The Game Beyond NFL!

Hardik Mitesh Kapadia
hkapad2@uic.edu

Niketan Doddamani
ndodd@uic.edu

Nishant Ragate
nragat2@uic.edu

1 Abstract

Football, the most played sport in the country, has been the subject of numerous studies since it's culmination over a century ago. Given its net worth of $6.49 billion, the efforts spent in studying are more than justified. However, most of the efforts are directed towards building the perfect team or formulating an ideal plan for a game. There is extremely limited research in the realm of what affects the selection of a player, from a player's perspective. We attempt to answer this question in our study. We have collected data regarding player's picked in the NFL drafts over the past decade by scraping the official NFL website. We then performed Exploratory data analysis on the data as described in Section 6. Using the results obtained from the EDA, we processed the data to ensure that we only use relevant features for the next step. The next step was to perform Structural learning. Using the resources at our disposal, we used the Peter-Clark (PC) algorithm to estimate a Structural Causal Model in the form of a Direct Acyclic Graph. Following these steps, we aim to attempt answering the question by calculating treatment effects, learning causal coefficients, etc.

2 Introduction

With a viewership of 123 million for the game & 47.5 million for the draft, the NFL's popularity drives intense interest in player performance and game analytics from fans, team management, and sponsors. It is a lucrative sports league globally — a $16 billion business. We want to explore more aspects of player valuation providing insights for team scouting departments and having more understanding of player value beyond traditional statistics which influence draft decisions along with the draft round in which they are picked.

3 Related Work

Prior research in NFL draft prediction has evolved significantly across multiple methodological streams, demonstrating both the progress and limitations in current approaches. Mulholland and Jensen (2014) [3] established an analysis of college performance statistics to predict NFL success, introducing a statistical approach to draft analytics and demonstrating that collegiate performance metrics could serve as meaningful predictors of professional potential. Building on this foundation, Brock Grassy's [2] research demonstrated the power of modern ensemble methods, leveraging Random Forests and XGBoost to enhance draft outcome predictions, though primarily through correlation-based analysis. While these machine-learning approaches showed promising accuracy, they highlighted a crucial limitation in understanding the causal mechanisms driving draft success. The Northwestern Sports Analytics Group [5] further advanced the field by introducing sophisticated metrics like Approximate Value per Game (APG) for position-specific analysis, particularly in quarterback evaluation, revealing complex patterns across draft categories and their relationship to NFL success. This offered new possibilities for revealing relationships that conventional matching methods often miss, particularly in complex sports analytics scenarios. Our research uniquely synthesizes these diverse methodological streams, integrating causal inference techniques with established machine-learning methods to create a more comprehensive analytical framework. By bridging the critical gap between correlation-based analytics and causation-driven decision-making in draft strategy, our approach aims to uncover the underlying causal mechanisms that drive both draft position and subsequent NFL success, while maintaining the predictive power demonstrated in earlier machine learning approaches. This integration of causal inference with the use of DAGs [6] with traditional machine learning techniques represents a significant advancement in draft analytics, offering teams more nuanced and actionable insights for their draft strategies.

4 Problem description

The problem we are trying to solve can be formalized verbally as: "Given the data related to a prospective NFL player, which parameters have a higher impact on the player's selection"

We can answer this question by calculating the causal coefficients in the Structural Causal Model by calculating the Average treatment Effect of an attribute on the output variable. We have two output variables: Y_0 (Pick number) and Y_1 (Pick round) where Y_1 denotes the inverse of the order in which they we picked.

$$Y_0 = (orderinwhichplayerwaspicked + 1)^{-1}$$

i.e., the player picked first was will have $Y_0 = \frac{1}{2}$, the player picked 100th will have a $Y_0 = \frac{1}{101}$. The second output is the inverse of the round in which the player was picked.

$$Y_1 = (roundinwhichplayerwaspicked + 1)^{-1}$$

As seen in the formula above, we also apply smoothening to avoid an output of infinity.

5 Initial Solution

To answer the question, we split the solution into three parts:

(a) Data Preparation
(b) SCM formulation
(c) Calculating the various treatment effects.

For the first step, we scrape the data, analyze it and process it for the second step. This process is described in more detail in Section 6. For the second step, we need to convert our tabular data into a Structural Causal Model, a Direct Acyclic Graph, which

denotes the relationships between the different features with an appropriate weight assigned to each edge. We contemplated various approaches to this including: the Peter-Clark Algorithm, the Hill Climb algorithm, and the Greedy Equivalent search. Each algorithm had it's advantages and disadvantages but we decided to proceed with the Peter-Clark (PC) algorithm. The PC algorithm, a constrain-based algorithm is the better option than the aforementioned the score-based approaches due to various reasons such as:

(a) The PC algorithm is best suited for mixed data – data with categorical and numerical values.
(b) Given the interconnected nature of the dataset, it is vital to capture the conditional dependencies that exist. (e.g. player speed has a varied impact on the selection conditioned on player's position)
(c) Although the dataset is interconnected, and a lot of the features have multiple depensencies, not all of these connections (edges) are relevant to answering our causal question. The PC algorithm excels in capturing local causal relationships while avoiding extremely dense overfit graphs.

We use the pgmpy[1] library to run the PC algorithm and obtain a direct acyclic graphs. For the sake of efficiency, we limit the number of conditional variables to 5. This DAG is an accurate representation of the data and the depicts the relations between the various features (nodes). However, while accurate, this can further be optimized, by manually eliminating certain edges based on domain-specific knowledge. We're still experimenting with different limitations given the resources at our hand and trying to perfect the DAG before moving to the next step. The third step is detailed further in Section 7.

6 Data Description

The dataset captures a comprehensive profile of NFL players, detailing their physical attributes, draft information, and performance metrics across their careers. It includes more than 50 features, with key data points such as players' college affiliations, positions, and draft details (round, pick number, and overall draft order). Physical attributes such as height, weight, 40-yard dash time, vertical jump, bench press repetitions, broad jump, shuttle run, and 3-cone drill results provide insights into players' athletic abilities. The dataset also encompasses career statistics, including solo and assisted tackles, tackles for loss, sacks, interceptions (with return yards and touchdowns), passes defended, forced fumbles, and fumble recoveries. Offensive and defensive statistics include passing completions, attempts, yards, touchdowns, completion percentage, passer rating, receiving yards, receptions, touchdowns, and rushing data (attempts, yards, touchdowns). Additionally, special teams statistics cover punt and kick returns, including yards, touchdowns, and kicking accuracy (field goals and extra points). These diverse metrics enable in-depth analysis of player performance and career trajectories.

Data was collected through web scraping from the NFL's official website [4], using the ZenRows API [7] to automate the extraction process. This allowed for efficient retrieval of detailed player statistics and draft information, ensuring that the dataset was both comprehensive and accurate. By leveraging the ZenRows API, the scraping process adhered to best practices for data collection, ensuring that the data captured was structured and organized for further analysis. The resulting dataset includes historical performance data for NFL players, spanning multiple seasons and draft years.

The data cleaning process involved several steps to ensure the dataset's accuracy and relevance. Given the over 50 features in the raw data, we first performed feature selection, discarding those with less than 10% data coverage to reduce noise and improve model interpretability. Missing values were identified and handled using the K-Nearest Neighbors (KNN) imputation method with k=5. This technique allowed us to fill in missing entries by considering the influence of the five nearest neighbors in the dataset, ensuring that imputed values were contextually relevant based on similar player profiles. The cleaning process ensured that the dataset was well-suited for further analysis, maintaining its integrity and enhancing its utility for predictive modeling and comparisons to state-of-the-art research on NFL player performance. Categorical variables like the `Position` column were encoded using `LabelEncoder`, converting positions into numerical values to make them interpretable by the model without implying any ordinal relationship between different roles. Non-predictive columns, including `Name`, `Stat URL`, `Year`, `College`, and `conf_abbr`, were removed, as they did not contribute to draft prediction and could introduce noise. To further optimize the model's performance, outcome variables `Round` and `Pick` were transformed with a reciprocal formula $\frac{1}{1+x}$, which helped smooth large values in later rounds, emphasizing significant distinctions in early draft rounds, where selection is most crucial. Additionally, physical and performance attributes such as `Height`, `Weight`, `40 Yard Dash`, and `Vertical Jump` were scaled using standardization techniques to ensure each feature contributed equally to the model, reducing the risk of biased learning due to varying magnitudes. These transformations and scaling steps collectively improved the dataset's quality and suitability for accurately predicting NFL draft outcomes.

The examination of positional draft distribution patterns reveals significant disparities across NFL positions, particularly among Offensive Tackles (OTs), Punters (P), and Kickers (K). This visualization demonstrates that OTs maintain the highest draft representation with approximately 190 players distributed across all rounds, while specialists (Punters and Kickers) show a markedly different pattern with a predominance of undrafted players. This positional disparity is particularly evident in early-round selections (rounds 1-3), where OTs are frequently selected while specialists rarely appear, suggesting position-specific draft strategies being considered in predictive modeling.

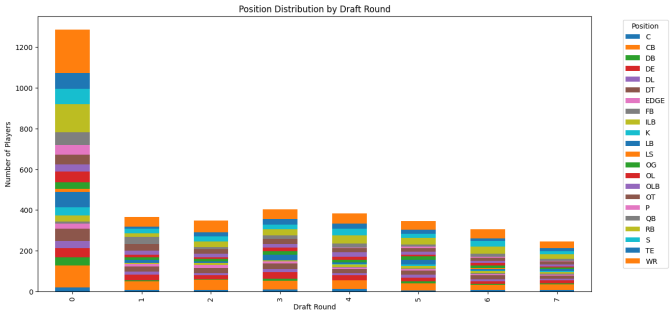

Figure 1: Position Distribution by Draft Round.

The temporal analysis of athletic performance metrics presents a comprehensive view of evolving NFL Combine standards through four key measurements. The 40-Yard Dash times show remarkable stability with a slight improvement from 4.75s to 4.70s over the studied period, while Bench Press performance displays more significant fluctuation, ranging from 18 to 21 repetitions with a general declining trend from 2012 to 2023. Notably, the Vertical Jump metric exhibits considerable variability within a 32-34 inch range, with a marked upward trajectory from 2020 to 2023, peaking at approximately 33.5 inches in 2022. Perhaps most significantly, the Broad Jump demonstrates the clearest positive progression, advancing from 114 to 118 inches over the period, indicating an increasing emphasis on lower body explosiveness in prospect evaluation.

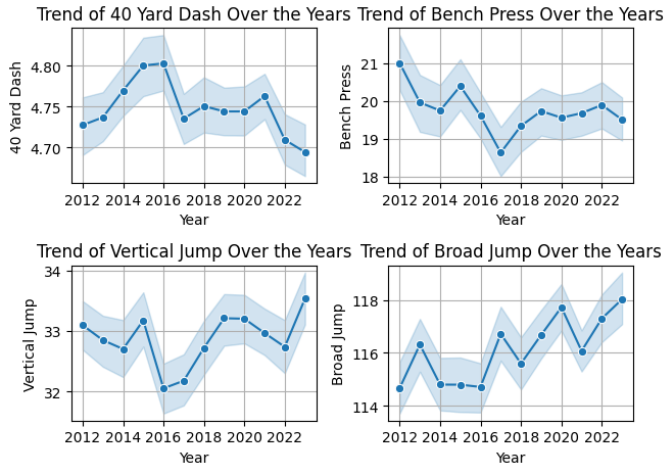

Figure 2: Combined Trends and Draft Implications.

7 Next Steps

For the next steps, we will proceed with step 3. With the structural model ready, we will attempt to answer the causal questions and identify the attributes with the highest impact on the output variables. The possible methods to do this are:

(a) Calculate the causal coefficients in the SCM.
(b) Calculate the average treatment effects of the parameters on the outputs

References

[1] Ankur Ankan and Abinash Panda. 2015. pgmpy: Probabilistic Graphical Models using Python. In *Proceedings of the Python in Science Conference (SciPy)*. SciPy. https://doi.org/10.25080/majora-7b98e3ed-001

[2] Brock Grassy. 2023. Applying Machine Learning to Predict the NFL Draft. https://www.brockgrassy.com/blog/applying-machine-learning-to-predict-the-nfl-draft/. Accessed: 2024-11-07.

[3] Jason Mulholland and Shane Jensen. 2016. Projecting the Draft and NFL Performance of Wide Receiver and Tight End Prospects. *CHANCE* 29 (10 2016), 24–31. https://doi.org/10.1080/09332480.2016.1263095

[4] Pro Football Reference. 2024. NFL Draft Data. https://www.pro-football-reference.com/draft/. Accessed: 2024-11-07.

[5] Samantha Sizemore and Raiber Alkurdi. 2019. Matching methods for causal inference: A machine learning update. In *Seminar Applied Predictive Modelling (SS19)*.

[6] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya like dags? a survey on structure learning and causal discovery. *Comput. Surveys* 55, 4 (2022), 1–36.

[7] ZenRows. 2024. ZenRows Scraper API. https://docs.zenrows.com/scraper-api/api-reference Accessed: 2024-11-07.