Analysing Football National Team Performance using Transfer Data from Top European Leagues

[1]Mihir Chaturvedi [2]Shivansh [3]Parth Vyas

1. Introduction

European football leagues have influenced their countries' international performance many times, although many football match predictors exist, our project takes the mentioned information in addition to other factors to determine if a country's national team is gonna win a match or not. Particularly, we aim to see if a country's major league expenditure on local players and youth translates to international competitiveness.

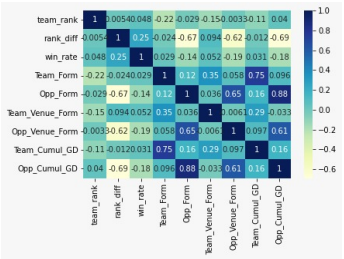
1.1. Related Works

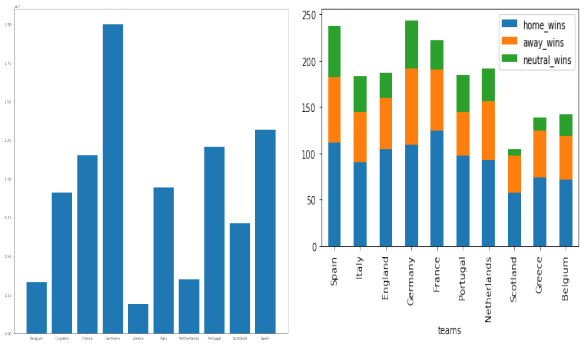
Team performances have been predicted by using several factors such as their previous performances, the most used metric is individual player performances but no related work uses the league transfers as a basis to check if the corresponding national team is gonna win or not. The most prominent work used probability conditioning in addition to random trees for drawing the world cup.

2. Dataset and Evaluation

Our initial dataset contained all the transfers from 2000-2018, but it did not contain the players' nationalities. Another dataset was used for the same to map the nationality correctly but out of the 4700 entries, only 2200 were found, so their nationalities were scraped from wikipedia. The data was narrowed down later to contain only the top 10 European leagues.

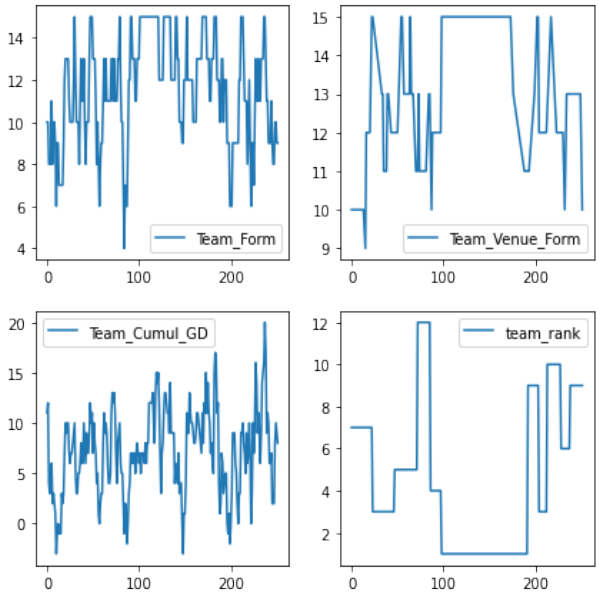
The features based on international performances were taken from two datasets (matches and shootouts) which were mapped to each other and filtered, followed by the addition of more features. And then the resulting dataset was split into 10 different parts corresponding to fixtures for an individual country





2.1. Exploratory Data Analysis

In the features extracted for international performance, a team's cumulative goal difference, their overall form, and the its form in the respective venue corresponding to the fixture (home,away,neutral), all taken over its last five matches, have relatively higher correlation. The rest of the chosen features show healthy correlation with each other.



The match distribution for each of the 10 European countries was uniform i.e, we had a balanced dataset, and we didn't require Principle Component Analysis (PCA) since the number of features were limited and convenient to visualize.

In the transfer data, we notice that the transfer fee has high correlation with the age and position of the player.

2.2. Data Preprocessing

- The features of all countries were scaled using min-max normalization
- output column was binarized (1 if the country wins, otherwise 0).

3. Analysis and Progress

3.1. Baseline Approach

For our baseline approach, we used Logistic Regression along with L1 and L2 regularization. The model was trained on the data for each of the 10 countries **separately**.

3.2. Splitting the Dataset

We applied 75% - 25% train-test split to each dataset. This is because the number of matches played by each of the 10 countries from 2000-2018 were between 200-250, and we needed an appreciable amount of data for testing.

3.3. Cross-Validation

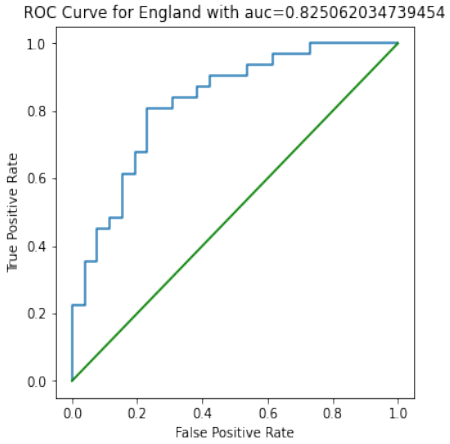
We used K-Fold Cross Validation on the training data with the number of splits being 3, due to smaller number of training samples (150-189), coming at the risk of higher bias.

3.4. Evaluation Metrics

Accuracy, Precision, Recall, F1, and ROC-AUC scores were used as evaluation metrics for each country's dataset. While training-validation, the scores were averaged across the validation sets.

Team	Accuracy	Precision	Recall	F1	ROC-AUC
Spain	0.71	0.72	0.95	0.82	0.78
Italy	0.63	0.73	0.61	0.66	0.65
England	0.77	0.76	0.84	0.80	0.82
Germany	0.58	0.64	0.76	0.70	0.64
France	0.61	0.62	0.84	0.71	0.70
Portugal	0.74	0.86	0.73	0.79	0.75
Netherlands	0.64	0.80	0.63	0.70	0.75
Scotland	0.70	0.76	0.53	0.63	0.80
Greece	0.62	0.70	0.50	0.58	0.75
Belgium	0.64	0.75	0.65	0.70	0.74

Table 1. L1 regularisation results for different countries



4. Error Analysis

After training the model without any regularization in the model, we applied L1 and L2 regularization. There was an improvement in the performance for most nations with L1 regularization compared to no regularization. Thus, our model was overfitting the data.

We compared the performance of our model for different nations. We noticed that performance of our model for nations with low win rates was not as good as compared to other nations, especially the recall scores (for countries like Greece and Scotland). Because of the low win rate, there were a higher number of false negatives reported. We also see that separability of data (wins vs losses) is poor for some countries (e.g. Germany) while it is great for England, as shown in the figure above. We are thinking about lowering the probability threshold and observing the changes to improve these.

5. Future Work

We are currently using only the statistics related to the national team performance to get the win predictions. Our next step is to incorporate the football transfer data as features and analyze its effect on the performance of our model. After that, we will try ensemble methods like random forests and advanced techniques such as neural networks to train our model and analyze any differences compared to the baseline model. We will use the error analysis techniques to gain insights into the model's failures and possible improvement solutions.

5.1. Division

Parth will be incorporating transfer-related features into our model, Mihir will try ensemble and advanced techniques and analyze the behavior, and Shivansh is responsible for the Error Analysis part.

References

[1] Andreas Groll (2018) *Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters*, Statistics Faculty, Technische Universit at Dortmund.