TALL
SOM FORTELLER

Outlier detection and imputation using ML

SUSIE.JENTOFT@SSB.NO

UNECE STATISTICAL DATA EDITING WORKSHOP

Statistisk sentralbyrå
Statistics Norway

Agenda

Earnings statistics in Norway

Outlier detection

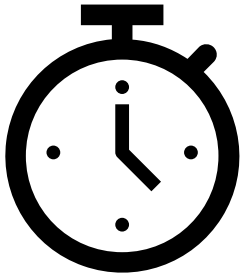
Imputation using xgboost

Earnings statistics

- Administrative data
- Published quarterly/yearly

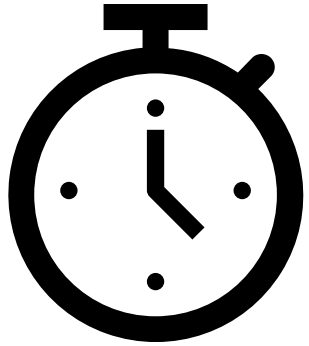
Statistisk sentralbyrå
Statistics Norway

Full-time equivalent earnings

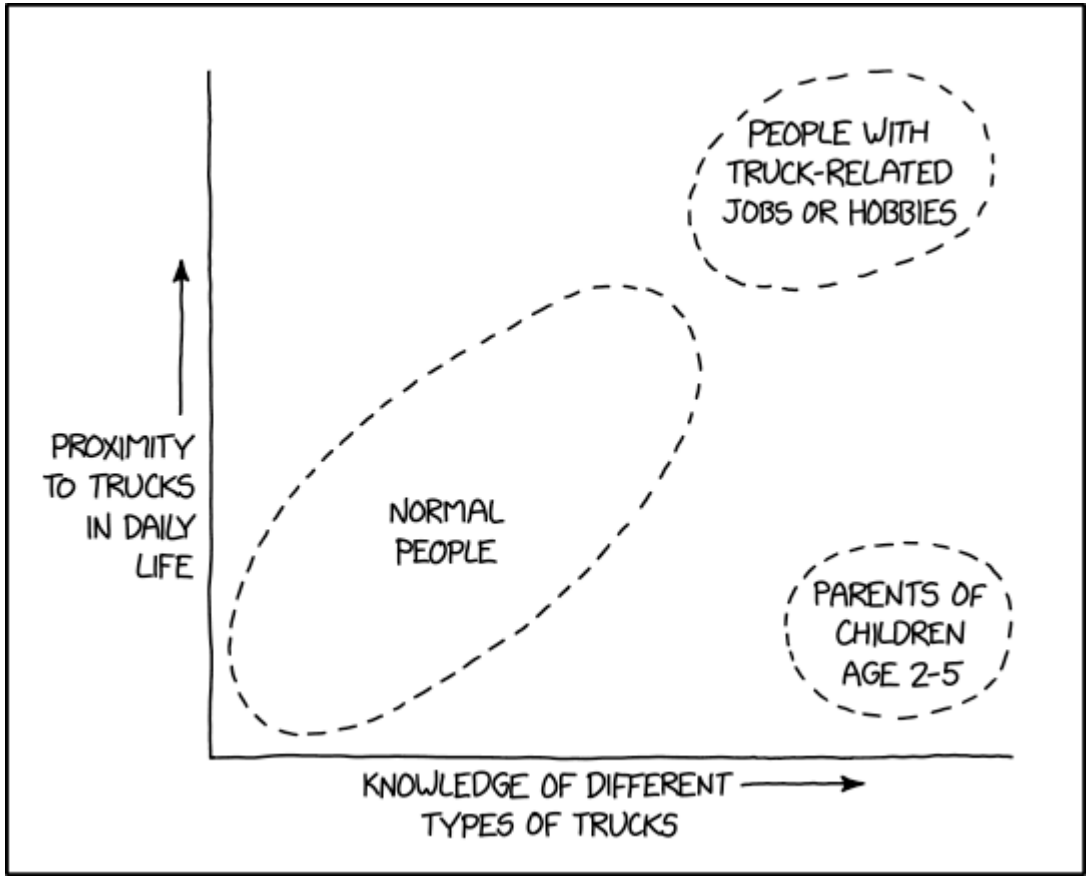
/  =

Outlier detection – contractual working hours

- Hourly based:
  - contractual working hours ~ paid working hours
- Fixed limits
  - FTE earnings
  - Hourly rate
- FTE earnings regression:
  - FTE earnings ~ paid earnings + …



Outliers



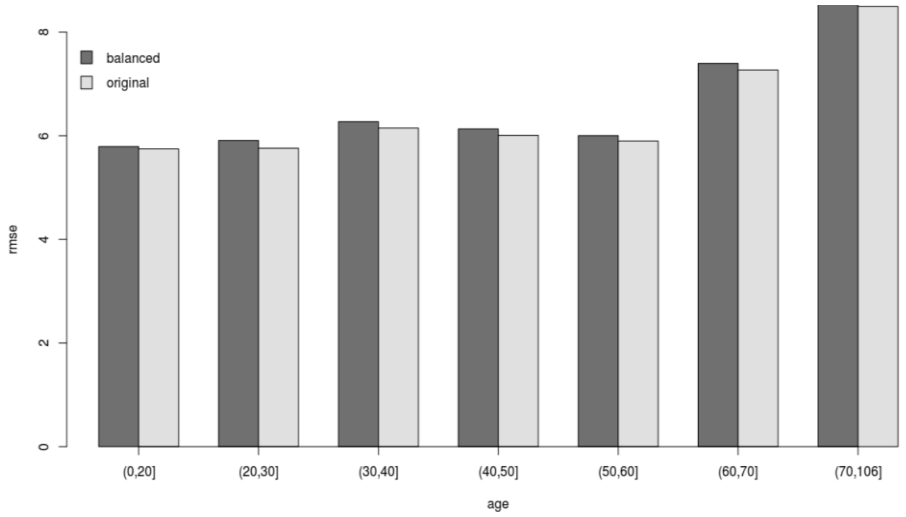
Source: xkcd.com

Imputation

Source: Analytics Vidhya

Results

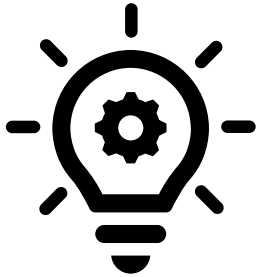
- Statistics are dependent on training dataset
- We don’t have a benchmark
- Test – «weighted» training dataset



Statistisk sentralbyrå
Statistics Norway

Future ideas

- Continue re-asssessing outlier detection rutines
- Use LFS for benchmarking certain groups



Aknowledgements

- Knut Håkon Grini
- Stine Bakke
- Ingvild Johansen

Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Takk!

Statistisk sentralbyrå
Statistics Norway