Masterarbeit

Daniel Siemmeister

2022 - 07 - 07

Titel der Arbeit

Erprobung unterschiedlicher Machine Learning Ansätze für die Vorhersage der Prüfungsaktivität von Studierenden

Wie viele prüfungsaktive Studierende wird es in drei Jahren geben?

Wie viele prüfungsaktive Studierende wird es in drei Jahren geben?

Ansätze des Leistungs- und Qualitätsmanagement (LQM)

Wie viele prüfungsaktive Studierende wird es in drei Jahren geben?

Ansätze des Leistungs- und Qualitätsmanagement (LQM)

Prädiktion der Wahrscheinlichkeit, in drei Jahren prüfungsaktiv zu sein - ohne konkrete Klassifizierung

Machine Learning

Machine Learning

$f(\cdot) \ldots$ mit $Y = f(\mathbf{X}) + \epsilon$

$\mathcal{A} \ldots$ Algorithmus mit $h_S = \mathcal{A}(S)$

$L_D(\mathcal{A}) =$
$\mathbb{E}[l(\mathcal{A}(S), (\mathbf{X}, Y))] \ldots$ wahre Risikofunktion

$L_S(h_S) =$
$\frac{1}{n} \sum_{i=1}^{n} l(h_S, (\mathbf{x}_i, y_i)) \ldots$ empirische Risikofunktion

Machine Learning

mit ϵ wird Verteilung von $\mathcal{D}_{Y|\mathbf{X}}$ festgelgt

Parameter von $\mathcal{D}_{Y|\mathbf{X}}$ soll mittels h_S approximiert werden

Sinvolle Wahlen: Erwartungswert, Median

loss-Funktion entscheidet darüber, welcher Parameter approximiert wird

- $l(h_S, (\mathbf{X}, Y)) = (Y - h_S(\mathbf{X}))^2$ approximiert $\mathbb{E}[Y|\mathbf{X}]$
- $l(h_S, (\mathbf{X}, Y)) = |Y - h_S(\mathbf{X})|$ approximiert $\mathrm{m}(Y|\mathbf{X})$ (Median)

Machine Learning

Das wahre Risiko kann folgendermaßen umgeformt werden:

$$\mathbb{E}[l(\mathcal{A}(S), (\mathbf{X}, Y))] =$$

$$\underbrace{\mathbb{E}[(h_S(\mathbf{X}) - \bar{h}(\mathbf{X}))^2]}_{\text{Variance}} + \underbrace{\mathbb{E}[(\bar{h}(\mathbf{X}) - \bar{y}(\mathbf{X}))^2]}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\bar{y}(\mathbf{X}) - Y)^2]}_{\text{Noise}}$$

Machine Learning

Lineare und logistische Regression

Machine Learning

Lineare und logistische Regression

Support Vector Machines

Machine Learning

Lineare und logistische Regression

Support Vector Machines

Random Forest Modelle

Machine Learning

Lineare und logistische Regression

Support Vector Machines

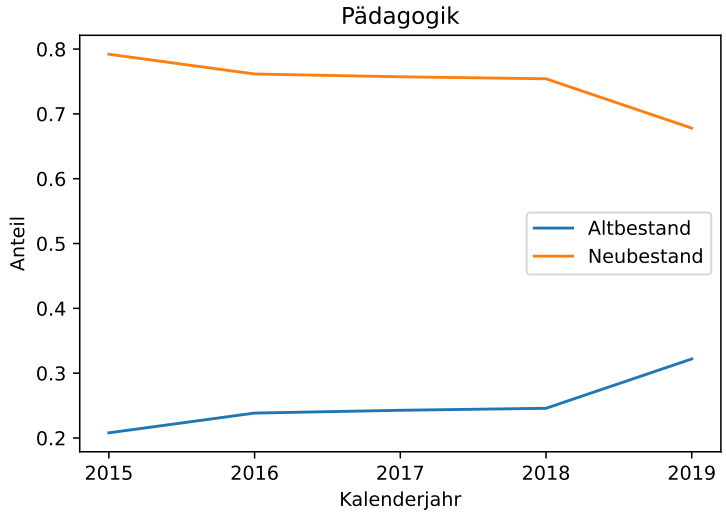
Random Forest Modelle

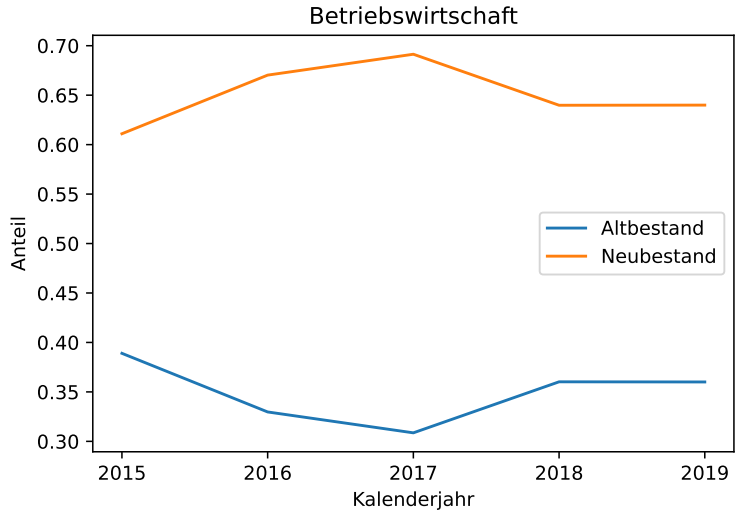
Künstliche Neuronale Netzwerke

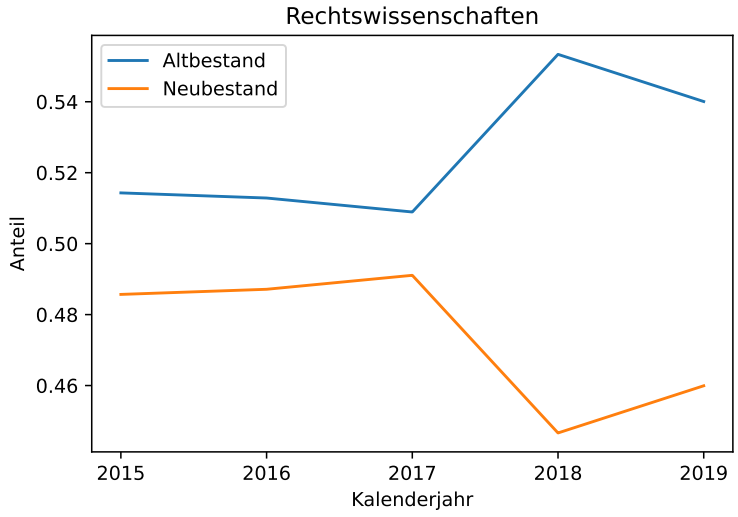
Problemstellung

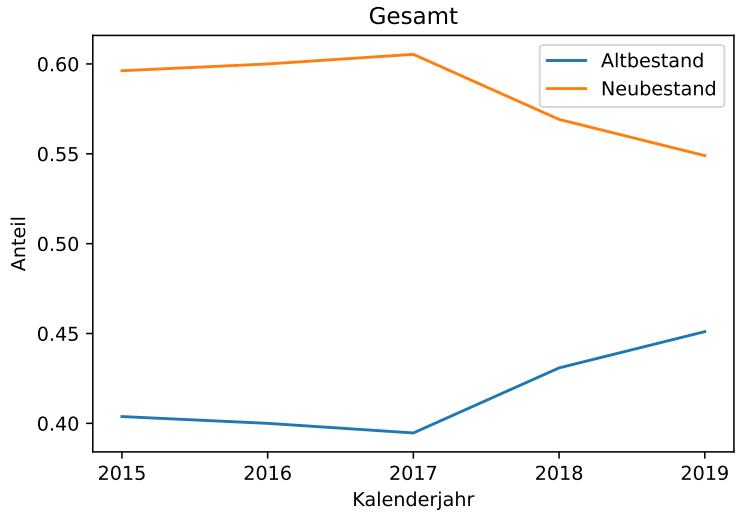
Daten von 2022 und davor	Daten von 2023, 2024, 2025
☐ Anzahl vorhanden	☐ Anzahl nicht vorhanden
☐ Merkmalskombinationen vorhanden	☐ Merkmalskombinationen nicht vorhanden

Relevanz der Problemstellung









Problem 1:
Schätzung der prüfungsaktiven Studierenden, von denen bereits Anzahl und Merkmalskombinationen vorhanden sind

Problem 1:
Schätzung der prüfungsaktiven Studierenden, von denen bereits Anzahl und Merkmalskombinationen vorhanden sind

Problem 2:
Schätzung der prüfungsaktiven Studierenden, von denen weder Anzahl noch Merkmalskombinationen vorhanden sind

Herangehensweise Problem 1

- Regression der ECTS

Herangehensweise Problem 1

- Regression der ECTS
- Markov Ketten Modell

Herangehensweise Problem 1

- Regression der ECTS
- Markov Ketten Modell
- Schätzung der Wahrscheinlichkeit aktiv zu sein, ohne zu klassifizieren

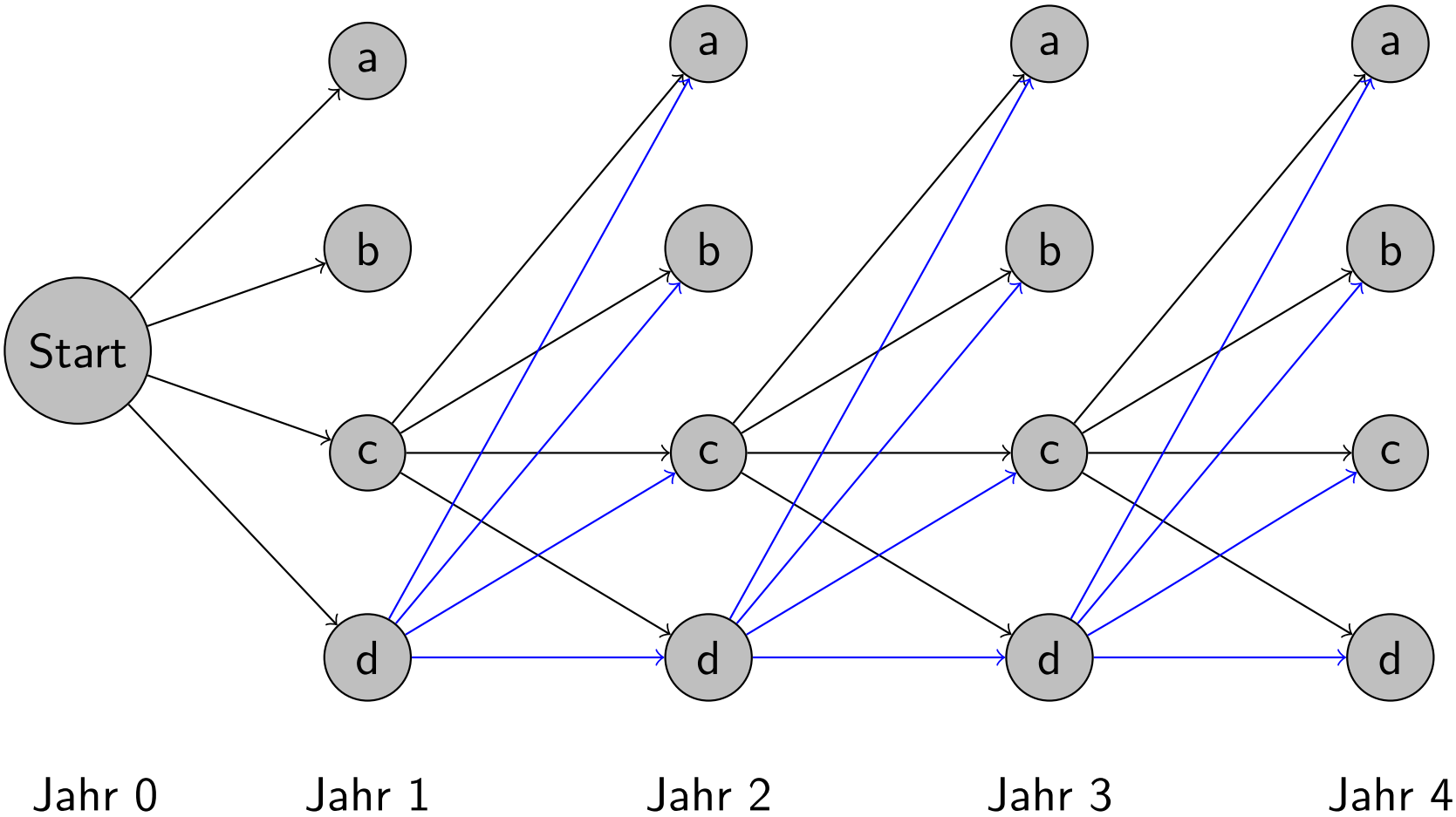
Ergebnisse für Ansatz 1 (P1)

Ergebnisse für Ansatz 1 (P1)

Metrik		lineare Regression	Random Forest	SVM	KNN
RMSE	1 Jahr	18.7 ± 0.2	19.2 ± 0.3	19.7 ± 0.4	18.7 ± 0.3
	$\geq$ 2 Jahre	16.8 ± 0.2	15.4 ± 0.2	19.2 ± 0.3	14.8 ± 0.2
MAE	1 Jahr	15.6	15.9	15.9	14.5
	$\geq$ 2 Jahre	13.3	11.7	16.2	10.4

Ergebnisse für Ansatz 2 (P1)

Ergebnisse für Ansatz 2 (P1)



Ergebnisse für Ansatz 2 (P1)

$$\begin{bmatrix} 0.05 & 0.19 & 0.53 & 0.23 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.02 & 0.13 & 0.79 & 0.07 \\ 0.02 & 0.49 & 0.19 & 0.29 \end{bmatrix},$$
$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.01 & 0.05 & 0.82 & 0.13 \\ 0.01 & 0.46 & 0.24 & 0.29 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.05 & 0.01 & 0.83 & 0.11 \\ 0.00 & 0.18 & 0.25 & 0.50 \end{bmatrix}.$$

Ergebnisse für Ansatz 3 (P1)

Ergebnisse für Ansatz 3 (P1)

		log. Reg.	RF	SVM	KNN
1 Jahr	Predicted	129.39	128.17	128.84	129.29
	Real	129	129	129	129
$\geq$ 2 Jahre	Predicted	121.25	117.46	120.59	120.9
	Real	121	121	121	121

Ergebnisse für Problem 1

Ergebnisse für Problem 1

- X Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS

Ergebnisse für Problem 1

- X Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS
- X Ansatz 2 benötigt mehr Daten, um ihn seriös zu erproben

Ergebnisse für Problem 1

- ✗ Ansatz 1 funktioniert nicht - zu große Fehler bei Schätzung der ECTS
- ✗ Ansatz 2 benötigt mehr Daten, um ihn seriös zu erproben
- ✓ Ansatz 3 funktioniert auf kleinem Datensatz (sehr!) gut - man benötigt mehr Daten um ihn noch besser zu erproben

Herangehensweise Problem 2

- Schätzung der Anzahl der Studierenden mit gleicher Merkmalskombination wie im Jahr zuvor

Herangehensweise Problem 2

- Schätzung der Anzahl der Studierenden mit gleicher Merkmalskombination wie im Jahr zuvor
- Clustering der Studierenden und anschließende Schätzung der Anzahl nach Cluster

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction dummy (Anzahl gegeben)	Daten	tatsächliche Anzahl
1 Jahr	2016	1105		1092
	2017	984		973
2 Jahre	2016	878		819
	2017	769		721

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction dummy (Anzahl gegeben)	Daten	tatsächliche Anzahl
1 Jahr	2016	1105		1092
	2017	984		973
2 Jahre	2016	878		819
	2017	769		721

- ∼ Legitimation von Ansatz 1 für Problem 2

Ergebnisse für Problem 2

Zeitspanne der Schätzung		Prediction dummy (Anzahl gegeben)	Daten	tatsächliche Anzahl
1 Jahr	2016	1105		1092
	2017	984		973
2 Jahre	2016	878		819
	2017	769		721

- ∼ Legitimation von Ansatz 1 für Problem 2
- ∼ Zu wenige Daten vorhanden, um Ansatz 2 für Problem 2 seriös zu erproben

Beiträge

Beiträge

- Klare Darstellung der Problemstellung

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Machine Learning Ansatz für Problem 1, der gute Ergebnisse liefert

Beiträge

Klare Darstellung der Problemstellung

Erprobung unterschiedlicher Ansätze

Machine Learning Ansatz für Problem 1, der gute Ergebnisse liefert

Notwendigkeit von mehr Daten, um Ansätze weiter zu Erproben