The Game Beyond NFL!

Hardik Mitesh Kapadia
hkapad2@uic.edu

Niketan Doddamani
ndodd@uic.edu

Nishant Ragate
nragat2@uic.edu

1 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 and team dynamics providing insights for team scouting departments and having more understanding of player value beyond traditional statistics.

2 Problem description

Our two key questions for the project:

1. Predicting draft selection: Determining if & when a player will be selected in the NFL draft based on their attributes. [2]

2. Evaluating individual impact: Assessing the influence of a player performance on team results and standings. [1] [3]

We aim to improve sports analytics and provide insights for NFL teams, analysts, and fans by leveraging causal inference techniques to address one of the approaches.

2.1 Datasets

This dataset spans 12 years of NFL draft data (2012-2023), containing approximately 3,600 entries (300 players per draft class), with over 50 features per player.

2.2 Feature Table

Category	Features
Player Info	Name, Position, College, Height, Weight
Draft Info	Round, Pick, Year
NFL Combine	40 Yard Dash, Bench Press, Vertical Jump, Broad Jump, 3 Cone Drill, Shuttle
College Stats	Conference, Games Played, Seasons
Defense	Tackles (Solo, Assists, Total, Loss), Sacks, Interceptions
Fumbles	Fumbles Recovered, Fumble Recovery Yards, Fumbles Forced
Receiving	Receptions, Receiving Yards, Receiving TDs
Rushing	Rush Attempts, Rush Yards, Rush TDs
Scrimmage	Scrimmage Yards, Scrimmage TDs

Table 1: Feature categories for the NFL draft dataset

3 Research plan

3.1 Data consolidation

The first step in the project will be to compute a precise and encompassing dataset for both aforementioned causal questions by integrating data from various sources.

(a) Dataset for Player Draft picking: The individual players' year-by-year statistics from their college performance and other relevant games which may influence draft decisions along with the draft round in which they were picked.

(b) Dataset for Team performance analysis: Merging the data from (a) with additional team-specific data such as the yearly team placement and composition.

3.2 Feature engineering

Once the appropriate dataset is prepared for each specific causal question, the next step will be to utilize the data along with domain-specific knowledge to perform feature engineering for both datasets This step will include - Initial data exploration, Handling missing data, Feature scaling, Feature Creation, Feature selection, etc.

3.3 Exploratory Data analysis

Once the dataset is finalized with pre-processed data and the new features, exploratory data analysis will be performed. The general methodologies behind this will be to analyze distributions of key variables, examine relationships between variables, and generate summary statistics. The vital part of this step will be determining the correlation between different variables to identify the probable edges between the nodes in the Structural Causal Model.

3.4 Comparing the two problems

Once the initial three steps are completed, the next step is to compare the two causal questions based on various parameters and determine which question will be answered in the scope of this project using factors like - Impact, Feasibility, The Scope of the problem in the domain of causal inference and the Research gap.

3.5 Structural Causal Model

Equipped with the data from the previous steps and with a single specific problem, a Direct acyclic graph will be constructed to represent the Structural Causal model. This will utilize the values extracted from the Exploratory data analysis to determine the edges, the nodes and the state of the nodes.

3.6 Causal Inference analysis

Once the SCM is constructed, various techniques will be employed to infer the impact of different parameters on the outcome and each other. Based on the SCM constructed, answers to various intermediary causal questions will also be derived. The results of the analysis will be documented and visualized.

3.7 Deliverables

The expected deliverables by the progress report are the complete pre-processed dataset, the results of the exploratory data analysis, and the speciic causal question this will project will attempt to answer.

The expected final outcome for each question is:

(a) Player Draft Picks analysis:

Identify the impact of different parameters on the draft round in which the player was picked as well as the most important parameters influencing a player's selection

(b) Team performance analysis

Identify the impact of the individual positions-specific player's stats and the specific positions on the team's success. Determine the most influential parameters affecting a player's selection.

References

[1] Caterina De Bacco, Yixin Wang, and David Blei. 2024. A causality-inspired plus-minus model for player evaluation in team sports. In *Causal Learning and Reasoning*. PMLR, 769–792.

[2] 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

[3] Jo ao Ribeiro, Pedro Silva, Rodrigo Duarte, Keith Davids, and Jo ao Garganta. 2017. Team sports performance analysed through the lens of social network theory: implications for research and practice. *Sports medicine* 47, 9 (2017), 1689–1696. https://doi.org/10.1007/s40279-017-0695-1