Projet de DBDM 2

Football Data Challenge: What is a probability that Roma wins against Napoli later today?

Simonaitis et Lajou

25 avril 2016

Our dataset : Football Data challenge

Problem

Roma vs. Napoli, starts at 15 : 00 in Stade olympique, Rome.
The odds from 6 different betting agencies : Bet365, Bet&Win... :

Home odds	Draw odds	Away odds
2.5	3.5	2.8
2.5	3.4	2.8
2.5	3.3	2.65
2.4	3.3	2.8
2.5	3.3	2.8
2.5	3.4	2.9

Data

For each match we have :

1. ID
2. Date
3. Home and away teams
4. Set of odds
5. Outcome for the train set

Data

We are given a train set of 1520 matches from the seasons :

2008-2014

And a test set from the seasons :

2014-2016

Predictions

You still have time to bet for today's matches :

1. Rome vs. Napoli, odds of winning are 46% vs. 32%
2. Verona vs. Milan, odds of winning are 8% vs. 71%

Features

We have tried different features.

The first class was either directly the odds or a ratio of odds :

$$\text{Home odds} \quad \text{Draw odds} \quad \text{Away odds}$$

vs

$$\frac{\text{Home odds}}{\text{Draw odds}} \qquad \frac{\text{Away odds}}{\text{Draw odds}} \qquad \frac{\text{Home odds}}{\text{Away odds}}$$

The second class was an optional additional feature representing the date :

$$\text{Progress within the season} \in [0, 1]$$

Features

The best features with cross validation :

	Without date	With date
Odds	54.44 %	53.95 %
Ratio	54.77 %	54.77 %

The regression model is Logistic Regression.
Final score on Kaggle : 47.541 %

Upsets findings

We also tried to detect upsets within the train set.
DEMO

Thank you for you attention. Your questions are welcome.