Feature Extraction and Aggregation for Predicting the Euro 2016

Maryam Tavakol[†], Hamid Zafartavanaelmi[‡], and Ulf Brefeld[†]

[†]Leuphana University of Lüneburg
[‡]Technical University of Darmstadt

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 train a linear model to compute scores of 24 participating countries. The obtained scores form {*win, lose, draw*} probabilities for all possible fixtures. The empirical evaluation until the semi-finals shows that the conceptually simple approach proves accurate for countries with 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. Although sports analytics has received much attention in the past few years (e.g., [5, 4, 3]), predicting the outcome of a single match (e.g., [1]) is largely understudied as it probably involves too many sources of randomness.

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*} for all possible fixtures.

We use provided the dataset which contains general information about countries as well as their players. In addition, we extract additional data from the history of all official games between countries by aggregating several data sources. The features are used together with a linear model to estimate a score for each country. The scores are then transformed into the desired probabilities for the mentioned challenge. We first describe the process of extracting features in Section 2. In Section 3, the proposed method for challenge 1 is introduced. The analysis of the results until the semi finals is presented in Section 4; Section 5 concludes.

2 Feature Extraction

The process of collecting relevant data and extracting significant features is key to the performance of any data-driven predictor. The organizer of the competition provided two data sets. The first one contains the overall ranking information for the participating countries of the Euro 2016, and the latter summarizes player specific properties. We select a subset of features from 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}

Additionally, a crawler collects the squads from the official website of UEFA and filters the **Players** data by the final list of players in every team. We then replace the number of appearances and goals of each player by their ratio, i.e., $\#goals/\#appearance$. Note that France, as the host of the tournament, has no appearance and goal data in the qualification phase. As a remedy, we use a ratio to 1 to introduce a host advantage for France. Features of different players are averaged, so that we end-up with a feature set on country-level.

Moreover, we extract extra data from a few public sources to create an auxiliary feature. We argue that if a team has more players who are playing for the same (successful) club, it is more likely that the harmony in the team leads to success of their national team as well. Thus, a list of the players' current football club together with the club ranking (top-200) is crawled from the International Federation of Football History & Statistics (IFFHS) for 2015[4]. For each country, the club with the highest number of national team players is chosen, where ties are broken by club rank. 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 the club ranking.

We aggregate the player related features such as Market Value and Age by averaging and the rest by summation. Lastly, the club rank times the number of players of the corresponding club for each national team has added to this aggregated intermediate dataset.

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

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

We then add the mean of the normalized features from 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.

The ultimate feature set is as follow: {FIFA ranking, FIFA points, UEFA ranking, UEFA coefficient, Elo ranking, Elo points, PlayerScore}

[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./

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.

3 The Proposed Approach

In this section, we present our algorithm for predicting the {*win, lose, draw*} probabilities from scores that are obtained by a linear model. These intermediate scores can be interpreted as indicators of the power of the teams in this tournament. In the remainder, we assume a linear relationship between the score of each country and the features. Let $s_i, i \in \{1, ..., 24\}$ be the score of ith country and $\mathbf{x}_i$ is the corresponding feature vector. Our linear model aims at learning a weight vector θ such that $s_i = \theta^T \mathbf{x}_i$.

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]. Unfortunately, available data were already aggregated and we only had access to each record of two countries which contains the numbers of wins, draws and losses. As a result we were not able to weight the recent data over the old ones that might be more representative of the current power of the countries The 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 desired probabilities. For instance, Italy and Sweden played 21 times against each other. Each won 6 times and they drew 9 times. The {*win, lose, draw*} probability is thus {0.28, 0.28, 0.43}. In the next part, we explain the conversion process of scores to 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 UEFA 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 until semi-finals, we can compute multi-class logarithmic loss of $\{$*win, lose, draw*$\}$ probabilities as follows,

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

where N is the number of games and M is equal to three classes as we are interested to calculate the loss for the predicted probabilities. Figure 1 summarizes the log loss error for the 51 matches. The average loss value is 1.32.

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 4 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 inconfident predictions. Figure 5 focuses on situations where more than four historical games are available; the average of logarithmic loss declines 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

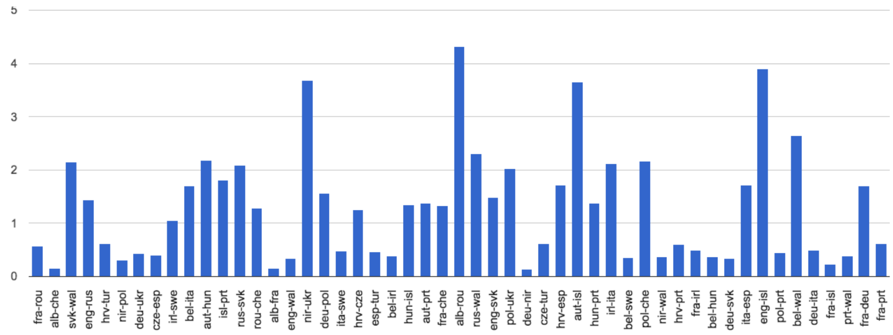

Fig. 1: Average of Logarithmic Loss for Challenge 1

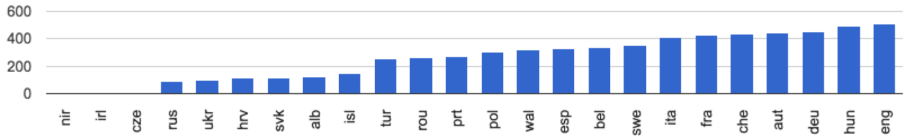

Fig. 2: The number of historical data (head-to-head) for each country

Wales did not qualify before. Austria and Ukraine had the chance to play in Euro championship just once. As shown in Figure 6, we observe a further decrease of the loss (0.9680) when we focus on pairs of teams with at least two previous appearances in Euro championship. In the presence of sufficient historical data, our approach is able to accurately predict the outcome of matches.

Additionally, we compare our prediction with a simple baseline which only takes the FIFA Ranking of the countries into account. For each pair of countries, we assign the winning probability of one for the country with the higher rank and zero for losing as well as draw. Figure 3 shows that the error of prediction for this simple strategy is very high compare to our approach which considers all the features.

5 Conclusion

We presented our solution for Euro 2016 competition. Our approach grounded on feature engineering and used linear models to predict the desired probabilities for all possible match outcomes. We showed the impact of the features on the resulting scores. We observed that accurate predictions are possible for pairs of teams that possess a long history of matches against each other while we faced cold-start-like problems for teams that hardly faced each other.

References

1. Knockout prediction. `http://trends.baidu.com/worldcup/events/knockout?locale=en`, 2014.

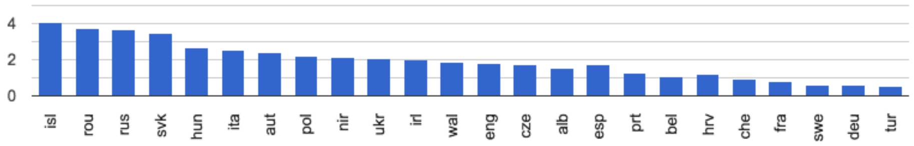

Fig. 3: Average of Logarithmic Loss of each country for Challenge 1: Sorted by loss value

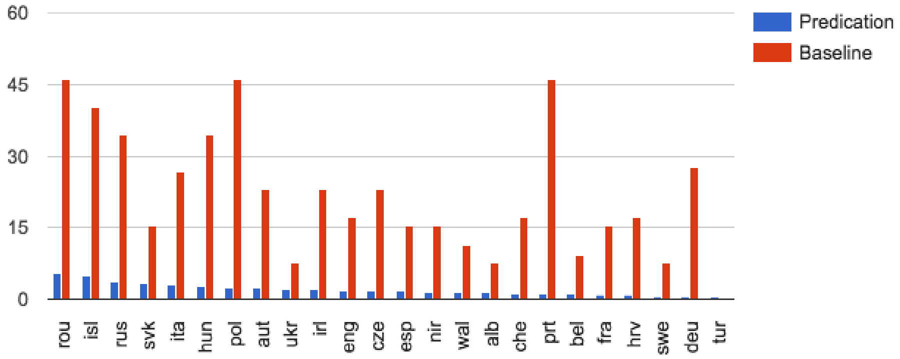

Fig. 4: Average of Logarithmic Loss of each country for Challenge 1: Compare to the Baseline

2. International football history and statistics. `www.11v11.com`, 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.

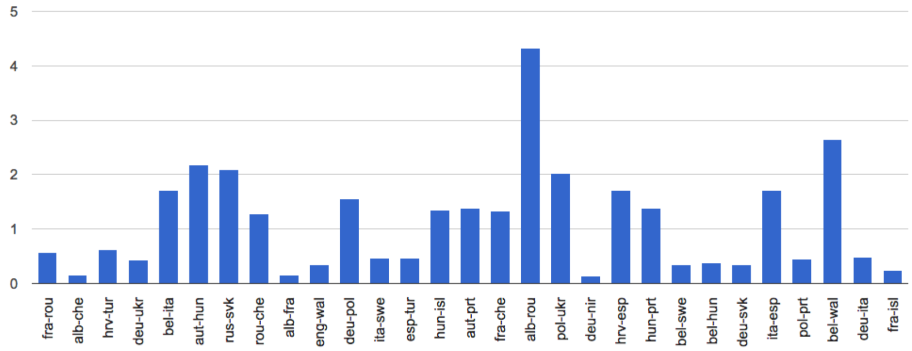

Fig. 5: Average of Logarithmic Loss for Challenge 1: after elimination of teams with less than five historical record

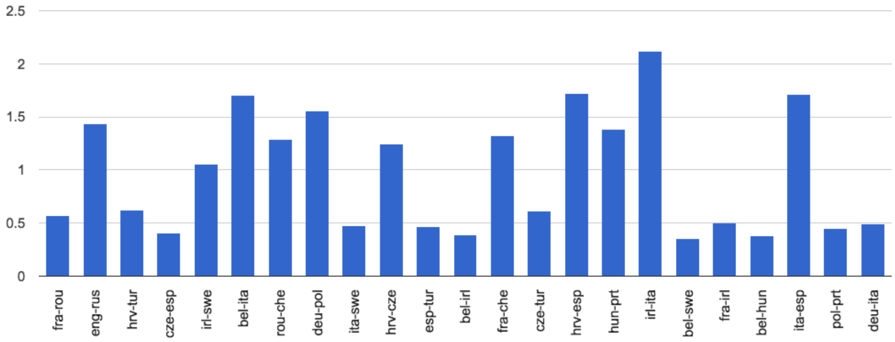

Fig. 6: Average of Logarithmic Loss for Challenge 1: after elimination of teams which qualified less than two times in Euro championship before Euro 2016