An Analysis of the American Ultimate Disc League

Ryan Drost
Shivam Ahuja
Chris Chen

(1) Introduction

The use of data analysis has risen exponentially in professional sports in the last twenty years, drastically altering the style of play in many cases and allowing teams to optimize their strategy in order to maximize their chances of winning. The changes have been widespread and have affected all aspects of an organization, including on-field strategy, player substitutions, timeout management, contract offers, and anything in between. In contrast, the American Ultimate Disc League (AUDL), a professional Ultimate Frisbee league, has done comparatively little in this regard, and so we were interested in seeing what understanding and insight we could gain into how the game is played.

In case you are not familiar with Ultimate Frisbee, the sport is played 7-on-7 with a plastic disc instead of a ball, with teams attempting to pass the disc into the end zone for a score. Critically, players cannot move when they are holding the disc. Any incomplete pass for any reason (whether it was thrown away, dropped, or blocked or intercepted by the defense) is a turnover and the defense immediately takes possession and attempts to score in the opposite end zone. Once one team scores, the teams reset and the team that just scored kicks off to their opponent.

We explored two datasets relating to on-field play, one of player statistics and one of team statistics, downloaded from https://theaudl.com/.[1] The team statistics dataset has information for about 1280 games played from 2014 to 2022. Each row has identifying information (such as teamID and gameID) and the team's statistics for that game, with a separate row for each of the two teams that played in a given game. Some games only had data for one team, and we removed these unpaired rows from the dataset, as well as any rows pertaining to exhibition games. There were many missing values for some statistics which had not been recorded in earlier seasons, so we removed these variables entirely since we could not make meaningful comparisons.

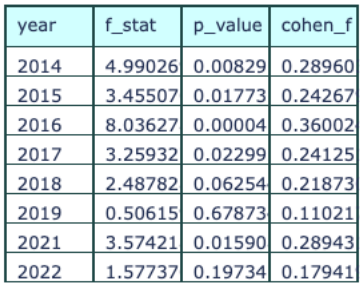
The player statistics dataset has information for games from 2012 to 2022, with a separate row for each game for each player, identified by playerID, teamID, and gameID, and displaying their statistics for that game. Once again, there were a lot of missing values because some statistics were not recorded in earlier seasons. We experimented with removing these variables, but opted to instead remove earlier seasons and only use the two most recent seasons (2021 and 2022) since these seasons have very few missing values and these variables appeared to have good predictive ability. For the remaining missing values in all variables, we either filled element-wise with the column average (the average over all player-games) or with zero, depending on what was more appropriate for the individual statistic.

(2) Inference

[1] Team statistics for each game are available here: https://theaudl.com/stats/team-game-stats. We scraped these as a json file from the following link, cycling through all pages: https://audl-stat-server.herokuapp.com/web-api/team-game-stats?limit=20&page=1.
Player career statistics are available here: https://theaudl.com/stats/player-stats. First we scraped these statistics in order to get a list of all player IDs, again cycling through all pages (https://audl-stat-server.herokuapp.com/web-api/player-stats?limit=20&page=1). Then we scraped each player's individual game statistics using the player ID (e.g.: https://theaudl.com/league/players/cbrock), through the json api (https://audl-stat-server.herokuapp.com/web-api/roster-game-stats-for-player?playerID=cbrock).
We manually created an additional dataset regarding divisions (with information from the same website) in order to label each game appropriately.

Within the AUDL, there are 4 divisions which each play almost all of their games against each other. Typically, the only games played between teams that are not in the same division are during the playoffs, and most teams do not play any opponents outside of their own division in a given season. Since there are very few direct data points comparing teams outside of the same division, we investigated whether the divisions are of varying quality.

In order to do this, we used turnovers as a proxy for quality of play (more turnovers indicates lower quality of play, and fewer turnovers indicates higher quality of play), and ran a one-way ANOVA test on the number of turnovers in each game.[2] We removed the handful of inter-divisional games, labelled each remaining game with the corresponding division, and calculated the total number of turnovers in the game. Because the overall quality of the league has increased drastically, we compared only games within each season; otherwise the test would be much less informative and some of the inherent assumptions, particularly homogeneity of variance, would be violated.

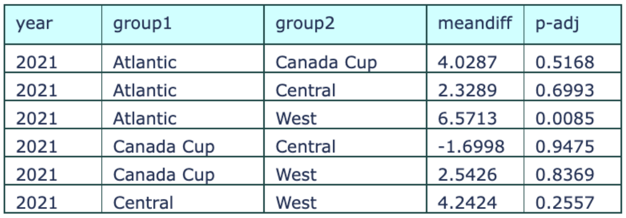


year	f_stat	p_value	cohen_f
2014	4.99026	0.00829	0.28960
2015	3.45507	0.01773	0.24267
2016	8.03627	0.00004	0.36002
2017	3.25932	0.02299	0.24125
2018	2.48782	0.06254	0.21873
2019	0.50615	0.67873	0.11021
2021	3.57421	0.01590	0.28943
2022	1.57737	0.19734	0.17941

Figure 1: ANOVA test results

Looking at each season, we found a significant difference ($\alpha = 0.05$) between divisions in each of the first 4 seasons (2014-2017), and also in 2021. However, for this past season $p = 0.197$, indicating that there is no evidence the divisions were of different quality.

In the next step, we looked at which pairs of divisions had a significant difference in their quality of play (for years where the ANOVA test produced a significant result), as determined by Tukey's Honestly Significant Difference (HSD) test ($\alpha = 0.05$). This controls for the family-wise error rate better than an independent samples t-test, which lowers the chance of a false positive and makes it a more natural follow-up to the ANOVA. Taking 2021 as an example, we observed that only the pairing of the Atlantic and West division had a statistically significant difference in quality ($p = 0.0085$).



year	group1	group2	meandiff	p-adj
2021	Atlantic	Canada Cup	4.0287	0.5168
2021	Atlantic	Central	2.3289	0.6993
2021	Atlantic	West	6.5713	0.0085
2021	Canada Cup	Central	-1.6998	0.9475
2021	Canada Cup	West	2.5426	0.8369
2021	Central	West	4.2424	0.2557

Figure 2: Tukey's HSD test results for 2021

[2] It could be argued that differing numbers of turnovers indicate a different style of play, rather than different levels of quality necessarily, but most knowledgeable people in the sport would agree that more turnovers is equivalent to lower quality (and potentially also a different style).

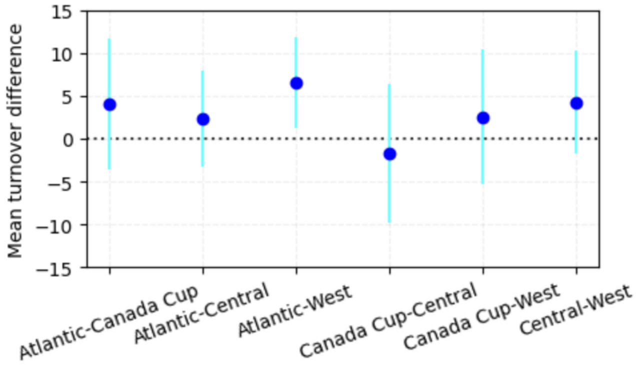


Figure 3: 95% CI of mean turnover difference between division pairs in 2021

Based on the results of these tests, we conclude that there is strong evidence the divisions were of differing quality in the seasons from 2014 to 2017. In particular, the ANOVA test revealed a fairly large effect size in 2016 ($f = 0.36$), and 3 of the 6 division pairings in that season showed significantly different quality in Tukey's HSD test (Central-South: $p = 0.0095$, Central-West: $p = 0.0001$, South-West: $p = 0.0078$). However, since 2017 only one total division pair has exhibited a significant difference in quality, and none did so in the most recent season, so we conclude that there is no evidence that there is currently a difference in the quality of the divisions.

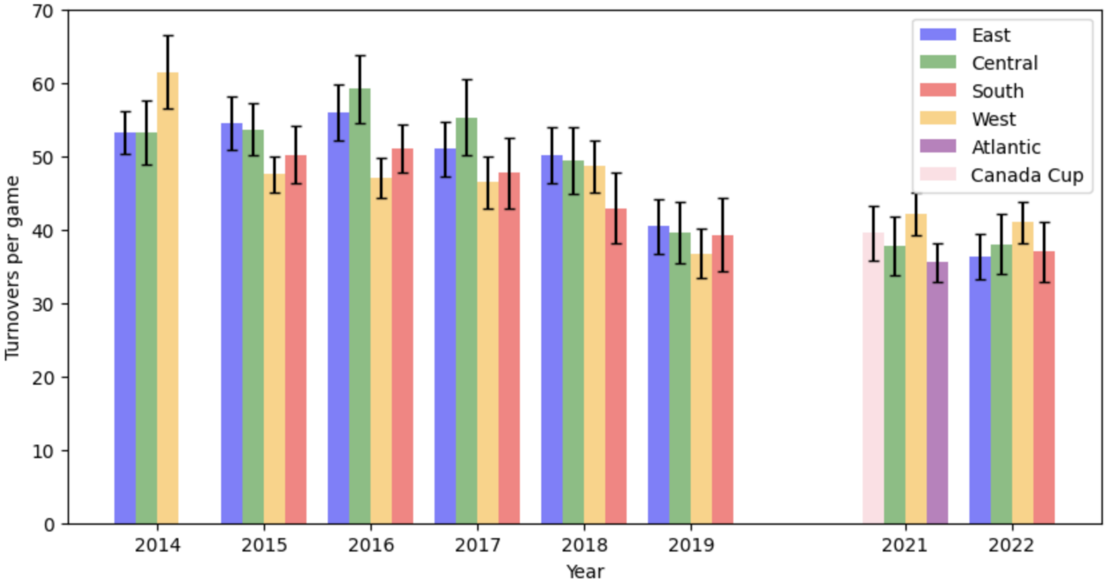


Figure 4: 95% CI of mean turnovers per game for each division (2014-2022)

(3) Prediction

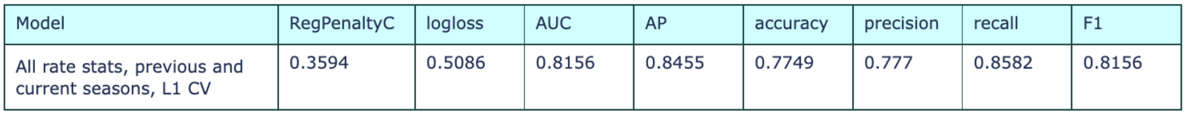
We attempted to predict game results (win or loss) based on data available prior to a given game, particularly based on each team's statistics for the previous season and to that point in the current season. Does past win percentage predict wins, while controlling for other past performance metrics?

We used logistic regression with L1 regularization to choose features and for cross validation. The LogisticRegressionCV function tests 10 logarithmically-spaced regularization penalty C's (larger

C's correspond to less L1 regularization), with 5-fold cross validation, maximizing accuracy, using liblinear as the solver/optimization algorithm. We controlled for confounds by including all variables as candidates in the logistic regression, so that each possible confound was incorporated with a coefficient in the final model.

In the processed dataset, we combined the two rows for each game so all data for a given game is in one row. We added the team's previous season's statistics as a variable in each row (replacing this with the column average if it was the team's first season) and the team's cumulative statistics to that point in the season (replacing this with previous season's statistics if it was the team's first game of the season). For each statistic, we then calculated the differences ($X_{Home} - X_{Away}$) and percentage differences ($(X_{Home} - X_{Away})/X_{Home}$) between the two teams to use as additional predictors. In total, there were 128 predictors.

Our target variable was whether the home team won (1) or lost (0). In the dataset, the home team wins 58% of the games. For our model, we used the data from 2015-2022, a total of 1151 games. We excluded 2014 because we did not have the previous season's data (and thus could not make meaningful imputations for the previous season's statistics variables) and also removed the two games that ended in a tie.[3] We performed a randomized 80/20 train/test split, but each row still only had data that was available prior to that game, so there should not be data or feature leakage. Finally, we standardized by scaling each variable to a mean of 0 and standard deviation of 1 using a scaler based on the training data.



Model	RegPenaltyC	logloss	AUC	AP	accuracy	precision	recall	F1
All rate stats, previous and current seasons, L1 CV	0.3594	0.5086	0.8156	0.8455	0.7749	0.777	0.8582	0.8156

Figure 5: Model results

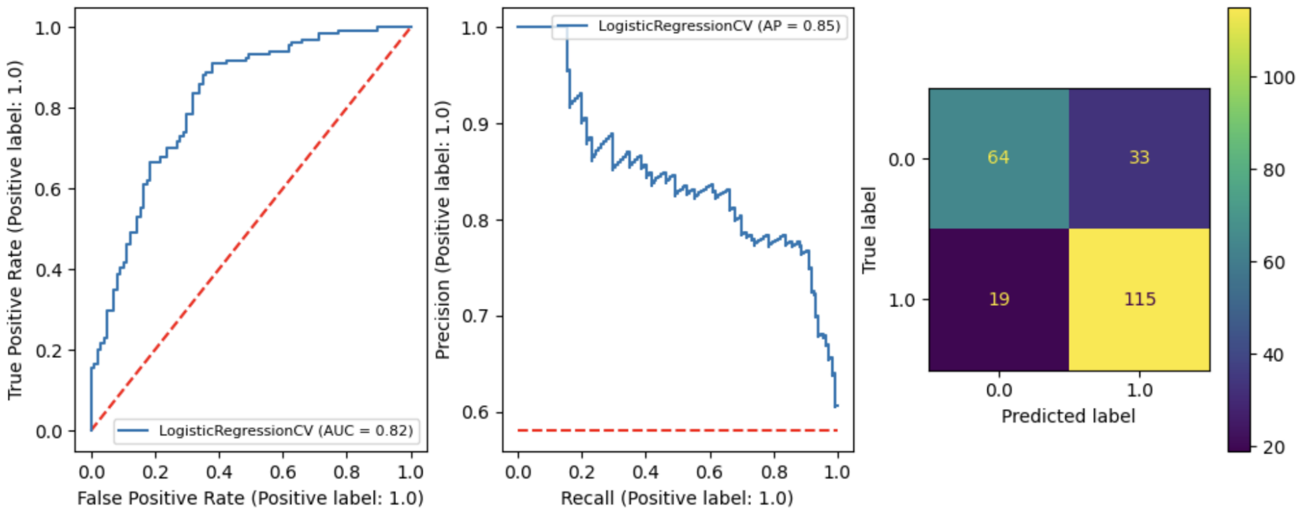


Figure 6: a) ROC curve, b) PR curve, and c) Confusion matrix

We are able to predict game results using the information about both team's statistics in the previous season and current season (prior to the given game). We achieved an AUC of 0.8156, an AP

[3] We were only interested in predicting a win or loss, as ties can only occur when games are ended early due to weather or other external factors (which explains why there are only two instances in the entire dataset).

of 0.8455 and an accuracy of 0.7749. The L1 regularization resulted in 30 nonzero coefficients, out of the original 128. Notably, win percentage (specifically the $(Home - Away)/Home$ percent difference) did have a strong positive predictive contribution, as the current season variable ($\beta = 0.3134$) and previous season variable ($\beta = 0.1806$) were among the ten strongest predictors as determined by the absolute value of the coefficients.

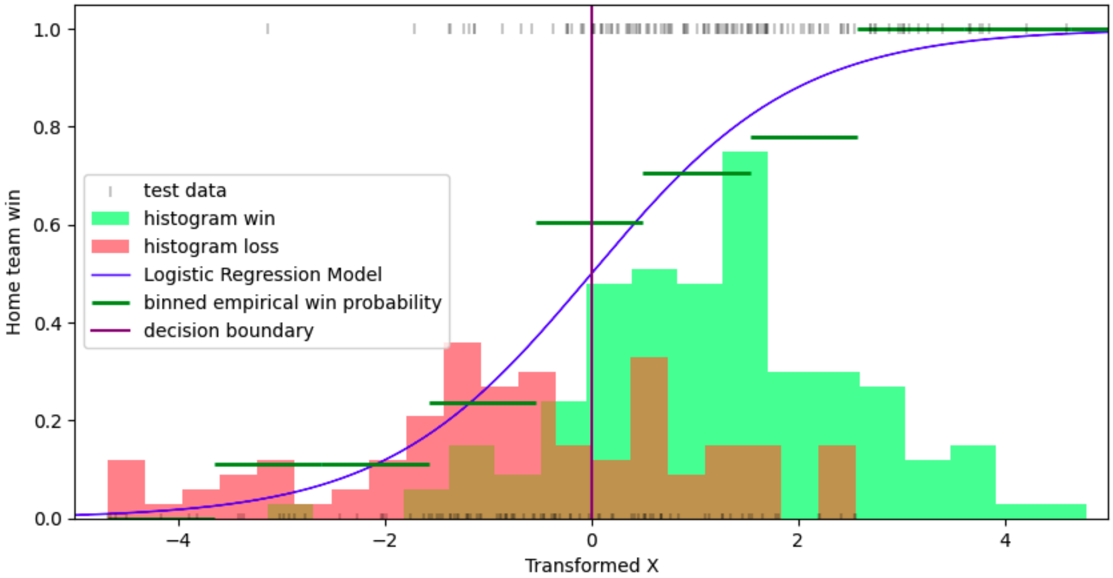


Figure 7: Logistic plot with transformed test data using coefficient matrix β and intercept $\beta_0 : X' = (\beta X^T + \beta_0)^T$

(4) Classification

While most sports have well-defined positions tagged to each player such as striker, midfielder, or defender in soccer, in the AUDL there are not as strict roles assigned to each player. Players likely still have these types of roles assigned to them internally (either explicitly or implicitly) as part of their team's overall strategy, and we wanted to see if we could identify distinct player types based off a player's statistics; and, following that, if we could build a classification model with sufficient accuracy to predict these roles for future players.

In order to do this, we first filtered the dataset to include only the most recent two seasons (2021 and 2022).[4] Then, we calculated average per game statistics by rolling up the game level data by player, excluding players who played fewer than 20 total points, since they likely don't have enough data to categorize effectively and would impact the model's performance negatively. The missing values in the player level database were imputed using either 0 if it was an individual statistic such as pull time, or the column average if it was a statistic which involved team performance, such as o_pct (which is the percent of offensive points the team scores when the player is on the field). We then ran a principal component analysis on the dataset to reduce our dimensions before running the k-means clustering algorithm. We chose 9 principal components since they covered 90% of the data variance.

[4] Some advanced statistics like hucks (Frisbee terminology for long passes), passing yards, and receiving yards were captured starting in 2021 and would significantly help in defining player roles, and thus it made sense to limit our dataset to just the years these statistics were available.

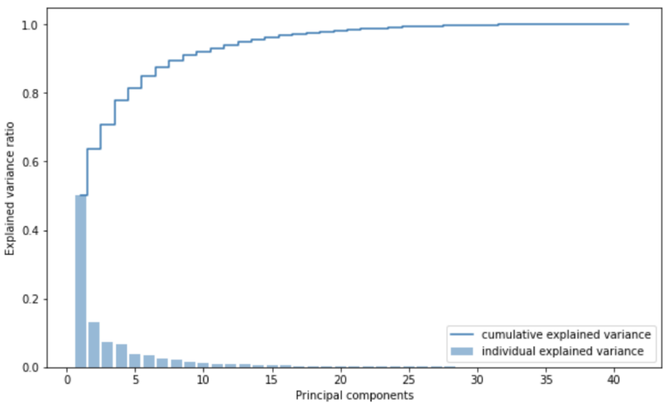


Figure 8: Principal Component Variance plot for player level data

Then we analyzed the loadings for each variable on each principal component to identify the ones which had the highest loading on the same principal component, and chose variables for each principal component as their representatives to further reduce the variable set. We performed a correlation analysis to remove variables which were highly collinear in order to remove similar information before running the clustering algorithm. Finally, on this reduced set of variables, we ran the PCA algorithm to represent the data in lower dimension and ran a k-means clustering algorithm on the PCA transformed data to return the final clusters.

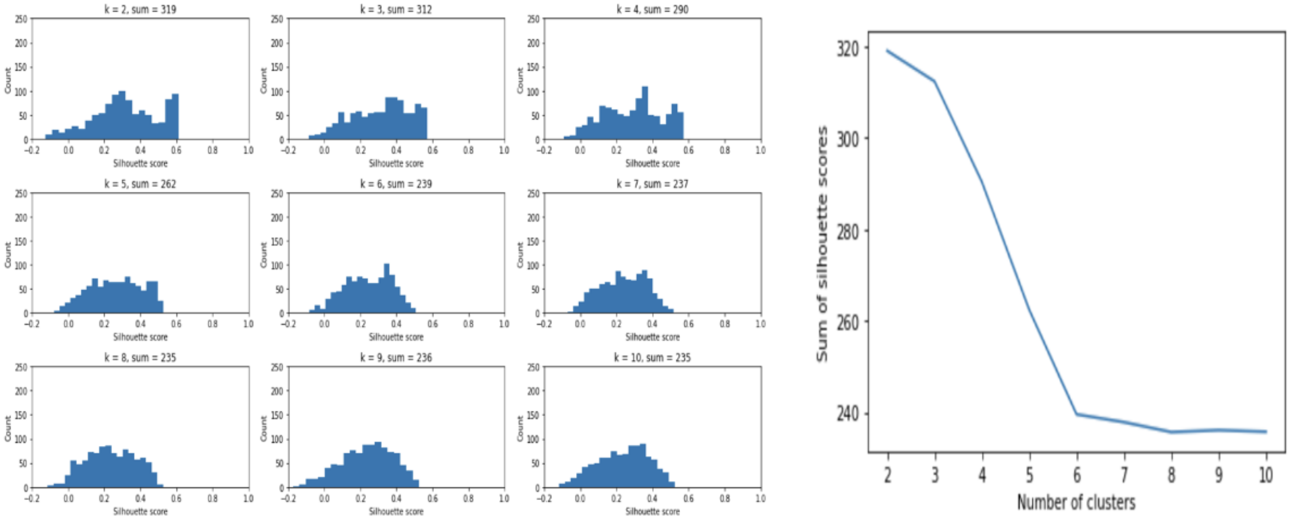


Figure 9: a) Silhouette score histogram for different values of k and b) Sum of silhouette score against number of clusters

We identified 4 final clusters using the silhouette method outlined above. The drop from 2 to 4 in silhouette score was minimal and from game understanding and expert opinion, we knew that there were more than 2 types of player roles in the game and thus we went with 4. K-means clustering was utilized with $k = 4$ to find the different types of player roles. There were 384 players in the first cluster, 146 in the second, 337 in the third, and 133 in the fourth. The clusters are explained in detail below.

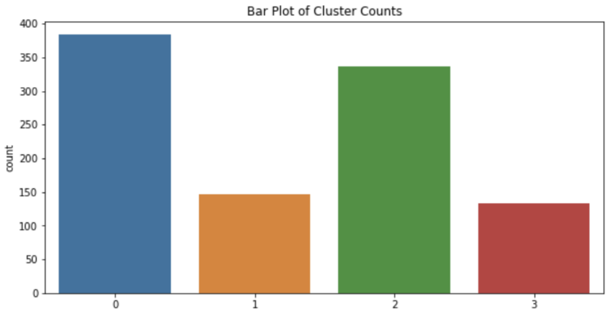


Figure 10: Size of each cluster

(a) Cluster 0: **Role players**
   - Have a relatively low average in most statistics
   - Second highest in yards per catch and higher yards received than thrown
   - Not the stars of the team, on the field these players fill in the gaps between the other players

(b) Cluster 1: **Receivers**
   - Very high in receiving yards, with more moderate throwing yards and more goals than assists, though they are high in both
   - High amount of yards per catch, indicating they are often catching long passes for goals
   - Critical offensive players who are attempting to score most often, akin to a striker in football

(c) Cluster 2: **Defenders**
   - High number of blocks and pulls
   - Low in most offensive statistics, notably goals, assists, receiving yards, and throwing yards
   - Critical defensive players who are are likely tasked with guarding the other team's star players

(d) Cluster 3: **Quarterbacks**
   - Very high in throwing yards, touches (catches), and hucks (long throws) attempted
   - More assists than goals, though they are high in both
   - Critical offensive players who keep the game flowing by distributing the disc to their teammates, and often also by throwing hucks to the receivers

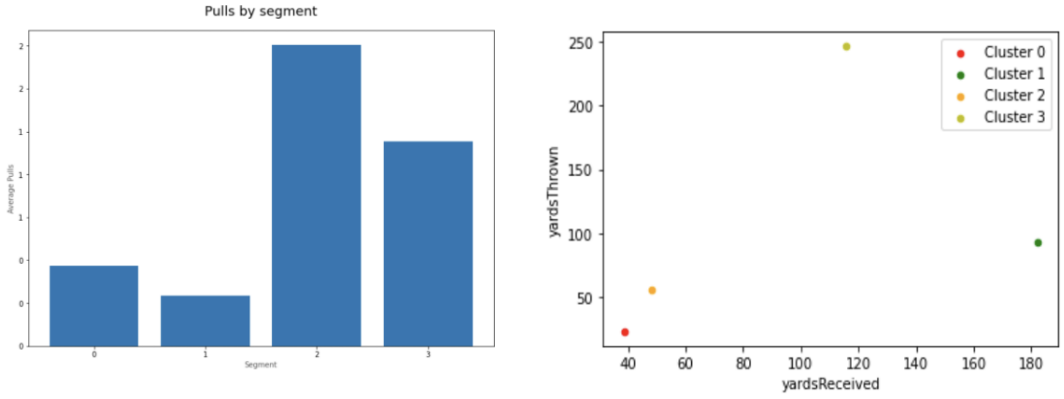


Figure 11: a) Average pulls (defensive statistic) by clusters and b) yards thrown vs yards received scatter plot by clusters

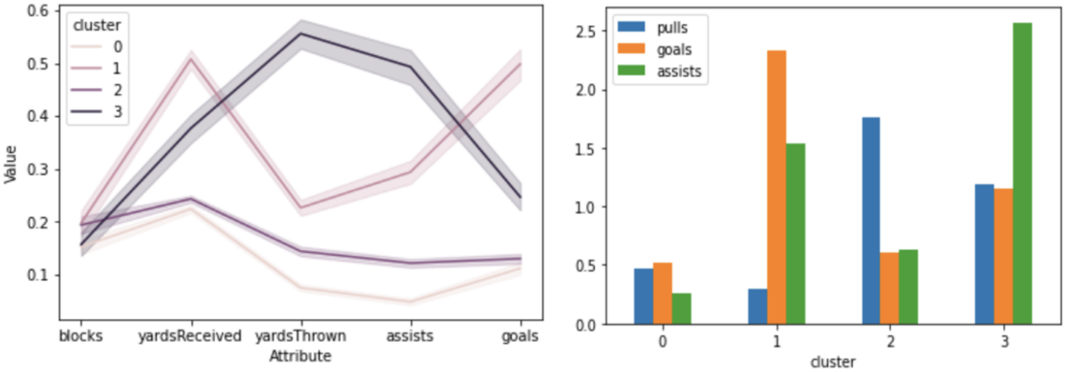


Figure 12: a) Line plot for key metrics by clusters and b) Goals, assists, and pulls by cluster

After identifying the player roles, we then ran a regularized multinomial logistic regression algorithm to check if we can predict these roles for any future player. Our target variable was the segment from the cluster analysis, and we performed a randomized 80/20 train/test split with each row still only containing data for that particular player in order to avoid data or feature leakage. The L2 regularization was utilized and cross validation was done for hyperparameter tuning and a range of penalty terms were passed to identify the optimal penalty term value (0.01).

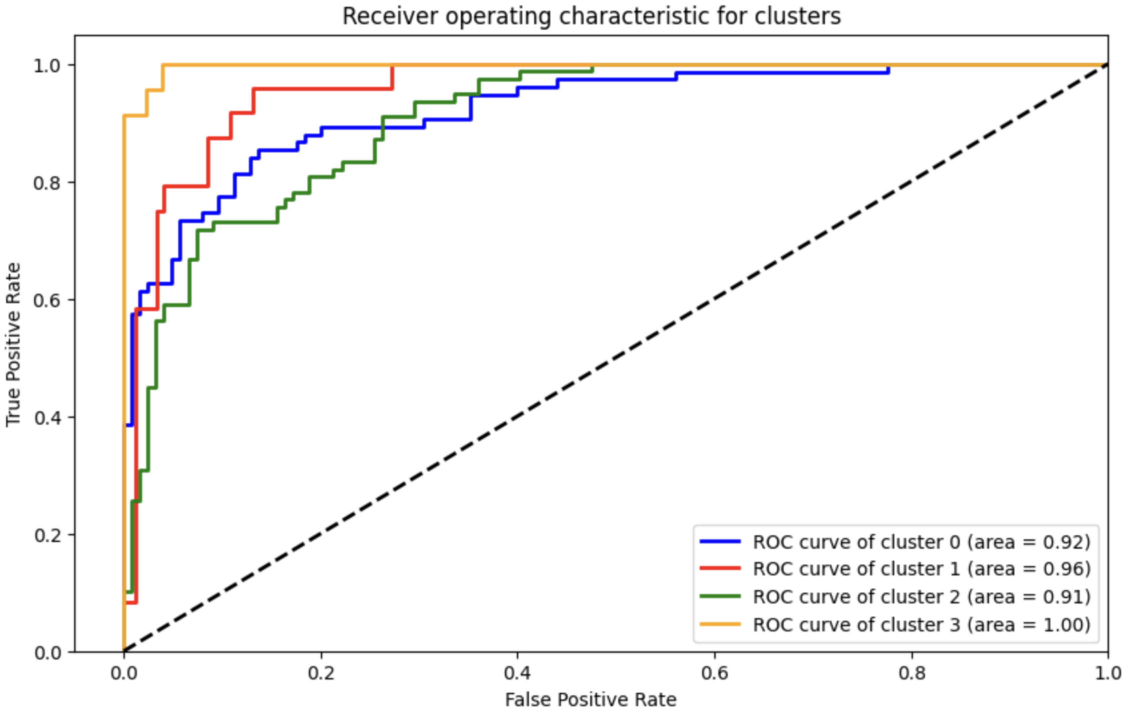


Figure 13: ROC Curve for each cluster

We were able to predict the cluster for each player using the per game average statistics of that player from the past two seasons. We achieved an accuracy of 0.845, an average precision of 0.86, and an average recall of 0.86. This model could be utilized to tag future players and also help current players identify their role in the team and help the team better.

(5) Conclusion

Overall, we have tested and quantified our intuitions about the AUDL. We can see that the quality of play in the league has increased over time, and the various divisions have become more equal as a result. We can also infer at a general level that this may make predicting game results more difficult going forward, as the teams become less stratified. Being able to factor in specific players based on their statistics and their role may become crucial to improving upon the model as this happens. By the same token, we could look to incorporate some of our win prediction features to identify how much each individual player has contributed to winning.

In terms of limitations, in our prediction model we have a large number of predictors that are collinear and not independent of each other. This also makes the model coefficients harder to interpret. While we wanted to include multiple transformations of the original data to see if one version is more predictive, in the future we may want to use fewer predictors which are more independent and would make the coefficients, and the model in general, easier to interpret. We can also observe how the model does with actual games in the upcoming 2023 season. For our clustering algorithm, in the future we could gather labelled data, potentially simply from observation, and use that to identify player types and help train our model.

More broadly, we could investigate how these aspects all interact. How can we integrate player statistics and player roles into our model for wins prediction? How can we use our analysis about divisional differences to investigate whether there are also different player types in different divisions (due to different styles of play)?

We could potentially answer these types of questions with the data we have currently, but certainly there are some pieces that simply aren't available because they aren't being recorded. For instance, there is no tracking data showing where players are on the field and how fast they are moving, which is now common in the major professional sports. There also is no data about the types of throws each player is making. Because the sport is played with a disc rather than a ball there are a wide array of types of throws, but in our dataset all passes are simply recorded the same, with no information on the type of throw. With data like this, we could undoubtedly improve on the analysis we have already done, and open up many new opportunities to evaluate all aspects of on-field play.