Elo: interpretable score of model predictive power

Alicja Gosiewska [1], Katarzyna Woźnica [1], Maciej Zwoliński [1], Przemysław Biecek [1,2]

[1]Faculty of Matchematics and Information Science, Warsaw University of Technology
[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw

Introduction

The measurement of performance is the foundation of the model selection and hyperparameter tuning. The choice of the algorithm strongly relies on the choice of an evaluation measure. But the most popular performance measures have some weakness. We introduce a new Elo rating system for predictive models, which handles these identified issues and enables new application and extentions in many areas of machine learning.

What is wrong with most common measures?

- There is no interpretation of differences in performance.
- There is no procedure for assessing the significance of the difference in performances.
- You cannot compare performances between data sets.
- You cannot assess the stability of the performance in cross-validation folds.

Elo-based Predictive Power score

Our novel idea is to transfer the way players are ranked in the Elo system to create rankings of models.
Let $p_{i,j}$ be the probability of model M_i wining with model M_j. Then we can specify formula

$$logit(p_{i,j}) = \beta_{M_i} - \beta_{M_j}. \quad (1)$$

Unknown β coefficients can be estimated with logistic regression. Once β coefficients are estimated, one can calculate $p_{i,j}$ from the following formula

$$p_{i,j} = invlogit(\beta_{M_i} - \beta_{M_j}) = \frac{e^{\beta_{M_i} - \beta_{M_j}}}{1 + e^{\beta_{M_i} - \beta_{M_j}}}. \quad (2)$$

Experiment setup - benchmark

We have used 4 machine learning algorithms (glmnet, kknn, and randomForest, ranger). Each algorithm has been studied for 100 different hyperparameters settings on 38 selected classification data sets from the OpenML100 benchmark. For each subset, we fitted models on train data, computed AUC on test data and then we applied methodology of calculating Elo.

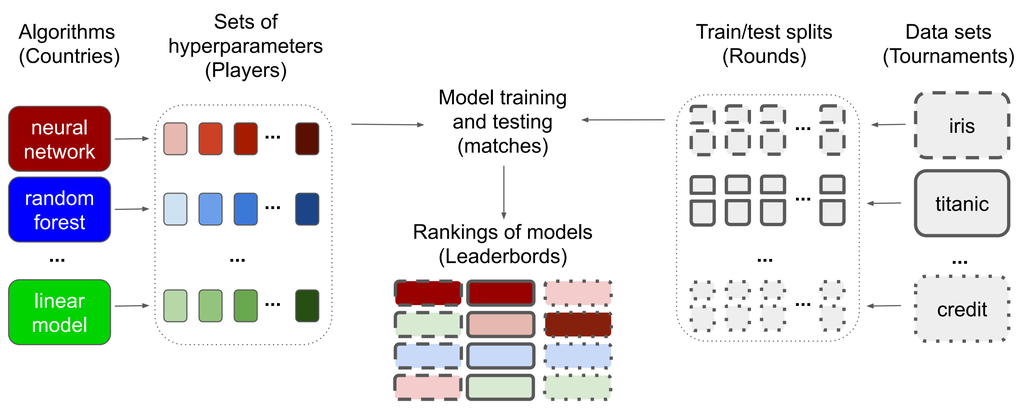


Figure 1: Plan of experiments

Tuning hyperparameters of algorithms - AutoML

The Elo score has a huge potential for supporting hyperparameter tuning - we can analyze performance measurses of different models with various hyperparameters and compare our results between different datasets.

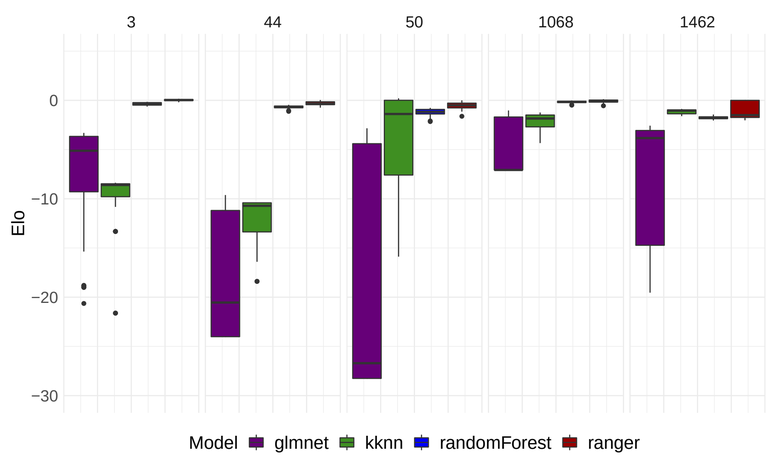


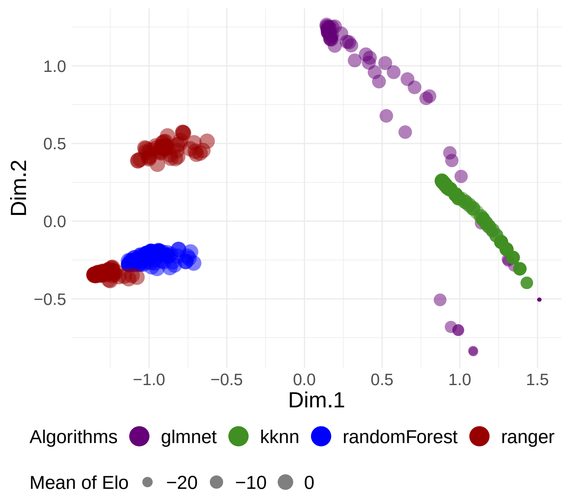
Figure 2: Boxplots of Elo scores for different models across data sets.

Embeddings

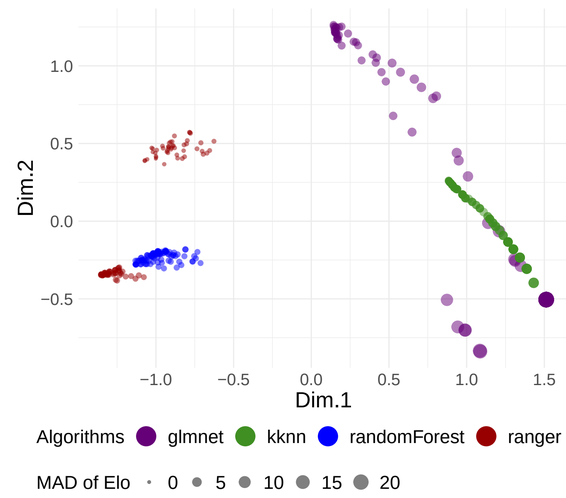
We gain opportunity to explore relationship between datasets and methods. We cluster data and models with hyperparameters simultaneously to find groups of algorithms which has close model performances for some type of datasets.

Figure 3: Results of biclustering of datasets and models. Clustering of models detect different methods. The most important conclusion is that some of the ranger hyperparameters are clustered with randomForest models. Glmnet are split into two groups.

To better insight into structure of Elo score vector space, there may be applied embeddings methods. We create map of relative positions of algorithms using multidimensional scaling.



(a) Size of points corresponds to mean Elo score in datasets.



(b) Size of points corresponds to mean absolute deviance around mean Elo score in datasets.

Figure 4: MDS plot for models. This is confirmation that some of ranger models are close to randomForest. Research reveals that these ranger models use specifing method of defining spliting rule.

Conclusions

Thanks to statistical properties, Elo opens many new way of application in machine learning models:

- navigated hyperparameters tuning,
- Explainable Artificial Intelligence (XAI),
- AutoML.

- ✉ woznicakatarzyna22@gmail.com

References

[1] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites and the OpenML100. 2017.

[2] Alicja Gosiewska, Mateusz Bąkała, Katarzyna Woźnica, Maciej Zwoliński, and Przemysław Biecek. EPP: interpretable score of model predictive power. 2018. URL https://arxiv.org/abs/1908.09213.

Acknowledgements

Alicja Gosiewska was financially supported by the grant of Polish Centre for Research and Development POIR.01.01.01-00-0328/17. Przemysław Biecek was financially supported by the grant NCN Opus grant 2017/27/B/ST6/01307.