Devin Wasilefsky and Christian Ingersoll

[Dwaz23@mit.edu](mailto:Dwaz23@mit.edu), [ingy23@mit.edu](mailto:ingy23@mit.edu)

Machine Learning Project Proposal

**Problem Summary:**

The problem that we are trying to address can be broken up into three separate parts. First, given statistics about a certain player in the NBA, we wish to accurately predict their salary. Secondly, we want to model a team’s success throughout the season based on the statistics of the team and its players; this will give us an idea of what aspects of the game most contribute to a winning season. Finally, using our salary predictions, we wish to create an optimization algorithm which can create the optimal team, maximizing the team’s performance, given budget constraints. This problem is not only applicable to current teams and their players, but also for future players that could join a team. For example, in the NBA draft, teams tend to try to bid for the best players overall; however, it may be more beneficial to get a lower ranked player who specializes in one area. In order to create the optimal team, and choose the best players within a given budget, it is crucial that salary values can be estimated accurately given player statistics.

**Datasets:**

The main data set that we plan on using is from Kaggle (<https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season/data>). This dataset is very extensive and has over 45 different statistics for NBA players. If this data is not enough, we plan to web scrape from various websites including https://data.world/datadavis/nba-salaries.

**Methods Used:**

From our class, we have learned many different techniques which combine machine learning and optimization. For the first part of our problem, predicting salaries, there are a few of these techniques that may perform the best. The first of these is holistic regression, then tree-based methods such as these is OCT, Random Forest and XGBoost, and finally we can use ensemble techniques to get the best possible predicted value.

For the second part of our problem, we will be creating a similar type of model to the first part of our problem, except rather than predicting player salaries we will be predicting team performance. We will investigate using statistics for the team as well as statistics for its individual players to isolate which aspects of the game have the highest contribution to a winning season.

For the third part of our problem, we plan to use MIO, along with our predicted salary values, to create an optimal team given a budget. Our objective function will be based on our model of team performance.

**Challenges:**

There are a few challenges that we may face during this project. One of the main challenges is that there may not be a strong correlation between the predictors in the dataset and their salaries. To solve this, we may have to use feature generation or web scrape for more data. Another challenge is that we are relying on the second part of our problem to create the MIO program in part 3. If the statistics that we deem important to winning a game are not accurate, our model will not be effective.