Data Science and Machine Learning

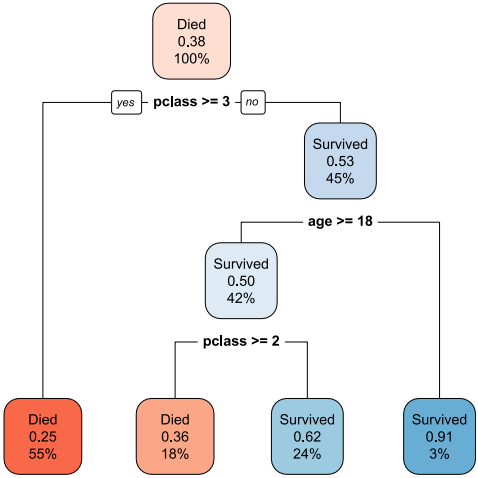
Info Session

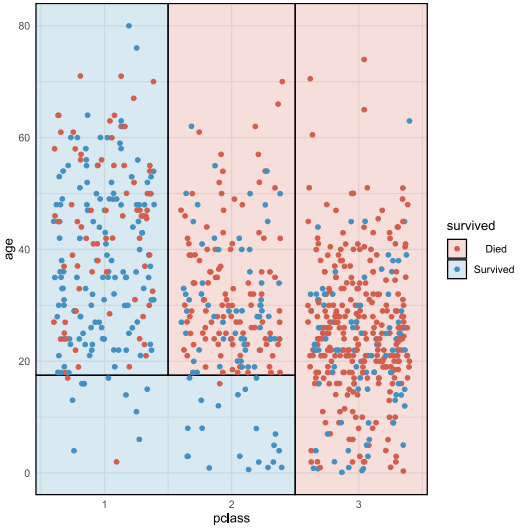
Katharina Fenz, Thomas Mitterling,
Lukas Schmoigl, Maximilian Thomasberger,
June 8, 2022

Covered Topics

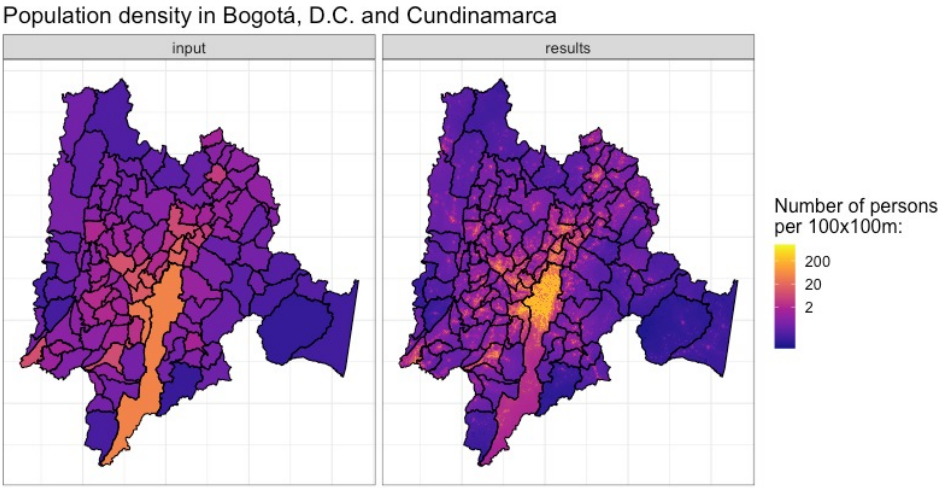
- Data Wrangling and Collaborative Tools
- Exploratory Data Analysis and Data Visualization
- Supervised Learning and Cross Validation
- Spatial Methods and Spatial Data Visualization
- Random Forest and Boosted Regression Trees
- (Spatial) Bayesian Model Averaging
- Unsupervised Learning and Clustering Algorithms
- Text Mining and Text Classification
- Additional Topics (e.g. web scraping, geocoding, interactive data viz)

Example: Classification Tree





Example: Spatial Methods



Class Design

Field Course:

- Introduction into data science and machine learning methods
- Grading
  - Exams (x2)
  - "Kaggle-style" competition assignment
  - Class participation

Seminar:

- Exercises and applications of methods using R
- Grading
  - Project paper and presentation
  - Assignments
  - Class participation

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 language 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 and no advanced mathematics
  - Ideally you bring a practical understanding of statistical modelling and working with data