DO WE UNDERSTAND RANDOM FORESTS?

Lasai Barreñada[1], Anne-Laure Boulesteix[2], Ben Van Calster[1,3]

[1]Department of Development and Regeneration, KU Leuven
[2]Biometry in Molecular Medicine, LMU Munich
[3]Department of Biomedical Data Sciences, LUMC Leiden

ABSTRACT

Introduction: In a case study on predicting ovarian malignancy with random forests, we observed training c-statistics close to 1. Although this suggests overfitting, performance was competitive on test data.
Objectives: Better understanding of this phenomenon by visualizing the predicted probabilities in the data space and running a simulation study to analyse the results in each scenario.
Methods: A simulation study with 192 scenarios varying random forest hyperparameters and exploration of the data space in a case study.
Results: Median training c-statistic was in most cases close to 1. Median test c-statistics were higher with higher events per variable, higher minimum node size, and binary predictors. Median test c-statistic was negatively correlated to median train c-statistic.
Conclusion: Random forests learn local probability peaks, often yielding near perfect training c-statistics. This peaks are local enough to not affect importantly the test performance. However, our results suggest going against the recommendation to use fully grown trees in random forest models.

CASE STUDY

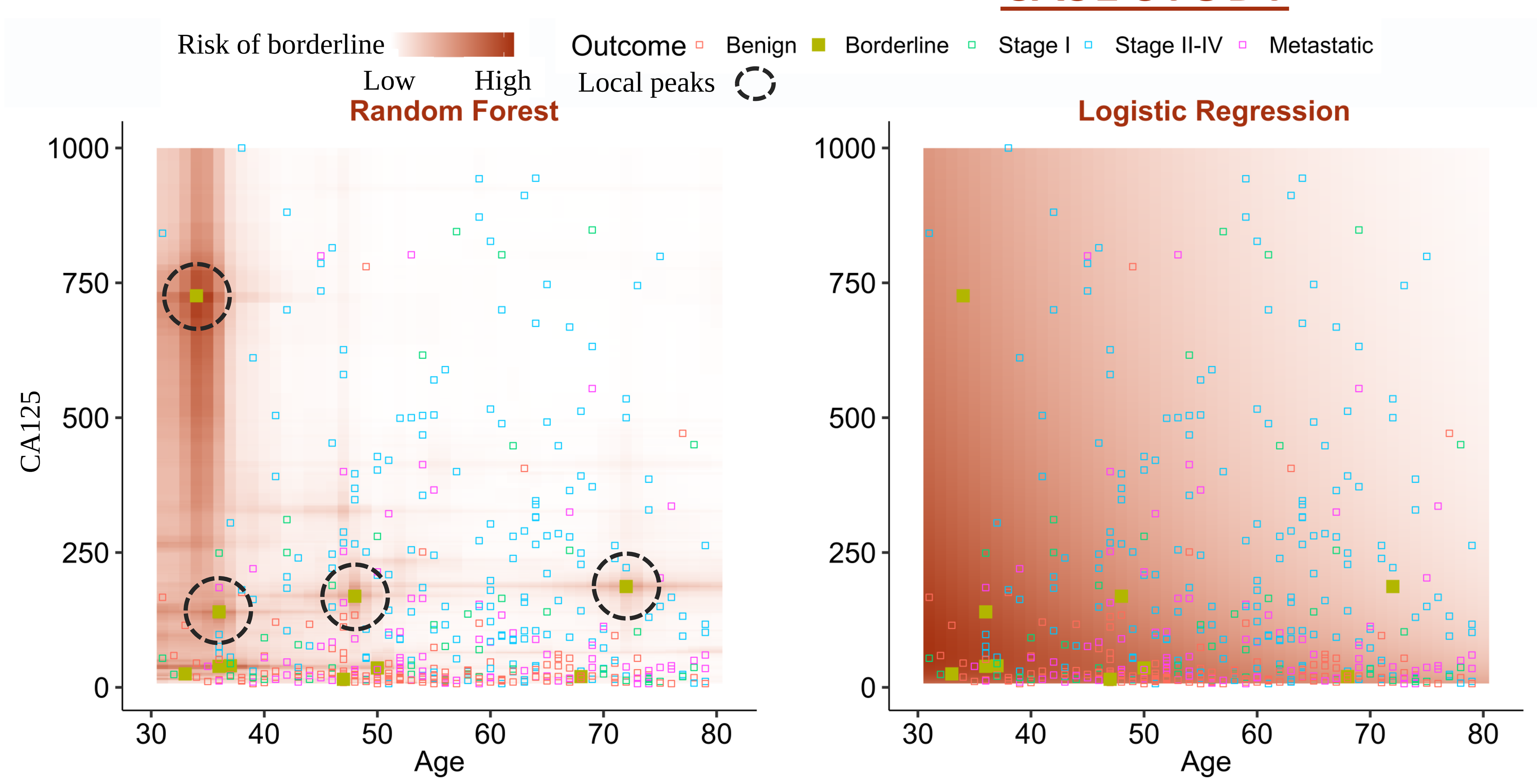

Figure 1. Risk probability estimates in data space for ovarian cancer diagnosis in random forest and logistic regression models.

SCAN ME!

EXPLORE THE DATA SPACE YOURSELF WITH THIS SHINY APP!

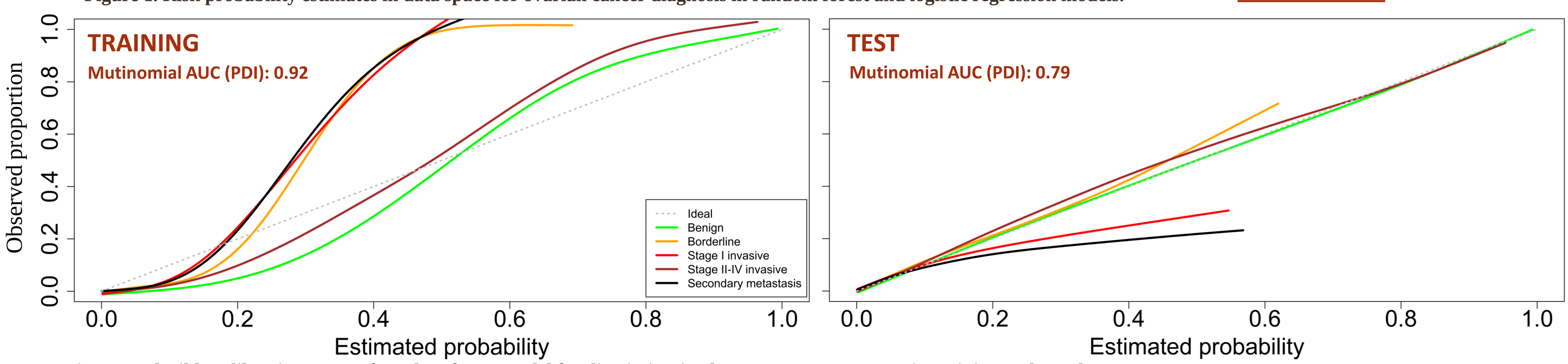

Figure 2. Flexible calibration curve of random forest model for discriminating between 5 tumour types in training and test data.

SIMULATION STUDY

Table 1. Simulation setup with 192 total scenarios.

Simulation factor	Values
Predictor distributions	Standard normal vs Binary (50% prevalence)
Number of predictors	4 vs 16 vs 16 (4 true 12 noise)
Correlation between predictors	0 vs 0.4
True c-statistic	0.75 vs 0.90
Strength of predictors	Balanced vs unbalanced
Training sample size	200 vs 4000
Minimum node size	2 vs 20

Methodology

- For each scenario we run 1000 simulations and compare the training performance in each simulation with the test performance in a unique test dataset (N = 100000).
- We compare performance and analyse it in terms of c-statistic, calibration and mean squared error.

CONCLUSIONS

- **Random forests learn local probability peaks, often yielding near perfect training c-statistics.**
- **Scenarios with higher training c-statistic tended to have poorer test performance.**
- **Higher minimum node size may often yield better test set performance.**
- **Training calibration slopes were always above 1, test slopes were above or below 1 (even for big training set) depending on the scenario.**
- **Further research is needed to better understand calibration performance and the convergence of the calibration slope.**

References

1. A. J. Wyner, M. Olson, J. Bleich, and D. Mease (2017). Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *Journal of Machine Learning Research 18*

2. Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. *Methods of Information in Medicine 51*

Funding: fwo

@BarrenadaLasai
lasai.barrenadataleb@kuleuven.be