients to zero, simplifying the

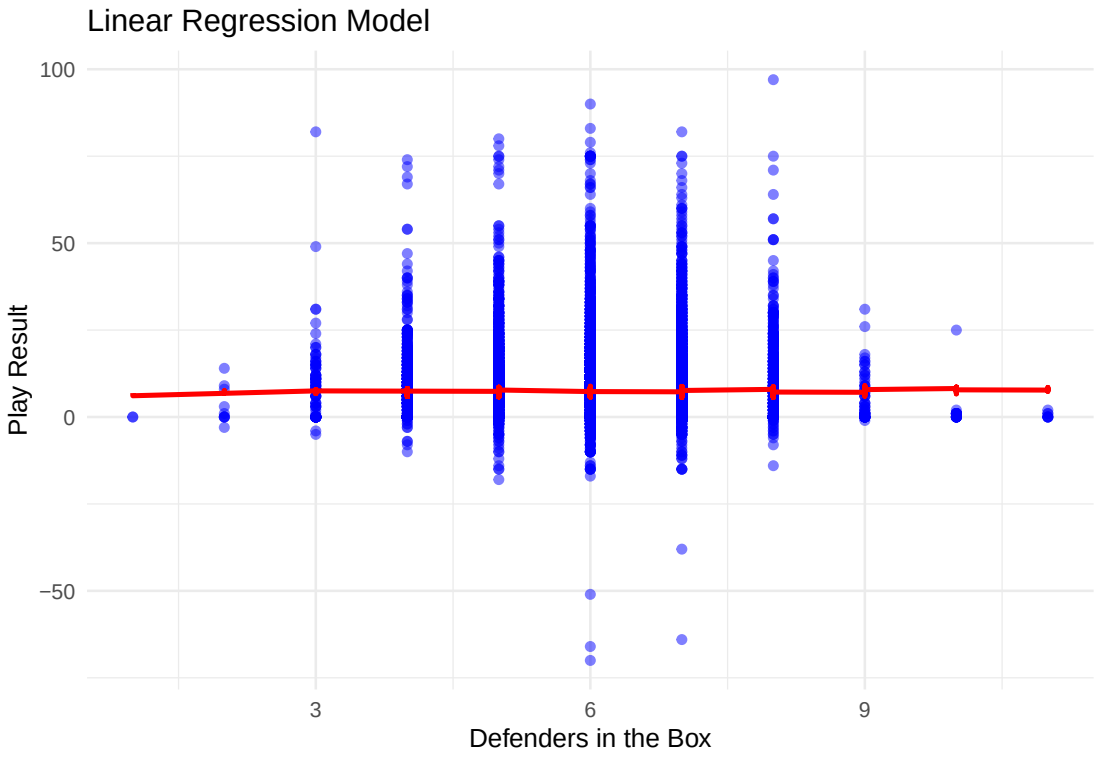
model while predicting EPA.

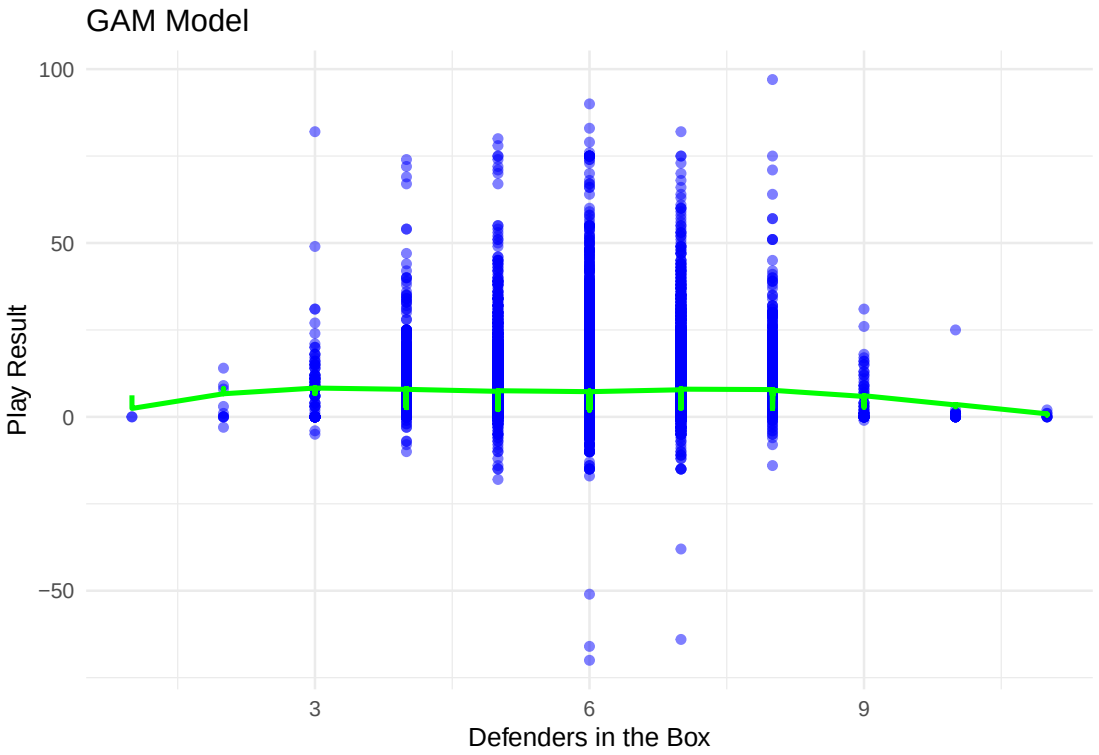
Conclusions & Results

The Random Forest model gave a prediction rate of 4.59% which was not particularly useful in predicting play result. A takeaway that was gained from this model was the higher statistical significance of number of pass rushers compared to number of defenders in the box. The XGBoost model followed a similar approach and had a test RMSE of 9.83 which was slightly better than the previous models. The overall takeaway was the importance of number of pass rushers on play result with number of pass rushers having a score of 0.6 and number of defenders in the box having a score of 0.4

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.





The results from out attempts to predict EPA based on pre-snap prediction variables showed very minimal improvement or even a reduction in MSE from our baseline. After experimenting with different seeds, we came to the conclusion that no model was significant enough to consider an improvement over the baseline. For example using seed 445 we found a -0.003 improvement for KNN, 0.01 improvement for SVM, 0.07 improvement for random forest, and a 0.06 improvement for both lasso and ridge in model MSE vs baseline MSE. This result aligns with expectations, given the sophistication within NFL teams. Teams employ highly skilled analysts, coordinators, and coaches who design offensive and defensive schemes to ensure they are as non-predictive as possible. The element of unpredictability is a cornerstone

of competitive advantage in the NFL. If a team's strategies were easily predictable, opponents could exploit this by making on-field adjustments or calling audibles to counteract the plays effectively, thereby increasing their chances of success. The absence of clear patterns suggests that teams are achieving their goal of keeping their strategies unpredictable to opponents. This non-predictability is a testament to the planning behind NFL strategists and coaches.

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