

24 Nächstgelegener Nachbar

Fortsetzung der Seiten in der 9. Auflage

24.2 Praktische Anwendung

Die Daten. Die Daten der Datei KREDIT.SAV sollen nun in der Anwendung von kNN als Klassifikationsverfahren für ein Demonstrationsbeispiel dienen.

Die Datei enthält mehrere kategoriale Variable. Um diese für eine Berechnung der Euklidischen bzw. Stadtblock-Distanz vorzubereiten, wird von SPSS automatisch für jede kategoriale Variable (auch für ordinal skalierte Variablen) mit mehr als zwei Kategorien eine Dummyvariable, d.h. eine Binärvariable (0/1-Variable) pro Kategorie erzeugt. Beispielsweise hat die Variable ZMORAL (Zahlungsmoral) insgesamt 5 Kategorien. Der kodierte numerische Wert 0 steht für „Keine Kredite bisher/alle bisherigen Kredite zurückgezahlt“, der Wert 1 für „Frühere Kredite bei der Bank einwandfrei abgewickelt“, der Wert 2 für „Noch bestehende Kredite bei der Bank bisher einwandfrei“ usw. Für jede Kategorie wird (aber nur temporär während des Verfahrens) eine Binärvariable gebildet. Die Binärvariable z.B. für die Kategorie „Keine Kredite bisher/alle bisherigen Kredite zurückgezahlt“ bekommt für einen Fall den Wert 1, wenn diese zutrifft und den Wert 0, wenn diese nicht zutrifft. Für die Variable ZMORAL entstehen also 5 Binärvariablen. Da typischerweise für einen Datensatz zur Bewertung von Kreditrisiken die Anzahl der kategorialen Variablen wesentlich höher ist als die der metrischen, wird der Variablenraum zur Berechnung der Distanzen in seiner Dimension durch die Binärvariablenbildung stark aufgebläht.¹ Dieses wirkt sich für den Klassifikationserfolg des Verfahrens sowie auch für die benötigte Rechenzeit ungünstig aus. Daher sollte man vor der Anwendung des Verfahrens die Anzahl der Kategorien kategorialer Variablen durch Zusammenfassungen verkleinern. Wir haben durch Umkodieren in neue Variablen die Kategorien der meisten kategorialen Variablen in zwei Kategorien zusammengefasst.² Dabei haben wir uns von der Plausibilität hinsichtlich des Zusammenhangs zur Zielvariable KRISIKO leiten lassen.

Metrische Variable wie z.B. ALTER und HOEHE (Kredithöhe) haben einen sehr unterschiedlichen Bereich auf der Messskala. Dadurch erhalten sie bei der Distanzberechnung eine unterschiedliche Gewichtung: Variable mit hohen Werten auf der Messskala haben bei der Distanzberechnung einen höheren Einfluss im Vergleich zu Variablen mit relativ kleinen Werten. Da dieses i.d.R. unerwünscht ist, werden die Variablen vor Anwendung des Verfahrens in einen gleichen Zahlenbereich transformiert (normalisiert). Übliche Transformationen sind die z-Transformation (⇒ Kap. 8.5), die Normierung in den Skalenbereich 0 bis 1 oder -1

¹ Ein ähnliches Problem zeigt sich bei der Nutzung von kategorialen Variablen in der Prozedur „Automatische Lineare Modellierung“ (⇒ Kap. 19.1).

² Die neuen umkodierten Variablen haben eine 1 am Ende der Namen der alten Variablen. Auf den Internetseiten zum Buch finden Sie die Beschreibung der Umkodierungen in Form der dafür verwendeten SPSS-Syntaxbefehle.

bis 1. In SPSS wird für das kNN-Verfahren die letzte der genannten Varianten genutzt. Dafür wird folgende Berechnung vorgenommen:

$$x_i^{\text{normalisiert}} = \frac{2(x_i - x_{\min})}{(x_{\max} - x_{\min})} - 1 \quad (24.7)$$

Durchführen der Analyse. Nach Öffnen der Datei KREDIT.SAV gehen Sie wie folgt vor:

- ▷ Klicken Sie die Befehlsfolge „Analysieren“ „Klassifizieren“, „Nächstgelegener Nachbar“. Es öffnet sich die in Abb. 24.2 dargestellte Dialogbox „Nächste-Nachbarn-Analyse“, die unterhalb des Namens der Dialogbox mehrere Registerkarten zum Spezifizieren des Verfahrens enthält. Geöffnet ist die Registerkarte „Variablen“.

Im Folgenden werden wir die Spezifizierungen auf den einzelnen Registerkarten für unser Anwendungsbeispiel erläutern und dabei auch auf die nicht genutzten Optionen eingehen.

- ▷ *Registerkarte „Variablen“.* Wir übertragen die Variable KRISIKO aus der Quellvariablenliste in das Eingabefeld „Ziel (optional):“ (⇒ Abb. 24.2). Wichtig ist, dass das Messniveau für die kategoriale Variable KRISIKO als „Nominal“ eingestellt ist. Hat man dieses in der Variablenansicht des Daten-Editors eventuell noch nicht gemacht, so kann man dies hier schnell temporär durch einen Rechtsklick auf die Zielvariable (hier KRISIKO) in der Quellvariablenliste nachholen. Es öffnet sich dann ein Kontextmenü und man wählt das erforderliche Messniveau. Wenn das Messniveau „Metrisch“ ist, wird das Verfahren kNN als Prognoseverfahren für eine metrische Variable genutzt.

Wird keine Zielvariable festgelegt, so werden die nächsten Nachbarn eines Falles bestimmt, aber es wird keine Klassifikation vorgenommen. Daher ist es auch nicht möglich, ein optimales k suchen zu lassen.

- ▷ In das Eingabefeld „Merkmale“ werden die für die Distanzberechnung zu verwendenden Variablen (auch Prädiktoren genannt) übertragen.

Je nachdem, ob man prädiktive Variablen (Variablen mit Einfluss auf die Zielvariable) für die Distanzberechnung durch das Verfahren auswählen lassen möchte oder nicht, muss man hier unterschiedlich vorgehen.

Will man prädiktive Variablen für die Nachbarschaftsbestimmung automatisch auswählen lassen [dieses wird auf der Registerkarte „Funktionen“ (Übersetzungsfehler!)³ angefordert], so kann man hier nach dem Motto vorgehen: eher eine Variable zu viel als eine zu wenig für die Analyse verwenden. Das Selektionsverfahren soll ja die relevanten Variablen auswählen und die irrelevanten außen vor lassen. Aber es gibt einen Nachteil bei einer zu großen Anzahl von Analysevariablen. Die Rechenzeit wird durch den höheren Aufwand für den Ausleseprozess erhöht. Daher sollte man i.d.R. schon vor Anwendung des Verfahrens eine gewisse grobe Vorauswahl für die zu nutzende Variablen durchführen. Hierfür bietet es sich für die kategorialen Variablen an, den

³ Auf der englischsprachigen Oberfläche (man kann die Sprache der Benutzeroberfläche auf der Registerkarte „Allgemein“ im Menü „Bearbeiten“, „Optionen“ umstellen) steht hier „Features“. Das ist ein im Data Mining und Machine Learning üblicher Begriff für Variable und ist mit Funktionen falsch übersetzt.

Zusammenhang zur Zielvariable KRISIKO mit einem Chi-Quadratstest (\Rightarrow Kap. 10.3) zu prüfen. Variable mit z.B. einem Signifikanz-Wert $> 0,05$ könnte man außen vor lassen. Für metrische Variablen könnte man eine grobe Vorauswahl vornehmen, indem man Korrelationskoeffizienten berechnet und Variable mit kleinem Korrelationskoeffizienten nicht einbezieht oder einen varianzanalytischer F-Test durchführt⁴.

Wir haben durch eine derartige Vorgehensweise folgende Variable vorselektiert: LAUFZEIT (Laufzeit des Kredits), HOEHE (Kredithöhe), LAUFKONTO1 (Laufendes Konto), ZMORAL1 (Zahlungsmoral), VERWENDG1 (Kreditverwendung); SPARKONTO1 (Spar- bzw. Wertpapierkonto), BESCHZEIT1 (gegenwärtige Beschäftigung seit), RATENHOEHE1 (Kreditratenhöhe), VERMOEGEN1 (Vermögen), WEITKREDITE1 (weitere Kredite) und WOHNUNG1 (Wohnungseigentum). Damit wird die Anzahl der Variablen erheblich reduziert.

Bei Anwendung des kNN-Verfahrens mit einer automatischen Selektion aus diesen Variablen (\Rightarrow Registerkarte „Funktionen“) kombiniert mit der Bestimmung des optimalen k (Registerkarte „Nachbarn“) haben wir sehr unterschiedliche Lösungsergebnisse hinsichtlich der zu bestimmenden Höhe von k , hinsichtlich der als relevant ausgewählten Variablen und auch hinsichtlich der erzielten Trefferquote erhalten. Um eine stabilere Lösung zu bekommen, haben wir uns daher gegen die automatische Auswahl der Variablen entschieden. Mit dieser Entscheidung verbunden ist, dass automatisch eine V -fache Kreuzvalidierung (\Rightarrow Registerkarte „Partitionen“) vorgenommen wird. Von dieser kann man erwarten, dass man eine stabilere Lösung erhält.

Möchte man auf das Auswählen von prädiktiven Variablen durch das Verfahren Nächste Nachbarn verzichten, so sollten in das Eingabefeld „Funktionen“ möglichst nur prädiktive Variable übertragen werden. Diese müssen also in einem vorausgehenden Analyseschritt gefiltert werden. Dazu kann man die oben im Zusammenhang mit einer Vorselektion genannte Vorgehensweisen nutzen. Um eine möglichst stabile Lösung zu bekommen, kann man das kNN-Verfahren auch wiederholt anwenden und bei den Wiederholungen jeweils die Variablen aus dem Modell entfernen, die in der vorherigen Ergebnisausgabe in der Rangfolge der Wichtigkeit am Ende stehen (\Rightarrow Abb. 24.9 rechtes Fenster). Für diese Vorgehensweise spricht, dass man möglichst eine in der Anzahl der Variablen „sparsame“ Lösung anstreben sollte.⁵ Auf diese Weise sind wir vorgegangen und haben die oben genannten vorselektierten Variable weiter auf folgende sechs reduziert: LAUFZEIT, HOEHE, LAUFKONTO1, ZMORAL1, SPARKONTO1 und BESCHZEIT1.

Mit diesen Variablen soll nun das kNN-Verfahren ohne Nutzung der automatischen Auswahl von Variablen demonstriert werden. Dazu übertragen wir

⁴ Seit SPSS 24 ist die Prozedur Naive Base (\Rightarrow Kap. 25) verfügbar. Mit diesem Verfahrens kann man für eine kategoriale Zielvariable prädiktive Variable selektieren. Seit SPSS 24 steht ein zweites Verfahren zur Selektion von Prädiktoren zur Verfügung (\Rightarrow Kap. 26).

⁵ Im Data Mining und Machine Learning wird dieses Grundprinzip, bei annähernd ähnlichen Klassifikationsergebnissen einfache Vorhersagemodelle zu bevorzugen, als Occams' razor bezeichnet. Es wird dafür auch eine Albert Einstein zugeschriebener Aussage verwendet: so einfach wie möglich, aber nicht einfacher.

die Variable LAUFZEIT, HOEHE, LAUFKONTO1, ZMORAL1, SPARKONTO1 und BESCHZEIT1 in das Eingabefeld „Merkmale“.

Die Option „Skalierungsmerkmale normalisieren“ ist voreingestellt. Mit dieser Option werden die metrischen Variablen entsprechend der Gleichung 24.7 in den Skalenbereich -1 bis 1 transformiert. Diese Option behalten wir bei, da die Variablen LAUFZEIT und HOEHE sehr unterschiedliche Bereiche auf der Messskala einnehmen.⁶

- ▷ In das Eingabefeld „Fokusfall-ID (optional)“ übertragen wird die Variable FOKUSFALL. Mit dieser optionalen Spezifizierung kann man bestimmte Fälle bei der Betrachtung der Analyseergebnisse herausheben, in den Fokus stellen (ID steht für identifizier). Zur Demonstration haben wir beispielhaft die Fälle mit den Fallzahlen 299 bis 302 als Fokusfälle angenommen. Die Fälle 299 und 300 sind Fälle mit KRISIKO = 0 und die Fälle 301 und 302 Fälle mit KRISIKO = 1.

In das Eingabefeld „Fallbeschriftung (optional)“ kann man eine Variable übertragen, deren Variablenwerte von Fällen in der Ergebnisausgabe angezeigt werden soll. Wir übertragen FALLNR in das Feld. Verzichtet man darauf werden standardmäßig die internen Fallnummern zur Kennzeichnung der Fälle verwendet.

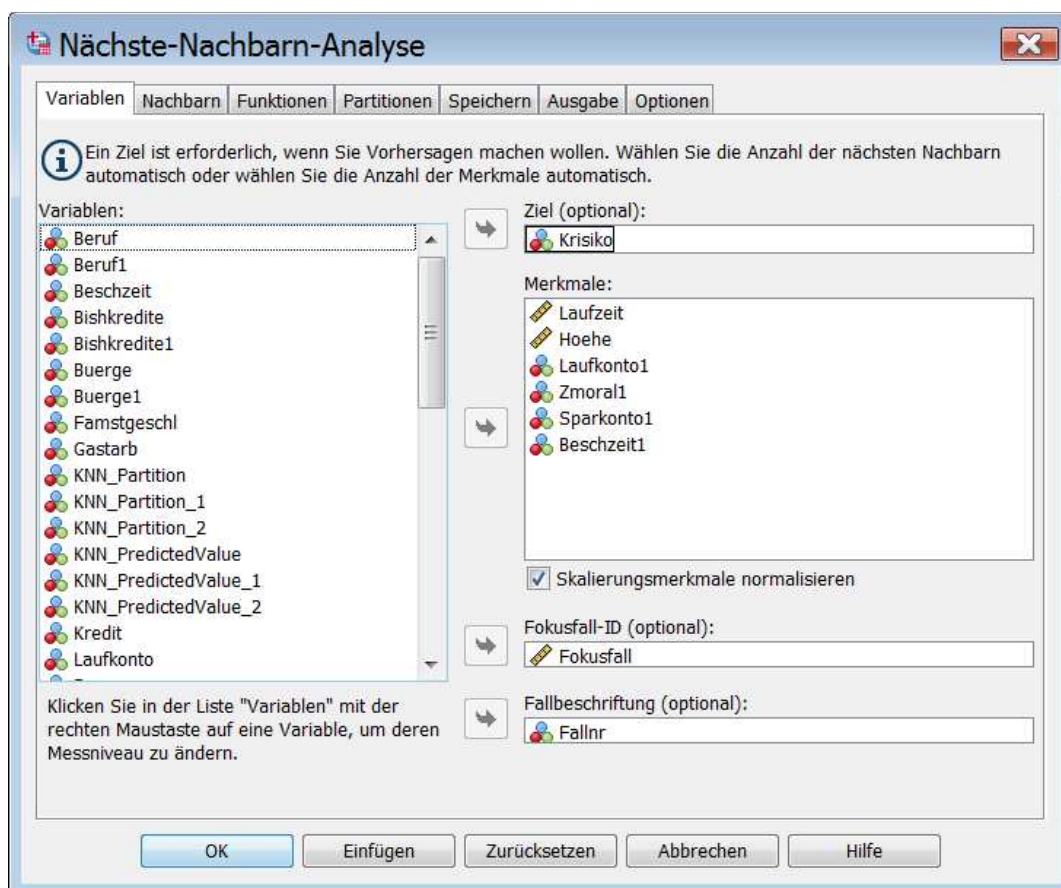


Abb. 24.2. Registerkarte „Variablen“

⁶ LAUFZEIT hat Werte von 4 bis 72 Monate, HOEHE Werte von 250 bis 18424 DM.

- ▷ Die Registerkarte „Nachbarn“. Nun wählen wir die Registerkarte „Nachbarn“ (⇒ Abb. 24.3). Im Feld „Anzahl der nächstgelegenen Nachbarn (k)“ wählen wir die Option „Automatisch k auswählen“. Man wird i.d.R. ein hinsichtlich des Klassifikationserfolgs richtiges k nicht kennen und daher von der Option einen „Festen k-Wert eingeben“ selten Gebrauch machen. Wir tragen „Minimum = 3“ und „Maximum = 6“ ein. Auf welche Weise SPSS das optimale k aus dem angegebenen Bereich zwischen $k = 3$ und $k = 6$ bestimmt, wird unten erläutert (⇒ Registerkarte „Partitionen“).
- ▷ Im Feld „Distanzberechnung“ wählen wir „Euklidische Metrik“. Des Weiteren aktivieren wird die Option „Merkmale bei Berechnung von Abständen nach Wichtigkeit gewichten“. Mit dieser Option werden die Variablen bei der Distanzberechnung gemäß Gleichung 24.4 bzw. 24.5 gewichtet. Wie SPSS die Gewichte der Variablen bestimmt, wird unten erläutert („Bestimmen der Variablen Gewichte“).

Die beiden Optionen im Feld „Vorhersagen für metrisches Ziel“ sind inaktiv geschaltet. Nur für den Fall, dass auf der Registerkarte „Variablen“ in das Eingabefeld „Ziel (optional):“ eine metrische Variable übertragen wird, kann man hier wählen, ob als Prognosewert \hat{y}_i eines Falles i der Mittelwert oder der Median der k Nachbarn genommen werden soll.

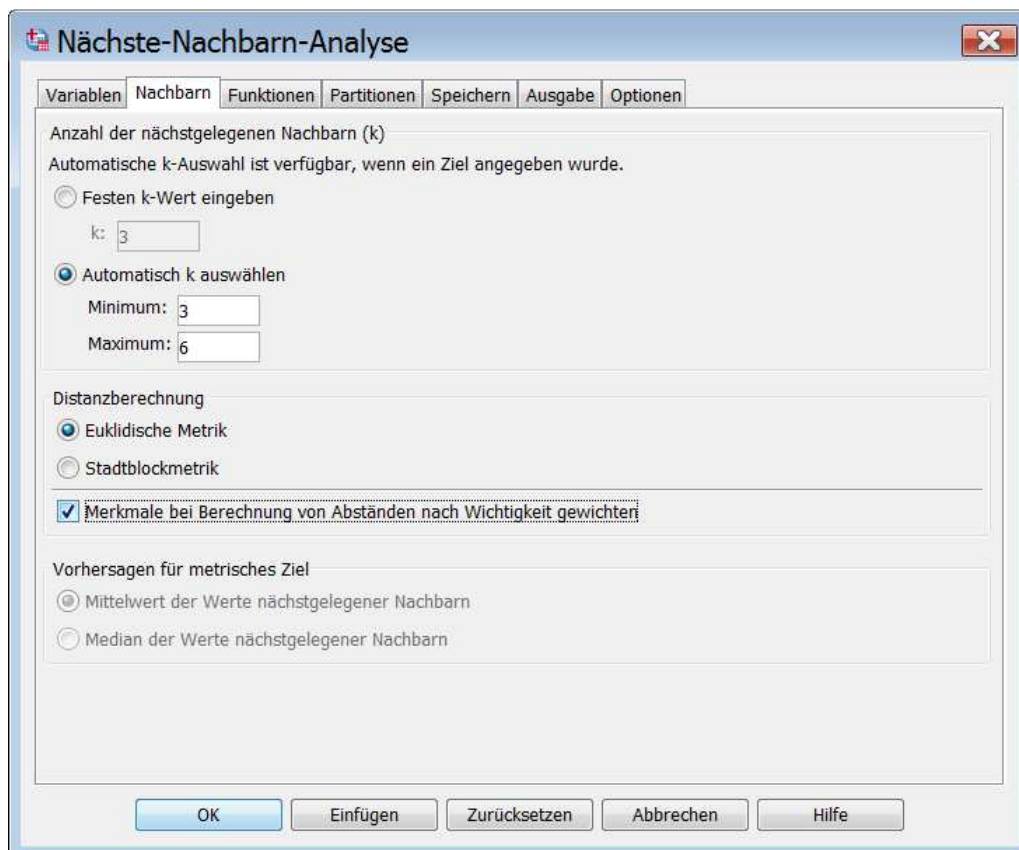


Abb. 24.3. Registerkarte „Nachbarn“

▷ *Die Registerkarte „Funktionen“*. Auf der Registerkarte „Funktionen“ (Übersetzungsfehler!)⁷ (⇒ Abb. 24.4) wird durch Einschalten des Optionsschalters „Merkmalsauswahl durchführen“ eine automatische Auswahl der für die Nachbarschaftsbestimmung relevanten Variablen angefordert. Alle auf der Registerkarte „Variablen“ in die Analyse ausgewählten Variable erscheinen hier in einer Liste und werden in das Auswahlverfahren einbezogen. Dabei wird eine Vorwärtsauswahl durchgeführt: als erstes wird die für den Klassifikationserfolg wichtigste Variable einbezogen, dann die zweitwichtigste usw.. Da diese Vorwärtsauswahl durch eine Abbruchbedingung gestoppt wird, kann verhindert werden, dass irrelevante Variablen einbezogen werden.

Möchte man sicherstellen, dass bestimmte Variablen sich diesem Ausleseprozess gar nicht stellen sollen, weil man sie unbedingt in die Distanzberechnung einbeziehen möchte, so kann man diese in das Eingabefeld „Erzwungener Eintrag“ übertragen.

Im Feld „Stoppkriterium“ hat man zwei Optionen, um den Auswahlprozess der Variablen durch eine Abbruchbedingung zu steuern. Man kann in das Eingabefeld „Auszuwählende Anzahl:“ eine maximale Anzahl von Variablen vorgeben. Ist diese Vorgabe erreicht, wird der Auswahlprozess abgebrochen.

Alternativ kann man den Auswahlprozess durch eine Vorgabe der maximalen Verringerung der Fehlerquote (⇒ Gleichung 24.2) stoppen lassen. Sobald das Auswählen einer weiteren Variable die Fehlerquote um weniger als z.B. den voreingestellten Wert 0,01 mindert, endet der Auswahlprozess. Im Eingabefeld „Minimale Änderung“ kann man den voreingestellten Wert ändern. Erhöht man den Wert, so führt das zu einer kleineren Anzahl ausgewählter Variablen.

Das Verfahren zur automatischen Auswahl von relevanten Variablen aus einer Liste von Variablen ist in den Prozess der Klassifikation eingebunden.⁸ In dem Auswahlprozess wird für jede Variable geprüft, um wieviel sich die Fehlerquote der Klassifikation verringert, wenn die Variable in die Nachbarschaftsbestimmung einbezogen wird. Aufgenommen wird jeweils die Variable, bei der sich die Fehlerquote am meisten verringert. Wird das automatische Auswählen von relevanten Variablen mit dem Bestimmen eines optimalen k kombiniert, so wird das Auswahlverfahren für jedes k (im vom Nutzer angegebenen Bereich) angewendet. Die optimale Lösung sucht die Kombination von k und Variablen-set mit der kleinsten Fehlerquote (⇒ Registerkarte „Partition“).

Für unser Anwendungsbeispiel verzichten wir auf die automatische Auswahl von Variablen, da wir eine Vorselektion von Variablen vorgenommen haben (⇒ Registerkarte „Variablen“) und die V -fache Kreuzvalidierung nutzen wollen (⇒ Registerkarte „Partitionen“).

⁷ S. Fußnote 3.

⁸ Im Machine Learning spricht man von einer Wrapper-Methode. Eine Alternative dazu ist die Filter-Methode. Hierbei werden die relevanten Variablen vor der Anwendung des Klassifikationsverfahrens mit speziellen Verfahren ausgewählt (⇒ P. Cunningham, S. J. Delany (2007), S. 8 ff.).

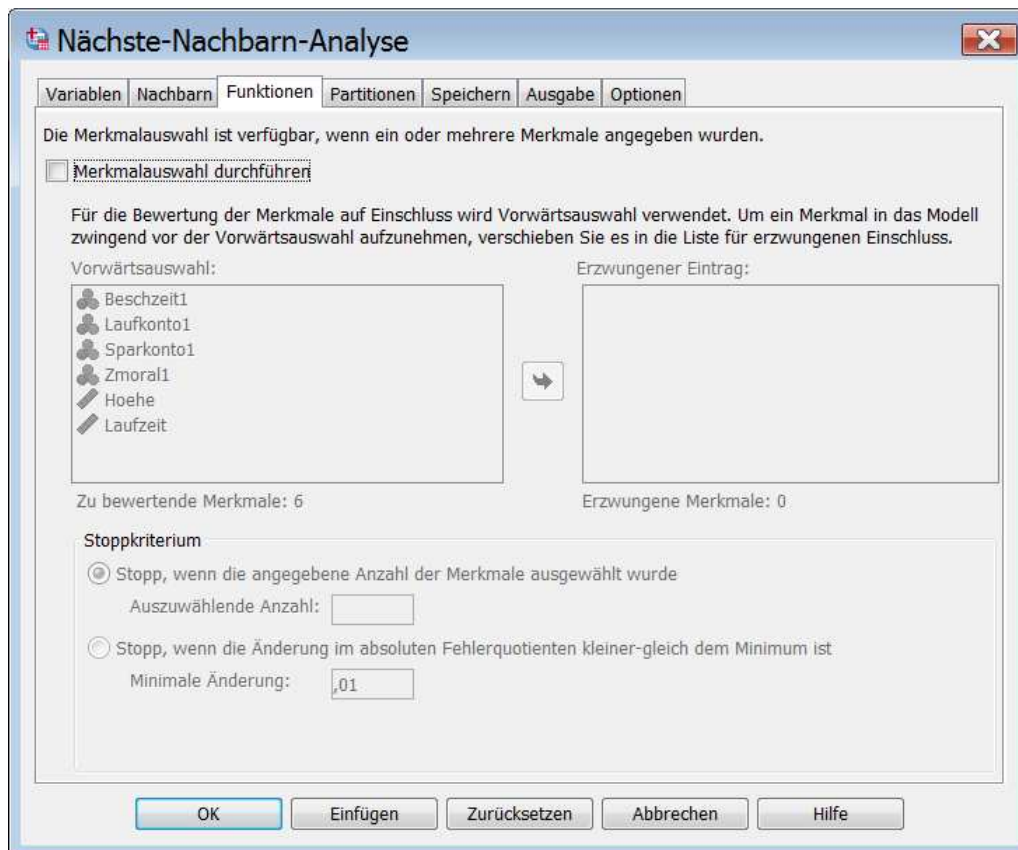


Abb. 24.4. Registerkarte „Funktionen“

Registerkarte „Partitionen“. Als grundlegende Methodik im Machine Learning und Data Mining hat sich als Standard etabliert, zuerst ein Modell (hier ein Klassifikationsmodell) anhand von Datenfällen (Trainingsdaten genannt) zu entwickeln und dann mittels anderer Datenfälle (Testdaten genannt) die Vorhersagegüte des Modells zu beurteilen.⁹ Die Vorhersagen eines Modells mittels „neuer“ Daten (im Vergleich zu den Daten, mit denen das Modell entwickelt wurde) bieten ein zuverlässigeres Bild für die Vorhersagegüte.¹⁰ In unserem Fall einer Klassifikationsaufgabe besteht die Entwicklung eines Modells darin: 1. das „beste“ k zu finden und 2. die „besten“ Variablen auszuwählen. Leitlinie ist dabei die kleinste Fehlerquote für die Testdaten. Sind k und die Variablen bestimmt, dann werden des Weiteren die Variablen Gewichte bestimmt. Auch diese Aufgabe ist mit den Trainingsdaten zu lösen. Mit den Testdaten kann man anschließend die Höhe der Fehlerquote des Modells überprüfen. Zur Gewinnung von Trainings- und Testdatenfällen aus verfügbaren Datenfällen werden den Teildateien Trainings- und Testdaten die Datenfälle aus der Gesamtdatei zufällig per Zufallsgenerator zugeordnet.

⁹ Ausführlicher \Rightarrow I. H. Witten und E. Frank, S. 144 ff.

¹⁰ Insbesondere soll eine Überanpassung des Modells an die Daten (Overfitting genannt) verhindert werden.

- ▷ Im Feld „Trainings- und Holdout-Partitionen“ (⇒ Abb. 24.5) wird festgelegt, mit welcher Quote die Datei zufällig in Trainings- und Testdaten (hier Holdout genannt) aufgeteilt werden soll. Die voreingestellte Quote 70 % Testdaten-Partition und 30 % Holdout-Partition ändern wir nicht.

Alternativ kann man auch die Option „Variable verwenden, um Fälle zuzuweisen“ nutzen, um die Aufteilung in Trainings- und Testdaten vorzunehmen. Diese Variable muss die gewünschte Aufteilung definieren (am besten mittels einer 0/1-Variable)¹¹.

Die zwei Optionen im Feld „Kreuzvalidierungsaufteilungen“ sind nur dann wählbar, wenn man 1. auf der Registerkarte „Nachbarn“ eine automatische Bestimmung von k und 2. auf der Registerkarte „Funktionen“ nicht „Merkmalsauswahl durchführen“ gewählt hat. Hat man diese Anwendungsform gewählt, wird eine V -fache Kreuzvalidierung (hier „ V -Fold-Kreuzvalidierung“ genannt) der Klassifikationsergebnisse durchgeführt.¹² Die Kreuzvalidierung dient zum Bestimmen des optimalen k . Die grundlegende Idee der Kreuzvalidierung haben wir schon im Zusammenhang mit der Aufteilung des Datensatzes in Trainings- und Testdaten kennengelernt. Hier wird sie in gewandelter Form genutzt.

Voreingestellt ist „Anzahl der Aufteilungen“ $V = 10$.¹³ Man kann als Anwender den Wert V festlegen. Für $V = 10$ wird der Trainingsdatensatz in 10 nicht-überlappende Teildateien mit etwa gleich vielen Fällen aufgeteilt. Die Zuordnung der Fälle zu den Teildateien erfolgt zufällig (per Zufallsgenerator). Nun wird für die Fälle ausschließlich der ersten Teildatei (der Holdout-Fälle der Trainingsdaten) das Klassifikationsverfahren angewendet und der Klassifikationsfehler FQ (⇒ Gleichung 24.2) für die Holdout-Fälle der Trainingsdaten berechnet. Danach setzt sich das Verfahren fort, indem die Fälle der zweiten Teildatei, der Trainingsdaten, von der Anwendung des Verfahrens ausgeschlossen werden und die Fehlerquote für die Fälle der zweiten Subdatei berechnet wird usw. Ergebnis sind im Fall der 10-fachen Kreuzvalidierung 10 Fehlerquoten für die jeweiligen Holdout-Fälle der Trainingsdaten. Aus diesen wird ein arithmetisches Mittel berechnet. Die Kreuzvalidierung wird für jedes k aus dem vom Anwender vorzugebenden Bereich von k durchgeführt. Das optimale k ergibt sich als das mit der kleinsten durchschnittlichen Fehlerquote der Kreuzvalidierung.

Wenn auf diese Weise das optimale k bestimmt ist, wird das Verfahren auf die Testdaten (Holdoutfälle der gesamten Datei)¹⁴ angewendet zur Abschätzung der zu erwartende Fehlerquote für neue Daten, bei denen die Klassenzugehörigkeit unbekannt ist. Um den Zufallseinfluss auf die geschätzte Fehlerquote weiter zu mindern wird empfohlen, die Kreuzvalidierung 10-fach zu wiederholen und aus den sich ergebenden 10 durchschnittlichen Fehlerquoten eine

¹¹ 1 bzw. positive Werte für die Trainingsdaten und 0 bzw. negative Werte für die Testdaten.

¹² Wird neben dem Bestimmen von k auch ein automatisches Auswählen von Variablen angefordert, so ist die V -fache Kreuzvalidierung ausgeschlossen, weil der Rechenaufwand zu hoch wird.

¹³ Eine 10-fache Kreuzvalidierung wird empfohlen (⇒ I. H. Witten und E. Frank, S. 150).

¹⁴ Diese werden auch Validierungsdaten genannt.

durchschnittliche zu berechnen.¹⁵ Für die Aufteilung in Teildateien kann auch eine Variable mit Werten von 1 bis V verwendet werden.

- ▷ Die Option „Start für Mersenne-Twister festlegen“ (⇔ „Auswählen einer Zufallsstichprobe“ in Kap. 7.4.2) erlaubt es, einen „Start“-Wert für den Zufallsgenerator festzulegen. Damit ist es möglich, eine kNN-Anwendung in gleicher Weise zu wiederholen.¹⁶ Damit Sie unser Ergebnis reproduzieren können, müssen Sie dort auch den von uns gewählten Wert 555 angeben.

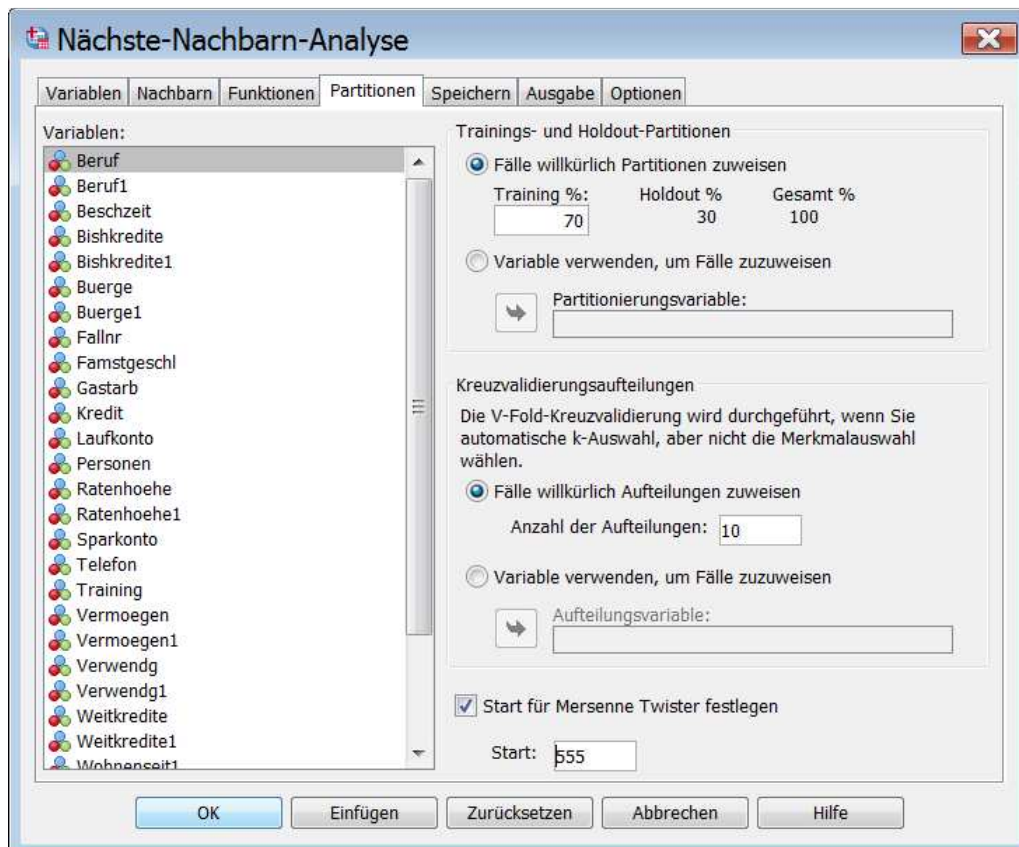


Abb. 24.5. Registerkarte „Partitionen“

Registerkarte „Speichern“. Auf dieser (⇔ Abb. 24.6) kann man festlegen, ob und auf welche Weise Ergebnisse des kNN-Verfahrens als Variable im Daten-Editor gespeichert werden sollen. Als „Zu speichernde Variable“ können folgende gewählt werden (in Klammern der Variablenname im Daten-Editor):

- *Vorhergesagte(r) Wert oder Kategorie* (KNN_PredictedValue): Die Kategorie (Klasse) der Zielvariablen (bzw. der vorhergesagte Prognosewert \hat{y} im Fall einer metrischen Zielvariablen).

¹⁵ Vergl. I. H. Witten und E. Frank (2005), S. 150.

¹⁶ Die zufällige Zuteilung der Fälle auf die Trainings- und Testdaten sowie die zufällige Fallzuweisungen für die V-fache Kreuzvalidierung kann so reproduziert werden.

- *Vorhergesagte Wahrscheinlichkeit (kategoriales Ziel)* (KNN_Probability_2 für KRisiko = 1 der vorhergesagten Klasse j für den Fall i.¹⁷ Bezeichnet man mit k_j die Anzahl der Fälle k, die der Klasse j angehören, so berechnet sich die Wahrscheinlichkeit für einen Fall i wie folgt:

$$P_i(j) = \frac{k_j + 1}{k + J} \quad (24.8)$$

Die Formel für die geschätzte Wahrscheinlichkeit (definiert als relative Häufigkeit k_j/k) ist im Zähler mit 1 und im Nenner mit J (der Anzahl der Klassen der Zielvariable, in unserem Beispiel ist $J = 2$) ergänzt. Mit dieser Laplace-Korrektur der Wahrscheinlichkeit soll für kleine k erreicht werden, dass die geschätzte Höhe der Wahrscheinlichkeit in Richtung $1/J$ bestimmt wird. Insbesondere sollen geschätzte Wahrscheinlichkeiten von 1 vermieden werden.

- *Trainings/Holdout-Partitionsvariable* (KNN_Partitionen). Die mit 1 kodierten Fälle sind die Trainings- und die mit 0 die Testdaten.
- *Kreuzvalidierung-Aufteilungsvariable* (KNN_Fold). Bei Nutzung der V-fachen Kreuzvalidierung sind die mit 0 kodierten Fälle die Trainingsdaten und die mit 1 bis V kodierten Fälle die Fälle der V Teildateien.

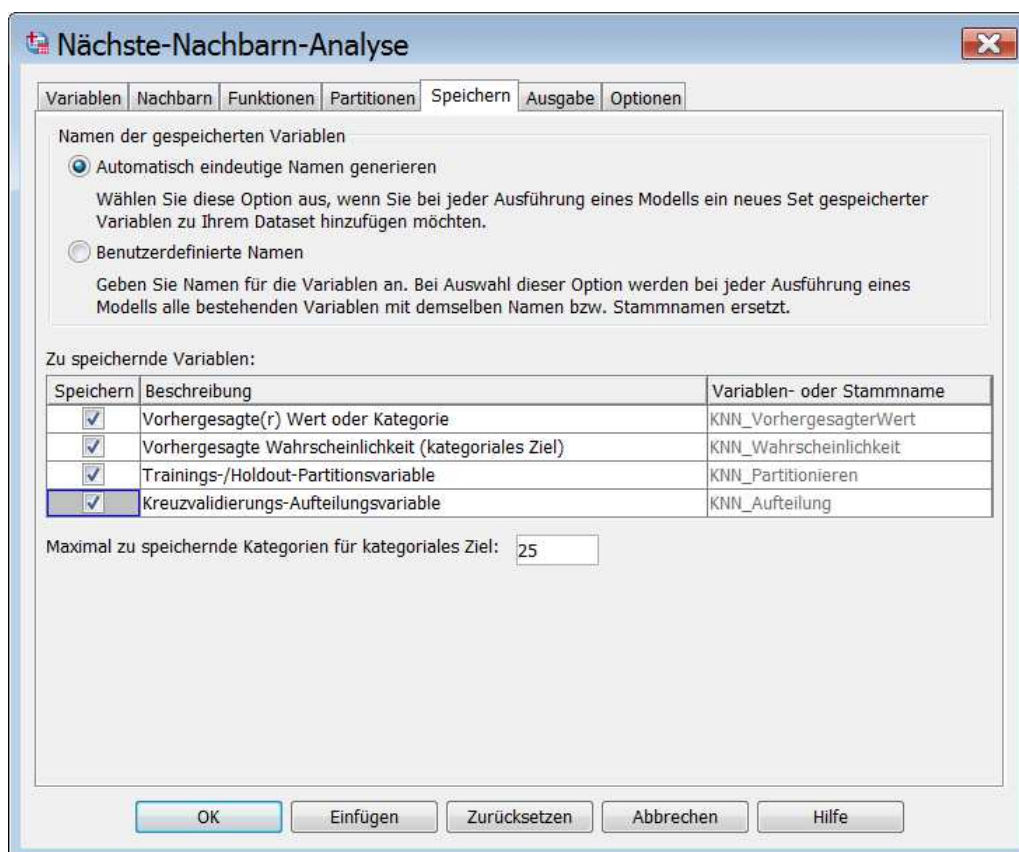


Abb. 24.6. Registerkarte „Speichern“

¹⁷ Entsprechend wird KNN_Probability_1 für KRISIKO = 0 in den Editor aufgenommen.

▷ *Registerkarte „Ausgabe“*. Für die Ausgabe kann man eine „Zusammenfassung der Fallverarbeitung“ sowie „Diagramme und Tabelle“ wählen. Wir wählen beide Optionen (⇒ Abb. 24.7).

Im Feld „Dateien“ kann man anfordern, die Modellspezifikationen des kNN-Modells in einer XML-Datei zu speichern (Extensible Markup Language). Wir haben die Datei kNN_Kreditdaten genannt und in einem Verzeichnis gespeichert (⇒ Abb. 24.7). Diese Datei wird zum Klassifizieren „neuer“ Daten genutzt (⇒ Klassifizieren neuer Daten).

Des Weiteren kann man sich für weitere Nutzungen die k Nächsten Nachbarn der Fokusfälle sowie deren Distanzen zu den k Nächsten Nachbarn in neue SPSS-Dateien bei Angabe eines Namens ausgeben lassen. Dabei hat man zwei Optionen:

- „*Neues Data-Set erstellen*“ erzeugt ein SPSS-Datenblatt mit diesen Daten.
- „*Neue Datendatei schreiben*“ erzeugt eine .SAV Datei mit diesen Daten.

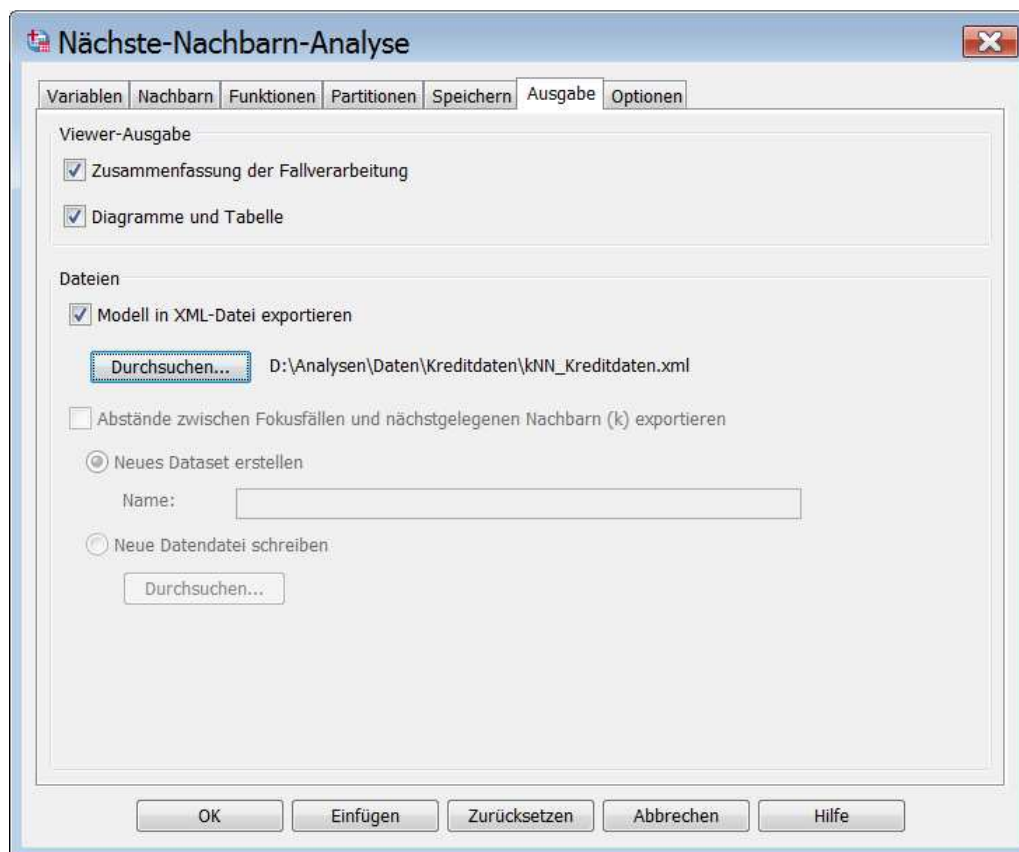


Abb. 24.7. Registerkarte „Ausgabe“

Registerkarte „Optionen“. Auf dieser kann man wählen, wie mit Fällen mit benutzerdefinierten fehlenden Werten in kategorialen Variablen umgegangen werden soll. Wählt man „Einschließen“, so werden diese Fälle als normale gültige Fälle einbezogen. Fälle mit systemdefinierten fehlenden Werten sowie für metrische Variablen werden stets ausgeschlossen.

Bestimmen der Variablen Gewichte. Hat man auf der Registerkarte „Nachbarn“ (\Rightarrow Abb. 24.3) die Option „Merkmale bei der Berechnung von Abständen nach Wichtigkeit gewichten“ gewählt, werden bei der Berechnung der Distanzen gemäß Gleichungen 24.4 und 24.5 die Variablen mit Wichtigkeitsgewichten g_h gewichtet. Wichtigere Variable sollen bei der Distanzberechnung ein höheres Gewicht erhalten als weniger wichtige. Verzichtet man auf diese Option, so werden alle Variable gleich gewichtet (alle Gewichte werden auf 1 gesetzt).

Um das Wichtigkeitsgewicht der in die Distanzberechnung einbezogenen Variablen zu bestimmen, wird wie folgt vorgegangen:

Sind im Vorwärtsauswahlprozess die Variablen $x_1, x_2, x_3, \dots, x_m$ ausgewählt, so wird zur Bestimmung z.B. des Wichtigkeitsgewichts der Variable x_2 diese Variable aus dem Set der verwendeten Variablen weggelassen und das Verfahren erneut angewendet. Die sich nun unter Ausschluss der Variable x_2 ergebende Fehlerquote FQ_2 für die Testdaten ist Basis zur Berechnung des Maßes Variablenwichtigkeit VW_2 für die Variable x_2 :

$$VW_2 = FQ_2 + \frac{1}{m} \quad (24.9)$$

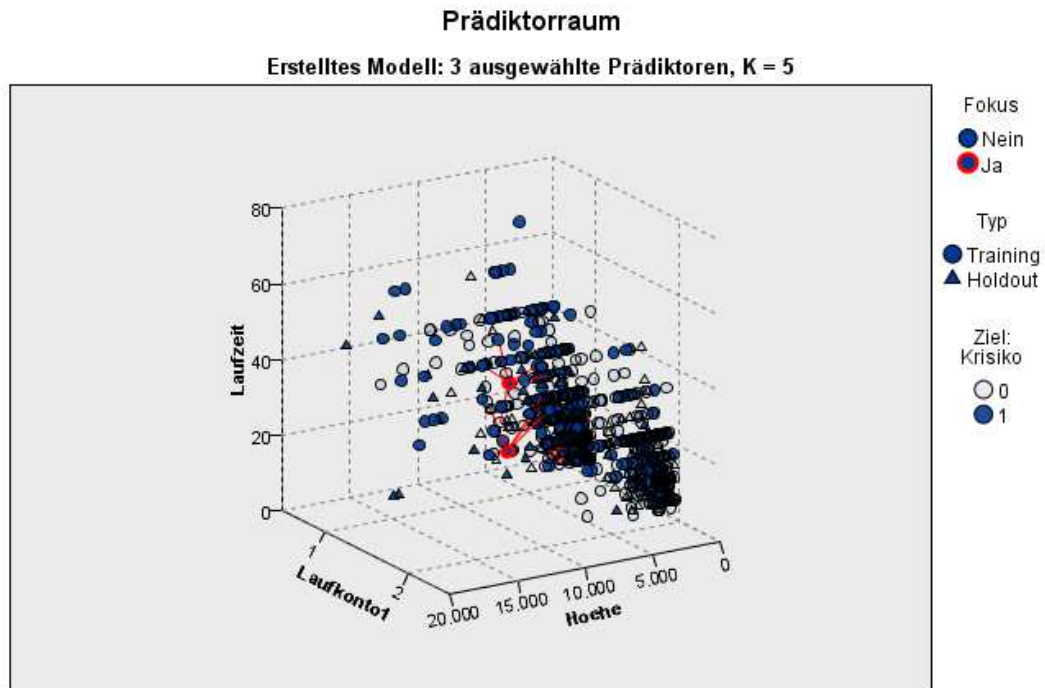
Sind auf diese Weise die Maße Variablenwichtigkeit für alle Variablen $x_1, x_2, x_3, \dots, x_m$ berechnet, wird in einem weiteren Schritt das Gewicht g_2 der Variable x_2 als relativer Anteil aller Variablenwichtigkeitsmaße der m Variablen berechnet:

$$g_2 = \frac{VW_2}{\sum_{h=1}^m VW_h} \quad (24.10)$$

Die Analyseergebnisse. Als Ergebnisse erhalten wir zwei Ausgaben im Ausgabefenster: erstens eine Tabelle zur Zusammenfassung der Fallbearbeitung. Wir verzichten wegen ihrer Einfachheit hier auf ihre Darstellung. Zweitens wird eine i.d.R. dreidimensionale Grafik ausgegeben. Es zeigt die Fälle der Trainings- und Testdaten im durch die Variablen LAUFZEIT, LAUFKONTO1 und HOEHE aufgespannten dreidimensionalen Raum (\Rightarrow Abb. 24.8). Diese drei Variablen sind vom Verfahren als die für die Klassifikation wichtigsten erkannt (\Rightarrow Abb. 24.9).

Bei der in Abb. 24.8 zu sehenden Grafik im Viewer handelt es sich aber nicht um eine herkömmliche Grafik, sondern um ein Modell-Objekt. Durch Doppelklicken auf das Modell-Objekt (alternativ: mit rechtem Maustaste auf das Modell-Objekt klicken und im sich öffnenden Kontextmenü „Inhalt bearbeiten im separaten Fenster wählen“) wird eine Modellanzeige („Modellviewer“) geöffnet (\Rightarrow Abb. 24.9). In dieser kann man sich in interaktiver Form detaillierte Informationen zum Modell anzeigen lassen. Es ist aber auch ein Bearbeiten der Grafik möglich. Es lassen sich z.B. die Variablen auf den Achsen der Grafik durch andere einbezoge-

ne Variablen austauschen und man kann Überarbeitungen im Layout der Grafik vornehmen¹⁸



Dieses Diagramm ist eine kleinere Projektion des Prädiktorraums, der insgesamt 6 Prädiktoren enthält.

Abb. 24.8. Das Modell-Objekt: die Fälle im 3D-Variablenraum

¹⁸ Für die Bearbeitungsmöglichkeiten verweisen wir auf das Hilfesystem. Die Vorgehensweise entspricht weitgehend der für die Bearbeitung von Grafiken im „Grafiktafel-Editor“ (⇨ Kap. 34.4).

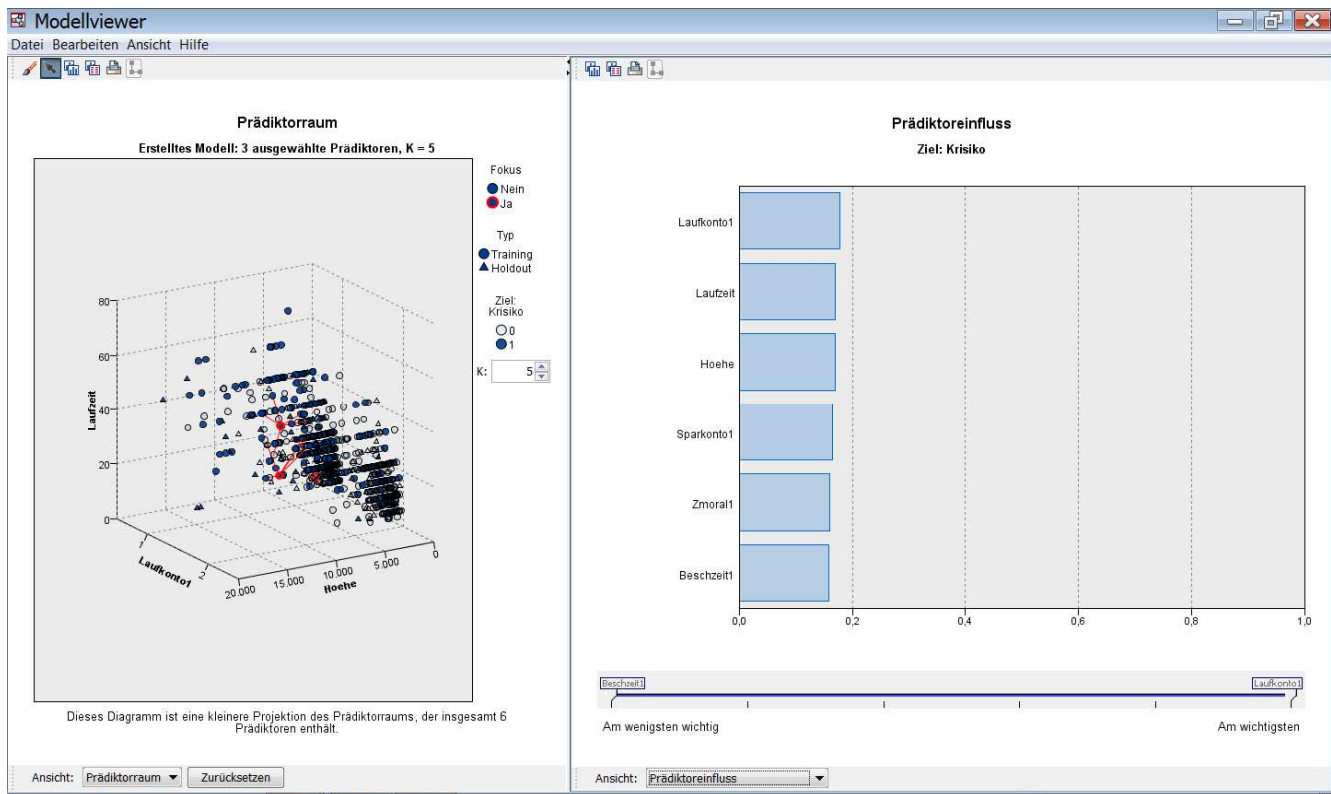



Abb. 24.9. Die Modellanzeige für Ansicht = Prädiktoreinfluss (Wichtigkeit)

Die Modellanzeige teilt sich in zwei Fenster. Im linken (Hauptansicht genannt) ist die dreidimensionale Grafik (als Variablen- bzw. Prädiktorraum) zu sehen und im rechten die Hilfsansicht. In der Hilfsansicht kann man sich mehrere detaillierte Informationen zum Modell in grafisch aufbereiteter Form oder in Tabellen anzeigen lassen.



Für einige der Hilfsansichten besteht eine Verknüpfung mit gewählten (markierten) Teilen der Grafik in der Hauptansicht.

Die Trennlinie zwischen den Ansichten kann man verschieben (mit dem Cursor auf die Linie gehen, mit der linken Maustaste festhalten und nach links bzw. rechts ziehen). Mit den Pfeilen  (oberhalb der Trennlinie) kann man je nach Wunsch sich nur die Haupt- bzw. Hilfsansicht in einem Fenster anzeigen lassen. Mit den Pfeilen kann man auch wieder in die Zweiteilung zurückkommen.

Die Hauptansicht. Ähnlich wie der Daten-Editor oder das Ausgabefenster hat auch die Modellanzeige eine eigene Menüleiste mit Menüs zum Aufrufen spezieller Befehle sowie Symbolleisten mit spezifischen Symbolen.

Datei. Der Befehl „Eigenschaften“ öffnet eine Dialogbox. Hier kann man wählen, ob man für eine Druckausgabe nur die Hauptansicht („nur sichtbare Ansicht ...“) oder alle Modellansichten einschließlich aller Ansichten der Hilfsansicht ausgedruckt werden sollen. Mit „Schließen“ wird die Modellanzeige geschlossen.

Bearbeiten. Mit dem Befehl „Hauptansicht kopieren“ bzw. „Zusatzansicht kopieren“ kann man die Hauptansicht bzw. die Hilfsansicht in die Windows-Zwischenablage kopieren und anschließend z.B. in Word oder das Ausgabefenster einfügen.

Ansicht. Mit  schaltet man den Bearbeitungsmodus für die Hauptansicht ein und mit  kommt man in den Explorations-(Anzeige-)modus zurück. Standardmäßig ist die Modellanzeige im Explorationsmodus. Mit „Paletten“ öffnet sich eine Palette mit wählbaren Elementen in der Hauptansicht (⇒ Abb. 24.10). Nur die Elemente „Allgemein“ und „Viewer“ sind im Sondierungsmodus aktiv geschaltet und in diesem wählbar. Alle anderen sind nur im Bearbeitungsmodus aktiv geschaltet und dann wählbar. Mit diesen kann man gezielt Elemente der Grafik für eine Layoutbearbeitung auswählen (markieren).

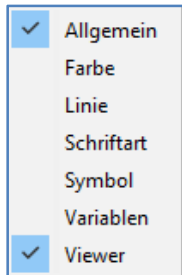








Abb. 24.10. Die Palette mit ihren Elementen

Hilfe. Es öffnet sich das Hilfesystem zum Umgang mit der Modellanzeige.

Ist in der „Palette“ im Menü „Ansicht“ der Explorationsmodus aktiv (entspricht der Standardeinstellung), so erscheint in der Modellanzeige die folgende Symbolleiste:



-  Einschalten des Bearbeitungsmodus.
-  Einschalten des Durchsuchungs-(Anzeige-)Modus (sonst Explorationsmodus genannt).
-  Visualisierung kopieren. (Kopieren der Grafik in die Windows-Zwischenablage).
-  Visualisierungsdaten kopieren. (Kopieren der Grafikdaten in die Windows-Zwischenablage).
-  Drucken der Grafik. Ist auch im Ausgabefenster möglich.
-  Visualisierungsbaum anzeigen. (Öffnen eines Visualisierungsbaumes mit Elementen der Grafik zum Überarbeiten der Grafik).

Schaltet man in den Bearbeitungsmodus und wählt ein Grafikelement (wie z.B. „Farbe“ oder „Schriftart“, so öffnen sich spezifische Symbolleisten zur Layoutgestaltung.¹⁹

¹⁹ Sie entspricht der im Grafiktabel-Editor (⇒ Kap. 34.4).

Mit ● bzw. ▲ werden die Fälle der Trainings- bzw. Testdatenfälle (Holdout genannt) im dreidimensionalen Variablenraum dargestellt. Fälle mit KRESIKO = 0 bzw. = 1 werden durch eine unterschiedliche Farbe sichtbar gemacht.

Fokusfälle werden rot umrandet herausgehoben. Temporär kann man auch jeden Fall als Fokusfall auswählen (markieren), indem man auf einen Fall mit der Maus klickt (für mehrere Fälle Strg-Taste verwenden). Nachbarn von Fokusfällen werden durch Verbindungslinien angezeigt. Zeigt man mit dem Cursor auf eine Verbindungslinie, so wird die Distanz angezeigt. Mit k : kann man steuern, wie viele Nachbarn von Fokusfällen angezeigt werden sollen.

Fährt man mit dem Cursor über die Fälle, so wird die Fallzahl („Id“) ausgewiesen und angezeigt, ob KRISIKO 1 oder 0 ist.²⁰ Auf diese Weise kann man sich ein Bild über den Zusammenhang zwischen KRISIKO und den Variablen machen.

Zeigt man mit dem Cursor auf die Grafik und zieht mit der linken Maustaste, so kann man die Grafik drehen und auf diese Weise verschiedene Blickwinkel auf die Grafik bekommen.

Ist in der Palette (⇒ Abb. 24.10) „Viewer“ eingeschaltet (entspricht der Standardeinstellung), so haben das Haupt- und das Hilfsfenster eine untere Leiste mit einer Dropdownliste für „Ansicht:“ zum Auswählen aus verfügbaren Ansichten.

In der Hauptansicht lässt sich die Dropdownliste jedoch nicht öffnen, weil das kNN-Verfahren nur die angezeigte Ansicht „Prädiktorraum“ (Variablenraum) hat.

am unteren rechten Rand der Hauptansicht ermöglicht es, den Anfangszustand der Grafik nach einer Bearbeitung wieder herzustellen.

Die Hilfsansicht. Im rechten Teil der Modellansicht (Hilfsansicht genannt) lassen sich detaillierte Informationen zum Modell in Form von Grafiken oder Tabellen anzeigen.

In der Hilfsansicht erscheint die gleiche Symbolleiste mit den gleichen Symbolen wie in der Hauptansicht, wenn im Menü „Ansicht“ in „Paletten“ „Allgemein“ aktiv geschaltet ist. Da die Symbole die gleiche Aufgabe haben (nur bezogen auf die Hilfsansicht), verweisen wir auf die Ausführungen oben.

Im Unterschied zur Hauptansicht, lassen sich in „Ansicht:“ auf der unteren Leiste des Hilfsfensters (sichtbar nur wenn „Viewer“ eingeschaltet ist ⇒ Abb. 24.10) per Dropdownliste eine Reihe von Ansichten für das Modell wählen.

Je nach Spezifizierung der Optionen für das kNN-Verfahren (automatische k-Auswahl, automatische Auswahl der Variablen) sind einige der Ansichten nicht immer verfügbar. Wir gehen auch kurz auf die für unsere Anwendungsspezifizierung nicht verfügbare Ansichten ein.

Ansicht „Prädiktoreinfluss“. In Abb. 24.9 sehen wir diese Modellansicht der Wichtigkeit der Variablen in der Hilfsansicht. Die Wichtigkeit der Variablen wird als Balkendiagramm dargestellt. Diese Modellanzeige gibt es nur, wenn man auf der Registerkarte „Nachbarn“ die Option „Funktionen bei Berechnung von Abständen nach Wichtigkeit gewichten“ nutzt.

²⁰ Für die Trainingsdaten werden die tatsächlichen und für die Testdaten die vorhergesagten Werte angezeigt.

Die Variable LAUFKONTO1 wird als wichtigste und BESCHZEI1 als unwichtigste ausgewiesen. Die relative Wichtigkeit der Variablen wird gemäß den Ausführungen zu den Gleichungen 24.9 und 24.10 berechnet. Zeigt man mit dem Mauszeiger auf einen der Balken, so wird die Höhe des relativen Gewichts g_h einer Variablen angezeigt. Bei unserer Modellentwicklung haben wir diese Informationen genutzt um ein „sparsames“ Modell zu finden.

Unter dem Balkendiagramm befindet sich eine Leiste mit einem (zunächst unsichtbaren) Schieberegler. Doppelklickt man mit der Maus auf den Namen der am wenigsten wichtigen Variablen auf der Leiste (hier BESCHZEIT1) und zieht bei gedrückter Maustaste, kann man die Zahl der angezeigten Variablen verkleinern (bzw. vergrößern)

Ansicht „Peers“. In dieser Ansicht des Klassifikationsmodells werden Fokusfälle und ihre Nachbarn in Punktsäulendiagrammen (\Rightarrow Kap. 32.10.4) dargestellt. Auf der senkrechten Achse (y-Achse) sind die Zielvariable KRISIKO bzw. die zur Distanzberechnung verwendeten Variable abgebildet. Fokusfälle (rot) und ihre k Nachbarn (blau) werden als Punkte (mit ihren Fallnummern) abgebildet.

Hat man Fokusfälle auf der Registerkarte „Variablen“ definiert, so werden standardmäßig die Punktdiagramme für die Zielvariable sowie für die fünf wichtigsten Variablen für die Klassifikation dargestellt. Abb. 24.13 zeigt das Peer-Diagramm für den Fokusfall 300.

Hat man keine Fokusfälle auf der Registerkarte „Variablen“ definiert, so erscheint in der Modellansicht „Peers“ die Meldung „keine Fokusdatensätze im Prädiktorraum ausgewählt“. Durch Klicken auf einen Fall in der 3D-Grafik wird dieser temporär zum Fokusfall und für diesen wird dann dynamisch verbunden das entsprechende Peer-Diagramm gezeigt.

Klicken auf den Schalter Prädiktoren auswählen... öffnet eine Dialogbox (\Rightarrow Abb. 24.12) zum Wählen von bisher nicht im Peer-Diagramm dargestellten Prädiktoren (Variablen). „Prädiktor 1“ bis „Prädiktor 5“ zeigen die momentanen fünf dargestellten Variablen in den Punktdiagrammen. Für jede Variable lässt sich eine Dropdownliste öffnen um eine dargestellte Variable durch eine bisher nicht dargestellte auszutauschen.



Abb. 24.12. Dialogbox zur Auswahl von Variablen im Peer-Diagramm

Mit dem Dropdown- Schalter von k : in der Hauptansicht lässt sich die Anzahl der angezeigten Nachbarfälle zum Fokusfall in den Punktsäulendiagrammen verkleinern.

Auch wenn man durch Klicken auf einen Fall in der 3D-Grafik einen Fall temporär zum Fokusfall erhebt, werden im Peerdiagramm die Nachbarn angezeigt. Wählt man im Diagramm in der Hauptansicht in k : ein kleines k , überträgt sich dieses sofort in die Peer-Diagramme.

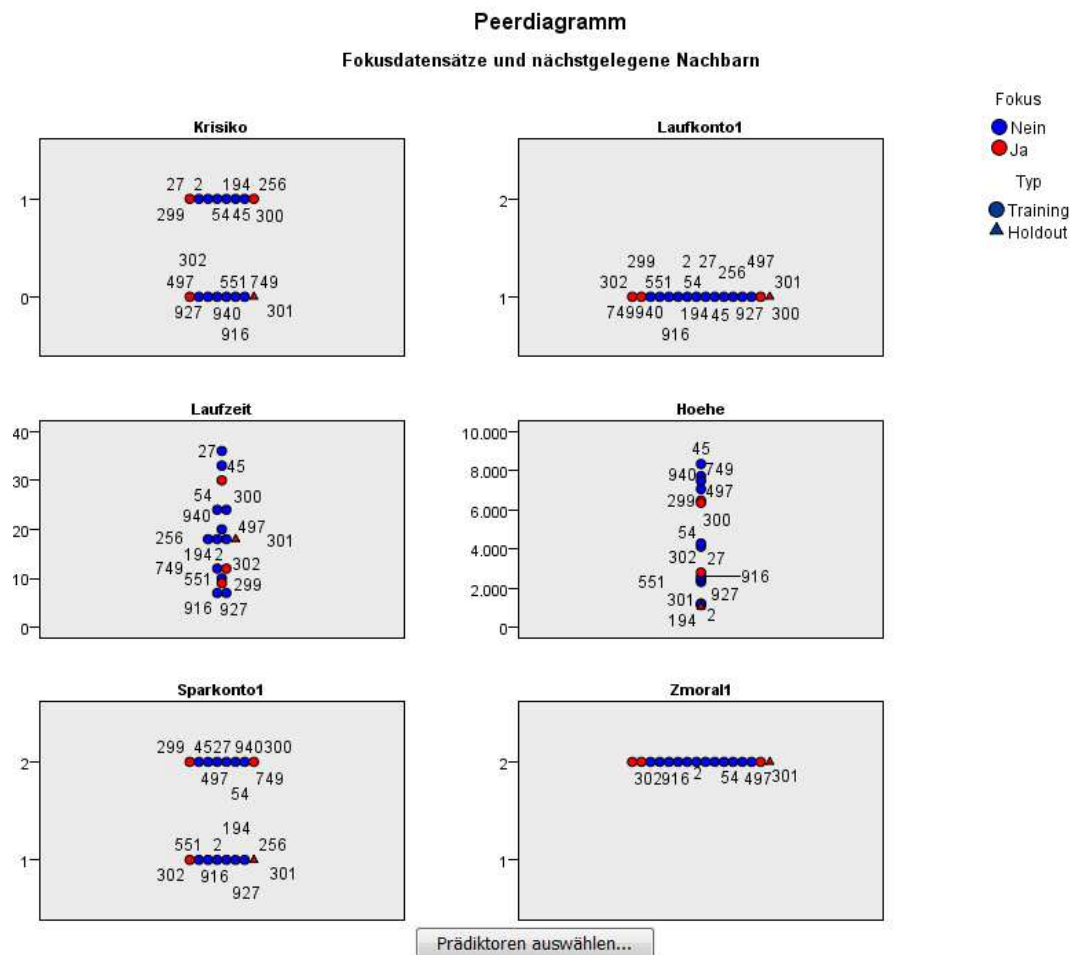


Abb. 24.13. Die Peers der Fokusfälle für die fünf wichtigsten Variablen für $k = 3$

Ansicht „Nachbar- und Abstandstabelle“. Diese Modellansicht gibt es in zwei Varianten. Erstens, wenn man auf der Registerkarte „Variablen“ (\Rightarrow Abb. 24.2) Fokusfälle definiert hat. Für unsere 4 Fokusfälle mit den Fallnummern 299 bis 302 werden die Fallnummern der $k = 5$ Nachbarn sowie deren Distanzen in einer Tabelle dargestellt (\Rightarrow Abb. 24.14). Diese Daten kann man sich auch in Form einer SPSS-Datendatei ausgeben lassen (\Rightarrow Registerkarte „Ausgabe“). Hat man keine Fokusfälle definiert, so erscheint in der Modellansicht die Meldung „Keine Fokusdatensätze im Prädiktorraum ausgewählt“.

Zweitens, wenn man in der 3D-Grafik im linken Fenster mit der Maus einen Fall anklickt und er so zu einem temporären Fokusfall wird. Dann werden rote Verbin-

dungslinien zu den nächsten k Nachbarn sichtbar und analog zu Abb. 24.14 werden Fallnummern und Abstände zum gewählten Fokusfall in einer dynamisch verbundenen Tabelle aufgeführt.

k Nächstgelegene Nachbarn und Abstände
Angezeigt für Anfangsfokusdatensätze

Fokusdatensatz	Nächstgelegene Nachbarn					Kürzeste Abstände				
	1	2	3	4	5	1	2	3	4	5
300	27	940	45	959	28	0,115	0,127	0,127	0,144	0,152
302	916	551	927	307	395	0,027	0,028	0,032	0,034	0,035
299	749	497	54	577	310	0,053	0,160	0,191	0,197	0,204
301	194	256	2	821	291	0,004	0,007	0,009	0,013	0,016

Abb. 24.14. Die Nachbar- und Distanztabelle der Fokusfälle

Ansicht „K-Auswahl“. In diesem Diagramm (\Rightarrow Abb. 24.15 für unsere Einstellungen auf den Registerkarten) wird die Fehlerquote („Fehlerrate“) gemäß Gleichung 24.2 auf der y-Achse und k auf der x-Achse abgebildet. Bei $k = 5$ ist die Fehlerquote am kleinsten. Das Diagramm gibt es nur, wenn man die „beste“ Höhe von k durch das kNN-Verfahren bestimmen lässt (bei Verzicht auf automatischer Variablenauswahl).

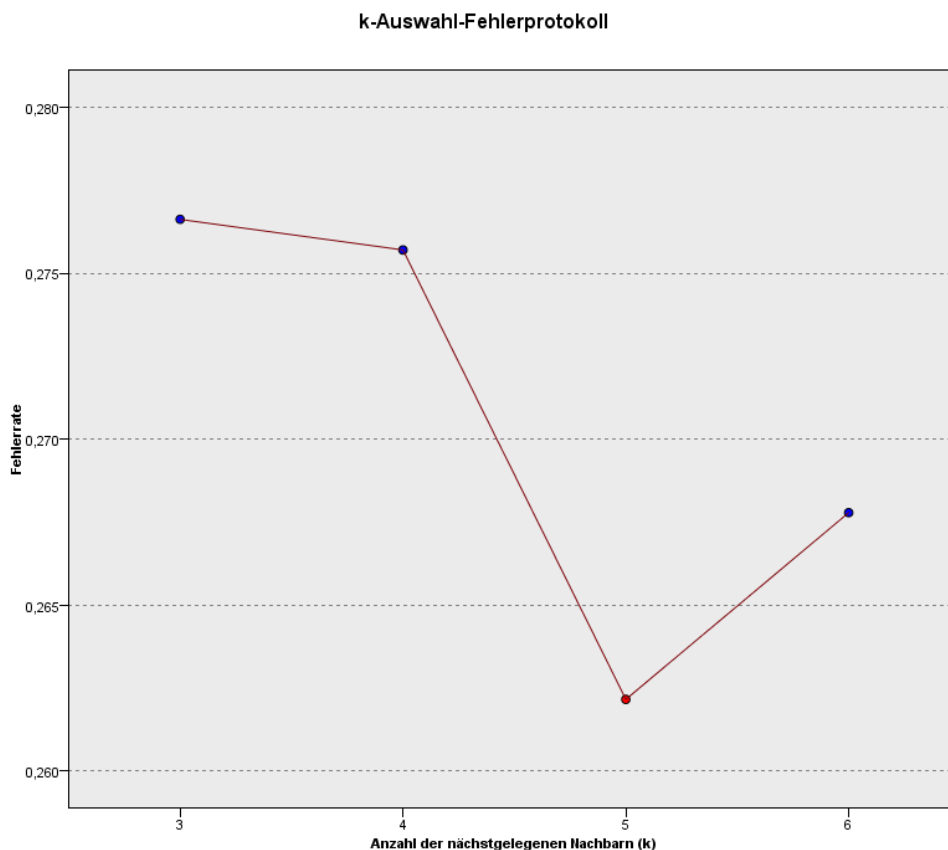


Abb. 24.15. Fehlerprotokoll zur k -Auswahl

Ansicht „Quadrantenkarte“. Quadrantenkarten sind Streudiagramme der Fokusfälle und ihren k Nachbarn mit der Zielvariable KRISIKO auf der y-Achse und den metrischen Variablen (hier HOEHE bzw. LAUFZEIT) auf der x-Achse (\Rightarrow Abb. 24.16). Analog wie bei der Peeransicht kann man sich auch hier in dynamischer Verknüpfung mit der Hauptansicht andere Fälle als temporäre Fokusfälle mit ihren Nachbarn im Streudiagramm anzeigen lassen.

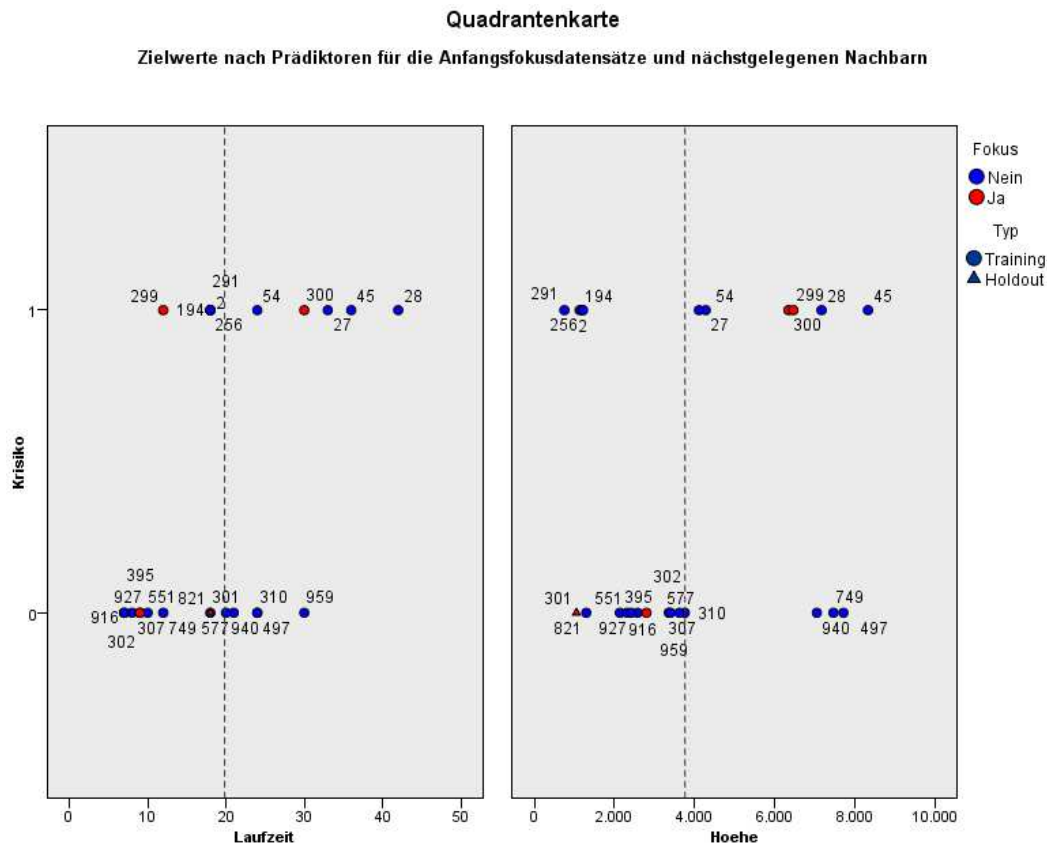


Abb. 24.16. Quadrantenkarte für $k = 5$

Ansicht „Prädiktorauswahl“. Diese Modellanzeige gibt es nur, wenn man auf der Registerkarte „Nachbarn“ einen festen Wert für k eingegeben hat und eine automatische Auswahl der Variablen anfordert. Auf der y-Achse wird mit „Fehlerrate“ die Fehlerquote gemäß Gleichung 24.2 und auf der x-Achse werden die Variablen dargestellt. Man kann ablesen in welcher Reihenfolge Variablen ausgewählt werden und wie sich die Fehlerquote dadurch mindert.

Ansicht „Prädiktor- und k -Auswahl“. Dieses Diagramm gibt es nur, wenn eine automatische Auswahl der Variablen (Registerkarte „Funktionsauswahl“) mit dem Bestimmen eines „besten“ k kombiniert wird. Man kann ablesen, in welcher Reihenfolge die Variablen für verschiedene k -Werte ausgewählt werden und wie sich dabei die mit „Fehlerrate“ angezeigte Fehlerquote verändert.

Ansicht „Klassifizierungstabelle“. In Abb. 24.17 ist die Klassifikationsmatrix als eine weitere Modellanzeige zu sehen. Diese wird sowohl für die Trainings- als

auch die Testdaten („Holdout“) aufgeführt. Als Maßstab für die Güte der Vorhersagequalität dienen die Ergebnisse für die Testdaten. Knapp 82 % der „guten“ und ca. 53 % der „schlechten“ Kredite werden richtig vorhergesagt. Diese Trefferquote lässt zu wünschen übrig. Bei der Bewertung dieser Ergebnisse sollte man aber berücksichtigen, dass alle 1000 Kredite schon einmal von Kreditsachbearbeitern der Bank begutachtet und akzeptiert worden sind.

In der Zeile „Fehlt“ wird ausgewiesen, dass in keinem Fall der Trainingsdaten ein fehlender Wert für die Zielvariable KRISIKO vorliegt.

Klassifikationstabelle

Partition	Beobachtet	Vorhergesagt		
		0	1	Prozent korrekt
Training	0	407	77	84,1%
	1	107	120	52,9%
	Prozent insgesamt	72,3%	27,7%	74,1%
Holdout	0	177	39	81,9%
	1	34	39	53,4%
	Fehlt	0	0	
	Prozent insgesamt	73,0%	27,0%	74,7%

Abb. 24.17. Klassifizierungstabelle

Ansicht „Fehlerzusammenfassung“. In dieser Modellansicht (\Rightarrow Abb. 24.18) wird die Fehlerquote gemäß Gleichung 24.1 differenziert nach Trainings- und Testdaten aufgeführt.

Fehlerzusammenfassung

Partition	Prozent falsch klassifizierte Datensätze
Training	25,9%
Holdout	25,3%

Abb. 24.18. Fehlerzusammenfassung

Klassifizieren neuer Daten. Wird bei der Analyse der Daten das Modell in einer XML-Datei gespeichert, so können mit der Befehlsfolge „Extras“; „Scoring-Assistent“, „neue“ Daten mit Hilfe dieser Datei klassifiziert werden. In Kap. 19.2 wird dies näher erläutert und demonstriert.