Machine Learning in der Berufsunfähigkeitsversicherung?

Eine Analyse von Risikofaktoren

Forum V-Versicherungsmathematisches Kolloquium an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Christian Eckert, Daniela Giesinger

29.06.2021

Data Science Challenge

Was ist das?

- Ins Leben gerufen vom Ausschuss „Actuarial Data Science“ der DAV
- Ziel: Mehr Aktuar*innen für die Beschäftigung mit Data-Science Fragen zu begeistern

Data Science Challenge

Anforderungen 2020

Anforderungen der Data Science Challenge 2020.

- Erstellung eines Python-Notebooks zu einem beliebigen Thema im Bereich Actuarial Data Science
- Nutzung eines öffentlichen Datensatzes

Data Science Challenge

Unser Team (2020 ein Team der NÜRNBERGER Versicherung)

Christian Eckert

Daniela Giesinger

Felix Müller

Antonia Schöning

Data Science Challenge

Weitere Infos über unser Projekt ...

Data Science Challenge

Auch 2021 gibt es wieder eine Data Science Challenge ...

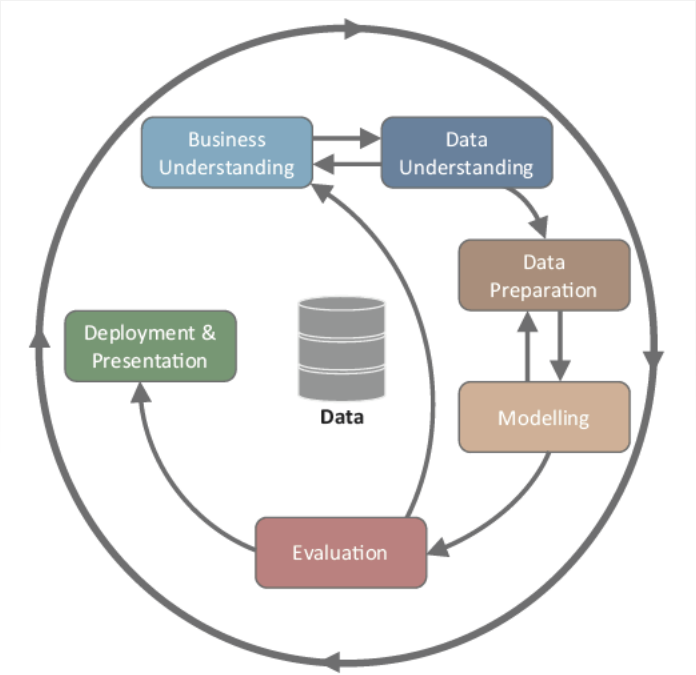
Thema der Data Science Challenge 2021.

„Interpretierbarkeit von Machine-Learning-Modellen und Tools"

Deadline 31.08.2021
Weitere Infos unter aktuar.de

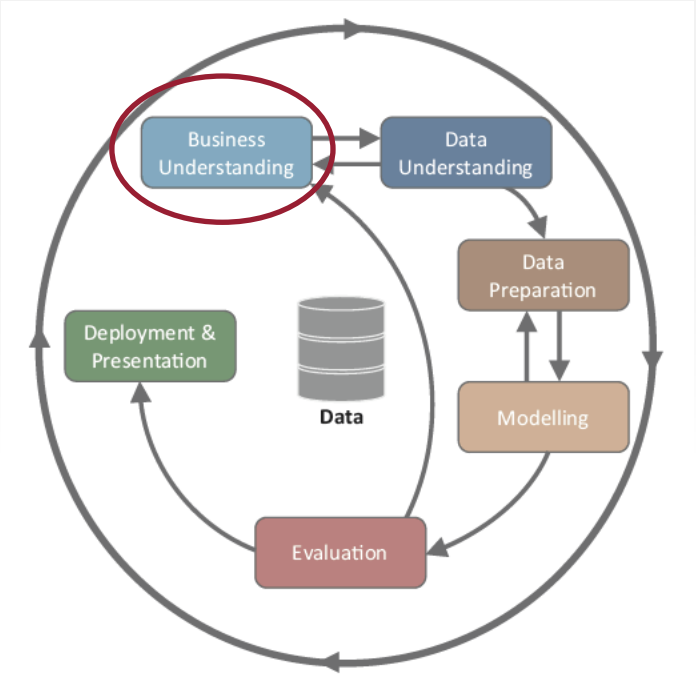
Unser Vorgehen

CRISP-DM = Cross Industry Standard Process for Data Mining



Quelle: Schnattinger (2020)

Business Understanding

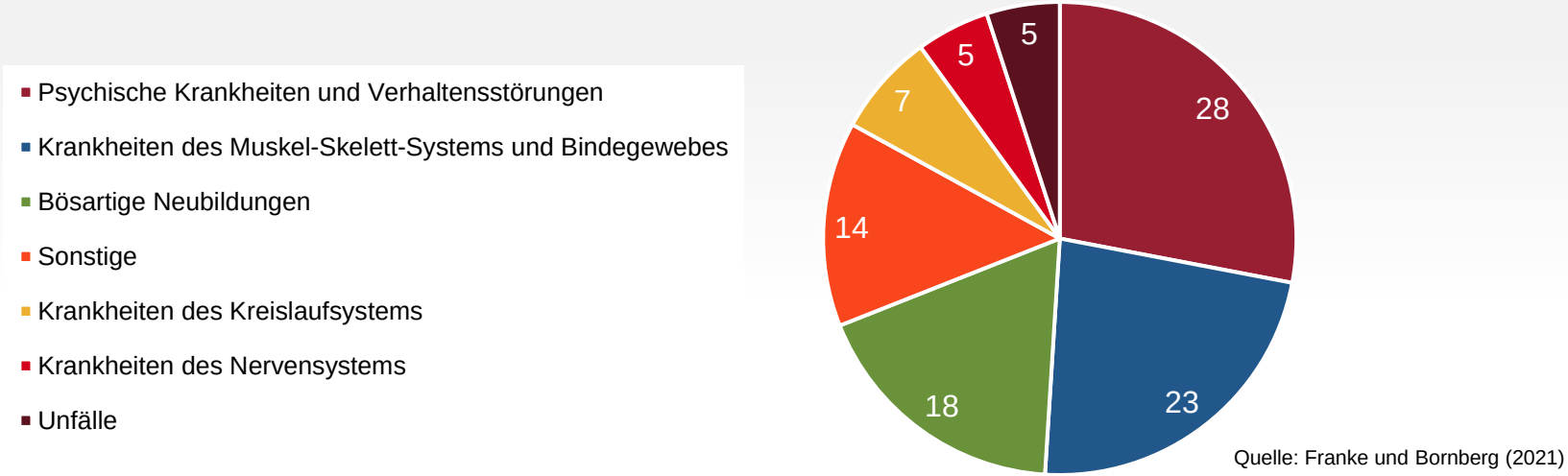


Quelle: Schnattinger (2020)

Business Understanding

Unsere Motivation

Krankheiten, die zur Anerkennung von BU führen



Quelle: Franke und Bornberg (2021)

Business Understanding

Unsere Motivation

Psychische Krankheiten sind ein sehr häufiger Grund für Berufsunfähigkeit

Forschungsfrage.

Welche Rahmenbedingungen begünstigen psychische Krankheiten, die dann zu einer Berufsunfähigkeit führen?

D. h. was sind hier relevante (und auch: was sind hier nicht relevante) Einflussfaktoren auf die Berufsunfähigkeit und welcher Zusammenhang besteht zur Berufsunfähigkeit?

Business Understanding

Unsere Motivation

Warum interessiert das Versicherungsunternehmen?

Tarifierung	Einflussfaktoren/Risikofaktoren sollten Einfluss auf die Prämie haben
Antragsprozess	Beschränkung auf wesentliche und relevante Einflussfaktoren für optimale Customer-Experience beim Antragsprozess
Präventiv-maßnahmen	Präventivmaßnahmen, die die Einflussfaktoren positiv beeinflussen, sind besonders erfolgsversprechend

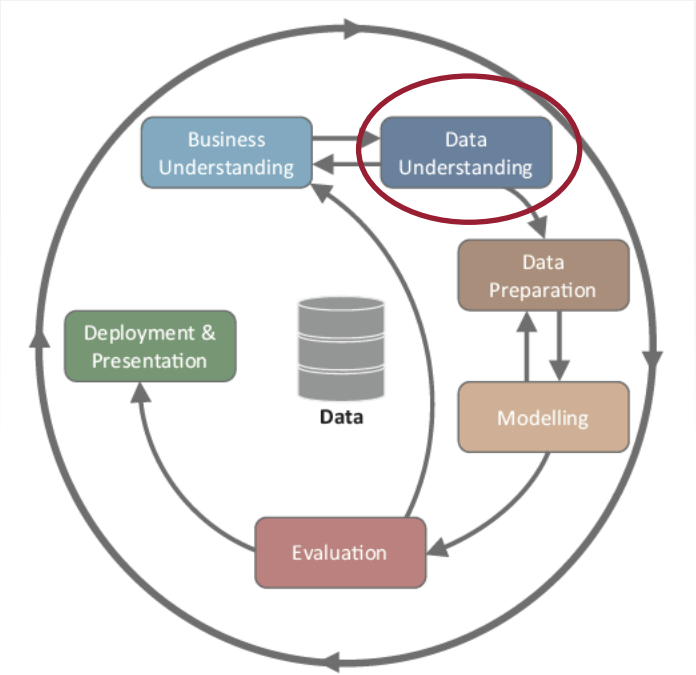
Business Understanding

Unsere Idee

Was ist dabei neu?

Wir wollen bewusst externe Daten beziehen, um auch Faktoren berücksichtigen zu können, die von VU ggf. bislang nicht erhoben wurden.

Data Understanding



Quelle: Schnattinger (2020)

Data Understanding

Welche Daten nutzen wir?

Wir nutzen Daten des National Health and Nutrition Examination Survey (NHANES).

- NHANES stellt ein für die USA repräsentatives Sample im Zeitraum von 1999 bis 2018 dar.
- Alle zwei Jahre werden äußerst umfangreiche Informationen über den Gesundheitszustand (insbesondere Psyche), die Lebensumstände und das Arbeitsverhältnis (inkl. Berufsunfähigkeit) von rund 5.000 Personen erhoben.

Data Understanding

Wie können wir damit unsere Forschungsfrage beantworten?

Wir untersuchen, wie die **abhängige Variable „Berufsunfähigkeit“** („Limitations keeping you from working“) durch in den Daten vorhandene unabhängige Variable erklärt werden kann.

Data Understanding

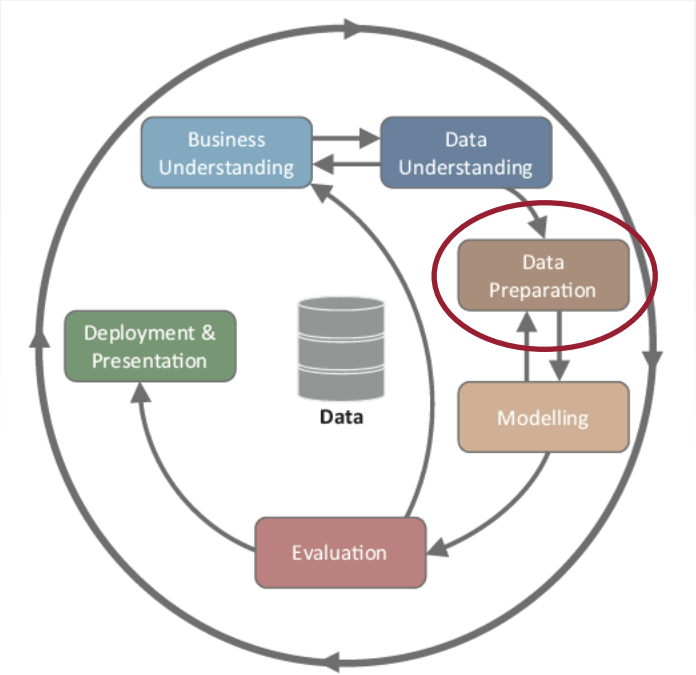
Wie können wir damit unsere Forschungsfrage beantworten?

Allerdings besitzen unsere Ergebnisse eine begrenzte Aussagekraft.

- Angaben werden von den Befragten gemacht und stellen eine Selbsteinschätzung dar.
- Es muss keine ärztlich festgestellte Berufsunfähigkeit vorliegen, wie sie für eine deutsche Versicherung notwendig wäre.

Unsere Ergebnisse geben daher nur Hinweise auf mögliche Risikofaktoren, sind aber nicht uneingeschränkt übertragbar!

Data Preparation



Quelle: Schnattinger (2020)

Data Preparation

Aufbereitung unserer Daten

U. a.

- Umgang mit fehlenden Werten
- One-Hot-Kodierung für kategoriale Variable
- Min-Max-Normalisierung
- Korrelationsanalyse

31 Variable und 21.555 Datensätze

Data Preparation

Aufbereitung unserer Daten – Umgang mit fehlenden Daten

Beobachtung.

In unserem Sample weisen einige Variable einen hohen Anteil fehlender Werte auf.

Maßnahmen.

- Variable löschen

  Möglichst wenig Variable löschen, um die Modellkomplexität nicht zu stark zu reduzieren!

- Datensätze entfernen

  Möglichst wenige Datensätze entfernen, um eine breite Datenbasis zu erhalten!

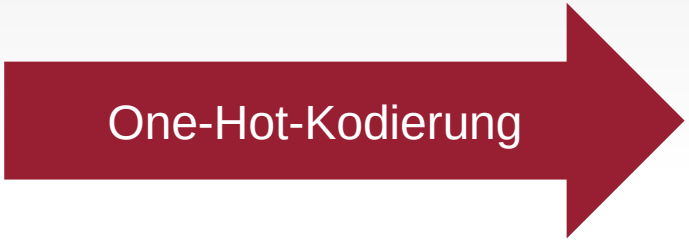
- Ergänzen der fehlenden Daten durch logisches Schließen
- Ergänzen der fehlenden Daten durch Schätzungen

Data Preparation

Aufbereitung unserer Daten – One-Hot-Kodierung

Beobachtung.

In unserem Sample sind auch kategoriale Variable, die wir für unsere Analysen nutzbar machen wollen.

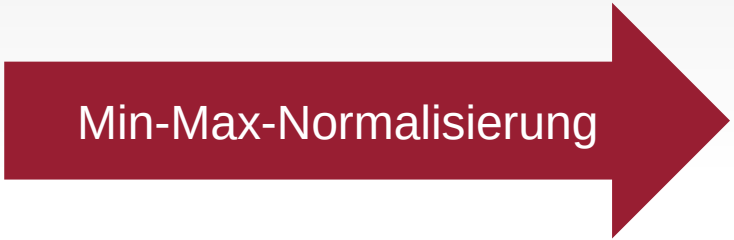


Data Preparation

Aufbereitung unserer Daten – Min-Max-Normalisierung

Beobachtung.

Die Wertebereiche unserer Variablen sind sehr heterogen und Algorithmen reagieren darauf teilweise sehr empfindlich.



Data Preparation

Aufbereitung unserer Daten – Korrelationsanalyse

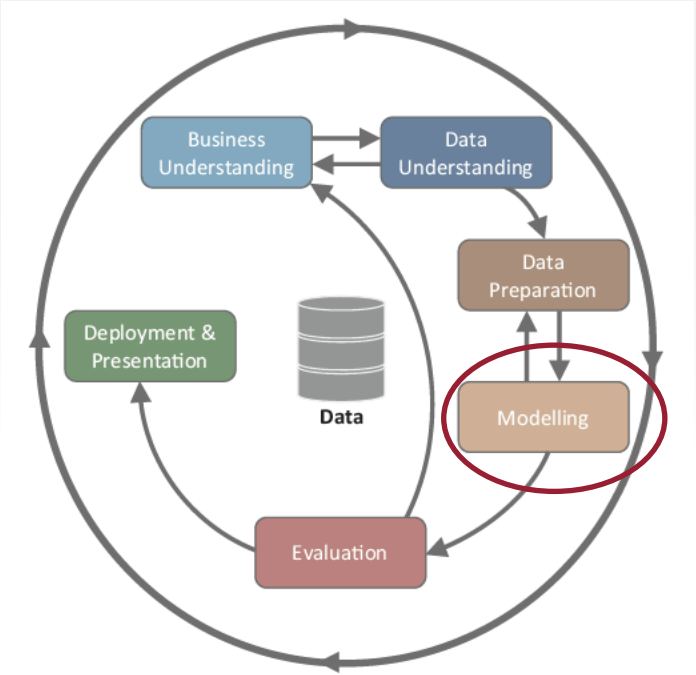
Motivation.

Weisen zwei oder mehr erklärende Variable eine sehr starke Korrelation auf, führt das zu Problemen in der Modellierung und Evaluation. (Multikollinearität)

Maßnahmen.

- Analyse der Korrelationen zwischen den erklärenden Variablen
- Entferne Variable, um die Multikollinearität zu reduzieren

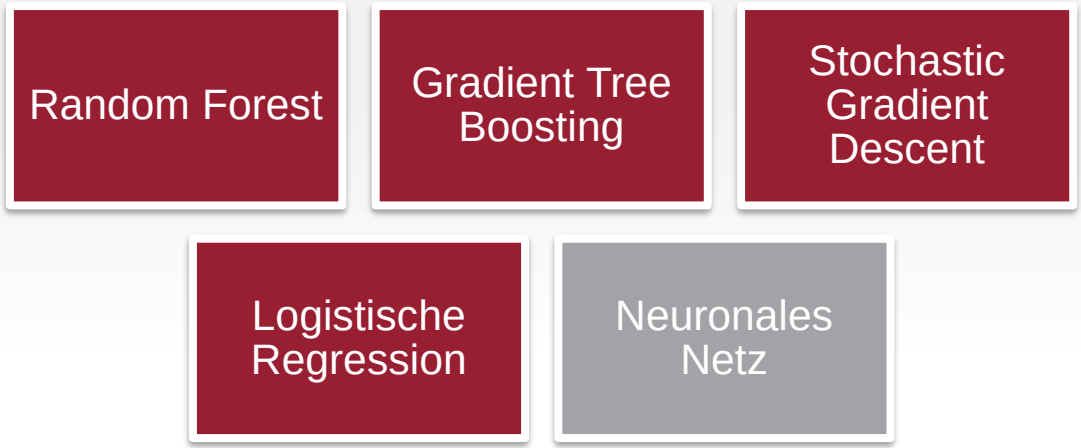
Modeling



Quelle: Schnattinger (2020)

Modeling

Angewendete Modelle

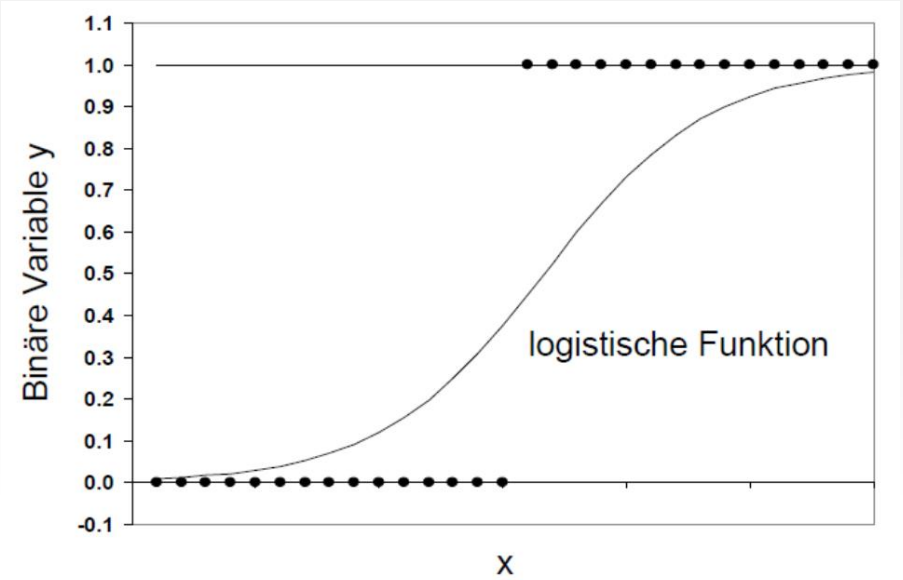


Modeling

Logistische Regression

$$P(y=1)=\frac{1}{1+e^{-z}}$$

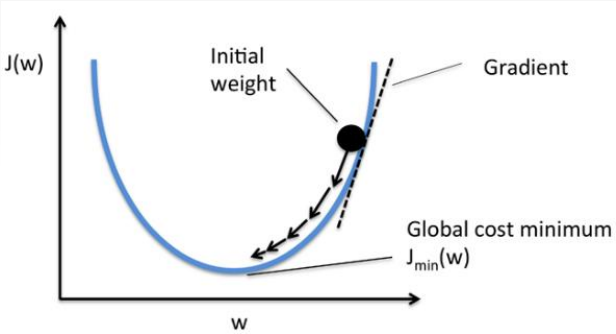
$$z=\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot x_3+\ldots+\beta_k\cdot x_k+\varepsilon$$

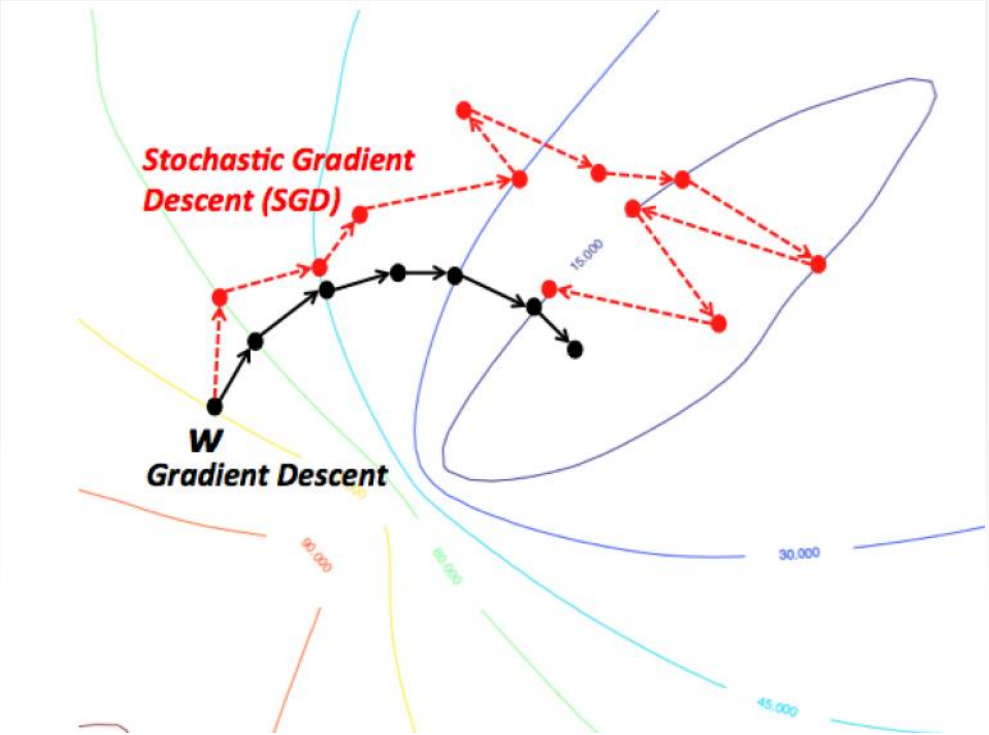


Quelle: Logistische Regression (hslu.ch)

Modeling

Stochastic Gradient Descent

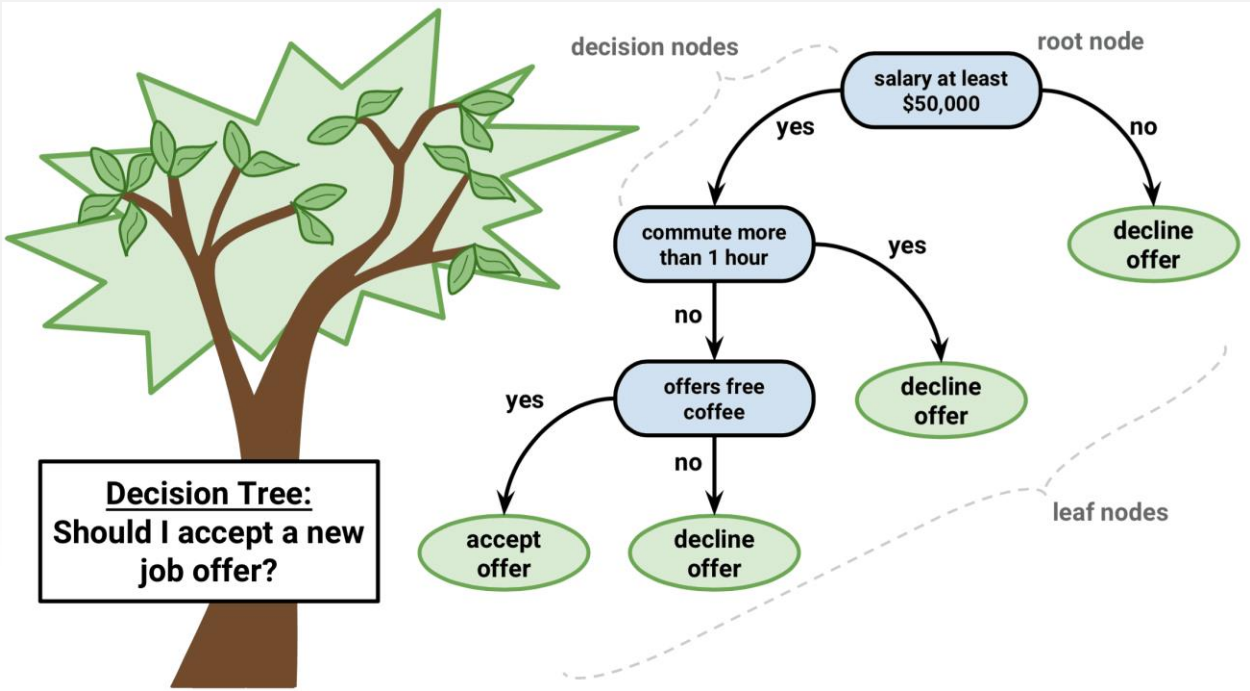




Quelle: scikit-learn: Batch gradient descent versus stochastic gradient descent - 2020 (bogotobogo.com)

Modeling

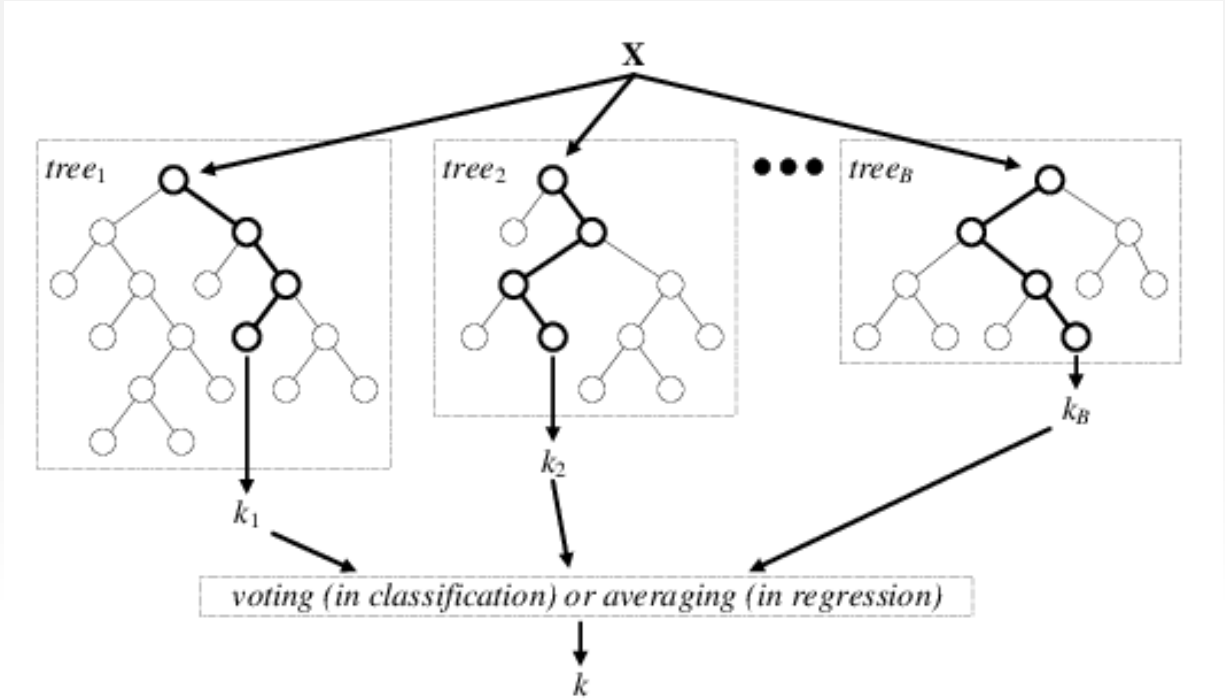
Decision Tree



Quelle: https://www.datacamp.com/community/tutorials/decision-trees-R

Modeling

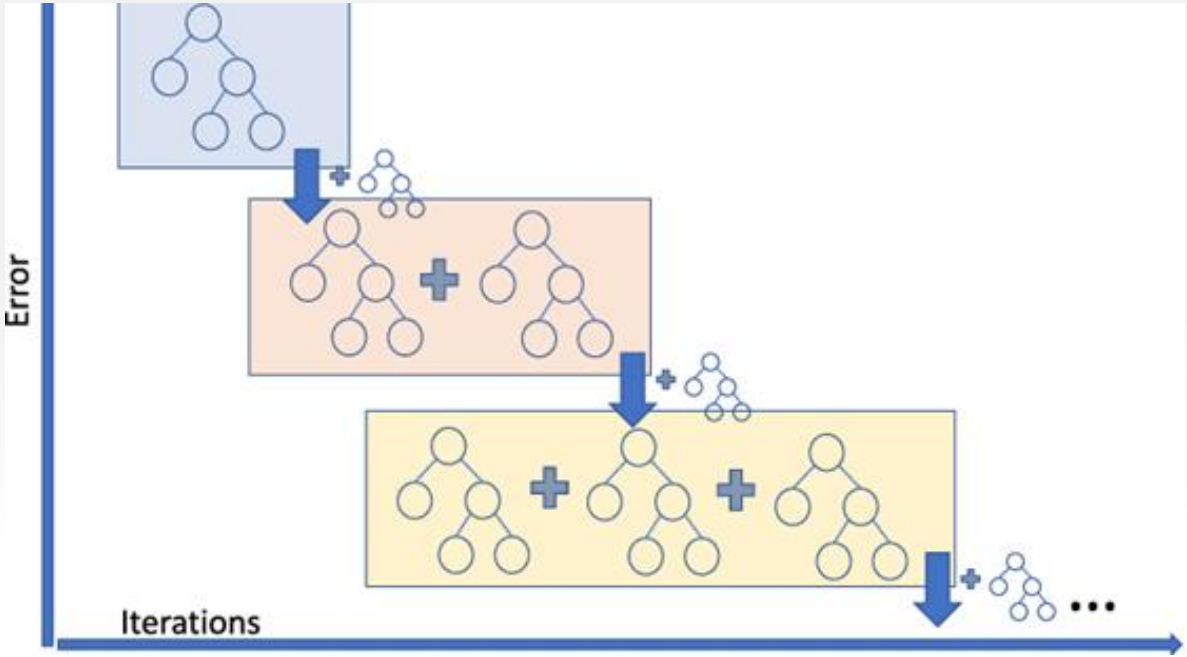
Random Forest



Quelle: https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643

Modeling

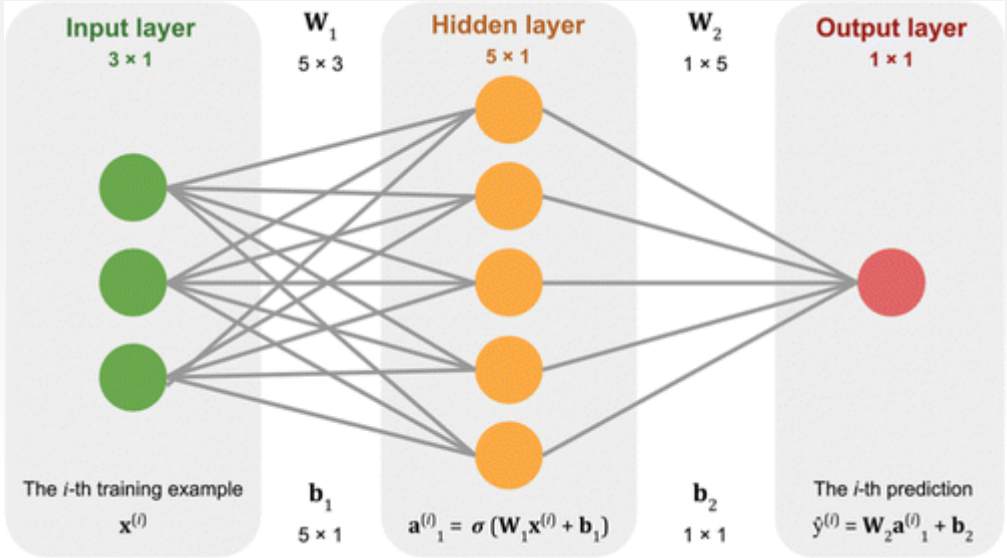
Gradient Boosting



Quelle: https://medium.com/analytics-vidhya/what-is-gradient-boosting-how-is-it-different-from-ada-boost-2d5ff5767cb2

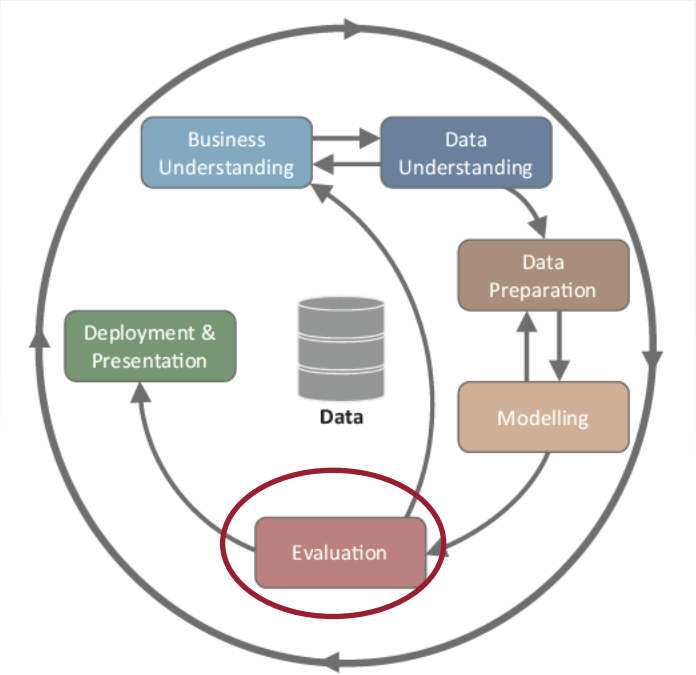
Modeling

Neuronales Netz



Quelle: Neural networks – SEG Wiki

Evaluation



Quelle: Schnattinger (2020)

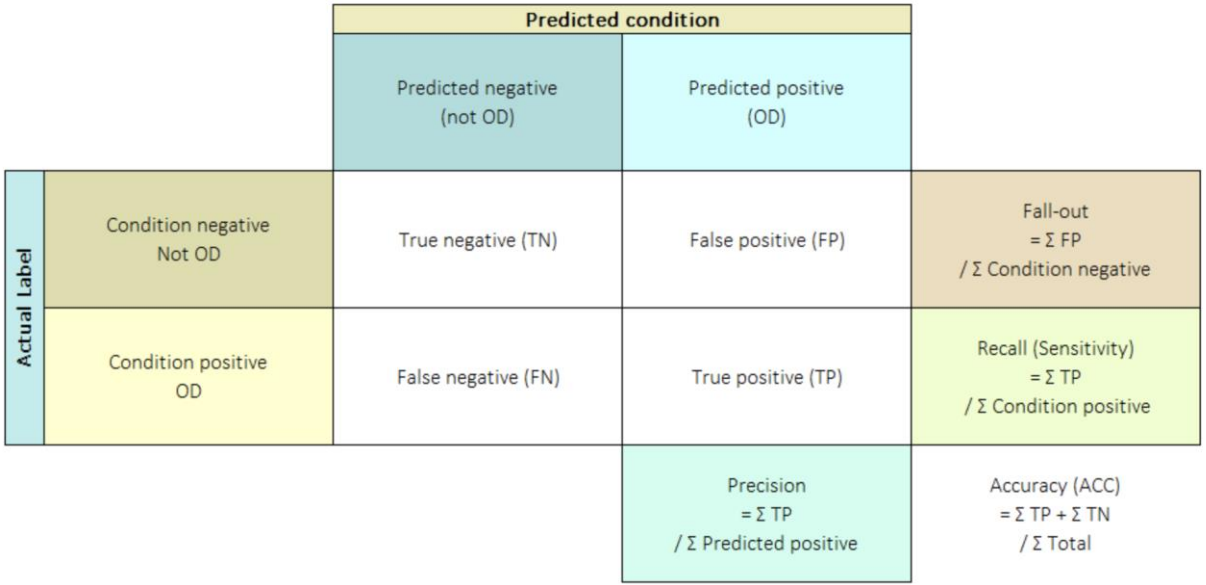
Modell-Evaluation

Überblick

- Da das Merkmal BU zwei Merkmalswerte hat (BU, nicht BU), treten bei Klassifizierungen vier verschiedene Fälle auf:
  - Korrekt negativ
  - Falsch negativ
  - Korrekt positiv
  - Falsch positiv
- Klassische Kennzahlen, um die Prognosefähigkeit des Modells zu analyieren: Accuracy, Precision, Recall, Fall-out, F-beta Maß

Modell-Evaluation

Kennzahlen



Actual Label	Predicted condition: Predicted negative (not OD)	Predicted condition: Predicted positive (OD)	
Condition negative Not OD	True negative (TN)	False positive (FP)	Fall-out = Σ FP / Σ Condition negative
Condition positive OD	False negative (FN)	True positive (TP)	Recall (Sensitivity) = Σ TP / Σ Condition positive
		Precision = Σ TP / Σ Predicted positive	Accuracy (ACC) = Σ TP + Σ TN / Σ Total

Modell-Evaluation

Kennzahlen

F-beta Maß (in unserem Modell: ß = 10)

Formel 1

$$F_\beta = (1 + \beta^2) \cdot \frac{Genauigkeit \cdot Sensitivität}{(\beta^2 \cdot Genauigkeit) + Sensitivität}$$

Modell-Evaluation

Benchmark

- Für einen Benchmark wird ein naiver Schätzer betrachtet
- Dieser Schätzer prognostiziert als Ergebnis immer „Gesund“
- Im Test-Sample befinden sich 6.467 Datensätze
  - davon sind 5.626 gesund
  - 841 ‚BU‘

- Der Naive Schätzer erreicht eine Accuracy (Genauigkeit) von $\frac{5.626}{6.467}$ ~ 87 Prozent

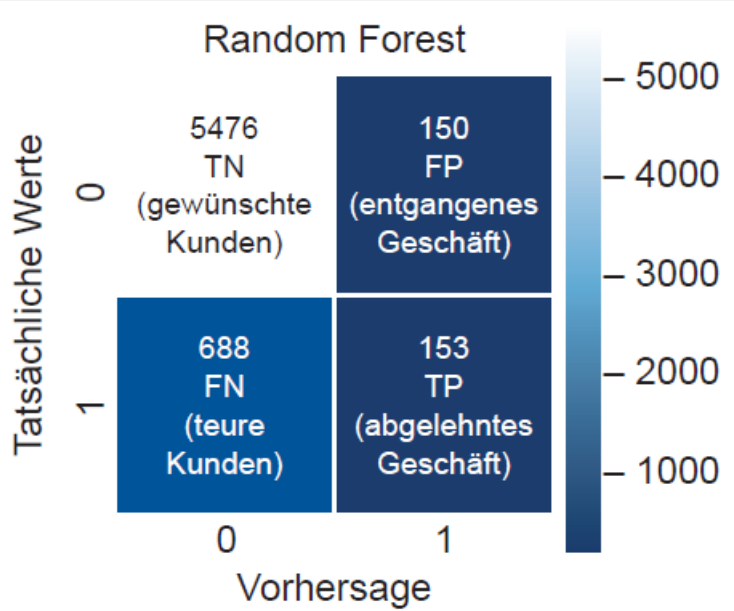
Modell-Evaluation

Confusion Matrix – Random Forest

Accuracy

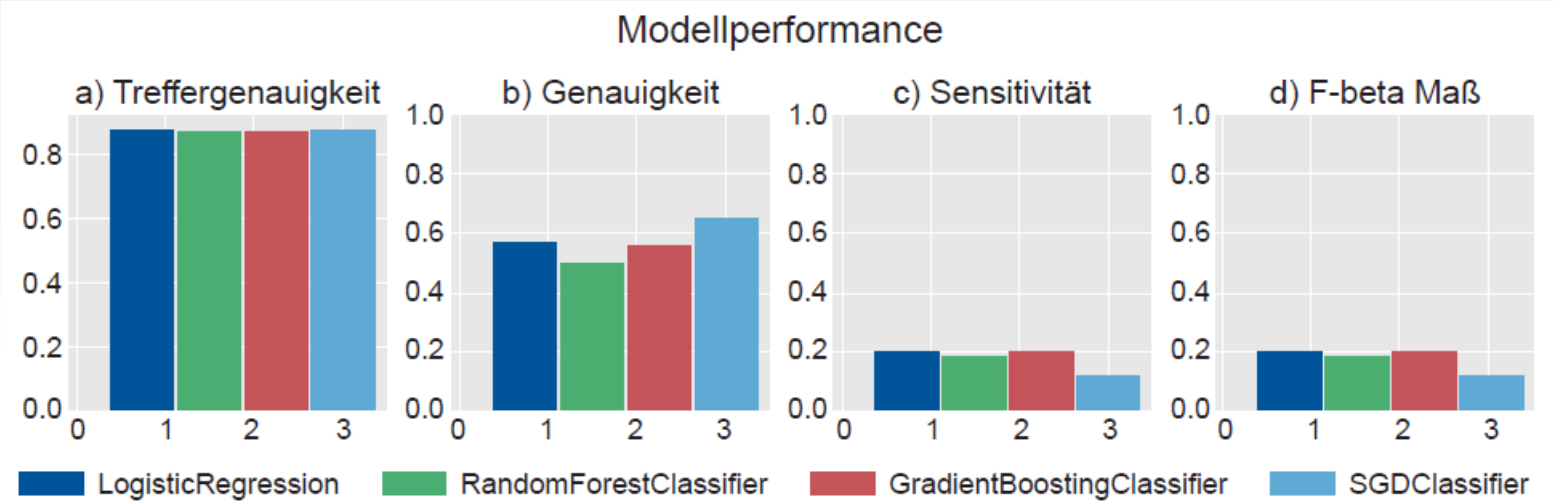
$$= \frac{153 + 5.476}{6.467}$$

~ 87 Prozent



Modell-Evaluation

Vergleich der verschiedenen Modelle



Wie können wir die Ergebnisse verbessern?

Idee: Resampling

- **Ausgangslage: Unbalancierte Daten**

  ~ 87 Prozent sind nicht BU

  ~ 13 Prozent sind BU

  → BU wird schlecht ‚gelernt'

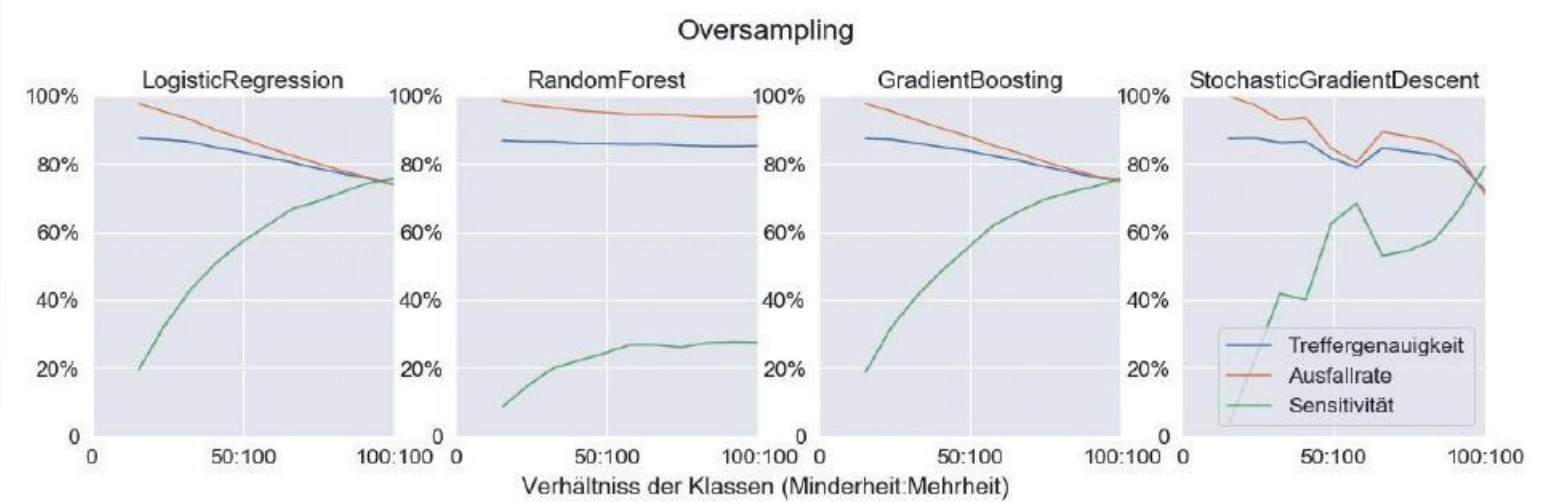
- **Idee: Klassengrößen annähern**

  – Oversampling der kleineren Klasse

  – Downsampling der größeren Klasse

Wie können wir die Ergebnisse verbessern?

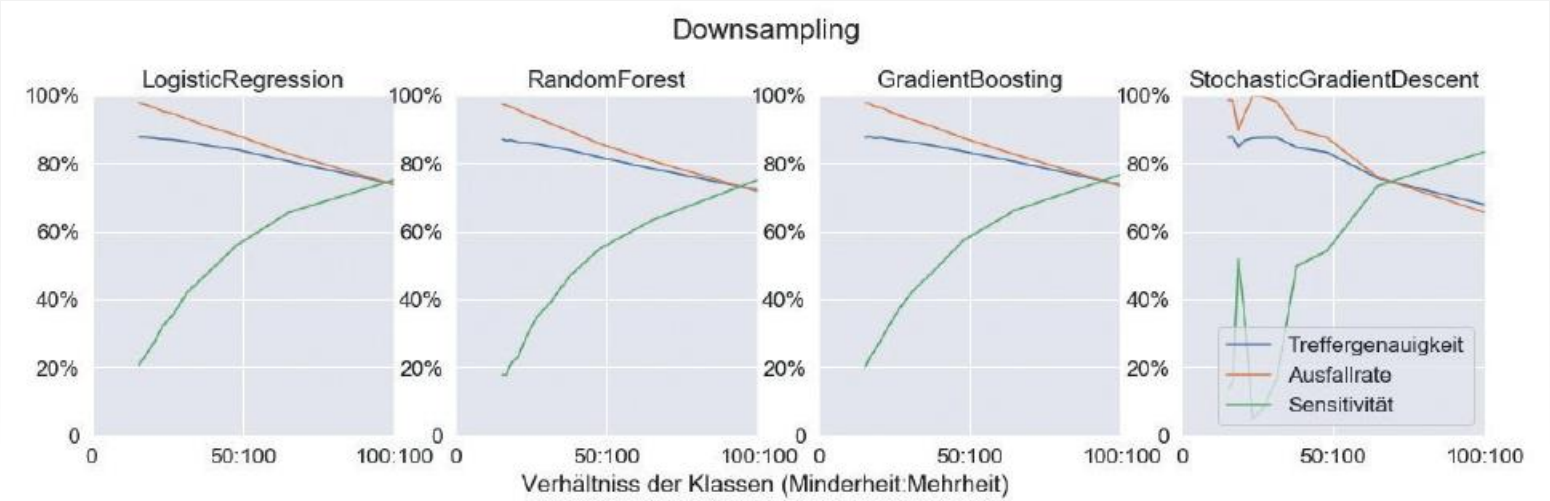
Oversampling



Auf der x-Achse sind verschiedene Verhältnisse der kleineren Klasse (BU-Fälle) zu der größeren Klasse (Gesunde) dargestellt bis zu einem Verhältnis von 100:100.

Wie können wir die Ergebnisse verbessern?

Downsampling



Auf der x-Achse sind verschiedene Verhältnisse der kleineren Klasse (BU-Fälle) zu der größeren Klasse (Gesunde) dargestellt bis zu einem Verhältnis von 100:100.

Wie können wir die Ergebnisse verbessern?

Downsampling

Nach dem Downsampling (Gradient Boosting)

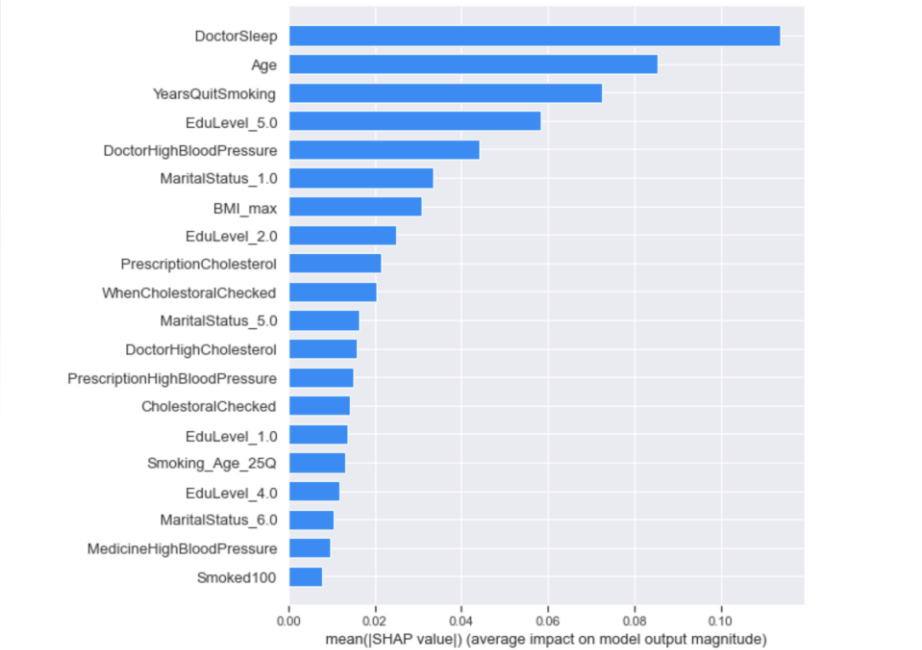
Im fiktiven Antragsprozess:

- von 6467 Anträgen
- würden 2148 (33,2 %) abgelehnt (FP + TP),
- 4319 angenommen und
- 196 davon berufsunfähig werden

-> **BU-Quote von 4,5%**

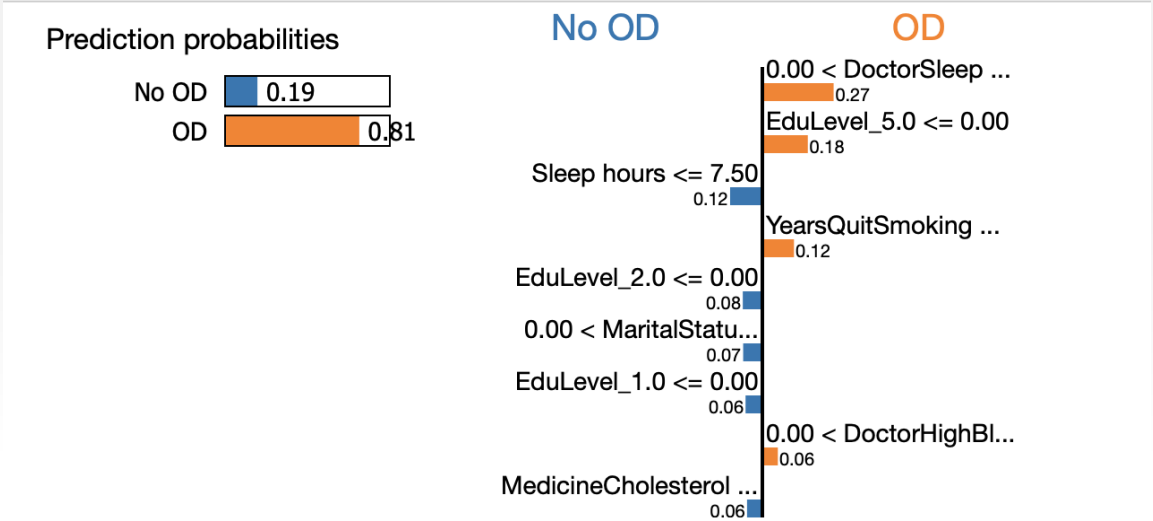
Erklärbarkeit der Modelle

Größte Einflussfaktoren – hier: SHAP-Werte (Gradient Boosting)



Erklärbarkeit der Modelle

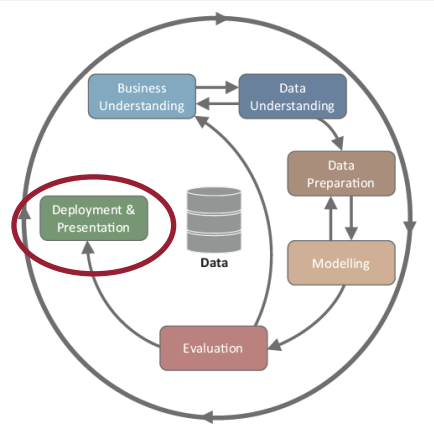
LIME (Local interpretable model-agnostic explanations)



Fazit

Anwendungsfall BU

- Laut unserer Analyse sind relevante Einflussfaktoren auf Berufsunfähigkeit (in diesem speziellen US-Datensatz) u. a.: BMI, Alter, Schlafprobleme
- Aber: Ungewisse Kausalitäten
- U. a. mögliche Berücksichtigung der Erkenntnisse bei Präventivmaßnahmen



Quelle: Schnattinger (2020)

Fazit

Methodik

- Datenbeschaffung und Datenaufbereitung stellen meist die Hauptarbeit dar
- Resampling-Verfahren können bei unausgewogenen Datensätzen zu einer deutlichen Verbesserung der Ergebnisse führen
- Ganzheitliche Evaluation der Modelle unter Berücksichtigung der Erklärbarkeit sehr bedeutend

Danke!