WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Data Science and Machine Learning

Info Session

Juan Caballero, Bettina Grün, Lukas Schmoigl,
June 6, 2024

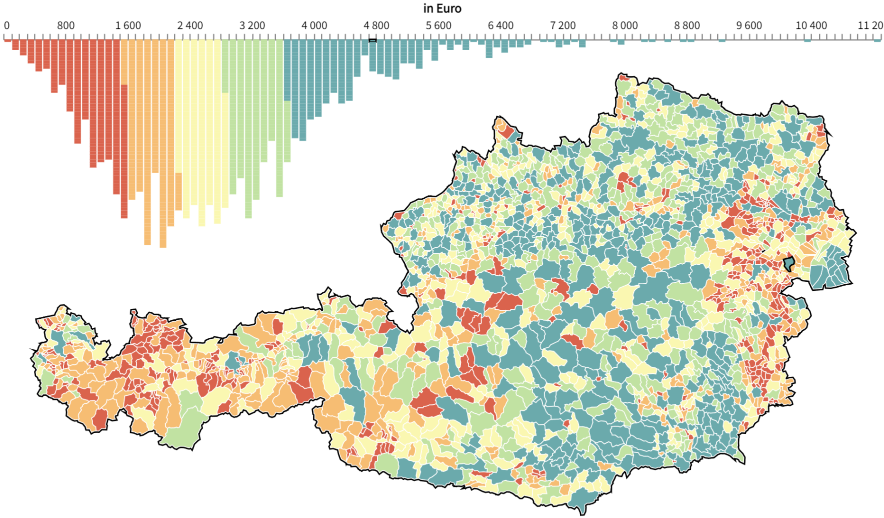
Covered Topics: Field Course

- Data Wrangling and Exploratory Data Analysis
- (Interactive) Data Visualization
- Introduction into Supervised Learning and Cross Validation
- Random Forest and Boosted Regression Trees
- Introduction to Unsupervised Learning
- Mixture Models
- PCA and Factor Models
- NLP and Text Classification
- Topic Models

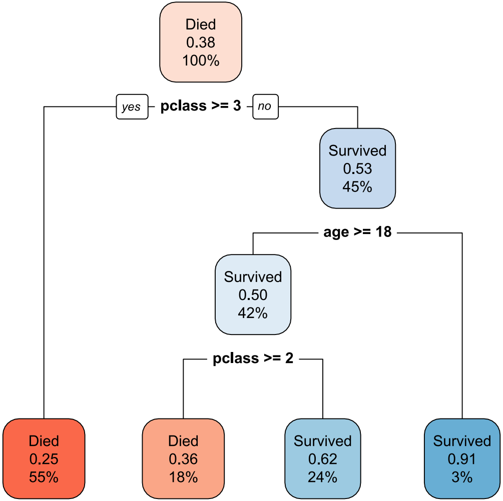
Additional Topics: Seminar

- Setup (Markdown + Git + Data Storage)
- Webscraping
- Spatial Data Visualization
- More Unsupervised Learning
- Dashboard Building
- Sentiment Analysis

Data Visualization



Machine Learning



Died
0.38
100%

yes — pclass >= 3 — no

Survived
0.53
45%

age >= 18

Survived
0.50
42%

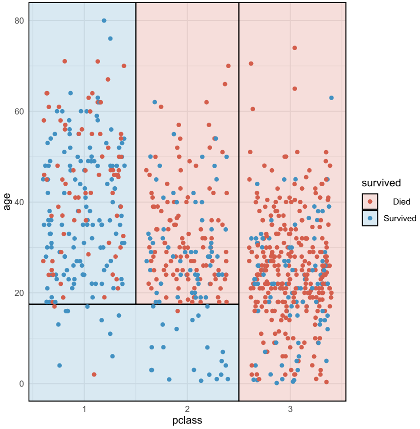
pclass >= 2

Died
0.25
55%

Died
0.36
18%

Survived
0.62
24%

Survived
0.91
3%



Class Design

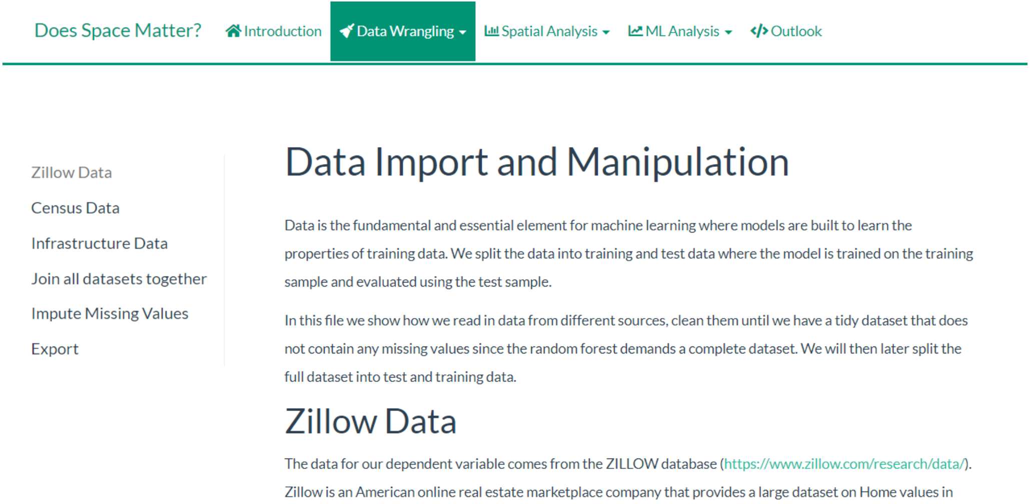
Field Course:

- Lectures and in-class applications
- Grading
  - First Exam (40%)
  - Second Exam (40%)
  - "Kaggle-style" competition presentation and results (20%)

Seminar:

- Exercises and applications of methods using R
- Grading
  - Project presentation 1 (30%)
  - Project presentation 2 (40%)
  - Assignments (30%)

Project Example



Data Import and Manipulation

Data is the fundamental and essential element for machine learning where models are built to learn the properties of training data. We split the data into training and test data where the model is trained on the training sample and evaluated using the test sample.

In this file we show how we read in data from different sources, clean them until we have a tidy dataset that does not contain any missing values since the random forest demands a complete dataset. We will then later split the full dataset into test and training data.

Zillow Data

The data for our dependent variable comes from the ZILLOW database (https://www.zillow.com/research/data/). Zillow is an American online real estate marketplace company that provides a large dataset on Home values in

Prerequisites

- Coding
  - Examples and assignments will be given using R.
  - Basic programming skills in R or a similar statistical programming language (e.g., Python, Julia) are required.
- Mathematics and Statistics
  - Basic understanding of probability, statistics, linear algebra and calculus is necessary.
  - However this is an applied class: no formal proofs.
  - Targeted learning outcome: practical understanding of statistical modeling and working with data.