Milestone II Team Project Proposal

Project Title: Rugby Match Prediction and Player Role Analysis

Adam Bakopolus, [Carolann Decasiano](mailto:cicasian@umich.edu)

**Introduction**

In the modern game of rugby, teams at every level from grassroots to international fixtures have redefined how the game is played. The game is faster, stronger and is a thinking person’s sport not relying solely on play calls but the skills to take advantage of opportunities. With the upcoming 2023 World Cup to be held in France this September, fans eagerly review the [world rankings](https://www.world.rugby/tournaments/rankings/mru) to glean any insights they have for how the event will go. The goal of this project is to apply data science to available rugby datasets to predict match results (supervised) and analyze player roles/skills (unsupervised).

Link to existing study similar to proposed approach: <https://www.sciencedirect.com/science/article/pii/S2210832717301485#b0040>

This article provides a framework for machine learning-based team and individual sport performance and score prediction. While this article did acknowledge some prior work around rugby score prediction, this study didn’t leverage any rugby-specific features and instead tried to globally predict Rugby Union, English Premier League, Australian Football, etc. With the goal of our project centered specifically around Rugby Union, this area of focus appears fairly novel following the literature review and will attempt to predict scores based on relevant rugby-specific features identified following exploratory data analysis.

**Supervised Learning**

The proposed dataset will be pulled from espnscrum.com and provide the necessary historical rugby game data needed to train supervised models to predict scores for a test set of games. To create this dataset, we will rely on an already developed scraper with publicly available code through github. While further work is needed around exploratory data analysis for appropriate feature selection, information around home team vs. away team, world ranking, points per game, etc. will likely form the necessary input needed to make a reliable prediction. Literature review thus far around sport score prediction all points to leveraging Artificial Neural Networks (ANNs) and, therefore, this will be the first learning approach taken. Additionally, we will also explore the effectiveness of Random Forest Regression and Support Vector Regression. A Random Forest was chosen over just a single Decision Tree to improve the generalization of the model, but the same benefits (easy visualization, non-required data pre-processing) will still be present. Lastly, Support Vector Regression is a suitable choice due to the ability to handle high dimensional data. This will give flexibility in this supervised learning task if the number of features does significantly increase. In terms of evaluation, the three supervised models will be trained on the historical rugby dataset. The models will predict a score for Team A and a score for Team B and a winning team will then be selected. The accuracy and performance of the models can then be assessed through 5-fold cross-validation. For visualizations, all three of these models could be plot in some manner, with SVR and Random Forests being traditional plots that would help present decision boundaries and/or the linear models leveraged for the prediction, whereas the neural network plot may instead show the weights tied to each of the neurons or the general structure of the network.

**Unsupervised Learning**

The proposed dataset for our unsupervised learning task will be scraped from https://www.sixnationsrugby.com/ containing player stats for six nation teams by match. The dataset contains 32 features. The goal of this task is to cluster players to analyze playing style and role on a roster. EDA will help us decide if feature reduction is necessary and whether any correlated features (e.g. tackles and missed tackles, position and scrum penalties) need to be explored. There is literature on applying k-means clustering to determine football playing style of teams[[1]](#footnote-0). We’ll look at applying k-means clustering and Principal Component Analysis for a more robust model on the playing style of individual players. For PCA, we’ll need to normalize the dataset in our data preprocessing. To get insight into the nature of the principal components we’ll create a heatmap to visualize the groupings in each component. We’ll want a scree plot of the eigenvalues from our PCA to choose to go with an elbow or kaiser rule when selecting the top components. For the purpose of evaluating our clustering, we’ll use silhouette scores and plots to determine the best k.

**Team Planning**

Adam will lead the Supervised Learning work. Carol will lead the Unsupervised Learning work. Both will collaborate around best practices for each of these components and will share an equal role in writing the proposal and final report. Following conclusion of the Comprehensive Exam, work will begin on scraping the ESPN and six nations rugby site to generate the raw dataset. Data cleaning and feature selection will begin and conclude by weeks 4 and 5 of the course.

1. P. Verma, B. Sudharsan, B. R. Chakravarthi, C. O'Riordan and S. Hill, "Unsupervised Method to Analyze Playing Styles of EPL Teams using Ball Possession-position Data," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 58-64, doi: 10.1109/ICACCS48705.2020.9074426. [↑](#footnote-ref-0)