Sohum Kaji
May 18th, 2017

Capstone Proposal

Domain Background

Soccer is a sport that has been played around the world for thousands of years. The casual observer would assume that given all of the popularity, obsession, and fanaticism involved in the sport there would exist a large amount of data and a huge amount of numerical analyses. Unfortunately, they would be wrong. In The Era Of Advanced Stats, Soccer Still Lags Behind, writes Chadwick Matlin of 538. The key issue, as mentioned by him, is a lack of data sources. Private data services have existed since the 2000 but they have always been priced prohibitively. Until recently there was no place for the armchair data explorer to go.

Enter nowgoal and soccerstats - providing a massive amount of data across a huge number of leagues, all at the low price of $0.00! The new issue is, what is important? Just browsing over to nowgoal's English Premier League 2016-2017 stats you're presented with tabs upon tabs of links each leading to another page full of descriptive numbers (not to mention this is one of ten leagues in one particular country).

There is a consensus in the soccer analysis community that the best metric for predicting the outcome of a match is a variation on the theme of expected goals (which also has quite a few haters). Expected goals ignores all but a few features and provides a serviceable prediction metric for the outcome of matches. I do not plan to walk on this well worn path.

Problem Statement

The problem I am looking to solve appears to be ignored by the general academic community on this subject: using an analysis of historical data combined with live match data (up until halftime) to predict the final score of the match. The reason this type of analysis is so unique is that most analyses are based on historical data but do not consider live data. Analyses like expected goals provide a trend that a team generally follows over a long period of time, but is not expected to be accurate in particular instances. Overall, it is not very useful for the general fan who watches matches and cares about what will occur in a particular match.

To restate: the problem is using previous data from the season as well as what has already occurred in the first half of this match, determining the final state of the

game as represented by the score. I have not found any resources that have attempted to do this in this manner.

Datasets and Inputs

As discussed above, historical data and halftime data will be used. The historical data is vast. It will be cleaned and have feature selection applied to it. I listed two sources for this data already above: nowgoal and soccerstats. This project will focus on the 5 largest European leagues, the first divisions of England, Spain, Italy, Germany, France. The data that will be collected includes: every halftime and fulltime result for every first division match in the top 5 leagues, standard halftime statistics for each of those matches, the formations used in each match, the possession and passing data for each match, and the overall numbers for the season, the home and away points table for each team, the overall season data for each team.

The key to developing an analysis that works in the short term is to consider an important perspective that is generally overlooked in the art: the work and impact/actions of the coaches and managers involved in the game. In the soccer world this is called tactics. Each game has different tactics or strategy employed by the coach. Looking at patterns in the tactics used in different situations (as embodied by formations as well as labels we will create using the historical data) will allow us to better understand games as they occur. The first half of the game can also give insight into this - we are able to see the strategies used in the first half of the game. To me, this is where approaches that only look at the previous performances of the team fall short - each game is it's own unique problem. This data is also available through the aforementioned websites.

Solution Statement

The solution to this problem is a model that provides insight into live soccer games at halftime. The first step is to acquire and clean the data, we will use Scrapy or Selenium to scrape the data and then the Python library pandas to clean/store/manipulate the data.

From there we need to find relationships within the data. This will be done using PCA and clustering. PCA will be applied to the historical data and clustering will be applied to the halftime data. Using the new information found through clustering, the halftime data will be used to train a stochastic gradient descent regressor. From there the outputs from this will be input into a feed-forward neural network to determine our actual results.

The output can be described as pair of integers: the predicted (most likely) goals for the home team, the predicted (most likely) goals for the away team. The format will be a list containing the 2 integers, for example:

[2,2]

would be one possible output for a match, indicated a predicted 2-2 scoreline.

Benchmark Model

The benchmark for this model will be creating a final score based on the halftime score:
gA = goals scored by Team A in the first half
gB = goals scored by Team B in the first half

[2*gA, 2*gB]

So, if a match was 1-1 at halftime the benchmark model would predict:

[2, 2]

The above describes two outputs of: double the home team first half score and double the away team first half score. This brings about a very rudimentary but baseline prediction that: the same result is most likely for the second half as occurred in the first. To outperform this type of prediction, the proposed project would have to some insight the second half.

Evaluation Metric

Rather than choose an evaluation function like sum of squared errors, I decided to award 1 point for 1 correctly predicted score (home or away team), 5 points for the correct number of total goals, and 10 points for the exact score. This is to ensure the "best" model is evaluated as the one with the most correctly predicted scores rather than the lowest error (which in an edge case, might have very few correct predictions).

Sohum Kaji
May 18th, 2017

Project Design

This project will begin with collecting the required data. That will be by using either Scrapy and/or Selenium to efficiently read the data from the websites: nowgoal and soccerstats.

After cleaning and scaling the data the analysis will split into three portions. The first portion will help organize the data. Clustering will be used to determine labels for the teams which describe their style of play. Examples of this include: 'possession based', 'counter-attacking', '4-3-3' (formations), 'pressing style', 'goals conceded'. The reason clustering will work for these particular labels is that each of these is directly represented by the data. Each teams possession per match is given, the formations are given, 'pressing style' will be given by turn-overs forced, etc.

These labels in tandem with the match results table are then used to create the reciprocal labels. Meaning, if many of the wins/losses are against teams clustered into a certain group, we can label this team as having superiority/trouble with that characteristic. Labels include: 'has trouble with possession based', 'has trouble with counter-attacking', 'plays well vs possession based', etc. by looking at the results of matches already played by the team.

The second portion is to train a regressor. In order to do that feature selection must be performed. Feature selection will be used to find significant features from the halftime scores, home and away table data, and the standard halftime statistics. Subsequently the most significant features will be used to train a regressor (the labels will be the final score of the match).

The third portion is to combine all of this information into an input for a standard feed-forward neural network. Using the output of the other two portions, we will attempt to predict the final score of a particular match from the halftime point.

Comparing this versus our baseline model and the objective perfection of 100% accurate answers will yield how powerful this model is.

Upon, completing this step there is a trained model that takes halftime scores and halftime statistics to predict the final score. For future matches we can input the halftime data into the model and gain insight from the output.