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

3 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

Training set included 1520 matches from the seasons 2008-2011 and 2013-2014 while test set consisted of matches from the seasons 2014-2016.

4 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 on average less than 1% of the matches were predicted to result in a tie by a betting agency and only 2 out of 1520 matches were predicted to result in a tie by all of the agencies. Thus 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.