Predicting early parenthood

Out-of-sample predictions

vs **In-sample predictions (R^2, regression coefficients…)**

- Reduce overfitting
- Evaluate strength of a theory
- Compare theories, importance of variables
- Compare theory-driven and data-driven models

Breiman 2001; Shmueli, 2010; Yarkoni & Westfall, 2017; Watts et al., 2018; Salganik et al. 2020; Hofman et al., 2021; Verhagen 2022; Stulp, Verhagen, Arpino (forthcoming)

university of groningen

Elizaveta Sivak
Department of Sociology
e.sivak@rug.nl

Predicting early parenthood

Out-of-sample predictions

vs **In-sample predictions (R², regression coefficients…)**

- Reduce overfitting
- Evaluate strength of a theory
- Compare theories, importance of variables,
- Compare theory-driven and data-driven models

Breiman 2001; Shmueli, 2010; Yarkoni & Westfall, 2017; Watts et al., 2018; Salganik et al. 2020; Hofman et al., 2021; Verhagen 2022; Stulp, Verhagen, Arpino (forthcoming)

How well can we predict early parenthood?

How might the results inform theory of fertility behaviour?

university of groningen

Elizaveta Sivak
Department of Sociology
e.sivak@rug.nl

DATA

Russian Longitudinal Panel Study of Educational and Occupational Trajectories (TrEC)

https://trec.hse.ru/en/

Nationally representative panel for one age cohort

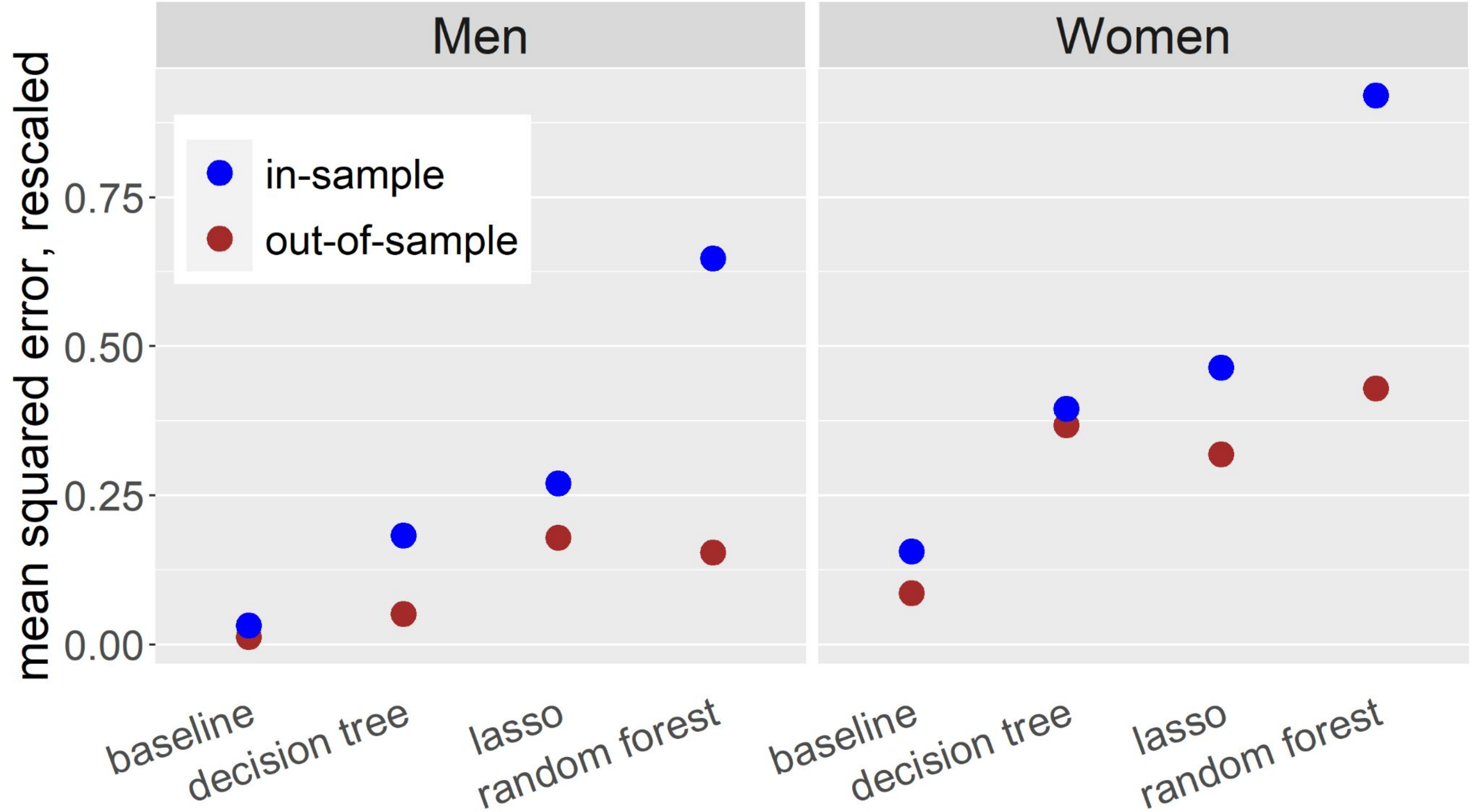
N ~ 4000, 15-16 years old in the first wave in 2012

METHOD

- Dependent variable: having at least one child at the age of 25 (22% overall, 28% women, 14% men). 10th wave, 2021
- Background variables: waves 1-7, 2012-2018
- 30%/70% test/train split
- The baseline model: logistic regression (level of education, partnership status, fertility intentions, job, income, mother's education, siblings)
- Data-driven models: decision tree, random forest, penalized logistic regression. All variables from the first 7 waves (~1700 variables)
- Cross-validation to tune the parameters of the models
- Performance:

$$MSE_{rescaled} = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - 0)^2}$$

(Salganik et al. 2020)

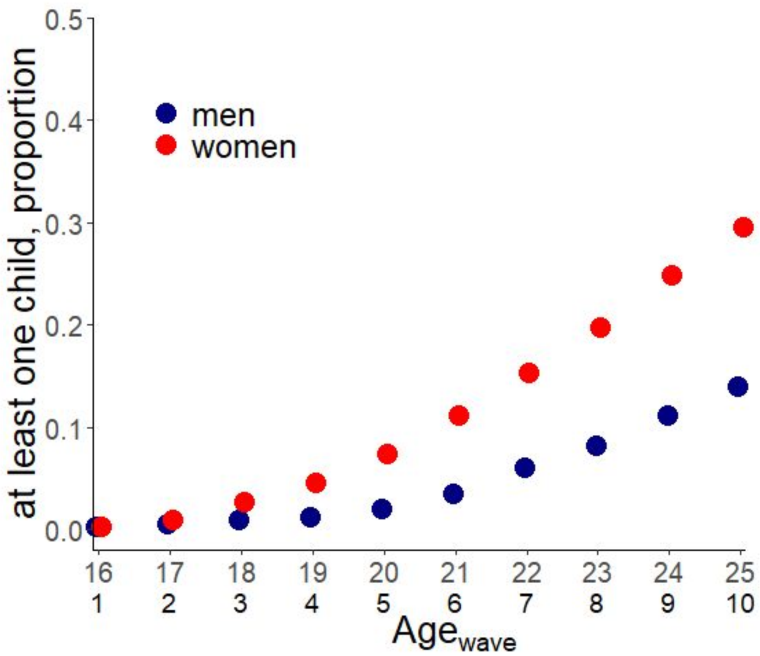

Men
Women
in-sample
out-of-sample
mean squared error, rescaled
0.75
0.50
0.25
0.00
baseline
decision tree
lasso
random forest
baseline
decision tree
lasso
random forest

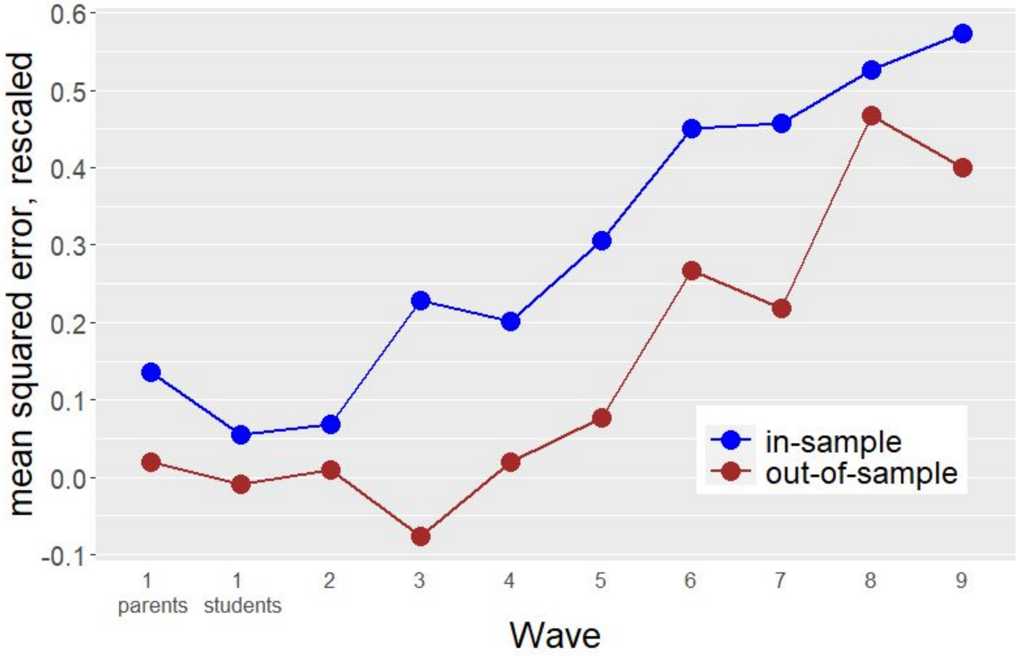
Predicting early parenthood

Most important predictors:

- Marital status, partnership trajectory
- Job trajectory
- Education trajectory

Data leakage?





Predicting early parenthood: conclusions

- Strongest predictors are related to long-term relationships, education and job trajectories — supports current theories on fertility
- Out-of-sample predictions are not very accurate —> theories are relevant, but predictors are weak?
- Richer dataset is needed to find unexpected predictors
- Better performance of the random forest model - non-linear relationships (?)

university of groningen

Elizaveta Sivak
Department of Sociology
e.sivak@rug.nl