Discovering the Determinant Features using Regression for Predicting Euro 2016

Maryam Tavakol, Hamid Zafartavanaelmi, and Ulf Brefeld

Leuphana University of Luneburg, Technical University of Darmstadt, Leuphana University of Luneburg, Germany {tavakol@leuphana.de,hamid.zafartavanaelmi@stud.tu-darmstadt.de,brefeld@leuphana.de}

Abstract. This paper is addressing the challenge of predicting Euro 2016 outcomes. A set of processed features alongside with a new proposed feature are used to construct a linear model in order to compute scores of 24 countries which are present in this tournament. The obtained scores form a {*win, lose, draw*} probability for every two countries against each other. The evaluation of the proposed approach until the end of quarter final shows that it is a promising and simple approach for countries with more historical data.

Keywords: Feature extraction, ridge regression, ranking.

1 Introduction

Football is among the most popular sports in the world. Big tournaments attract the interest of different groups of people, every year. The ability to predict the outcome of matches is very challenging as they are highly uncertain. Data Mining and machine learning techniques however are eligible candidates for extracting associated features, learning models, and predicting the outcomes. Several researches have been conducted either regarding sport analysing in general; such as extracting similar trajectories [5] and performance of physical activities [4], or specifically on soccer [3]. Nevertheless, predicting the outcome of a single match is quite new. In the world cup 2014, Andrew Ng introduced the success of a deep learning approach for predicting the winner of knockout games [1].

The UEFA European Championship is occurring every four years between 24 qualified national teams from all over Europe. This paper is attached to the prediction competition for Euro 2016. We introduce a method for predicting the scores of the countries which are present in this tournament. The scores are further used to address the first challenge of the competition; the probability of {*win, lose, draw*} of every two countries against each other.

We use the available dataset provided by the organization which contains general information about countries as well as their players. In addition, we extract a new feature set by aggregating several data sources to create a new informative feature. The features are processed and engineered to be fed into a linear model to estimate a score for each country. The desired probabilities for the mentioned challenge are therefore obtained from the computed scores. In order to learn the model, we extract additional data from the history of all official games between countries. Finally, ridge regression is applied to learn the weights of feature set in the linear model.

In this paper, we first describe the process of extracting features in section 2. In section 3, the proposed method is introduced and the challenge 1 is addressed. The analysis of the results until the end of quarter final is characterized in section 4, and section 5 concludes.

2 Feature Extraction

One of the main stage of such a prediction problem is extracting related features. The process of collecting significant data and processing features has a major impact on the performance of prediction. The competition organization has provided two sets of data. The first dataset contains the overall ranking information per each twenty four participants in Euro 2016, and the other one summarizes the players specific properties. We select a subset of features from the both datasets as follows:

- **Countries**: {FIFA[1] ranking, FIFA points, UEFA[2] ranking, UEFA coefficient, ELO[3] ranking and ELO points}
- **Players**: {Market value, Age, Euro 2016 matches and goals, All time matches and goal, Career matches and goals}

We first operate a simple crawler to get the teams' squad from the official website of UEFA and filter the **Players** data by the final list of players in every team. We then replace the number of appearance and goals of each player by their ratio, e.g., $\frac{num\ of\ goals}{num\ of\ appearance}$. Note that France has no data on Euro 2016 appearance and goal as it is the host and did not participate in the qualification phase. Therefore, we set its ratio to 1 in order to consider host advantage for France. Afterwards, the values of players for each feature are averaged per country; e.g., average age.

In addition, we extract extra data from a few public sources to create an innovative feature. We argue that if a team has more players who are member of a same club and that club has a good reputation, it is more likely that the harmony among them leads to success of their national team as well. For instance, there are five players of Spain who are a member of Barcelona FC. which it has a positive effect on the team. Thus, a list of players' current football club plus the club ranking (top-200) is created from crawling International Federation of Football History & Statistics (IFFHS) for year 2015[4]. Afterwards, we choose for each country the club which has the most of national team players, and in case

[1] Fdration Internationale de Football Association

[2] Union of European Football Associations

[3] World Football Elo Ratings web site, http://www.eloratings.net/

[4] http://iffhs.de/club-world-ranking-2015./

of more than one club, the one with the higher rank is selected. Table 1 shows the short list of statistics for the teams in quarter final. As can be seen in Table 1, this gave us an engineered dataset that contains country, club, number of national team members who play in that club and club world ranking.

Table 1: Club ranking associated with each national team in Quarter-Final

Country	Number Of Players	Club	Club Rank
Spain	5	Barcelona	1
Italy	6	Juventus	2
France	2	Juventus	2
Germany	5	Bayern Munich	4
Belgium	3	Liverpool	42
Poland	3	Legia	52
Portugal	4	Sporting CP	179
Wales	3	Crystal Palace	0*
Iceland	2	Hammarby	0*

* club was not in the top 200 of year 2015.

Once the player related features are obtained per country, we normalize them using feature scaling as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where we further use $1 - x'$ for *age* as the lower the average age of a team, the better the performance. We then add the mean of normalized features form the **Players** data as a new single feature (PlayersScore) to the **Countries** data. Finally, the remaining features of **Countries** data are also normalized by feature scaling.

3 The Proposed Approach

In this section, we present our algorithm for predicting the $\{$*win, lose, draw*$\}$ probability of the teams against each other. The features are extracted, preprocessed and normalized in the previous section and we aggregate them to calculate a score for each country. The scores are obtained as they represent the power of teams in this tournament which are further applied in the algorithm for assigning the required probabilities in the first challenge.

We assume that the score of each country is a linear model of its features (including players features). Let $s_i, i \in \{1, ..., 24\}$ be the score of ith country, and $\mathbf{x}$ is the corresponding feature vector. Our linear model aims at learning a weight vector θ such that $s_i = \theta^T \mathbf{x}$.

3.1 Learning

For learning the parameters θ, we describe a method to compute an estimation of scores, $\hat{s_i}$, for every country. The head to head records of national teams against each other is gathered from [2]. Each record of two countries contains the number of win, draw and lose for them. This historical data is captured for the training purpose, hence, we are able to manipulate the dataset format to a more convenient form. The counts of win, draw and lose are converted to the probabilities of respected fields. For instance, Italy and Sweden played 21 times against each other. There is 6 times winning for each and they drew 9 times. The $\{$*win, lose, draw*$\}$ probability is $\{0.28,\ 0.28,\ 0.43\}$. In the next part, we explain the conversion process of scores to the probabilities. Therefore, connecting the probabilities from historical records to the scores is the other way around.

By applying ridge regression on the data, the weight vector is optimized as follows:

$$\hat{\theta} = (X^T X + I)^{-1} X^T \hat{\mathbf{s}},$$

where $X \in R^{24\times 7}$ is the matrix of seven final features for 24 countries, I is identity matrix, and $\hat{\mathbf{s}}$ is vector of their scores. Each entry in $\hat{\theta}$ shows the importance of the feature in that position of the vector; e.g., FIFA ranking and PlayersScore are the most important features, while EUFA coefficient is the least important feature. The obtained scores are used for prediction in the following challenge.

3.2 Challenge 1: Predicting Match Outcome

In the first challenge, a prediction of $\{$*win (P_w), lose (P_l), draw (P_d)*$\}$ probability for each country against every other country is required. A single score is used for defining the desired probabilities. For two countries i and j with scores of s_i and s_j, respectively, the probabilities are computed as follows.

```
if $s_i \geq s_j$:
    $P_{w_i} = \frac{s_i}{(s_i+s_j)}$
    $P_{w_j} = (1 - P_{w_i}) * s_j = P_{l_i}$
else:
    $P_{w_j} = \frac{s_j}{(s_i+s_j)}$
    $P_{w_i} = (1 - P_{w_j}) * s_i = P_{l_j}$
$P_d = 1 - P_{w_i} - P_{w_j}$
```

In this setting, the country with the higher score is more likely to win.

4 Performance Analysis

We evaluate the performance of our algorithm by comparing the predicted values to the actual results. As the results are determined till semi-final, we easily compute multi-class logarithmic loss of $\{$*win, lose, draw*$\}$ probabilities as below:

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * log(p_{ij}),$$

where N is the number of games and M is equal to three classes as we are interested to calculate the loss value for the predicated probability of win, draw and lose. Figure 1 summarizes the log loss error for 45 matches. The average loss value is 1.3187.

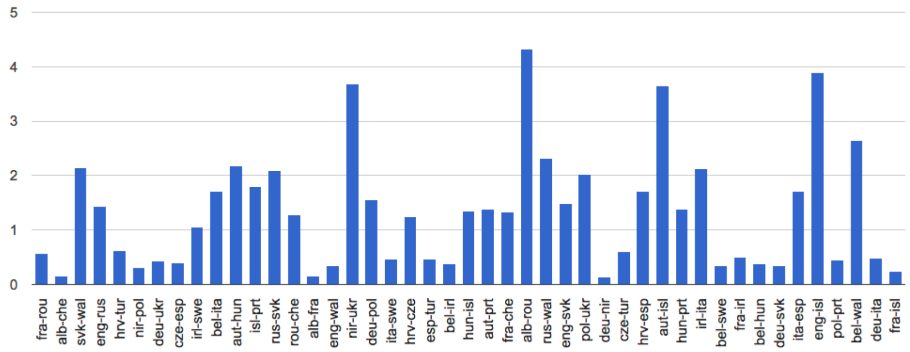

Fig. 1: Average of Logarithmic Loss for Chanllenge 1

Although the average number of head-to-head matches is 13.7746, historical data for several countries are not adequate for a justifiable predication. Figure 2 shows the number of matches of each team with all other teams in this tournament. It can be spotted in Figure 3 that most of the countries with high loss value, were provided by small number of historical data on previous matches. For example, the number of matches between Wales and Russia were limited to four which leads to a deficient predication. Figure 4 shows matches when more than four historical data was available for each match. The average of logarithmic loss declined to 1.1129.

Moreover, the loss for the teams which have no record or just one appearance in previous Euro championship are relatively higher than the rest of the teams. Among all twenty four teams, Albania, Iceland, Northern Ireland, Slovakia and Wales had not been qualified before. Austria and Ukraine had the chance to play in Euro championship just once. We get a lower loss value (0.9680) when we only consider the matches conditioned on at least two appearance in Euro championship. It is depicted in Figure 5. Therefore, when there is enough historical data available, our simple approach is able to promisingly predict the outcome of matches.

5 Conclusion

We discussed creating an engineered feature set for the predication of Euro 2016 football championship. We also represented a linear model that aimed to find out the probability of each match outcome. By applying the model on the features, we showed that features such as similar arrangement of football club in national

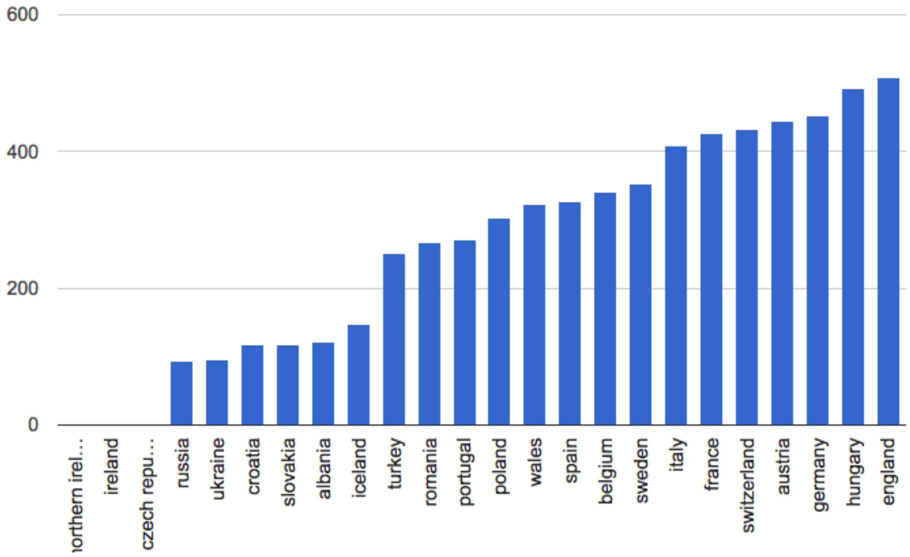

Fig. 2: The number of historical data (head-to-head) for each country

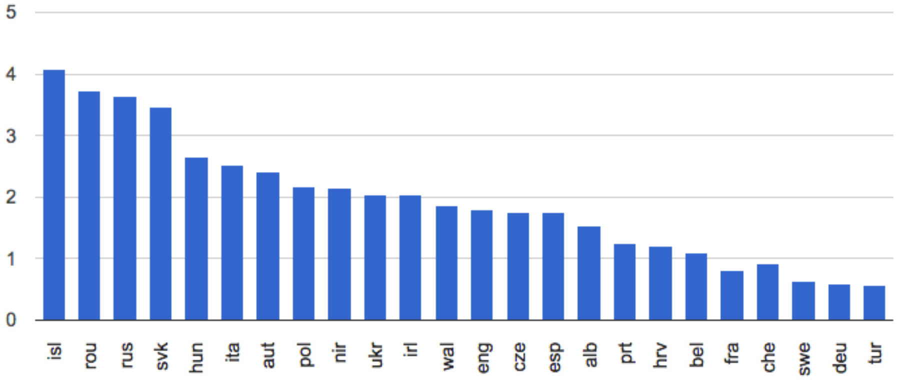

Fig. 3: Average of Logarithmic Loss of each country for Challenge 1: Sorted by loss value

teams are useful to predicate the result of a match. In addition, historical data shown to be beneficial to improve the accuracy of a football match predication.

References

1. Knockout prediction. `http://trends.baidu.com/worldcup/events/knockout?locale=en`, 2014.
2. International football history and statistics. `www.11v11.com`, 2016.

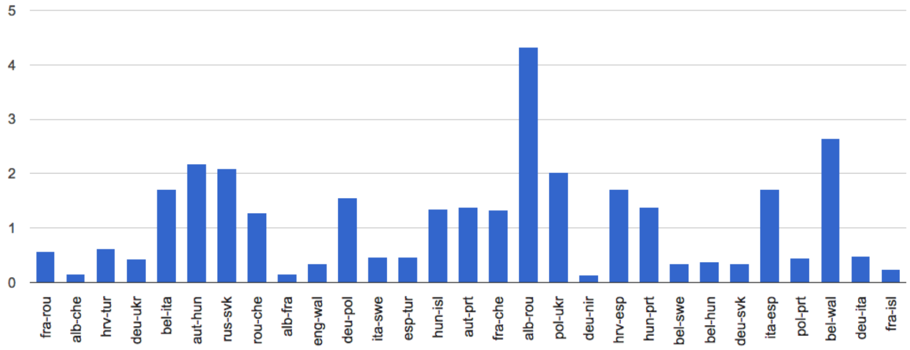

Fig. 4: Average of Logarithmic Loss for Challenge 1: after elimination of teams with less than five historical record

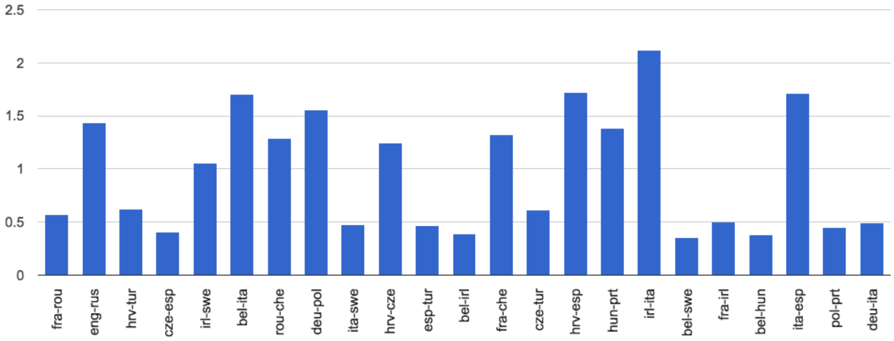

Fig. 5: Average of Logarithmic Loss for Challenge 1: after elimination of teams which qualified less than two times in Euro championship before Euro 2016

3. M. Brandt and U. Brefeld. Graph-based approaches for analyzing team interaction on the example of soccer.
4. B. Gabin, O. Camerino, M. T. Anguera, and M. Castañer. Lince: multiplatform sport analysis software. *Procedia-Social and Behavioral Sciences*, 46:4692–4694, 2012.
5. J. Haase and U. Brefeld. Mining positional data streams. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 102–116. Springer, 2014.