Draft Success: Using Data to Finding Franchise Players

Proposal Report

William Vernon III  
 Data Science  
 University of Colorado Boulder  
 Boulder, CO, USA

ABSTRACT

One of the most important aspects of the NFL is the collegiate draft where players from around the country and the world are picked by the 32 NFL teams in between each season. The core objective of each team is finding those players who can contribute to their teams in a significant way for as long as possible. This is done by analyzing lots of data of many collegiate players throughout their career stats, film from games, and stats from the Draft Combine. This project will go through a smaller version of what each team’s scouting and personnel do for each draft to find franchise players. My goal is to find patterns in data from these players to find players who have a upward progression with stats and abilities like franchise players already in the NFL. I will be focusing my attention to arguable the most important position groups in the sport the “Tranches” aka Offensive and Defensive Line. These players can really establish the tone for their respective side of the ball and the foundation needed to build a championship caliber team.

KEYWORDS

NFL, NFL Draft Combine, Film, Machine Learning, Data Mining, Hall of Famer, All-Pro, Pro-Bowl

1 Introduction

Teams like the Dallas Cowboys, have exceptional record of drafting Hall of Famers, All-Pros, and Pro-Bowlers. They achieved this by not trading away valuable draft picks in blockbuster trades or relying on free agency to acquire key franchise players. The Dallas Cowboys draft team lead by Will McClay have drafted 12 All-Pro selections (seven unique) and 24 Pro Bowl selections (13 unique) and including potentially future Hall of Famers. This is what has led me to look at this as my project subject, can you use data and machine learning to help add another layer of analysis for picking players in the draft. There are certain things that all great O-Line & D-Line players have in common with each other in their respective position groups like size (height, weight), lateral quickness, (ADD IN CORRELATIONS OF ELITE PLAYERS AND THEIR COMBINE STATS). This can be used to feed into different Machine Learning techniques like clustering to group High, Medium, and Low players based on their own attributes. Draft Success will cluster groups of players who meet the criteria of a possible franchise player.

The limitation in traditional scouting is human bias with overweighing a feel for a player, focusing too much on film and not on any of the data, and outsourcing too much to outside contractors for opinions. Old school football guys believe they have a sacred ability to tell just by looking at a guy if he will be a super star or not and most of the time there is data to either reinforce that belief or negate the belief. In modern NFL with so much data collection and tools available to you why would you not use it to at a minimum supplement idea, opinions, and strategies.

2 Related Work

2.1 NFL Virtual Athletes

Looking for more related work like NFL using ML for player Eval like they are using for digital athletes

2.2 Research Opportunities

The objective is to introduce a non-human evaluation into this process to mitigate potential human biases. The process human scouts take to evaluate players leads to excessive human influence in player evaluation. It is challenging to eliminate human bias from such processes, especially when scouts have long-term relationships with staff and players at these schools. Even a small amount of bias can significantly impact the evaluation, potentially leading to the selection of Player B over Player A. Data collected with a rigorous scientific standard can effectively mitigate human bias to near-zero levels, enabling the identification of players with a higher probability of success. This data can serve as a decisive factor and not at a complete replacement of the current system just a small added feature to help make decisions.

3 Data and Methods

3.1 Data Mining Pipeline:

Gather Data --> Process Data --> Feature Engineering --> Modeling --> Evaluation --> Report

3.2 Datasets:

3.2.1 Pro-Football-Reference NFL Stats

Pro-Football-Reference is a complete source for current and historical NFL, AFL, and AAFC players, teams, and scores. It will be the source of most of the data used in this research because of its accuracy and ease of use offering many methods to download the data specifically for research.

3.2.2 NFL 100 List

The NFL 100 is created during each off season where current NFL players rank their peers to get a well-rounded list of the current seasons top 100 players regardless of position. This is a great list to get a group of players that are considered the best of the best. I will be using this to look at where good players are drafted.

3.2.3 NFL Draft Combine

The NFL Draft Combine happens each off season where each collegiate player that has declared for the draft can complete a series of drills and benchmarks to show overall strength, speed, and athleticism. It can be a great source of raw athletic ability, but not all players opt into this and some who do opt out of some of the workouts and drills.

3.2.4 Sports-Reference CFB Stats

This is the same maker of Pro-Football-Reference but for collegiate football. This will be where all the college career stats come from for training and analysis.

3.3 Data Processing:

3.3.1 Data Cleaning:

Most of the data from Pro-Football-Reference is clean but they have some columns that are not needed like the player’s social media which is removed. Along with any missing data and any wrong data types. Here I will also take an important category “Awards” and encode it to either 1 or 0.

3.3.2 Data Aggregation:

Since I’m using multiple years of data they are all separated from Pro-Football-Reference so after cleaning the data I make sure all the columns are correctly named and matching. Then I will do a merge for each year. For data like career stats, I will take the mean of every column so that it creates a dataset of their career average except for “Awards” will be a sum.

3.3.3 Feature Engineering:

Now finding the best stats to use for analysis mostly all performance stats are going to be useful, but games played, and some other less important features will be removed. Some of the more important defensive line stats will be Sacks, Tackles for Loss, QB Hits, Solo Tackles, and Combo Tackles.

3.4 Defining Target Variable and Modeling:

3.4.1 K-Means Clustering:

I’m using K-means clustering on the NFL players to group each player into three clusters (1) Bad, which will be the players who perform at a lesser level than most. (2) Good, players whom are about average or slightly above average good role players. (3) Elite, players who make the most impact when they are on the field that can seriously separate themselves from the others.

3.4.2 Linear Regression:

Using linear regression to track a player’s progression through their career in college. You want players who increase their level of performance each year which can hopefully continue to increase into the NFL as well. We can see what the projection for their progression is.

3.4.3 XGBoost: Player Success Prediction

XGBoost offers a good way to model projections for future ability based on training data from past productivity. I can use this to train on the college career and try to predict what their NFL will be.

4 Evaluation Plan:

5 Update on Progress:

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00