EPP: Elo-based predictive power score

Alicja Gosiewska
COSEAL, Potsdam, 26 August 2019

About me

Alicja Gosiewska

PhD candidate in Computer Science

- Explainable Artificial Intelligence (XAI)
- Automated Machine Learning (AutoML)

MSc in Mathematics (Mathematical Statistics and Data Analysis)

Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

MI

MI^2 Data Lab

MI

EPP: INTERPRETABLE SCORE OF MODEL PREDICTIVE POWER

A PREPRINT

Alicja Gosiewska
Faculty of Mathematics and Information Science
Warsaw University of Technology
alicjagosiewska@gmail.com
https://orcid.org/0000-0001-6563-5742

Mateusz Bąkała
Faculty of Mathematics and Information Science
Warsaw University of Technology

Katarzyna Woźnica
Faculty of Mathematics and Information Science
Warsaw University of Technology

Maciej Zwoliński
Faculty of Mathematics and Information Science
Warsaw University of Technology

Przemysław Biecek
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
Faculty of Mathematics and Information Science
Warsaw University of Technology
przemyslaw.biecek@gmail.com
https://orcid.org/0000-0001-8423-1823

Available on arXiv from Tuesday.

What is wrong with AUC?

Weaknesses of most popular measures 1/2

1. What is the interpretation of a difference in performance for the two models?

	score
team	
Erkut & Mark,Google AutoML	0.618492
Erkut & Mark	0.616913
Google AutoML	0.615982
Erkut & Mark,Google AutoML,Sweet Deal	0.615858
Sweet Deal	0.615766

source: https://www.kaggle.com/antgoldbloom/analyzing-kaggledays-sf-competition-data

MI

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

#	Team Name	Score
1	alijs	0.9562
2	7777777777777777777777777...	0.9559
3	ML Keksika	0.9546
4	krivoship	0.9544
5	2 old mipt dogs	0.9543

IEEE-CIS Fraud Detection
https://www.kaggle.com/c/ieee-fraud-detection/leaderboard

Springleaf Marketing Response

#	△pub	Team Name	Score
1	—	Asian Ensemble	0.80426
2	▲1	.baGGaj.	0.80393
3	▲1	Merging the Mundane and th...	0.80389
4	▼2	ARG eMMSamble	0.80367
5	—	n_m	0.80208

Springleaf Marketing Response
https://www.kaggle.com/c/springleaf-marketing-response/leaderboard

MI

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

diff = 0.0003

#	Team Name	Score
1	alijs	0.9562
2	7777777777777777777777777...	0.9559
3	ML Keksika	0.9546
4	krivoship	0.9544
5	2 old mipt dogs	0.9543

IEEE-CIS Fraud Detection
https://www.kaggle.com/c/ieee-fraud-detection/leaderboard

Springleaf Marketing Response

diff = 0.0003

#	△pub	Team Name	Score
1	—	Asian Ensemble	0.80426
2	▲1	.baGGaj.	0.80393
3	▲1	Merging the Mundane and th...	0.80389
4	▼2	ARG eMMSamble	0.80367
5	—	n_m	0.80208

Springleaf Marketing Response
https://www.kaggle.com/c/springleaf-marketing-response/leaderboard

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

diff = 0.0003

#	Team Name	Score
1	alijs	0.9562
2	77777777777777777777777...	0.9559
3	ML Keksika	0.9546

Springleaf Marketing Response

diff = 0.0003

#	△pub	Team Name	Score
1	—	Asian Ensemble	0.80426
2	▲1	.baGGaj.	0.80393
3	▲1	Merging the Mundane and th...	0.80389

Is 0.0003 the same increase for both data sets?

MI

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

diff = 0.0003

#	Team Name	Score
1	**alijs**	0.9562
2	**77777777777777777777777...**	0.9559
3	**ML Keksika**	0.9546

Springleaf Marketing Response

diff = 0.0003

#	∆pub	Team Name	Score
1	—	**Asian Ensemble**	0.80426
2	▲1	**.baGGaj.**	0.80393
3	▲1	**Merging the Mundane and th...**	0.80389

- The gaps are almost **the same for both** data sets, because the differences in AUC are almost similar.

MI

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

diff = 0.0003

#	Team Name	Score
1	**alijs**	0.9562
2	**7777777777777777777777777...**	0.9559
3	**ML Keksika**	0.9546

Springleaf Marketing Response

diff = 0.0003

#	∆pub	Team Name	Score
1	—	**Asian Ensemble**	0.80426
2	▲ 1	**.baGGaj.**	0.80393
3	▲ 1	**Merging the Mundane and th...**	0.80389

- The gaps are almost **the same for both** data sets, because the differences in AUC are almost similar.
- The gap in the IEEE-CIS Fraud Competition is larger as AUC is closer to 1.
  Therefore, relative improvement for **IEEE-CIS is larger** than relative improvement for Springleaf.

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

diff = 0.0003

#	Team Name	Score
1	**alijs**	0.9562
2	**7777777777777777777777...**	0.9559
3	**ML Keksika**	0.9546

Springleaf Marketing Response

diff = 0.0003

#	△pub	Team Name	Score
1	—	**Asian Ensemble**	0.80426
2	▲1	**.baGGaj.**	0.80393
3	▲1	**Merging the Mundane and th...**	0.80389

- The gaps are almost **the same for both** data sets, because the differences in AUC are almost similar.
- The gap in the IEEE-CIS Fraud Competition is larger as AUC is closer to 1.
  Therefore, relative improvement for **IEEE-CIS is larger** than relative improvement for Springleaf.
- Improvement for **Springleaf is larger** than for IEEE-CIS.
  The gap between first and second place for Springleaf is larger that the difference between the second and the third place. The opposite is true for IEEE-CIS Fraud detection.

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

#	Team Name	Score
1	alijs	0.9562
2	77777777777777777777777...	0.9559
3	ML Keksika	0.9546

diff = 0.0003

Springleaf Marketing Response

#	△pub	Team Name	Score
1	—	Asian Ensemble	0.80426
2	▲ 1	.baGGaj.	0.80393
3	▲ 1	Merging the Mundane and th...	0.80389

diff = 0.0003

diff = 0.00004

- The gaps are almost **the same for both** data sets, because the differences in AUC are almost similar.
- The gap in the IEEE-CIS Fraud Competition is larger as AUC is closer to 1.
  Therefore, relative improvement for **IEEE-CIS is larger** than relative improvement for Springleaf.
- Improvement for **Springleaf is larger** than for IEEE-CIS.
  The gap between first and second place for Springleaf is larger that the difference between the second and the third place. The opposite is true for IEEE-CIS Fraud detection.

Weaknesses of most popular measures 2/2

2. How to compare performances of models between data sets?

IEEE-CIS Fraud Detection

#	Team Name	Score
1	**alijs**	0.9562
2	**7777777777777777777777777...**	0.9559
3	**ML Keksika**	0.9546

diff = 0.0003

diff = 0.0013

Springleaf Marketing Response

#	∆pub	Team Name	Score
1	—	**Asian Ensemble**	0.80426
2	▲ 1	**.baGGaj.**	0.80393
3	▲ 1	**Merging the Mundane and th...**	0.80389

diff = 0.0003

diff = 0.00004

- The gaps are almost **the same for both** data sets, because the differences in AUC are almost similar.
- The gap in the IEEE-CIS Fraud Competition is larger as AUC is closer to 1.
  Therefore, relative improvement for **IEEE-CIS is larger** than relative improvement for Springleaf.
- Improvement for **Springleaf is larger** than for IEEE-CIS.
  The gap between first and second place for Springleaf is larger that the difference between the second and the third place. The opposite is true for IEEE-CIS Fraud detection.

EPP:
Elo-based Predictive Power performance score

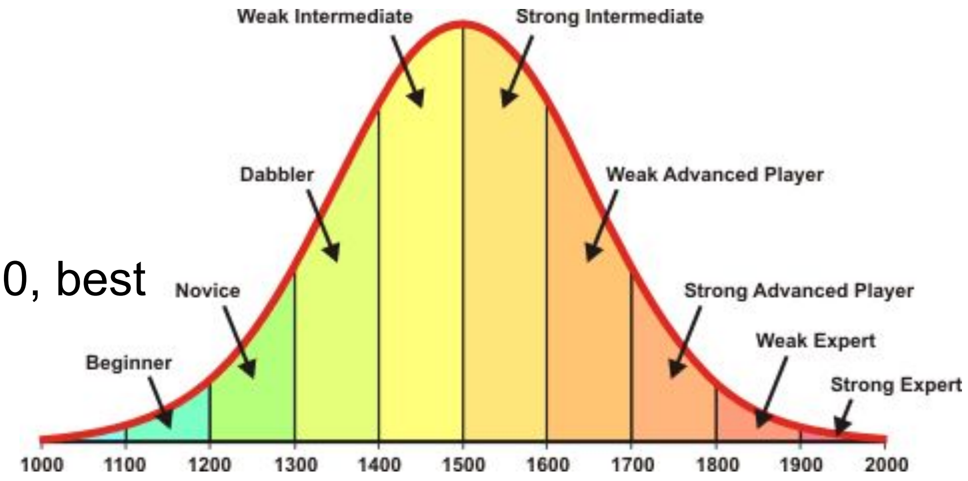
ELO rating system

https://www.365chess.com/players/Garry_Kasparov

http://www.stationgossip.com/2017/08/the-history-of-football-100-pics_9.html

ELO

- **Meaningful values.**
  An Average player have a rating of 1500, best players obtain rating over 2000.

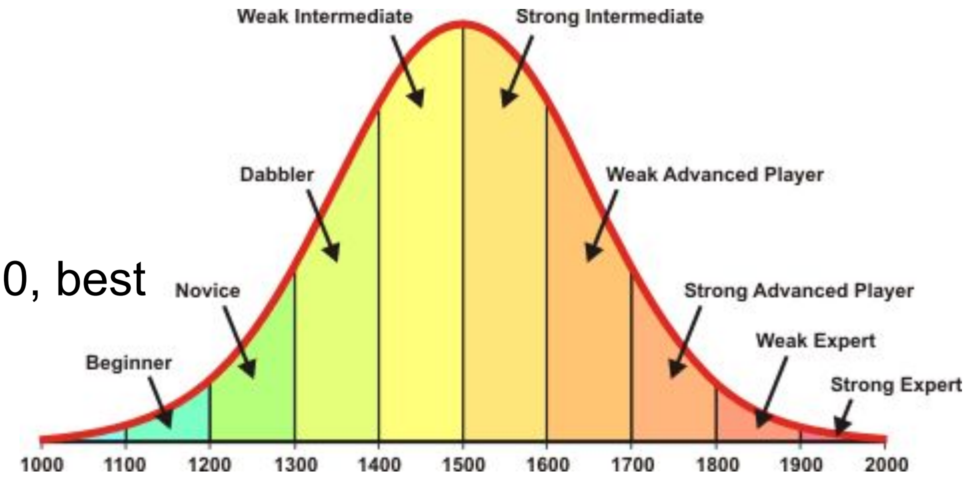


https://bkgm.com/faq/Ratings.html

ELO

- **Meaningful values.**
  An Average player have a rating of 1500, best players obtain rating over 2000.

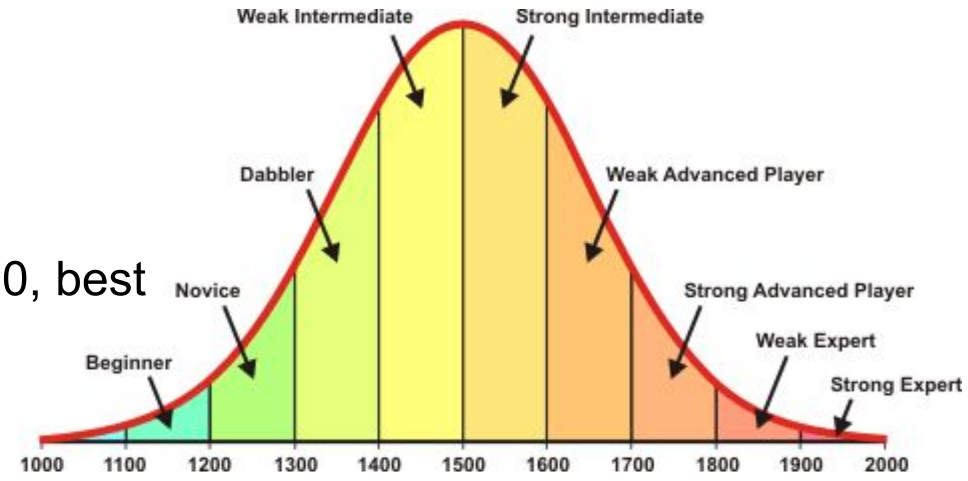
- **Probabilistic interpretation.**
  The difference between Elo scores of two players can be transferred into probabilities of winning when they play against each other.



https://bkgm.com/faq/Ratings.html

ELO

- **Meaningful values.**
  An Average player have a rating of 1500, best players obtain rating over 2000.



https://bkgm.com/faq/Ratings.html

- **Probabilistic interpretation.**
  The difference between Elo scores of two players can be transferred into probabilities of winning when they play against each other.

- **Partial results are enough.**
  It is not necessary for each player to play with each other player.

Elo-based Predictive Power score (EPP)

- There is an interpretation of differences in performance.

$$\text{diff} = EPP_A - EPP_B$$

Elo-based Predictive Power score (EPP)

- There is an interpretation of differences in performance.

$$\text{diff} = EPP_A - EPP_B$$

$$P\left(\begin{array}{l}\text{A achieves better} \\ \text{performance than B}\end{array}\right) = invlogit(\text{diff}) = \frac{e^{\text{diff}}}{1 + e^{\text{diff}}}$$

Elo-based Predictive Power score (EPP)

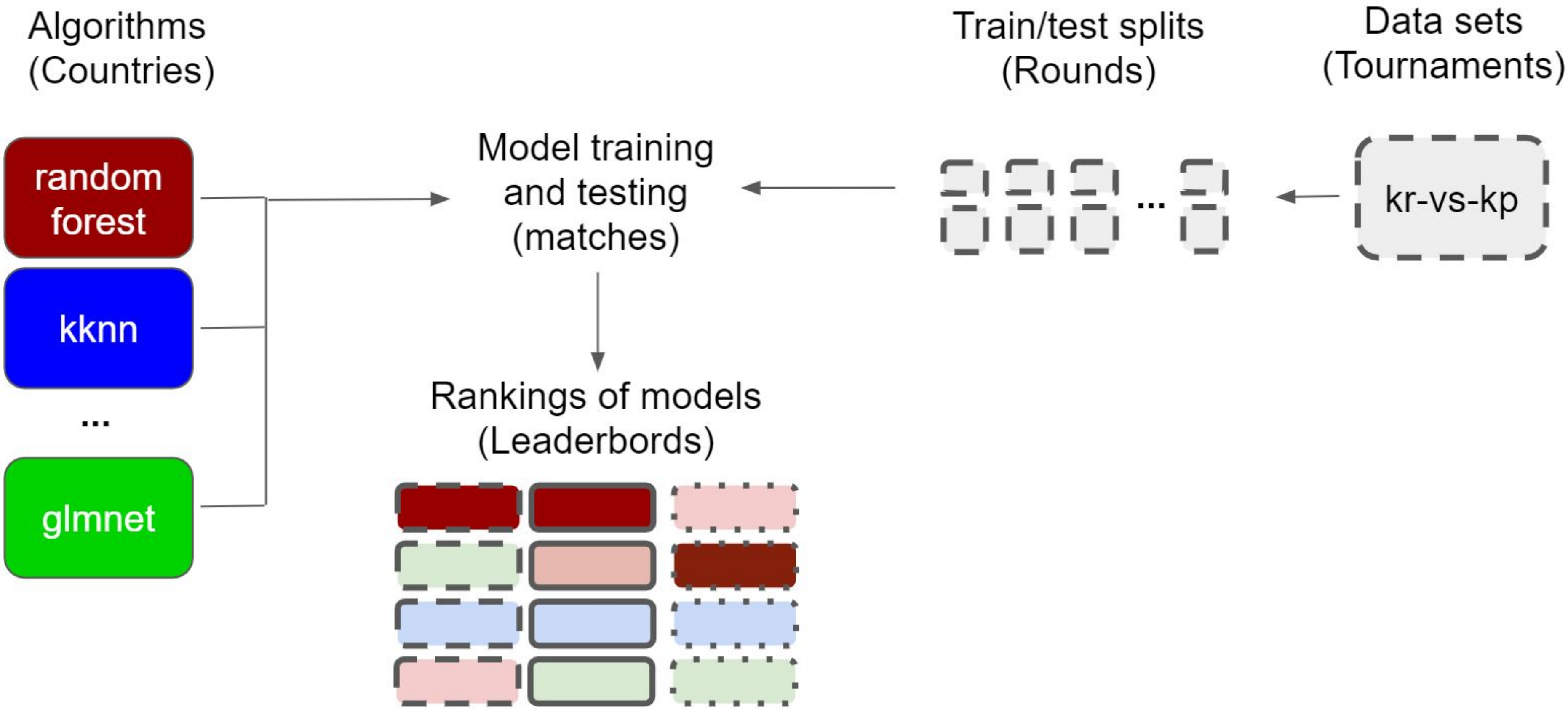
- There is an interpretation of differences in performance.

$$\text{diff} = EPP_A - EPP_B$$

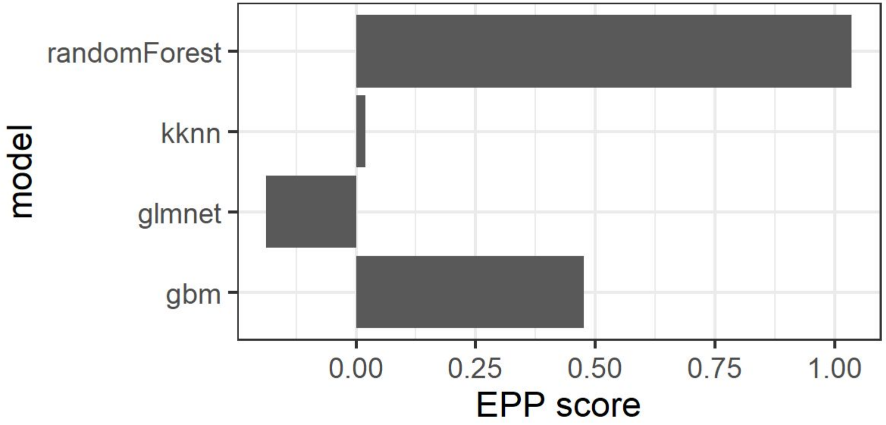
$$P\left(\begin{array}{c}\text{A achieves better} \\ \text{performance than B}\end{array}\right) = invlogit(\text{diff}) = \frac{e^{\text{diff}}}{1 + e^{\text{diff}}}$$

- One can compare performances between data sets.

The analogy between Elo and EPP

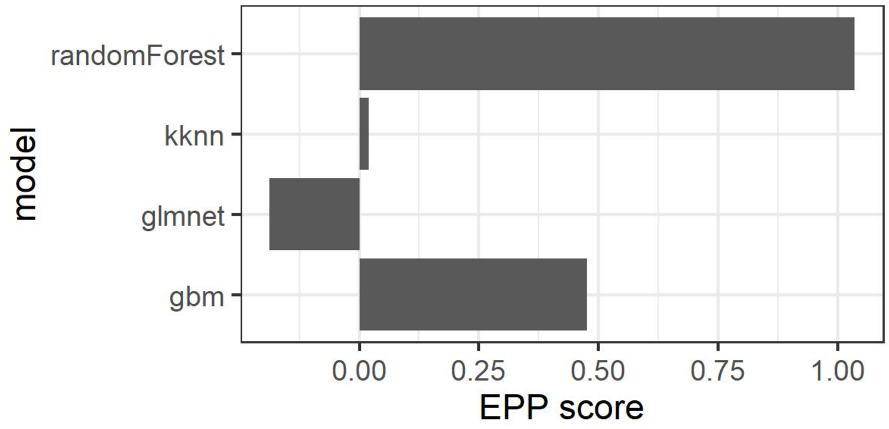


EPP scores are interpretable!



Model	EPP
randomForest	1.03
kknn	0.0195
glmnet	-0.187
gbm	0.476

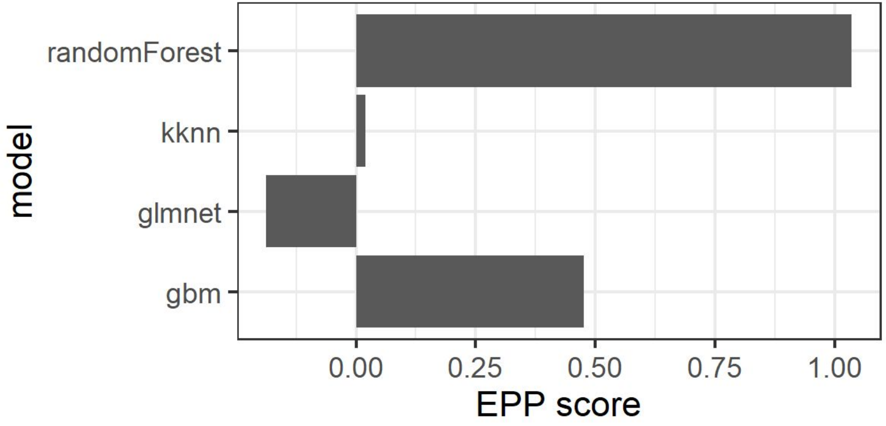
EPP scores are interpretable!



Model	EPP
randomForest	1.03
kknn	0.0195
glmnet	-0.187
gbm	0.476

$$\mathrm{diff} = EPP_{RF} - EPP_{GMB} = 1.03 - 0.476 = 0.554$$

EPP scores are interpretable!

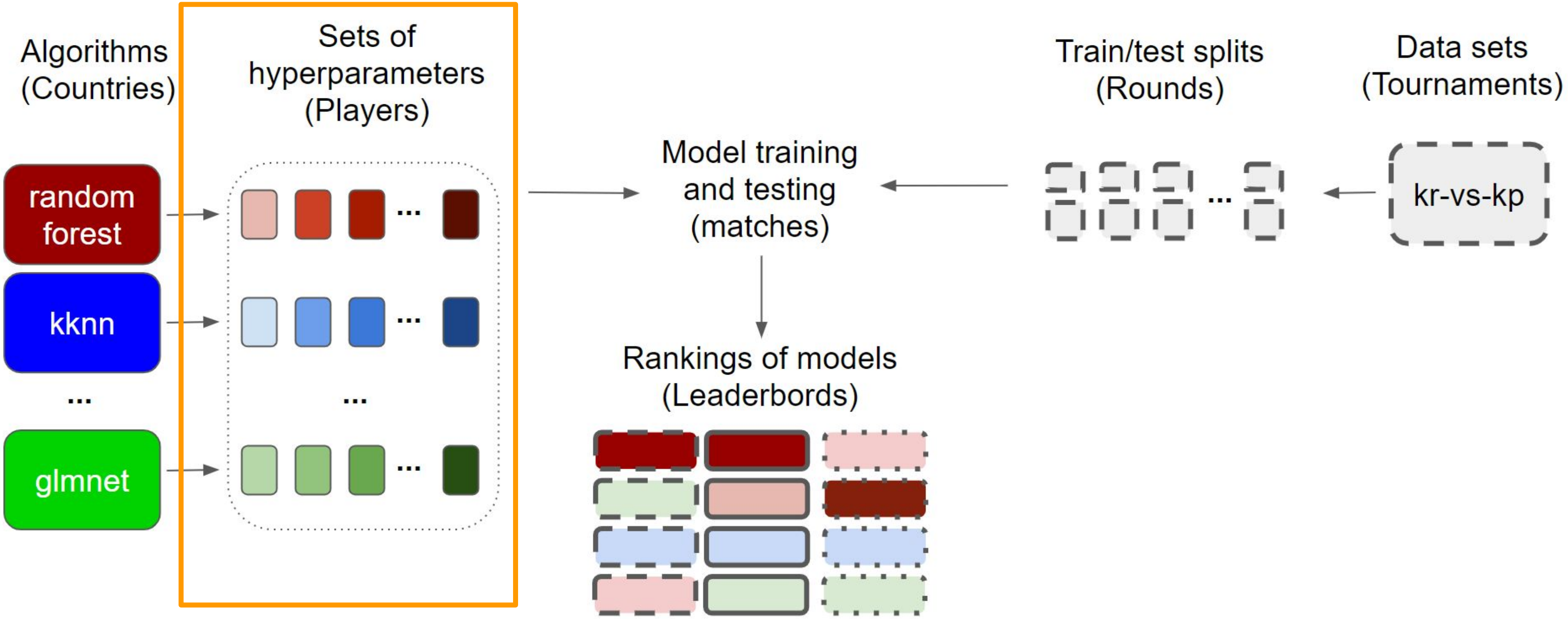


Model	EPP
randomForest	1.03
kknn	0.0195
glmnet	-0.187
gbm	0.476

$$\text{diff} = EPP_{RF} - EPP_{GMB} = 1.03 - 0.476 = 0.554$$

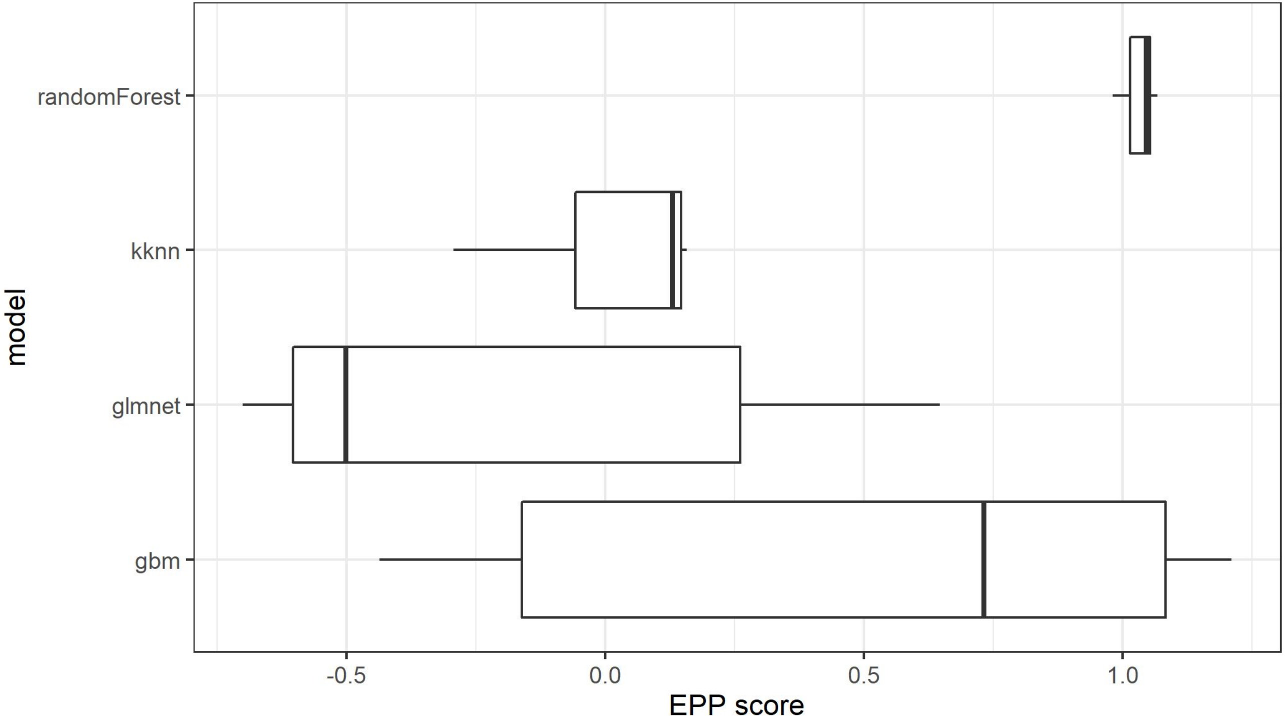
$$P\left(\begin{array}{c}\text{randomForest}\\ \text{wins with gbm}\end{array}\right) = invlogit(\text{diff}) = \frac{e^{\text{diff}}}{1+e^{\text{diff}}} = \frac{e^{0.554}}{1+e^{0.554}} = 0.635$$

Tunability of the algorithms

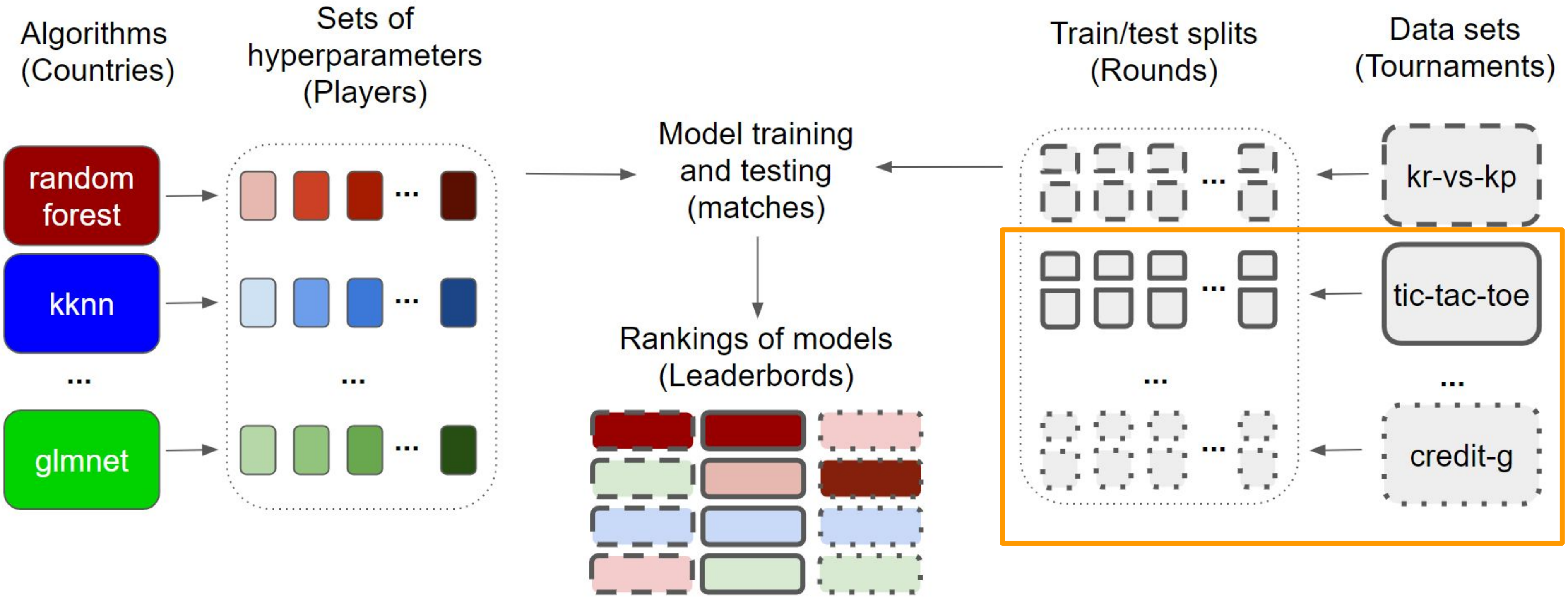


Tunability of the algorithms

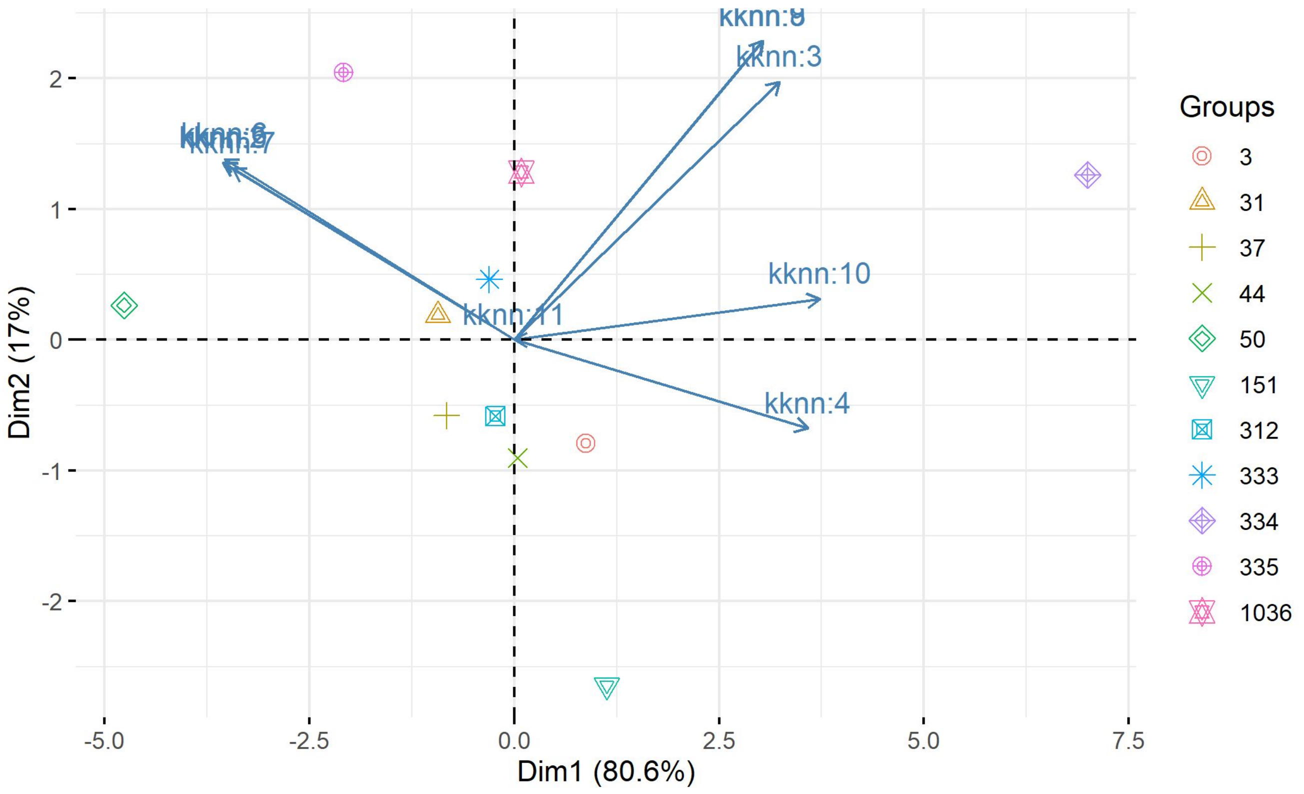
EPP scores for different hyperparameters



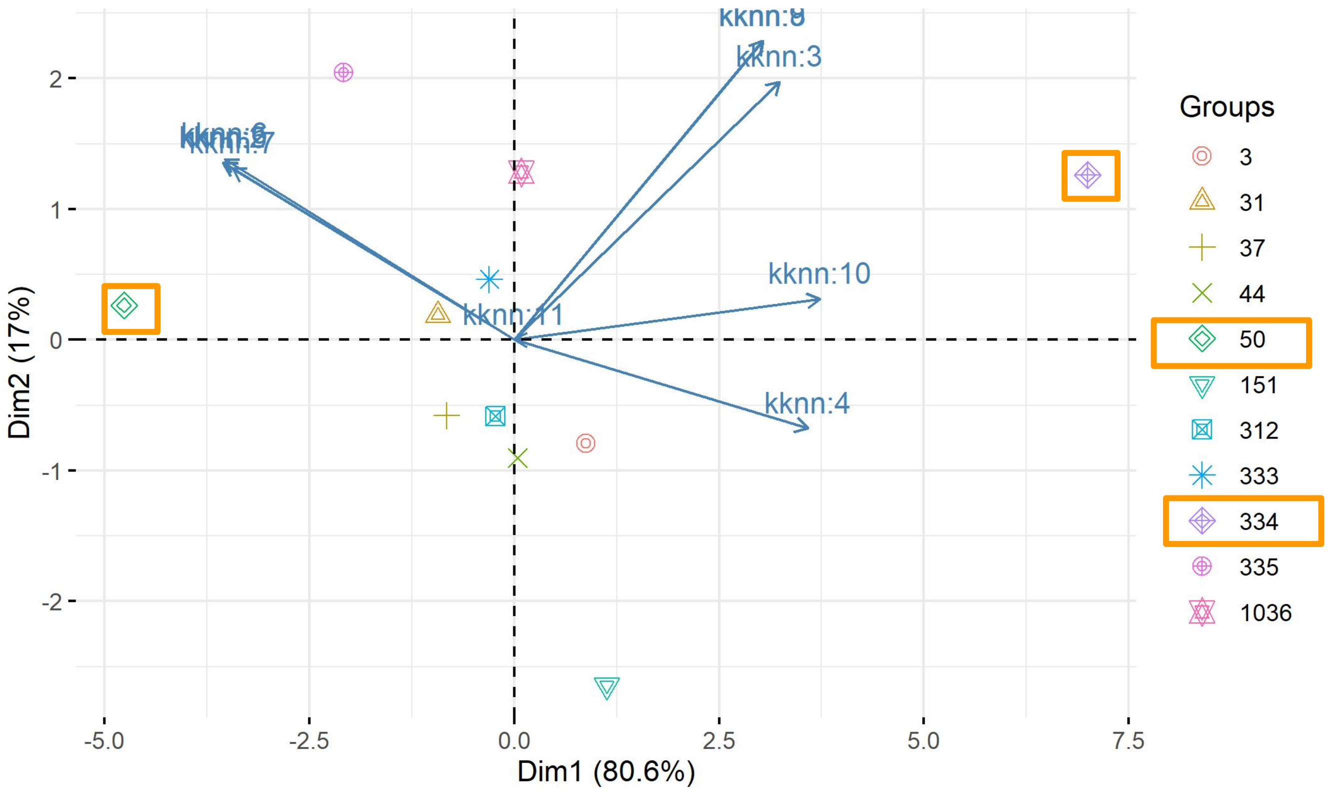
EPP-based embeddings of data sets



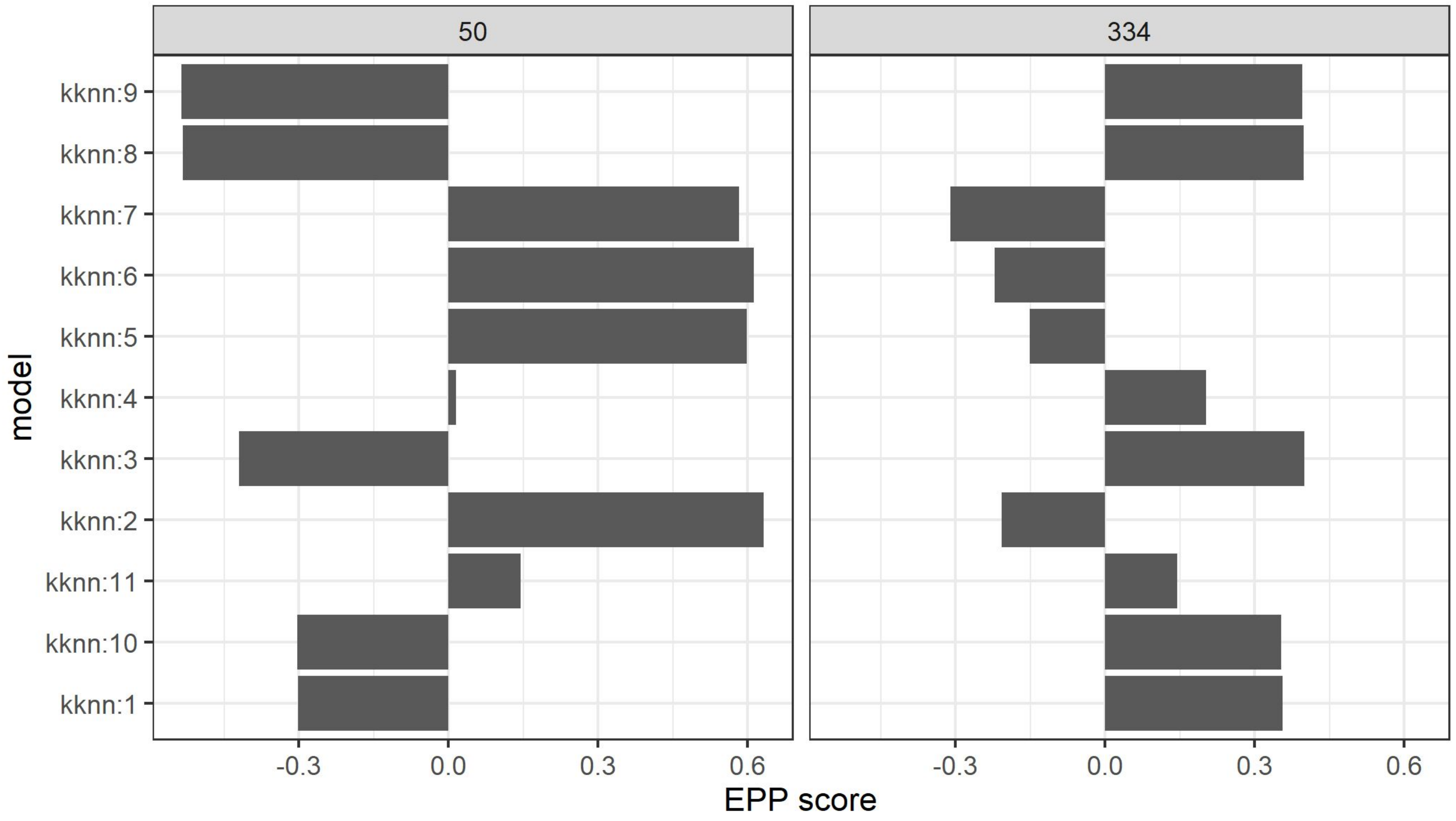
PCA on EPP scores across data sets



PCA on EPP scores across data sets



EPP scores for different hyperparameter settings



Takeouts

Takeouts

EPP: Elo-based Predictive Power score:

1) There is a probabilistic interpretation of differences in performance.

2) You can use EPP score to compare models across different hyperparameters and different data sets.

EPP: INTERPRETABLE SCORE OF MODEL PREDICTIVE POWER

A PREPRINT

http://gosiewska.com/

alicjagosiewska@gmail.com

agosiewska

NCN Opus grant 2017/27/B/ST6/01307

MI

Weaknesses of most popular measures 3/3

3. How stable is the performance for different CV folds?

k	AUC AutoML_1	AUC AutoML_2
1	0.8	0.9
2	0.8	0.78
3	0.8	0.78
4	0.8	0.78
Mean AUC	**0.8**	**0.81**

Weaknesses of most popular measures 3/3

3. How stable is the performance for different CV folds?

k	AUC AutoML_1	AUC AutoML_2
1	0.8	0.9
2	0.8	0.78
3	0.8	0.78
4	0.8	0.78
Mean AUC	**0.8**	**0.81**

Comparing just means across folds creates false impression that the AutoML_2 model is better than the AutoML_1.

Yet, we can see that AutoML_1 wins in 3 out of 4 folds.