

DATA SCIENCE 2

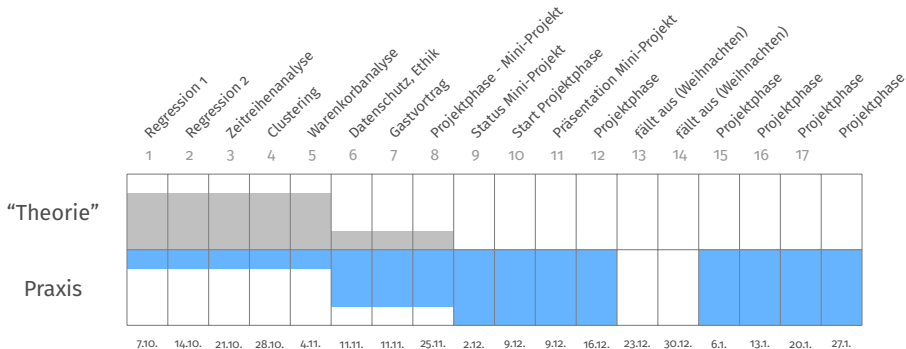
PROJEKTPHASE 2

PROF. DR. CHRISTIAN BOCKERMANN

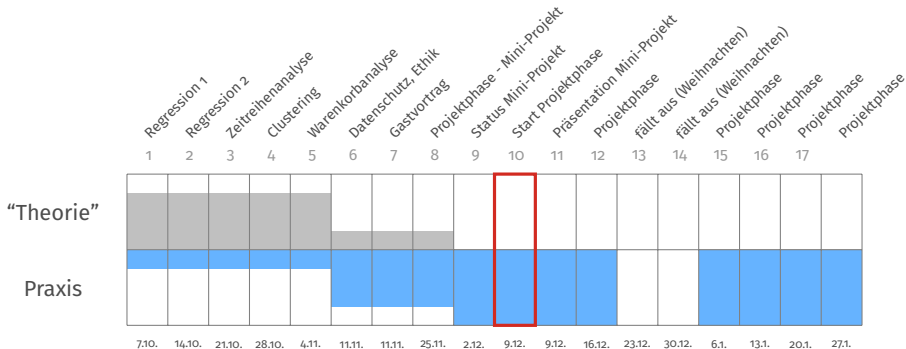
HOCHSCHULE BOCHUM

WINTERSEMESTER 2021/2022

Themen der Vorlesung



Themen der Vorlesung



- 1 Mini-Projekt: Feedback + Erfahrungsaustausch
- 2 Projektphase – Teil 2
- 3 Weitere Vorlesungen

Mini-Projekt: Feedback + Erfahrungsaustausch

Gruppe 1

- Kim Sophie Schreiber, Jan Cornehl, Christian Wolter

Gruppe 2

- Vivien Huynh, Harjot Kaur

Gruppe 3

- Gregor Schumacher, Muhammad Isa Wirasomantri

Gruppe 4

- Leon Kuntz, Tom Magiera

Gruppe 5

- Said Albayram, Biranavan Sakthivel, Sefa Akhoy

Diskussion

- Welche Lernverfahren haben Sie ausprobiert?

Diskussion

- Welche Lernverfahren haben Sie ausprobiert?
- Welche Merkmale wurden benutzt?
- Umgang mit nicht-numerischen Merkmalen?
- Quellen für weitere Lernverfahren (RF, LogReg,...)?

Diskussion

- Welche Lernverfahren haben Sie ausprobiert?
- Welche Merkmale wurden benutzt?
- Umgang mit nicht-numerischen Merkmalen?
- Quellen für weitere Lernverfahren (RF, LogReg,...)?
- Welche **Probleme** hat es gegeben?

Projektphase – Teil 2

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: **Individuelle Prüfungsleistung?**

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: **Individuelle Prüfungsleistung?**

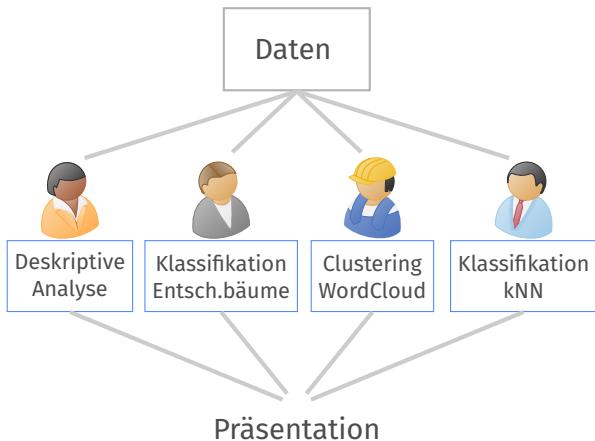
Modus B

- Freie Wahl des Datensatzes
- Je Gruppe unterschiedliche Fragen/Aspekte
- Jeder Teilnehmer bearbeitet u.a. eigenständigen Aspekt
- Gemeinsame Präsentation der Analyse des Datensatzes

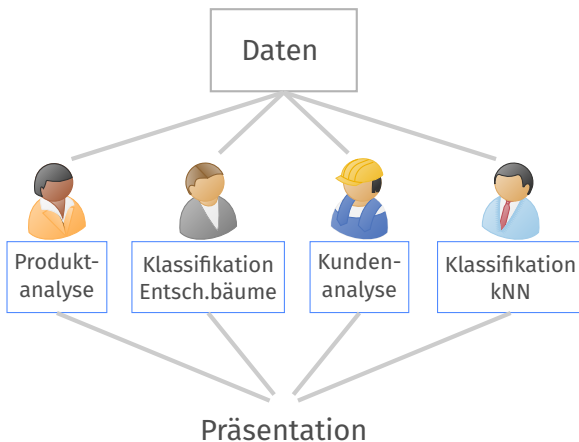
Abschlussprojekte

- Daten-/Themenvergabe heute
- Mögliche Themen/Datensätze stelle ich gleich vor
- Am Ende 1 Präsentation je Gruppe
- Jeder Teilnehmer lädt Notebook mit Analyse (ausführlich!)
bis 13.2. 23:59 Uhr in Moodle-Kurs hoch
- Präsentationen am 14.2.2022

Abschlussprojekte



Abschlussprojekte



Abschlusspräsentation

- Termin: 14.2.2022

Abschlusspräsentation

- Termin: 14.2.2022

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Ordentliches Notebook (Visualisierungen!)
- Auch erfolglose Modelle (mit Begründung!) gut
- Bewertet wird Präsentation + Hausarbeit
- “Leitfaden” für Projektphase 2:

<https://datascience.hs-bochum.de/Vorlesung/WS2122/>

Mögliche Datensätze

- California House Pricing, House Prices Advances

<https://www.kaggle.com/c/california-house-prices/overview>

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

- Predict Future Sales for Store/Product

<https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview>

<https://www.kaggle.com/c/instacart-market-basket-analysis/data>

- Natural Language Processing with Disaster Tweets

<https://www.kaggle.com/c/nlp-getting-started>

- Inside AirBnB

<http://insideairbnb.com/get-the-data.html>

- RKI Covid19

[https:](https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0)

[//npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0](https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0)

Vorhersage von Immobilienpreisen

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 11,285 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules [Join Competition](#)

Overview

Description

Evaluation

Tutorials

Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Vorhersage von Verkäufen

Featured Prediction Competition

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

3,298 teams · 6 years ago

\$35,000
Prize Money

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Late Submission](#)

Overview

Description

Evaluation

Prizes

Timeline

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

<https://www.kaggle.com/c/rossmann-store-sales/overview>

Meint der Tweet eine Katastrophe? Ja/Nein

Getting Started Prediction Competition

Natural Language Processing with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not

Kaggle · 3,114 teams · Ongoing

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Predictions

Overview

Description

Evaluation

FAQ

Welcome to one of our "Getting Started" competitions 🙌

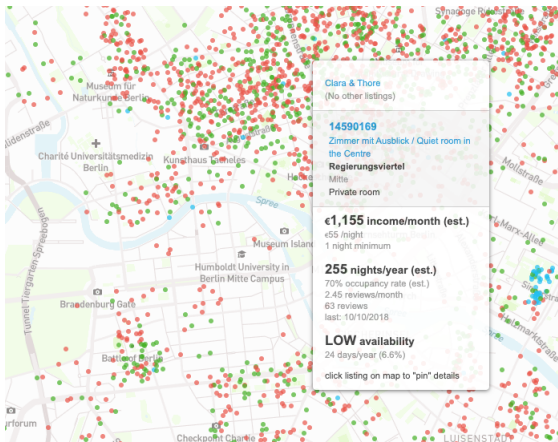
This particular challenge is perfect for data scientists looking to get started with Natural Language Processing. The competition dataset is not too big, and even if you don't have much personal computing power, you can do all of the work in our free, no-setup, Jupyter Notebooks environment called [Kaggle Notebooks](#).

Competition Description

Twitter has become an important communication channel in times of emergency.

<https://www.kaggle.com/c/nlp-getting-started>

Analyse von AirBnB-Daten (z.B. Berlin)



<http://insideairbnb.com/get-the-data.html>

Mögliche Fragestellungen

- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter?
(Hosts mit mehr als 2 oder 3 Angeboten)

Mögliche Fragestellungen

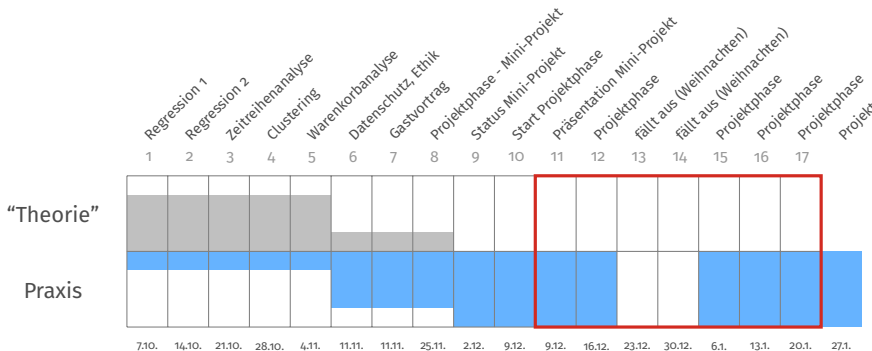
- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter? (Hosts mit mehr als 2 oder 3 Angeboten)

Advanced

- Wie unterscheiden sich Preise/Angebote von Stadt A und Stadt B?

Weitere Vorlesungen

Die weiteren Vorlesungstermine



Zeitplan Projektphase – 2021

9.12.2021 10 Uhr – **Start der Projektphase**

- “Präsentation” Mini-Projekte
- Vorstellung Datensätze für Projektphase

16.12.2021, 10 Uhr – **Visualisierung mit Python**

- Status Bericht aus den Gruppen
- Vortrag: Visualisierungen mit Seaborn

Zeitplan Projektphase – 2022

6.1.2022, 10 Uhr – Was ist Big Data?

- Status: Bericht aus den Gruppen
- Vortrag: Was ist Big Data?

13.1.2021, 10 Uhr – Use-Case: Bot-Detection

- Status: Bericht aus den Gruppen
- Vortrag: Bot-Detection Use-Case

20.1.2021, 10 Uhr – ??

- Status: Bericht aus den Gruppen

27.1.2021, 10 Uhr – ??

- Status: Bericht aus den Gruppen

Zeitplan Projektphase (Abschlussprojekt)

- 16.12.2021 Besprechung der finalen Datensätze je Gruppe
- 13.2.2022 bis 23:59 Uhr Abgabe Hausarbeit (Jupyter Notebook)
- 14.2.2022 ab 10 Uhr: Präsentationen der Projekte