Designing the Optimal Lineup in English Premier League Football

Derek Caramella & Miguel Novo Villar

Department of Computer Science, University of Rochester

CSC440.2: Data Mining

Dr. Jiebo Luo

13 December 2021

Abstract

The aim of this project is to propose an algorithm that demonstrates patterns that help teams allocate their budget in order to achieve a higher position in the Barclays Premier League season. Following clustering players by position and various other features derived from the 2016-2017 to 2020-2021 seasons - the researchers identified the frequent player type combinations in top-flight and relegation class clubs. The frequent player type combinations in a top-flight club are: {DefenderGold, DefenderGold, MidfielderGold}, {DefenderGold, MidfielderSilver, MidfielderGold, MidfielderGold}, {DefenderGold, MidfielderSilver, MidfielderGold}, and {MidfielderGold, MidfielderSilver, MidfielderSilver}; while the frequent squad compositions among the bottom teams are: {DefenderSilver}, {DefenderGold, MidfielderSilver, MidfielderSilver}. Two distinct groups separate the similarity between top-flight and relegation class clubs; however, a few clusters exist such that top-flight and relegation teams overlap – exhibiting that team management and tactic formation affect club success.

1. Introduction

Expending minimal capital to achieve the championship is the chief goal of many club owners' ambitions in the British Premier League (BPL). Although a few clubs have enormous capital reserves for acquiring players – Manchester City, Liverpool, Manchester United, and Chelsea, most of the clubs' talent is constrained by a limited wage budget. Total wage budget contributes to league position, according to the ridge regression in *Figure 1.1* below[1]; some privileged clubs discover success without extensive spending such as Leicester City and Ajax.

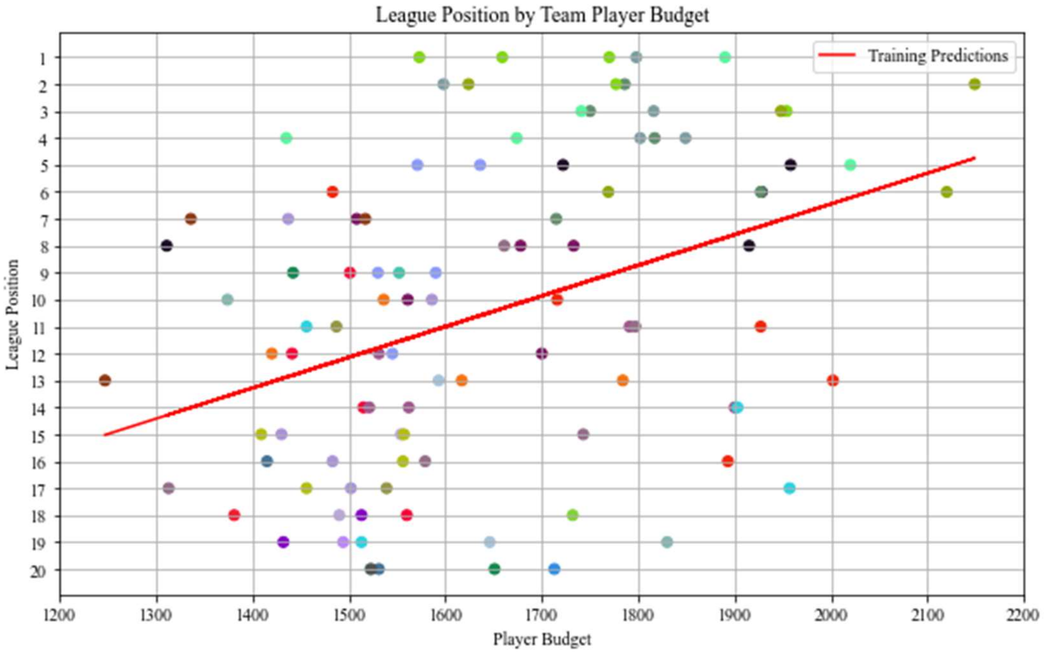

Figure 1.1: A ridge regression exhibiting the relationship between player budget and league position by team. On average, a higher budget increases league ranking.

Our research questions are as follows:

- What player types do top-flight clubs invest in to differentiate themselves from relegation level clubs?
- How are the capital wage budgets allocated to improve athletic performance?
- What player types contribute the most to maximize winning percentage?

We seek to answer these questions using clustering, frequent-itemset analysis, and other statistical techniques to demonstrate that winning teams share similar characteristics and playing styles in the BPL.

[1] Each color represents a club.

2. Related work

We examined literature related to pattern generation, budget optimization and clustering – all related to football lineup generation.

Gangal, Talnikar, Dalvi, Zope & Kulkarni (2015) creates a point system algorithm that aims to improve the fantasy premier league model by creating a system of predictions that potentially will increase the engagement of fans to the game. The proposed algorithm is an equation that creates an "AlgorithmScore" for each team based on predictions from the following fluctuating weighted factors: current form, injuries, suspensions, previous games and potential effects of other tournaments.

Bloomfield, Polman, & O'Donoghue (2007) aims at identifying and outlining the physical demands of soccer through computerized time-motion video-analysis by position. Proves that every spot has different movements, patterns and required physical conditioning. The main takeaway is the classification methodology used to divide the team in defenders, midfielders, and strikers - based on empirical evidence.

Gollan, Bellenger & Norton (2020) evaluates how different factors such as formation or tactics affect performance on professional football teams. The researchers used logistic regression and odds ratios across different cluster teams. Results demonstrate similar patterns across top-flight teams and relegation teams on how opposition patterns affected their results.

There are several studies that create a hypothetical algorithm aiming to generate the best possible team within certain characteristics. The truth is that several characteristics such as coaching, communication or other intangible attributes occur outside the scope of these projects. Nevertheless, they help us to understand different techniques used to make an approximation and quantification of this phenomenon.

3. Methodology

The researchers utilized centroid-based clustering and frequent itemset analysis to identify frequent team composition among the top 25 percent of teams (top-flight clubs) relative to the bottom 25 percent of teams (relegation class clubs) in the BPL. The researchers classified players into gold, silver, and bronze tiers utilizing *k*-means++ clustering with minutes played, number of assists, goals scored, goals conceded, clean sheets, yellow cards, and market price features. In *Figure 3.1*, the correlation matrix exhibits that market price is heavily correlated to assists and goals scored; moreover, goals conceded, scoreless games, and yellow cards are correlated to the number of appearances.

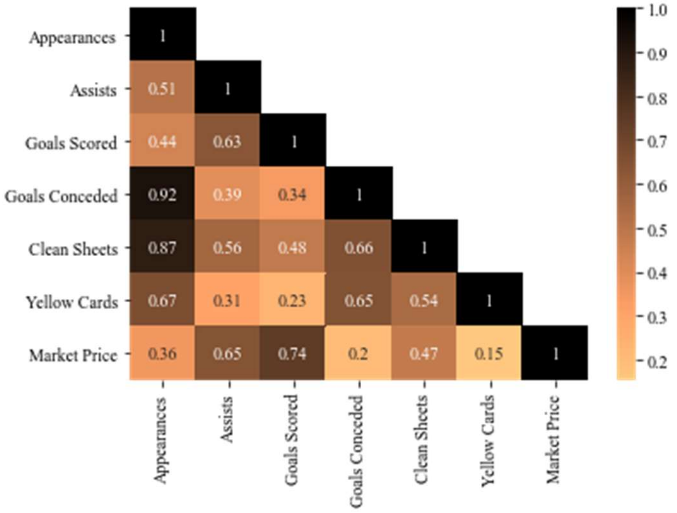

Figure 3.1: The correlation matrix displays the cluster features. Market price is correlated with assists and goals scored; furthermore, goals conceded, scoreless games, and yellow cards are correlated.

After clustering, the game week data is constructed at a less granular level, which enables effective frequent pattern itemset mining. We partitioned the data set into two data frames: (1) the top five teams and (2) the bottom five teams in the league table. Then, we treated each game week, between 2016-2017 and 2020-2021, as an itemset producing approximately 730 itemsets for each partition. Lastly, we evaluated the dissimilarity among top-flight and relegation class clubs utilizing *t*-distributed stochastic neighborhood embedding (*t*-SNE) (Maaten & Hinton, 2008). We identified the frequent itemsets present in top-flight and relegation class clubs using clustering and frequent itemset mining, then compared the similarity between top-flight teams and relegation class teams.

The researchers utilized the *k*-means++ clustering methodology to identify general-purpose, flat geometric clusters to appropriate players into a three-tiered ranking system. The *k*-means++ algorithm chooses the initial features first by selecting one center uniformly at random from the objects. Then iteratively, for each object other than the chosen center, the algorithm calculates the distance relative to the initially selected center and maximizes the distance between clusters. Following initialization, the algorithm calculates the mean of the cluster, then assigns the mean as the new center. When the mean does not change across iterations, the cluster is solidified. *K*-means++ speeds up convergence and ensures higher quality results because the initialization is dispersed across the domain relative to an unfortunate initialization from a vanilla *k*-means algorithm. The researchers utilized *k*-means++ to analyze player data at a less granular level and to conduct frequent itemset analysis.

The researchers utilized the Apriori algorithm to conduct frequent itemset analysis (Toivonen, 2017). The Apriori algorithm creates a table containing the support count of each item in the dataset, then compares the support with the hyperparameter minimum support, dropping the itemsets that do not satisfy the minimum support. Then, from the remaining itemsets, the algorithm generates candidate itemsets by combining the candidates from the previous step. The algorithm scans the database to identify the candidate support. The algorithm removes candidates that do not satisfy the minimum support. The process is continued iteratively until no frequent itemsets are found in the last iteration.

We utilized the *t*-SNE dimensionality reduction technique to visually represent the dissimilarity between top-flight and relegation class teams. Unlike Principal Component Analysis (PCA), *t*-SNE attempts to preserve local structure by minimizing the Kullback-Leibler divergence (KL divergence) between two distributions respective to the location of the points on the map (Minka, 2000). Moreover, *t*-SNE is less sensitive to outliers relative to PCA. We identified the frequent itemsets for top-flight and relegation teams, then one-hot encoded the clubs from 2016-2017 to 2020-2021, such that an array identifies if a team possesses all, a portion, or no frequent itemsets[2]. The researchers utilized *t*-SNE to compare the dissimilarity between the top-flight and relegation class clubs on a cartesian plane since the data is multidimensional.

[2] [{DefenderGold, DefenderGold, MidfielderGold}, {DefenderGold, MidfielderGold, MidfielderGold, MidfielderSilver}, {DefenderGold, MidfielderGold, MidfielderSilver}, {MidfielderGold, MidfielderSilver, MidfielderSilver}, {DefenderSilver}, {DefenderGold, MidfielderSilver, MidfielderSilver}]

4. Experiment

The researchers utilized the Anand (2020) GitHub repository to develop a game week data set for analysis. For each season between 2016-2017 and 2020-2021, the researchers stitched tuples together utilizing primary and composite keys to acquire player records for each game week with the features in *Table 4.1*.

Table 4.1: All features available in the dataset.

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

We labeled players as gold, silver, or bronze tier relative to their position, for example, Goalkeeper Gold, Defender Silver, Midfielder Gold, and Forward Gold. The researchers relaxed the model assumptions by partitioning the data set into five projections for each season because a gold tier player in one season may not be considered a gold tier player in another season due to game progression, off-season transfers, and/or league bolstering. Moreover, the researchers partitioned the data set into four dataframe subsets to cluster by position. Lastly, the researchers preprocessed the data by utilizing the min-max methodology to scale the features into a range between zero and one.

$$x'_i = \frac{v_i - min_A}{max_A - min_A}$$

The radar plots (*Figure 4.2*) below depict position-specific characteristics relative to the clusters. Moreover, the figures are plotted on a natural logarithmic cartesian plane. As expected, the gold tier for each position overtakes the lesser tiers, while the silver tier encompasses the bronze tier.

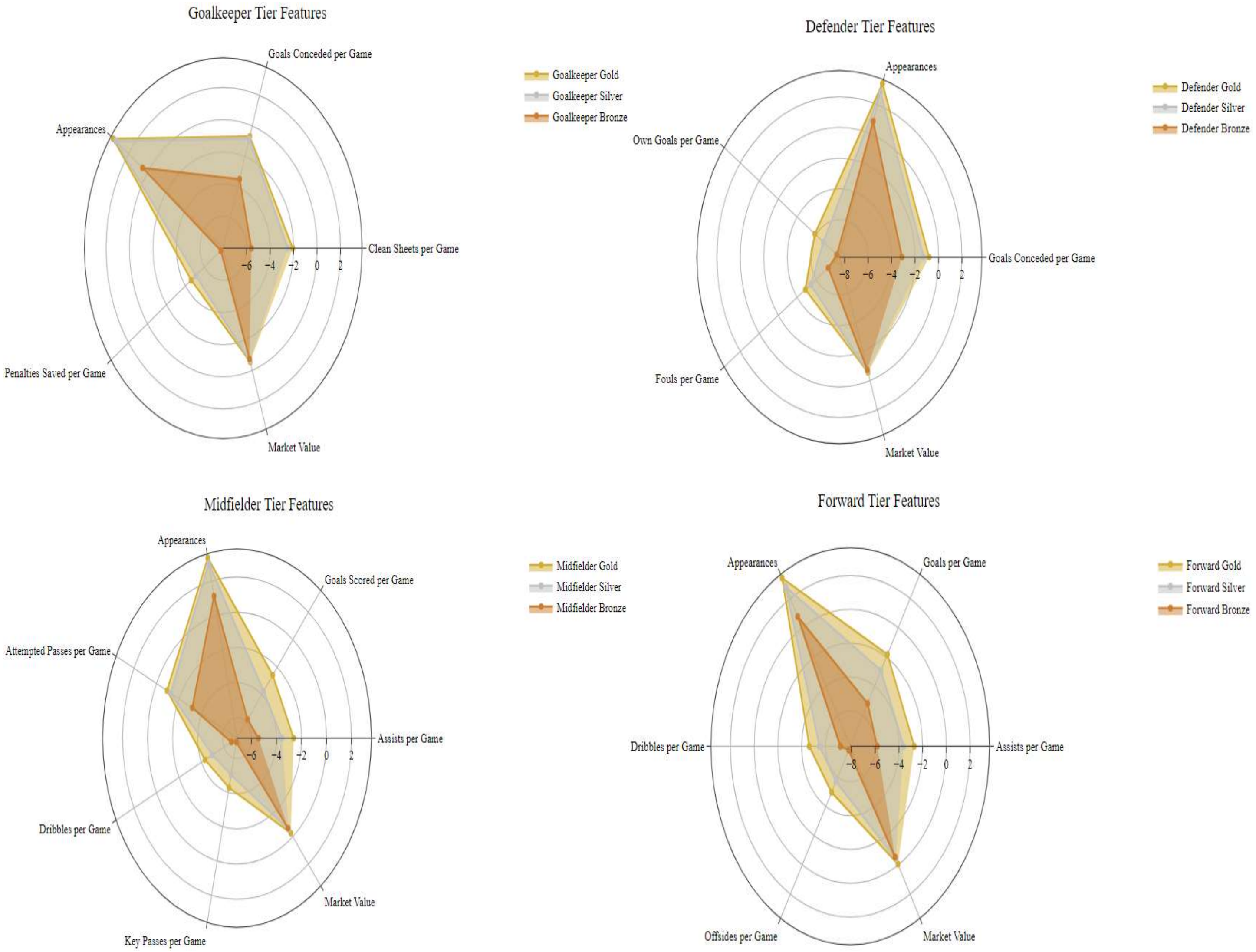

Figure 4.2: The average player in the gold, silver, and bronze tiers for each position with relative position features (natural log theta). Since gold exceeds silver and silver exceeds bronze in each radar plot, the clustering was effective identifying gold, silver, and bronze tier players for each position.

The plot (*Figure 4.3*) below depicts club position, assists, and goals where assists and goals are scaled with the natural logarithm. The *x* markers represent midfielders, and the circle markers represent forwards. The plots show all teams possess bronze tier players regardless of league position. Contrary to expectation, the midfielders coexist in the upper echelons of goal scoring with forwards, while forwards rarely contribute the number of assists midfielders provide. Lastly, it is evident that as the league position increases the higher the tier is present in the team sheet.

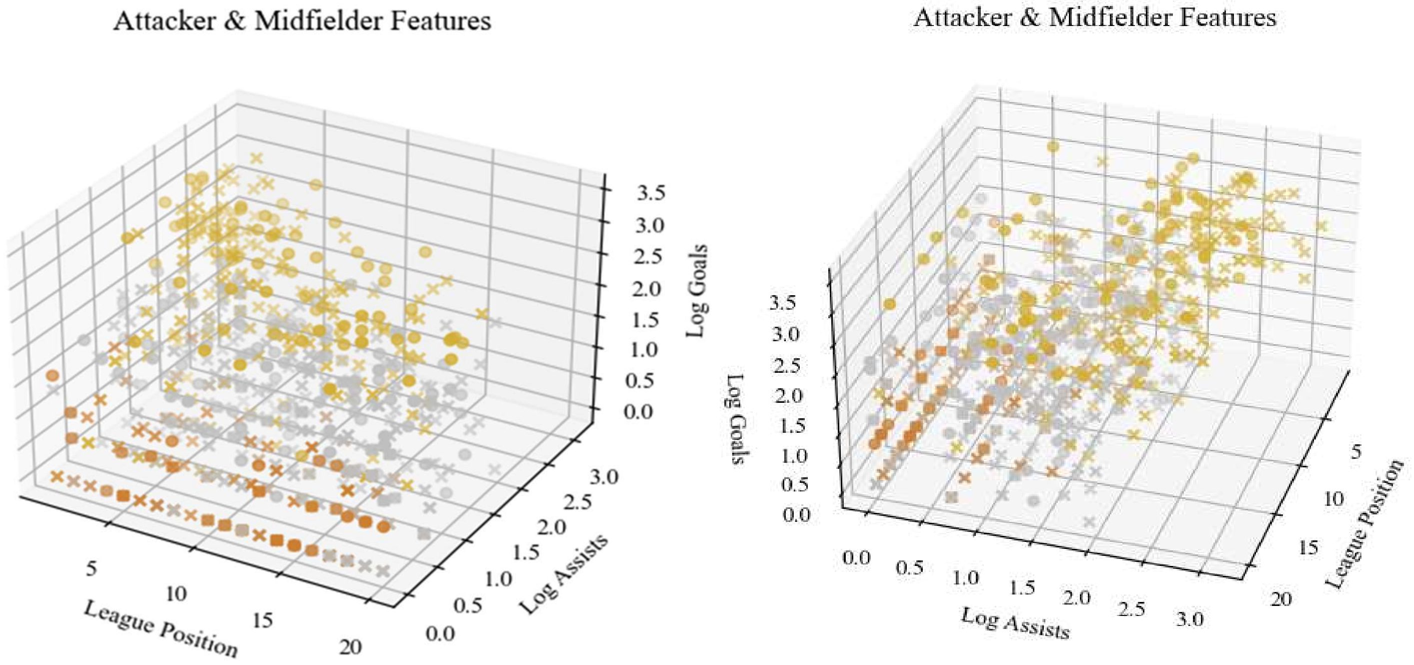

Figure 4.3: A three-dimensional plot displaying league position, assists (natural log), and goals scored (natural log); additionally, x markers represent midfielders while o markers represent forwards. The plot indicates that all teams possess bronze players regardless of league position and some gold tier midfields share the high goal scoring space with gold forwards. Lastly, gold-tiered midfielders dominate the high assist space.

The boxen plots (*Figure 4.4*) below compare midfielders and forwards on a univariable scale: goals scored, and assists completed. The upper tier of a gold midfielder contributes the same expected goal count as a gold forward, which was present in the three-dimensional plot above. However, within the silver tier, the forwards and midfielders provide nearly identical assists. In addition, there is significant overlap between the gold and silver tiers of goalkeepers on the scoreless-game feature. The similarity is mirrored between the defender gold and silver tiers.

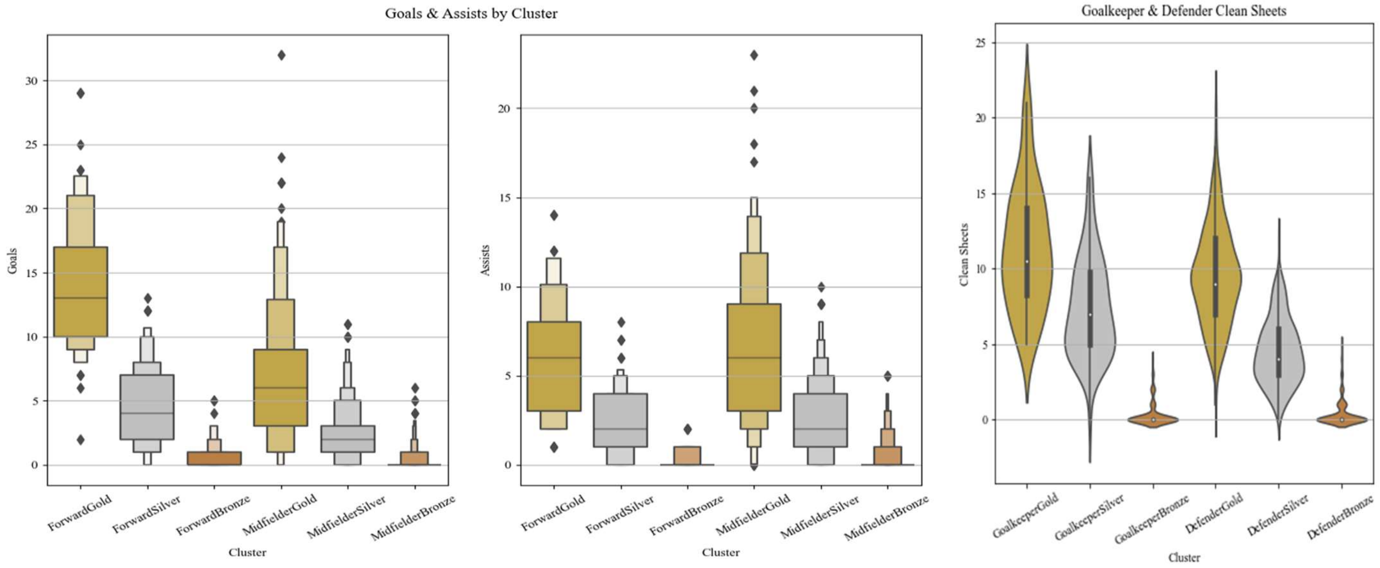

Figure 4.4: The boxen plots compare midfielders and forwards on goals and assists. The plots exhibit that above average gold-tiered midfielders score the expected value of gold-tiered forwards; moreover, gold-tiered midfielders contribute more assists relative to a gold-tiered forward. The violin plot exhibits the relationship between goalkeepers and defenders on scoreless games. Goalkeepers at every tier, save for bronze, overtake defenders in clean sheets.

The researchers partitioned the dataset into two discrete dataframes, (1) the top five teams and (2) the bottom five teams between the 2016-2017 to 2020-2021 seasons. We consider each game week as an itemset; thus, each itemset is composed of cluster labels. For example, a game week may be represented as {GoalkeeperGold1, DefenderBronze1. DefenderSilver2, DefenderGold1, MidfielderGold4, MidfielderSilver3, ForwardGold2}. After executing the Apriori algorithm (75 percent minimum support) the frequent squad compositions among the top teams are: {DefenderGold, DefenderGold, MidfielderGold}, {DefenderGold, MidfielderSilver, MidfielderGold, MidfielderGold}, {DefenderGold, MidfielderSilver, MidfielderGold}, and {MidfielderGold, MidfielderSilver, MidfielderSilver} while the frequent squad compositions among the bottom teams are: {DefenderSilver}, {DefenderGold, MidfielderSilver, MidfielderSilver}. The frequent itemsets are applicable to the BPL, but the results are not applicable to soccer as a whole due to various playing styles.

Figure 4.5 exhibits the dissimilarity between top-flight and relegation class clubs. For each season, each top-flight and relegation club was one-hot encoded by the frequent itemset features; applicably, each array is a club that exhibits the frequent itemsets present in the roster. We implemented the *t*-SNE technique[3] to reduce the multidimensional arrays and maintain the local data structure, then plotted the clubs by the league position class. Two defined clusters appeared on the cartesian plane, one containing top-flight clubs the other containing relegation class teams. Moreover, there are sparse, sporadic clusters of relegation class clubs. Lastly, there exists a few clusters containing top-flight and relegation class teams, these clubs contain sufficient player type combinations on their roster. However, team management and player wellness inhibited team performance. Hence, features such as management style and formation impact team performance, in addition to the model's features. The *t*-SNE results yielded that two well-defined clusters exist and separate top-flight and relegation class clubs; however, a few spaces exist such that team management and playing style are integral components to achieve success.

[3] Perplexity: 30, Learning Rate: 200, Number of Steps: 1,000

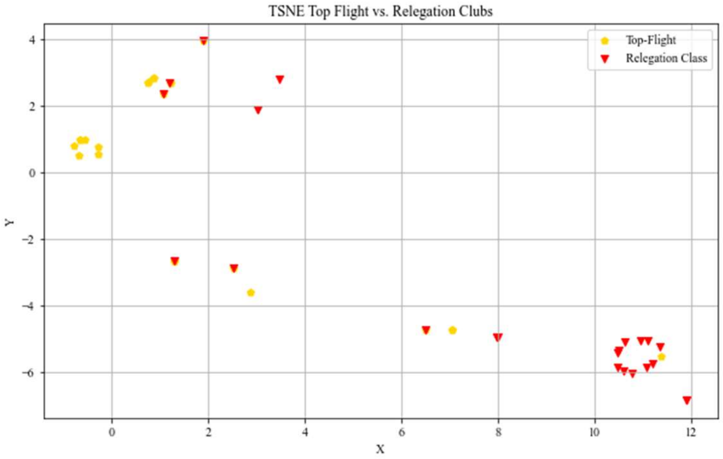

Figure 4.5: A t-SNE plot exhibiting the similarity between top-flight clubs and relegation class teams. The t-SNE dimensionally reduces bit arrays where each array represents a team, and the features are the frequent itemsets. Two distinct clusters exist, (1) top-flight teams and (2) relegation class clubs. However, some sparse clusters exist that contain top-flight and relegation class clubs; these clusters depict teams that contain a winning squad, but do not receive appropriate training, tactics, and coaching to be successful.

According to the frequent itemsets, top teams focus capital on their midfield, which contain at least one gold-tiered midfielder. Gollan, Ferrar, and Norton (2018) asserted that top-flight BPL teams dominate transitions to offense, transitions to defense, and establish an offensive presence (p. 1001). Moreover, top-ranked clubs employ a modern wing-back, the defensive winger joins the attack to maintain possession and generate chances, while maintaining the primary responsibilities as a defender. The frequent itemsets reflect these qualities because the midfielders are integral for efficient, rapid transitions from the back line towards the opposition and vice versa. Additionally, the clustering results conveyed that above-average gold-tiered midfielders notched the expected value as a gold-tiered forward; therefore, gold-tiered midfielders are an optimal investment. The gold-tiered defenders are the wing-backs that are utilized for counter-attacks, fast breaks, and early crosses. According to Bloomfield, Polman, and O'Donoghue (2007), forwards expend approximately 40 percent of the match performing purposeful movement and partake in high contact competitions. The forward's purpose is to finish the ball after a wing-back (gold-tiered defender) delivers a cross or a midfielder delivers the ball into a dangerous area. If a team concedes possession, the gold-tiered midfielders should be competent to regain possession and develop an attack, such as Manchester City. The top-flight frequent itemsets are conducive to the BPL playing style and contemporary tactics. The relegation grade clubs exhibit a weak midfield, which limits possession and restricts scoring opportunities.

Evaluating the relegation class frequent itemsets, each set contains at least a silver or gold tier defender and at least one silver midfielder. According to Gollan, Ferrar, and Norton (2018), lower-performing teams employ well equipped defenders to maintain the score-line. The gold tier defenders exhibit this characteristic because the defenders are center-backs, not wing-backs relative to top-flight clubs. The top-flight clubs may default to silver tiered center-backs because the clubs can rely on possession from a strong midfield and rapid wingbacks. However, lower tiered teams require a gold-tiered center-back to lead and defend incessant attacks from the opposition. A prime example of capital misallocation from a relegation class club is Swansea City, the club purchased a gold-tiered forward, Wilfred Bony; however, the midfield behind the forward was silver, consequently, Bony was unable to receive the ball due to an inapt midfield. Moreover, since the midfield was weak, a gold-tiered center-back, Ashley Williams, was needed to keep the team from being relegated. The frequent itemsets for bottom-class clubs in the BPL reflect the necessity for defense without a competent midfield to ensure seamless transitional moments within a game.

5. Conclusion

Our results establish that a club should invest in a strong gold-tiered midfield accompanied by at least one strong wing-back to optimize a club's wage budget. Moreover, a club should seek a strong, physical, and talented finisher as a forward to complete chances provided by the midfield and wingers. A club heavily investing in gold-tiered center-backs and average midfielders will face challenges winning and maintaining possession; therefore, posting subpar score-lines. Above-average gold-tiered midfielders score the expected value as gold-tiered forwards and provide more assists compared to gold-tiered forwards; therefore, the above-average gold-tiered midfielders are effective on both ends of the attacking end. Our research shows that a well-allocated budget focusing on possession and tactic transition are prevalent in top-flight BPL clubs. Lastly, although a club may possess all frequent player-type combinations of top-flight clubs, the roster does not solely determine league success; in addition to the model's features, team management and tactics determine a club's success in the BPL. Future research should include a study of LaLiga relative to the BPL because the tactics rival each other. The expected result would be LaLiga strives for gold-tiered forwards while lessening the demand for gold-tiered defenders. Moreover, a study evaluating optimal formations would enhance the frequent itemset analysis. According to our frequent itemset analysis of top-flight clubs in the BPL, a team should invest in gold-tiered midfielders to reap the highest return on capital investment.

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

Bloomfield, J., Polman, R., & O'Donoghue, P. (2007). Physical Demands of Different Positions in FA Premier League Soccer. Journal of sports science & medicine, 6(1), 63–70.

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

F. Perez-Cruz, “Kullback-Leibler divergence estimation of continuous distributions,” 2008 IEEE international Symposium on Information Theory, 2008, pp. 1666-1670, doi: 10.1109/ISIT.2008.4595271.

Gangal, A., Talnikar, A., Dalvi, A., Zope, V., & Kulkarni, A. (2015). Analysis and Prediction of Football Statistics using Data Mining Techniques. International Journal of Computer Applications, 132(5), 8-11.

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.

Gollan, Stuart & Bellenger, Clint & Norton, Kevin. (2020). Contextual Factors Impact Styles of Play in the English Premier League. Journal of sports science & medicine. 19. 78 - 83.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T.E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Jin X., Han J. (2011) K-Means Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Minka, T. (2000). Automatic choice of dimensionality for PCA. Advances in neural information processing systems, 13, 598-604.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

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.

Toivonen H. (2017) Apriori Algorithm. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_27.

Van der Maaten, L.J.P.; Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605.