19 Automatische Lineare Modellierung

Hier wird in Ergänzung zum Buch in der 9. Auflage die praktische Anwendung der Modellbildungsverfahren Boosting und Bagging sowie „Beste Untergruppen" behandelt.

19.2 Praktische Anwendung

Vorbemerkungen. In Kap. 19.1 wird das Modellbildungsverfahren „Standard" mit der Variablenauswahlmethode „Schrittweise vorwärts" mit dem Prüfkriterium „AICC" zur Vorhersage von Preisen für gebrauchte Chevrolet-Modelle genutzt (Datendatei AUTOPREISE.SAV). Ergänzend soll hier die Anwendung der Verfahren „Boosting" und Bagging" sowie der Methode „Beste Untergruppen" zur Modellauswahl mit diesem Datensatz erläutert werden.

Eine Verwendung dieser Verfahren verspricht eventuell eine bessere Vorhersage der abhängigen Variable PREIS.

Zur Erläuterung der Verfahren „Boosting", Bagging" und „Beste Untergruppen" sei auf Kap. 19.1 im Buch verwiesen.

Um die Vorhersagewerte für die Variable PREIS bei Nutzung der verschiedenen Verfahren vergleichen zu können, wiederholen wir als erstes die im Buch vorgeführte Berechnung mit dem Verfahren „Standard" mit der Variablenauswahlmethode „Schrittweise vorwärts" und dem Prüfkriterium „AICC". Wir gehen wie im Buch vor und wählen (abweichend von der Vorgehensweise im Buch) auf der Registerkarte „Modelloptionen" (⇨ Abb. 19.5) die Option „Vorhergesagte Werte in Daten-Set speichern" und vergeben als Variablennamen für den vorhergesagten Preis VStandard. Anschließend nutzen wir die Verfahren Boosting und Bagging (zur Vergleichbarkeit mit der Variablenauswahlmethode „Schrittweise vorwärts" und dem Prüfkriterium „AICC") sowie das Modellauswahlverfahren „Beste Untergruppen" (ebenfalls mit dem Prüfkriterium „AICC"). Für die Vorhersagewerte dieser Verfahren vergeben wir die Namen VBoosting, VBagging und VUntergruppen.

Das Modellbildungsverfahren „Boosting". Das Verfahren kann durch Bilden von mittleren Werten aus den Vorhersagewerten mehrerer Modelle des Ensembles die Genauigkeit der Vorhersage verbessern [durch Verringern der Varianz und der Verzerrung (bias) der Vorhersagewerte].

Nach Öffnen der Datei AUTOPREISE.SAV gehen wir wie folgt vor.[1]

▷ Wir klicken die Befehlsfolge „Analysieren" „Regression", „Automatische Lineare Modellierung". Es öffnet sich die in Abb. 19.14 dargestellte Dialogbox

1 Es wird davon ausgegangen, dass Sie auf der Registerkarte „Felder" (⇨ Abb. 19.14) die Option „Benutzerdefinierte Feldzuweisungen" gewählt haben. Zur Option „Vordefinierte Rollen verwenden" verweisen wir auf Kap. 3.1 sowie auf „Neue Benutzeroberfläche und Ergebnisausgabe in der Modellanzeige" in Kap. 30.1

mit geöffneter Registerkarte “Felder“. In das Eingabefeld „Ziel“ übertragen wird die Variable PREIS und in das Eingabefeld „Prädiktoren (Eingaben)“ alle anderen (unabhängigen) Variablen.

- ▷ Auf der Registerkarte „*Erstellungsoptionen*“ wählen wir für das Element „Ziele“ die Option „Modellgenauigkeit verbessern (Boosting)“ (⇨ Abb. 19.15).
- ▷ Für das Element „*Basis*“ der gleichen Registerkarte ist die automatische Datenvorbereitung sowie ein Konfidenzintervall von 95 % voreingestellt (⇨ Abb. 19.16). Diese Voreinstellungen behalten wir bei.
- ▷ Für das Element „*Modellauswahl*“ der gleichen Registerkarte ist als Modellauswahlmethode „Schrittweise vorwärts“ und als Kriterium für „Aufnahme bzw. Ausschluss“ einer Variablen“ „Informationskriterium (AICC)“ voreingestellt (⇨ Abb. 19.17). Diese Voreinstellungen behalten wir bei.
- ▷ Für das Element „*Ensembles*“ der gleichen Registerkarte kann man nur scheinbar für die „Standard-Kombinationsregel für stetige Ziele“ wählen, ob der Vorhersagewert der abhängigen Variables des Ensembles als Mittelwert oder als Median der Vorhersagewerte der einzelnen Modelle des Ensembles berechnet werden soll (⇨ Abb. 19.18). Tatsächlich wird die Berechnung eines Mittelwerts ignoriert und der (gewichtete) Median berechnet. Die Anzahl der gerechneten Modelle ist auf „10“ voreingestellt. Wir belassen beide Voreinstellungen.
- ▷ Für das Element „*Erweitert*“ der gleichen Registerkarte ist voreingestellt, dass man die Ergebnisse mit dem gleichen Startwert des Zufallsgenerators replizieren kann. Wir belassen die Voreinstellungen (⇨ Abb. 19.19).
- ▷ Wir wählen auf der Registerkarte „*Modelloptionen*“ „Vorhergesagte Werte in Daten-Set speichern“ und vergeben als „Feldname“ (Variablennamen) VBoosting (⇨ Abb. 19.20).
- ▷ Mit Klicken von „Ausführen“ starten wir die Berechnung.

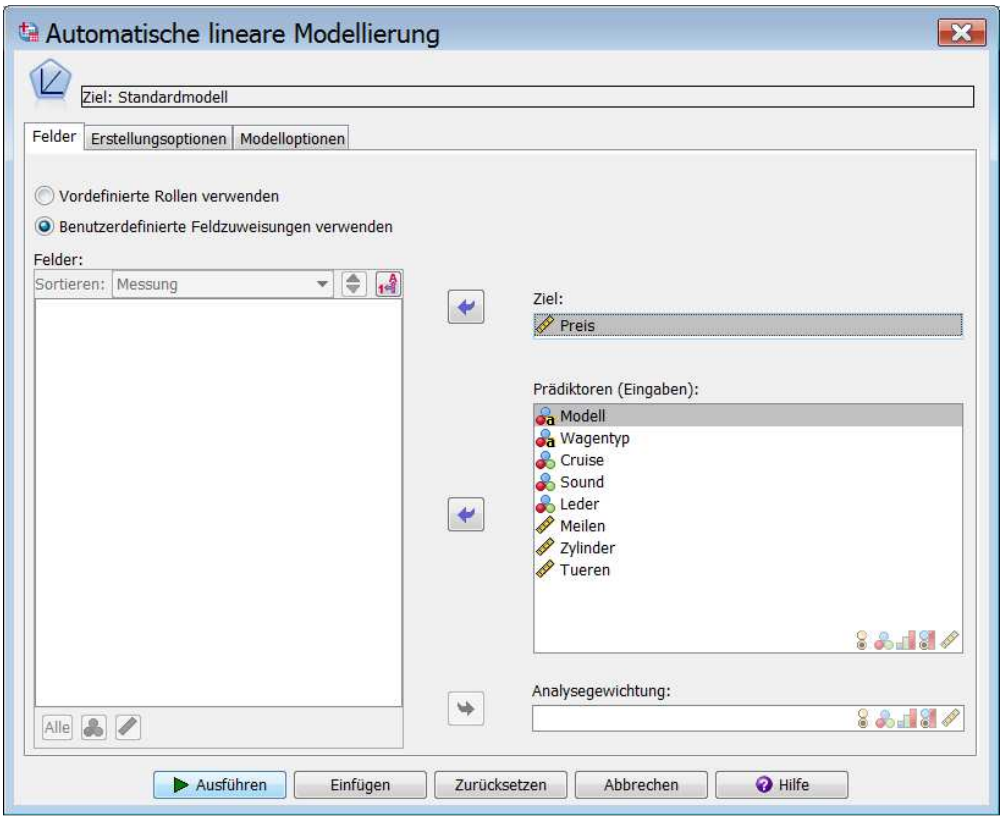


Abb. 19.14. Registerkarte „Felder“

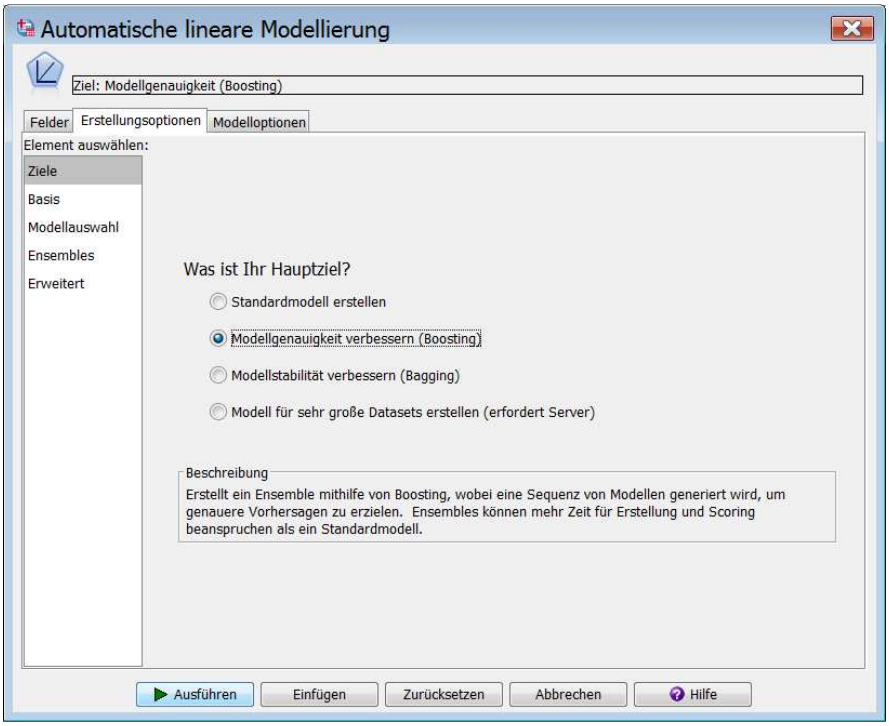


Abb. 19.15. Registerkarte „Erstellungsoptionen“ für das Element „Ziele“

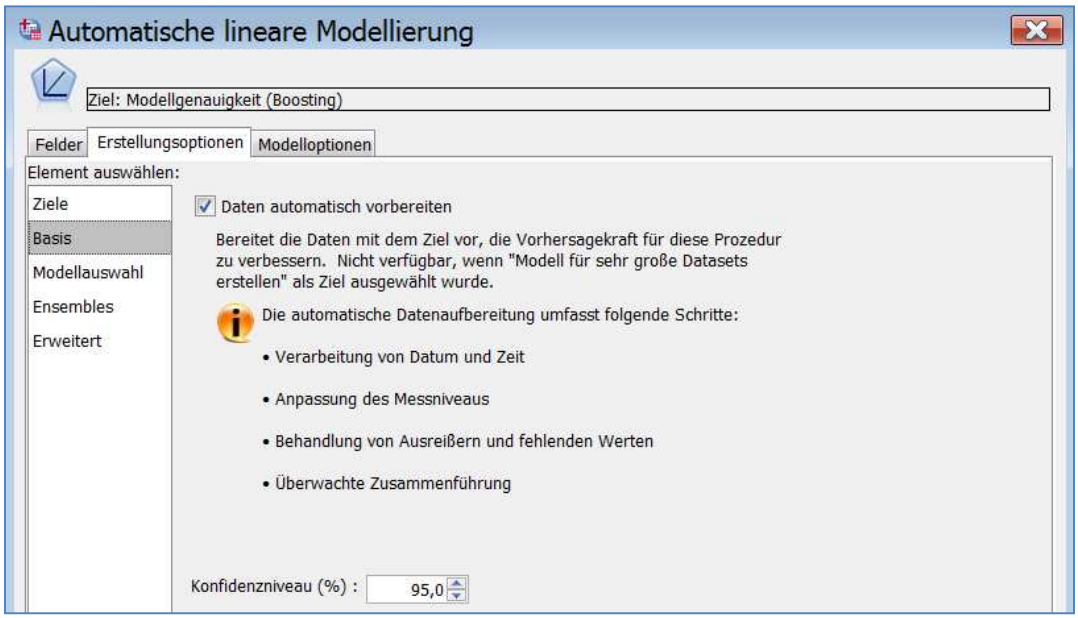


Abb. 19.16. Registerkarte „Erstellungsoptionen„ für das Element „Basis“

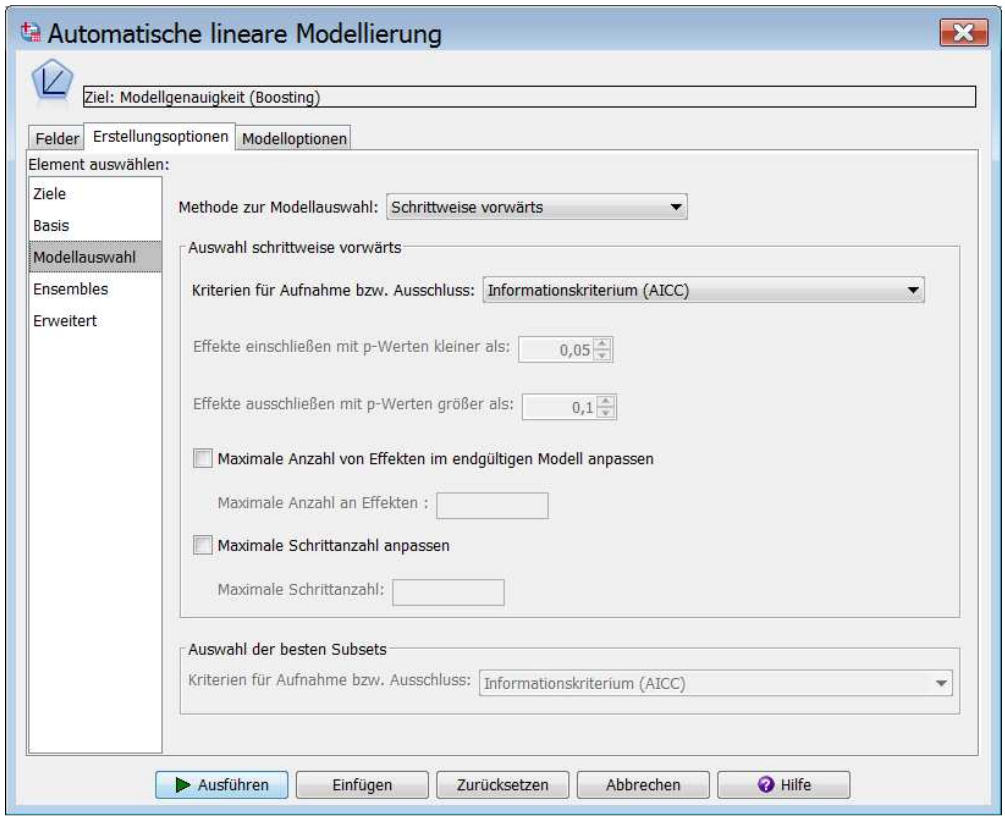


Abb. 19.17. Registerkarte „Erstellungsoptionen“ für das Element „Modellauswahl“

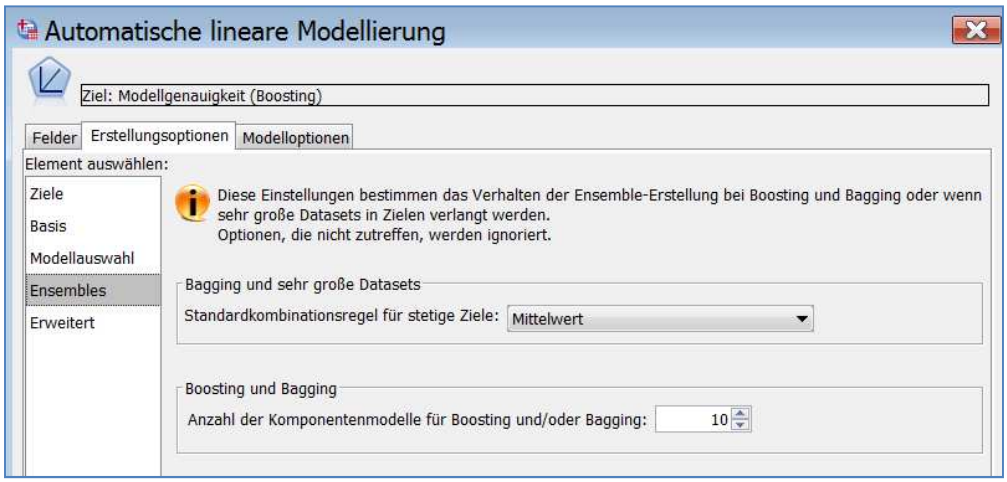


Abb. 19.18. Registerkarte „Einstellungsoptionen“ für das Element „Ensembles

Abb. 19.19. Registerkarte „Erstellungsoptionen“ für das Element „Erweitert“

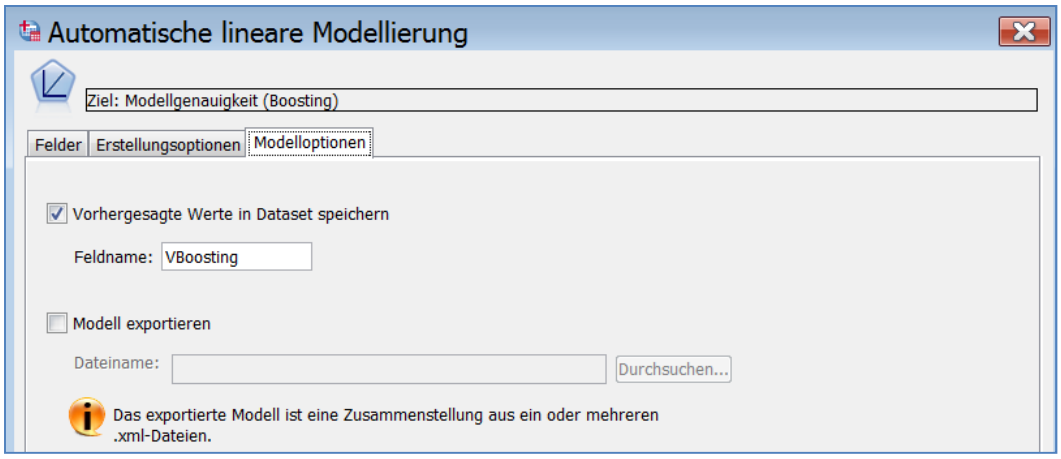


Abb. 19.20. Registerkarte „Modelloptionen“

Als erstes Ergebnis erhält man eine Übersicht über die Modelle des Ensembles (⇨ Abb. 19.21). In einer Warnung über der Tabelle wird erläutert, warum im Ensemble nicht alle Modelle enthalten sind ─ nur 8 Modelle, obwohl 10 (die Standardein-

stellung) gemäß der Registerkarte in Abb. 19.18 angefordert worden sind. Zu Beginn der Tabelle wird als Referenzmodell das Modell ohne Einsatz von Boosting aufgeführt.

Im Verfahren Boosting werden bei der sukzessiven Berechnung der Modelle des Ensembles die Fälle unterschiedlich nach der jeweiligen Höhe der Residualwerte ($= y - \hat{y}$) im zuvor berechneten Modells gewichtet. Die Berechnung des 1. Modells („Komponente 1“) beruht auf allen 320 Fällen der Datendatei. Für jedes nachfolgende Modell gehen die Fälle mit einem Gewicht in die Berechnung ein, das sich nach der Höhe des Residualwerts dieses Falles im zuvor berechneten Modell bemisst. Fälle mit hohen Residualwerten im zuvor berechneten Modell werden also stärker gewichtet als Fälle mit kleinen Residualwerten, so dass in diesem Prozess der Modellbildung „gelernt“ wird, hohe Residualwerte zu verringern.

Durch die Gewichtung bekommen die einzelnen Modelle des Ensembles unterschiedliche Fallzahlen. Die Fallanzahl in den Modellen 2 bis 8 schwankt zwischen 154 und 195.

Zur Bestimmung der Vorhersagewerte des Ensembles wird für jeden Fall der (gewichtete) Median aus den Vorhersagewerten der 8 Modelle gebildet.

Warnungen

During boosting some base models were not appropriate and have been removed from the ensemble

Zusammenfassung der Fallverarbeitung

Modell		N	Prozent
Referenz	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 1	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 2	Eingeschlossen	195	100,0%
	Ausgeschlossen	125	0,0%
	Gesamt	320	100,0%
Komponente 3	Eingeschlossen	193	100,0%
	Ausgeschlossen	127	0,0%
	Gesamt	320	100,0%
Komponente 4	Eingeschlossen	168	100,0%
	Ausgeschlossen	152	0,0%
	Gesamt	320	100,0%
Komponente 5	Eingeschlossen	154	100,0%
	Ausgeschlossen	166	0,0%
	Gesamt	320	100,0%
Komponente 6	Eingeschlossen	161	100,0%
	Ausgeschlossen	159	0,0%
	Gesamt	320	100,0%
Komponente 7	Eingeschlossen	155	100,0%
	Ausgeschlossen	165	0,0%
	Gesamt	320	100,0%
Komponente 8	Eingeschlossen	159	100,0%
	Ausgeschlossen	161	0,0%
	Gesamt	320	100,0%

Abb. 19.21. Die Fallzahlen des Basismodells und der Modelle des Ensembles.

Abb. 19.22 zeigt das Ergebnis als Modellobjekt. Verglichen wird die Güte der Vorhersage des Referenzmodells (berechnet mit allen 320 Fällen ohne Boosting) und des durch Boosting generierten Ensembles. Die Güte („Genauigkeit“) beider Modelle beträgt 96,4 %. Sie ist hoch: beide Modelle sind insofern gleichwertig.

Boosting hat hier also keine Verbesserung der Güte der Vorhersage gebracht. Beide Modelle sind für Vorhersagen gleich gut geeignet.

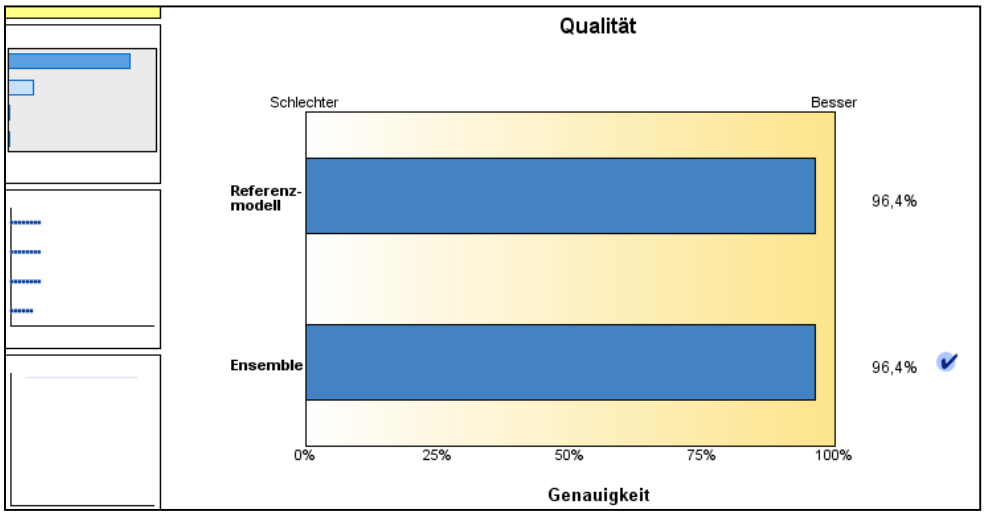


Abb. 19.22. Das Ergebnis als Modellobjekt („Boosting“)

Auf den ersten Blick ist verwirrend, dass die Güte des Referenzmodells (es entspricht dem „Standard“-Modell) mit 96,4 % und die des „Standard“-Modells (⇨ Abb. 19.6) mit 96,3 % ausgewiesen wird. Der Grund ist, dass die Güte des Referenzmodells wegen der Vergleichbarkeit mit der Güte des Ensemble mit R^2 (⇨ Gleichung 17.7) und die Güte des „Standard“-Modell mit R^2_{korr} (⇨ Gleichung 17.30) gemessen wird.

Unterhalb der Befehlsmenüleiste in der Modellanzeige befindet sich ein Dropdownschalter: ⇨ Neue Daten scoren mit: Referenz. Hier kann man wählen, ob eine Vorhersage mit dem Ensemble oder dem Referenzmodell vorgenommen werden soll. Ein Häkchen hinter den Balken zum Ausweis der Genauigkeit in der Modellzusammenfassung (⇨ Abb. 19.22) zeigt die aktuelle Einstellung.

In Abb. 19.23 wird die Wichtigkeit[2] der im Modell enthaltenen Prädiktoren vergleichend dargestellt. Zunächst ist anzumerken, dass ─ wie im „Standard“-Verfahren ─ die Variablen ZYLINDER, TUEREN, CRUISE und LEDER in keinem der Ensemble-Modelle einbezogen worden sind. Die Höhe der Wichtigkeit der Prädiktoren hat sich gegenüber dem „Standard“-Verfahren verändert. Dadurch hat sich auch die Reihenfolge in der Wichtigkeit gegenüber dem „Standard“-Verfahren verändert. MEILEN („Tachostand“) ist von Platz 2 im „Standard“-Verfahren auf Platz 3 gefallen. Nun ist WAGENTYP der zweitwichtigste Prädiktor.

2 Die Berechnung des Maßes Wichtigkeit wird im Buch erläutert.

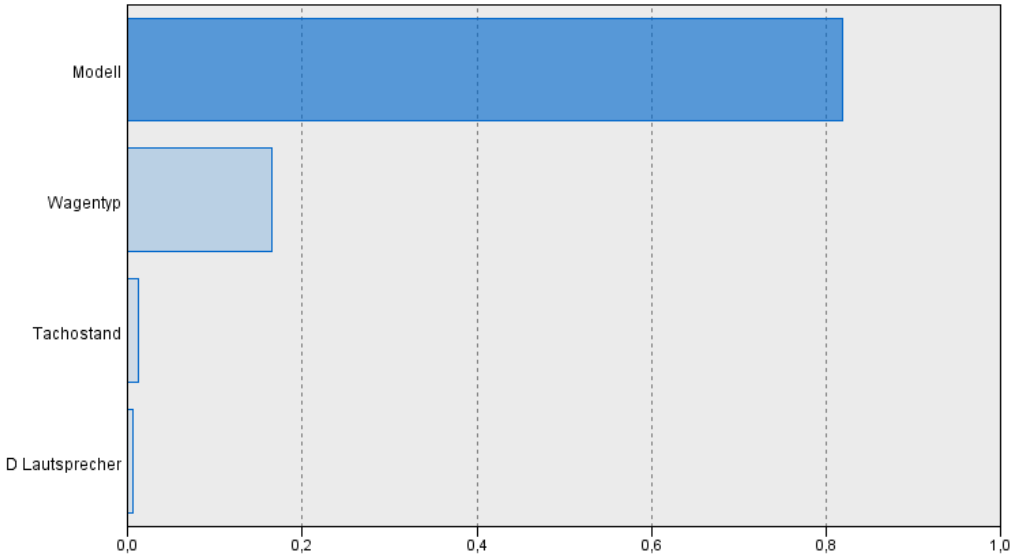


Abb. 19.23. Die Wichtigkeit der Prädiktoren in der Modellanzeige („Boosting“)

In Abb. 19.24 ist in einem Punktdiagramm (⇨ Kap. 32.10.4) für jeden Prädiktor dargestellt, in welchen der Ensemble-Modelle er enthalten ist. Geht man mit dem Mauszeiger auf einen der Punkte, so wird die Nummer des Modells angezeigt. In Abb. 19.24 zeigt der Mauszeiger auf den letzten Punkt von Prädiktor SOUND („D Lautsprecher“). Es erscheint die Nummer 8. Per Mauszeiger kann man auch sehen, dass dieser Prädiktor nicht in den Modellen 3 und 5 aufgenommen worden ist.

In einer weiteren (hier nicht aufgeführten) Grafik wird die „Kumulative Genauigkeit“ dargestellt, angezeigt durch eine Linie. Der Sinn dieser Grafik erschließt sich nicht so recht, weil die Genauigkeit schon in der Modellzusammenfassung (⇨ Abb. 19.22) zu sehen ist. Geht man mit der Maus auf die Linie, so wird die Genauigkeit angegeben (aber abweichend von der Modellzusammenfassung mit drei Stellen hinter dem Komma). Nicht klar ist, ob hier eventuell eine anders gemessene Genauigkeit gemeint ist und warum diese für jede Komponente gleich ist.

In Abb. 19.25 werden weitere Informationen zu den Ensemble-Modellen gegeben. Die Genauigkeit der Modelle schwankt um den Wert 95 %. Die Anzahl der Prädiktoren beträgt 3 oder 4 und die Anzahl der Regressionskoeffizienten 9 oder 10.[3] Im „Standard“-Verfahren gibt es 4 Prädiktoren und 10 Koeffizienten (einschließlich der Konstanten, ⇨ Abb. 19.9 im Buch). Wie schon oben angesprochen, ist in den Modellen 3 und 5 der am wenig wichtigste Prädiktor SOUND („D Lautsprecher“) nicht einbezogen, so dass in diesen Modellen nur 9 Koeffizienten geschätzt werden.

3 Da hier kategoriale Prädiktoren verwendet werden gehen diese in Form von Dummy-Variablen in die Regressionsgleichung ein (⇨ Kap. 18.3).

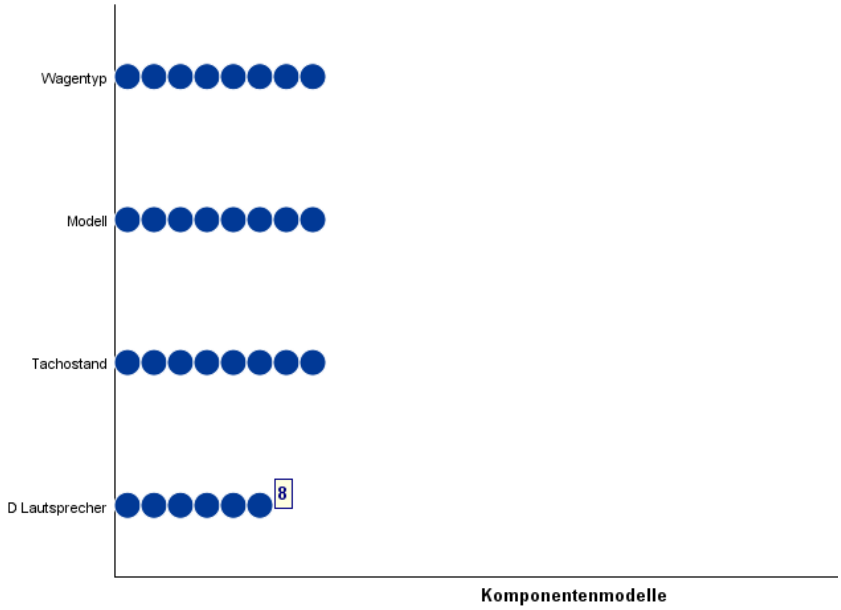


Abb. 19.24. Die Prädiktorhäufigkeit in der Modellanzeige („Boosting“)

Komponentenmodelldetails

Modell	Genauigkeit	Methode	Prädiktoren	Modellgröße (Koeffizienten)	Datensätze
1	96,4%		4	10	320
2	95,7%		4	10	320
3	95,7%		3	9	320
4	95,6%		4	10	320
5	95,8%		3	9	320
6	95,0%		4	10	320
7	95,1%		4	10	320
8	94,8%		4	10	320

Abb. 19.25. Komponentenmodelldetails in der Modellanzeige („Boosting“)

Als weiteres Ergebnis erscheint in der Modellanzeige eine Übersicht über die Transformationen bei der automatischen Datenvorbereitung. Diese entsprechen der im „Standard“-Verfahren (⇨ Abb. 19.7 links im Buch). Schließlich wird mit dem Ergebnis eine detaillierte Zusammenfassung der Modellschätzung geboten.

Das Modellbildungsverfahren Bagging. Bagging soll die Stabilität des Modells verbessern. Daher eignet sich das Verfahren, wenn der bestehende lineare Zusammenhang in den Daten der Grundgesamtheit sich in unterschiedlichen Stichprobendaten unterschiedlich (nicht stabil) darstellt. Das Verfahren kann durch

die Bildung von mittleren Werten aus den Vorhersagewerten der Modelle des Ensembles die Varianz der Vorhersagewerte verringern.

Man geht wie bei dem Verfahren „Boosting“ vor, mit folgenden Unterschieden:

- Auf der Registerkarte „Erstellungsoptionen“ mit dem Element „Ziele“ (⇨ Abb. 19.15) wählen wir die Option „Modellstabilität verbessern (Bagging)“
- Auf der gleichen Registerkarte mit dem Element „Ensembles“ wählen wir als „Standardkombinationsregel für stetige Ziele“ die Option „Median“. Man kann hier wählen, ob für die Vorhersagewerte der abhängigen Variable des Ensembles der Median oder der Mittelwert aus den Vorhersagewerten der einzelnen Modelle des Ensembles berechnet werden soll. Da wir die Vorhersageergebnisse der verschiedenen Modelle vergleichen wollen und im Modellbildungsverfahren „Boosting“ unabhängig von der dort gewählten Option immer der Median berechnet wird, wählen wir „Median“.
- Auf der Registerkarte „Modelloptionen (⇨ Abb. 19.20) markieren wir „Vorhergesagte Werte im Dataset speichern“ und geben als „Feldname“ VBagging ein.

Nach Klicken auf „Ausführen“ erscheint als erstes Ergebnis wie beim Verfahren „Boosting“ eine tabellarische Übersicht (⇨ Abb. 19.26). Neben dem Referenzmodell sind 10 Ensemble-Modelle aufgeführt. Jedes Modell beruht auf aus den 320 Datenfällen ausgewählte Stichproben. Die Stichproben nach dem Urnenmodell „Ziehen mit Zurücklegen“ haben jeweils einen Stichprobenumfang von 320.[4]

4 Dieses Urnenmodell mit Zurücklegen hat zur Folge, dass es sein kann, dass einige Fälle mehrfach und andere gar nicht in eine Stichprobe kommen.

Zusammenfassung der Fallverarbeitung

Modell		N	Prozent
Referenz	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 1	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 2	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 3	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 4	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 5	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 6	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 7	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 8	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 9	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%
Komponente 10	Eingeschlossen	320	100,0%
	Ausgeschlossen	0	0,0%
	Gesamt	320	100,0%

Abb. 19.26. Die Fallzahlen des Basismodells und der Modelle des Ensembles.

In Abb. 19.27 ist das Ergebnis in Form eines Modellobjekt zu sehen. Wie beim „Boosting“ wird die Güte („Genauigkeit“) des Referenzmodells mit der des Ensembles verglichen. Die Gütehöhen unterscheiden sich nicht von denen des Boosting (⇨ Abb. 19.22). Die angeführte Güte des Ensembles in Höhe von 96,4 % übersteigt die der einzelnen Ensemble-Modelle (⇨ Abb. 19.31).[5]

In Abb. 19.28 wird die Wichtigkeit der Prädiktoren vergleichend dargestellt. Im Vergleich zum „Standard“- sowie dem Boosting“-Verfahren kommt hier der Prädiktor CRUISE hinzu. Die Rangfolge der Wichtigkeit der in Modellen einbezogenen Prädiktoren MODELL, WAGENTYP, MEILEN und SOUND hat sich gegenüber dem Boosting-Verfahren nicht verändert (⇨ Abb. 19.23), wohl aber die Höhe der Gewichte. Erstaunlich ist, dass CRUISE („Cruiseschaltung“) in der Wichtigkeit noch vor der von MEILEN („Tachostand“) steht.

5 Dies liegt wohl an der unterschiedlichen Messung der Güte (R^2 bzw. R^2_{korr}).

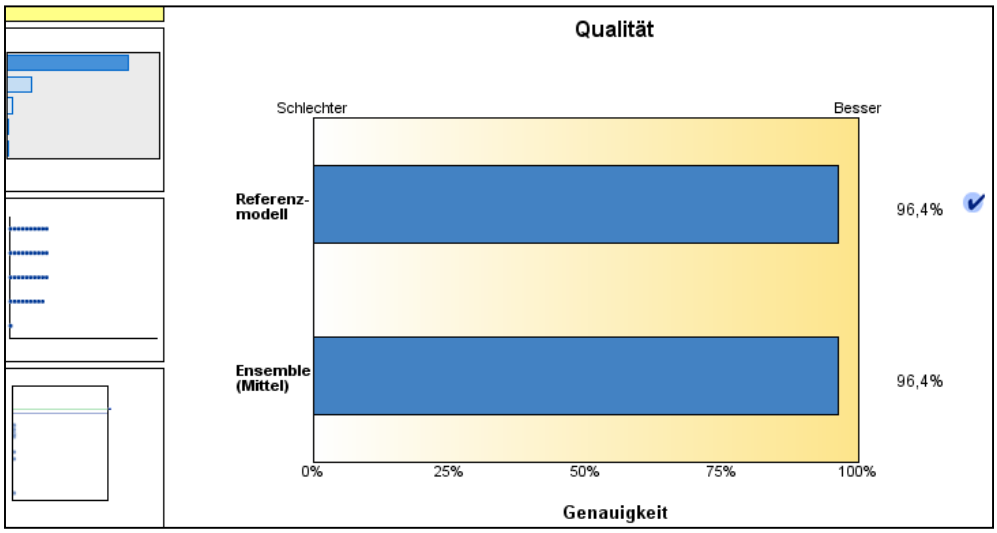


Abb. 19.27. Das Ergebnis als Modellobjekt („Bagging“)

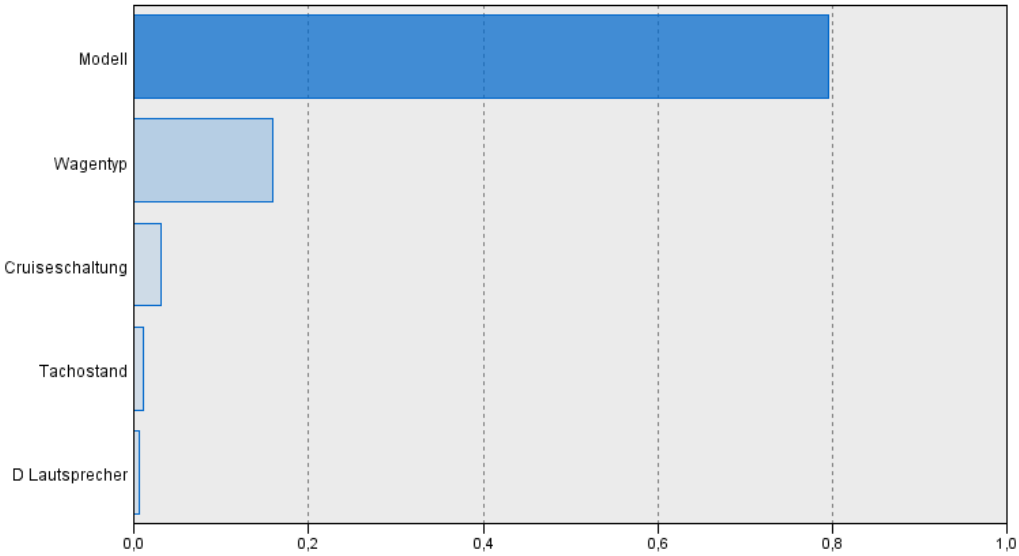


Abb. 19.28. Die Wichtigkeit der Prädiktoren in der Modellanzeige („Bagging“)

Geht man in der Übersichtsgrafik zur Prädiktorhäufigkeit in der Modellanzeige (⇨ Abb. 19.29) mit der Maus auf die Punkte, so wird angezeigt, dass die Prädiktoren WAGENTYP, MODELL und MEILEN (“Tachostand“) in allen 10 Ensemble-Modellen enthalten sind. Der Prädiktor SOUND („D Lautsprecher“) ist nicht im Modell 4 enthalten. Der neu hinzu gekommene Prädiktor CRUISE („Cruise Schaltung“) ist nur im Modell 3 enthalten (⇨ Abb. 19.29). In diesem Grafiktyp werden nur die 10 wichtigsten Prädiktoren angezeigt.

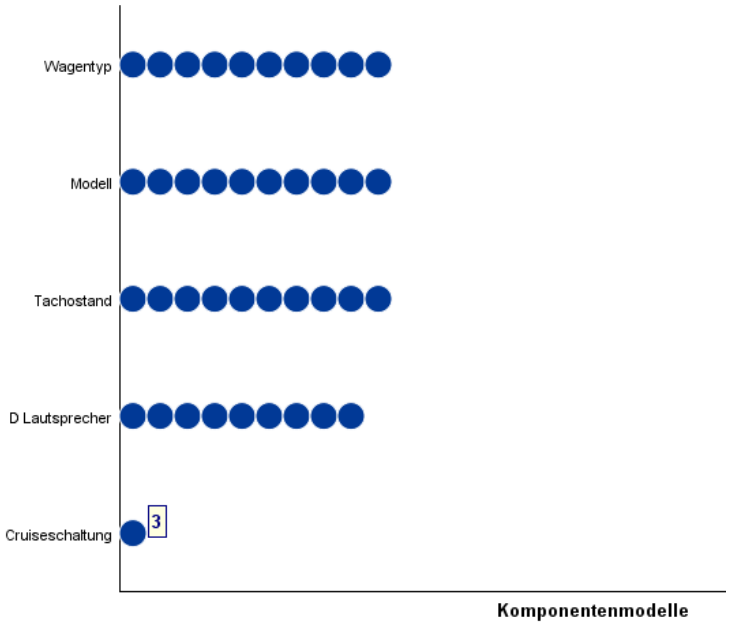


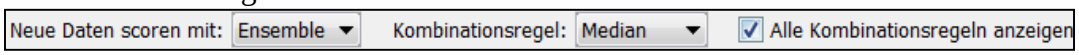
Abb. 19.29. Die Prädiktorhäufigkeit in der Modellanzeige („Bagging“)

In der Darstellung der „Komponentenmodellgenauigkeit“ in der Modellanzeige (⇨ Abb. 19.30) wird für jedes Ensemble-Modell die Genauigkeit (Güte) in Form eines Punktes angezeigt. Zeigt man mit der Maus auf einen Punkt, so wird die Modellnummer angezeigt. Die Güte einzelner Modelle kann man aber genauer aus der Darstellung der „Komponentenmodelldetails“ in der Modellanzeige ablesen (⇨ Abb. 19.31).

Die Genauigkeit des Referenzmodells sowie des Ensembles wird in Abb. 19.30 durch eine blaue und grüne horizontale Linie dargestellt. Geht man mit der Maus auf diese Linien, so wird die Höhe der Genauigkeit angezeigt. Obwohl die Linien nicht übereinander liegen, kann man per Mauszeiger sehen, dass die Genauigkeit gleich ist (wie auch in der Modellzusammenfassung in Abb. 19.27 zu sehen ist).

Auf der Leiste unterhalb der Menübefehle der Modellanzeige kann man im Dropdownschalter „Kombinationsregel:“ wählen, ob die in der Grafik angezeigte Modellgüte des Ensembles auf Basis der Berechnung des Mittelwerts oder des Medians der Vorhersagewerte der Ensemble-Modelle erfolgen soll. Wechselt man die „Kombinationsregel, so werden die von der Regel beeinflussten Informationen in den Modellansichten dynamisch angepasst.

Wählt man „Alle Kombinationsregeln anzeigen“, so werden beide Ensemble-Gütewerte in der Grafik angezeigt. Ein Häkchen hinter einer der Linien zeigt die aktuelle Einstellung.



Der Dropdownschalter „Neue Daten scoren mit):“ ist schon im Zusammenhang mit dem Boosting-Verfahren erläutert worden.

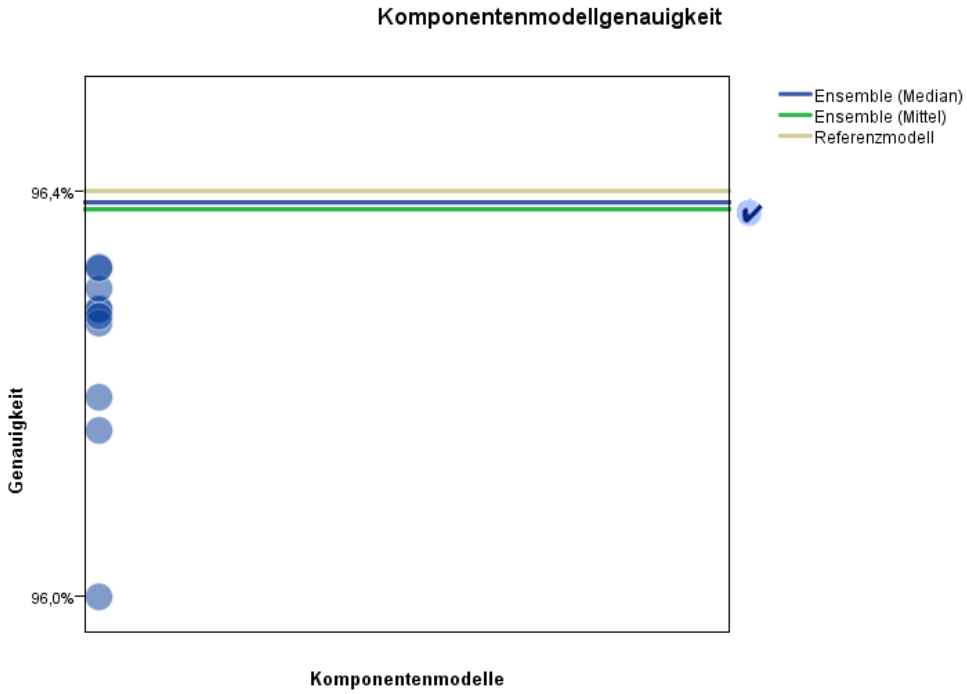


Abb. 19.30. Die Komponentenmodellgenauigkeit in der Modellanzeige („Bagging“)

In einer weiteren Modellanzeige werden “Komponentendetails“ aufgeführt (⇨ Abb. 19.31). Das Modell 3 z.B. enthält (wie schon oben angesprochen) 5 Prädiktoren und benötigt zur Modellschätzung 11 Koeffizienten.

Klickt man auf auf einen de Spaltenköpfe (z.B. Genauigkeit, so wird die Reihenfolge der Modelle in der Tabelle nach den Kriterien dieser Spalte neu sortiert. Geht man mit dem Mauszeiger auf das Symbol, dann wird mit ALM die verwendete SPSS-Prozedur (Automatic linear Modeling) angezeigt.

Das Variablenauswahlverfahren „Beste Untergruppen“. Das Verfahren „Beste Untergruppen“ ist eine Alternative zum Variablenauswahlverfahren „Schrittweise vorwärts“. Es bezieht weitere Untergruppen von potentiellen Prädiktoren in die Prüfung ein, ob sie für die gesuchte Modellgleichung in Frage kommen.[6] Bei bis zu 20 Regressionskoeffizienten (einschließlich derjenigen für Dummies) werden alle möglichen Untergruppen in die Prüfung einbezogen. Ist die Zahl der Regressionskoeffizienten größer, so wird ein Verfahren verwendet das „Schrittweise vorwärts“ und „Beste Untergruppen“ kombiniert, um den rechenintensiven Suchprozess in angemessener Zeit zu bewältigen.

Die Variablenauswahlmethode „Beste Untergruppen“ kann sowohl mit dem Modellbildungsverfahren „Standard“ als auch mit „Boosting“ sowie „Bagging“

6 Hat man z.B. als potentielle Prädiktoren die Variablen x_1, x_2 und x_3, so bilden die Gleichungen $\hat{y} = b_0 + b_1x_1$, $\hat{y} = b_0 + b_2x_2$, $\hat{y} = b_0 + b_3x_3$, $\hat{y} = b_0 + b_1x_1 + b_2x_2$, $\hat{y} = b_0 + b_1x_1 + b_3x_3$, $\hat{y} = b_0 + b_2x_2 + b_3x_3$ sowie $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$ alle möglichen Regressionsmodelle (Untergruppen, subsets). Bei m Prädiktoren gibt es $2^m - 1$ Modellgleichungen als Untergruppen. Ziel ist es, eine gute Vorhersage mit möglichst wenigen Prädiktoren zu erreichen (⇨ Kap. 19.1 im Buch).

kombiniert werden. Dabei können alternativ die Prüfkriterien AICC, R^2_{korr} und ASE für die Modellauswahl genutzt werden (⇨ Tabelle 19.1 im Buch).

Komponentenmodelldetails

Modell	Genauigkeit	Methode	Prädiktoren	Modellgröße (Koeffizienten)	Datensätze
1	96,3%		4	10	320
2	96,3%		4	10	320
3	96,2%		5	11	320
4	96,0%		3	9	320
5	96,3%		4	10	320
6	96,3%		4	10	320
7	96,3%		4	10	320
8	96,3%		4	10	320
9	96,3%		4	10	320
10	96,2%		4	10	320

Abb. 19.31. Komponentenmodelldetails in der Modellanzeige („Bagging“)

Zur Schätzung eines Modells geht man wie beim „Standardverfahren“ vor (⇨ Kap. 19.1 im Buch) mit dem Unterschied, dass wir auf der Registerkarte „Erstellungsoptionen“ für das Element „Modellauswahl“ nun „Beste Subsets“ wählen (⇨ Abb. 19.32). Für das Element „Ensembles“ wählen wir als „Standardkombinationsregel für stetige Ziele“ die Option „Median“ (⇨ Abb. 19.18). Auf der Registerkarte „Modelloptionen“ (⇨ Abb. 19.20) wird der „Feldname“ VUntergruppen eingegeben.

Das Ergebnis erscheint im Ausgabefenster als Modellobjekt (⇨ Abb. 19.33). Die Höhe des „Informationskriteriums“ (AICC) und auch die des Gütemaßes entsprechen denen der Modellauswahlmethode „Schrittweise vorwärts“ (⇨ Abb. 19.6). Auch alle anderen Ergebnisse sind gleich.

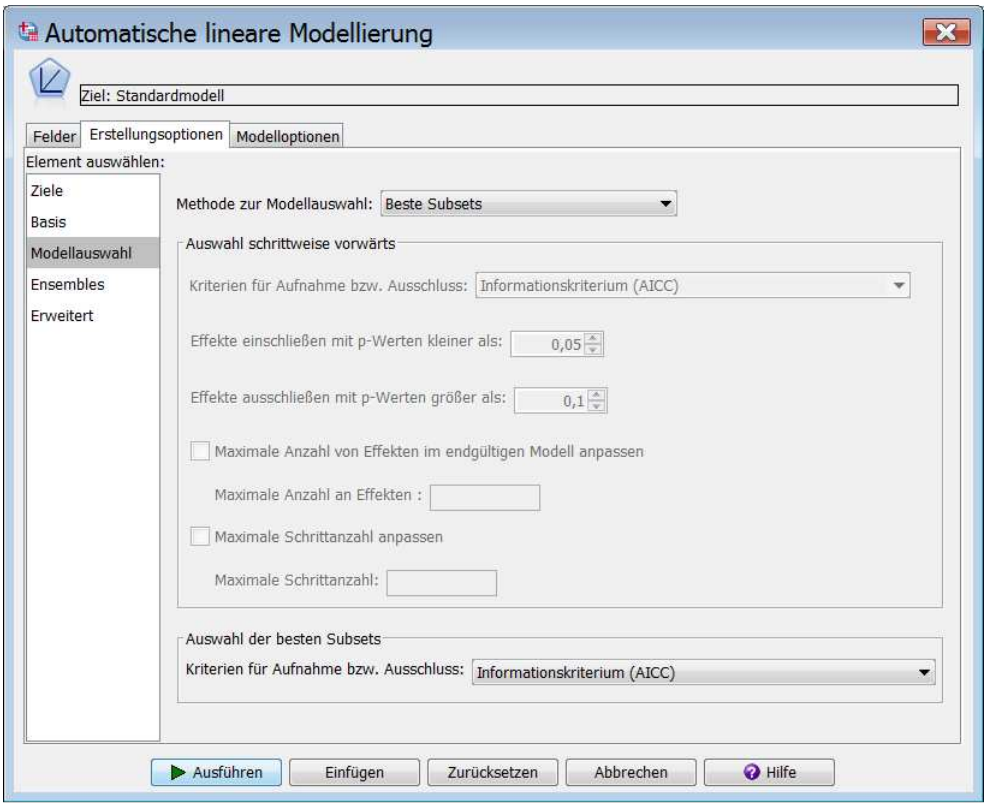


Abb. 19.32. Registerkarte „Erstellungsoptionen“ für das Element „Modellauswahl“

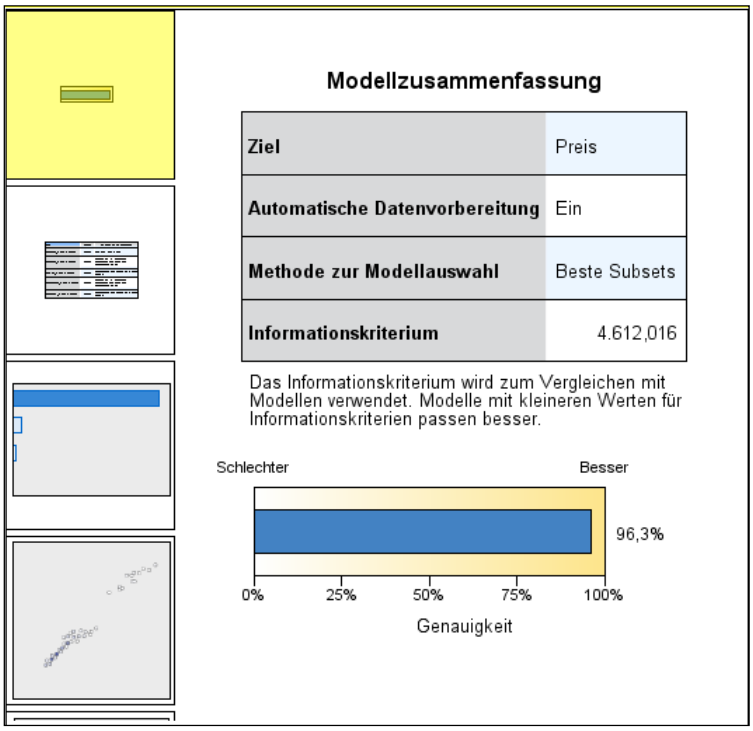


Abb. 19.33. Das Ergebnis als Modellobjekt („Beste Subsets“)

Die Vorhersagewerte im Dateneditor. In Abb. 19.34 sind die Vorhersagewerte für den Preis von Chevrolet-Modellen bei Anwendung der vier verschiedenen Berechnungsmodelle für die ersten 10 Fälle zu sehen. Die Vorhersagewerte der Verfahren „Standard“, „Bagging“ und „Beste Untergruppen“ unterscheiden sich nicht. Nur das Verfahren „Boosting“ führt zu unterschiedlichen Vorhersagewerten. Dies ist auch anhand der in Tabelle 19.2 angeführten statistischen Kennzahlen der Vorhersagewerte zu sehen.[7]

Bei Verzicht auf die automatische Datenvorbereitung kommt man bei allen vier Berechnungsmodellen zu gleichen Vorhersagewerten. Bei einigen anderen Datendateien sind wir zu ähnlichen Ergebnissen gekommen. Um das Vertrauen in die Prozedur zu verbessern, sollten weitere Berechnungen mit einer Vielzahl von Datendateien und den unterschiedlichen Optionen gemacht werden.

VStandard	VBoosting	VBagging	VUntergruppen
11657	11837	11657	11657
11383	11537	11383	11383
10388	10396	10388	10388
11034	10860	11034	11034
10305	10305	10305	10305
10057	10057	10057	10057
10722	10518	10722	10722
9750	9746	9750	9750
9313	9267	9313	9313
9081	9012	9081	9081

Abb. 19.34. Die Vorhersagewerte der ersten 10 Fälle im Dateneditor

Tabelle 19.2. Statistische Kennziffern der Vorhersagewerte

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung
VStandard	320	8345	43764	16427,60	6776,552
VBoosting	320	8205	44084	16478,54	6830,187
VBagging	320	8345	43764	16427,60	6776,552
VUntergruppen	320	8345	43764	16427,60	6776,552
Gültige Werte (Listenweise)	320				

Für unsere Berechnungsbeispiele haben wir das Kriterium AICC für die Auswahl der unabhängigen Variablen in die Gleichung gewählt. Wählt man ASE als Kriterium, so wird das Modellbildungsverfahren um eine moderne Technik der Datenanalyse ergänzt, die im Maschinellen Lernen und Data Mining zum Standard gehört. Bei diesem Verfahren wird der Datensatz per Zufallsauswahl im Verhältnis von ca. 2/3 zu 1/3 in eine Lern- und Teststichprobe aufgeteilt. Für die Modellberechnung werden die Daten der Lernstichprobe verwendet, ASE wird für die Testdaten berechnet. Diese Verfahrensweise soll Overfitting vermeiden.[8]

7 Erstellt im Menü „Deskriptive Statistiken“, „Deskriptive Statistik“

8 Für die Nutzung dieses Verfahrens sollte der verfügbare Datensatz aber nicht zu klein sein. Zu Overfitting ⇨ Kap. 19.1 im Buch. Im Analyseverfahren „Nächstgelegener Nachbar“ (⇨ Kap. 24.2) ist die Technik integriert.

Bei Verwendung von ASE als Kriterium für die Modellauswahl kommt man zu anderen Vorhersagewerten als bei Verwendung von AICC. Auch unterscheiden sich die Vorhersagewerte der verschiedenen Modellberechnungsmethoden.