KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Prognosemodelle (Klassifikation und Regression):
02_06Ergebnis_Prognosemodell

Prognosemodelle und Evaluation

Erarbeitet von
Dr. Katja Theune

Lernziele

- Du kannst die Idee der k-fold cross-validation erläutern
- Du kannst eine confusion matrix eines Prognosemodells erläutern
- Du kannst die resultierenden Evaluationsmetriken eines Prognosemodells erläutern
- Du kannst erläutern, warum der Testfehler (anstatt des Trainingsfehlers) für eine Modellevaluation wichtig ist
- Du kannst anhand von Evaluationsmetriken ein Prognosemodell auswählen

Inhalt

Einstieg

Wie präzise ist jetzt eigentlich eine Prognose für mögliche Studienabbrüche und welches Verfahren ist das beste? Und welche Schlussfolgerungen kann ich aus meinen Ergebnissen ziehen? Das wollen wir jetzt im Folgenden besprechen. Dafür nutzen wir die uns vorliegenden realen Daten und trainieren, evaluieren und vergleichen verschiedene Prognosemodelle.

Quelle [1]

Vorbereitung der Analysen

Wir verwenden als Inputs nur diejenigen, die wir bei einer Datenexploration ausgesucht haben. Das sind folgende (siehe Grafik):

Verwendete Daten

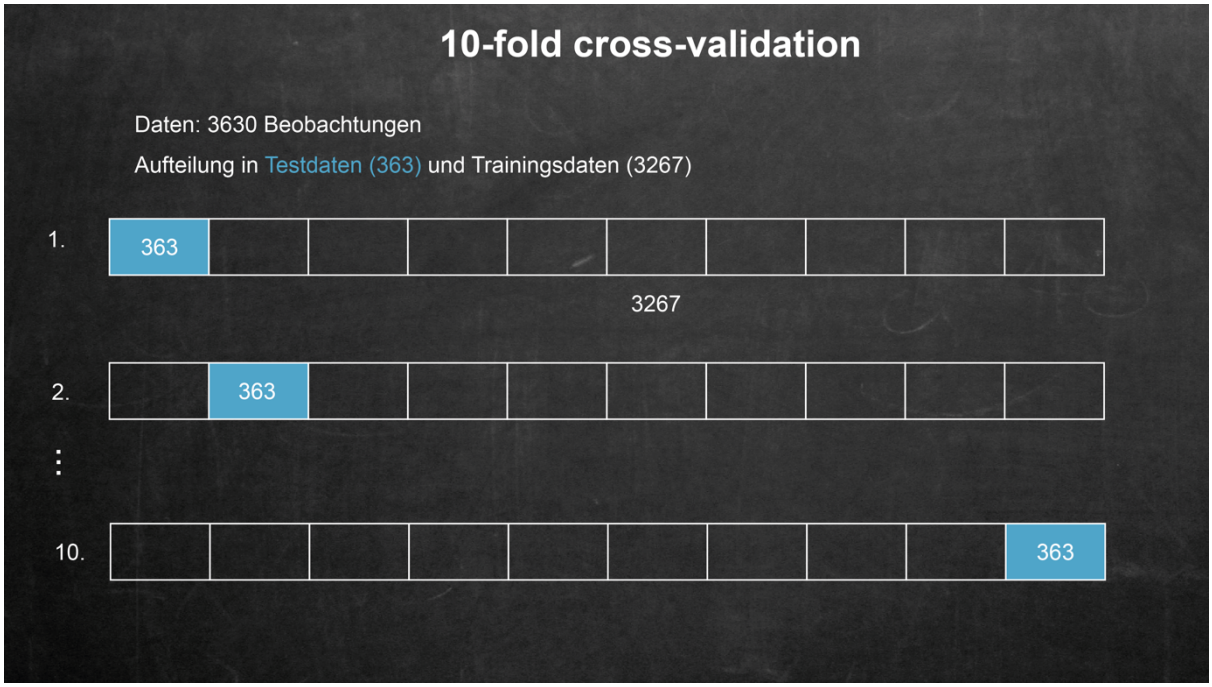
Inputs	application mode
	course
	previous qualification
	mothers qualification
	tuition fees
	gender
	scholarship holder
	age at enrollment
	curricular units 2nd semester (approved)
	curricular units 2nd semester (grade)
Output	target (dropout vs. graduate)

Unser Output heißt hier „Target" und beinhaltet die beiden Klassen „Dropout" und „Graduate", also Studienabbruch und Studienerfolg. Unsere interessierende Klasse ist „Dropout", da wir insbesondere mögliche Studienabbrüche prognostizieren wollen. Daher sind das im Folgenden unsere Positives.

Da wir klassifizieren wollen, benötigen wir überwachte Lernverfahren. Als Verfahren wählen wir daher zunächst den random forest, da er sehr beliebt ist und die bereits besprochenen Vorteile mit sich bringt. Wir vergleichen ihn aber auch mit einem einfachen decision tree als Benchmark und der sehr häufig angewendeten logistischen Regression. Beginnen wir aber mit einem random forest.

Cross-validation und confusion matrix

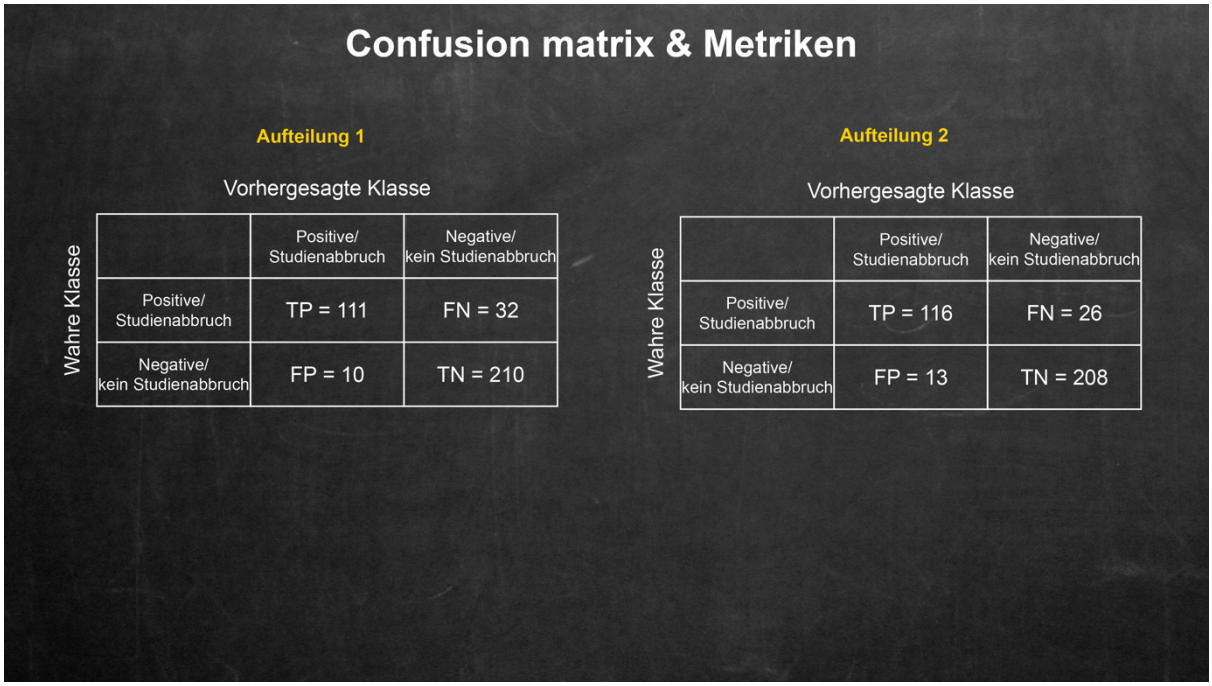
Um unser Modell zu evaluieren, verwenden wir eine k-fold cross-validation mit k=10. Wir erinnern uns: wir haben durch die cross-validation dann 10 verschiedene Aufteilungen und für jede dieser Aufteilungen wird ein Modell auf den Trainingsdaten trainiert und auf den Testdaten evaluiert.



Insgesamt haben wir 3630 Studierende im Datensatz. Für jede Aufteilung umfassen die Testdaten also 363 Studierende, was den Beobachtungen in einem fold entspricht. Die Trainingsdaten enthalten dann die restlichen 3267 Studierenden aus den anderen 9 folds.

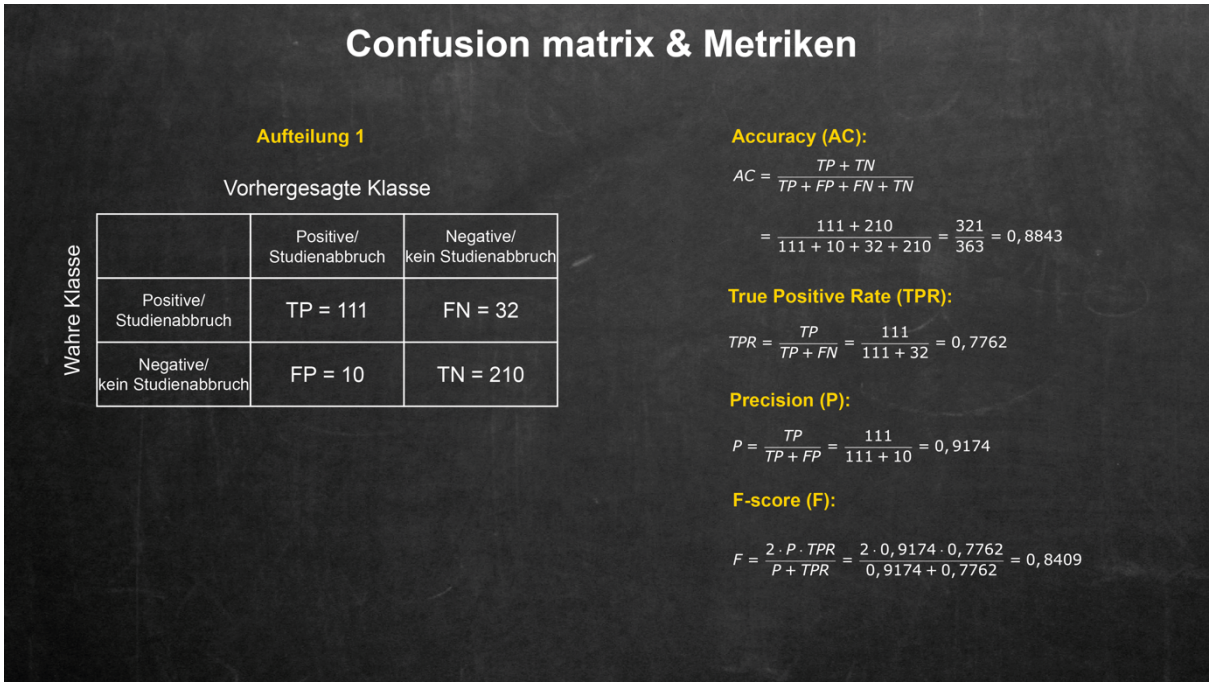
Zunächst sehen wir hier eine confusion matrix, die sich für eine konkrete Aufteilung in Trainings- und Testdaten ergeben hat (siehe folgende Grafik, Aufteilung 1). Wir sehen z. B., dass wir 111 True Positives haben. Also 111 wahre Studienabbrüche wurden vom Modell auch als solche erkannt. Darüber hinaus haben wir 10 False Positives, 32 False Negatives und 210 True Negatives. Insgesamt sind das unsere 363 Studierenden in den Testdaten.

Zum Vergleich sehen wir hier eine weitere confusion matrix für eine andere Aufteilung (siehe folgende Grafik, Aufteilung 2). Es gibt hier einige Unterschiede zu der ersten Matrix. Das veranschaulicht gut, dass wir bei jeder der 10 Aufteilungen andere Ergebnisse erhalten.



Evaluationsmetriken

Auf Basis solch einer confusion matrix können wir nun verschiedene Evaluationsmetriken berechnen. Hier schauen wir uns für die erste Aufteilung die Accuracy, die True Positive Rate bzw. Sensitivität, die Precision und den F-score an.



Wir sehen, dass wir eine Accuracy von 0,8843 erreichen. Wir klassifizieren mit unserem Modell also 88,43 % aller Beobachtungen richtig. Die True Positive Rate beträgt 0,7762. 77,62 % aller Studienabbrüche wurden damit auch vom Modell erkannt. Die Precision, das Maß der Exaktheit des Modells, beträgt hier 0,9174 und der F-Score 0,8409.

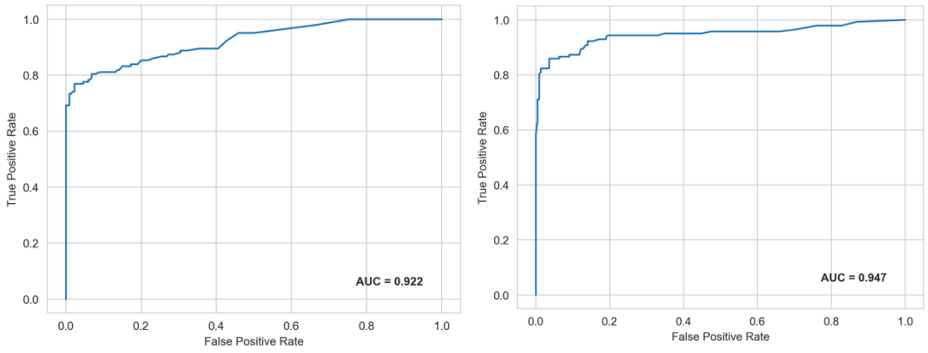
Schauen wir zum Vergleich auch mal die Metriken für alle 10 Aufteilungen und den Durchschnitt über alle Aufteilungen an. So lässt sich dann auch gut nochmal die Idee und der Sinn der cross-validation veranschaulichen.

Übersicht Metriken für verschiedene Aufteilungen

Aufteilung	Accuracy	TPR	Precision	F-Score
1	0,8843	0,7762	0,9174	0,8409
2	0,8926	0,8169	0,8992	0,8561
3	0,9063	0,8592	0,8971	0,8777
...	...	...	...	...
10	0,8926	0,831	0,8872	0,8582
Durchschnitt	0,8898	0,8100	0,8990	0.8518

Wir sehen hier, dass sich für jede der 10 Aufteilungen andere Werte für die Metriken ergeben. Dabei sind manche Werte schlechter oder besser als der jeweilige Durchschnitt über alle Werte einer bestimmten Metrik.

Ähnliches erkennen wir auch, wenn wir uns die ROC-Kurven bzw. das AUC-Maß z. B. für Aufteilung 1 (links) und 3 (rechts) ansehen. Das Modell, dass auf der dritten Aufteilung basiert, schneidet deutlich besser ab als das Modell, das auf der ersten Aufteilung basiert. Die Kurve liegt hier weiter links.



Der Durchschnitt gibt uns damit ein realistischeres Bild des Fehlers als nur ein einziger Wert basierend auf einem einzigen z. B. zufälligen Test-Datensatz wie bei dem validation set Ansatz.

Einige dieser Ergebnisse wollen wir jetzt noch ein wenig diskutieren und einordnen.

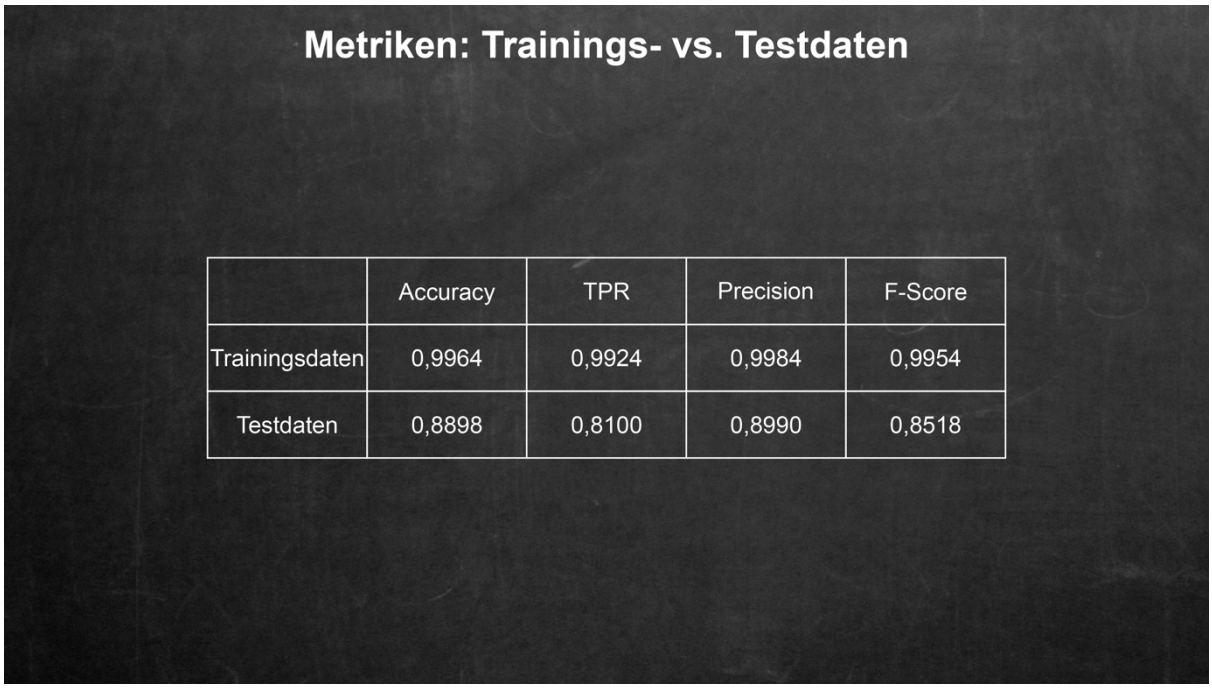
Wir sehen z. B., dass die durchschnittliche Accuracy mit knapp 89 % (siehe Grafik oben) recht gut ist. Wir prognostizieren also sehr viele Beobachtungen richtig. Allerdings müssen wir hier im Auge behalten, dass wir eine etwas ungleiche Klassenverteilung von 60 % Graduates und 40 % Dropouts haben und damit gerade die uns interessierende Klasse unterrepräsentiert ist. Daher sind auch andere Metriken wichtig, die sich auf diese Klasse fokussieren.

Z. B. besagt die durchschnittliche True Positive Rate bzw. die Trefferquote hier, dass 81 % aller Studienabbrüche vom Modell richtig erkannt wurden. Das klingt erstmal gar nicht schlecht. Allerdings heißt das auch, dass wir 19 %, das wäre dann die False Negative Rate, der gefährdeten Studierenden nicht erkennen und damit auch nicht unterstützen, was viele negative Folgen nach sich ziehen kann.

Wir sehen also insgesamt auch bei unserem realen Beispiel, dass es sinnvoll ist, sich mehrere Metriken anzuschauen und die Ergebnisse kontextspezifisch einzuordnen.

Trainings- vs. Testfehler

Wir können uns auch nochmal ansehen, wie diese Evaluationsmetriken aussehen, wenn wir sie auf den Trainingsdaten anstatt auf den Testdaten berechnen.

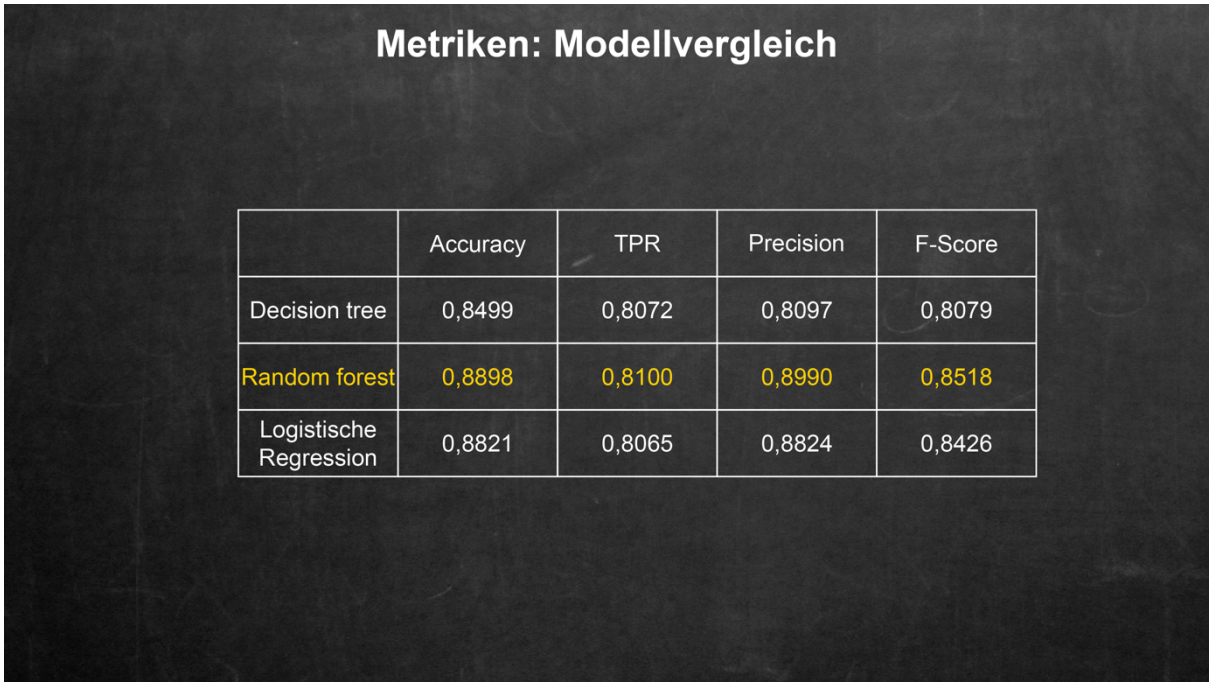


	Accuracy	TPR	Precision	F-Score
Trainingsdaten	0,9964	0,9924	0,9984	0,9954
Testdaten	0,8898	0,8100	0,8990	0,8518

Auch hier wird nochmal deutlich, wie wichtig es ist, einen Testfehler anstatt eines Trainingsfehlers zu bestimmen. Die Werte der Metriken auf den Trainingsdaten sind deutlich höher und damit besser als die Werte auf dem Testdatensatz. Das veranschaulicht noch einmal das Problem des Overfittings. Das Modell ist meist zu sehr auf die Trainingsdaten angepasst, so dass es zu einer Unterschätzung des wahren Fehlers kommt, wenn wir auf denselben Daten den Fehler bestimmen, auf denen das Modell trainiert wurde.

Modellvergleich

Schauen wir uns jetzt an, wie gut der random forest im Vergleich zu einem einfachen decision tree und der häufig angewendeten logistischen Regression ist.



	Accuracy	TPR	Precision	F-Score
Decision tree	0,8499	0,8072	0,8097	0,8079
Random forest	0,8898	0,8100	0,8990	0,8518
Logistische Regression	0,8821	0,8065	0,8824	0,8426

Bei allen betrachteten Evaluationsmetriken schneidet der random forest am besten ab. Danach folgt die logistische Regression. Der decision tree schneidet bei fast allen Metriken am schlechtesten ab.

Wir sehen aber z. B. auch, dass die logistische Regression im Vergleich zum decision tree zwar eine höhere Precision, aber eine, hier nur etwas, kleinere True Positive Rate hat. Hier wird deutlich, dass kombinierte Metriken wie der F-score sehr hilfreich sind. Und auch, dass verschiedene Modelle bei verschiedenen Metriken besser oder schlechter als andere Modelle sein können und wir daher die für uns wichtigsten Metriken vorher identifizieren sollten.

Abschluss

Wir haben nun auf einem realen Datensatz verschiedene Prognosemodelle trainiert und evaluiert und dabei das Verfahren und die Bedeutung der cross-validation und verschiedene Metriken zur Evaluation nachvollzogen und eingeordnet. Zudem haben wir nochmal die Wichtigkeit der Verwendung von Testdaten im Gegensatz zu den Trainingsdaten für eine realistische Evaluation unserer Modelle besprochen.

Quellen

Quelle [1] https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data

Berechnungen (Python): Dr. Ludmila Himmelspach

Disclaimer

Transkript zu dem Video „Prognosemodelle (Klassifikation und Regression): Prognosemodelle und Evaluation“, Dr. Katja Theune.