Milestone II Team Project Proposal

Project Title: Rugby Match Prediction and Player Analysis

Adam Bakopolus, Carolann Decasiano

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 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 was already available, this study did not leverage any rugby-specific features and instead tried to globally predict scores across multiple leagues (e.g. Rugby Union, English Premier League, Australian Football). With the goal of our project centered specifically around Rugby Union, this specific area of focus appears novel following the literature review.

Supervised Learning

International Rugby Union match data, following ESPN's discontinuation of their espnscrum domain, has become difficult to track down. Fortunately, match data was available through github for matches between 2004 and 2022 for International Men's Rugby Union. Data cleaning, preparation, and transformation is ongoing but, most notably, the current dataset includes the country where the match was played, the two countries involved, the final score of the match, whether one of the teams had homefield/country advantage, and a three-match moving average of points for and points allowed. Additionally, to better inform the skill of the two teams, the points-based world ranking for both teams, available via world.rugby, was pulled into this dataset. These datasets and features provide the necessary historical International Men's Rugby game data needed to train supervised models to predict match scores for both competing teams in the match. 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 explore the effectiveness of Random Forest Regression and Support Vector Regression. A Random Forest was chosen over 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 relative to Linear Regression in case non-linear data relationships exist and for its greater generalizability via hyperparameter tuning. 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. In addition to selecting a winner, we will also evaluate the predicted point differential to the actual point differential of the match in determining the strongest model. The accuracy and performance of the models can be assessed through 5-fold cross-validation. In addition to testing on this created dataset, we will also attempt to use the “best” (as determined above) model to predict how the 2023 Rugby Union World Cup will progress, from group play to the final. Following completion of training and testing the models, further consideration will then center around how to best visualize the models. For example, a plot of the neural network may show the weights tied to each of the neurons or the general structure of the network, but concrete visualization determinations will be made as the project progresses.

Unsupervised Learning

The proposed dataset for our unsupervised learning task will be scraped from rugbypass.com containing player stats of qualifying world cup teams. The dataset contains 23 features about players position, weight, attacking and defensive 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 penalties) need to be explored. There is literature on applying k-means clustering to determine football playing style of teams[1]. 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. Will also use t-SNE and other methods to get a sense of global and local clustering.

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.

Timeline

Course Week	Task
4-5	Data Cleaning and Feature selection from scraped and gathered data
6	EDA and begin Supervised Models (Random Forest, Support Vector Regression) Visualizations and Unsupervised PCA for Dimensionality Reduction and Visualizations
7	Supervised Model evaluations and Unsupervised k-means clustering with evaluations.
8	Complete and Submit Milestone II Report

[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.