How to Win Your League

Kaelynn Lackey, Harsh Khot, Dengfeng Jiang, and Zachary Kaufman

Project Overview:

The goal of this project is to figure out which players are undervalued, correctly valued, and overvalued relative to their average draft position (ADP) for fantasy sports. Fantasy football has a massive following, with the two largest fantasy football platforms being ESPN (12 million users) and Yahoo (10 million users). Very few of these users are functioning with more than a gut feeling, some advice from podcasts or articles, and the ADP given by their respective platform. To give our user an advantage we will use factors from at least three main categories. These will be:

- Base stats: age, raw stats (ex: rushing yards in football), etc. with a player's position in mind (running backs have rushing yards weighted more heavily than passing yards)

- ADP to result: We can take previous ADPs and see how the player turned out, allowing us to create a baseline expectation (A player with a calf injury performed below ADP, while a player coming off an injured riddled previous season outperforms ADP)

- News rankings: Reporters are constantly writing articles, tweeting information, and giving analysis online. These rankings will be collected and ranked on a scale of positive to negative. (A player with a calf injury that forces him to sit out of practice will be given a more negative score than a player with an injury that they practice through) This will be restricted to larger sites like ESPN, The Ringer, The Athletic, etc.

Proposed Method:

This project aims to have a holistic look at various metrics defining the players' performance in the past seasons and calculating a cumulative score. Based on this score we intend to classify the players into various deciles or lists which could help fantasy football users pick and drop players in their respective teams.

The dataset creation would be more of an intuitive process, where, based on our understanding of the game we would create certain features which would be relevant in calculating the aforementioned performance score. Example: age, height, weight, etc. The data pertaining to these features will be web-scraped or manually fetched from reliable online sources.

The two novelties in our approach are, first, the usage of "ensemble learning." Most of the pre-existing literature revolving around this application involves the usage of some kind of classical ML technology that lacks a loss function and/or regularization term. These are essential in punishing the model based on the deviation occurring between the actual and the predicted values. To do this, we intend to use **the XGBoost regression learning algorithm**. This algorithm employs a linear loss function to perform the mentioned "correction mechanism". It also uses several regularization functions which "punish" the model if it becomes overcomplicated (over-fits)

Second, the usage of NLP to generate some kind of numeric score based on analyzing several news headlines, tweets, or textual information about the players in the last couple of months. We will understand the correlation between this "News-score" and the score generated by our regression model mentioned earlier and involve this as a feature variable in our model. This is more of an additional feature in our project and we will start working on it once the skeletal regression model is fully built.

Related Work:

Machine learning has been applied to various fields and has recently gained traction in Fantasy Football analytics. Predicting scores based on historical performance metrics and player statistics has emerged as an interesting area of exploration.

In [Bangdiwald et al., 2022], the authors test AI algorithms such as Linear Regression, Decision Tree, and Random Forest to evaluate which algorithm performs the best for predicting Fantasy Premier League points. Their models incorporate factors like fixture difficulty, player form, and opposition difficulty, with results demonstrating that the Linear Regression model performs the best overall in terms of predictive accuracy.

In [Baughman et al., 2021], the authors describe a machine learning pipeline that integrates deep learning and natural language processing to classify Fantasy Football teams based on their likelihood of success or failure. They processed data from over fifty thousand sources, including news articles, podcasts, and videos to provide player classifications with up to a sixty-seven percent accuracy.

In [Yang et al., 2021], the authors conducted research on sports performance predictions using BP neural networks. Their study explores the application of neural networks in capturing relationships between various athlete performance metrics. The high accuracy of their model demonstrates the potential of neural networks in sports analytics, suggesting that similar techniques could be effectively applied to other areas such as Fantasy Football.

Timeline:

- Base Dataset processing - Data collection and Data engineering, two weeks by end of Oct 11th

- Model - preTraining - Base data simulation/feature engineering and XG-boost regression, two weeks by end of Oct 25th

- Model - postTraining - News dynamic tuning with NLP tech, one week by end of Nov 1st

- Interface implementation and End to End pilot run, two weeks Nov 15th

Sep 28th - Nov 20th , aiming at delivering the go-live version before Nov 20th.

References

[Bangdiwala et al., 2022] Bangdiwala, M., Choudhari, R., Hegde, A., and Salunke, A, "Using ML Models to Predict Points in Fantasy Premier League," in *2022 2nd Asian Conference on Innovation in Technology,* Pune, India, pages 1-6.

[Baughman et al., 2021] A. Baughman, M. Forester, J. Powell, E. Morales, S. McPartlin, and D. Bohm, "Deep Artificial Intelligence for Fantasy Football Language Understanding," arXiv, 2021.

[Yang et al., 2021] S. Yang, L. Luo, and B. Tan, "Research on Sports Performance Prediction Based on BP Neural Network," *Mobile Information Systems*, 8 pages, 2021.