DESIGNING THE OPTIMAL LINEUP IN ENGLISH PREMIER LEAGUE FOOTBALL

Derek Caramella & Miguel Novo Villar

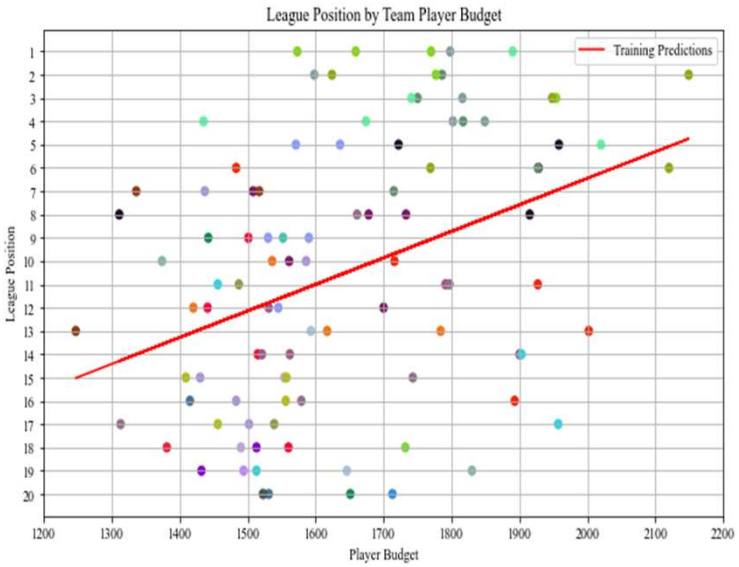
02 December 2021

Why do Managers Care About Team Composition?

- Wage Capital Scarcity
- Transfer Windows & Deadlines
- Dodging relegation is top priority
- **Who do I buy?**
- **Who do I play?**
- **What combination of player types perform successfully?**

Why do Managers Care About Team Composition?

- Wage Capital Scarcity
- Transfer Windows & Deadlines
- Dodging relegation is top priority
- **Who do I buy?**
- **Who do I play?**
- **What combination of player types perform successfully?**

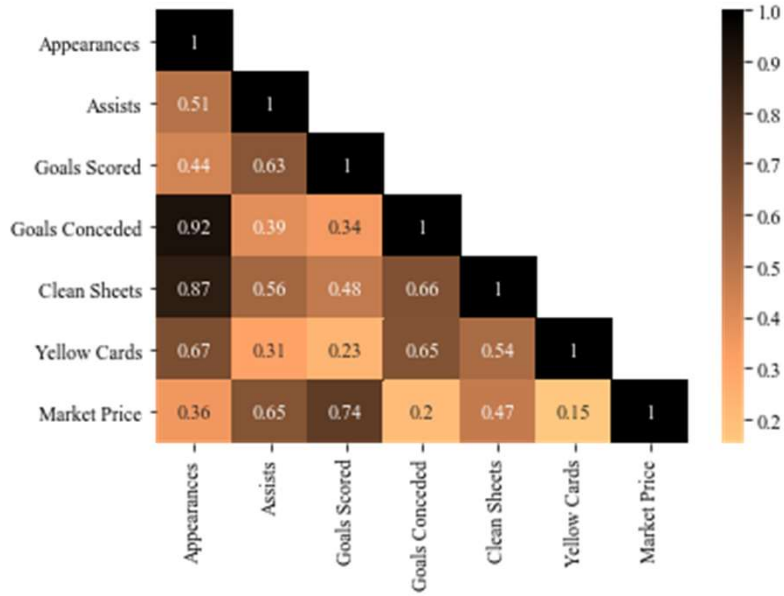


Does Money Impact League Position?

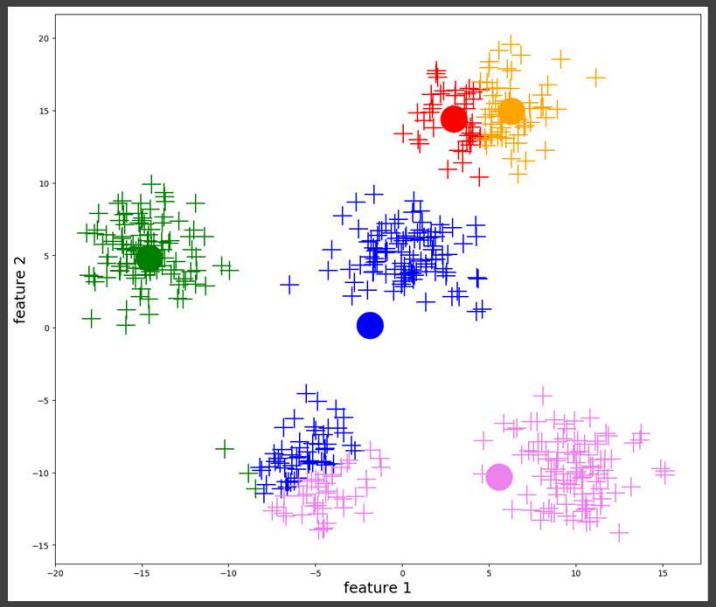
- Yes
  - Ridge Regression
    - As wage capital expenses increase, a club, on average, performs better
    - From 2016-2017 to 2020-2021 season
    - Each color represents a team throughout the time interval

What Features Determine a Player Tier?

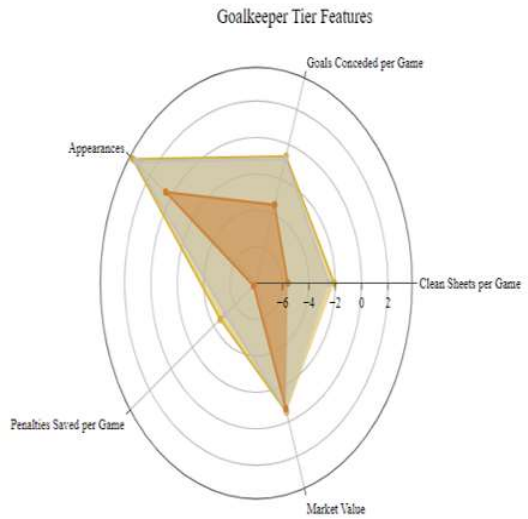
- Market price correlated with assists & goals scored
- Goals conceded, clean sheets, & yellow cards are correlated with appearances

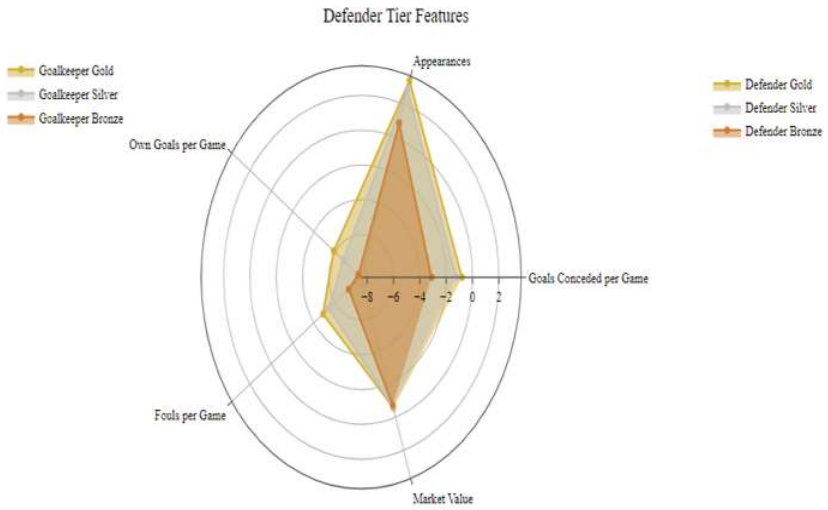


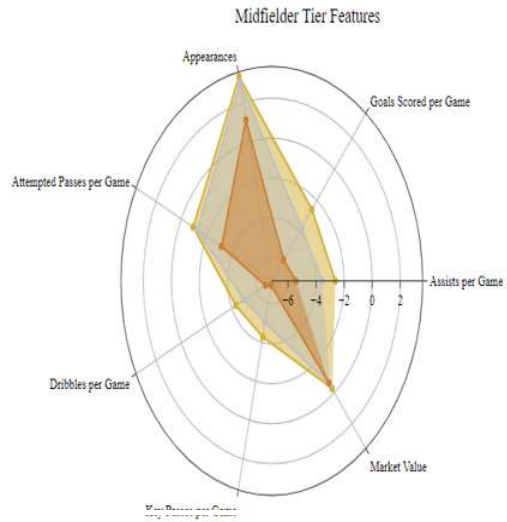
How to Identify Player Type?

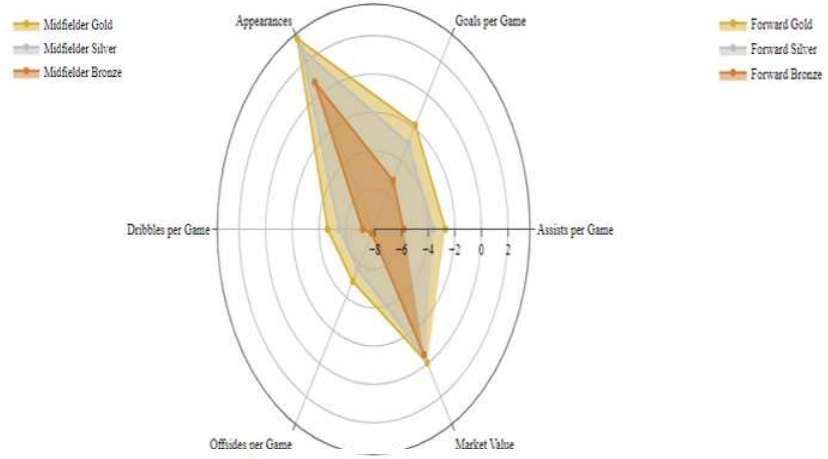


- *K*-means++ Clustering
  - *Centroid Based Clustering*
  - *Initialize by choosing a random centroid, then identify other centroids by maximizing the distance between clusters*
  - *Iteratively calculate the mean of the cluster, then adjust the centroid until the mean is constant across iterations*
  - *Higher quality results: initialization dispersed across the domain relative to a randomized initialization of a vanilla k-means algorithm*
- *Features:*
  - *Appearances*
  - *Assists*
  - *Goals scored*
  - *Goals conceded*
  - *Clean sheets*
  - *Yellow cards*
  - *Market Value*
- *Three Clusters*
  - *Gold, Silver, & Bronze tiers*







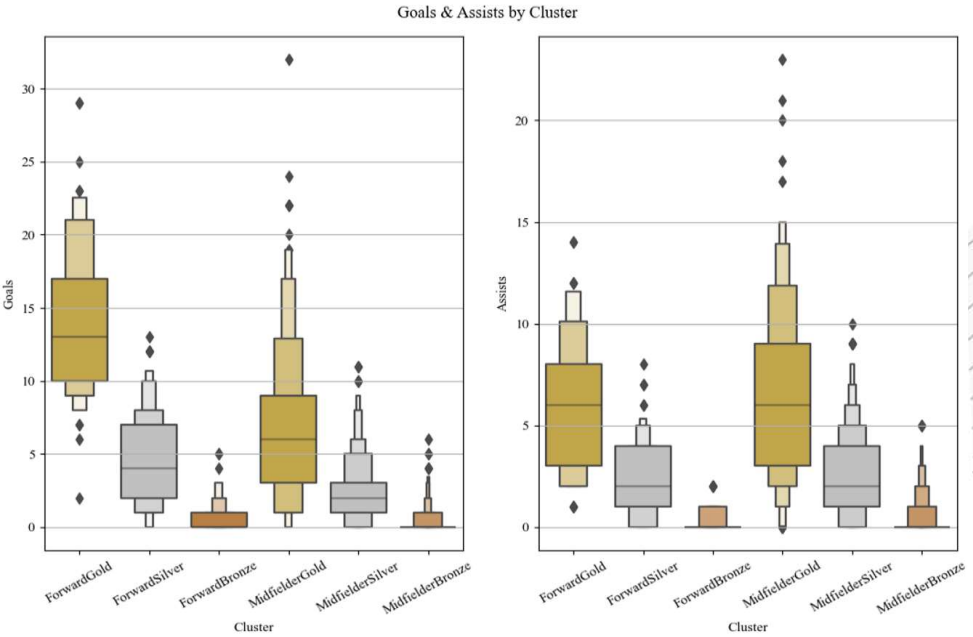


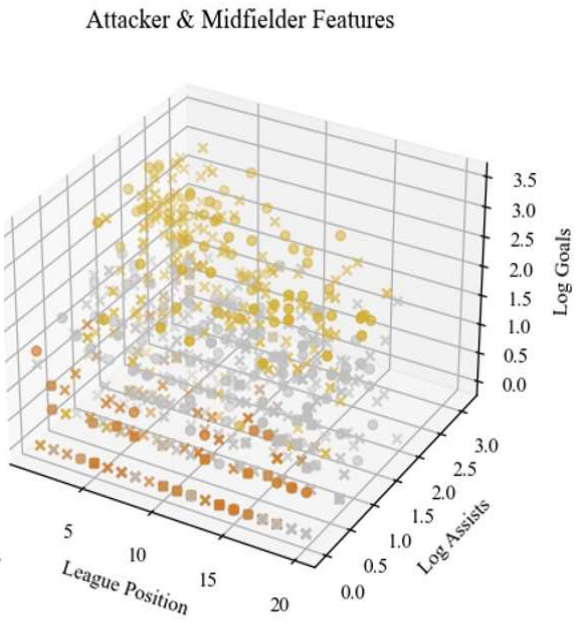
How to Correct for Season & Position?

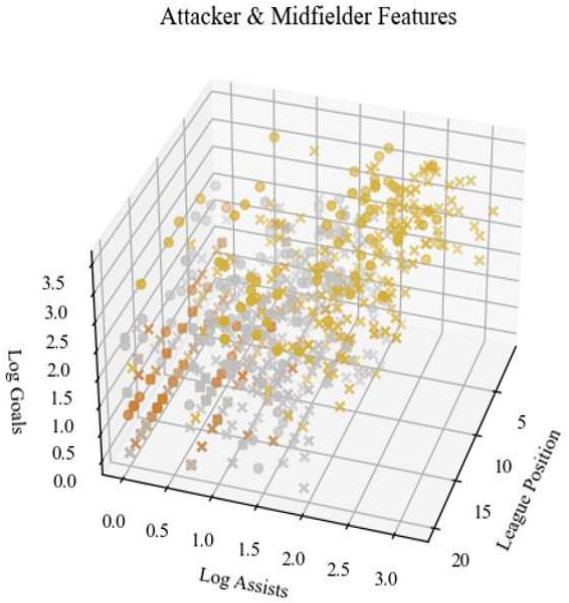
- *Partition Data Set by Season*
  - *Game Progression*
  - *Off Season Transfers*
  - *League Bolstering*
- *Partition Data Set by Position*

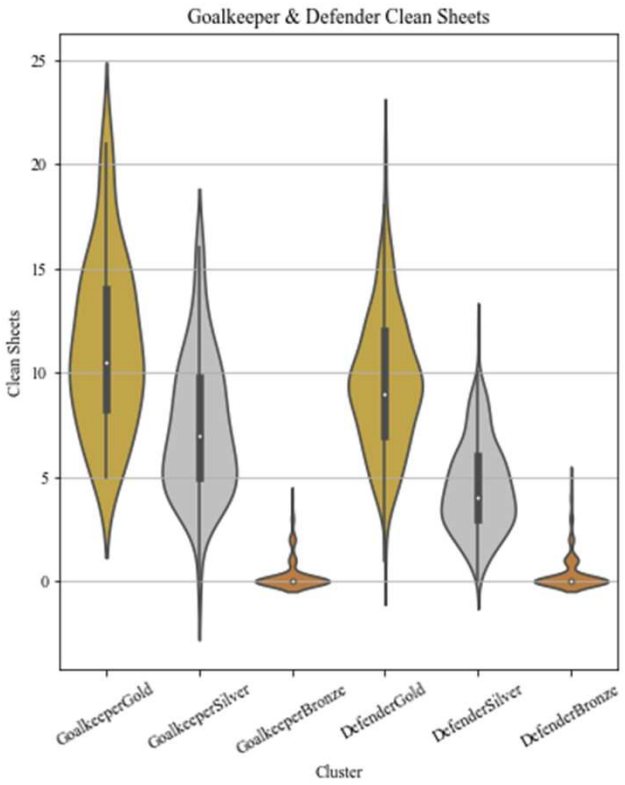
Are Forwards & Midfielders Similar?

- *x* represents midfielders & *o* exhibits forwards
- *Key Findings:*
  - *Upper tier Gold Midfielder contribute the same expected goal count as a Gold Forward*
  - *Forwards rarely contribute the number of assists midfielders provide*
  - *All teams possess Bronze players regardless of league position*









Are Goalkeepers & Defenders Similar?

- Gold Goalkeepers outperform Gold Defenders
- Gold Goalkeepers are scarcer than Gold Defenders
- Silver Goalkeepers outperform Silver Defenders
- Bronze Goalkeepers & Defenders contribute roughly congruent clean sheets

What Separates Top-Flight & Relegation Class Clubs?

- Team Composition & Tactics
- Frequent itemset analysis (Apriori) to identify frequent player combinations amongst Top-Flight clubs & Relegation Class Clubs
- Project data set into two sets: The top 25% of teams & bottom 25% of teams
- For each partition, consider each game week as a transaction
  - {GoalkeeperGold1, DefenderBronze1, DefenderSilver2, DefenderGold1, MidfielderGold4, MidfielderSilver3, ForwardGold2}

Top-Flight Clubs

- {DefenderGold2, MidfielderGold1}
- {DefenderGold1, MidfielderSilver1, MidfielderGold1}
- {DefenderGold1, MidfielderSilver1, MidfielderGold2}
- {MidfielderGold1, MidfielderSilver2}

Relegation Class Clubs

- {DefenderSilver1}
- {DefenderGold1, MidfielderSilver2}

How Do Top-Flight & Relegation Clubs Play?

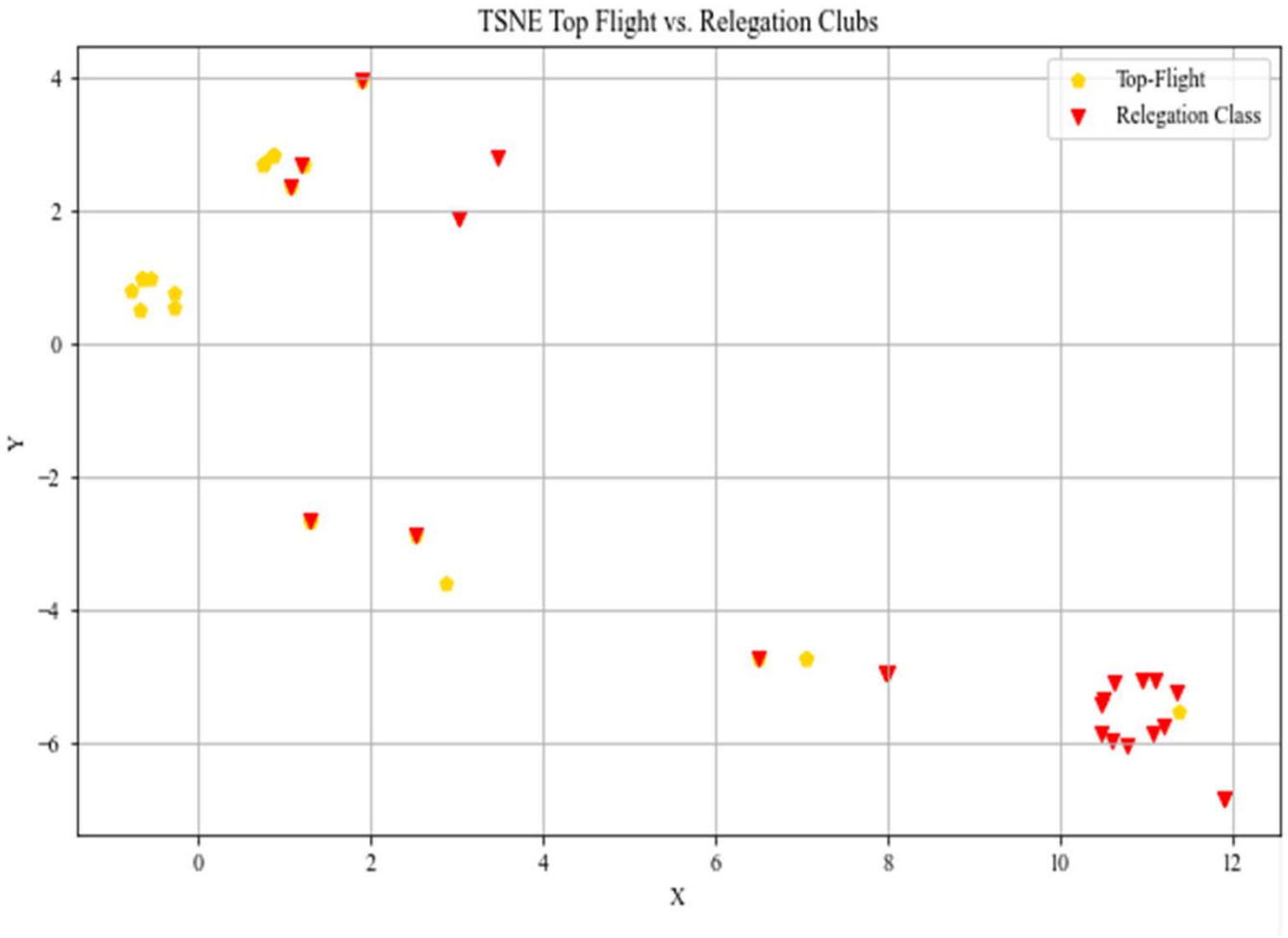
Top-Flight Clubs

- Offense transitions
- Defense transitions
- Establish offensive presence
- Heavy emphasis on possession
- Gold tier wingback defenders to maintain possession & deliver crosses
- Gold tier midfielders to serve a strong, resilient forward

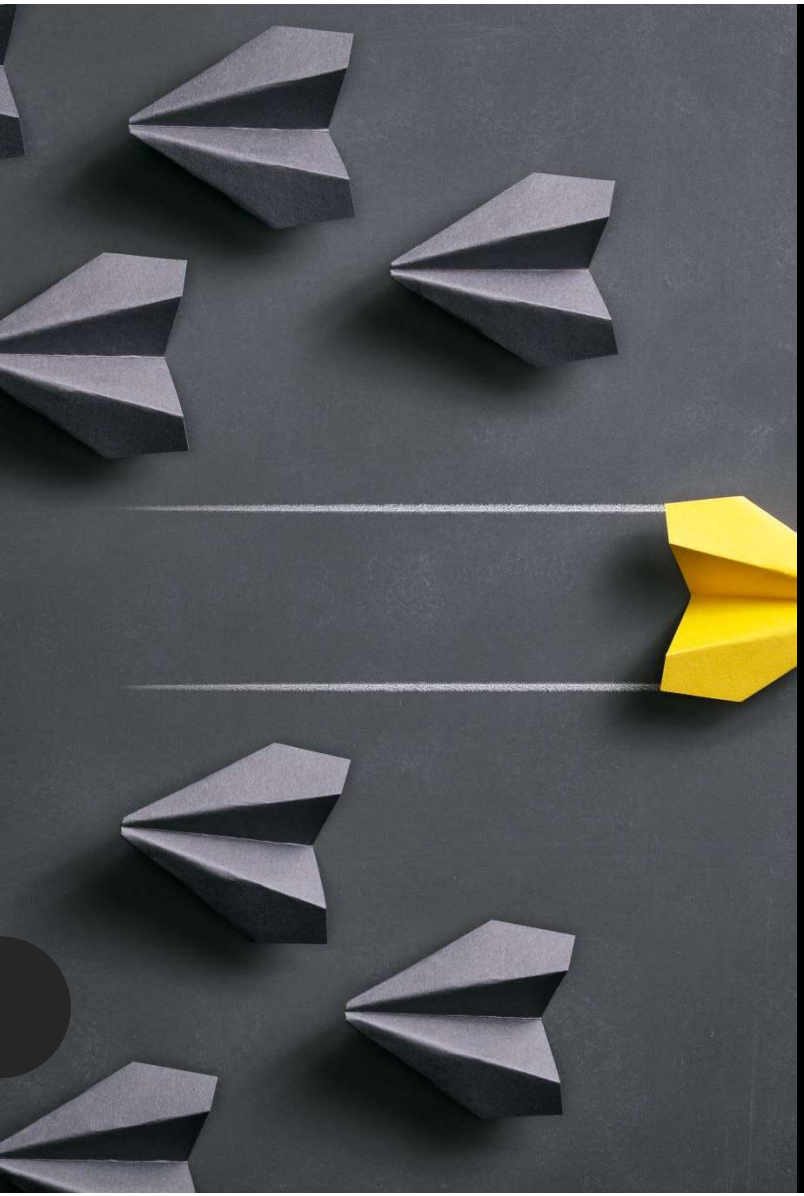
Relegation Class Clubs

- Establish defensive presence
- Heavy emphasis on gold tier center back
- Keep shots away from the middle
- Attempt to score on counter attacks & mistakes

How Similar are Top-Flight & Relegation Class Clubs?



- One-Hot Encoding
  - An array determining the frequent itemsets present in the club
- *T*-SNE
  - *t*-Distributed Stochastic Neighborhood Embedding
  - Non-Linear Dimensionality Reduction
  - Preserve Local Structure
  - Handles Outliers
- Results
  - Two distinct groups containing Top-Flight and Relegation Class clubs
  - Sporadic relegation class clusters
  - Sometimes Relegation Class clubs contain the correct player type combinations, but cannot get it done



What are the Key Points?

1. Invest in a strong gold tiered midfield accompanied by at least one strong wing back
2. Tactically focus on transitional moments relative to establishing defense
3. Above average Gold Midfielders score the expected value of Gold Forwards & provide more assists
4. Gold Midfielders reap the highest return on capital investment
5. Ensure tactics conform to the player combinations to generate fortunate results

What's Next?

- Similar analysis with La Liga
- Various formations with frequent itemsets
- Possession & passing decision during attack buildup

QUESTIONS?

References

[1] 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.

[2] 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.

[3] 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.

[4] 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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.