AML Project Proposal
Group 15

Background:
The market size of the fantasy sports services sector in the United States reached 8.88 billion U.S. dollars in 2021[1]. A major chunk of it is held by fantasy soccer leagues with over 9 million players and more[2]. In fantasy soccer, you select exactly 15 players to start with at the beginning of the season to maximize the total number of fantasy points accumulated by your team at the end of the season. Team formation is restricted to the following constraints:

- Budget - no more than $100 million fantasy money can be spent on selecting players.
- Team - no more than 3 players can be selected from one particular soccer team.
- Position - limitations on the number of players that can be selected per position.

Datasets and Objectives:
The dataset (https://github.com/vaastav/Fantasy-Premier-League) is a scraped dataset of the fantasy statistics of each player for each year/season stretching back to 2016-2017.

- Overview of the data
    - Complete data is available for the past 6 years and partial data for the ongoing season
    - For each year, there are approximately 600-700 players whose data are monitored
    - For each player, there are nearly 15 numerical and categorical features that are tracked
- The **features** include statistics like goals scored, assists, minutes, red/yellow cards, threat index, etc. An important distinction is that fantasy league statistics are measured slightly differently from real-world statistics (what counts as an *assist* in fantasy soccer might not be considered an *assist* in reality). For the purpose of this project, we will be considering only fantasy soccer statistics.
- The **target** is the total number of fantasy points a player accumulates at the end of a given season (a continuous integer value).
- Our **objective** is to predict the end-of-season fantasy points each player will bring in based on their statistics from the prior seasons.

Machine Learning Techniques:

- Regression analysis
    - Train a model to predict how many fantasy points a player will accumulate at the end of the season based on their statistics coming into the season.
    - Potential models include ensemble techniques like Random Forests, Gradient Boosting, etc. Also multivariate linear regression with different loss functions and regularization parameters.
    - Once the models have been trained, tuned, and evaluated, make predictions using the best model.
- Constrained optimization for team formation (for application purposes)
    - Combine the predictive model with a deterministic algorithm to build the squad adhering to the constraints of squad selection while maximizing the expected number of total points.
- Extensions:
    - If time allows we can experiment with other regression techniques learned later on in the course, for instance, neural networks, and see if those outperform our models or not.

[1] https://www.statista.com/statistics/1175890/fantasy-sports-service-industry-market-size-us/
[2] https://bit.ly/fantasysportsmarket

Shankh Suri (ss6625)
Eklavya Jain (ej2487)
Marvin Limpijankit (ml4431)
Qifan Jiang (qj2172)