Appendix

Further dimensions of model performance

Dimensions of Model Performance

Many factors determine the value of an analytical model

Accuracy

How well does the model predict? For example, is it able to distinguish good and bad risks with high accuracy?

Scalability

How much time is needed to build and to apply the model? Does it scale to large data sets?

Robustness

Can the model cope with noise and missing values? How about irrelevant and correlated attributes?

Comprehensibility

Can we understand the model? Is it clear how it transforms attribute values into predictions of the response variable?

Justifiability

Is the use of attributes within the model in line with business rules/ understanding?

Calibration

For probability forecasts!
Out of 100 events predicted to have 90% chance, about 90 should have occurred. True?

Dimensions of Model Performance
Scalability

- **Consumption of time resources**
- **Time needed to build model (training time)**
  - Depends on number of cases and attributes
  - Run-time complexity
  - Importance depends on update frequency
- **Time needed to generate predictions**
  - Much less than training time
  - Critical in real-time settings (e.g., E-Commerce)
- **Both time factors differ substantially across algorithms**

- **Consumption of memory resources**
  - During model building
  - When storing final model
  - Big data prohibits keeping all training data in memory
- **Sensitivity with respect to hyperparameters**
  - Building one model is never enough
  - Some models need a lot more tuning than others
- **Parallelization important**
  - Model building
  - Model tuning

Dimensions of Model Performance
Robustness

- **Real-world data is noisy**
  - Missing values
  - Erroneous data entries
  - Wrong labels
  - Irrelevant / correlated attributes
- **Real-world phenomena change over time**
  - Concept drift
  - Model recalibration versus re-estimation
- **How to these factors affect the model?**
  - During model building
  - After model building

Dimensions of Model Performance

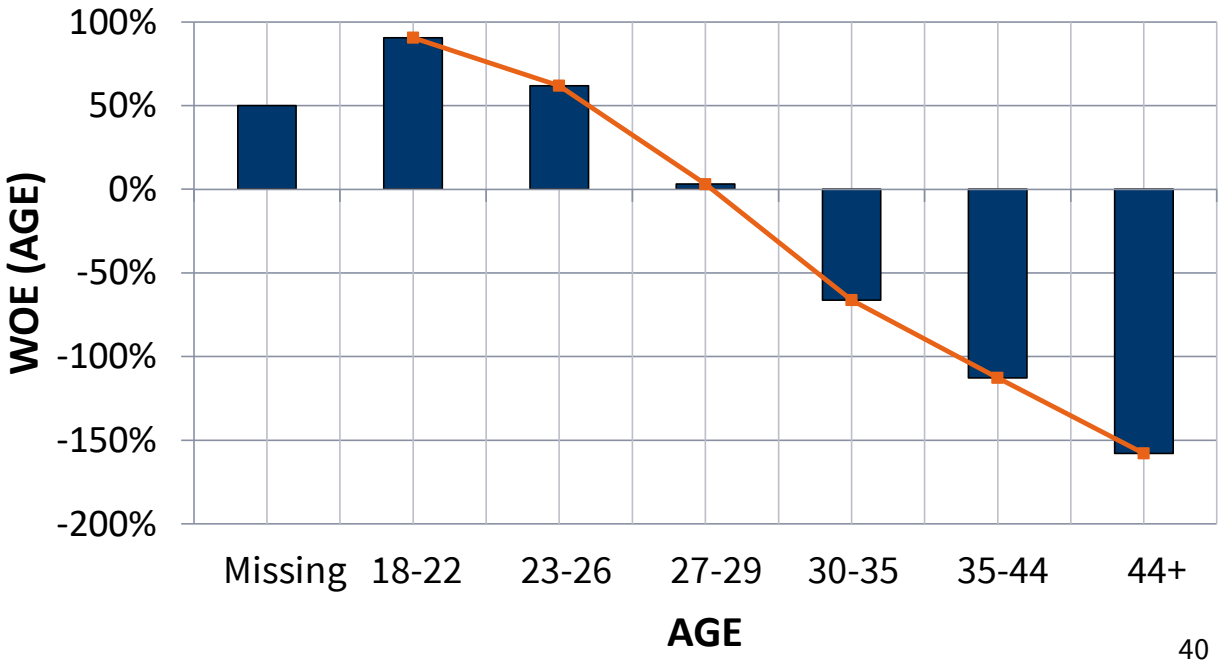
Comprehensibility: crucial and challenging to measure

- **Is it possible to understand how a model translates attribute values into prediction?**
  - Alternative terms: interpretability, transparency, white-box (vs. black-box) model
  - Becoming increasingly relevant with the raising popularity of machine learning
  - “Managers don’t trust black-box models”
- **New research fields on interpretable machine learning (see subsequent sessions)**
  - Global interpretability: equivalent to above point. How do covariates govern predictions
  - Local interpretability: how was the prediction of a specific observation determined by covariate values
- **Prediction versus insight and correlation versus causality**
  - Prediction: “Next month, we sell 100 laptops”
  - Insight: “Sales increase by 2% if we lower prices by €50”
  - Standard machine learning models are correlational

Dimensions of Model Performance

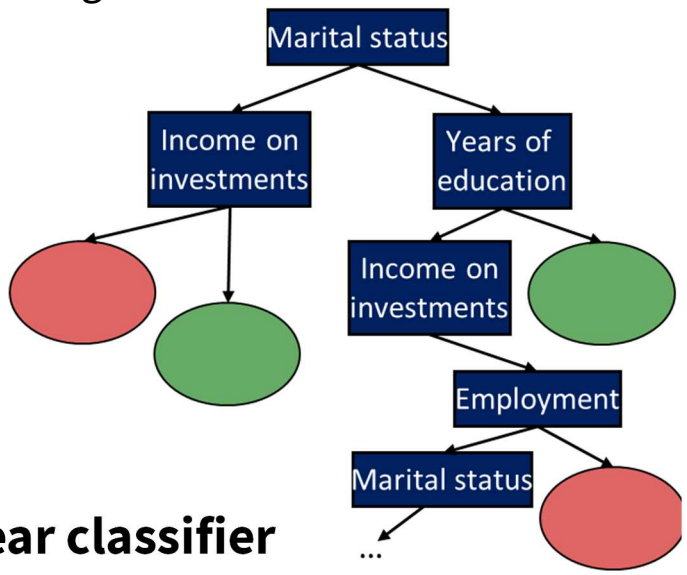
Justifiability: a key driver of model acceptance in industry

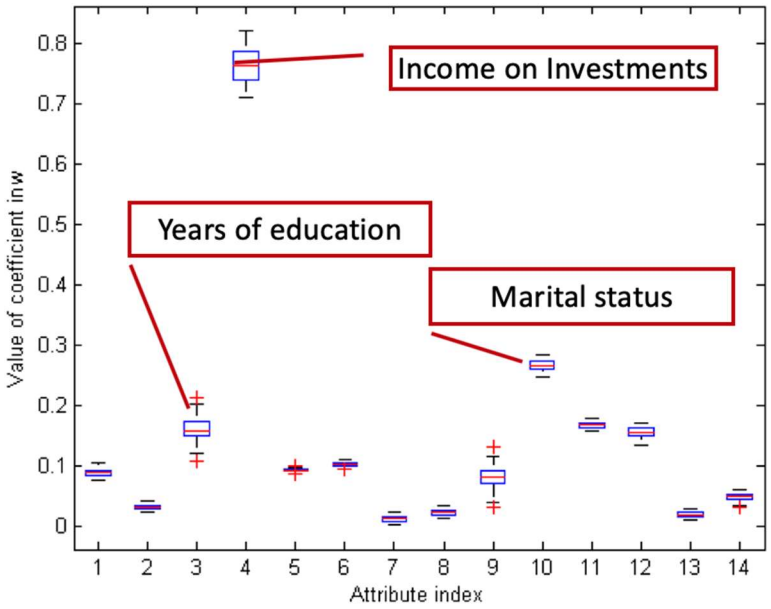
- **Does the way in which attribute values affect predictions agrees with prior beliefs or business rules?**
  - Exemplary business rules: sales decrease with price, long-term customers are more profitable than new customers, etc.
  - Requires interpretability
- **Credit risk example**
  - Business rule: credit risk decreases with age
  - Test: does WOE show this trend



Dimensions of Model Performance
Comprehensibility / Justifiability Example

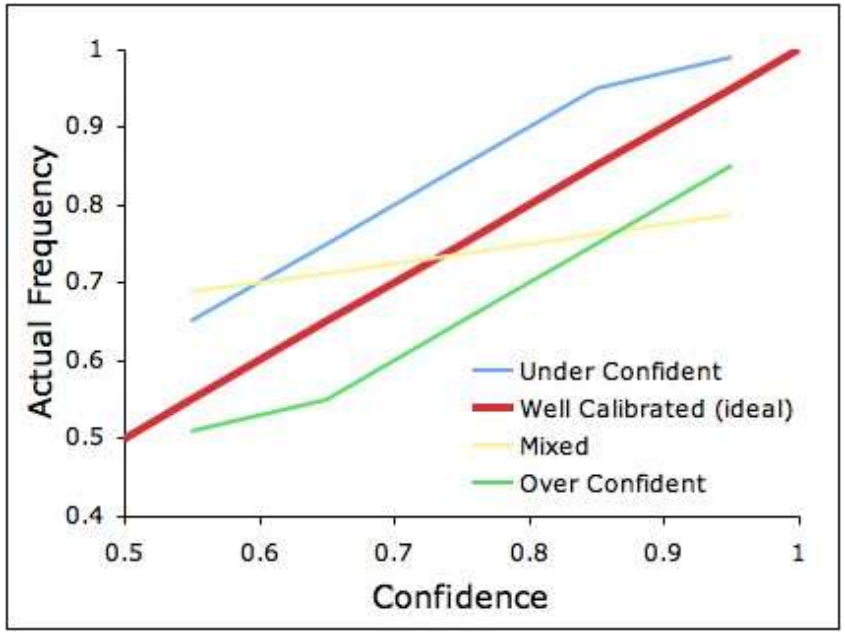
- **US Census data set from UCI library** (https://archive.ics.uci.edu/ml/datasets/Adult)
- **Classification task**
  - Is household income below or above $50,000 p.a.?
  - Fourteen attributes describing a household
    - Marital status
    - Working hours
    - Academic degree
    - Years of education
    - Country of origin
    - Income on investments
    - Employment
    - …
- **Result of tree and linear classifier**





Dimensions of Model Performance
Calibration

- **Feature of probabilistic predictions**
- **Credit Scoring Example**
  - Model makes risk forecasts for 100 credit applications
  - Forecasts are all the same and predict default of 90%
  - Then, we should eventually observe 90 actual defaults
- **For prediction models**
  - Calibration can be poor
  - Special treatment needed
  - See, e.g., Bequé et al. (2017)



[https://goodmorningeconomics.wordpress.com/2008/07/11/calibrated-probability-assessmentorg/]