Tracking and Predicting Football Performance

Team 128: Tarik Ouradi, Priya Krishnamurthy, Nidhi Anand, Jackson Duke, Matt Parsons

1. Introduction

During the FIFA World Cup 2018, an estimated $160 billion was taken in by betting sites and bookmakers from around the world[1], and with the rise in popularity of online sports betting in recent years the amount will predictably increase for the upcoming FIFA World Cup in 2022. Our goal is to create an interface which can showcase national team performance and predict the outcome of upcoming matchups based on recent performances. While there is no shortage of sports models working today, we aim to innovate by providing understandable and dynamic visualizations to accompany our model, allowing the end user to better comprehend the underlying analytics at work.

The 2022 FIFA World Cup in Qatar is estimated to be viewed by 5 billion people[2], and fans of the sport will be interested to know how their team has been performing coming into the tournament. While there is financial incentive for those who choose to bet on games to follow our application, it is also useful for everyday fans to research and follow their team.

2. Literature Survey

Most of the current academic work undertaken and published researched, utilizes Poisson regression models [3], or Markov chain Monte Carlo [4] (computationally expensive). Some use machine learning based qualitative methods, or GB machine learning technique [5]. Incorporating domain knowledge in machine learning is one of the key aspects [6]. The drawback of existing methods and their supporting datasets are that they are too complex for the public and not generally available for public such as proprietary data owned and controlled by soccer teams [7, 8, 9]. Also, some models do not consider "sophisticated" match statistics, such as fouls committed, corners conceded by each team, trajectories or relevant data about players and teams [10].

Recent studies suggest that statistical models are superior to lay and expert predictions but have less predictive power than the bookmaker odds [11]-[14]. This observation strongly suggests that either the information, used by the bookmakers, is more powerful or, alternatively, the inference process, based on the same information, is more efficient. Probably, both aspects may play a role. [11,12, 13]

Some models consider the goals scored and conceded by each team; or modelling win–draw–lose match results, they do not consider both the measurements [15]. Couple of papers focus on human behaviors especially goal keepers' [16,17], adding human bias in the mix that goalies have a behavioral-bias tendency, but the findings are based on a subset of a soccer-game environment (as opposed to a full-game scenario). Extending the focus towards the whole-game environment could augment the paper's findings and offer insights that could improve the awareness of players' non-optimal behavioral tendencies during a game. The main idea in Decroos [18] is that value latency and predictive quality has more to do with pre-goal player actions (i.e., physical movements and actions taken) than actual outcomes (e.g., goal scores). In particular, the paper describes how goals result from a chain of pre-goal events. The paper's shortcomings relate to definitions. Strategic features are confusing because "strategic" relates to a "how to plan" to achieve a set goal or suite of objectives. We inferred that strategic features were X-predictor variables.

3. Methodology and Data

3.1 Data

We collected two datasets

1. **SoccerBase Match Results** – national team results and key statistics from 2015-2022
2. **FIFA Player Ratings** – ratings across 15000+ players ranging from 0-100 in key football areas (attacking, defending, etc.) for 2015-2022

The primary dataset used to train and test our models is from SoccerBase.com, a website which hosts statistics and results for football matches in both club and national team competitions. This is a key

[1] 2018 World Cup generated €136 billion in bets worldwide, FIFA says. Casino.org. (2018, August 10). link

[2] Qatar 2022 to be watched by 5bn people, says Gianni Infantino. SportsPro. (2022, May 25). link

innovation in our project, as we are using publicly available data to keep our costs low. While there is no database or API available for this data, we leveraged Python libraries (Selenium, BeautifulSoup, and SQLite) to scrape, parse, and store the match results. Our final cleaned dataset consists of 1,372 matches across 168 national team soccer teams from 2015-2022. Bringing this data down to the player grain (one row for each player in each match), the total number of rows is 61,073 in the dataset. Please see appendix for the structure of the base dataset.

Next, we pulled player ratings from the video game FIFA for the years 2015-2022. This dataset is at the player level and has detailed ratings for attacking, defending, passing characteristics and more. While we acknowledge these ratings are subjective, we believe that providing our model with the average "strength" of a squad in various categories will give context to whether a strong team is expected to beat a less strong team. This is an innovation in our project, since in our research we have not seen player ratings as input. Please see appendix for the structure of the FIFA dataset.

We then aggregated the 2 datasets into one final dataset by:

1. Averaging the ratings from the FIFA ratings at the player level to the squad level
2. Performing weighted averages of match statistics from the SoccerBase results to give a look at recent performance (e.g. Team A Weighted Average Possession)

Attribute	**Grain**	**Description**
Match Date	Match	Date of the match between Team A and Team B.
Match Comp	Match	Competition in which the match took place
Team Name	Team	Name of Team (e.g. "Brazil")
Team Score	Team	Weighted average goals scored by the team in the last 3 matches
Team Possession	Team	Weighted average percent of time that the team had the ball in the last 3 matches
Team Shots On	Team	Weighted average number of shots on target for the team in the last 3 matches
Team Shots Off	Team	Weighted average number of shots off target for the team in the last 3 matches
Team Corners	Team	Weighted average number of corner kicks for the team in the last 3 matches
Avg Overall	Team	Average player "Overall" rating for the team in the match
Avg Crossing	Team	Average player "Crossing" rating for the team in the match
Avg ShotPower	Team	Average player "ShotPower" rating for the team in the match
Avg Dribbling	Team	Average player "Dribbling" rating for the team in the match
Avg SprintSpeed	Team	Average player "SprintSpeed" rating for the team in the match
Avg Tackle	Team	Average player "Tackle" rating for the team in the match
…	…	…
Result	Match	Win/Loss/Draw column for the team in the match (target y-metric for model)

Figure 1. Layout of final dataset.

4. Model Evaluation

For our model evaluation, we focused on comparing a range of key performance metrics across each model. Specifically, after constructing each model (based on training data), we evaluated the relative models' performances by utilizing validation/test data.

For model construction, we selected ~ circa 70% of our data for model training (i.e., building models). The remaining 30% of our data was then used for model validation/testing (to understand how the model would generalize to new data).

Once the best-performing model was identified, we selected that model as our final model to apply to new data for making predictions.

The key performance metrics we reviewed for model-evaluation purposes were as follows:

1. **Model accuracy**: which is the ratio of the number of times the predicted value (e.g., predicted game win or loss) equals the actual value (actual game win or loss) to the total sample values:
    - True positives (TP) + true negatives (TN) / (true positives (TP) + true negatives (TN) + false positives (FP) + false negatives (FN)).
2. **Model sensitivity (recall)**: this measure calculates the proportion of game wins that are correctly predicted. The formula is:
    - True positives (TP) / (true positives (TP) + false negatives (FN)).
3. **Model precision**: this evaluates the number of noisy positives (FP) that are in the model. It's defined as:
    - True positives (TP) / (true positives (TP) + false positives (FP)).

The above evaluation framework describes a **confusion matrix**, which can be used to evaluate **categorical** responses. Ultimately, the measure we focused on most was **accuracy** because accuracy was the most balanced performance metric in predicting across all three categories: Win, Loss, and Draw. Additionally, we didn't prefer adding more significant weightings to either true positives or true negatives. Therefore, the most meaningful performance metric was **accuracy**.

4.1 Model Evaluation Results:

Five models were ultimately developed: Random Forrest, Decision Tree (CART), K-Nearest Neighbors (kNN), Gradient Boosted Classifier, and AdaBoost Classifier.

As mentioned in the above methodology, our models were developed using training data. Then, using validation/test data, we were able to see which models performed best relative to each other. The best performing models were the **AdaBoost Classifier** and **Gradient Boosted Classifier** (using **1 – test accuracy** as the definition for model test error). The model results are as follows:

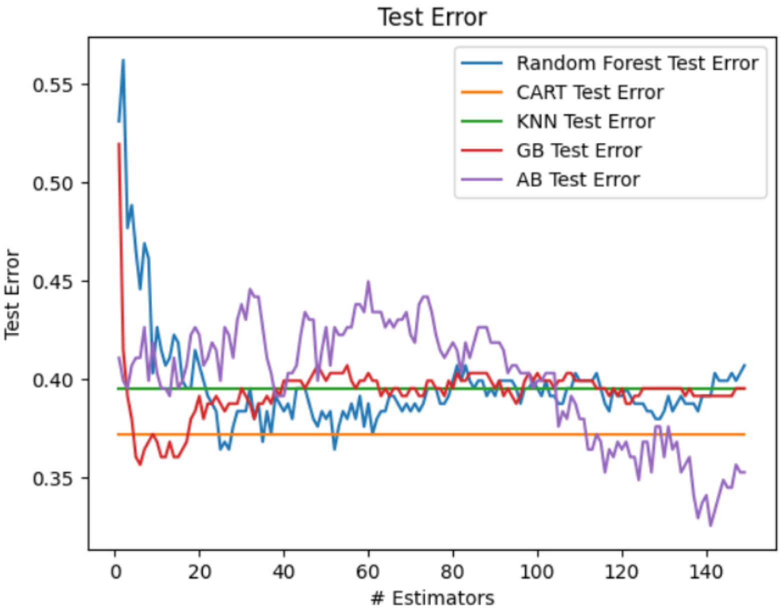


Figure 2. Model testing error as a function of number of estimators.

As can be seen from the tables below, the best performing models were:

1. AdaBoost Classifier (test accuracy: 67.44%); and
2. Gradient-Boosted Classifier (test accuracy: 64.34%).

The detailed results of all models based on training and validation/test data are as follows (results are based on the confusion matrix of each model using training data):

Random Forrest:

TRAINING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**263**	0	0
	Draw	0	**149**	0
	Loss	0	0	**190**

Acc.		Sens.	Spec.
100.0%	Win	100.0%	100.0%
	Draw	100.0%	100.0%
	Loss	100.0%	100.0%

TESTING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**99**	12	13
	Loss	28	**15**	21
	Draw	10	10	**50**

Acc.		Sens.	Spec.
63.57%	Win	79.84%	71.64%
	Draw	23.44%	88.66%
	Loss	71.43%	81.91%

Decision Tree (CART):

TRAINING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**243**	5	15
	Draw	84	**22**	43
	Loss	49	7	**134**

Acc.		Sens.	Spec.
66.28%	Win	92.40%	60.77%
	Draw	14.77%	97.35%
	Loss	70.53%	85.92%

TESTING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**110**	5	9
	Draw	36	**5**	23
	Loss	20	3	**47**

Acc.		Sens.	Spec.
62.79%	Win	88.71%	58.21%
	Draw	07.81%	95.88%
	Loss	67.14%	82.98%

K-Nearest Neighbors (kNN):

TRAINING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**213**	1	49
	Draw	82	**2**	65
	Loss	48	3	**139**

Acc.		Sens.	Spec.
58.80%	Win	80.99%	72.33%
	Draw	01.34%	99.12%
	Loss	73.16%	61.65%

TESTING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**98**	0	26
	Draw	40	**1**	23
	Loss	13	0	**57**

Acc.		Sens.	Spec.
60.47%	Win	79.03%	60.45%
	Draw	01.56%	100.0%
	Loss	81.43%	73.94%

Gradient Boosted Classifier:

TRAINING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**253**	0	10
	Draw	94	**34**	21
	Loss	43	1	**146**

Acc.		Sens.	Spec.
71.93%	Win	96.20%	59.59%
	Draw	22.82%	99.78%
	Loss	76.84%	92.48%

TESTING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**112**	4	8
	Draw	35	**6**	23
	Loss	16	6	**48**

Acc.		Sens.	Spec.
64.34%	Win	90.32%	61.94%
	Draw	09.38%	94.85%
	Loss	68.57%	83.51%

AdaBoost Classifier:

TRAINING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**242**	20	1
	Draw	32	**107**	10
	Loss	1	17	**172**

TESTING

		Prediction		
		Win	**Draw**	**Loss**
True	**Win**	**98**	21	5
	Draw	25	**28**	11
	Loss	6	16	**48**

Acc.		Sens.	Spec.
86.54%	Win	92.02%	90.27%
	Draw	71.81%	91.83%
	Loss	90.53%	97.33%

Acc.		Sens.	Spec.
67.44%	Win	79.03%	76.87%
	Draw	43.75%	80.93%
	Loss	68.57%	91.49%

Based on **accuracy** as our relative performance measure for comparing models, the best performing model (based on training data) was the **Random Forrest**, followed by the **AdaBoost Classifier**.

However, comparing models based on training data is not the best approach for model selection because training data tends to overestimate model performance. The reason for this overperformance tendency is due to trained models fitting random effects as though those random effects are real effects. What this means is that training data results in a higher degree of overfitting relative to models that use new data (e.g., validation/test data). When models overfit data, they tend to not generalize well to new data. Specifically, overfit models have issues with high variance.

To that end, comparing based on validation/test data the best performing models were:

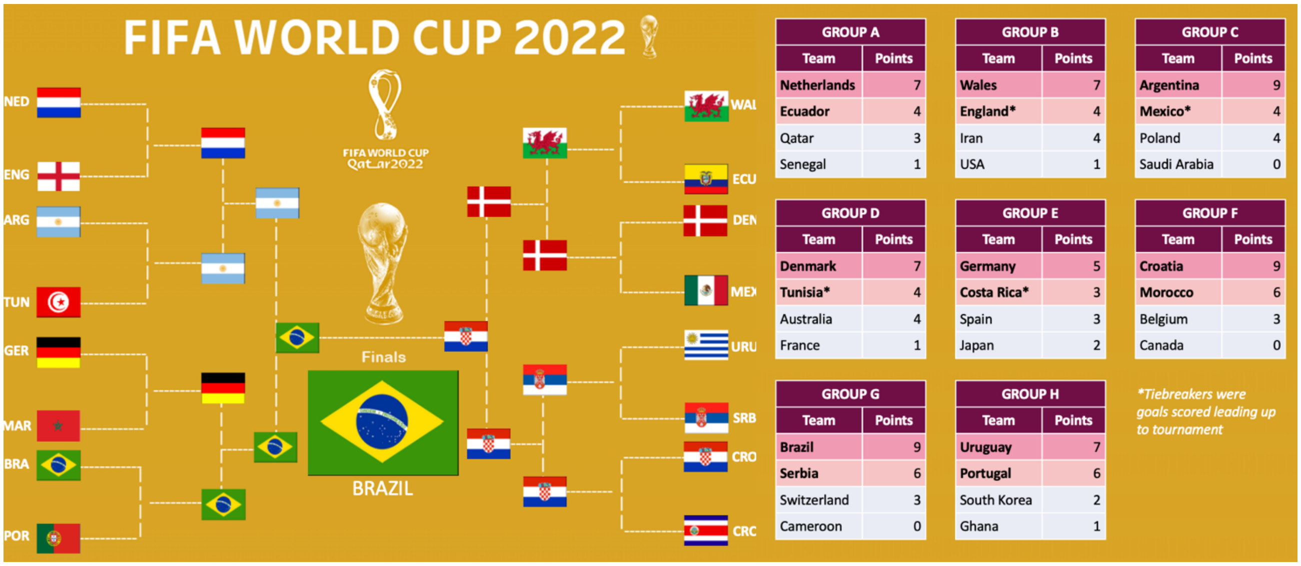
- AdaBoost Classifier (accuracy: 67.44%); and
- Gradient Boosted Classifier (accuracy 64.34%).

Based on validation/test data we selected the **AdaBoost Classifier** as our final model because it resulted in highest accuracy on validation/test data. The feature importance in our final AdaBoost Classifier model is as follows:

Feature	Importance	Description
B_roll_team_score	6.38%	Weighted average goals scored from previous 3 matches
A_roll_team_score	5.67%	Weighted average goals scored from previous 3 matches
A_avg_finishing	4.26%	Average finishing rating for Team A
A_avg_def_awareness	4.26%	Average defensive awareness rating for Team A
B_avg_age	3.55%	Average age in years of Team B
B_roll_team_shots_off	3.55%	Weighted average shots off target from previous 3 matches
B_avg_penalties	3.55%	Average penalty rating for Team B
A_avg_shotPower	2.84%	Average show power rating for Team A
B_avg_reactions	2.84%	Average reaction score for Team B
B_avg_strength	2.84%	Average strength score for Team B

5. Visualizations

5.1 Tournament View – summary view of the predictions our top-performing model



GROUP A	
Team	Points
Netherlands	7
Ecuador	4
Qatar	3
Senegal	1

GROUP B	
Team	Points
Wales	7
England*	4
Iran	4
USA	1

GROUP C	
Team	Points
Argentina	9
Mexico*	4
Poland	4
Saudi Arabia	0

GROUP D	
Team	Points
Denmark	7
Tunisia*	4
Australia	4
France	1

GROUP E	
Team	Points
Germany	5
Costa Rica*	3
Spain	3
Japan	2

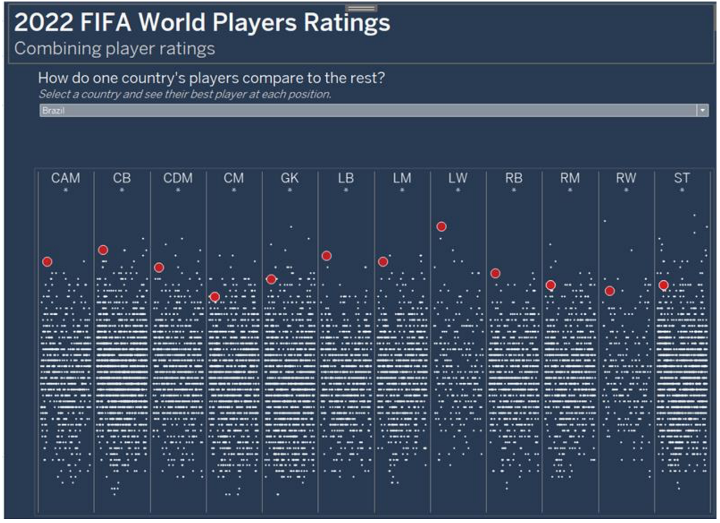
GROUP F	
Team	Points
Croatia	9
Morocco	6
Belgium	3
Canada	0

GROUP G	
Team	Points
Brazil	9
Serbia	6
Switzerland	3
Cameroon	0

GROUP H	
Team	Points
Uruguay	7
Portugal	6
South Korea	2
Ghana	1

*Tiebreakers were goals scored leading up to tournament

5.2 Team Detailed View – view with dropdown, detailing the world cup team's strength by position



2022 FIFA World Players Ratings

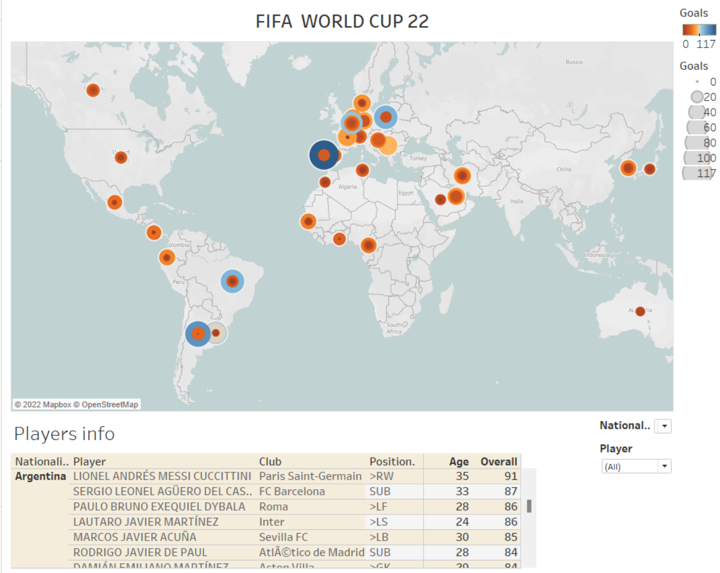
Combining player ratings

How do one country's players compare to the rest?

Select a country and see their best player at each position.

Brazil

5.3 World Map View – view showing how one country's goal scoring prowess compares to its opponents



FIFA WORLD CUP 22

Players info

Nationali..	Player	Club	Position.	Age	Overall
Argentina	LIONEL ANDRÉS MESSI CUCCITTINI	Paris Saint-Germain	>RW	35	91
	SERGIO LEONEL AGÜERO DEL CAS..	FC Barcelona	SUB	33	87
	PAULO BRUNO EXEQUIEL DYBALA	Roma	>LF	28	86
	LAUTARO JAVIER MARTÍNEZ	Inter	>LS	24	86
	MARCOS JAVIER ACUÑA	Sevilla FC	>LB	30	85
	RODRIGO JAVIER DE PAUL	AtlÃ©tico de Madrid	SUB	28	84
	DAMIÁN EMILIANO MARTÍNEZ	Aston Villa	>GK	29	84

6. Conclusion

With the public data available on FIFA players and matches, we were able to predict winner of the FIFA game with 67% accuracy with AdaBoost Classifier. The tableau report of the world map displays how each country stack up against other countries and how each country's players stack up against their counterparts of other countries.

Future Improvements: while this model trained using the data available before the tournament started, we believe the model would produce more accurate results if:

- We run the predictions after each game.
- Instead of full roster, use only the players' stats who are on the field, which can change any time due to sub-outs or injury etc.
- Using rolling averages from various time frames (5 games, 10 games, etc.)
- Imputing stats for players who were not in the FIFA database.
- Enhance UI to present the model predictions real time (refresh on user input real time).
- Enhance UI to add or remove players at run time and predict games with new team roster.

All members have contributed similar effort towards this project work.

Appendix A: Dataset Layouts

Attribute	Grain	Description
Match Date	Match	Date of the match between Team A and Team B.
Match Competition	Match	Competition in which the match took place (e.g. World Cup, International Friendly, Euro Cup)
Team A Name	Team	Name of Team A (e.g. “Brazil”)
Team A Score	Team	Goals scored by Team A in the match
Team A Possession	Team	Percent of time that the Team A had the ball in the match
Team A Shots On	Team	Number of shots on target for goal for the Team A in the match
Team A Shots Off	Team	Number of shots off target for goal for the Team A in the match
Team A Corners	Team	Number of corner kicks for the Team A in the match
Team B Name	Team	Name of Team B (e.g. “Brazil”)
Team B Score	Team	Goals scored by Team B in the match
Team B Possession	Team	Percent of time that the Team B had the ball in the match
Team B Shots On	Team	Number of shots on target for goal for the Team B in the match
Team B Shots Off	Team	Number of shots off target for goal for the Team B in the match
Team B Corners	Team	Number of corner kicks for the Team B in the match
Player Name	Player	Full name of the player in the match
Player Team	Player	Team which the player belongs to (either Team A or Team B)
Player Role	Player	Starter/Reserve
Subbed Minute	Player	Minute of the game when the player was subbed on/off
Yellow Card	Player	“1” if the player received a yellow card for a foul in the match
Red Card	Player	“1” if the player received a red card for a foul in the match

Figure A1. SoccerBase Dataset Layout.

Attribute	Grain	Description
Name	Player	Full name of the player
Overall	Player	Overall quality rating of the player
Crossing	Player	Rating of the player’s cross passing ability
ShotPower	Player	Rating of how powerful the player’s shot is
Dribbling	Player	Rating of how skilled the player is a dribbling the ball
SprintSpeed	Player	Rating of how fast the player can run at full speed
Tackle	Player	Rating of how strong the player is at tackling (defending)
…	…	…

Figure A2. Example Features from FIFA Player Dataset Layout.

References

1. Eryarsoy, E., & Delen, D. Predicting the outcome of a football game, 2009.
2. Qatar 2022 to be watched by 5bn people, says Gianni Infantino. SportsPro. (2022, May 25)
3. S. J. Koopman, R. Lit, “A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League”, Journal of the Royal Statisitical Society: Series A, Vol 178, Issue 1, pp.167- 186.
4. Martin Crowder, Mark Dixon, Anthony Ledford, Mike Robinson, Dynamic modelling and prediction of English Football League matches for betting, 20 June 2002
5. YoonjaeCho, JaewoongYoon, Sukjun Lee, Using social network analysis and gradient boosting to develop a soccer win–lose prediction model, Engineering Applications of Artificial Intelligence, Volume 72, June 2018, Pages 228-240
6. Daniel Berrar, Philippe Lopes , Werner Dubitzky Incorporating domain knowledge in machine learning for soccer outcome prediction, 2018, Machine Learning (2019) 108:97–126
7. P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews. Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In 9th Annual MIT Sloan Sports Analytics Conference, 2015
8. J. Fernandez and L. Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In MIT Sloan Sports Analytics Conference, 2018
9. P. Power, H. Ruiz, X. Wei, and P. Lucey. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In ACM SIGKDD, 2017
10. Julen Castellano., David Casamichana and Carlos Lago, The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams, 2012 Apr 3
11. Hvattum LM, Arntzen H (2010) Using elo ratings for match result prediction in association football. International Journal of Forecasting 26: 460–470.
12. Andersson P, Edman J, Ekman M (2005) Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. International Journal of Forecasting 21: 565–576.
13. Song CU Boulier BL, Stekler HO (2007) The comparative accuracy of judgmental and model forecasts of american football games. International Journal of Forecasting 23: 405–413.
14. Forrest D, Simmons R (2000) Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting 16: 317–331.
15. JohnGoddard, Regression models for forecasting goals and match results in association football, International Journal of Forecasting, Volume 21, Issue 2, April–June 2005, Pages 331-340
16. Michael Bar-Eli, Ofer H. Azar, Ilana Ritov, Yael Keidar-Levin, and Galit Schein. Action bias among elite soccer goalkeepers: The case of penalty kicks. Journal of Economic Psychology, 28(5), 2007, 606–621
17. P.-A. Chiappori, Steven Levitt, and Timothy Groseclose. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. American Economic Review, 2002, 1138–1151
18. T. Decroos, L. Bransen, J. Van Haaren, and J. Davis. Actions speak louder than goals: Valuing player actions in soccer. In the 25th ACM SIGKDD, 2019, pp. 1851–1861.