Soccer Predictions - ISyE 7406 Data Mining Project

Group C9

April 23, 2017

Team

Parit Burintrathikul - 48129 - paritb@gatech.edu

Spencer Collins

Johnny Humphrey - 15569 - johnny.humphrey@gatech.edu

Shoili Pal - 43987 - spal41@gatech.edu

Steven Yeh - 28005 - syeh35@gatech.edu

Abstract

The goal of this project is to predict soccer match outcomes. Using non-statistical heuristic models as baselines, we develop predictive models based on data mining techniques. We try to determine the factors having the greatest impact on outcomes. We compare our results to professional gambling websites' odds. Finally, we suggest ways that could possibly lead to improving our predictions further.

Introduction

Motivation and Description

Sports are a popular form of entertainment. A form of entertainment parallel to the playing or watching matches themselves is to predict the outcomes. Much time, effort, and money are spent trying to make accurate predictions, especially at a rate better than everyone else. With this project we enter that highly competitive parallel stream. The sport we chose to model was soccer, with some general statistics coming from European soccer and our predictive models built based on data from the English Premier League (EPL). We decided to concentrate on one league since we had more data for it and since different leagues tend to have different styles of play which could affect the accuracy of our models.

Challenges

The variability in sporting outcomes, and the fact that it is hard to quantify the variability given that athletes have good days and bad days, is what adds novelty to sporting events, and what helps make them so entertaining. However, it makes modeling sports with a high degree of accuracy much more difficult.

While we decided to model the outcome of a match instead of the scores, soccer has three possible outcomes, not just win and loss but also ties. In the English Premier League there are no tie-breakers. This makes the task of prediction more complex since the outcome is not binary. We tried several different model types and features to see which did the best job of accounting for the variation in outcomes, as measured by the accuracy of our predictions.

Data

Data Source and Description

The data we are using is a dataset obtainable at https://www.kaggle.com/hugomathien/soccer in the form of a SQL database. It was originally created by scraping several websites and crowdsourcing. It contains information about soccer matches in the top tier leagues from 11 different European countries. The following are the tables in the database and a description of the data that they hold.

- Country: Country name and ID for 11 European countries.
- League: League name, country, and ID for the top league in the 11 different countries.
- Team: Team name, league, IDs for this database, for the match information web site, and for the FIFA video game from Electronic Arts (EA). 299 records total, 34 in the EPL.
- Match: Season, match IDs, home and away team goals, home and away player starting position and IDs, match events(goal types, possession, shots on/off, penalty cards, fouls committed, corner kicks), betting odds from up to 10 providers. 25979 matches total, 3040 total for the EPL spread across 8 seasons.
- Player: ID, name, FIFA ID, birthday, height, and weight for 11060 total players.
- Player Attributes: Player ID and FIFA ID, date updated, 38 FIFA attributes (overall rating, potential, attacking and defensive work rate, crossing, finishing, reactions, shot power, etc.). Each row represents an updated FIFA profile for a given player, so multiple rows per player are possible.
- Team Attributes: Team ID and FIFA ID, date updated, 21 FIFA team attributes (build-up classes and values, chance creation classes and values, defense classes and values)

Data Pre-processing

In our initial exploration of the database, we quickly realized that many of the variables in the dataset had a lot of missing values, mostly in the betting data and players in each match variables. Many of the player's data were crowdsourced by individuals and thus were only filled in haphazardly. This may lead to being biased towards players who are popular and outstanding. Therefore, we decided to remove all player data and focused solely on the team data like overall style of play and match related data like goal differentials, specifically in the English Premier League as we had more data for that.

We joined all the separate tables except the players table into one aggregated dataset, cleaning out N/A values as well as removing all the duplicated identifiers. We built two separate datasets, one including and one excluding betting data. Our assumption was that since the betting data were modelled and calculated by betting companies, they would be highly correlated to the match outcomes and since our objective is to predict match outcomes, building our predictions using other prediction data would not be ideal. However, we also decided to include them in separate models to compare the performances of models with and without the betting data. The dataset ultimately contains each match as one row, with predictor variables pertaining to the teams and the response variable being win, loss, or draw for the home team.

While the Match database table had the home and away goal totals, it did not explicitly have the goal differential, which we were interested in. We created a custom table to hold the results of a team's home or away goal differential going into each match, with the time horizons for calculating the differential of 1, 2, 3, 5, and 10 games, as well as season-to-date.

Exploratory Data Analysis

Looking at each league across all of their seasons in the database, we can find the average percentage outcomes of a home win, an away win, or a draw for the league. We found that there was some variation. [Appendix Tables 1 & 2]

x	Home Win %	Away Win %	Draw %
High	48.8	33.8	28.3
Low	41.7	27.0	23.2
Average	45.9	28.7	25.4

Looking specifically at the EPL, and looking at per-season statistics, we notice that the average EPL season is a lot like the average across all European leagues, but there is still some fluctuation in the home-win/away-win/draw percentages across years. [Appendix Tables 3 & 4]

x	Home Win %	Away Win %	Draw %
High	50.8	32.4	29.2
Low	41.3	23.7	20.5
Average	45.7	28.5	25.8

Finally, as a reference for a good prediction accuracy, we note that the professional oddsmakers correctly predict home-win/away-win/draw for a match in the EPL 53% of the time. [Appendix Table 5]

Methods

Baseline

First we establish a baseline against which we will compare our models. We used a couple of prediction heuristics for this.

Home Wins

This heuristic for predicting winning outcomes is correct 46% of the time, as we see from the data distribution tables. That is better than random picks based on the 46% - 28% - 26% distribution of home-win/away-win/draw, which would be about 36%, but it is still significantly behind the oddsmakers' accuracy at 53%, so we hope to do better with a statistical model.

Goal Differential

The Home Wins rule for picking the winner is a good heuristic only for predicting home team wins. For a simple way to pick between the three outcomes goal differential is a good model.

The correlation of goal differential with wins is common knowledge in the sport community. It is obvious that, when wins are determined by scoring more than your opponent, something that measures that ability should correlate well with wins. For this reason, web sites like ESPN list goal differentials along with the other statistics they provide in their standings. Once again, soccer has draws as an outcome, and they are not as good as a win but are better than a loss, the English Premier League weighs wins as 3 points and draws as 1 point when calculating total points to determine a team's position in the league table.

As an example, when we use that metric for points and compare a team's point total with their goal differential for the 2015 - 2016 EPL season, we find a correlation between points and goal differential of 0.961, which is very strong.

While we have seen a strong correlation between a season-long cumulative goal differential and total points, these are both aggregate measures, and so we still needed to determine how well goal differential helps

predict individual match outcomes. To make those predictions we need to decide upon rules for mapping goal differential information into a match outcome. We use the following:

1. As we've seen in the outcome distributions, having the home field is an advantage, so we keep track of home goal differential and away goal differential for teams separately.
2. We treat seasons separately. It is possible that there could be some carry-over from season to season, but we don't try to determine this. For our calculations, goal differential starts back at 0 at the start of every season.
3. It's possible that goal differential is a stronger predictor over some number of the most recent games than season-to date. We checked results using goal differentials calculated from the most recent single game, and the most recent 2, 3, 5, and 10 games.
4. We need to compare average per-game goal differentials even if the total is determined over some number of games. For example, if the home team has played 3 home games but the away team has only played 1 away game, it wouldn't be a fair comparison to use a 3-game home total against a 1-game away total.
5. The rule we use to determine a draw is if the home team's home goal differential average is within a half game of the away team's away goal differential average, then we predict a tie. If the difference is greater than a half point, then the team with the greatest goal differential is picked as the predicted winner.

We note that these rules are not a Data Mining or Machine Learning model. We aren't estimating any parameters to minimize error. For example, it is possible that we would have greater accuracy if we used some number higher or lower than 0.5 as the cutoff for draws. We did not seek to find out what that number might be. We are only trying to see if we can improve on our baseline Home Wins rule. Also, even if goal differential isn't enough by itself for predictions, we may find that it is a useful feature to have in a model.

After calculating the goal differentials and then the predicted outcomes, we compared the actual outcomes and found that the accuracy in the EPL of the goal differential predictions was about 43%. [Appendix Table 7] This is actually slightly less than simply predicting that Home Wins.

Models

We tackle the problem in two ways -

- as a 3-class classification problem into wins, losses and draws for the home team.
- as a 2-class classification problem where the response is home win or not.

We use both parametric and non-parametric methods to model the problem. While the parametric models of logistic regression and SVM will give us a clear idea of the relation between the predictor variables and the response, sometimes non-parametric models like random forests can learn patterns that are not obvious to the human eye and give good results, especially in an area as unpredictable as sports. Hence we try both.

We use the same models for the binary classification and the 3-class classification.

Ten fold cross validation is used on the regression and SVM models to establish the testing error.

Parametric

Logistic and Multinomial Regression

Logistic regression was used in the two-class prediction problem and Multinomial regression for the 3-class problem. Relying on the logit function as the link function between the calculated probability and the linear regression function, the model fits coefficients for each predictor variable. Logistic regression does not make any assumptions about common variance, and can be used as a robust estimator in this problem.

Support Vector Machines

Support Vector Machines was chosen as a model for predicting both the two-class (Win/Not) and three-class (Win/Loss/Draw) problems. SVMs are useful in high dimensional space and can be tuned using different kernel functions for the decision function and different decision boundaries 'gamma'. A grid search was performed to find the best value of gamma and several kernels were tried like the radial and sigmoid kernels.

Ensemble of Regression and SVM

Bradley Terry Model

Non-Parametric

Random Forest

Random Forest was chosen as one of the models to try as it is known to give good accuracy in many cases and is not too computationally intense. It was used for both Win/Not Win and Win/Draw/Loss prediction.

Results

For Binary Classification

Name	Testing Error
Bradley Terry Model	
Logistic Regression (with Betting data)	
Logistic Regression (without Betting data)	
SVM (with Betting data)	
SVM (without Betting data)	
Random Forest (without Betting data)	0.53
Random Forest (with Betting data)	

For 3-Class Classification

Name	Testing Error
Logistic Regression (with Betting data)	
Logistic Regression (without Betting data)	
SVM (with Betting data)	
SVM (without Betting data)	
Random Forest (without Betting data)	0.51
Random Forest (with Betting data)	
Ensemble Method	

Discussion

Conclusions

Appendices

Bibliography and Credits

- European Soccer Database

https://www.kaggle.com/hugomathien/soccer

- SQLite

https://www.sqlite.org/index.html

- R

https://www.r-project.org/

https://cran.r-project.org/doc/manuals/r-patched/R-intro.pdf