Various data imputation techniques in R

Jan Borowski Filip Chrzuszcz Piotr Fic

Warsaw University of Technology

Agenda

01 Introduction

02 Methodology

03 Results

Motivation

- Many datasets have missing values.
- This causes problems in the implementation of machine learning models.

Aim of the study

Comparison of data imputation packages
in the context of supervised machine learning.

Compared imputation techniques

- median and mode imputation
- softImpute
- VIM
- missForest
- missMDA
- mice

Data sets

14 examples from OpenML library

Percentage of missing data

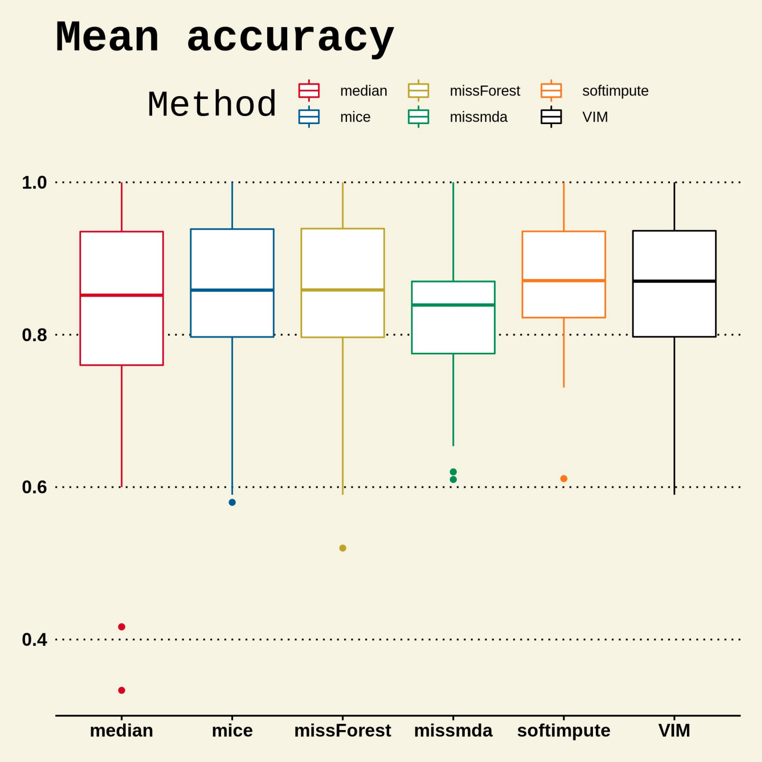
from 0.7%

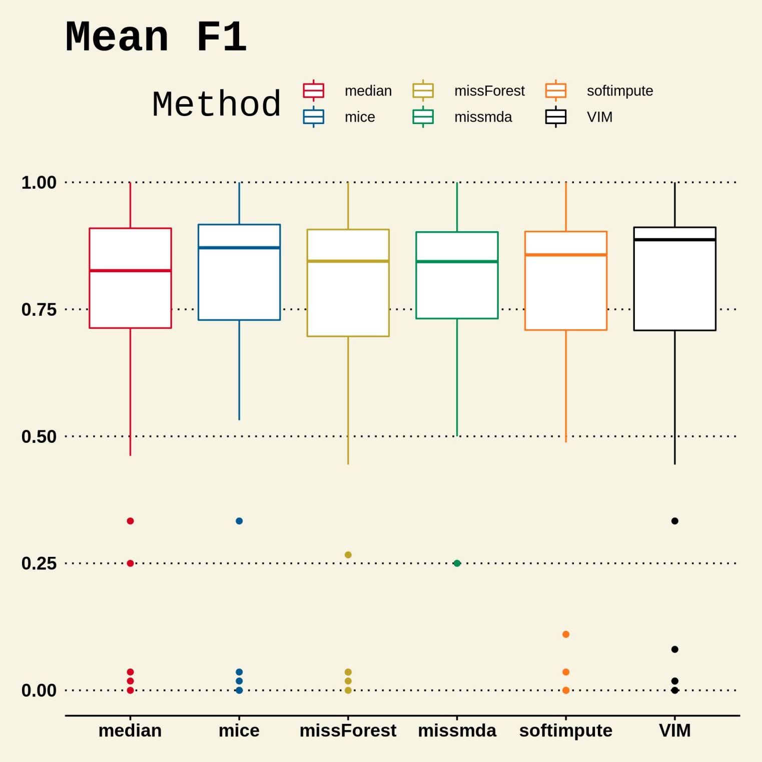
to 35.8%

Models

- logistic regression
- random forest
- SVM
- XGBoost

Scores


Mean accuracy
Method
median
mice
missForest
missmda
softimpute
VIM
1.0
0.8
0.6
0.4
median
mice
missForest
missmda
softimpute
VIM


Mean F1
Method
median
mice
missForest
missmda
softimpute
VIM
1.00
0.75
0.50
0.25
0.00
median
mice
missForest
missmda
softimpute
VIM

Another comparing approach

Mean F1 difference

- We treat the median as the basic form of imputation and we will compare it to all other methods. We want to check how much the average prediction measured made by other algorithms differes from the median

Method	Score
MissForest	0.039
SoftImpute	0.013
Mice	-0.042
VIM	-0.105
MissMDA	-0.112

Conclusion

- Different algorithms shine with different metrics

Possible gain vs lose

Score can fluctuate from *+4%*

to *-11%*