Mitchell Gil

Final Report

Problem Statement

Soccer is a sport that is as equally reliant on individual player skill as well as the tactics, formations, and attributes of the team as a whole. Coaches are always trying to one up each other in order to improve their chances of winning their respective games.

We try to use classification models to best predict which team will won its respective match using average'd player attributes per team, as well as team attributes such as formations, how much was placed on attack, vs defence, team style of play, etc.

Using the model, we determine which attributes the winning team had which best contributed to their success, giving pointers for coaches on how best to train their teams.

Data Wrangling

Wrangling the data proved to be a challange. The data was provided in the form of multiple SQL tables that needed to be merged into a single dataframe for modeling. It had multiple NaN values, only provided player positions by their positional coordinates in the field, which needed to converted into usable player positions. Some of the players attributes were incorrectly reported, which was worked around by using the max values of the team and assigning that as the global value across the board per-position. Match events themselves were provided inside cells as XML data, and needed to be extracted and averaged or summed to the main dataframe as well.

Due to the large size of the data, and to include as many variables as possible to the models, the data was reduced to around 45% of its original size after discarding all the rows which contained multiple instances of NaN values across them.

Exploratory Data Analysis

All the numerical features with the execption of some attributes were normally distributed when plotted as a histogram. Home team won a little over half of the 12,000 games played, with draws around 3,500, and Away team winning around 4,000 of the games. There seems to be a home team advantage when playing. The goalkeeping attributes also included data from all non-goalkeeping players so the data has two normal distributions, one presumably with numbers of actual goalkeepers, and one without which skews it.

Pre-processing

All the SQL tables were merged into a single dataset and labeled corresponding to whether they describe the home, or the away team. A basic heatmap showing correlation among variables was also created. The features that were sourced from the same table seemed to have a siazable degree of correlation, but nothing significant.

The training and test data was split at 70-30 train/test size. The y variable was the outcome of the game:A win, draw, or loss for the home team, which could be interpreted vice versa for the away team.

Modeling

Models chosen were Random Forest, Logistic Regression, Ridge Classifier, Support Vector Machines, and KnearestNeighbors. The untuned models only had an accuracy score of around 45-50%. The tuned models all equally performed at 49-50% score. The models all heavily favored a home team win in a more lopsided manner than the actual number of home games won were.

Therefore all models were chosen and their important features extracted to be cross-vetted against each other, and see if any repeating features turn up across multiple different models and what story they tell.

Findings and Conclusion

The most common features that showed up across different models were short_passing, player vision, defence pressure, interceptions and buildup play passing. Interestingly, ball possession, which has been a standard metric all sports networks report when evaluating a team’s performance in a game, did not make any significant contributions to the models.

Although the models were not too accurate, the little they did tell seem to point towards a team’s passing ability to be crucial towards their success. Defence pressure and vision seem to paint a bigger picture of tight coordination being a major requirement for teams to win their games, and their individual players need to have the vision to be able to take advantage of the spaces and chances created to maintain that tight coordination.

There is also numerical evidence that teams playing in their home stadium does provide an advantage. This has always been common convention among even casual sports fans and is not specific to just soccer.