England Premier League Match Outcome Prediction

Team F: Jiayi Dong, Puthiwinyoo Thachakorn, Yanming Xue, Ziqiu Zang, Mengran Zhang

Introduction

Soccer is one of the most popular sport games in the world, and the most famous soccer league is the English Premier League (EPL). The large population of the audience and soccer fans makes each match important and valuable.

We are representatives from a consulting company which is working with a betting company. As plenty of factors play critical roles in soccer, it tends to be difficult to accurately predict match outcomes and the final winner, which contributes to the soccer betting industry. Moreover, the effect of luck will make the prediction even harder. Our goal is to generate more accurate model to predict probability of each outcome for future games played within the EPL, and to improve the expected return of the betting system.

For EPL, a season runs from August to May next year, and 20 teams contest for the first place, which the bottom three teams will be replaced by top three teams from lower leagues. Because each team plays every other team twice as home and away, there are totally 380 games per season.

Data Collection

EPL data is ranging from season 2008/09 to season 2015/16. We selected the matches that played between the current 20 active teams (1375 out of 1729 matches). Original datasets contain around 160 attributes for away and home teams, however, most of them could not be directly used, such as team formation and red cards.

Details regarding used datasets and 20 active teams:

Dataset	Description
Match.csv	Main dataset, contains team ID and Player ID, scores, team formation, goal types, corner, cross and cards of each match
Player.csv	Player names and ID
Player_attribute.csv	Player ID and overall score
Team.csv	Team ID

Team Name			
Arsenal	Everton	Manchester United	Swansea City
Bournemouth	Hull City	Middlesbrough	Tottenham Hotspur
Burnley	Leicester City	Southampton	Watford
Chelsea	Liverpool	Stoke City	West Bromwich Albion
Crystal Palace	Manchester City	Sunderland	West Ham United

Data Organization

Step 1. Clean rows with missing values and columns with identical values
Step 2. Extract values (number of red cards, number of shoot on targets etc.) from html structure
Step 3. Calculate average score by players' position
Player score: we assume player's score does not change very much during this period.
From each player's X, Y coordinates, get information about each team lining up with squad formation (ex.4-4-2) and split all positions (11 for each team) into four major areas: GoalKeeper, Strikers(Front), Midfielders(Middle) and Defenders(Back) (**Figure 1**). Calculate the team area scores in each match by averaging players' overall scores in each area given by FIFA database.

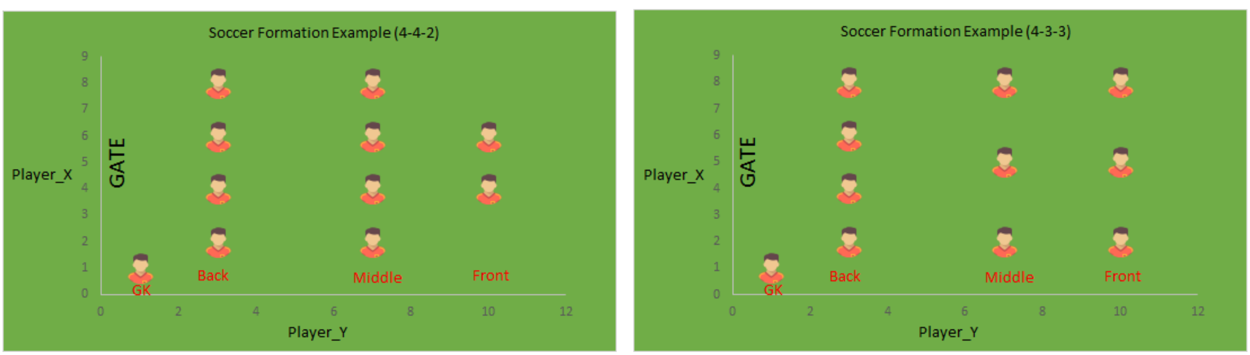

Figure 1. Player Positions

Step 4. Determine Home Advantage
For each team, we calculate the difference of the scores when this team plays home and away, and average the difference of the scores for the past years. Finally there are 20 unique home advantages.

$$Home\ Advantage = \frac{Team's\ Total\ Home\ Scores - Team's\ Total\ Away\ Scores}{Number\ of\ Matches\ Team\ Attended\ /2}$$

Step 5. Generate response variables: goal difference and game outcome (win, draw and loss)
After cleaning and organizing the dataset, there are 19 attributes.

Attribute	Description	Source
RedCard (Home/Away)	Red Card Amount	Step 2 From HTML code in Match.csv
ShotOn (H/A)	Shot On Target Quantity	Step 2 From HTML code in Match.csv
ShotOff (H/A)	Shot Off Target Quantity	Step 2 From HTML code in Match.csv
Corner (H/A)	Corner Quantity	Step 2 From HTML code in Match.csv
Cross (H/A)	Cross Quantity	Step 2 From HTML code in Match.csv
GK_AVG (H/A)	Goal Keeper's Rating	Step 3 From Player_Attribute.csv
Back_AVG (H/A)	Defenders' Average Rating	Step 3 From Player_Attribute.csv
Middle_AVG (H/A)	Midfielders' Average Rating	Step 3 From Player_Attribute.csv
Front_AVG (H/A)	Strikers' Average Rating	Step 3 From Player_Attribute.csv
Home_Adv	Home Advantage	Step 4 Calculated in Match.csv

Response Variable	Description	Source
Result	Match Result (Win, Draw, Loss)	Step 5
Goal_diff	Goal Difference between Home and Away Team	Step 5

Model Selection

The real-time match data cannot be obtained before the match starts. In addition, one single match data can only reflect limited information about a team's performance. Hence, we average the data of certain amount of previous matches played by a team, and use these new data to predict the performance of the team in the next match.

We generate the new dataset by doing the following:

- Average past k games' data on ShotOn(H/A)/ ShotOff(H/A)/ Corner(H/A)/ Cross(H/A).
- As for the number of red cards received, we think only considering the number of red card from the very last matches is biased, so we use lasso to find how many previous matches should be used to average for the number of red cards.
- Other features remain the same.

Then we randomly select 75% of the total data as train dataset and left 25% as test dataset.
We train and test the following models and compare the error rate of each (**Figure 2**), finally we conclude the LDA model perform the best:

Method/Prediction	Prediction Type	Error Rate	k
Linear Regression	Goal Difference	2.427582	---
Linear Regression + Forward Subset Selection	Goal Difference	2.430666	---
PLS	Goal Difference	2.422217	---
PCR	Goal Difference	2.505872	---
Multinomial Logistic Regression	Match Outcome	0.4534884	4
Multinomial Logistic Regression + Lasso	Match Outcome	0.433195	4
LDA	Match Outcome	0.3372093	2
KNN	Match Outcome	0.6226964	4
Decision Tree	Match Outcome	0.4622093	2

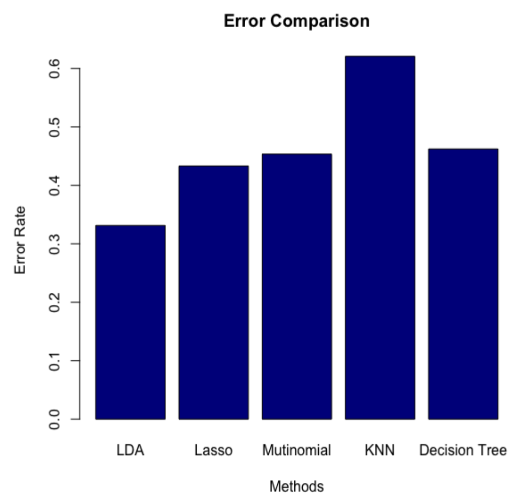

Figure 2. Error Comparison of Different Models

Column k shows the number of previous matches we should take average of. We conclude that since it will give us the least error rate. Also, after training and testing the model, we decide to abandon the goal-difference approach of prediction because the mean squared errors are high comparing to the actual score margin. Also, we do not include luck in the model because it will have same effects on both teams.

Final Models:

LDA Model for Outcome Probabilities:

```
Coefficients of linear discriminants:
                      LD1         LD2
H_GK_AVG       0.02294249  0.02215330
H_Back_AVG     0.12006513 -0.13295535
H_Middle_AVG   0.09610578  0.01034457
H_Front_AVG   -0.01673923  0.01575037
A_GK_AVG      -0.01453823 -0.04893227
A_Back_AVG    -0.09804280 -0.09330187
A_Middle_AVG  -0.06217975  0.10803546
A_Front_AVG   -0.01817038 -0.08926917
Home_Adv       1.19581001  2.39568101
redCard_Home  -2.94832802  3.66986680
redCard_Away  -0.38351609  1.80193591
shotOn_Home    0.01567463  0.01357106
shotOn_Away    0.03693227  0.26801521
shotOff_Home   0.06469156  0.20354291
shotOff_Away  -0.02051939 -0.08829614
corner_Home   -0.06507730 -0.05708049
corner_Away   -0.15719943  0.12374764
cross_Home    -0.08309752  0.06504712
cross_Away     0.03458597  0.09663582
```

```
Group means:
    H_GK_AVG H_Back_AVG H_Middle_AVG H_Front_AVG A_GK_AVG A_Back_AVG
0   76.67616   74.53320     75.21128    75.64189 79.19217   77.07060
0.5 78.00000   75.89010     76.50663    76.69108 78.52778   76.16201
1   79.23593   76.76028     77.47939    77.60435 77.08225   74.96545
    A_Middle_AVG A_Front_AVG  Home_Adv redCard_Home redCard_Away
0       77.73797    78.24203 0.3720052   0.15658363   0.06049822
0.5     76.60125    77.51271 0.3736138   0.04861111   0.06944444
1       75.47305    76.04693 0.4272953   0.02597403   0.11471861
    shotOn_Home shotOn_Away shotOff_Home shotOff_Away corner_Home
0      6.245552    5.686833     6.341637     5.555160    5.975089
0.5    6.968750    5.302083     6.392361     5.194444    6.368056
1      6.969697    5.266234     6.696970     4.974026    5.948052
    corner_Away cross_Home cross_Away
0      4.975089   18.30961   12.23843
0.5    4.736111   17.77431   13.01736
1      4.675325   14.59957   13.57359
```

Prediction Future Match using LDA

Using our LDA model, we predict the outcome of the following two matches which were just completed recently. We collect new data of the past two matches played by those four teams in March and April, such as number of corner points, number of red cards received etc., and plug those numbers as variables into our model. We compare our predictions with the actual outcomes, and also compare it with the predictions provided by a famous sport website, FiveThirtyEight.com:

Match	Actual Outcome	Website Prediction	Our Prediction
Chelsea vs Southampton (4/25/2017)	**4:2**	***Win: 70%*** Lost: 10% Draw: 20%	***Win: 67.68%*** Lost: 5.61% Draw: 26.71%
Sunderland vs West Ham United (4/15/2017)	**2:2**	Win: 33% Lost: 39% ***Draw: 28%***	Win: 34.98% Lost: 33.73% ***Draw: 31.29%***

As shown in the table above, for the 1st match, our model predicts the outcome correctly. For the 2nd match, although Win has the highest probability in our model, while the actual outcome is draw, the probabilities of the three outcomes in our model are very similar. Comparing the prediction from FiveThirtyEight.com, our prediction is much closer to the real outcome.

Poisson Distribution in Predicting Scoreline Probability:

In the poisson model, we first determine the “Attack Strength” and “Defence Strength” for each team. The 38 games played by each team in the 2015/16 EPL season will provide a sufficient sample size to calculate these attributes.

Here’s how we calculate the expected goal for the home team:

- Home Team’s Attack Strength: Average number of goals scored at home by home team/Average number of goals scored at home by the league
- Away Team’s Defence Strength: Average number of goals conceded away from home by away team/Average number of goals conceded away from home by the league
- Home Team’s Goals Expectancy: Home Team’s Attack Strength * Away Team’s Defence Strength * Average number of goals scored at home by the league

Similarly, we can predict the away team’s goals expectancy.

After calculating the average value of goals scored per team, we can use Poisson Distribution to distribute 100% of probability across a range of goal outcomes for each team.

Poisson Distribution for predicting West Ham vs Tottenham on 05/05/2017:

	0	1	2	3	4	5	6
West Ham	52.86%	33.70%	10.74%	2.28%	0.36%	0.05%	0.00%
Tottenham	14.43%	27.93%	27.04%	17.45%	8.45%	3.27%	1.11%

We assume both scores are independent, then the expected score will be 0:1 with the probability of (0.5286 * 0.2793) = 14.76% - pairing the most probable outcomes for each team.

Real-life Applications in Betting Industry

For the bookmakers, the odds are calculated by the formula (1/probability) and plus the edge to make profits. By reducing error rates of predicting outcomes, the betting company could formulate a more accurate betting system and improved expected profits.

Reference

Cronin, Benjamin. "Poisson Distribution: Predict the Score in Soccer Betting." *Pinnacle*. N.p., n.d. Web. 26 Apr. 2017.

Jayboice. "2016-17 Club Soccer Predictions." *FiveThirtyEight*. N.p., 25 Apr. 2017. Web. 26 Apr. 2017.

"Learn How To Convert Odds - Conversion Calculator." *Bettingexpert.com.* N.p., n.d. Web. 26 Apr. 2017.

Ollie. "Home-Field Advantage Doesn’t Mean What It Used To In English Football." *FiveThirtyEight*. FiveThirtyEight, 08 Oct. 2014. Web. 26 Apr. 2017.

Data Sources:

football-data.co.uk
football-data.mx-api.enetscores.com
kaggle.com/hugomathien/soccer
EA Sports' FIFA games website