Use FIFA Game's Rating To Predict Football Matches

YE Rougang

The Hong Kong University of Science and Technology

Abstract

The world of sports is abundant in quantifiable elements, making it ideal for using data analytics to predict match outcomes. So I try to analyze historical football data from season 2008 to 2016 to check whether one can build a predictive model. My work mainly includes:

- Analyze the data and extract features.
- Apply machine learning models and compare results with betting odds.
- Derive the MM algorithm of a generalized Bradley-Terry model and apply it on the data.

Introduction

The movie *Moneyball*, among many things, can be considered as the prime example of data-driven performance optimization in sports. It depicts the story of how a MLB team's general manager used statistical data and analytics to build a competitive team despite the team's small budget.
Here I use a data set from Kaggle [1]. The data set includes **25979 matches** of 11 European leagues including England Premier League, France Ligue 1, Germany Bundesliga, Italy Serie A etc. Matches start from 2008/07/18 to 2016/05/25.

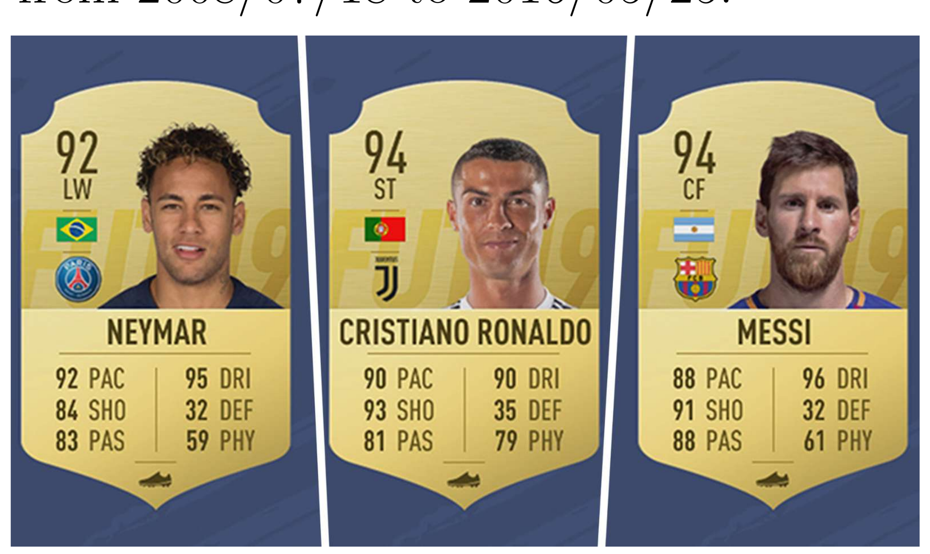

Figure 1: FIFA ratings

Especially, the data set has **attributes of 11060 players**. They are crawled from EA Sports' FIFA video game series including the weekly updates. Specifically, each player has abilities of different dimensions: attacking, skill, movement, power etc. Figure 1 gives an example of FIFA ratings of players.

Feature Engineering

I create features from the data based on the idea that a team's rating should combine FIFA ratings of its players. Recent performance of teams also need to be considered.
Finally, there are 30 features for one match (sample), i.e.,

- Home team's 7-dim ratings of [attacking, skill, movement, power, mentality, defending, goalkeeping]
- The mean difference of goals and 7-dim ratings of home team and its opponent of last 10 matches
- Away team's 7-dim ratings of [attacking, skill, movement, power, mentality, defending, goalkeeping]
- The mean difference of goals and 7-dim ratings of away team and its opponent of last 10 matches

Generalized Bradley-Terry Model

The Bradley-Terry model [2] is a simple and much-studied method to describe the probabilities of possible outcomes when subjects are judged against one another in pairs.
In order to apply it on the football matches, I extend it to **involve home-field advantage and allow ties happen between 2 subjects:**

$$P(i \text{ beats } j \text{ at } i\text{'s home}) = \frac{\alpha r_i}{\alpha r_i + \theta r_j}$$

$$P(j \text{ beats } i \text{ at } i\text{'s home}) = \frac{r_j}{\alpha\theta r_i + r_j}$$

$$P(i \text{ ties } j \text{ at } i\text{'s home}) = \frac{\alpha(\theta^2-1)r_i r_j}{(\alpha r_i + \theta r_j)(\alpha\theta r_i + r_j)}$$

where r_i is rating of the subject i, $\alpha > 0$ measures the strength of the home-field advantage or disadvantage, $\theta > 1$ is the threshold of draw.
Based on the above model, log-likelihood function is

$$l(\boldsymbol{r},\theta,\alpha) = \sum_{i=1}^{n}\sum_{j=1}^{n}\Big[a_{ij}\ln\frac{\alpha r_i}{\alpha r_i + \theta r_j} + b_{ij}\ln\frac{r_j}{\alpha\theta r_i + r_j} + t_{ij}\ln\frac{\alpha(\theta^2-1)r_i r_j}{(\alpha r_i + \theta r_j)(\alpha\theta r_i + r_j)}\Big]$$

where $a_{ij}/b_{ij}/t_{ij}$ is the number of times i beats/loses to/ties j at i's home.

Methods

After sorting the dataset by match dates, I split the data set into training set and test set by the ratio of 75%/25% (19484 games and 6495 games). Training matches start from 2008/07/18 to 2014/08/17. Test matches start from 2014/08/17 to 2016/05/25.
Question now becomes a 3-class(home win, draw, away win) classification.
Then I apply different machine learning models (random forest, XGBoost, logistic regression, SVM) on the training set and compute several metrics to check models' performance on the test set. **In order to compare with odds implied probability, the metrics are computed with respect to 5698 games of the test set, which have non-missing PS odds.**
The baseline is to use [1/3, 1/3, 1/3] as the predicted probabilities of 3 outcomes.

MM Algorithm

Hunter [3] proposed MM algorithm for generalized Bradley-Terry models. From the strict concavity of the logarithm function, that is, for positive x and y,

$$-\ln x \geq 1 - \ln y - (x/y)$$

where the equality is obtained if and only if $x = y$, we can construct a minorizing function $Q_k(\boldsymbol{r})$ which satisfies

$$Q_k(\boldsymbol{r}) \leq l(\boldsymbol{r}) \quad \text{with equality if } \boldsymbol{r} = \boldsymbol{r}^{(k)}.$$

And then maximize $Q_k(\boldsymbol{r})$ as

$$Q_k(\boldsymbol{r}) \geq Q_k(\boldsymbol{r}^{(k)}) \quad \text{implies} \quad l(\boldsymbol{r}) \geq l(\boldsymbol{r}^{(k)}).$$

In this way, we can **iteratively create a minorizing function and then maximize it to update all parameters in BT model**.

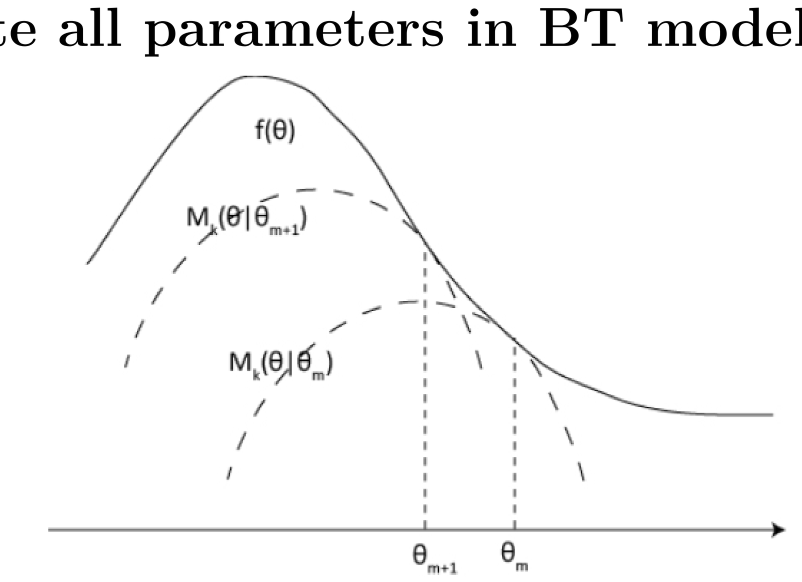

Figure 2: MM algorithm

Results

The following table presents the prediction accuracy, mean squared error and F1 scores of 3 outcomes of different methods. PS(Pinnacle) odds can be regarded as the goal to beat the bookmaker.

Method	Accuracy	MSE	F1 score
Baseline	/	0.222	/
PS odds	0.526	0.193	[0.653, 0, 0.503]
Random forest	0.511	0.198	[0.644, 0.02, 0.453]
XGBoost	0.505	0.198	[0.639, 0.03, 0.442]
Logistic regression	0.512	0.197	[0.644, 0, 0.468]
Linear SVM	0.515	0.199	[0.649, 0, 0.448]
RBF SVM	0.512	0.203	[0.647, 0, 0.433]
Generalized BT	0.445	0.214	[0.636, 0.268, 0]

Conclusion

Although predictions of all models are less accurate than the implied probabilities of Pinnacle odds, **some of them, for example, logistic regression, have performance very close to it.** This is not trivial as odds usually include far more information about matches.
This proves the value of ratings of the FIFA game. And the reason they have such power is because a lot of resources have been spent on it to make video games as realistic as possible. It requires collecting and curating a lot of real world data.
Thus, Bradley-Terry model only using match results did not obtain such good result. Bradley-Terry model with covariates should be the next improvement.

References

[1] European Soccer Database.
`https://www.kaggle.com/hugomathien/soccer`.

[2] Ralph Allan Bradley and Milton E. Terry.
Rank analysis of incomplete block designs. I. The method of paired comparisons.
Biometrika, 39:324–345, 1952.

[3] David R. Hunter.
MM algorithms for generalized Bradley-Terry models.
Ann. Statist., 32(1):384–406, 2004.