Learning Predictive Analytics From The NFL Big Data Bowl

Parker Hedrick, Vineel Devalapalli, Conor Sheridan, Matthew A. Lanham

Purdue University Krannert School of Management

phedric@purdue.edu; vdevalap@purdue.edu; sheridac@purdue.edu; lanhamm@purdue.edu

Abstract

We provide an undergraduate business student's perspective on learning and doing predictive analytics by participating in a challenging Kaggle sports analytics competition sponsored by The National Football League. We discuss the goals of this competition, our approach to predicting how many yards a player will gain after a handoff, and how this problem could be related to other common business problems.

Parameters impacting rushing yards?

Does weather impact the rushing game?

Introduction

NFL teams are struggling to place a value on the running back position.
More and more, the teams let star running backs walk away or trade them to another team in hopes of avoiding another overvalued contract. Today's game is dominated by high-octane passing attacks in which few running backs are drafted in the first round and teams are rushing a lot less.

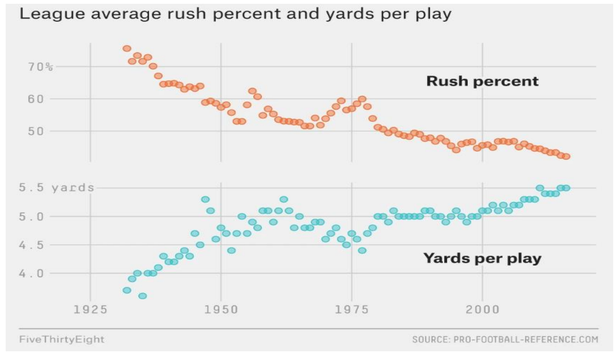


Figure 1. Decline of rushing game over the years

It is ever important to pull the most value from running backs by diving deep into the data behind what makes a more efficient running back.

Literature Review

We looked at papers that discussed the age of players, the effects of weather on different teams and overall play, if collegiate and combine performance are good indicators for professional performance, and if a team is more inclined to utilize their running backs if they are better than average. These specific studies have allowed us to get outside perspectives and additional analysis on some of the most importance variables we ourselves are analyzing.

Important Topics on Discussion Board:

- Regression vs Classification
- How the orientation of the running backs was recorded?
- What is defined as a "rush play"?
- What is the best model to utilize?

Methodology

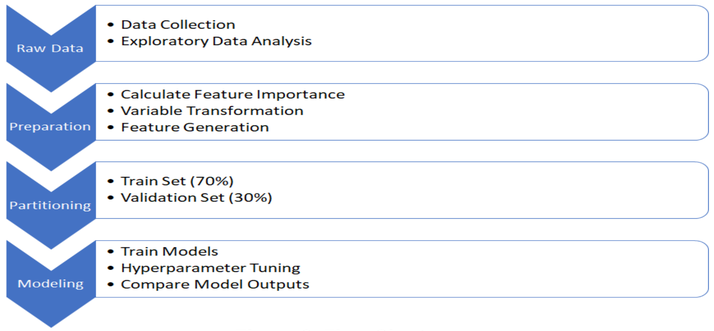


Figure 2. Flow Chart

Data

This dataset contains Next Gen Stats tracking data for running plays. You must use features known at the time when the ball is handed off to forecast the yardage gained on that play.

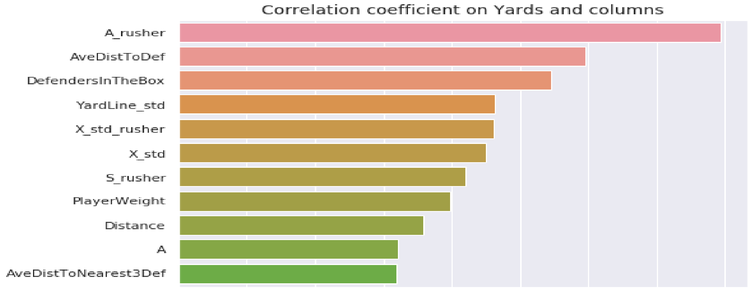
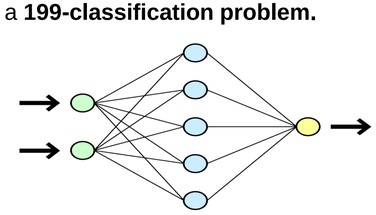
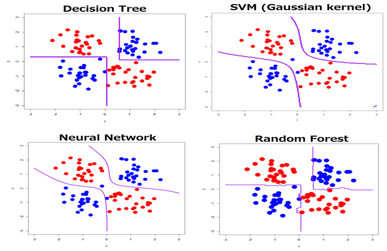


Figure 3. Something useful that helps me discuss the models we used….

Methodology (Approach) Selection

After analyzing the variables and records, as well as focusing on the problem at hand, we came to the conclusion that this problem is a multi-classification. We came to this by understanding the data in regards to the yards. This is a **199-classification problem.**





Learnings/Research

- Neural network = Less feature engineering
- Keras sequential classifier -> most used
- Continuous ranked probability score

Results

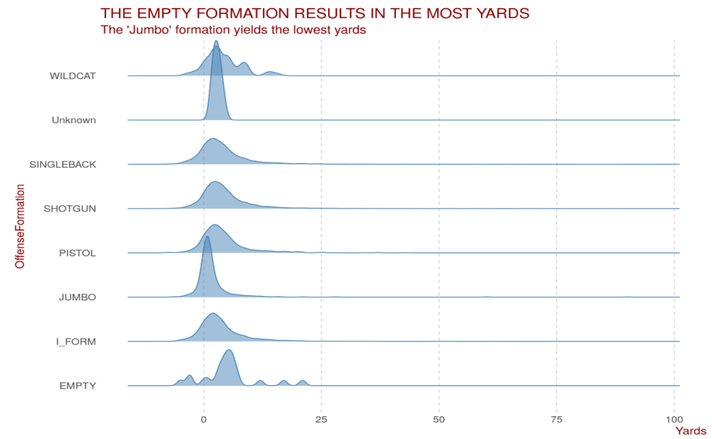


Figure 4. More packed backfields result in less yards

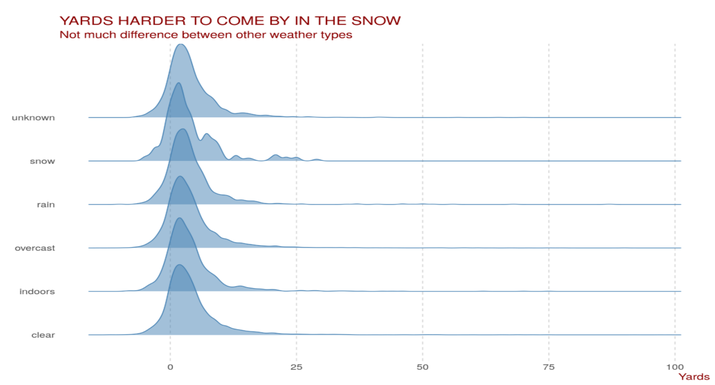


Figure 5. Snow makes yards a lot harder to come by

Conclusions

Today's game is dominated by high-octane passing attacks in which few running backs are drafted in the first round and teams are rushing a lot less.
Most Significant Factors:

- Biggest Weather Impact: Snow
- Biggest Non-Weather Impact: Rusher Acceleration
- Top 3 Correlated Parameters

By analyzing these prominent attributes coaches, GMs, and owners will have a better insight as to who they want to award a contract too. Figuring out which players fit best into their offensive scheme is crucial to the success of the organization. With this knowledge these leaders will be better informed to make the best decision possible.