EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

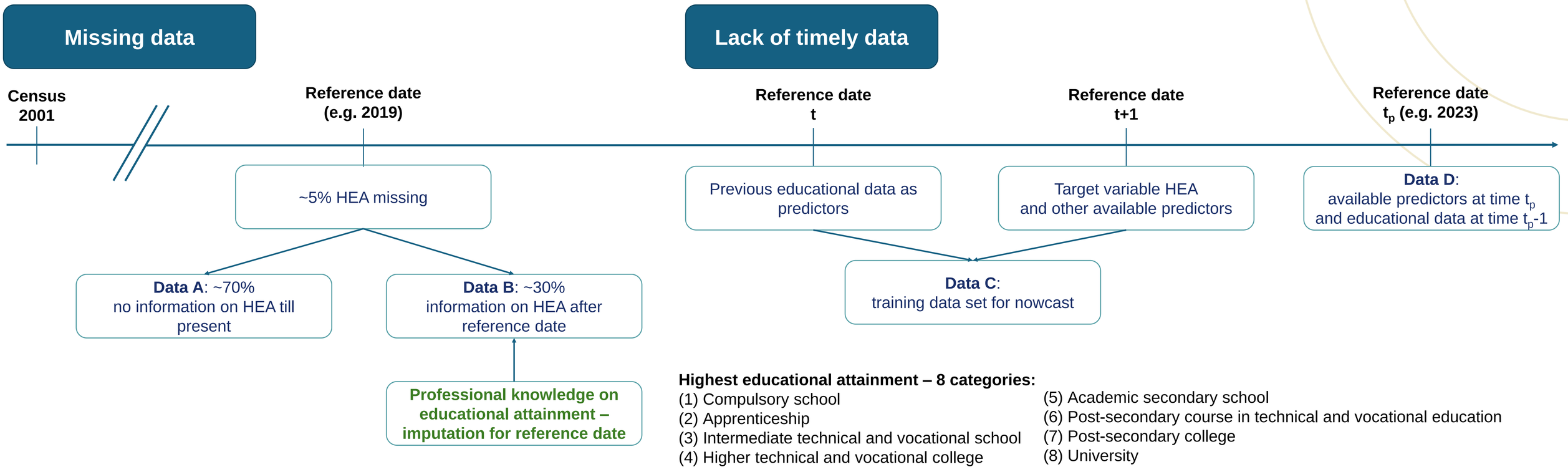
The conference is partly financed by the European Union

Imputation and nowcast of highest educational attainment: Combining professional knowledge and machine learning techniques

Christine Ning, Daniel Reiter
Statistik Austria, Vienna, Austria

Introduction

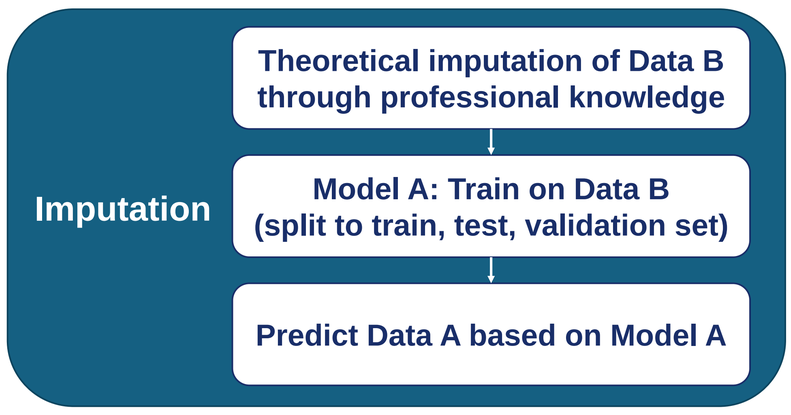
Highest educational attainment (HEA) of Austrian population aged 15 and above on reference date (31.10)

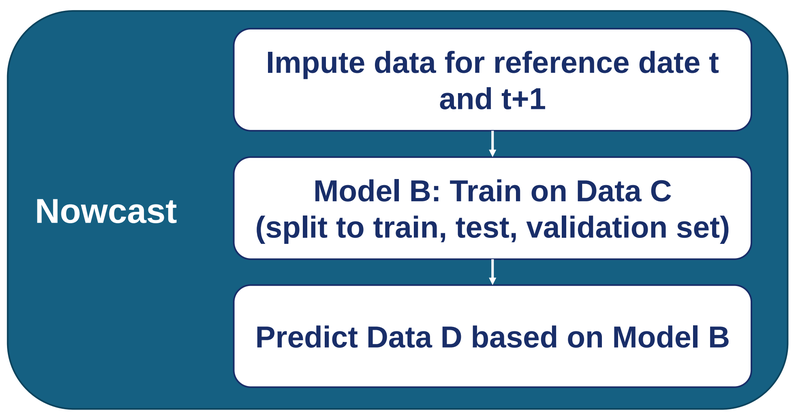


Highest educational attainment – 8 categories:

(1) Compulsory school
(2) Apprenticeship
(3) Intermediate technical and vocational school
(4) Higher technical and vocational college
(5) Academic secondary school
(6) Post-secondary course in technical and vocational education
(7) Post-secondary college
(8) University

Method





Models:

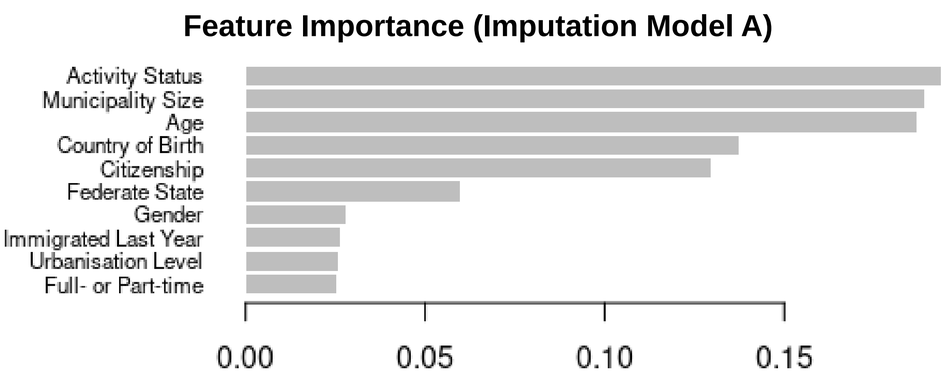
- XGBoost with 5-fold target encoding
- XGBoost with label encoding
- LightGBM with 5-fold target encoding
- LightGBM with label encoding

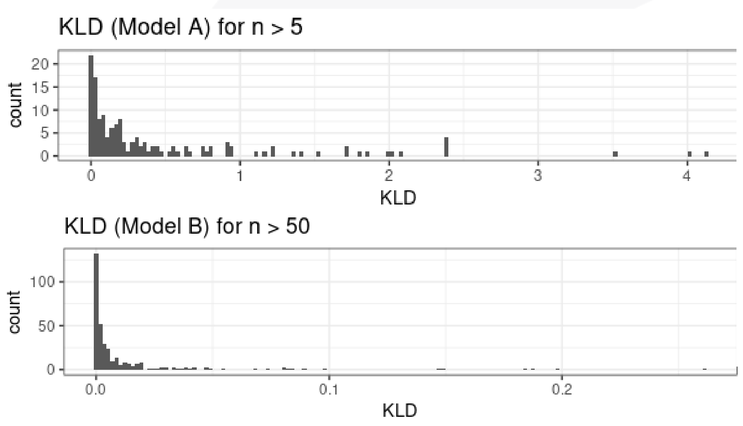
Model Selection:

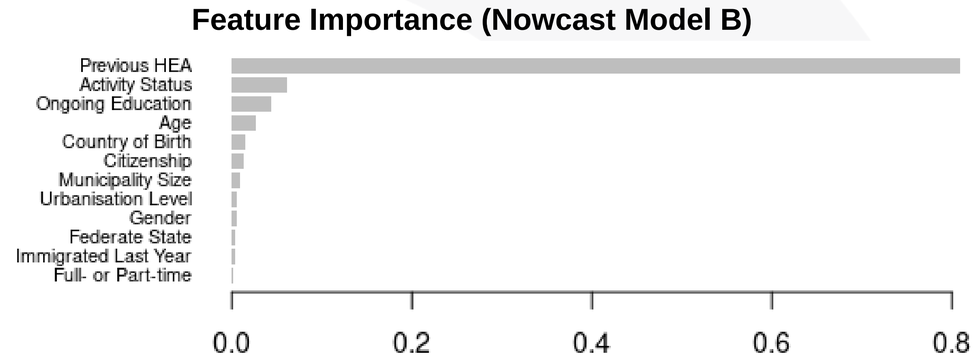
- Hyperparameter tuning with random grid search
- Evaluation metric: Weighted Kullback-Leibler divergence (KLD) for feature combinations

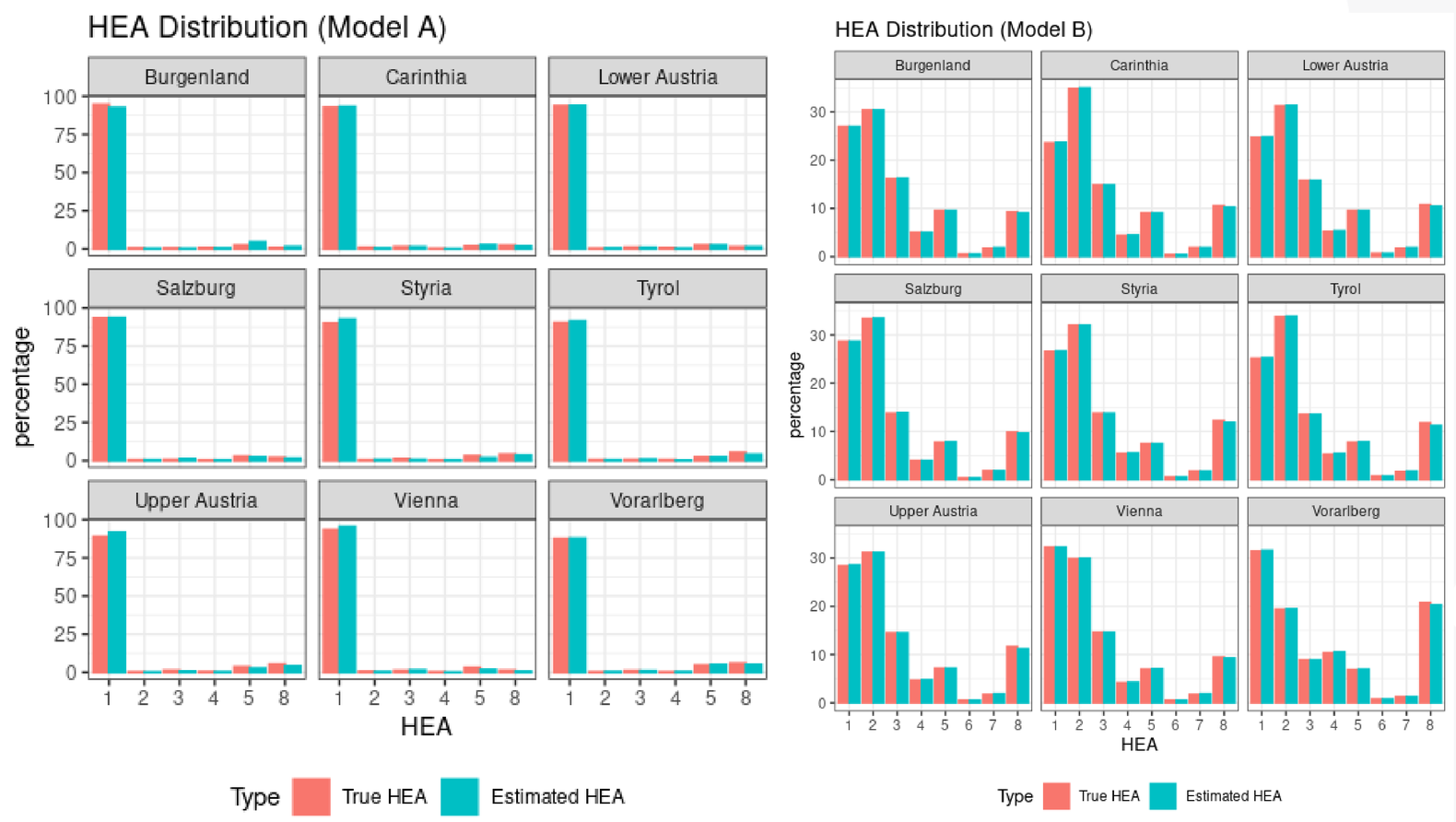
Results

Models	Weighted KLD (Model A)	Weighted KLD (Model B)
LightGBM label encoding	0.147	0.0012
LightGBM target encoding	0.176	0.1174
XGBoost label encoding	0.126	0.0010
XGBoost target encoding	0.195	0.0916









Conclusion

- Models with label encoding performed better than target encoding
- XGBoost performed better for Model A and B compared to LightGBM
- Kullback-Leibler divergence used for evaluation – goal: good prediction of HEA distribution within feature combinations
- Model updates necessary in the long term