h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Fachbereich Mathematik und Naturwissenschaften
Studiengang Business Mathematics (Master of Science)

Data Mining
Booklet

Vorgelegt von Büsra Karaoglan am 22. September 2017
Matrikelnummer : 754331

Referent Prof. Dr. Werner E. Helm

Inhaltsverzeichnis

1 Einleitung

Diese Ausarbeitung dient der Vorleistung des Fachbereichs Business Mathematics an der Hochschule Darmstadt für das Modul *Data Mining 1*, gelesen von Prof. Dr. Werner Helm. Sie beinhaltet zwei Teile: zum einen Aufgaben die nur mit SAS Prozeduren und zum anderen Aufgaben die mit Hilfe von SAS Enterprise Miner bearbeitet wurden. Die lauffähigen SAS und SAS Enterprise Miner-Dateien werden zusätzlich an Prof. Dr. Werner Helm ausgehändigt. Es werden in dieser Ausarbeitung Themenbereiche wie logistische Regression, Diskriminanzanalyse, Entscheidungsbäume und Support Vector Machines erarbeitet. Diese Themen werden auch im Folgenden kurz erläutert. Das theoretische beziehungsweise mathematische Hintergrundwissen der Themen sind Voraussetzungen für das Verständnis der Aufgaben. Allerdings werden an erforderlichen Stellen ergänzende Bemerkungen bezüglich der Thematik gemacht.

Bei vielen Fragestellungen, bei denen es um die Analyse des Einflusses einer oder mehrerer unabhängiger Variablen auf eine abhängige Variable geht, liegt als abhängige Variable keine stetige, sondern eine kategoriale oder qualitative Variable vor. Verfügt die abhängige Variable über zwei mögliche Ausprägungen (z.B. Therapieerfolg = Ja oder Nein), dann spricht man von einer binären logistischen Regression. Die Ereignisse werden dabei als 0/1-Ereignisse oder auch Komplementärereignisse bezeichnet. Die abhängige Variable Y kann also die beiden Ausprägungen 0 und 1 annehmen. Die unabhängigen Variablen, oftmals auch als Kovariate oder Regressoren bezeichnet, können sowohl metrisch als auch kategorial skaliert sein. Hier in dieser Arbeit wird die binäre logistische Regression genauer dargestellt [AL05].

Die Diskriminanzanalyse wird verwendet um die Unterschiedlichkeit von zwei oder mehreren Gruppen, hinsichtlich einer Mehrzahl von Variablen zu untersuchen. Das Ziel ist es, mit Hilfe der ermittelten Diskriminanzfunktionen die Gruppen optimal zu trennen. Dabei soll die Streuung zwischen den Gruppen möglichst groß und innerhalb der Gruppen möglichst klein sein. Die Gruppierungsvariable, also die abhängige Variable, muss nominal und die Prädiktorvariablen bzw. Regressoren, also die unabhängigen Variablen, sollen metrisch skaliert sein. Dabei kann die Diskriminanzanalyse in *analysierende Diskriminanznalyse:* „Welche Variablen sind zur Unterscheidung zwischen den Gruppen geeignet bzw. ungeeignet?“ und *klassifizierende Diskriminanzanalyse:* „Unterscheiden sich die Gruppen signifikant voneinander hinsichtlich der Variablen?“ unterteilt werden [KB00].

Die Entscheidungsbäume dienen der Aufteilung von Objekten. Dies geschieht anhand geeigneter Merkmale in Gruppen in Hinblick auf eine vorgegebene Zielgröße. Ein Entscheidungsbaum hat eine baumartige Struktur mit einer Wurzel, mehreren Blattknoten, inneren Knoten und Kanten. Jedem Blattknoten ist eine Klasse zugeordnet; pro Klasse sind mehrere Blattknoten möglich. Jedem inneren Knoten ist ein Merkmal zugeordnet; pro Merkmal sind mehrere innere Knoten möglich. Die Entscheidungsbäume lassen sich zusätzlich in zwei Varianten unterteilen: Klassifikationsbäume und Regressionsbäume. Klassifikationsbäume werden bei nominal skalierten Variablen als abhängige Zielgröße eingesetzt, während bei Regressionsbäumen eine quantitative Variable als abhängige Zielgröße vorliegt [UB08].

Eine Support Vector Machine (SVM) ist ein Verfahren bei dem ein gegebener Datensatz (welcher durch Vektoren in einem Vektor-Raum repräsentiert ist) durch eine Hyperebene in der Art zu teilen, dass den Vektoren auf derselben Seite die gleiche Klasse zugeordnet ist. Zudem wird die Größe des Randes der Hyperebene maximiert. Der Rand der Hyperebene ist durch die Vektoren gegeben, die den geringsten Abstand von dieser Ebene besitzen. Diese Vektoren werden als Support-Vektoren bezeichnet. Eine Klassifizierung mit einer Support Vector Machine ist ein überwachtes Lernverhalten, wobei die Support Vector-Klassifizierung durch das Lösen eine dualen Optimierungsproblems erfolgt [TB14].

2 Teil 1

2.1 Aufgabe 3

Aufgabenstellung:

Arbeiten Sie das Paper durch: *Tom Fawcett: An introduction to ROC analysis.* [TF06]

Schreiben Sie eine deutsche Kurzzussammenfassung. Bringen Sie die Inhalte in Zusammenhang und zum Einsatz bei den folgenden Analyseaufgaben.

Lösung:

Die ROC-Kurve (ROC = Receiver Operating Characteristics) ist eine Methode für die Visualisierung und die Prüfung der Leistung eines (binären) Klassifikators. ROC-Graphen wurden lange Zeit in der Signalentdeckungstheorie verwendet, um den Kompromiss zwischen Trefferquoten und falschen Alarmraten von Klassifikatoren darzustellen. Die ROC-Kurven werden nun häufig in der medizinischen Entscheidungsfindung verwendet und wurden in den letzten Jahren zunehmend in der maschinellen Learn- und Data-Mining-Forschung eingesetzt.

Im Folgenden werden Prüfungen von Klassifizierungsproblemen nur zweier Klassen vorgeführt. Jede Instanz I wird auf ein Element $\{p, n\}$ der positiven und negativen Klassenlabels abgebildet. Um zwischen der eigentlichen Klasse und der vorhergesagten Klasse zu unterscheiden, werden $\{Y, N\}$ für die von dem Modell erzeugten Klassenvorhersagen verwendet. Somit werden die tatsächlichen Beobachtungen in die Klassen p für *positiv* und n für *negativ* und die prognostizierten Beobachtungen in die Klassen Y für *Yes* und N für *No* zugeordnet. Somit ergibt sich vier mögliche Ergebnisse:

- True Positives-TP: Beobachtung wird als positiv klassifiziert, ist tatsächlich positiv
- True Negatives-TN: Beobachtung wird als negativ klassifiziert, ist tatsächlich negativ
- False Positives-FP: Beobachtung wird als positiv klassifiziert, ist tatsächlich negativ
- False Negatives-FN: Beobachtung wird als negativ klassifiziert, ist tatsächlich positiv

Die unten abgebildete Grafik zeigt eine 2×2 *Confusion Matrix* (auch Kontingenztafel genannt).

		True class p	True class n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		**P**	**N**

Es werden nun folgende Maßzahlen definiert:

$$\text{tp rate} \approx \frac{\text{Postives correctly classified}}{\text{Total positives}} = \frac{TP}{P}$$

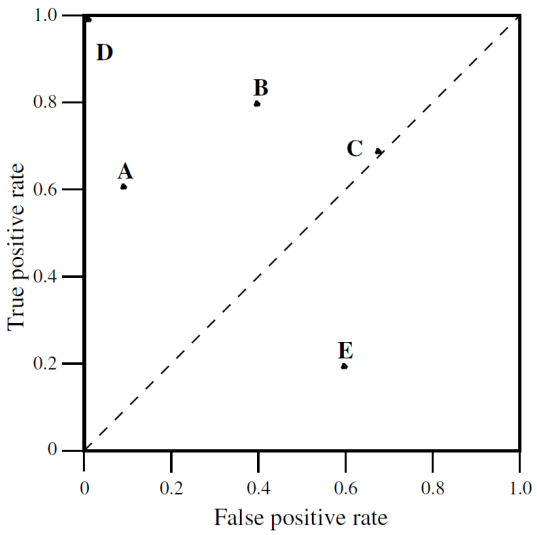
$$\text{fp rate} \approx \frac{\text{Negatives incorrectly classified}}{\text{Total negatives}} = \frac{FP}{N}$$

Zusätzlich werden im Zusammenhang mit der ROC-Kurven-Analyse die Begriffe *Sensitivität* und *Spezifität* eingeführt. Dabei gilt

$$\text{Sensitivität} = \frac{TP}{P} \text{ und Spezifität} = 1 - \text{fp rate} = 1 - \frac{FP}{N}.$$

Die ROC-Kurven sind zweidimensionale Graphen, in denen tp rate von 0 bis 1 (0% bis 100%) auf der Y-Achse und die fp rate von 0 bis 1 (0% bis 100%) auf der X-Achse aufgetragen ist. Eine ROC-Grafik stellt entsprechende Kompromisse zwischen Nutzen (true positives) und Kosten (false positives) dar.

Im Folgenden wird als Beispiel ein diskreter Klassifikator vorgeführt. Jeder diskrete Klassifikator erzeugt ein (fp rate, tp rate) Paar entsprechend einem einzigen Punkt im ROC-Raum. Die fünf Klassifikatoren (A bis E) in den untenstehenden Grafik sind alle diskrete Klassifikatoren. Es wird ein Kompromiss zwischen true positives und false positives gesucht.

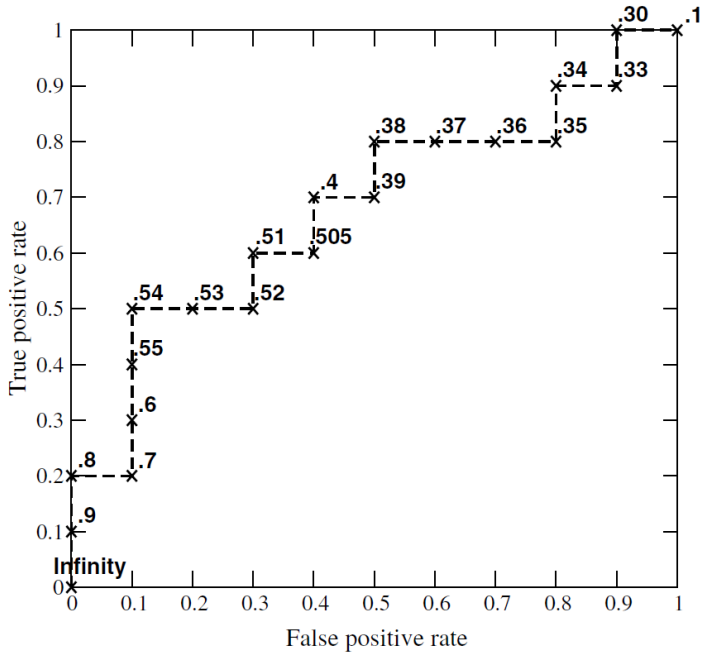


Der Punkt auf der ROC-Kurve, welche den geringsten Abstand zur linken oberen Ecke hat, stellt die optimale Kombination aus true positives und false positives. Somit ist auch der Punkt D der bestmögliche Klassifikator. Die diagonale Linie $y = x$ deutet auf eine rein zufällige Zuordnung bzw. einen Zufallsprozess hin. Der Punkt C wird als Zufallsgröße betrachtet, weil er auf der Diagonalen liegt.

Wenn ein diskreter Klassfierer zu einem Test-Set angewendet wird, ergibt eine einzige Confusions Matrix, die wiederum einem ROC-Punkt entspricht. Also produziert ein diskreter Klassifikator nur einen einzigen Punkt im ROC-Raum. Daher wird nun auch ein skalarer Wert (Score) betrachtet und mit Wahrscheinlichkeits-Klassifierer gearbeitet. Ein solches Ranking oder Scoring Klassifikator verwendet einen Schwellenwert um einen diskreten (binären) Klassifikator zu erzeugen: Wenn der Klassifikatorausgang über dem Schwellenwert liegt, erzeugt der Klassifikator ein Y, sonst ein N. Jeder Schwellenwert erzeugt somit einen anderen Punkt in ROC-Raum. Die genaue Vorgehensweise wird am untenstehenden Beispiel erläutert.

Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1

Die Tabelle beinhaltet 10 Positiv und 10 Negativ Klassen und die dazugehörigen Scorewerte. Die 20 Beobachtungen werden anhand der Scorewerte absteigend sortiert. Bei einer Schwelle von $+\infty$ wird der Punkt $(0; 0)$ erzeugt. Wird der Schwellenwert auf $0,9$ reduziert, so wird erkannt, dass erste positive Instanz positiv klassifiziert wurde $(0; 0,1)$. Im nächsten Schritt wird die Schwelle auf 0,8 gesetzt und wieder wurde bei der zweiten Beobachtung die positive Instanz als positiv klassifiziert $(0; 0,2)$. Beim darauffolgenden Schritt wird der Schwellenwert auf 0,7 gesetzt und diesmal wurde die dritte Beobachtung als positiv klassifiziert obwohl es negativ sein sollte. Daher gilt hier $(0,1; 0,2)$. So wird das ganze weitergeführt, bis der Schwellenwert von 0,1 und $(1; 1)$ erreicht wird.



In Bezug auf die ROC-Kurve existiert noch die AUC (Area Under the Curve) Maßzahl. AUC ist definiert als die Fläche unter der ROC-Kurve. Dabei gilt ein AUC nahe 1 als zuverlässiges und AUC weniger als 0,5 als wertloser bzw. als zufälliger Klassifikator.

2.2 Aufgabe 5u

Aufgabenstellung:

Bearbeiten Sie die Fallstudie SPESSART aus dem SMB (Kapitel 4, Seite 149 ff.) mit

- Logistischer Regression
- Schrittweiser logistischer Regression.

Erläutern Sie Ihre Ergebnisse (Einzelschritte und Fazit) und ordnen Sie diese auch den in dem SMB dargestellten Ergebnissen zu. Stellen Sie auf nachprüfbare Art und Weise den Zusammenhang zwischen mathematischen Formeln (SMB) und SAS-Input/Output her.

Lösung:

Es werden in der Fallstudie SPESSART an 82 verschiedenen Standorten ($n = 82$) der Blattverlust von Buchen untersucht. Im Folgenden wird eine (binäre) logistische Regressionsanalyse für die Betrachtung des Blattverlustes angewendet. Es wird also geprüft, ob ein Zusammenhang zwischen einer abhängigen binären Variable (hier Blattverlust) und mehreren unabhängigen Variablen (die unten aufgeführten acht Regressoren) besteht. Als Regressoren fungieren dann:

- x_1 = NG: Neigung [Grad] des Hanges
- x_2 = HO: Meereshöhe
- x_3 = AL: Alterswert
- x_4 = BS: Beschirmungsgrad
- x_5 = BT: Bestandstyp
- x_6 = DU: Düngung
- x_7 = HU: Humusstärke
- x_8 = PHo: pH-Wert in 0 - 2 cm Tiefe

Den Blattverlust, der in den Kategorien

$$\text{BuDef} = 0 \approx 0\%,\ 1 \approx 12,5\%,\ 2 \approx 25\%, ...$$

gemessen wurde, wird hier reduziert auf die Alternativen

$$\begin{aligned} Y &= 0 \quad \text{kein Blattverlust [d.h. BuDef} = 0], \\ Y &= 1 \quad \text{Blattverlust [d.h. BuDef} \geq 1]. \end{aligned}$$

Es wird also angenommen, dass hier eine dichotome (binäre) Variable Y vorliegt, die nur die Werte 0 und 1 annimmt. Man möchte nun beim logistischen Regressionsmodell mithilfe der logistischen Verteilungsfunktion den Effekt der erklärenden Variablen $x_1, ..., x_k$ (hier $k = 8$ Regressoren) auf die Wahrscheinlichkeit für $Y_i = 0$ bzw. $Y_i = 0$ bestimmen. Hierbei wird der Erwartungswert von Y auf das Intervall $[0, 1]$ beschränkt, durch das Aufsetzen einer *Response*-Funktion $F(x),\ x \in \mathbb{R}$ wird hier die Linearkombination $\eta = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_8 x_8$ der Regressoren ebenfalls auf dieses Intervall begrenzt. Die logistische Regression benutzt als Responsefunktion die logistische Funktion

$$\Pi_i = P(Y_i = 1 | x_1, ..., x_k) = F(\eta) = \frac{e_i^\eta}{1 + e_i^\eta} = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_8 x_8)}}.$$

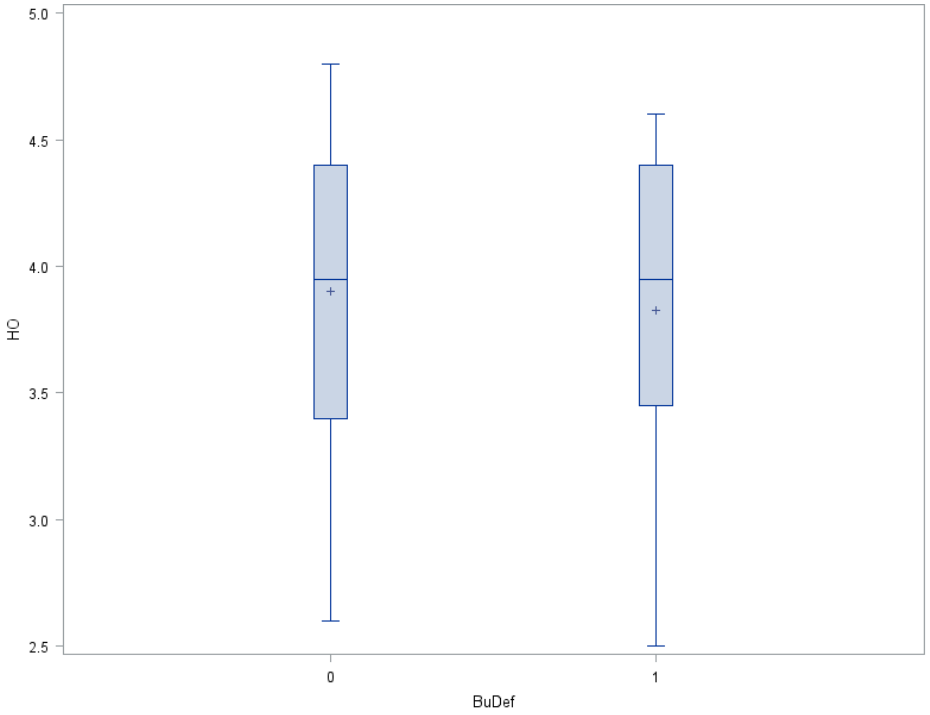
Im Vergleich zur linearen Regression wird also hier die Wahrscheinlichkeit für $Y = 1$ nicht direkt aus den erklärenden Variablen modelliert, sondern indirekt über das sogenannte Logit-Modell. Das Logit ist die logarithmierte Chance für das Auftreten von $Y = 1$ und ist wie folgt definiert:

$$\eta = Logit(Y_i = 1|x_1, ..., x_k) = ln\left(\frac{\Pi_i}{1 - \Pi_i}\right) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

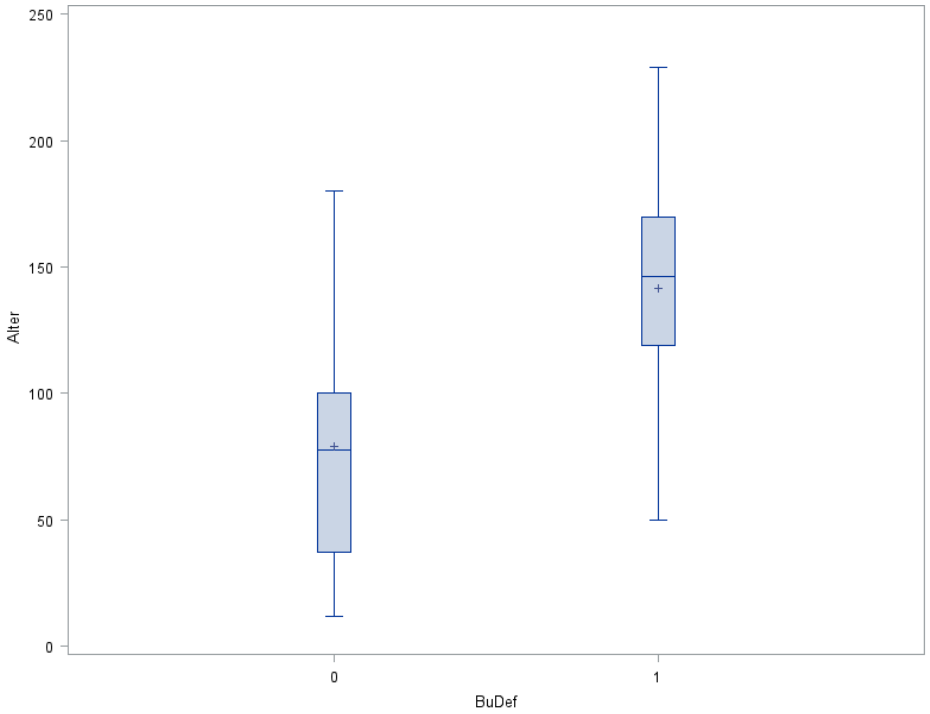
Zusätzlich wird die Chance $\frac{\Pi_i}{1-\Pi_i} = \frac{P(Y_i=1)}{P(Y_i=0)}$ auch als Odds bezeichnet und wird unten an der entsprechenden Stelle weiter ausgeführt.

Zunächst wurden im *Statistischen Methodenbuch* [HP06] von Helmut Pruscha die drei Regressoren Meereshöhe HO, Alter AL und Humusstärke HU aufgespalten in die Werte $Y = 0$ (ohne Defoliation) und $Y = 1$ (min Defoliation) und in Form von Boxplots dargestellt.

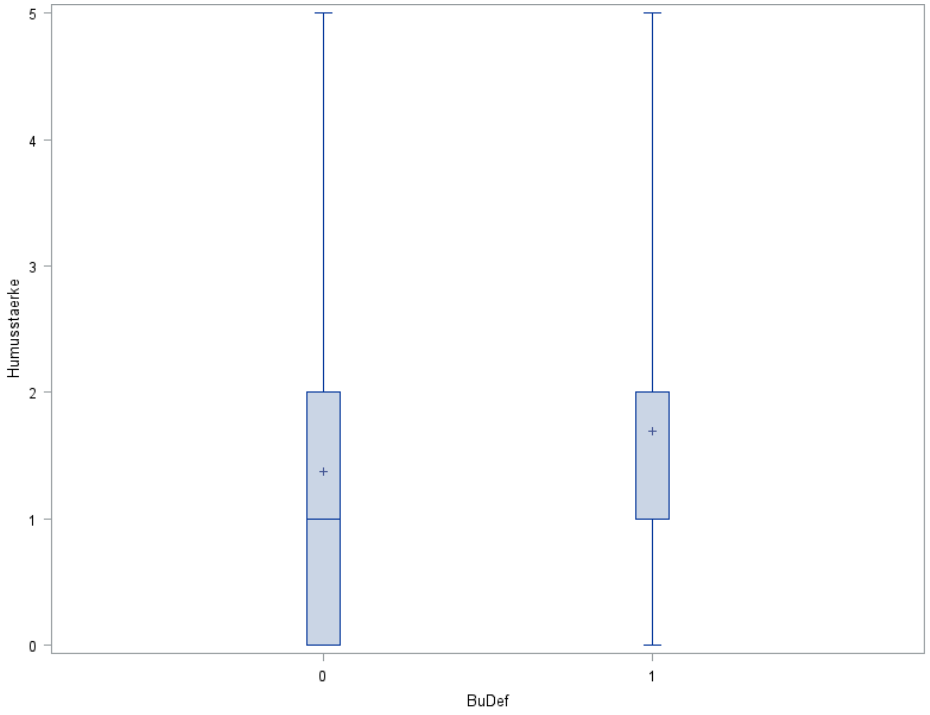
Zunächst erkennt man, dass der Regressor Meereshöhe HO keinen Einfluss auf den Blattverlust hat.



Bei der Regressor Alter AL fällt auf, dass höheres Alter eher mit Blattverlust einhergeht als geringeres Alter.



Zuletzt wird noch Humusstärke als Bloxplot dargestellt. Hier hat der Regressor eine geringe Auswirkung auf den Blattverlust.



Nun wird eine **binäre logistische Regressionsanalyse** für die $n = 82$ Buchen-Standorte durchgeführt. Die Wald-Statistik testet die Nullhypothese, dass der jeweilige Regressionskoeffizient β in der Grundgesamtheit 0 ist. Somit wird nun überprüft, ob die unabhängigen Variablen einen Einfluss haben oder nicht. Falls die p-Werte des Wald-Tests kleiner als die konventionelle Signifikanzgrenze von 5% sind, wird es als signifikant eingeschätzt und die Regressoren leisten einen Beitrag zur Erklärung der abhängigen Variablen. Die Hypothesen lauten:

$$\begin{aligned} H0 &: \text{Der Regressionskoeffizient} \beta_i \text{ ist Null.} \\ H1 &: \text{Der Regressionskoeffizient} \beta_i \text{ ist ungleich Null.} \end{aligned}$$

Hier bei der globale Test auf das Modell schreibt SAS nur BETA=0. Dies kann man wie folgt verstehen:

$$\begin{aligned} H0 &: \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H1 &: \text{Mindestens ein } \beta_i \neq 0,\ 1 \leq i \leq k \end{aligned}$$

Die Ergebnisse des Wald-Tests und deren Signifikanz können dann unten in der Abbildung entnommen werden. Hierbei stellt sich heraus, dass die Regressoren einen signifikanten Einfluss auf die abhängigen Variablen, hier den Blattverlust, haben.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	54.0243	8	<.0001
Score	42.0015	8	<.0001
Wald	23.0504	8	0.0033

Die folgende Tabelle gibt dabei die Schätzer der Regressionskoeffizienten wieder. SAS schätzt die Koeffizienten mittels der Maximum-Likelihood-Funktion. Zusätzlich sind ihre Standardfehler und die dazugehörigen Teststatistiken angegeben.

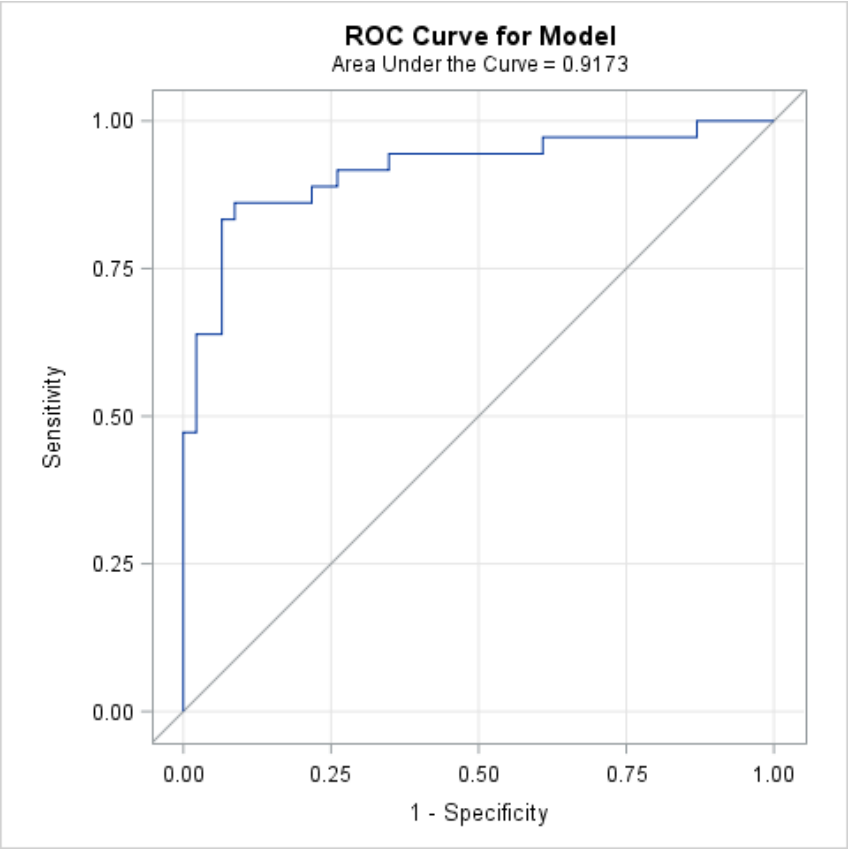
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-20.9703	8.7425	5.7536	0.0165
NG	1	0.0382	0.0446	0.7333	0.3918
HO	1	0.6142	0.7282	0.7114	0.3990
AL	1	0.0271	0.00850	10.1313	0.0015
BS	1	-0.6203	0.2161	8.2365	0.0041
BT	1	1.0301	0.7985	1.6640	0.1971
DU	1	-1.9116	1.1330	2.8467	0.0916
HU	1	0.5571	0.2845	3.8336	0.0502
PHo	1	3.7407	1.5516	5.8124	0.0159

Aufgrund der P-Werte kann behauptet werden, dass die Regressoren AL, BS und PHo einen signifikanten Einfluss auf den Blattverlust haben. Die restlichen Regressoren haben keinen signifikanten Einfluss und können aus der obigen Modellgleichung entfernt werden. Somit folgt

$$F(\eta) = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3)}}$$

mit x_1 =AL, x_2 =BS und x_3 =PHo.

Am Schluss kann noch die ROC-Kurve betrachtet werden. Hier hat die ROC-Kurve eine AUC von 0,9173. Trotz nicht signifikanter Regressoren liefert die logistische Regression somit ein gutes Ergebnis.



Im Folgenden wird eine **schrittweise logistische Regressionsanalyse** durchgeführt. Die mit der *forward* Methode durchgeführte schrittweise logistische Regression liefert zunächst (in Kurzfassung) die folgenden fünf Tabellen. Hierbei wurde jedes Mal ein Regressor überprüft, ob er signifikant genug ist um in das Modell aufgenommen zu werden. Als Voreinstellung hatte das Programm ein Signifikanzlevel von 10% für forward Mehtode (slentry=0,1). Diese wurde auch hier weiterhin beibehalten.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	33.3602	1	<.0001
Score	28.5513	1	<.0001
Wald	20.5781	1	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.7434	2	<.0001
Score	34.8779	2	<.0001
Wald	22.4062	2	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	45.8650	3	<.0001
Score	37.1092	3	<.0001
Wald	22.0036	3	<.0001

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.6731	4	<.0001
Score	38.8637	4	<.0001
Wald	22.1038	4	0.0002

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.3986	5	<.0001
Score	40.6616	5	<.0001
Wald	22.5779	5	0.0004

Zusätzlich gibt SAS eine Zusammenfassung (Summary) die mit der forward Methode in das Modell aufgenommene Regressoren.

Summary of Forward Selection						
Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	AL	1	1	28.5513	<.0001	Alter
2	BS	1	2	8.4852	0.0036	
3	PHo	1	3	4.3314	0.0374	PH-Wert oben
4	DU	1	4	2.7234	0.0989	
5	HU	1	5	2.7090	0.0998	Humusstaerke

Demzufolge sind die Regressoren AL, BS, PHo, DU das optimale Ergebnis. Zu diesen Regressoren werden in der folgenden Tabelle die Koeffizienten geschätzt. Zuzüglich werden ihre Standardfehler und Teststatistiken mit ausgegeben.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-14.6451	5.6551	6.7067	0.0096
AL	1	0.0309	0.00838	13.5778	0.0002
BS	1	-0.5616	0.1915	8.5999	0.0034
DU	1	-1.9325	0.9978	3.7513	0.0528
HU	1	0.3990	0.2495	2.5578	0.1097
PHo	1	3.4214	1.2843	7.0974	0.0077

Folglich hat die schrittweise logistische Regression zwei weitere Regressoren als die normale logistische Regression ausgegeben. Daher gilt

$$F(\eta) = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5)}}$$

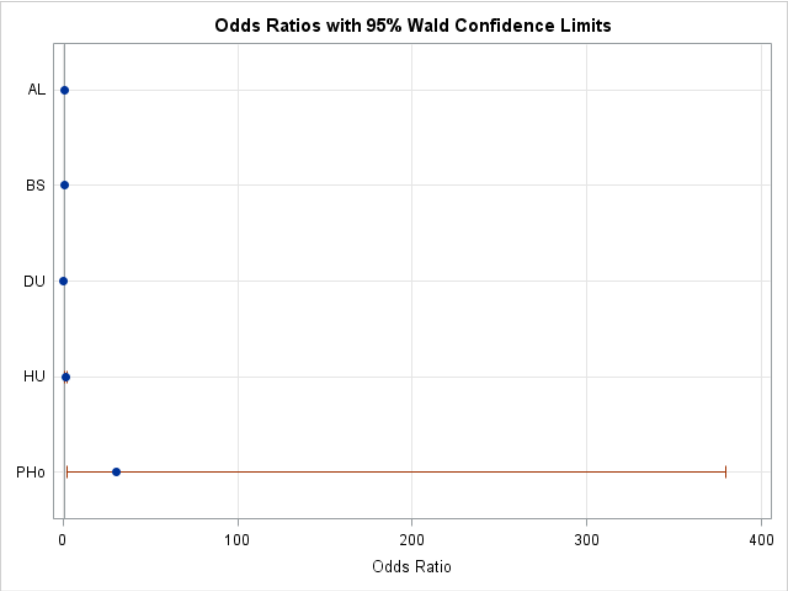
mit x_1 =AL, x_2 =BS, x_3 =DU, x_4 =HU und x_3 =PHo.

Zusätzlich werden nun die Schätzer des Quotenverhältnis, auch Odds-Ratio (kurz OR) genannt, betrachtet. Dabei drücken die Chance, das Eintreten eines Ereignisses, im Verhältnis zu dem Nicht-Eintreten des Ereignisses auf. In der Regel werden zu der Odds-Ratio-Schätzern die 95% Konfidenzintervalle angegeben. Das bedeutet, dass der gesuchte Parameter mit einer Wahrscheinlichkeit von 95% im Konfidenzintervall liegt. Überdies erlaubt der Konfidenzintervall Schlüsse bezüglich der statistischen Signifikanz zu ziehen. Beinhaltet der 95%-ige Konfidenzintervall nicht den Wert der Nullhypothese, welche bei Odds Ratio 1 wäre (also H0: OR =1), dann bedeutet das ein signifikantes Ergebnis zum Niveau $\alpha = 5\%$ für den Ausschluss eines Nulleffekts ist. Zusammenfassend kann also die statistische Signifikanz durch Konfidenzintervallen geprüft werden. Dabei müssen die komplett oberhalb oder unterhalb der 1 liegen.

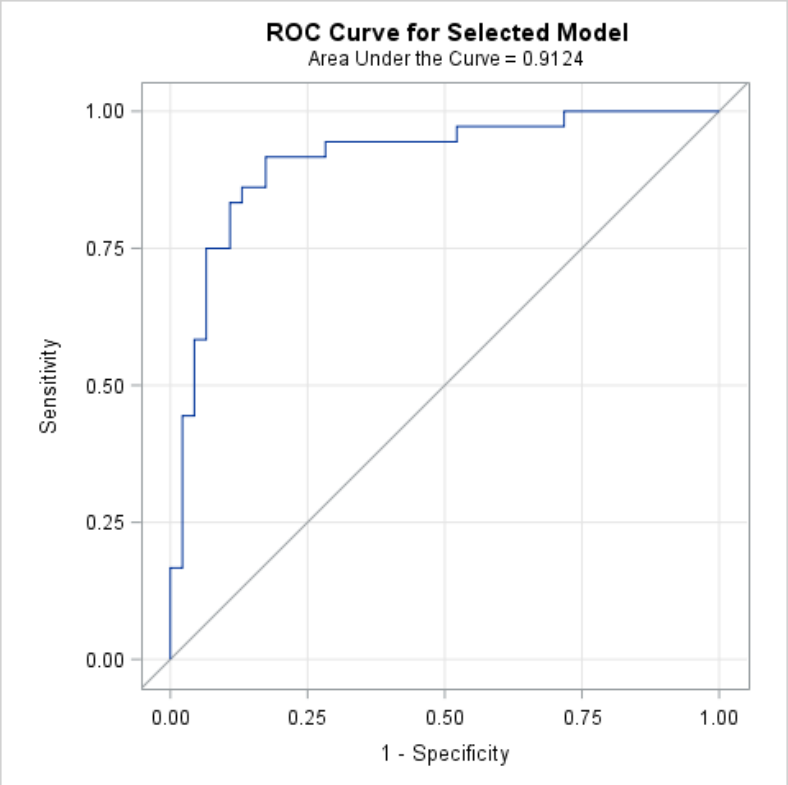
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AL	1.031	1.015	1.048
BS	0.570	0.392	0.830
DU	0.145	0.020	1.023
HU	1.490	0.914	2.430
PHo	30.613	2.470	379.388

Es lässt sich ablesen, dass die Regressoren AL, BS und PHo statistisch signifikant sind. Somit haben die Regressoren DU und HU keinen signifikanten Einfluss.

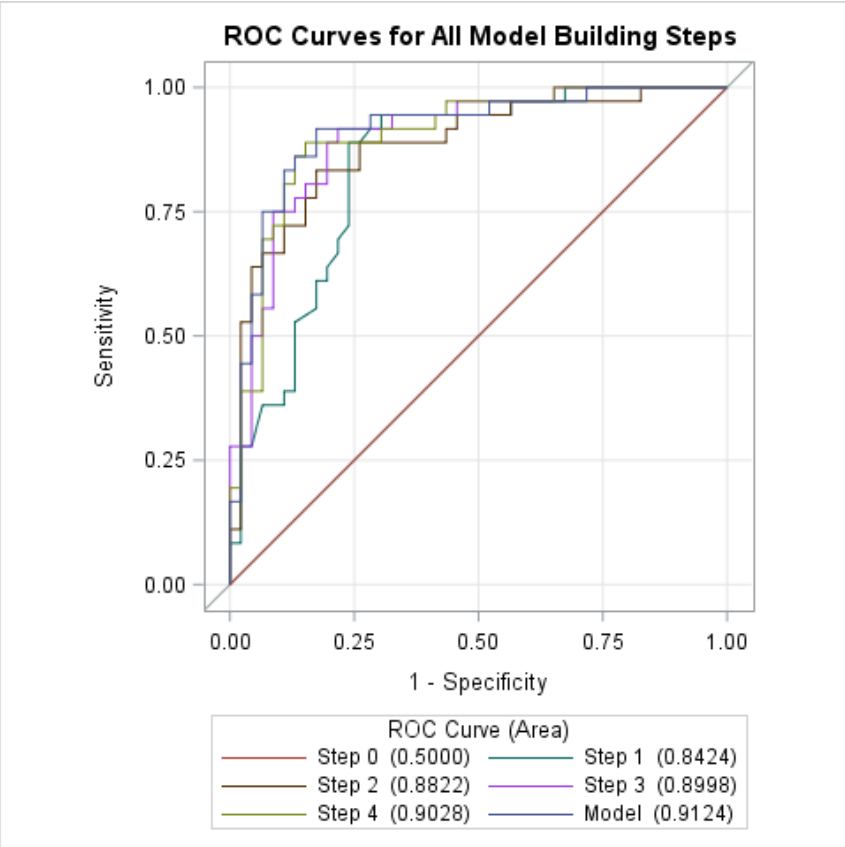
In der untenstehenden Grafik ist ferner das stark ausgeprägte Konfidenzintervall von PHo abgebildet.



Die, dem Modell zugehörige ROC-Kurve hat einen sehr guten AUC-Wert.



Zudem kann in der untenstehenden Grafik das schrittweise verbesserte Modell anhand der jeweiligen ROC-Kurven und deren AUC-Werte betrachtet werden.



Bei der logistischen Regressionsanalyse werden wie schon erläutert Wahrscheinlichkeiten berechnet, dass die abhängige Variable den Wert 1 (SAS: BuDef (event='1') annimmt. Diese Wahrscheinlichkeiten variieren natürlich zwischen 0 und 1. Der Trennwert bzw. der Schwellenwert liegt bei der **Klassifikationstabelle bzw. Confusion Matrix** bei 0,5. Liegt die berechnete Blattverlustwahrscheinlichkeit unter 0,5 so wird es in Y=0 eingeteilt, sonst in Y=1.

Table of _INTO_ by _FROM_				
		FROM(Formatted Value of the Observed Response)		
		0	1	Total
INTO(Formatted Value of the Predicted Response)				
0	Frequency	40	6	46
	Percent	48.78	7.32	56.10
	Row Pct	86.96	13.04	
	Col Pct	86.96	16.67	
1	Frequency	6	30	36
	Percent	7.32	36.59	43.90
	Row Pct	16.67	83.33	
	Col Pct	13.04	83.33	
Total	Frequency	46	36	82
	Percent	56.10	43.90	100.00

Es ergeben hier 70 (40 + 30) korrekt klassifizierte Fälle. Außerdem ist zu erkennen, dass Blattverlust (BuDef = 1) bei 40 von 46 und kein Blattverlust (BuDef = 0) bei 30 von 36 Beobachtungen richtig klassifiziert wurde.

Zusätzlich könnte man noch das Modell ohne PHo ausführen um die Konfidenzintervalle der restlichen Regressoren besser zu betrachten. Daher wird das Modell nun ohne PHo ausgeführt.

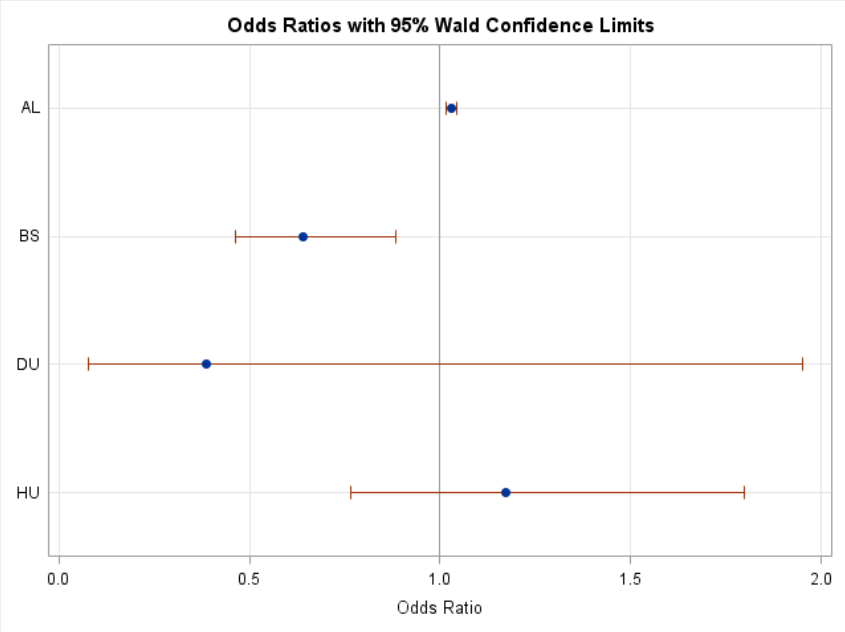
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2656	1.5743	0.0285	0.8660
AL	1	0.0291	0.00757	14.7674	0.0001
BS	1	-0.4470	0.1640	7.4318	0.0064
DU	1	-0.9491	0.8251	1.3232	0.2500
HU	1	0.1601	0.2179	0.5397	0.4625

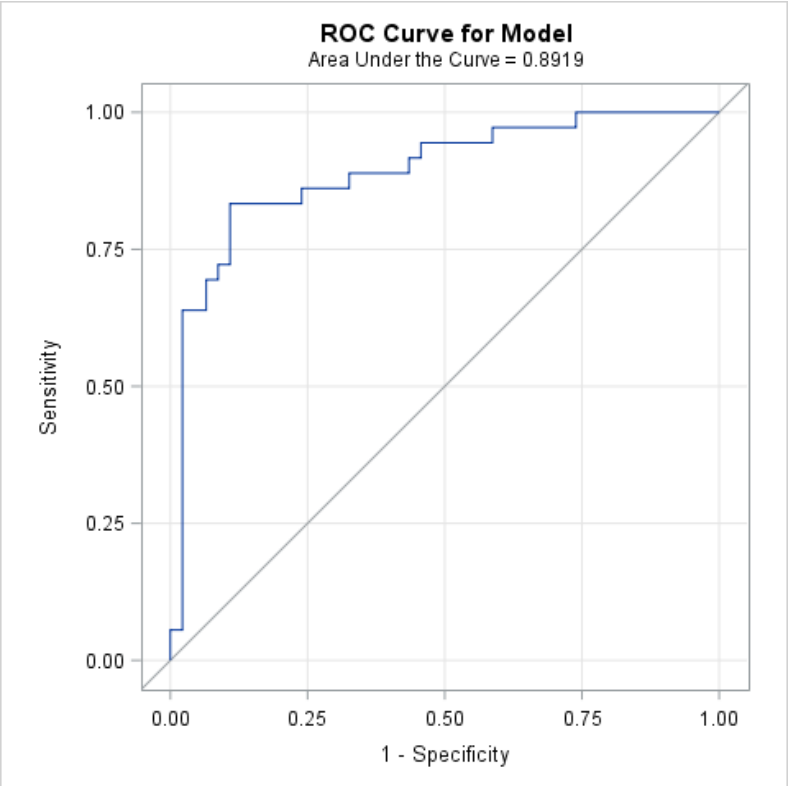
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AL	1.029	1.014	1.045
BS	0.640	0.464	0.882
DU	0.387	0.077	1.950
HU	1.174	0.766	1.799

Anhand der unteren Grafik ist deutlich zu erkennen, dass die Regressoren DU und HU nicht signifikant sind.

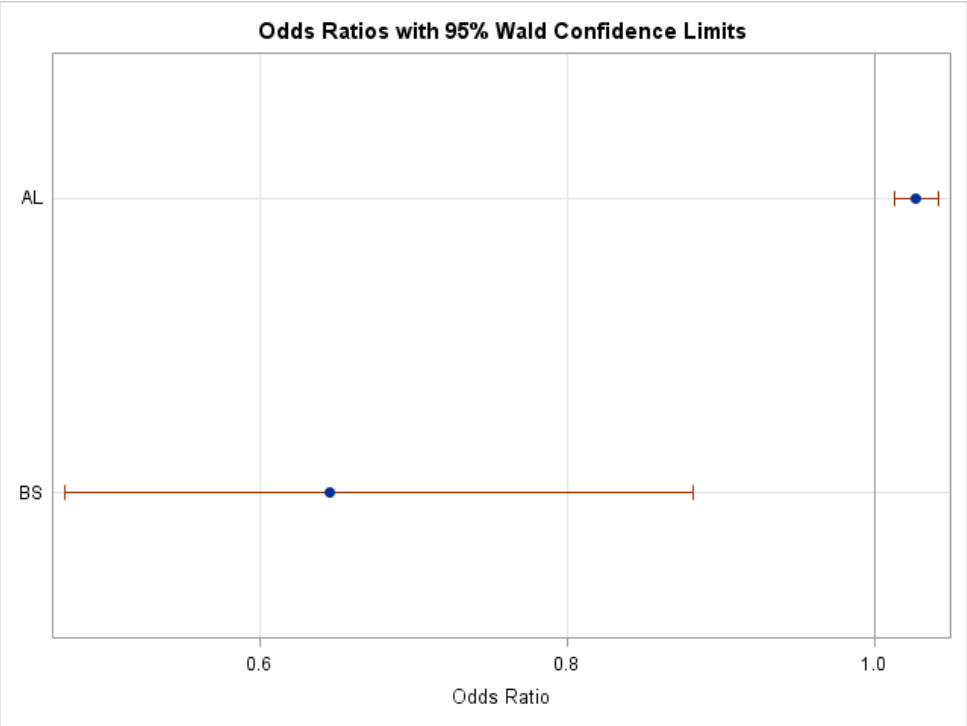


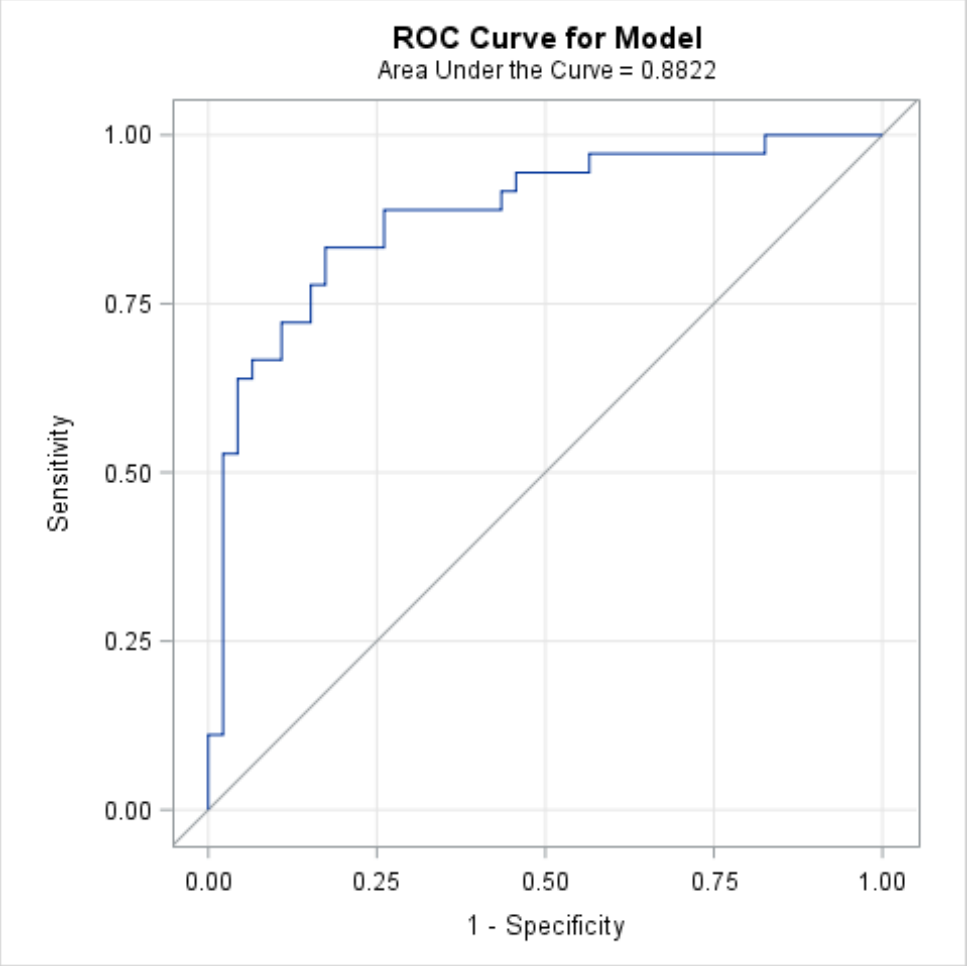
Desweiteren wird die ROC-Kurve abgebildet. Hier ergibt sich ein AUC Wert von 0,8919. Der Wert ist etwas geringer als bei der normalen logistischen Regression.



Als Schlussbetrachtung werden die nicht signifikanten Regressoren DU und HO aus dem Modell rausgenommen.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.00648	1.4913	0.0000	0.9965
AL	1	0.0269	0.00710	14.3763	0.0001
BS	1	-0.4373	0.1589	7.5712	0.0059





Der AUC Wert ist zwar ein bisschen geringer jedoch wurden hier auch nur zwei Regressoren, AL und BS, für das Modell eingesetzt.

2.3 Aufgabe 5v

Aufgabenstellung:

Bearbeiten Sie die Daten aus dem SAS Programm UEBUNG_61_START.sas oder einem anderen bekannten Programm, welche in SDA/MVA zur Logistischen Regression in VL oder als Aufgaben verwendet worden sind. Und zwar mit

- Logistischer Regression
- Schrittweiser logistischer Regression.

Erläutern Sie Ihre Ergebnisse (Einzelschritte und Fazit) und ordnen Sie diese auch den in der SDA VL dargestellten Ergebnissen zu. Stellen Sie auf nachprüfbare Art und Weise den Zusammenhang zwischen mathematischen Formeln (SDA VL) und SAS-Input/Output her.

Erweitern Sie Ihr Wissen um einige Elemente, z.B. Grafiken zur Bewertung der Regression, statistische Test, Klassifikationstabellen.

Lösung:

In dieser Aufgabe wird der Therapieerfolg betrachtet. Die abhängige Variable, also Erfolg einer Therapie, wird wieder in zwei Kategorien eingeteilt:

- Y=0 kein Therapieerfolg
- Y=1 Therapieerfolg

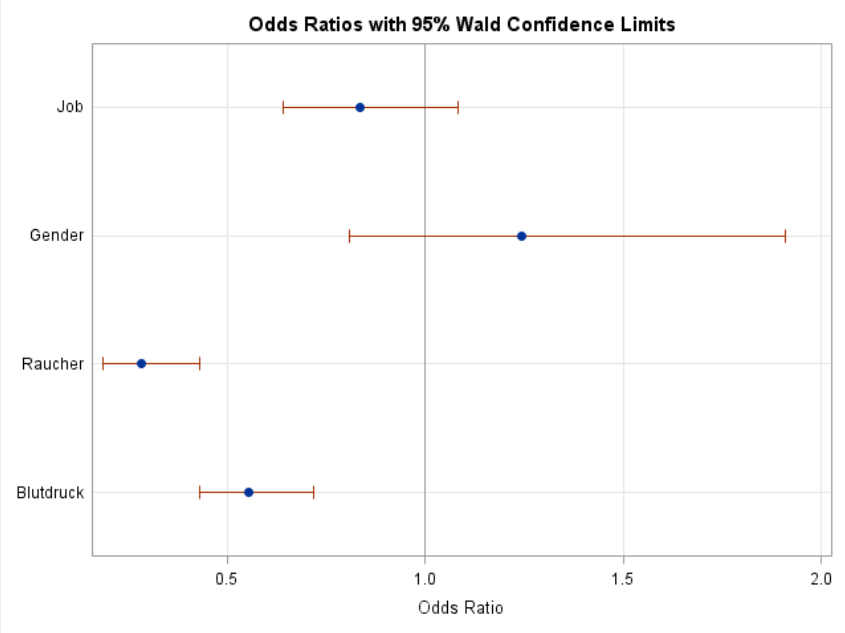
Hier werden die vier unabhängigen Variablen Job, Gender, Raucher und Blutdruck als Regressoren untersucht. Dabei soll es wieder zunächst mit der **logistischen Regression** untersucht werden, ob die vier Regressoren einen signifikanten Einfluss auf den Therapieerfolg haben oder nicht. Zunächst wird wie in Aufgabe 5u ein globaler Test durchgeführt, ob die Regressoren einen signifikanten Einfluss auf die abhängige Variablen haben.

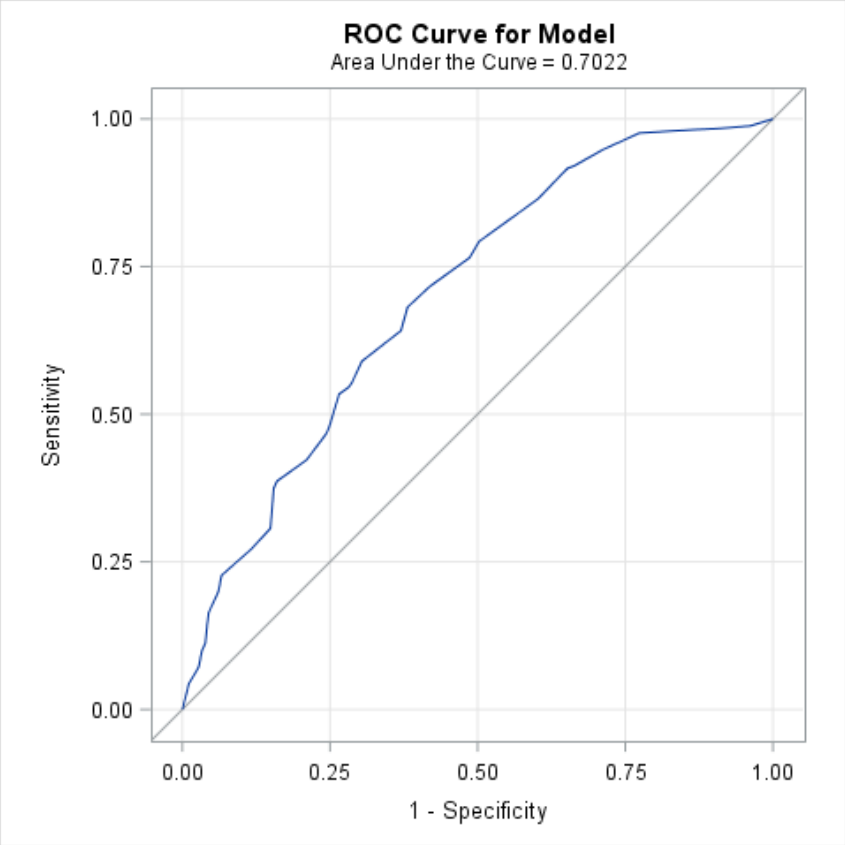
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	56.3874	4	<.0001
Score	53.7873	4	<.0001
Wald	48.5781	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8664	0.3907	4.9190	0.0266
Job	1	-0.1807	0.1328	1.8530	0.1734
Gender	1	0.2180	0.2193	0.9888	0.3200
Raucher	1	-1.2579	0.2142	34.4803	<.0001
Blutdruck	1	-0.5869	0.1315	19.9306	<.0001

Laut der p-Werte sind nur die Regressoren Raucher und Blutdruck signifikant. Dies wird zusätzlich dank der Odds Ratio Schätzer und der grafische Darstellung der Konfidenzintervalle bestätigt.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Job	0.835	0.643	1.083
Gender	1.244	0.809	1.911
Raucher	0.284	0.187	0.433
Blutdruck	0.556	0.430	0.719





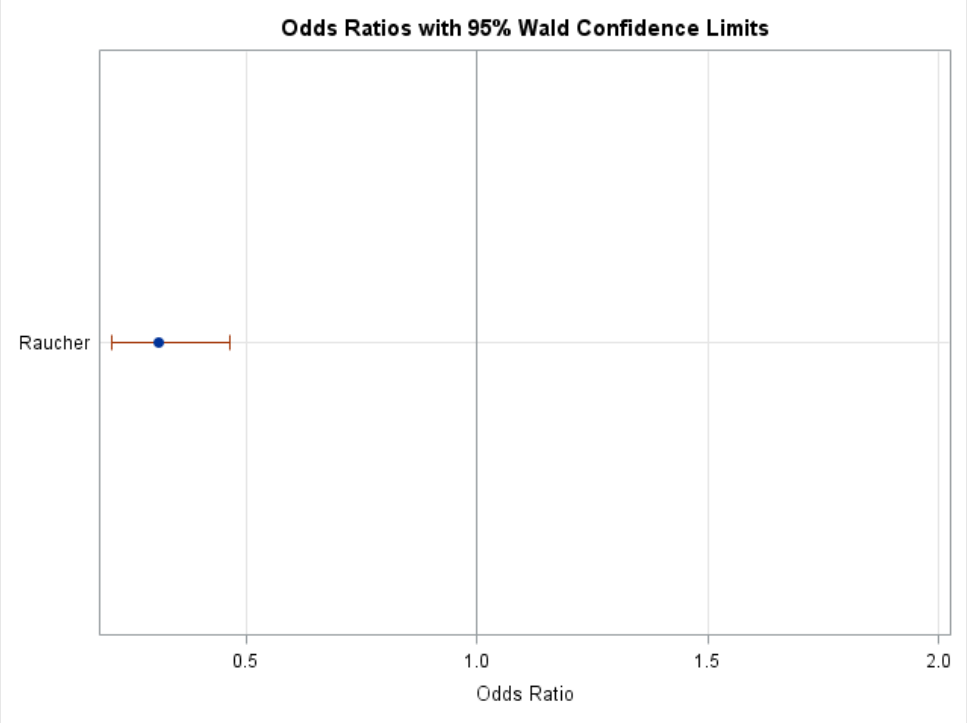
Der AUC Wert von 0,7022 der ROC-Kurve ist zwar nicht sehr nah an 1 dennoch größer als 0,5.

Nun wird die **schrittweise logistische Regressionsanalyse** durchgeführt. Zunächst wird getestet, ob der Regressor Raucher aufgenommen werden soll.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.2671	1	<.0001
Score	33.9061	1	<.0001
Wald	32.8974	1	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	0.8952	0.1450	38.1172	<.0001	2.448
Raucher	1	-1.1655	0.2032	32.8974	<.0001	0.312

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Raucher	0.312	0.209	0.464



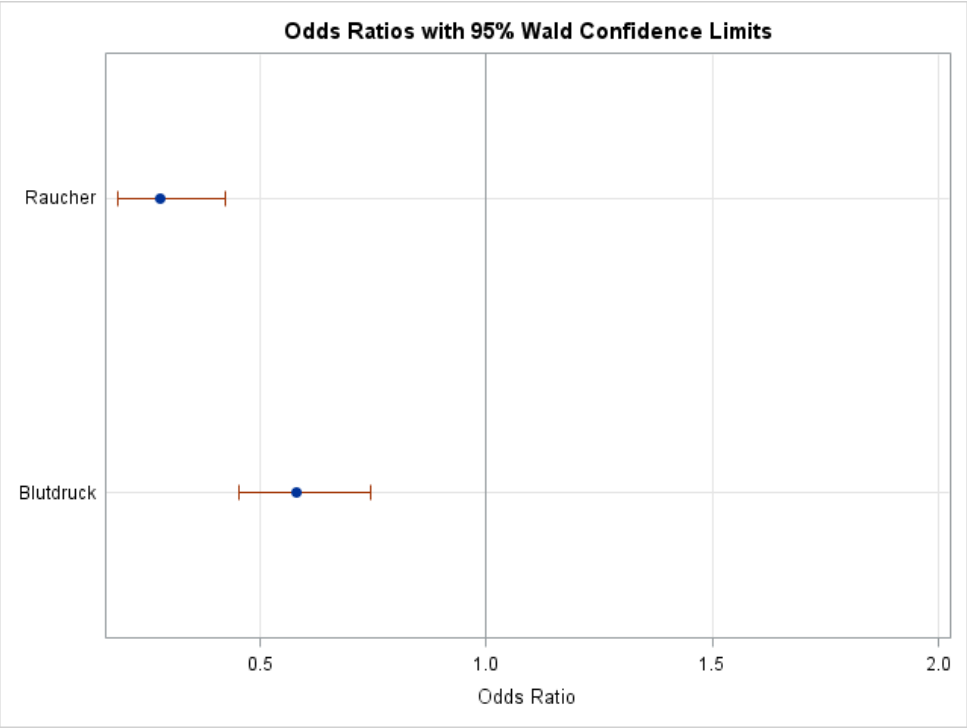
Der Regressor Raucher wird in das Modell aufgenommen, da er laut des Konfidenzintervalles der Odds Ratio Schätzer signifikanten Einfluss auf den Therapieerfolg hat.

Im zweiten Schritt wird der Regressor Blutdruck geprüft.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	53.6920	2	<.0001
Score	51.6741	2	<.0001
Wald	47.1063	2	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	0.9956	0.1511	43.4039	<.0001	2.706
Raucher	1	-1.2762	0.2114	36.4272	<.0001	0.279
Blutdruck	1	-0.5432	0.1260	18.5850	<.0001	0.581

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Raucher	0.279	0.184	0.422
Blutdruck	0.581	0.454	0.744



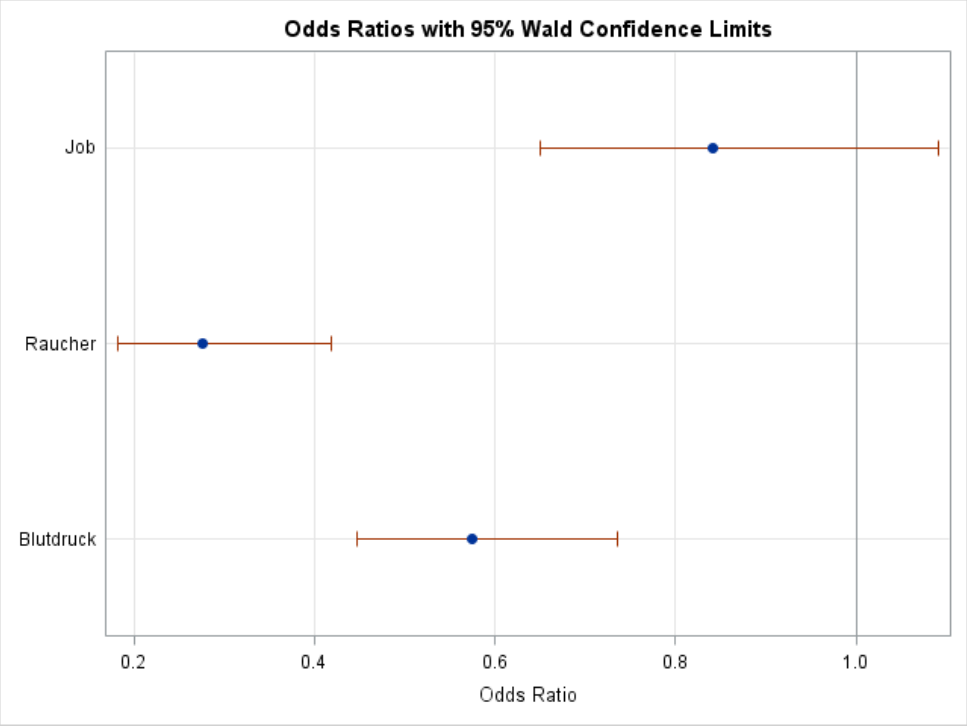
Hier ist auch der zweite Regressor signifikant und wird in das Modell aufgenommen.

Im letzten Schritt wird der Regressor Job betrachtet.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	55.3963	3	<.0001
Score	52.9953	3	<.0001
Wald	47.9962	3	<.0001

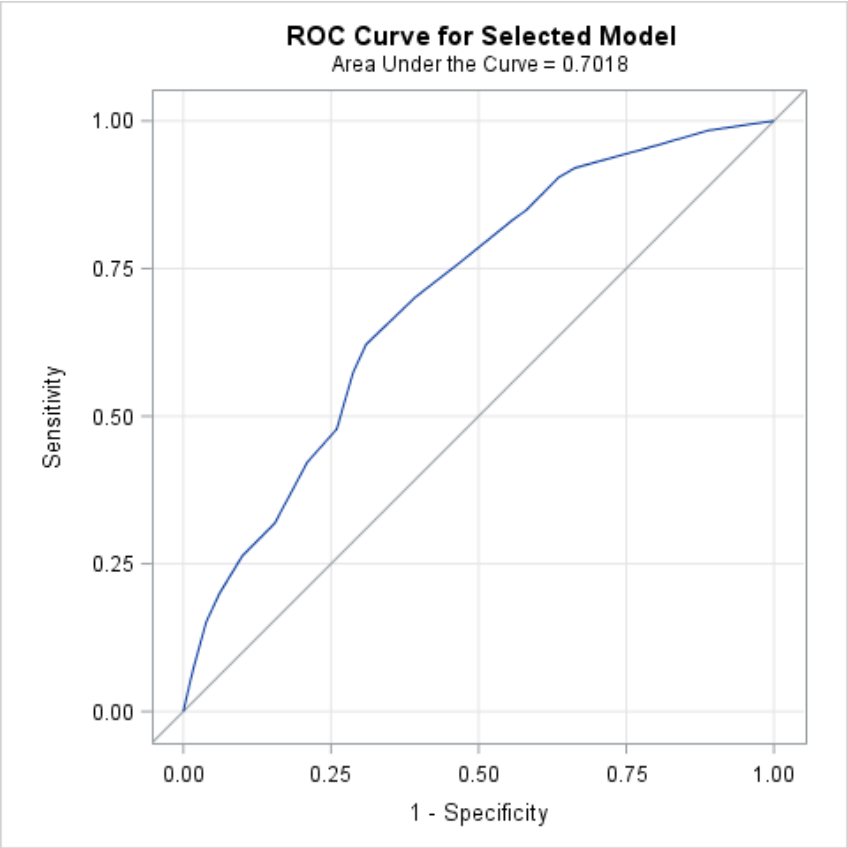
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	1.1922	0.2167	30.2758	<.0001	3.294
Job	1	-0.1722	0.1323	1.6939	0.1931	0.842
Raucher	1	-1.2871	0.2124	36.7366	<.0001	0.276
Blutdruck	1	-0.5548	0.1269	19.1147	<.0001	0.574

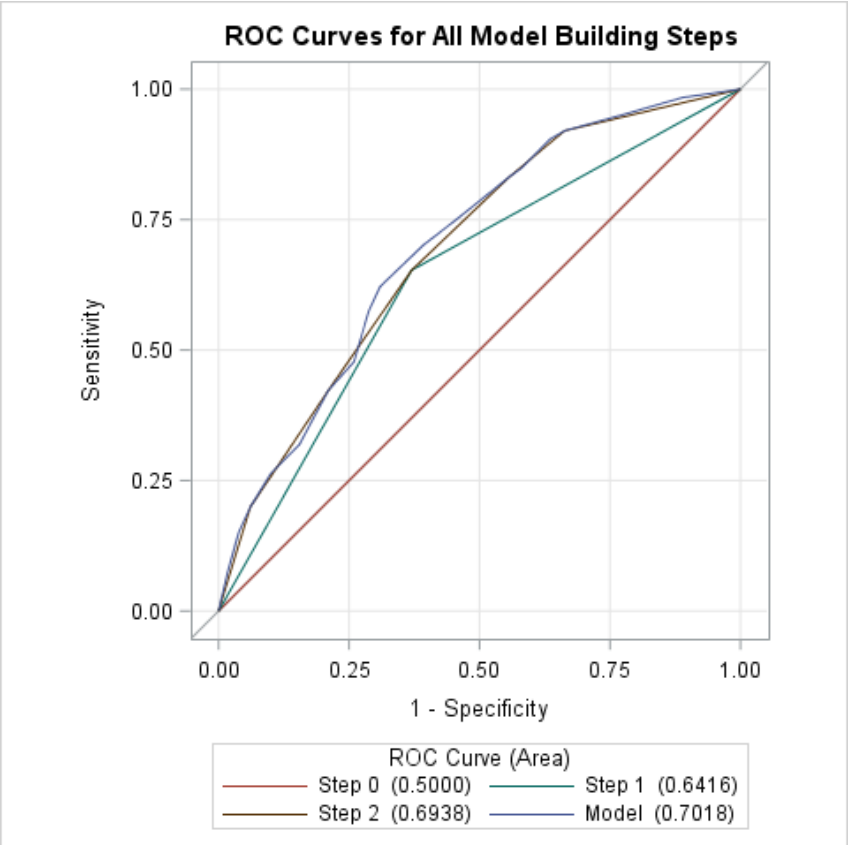
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Job	0.842	0.650	1.091
Raucher	0.276	0.182	0.419
Blutdruck	0.574	0.448	0.736



Hierbei erkennt man, dass der Regressor Job eigentlich als nicht signifikant bezeichnet werden sollte und auch nicht in die Modellgleichung aufgenommen werden sollte. Hier wurde aber die in der Vorlesung und in dem SAS Programm UEBUNG_61_START.sas vordefinierte slentry=0.3 und slstay=0.35 für die schrittweise logistische Regression (bei stepwise Methode) übernommen. Daher wird auch hier der Regressor Job als signifikant eingestuft und in das Modell aufgenommen.

Die dazugehörige ROC-Kurve liefert ein AUC Wert von 0,7018 was annähernd gleich den AUC Wert von der logistischen Regression ist.





Im Anschluss können auch die ROC-Kurven der schrittweise logistischen Regression in einer Grafik dargestellt werden. Hier wird auch der Anstieg der AUC Werte sichtbar, die durch Aufnahme der signifikanten Regressoren zustande kommen.

In der untenstehenden Tabelle können auch die p-Werte der jeweiligen Regressoren beobachtet werden. Hier wurde wie oben schon erwähnt der Signifikanzlevel bei der schrittweise logistischen Regression (forward Methode) für die Aufnahme der Regressoren in das Modell auf 0,3 gesetzt. Daher wird auch der Regressor Job mit in das Modell aufgenommen.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Raucher		1	1	33.9061		<.0001
2	Blutdruck		1	2	19.0822		<.0001
3	Job		1	3	1.6986		0.1925

Zuletzt wird noch eine **Klassifikationstabelle** erstellt, um zu erkennen wie viele Therapieerfolge oder ob keine Therapieerfolge richtig zugeordet sind. Der Trennwert bzw. der Schwellenwert liegt, wie schon in Aufgabe 5u erläutert, bei der Klassifikationstabelle bzw. Confusion Matrix bei 0,5. Das bedeutet, dass bei berechneter Therapieerfolgwahrscheinlichkeit von 0,49 auf Y=0 (kein Therapieerfolg) und ab 0,5 auf Y=1 (Therapieerfolg) gesetzt wird.

Table of _INTO_ by _FROM_

INTO(Formatted Value of the Predicted Response)		_FROM_(Formatted Value of the Observed Response) 0	1	Total
0	Frequency	80	42	122
	Percent	18.52	9.72	28.24
	Row Pct	65.57	34.43	
	Col Pct	44.20	16.73	
1	Frequency	101	209	310
	Percent	23.38	48.38	71.76
	Row Pct	32.58	67.42	
	Col Pct	55.80	83.27	
Total	Frequency	181	251	432
	Percent	41.90	58.10	100.00

Es ist zu erkennen, dass die Fälle aus Gruppe 0 (keine Therapieerfolg) zu $44,20\%$ und die Fälle aus Gruppe 1 (Therapieerfolg) zu $83,27\%$ richtig klassifiziert werden. Die Korrektklassifikationsrate beträgt hier

$$\frac{80+209}{432} = 0,669.$$

Dies bedeutet, dass hier $66,9\%$ alle Therapieerfolge bzw. keine Therapieerfolge richtig klassifiziert wurden. Das Modell ist somit gut. Für eine ausführlichere Klassifikation (mit verschieden Schwellenwerte) können die obige ROC-Kurve und die AUC-Werte betrachtet werden.

2.4 Aufgabe 7u

2.4.1 7u A

Aufgabenstellung:

Sie erhalten eine SAS Datei <DEPRESS>, dazu ein sogenannten **Codebook**, welches die Bedeutung der Variablen erklärt und auch ein Programm, welches den codierten Werten SAS Formate mit den Orginal-Antwort-Texten zuordnet. Möglicherweise ist es besser und einfacher, weder die Daten dauerhaft in Texte zu konvertieren, noch sie dauerhaft mit SAS-Formaten (fest) zu assoziieren, sondern nur gelegentlich wegen der Bedeutung in der Formatauflistung nachzuschauen (Geschmacksache).

Die wesentliche ZIELGRÖßE (TARGET) ist die Variable <CASES>, welche Depressionsfälle beschreibt, die als CESD >= 16 definiert werden. CESD ist ein in den USA verwendeter medizinischer Score, bei dem große Werte das Vorliegen einer Depression anzeigen.

Sie sollen verschieden Modelle erstellen, begründen und bewerten, welche die Depression == binäre Variable <CASES> aus anderen Einflussgrößen vorhersagen kann.

Dabei macht es keinen Sinn, die Variable CESD als Einflussgröße verwenden zu wollen!

- Arbeiten Sie mit Logistischer Regression (SAS Proc Logistic).
- Arbeiten Sie mit Linearer Diskriminanzanalyse [LDA] (SAS Procs CANDISC/DISCRIM).

Was ist das einfachste nicht-triviale Modell, was sind Ihre zwei besten Modelle und warum?

Wichtig ist das Demonstrieren, dass Sie die wesentlichen Schritte und Elemente bei logistischer Regressionen/Diskriminanzanalysen verstanden haben und erklären können, insbesondere bei einfachen Modellen.

Lösung:

Die Aufgabe beschäftigt sich mit den <DEPRESS> Daten. Anhand dieser soll nun zunächst **logistische Regression** durchgeführt werden. Mit den folgenden 14 Regressoren wird versucht die Eintrittswahrscheinlichkeit eines Depressionszustandes (abhängige Variable: CASES) vorherzusagen.

- SEX: Geschlecht
- AGE: Alter der Person
- MARTIAL: Familienstand
- EDUCAT: Schulbildung
- EMPLOY: Beschäftigungsstatus
- INCOME: Jahreseinkommen
- RELIG: Religionszugehörigkeit
- DRINK: Gibt an, ob die Person regulär Alkohol trinkt oder nicht
- HEALTH: Zustand der Gesundheit
- REGDOC: Gibt an, ob die Person regulär zum Arzt geht oder nicht
- TREAT: Gibt an, ob die Person Medikamente nimmt oder nicht
- BEDDAYS: Gibt an, ob die Person in den letzten zwei Monaten einen oder mehrere Tage in Bett verbracht hat oder nicht

- ACUTEILL: Gibt an, ob die Person innerhalb der zwei Monate an einer akuten Krankheit gelitten hat oder nicht
- CHRONILL: Gibt an, ob die Person in der letzten Jahr an einer chronischen Krankheit gelitten hat oder nicht

Dabei wird der Depressionszustand (CASES) in zwei Kategorien aufgeteilt:

- CASES = 0 keine Depression
- CASES = 1 Depression

Zunächst prüft SAS, ob die Regressionen einen Einfluss auf die abhängige Variable haben.

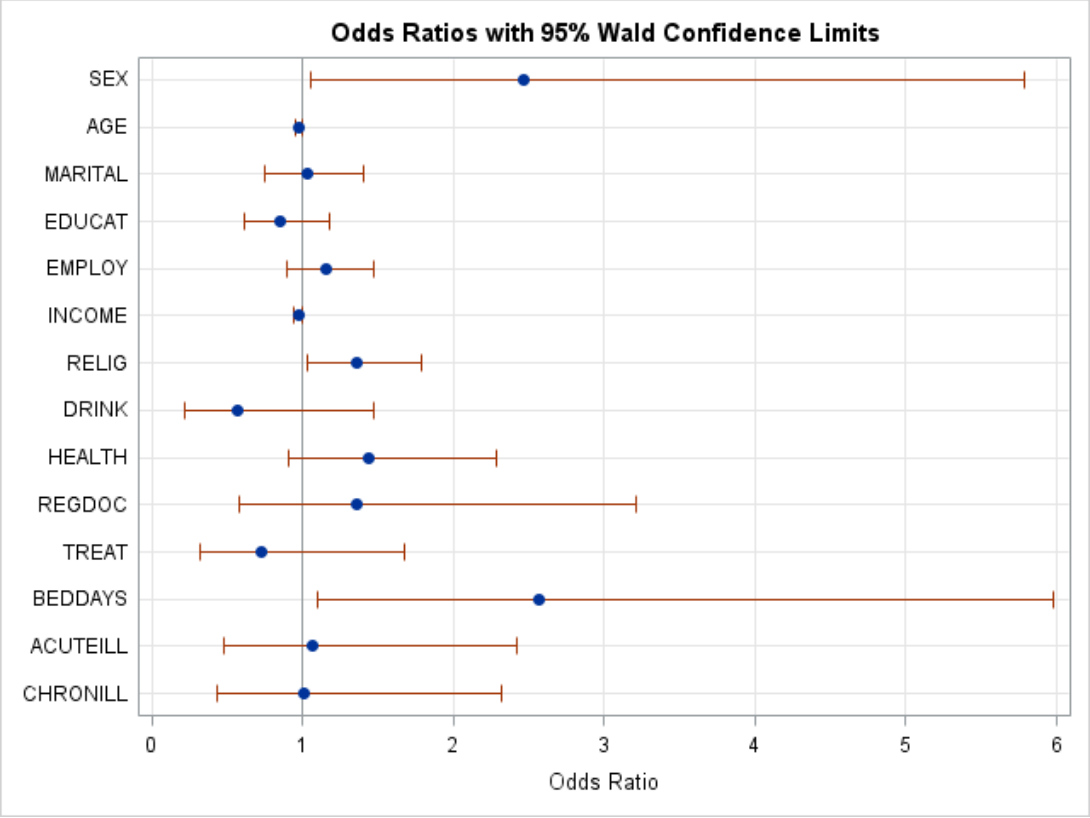
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	45.6259	14	<.0001
Score	43.4073	14	<.0001
Wald	34.2064	14	0.0019

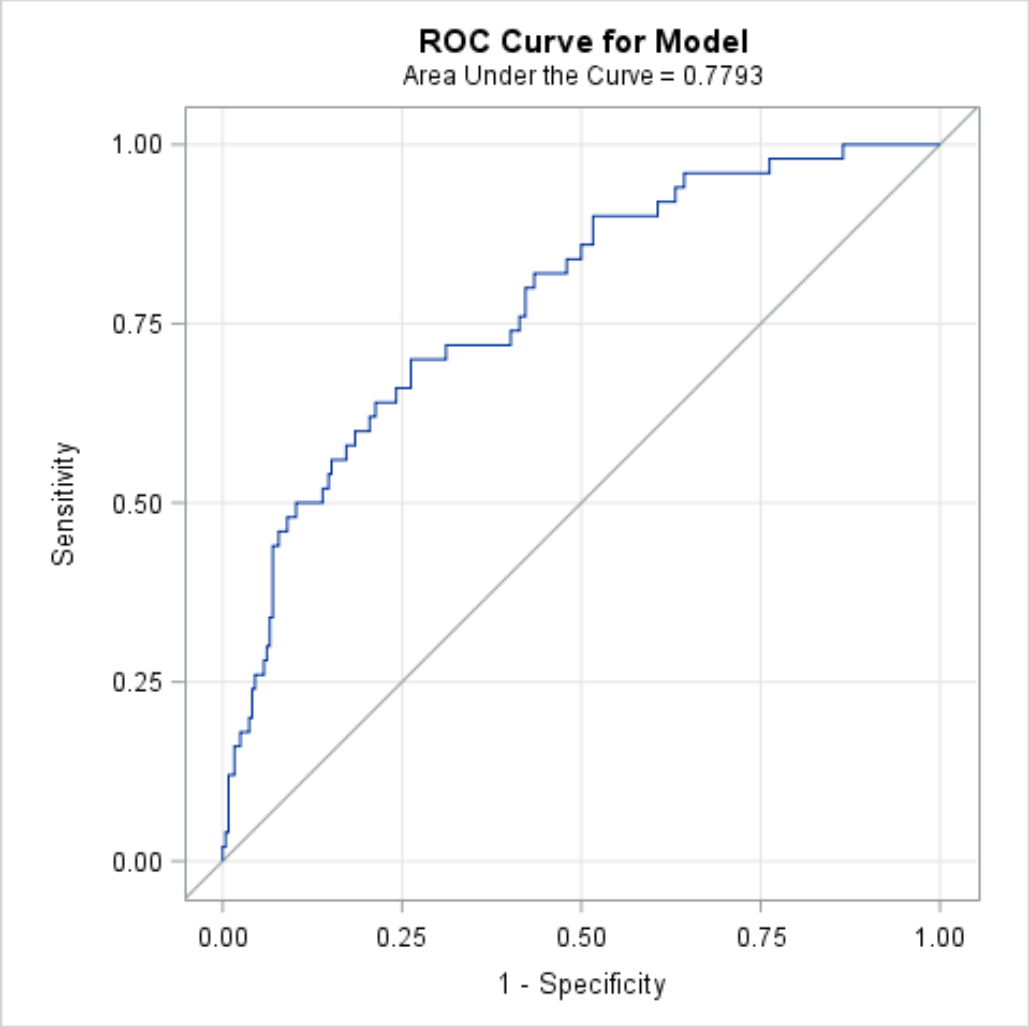
Da der p-Wert vom Wald Test signifikant ausgefallen ist, können nun die einzelne Regressoren und deren Signifikanz betrachtet werden.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.1194	1.6372	1.6757	0.1955
SEX	1	0.9031	0.4344	4.3228	0.0376
AGE	1	-0.0262	0.0134	3.8022	0.0512
MARITAL	1	0.0260	0.1596	0.0266	0.8706
EDUCAT	1	-0.1646	0.1676	0.9646	0.3260
EMPLOY	1	0.1382	0.1251	1.2197	0.2694
INCOME	1	-0.0306	0.0152	4.0435	0.0443
RELIG	1	0.3050	0.1389	4.8205	0.0281
DRINK	1	-0.5664	0.4873	1.3509	0.2451
HEALTH	1	0.3648	0.2339	2.4331	0.1188
REGDOC	1	0.3061	0.4391	0.4858	0.4858
TREAT	1	-0.3252	0.4272	0.5796	0.4465
BEDDAYS	1	0.9420	0.4312	4.7723	0.0289
ACUTEILL	1	0.0650	0.4175	0.0242	0.8764
CHRONILL	1	0.00483	0.4275	0.0001	0.9910

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SEX	2.467	1.053	5.780
AGE	0.974	0.949	1.000
MARITAL	1.026	0.751	1.403
EDUCAT	0.848	0.611	1.178
EMPLOY	1.148	0.898	1.467
INCOME	0.970	0.941	0.999
RELIG	1.357	1.033	1.781
DRINK	0.568	0.218	1.475
HEALTH	1.440	0.911	2.278
REGDOC	1.358	0.574	3.212
TREAT	0.722	0.313	1.669
BEDDAYS	2.565	1.102	5.973
ACUTEILL	1.067	0.471	2.419
CHRONILL	1.005	0.435	2.323

Nur die Regressoren SEX, INCOME, RELIG und BEDDAYS haben einen signifikanten Einfluss. Dies kann auch mit Hilfe der Odds Ratio Schätzer und deren Konfidenzintervalle bestätigt werden.





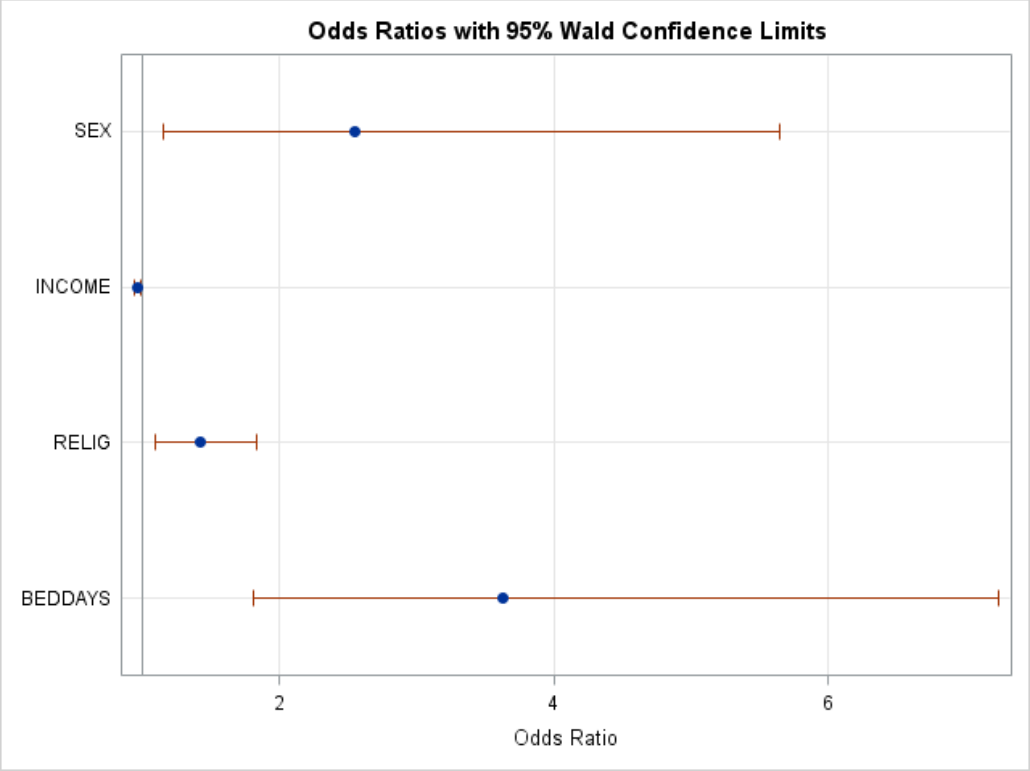
Der AUC-Wert beträgt 0,7793. Es wird erhofft, dass durch Reduktion der Regressoren ein besserer AUC-Wert erzielt wird.

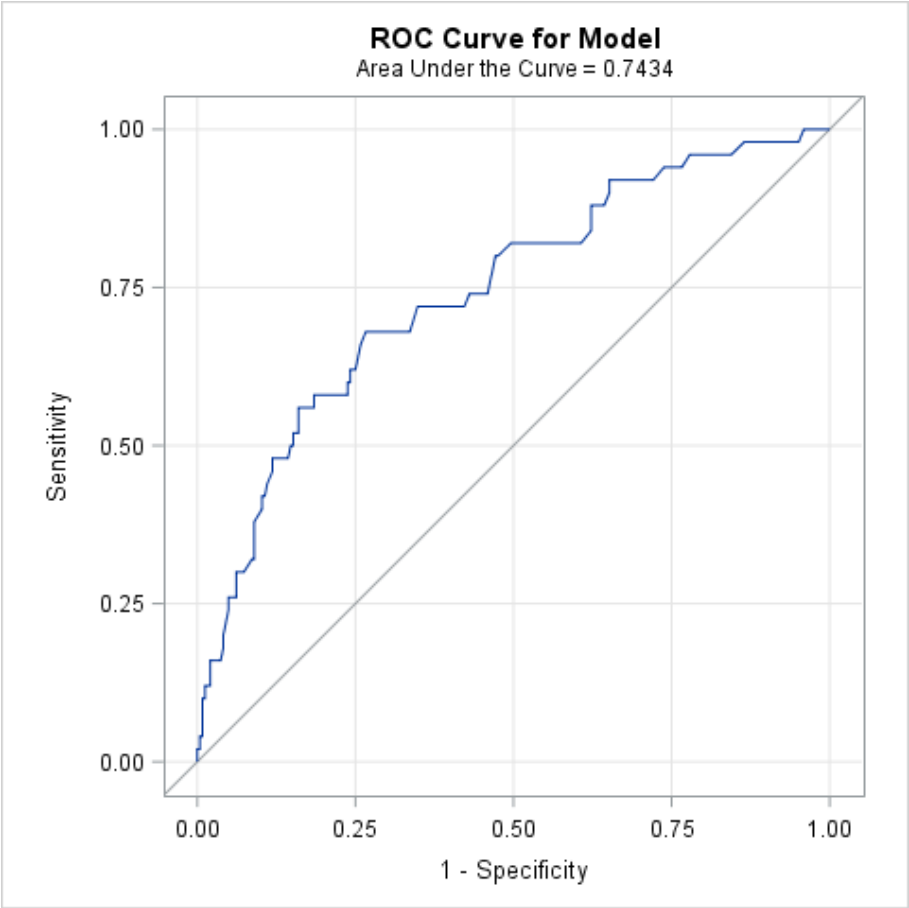
Beim reduzierten Modell wird die logistische Regression nur mit vier Regressoren durchgeführt.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.5676	4	<.0001
Score	33.2976	4	<.0001
Wald	28.4927	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6397	0.8899	16.7285	<.0001
SEX	1	0.9359	0.4054	5.3283	0.0210
INCOME	1	-0.0355	0.0138	6.6298	0.0100
RELIG	1	0.3477	0.1300	7.1581	0.0075
BEDDAYS	1	1.2875	0.3530	13.3041	0.0003

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SEX	2.549	1.152	5.644
INCOME	0.965	0.939	0.992
RELIG	1.416	1.097	1.827
BEDDAYS	3.624	1.814	7.237





Es folgt hier, dass der AUC-Wert sogar sich minimal verschlechtert.

Nun wird eine **Diskriminanzanalyse** durchgeführt. Es ist ein struktur-prüfendes Verfahren, mit dem die Abhängigkeit einer nominal skalierten Variable (Gruppierungsvariable, hier CASES = 0 oder 1) von metrisch skalierten unabhängigen Variablen (Merkmalsvariablen, hier 14 Regressoren, die oben ausgeführt sind) untersucht wird. Bei der logistischen Regression wurde die Wahrscheinlichkeit des Eintretens eines bestimmten Ereignisses in Abhängigkeit der Regressoren ermittelt. Hier dient die Diskriminanzanalyse zur Bestimmung oder Prognose der Gruppenzugehörigkeit von Elementen (Klassifizierung).

Im Rahmen der Diskriminanzanalyse wird eine Diskriminanzfunktion (Trennfunktion) formuliert und geschätzt, die so dann eine optimale Trennung zwischen den Gruppen und eine Prüfung der diskriminatorischen Bedeutung der Merkmalsvariablen ermöglichen soll. Die Diskriminanzfunktion hat die Form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j$$

mit Y als Diskriminanzvariable, X_j als Merkmalsvariable j, β_j als Diskriminanzkoeffizient für Merkmalsvariable j und β_j als konstantes Glied.

Hier gilt dann eine Diskriminanzfunktion

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

mit X_1 als Streichfähigkeit und X_2 als Haltbarkeit.

Zunächst wird die Prozedur DISCRIM ausgeführt. Es werden insgesamt 294 Beobachtungen (TOTAL SAMPLE SIZE) in zwei Gruppen (CLASSES) eingeteilt.

Total Sample Size	294	DF Total	293
Variables	14	DF Within Classes	292
Classes	2	DF Between Classes	1

Number of Observations Read	294
Number of Observations Used	294

Bei der DISCRIM Prozedur wird die Anweisung PRIORS mit angegeben. Gibt man PRIORS PROP (gleichbedeutend mit PROPORTIONAL) an, so wird SAS die a-priori-Wahrscheinlichkeiten, also das Vorwissen über die Klassenzugehörigkeiten, verwenden, das heißt (Anzahl Beobachtungen in der Gruppe) / (Anzahl Beobachtungen in den Daten). Die Gewichtung, also PRIOR PROBABILITY, wird anhand der Class Level Information Tabelle oben abgelesen werden. Hier zeigt es an, dass es mehr CASES = 0 (keine Depressionszustände, hier 83%) in den Daten existieren als CASES = 1 (Depressionszustände, hier 17,01%). Zusätzlich können noch die absoluten Werte, hier CASES = 0 kam 244 und CASES = 1 kam 50 mal vor, abgelesen werden.

Class Level Information					
CASES	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	244	244.0000	0.829932	0.829932
1	1	50	50.0000	0.170068	0.170068

Es folgen nun zwei Klassifikationstabellen bzw. Confusion Matrizen. Zum einen wird es mit der Resubstitution Methode (auch R-Methode genannt) und zum anderen mit der Kreuzvalidierungs-Methdode Confusion Matrix erstellt.

Zunächst wird die Confusion Matrix mit der Resubstitution Methode interpretiert. Hier erkennt man, dass insgesamt 50 (43+7) Beobachtungen falsch klassifiziert wurden. Zudem fällt auf, dass bei 43 von 50 Depressionszustände diese als keine Depression klassifiziert wurde.

Number of Observations and Percent Classified into CASES			
From CASES	0	1	Total
0	237	7	244
	97.13	2.87	100.00
1	43	7	50
	86.00	14.00	100.00
Total	280	14	294
	95.24	4.76	100.00
Priors	0.82993	0.17007	

Insgesamt wurden also 17,01% der Beobachtungen fehlerhaft klassifiziert.

Error Count Estimates for CASES			
	0	1	Total
Rate	0.0287	0.8600	0.1701
Priors	0.8299	0.1701	

Zusätzlich wird eine Confusion Matrix mit der Kreuzvalidierungs-Methode erstellt. Hierbei steigt sogar die Fehlklassifikationen von 50 auf 56 und die Fehlerquote auf 19,05%.

Number of Observations and Percent Classified into CASES			
From CASES	0	1	Total
0	232	12	244
	95.08	4.92	100.00
1	44	6	50
	88.00	12.00	100.00
Total	276	18	294
	93.88	6.12	100.00
Priors	0.82993	0.17007	

Error Count Estimates for CASES			
	0	1	Total
Rate	0.0492	0.8800	0.1905
Priors	0.8299	0.1701	

Bei der logistischen Regression wurde herausgefunden, dass nur vier Regressoren einen signifikanten Einfluss auf CASES haben. Daher wird nun die DISCRIM Prozedur auch auf das reduziertes Modell angewendet und interpretiert.

Total Sample Size	294	DF Total	293
Variables	4	DF Within Classes	292
Classes	2	DF Between Classes	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information

CASES	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	244	244.0000	0.829932	0.829932
1	1	50	50.0000	0.170068	0.170068

Bei der Confusion Matrix mit der Resubstitution Methode ergibt nur 48 (43 + 5) fehlerhafte Klassifikation und eine Fehlerrate von $16,33\%$.

Number of Observations and Percent Classified into CASES

From CASES	0	1	Total
0	239	5	244
	97.95	2.05	100.00
1	43	7	50
	86.00	14.00	100.00
Total	282	12	294
	95.92	4.08	100.00
Priors	0.82993	0.17007	

Error Count Estimates for CASES

	0	1	Total
Rate	0.0205	0.8600	0.1633
Priors	0.8299	0.1701	

Number of Observations and Percent Classified into CASES

From CASES	0	1	Total
0	239	5	244
	97.95	2.05	100.00
1	44	6	50
	88.00	12.00	100.00
Total	283	11	294
	96.26	3.74	100.00
Priors	0.82993	0.17007	

Error Count Estimates for CASES

	0	1	Total
Rate	0.0205	0.8800	0.1667
Priors	0.8299	0.1701	

Und bei der Kreuzvalidierungs-Methode werden 49 (44 + 5) Beobachtungen falsch klassifiziert und es ergibt eine Fehlerquote von $16,67\%$.

Hierbei soll noch erwähnt werden, dass die Prozedur CANDISC die sogenannten *Mahalanobis-Distanzen* (näheres auf Seite 183 in [KB00]) verwendet. Dies geschieht mit der Anweisung DISTANCE.

Zunächst wird, wie bei der DISCRIM Prozedur, die CLASS LEVEL INFORMATION ausgegeben.

Total Sample Size	294	DF Total	293
Variables	14	DF Within Classes	292
Classes	2	DF Between Classes	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information

CASES	Variable Name	Frequency	Weight	Proportion
0	0	244	244.0000	0.829932
1	1	50	50.0000	0.170068

In der untenstehenden Tabelle werden die Merkmalsvariablen bzw. die Regressoren auf ihre Signifikanz für die Unterscheidung der Gruppen geprüft. Somit wird zum einen probiert die Unterschiedlichkeit der Gruppen zu erklären und zum anderen möchte man die unwichtigen Variablen aus der Diskriminanzfunktion entfernen.

Hier wird deutlich, dass die Regressoren SEX, INCOME, RELIG, HEALTH und BEDDAYS signifikanten Einfluss auf die Diskriminanzfunktion haben. Jedoch wird später noch das reduzierte Modell von der logistischen Regression weiter betrachtet. Dort ist der Regressor HEALTH nicht vertreten. Eine ausführlichere Version befindet sich in Teil B der Aufgabe 7v.

Univariate Test Statistics

F Statistics, Num DF=1, Den DF=292

Variable	Label	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
SEX		0.4856	0.4797	0.1137	0.0275	0.0283	8.25	0.0044
AGE	age in years at last birthday	18.0854	18.0234	2.5831	0.0102	0.0103	3.02	0.0833
MARITAL		1.2650	1.2670	0.0219	0.0001	0.0001	0.04	0.8344
EDUCAT		1.3107	1.3049	0.2046	0.0122	0.0124	3.61	0.0583
EMPLOY		1.5506	1.5441	0.2376	0.0118	0.0119	3.48	0.0631
INCOME	thousands of dollars per year	15.2901	15.1205	3.4409	0.0254	0.0261	7.61	0.0062
RELIG		1.2267	1.2192	0.2157	0.0155	0.0158	4.60	0.0327
DRINK	regular drinker?	0.4037	0.4043	0.0154	0.0007	0.0007	0.21	0.6442
HEALTH	general health	0.8379	0.8291	0.1843	0.0243	0.0249	7.26	0.0074
REGDOC	have a regular doctor?	0.3906	0.3899	0.0467	0.0072	0.0072	2.11	0.1476
TREAT	Has a doctor prescribed or recommended that you take medicine, medical treatments, or change your way of living in such	0.5008	0.4998	0.0618	0.0076	0.0077	2.25	0.1346
BEDDAYS	spent entire day(s) in bed in last two months?	0.4110	0.4010	0.1317	0.0515	0.0543	15.86	<.0001
ACUTEILL	any acute illness in last two months?	0.4572	0.4564	0.0538	0.0070	0.0070	2.04	0.1538
CHRONILL	any chronic illness in last year?	0.5008	0.4990	0.0725	0.0105	0.0106	3.10	0.0793

Einer der Kriterien zur Prüfung der Diskriminanzfunktion ist die Wilks' Lambda. Es ist ein inverses Gütemaß, das heißt kleinere Werte bedeuten höhere Trennkraft der Diskriminanzfunktion und umgekehrt. Die Wilks' Lambda ist wie folgt definiert:

$$\Lambda = \frac{1}{1+\lambda} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}}$$

Die folgende Transformation

$$\chi^2 = -\left[N - \frac{J+G}{2} - 1\right] \cdot ln(\Lambda)$$

mit N: Anzahl der Fälle, J: Anzahl der Variable, G: Anzahl der Gruppen und Λ: Wilks' Lambda liefert eine Variable, die angenähert wie χ^2 verteilt ist mit $J \cdot (G-1)$ Freiheitsgraden. Die Signifikanz der Gruppentrennung wird mittels eines χ^2-Tests geprüft.

Dabei werden die folgende Hypothesen definiert und geprüft:

$$H0 \quad : \quad \text{Die beiden Gruppen unterscheiden sich nicht.}$$
$$H1 \quad : \quad \text{Die beiden Gruppen unterscheiden sich.}$$

Multivariate Statistics and Exact F Statistics					
	S=1	M=6	N=138.5		
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.85235617	3.45	14	279	<.0001
Pillai's Trace	0.14764383	3.45	14	279	<.0001
Hotelling-Lawley Trace	0.17321847	3.45	14	279	<.0001
Roy's Greatest Root	0.17321847	3.45	14	279	<.0001

Hier beträgt der Wert für Wilks' Lambda $0,85$. Somit ist die Trennkraft der Diskriminanzfunktion nicht hoch. Jedoch kann gesagt werden, dass H0 abzulehnen und H1 anzunehmen ist. Dies geschieht aufgrund des signifikanten p-Wertes. Somit ist H0 abzulehnen und H1 anzunehmen. Es lässt sich also sagen, dass die beiden Gruppen sich signifikant unterscheiden.

Bei der Diskriminanzanalyse wird zudem die kanonische Korrelationskoeffizient betrachtet:

$$c = \sqrt{\frac{\gamma}{1+\gamma}} = \sqrt{\frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}}$$

Im Zwei-Gruppen-Fall (wie hier CASES = 0 oder 1) ist die kanonische Korrelation identisch mit der einfachen Korrelation zwischen den geschätzten Diskriminanzwerten und der Gruppierungsvariablen. Somit ist dies ein Maß für die Stärke des Zusammenhanges zwischen der Gruppierungsvariablen und der Diskriminanzfunktion. Der kanonische Korrelationskoeffizient nimmt, ebenso wie Wilks' Lambda, die Werte zwischen 0 und 1. Hier gilt, dass hohe Korrelationen auf starke Zusammenhänge schließen. Beim vollständigen Modell ergibt ein kanonischer Korrelation von $0,38$.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.384244	0.335593	0.049795	0.147644	0.1732		1.0000	1.0000	0.85235617	3.45	14	279	<.0001

Zusätzlich gibt SAS natülich weitere ausfürlichere Tabellen zur CANDISC Prozedur an. Darunter zum Beispiel die quadrierten Distanzen (SQUARED DISTANCE), die Mahalanobis-Distanzen (MAHALANOBIS DISTANCE FOR SQUARED DISTANCE) oder auch LINEAR DISCRIMINANT FUNCTION FOR CASES. Diese Tabellen werden in dieser Ausarbeitung nicht weiter erwähnt, da diese für die interne Berechnungen gebraucht werden (zum Beispiel für Wilks' Lambda oder kanonischer Korrelationskoeffizienten). Bei Interesse kann die expliziten Interpretationsmöglichkeiten in [KB00] nachgelesen werden.

Nun wird die CANDISC Prozedur auf das reduzierte Modell mit vier Regressoren angewendet.

Total Sample Size	294	DF Total	293
Variables	4	DF Within Classes	292
Classes	2	DF Between Classes	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information

CASES	Variable Name	Frequency	Weight	Proportion
0	0	244	244.0000	0.829932
1	1	50	50.0000	0.170068

Wie erwartet sind die Regressoren signifikant.

Univariate Test Statistics

F Statistics, Num DF=1, Den DF=292

Variable	Label	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
SEX		0.4856	0.4797	0.1137	0.0275	0.0283	8.25	0.0044
INCOME	thousands of dollars per year	15.2901	15.1205	3.4409	0.0254	0.0261	7.61	0.0062
RELIG		1.2267	1.2192	0.2157	0.0155	0.0158	4.60	0.0327
BEDDAYS	spent entire day(s) in bed in last two months?	0.4110	0.4010	0.1317	0.0515	0.0543	15.86	<.0001

Zudem ist der Wilks' Lambda auch signifikant.

Multivariate Statistics and Exact F Statistics

S=1 M=1 N=143.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.88674270	9.23	4	289	<.0001
Pillai's Trace	0.11325730	9.23	4	289	<.0001
Hotelling-Lawley Trace	0.12772284	9.23	4	289	<.0001
Roy's Greatest Root	0.12772284	9.23	4	289	<.0001

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.336537	0.323557	0.051804	0.113257	0.1277		1.0000	1.0000	0.88674270	9.23	4	289	<.0001

Der einzige Unterschied wird erst beim kanonischen Korrelationskoeffizienten sichtbar. Hier beträgt die $0,3365$. Dies ist sogar geringer als beim vollen Modell.

2.4.2 7u B

Aufgabenstellung:

RÜCKGRIFF auf elementare Analyse von Kontingenztafeln:
Analysieren Sie den Spezialfall der folgenden Datenreduktion vergleichend mit Logistischer Regression und mit Tabellenanalyse (SAS Proc FREQ - elementare Analyse einer 2×2 Tabelle (Vierfeldertafel).

Hat das Geschlecht (Variable SEX) einen Einfluss? Wie hoch ist dieser? Begründungen!!!

F = Female $\Rightarrow$ X=1; M = Male $\Rightarrow$ X=0 (kategorielle, sogar binäre Einflussgröße!).

SEX	DEPRESSION=JA	DEPRESSION=NEIN
F (X=1)	40	143
M (X=0)	10	101

Lösung:

Laut dem ersten Teil der Aufgabe 7u hat das Geschlecht, also der Regressor SEX, einen signifikanten Einfluss auf den Depressionszustand. Nun wird dies im einzelnen an einem oben aufgeführten kleinen Datensatz mit Hilfe der Prozedur FREQ untersucht.

Table of DEPRESSION_NEU by SEX

		SEX		
		F	M	Total
DEPRESSION_NEU				
JA	Frequency	40	10	50
	Percent	13.61	3.40	17.01
	Row Pct	80.00	20.00	
	Col Pct	21.86	9.01	
NEIN	Frequency	143	101	244
	Percent	48.64	34.35	82.99
	Row Pct	58.61	41.39	
	Col Pct	78.14	90.99	
Total	Frequency	183	111	294
	Percent	62.24	37.76	100.00

Die oben abgebildete Tabelle zeigt an, dass $21,86\%$ der beobachteten Frauen und $9,01\%$ der Männer an einer Depression leiden. Somit kann gesagt werden, dass Frauen mit einer höheren Wahrscheinlichkeit an einer Depression leiden werden als Männer.

2.5 Aufgabe 7v

2.5.1 7v A

Aufgabenstellung:

Sie (haben) erhalten zwei SAS Programme. Wählen Sie **eines** davon aus:

- DISKRIMINANZ_1_2016.sas
- SMB_TOSKANA_PCA_CCA_3.sas

Das erste Programm analysiert die MARGARINE-Daten von Backhaus et al (vgl. Kopien).

Das zweite analysiert aus dem SMB die TOSKANA-Daten mit 2 bzw. 3 Gruppen.

A.

Ordnen Sie die numerischen Zahlenangaben aus den beiden Quellen und die Bilddarstellungen aus den beiden Quellen den Ergebnissen von SAS (Tabellen und Grafiken) zu und erläutern Sie deren Bedeutung.

Sie können sich dabei auf die wesentlichen Elemente konzentrieren.

In Bezug auf den (umfangreichen) Output von SAS gilt <Mut zur Lücke>: wenn Sie in einem Projekt einige Diskriminanzanalysen durchgeführt und interpretiert haben, werden Ihnen die weiteren Elemente schnell klargeworden sein!

Lösung:

Im Folgenden wird die MARGARINE-Daten aus dem *Multivariate Analysemethoden* von Klaus Backhaus [KB00] bearbeitet. Ein Margarinehersteller möchte herausfinden, welche Bedeutung die Merkmale Streichfähigkeit und Haltbarkeit, also zwei Regressoren, für die Markenwahl (hier der abhängige nominale Variable: MARKE = A oder B) haben. Es werden jeweils 12 Stammkäufer der Marken A und B befragt und die insgesamt 24 Personen werden gebeten, die empfundene Wichtigkeit der beiden Merkmale auf einer siebenstufigen Rating-Skala zu beurteilen.

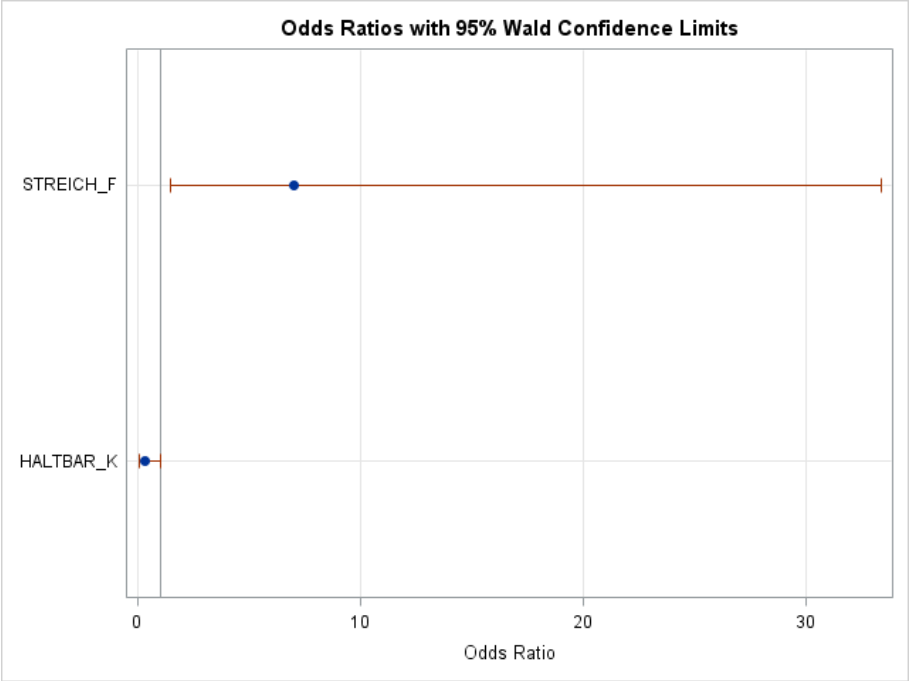
Zunächst folgt eine **logistische Regressionsanalyse** der oben ausgeführten Aufgabenstellung.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.6805	2	0.0006
Score	11.4493	2	0.0033
Wald	5.9431	2	0.0512

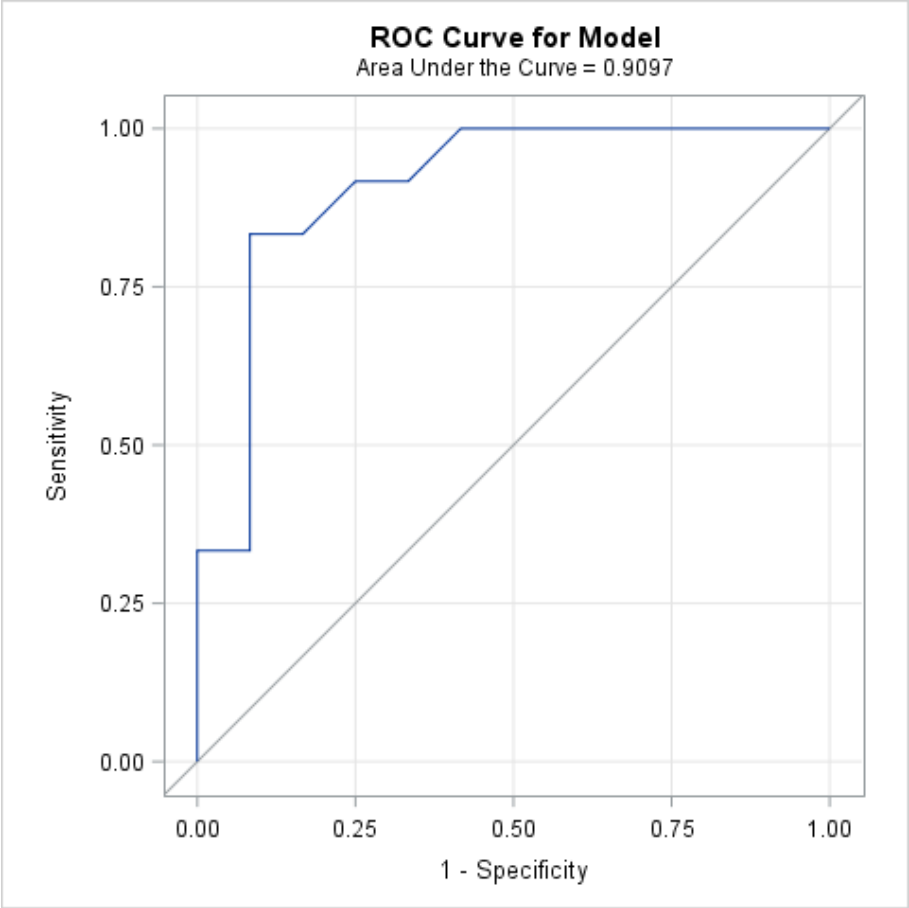
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5277	2.3384	2.2759	0.1314
STREICH_F	1	1.9427	0.7982	5.9240	0.0149
HALTBAR_K	1	-1.1194	0.5863	3.6453	0.0562

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
STREICH_F	6.978	1.460	33.353
HALTBAR_K	0.326	0.103	1.030

Hier ist erkennbar, dass der Regressor HALTBAR_K die 1 innerhalb des Konfidenzintervalles sich befindet und dass der p-Wert 0,0562, also minimal größer als der Signifikanzlevel von 0,5, beträgt. Nichtsdestotrotz wird der Regressor beibehalten. Dies wird durch die Beobachtung der AUC-Werte von ROC Kurven begründet.



Der AUC-Wert beträgt beim vollen Modell (also Regressoren STREICH_F und HALTBAR_K) 0,9097. Wird aber nur der Regressor STREICH_F als signifikant eingestuft und eine logistische Regression durchgeführt, so folgt ein AUC-Wert von 0,8229 (wird nicht in der Ausarbeitung ausgeführt, kann aber im beigefügten Programm nachgesehen werden). Daher wird das volle Modell als richtig und die beiden Regressoren als signifikant genug empfunden.



Zuletzt werden die Daten mit Hilfe der **Diskriminanzanalyse** untersucht. Hierbei werden die Ergebnisse mit den Ausführungen von Klaus Backhaus aus [KB00] verglichen. Wiederholend kann gesagt werden, dass es g=2 Gruppen (MARKE A und B) und 2 Merkmale bzw. Regressoren (STREICH_F und HALTBAR_K) vorliegen. Die beiden Gruppen sollen nun aufgrund der Regressoren optimal getrennt (diskriminiert) werden.

Zunächst werden unter Anderem die Mittelwerte (vgl. Abbildung 4.7) der Regressoren, sortiert nach MARKE = A oder B, ausgegeben.

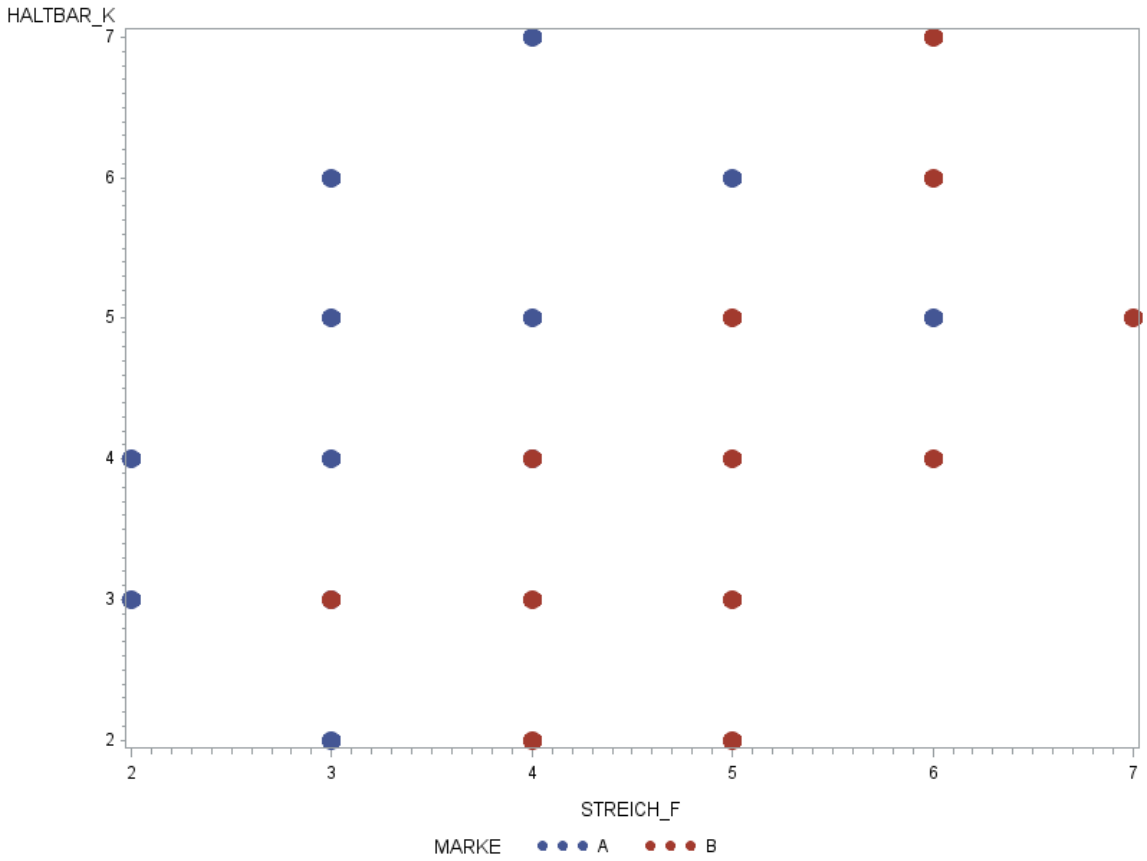
MARKE=A

Variable	N	Mean	Std Dev	Minimum	Maximum
STREICH_F	12	3.5000000	1.1677484	2.0000000	6.0000000
HALTBAR_K	12	4.5000000	1.4459976	2.0000000	7.0000000

MARKE=B

Variable	N	Mean	Std Dev	Minimum	Maximum
STREICH_F	12	5.0000000	1.1281521	3.0000000	7.0000000
HALTBAR_K	12	4.0000000	1.5374122	2.0000000	7.0000000

In folgender Abbildung ist das Ergebnis der Befragung als Streudiagramm dargestellt. Jede der 24 befragten Personen ist entsprechend der abgegebenen Urteilswerte im Raum der beiden Variablen als Punkt repräsentiert. Dabei sind die Käufer von Marke A durch blaue Punkte und die der Marke B durch rote Punkte markiert.



Hierbei ist deutlich, dass die Stammkäufer von Marke B die Wichtigkeit der Streichfähigkeit tendenziell höher einstufen als die Stammkäufer von Marke A. Doch dagegen ergeben sich für die Käufer der Marke A im Durchschnitt etwas höhere Werte bei der Einstufung der Haltbarkeit.

Zudem können auch die Gesamtmittelwerte der Merkmalsvariablen (vgl. Abbildung 4.9) angegeben werden.

Simple Statistics		
	STREICH_F	HALTBAR_K
Mean	4.250000000	4.250000000
StD	1.359347670	1.481773321

Nun wird die CANDISC Prozedur angewendet. Zunächst folgen die grundlegende Informationen. Hierbei erkennt man auch, dass die Gewichte gleich verteilt sind, da ja jeweils 12 Personen für MARKE = A oder B beurteilt haben.

Total Sample Size	24	DF Total	23
Variables	2	DF Within Classes	22
Classes	2	DF Between Classes	1

Number of Observations Read	24
Number of Observations Used	24

Class Level Information				
MARKE	Variable Name	Frequency	Weight	Proportion
A	A	12	12.0000	0.500000
B	B	12	12.0000	0.500000

Zusätzlich können in der folgenden Tabelle die Quadratsumme der Abweichungen vom Mittelwert und die Kreuzproduktsumme der Abweichungen (vgl. Abbildung 4.7) abgelesen werden.

MARKE = A		
Variable	STREICH_F	HALTBAR_K
STREICH_F	15.00000000	9.00000000
HALTBAR_K	9.00000000	23.00000000

MARKE = B		
Variable	STREICH_F	HALTBAR_K
STREICH_F	14.00000000	12.00000000
HALTBAR_K	12.00000000	26.00000000

Pooled Within-Class SSCP Matrix		
Variable	STREICH_F	HALTBAR_K
STREICH_F	29.00000000	21.00000000
HALTBAR_K	21.00000000	49.00000000

Between-Class SSCP Matrix		
Variable	STREICH_F	HALTBAR_K
STREICH_F	13.50000000	-4.50000000
HALTBAR_K	-4.50000000	1.50000000

In den obigen Tabellen können zum einen die Innergruppen-Streuungsmaße der Merkmalsvariablen, hier POOLED-WITHIN-CLASS SSCP MATRIX, und zum anderen die Zwischengruppen-Streuungsmaße der Merkmalsvariablen, hier BETWEEN-CLASS SSCP MATRIX, betrachtet werden (vgl. Abbildungen 4.8 und 4.10).

Bei der Signifikanzprüfung der Merkamlsvariablen folgt auch hier, dass die Streichfähigkeit einen signifikanten Einfluss auf das Ergebnis hat. Für die Diskriminanz von Variable Streichfähigkeit ergibt sich ein Signifikanzniveau (also Irrtumswahrscheinlichkeit) von $0,4\%$ und für Variable Haltbarkeit von $42,06\%$ (vgl. Abbildung 4.19).

Univariate Test Statistics

F Statistics, Num DF=1, Den DF=22

Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
STREICH_F	1.3593	1.1481	1.0607	0.3176	0.4655	10.24	0.0041
HALTBAR_K	1.4818	1.4924	0.3536	0.0297	0.0306	0.67	0.4206

Mit Hilfe Wilks' Lambda wird nun auf die Unterschiedlichkeit der Gruppen getestet. Hierbei ergibt sich ein Wert von $0,52$. Somit ist die Trennkraft der Diskriminanzfunktion hoch. Auch mit dem p-Wert lässt sich argumentieren, dass die beiden Gruppen unterschiedlich sind.

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=9.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.52294557	9.58	2	21	0.0011
Pillai's Trace	0.47705443	9.58	2	21	0.0011
Hotelling-Lawley Trace	0.91224490	9.58	2	21	0.0011
Roy's Greatest Root	0.91224490	9.58	2	21	0.0011

Zuletzt wird noch der kanonische Korrelationskoeffizient angegeben. Dieser beträgt hier $0,691$.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.690691	0.682084	0.109042	0.477054	0.9122		1.0000	1.0000	0.52294557	9.58	2	21	0.0011

Nun wird noch die Prozedur DISCRIM angewendet. Hierbei wird zunächst die CLASS LEVEL INFORMATION geliefert.

Total Sample Size	24	DF Total	23
Variables	2	DF Within Classes	22
Classes	2	DF Between Classes	1

Number of Observations Read	24
Number of Observations Used	24

Class Level Information

MARKE	Variable Name	Frequency	Weight	Proportion	Prior Probability
A	A	12	12.0000	0.500000	0.500000
B	B	12	12.0000	0.500000	0.500000

Anschließend wird eine Confusion Matrix (vgl. Abbildung 4.17) erstellt. Insgesamt wurden 21 (11+10) von 24 Beobachtungen richtig klassifiziert und die Trefferquote liegt bei 87,5%. Es sollte noch beachtet werden, dass SAS die Fehlerquote, hier 12,5%, ausgibt.

Number of Observations and Percent Classified into MARKE

From MARKE	A	B	Total
A	11	1	12
	91.67	8.33	100.00
B	2	10	12
	16.67	83.33	100.00
Total	13	11	24
	54.17	45.83	100.00
Priors	0.5	0.5	

Error Count Estimates for MARKE

	A	B	Total
Rate	0.0833	0.1667	0.1250
Priors	0.5000	0.5000	

Im [KB00] ist nur die oben angegebene Confusion Matrix mit Resubstitutions-Verfahren erläutert worden. Zusätzlich kann auch die Confusion Matrix mit Kreuzvalidierungs-Verfahren ausgegeben werden. Hierbei liefert dies sogar ein schlechteres Ergebnis als mit dem Resubstitutions-Verfahren.

Number of Observations and Percent Classified into MARKE

From MARKE	A	B	Total
A	10	2	12
	83.33	16.67	100.00
B	3	9	12
	25.00	75.00	100.00
Total	13	11	24
	54.17	45.83	100.00
Priors	0.5	0.5	

Error Count Estimates for MARKE

	A	B	Total
Rate	0.1667	0.2500	0.2083
Priors	0.5000	0.5000	

2.5.2 7v B

Aufgabenstellung:

B.

Analysieren Sie die **(DEPRESSions Daten)** aus **Aufgabe 7u** mit Methoden der Diskriminanzanalyse:

(SAS Proc Candisc zur Visualisierung; SAS Proc Discrim zur Analyse).

Sie können ein schrittweises Modell durch SAS aufbauen lassen und bewerten.

Und/oder nach eigenen Überlegungen "ein bestes" Modell konstruieren.

Wo sehen Sie Übereinstimmungen zwischen den Ergebnissen von Logistischer Regression und Diskriminanzanalyse und wo Unterschiede?

Was sind diejenigen via Modell geschätzten Größen, die sich am direktesten vergleichen lassen? Tun Sie dies.

Lösung:

In der Aufgabe 7u wurde schon die DEPRESSions Daten ausführlich mit der logistischen Regression und mit der Diskriminanzanalyse bearbeitet. Hier wird nun zusätzlich bei der logistischen Regression die Confusion Matrix mit Hilfe der Prozedur FREQ ausgegeben und mit der Confusion Matrix der Diskriminanzanalyse verglichen.

Die untenstehende Tabelle zeigt das reduzierte Modell, also nur vier Regressoren, angewendet auf die logistische Regression. Hierbei wurden 243 (238+5) Beobachtungen richtig klassifiziert.

Table of _INTO_ by _FROM_

		FROM(Formatted Value of the Observed Response)		
		0	1	Total
INTO(Formatted Value of the Predicted Response)				
0	Frequency	238	45	283
	Percent	80.95	15.31	96.26
	Row Pct	84.10	15.90	
	Col Pct	97.54	90.00	
1	Frequency	6	5	11
	Percent	2.04	1.70	3.74
	Row Pct	54.55	45.45	
	Col Pct	2.46	10.00	
Total	Frequency	244	50	294
	Percent	82.99	17.01	100.00

Und bei der Diskriminanzanalyse mit reduziertem Modell ergibt die folgende Confusion Matrix. Hier wurde die Resubstitution Methode verwendet. Bei diesem Verfahren wurden 246 (239+7) Beobachtungen korrekt klassifiziert.

Number of Observations and Percent Classified into CASES			
From CASES	0	1	Total
0	239	5	244
	97.95	2.05	100.00
1	43	7	50
	86.00	14.00	100.00
Total	282	12	294
	95.92	4.08	100.00
Priors	0.82993	0.17007	

Error Count Estimates for CASES			
	0	1	Total
Rate	0.0205	0.8600	0.1633
Priors	0.8299	0.1701	

Die binäre logistische Regression weist insbesondere zur Zwei-Gruppen-Diskriminanzanalyse eine recht hohe Ähnlichkeit auf. Der zentrale Unterschied ist jedoch darin zu sehen, dass die logistische Regression im Vergleich zur Diskriminanzanalyse als wesentlich robuster angesehen werden kann, da sie an weniger starke Prämissen geknüpft ist. Die Diskriminanzanalyse setzt zum Beispiel multinormalverteilt unabhängige Variablen sowie gleiche Varianz-Kovarianz-Matrizen in den betrachteten Gruppen voraus, während die logistische Regression diese Voraussetzungen nicht benötigt [AL05]. Laut [HB06] würden viele Forscher selbst wenn alle genannten Anforderungen erfüllt sind die logistische Regression bevorzugen. Dies geschieht, weil sie der multiplen Regression sehr ähnlich ist und hierbei einfache statistische Tests verwendet werden und ähnliche Ansätze anwendet bei der Einbeziehung metrischer und nicht-metrischer Variablen sowie nicht-linearer Effekte.

3 Teil 2

3.1 Aufgabe 8

Aufgabenstellung:

Bearbeiten (Lösen) Sie die Aufgaben 4 , 5, 6 und 7 as is, d.h. noch nicht mit der (neuen) Mining Philosophie von Data Partitioning.

Präsentieren Sie **kurz** Ihre Ergebnisse und ordnen bitte gleichartige oder gleichwertige Ergebnisse zwischen SAS Prozeduren und EMiner Ergebnissen einander zu.

Beleuchten Sie eventuelle Unterschiede.

Anmerkung: Es reicht aus, **eine** der genannten Aufgaben zu bearbeiten, z.B. von

Aufgabe 7_u: (DEPRESSions Daten) den Teil mit Logistischer Regression.

Lösung:

Im Folgenden können die Ergebnisse der SAS Enterprise Miner für die logistische Regression aus **Aufgabe 7u** betrachtet werden. Als erster Schritt werden alle Regressoren bzw. unabhängigen Variablen als metrische Variablen definiert. Dies geschieht, weil der normale SAS PROC LOG Prozedur diese Variablen intern als metrisch deklariert hat und es soll am Ende ein Vergleich zwischen SAS und SAS Enterprise Miner Prozessen gezogen werden. Die Variablen C1 bis C20 und CESD wurden auf REJECTED gesetzt, da diese nicht als Regressoren in das Modell aufgenommen werden.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ACUTEILL	Input	Interval	No		No	.	.
AGE	Input	Interval	No		No	.	.
BEDDAYS	Input	Interval	No		No	.	.
C1	Rejected	Interval	No		No	.	.
C10	Rejected	Interval	No		No	.	.
C11	Rejected	Interval	No		No	.	.
C12	Rejected	Interval	No		No	.	.
C13	Rejected	Interval	No		No	.	.
C14	Rejected	Interval	No		No	.	.
C15	Rejected	Interval	No		No	.	.
C16	Rejected	Interval	No		No	.	.
C17	Rejected	Interval	No		No	.	.
C18	Rejected	Interval	No		No	.	.
C19	Rejected	Interval	No		No	.	.
C2	Rejected	Interval	No		No	.	.
C20	Rejected	Interval	No		No	.	.
C3	Rejected	Interval	No		No	.	.
C4	Rejected	Interval	No		No	.	.
C5	Rejected	Interval	No		No	.	.
C6	Rejected	Interval	No		No	.	.
C7	Rejected	Interval	No		No	.	.
C8	Rejected	Interval	No		No	.	.
C9	Rejected	Interval	No		No	.	.
CASES	Target	Binary	No		No	.	.
CESD	Rejected	Interval	No		No	.	.
CHRONILL	Input	Interval	No		No	.	.
DRINK	Input	Interval	No		No	.	.
EDUCAT	Input	Interval	No		No	.	.
EMPLOY	Input	Interval	No		No	.	.
HEALTH	Input	Interval	No		No	.	.
ID	Rejected	Nominal	No		No	.	.
INCOME	Input	Interval	No		No	.	.
MARITAL	Input	Interval	No		No	.	.
REGDOC	Input	Interval	No		No	.	.
RELIG	Input	Interval	No		No	.	.
SEX	Input	Interval	No		No	.	.
TREAT	Input	Interval	No		No	.	.

Anhand der folgenden Ergebnisse kann man sagen, dass SAS und SAS Enterprise Miner die identischen Lösungen geliefert haben. Der einzige Unterschied ist, dass der Enterprise Miner keine Konfidenzintervalle bezüglich der Odds Ratio Schätzer ausgegeben hat.

Likelihood-Ratio-Test - Globale Nullhypothese: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Quadrat	DF	Pr > ChiSq
268.125	222.499	45.6259	14	<.0001

Analyse Maximum-Likelihood-Schätzer

Parameter	DF	Schätzung	Standard Fehler	Waldsches Chi-Quadrat	Pr > ChiSq	Standardisierter Schätzer	Exp(Est)
Intercept	1	-2.1194	1.6372	1.68	0.1955		0.120
ACUTEILL	1	0.0650	0.4175	0.02	0.8764	0.0164	1.067
AGE	1	-0.0262	0.0134	3.80	0.0512	-0.2609	0.974
BEDDAYS	1	0.9420	0.4312	4.77	0.0289	0.2135	2.565
CHRONILL	1	0.00483	0.4275	0.00	0.9910	0.00133	1.005
DRINK	1	-0.5664	0.4873	1.35	0.2451	-0.1261	0.568
EDUCAT	1	-0.1646	0.1676	0.96	0.3260	-0.1189	0.848
EMPLOY	1	0.1382	0.1251	1.22	0.2694	0.1181	1.148
HEALTH	1	0.3648	0.2339	2.43	0.1188	0.1685	1.440
INCOME	1	-0.0306	0.0152	4.04	0.0443	-0.2582	0.970
MARITAL	1	0.0260	0.1596	0.03	0.8706	0.0181	1.026
REGDOC	1	0.3061	0.4391	0.49	0.4858	0.0659	1.358
RELIG	1	0.3050	0.1389	4.82	0.0281	0.2062	1.357
SEX	1	0.9031	0.4344	4.32	0.0376	0.2418	2.467
TREAT	1	-0.3252	0.4272	0.58	0.4465	-0.0898	0.722

Odds-Ratio-Schätzer

Effekt	Punktschätzwert
ACUTEILL	1.067
AGE	0.974
BEDDAYS	2.565
CHRONILL	1.005
DRINK	0.568
EDUCAT	0.848
EMPLOY	1.148
HEALTH	1.440
INCOME	0.970
MARITAL	1.026
REGDOC	1.358
RELIG	1.357
SEX	2.467
TREAT	0.722

Des Weiteren wird die reduzierte logistische Regression ausgeführt. Hierfür werden von den 14 Regressoren nur die vier signifikanten verwendet und die nicht signifikanten Regressoren auf REJECTED gesetzt.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ACUTEILL	Rejected	Interval	No		No	.	.
AGE	Rejected	Interval	No		No	.	.
BEDDAYS	Input	Interval	No		No	.	.
C1	Rejected	Interval	No		No	.	.
C10	Rejected	Interval	No		No	.	.
C11	Rejected	Interval	No		No	.	.
C12	Rejected	Interval	No		No	.	.
C13	Rejected	Interval	No		No	.	.
C14	Rejected	Interval	No		No	.	.
C15	Rejected	Interval	No		No	.	.
C16	Rejected	Interval	No		No	.	.
C17	Rejected	Interval	No		No	.	.
C18	Rejected	Interval	No		No	.	.
C19	Rejected	Interval	No		No	.	.
C2	Rejected	Interval	No		No	.	.
C20	Rejected	Interval	No		No	.	.
C3	Rejected	Interval	No		No	.	.
C4	Rejected	Interval	No		No	.	.
C5	Rejected	Interval	No		No	.	.
C6	Rejected	Interval	No		No	.	.
C7	Rejected	Interval	No		No	.	.
C8	Rejected	Interval	No		No	.	.
C9	Rejected	Interval	No		No	.	.
CASES	Target	Binary	No		No	.	.
CESD	Rejected	Interval	No		No	.	.
CHRONILL	Rejected	Interval	No		No	.	.
DRINK	Rejected	Interval	No		No	.	.
EDUCAT	Rejected	Interval	No		No	.	.
EMPLOY	Rejected	Interval	No		No	.	.
HEALTH	Rejected	Interval	No		No	.	.
ID	Rejected	Nominal	No		No	.	.
INCOME	Input	Interval	No		No	.	.
MARITAL	Rejected	Interval	No		No	.	.
REGDOC	Rejected	Interval	No		No	.	.
RELIG	Input	Interval	No		No	.	.
SEX	Input	Interval	No		No	.	.
TREAT	Rejected	Interval	No		No	.	.

Auch diesmal liefert SAS Enterprise Miner die selben Ergebnisse wie SAS PROC LOG.

Likelihood-Ratio-Test - Globale Nullhypothese: BETA=0

-2 Log Likelihood Intercept Only	Intercept & Covariates	Likelihood Ratio Chi-Quadrat	DF	Pr > ChiSq
268.125	233.557	34.5676	4	<.0001

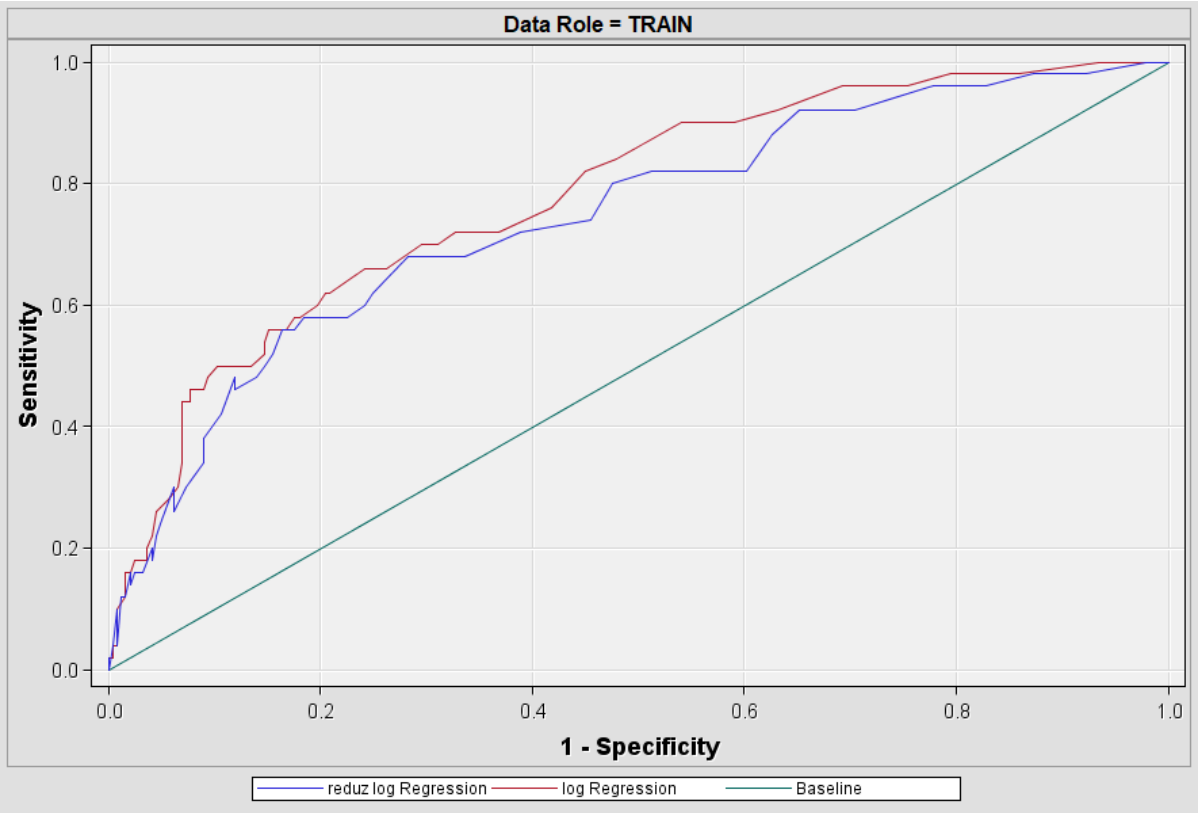
Analyse Maximum-Likelihood-Schätzer

Parameter	DF	Schätzung	Standard Fehler	Waldsches Chi-Quadrat	Pr > ChiSq	Standardisierter Schätzer	Exp(Est)
Intercept	1	-3.6397	0.8899	16.73	<.0001		0.026
BEDDAYS	1	1.2875	0.3530	13.30	0.0003	0.2917	3.624
INCOME	1	-0.0355	0.0138	6.63	0.0100	-0.2992	0.965
RELIG	1	0.3477	0.1300	7.16	0.0075	0.2352	1.416
SEX	1	0.9359	0.4054	5.33	0.0210	0.2506	2.549

Odds-Ratio-Schätzer

Effekt	Punktschätzwert
BEDDAYS	3.624
INCOME	0.965
RELIG	1.416
SEX	2.549

Zuletzt werden die beiden Modelle als ROC-Kurven dargestellt.



Zudem gibt SAS Enterprise Miner noch FIT STATISTICS Tabellen aus. Somit können die beiden Modell zum Beispiel anhand der AVERAGE SQUARED ERROR miteinander verglichen werden.

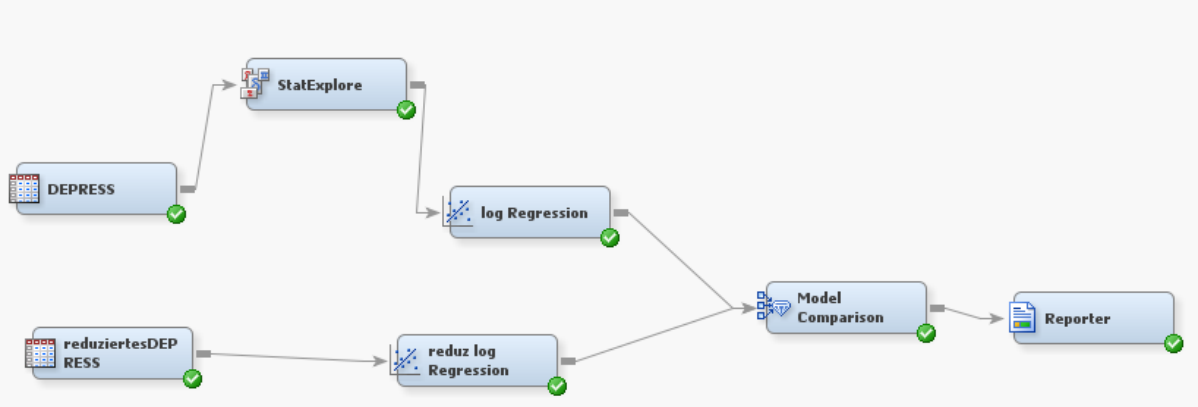
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Reg	Reg	log Regression	CASES	depressed is cesd >= 16	0.159864	252.49...	0.1173...
	Reg2	Reg2	reduz log Regression	CASES	depressed is cesd >= 16	0.159864	243.557	0.1222...

Fit Statistics
Model Selection based on Train: Misclassification Rate (_MISC_)

Selected Model	Model Node	Model Description	Train: Misclassification Rate	Train: Average Squared Error	Train: Roc Index	Train: Gini Coefficient
Y	Reg	log Regression	0.15986	0.11739	0.779	0.559
	Reg2	reduz log Regression	0.15986	0.12227	0.743	0.487

Zum Schluss kann noch das ganze Diagramm der Aufgabe 8 betrachtet werden.



3.2 Aufgabe 9

Aufgabenstellung:

Den Teil, den der Dozent vorführt, relativ knapp. Ihre eigenen Teile ausführlich!

Bringen Sie die Miner Daten in eine bereits definierte SAS Lib oder definieren Sie für diese Daten eine neue | alternative SAS Lib.

Checken Sie es via SAS. Schließen Sie SAS.

Starten Sie den SAS Enterprise Miner Workstation 14.2.

Geben Sie dem Miner den Ort Ihrer Daten und den Namen der zugehörigen SAS Lib bekannt.

In **Aufgabe 9** soll die Datei **ORGANICS** bearbeitet werden.

Lesen und verstehen Sie die Beschreibung der Merkmale.

Schauen Sie sich die Daten von **ORGANICS** (organics | scoreorganics) an.

U. Folgen Sie der Vorgehensweise Ihres Dozenten / dem Getting Started (Mut zur Lücke!). Vervollständigen Sie alle Elemente der zugehörigen ÜBUNG.

V. Bauen Sie die Übung dahingehend aus, dass Sie weitere Analyse Knoten aktivieren. Welche geeignet sind, darüber entscheidet das SKALEN-Niveau der Targets. Ein bekannter heißer Kandidat ist die logistische Regression. Es gibt einige weitere.

Vergleichen und bewerten Sie verschiedene Modelle (mit ROC-, Lift-, Gain- Charts).

Was halten Sie für das beste / ein bestes Modell? Charakterisieren Sie dieses.

Geben Sie an, was die wesentlichen Aussagen dieser Modelle sind.

Formulieren Sie Ihre Erkenntnisse (Was/Warum/Wieso/Wozu?).

Bilden Sie Ihren Finalen Workflow ab.

Setzen Sie versuchsweise im DT Decision Tree / HPSPLIT einmal Maximum Branch von 2 auf 3 oder 4.

Lösung:

Ein Supermarkt bietet eine neue Reihe von Bio-Produkten an. Das Management des Supermarktes möchte bestimmen, welche Kunden diese neue Produkte wahrscheinlich kaufen würden. Dabei ist die Zielvariable bzw. abhängige Variable TargetBuy und die folgenden 8 unabhängige Variablen werden als Regressoren betrachtet:

- DemAffl: Wohlstandsklasse
- DemAge: Alter der Kunden
- DemClusterGroup: Nachbarschaftsgruppe
- DemGender: Geschlecht der Kunden
- DemRegion: Geografische Region
- DemTVReg: Television Region
- PromClass: Treuestatus
- PromSpend: ausgegebene Gesamtbetrag
- PromTime: Zeit als Treuekartenmitglied

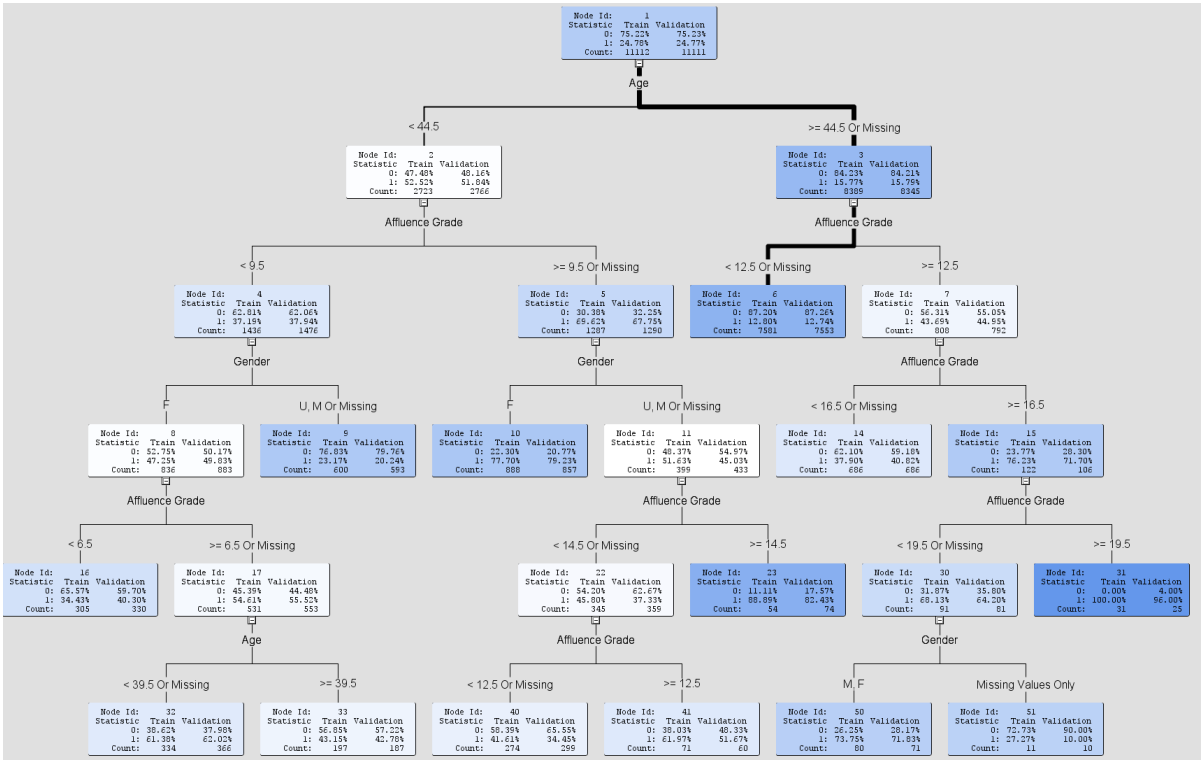
Der ORGANICS-Datensatz enthält insgesamt 13 Variablen und über 22.000 Beobachtungen. Dabei werden diese 50% als Trainings- und 50% als Validierungsdaten aufgeteilt. Die Variablen im Datensatz sind unten (wie in der Vorlesung ausgeteilten PDF) mit den entsprechenden Rollen und Ebenen dargestellt:

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DemAffl	Input	Interval	No		No	.	.
DemAge	Input	Interval	No		No	.	.
DemCluster	Rejected	Nominal	No		No	.	.
DemClusterG	Input	Nominal	No		No	.	.
DemGender	Input	Nominal	No		No	.	.
DemReg	Input	Nominal	No		No	.	.
DemTVReg	Input	Nominal	No		No	.	.
ID	ID	Nominal	No		No	.	.
PromClass	Input	Nominal	No		No	.	.
PromSpend	Input	Interval	No		No	.	.
PromTime	Input	Interval	No		No	.	.
TargetAmt	Rejected	Interval	No		No	.	.
TargetBuy	Target	Binary	No		No	.	.

Zunächst wird mit Hilfe der SAS Enterprise Miner ein Entscheidungsbaum mit zwei als maximale Astzahl (MAXIMUM BRANCH = 2) ausgeführt. Hierbei ergibt eine Tabelle indem die Wichtigkeit der Variablen aufgelistet ist. Es ist zu erkennen, dass zuerst DemAge dann DemAffl und zuletzt DemGender als Trennungskriterium der Decision Tree verwendet werden können.

Variable Importance

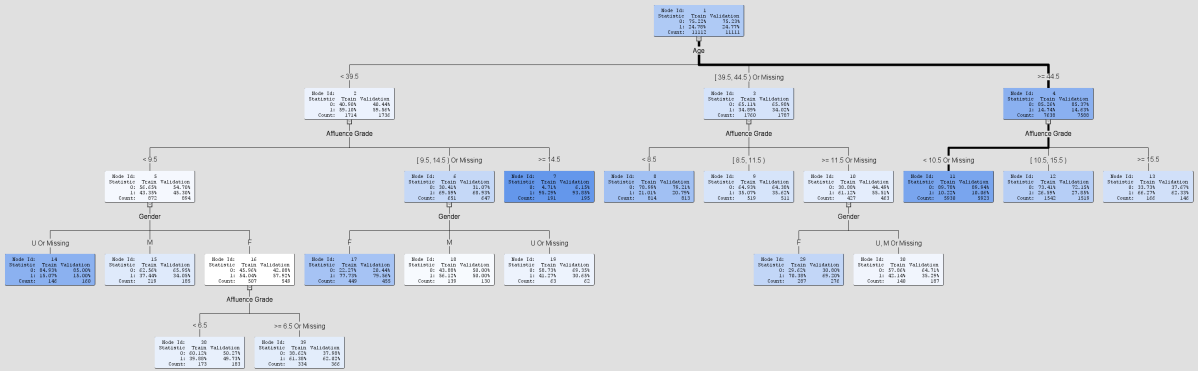
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
DemAge	Age	2	1.0000	1.0000	1.0000
DemAffl	Affluence Grade	7	0.7939	0.7698	0.9696
DemGender	Gender	3	0.3818	0.4874	1.2767



Zusätzlich wird noch ein Decision Tree mit MAXIMUM BRANCH = 3 ausgeführt. Dabei bleibt der VARIABLE IMPORTANCE fast identisch zu der obigen Entscheidungsbaum.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
DemAge	Age	1	1.0000	1.0000	1.0000
DemAffl	Affluence Grade	4	0.7830	0.7110	0.9081
DemGender	Gender	3	0.3497	0.4298	1.2290



Mit Hilfe der INPUTE und TRANSFORM VARIABLES kann nun die schrittweise logistische Regression auf die Daten angewendet werden. Dabei wurden nur die folgenden vier Regressoren als signifikant ermittelt.

Analyse Maximum-Likelihood-Schätzer

Parameter		DF	Schätzung	Standard Fehler	Waldsches Chi-Quadrat	Pr > ChiSq	Standardisierter Schätzer	Exp(Est)
Intercept		1	-1.0240	0.1459	49.25	<.0001		0.359
DemGender	F	1	0.9352	0.0576	263.49	<.0001		2.548
DemGender	M	1	0.0719	0.0631	1.30	0.2542		1.075
IMP_DemAffl		1	0.2433	0.00884	757.72	<.0001	0.4504	1.275
IMP_DemAge		1	-0.0548	0.00221	611.50	<.0001	-0.3867	0.947

Zuletzt werden die drei Modelle mit Hilfe der MODEL COMPARISON miteinander verglichen. Es fällt auf, dass der Entscheidungsbaum mit maximaler Astanzahl von 3 bei Trainings- und Validierungsdaten das bestmögliche AVERAGE SQUARED ERROR liefert. Die ROC-Kurve bestätigt auch, dass der Entscheidungsbaum mit maximal 3 Splits das bestmögliche Modell ist.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

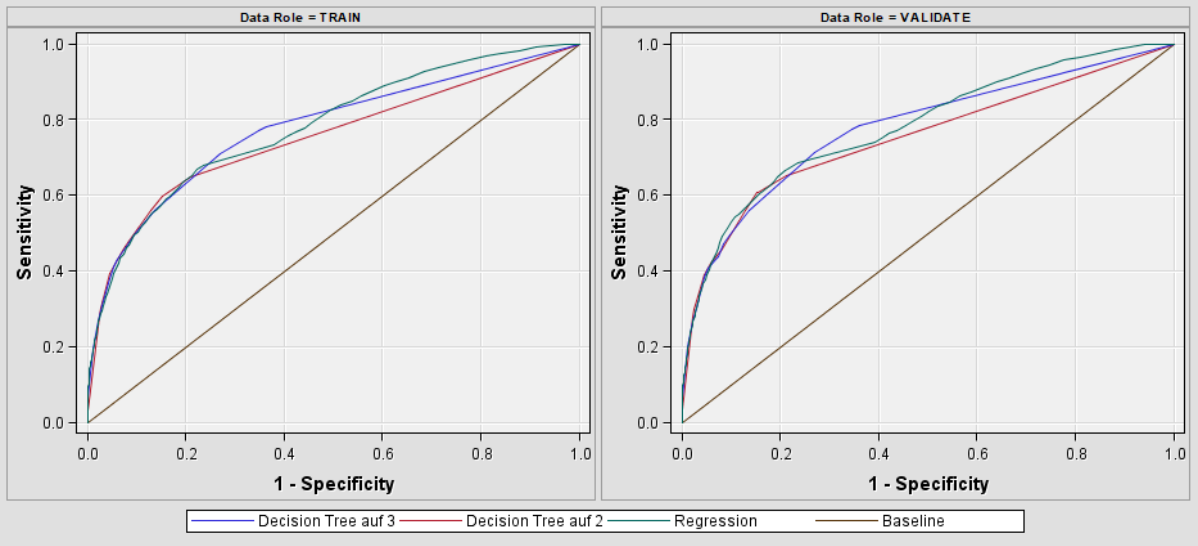
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree auf 2	0.18531	0.14135	0.18512	0.14113
	Tree2	Decision Tree auf 3	0.18738	0.14001	0.18602	0.14020
	Reg	Regression	0.19017	0.14358	0.19150	0.14268

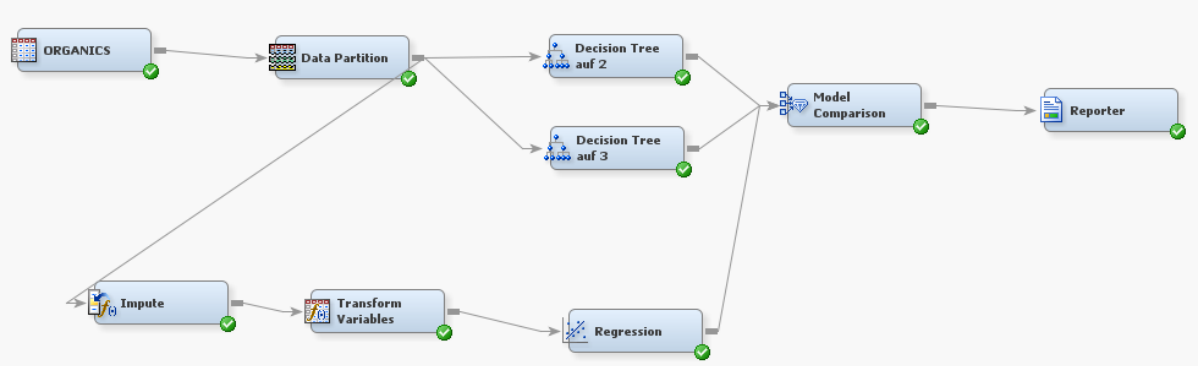
Data Role=Train

Statistics	Tree	Tree2	Reg
Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.29	0.27	0.29
Train: Kolmogorov-Smirnov Statistic	0.44	0.44	0.45
Train: Akaike's Information Criterion	.	.	10052.86
Train: Average Squared Error	0.14	0.14	0.14
Train: Roc Index	0.75	0.78	0.78
Train: Average Error Function	.	.	0.45
Train: Cumulative Percent Captured Response	31.40	31.04	30.72
Train: Percent Captured Response	15.24	14.13	13.48
Selection Criterion: Valid: Misclassification Rate	0.19	0.19	0.19
Train: Degrees of Freedom for Error	.	.	11107.00
Train: Model Degrees of Freedom	.	.	5.00
Train: Total Degrees of Freedom	11112.00	11112.00	11112.00
Train: Divisor for ASE	22224.00	22224.00	22224.00
Train: Error Function	.	.	10042.86
Train: Final Prediction Error	.	.	0.14
Train: Gain	213.82	210.22	206.99
Train: Gini Coefficient	0.50	0.55	0.56
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.44	0.44	0.44
Train: Kolmogorov-Smirnov Probability Cutoff	0.24	0.22	0.28
Train: Cumulative Lift	3.14	3.10	3.07
Train: Lift	3.05	2.82	2.69
Train: Maximum Absolute Error	0.89	0.95	0.99
Train: Misclassification Rate	0.19	0.19	0.19
Train: Mean Square Error	.	.	0.14
Train: Sum of Frequencies	11112.00	11112.00	11112.00
Train: Number of Estimate Weights	.	.	5.00
Train: Root Average Squared Error	0.38	0.37	0.38
Train: Cumulative Percent Response	77.75	76.86	76.06
Train: Percent Response	75.46	69.96	66.75
Train: Root Final Prediction Error	.	.	0.38
Train: Root Mean Squared Error	.	.	0.38
Train: Schwarz's Bayesian Criterion	.	.	10089.44
Train: Sum of Squared Errors	3141.31	3111.65	3190.84
Train: Sum of Case Weights Times Freq	.	.	22224.00

Data Role=Valid

Statistics	Tree	Tree2	Reg
Valid: Kolmogorov-Smirnov Statistic	0.45	0.44	0.46
Valid: Average Squared Error	0.14	0.14	0.14
Valid: Roc Index	0.75	0.78	0.78
Valid: Average Error Function	.	.	0.45
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.29	0.24	0.29
Valid: Cumulative Percent Captured Response	31.30	30.95	30.98
Valid: Percent Captured Response	15.06	13.87	13.78
Valid: Divisor for VASE	22222.00	22222.00	22222.00
Valid: Error Function	.	.	10012.89
Valid: Gain	212.79	209.27	209.51
Valid: Gini Coefficient	0.50	0.56	0.56
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.44	0.44	0.45
Valid: Kolmogorov-Smirnov Probability Cutoff	0.28	0.22	0.29
Valid: Cumulative Lift	3.13	3.09	3.10
Valid: Lift	3.01	2.77	2.75
Valid: Maximum Absolute Error	1.00	0.95	0.99
Valid: Misclassification Rate	0.19	0.19	0.19
Valid: Mean Square Error	.	.	0.14
Valid: Sum of Frequencies	11111.00	11111.00	11111.00
Valid: Root Average Squared Error	0.38	0.37	0.38
Valid: Cumulative Percent Response	77.47	76.60	76.66
Valid: Percent Response	74.54	68.63	68.20
Valid: Root Mean Square Error	.	.	0.38
Valid: Sum of Squared Errors	3136.11	3115.44	3170.58
Valid: Sum of Case Weights Times Freq	.	.	22222.00





3.3 Aufgabe 10

Aufgabenstellung:

(Bearbeitung einer kleinen FALLSTUDIE)

Bringen Sie die Miner Daten in eine bereits definierte SAS Lib oder definieren Sie für diese Daten eine neue | alternative SAS Lib.

[Checken Sie es via SAS . Schließen Sie SAS.]

Starten Sie den SAS Enterprise Miner Workstation 14.2.

Geben Sie dem Miner den Ort Ihrer Daten und den Namen der zugehörigen SAS Lib bekannt.

In **Aufgabe 10** soll die Datei **CREDIT** bearbeitet werden.

Die Zielgröße heißt TARGET .

Verwenden Sie 50% der Daten zum TRAINING und die anderen 50% zur VALIDIERUNG.

Wenn Sie sich sicher fühlen, wählen Sie eine Gruppen-eigene SEED aus (Startzahl für den ZZ-Generator).

Lesen und verstehen Sie die Beschreibung der Merkmale. Schauen Sie sich die Daten von **CREDIT** an.

U. Folgen Sie dem Muster von Aufgabe 9. Erweitern Sie dieses in verschiedene Richtungen! Vervollständigen Sie alle Elemente der zugehörigen Aufgabenstellung.

V. Bauen Sie die Aufgabenstellung dahingehend aus, dass Sie weitere passende Analyse Knoten aktivieren. Welche geeignet sind, darüber entscheidet das SKALEN-Niveau der Targets.

Vergleichen und bewerten Sie verschiedene Modelle (mit ROC-, Lift-, Gain- Charts).

Was halten Sie für das beste / ein bestes Modell? Charakterisieren Sie dieses.

Geben Sie an, was die wesentlichen Aussagen dieser Modelle sind.

Was sind Ihrer Meinung nach die besten Modelle? Warum? Wie sehen diese aus?

Was sagen die Fit Statistics Größen? (Bewertung immer auf der VALIDierungsmenge!). Führen alle Fit Statistics zu demselben CHAMPION-Modell?

Wie können diese Erkenntnisse an BWLer kommuniziert werden? Wo am einfachsten ?

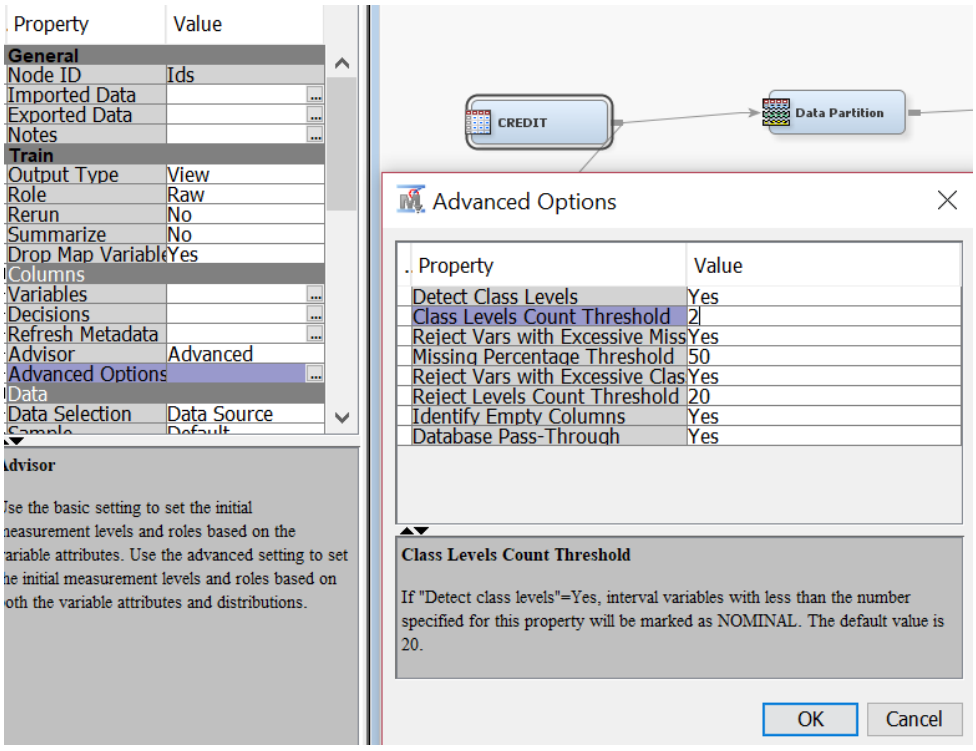
Lösung:

Eine Bank möchte ein neues Kreditprodukt anbieten. Um zukünftige Kreditentscheidungen zu treffen wird ein Risikomodell erstellt. Dabei werden eine Stichprobe von Bewerbern für das ursprüngliche Kreditprodukt als Daten ausgewählt. Es wird nun ein Enterprise Miner Project erstellt, um die besten Vorhersagemodelle zu finden. Zunächst wird die Aufgabe nach den Schritten der in der Vorlesung ausgeteilten PDF ausgearbeitet.

Es folgt ein Ausschnitt der Variablen. Hierbei werden 28 Regressoren für die Vorhersage der abhängigen Variable TARGET verwendet.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
BanruptcyInd	Input	Binary	No		No	.	.
CollectCnt	Input	Interval	No		No	.	.
DerogCnt	Input	Interval	No		No	.	.
ID	ID	Nominal	No		No	.	.
InqCnt06	Input	Interval	No		No	.	.
InqFinanceCr	Input	Interval	No		No	.	.
InqTimeLast	Input	Interval	No		No	.	.
TARGET	Target	Binary	No		No	.	.
TL50UtilCnt	Input	Interval	No		No	.	.
TL75UtilCnt	Input	Interval	No		No	.	.
TLBadCnt24	Input	Interval	No		No	.	.
TLBadDerogC	Input	Interval	No		No	.	.
TLBalHCPct	Input	Interval	No		No	.	.
TLCnt	Input	Interval	No		No	.	.
TLCnt03	Input	Interval	No		No	.	.
TLCnt12	Input	Interval	No		No	.	.
TLCnt24	Input	Interval	No		No	.	.
TLDel3060Cr	Input	Interval	No		No	.	.
TLDel60Cnt	Input	Interval	No		No	.	.
TLDel60Cnt2	Input	Interval	No		No	.	.
TLDel60CntA	Input	Interval	No		No	.	.
TLDel90Cnt2	Input	Interval	No		No	.	.
TLMaxSum	Input	Interval	No		No	.	.
TLOpen24Pct	Input	Interval	No		No	.	.
TLOpenPct	Input	Interval	No		No	.	.
TLSatCnt	Input	Interval	No		No	.	.
TLSatPct	Input	Interval	No		No	.	.
TLSum	Input	Interval	No		No	.	.
TLTimeFirst	Input	Interval	No		No	.	.
TLTimeLast	Input	Interval	No		No	.	.

Die ADVANCED OPTIONS werden wie folgt geändert:



Der StatExplore-Knoten wurde verwendet, um vorläufige Statistiken über die Zielvariable bereitzustellen. Die folgenden zwei Abbildungen und deren Ergebnisse sind auch in der PDF zu finden.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	BanruptcyInd	INPUT	2	0	0	84.67	1	15.33
TRAIN	TARGET	TARGET	2	0	0	83.33	1	16.67

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CollectCnt	INPUT	0.857	2.161352	3000	0	0	0	50	7.556541	111.8365
DerogCnt	INPUT	1.43	2.731469	3000	0	0	0	51	5.045122	50.93801
InqCnt06	INPUT	3.108333	3.479171	3000	0	0	2	40	2.580016	12.82077
InqFinanceCnt24	INPUT	3.555	4.477536	3000	0	0	2	48	2.806893	13.05141
InqTimeLast	INPUT	3.108108	4.637831	2812	188	0	1	24	2.386563	5.626803
TL50UtilCnt	INPUT	4.077904	3.108076	2901	99	0	3	23	1.443077	3.350659
TL75UtilCnt	INPUT	3.121682	2.605435	2901	99	0	3	20	1.50789	3.686636
TLBadCnt24	INPUT	0.567	1.324423	3000	0	0	0	16	4.376858	28.58301
TLBadDerogCnt	INPUT	1.409	2.460434	3000	0	0	0	47	4.580204	48.24276
TLBalHCPct	INPUT	0.648178	0.266486	2959	41	0	0.6955	3.3613	-0.18073	4.015619
TLCnt	INPUT	7.879546	5.421595	2997	3	0	7	40	1.235579	2.195363
TLCnt03	INPUT	0.275	0.582084	3000	0	0	0	7	2.805575	12.66839
TLCnt12	INPUT	1.821333	1.925265	3000	0	0	1	15	1.623636	3.684793
TLCnt24	INPUT	3.882333	3.396714	3000	0	0	3	28	1.60771	4.379948
TLDel3060Cnt24	INPUT	0.726	1.163633	3000	0	0	0	8	1.381942	1.408509
TLDel60Cnt	INPUT	1.522	2.809653	3000	0	0	0	38	3.30846	17.76184
TLDel60Cnt24	INPUT	1.068333	1.806124	3000	0	0	0	20	3.080191	14.35044
TLDel60CntAll	INPUT	2.522	3.407255	3000	0	0	1	45	2.564126	12.70062
TLDel90Cnt24	INPUT	0.814667	1.609508	3000	0	0	0	19	3.623972	19.7006
TLMaxSum	INPUT	31205.9	29092.91	2960	40	0	24187	271036	2.061138	8.093434
TLOpen24Pct	INPUT	0.564219	0.480105	2997	3	0	0.5	6	2.779055	18.5329
TLOpenPct	INPUT	0.496168	0.206722	2997	3	0	0.5	1	0.379339	-0.01934
TLSatCnt	INPUT	13.51168	8.931769	2996	4	0	12	57	0.851193	0.690344
TLSatPct	INPUT	0.518331	0.234759	2996	4	0	0.5263	1	-0.12407	-0.48393
TLSum	INPUT	20151.1	19682.09	2960	40	0	15546	210612	2.276832	10.96413
TLTimeFirst	INPUT	170.1137	92.8137	3000	0	6	151	933	1.031307	2.860035
TLTimeLast	INPUT	11.87367	16.32141	3000	0	0	7	342	6.447907	80.31043

Als erstes wird eine schrittweise logistische Regressionsanalyse durchgeführt. Hierbei werden 50% der Daten für das Training und 50% für die Validierung verwendet.

Übersicht schrittweise Auswahl

Schritt	Effekt Eingegeben	DF	Anzahl ein	Score Chi-Quadrat	Waldsches Chi-Quadrat	Pr > ChiSq
1	TLDel60Cnt24	1	1	87.6784		<.0001
2	IMP_TLBalHCPct	1	2	47.3054		<.0001
3	TLDel3060Cnt24	1	3	46.1590		<.0001
4	IMP_TLSatPct	1	4	18.2931		<.0001
5	InqFinanceCnt24	1	5	17.0563		<.0001
6	TLOpenPct	1	6	15.6759		<.0001
7	TLTimeFirst	1	7	7.6235		0.0058
8	IMP_TL75UtilCnt	1	8	6.6041		0.0102
9	BanruptcyInd	1	9	4.4758		0.0344
10	TLCnt03	1	10	4.5268		0.0334
11	TLOpen24Pct	1	11	4.9244		0.0265

Analyse Maximum-Likelihood-Schätzer

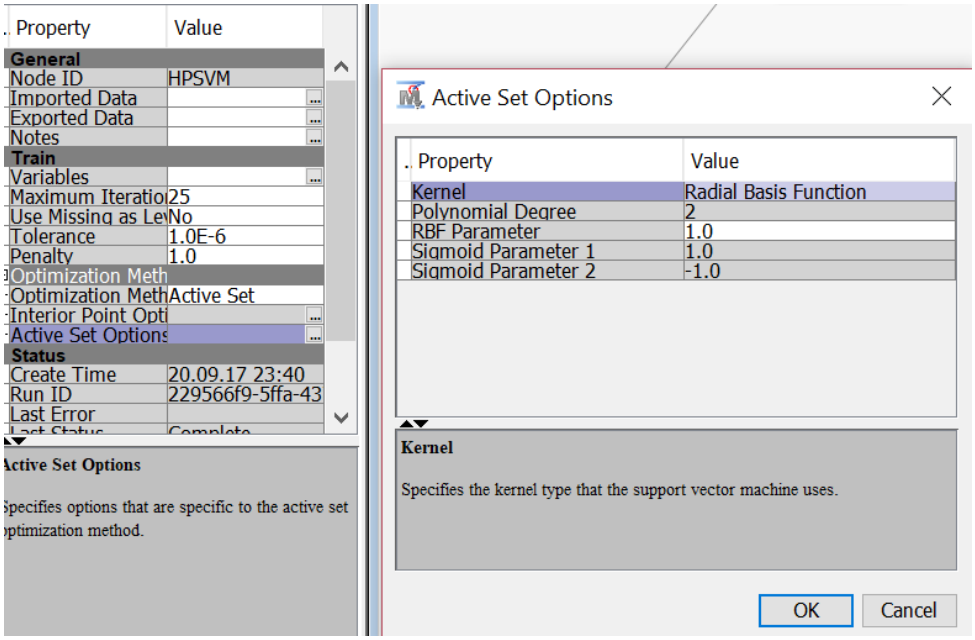
Parameter		DF	Schätzung	Standard Fehler	Waldsches Chi-Quadrat	Pr > ChiSq	Standardisierter Schätzer	Exp(Est)
Intercept		1	-2.7983	0.4470	39.19	<.0001		0.061
BanruptcyInd	0	1	0.2656	0.1172	5.13	0.0235		1.304
IMP_TL75UtilCnt		1	0.0942	0.0318	8.79	0.0030	0.1420	1.099
IMP_TLBalHCPct		1	1.5836	0.3456	21.00	<.0001	0.2340	4.872
IMP_TLSatPct		1	-3.3490	0.4992	45.01	<.0001	-0.4316	0.035
InqFinanceCnt24		1	0.0576	0.0157	13.38	0.0003	0.1441	1.059
TLCnt03		1	-0.3939	0.1613	5.96	0.0146	-0.1163	0.674
TLDel3060Cnt24		1	0.2639	0.0653	16.31	<.0001	0.1656	1.302
TLDel60Cnt24		1	0.1167	0.0426	7.51	0.0061	0.1143	1.124
TLOpen24Pct		1	0.3747	0.1697	4.87	0.0273	0.1009	1.454
TLOpenPct		1	1.6519	0.4983	10.99	0.0009	0.1887	5.217
TLTimeFirst		1	-0.00259	0.000975	7.05	0.0079	-0.1343	0.997

Odds-Ratio-Schätzer

Effekt	Punktschätzwert
BanruptcyInd 0 vs 1	1.701
IMP_TL75UtilCnt	1.099
IMP_TLBalHCPct	4.872
IMP_TLSatPct	0.035
InqFinanceCnt24	1.059
TLCnt03	0.674
TLDel3060Cnt24	1.302
TLDel60Cnt24	1.124
TLOpen24Pct	1.454
TLOpenPct	5.217
TLTimeFirst	0.997

Es stellt sich heraus, dass nur 11 der Regressoren einen signifikanten Einfluss auf die abhängige Variable TARGET haben.

Nun wird die Vorhersage mit Hilfe der Support Vector Machine (SAS Enterprise Miner: HP SVM) ermittelt. Support Vector Machines sind, wie schon in der Einleitung erwähnt, leistungsstarke maschinelle Lerntechniken für die Klassifizierung und Regression. Hierbei kann man mit INTERIOR POINT und mit ACTIVE SET arbeiten. Näheres kann in [SA10] nachgelesen werden. Nachdem HPSVM-Knoten mit verschiedene Möglichkeiten (INTERIOR POINT und ACTIVE SET mit verschiedene Kernfunktionen) ergab ACTIVE SET mit radialer Basisfunktion das bestmögliche Vorhersagemodell. Daher werden auch nur die Ergebnisse der SVM mit radialer Basisfunktion als Kern untersucht.



Zunächst wird ein Klassifizierungsmatrix angegeben. Hierbei sollte dies mit Vorsicht betrachtet werden, da die Spalten und Zeilen (Beobachtet = True Class und Training-Prognose = Hypothesized Class) vertauscht sind. Zudem werden viele Statistiken wie zum Beispiel AVERAGE SQUARED ERROR ausgegeben.

Klassifizierungsmatrix

	Training-Prognose			Validierungs-Prognose		
Beobachtet	1	0	Summe	1	0	Summe
1	0	250	250	0	249	249
0	0	1250	1250	0	1248	1248
Total	0	1500	1500	0	1497	1497

Fit Statistics

Target=TARGET Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.14	0.15
DIV	Divisor for ASE	3000.00	3000.00
MAX	Maximum Absolute Error	0.98	1.00
NOBS	Sum of Frequencies	1500.00	1500.00
RASE	Root Average Squared Error	0.38	0.38
SSE	Sum of Squared Errors	428.99	440.69
DISF	Frequency of Classified Cases	1500.00	1500.00
MISC	Misclassification Rate	0.17	0.17
WRONG	Number of Wrong Classifications	250.00	250.00

Nun werden die beiden Modelle miteinander verglichen.

Fit Statistics Table
Target: TARGET

Data Role=Train

Statistics	Reg	HPSVM
Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.24	0.05
Train: Kolmogorov-Smirnov Statistic	0.43	0.57
Train: Akaike's Information Criterion	1134.91	.
Train: Average Squared Error	0.11	0.14
Train: Roc Index	0.79	0.82
Train: Average Error Function	0.37	.
Train: Cumulative Percent Captured Response	34.00	42.00
Train: Percent Captured Response	15.60	18.80
Selection Criterion: Valid: Misclassification Rate	0.16	0.17
Train: Degrees of Freedom for Error	1488.00	.
Train: Model Degrees of Freedom	12.00	.
Train: Total Degrees of Freedom	1500.00	.
Train: Frequency of Classified Cases	.	1500.00
Train: Divisor for ASE	3000.00	3000.00
Train: Error Function	1110.91	.
Train: Final Prediction Error	0.12	.
Train: Gain	240.00	320.00
Train: Gini Coefficient	0.58	0.64
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.43	0.58
Train: Kolmogorov-Smirnov Probability Cutoff	0.22	0.05
Train: Cumulative Lift	3.40	4.20
Train: Lift	3.12	3.76
Train: Maximum Absolute Error	0.98	0.98
Train: Misclassification Rate	0.15	0.17
Train: Mean Square Error	0.11	.
Train: Sum of Frequencies	1500.00	1500.00
Train: Number of Estimate Weights	12.00	.
Train: Root Average Sum of Squares	0.34	0.38
Train: Cumulative Percent Response	56.67	70.00
Train: Percent Response	52.00	62.67
Train: Root Final Prediction Error	0.34	.
Train: Root Mean Squared Error	0.34	.
Train: Schwarz's Bayesian Criterion	1198.67	.
Train: Sum of Squared Errors	341.41	428.99
Train: Sum of Case Weights Times Freq	3000.00	.
Train: Number of Wrong Classifications	.	250.00

Data Role=Valid

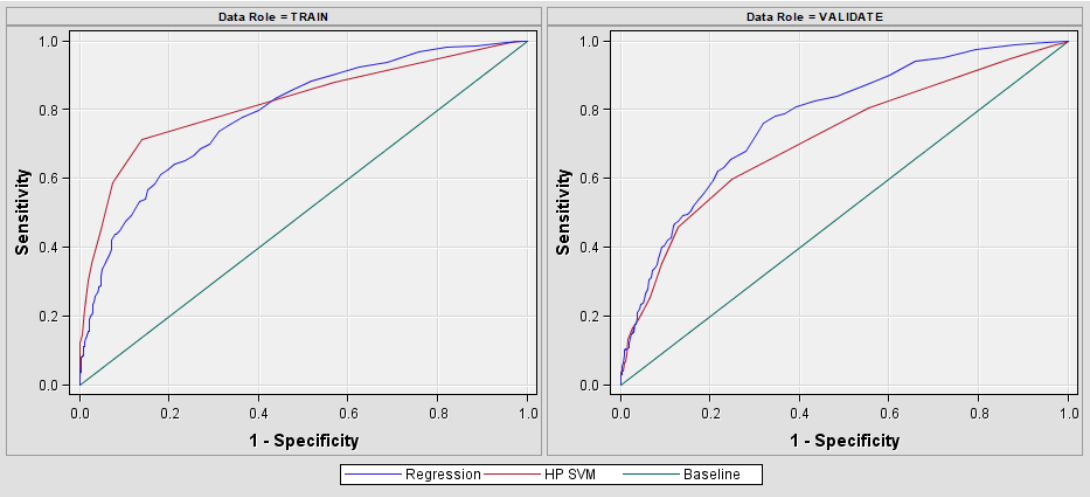
Statistics	Reg	HPSVM
Valid: Kolmogorov-Smirnov Statistic	0.44	0.35
Valid: Average Squared Error	0.12	0.15
Valid: Roc Index	0.77	0.71
Valid: Average Error Function	0.38	.
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.15	0.05
Valid: Cumulative Percent Captured Response	28.40	25.20
Valid: Percent Captured Response	13.60	8.80
Valid: Frequency of Classified Cases	.	1500.00
Valid: Divisor for VASE	3000.00	3000.00
Valid: Error Function	1143.41	.
Valid: Gain	184.00	152.00
Valid: Gini Coefficient	0.55	0.43
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.44	0.36
Valid: Kolmogorov-Smirnov Probability Cutoff	0.14	0.05
Valid: Cumulative Lift	2.84	2.52
Valid: Lift	2.72	1.76
Valid: Maximum Absolute Error	0.98	1.00
Valid: Misclassification Rate	0.16	0.17
Valid: Mean Square Error	0.12	.
Valid: Sum of Frequencies	1500.00	1500.00
Valid: Root Average Squared Error	0.34	0.38
Valid: Cumulative Percent Response	47.33	42.00
Valid: Percent Response	45.33	29.33
Valid: Root Mean Square Error	0.34	.
Valid: Sum of Square Errors	355.85	440.69
Valid: Sum of Case Weights Times Freq	3000.00	.
Valid: Number of Wrong Classifications	.	250.00

Unten ist das CLASSIFICATION TABLE abgebildet. Dabei werden direkt die True Negativ, True Positiv, False Positive und False Negative ausgegeben.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Regression	TRAIN	TARGET		200	1219	31	50
Reg	Regression	VALIDATE	TARGET		206	1210	40	44
HPSVM	HP SVM	TRAIN	TARGET		250	1250	0	0
HPSVM	HP SVM	VALIDATE	TARGET		250	1250	0	0

Ferner werden die ROC-Kurven beider Modelle abgebildet.

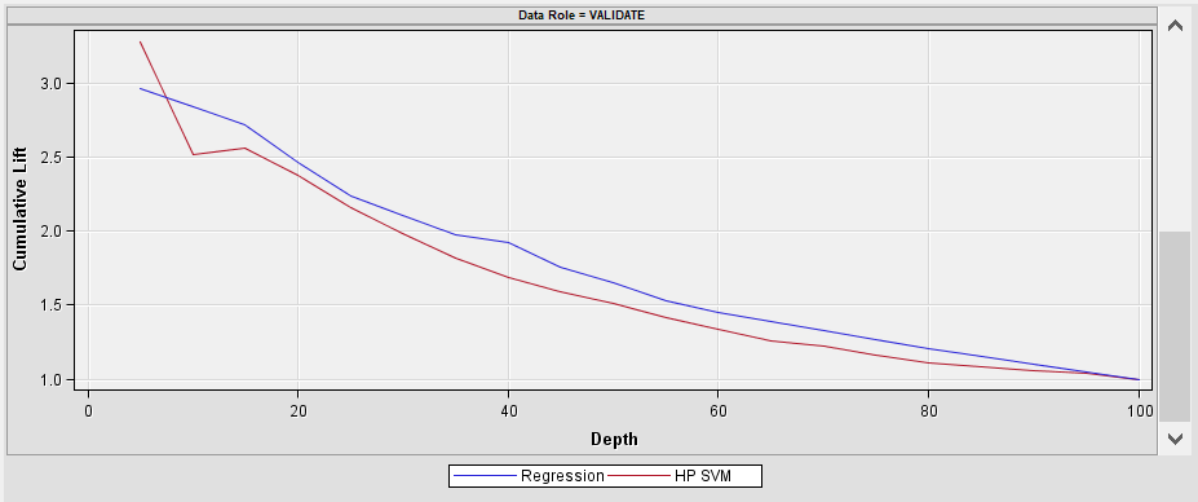


Laut der Aufgabenstellung soll mit der Validierungsdaten argumentiert werden. Hierbei ergab, dass der Support Vector Machines einen höheren AVERAGE SQUARED ERROR hatte als bei der schrittweisen logistischen Regression. Zusätzlich erkennt man anhand der ROC-Index und ROC-Kurve, dass der schrittweise logistische Regression besser als Vorhersagemodell für die Kreditwürdigkeit passt als SVM.

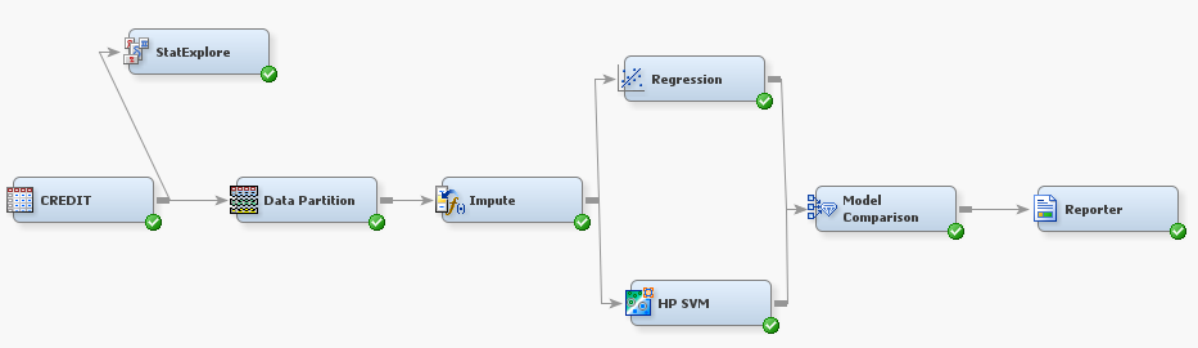
Lift und Gain Charts visualisieren wie gut ein Vorhersagemodell ist. Da in der Vorlesung vermehrt die kumulativen Lift Charts betrachtet wurde, werden auch hier nur diese interpretiert. Ferner werden auch hier nur die Validierungsergebnisse der Modelle abgebildet. Die Frage ist wie viele Kunden müssten mit der neuen Kreditprodukt bekannt gemacht werden, damit diese sich dafür interessieren. Der kumulative Lift Chart zeigt visuell den Vorteil der Verwendung eines der Vorhersagemodelle und hilft bei der Frage, wie viel wahrscheinlicher kann ein Kunde auf das neue Kreditprodukt aufmerksam gemacht werden im Vergleich einer zufällige Kontaktaufnahme der Kunden.

Es ist hier zu erkennen, dass bei beiden Modellen etwa 50% der möglichen Neukunden erreicht wird, wenn die Top 20% der Kunden kontaktiert werden - da ein Lift von etwa 2,5 herrscht.

Des Weiteren kann anhand der Grafik auch gesagt werden, dass bei der schrittweisen logistischen Regression mehr möglichen Neukunden erreicht werden können.



Zum Schluss folgt die Abbildung der Aufgabe 10 im Ganzen.



Literaturverzeichnis

[AL05] Alexander Linder: Web Mining - Die Fallstudie Swarovski: Theoretische Grundlagen und praktische Anwendung, Springer-Verlag, 2005

[HB06] J.F. Hair, W.C. Black: Multivariate data analysis, Prentice Hall, 2006

[HP06] Helmut Pruscha: Statistisches Methodenbuch - Verfahren, Fallstudien, Programmcodes, Springer-Verlag, 2006

[KB00] Klaus Backhaus, Bernd Erichson, Wulff Plinke und Rolf Weiber: Multivariate Analysemethoden, Springer-Verlag, 2000

[SA10] Shigeo Abe: Support Vector Machines for Pattern Classification, Springer Science & Business Media, 2010

[TB14] Tim vor der Brück: Wissensakquisition mithilfe maschineller Lernverfahren auf tiefen semantischen Repräsentationen, Springer-Verlag, 2014

[TF06] Tom Fawcett: An introduction to ROC analysis, Pattern Recognition Letters 27, 2006

[UB08] Udo Bankhofer und Jürgen Vogel: Datenanalyse und Statistik: Eine Einführung für Ökonomen im Bachelor, Springer-Verlag, 2008

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Soweit ich auf fremde Materialien, Texte oder Gedankengänge zurückgegriffen habe, enthalten meine Ausführungen vollständige und eindeutige Verweise auf die Urheber und Quellen. Alle weiteren Inhalte der vorgelegten Arbeit stammen von mir im urheberrechtlichen Sinn, sowie keine Verweise und Zitate erfolgen. Mit ist bekannt, dass ein Täuschungsversuch vorliegt, wenn die vorstehende Erklärung sich als unrichtig erweist.

Darmstadt, den 22. September 2017

Unterschrift