Building meaningful machine learning models for disease prediction

Dr Shirin Glander

Dep. of Genetic Epidemiology
Institute of Human Genetics
University of Münster

shirin.glander@wwu.de

https://shiring.github.io
https://github.com/ShirinG

31.03.2017

Table of contents

About me

2005 - 2011 BSc and MSc of Science in Biology
Evolutionary genetics,
immune memory in Drosophila

2011 - 2015 PhD in Biology
Is the immune system of plants required to adapt to flowering time change?

since 2015 Bioinformatics Postdoc
Autoinflammatory diseases & innate immunity
Next Generation Sequencing

What makes a model meaningful?

What makes a model meaningful?

Machine learning is a powerful approach for developing sophisticated, automatic, and objective algorithms for analysis of high-dimensional and multimodal biomedical data.

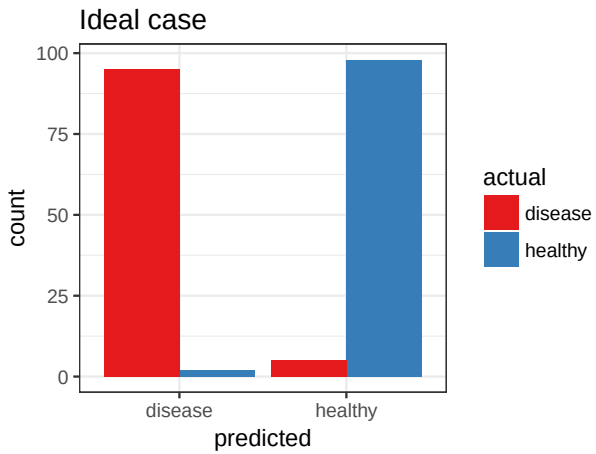
A key aspect of the precision medicine effort is the development of informatics tools that can analyze and interpret 'big data' sets in an automated and adaptive fashion while providing accurate and actionable clinical information.

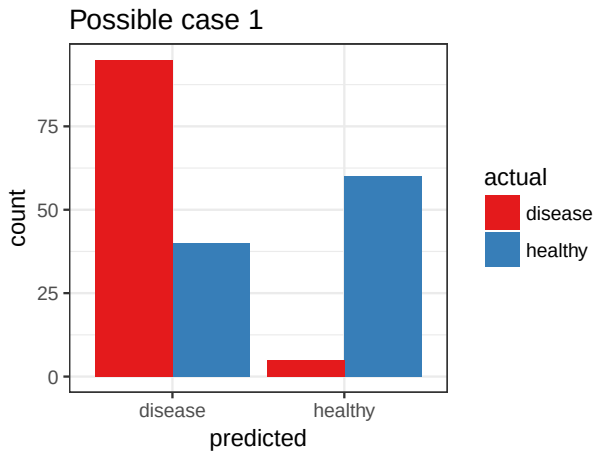
ML based model can improve detection, diagnosis, and therapeutic monitoring of disease.

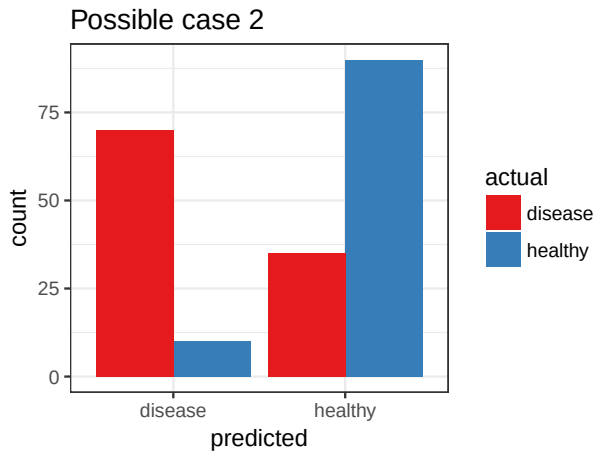
Meaningful models

- answer the question(s) posed...
- ... with sufficient accuracy to be trustworthy

Accuracy depends on the problem!

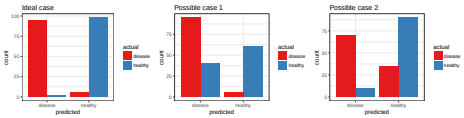






Meaningful models
- answer the question(s) posed...
- ... with sufficient accuracy to be trustworthy

Accuracy depends on the problem!



I want to begin with an introduction to the main question of my talk: what makes a model *good* or *meaningful*?

A meaningful model

- answers the questions or questions posed...
- ... with sufficient accuracy to be trustworthy

But what we mean exactly with *accuracy* can not be defined with a one-size-fits-all approach: it depends on the problem we want to model.

Let me illustrate what I mean with the following examples:

- **Ideal case:** Of course, we all want to achieve ideal modeling results where overall prediction accuracy is very high. With a model like that, we can be very confident that a healthy person is indeed healthy and a sick person is not.
- But in reality, we often achieve prediction accuracies that are much less nice.
- This forces us to evaluate how we want to define a good model
- Scenarios 1 and 2: Le's consider two possible scenarios:
- in case 1, we can be very confident that a person who got classified as "healthyïs indeed healthy,
  while a person who has been diagnosed as diseased might as well be healthy based on these prediction accuracies
- in case 2, it is the other way around.
- We now need to make a decision which scenario is better and in which direction we want to optimize our model:
  do we rather want to refer a few healthy people for further checking because the model predicted them as diseased?
  Or do we rather want to be as certain as possible that a predicted disease is actually true
  and accept that we might miss a few disease cases?

Machine Learning (ML) in disease modeling

ML in disease modeling

- falls into the field of artificial intelligence (AI)
- algorithms learn by being trained on observed data
- learned models can predict unknown data
- ML concepts are not new, but the increase in computational capacity has made them more accessible

Examples:

- Computer-aided diagnosis of breast cancer from mammograms,
- early diagnosis of osteoporosis from chest radiographs, etc.[1]
- Identifying signatures of Brain Cancer from MRSI[2]

---

[1] K. Doi (2007). "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31.4-5, S. 198–2011.

[2] P. Sadja (2006). "Machine Learning for Detection and Diagnosis of Disease". In: *Annual Review of Biomedical Engineering* 8.1. PMID: 16834566, S. 537–565.

ML in disease modeling

- falls into the field of artificial intelligence (AI)
- algorithms learn by being trained on observed data
- learned models can predict unknown data
- ML concepts are not new, but the increase in computational capacity has made them more accessible

Examples:

- Computer-aided diagnosis of breast cancer from mammograms,
- early diagnosis of osteoporosis from chest radiographs, etc.[1]
- Identifying signatures of Brain Cancer from MRSI[2]

[1] K. Doi (2007). "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31.4-5, S. 198–2011.

[2] P. Sadja (2006). "Machine Learning for Detection and Diagnosis of Disease". In: *Annual Review of Biomedical Engineering* 8.1. PMID: 16834566, S. 537–565.

Computer-aided diagnosis (CAD) has become one of the major research subjects in medical imaging and diagnostic radiology
With CAD, radiologists use the computer output as a 'second opinion' and make the final decisions.

In vivo magnetic resonance spectroscopy imaging (MRSI) allows noninvasive characterization and quantification of molecular markers of potentially high clinical utility for improving detection, identification, and treatment for a variety of diseases, most notably brain cancers.

1. Unsupervised matrix decomposition methods, such as nonnegative matrix factorization, which impose general, although physically meaningful, constraints, are able to recover biomarkers of disease and toxicity, generating a natural basis for data visualization and pattern classification.

2. Supervised discriminative models that explicitly address the bias-variance trade-off, such as the support vector machine, have shown great promise for disease diagnosis in computational biology, where data types are disparate and high dimensional.

3. Generative models based on Bayesian networks offer a general approach for biomedical image and signal analysis in that they enable one to directly model the uncertainty and variability inherent to biomedical data as well as provide a framework for an array of analysis, including classification, segmentation, and compression.

A quick recap of ML basics

Supervised vs Unsupervised algorithms

Classification vs Regression

Features

- feature selection
- feature extraction
- feature engineering

Features

- feature selection
- feature extraction
- feature engineering

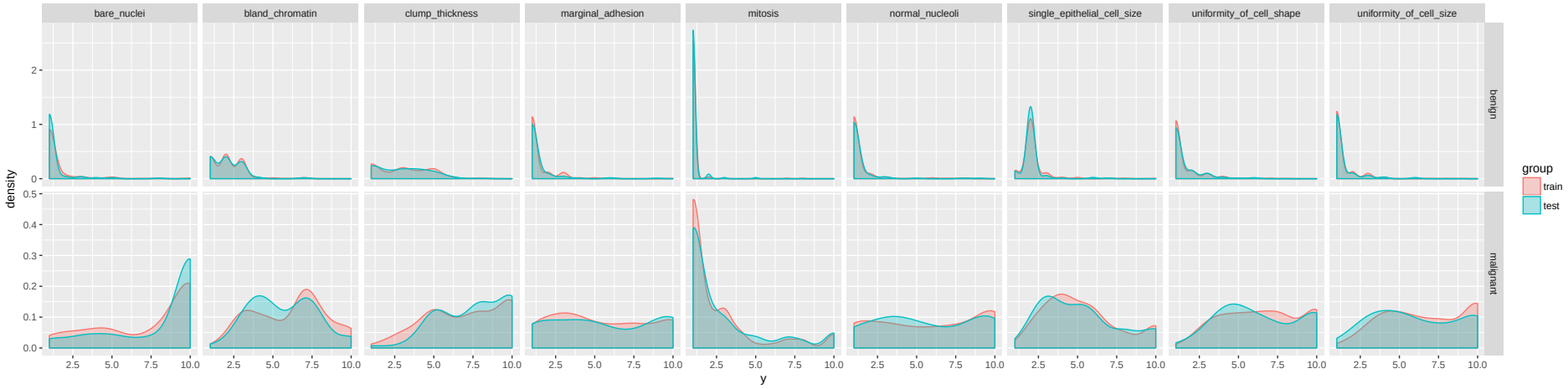
extraction of salient structure in the data that is more informative than the raw data itself (the feature extraction problem)

How to build ML models in R

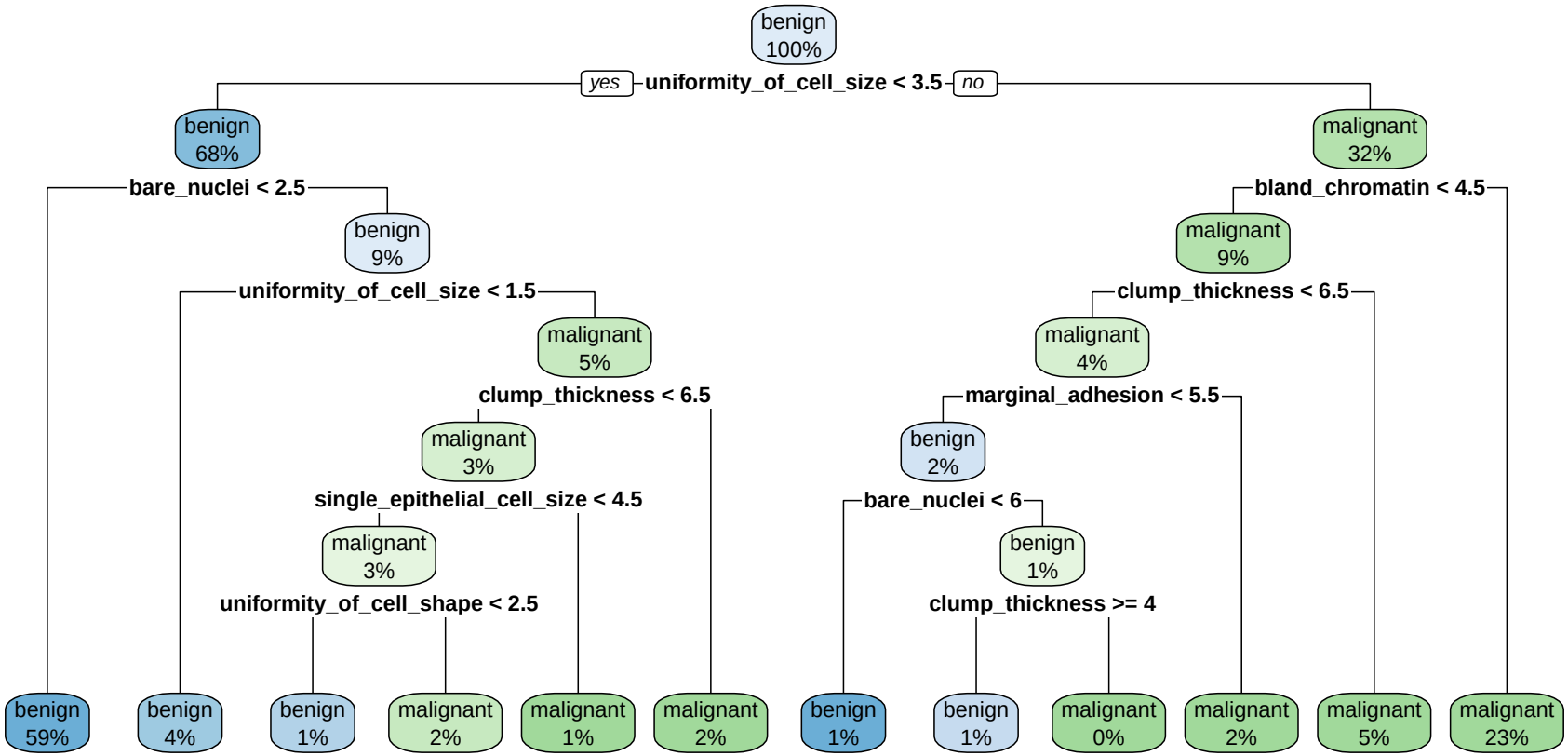
Session setup

Code will be available on my website and on Github
Breast cancer Wisconsin dataset
caret package
h2o package

Distribution

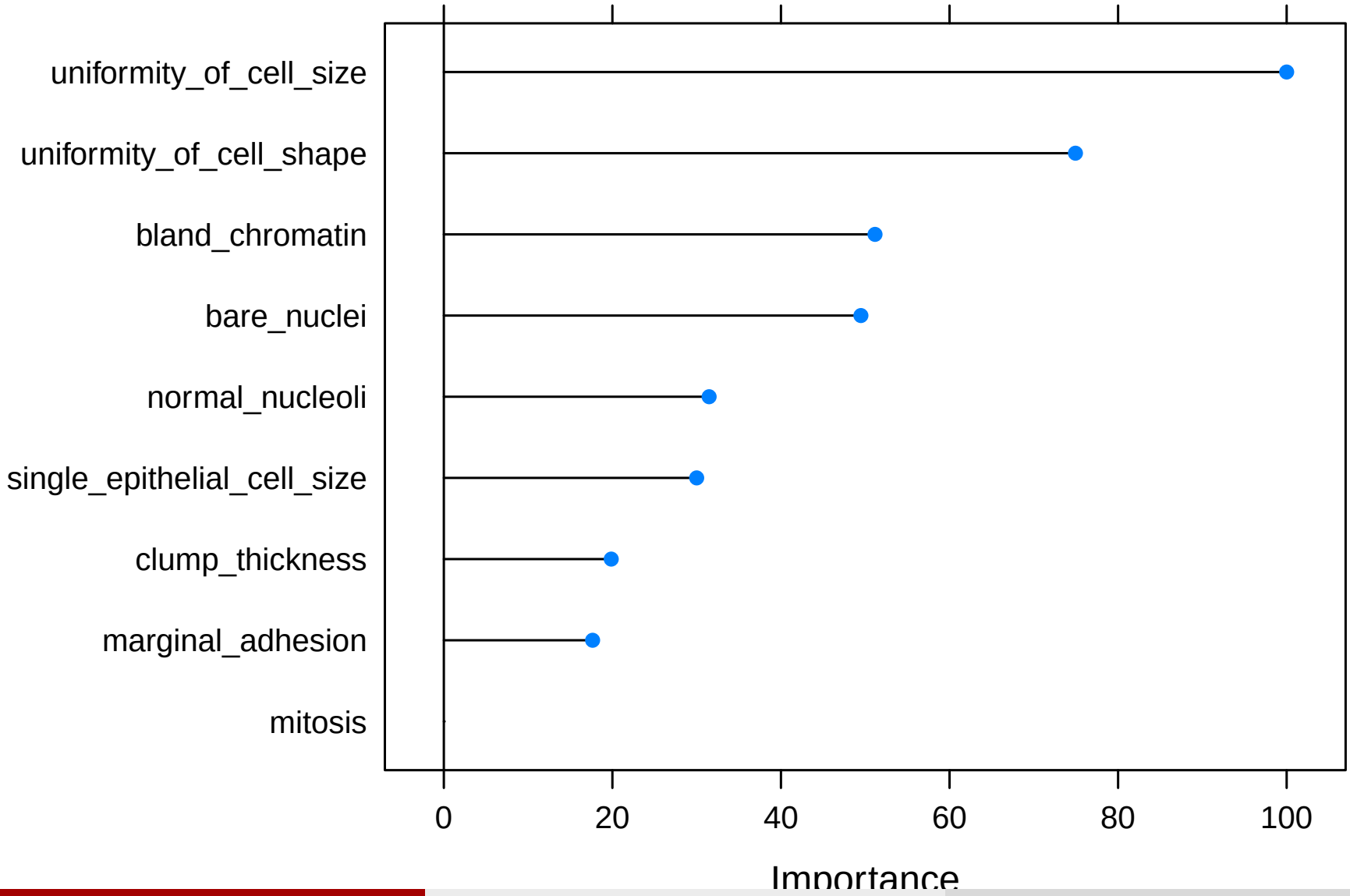


Decision trees

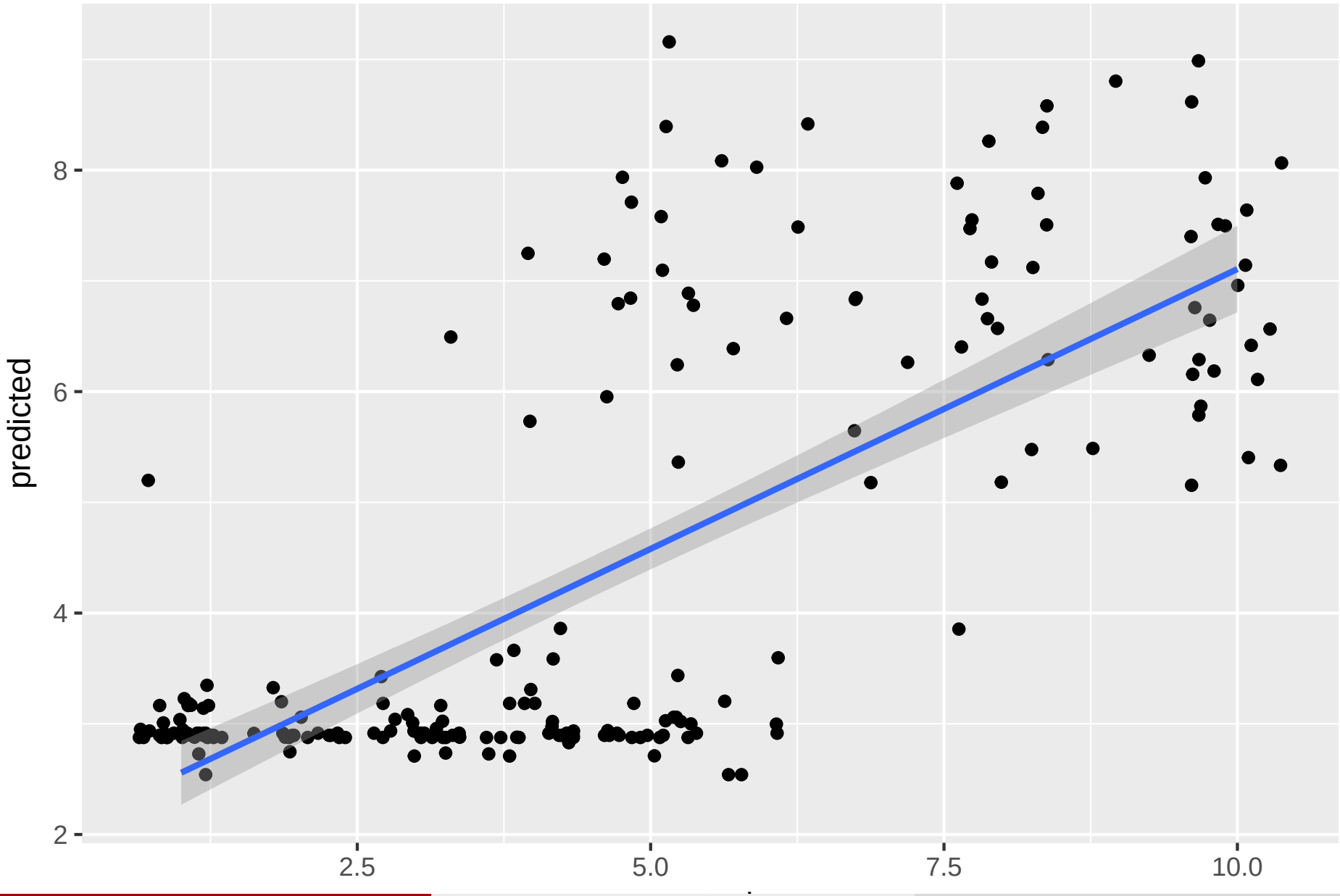


Random Forest

Feature importance

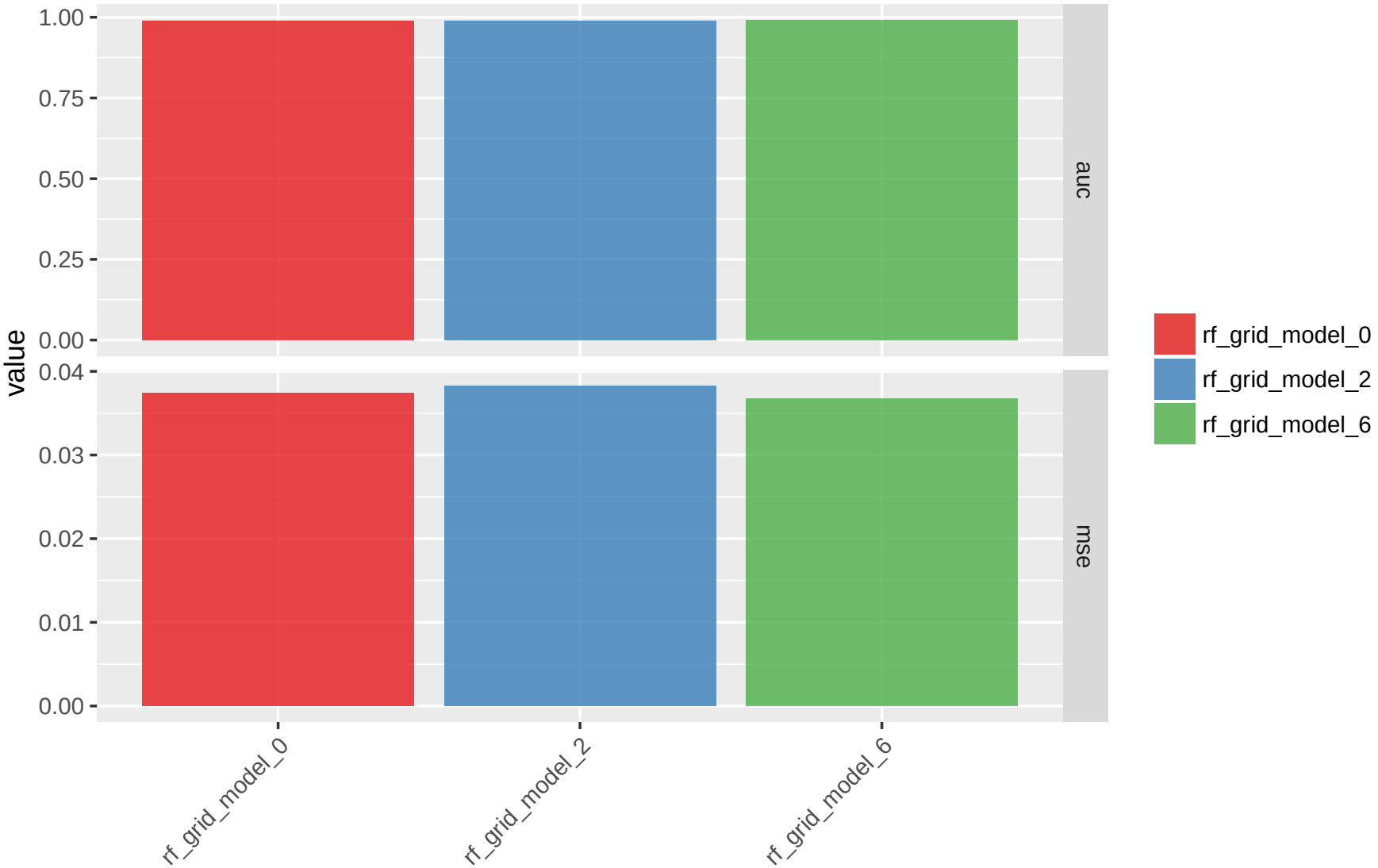


(Generalized) Linear Models



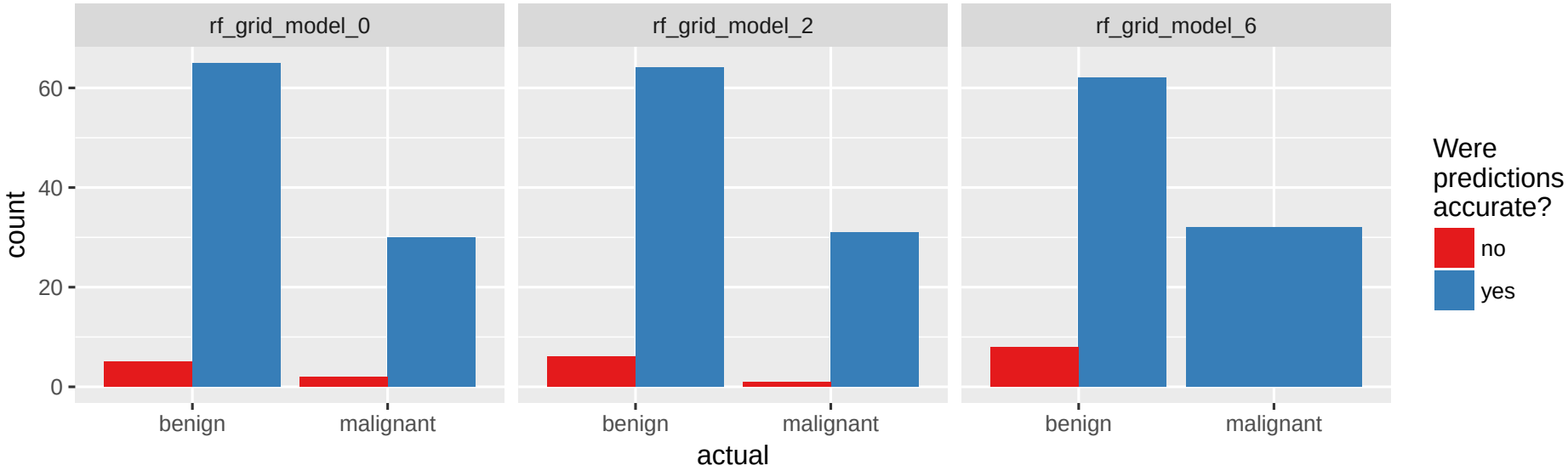
Evaluating ML model performance

AUC and MSE

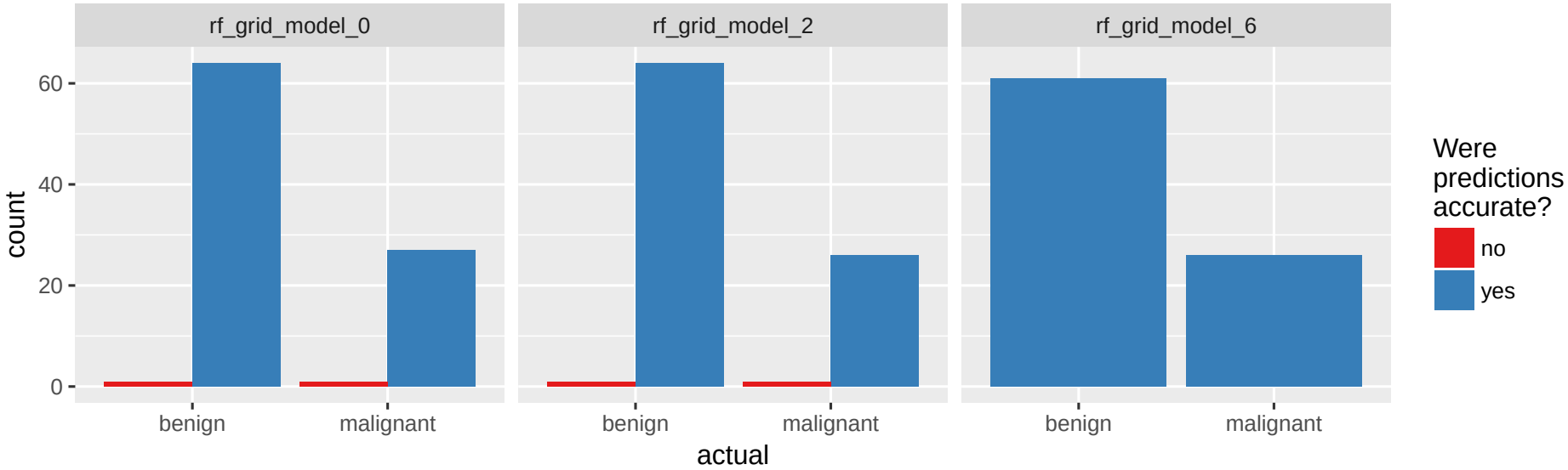

value
1.00
0.75
0.50
0.25
0.00
0.04
0.03
0.02
0.01
0.00
auc
mse
rf_grid_model_0
rf_grid_model_2
rf_grid_model_6
rf_grid_model_0
rf_grid_model_2
rf_grid_model_6

Predictions on test data

Default predictions



Stringent predictions



Thank you for your attention!

Questions?

Slides and code will be available on Github:
https://github.com/ShirinG/Webinar_ML_for_disease

Code will also be on my website: https://shiring.github.io

shirin.glander@wwu.de