ESRA Conference 2019

Predictive Modeling and Machine Learning in Survey Research

Zagreb, 18 Jul 2019

SHARE
SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

UPDATE ON MACHINE LEARNING IN DATA ANALYSIS FOR SOCIAL RESEARCH

Arne Bethmann (MEA)
Jonas Beste (IAB)
Giuseppe Casalicchio (LMU)
Bernd Bischl (LMU)

mea | MAX PLANCK INSTITUTE FOR SOCIAL LAW AND SOCIAL POLICY
Munich Center for the Economics of Aging

This project has received funding from the European Union under grant agreement VS/2018/0285 and the European Union's Horizon 2020 research and innovation programme under grant agreements No 676536, No 654221

SPONSORED BY THE
Federal Ministry of Education and Research

NIH National Institute on Aging

ML tasks in substantive social research

Machine Learning Tasks

	Numerical outcome	Categorical outcome
Supervised outcome observed	**Regression task** • Predict the average fiancial wealth of migrants and non-migrants • Predict blood measures from dried blood spots	**Classification tasks** • Predict propensity scores to analyse the difference in health between migrants and non-migrants • Analyze the effect of education on old age poverty
Unsupervised outcome unobserved	**Factor analysis tasks** • Predict individual factor scores on the EURO-D scale • Predict movement intensity from accelerometry data	**Clustering tasks** • Find similar groups of labour market trajectories among SHARE individuals • Find typology of SHARE interviewers in order to identify suspicious clusters

Some algorithms/models

	Numerical outcome	Categorical outcome
Supervised outcome observed	**Regression tasks** • Ordinary Least Squares • Linear Models • Splines, Lasso, Ridge • Regression Trees, Random Forests • Support Vector Machines • k-Nearest-Neighbors • Neural Nets, „Deep learning“	**Classification tasks** • Generalized Linear Models (Logistic or probit, …) • Splines, Lasso, Ridge • Classification Trees, Random Forests • Support Vector Machines • k-Nearest-Neighbors • Neural Nets, „Deep learning“ • Naive bayes
Unsupervised outcome unobserved	**Factor analysis tasks** • Principal Components Analysis (PCA) • Exploratory Factor Analysis (EFA) • Confirmatory Factor Analysis (CFA)	**Clustering tasks** • Hierarchical clustering (k-Means, Linkage, Ward …) • Model based (Finite mixtures, …) • Latent Class Analysis (for ordinal outomes)

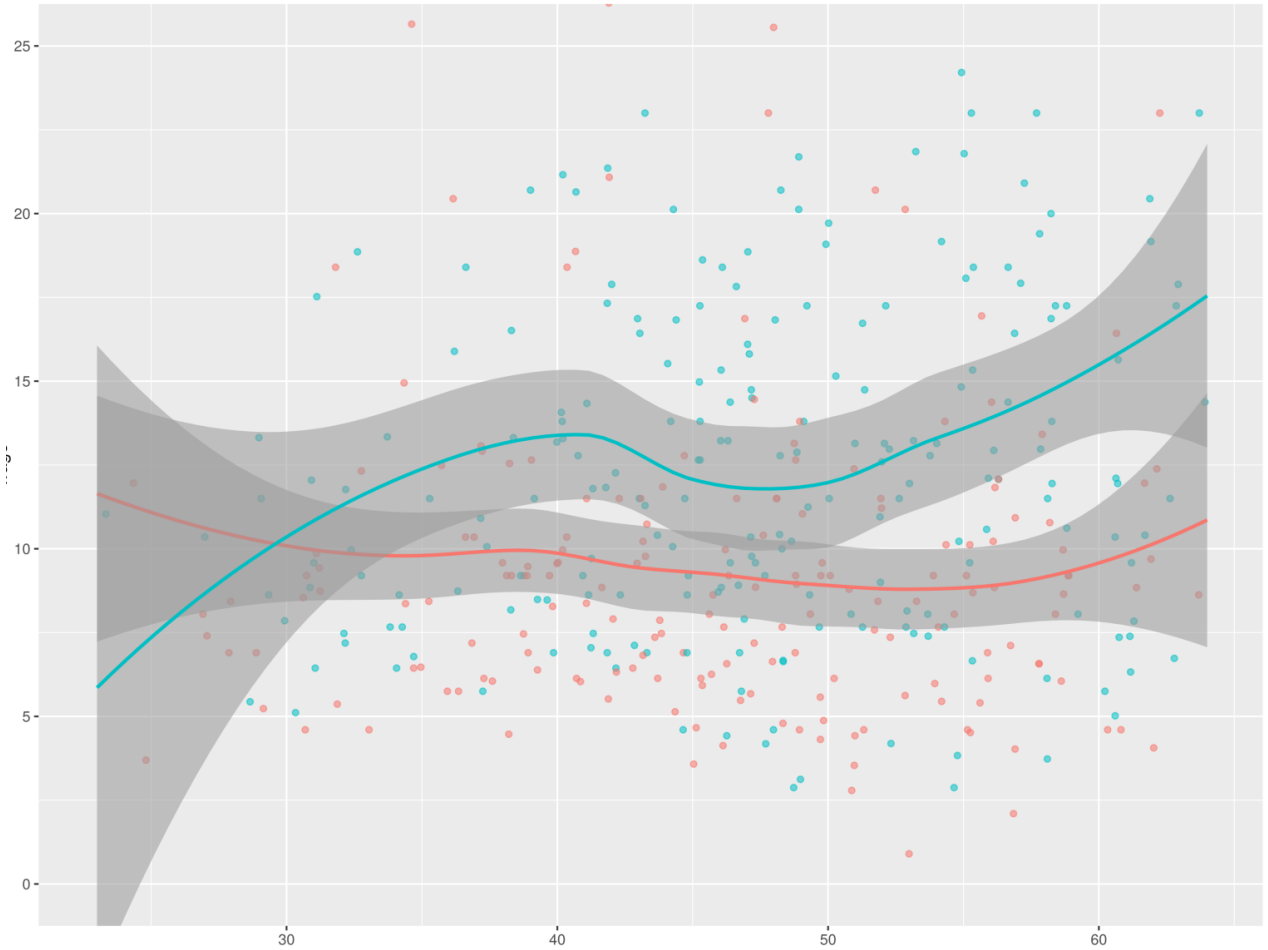
… among many others

Estimating effects as a regression tasks

- Estimate effect of variable on outcome
- Produce easily interpretable statistics
- Applicable to any (ML) model that yields individual predictions (i.e. model agnostic)

Regression tasks

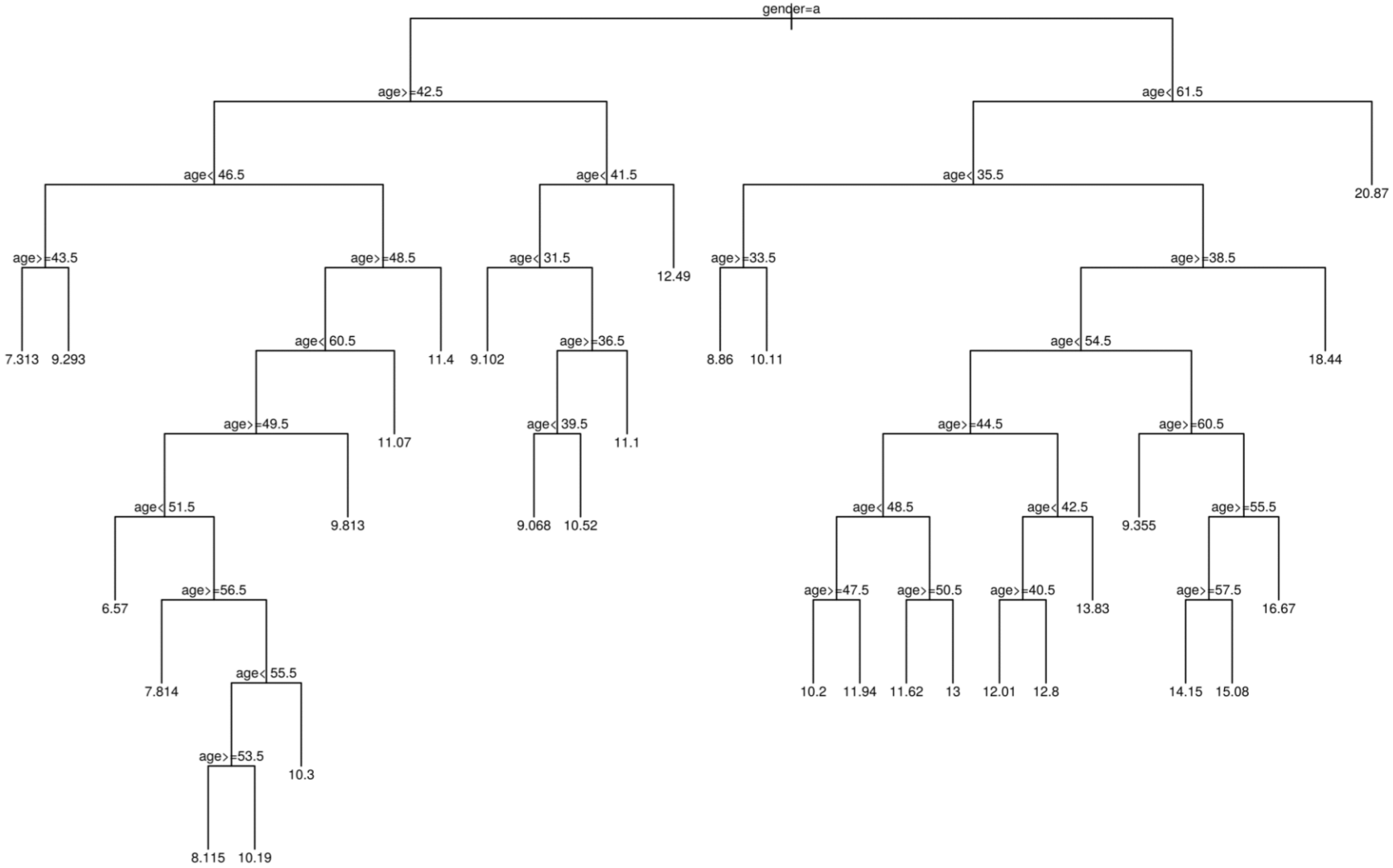
Example data



Hourly wage by gender and age (ISSP Germany 2012)

Regression tasks

Example learner/model: regression tree



Average marginal effects

1. Estimate (complex, multivariate) model
2. Set all cases in sample to first value for variable (e. g. gender = 0 “female”)
3. Predict response for all cases
4. Set all cases in sample to next value for variable (e. g. gender = 1 “male”)
5. Predict response for all cases – again
6. Calculate AME / APE as mean difference between two predictions for all cases

Average marginal effects

	Linear Model	Regression Tree
Gender (male vs. female)		
AME / APE	3.61 €	3.68 €
95 % CI Lower Bound	2.43 €	2.42 €
95 % CI Upper Bound	4.65 €	4.91 €
Age (per year)		
AME / APE	0.32 €	0.75 €
95 % CI Lower Bound	-0.05 €	0.09 €
95 % CI Upper Bound	0.66 €	1.32 €

Data: ISSP Germany 2012; CIs: bootstrap estimates

Recent and next steps

- **„ame“ R package (Average Marginal Effects)**

  Giuseppe Casalicchio (Comp. Statistics, LMU)

  https://github.com/compstat-lmu/ame

- Integrates with mlr – „Machine Learning in R“ – a very comprehensive ML framework

  Bernd Bischl et al. (*Comp. Statistics, LMU)*

  https://github.com/mlr-org/mlr

- (Still) Work in progress:

  Paper discussing approach and implementation (Beste, Bethmann & Casalicchio)

Pass by at our booth