Outlier detection and imputation using ML

SUSIE.JENTOFT@SSB.NO

UNECE STATISTICAL DATA EDITING WORKSHOP
Agenda

Earnings statistics in Norway

Outlier detection

Imputation using xgboost
Earnings statistics

- Administrative data
- Published quarterly/yearly
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 + ...
Imputation

Cache awareness and out-of-core computing

Regularization for avoiding overfitting

Tree pruning using depth-first approach

Efficient handling of missing data

Parallelized tree building

In-built cross-validation capability

Source: Analytics Vidhya
Results

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

• Continue re-assessing outlier detection routines
• Use LFS for benchmarking certain groups
Aknowledgements

• Knut Håkon Grini
• Stine Bakke
• Ingvild Johansen
Takk!