Your **project paper** should be written with a technical audience in mind. Here are the components you should cover:

* How you chose which features to use in your analysis
* Details of your modeling process, including how you selected your models and validated them
* Your challenges and successes
* Possible extensions or business applications of your project
* Conclusions and key learnings

Every sport attempts to rank teams in some fashion. In many, it’s a simple win-loss comparison mixed with some head-to-head deciders when there’s a draw. In soccer and rugby, rankings – or tables – are set based on a points system. For rugby, teams receive 4 points for a win, 2 points for a draw, 1 point for losing by 7 points or less, and 1 point for scoring 4 tries (the equivalent of an NFL touchdown) in a single game. But, how valuable are any of these ranking systems?

Ultimately, the value of the ranking is their ability to determine what will be the best game of the year: the championship. The NFL attempts to rank teams so that the best team in each conference makes it the Superbowl. Similarly, rugby leagues use the ranking systems to select the top four teams in the league. Those teams are then matched up against one another such that the top two teams should theoretically end up in the final. Thus, the best system for ranking teams is the one that is most accurate in predicting the outcome of future games. That is, the one where the highest rank team has the best odds of beating the teams beneath it, and so on down the table.

A simple win-loss ratio does not take into account many factors that could indicate a team’s odds of winning a future match. Rugby attempts to take certain factors into account: teams that score a lot of points (four tries or more) and teams that lose by only a small margin (7 points or less). By rewarding teams that earn these “bonus points,” and factoring those points into the ranking system, rugby is ultimately attempting to improve the accuracy of their tables. The question is, can we use simple indicators based on previous matches to predict the outcome of future games?

For this project, I am focusing on the sport of rugby, and specifically on the Aviva Premiership, which is England’s top league. It is made up of 12 teams, each of which plays each other team twice during the regular season. This accounts for 22 rounds of play, with 12 observations per round (6 games, two teams per game). This creates a total of 264 observations per season. I’ll be working with 5 seasons worth of data, or 1320 observations.

Just like previous ranking models, mine will be designed to predict the outcome of future games. I’ll be using data from previous games in a season to predict future games in the same season. When my model can accurately predict the outcome of future games significantly better than either a simple win-loss ratio comparison or the current points-system, I’ll be happy with my model.

I am collecting my data from ESPNscrum.com, ESPN’s domain for rugby-related news. Because all of the information is located in tables that I am unable to scrape, I collected full data for one season of the Premiership by hand, and partial data for the other four seasons has been collected in the same manner. ESPN keeps a relatively detailed record of rugby games going back a number of years. For the most recent games, there is a lot of interesting information, including statistics about percentage of possession, number of tackles made or missed, and the number of passes made by either team. While all of this information is very interesting, and could certainly be used for some highly detailed analysis, I’m relying on simpler information.

Most ranking systems start with the win-loss record of each team. I hypothesize that the best model will include that information as well. The four-try and lose-by-less-than-7 bonus points are designed in part to account for point differential. I hypothesize that the best model will be account for it more specifically by factoring in the average point differential at that point in the season when predicting the outcome of the upcoming game. Finally, I hypothesize that a good model should factor home field advantage into its ranking. That is, a win at home should count less toward the table than the same win away. Additional variables will be considered, but this is what I hypothesize will be enough to be confident in my model.

**Exploration**

Initially, I was thinking I would use many of the diverse variables included on ESPN’s website. However, I wanted to create simple model that relied on things that any fan could understand. I did not want, yards carried or number of scrums won to weigh heavily on the system. More like the current points, I wanted there to be a simple way to understand the rankings (rather than officials simply relying on “an algorithm”).

So, I explored all of the stats for the current season of the Premiership (14-15). I saw the strongest correlations between points scored: meters run per game, meters-per-run, and overall clean breaks. It would seem that activity with the ball (and not just possession) is most important.

After deciding to work with simpler variables, I did more exploration. I found that, generally, home teams have some level of advantage. One of my analysis suggested that home field advantage could account for about a 5.95 change in the point difference for that match (with the home team benefitting). After training on the first 19 rounds of the games, the model predicted the last 3 rounds right with 75% accuracy based on the “home” variable alone. However, this is a highly variable model, as the home team does not win 75% of the time overall (closer to 65% for the full 5 years).

**Models**

So far, I’m only working with logistic regression. I am hoping to work through a few more types of models and hopefully ensemble them. The aim was to refine my model by training on the n1 rounds in a season and then testing on the next round, and then training on n2 rounds and testing on the next, and so on. The idea was to give my model enough data to train and test on. At this stage, I’m not sure how to implement that type of strategy. More likely than not, I will simply train on the first 16 rounds of each season, and test that model on the following 6 rounds. However, it is possible that I could run the type of staggered predictions laid out above, and ensemble those models before running the model again on my final data set.

Currently, I’m still working through feature selection for models. I’m having a bit of trouble analyzing the value of average point differentials versus number of wins. My current logistic regression seems to predict equally as well when I switch them out for one another. This is slightly troubling, as my hypothesis indicates that the value of the win, and not just the win itself, should lend toward better predicting.

Extensions

My ultimate hope is to create a system that assigns an actual “strength” to each team. The strength should be used to rank the teams (strongest to weakest), and that rank can be used to predict the outcome between any two teams in an upcoming match (within a certain percentage of confidence, at least).

This is difficult to do, because the current models I’m working with do not understand the relationships between observations, only within the individual observation. That is, the model does not understand that it is giving me the prediction for Bath versus Northampton, only that for observation x, the predicted outcome is a 1 (win for Bath). Thus, it can’t necessarily tell me the “strength” of Bath versus Northampton, only predict which team will win at a certain point in the season.