Final Project Report

DSCI 510

Kavin Phabiani

Peter Park

GitHub Repository:

https://github.com/kavinphab/DSCI_510_Final_Project

Our final project, "Bear Necessities: Finding the Chicago Bears' Next Draft Gem", took a data driven analysis approach to determine the 2026 draft pick for the Chicago Bears. As the two members of the project (Kavin Phabiani (kphabian@usc.edu USCID:6414233457) and Peter Park (pspark@usc.edu USCID:4240400712)), we decided to take a look at one of the most fascinating areas of sports this year. The Chicago Bears are having a breakout season with great players and a new head coach, surprising everyone with a winning record. The last time the Bears had a winning record was in 2018, with their only Super Bowl win ever in 1986. For the Bears to maintain their success, we decided to take one step ahead and see what they could do in the next season. Although they are having a phenomenal season, statistically, there are some areas they lack in comparison to the rest of the league. To identify the deficiencies, we analyzed the Bears' 2024 season performance across all positional groups using publicly available NFL (National Football League) statistics. The goal was to identify which position groups are underperforming relative to league averages and predict (1) which position the Bears most urgently need to draft in 2026, and (2) which college prospects from the NCAA (National Collegiate Athletic Association) best fit that positional need based on performance metrics.

We collected data from two main datasets. We needed data on the Chicago Bears as well as the rest of the NFL, which we gathered from SportsDataIO NFL Stats API. We used the requests library to authenticate through API key, downloaded the JSON formatted team statistics, and saved them locally for further cleaning. This dataset collection consisted of 32 teams (rows) with 225 statistical fields (columns). This allowed us to get a better insight on how the Bears were performing compared to other teams in the league. The next bit of information we needed were the stats on the college players. For this, we used the CollegeFootballData (CFBD) API, and using the same approach (API) as we did for the first dataset, we were able to gather information about our next NFL draft prospect from college football. This dataset provided 13,691 college players (rows) and their stats in 35 different categories (columns).

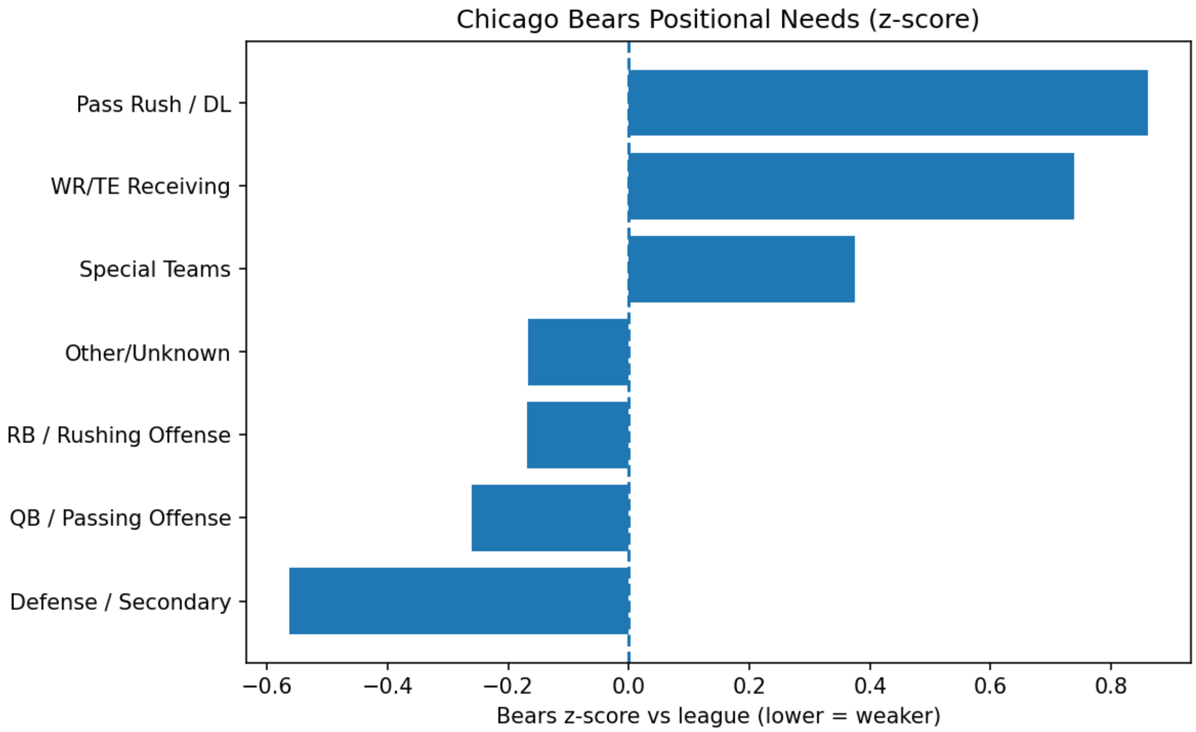
After we collected our data, we had to clean the data for analysis and visualization. To clean our data, we flattened the JSON data, making it tabular and organizing key variables. After flattening the file, we then saved the tables as csv files for processing. Cleaning our data also involved pivoting long form NCAA statistics into wide format using the pivot function, as well as removing any inconsistencies, where the API might not use the same column name to describe the same attribute. We also extracted the Chicago Bears specific data to zoom in on the team and compare to the rest of the NFL.

Originally, we planned to scrape NFL statistics directly from public websites. However, many NFL sites dynamically generate content using JavaScript, which made static scraping unreliable for this project. To ensure accurate and structured data, we switched to using SportsDataIO's authenticated NFL API. A second challenge involved NCAA stat formatting. The CFBD API returns statistics in a long format, where each row represents a single player-stat pair (e.g., "Player X – Tackles = 8"). To analyze players effectively, we needed to pivot this into a wide format. One row per player with columns for each stat was used for this. This required additional data cleaning, deduplication, and handling of inconsistent naming schemes.

Analysis of the Chicago Bears' season metrics was the most important part in being able to understand the deficiencies of the team. The method to accomplish this task was the Z-score standardization method. From the given metrics from the NFL dataset, the Bears' performance in each category was calculated against the league. The equation used was:

$$z = \frac{Bears\_Metric - League\_Mean}{League\_STD}$$

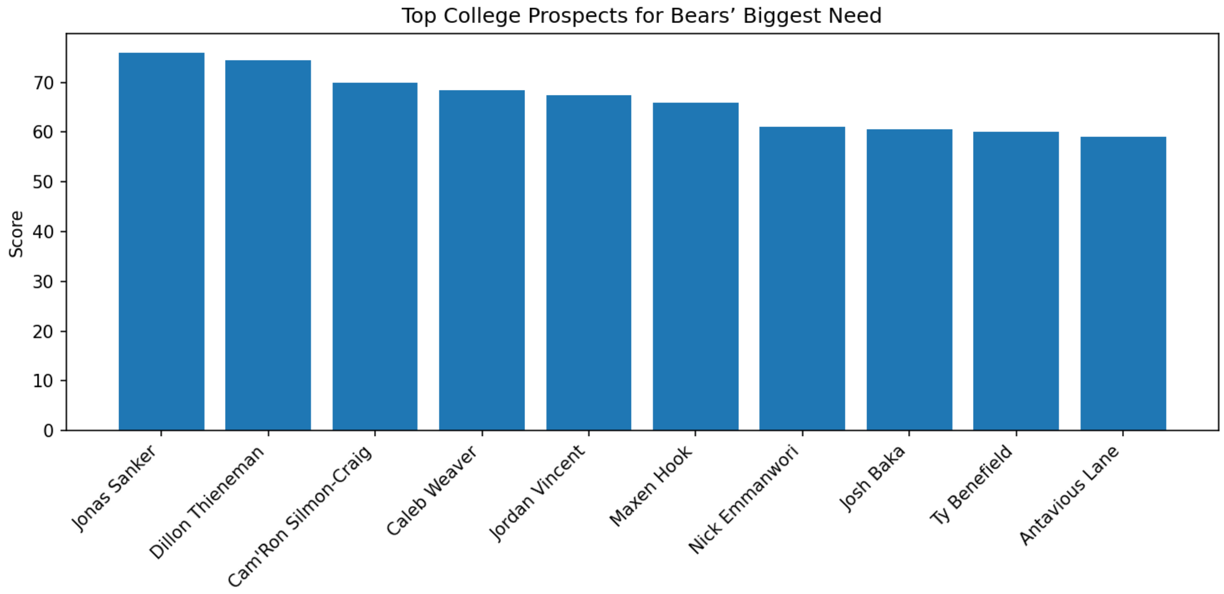
Which allowed the objective measurement of the Bears' performance:



This was then taken into evaluating which player in the NCAA is performing the best in pass defense and tackles. Knowing that the required position is identified, the calculation for the best pass defender (usually a safety (S), defensive back (DB), or cornerback (CB)) was performed with the NCAA dataset. A multitude of metrics were used to calculate a composite score for each defensive player. The metrics used were sacks (SACK), solo tackles (SOLO), total tackles (TKL), interceptions (INT), passes defended (PDEF), and tackles for loss (TFL). These are all metrics that are critical to playing a defensive back position in football. With these metrics, a composite score was calculated where every player gets a single numeric score that adds up all the critical metrics. Missing stats for these metrics were handled as zeros for fair processing. The math for the composite score was as follows:

$$Score = SACK + SOLO + TKL + INT + PDEF + TFL$$

The analysis and mathematics provided the scores for all potential players in the NCAA, with the best prospects were graphed in order:



The two plots shown above are the best visualizations for the analysis, as it shows clearly what message needs to be conveyed. The first plot is a horizontal bar chart that visualizes the Bears’ positional needs. It plots the z scores for different performance metrics for the Bears. This very easily shows that the lowest z score for the Chicago Bears. The second plot is the top college prospects for the Bears. It plots the composite scores on the y-axis, and the x-axis is all the prospects with the highest composite scores. This quickly shows all the top prospects, starting with the number one prospect.

Once the z score for the lowest value in all the categories was calculated and the data was visualized, it was evident that defense/secondary had the lowest score. This meant that the Chicago Bears were struggling in the secondary defense more than other areas. The Bears were not performing well in pass defense, guarding wide receivers against catching passes from the quarterbacks. To fill the gap, it would be beneficial for the Chicago Bears to draft a defensive back or a safety. From the second plot, it is evident that Jonas Sanker of the University of Virginia is the best prospect to fill the defensive gap the Chicago Bears currently suffer. However, from the current season, the Bears are projected to be the 25th pick in the draft next year. All the prospects shown on the plot are good options, in the case that another team would pick the best player before the Bears. This concludes that the Chicago Bears should keep their eyes on Jonas Sanker to fill their gap in pass defense, but because of the draft order, it would benefit them to keep their options open for other strongly qualified defensive backs such as Dillon Thieneman or Cam’Ron Silmon-Craig. This provides a critical, impactful and analytics-driven approach to the NFL draft rather than a speculative one. The biggest impact of this analysis project is the strong reliance on real life data, trends, and analytics for some of the most major decision such as an NFL draft. This is perhaps the biggest decision teams make every year, and the first pick of the draft for each team often go on to become hall of fame players that level up the performance of each team.

If we had more dedicated time towards this project, it would be fascinating to analyze the data based on the Bears' playbook as well. Having good players and executing plays is a key aspect of winning football games, but another big part of the game that people tend to overlook is the play calling. Football is comparable to a chess game between the defense and offense, and it would be interesting to see what offensive plays and defensive plays the Bears ran and how successful they were. This would require more extensive research on every single play the Bears ran both on offense and defense. The analysis would become more multi-layered, as the defensive and offensive plays would be matched to see if there are any correlations between plays, and it would be interesting to set up a clustering model to visualize where the successes and failures are.