ELO: interpretable score of model predictive power

Alicja Gosiewska
Mateusz Bakała
Katarzyna Woźnica
Maciej Zwoliński
dr hab. inż. Przemysław Biecek

October 9, 2019

Contents

1. Motivation: Problems of used performance measure
2. What is Elo ranking system?
3. EloML

Problems

Team	AUC
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
Arno Candel @ H2O.ai	0.615492
ALDAPOP	0.615040
9hr Overfitness	0.614371
Shlandryn	0.614132
Erin (H2O AutoML 100 mins)	0.612657

Table: Top 10 results of KaggleDays SF competition in 2019. https://www.kaggle.com/antgoldbloom/analyzing-kaggledays-sf-competition-data/notebook

Problems

Team	AUC
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
Arno Candel @ H2O.ai	0.615492
ALDAPOP	0.615040
9hr Overfitness	0.614371
Shlandryn	0.614132
Erin (H2O AutoML 100 mins)	0.612657

Table: Top 10 results of KaggleDays SF competition in 2019. https://www.kaggle.com/antgoldbloom/analyzing-kaggledays-sf-competition-data/notebook

Weakness 1: There is no interpretation of differences in performance

Problems

Team	AUC
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
Arno Candel @ H2O.ai	0.615492
ALDAPOP	0.615040
9hr Overfitness	0.614371
Shlandryn	0.614132
Erin (H2O AutoML 100 mins)	0.612657

Table: Top 10 results of KaggleDays SF competition in 2019. https://www.kaggle.com/antgoldbloom/analyzing-kaggledays-sf-competition-data/notebook

Weakness 1: There is no interpretation of differences in performance

Weakness 2: There is no procedure for assessing the significance of the difference in performances

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**

Table: Artifficial results from 4-fold cross-validation.

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**

Table: Artifficial results from 4-fold cross-validation.

Weakness 3: You cannot assess the stability of the performance in cross-validation folds

Team Name	AUC
Asian Ensemble	0.8043
.baGGaj.	0.8039
Erkut & Mark,Google AutoML	0.8039
ARG eMMSamble	0.8037
n_m	0.8021

Table: Springleaf Marketing Response Kaggle Competition, https://www.kaggle.com/c/springleaf-marketing-response

Team Name	AUC
alijs	0.9562
7777777777777...	0.9559
ML Keksika	0.9546
krivoship	0.9544
2 old mipt dogs	0.9543

Table: IEEE-CIS Fraud Detection Kaggle Competition, https://www.kaggle.com/c/ieee-fraud-detection

Team Name	AUC
Asian Ensemble	0.8043
.baGGaj.	0.8039
Erkut & Mark,Google AutoML	0.8039
ARG eMMSamble	0.8037
n_m	0.8021

Table: Springleaf Marketing Response Kaggle Competition, https://www.kaggle.com/c/springleaf-marketing-response

Team Name	AUC
alijs	0.9562
7777777777777...	0.9559
ML Keksika	0.9546
krivoship	0.9544
2 old mipt dogs	0.9543

Table: IEEE-CIS Fraud Detection Kaggle Competition, https://www.kaggle.com/c/ieee-fraud-detection

Weakness 4: You cannot compare performances between data sets

What is Elo ranking system?

Elo is used in:

- chess
- football and basketball ratings

Pros: The difference between Elo ratings of two players can be transferred into probabilities of winning when they play against each other.

What is Elo ranking system?

Elo is used in:

- chess
- football and basketball ratings

Pros: The difference between Elo ratings of two players can be transferred into probabilities of winning when they play against each other.

- rating is calculated on the basis of two components, result of match and rating of the opponent, The scores are updated after each match

$$E_1 = \frac{1}{1 + 10^{\frac{(S_1 - S2)}{400}}}.$$

$$S_1' = S_1 + K(A_1 - E_1),$$

Calculating ELO for predictive power

Let $p_{i,j}$ be the probability of model M_i wining with model M_j

$$logit(p_{i,j}) = \beta_{M_i} - \beta_{M_j}.$$

For larger number of models:

$$logit(p_{i,j}) = \beta_{M_1} x_{M_1} + \beta_{M_2} x_{M_2} + ... + \beta_{M_k} x_{M_n}$$

where

$$x_{M_a} = \begin{cases} 1 & \text{if } a = i \\ -1 & \text{if } a = j \\ 0 & \text{otherwise} \end{cases}$$

Calculating ELO for predictive power

Let $p_{i,j}$ be the probability of model M_i wining with model M_j

$$logit(p_{i,j}) = \beta_{M_i} - \beta_{M_j}.$$

For larger number of models:

$$logit(p_{i,j}) = \beta_{M_1} x_{M_1} + \beta_{M_2} x_{M_2} + ... + \beta_{M_k} x_{M_n}$$

where

$$x_{M_a} = \begin{cases} 1 & \text{if } a = i \\ -1 & \text{if } a = j \\ 0 & \text{otherwise} \end{cases}$$

We use logistic regression with contrast matrix.

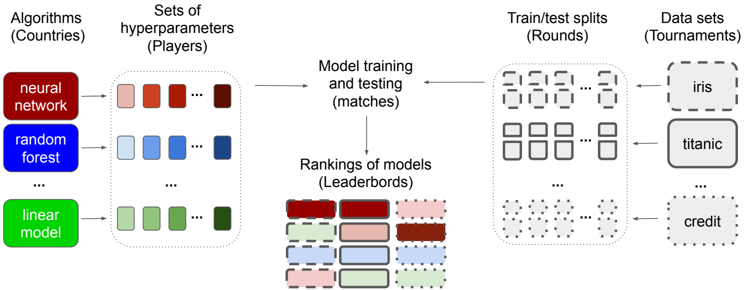
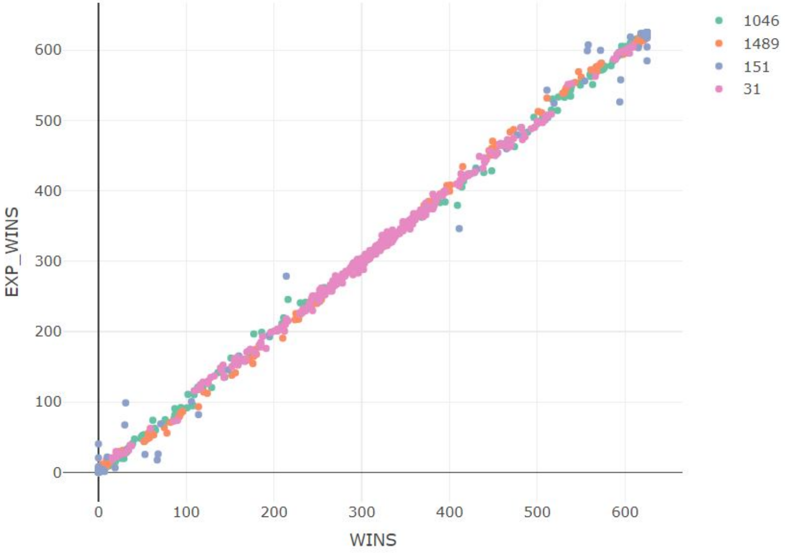


Figure: Our novel concept of Elo-based model ranking. Colors represent machine learning algorithms, gradients represent sets of hyperparameters, border styles represent data set.

Expected vs. Actual Results



EloML

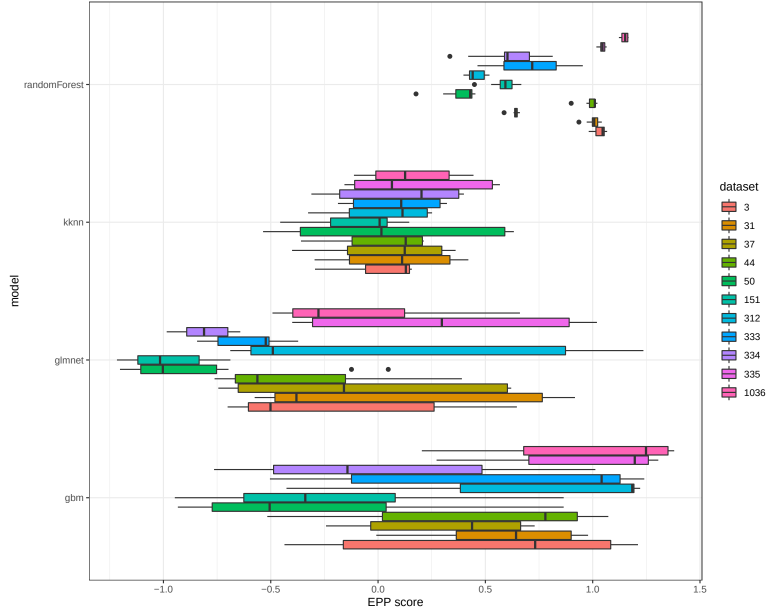
The advantages of ELO

1. ELO score provides the direct interpretation in terms of probability

$$p_{i,j} = invlogit(\beta_{M_i} - \beta_{M_j}) = \frac{e^{\beta_{M_i} - \beta_{M_j}}}{1 + e^{\beta_{M_i} - \beta_{M_j}}}.$$

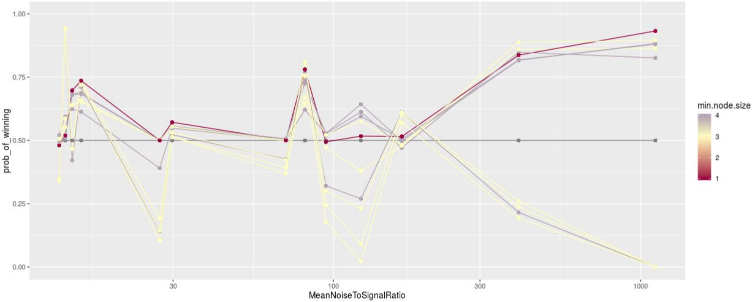
2. There is a procedure for assessing the significance of the difference in performances
3. You can assess the stability of the performance in cross-validation folds
4. You can compare performances between data sets

Tunability



randomForest
kknn
glmnet
gbm
model
−1.0
−0.5
0.0
0.5
1.0
1.5
EPP score
dataset
3
31
37
44
50
151
312
333
334
335
1036

Comparison between datasets



https://github.com/ModelOriented/EloML