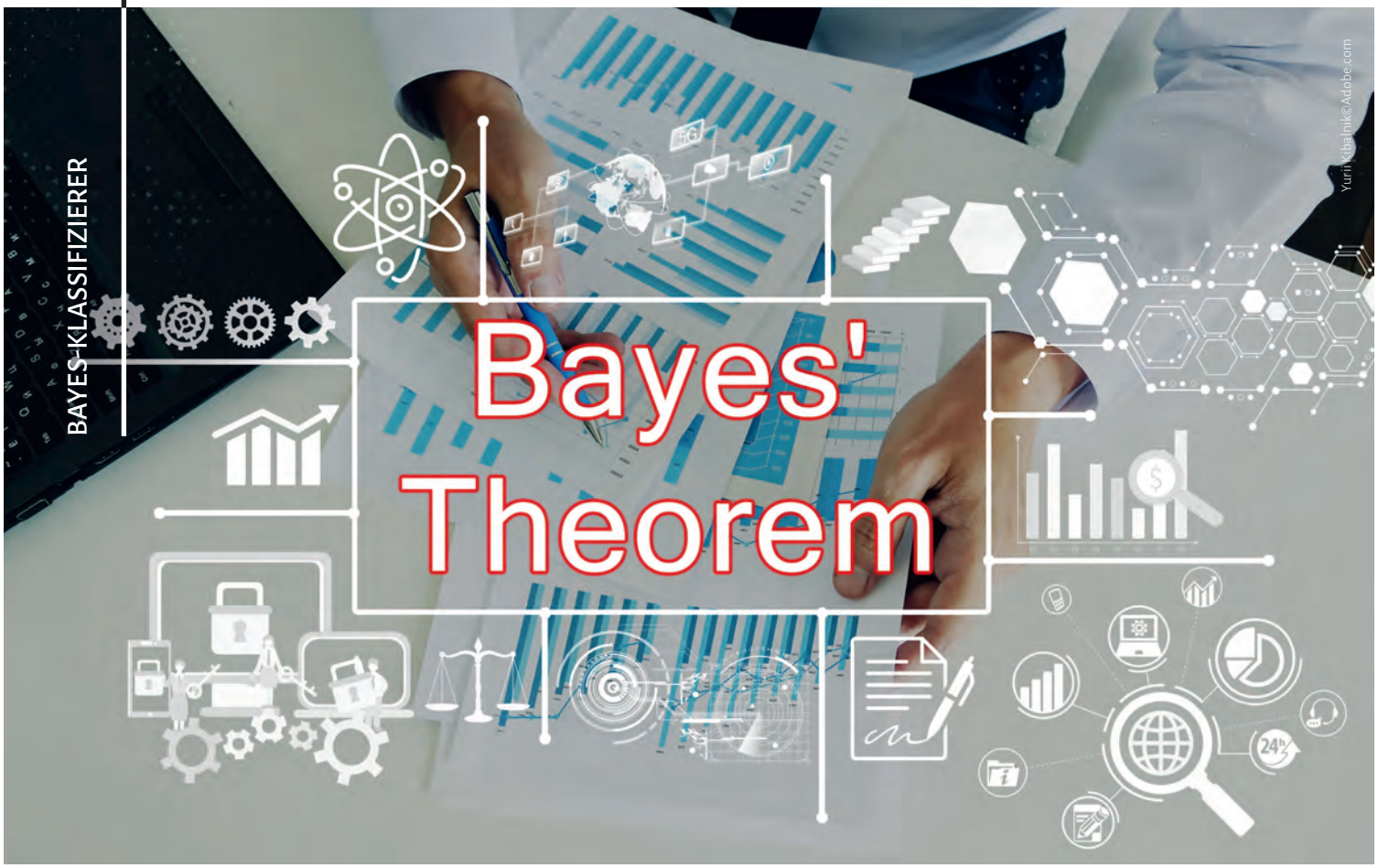

Predictive Analytics mit dem Bayes-Klassifizierer entdecken

Praxisbeispiel zur Ermittlung des Ausfallrisikos von Kunden. Florian Bliefert

Predictive Analytics hat viele Einsatzmöglichkeiten im Controlling. Eine davon ist der Bayes-Klassifizierer, der Entscheidungen anhand bedingter Wahrscheinlichkeiten trifft. Die bedingte Wahrscheinlichkeit beschreibt die Wahrscheinlichkeit des Eintretens eines Ereignisses A unter der Bedingung, dass ein anderes Ereignis B bereits eingetreten ist. Sie wird als P(A|B) dargestellt und berechnet sich durch das Verhältnis der Wahrscheinlichkeit des gemeinsamen Eintretens von A und B zur Wahrscheinlichkeit des Eintretens von B (siehe **Abb. 1**).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Abb. 1: Berechnung der Wahrscheinlichkeit A unter der Bedingung B

Ein Beispiel soll das Prinzip verdeutlichen: angenommen, wir möchten die Wahrscheinlichkeit berechnen, dass ein Kunde seine Rechnung nicht begleicht, wenn er vorher schon einmal eine Rechnung zu spät bezahlt hat. Dazu benötigen wir folgende Wahrscheinlichkeiten:

- P(L): die Wahrscheinlichkeit, dass der Kunde eine Rechnung zu spät begleicht (Late)
- P(D ∩ L): die Wahrscheinlichkeit, dass ein Kunde sowohl zahlungsunfähig (Default) ist als auch eine Rechnung zu spät beglichen hat.

Um auszurechnen, wie hoch die Wahrscheinlichkeit ist, dass der Kunde seine Rechnung nicht begleicht unter der Voraussetzung, dass er eine Rechnung zu spät gezahlt hat (geschrieben als P(D|L)), verwenden wir die Formel aus Abb. 1.

Doch wie kommen wir an die jeweiligen Wahrscheinlichkeiten? Die können wir aus

Summary

Fortschreitende Digitalisierung und exponentieller Anstieg der verfügbaren Datenmengen haben Predictive Analytics zu einem mächtigen Werkzeug gemacht. Gerade im Controlling haben quantitative Analyse und Vorhersagemodelle schon immer eine entscheidende Rolle gespielt, daher bietet es sich an, Predictive Analytics auch hier zu benutzen. Dabei gibt es viele unterschiedliche Werkzeuge, eines davon ist die bedingte Wahrscheinlichkeit. Sie ist die Grundlage für einen Entscheidungsalgorithmus, den Bayes-Klassifizierer, mit dessen Hilfe zum Beispiel das Ausfallrisiko eines Kunden bewertet werden kann.

unseren vorhandenen Kundendaten herleiten. In einer Stichprobe von 1.000 Kunden

- haben 200 eine Rechnung zu spät gezahlt (P(L) = 0,2)
- sind 50 zahlungsunfähig (P(D) = 0,05)
- und von den 50 haben 30 vorher eine Rechnung zu spät bezahlt (P(D ∩ L) = 0,03)

Wenn wir das in die Formel einsetzen, ergibt sich eine Wahrscheinlichkeit von 15% , dass ein Kunde zahlungsunfähig wird, nachdem er eine Rechnung nicht bezahlt hat (siehe **Abb. 2**). Für die Risikobewertung ergibt sich damit folgendes Bild: ein „unauffälliger" Kunde hat eine Ausfallwahrscheinlichkeit P(D) von 5%. Sobald dieser Kunde jedoch eine Rechnung zu spät bezahlt, springt sein Ausfallrisiko auf das Dreifache. Das sollte auf jeden Fall Maßnahmen auslösen!

$$P(D|L) = \frac{P(D \cap L)}{P(L)} = \frac{0,03}{0,2} = 0,15$$

Abb. 2: Berechnung der Wahrscheinlichkeit eines Zahlungsausfalls unter der Bedingung einer versäumten Zahlung

Der Bayes-Klassifizierer

Der Bayes-Klassifizierer ist ein statistischer Algorithmus, der auf dem Bayes-Theorem und der bedingten Wahrscheinlichkeit basiert. Er wird mit Daten trainiert, die bereits bekannte Zusammenhänge zwischen den Eingabevariablen und den zu entscheidenden Klassen enthalten. In unserem Fall wären das die Daten über die Zahlungsmoral unserer Kunden. Dabei würde jedoch nicht nur eine einzelne Eingabevariable berücksichtigt werden, sondern idealerweise alle relevanten Informationen über den Kunden. Das könnten zusätzlich das Alter, die Dauer der bestehenden Kundenbeziehung, vorherige Bestellungen oder andere demographische Informationen sein. Aus den unterschiedlichen Ausprägungen der Eingabevariablen und der bekannten Klassen „zahlungsfähig/zahlungsunfähig" berechnet der Bayes-Klassifizierer nun die jeweiligen Wahrscheinlichkeiten. Die ermittelten Wahrscheinlichkeiten sind dann auch das Entscheidungskriterium, in welche Klasse der Bayes-Klassifizierer den jeweiligen Kunden einordnen würde. Das funktioniert nach dem Training auch mit Kunden, die nicht in der Stichprobe bzw. dem Trainingsdatensatz waren. Je nach Ausprägungen der Eingangsvariablen berechnet der Algorithmus eine Wahrscheinlichkeit pro Klasse und sortiert ihn in die Klasse, in die er mit der höchsten Wahrscheinlichkeit gehört.

Natürlich gibt es auch Nachteile – die Qualität der Vorhersagen hängt stark von der Qualität der Trainingsdaten ab. Das ist weiterer ein Grund, warum Unternehmen Wert auf eine möglichst gute Datenqualität legen sollten, weil Modelle nur so gut sein können wie die zugrundeliegende Datenbasis. Darüber hinaus ist die Qualität der Stichprobe ebenfalls entscheidend. Generell neigt der Bayes-Klassifizierer dazu, kleinere Stichproben schlechter zu modellieren als größere. Eine weitere Annahme des Bayes-Klassifizierers ist die Unabhängigkeit der Eingangsvariablen. In der Realität gibt es jedoch üblicherweise Abhängigkeiten, wie zum Beispiel zwischen Alter und Einkommen. Aufgrund dieser Annahme wird der Algorithmus oft auch „naiver Bayes-Klassifizierer" genannt.

Anwendungsbeispiel

Die Universität von Kalifornien in Irvine (UCI) stellt auf ihrer Internetseite verschiedene Datensätze für Data Science Anwendungen frei zur Verfügung, einer davon ist das „Credit Card Dataset". Er enthält Informationen über Kreditkarteninhaber und deren Kreditverhalten und wird häufig in der Forschung und im maschinellen Lernen verwendet, um verschiedene Aspekte der Kreditrisikobewertung zu untersuchen und Vorhersagemodelle zu entwickeln.

Der Datensatz enthält eine Vielzahl von Merkmalen, darunter demografische Informationen wie Alter, Geschlecht und Bildungsstand, finanzielle Informationen wie Kreditlimit, ausstehender Saldo und Zahlungsverhalten sowie Informationen über vergangene Transaktionen. Das Ziel besteht darin, anhand dieser Merkmale Vorhersagen über das Kreditrisiko eines Kunden zu treffen, zum Beispiel ob er zahlungsunfähig wird oder seine Rechnungen rechtzeitig bezahlt. Damit ist er gut geeignet, das Prinzip des Bayes-Klassifizieres auszuprobieren.

Die Beispieldaten der UCI können Sie unter folgendem Link downloaden:
https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

Die quelloffene und kostenfreie KNIME Analytics Plattform können Sie unter folgendem Link downloaden:
https://www.knime.com/downloads

Die Datei mit dem erstellten Workflow stellt der Autor auf Anfrage gerne zur Verfügung.

Florian Bliefert

ist Manager bei der CA Akademie AG und sowohl als Trainer als auch Berater tätig. Seine Schwerpunkte liegen in den Bereichen Kosten- und Leistungsrechnung, Planung und Reporting sowie in Digitalisierungsthemen. Vor der CA war er in verschiedenen Controller-Funktionen u. a. beim Flughafen München und dem ADAC sowie als Leiter Controlling tätig.
f.bliefert@ca-akademie.de

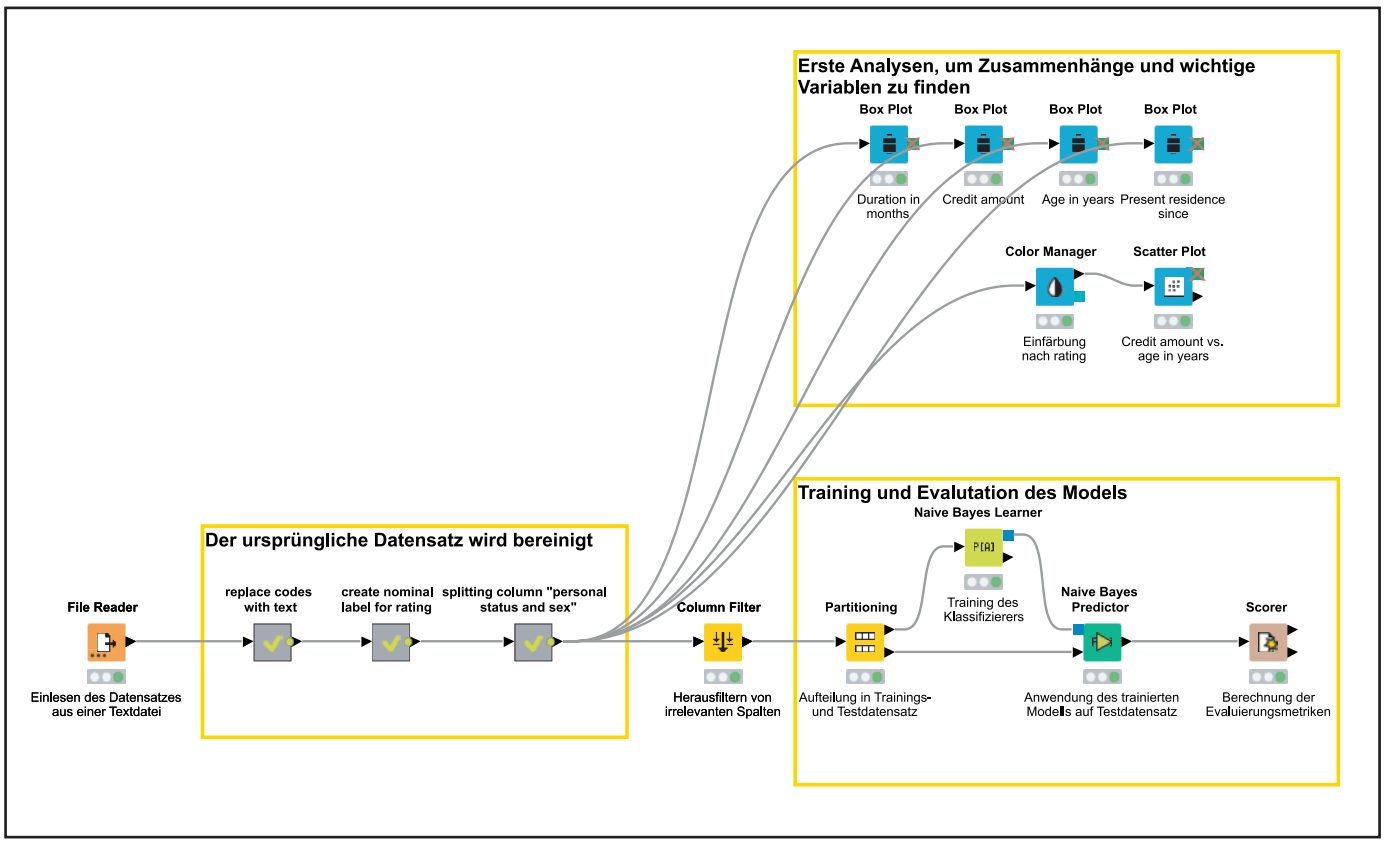

Abb. 3: Workflow zu Training und Anwendung des Bayes-Klassifizierers in KNIME

Dazu müssen zunächst die Daten aus dem Datensatz geladen und in ein geeignetes Format gebracht werden. Die Merkmale (Features) und die Zielvariable (Kreditrisiko: zahlungsunfähig oder nicht) müssen identifiziert und in Trainings- und Testdatensätze aufgeteilt werden. Mit den vorbereiteten Trainingsdaten wird dann der Bayes-Klassifizierer trainiert. Er berechnet dazu für jedes Merkmal die bedingten Wahrscheinlichkeiten unter der Voraussetzung des gegebenen Kreditrisikos. Mit Hilfe von Bayes Theorem wird das für die Vorhersage (im Knoten Naive Bayes Predictor) umgedreht: er berechnet die Wahrscheinlichkeit, dass ein Kunde zahlungsunfähig ist unter der Voraussetzung seiner Merkmale und Kreditverhalten. In **Abb. 3** ist zu sehen, wie dieser komplette Workflow in der Software KNIME aussieht.

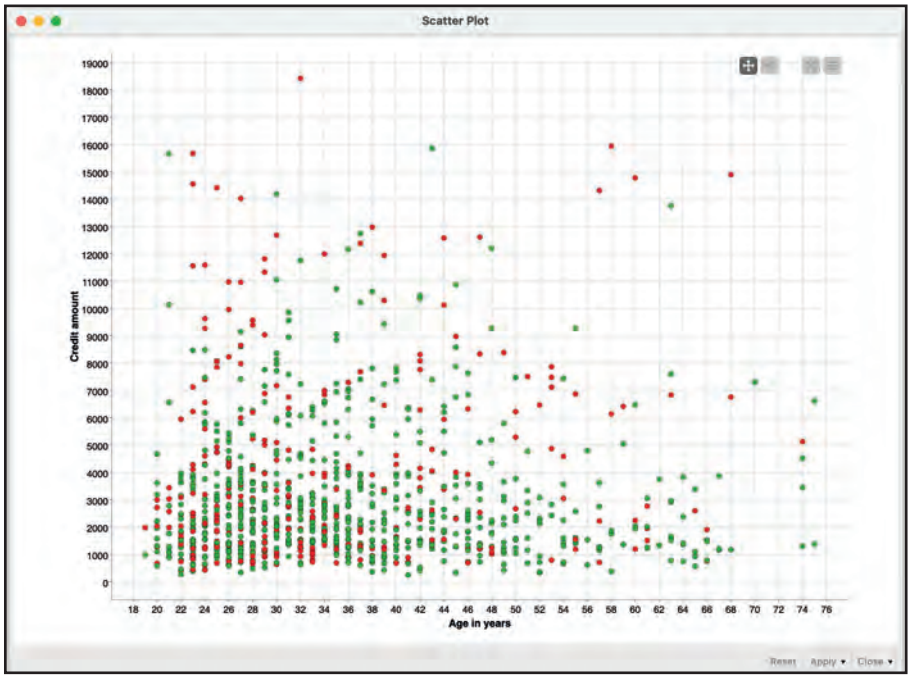

Abb. 4: Es scheint keinen Zusammenhang zwischen Alter, Kredithöhe und Zahlungsausfall zu geben

Im oberen rechten Kasten wurden auch erste explorative Analysen durchgeführt, um ein grundlegendes Verständnis des Datensatzes zu erhalten. Dabei können Zusammenhänge zwischen den Variablen gefunden werden, die die Qualität des Klassifizierers verbessern oder Merkmale aussortiert werden, die für eine Vorhersage nicht relevant sind. So scheint es keinen offensichtlichen Zusammenhang zwischen Alter, Kredithöhe und Zahlungsausfall zu geben (siehe **Abb. 4**). Diese Veränderungen der Merkmale, um die Vorhersagequalität zu steigern, werden auch Feature Engineering genannt.

Um die Qualität des Bayes-Klassifizierers zu bewerten, müssen wir die Vorhersagen des Modells mit den tatsächlichen Kreditrisiken vergleichen. Dazu wurde der Datensatz vor dem Training bereits in ein Trainings- und ein Testset aufgeteilt (im Knoten Partitionierung). Die Kunden des Testsets werden nun durch das trainierte Modell klassifiziert und

	Ausfall vorhergesagt	Kein Ausfall vorhergesagt
Ausgefallen	49	41
Nicht ausgefallen	28	182

Abb. 5: Evaluierungsmatrix („Confusion Matrix") des Credit Card Datensatzes

die Ergebnisse mit den tatsächlichen Klassen verglichen (siehe **Abb. 5**).

Daraus ergeben sich verschiedene Evaluierungsmetriken (im Knoten Scorer), wie zum Beispiel:

- Genauigkeit von 77%: Anteil der korrekten Vorhersagen (richtig klassifizierte Kunden) im Verhältnis zur Gesamtzahl der Kunden.
- Präzision von 64%: Anteil der tatsächlich zahlungsunfähigen Kunden im Verhältnis zu den vom Modell als zahlungsunfähig klassifizierten Kunden.
- Recall (auch Sensitivität genannt) von 54%: Anteil der vom Modell korrekt als zahlungsunfähig erkannten Kunden im Verhältnis zur Gesamtzahl der tatsächlich zahlungsunfähigen Kunden.

Die Präzision gibt an, wie zuverlässig das Modell ist, wenn es eine Zahlungsunfähigkeit vorhersagt. Eine hohe Präzision bedeutet, dass die meisten Kunden, die vom Modell als zahlungsunfähig eingestuft werden, tatsächlich zahlungsunfähig sind. Eine niedrige Präzision deutet hingegen darauf hin, dass das Modell eine beträchtliche Anzahl von falsch positiven Vorhersagen macht. Mit anderen Worten gibt die Präzision an, wie viele der positiven Vorhersagen des Modells tatsächlich korrekt sind.

Die Sensitivität misst, wie gut das Modell in der Lage ist, die tatsächlich zahlungsunfähigen Kunden zu identifizieren. Mit anderen Worten gibt sie an, wie viele der zahlungsunfähigen Kunden vom Modell richtig gefunden wurden.

In Bezug auf das Kreditrisiko-Klassifizierungsproblem ist es wichtig, sowohl die Präzision als auch die Sensitivität zu berücksichtigen. Eine hohe Präzision ist wünschenswert, um sicherzustellen, dass die vom Modell als zahlungsunfähig klassifizierten Kunden tatsächlich ein höheres Risiko haben. Eine hohe Sensitivität ist ebenfalls wichtig, um sicherzustellen, dass das Modell die meisten zahlungsunfähigen Kunden erkennt, um Risiken zu minimieren. Hier kann auch die Strategie eines Unternehmens eine Rolle spielen: möchte man lieber einen zahlungsfähigen Kunden abweisen und erscheint abweisend oder möchte man kundenfreundlich auftreten mit dem erhöhten Risiko für Zahlungsausfälle.

Häufig gibt es jedoch einen Trade-off zwischen Präzision und Sensitivität. Das bedeutet, dass die Verbesserung der Präzision dazu führen kann, dass die Sensitivität abnimmt und umgekehrt Die Wahl eines geeigneten Schwellenwerts oder die Anwendung von Techniken wie der Anpassung der Klassifikationsschwelle kann dazu beitragen, ein ausgewogenes Verhältnis zwischen Präzision und Sensitivität zu erreichen, das den Anforderungen des Unternehmens entspricht. In unserem Fall würde man nun versuchen, mit weiteren Informationen das Vorhersagemodell zu verbessern.

Einordnung und Fazit

Der Bayes-Klassifizierer ist ein leistungsstarkes Werkzeug, um Kunden in Kategorien einteilen zu können. Im Beispiel des „Credit Card Dataset" der UCI wurde er genutzt, um die wahrscheinlich zahlungsunfähigen Kunden zu identifizieren. Durch die Anwendung des Bayes-Klassifizierers können Unternehmen fundierte Entscheidungen im Kreditrisikomanagement treffen und Ressourcen effizienter nutzen.

Dabei ist es wichtig, die Unterschiede zwischen Präzision und Sensitivität zu verstehen und die Qualität des Klassifizierers anhand der Evaluierungsmetriken zu bewerten. Eine hohe Präzision gewährleistet die Zuverlässigkeit der Vorhersagen, während eine hohe Sensitivität sicherstellt, dass zahlungsunfähige Kunden nicht übersehen werden. Die wichtigsten Voraussetzungen für genaue und zuverlässige Ergebnisse sind jedoch eine hohe Qualität der Daten und die Auswahl geeigneter Merkmale und Werkzeuge. ■