Does imputation matter? Benchmark for real-life classification problems

Katarzyna Woźnica

Warsaw University of Technology
Faculty of Mathematics and Information Science

k.woznica@mini.pw.edu.pl

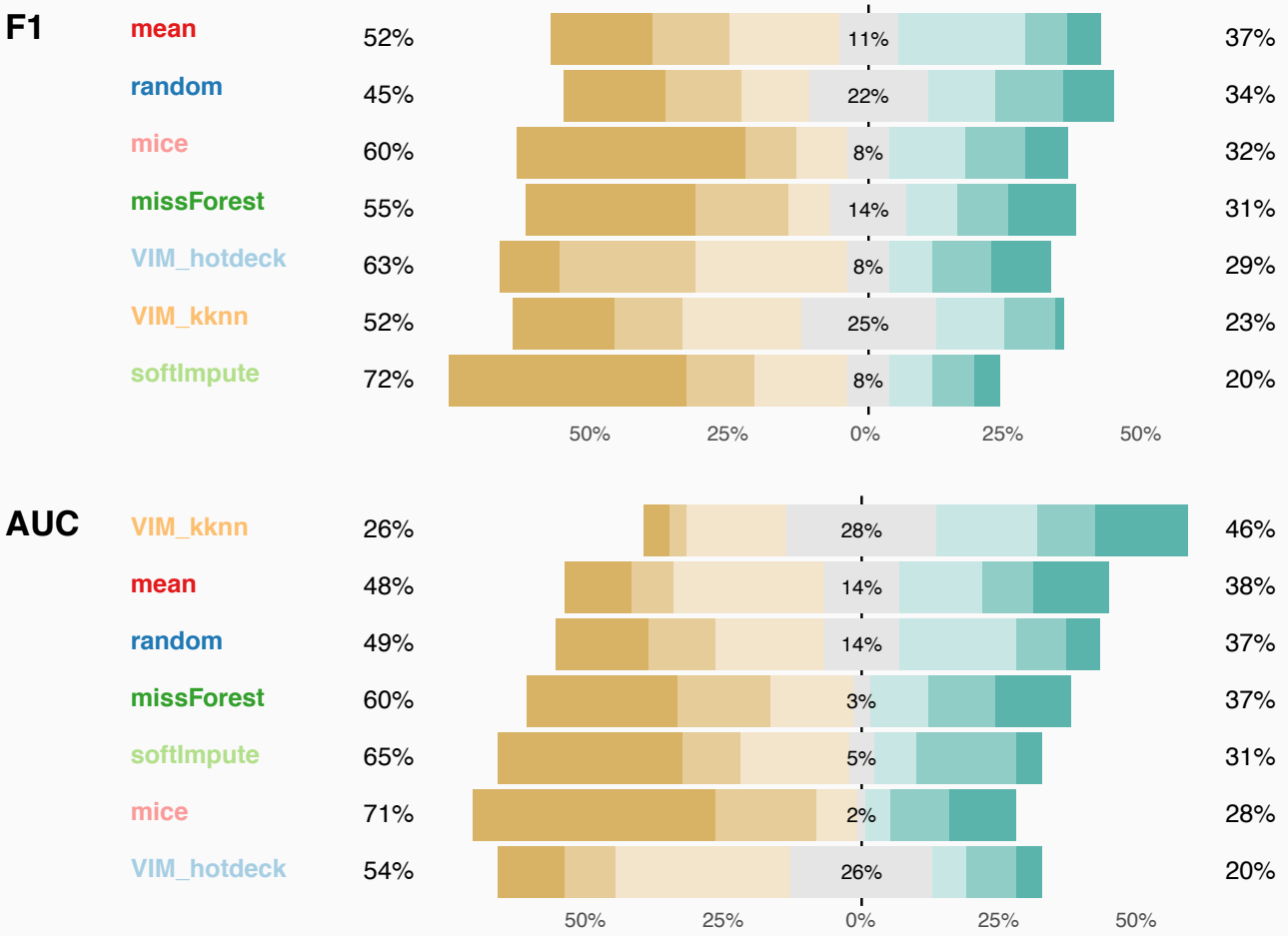
Settings for benchmark

- **7 imputation methods**: random, mean, kknn, hotdeck, mice, softImpute, missForest
- **5 ML classification algorithms**: glmnet, rpart, ranger, kknn, xgboost
- 2 measures of model performace: AUC and F1
- 14 real-world data sets

dataset name (dataset ID)	# obs	prc of missings
ipums_la_99-small (1018)	8844	7%
adult (1590)	48842	1%
eucalyptus (188)	736	3.9%
dresses-sales (23381)	500	14.7%
colic (27)	368	16.3%
credit-approval (29)	690	0.6%
sick (38)	3772	2.2%
labor (4)	57	33.6%
SpeedDating (40536)	8378	1.8%
hepatitis (55)	155	5.4%
vote (56)	435	5.3%
cylinder-bands (6332)	540	5.1%
echoMonths (944)	130	7.5%

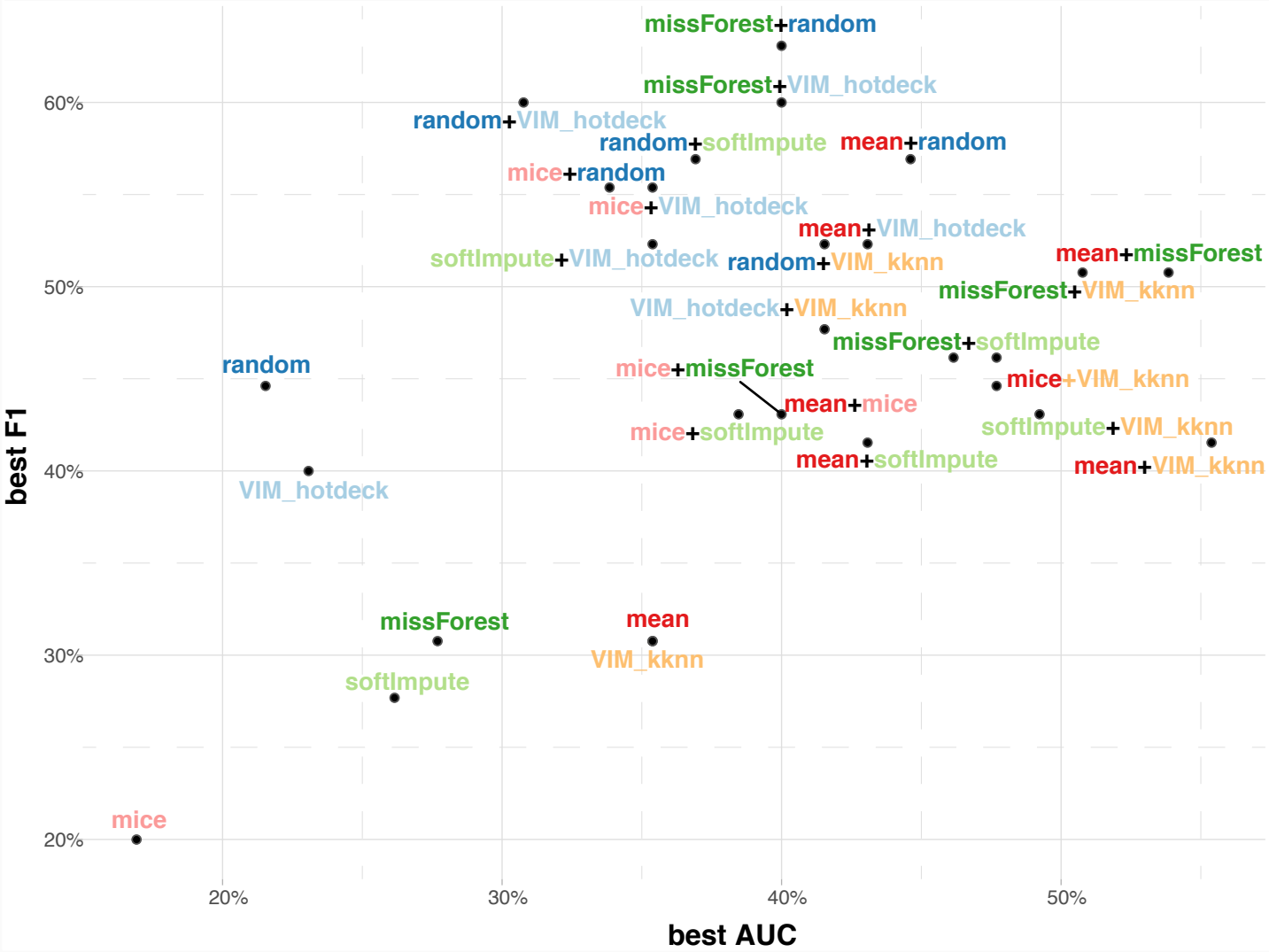
Does exist the best universal imputation method?

For F1 the best results are on average for *mean* imputations. For AUC for *kknn* imputations. Simple methods are surprisingly good!



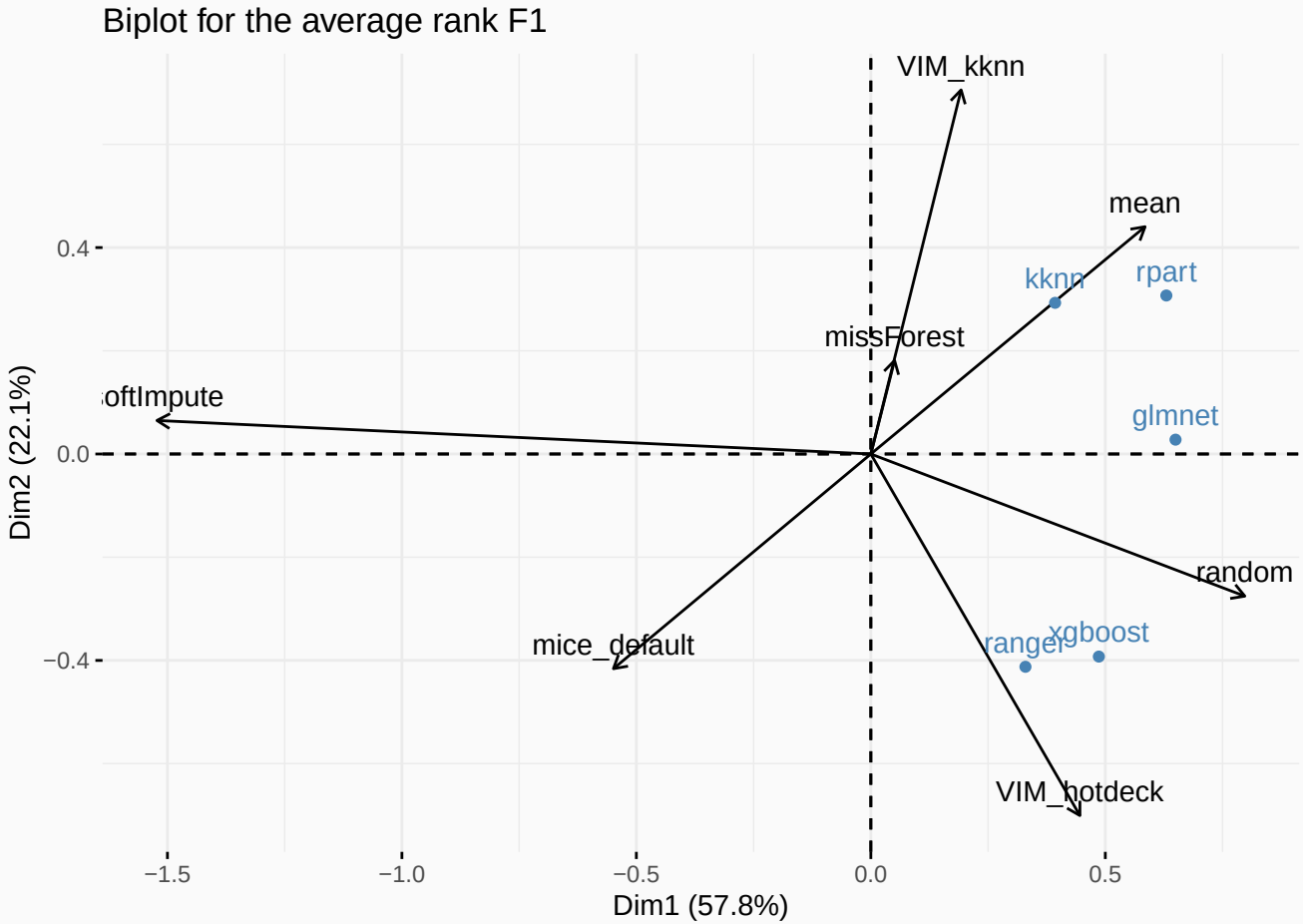
Do simple methods work effectively on similar tasks and ML algorithms?

Combinations of two methods are able to cover above 50% of best results. For F1 measure optimal pair is *missForest* and *random*. For AUC this is *mean* and *VIM_kknn* imputation.



What are the interactions between ML algorithms and imputation methods?

Mean, *missForest* and *VIM_kknn* methods cooperate with rpart and kknn while *mice* works with ranger and xgboost.



Conclusions

1. We perform the first empirically benchmark of imputation methods in terms of their impact on the predictive power of classifier algorithms.
2. We propose the general plan of the experiment.
3. Simple imputation methods achieve surprisingly good results. There are some trends in results but generally their structure is very complex. This is the area to employ meta-learning model.
4. We plan to extend this benchmark for other methods for imputations and more data sets.

Github repository: *https://github.com/ModelOriented/EMMMA*