An Analysis of European Soccer Database Using R

Team 1
Assem, John, Joshua, Ern Chern

Research Questions

PLAYERS

What are the most important features for successful **player**?

TEAM

What are the most important features for successful **team**?

MATCH

How to predict a probability of a **victory** in a match?

INDIVIDUAL / TEAM ?

Does **individual** talent directly translate to **team** success?

Dataset

- 3 leagues (England Premier League, Germany 1. Bundesliga, Spain LIGA BBVA)
- Season 2014/15 for training data (from 2014-07-18 till 2015-05-31)
- Season 2015/16 for testing data (from 2015-07-17 till 2016-05-25)

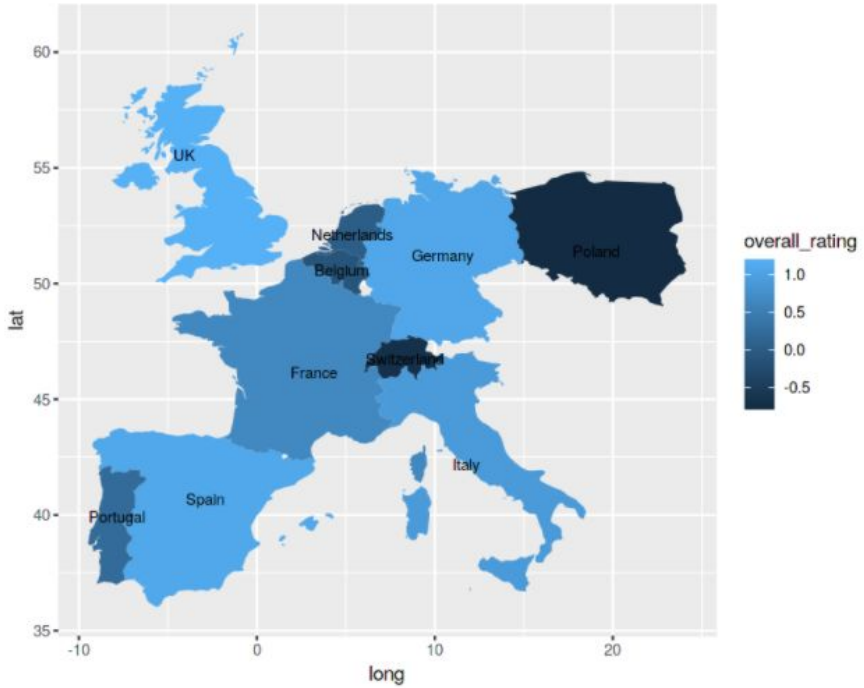
Preprocessing Data

PLAYER & TEAM

TEAM & MATCH

Remove negative values + Remove outliers (mean ± 2 sd) + Scale numeric data + Add dummy variables for categorical data + Aggregate values for same player in same year	Calculate winning rate for each team + Aggregate team with player via match + Calculate the average individual features for each team

Countries have different levels



Results: Players Feature Extraction

PLAYERS

Linear Regression

Overall rating ~ Players features

Significance level = 0.95

*colored for unique significant features for each league

England	Germany	Spain
height, heading_accuracy, long_passing, ball_control, acceleration, reactions, stamina, long_shots, aggression, positioning, gk_positioning	heading_accuracy, short_passing, long_passing, acceleration, reactions, balance, strength, positioning, gk_diving, defend_medium	heading_accuracy, short_passing, free_kick_accuracy, long_passing, ball_control, sprint_speed, reactions, strength, positioning, marking, gk_diving, gk_positioning, gk_reflexes, defend_high, defend_medium
R^2= 0.81	R^2= 0.87	R^2= 0.83

Results: Team Feature Extraction

TEAM

Linear Regression

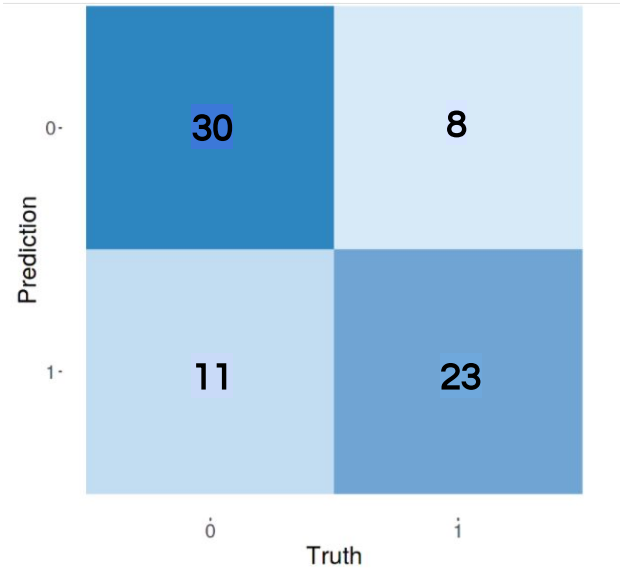
Winning rate ~ Team features

Significance level = 0.90

England	Germany	Spain
Build up play Passing 0.0135 Build up play Dribbling 0.0490 Defence pressure 0.0580	Build up play Dribbling 0.0193 Defence pressure 0.0941 Chance creation Crossing 0.0965	Defence team width 0.0905 Build up play Dribbling 0.0734
R^2= 0.73	R^2= 0.71	R^2= 0.43

Results: Match prediction

MATCH



- Logistic regression model
- Home team win ~ [Team features] + [Average player features] (after removal of correlated features)
- Accuracy for Spain: 73.6%
- HLtest: 0.6935591

Results: Individual Talent → Team Success?

INDIVIDUAL / TEAM ?

Rank each team by:

1. Average player rating
2. Win rate

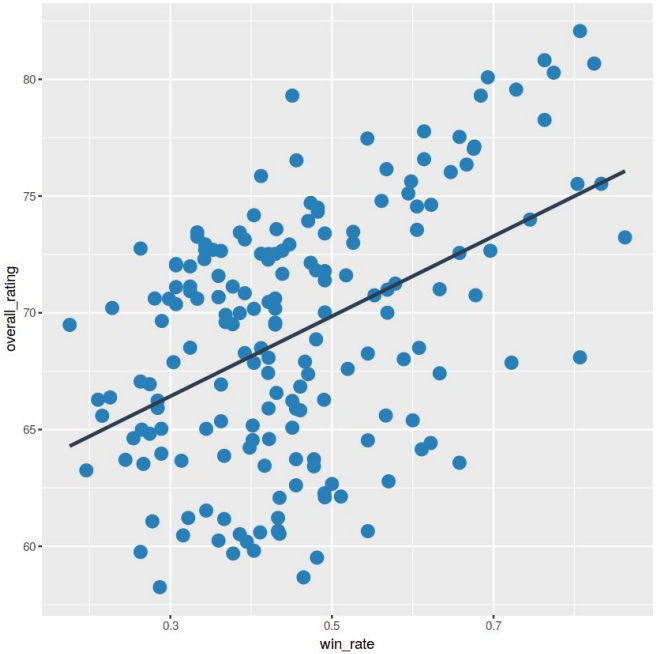
Spearman Correlation test:

- p-value is 3.141e-08
- Rho is 0.39

Weak monotonic **correlation**

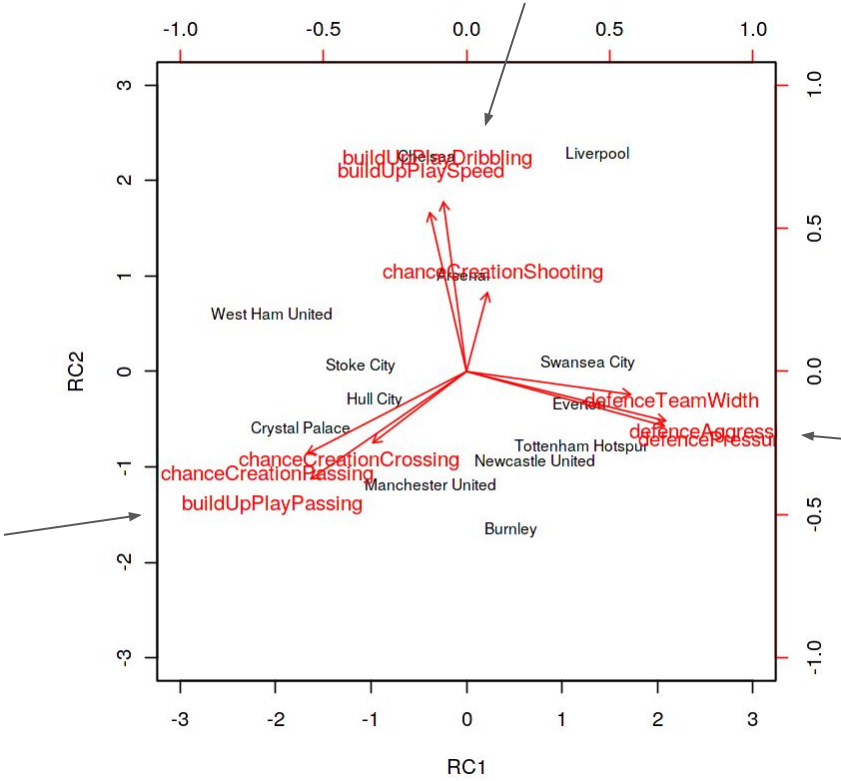
Interpretation:

Individual talent does not guarantee team success



Results: PCA on England

Individual contribution



Team work

Defence strategy

Conclusions

- Different leagues have different strategies → different requirements for player



- We can predict match results for Spain with ~74 % accuracy
- Individual talent has effect on the winning rate (≠ 0) but does not guarantee the success of the whole team

https://doi.org/10.1016/j.aci.2017.09.005

Limitation and feedback on future plan

- Prediction for England and Germany needs more in-depth curation
- We want to have more comparison between teams of different country, but this dataset does not have international match
- We would like to have a clearer evidence of teamwork, but unfortunately teamwork is very implicit
- Future work: debunk/confirm some famous assumption (home advantage, left advantage)

https://doi.org/10.1016/j.aci.2017.09.005

Thank you