Team Members: Daniel Lisko, Calvin Yu, Ruchi Bhavsar, and Toan Vang

Project Overview and Plans

The goal of all sports franchises, no matter the sport, is to win a championship in that respective sport. Looking into the NBA, teams can acquire players to create the perfect team that can help them achieve that goal, either by acquiring free agents or through the NBA draft. However, franchises have to take into consideration a roster limit of 15 players as well as a salary cap.

Therefore, we plan on generating a team of players whose combined season salary fits within a range. Keeping this salary within range allows a team that is limited financially to create a potential playoff team. As a team performs better, more revenue is generated, leading to increases in salary caps and overall growth in a franchise.

The following assumptions are made to formulate the problem. We assume that a team is deemed good when it has a high number of wins in the regular season, increasing its chance of making the playoffs and hopefully winning a championship. Another primary assumption is that players who meet a minimum number of games and minutes played contribute to a team's total number of wins. Lastly, a player's advanced statistics contribute mostly to wins and we can average these statistics across an entire team to get a feature vector.

Project Approach

We will approach this problem in five steps. The first is data aggregation and preprocessing. Because we wanted more freedom in the feature selection process, we will resort to downloading individual tables from https://www.basketball-reference.com/ and combining them accordingly. The advanced statistics, player salary for the 2021-22 season, and the number of wins for each team in the 2021-22 season are among the required data. We will preprocess the data by removing players who do not meet a minimum number of games and minimum number of minutes in addition to some unnecessary features.

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.

Furthermore, we can perform additional experiments such as predicting if a player is an all-star or what a player’s salary should be using classification models since our data would easily allow for these tasks. For these classification models, we can use ROC curves, precision, recall, and f-score to easily evaluate performance.

Deliverables

Our final deliverable will be a python notebook file that contains the results of our experiments. After running the notebook, the top 10 teams generated by the model will be displayed along with their averaged statistics and salaries. Results such as how many players of a particular team were in top 10 or bottom 10 teams will also be shown. Additionally, another model will generate the probabilities of a draft prospect making the All-Star team at least once in their career. Results for this will be shown as the predicted vs actual all-star results for each year to show the accuracy of the model.

Brainstorming

Data

- Adding data (Toan)
- **Multiple years of stats**
- **Standings for those years**
- **Summary statistics**
- **Feature selection (Ruchi)**
  - **Correlation of features**
  - **Contribution of features**

Evaluation and Analysis

- **MSE on regression outputs**
- Domain based
  - Players in top 10 teams, bottom 10 teams

Additional experiments

- Predicting all-star players, salaries
- Incorporating positions
- Explore different classification and regression models