Calvin Yu, Daniel Lisko, Toan Vang, Ruchi Bhavsar

CS6220 Draft of the Final Report

a. Abstract - This is a brief summary not exceeding 500 - 750 words describing the problem, the solution method, and the results.

- Advanced statistics in sports drive the selection of the best players, leading to wins, fan culture, and overall economic growth for cities
- This is often easy for already successful and “rich” teams given their power and resources, but what if we want a less prosperous/successful team with a limited salary cap to find the same success
    - This increases competition, overall viewership of the sport, and success to the city
- Selection based on numbers (introduce case study of the oakland a’s) gives a formulaic way to outperform a scout’s intuition of a good player
- How do we know that certain advanced statistics lead to recruiting better players
    - We propose a unsupervised hierarchical clustering model to determine if certain players are grouped together according to certain statistics
- If we show that our intuition that better players have better advanced statistics, how can we form a team given limited resources?
    - We propose a supervised regression model that

b. Introduction – Approximately one to two pages long and contains

(i) Statement of the problem you are trying to solve

- Determine if there is a pattern in advanced statistics for successful players
- Form a team using advanced statistics given the limited salary cap a team may have

(ii) Why is it important to solve this problem?

- Better players means more success for a team, leading to wins, fan culture, and overall economic growth for cities
- Improving the quality of teams with limited salary caps improves competitions and leads to higher overall viewership of the sport and ultimately the success of a city

(iii) Background information and a bit of literature survey to present what’s already known about this problem.

c. Methodology - What's your solution, how do you propose to solve the identified problem?

- Extract player advanced statistics and historical team data
- Use a hierarchical clustering method to determine if certain types of players will be grouped based on advanced stats
  - Daniel can add how he evaluated the performance
- Use a regression method

d. Code - Brief explanation of the code.

- Cluster model
  - For this analysis, we will collect player stats from the 2010 - 2022 season and then average those stats for each player. For the clustering model, we will use advanced players' stats since many of these advanced statistics consider a player's contribution to a team's success. Since there are 20 features to consider, we first perform a principal components analysis (PCA) and select the components with the highest explained percentage. Before performing the PCA, we will also perform standard scaling so that all the features have the same weight when constructing the PCA.
  - Since we want to group players with similar feature characteristics, we will perform hierarchical clustering to group players into different tier rankings (i.e., tier 1 being the best players and tier 5 bench players).
  - We will visually inspect the hierarchical tree to evaluate cluster separation. Additionally, we will look to see if each group has differences in advanced stats. For example, we expect a higher-ranking group to produce more all-star players with similar skill sets. Finally, we will look at the top 20 players, based on player efficiency ratings, to see if the grouping of players makes sense since top-tier players should have the most significant impact on a game.

- Regression model
  - After acquiring the data, we will generate training data. This data will consist of teams of players and their averaged statistics as the independent variables and that team's number of wins in the 2021-22 season. We will reduce the size of teams to say, 5, to allow for different combinations of players within the same team and an increase in our training data size.
  - We will then generate a test set by picking random players and averaging their statistics. This set will be different from a typical test set in a machine learning setting since it is not derived from the train set. We do this to make random teams that we have not seen before and do not have win labels for.
  - Lastly, we will train and apply a regression model to produce the number of wins each random team has and select the highest output as our "best" team. To evaluate performance, we can first get a test set from our training data and then

output the mean squared error. We can also look at metrics that are specific to our problem such as how many players in that team are from top 10 teams in the 2021-22 season and how many are in the bottom 10.

e. Results – here you just state the results as you see them. E.g., In a fictitious study you were studying data collected over a period of 50 years of height of children aged 12 years. You can state the trends shown in the data, like in all countries the average height of 12-year-olds seem to be increasing by 3 mm every decade, except in countries C and D. In these two countries it seemed to decrease by 2 mm in the 1980s.

f. Discussion - interpretation of the results - what do the numbers really mean, do they make sense. E.g., In the fictitious study we used to illustrate the results section, Does it make sense that in countries C and D the average height decreased. Probably, there was a famine in Country C and Country D had an epidemic that affected the health of children in the 1980s.

g. Future Work - Is this study conclusive or does it lead to some future work? h. Conclusion - what conclusions can you draw.

i. References – Research papers, articles, and Internet resources referred to in the rest of the report.

j. Annex-A – This Annex has information about contribution of each individual in the project group.