ML – Capstone Project

Predicting the Results of Soccer Matches

1. Definition

Project Overview

Sports Analytics in the past 10-15 years has increasingly become a part of every sport as teams begin to make analytical, data-driven decisions rather than a conventional or instinctual feeling that coincides with traditional beliefs. As a result, statistics about games have become increasingly available to the average fan. With these statistics, we’re looking to use machine learning to help predict the results of a soccer match.

Problem Statement

Predicting soccer matches is unique compared to other sports because soccer can have one out of three results, win, lose, or draw. The result of a draw happens very often in the sport where as with other sports if a draw is possible it happens very rarely. In the other top 4 sports in the US NBA and NHL games cannot end in a tie. There have only been 3 ties in the [NFL](https://en.wikipedia.org/wiki/List_of_NFL_tied_games) since the 2008 season and in the [MLB](https://en.wikipedia.org/wiki/Tie_(draw)) ties only occur due to weather or other extremely rare cases. Having ties as an additional result increases the complexity of creating a predictive model for soccer matches. In doing research on the topic I found this [project](https://github.com/GoogleCloudPlatform/ipython-soccer-predictions/blob/master/predict/wc-final.ipynb) by Felipe Hoffa and Jordan Tigani of Google during the 2014 World Cup. They looked to predict the winner of each match in the tournament and in their initial run of the data they don’t train on results that end in a draw since ‘they have less signal’ so all of their matches end up with either of the two teams winning. Which on some levels invalidates lowers their accuracy percentage since they are assuming that the winner of the penalty kicks (deciding factor on who continues to the next match) is considered to be the winning result of the match when in actuality the match result is a draw.

Not only does soccer have an extra result that makes predicting matches difficult, it’s also a difficult sport to return statistical analysis on because it lack statistical history outside of standard stats and because of it’s non-stop, free flowing nature. Other sports such as baseball naturally has more stats to utilize since box scores have been published for decades now and these stats can break a game down to the pitch. Also, due to licensing terms of the data on [Sportradar](http://sportradar.us/) I was only able to pull a minimal amount of games and it’s data. There may be more data that they provide for a paid version but I’

Having a minimal amount of stats and having one more outcome to predict makes predicting soccer matches more difficult than other sports. A combination of approaches might have to be taken as we explore the data and begin to break the data down to what is needed.

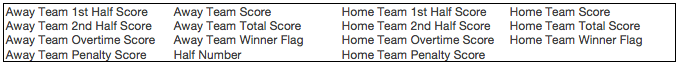
Metrics

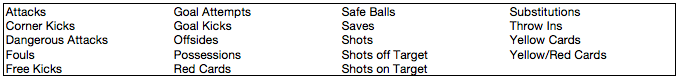
Teams can win, draw, or lose a soccer match meaning they can earn 3, 1, or 0 points respectively. This will be the target or label of the dataset. We are going to test a variety of models but initial assumption is that accuracy will need to be determined based off of a combination of a few models and not just one. One model might not be able to clearly predict wins/losses and ties (as the project above suggested) so we’ll need to identify and determine which matches might end in a draw and which matches will clearly have an outright winner.

1. Analysis

Data Exploration

Using data pulled in from [SportRadar’s](http://sportradar.us/) API I was able to pull Boxscore Information and Team Match Statistics in all previous MLS matches in the 2016 season.



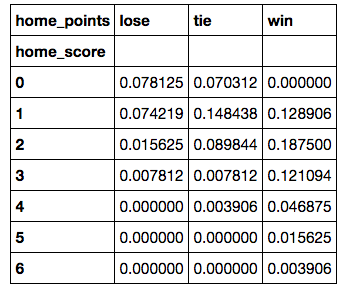
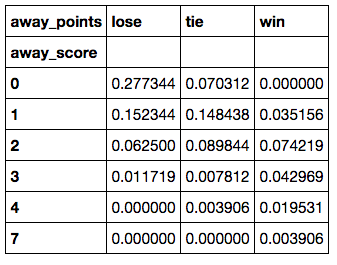


From this data, I’m able to pull all matches in the current MLS season. With this data, I’m able to predict upcoming matches based on all previous matches that the two teams have previously played. The data inputted into the database wasn’t filled 100% and had some missing data. But since averages are taken on the previous games of a team, we use numpy to ignore that ‘null’ space and not factor in that game for that feature. Also note that because our model uses previous game stats to determine the outcome of the current game, we need previous data so therefore the matches start in our data at week 2.

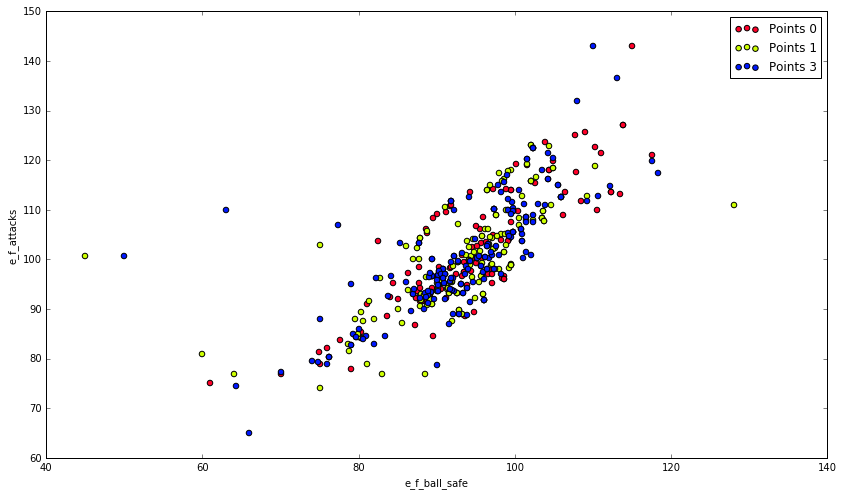
Because all matches are contained within one MLS season, all data should be relevant and ‘outliers’ should still be considered. There are some adjustments that could be made in the future to the data that could help balance out ‘blowout’ games. For instance, you could lessen the weight of ‘Goals Scored’ if the margin is greater than 2 goals. Since this means the other team likely isn’t trying as hard the goal isn’t as significant as a goal to put a team ahead.

With this data from the API, I looked to modify and enhance the given statistics into relevant features that can help predict the result of an upcoming soccer match. Essentially these features can be grouped into 3 categories (not including the Target and Non-Feature Columns): Standard/Cumulative Features, Home/Away Features, and Extended Features Columns.

* Standard/Cumulative Features – Features that include situational statistics which include if a team is home/away or statistics that are standard but averaged out over all the teams previous games, i.e. a team’s winning percentage or a teams average goal scored.
* Might need to adjust goals for and goals against
* Home/Away Features – There is an assumption that can be made that a team may play differently as an away team compared to how they play as a home team. In order to enhance this assumption, I split out the statistics of a team given the home/away status of the team in hopes that in the model can pick up a correlation in the change of numbers and attribute it to the prediction. Applied to not just the current team being tested but also the opponent of the current team and the previous opponents of the current team and the opponents of the opponent of the current team.
* Extended Features - These features were created based off the assumption that certain statistics impacted the decision model more than of the given statistics. With that in mind, we applied the accumulation of these features across all previous games to the current team, the opponent of the current team, the current teams previous opponents, and the opponents of the current team’s opponent.
* The below images show a breakdown of the score of both Home/Away teams and the result of the match. Some interesting points to note is that if an away team does not score any points so far this season those teams have lost 28% (71 games) of the time. Where on the other hand there have been 20 games where the home team has not scored and lost the game. Also, there have been 175 games (68%) where the away team has not scored or has only scored one goal and of those games they have only won 9 of those games. There have been 218 (85%) games where the home team has scored 1 goal or more and of those 218 games they have only lost 25 (11.4%) of those games.

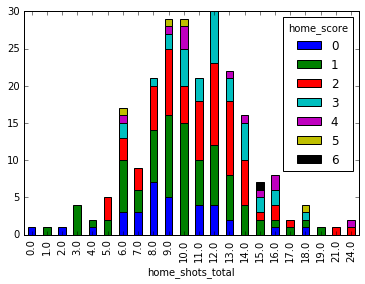


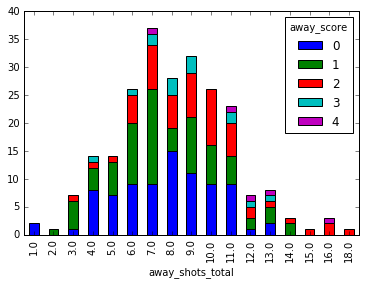
* An interesting observation when exploring the data shows a direct relationship between ‘Attacks’ and ‘Ball Safe’ (please see PDF of the Extended\_Features\_SP). Obviously there should be some direct relationships between some of the extended features (‘possession’, ‘ball\_safe’, ‘attacks’, ‘dangerous\_attacks’, ‘shots\_on\_total’, ‘shots\_on\_target’) such as ‘attacks’ and ‘dangerous attacks’ and also ‘shots\_on\_total’ and ‘shots\_on\_target’. But ‘Ball Safe’, which SportRadar defines as ‘a ball controlled by a team on their end of the field’, influences ‘Attacks’ which consists of a team playing the ball in the offensive third (opposite side) of the field.



Also interesting to note is the relationships of the feature ‘possession’ or lack thereof. Possession is ‘the’ stat for soccer. Essentially the thought is the longer a team holds possession of the ball the higher chance they have to win but based on this subset of data Possession has a very loose relationship with Attacks, Dangerous Attacks, and Shots Total. It’s not to say that these determine the outcome of the game but it’s noteworthy in that Possession might not control all aspects of the game as previously thought.

* Looking at the Total Shots and the Scores Values we can see that in the majority of the games, home teams generally get more shots off (8-12) (203/373 = 54%) where as the away teams shots are lower, between 6-10 224/373 = 60%. The home/away data show a significant difference between home and away games. Home games are more evenly distributed amongst the data compared to the away games where it’s apparent there is a skew in the data to the left. This essentially supports the theory of home field advantage and the teams are ‘weaker’ when they play away.
* Also to note is that within these normal ranges for both home and away the home 18.2% (37/203) of the time never score when they take these range of shots where the away teams never score 31.2% (70/224) of the time. Which confirms the slight shift between the home and away data and also suggests a relationship (though weak it may be) between scoring and shots taken.





1. Initial Hypothesis

In order to predict the results of the match and begin to understand when a team can draw, win, or lose we need to try break down the match into more detail so that we can analyze the results of the two teams in a match individually and then compare those results together to determine a result. To do this I propose we first run a predictive algorithm to try to predict the number of goals a team scores in a match. We can either try a classification model or a regression model. Once we have our determined our model and are satisfied with the predictions, we can then either use the score predictions (expected score) in the final classification model to help predict the result of the match. Or we can combine the results of the score predictions with the results of the classification model that predicts the results of the match to help determine the final results of the match.

After analyzing the data one can begin to understand the relationship between the features and how we can begin to determine the results of a game. There is a strong correlation between the features ‘Ball Safe’ and ‘Attacks’. The more ‘Attacks’ there are the more ‘Dangerous Attacks’ there will be. The more ‘Dangerous Attacks’ there are increases the likelihood of ‘Shots Total’. The more ‘Shots Total’ there is increases the chances of there being a high number of ‘Shots on Target’. And the more ‘Shots on Target’ there is lead to more ‘opportunities’ to score goals for a team (and not allow the other team to score goals).

Ball Safe 🡪 Attacks 🡪 Dangerous Attacks 🡪 Shots Total 🡪 Shots on Target 🡪 Goals

Essentially we should be able to use these features in order to help predict how many goals a team should score. And inversely how many goals a team will let in against them.