# NFL Cornerback Draft Round Prediction Project

## Project Overview

This project explores whether NFL Combine and body measurement data can be used to predict the draft round of college cornerbacks (CBs). The motivation stems from an interest in performance forecasting, player valuation, and personnel decision-making processes. While this is a stand-alone data science project, it also serves as the foundation for a doctoral dissertation that will evolve to include more sophisticated features and modeling approaches over time.

## Objective

To build a reproducible and explainable classification model that estimates the NFL Draft round of a college cornerback based solely on their pre-draft measurable attributes. This version focuses only on Combine drills and body measurements to simulate a realistic scouting scenario where game performance data is unavailable or withheld.

## Dataset Summary

The dataset includes all drafted cornerbacks from the 2015 to 2021 NFL Drafts with complete Combine and measurement data. Data was aggregated from publicly available sources and includes the following features:

* 40-Yard Dash (sec)
* 20-Yard Shuttle (sec)
* 3-Cone Drill (sec)
* Vertical Jump (in)
* Broad Jump (in)
* Bench Press (reps)
* Height (in)
* Weight (lb)
* Arm Length (in)
* Hand Span (in)

Target Variable: Draft Round (1 through 7)

## Combine and Measurement Descriptions

* **40-Yard Dash:** Measures straight-line speed and acceleration.
* **20-Yard Shuttle:** Assesses lateral quickness and agility.
* **3-Cone Drill:** Tests body control, change of direction, and agility.
* **Vertical Jump:** Evaluates lower-body explosiveness.
* **Broad Jump:** Gauges balance and leg strength.
* **Bench Press:** Measures upper body strength and stamina.
* **Height / Weight:** General body size measurements.
* **Arm Length:** Important for press coverage and disrupting passes.
* **Hand Span:** May impact ball security and interception ability.

## Methods

### Data Preparation

* Missing values imputed using median values grouped by draft round
* Features standardized using scikit-learn’s StandardScaler

### Modeling

* **Logistic Regression (Multinomial):** Chosen for its interpretability and multi-class capabilities
* **Neural Network (MLPClassifier):** Used to test non-linear classification capabilities
* **Unsupervised Learning:** K-Means and K-Prototypes clustering were used to identify groupings among CBs

### Simulation Module

A simulation module was created to allow users to input test values for each Combine metric and body measurement. The logistic regression model then runs 1,000 Monte Carlo simulations to generate a distribution of possible draft rounds, enabling probability-based interpretation and uncertainty quantification.

## Model Evaluation Summary

### Logistic Regression

* Moderate classification performance
* Best predictive recall observed for Rounds 6 and 1
* Common misclassifications often occurred between adjacent rounds

### Neural Network

* Overall lower accuracy (approx. 21%)
* High recall for Round 6 but extremely poor precision and recall for other rounds
* Round 3 received zero correct predictions

### Comparison

The logistic regression model outperformed the neural network in terms of precision, recall, and F1-score across most classes. It also provided a more interpretable structure for use in the simulation module.

## Key Insights

* The Combine data alone is insufficient to reliably distinguish all draft rounds
* Some rounds (e.g., 1 and 6) show higher model confidence, possibly due to more distinct physical profiles
* Other contextual variables such as college PFF grades may improve model fidelity

## Project Value

This project demonstrates the following:

* End-to-end development of a machine learning pipeline
* Experience with classification, simulation, and clustering
* Skills in data cleaning, feature engineering, model evaluation, and interpretability
* Application of predictive modeling to a real-world, decision-making scenario
* Capability to communicate findings with clarity and transparency

## Future Work

* Incorporate Pro Football Focus (PFF) grades and snap counts to improve predictive resolution
* Incorporate target variables of NFL career performance (e.g., NFL PFF grades, Earnings, snap counts)
* Introduce time-series modeling (e.g., LSTM) using college season-by-season data
* Refine simulation logic with calibrated probabilistic outputs

## Limitations

* **Clustering Diagnostics:** This version did not apply the elbow method or silhouette scores to formally validate the optimal number of clusters in K-Means.
* **Centroid Interpretability:** The K-Prototypes clustering outputs were not deeply analyzed, and centroids were not interpreted or visualized.
* **Model Weakness:** Both classification models showed limited ability to separate rounds with overlapping Combine profiles, particularly in the mid-rounds (Rounds 3–5). The neural network struggled significantly with recall and failed to correctly classify some rounds altogether.
* **Single-Modality Inputs:** The model only uses physical measurements and Combine performance, which do not sufficiently capture player skill or game impact.

## How to Use

* Clone the repository and open the Jupyter notebook
* Run cells to process data, train models, and evaluate performance
* Use the simulation cell to test custom profiles and analyze draft probabilities