**Final Project: Proposal**

Derek Caramella & Miguel Novo Villar

Department of Computer Science, University of Rochester

CSC440.2: Data Mining

Dr. Luo

28 October 2021

###### Executive Summary

The researchers seek to uncover the player-type combinations that will generate the most points within a Barclays Premier League (BPL) season. Players are clustered by market value and points generated; therefore, categorizing players as gold, silver, and bronze level players. First, the researchers will conduct linear optimization analyzing game week data from the 2016/2017 to the 2020/2021 season that will produce maximum yielding player combinations within the financial constructs. Second, the researchers will replace the players with their corresponding cluster and conduct frequent pattern analysis to identify frequent line-up types that yielded maximum points. Third, the investigators will generalize their results by testing if previous BPL championship winning teams hold the frequent point maximizing combinations identified. Lastly, the researchers will conduct an inverse investigation, the researchers will minimize the number of points, which will produce the worst line-up combination. Finally, the investigators will create a map (network) exhibiting the dissimilarity between the maximum frequent line-ups and the minimum frequent line-ups.

**Motivation**

An all-inclusive 7-night vacation in the United Kingdom with VIP hospitality to two Barclays Premier League (BPL) football fixtures appears attractable, two collegiate football players. Twenty teams comprise the BPL, the top-flight football league in Great Britain. The BPL hosts an open “fantasy” league challenge to select a 15-player squad that earns points dependent on real-game performance; the winner is awarded the all-inclusive prize. The goal is to optimize player selection at the beginning of the season and effectively construct starting lineups to maximize points week-over-week.

**Model Constraints**

At the start of the season, the fan is allocated £100 million to purchase two goalkeepers, five defenders, five midfielders, and three forwards; moreover, the fan may not exceed three players from a single BPL team. Each week, the fan selects a starting 11 squad from the 15-player roster, additionally, the fan elects a captain – the captain’s points are doubled in the game week. If a starting player in our line-up does not play, it is swapped by a bench player. The following table outlines the scoring system:

|  |  |
| --- | --- |
| Action | Points |
| For playing up to 60 minutes | 1 |
| For playing 60 minutes or more (excluding stoppage time) | 2 |
| For each goal scored by a goalkeeper or defender | 6 |
| For each goal scored by a midfielder | 5 |
| For each goal scored by a forward | 4 |
| For each goal assist | 3 |
| For a clean sheet by a goalkeeper or defender | 4 |
| For a clean sheet by a midfielder | 1 |
| For every 3 shot saves by a goalkeeper | 1 |
| For each penalty save | 5 |
| For each penalty miss | -2 |
| Bonus points for the best players in a match | 1-3 |
| For every 2 goals conceded by a goalkeeper or defender | -1 |
| For each yellow card | -1 |
| For each red card | -3 |
| For each own goal | -2 |

**Data Source**

The researchers will develop an algorithm that will identify sets of player-type combinations that maximize points within the fantasy league’s constructs. The investigators will utilize the Anand (2020) GitHub repository for a data source. The repository contains game week specific and historical season data for each player in the BPL spanning from the 2016/2017 season to the 2020/2021 season. The table below outlines the features of the data set per season:

|  |
| --- |
| Attribute |
| Name |
| Position |
| Team |
| Experience |
| Assists |
| Bonus |
| Bps |
| Clean Sheets |
| Creativity |
| Element |
| Fixture |
| Goals Conceded |
| Goals Scored |
| Influence |
| Opponent Team |
| Own Goals |
| Penalties Missed |
| Penalties Saved |
| Yellow Cards |
| Red Cards |
| Saves |
| Selected |
| Away Team Score |
| Home Team Score |
| Total Points |
| Transfer Balance |
| Was Home |

**Proposed Solution**

After data cleaning and integration, the researchers will obtain week-by-week player performance from the 2016/2017 to the 2020/2021 season. Although the data breadth is valuable, the data is too granular to produce generalizable teams for future years. As seasons ensue, players depart the club, retire, or become injured; therefore, we will analyze the week progressions on a less granular level. The investigators will conduct kernel clustering by partitioning the data into four data sets: goalkeeper, defender, midfielder, and attacker. Then, create clusters on market value and points generated to group players. Subsequently, the investigators can analyze the data set from a less granular perspective.

The researchers will implement a linear optimization approach to generate optimal line-ups week-over-week within the constructs of the fantasy league’s rules. After the line-ups are generated, the investigators will replace the player names with the corresponding cluster. Therefore, the researchers will obtain a transactional data frame that has a set of 15 cluster labels for each transaction (game week). The investigators will analyze the transactional data frame utilizing the Apriori pattern mining algorithm. After the Apriori algorithm, a list of frequent subsets of optimal line-ups by player-type will be generated.

Although obtaining frequent optimal line ups by player type is an intriguing finding, the researchers will test their results by evaluating if the BPL title winning teams from the 2016/2017 to 2020/2021 seasons contain the frequent line-ups discovered. The results will display the research’s applicability because the championship winning teams should hold the exact type of player combinations previously identified

Lastly, an inverse investigation will be conducted, seeking to answer the question: What is the worst combination within the fantasy league constraints? Instead of conducting linear optimization to maximize points, the goal will be to minimize points. Then, the investigators will conduct pattern mining to identify the frequent itemsets. Lastly, the researchers will construct a map (network) depicting the distance between sets to exhibit dissimilarity.

**Project Expectations**

The researchers seek to uncover a new pattern for purchasing players within a constrained budget to generate maximum points and identify player-types that are frequently paired, that generate maximum points. Additionally, the researchers strive to verify that point maximining player-type sets are present on championship-winning teams. Lastly, the investigators will evaluate the dissimilarity between the point maximining line-ups and minimizing line-ups.

**References**

Anand, V. (2020, September 31). *Cleaned players* [Data set]. Retrieved from https://github.com

/vaastav/Fantasy-Premier-League/blob/master/data/2021-22/cleaned\_players.csv

Cay, A. (2021, May 12). Hindsight optimization for FPL. *Alpa Code*. Retrieved from https://alps

code.com/blog/hindsight-optimization/

Ge, T., An, Z., Cai, H., & Wang, Y. (2020, August). An analysis on the effectiveness of

cooperation in a soccer team. *2020 15th International Conference on Computer Science & Education (ICCSE)* 787-794.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2008). Clustering and sequential

pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge*

*and Data Engineering, 21*(6), 759-772.

Sæbø, O. D., & Hvattum, L. M. (2019). Modelling the financial contribution of soccer players to

their clubs. *Journal of Sports Analytics, 5*(1), 23-34.