Football Data Challenge

prepared by Dimitri Lajou, Pijus Simonaitis
ENS Lyon, DBDM M1

April 25, 2016

1 Problem

In our project we tried to predict an outcome of the football match from the betting odds provided by the betting agencies. Football data challenges data set from kaggle in Class provided us with information on the betting odds from Serie A, Italian football league. We were attracted by this problem because of our interest in football and curiosity about how predictable the outcomes of the football matches are.

2 Description of the data set

2.1 Dataset

Another interest of the data set was that it was easy to comprehend and required little of preprocessing as it consisted of a few factors:

1. Unique ID associated to every match
2. Date of the match
3. Home and away teams
4. Sets of odds proposed by six betting agencies: Bet365, Bet&Win, Interwetten, Ladbrokes, VC Bet and William Hill

A set of odds being the scores for:

1. Home win
2. Away win

3. Draws

The odds represents the amount of money that you will gain from beting on the team. Thus a lower odd means a higher chance of victory.
Training set included 1520 matches from the seasons 2008-2011 and 2013-2014 while test set consisted of matches from the seasons 2014-2016.

2.2 Patterns in the Data Set

In the training set 48.5% of the matches were won by home team, 26,2% by away team and 25,3% resulted in a tie. However only 1.5% of the matches were predicted to result in a tie by at least one of the betting agencies and only 2 out of 1520 matches were predicted to result in a tie by all of the agencies. The rarity of the predicted draws was a handicap that predertimined the fact that almost none of the draws were predicted by our algorithm as well, meaning that outcomes of 25% of the matches were deemed to be predicted falsely.

3 Features

We have tried different features.
The first class was either directly the odds or a ratio of odds:

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

The second class was an optional additional feature representing the date corresponding to the progress within the season. This is a number $\in [0, 1]$. This feature is not precise since some matches belonging to the same football day will not be played on the same day in reality.

4 Model

We tried differents learning models. The two that stands out with the best scores are Logistic regresion and random forest. However the later one is always worst than the first and can at best tie with chance. That is why we choose to use the logistic regression during our tests.

5 Results

We used cross validation on the train set to get an idea of our performances.

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

Those scores are realy close and the test set can make those number unacurate. On Kaggle the best score we could get is 47.541 % with on the ratio features. The best score at the moment is just above 50 %. There is a huge difference in the cross validation scores and the Kaggle score, we do not have a good explanation for it. The only thing that could explain this is a change in the way the odds are computed.

6 Artefact Detection

We also implemented an artefact detection option in our script. This finds a set of matches that didn't go as planed. Concretely it predicts the matches and store the ones where the result has a low probability to have happen. The matches in the train set are too recent to be linked with the fixed matches affair in Italia, so those matches are only regular upsets.