NFL Draft Classification Model

By Aneesh Sunkarapalli and Rudra Patel

Introduction

Our project aims to predict the chances of being drafted in the NFL based on NFL combine data and player performance.

Our main dataset that we split into cross validation sets utilized 2010-2023 combine data, and we also web scraped 2024 combine data to test our prediction model on.

Our dataset contains various combine performance metrics such as 40 yd dash times, vertical jump height, bench press reps, broad jump length, and 3 cone and shuttle times as well as basic player information such as name, college, position, and a binary classification for whether they were drafted or not.

Related Works

Although our project is focused on a different topic, we found various similar projects online such as predicting NFL success from the Northwestern sports analytics group, articles and work from Google Scholar related to predicting the results of NHL games, and even monitoring public sentiment of NFL draft picks using machine learning.

Other similar sources we found online were more informal. These projects were on blogs, personal websites, and GitHub projects. We did not really consult any informal sources before starting our project.

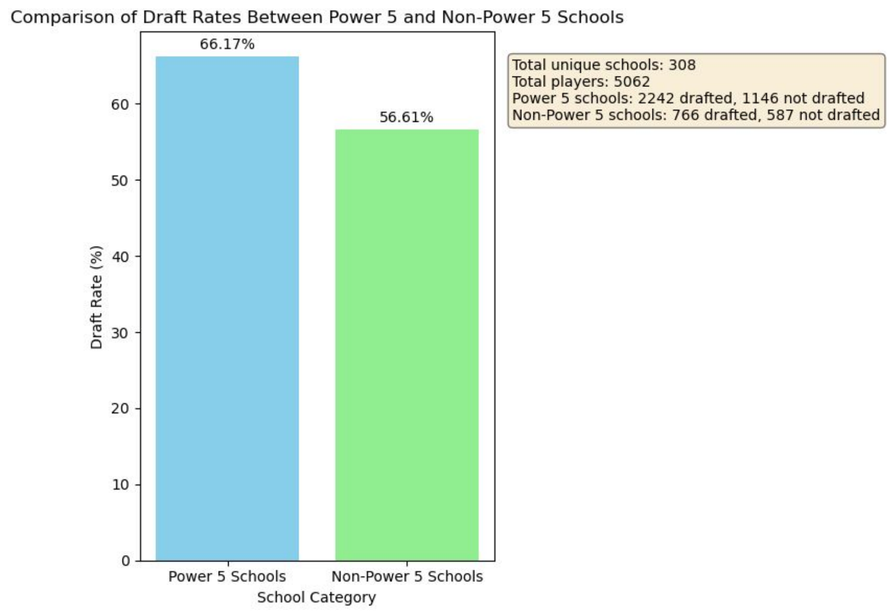
Data Analysis

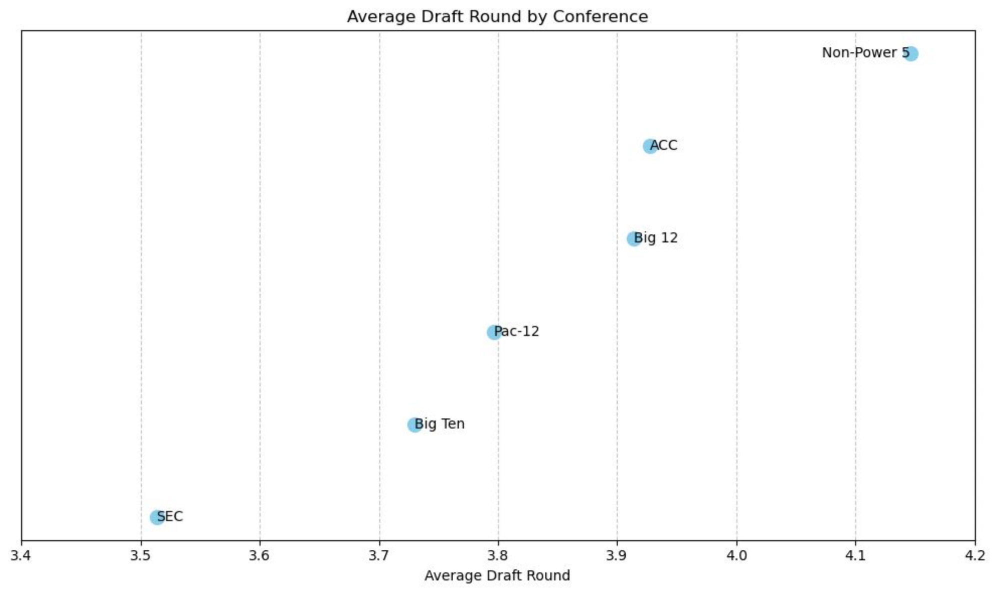
Key Points:

- Power 5 school players have a higher draft rate (66.17%) than Non-Power 5 players (56.61%).

- Suggests Power 5 affiliation offers an advantage in draft selection.

Power 5 - SEC, ACC, Big Ten, Big 12, Pac-12



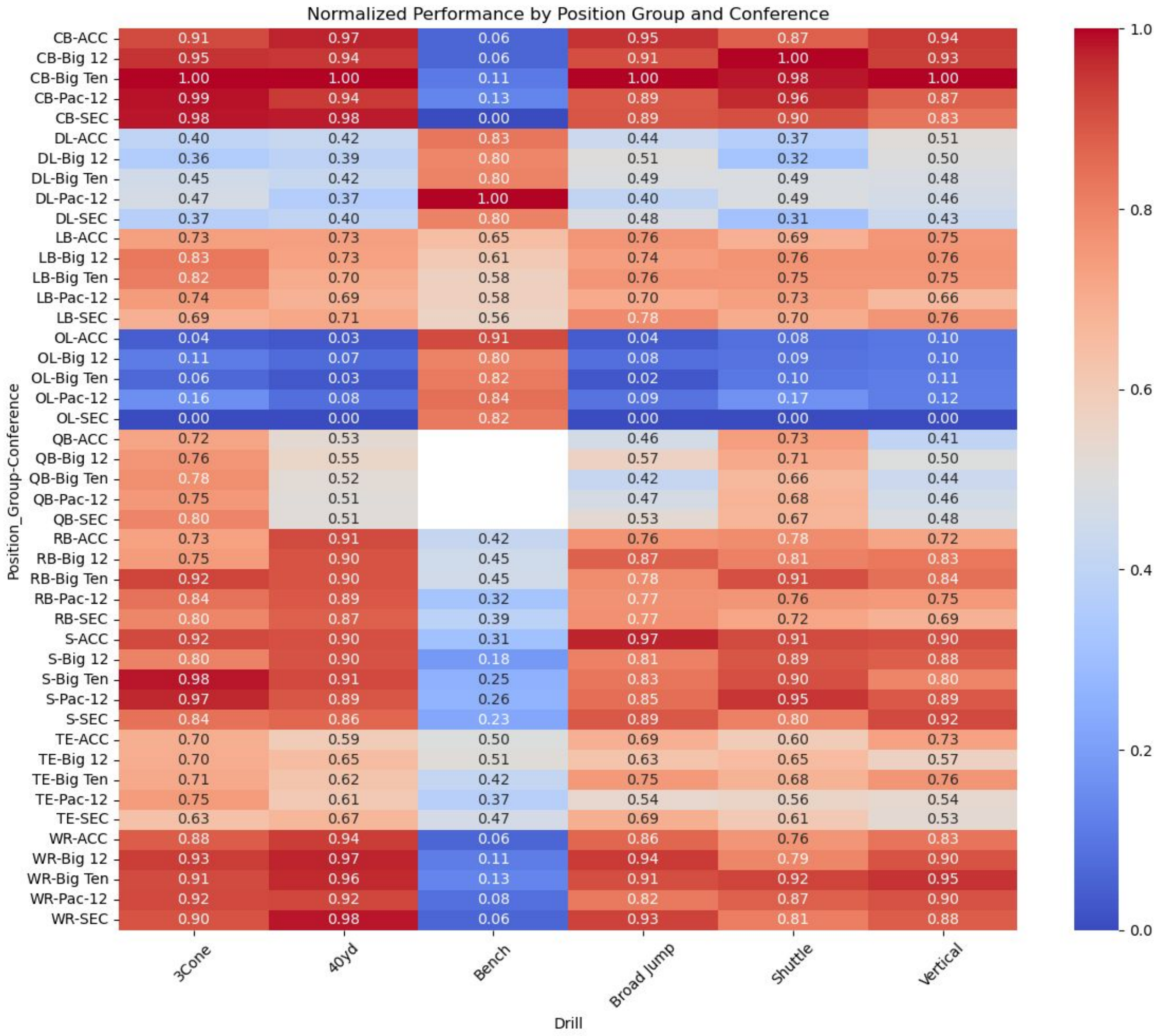


- SEC players are drafted earlier on average (3.51 round) compared to other Power 5 conferences and Non-Power 5 schools (4.15 round).

- Indicates a possible correlation between conference prestige and draft position.

- Visualization: Heatmap showcasing normalized combine performance across positions and conferences.

- Diverse performance profiles across different Power 5 conferences for various position groups, there is no standout conference that dominates others in all aspects (ex. Pac-12 DL perform exceptionally well on bench press, ACC safeties performed best on broad jump, Big Ten corners verticals, SEC WRs 40 yd times, etc.)



Machine Learning Models

Logistic Regression

Model Performance:

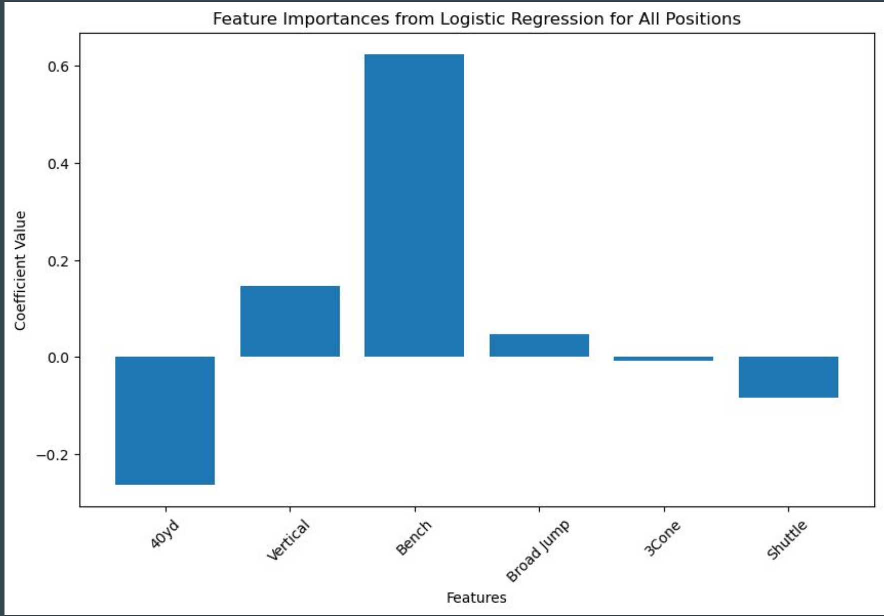
- Accuracy: 69% — The model correctly predicted draft outcomes for 69% of players.
- Precision (Not Drafted): 72% — High precision for predicting non-drafted players.
- Recall (Not Drafted): 21% — Low recall, indicating many undrafted players were missed.
- Precision (Drafted): 69% — Reasonable precision for predicting drafted players.
- Recall (Drafted): 95% — High recall, most drafted players were correctly identified.

Feature Importance:

- Bench - Most positive impact on draft prediction.
- 40yd & Shuttle: Negative coefficients, but interpretation is not straightforward.
- Vertical & Broad Jump & 3Cone: Marginal positive influence.

Note: The model treats all player positions equally; specific positions like linemen might show different feature importance patterns. Position-based filtering could refine the model's predictive power.

	precision	recall	f1-score	support
0	0.72	0.21	0.33	145
1	0.69	0.95	0.80	261
accuracy			0.69	406
macro avg	0.70	0.58	0.56	406
weighted avg	0.70	0.69	0.63	406



Random Forest

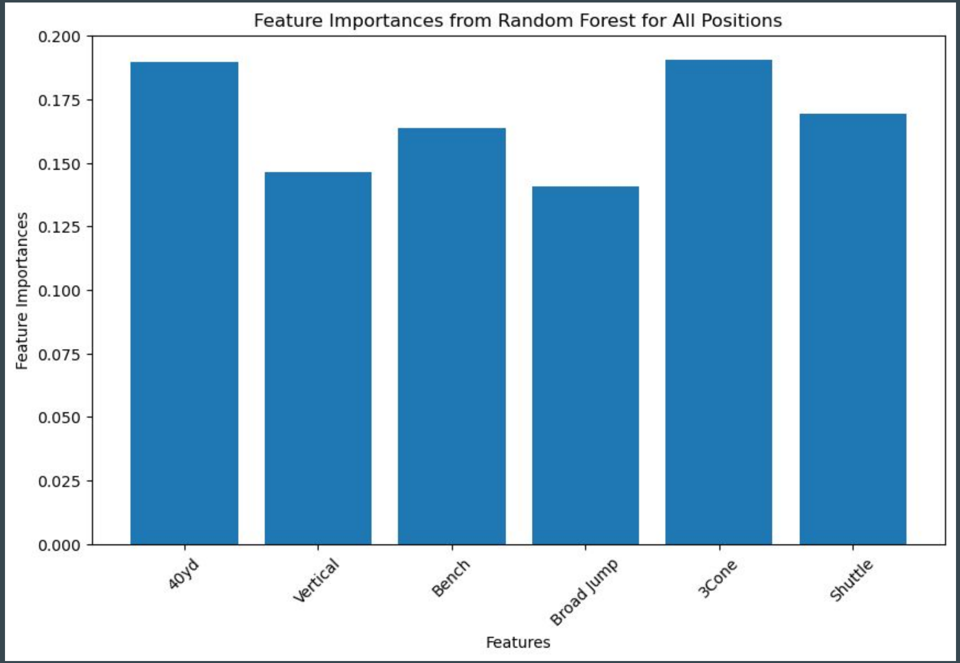
Model Performance:

- Accuracy: 68% — Slightly lower overall prediction accuracy compared to Logistic Regression.
- Precision (Not Drafted): 58% — Moderate precision, with some false positives.
- Recall (Not Drafted): 34% — An improvement in identifying non-drafted players compared to Logistic Regression.
- Precision (Drafted): 70% — Comparable to Logistic Regression.
- Recall (Drafted): 86% — Most drafted players were correctly predicted but lower than Logistic Regression.

Feature Importance:

- 40yd Dash: Identified as the most influential feature for predicting drafts.
- Vertical, Bench, & Shuttle: Important, but less so than the 40yd dash.
- 3Cone: Least important feature in this model.

	precision	recall	f1-score	support
0	0.58	0.34	0.43	145
1	0.70	0.86	0.77	261
accuracy			0.68	406
macro avg	0.64	0.60	0.60	406
weighted avg	0.66	0.68	0.65	406



K-Means Clustering

```
Cluster 0 Draft Rate: 0.5344827586206896
Cluster 1 Draft Rate: 0.5616438356164384
Cluster 2 Draft Rate: 0.6538461538461539
Cluster 3 Draft Rate: 0.7151515151515152
Cluster 4 Draft Rate: 0.7723214285714286
```

Clustering Output

We applied K-Means clustering to our NFL Combine data from 2010-2023 to group players into five distinct clusters based on their performance metrics (feature selection: 40yd, Vertical, Bench, Broad Jump, 3Cone, Shuttle)

The K-Means algorithm helped group players with similar combine performance profiles, and the clusters show their respective likelihoods of being drafted.

Our clustering output is interactive and provides key summary statistics for each player when you hover over each point. Click on the link above to interact with our visualization.

Decision Tree

Decision Tree Visual

We utilized the decision tree model with the NFL Combine data from 2010-2023 to group players into five distinct clusters based on their performance metrics (same features of: 40yd, Vertical, Bench, Broad Jump, 3Cone, Shuttle)

Much like other decision trees, each node pivots around a feature's computed threshold (calculated in the data training process). In our case, the splitting criteria was based on each drill statistic.

The goal of the decision tree was to determine draft status based on the class of each node; as the depth of the tree increased, so did the specialization of players that would fall into different sections of the tree based on their respective performance in the drills (our features). The nodes would be classified as drafted or not drafted, ultimately providing a visual representation of the ratio of players drafted or not drafted into the NFL once the last level of .

Since the max depth of the tree was initialized at a level of 4, the visual representation of the tree (output) is quite expansive. Click on the link above to view our visualization.

Conclusion and Future Work

Overall Conclusion:

Power 5 school players demonstrate higher draft rates and are selected earlier in the draft, especially from the SEC.

Performance data indicates strengths vary across conferences and positions, and the competitive nature of Power 5 conferences contributes meaningfully to players' NFL draft prospects.

Suggestions for Future Analysis:

Use NLP to investigate sentiment analysis surrounding draft picks and the NFL in general. News articles, social media, and commentary could help gauge public and expert opinions on players to identify undervalued talent.

Neural Networks can be used to analyze movement patterns from video data (game/practice film). This could help identify and evaluate player techniques, agility, and decision-making in real situations and predict player performance. We can also compare with historically successful players and their movement patterns.