Abschlussarbeit zur Erlangung des
akademischen Grades

Master of Science

Interpretierbares Machine-Learning. Post-hoc modellagnostische Verfahren zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen

---

Christian Alexander Scholbeck

betreut von

Prof. Dr. Christian Heumann

Giuseppe Casalicchio

Ludwig-Maximilians-Universität München

Fakultät für Mathematik, Informatik und Statistik

Institut für Statistik

November 2018

Abstract

Machine-Learning-Modelle besitzen ein sehr viel höheres prädiktives Potential als herkömmliche statistische Verfahren. Aufgrund ihrer mangelnden Interpretierbarkeit wird jedoch häufig auf traditionelle Modellierungsansätze zurückgegriffen. Der Ansatz des Interpretierbaren Machine-Learnings eröffnet neuartige Möglichkeiten, das Verhalten von Machine-Learning-Modellen zu analysieren und zu interpretieren. Die vorliegende Arbeit stellt etablierte post-hoc modellagnostische Verfahren zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen vor. Dabei werden Zusammenhänge und Parallelen zwischen marginalen Effekten, der Individual Conditional Expectation & Partial Dependence, sowie Accumulated Local Effects aufgezeigt und diskutiert. Wir entwickeln ein generalisiertes System zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen, das jedes Verfahren auf gemeinsame Arbeitsschritte reduziert. Anschließend erweitern wir das generalisierte System auf die Bestimmung der Feature-Importance. Zum Abschluss werden die vorgestellten Verfahren auf einem realen Datensatz demonstriert.

Inhaltsverzeichnis

Prolog

„No computer has ever been designed that is aware of what it's doing. But most of the time, we aren't either.“

Marvin Minsky

In der Originalfassung von *Perceptrons* beschrieb Minsky (1969) noch eine einzige simple Variante von Machine-Learning-Modellen, das von Rosenblatt (1958) vorgestellte Perzeptron. Seit der Veröffentlichung ist die Anzahl verfügbarer Algorithmen des Machine-Learnings exorbitant gewachsen. Ein Ende dieser Entwicklung ist nicht absehbar. Machine-Learning bildet gegenwärtig die Basis für die Entwicklung künstlicher Intelligenzen [KI]. Das *Paradoxon von Moravec* (Moravec, 1988) verkörpert die zentrale Problematik der KI-Entwicklung. So seien diejenigen Aufgaben für Maschinen besonders einfach zu lösen, die von erwachsenen Menschen als schwer erachtet werden, wie die Lösung algebraischer Gleichungen. Es sei jedoch schwer oder unmöglich, einer Maschine beispielsweise die kognitiven Fähigkeiten eines Kleinkindes zu verleihen. Pinker (2003) schreibt: „The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems ever conceived... As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines."

Machine-Learning war in der Menschheitsgeschichte der erste Lichtblick, um das scheinbare Paradoxon von Moravec aufzulösen. Maschinen sind gegenwärtig in der Lage, selbstständig zu zeichnen, Fahrzeuge zu führen, oder sich in Sprache und Schrift zu verständigen. Die Basis für derartige Algorithmen ist die Auswertung massiver Datenmengen, aus denen die Maschine lernen kann. Das menschliche Gehirn ist nicht mehr in der Lage, die Verarbeitung dieser massiven Datenmengen nachzuvollziehen. Bisweilen mussten wir den Entscheidungen der Maschine Glauben schenken, bzw. die Präzision ihrer Entscheidungen auf neuen Daten validieren. Dies führte zur Tendenz, den entwickelten Verfahren außerhalb theoretischer Betrachtungen mit Misstrauen zu entgegnen. Der neuartige Ansatz des *Interpretierbaren Machine-Learnings* stellt Methoden bereit, mit Hilfe derer die Entscheidungen eines Machine-Learning-Modells nachvollzogen werden können.

Die vorliegende Arbeit liefert einen Einblick in die Theorie des Interpretierbaren Machine-Learnings. Im ersten Kapitel werden konzeptionelle Hintergründe erklärt. Wir legen den Fokus auf post-hoc modellagnostische Verfahren zur Bestimmung von Prädiktoreffekten. Diese werden in Kapitel 2 im Detail behandelt. Zum Anschluss wird in Kapitel 3 ein *generalisiertes System zur Bestimmung von Prädiktoreffekten* entwickelt, in das die vorgestellten Verfahren eingebettet werden können. Das System wird anschließend auf die Bestimmung der Feature-Importance erweitert. Im vierten Kapitel wenden wir die vorgestellte Theorie auf einen realen Datensatz an. Im abschließenden fünften Kapitel wird das vorgestellte Material diskutiert und ein Ausblick in die Zukunft des Machine-Learnings und des Interpretierbaren Machine-Learnings gegeben.

KAPITEL 1.

Interpretierbares Machine-Learning

„If you can't describe what you're doing as a process, you don't know what you're doing."

W. Edwards Deming

Interpretierbare Modelle existieren in der statistischen Wissenschaft bereits sehr lange. Diese leiden jedoch unter diversen Schwächen. Viele traditionelle Modellierungsansätze sind nicht für große Datenmengen geeignet. Des Weiteren werden sie der Nichtlinearität vieler Datensätze nicht gerecht und bieten keine ausreichende Vorhersagekraft. Viele Aussagen zur Modellinterpretation sind zudem nur asymptotisch und unter bestimmten Annahmen valide. In der Realität sind Annahmen über datengenerierende Prozesse grundsätzlich anzuzweifeln. Machine-Learning stellt einen Ansatz dar, prädiktive Modelle zu konstruieren, die auf keinen Annahmen über den datengenerierenden Prozess beruhen.

1.1. Statistik und Machine-Learning

Definition 1.1.1 (Machine-Learning)**.** *Machine-Learning [ML] erforscht die Konstruktion von Algorithmen, die aus Daten lernen und Vorhersagen tätigen können (Kohavi, 1998).*

Definition 1.1.2 (Machine-Learning-Algorithmus)**.** *Für die Terminologie des Algorithmus existiert eine Vielzahl an Definitionen. Markov (1957) definiert einen Algorithmus im mathematischen Kontext als die exakte Vorschrift eines computationalen Prozesses, der von einer Vielzahl an initialen Daten zum gewünschten Ergebnis führt. Ein Machine-Learning-Algorithmus ist weiterführend ein Algorithmus, der „lernt", aus einer Vielzahl an initialen Daten über diese hinaus zu generalisieren (Kohavi, 1998). Die Ausgabe des ML-Algorithmus besteht aus einem Modell, das den Zusammenhang der Eingangsdaten „gelernt" hat und als trainiertes Modell (Trained Model) bezeichnet wird. Aufgrund des Lernprozesses werden ML-Algorithmen auch Learner oder Inducer genannt (Kohavi, 1998).*

Breiman (2001b) präsentiert einen Überblick über die Gemeinsamkeiten und Gegensätze der statistischen Wissenschaft und des Machine-Learnings. Die Statistik verfolge eine Kultur der Datenmodellierung, während Machine-Learning eine Kultur der algorithmischen Modellierung verfolge. Breiman (2001b) führt drei Negativpunkte an, die aus dem Festhalten an der datenmodellierenden Kultur der statistischen Wissenschaft resultieren:

(1) Sie führe zu irrelevanter Theorie und falschen wissenschaftlichen Schlussfolgerungen.

(2) Statistiker würden davon abgehalten, angemessenere algorithmische Modelle zu verwenden.

(3) Interessante Probleme würden vom Tätigkeitsbereich von Statistikern ferngehalten.

Das Modell repräsentiert die Ausgabe des Algorithmus. Es kann beispielsweise aus einer Menge an Koeffizienten eines linearen Modells oder an Gewichten eines neuronalen Netzes bestehen (Molnar, 2018). Als Eingaben des Algorithmus fungieren die Daten, sowie Hyperparameter, die als Stellschrauben dienen, um die Leistung und Funktionsweise des Algorithmus zu verändern und zu optimieren. ML-Modelle besitzen das Potential einer erheblich erhöhten Vorhersagekraft im Vergleich zu herkömmlichen statistischen Ansätzen. Die Verwendung von ML-Algorithmen scheitert jedoch bisweilen hauptsächlich an der fehlenden *Interpretierbarkeit* des ausgegebenen Modells.

Definition 1.1.3 (Interpretierbarkeit)**.** *Doshi-Velez und Kim (2017) definieren den Vorgang des Interpretierens als das Erklären oder Darstellen in verständlichen Begriffen. Im Machine-Learning-Kontext definieren sie die Interpretierbarkeit als die Erklärung oder Darstellung in für einen Menschen verständlichen Begriffen.*

Doshi-Velez und Kim (2017) gehen detailliert auf etwaige psychologische und philosphische Hintergründe der Diskussion ein. Auf eine formale und einheitliche Definition stoßen sie nicht. Für die vorliegende Arbeit genügt eine ad-hoc-Definition:

Definition 1.1.4 (Interpretierbarkeit eines Modells)**.** *Die Interpretierbarkeit eines Modells ist sichergestellt, wenn der menschliche Anwender nachvollziehen kann, wie das Modell zu einer Entscheidung bzw. Vorhersage gekommen ist.*

1.2. Zur Relevanz des Interpretierbaren Machine-Learnings

Weshalb sollte die Interpretierbarkeit eines Modells gewünscht sein? In manchen Fällen ist die bestmögliche Vorhersagekraft des Modells tatsächlich die einzige Zielsetzung. Doshi-Velez und Kim (2017) geben zwei Gründe hierfür an. Einerseits ist die Erklärung der Entscheidungsfindung eines Modells irrelevant, wenn es keine signifikanten Konsequenzen für die Modellentscheidung gibt. Andererseits kann das zugrundeliegende Problem in realen Applikationen ausreichend gut verstanden sein, sodass der Entscheidung des Modells grundsätzlich vertraut werden kann. Doshi-Velez und Kim (2017) argumentieren, dass die Interpretierbarkeit eines Modells genau dann gewünscht ist, wenn die Formalisierung des Problems nicht vollständig ist und somit eine fundamentale Barriere zwischen Optimierung und Evaluierung des Modells liegt.

1.3. Gruppierung der Verfahren

Verfahren zur Interpretation von ML-Modellen lassen sich konsekutiv gruppieren. Es können vier Fragestellungen abgeleitet werden, um zu einer Verfahrensart zu gelangen:

(1) Intrinsische oder post-hoc Interpretierbarkeit.

(2) Modellspezifische oder modellagnostische Verfahren.

(3) Lokale oder globale Interpretation.

(4) Feature-Effects oder Feature-Importance.

Die Menge der zur Verfügung stehenden Modelle wird auf *Überwachtes Lernen (Supervised-Learning)* beschränkt.

Definition 1.3.1 (Supervised-Learning)**.** *Supervised-Learning wird als diejenige Teilmenge des Machine-Learnings definiert, deren Lernprozess auf der Grundlage von Eingabe-Ausgabe Paaren in den Trainingsdaten beruht. Hierzu sind Trainingsdaten notwendig, die mit den wahren Werten der Zielvariable ausgestattet sind (Labeled Data).*

1.3.1. Intrinsische oder post-hoc Interpretierbarkeit

Wird ausschließlich die bestmögliche Vorhersage durch das konstruierte Modell gewünscht, befinden wir uns im Bereich des klassischen Machine-Learnings. Nach der Wahl eines Algorithmus wird ein Modell an die Trainingsdaten angepasst, das über die Veränderung von Hyperparametern optimiert werden kann (*Tuning*). Wird die Interpretierbarkeit des angepassten Modells gewünscht, kann auf intrinsisch interpretierbare Modelle zurückgegriffen werden. Hier befinden wir uns zunächst im Bereich der klassischen Statistik. Das *allgemeine lineare Modell*, *generalisierte lineare Modelle*, oder *generalisierte additive Modelle* besitzen interpretierbare Koeffizienten. Den klassischen Fall stellt das in vielen Wissenschaften verwendete *multiple lineare Regressionsmodell* dar. Hier erlauben die geschätzten Koeffizienten Einsicht in die Arbeitsweise des Modells. Über Verteilungsannahmen können zusätzlich Unsicherheitsmaße der geschätzten Koeffizienten wie p-Werte und Konfidenzintervalle konstruiert werden. Darüber hinaus existieren intrinsisch interpretierbare *Klassifikations-/ und Regressionsbäume (Classification and Regression Trees [CART])*, die den Datenraum rekursiv partitionieren.

Ist eine bestmögliche Vorhersage in Kombination mit der Interpretation des Modells gefragt, kann auf *post-hoc* Verfahren zurückgegriffen werden. Der Interpretationsprozess geschieht hierbei im Anschluss an den Modellanpassungsprozess. Das verwendete Modell selbst ist üblicherweise nicht intrinsisch interpretierbar.

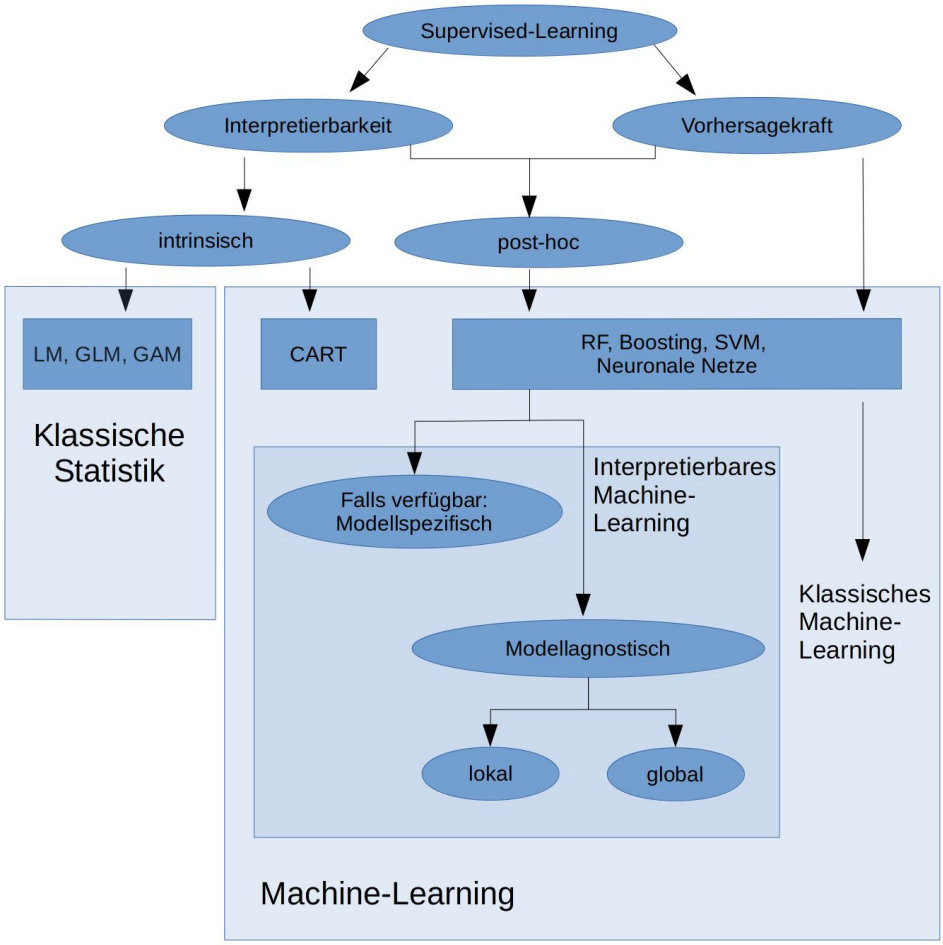


Abbildung 1.3.1.: Entscheidungsflussdiagramm zur Findung einer anwendungsgerechten Verfahrensart.

1.3.2. Modellspezifische oder modellagnostische Verfahren

Bei Anwendung eines post-hoc Verfahrens hängt die weitere Vorgehensweise vom gewählten Algorithmus ab. Für manche Algorithmen existieren *modellspezifische* Interpretationsmethoden. Diese können ausschließlich auf bestimmte Modelle angewandt werden. Für neuronale Netze existieren beispielsweise modellspezifische Interpretations-/ und Visualisierungsmöglichkeiten, wie von Zeiler und Fergus (2013) beschrieben.

Falls kein modellspezifisches Verfahren existiert, kann auf *modellagnostische* Verfahren zurückgegriffen werden. Diese sind auf alle Supervised-Learning-Modelle anwendbar und stellen den Großteil der zur Verfügung stehenden Methoden dar.

Beispiele hierfür sind die in Kapitel 2 vorgestellten *Marginal Effects*, die *Individual Conditional Expectation & Partial Dependence*, sowie *Accumulated Local Effects*. Modellagnostische Verfahren eignen sich darüber hinaus zum *Modellvergleich*.

1.3.3. Lokale oder globale Interpretation

Post-hoc modellagnostische Interpretationsverfahren erfordern schließlich die Entscheidung zwischen lokalen oder globalen Betrachtungen. Eine strikt lokale Interpretation erklärt die Vorhersage einzelner Datenpunkte. Globale Betrachtungen interpretieren die Funktionsweise des Modells für den gesamten Datensatz. Es ist weiterhin möglich, Aussagen über Gruppierungen von Beobachtungen zu treffen, beispielsweise auf Intervallen. Häufig können lokale und globale Interpretationsverfahren über eine (Dis-)Aggregation ineinander überführt werden, wie beispielsweise bei der *Individual Conditional Expectation* und der *Partial Dependence*.

1.3.4. Feature-Effects oder Feature-Importance

Darüber hinaus erfolgt die Interpretation des Modells über zwei Zielsetzungen. Einerseits können Effektrichtungen und -/größen (*Feature-Effects*) eruiert werden. Hierbei wird bestimmt, wie sich die Änderung der Werte einer Prädiktorvariable auf die Vorhersage der Zielvariable auswirkt. Andererseits ist die Bestimmung der *Feature-Importance* möglich. Die Importance ist ein Indikator für den Beitrag eines Prädiktors zur *Modellgüte (Performance)*.

1.4. Eingrenzung der Verfahren

Die Fülle an zur Verfügung stehenden Methoden erfordert die Konzentration auf eine Teilmenge. Der Fokus der vorliegenden Arbeit liegt auf der Bestimmung von Effektrichtungen und -/größen in nicht-intrinsisch interpretierbaren Supervised-Learning-Modellen. Der Interpretationsprozess geschieht post-hoc und soll auf jede Modellart anwendbar sein (modellagnostisch). Aufgrund der Interdependenz von lokalen und globalen Effekten werden beide Varianten betrachtet.

KAPITEL 2.

Post-hoc modellagnostische Verfahren zur Bestimmung lokaler und globaler Prädiktoreffekte

„Truth ... is much too complicated to allow anything but approximations."

John von Neumann

Den Idealfall für die Bestimmung von Prädiktoreffekten stellt das vollständig und korrekt spezifizierte allgemeine lineare Modell (*General Linear Model* [LM]) dar. Es gehört zur Gruppe der parametrischen Modelle und wird über die Modellgleichung

$$Y = f(X) = X^\top \beta + \epsilon$$
$$Y_i = f(x_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2} + \beta_4 x_{i,3} + ... + \epsilon_i$$

spezifiziert. Für jeden inkludierten Prädiktor $j = 1, \ldots, P$ wird der Koeffizient β_j geschätzt. Jeder Koeffizient ist modellübergreifend und für alle Beobachtungen identisch. Der lokale Prädiktoreffekt stellt gleichzeitig den globalen Prädiktoreffekt dar.

Die Modellierung anhand eines allgemeinen linearen Modells ist restriktiv. Für das geschätzte Modell $\hat{f}(X)$ wird der additiv lineare datengenerierende Prozess $f(X)$ angenommen. Nichtlineare Datenverläufe können nur über die exakte Spezifikation von Termen höheren Grades und Interaktionseffekten modelliert werden. Die analytischen Formen von hochdimensionalen datengenerierenden Prozessen sind jedoch in der Regel nicht bekannt.

Das geschätzte Modell $\hat{f}(X)$ wird im Folgenden durch ein allgemeines Supervised-Learning-Modell repräsentiert und stellt eine Approximation des wahren Modells $f(X)$ dar. Voraussetzung für das gewählte Modell ist, dass es auf keinen Annahmen über den datengenerierenden Prozess basiert.

Definition 2.0.1 (Supervised-Learning-Model)**.** *Wir folgen der Notation von Casalicchio, Molnar und Bischl (2018). Es existiere ein p-dimensionaler Prädiktorraum $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p$ mit Indexmenge $p = 1, \ldots, P$ und ein Zielvariablenraum $\mathcal{Y}$. Das Supervised-Learning-Modell lernt den Zusammenhang anhand einer i.i.d.-Stichprobe aus der gemeinsamen Verteilung $\mathcal{X}_p \times \mathcal{Y}$. Die korrespondierenden Zufallsvariablen aus dem Prädiktorraum seien mit $x_1, \ldots, x_p$ notiert, die Zufallsvariable aus dem Zielvariablenraum mit y. Der Vektor $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)})^\top$ repräsentiert die i-te Beobachtung, die mit dem Zielvariablenwert $y^{(i)} \in \mathcal{Y}$ assoziiert ist. Der Vektor $x_j = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(n)})^\top$ repräsentiert die Werte des j-ten Prädiktors.*

Supervised-Learning-Modelle besitzen eine hochdimensionale Prädiktionsfunktion, die nicht in einer ähnlichen algebraischen Form darstellbar ist wie parametrische Modelle und als „Black-Box“ verstanden werden kann (Hooker, 2007). Unser Ziel ist die Dekomposition der geschätzten Black-Box-Responsefunktion $\hat{f}(X)$ in Effekte erster, zweiter und höherer Ordnung. Roosen (1995), Hooker (2007) und Muehlenstaedt et al. (2012) geben eine Dekomposition der gelernten Funktion $\hat{f}(X)$ in eine Summe aus niedrigdimensionalen Funktionen an, die zueinander orthogonal sind.

Die Gesamtvarianz der Funktion kann ebenfalls additiv zerlegt werden, weshalb die Dekomposition *Functional Analysis of Variance [Functional ANOVA]* genannt wird.

Definition 2.0.2 (Functional-ANOVA-Dekomposition)**.** *Betrachte $f : \triangle \to \mathcal{R}, f \in L_2(\triangle, \mathcal{R})$ mit $\triangle = \triangle_1 \times, \ldots, \triangle_P$. Sei x ein Zufallsvektor auf dem Definitionsbereich $\triangle$ mit Integrationsmaß dv. $x_1, \ldots, x_P$ werden als unabhängig angenommen. Daraus folgt $dv = dv_1\ dv_2\ \ldots\ dv_P$. Sei weiterhin die Funktion f quadratintegrierbar. Die Functional-ANOVA-Dekomposition ist definiert als*

$$f(X) = g_0 + \sum_{i=1}^{P} g_i(x_i) + \sum_{j<k} g_{jk}(x_j, x_k) + \sum_{j<k<l} g_{jkl}(x_j, x_k, x_l) + \cdots + \\ + g_{12\ldots P}(x_1, x_2, x_3, \ldots, x_P)$$

Jeder Term ist zentriert und orthogonal:

$$\mathbb{E}\left[g_j(x_j)\right] = 0$$
$$\forall i \neq j : \mathbb{E}\left[g_j(x_j)\, g_k(x_k)\right] = 0$$

$j = 1, 2, \ldots, P$ repräsentieren die Indices der Prädiktorvariablen $x_1, x_2, \ldots, x_P$. Es existiert ein konstanter Term g_0. Die Funktionen $g_1(x_1), \ldots, g_P(x_P)$ können als Haupteffekte (Effekte erster Ordnung) der Prädiktorvariablen $x_1, \ldots, x_P$ interpretiert werden, die Terme $g_{jk}(x_j, x_k), j < k$ als zweifache Interaktionen zwischen den Prädiktoren x_j und x_k (Effekte zweiter Ordnung), etc.

Effekte der Ordnung ≥ 3 sind zunehmend schwerer interpretierbar (Apley, 2016). Eine intuitive Visualisierung ist zudem nur bis zum zweiten Effektgrad möglich, weshalb der Fokus im Folgenden auf Effekte erster und zweiter Ordnung gelegt wird. Die Notation folgt Goldstein et al. (2013), sowie Casalicchio, Molnar und Bischl (2018).

Definition 2.0.3 (Selektierte und nicht selektierte komplementäre Prädiktoren)**.** *Jedem Prädiktor wird eine eindeutige Indexnummer i zugewiesen.*

$$i \in \{1, \ldots, K\}, \quad P = \textit{Anzahl der Prädiktoren}$$

Die Menge S repräsentiert die Indices der selektierten Prädiktorvariablen. Nicht selektierte Prädiktoren werden in der Menge C zusammengefasst. Falls nicht anders angegeben, ist S einelementig. Für P = Anzahl der Prädiktoren gilt:

$$S \in \{1, \ldots, P\}, \quad C = \{1, \ldots, P\} \setminus S$$

Die jeweiligen Variablenwerte werden durch x_S und x_C repräsentiert.

Definition 2.0.4 (Effektordnungen)**.** *Der Effekt erster Ordnung des Prädiktors x_i sei mit $g_i(x_i)$ notiert. Der Effekt zweiter Ordnung der Prädiktoren x_i und x_j mit $g_{ij}(x_i, x_j)$, etc.*

$$\textit{Effekt erster Ordnung} = g_i(x_i)$$
$$\textit{Effekt zweiter Ordnung} = g_{ij}(x_i, x_j)$$
$$\textit{Effekt dritter Ordnung} = g_{ijk}(x_i, x_j, x_k)$$

Definition 2.0.5 (Prädiktoreffekt)**.** *Der totale Effekt des Prädiktors x_i ist durch diejenigen Terme der Functional-ANOVA-Dekomposition gegeben, die von dessen Werten abhängen. Der Term Prädiktor steht stellvertretend für Variable, Prädiktorvariable oder die im Machine-Learing übliche Terminologie Feature. Der Prädiktoreffekt ist synonym für den Feature-Effect.*

$$f(x_i) = g_0 + g_i(x_i) + g_{ij}(x_i, x_j) + g_{ijk}(x_i, x_j, x_k) + ...$$

Die Unterteilung von Prädiktoreffekten in Effektordnungen ist von zentraler Bedeutung. Wird beispielsweise eine marginale Änderung der Vorhersage bei Variation eines Inputs betrachtet, stellt der geschätzte Prädiktoreffekt eine Summation über alle Effektordnungen dar und kann nicht als isolierter Effekt erster Ordnung betrachtet werden. Zur Illustration soll folgendes hypothetisches Beispiel zur Erkrankungswahrscheinlichkeit an Typ II Diabetes dienen, angelehnt an Molnar (2018).

Tabelle 2.0.1.: Hypothetisches Beispiel zu Effekten k-ter Ordnung

Alter ≥ 50	hohe Triglyceridwerte	Wahrscheinlichkeit für Typ II Diabetes
nein	nein	0.05
ja	nein	0.10
nein	ja	0.15
ja	ja	0.40

Es existiere ein konstanter Term von 0.05 für die Wahrscheinlichkeit eines Menschen, an Typ II Diabetes zu erkranken. Der Haupteffekt der Dummy-Variable *Alter* ≥ 50 betrage +0.05, der Haupteffekt *erhöhter Triglyceridwerte* im Blutbild des Probanden betrage +0.10. Für eine Person, die sowohl älter als 50 ist, als auch erhöhte Triglyceridwerte aufweist, steigt die Wahrscheinlichkeit nicht nur um die Summe beider

Haupteffekte um $+0.05 + 0.10 = +0.15$ auf 0.20, sondern um einen zusätzlichen Effekt zweiter Ordnung der Größe $+0.20$ auf insgesamt 0.40.

Betrachten wir nun eine Person, die jünger als 50 Jahre ist und erhöhte Triglyceridwerte aufweist. Die hypothetische Erkrankungswahrscheinlichkeit für Typ II Diabetes beträgt in diesem Fall 15 %. Eine marginale Erhöhung der Dummy-Variable *Alter* ≥ 50 erhöht die Erkrankungswahrscheinlichkeit um 25 % auf 40 %. Der Haupteffekt des Prädiktors beträgt jedoch nur 5%. Wir haben einen additiven Prädiktoreffekt von Hauptfaktoren und Effekten zweiter Ordnung erhalten. Problematisch wird diese Tatsache vor allem, wenn Prädiktoren stark über Effekte zweiter Ordnung auf die Zielvariable Einfluss nehmen. Im Extremfall wirken diese nur in Interaktion und es existieren keine Haupteffekte.

Ein Verfahren zur Schätzung des Prädiktoreffektes ist im Idealfall unverzerrt. Bezüglich der Functional-ANOVA-Dekomposition bedeutet dies, dass der Schätzer den Prädiktoreffekt in Def. 2.0.5 auf der vorherigen Seite identifiziert.

Definition 2.0.6 (Additive Unverzerrtheit eines Effektschätzers)**.** *Sei $H(\hat{f}(x_S, x_C))$ ein Prädiktoreffektschätzer bezüglich x_S. Die Functional-ANOVA-Dekomposition der geschätzten Black-Box-Responsefunktion sei durch Effekte erster und zweiter Ordnung aller Prädiktorvariablen gegeben.*

$$f(x_S, x_C) = h_0 + g_S(x_S) + \sum_{j \in C} g_{Sj}(x_S, x_j) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j), \quad i, j \in C$$

Der Effektschätzer ist additiv unverzerrt, wenn er den totalen Effekt des Prädiktors identifiziert und additiv verknüpfte Effekte anderer Prädiktorvariablen ausblendet.

$$\mathbb{E}\left[H(\hat{f}(x_S, x_C))\right] = g_S(x_S) + \sum_{j \in C} g_{Sj}(x_S, x_j)$$

Multiplikativ verknüpfte Interaktionseffekte sind durch einen univariaten Einfluss umständlich darzustellen. Wird beispielsweise eine Interaktion der Form $\hat{f}(x_1, x_2) = x_1\, x_2$ betrachtet und das Ziel ist die Effektbestimmung von x_1, dann ist der Prädiktoreffekt von x_1 abhängig von x_2. Wir definieren die multiplikative Unverzerrtheit eines Effektschätzers wie folgt.

Definition 2.0.7 (Multiplikative Unverzerrtheit eines Effektschätzers)**.** *Sei $H(\hat{f}(x_S, x_C))$ ein Prädiktoreffektschätzer bezüglich x_S. Die Functional-ANOVA-Dekomposition der geschätzten Black-Box-Responsefunktion sei durch alleinige multiplikativ verknüpfte Interaktionseffekte der Prädiktorvariablen gegeben. Sei der Effekt zweiter*

Ordnung $g_{ij}(x_i, x_j)$ durch die multiplikative Verknüpfung $g_{ij}(x_i, x_j) = v_i(x_i)\, v_j(x_j)$ gegeben.

$$\begin{aligned} f(x_S, x_C) &= h_0 + \sum_{i \neq j} g_{ij}(x_i, x_j) \\ &= h_0 + \sum_{i \neq j} \left[v_i(x_i)\, v_j(x_j) \right] \end{aligned}$$

Der Effektschätzer von x_i ist multiplikativ unverzerrt, wenn er $v_i(x_i)$ bis auf eine multiplikative Konstante schätzt.

$$\mathbb{E}\left[H(\hat{f}(x_S, x_C)) \right] = v_S(x_S) \sum_{j \in C} ...$$

Es ist jedoch fragwürdig, inwiefern die geschätzten Haupteffekte eines Interaktionseffektes eine Aussagekraft besitzen (Apley, 2016). Treten Interaktionseffekte auf, sollten Effekte erster Ordnung grundsätzlich nie in Isolation betrachtet werden (Apley, 2016). Bei Haupteffektschätzungen ist vor allem die additive Unverzerrtheit des Schätzers relevant, sodass der geschätzte Prädiktoreffekt von x_S bereinigt von additiv verknüpften Effekten anderer Prädiktorvariablen ist.

2.1. Marginale Effekte

Nelder und Wedderburn (1972) stellten *generalisierte lineare Modelle (Generalized Linear Models) [GLM]* als eine flexiblere Verallgemeinerung des linearen Modells vor. Zwar stellt die Prädiktionsfunktion eines GLM keine Black-Box dar, jedoch wird der Erwartungswert der Zielvariable über eine nichtlineare Linkfunktion modelliert. Die Prädiktoren gehen über die Linkfunktion in den Erwartungswert der Zielvariable ein, weshalb auch die geschätzten Prädiktoreffekte nichtlinear sind. Im Logitmodell werden beispielsweise die Log-Odds verwendet. Aufgrunddessen erfolgt die Interpretation über *marginale Effekte (Marginal Effects) [ME]*, welche die Responsefunktion an spezifizierten Stellen numerisch differenzieren. Marginale Effekte können prinzipiell auf jegliche (differenzierbare) Responsefunktion angewandt werden. Eine rechentechnische Implementation steht beispielsweise über den *margins*-Befehl des Softwarepakets *Stata* zur Verfügung.

Definition 2.1.1 (Marginaler Effekt)**.** *Der marginale Effekt der Prädiktorvariablen x_S auf die Responsefunktion $f(x_S, x_C)$ stellt die partielle Ableitung $\frac{\partial f(x_S, x_C)}{\partial x_S}$ dar.*

Definition 2.1.2 (Schätzung des marginalen Effektes)**.** *Der marginale Effekt wird durch die numerische Differenzierung der Responsefunktion nach der Variablen x_S an lokal spezifizierten Stellen geschätzt. Falls x_S eine kategoriale Variable ist, wird die Änderung der Response im Vergleich zu einer Basiskategorie (ceteris paribus) ermittelt.*

2.1.1. Die numerische Differenzierung

Die numerische Differenzierung stellt das rechengestützte Verfahren dar, die Ableitung einer Funktion $f(X)$ numerisch zu approximieren. Da die zu differenzierenden Funktionen i.d.R. nicht analytisch zugänglich sind, ist die Anwendung von Differenzierungsregeln unmöglich.

Definition 2.1.3 (Differentialquotient)**.** *Der Differentialquotient an der Stelle x_0 ist definiert als Grenzwert des Differenzenquotienten im Intervall $[x_0, x_0 + h]$:*

$$\frac{\partial f(x)}{\partial x}_{|x=x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

Der Differentialquotient wird numerisch über die lokale Variation der Eingangsgröße x und der Betrachtung der marginalen Änderung der Ausgangsgröße y approximiert. Im *einseitigen* Ansatz bzw. dem *Vorwärtsdifferenzieren* basiert die Approximation auf dem *Vorwärtsdifferenzenquotient*. Der Vorwärtsdifferenzenquotient wird auch *Newton'scher Differenzenquotient* genannt. Analog wird der *Rückwärtsdifferenzenquotient* für eine Rückwärtsdifferenz definiert.

Definition 2.1.4 (Vorwärtsdifferenzenquotient)**.**

$$\frac{\partial f(x)}{\partial x}_{|x=x_0} \approx \frac{f(x_0+h)-f(x_0)}{h}, \quad h>0$$

Definition 2.1.5 (Rückwärtsdifferenzenquotient)**.**

$$\frac{\partial f(x)}{\partial x}_{|x=x_0} \approx \frac{f(x_0)-f(x_0-h)}{h}, \quad h>0$$

Der *zweiseitige Ansatz* kombiniert Vorwärts-/ und Rückwärtsdifferenzenquotient zum *Zentralen Differenzenquotient* und wird üblicherweise zur numerischen Differenzierung über finite Differenzen verwendet.

Definition 2.1.6 (Zentraler bzw. Symmetrischer Differenzenquotient)**.**

$$\frac{\partial f(x)}{\partial x}_{|x=x_0} \approx \frac{f(x_0+h)-f(x_0-h)}{2h}, \quad h>0$$

Aufgrund der physischen Limitationen von binär arbeitenden Rechensystemen müssen Zahlen in Fließkommadarstellung konvertiert werden und unterliegen hierbei diversen Ungenauigkeiten. Dabei ist zu beachten, dass jede Art von numerischer Differenzierung, die auf der Berechnung von finiten Differenzen basiert, schlecht konditioniert ist (Fornberg, 1981). Zu kleine Werte von h führen zum Problem der Auslöschung (Squire und Trapp, 1998). Hierbei unterscheiden sich $f(x+h)$ und $f(x-h)$ im Zähler des Differenzenquotienten nur durch sehr hohe Nachkommastellen. Aufgrund der notwendigen Rundung resultiert eine finite Differenz von 0. Zu große Werte führen zu einer schlechteren Approximation. Dabei kann h desto größer gewählt werden, je flacher die Funktion im Intervall $[x-h,\, x+h]$ verläuft. Stark gekrümmte Funktionen erfordern kleinere Intervalle. Lindfield und Penny (1989) und Fornberg und Sloan (1994) beschreiben die theoretischen Hintergründe zur Findung einer optimalen Approximation und zur Fehleranalyse. Wir halten zwei Ungenauigkeiten der Approximation des (zentralen) Differenzenquotienten an die Ableitung der Funktion fest:

(1) Einen analytischen Fehler, der aus der Differenz des wahren Differenzenquotienten von der wahren Ableitung resultiert und von der Schrittweite h bestimmt wird.

(2) Einen numerischen Fehler, der aus der Fließkommadarstellung der zugrundeliegenden Zahlen resultiert.

Fehler (1) ist eine Funktion der Schrittweite h. Wir konstruieren die Taylor-Expansion für die Vorwärts-/ und Rückwärtsdifferenz der Vorhersage:

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)\,h^2}{2!} + \frac{f'''(x)\,h^3}{3!} + \frac{f''''(x)\,h^4}{4!} + \dots \quad \text{(A)}$$

$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)\,h^2}{2!} - \frac{f'''(x)\,h^3}{3!} + \frac{f''''(x)\,h^4}{4!} - \dots \quad \text{(B)}$$

$$(A) - (B) = 2f'(x)h + 2\frac{f'''(x)h^3}{3!} + 2\frac{f'''''(x)h^5}{5!} + \dots$$

$$\frac{(A)-(B)}{2h} = f'(x) + \frac{f'''(x)h^2}{3!} + \frac{f'''''(x)h^4}{5!} + \dots$$

$$= f'(x) + b_1h^2 + b_2h^4 + b_3h^8 \dots$$

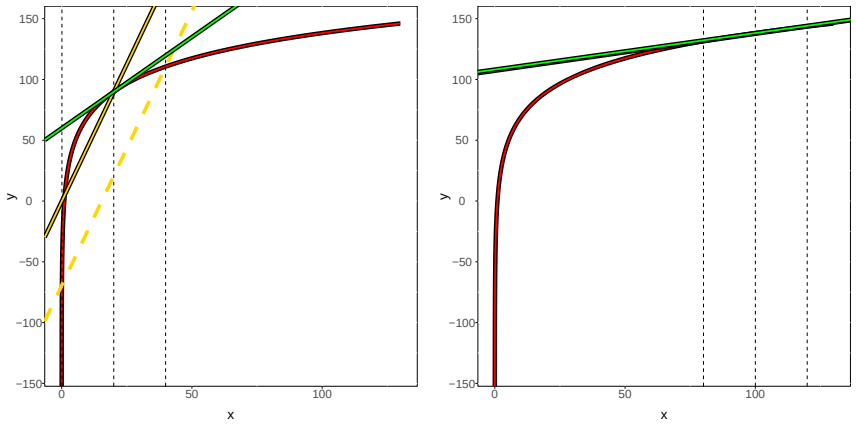
Für den absoluten Fehler (1) folgt

$$\frac{f(x+h)-f(x-h)}{2h} - f'(x) = f'(x) + b_1h^2 + b_2h^4 + b_3h^8 \cdots - f'(x)$$

$$= b_1h^2 + b_2h^4 + b_3h^8 + \dots$$

$$\in \mathcal{O}(h^2)$$

Der absolute Fehler (1) der Approximation des zentralen Differenzenquotienten an die wahre Ableitung wächst quadratisch mit h.



(a) (b)

Abbildung 2.1.1.: Der zentrale Differenzenquotient (gelb) liefert auch bei großen Werten von h eine gute Approximation an den Differentialquotienten (grün), falls die Funktion (rot) nicht stark gekrümmt ist. Gekrümmte Verläufe erfordern kleinere Werte von h.

Betrachten wir nun den numerischen Fehler (2). Wir notieren die fehlerbehaftete numerische Approximation der wahren Funktion $f(x)$ mit $\bar{f}(x)$. Der fehlerbehaftete

zentrale Differenzenquotient wird repräsentiert durch

$$\frac{\bar{f}(x+h) - \bar{f}(x-h)}{2h}$$

Mit ϵ als *Maschinen-Epsilon* gilt für den absoluten Fehler (2):

$$\begin{aligned}
&\frac{\bar{f}(x+h) - \bar{f}(x-h)}{2h} - \frac{f(x+h) - f(x-h)}{2h} \\
=& \frac{\bar{f}(x+h) - \bar{f}(x-h) - [f(x+h) - f(x-h)]}{2h} \\
=& \frac{\bar{f}(x+h) - f(x+h) - \left[\bar{f}(x-h) - f(x-h)\right]}{2h} \\
\leq& \frac{|\bar{f}(x+h) - f(x+h)| + |\bar{f}(x-h) - f(x-h)|}{2h} \\
\leq& \frac{\epsilon + \epsilon}{2h} \\
=& \frac{2\epsilon}{2h} \\
=& \frac{\epsilon}{h}
\end{aligned}$$

Der durch die Fließkommadarstellungsweise resultierende Fehler wächst invers mit h. Daraus folgt: Fehler (1) steigt, je größer die Schrittweite h. Fehler (2) sinkt, je größer die Schrittweite h. Für die simultane Minimierung beider Fehler muss h ausreichend klein für die Minimierung von Fehler (1) und ausreichend groß für die Minimierung von Fehler (2) sein. Der totale Fehler R_{total} ist durch die Addition beider absoluter Fehler gegeben. Wir schätzen ihn ab durch:

$$R_{total} \leq \frac{f'''(x)h^2}{3!} + \frac{\epsilon}{h}$$

Die optimale Schrittweite h minimiert R_{total}. Zur Findung einer optimalen Schrittweite h muss die Funktion $f'''(x)$ zunächst numerisch approximiert werden, weshalb das Verfahren zur Bestimmung von $f'(x)$ nicht praktikabel ist.

Die *Richardson-Extrapolation [RE]* ist ein alternatives Verfahren, das die numerische Approximation der Ableitung durch die Linearkombination zweier Differenzenquotienten verbessert. Eine initiale Schrittweite h_0 kann sukzessive verringert

und die Approximation durch rekursive RE verbessert werden. Sei der zentrale Differenzenquotient der Funktion $f(x)$ zur Schrittweite h gegeben durch $D(h)$. Wir haben bereits gezeigt, dass:

$$D(h) = \frac{f(x+h) - f(x-h)}{2h} = f'(x) + b_1 h^2 + \mathcal{O}(h^4)$$

Daraus folgt, dass für die doppelte Schrittweite gilt:

$$D(2h) = \frac{f(x+2h) - f(x-2h)}{4h} = f'(x) + b_1 4h^2 + \mathcal{O}(h^4)$$

Die Differenz beider Differenzenquotienten ergibt:

$$D(2h) - D(h) = 3b_1 h^2 + \mathcal{O}(h^4)$$

Die Differenz kann folgendermaßen umgeformt werden:

$$\begin{aligned}
\frac{D(2h) - D(h)}{3} &= b_1 h^2 + \mathcal{O}(h^4) \\
\frac{D(2h) - D(h)}{3} &= D(h) - f'(x) + \mathcal{O}(h^4) \\
f'(x) &= D(h) - \frac{D(2h) - D(h)}{3} + \mathcal{O}(h^4) \\
f'(x) &= \frac{4D(h) - D(2h)}{3} + \mathcal{O}(h^4)
\end{aligned}$$

Definition 2.1.7 (Richardson-Extrapolation zur Approximation der Ableitung einer Funktion)**.** *Für zwei Approximationen der ersten Ableitung von $f(x)$ durch die zentralen Differenzenquotienten $D(h)$ und $D(2h)$ ist eine verbesserte Approximation des Differenzenquotienten $D(h)$ mit Trunkierungsfehler $\in \mathcal{O}(h^4)$ gegeben durch:*

$$\begin{aligned}
f'(x) &= \frac{4D(h) - D(2h)}{3} + \mathcal{O}(h^4) \\
f'(x) &\approx \frac{4D(h) - D(2h)}{3}
\end{aligned}$$

Wir verwenden für numerische Ableitungen im Folgenden die *grad*-Funktion des R-Paketes *numDeriv*, das die RE als Standardmethode zur Gradientenbildung verwendet. Als initiale Schrittweite verwendet die *grad*-Funktion an der Stelle x_0 einen Wert von $h = 0.0001\, x_0$. Für $x_0 = 0$ oder Werte, die kleiner als eine Toleranzgrenze

sind, wird auf eine initiale Schrittweite von e^{-4} zurückgegriffen. Anschließend wird über mehrere Iterationen die Schrittweite halbiert und die RE rekursiv angewandt, um eine verbesserte Approximation zu erzielen.

Für $x_0 = 1$ erhalten wir die Schrittweiten $h_0 = 0.0001, h_1 = 0.00005, h_2 = 0.000025, h_3 = 0.0000125, h_4 = 0.00000625$. Auf die resultierenden Differenzenquotienten $D_0, \ldots, D_4$ wird nun eine rekursive RE angewandt. Die Extrapolation des i-ten und j-ten Differenzenquotienten sei mit $R(D_i, D_j)$ notiert.

$$
\begin{aligned}
R(D_0, D_1) &= D_1' \\
R(D_1', D_2) &= D_2' \\
R(D_2', D_3) &= D_3' \\
R(D_3', D_4) &= D_4'
\end{aligned}
$$

D_4' stellt die endgültige Approximation der ersten Ableitung dar. Die *grad*-Funktion verwendet per Standard vier Iterationen. Im Folgenden werden diverse Eigenschaften von marginalen Effekten gezeigt, die auf der Berechnung finiter Differenzen beruhen. Da die Richardson-Extrapolation auf einer Linearkombination zweier Differenzenquotienten beruht, bleiben die diskutierten Eigenschaften erhalten.

2.1.2. Additive & multiplikative Unverzerrtheit des marginalen Effektes

Über die Bildung einer finiten Differenz im Zähler des zentralen Differenzenquotienten werden Effekte anderer Prädiktoren ausgeblendet. Der marginale Effekt der Variable x_S entspricht der Summe marginaler Effekte des Haupteffektes von x_S und von Interaktionseffekten mit x_S. Eine beispielhafte Responsefunktion $f(x_1, x_2)$ mit x_2 als Störvariable (*Nuisance Variable*) besitze folgende Effektstruktur (siehe Def. 2.0.5 auf Seite 15).

$$
\hat{f}(x_1, x_2) = h_0 + g_1(x_1) + g_2(x_2) + g_{12}(x_1, x_2)
$$

Der zentrale Differenzenquotient lautet:

$$
\begin{aligned}
\frac{\partial f(x_S, x_C)}{\partial x_S} &\approx \frac{\hat{f}(x_1 + h, x_2) - \hat{f}(x_1 - h, x_2)}{2h} \\
&= \frac{[h_0 + g_1(x_1 + h) + g_2(x_2) + g_{12}(x_1 + h, x_2)]}{2h} - \\
&\quad - \frac{[h_0 + g_1(x_1 - h) + g_2(x_2) + g_{12}(x_1 - h, x_2)]}{2h} =
\end{aligned}
$$

$$= \frac{g_1(x_1 + h) - g_1(x_1 - h)}{2h} + \frac{g_{12}(x_1 + h, x_2) - g_{12}(x_1 - h, x_2)}{2h}$$

Das Szenario kann auf eine beliebige Menge an Prädiktoren und Effektordnungen erweitert werden.

Proposition 2.1.1 (Additive Unverzerrtheit des marginalen Effektes)**.** *Der zentrale Differenzenquotient der geschätzten Responsefunktion $\hat{f}(x_S^{(i)}, x_C^{(i)})$ bezüglich einer Prädiktorvariablen x_S approximiert sowohl die partielle Ableitung des Haupteffekts von x_S, als auch die partielle Ableitung von Effekten höherer Ordnung von x_S mit Prädiktorvariablen in C. Additiv verknüpfte Effekte von Prädiktorvariablen in C werden ausgeblendet.*

$$\begin{aligned}\frac{\partial f(x_S, x_C)}{\partial x_S} \approx \frac{g_S(x_S + h) - g_S(x_S - h)}{2h} + \\ + \sum_{j \in C} \frac{g_{Sj}(x_S + h, x_j) - g_{Sj}(x_S - h, x_j)}{2h} + \\ + \sum_{j,\, l \in C,\, j \neq l} \frac{g_{Sjl}(x_S + h, x_j, x_l) - g_{Sjl}(x_S - h, x_j, x_l)}{2h} + ...\end{aligned}$$

Proposition 2.1.2 (Multiplikative Unverzerrtheit des marginalen Effektes)**.** *Sei ein multiplikativ verknüpfter Interaktionseffekt $g_{12}(x_1, x_2) = v_1(x_1)\, v_2(x_2)$ gegeben, sodass $\hat{f}(x_1, x_2) = g_{12}(x_1, x_2)$. Der zentrale Differenzenquotient des Interaktionseffektes entspricht dem zentralen Differenzenquotienten von $v_1(x_1)$ bezüglich x_1 bis auf eine multiplikative Konstante $v_2(x_2)$:*

$$\begin{aligned}\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} &\approx \frac{v_1(x_1 + h)v_2(x_2) - v_1(x_1 - h)v_2(x_2)}{2h} \\ &= \frac{v_1(x_1 + h) - v_1(x_1 - h)}{2h}\, v_2(x_2)\end{aligned}$$

2.1.3. Varianten marginaler Effekte

Da die Differenzierung lokal definiert ist, müssen Ableitungsstellen des Definitionsraums gewählt werden. Dieser ist im Allgemeinen mehrdimensional. Es existieren drei etablierte Verfahren für die Spezifikation der Differenzierungsstellen (Leeper, 2018).

(i) Durchschnittliche marginale Effekte (Average Marginal Effects) [AME]

(ii) Marginale Effekte am Mittelwert (Marginal Effects at the Mean) [MEM]

(iii) Marginale Effekte an repräsentativen Werten (Marginal Effects at Representative Values) [MER]

Definition 2.1.8 (Average Marginal Effect)**.** *Der durchschnittliche marginale Effekt von x_S an den beobachteten Stichprobenwerten (Bartus, 2005).*

Average Marginal Effects berechnen den marginalen Effekt von x_S auf die Zielvariable für jeden beobachteten Variablenvektor. Der marginale Effekt je Beobachtung wird anschließend gemittelt. AME produzieren im Gegensatz zu MEM eine skalare Metrik für den marginalen Effekt, die als Approximation für den globalen Effekt dienen kann. Die Interpretation eines AME würde also dem durchschnittlich zu erwartenden Effekt von x_S auf die Zielvariable entsprechen, wäre das Modell eine akkurate Beschreibung des datengenerierenden Prozesses (Leeper, 2018) und X eine repräsentative Stichprobe der Grundgesamtheit. Der AME eignet sich zum *Modellvergleich* (Best und Wolf, 2012) und wird unter anderem häufig in der Ökonometrie angewandt (Kleiber und Zeileis, 2008).

Algorithmus 1: Algorithmus zur Berechnung des AME

Data: Datenmatrix X, angepasstes Modell $\hat{f}$
Result: Skalarer Wert des AME
Initialisiere;
N = Anzahl der Beobachtungen ;
forall $i \in N$ **do**
    `// Durchlaufe alle Beobachtungen`
    $ME^{(i)} = \text{Gradient}_{x_S}\left[\hat{f}(X[i,])\right]$ `// Differenziere die geschätzte Responsefunktion am i-ten beobachteten Variablenvektor nach` x_S
end
$AME = \frac{1}{N}\sum_{i=1}^{N} ME^{(i)}$;
return AME

Definition 2.1.9 (Marginal Effects at the Mean)**.** *Der durchschnittliche marginale Effekt von x_S auf die Zielvariable, wobei x_C durch die Stichprobenmittelwerte ersetzt wird (Bartus, 2005).*

Marginal Effects at the Mean substituieren x_C durch die Stichprobenmittelwerte. Anschließend wird die Prädiktionsfunktion am jeweiligen beobachteten Wert von x_S numerisch differenziert und der Effekt gemittelt. Der MEM ist eine Abwandlung des

AME. Die beobachteten Stichprobenmittelwerte können je nach Stichprobenziehung von Populationsmittelwerten abweichen. MEM stellen also ein potentiell irreführendes Verfahren dar, da der mehrdimensionale Mittelpunkt von X, besonders im Fall von Dummy-Variablen, unbeobachtet oder sogar unbeobachtbar sein kann (Bartus, 2005). MEM werden im Folgenden nicht mehr behandelt.

Definition 2.1.10 (Marginal Effects at Representative Values)**.** *Marginale Effekte von x_S auf die Zielvariable, wobei x_C durch manuell spezifizierte Werte ersetzt wird.*

Marginal Effects at Representative Values evaluieren die Prädiktionsfunktion ebenfalls auf einem modifizierten Datensatz, wobei die Werte bestimmter Variablen durch manuell spezifizierte Werte ersetzt werden. MER können als *bedingte marginale Effekte* betrachtet werden. Vor allem bei der Analyse von *Counterfactuals* sind MER hilfreich. MER erlauben eine Vorhersage, bedingt auf die Ausprägung einer Variable in C, z.B. dass die Ausprägung „Geschlecht“ aller Beobachtungen den Wert „männlich“ annimmt. Diese Vorgehensweise ist bei der Verwendung von Supervised-Learning-Modellen problematisch. Es kann nicht davon ausgegangen werden, dass diese zuverlässig über die Trainingsdaten hinaus extrapolieren können, wie es beispielsweise in einem korrekt spezifizierten GLM der Fall wäre (Apley, 2016). Bei der Betrachtung von MER und Verwendung eines Supervised-Learning-Modells muss daher darauf geachtet werden, die spezifizierten Werte nicht außerhalb des Trainingsdatenraumes zu wählen.

2.1.4. Informationsverlust bei AME

Die Aggregation zum AME führt zu einem Informationsverlust, falls die geschätzten marginalen Effekte heterogen sind. Zur Illustration wird eine SVM auf folgendem Simulationsdatensatz trainiert und in Abb. 2.1.2 auf der nächsten Seite visualisiert.

Modell 2.1.1.

$$Y = -x^2 + \varepsilon$$
$$\varepsilon \overset{iid}{\sim} \mathcal{N}(0,3), \quad x \overset{iid}{\sim} \mathcal{U}(-5,5), \quad N = 1000$$

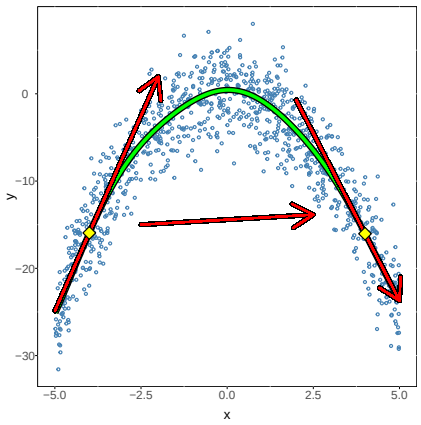


Abbildung 2.1.2.: Die Aggregation marginaler Effekte zum AME verschleiert den quadratischen Effekt der Prädiktorvariable auf die Zielvariable.

Die grüne Linie repräsentiert die Vorhersagen durch das Modell. Die Steigung der horizontalen Linie ist der geschätzte AME von x auf y. Die Steigungen der roten Tangenten enstprechen den geschätzten marginalen Effekten an den Stellen $x = \{-4, 4\}$. Aufgrund des quadratischen Effektes von x auf y stellt der AME ein nicht repräsentatives Maß für den marginalen Effekt dar. Es ist möglich, den AME intervallbasiert anzugeben. Apley (2016) verwendet für das Verfahren der *Accumulated Local Effects* eine Intervallschachtelung, die auf den Quantilen von x_S beruht. Eine solche Intervallschachtelung kann auch für einen intervallbasierten AME genutzt werden.

```
[R]>     interval.left inteval.right       AME
[R]> 0%    -4.98268343   -4.17796151  9.385133
[R]> 10%   -4.17796151   -3.19937736  8.125646
[R]> 20%   -3.19937736   -1.99021930  4.907241
[R]> 30%   -1.99021930   -0.98693663  2.977691
[R]> 40%   -0.98693663    0.03398477  1.113129
[R]> 50%    0.03398477    1.09074704 -1.482359
[R]> 60%    1.09074704    1.94001951 -3.390010
[R]> 70%    1.94001951    2.88875373 -4.938731
[R]> 80%    2.88875373    4.02677317 -6.804336
[R]> 90%    4.02677317    4.99524740 -7.593892
```

2.1.5. Marginale Effekte auf Treppenfunktionen

Weiterhin besitzt die numerische Differenzierung auf Treppenfunktionen keine Aussagekraft. Baumbasierte Verfahren wie *Classification and Regression Trees [CART]* oder der darauf basierende *Random Forest* besitzen beispielsweise eine treppenartige Responsefunktion. Diese resultiert aus der rekursiven Partitionierung des Definitionsraums in disjunkte Teilmengen. Abgesehen von den Sprungstellen zur nächsten Partition verläuft die Responsefunktion konstant. Für den Differenzenquotient an der Stelle x_S und somit den marginalen Effekt gilt:

$$\frac{\partial f(x_S, x_C)}{\partial x_S} \approx \frac{f(x_S + h, x_C) - f(x_S - h, x_C)}{2h} = 0 \quad , x_S \text{ keine Sprungstelle}$$

Der Random Forest stellt eine Ensemble-Methode dar, die die Vorhersage einer Vielzahl an Bäumen mittelt, die auf einer *Bootstrap*-Stichprobe geschätzt wurden. Die Eigenschaft einer intervallweise konstanten Responsefunktion bleibt durch die Mittelwertbildung erhalten. Die Prädiktionsoberflächen einer trainierten SVM, eines Regressionsbaumes und eines Random Forest werden anhand eines simulierten Datensatzes visualisiert (Abb. 2.1.3 auf der nächsten Seite).

Modell 2.1.2.

$$y = -(x_1^2 + x_2^2)$$
$$x_1, x_2 \overset{iid}{\sim} \mathcal{U}(-5, 5), \quad N = 100$$

2.1.6. Marginal Tree Split Effect

Es ist möglich, dennoch sinnvolle marginale Effekte für Treppenfunktionen anzugeben, wenn man diese nur auf die Sprungstellen begrenzt. Die Sprungstellen der Vorhersagen eines Baum-Modells werden *Splits* genannt. Wir schlagen den *Marginal Tree Split Effect [MTSE]* vor, der einen marginalen Effekt je Intervall für CART-Modelle angibt.

Sei die rechte Intervallgrenze durch z_j und die linke Intervallgrenze durch z_{j-1} gegeben. Je Intervall j wird der Vorwärtsdifferenzenquotient (Def. 2.1.4 auf Seite 19) gebildet. Der MTSE ist gegeben durch den Vorwärtsdifferenzenquotient im Intervall j mit Intervallbreite $h = z_j - z_{j-1}$:

$$\begin{aligned} MTSE_j &= \frac{f(z_{j-1} + z_j - z_{j-1}) - f(z_{j-1})}{z_j - z_{j-1}} \\ &= \frac{f(z_j) - f(z_{j-1})}{z_j - z_{j-1}} \end{aligned}$$

Der MTSE enstpricht einer Interpolation der Splits durch eine stückweise Linie. Der marginale Effekt wird anschließend intervallweise und nicht mehr punktweise interpretiert und enstpricht der Liniensteigung im jeweiligen Intervall. Für den *Average Marginal Tree Split Effect [AMTSE]* werden die intervallweisen Steigungen anschließend gemittelt.

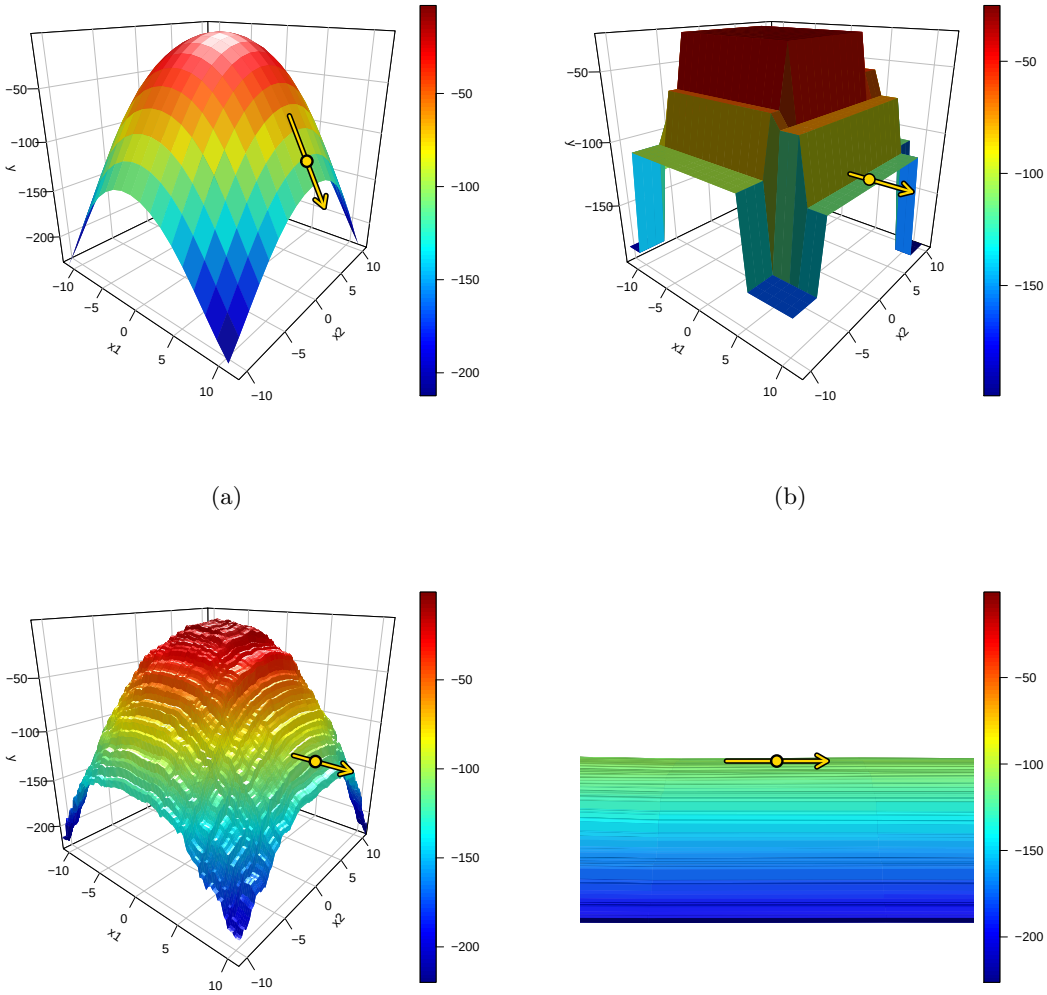


Abbildung 2.1.3.: Responsefunktionen einer SVM (a), eines Regressionsbaumes (b), eines Random Forest (c), sowie ein vergrößerter Ausschnitt des Random Forest (d). Der marginale Effekt von x_1 auf y (Steigung des Pfeils) ist jeweils am Punkt $(x_1 = 10, x_2 = 0)$ gegeben. Aufgrund der stückweisen Konstantheit der Prädiktionsfunktion von Random Forest und CART sind die berechneten marginalen Effekte kein sinnvolles Maß für den Prädiktoreffekt.

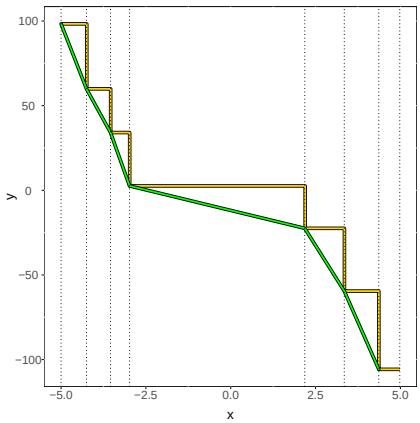


Abbildung 2.1.4.: Die inverse Gewichtung eines Sprungeffektes mit der vorangehenden Intervallbreite ermöglicht die Angabe eines intervallweiten marginalen Effektes.

MTSE und AMTSE lauten:

```
[R]> [1] -51.271231 -36.459790 -56.134794  -4.848847 -31.820173 -45.450690
[R]> [1] -37.66425
```

Im Fall von intrinsisch interpretierbaren CART-Modellen stellt sich die Frage, ob eine Approximation von lokalen Prädiktoreffekten notwendig ist. Als globaler Prädiktoreffekt eignet sich der AMTSE. Für den Random Forest wäre eine Variante aussagekräfter marginaler Effekte wünschenswert, jedoch muss hier für jeden Baum des Ensembles der MTSE ermittelt und anschließend ein Verfahren zur Mittelung gefunden werden.

2.1.7. Schätzgenauigkeit des marginalen Effektes

Die Schätzgenauigkeit des marginalen Effektes hängt maßgeblich von der Qualität des zugrundeliegenden angepassten Modells ab. Nur falls die angepasste Responsefunktion eine ausreichende Approximation an das wahre Modell ist, identifiziert der marginale Effekt auch die partielle Ableitung des wahren Prädiktoreffektes. Eine 10-malige Simulation für drei verschiedene Effekte erster Ordnung der Response $f(x_1, x_2) = \beta x_1 - 5x_2 + \epsilon$ und einer darauf angepassten SVM ohne Tuning zeigt die Fluktuation der Schätzungen:

```
[R]>    beta = -3 beta = 5 beta = 10 iteration
[R]> 1  -2.915887 4.858402  9.619924         1
[R]> 2  -2.931889 4.773995  9.455263         2
[R]> 3  -2.942376 4.876074  9.592102         3
[R]> 4  -2.928268 4.850098  9.463395         4
[R]> 5  -2.915479 4.857313  9.556795         5
[R]> 6  -2.919194 4.907548  9.630847         6
```

```
[R]> 7  -2.946981 4.825488  9.716955         7
[R]> 8  -2.951578 4.825350  9.614363         8
[R]> 9  -2.951502 4.882933  9.592012         9
[R]> 10 -2.955314 4.843125  9.594067        10
```

2.1.8. Counterfactuals zur Entdeckung von Interaktionseffekten

Der marginale Effekt kann sowohl Haupteffekte, als auch Effekte höherer Ordnung (Interaktionseffekte) enthalten. Über die Spezifikation von MER können Interaktionseffekte entdeckt werden. Sei die Differenz zweier marginaler Effekte von x_1 (Prop. 2.1.1 auf Seite 24) für zwei Counterfactuals der Störvariable x_2 gegeben.

$$
\begin{aligned}
&\triangle MER = MER(x_2 = k_1) - MER(x_2 = k_2) = \\
&= \left[\frac{g_1(x_1 + h) - g_1(x_1 - h)}{2h} + \frac{g_{12}(x_1 + h, x_2 = k_1) - g_{12}(x_1 - h, x_2 = k_1)}{2h}\right] - \\
&\quad - \left[\frac{g_1(x_1 + h) - g_1(x_1 - h)}{2h} + \frac{g_{12}(x_1 + h, x_2 = k_2) - g_{12}(x_1 - h, x_2 = k_2)}{2h}\right] = \\
&= \left[\frac{g_{12}(x_1 + h, x_2 = k_1) - g_{12}(x_1 - h, x_2 = k_1)}{2h}\right] - \\
&\quad - \left[\frac{g_{12}(x_1 + h, x_2 = k_2) - g_{12}(x_1 - h, x_2 = k_2)}{2h}\right]
\end{aligned}
$$

Die Änderung des marginalen Effektes durch Setzen zweier Counterfactuals reflektiert die Differenzen des Effektes zweiter Ordnung.

Proposition 2.1.3 (Effektidentifikation einer Differenz von MER)**.** *Die Differenz zweier MER reflektiert ausschließlich diejenigen Effekte höherer Ordnung, die von Counterfactuals betroffene Prädiktoren enthalten.*

$$
\begin{aligned}
&\triangle MER = MER(x_i = k_1) - MER(x_i = k_2) = \\
&= [\frac{g_S(x_S + h) - g_S(x_S - h)}{2h} + \\
&\quad + \frac{g_{Si}(x_S + h, x_i = k_1) - g_{Si}(x_S - h, x_i = k_1)}{2h} + \\
&\quad + \sum_{j \in C \setminus i} \frac{g_{Sij}(x_S + h, x_i = k_1, x_j) - g_{Sij}(x_S - h, x_i = k_1, x_j)}{2h} + \\
&\quad + \dots] \\
&- \\
&\quad [\frac{g_S(x_S + h) - g_S(x_S - h)}{2h} +
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{g_{Si}(x_S + h, x_i = k_2) - g_{Si}(x_S - h, x_i = k_2)}{2h} + \\
&+ \sum_{j \in C \setminus i} \frac{g_{Sij}(x_S + h, x_i = k_2, x_j) - g_{Sij}(x_S - h, x_i = k_2, x_j)}{2h} + \\
&+ \dots] \\
= \\
&[\frac{g_{Si}(x_S + h, x_i = k_1) - g_{Sij}(x_S - h, x_i = k_1)}{2h} \\
&- \frac{g_{Si}(x_S + h, x_i = k_2) - g_{Sij}(x_S - h, x_i = k_2)}{2h}] \\
+ \\
&\sum_{j \in C \setminus i} [\frac{g_{Sij}(x_S + h, x_i = k_1, x_j) - g_{Sij}(x_S - h, x_i = k_1, x_j)}{2h} - \\
&- \frac{g_{Sij}(x_S + h, x_i = k_2, x_j) - g_{Sij}(x_S - h, x_i = k_2, x_j)}{2h}] \\
&+ \dots
\end{aligned}
$$

Im Folgenden wird eine SVM an den *Boston Housing*-Datensatz angepasst. Die Zielvariable stellt der Median-Häuserpreis einer Wohngegend dar. Die marginalen Effekte der Variable *age (Anteil der Häuser einer Wohngegend, die vor 1940 gebaut wurden)*, sowie dreier Counterfactuals der Variable *ptratio (Schüler-Lehrer-Verhältnis der Wohngegend)* werden untersucht.

```
[R>       age
[R> 1 -0.0485
[R>   at(ptratio)      age
[R> 1           0 -0.00006
[R> 2          20 -0.05260
[R> 3          40 -0.00005
```

Ein höherer Anteil alter Häuser ist zunächst negativ mit der Vorhersage des Median-Häuserpreises einer Wohngegend assoziiert. Eine Counterfactual-Schätzung verrät, dass eine Interaktion zwischen den Variablen *age* und *ptratio* vorliegt. Der marginale Effekt kann hier nicht als Effekt erster Ordnung betrachtet werden. Aufgrund der Black-Box-Natur des verwendeten Modells kann der Effektunterschied auch durch Effekte höherer Ordnung, beispielsweise mit einem dritten Prädiktor zustande kommen. Solche können durch die Spezifikation eines *Gitters (Grid)* entdeckt werden.

2.1.9. Gitterstrukturen für Counterfactual-Schätzungen

Bei Verwendung von P Prädiktoren und N verschiedenen Werten pro Variable beträgt die Gridgröße bzw. die Menge an Wertkombinationen N^P. Die Gitterstruktur für MER unterliegt dem *Fluch der Dimensionalität*. Die Anfertigung eines maximalen Gitters erscheint aus rechentechnischen Gründen wenig sinnvoll. Zudem kann eine Vielzahl an Wertkombinationen inhaltliche Widersprüche darstellen. Des Weiteren wird das Supervised-Learning-Modell hierbei zwangsläufig extrapolieren.

Durch die Zerlegung der Prädiktorenmenge $x_1, \ldots, x_P$ in den selektierten Prädiktor x_S und dessen Komplement x_C (Def. 2.0.3 auf Seite 14) kann ein Grid der maximalen Größe N^2 konstruiert werden. In diesem Fall bedingt der marginale Effekt auf den *gesamten* Vektor der komplementären Prädiktoren. Anhand einer beispielhaften Beobachtungsmatrix mit $N = 3$, $P = 3$ soll das Wertegitter schematisiert werden (Tabelle 2.1.1)

In der vorliegenden Permutations-/Prädiktionsmatrix (Tabelle 2.1.2) sind die Vorhersagen des angepassten Modells $\hat{f}(x_S, x_C)$ mit permutierten Inputkombinationen aus x_S und x_C gegeben. Die Diagonalelemente stellen dabei die tatsächlich beobachteten Wertekombinationen dar. Der AME entspricht dem Durchschnitt der numerischen Differenzierungen aller Diagonalelemente. Da nur die beobachteten Kombinationen von x_C verwendet werden, ist die Möglichkeit zur Extrapolation oder zu semantischen Widersprüchen eingedämmt. Apley (2016) & Hooker (2007) demonstrieren, dass bei Verwendung eines solchen Gitters jedoch immer noch die Möglichkeit einer Extrapolation über die Trainingsdaten hinaus besteht.

Tabelle 2.1.1.: Beispielhafte Beobachtungsmatrix

$$\begin{matrix} Y & A & B & C \end{matrix}$$
$$\begin{pmatrix} Y_1 & A_1 & B_1 & C_1 \\ Y_2 & A_2 & B_2 & C_2 \\ Y_3 & A_3 & B_3 & C_3 \end{pmatrix}$$

Tabelle 2.1.2.: Permutations-/Prädiktionsmatrix

$$\begin{matrix} & & x_C^{(1)} & x_C^{(2)} & x_C^{(3)} \\ & & (B = B_1, C = C_1) & (B = B_2, C = C_2) & (B = B_3, C = C_3) \\ x_S^{(1)} & (A = A_1) & \hat{f}(A_1, B_1, C_1) & \hat{f}(A_1, B_2, C_2) & \hat{f}(A_1, B_3, C_3) \\ x_S^{(2)} & (A = A_2) & \hat{f}(A_2, B_1, C_1) & \hat{f}(A_2, B_2, C_2) & \hat{f}(A_2, B_3, C_3) \\ x_S^{(3)} & (A = A_3) & \hat{f}(A_3, B_1, C_1) & \hat{f}(A_3, B_2, C_2) & \hat{f}(A_3, B_3, C_3) \end{matrix}$$

Das Gitter kann zeilenweise (1) oder spaltenweise (2) interpretiert werden. Für eine Beobachtung i existiere jeweils der Wert der selektierten Variable $x_S^{(i)}$ und der Wert der nicht selektierten komplementären Prädiktoren $x_C^{(i)}$.

(1) Zeilenweise Interpretation: Für jede Beobachtung i, fixiere den Wert des selektierten Prädiktors $x_S^{(i)}$ und permutiere über alle beobachteten Vektoren der nicht selektierten komplementären Prädiktoren x_C.

(2) Spaltenweise Interpretation: Für jede Beobachtung i, fixiere die Werte der nicht selektierten komplementären Prädiktoren $x_C^{(i)}$ und permutiere über alle beobachteten Werte des selektierten Prädiktors x_S.

Die *Individual Conditional Expectation* und *Partial Dependence* basieren auf der Idee des oben genannten Gitters und werden in Sektion 2.2 auf der nächsten Seite vorgestellt.

2.1.10. Laufzeitverhalten von marginalen Effekten

Der Rechenaufwand des AME ist für N Beobachtungen durch die N-malige Differenzierung der Responsefunktion, sowie der Durchschnittsbildung von N Werten gegeben. Die Worst-Case Laufzeit in Landau-Notation lautet:

$$f_{\text{AME}}(x) \in \mathcal{O}(N)$$

Für MER hängt die Laufzeit von der Gittergröße ab. Für ein vorgeschlagenes Grid aus x_S und x_C lautet die Worst-Case-Laufzeit:

$$f_{\text{MER}_{Grid}}(x) \in \mathcal{O}(N^2), \quad Grid = \{x_S, x_C\}$$

2.2. Individual Conditional Expectation & Partial Dependence

Die *Partial Dependence* nach Friedman (2001) definiert eine partielle Abhängigkeit der Prädiktionsfunktion $\hat{f}(x_S, x_C)$ von einer oder mehreren selektierten Variable(n) x_S, nachdem über die marginale Verteilung von x_C integriert wurde (Goldstein et al. 2013).

Definition 2.2.1 (Partial Dependence)**.**

$$f_{PD}(x_S) = \mathbb{E}_{x_c}\left[f(x_s, x_c)\right] = \int f(x_s, x_c) dP(x_c)$$

Jede Untermenge von Prädiktoren x_S besitzt eine eigene Partial Dependence Funktion, die den durchschnittlichen Wert der vorhergesagten Response für verschiedene Werte des selektierten Prädiktors x_S angibt, während der Vektor komplementärer Prädiktoren x_C über dessen marginale Verteilung $dP(x_C)$ variiert. Weder das wahre Modell, noch die marginale Verteilung $dP(x_C)$ sind im Regelfall bekannt. Die Schätzung erfolgt anhand der Stichprobenverteilung durch Monte-Carlo-Integration:

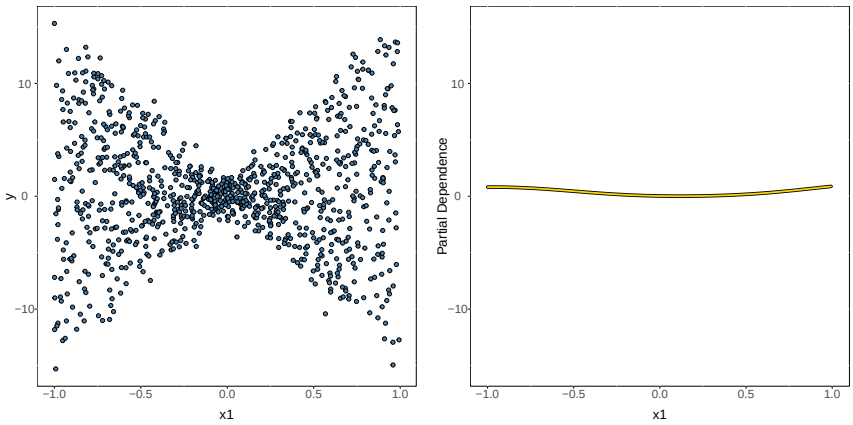
Definition 2.2.2 (Schätzung der Partial Dependence)**.**

$$\widehat{f_{PD}}(x_S) = \sum_{i=1}^{n} \hat{f}_S{}^{(i)}(x_S) = \frac{1}{N} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

Die Schätzung der Partial Dependence [PD] am Punkt $x_S^{(i)}$ entspricht dem arithmetischen Durchschnitt der i-ten Zeilenelemente der Permutations-/Prädiktionsmatrix. Die geschätzte Funktion der Partial Dependence wird durch die Vereinigung der Zeilendurchschnitte repräsentiert. Ihre graphische Visualisierung wird *Partial Dependence Plot [PDP]* genannt (Friedman, 2001). Die PD stellt ein aggregiertes Maß für den Prädiktoreffekt von x_S dar. Analog zum AME geht bei der Aggregation der lokalen Effekte Information verloren, falls diese heterogen sind. Als Beispiel soll analog zu Goldstein et al. (2013) folgender datengenerierender Prozess, sowie eine darauf trainierte SVM dienen.

Modell 2.2.1.

$$Y = x_1^2 - 15x_1x_2 + \varepsilon$$
$$\varepsilon \overset{iid}{\sim} \mathcal{N}(0, 0.3), \quad x_1, x_2 \overset{iid}{\sim} \mathcal{U}(-1, 1), \quad N = 1000$$



(a) (b)

Abbildung 2.2.1.: Der Partial Dependence Plot verschleiert den wahren Effekt des Prädiktors.

Trotz der eindeutigen Korrelation zwischen x_1 und der Zielvariable y suggeriert der PDP einen flachen Zusammenhang (Abb. 2.2.1). Friedman (2001) schlägt den PDP als nützliches Maß des Prädiktoreffekts vor, falls keine nennenswerten Interaktionseffekte existieren. Im Falle von Interaktionseffekten zwischen den Prädiktoren kann dieser irreführend sein. Goldstein et al. (2013) empfehlen stattdessen die *Individual Conditional Expectation.*

Definition 2.2.3 (Individual Conditional Expectation)**.** *Erwartete Vorhersage der Zielvariable je Beobachtung i in Abhängigkeit von x_S, bedingt auf den beobachteten Vektor $x_C^{(i)}$.*

$$\widehat{f_{ICE}}(x_S) = \mathbb{E}[\hat{f}(x_S, x_C^{(i)})]$$

Die *Individual Conditional Expectation [ICE]* disaggregiert die Schätzung der Partial Dependence. Goldstein et al. (2013) definieren diese wie folgt. Statt des durchschnittlichen partiellen Effektes von x_S auf die vorhergesagte Zielvariable werden die N geschätzten bedingten Erwartungskurven der Zielvariable geschätzt und visualisiert. Jede Kurve repräsentiert die vorhergesagte Zielvariable einer Beobachtung als Funktion von x_S, bedingt auf den beobachteten Vektor der komplementären Prädiktorvariablen x_C. In der Permutations-/Prädiktionsmatrix (Abb. 2.2.1 auf Seite 39) stellt die j-te Spalte die j-te ICE-Kurve dar, während die Partial Dependence am Punkt i den i-ten *Zeilendurchschnitt* darstellt. Jede ICE-Kurve repräsentiert eine individuelle Beobachtung.

Es ist möglich, eine Teilmenge an Beobachtungen ($n < N$) auszuwählen, sowie die Anzahl der Knotenpunkte ($k < N$) zu begrenzen. Die Auswahl der Beobach-

tungen geschieht üblicherweise zufällig, die Knotenzahl wird gleichverteilt gewählt. Dies geschieht zu Lasten der Schätzgenauigkeit. Stehen sehr viele Beobachtungen zur Verfügung, ist ein solches *Sampling* der Werte jedoch zwangsläufig notwendig. Wir betrachten den Standardfall $k = N$ und $n = N$.

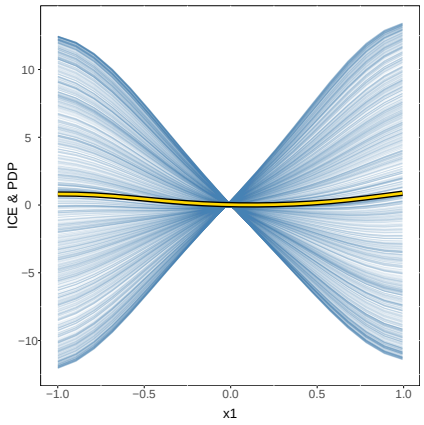


Abbildung 2.2.2.: Die Disaggregation der Partial Dependence zur Individual Conditional Expectation verrät den wahren Prädiktoreffekt.

Die Disaggregation der Partial Dependence zur Individual Conditional Expectation (Abb. 2.2.2) stellt den Zusammenhang korrekt dar. Die Aggregation zur Partial Dependence ist genau dann sinnvoll, wenn die Verläufe der ICE-Kurven formähnlich und nur vertikal verschoben sind. Divergierende Verläufe sind indikativ für Interaktionseffekte zwischen den Prädiktorvariablen (Goldstein et al. 2013).

2.2.1. Partial Dependence und Interaktionseffekte

Die beobachteten Vektoren komplementärer Prädiktorwerte $x_C^{(i)}$ sind für unterschiedliche Verläufe der ICE-Kurven verantwortlich. Weshalb divergente Verläufe indikativ für Interaktionseffekte sind, wird an einer dreidimensionalen Visualisierung (Abb. 2.2.3 auf Seite 40) von Datenmodell 2.2.1 auf Seite 35 deutlich.

Es sind zwei konträr verlaufende ICE-Kurven in roter Farbe hervorgehoben. Diese repräsentieren jeweils eine Beobachtung am Minimum bzw. Maximum der marginalen Verteilung von x_2. Die Partial Dependence ist durch gelbe Farbe gekennzeichnet. In der zur Achse von x_1 perpendikulären Perspektive (links) sind keine Interaktionseffekte an den Datenpunkten zu erkennen. Wird die Perspektive jedoch rotiert (rechts), offenbart sich der Interaktionseffekt zwischen x_1 und x_2 auf die Zielvariable y. Die konträr verlaufenden ICE-Kurven in zweidimensionaler Darstellung sind ein gutes Indiz für vorhandene Interaktionseffekte. Der PDP ist nicht repräsentativ für den Prädiktoreffekt von x_1 auf y.

Algorithmus 2: Berechnung von ICE und PD mit einer Teilmenge an Beobachtungen und Knotenpunkten

Data: Datenmatrix $X = \{x_S, x_C\}$, angepasstes Modell $\hat{f}$, selektierter Prädiktor S, Anzahl der Beobachtungen n, Anzahl der Knotenpunkte k

Result: n Vektoren der ICE und 1 Vektor der PD

Initialisiere;
Zufällige Stichprobe N* der Größe n aus $1, \ldots, N$ für die Indices der ICE;
Gleichverteilte Stichprobe K* der Größe k aus $1, \ldots, N$ für die Indices der Stichprobenwerte von x_S ;
X_{ICE} = Matrix mit n Zeilen und k Spalten;
PD = Vektor der Länge k;
forall $i \in N^*$ **do**
    `// Berechne Vektor der ICE für jede gezogene Beobachtung`
    ICE = Vektor der Länge k;
    forall $j \in K^*$ **do**
        $ICE[j] = \hat{f}(x_S = x_S^{(j)}, x_C^{(i)})$ `// Je Beobachtung i: Durchlaufe alle gezogenen Werte des selektierten Prädiktors und berechne Vorhersage mit substituiertem Wert`
    end
    $X_{ICE}[i,] = ICE$
end
forall $j \in K^*$ **do**
    $PD[j] = \frac{1}{n} \sum_{i \in N^*} \hat{f}_{ICE}^{(i)}(x_S = x_S^{(j)}, x_C = x_C^{(i)})$ `// Je Knotenpunkt j: Bilde arithmetischen Mittelwert aller ICE am jeweiligen Knotenpunkt zur Partial Dependence`
end
return X_{ICE}, PD

Tabelle 2.2.1.: Die Permutations-/Prädiktionsmatrix erzeugt ICE und Partial Dependence

$$
\begin{array}{cc|ccc|c}
 & & x_C^{(1)} & x_C^{(2)} & x_C^{(3)} & \\
 & & (B=B_1, C=C_1) & (B=B_2, C=C_2) & (B=B_3, C=C_3) & \\
x_S^{(1)} & (A=A_1) & \hat{f}(A_1, B_1, C_1) & \hat{f}(A_1, B_2, C_2) & \hat{f}(A_1, B_3, C_3) & \frac{1}{3} * \sum_{i=1}^{3} \hat{f}(x_S^{(1)}, x_C^{(i)}) = \mathrm{PD}(x_S = x_S^{(1)}) \\
x_S^{(2)} & (A=A_2) & \hat{f}(A_2, B_1, C_1) & \hat{f}(A_2, B_2, C_2) & \hat{f}(A_2, B_3, C_3) & \frac{1}{3} * \sum_{i=1}^{3} \hat{f}(x_S^{(2)}, x_C^{(i)}) = \mathrm{PD}(x_S = x_S^{(2)}) \\
x_S^{(3)} & (A=A_3) & \hat{f}(A_3, B_1, B_1) & \hat{f}(A_3, B_2, B_2) & \hat{f}(A_1, B_3, C_3) & \frac{1}{3} * \sum_{i=1}^{3} \hat{f}(x_S^{(3)}, x_C^{(i)}) = \mathrm{PD}(x_S = x_S^{(3)})
\end{array}
$$

$$
\begin{aligned}
&\bigcup_{i\in[1,2,3]} \hat{f}(x_S^{(i)}, x_C^{(1)}) \quad \bigcup_{i\in[1,2,3]} \hat{f}(x_S^{(i)}, x_C^{(2)}) \quad \bigcup_{i\in[1,2,3]} \hat{f}(x_S^{(i)}, x_C^{(3)}) \\
&= \mathrm{ICE}(x_S^{(1)}) \qquad = \mathrm{ICE}(x_S^{(2)}) \qquad = \mathrm{ICE}(x_S^{(3)})
\end{aligned}
$$

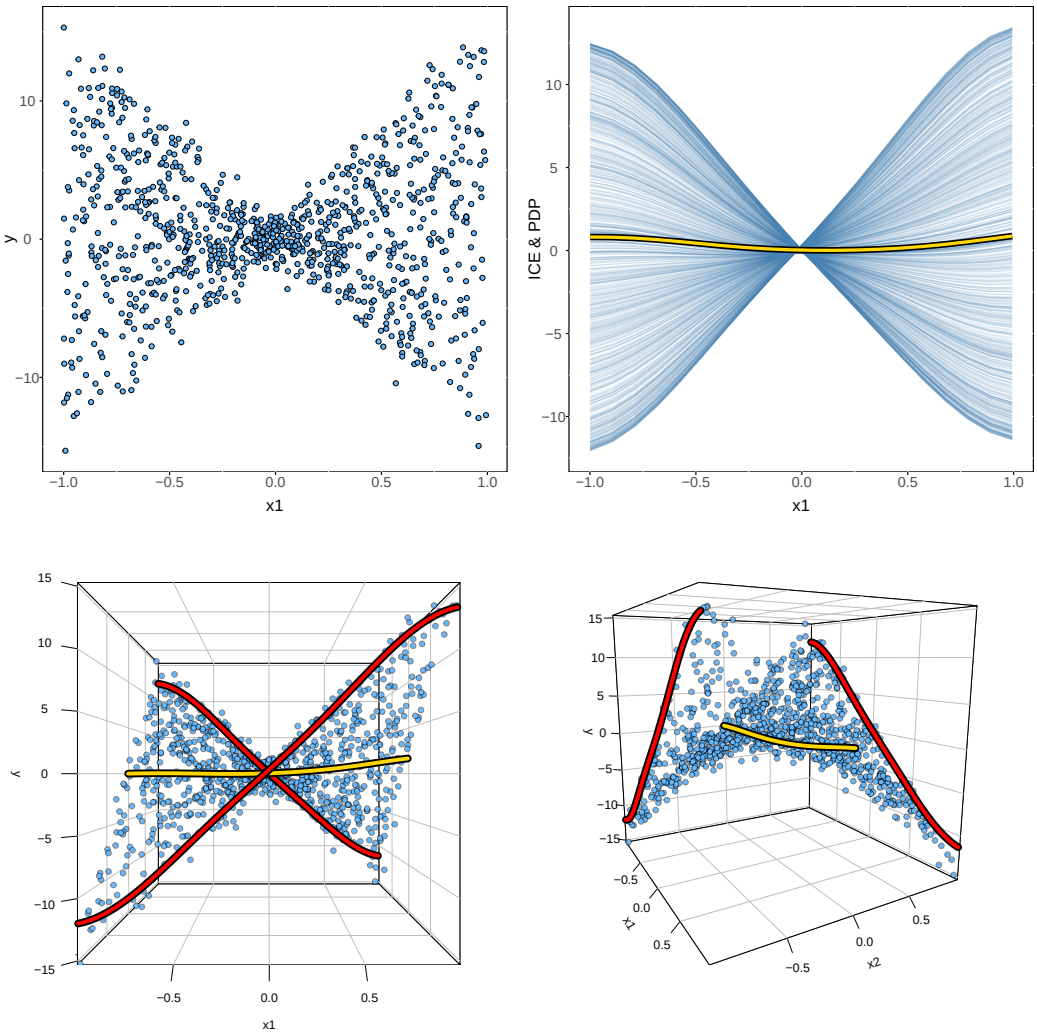


Abbildung 2.2.3.: Divergierende ICE-Kurven sind indikativ für Interaktionseffekte zwischen den Prädiktoren. Die dreidimensionale Darstellung bezeugt die vermuteten Interaktionseffekte. Die Partial Dependence ist nicht repräsentativ für den Prädiktoreffekt von x_1.

Zur Gegenüberstellung wird eine SVM auf einem Datenmodell ohne Interaktionseffekte trainiert und in gleicher Darstellungsweise visualisiert (Abb. 2.2.5 auf der nächsten Seite):

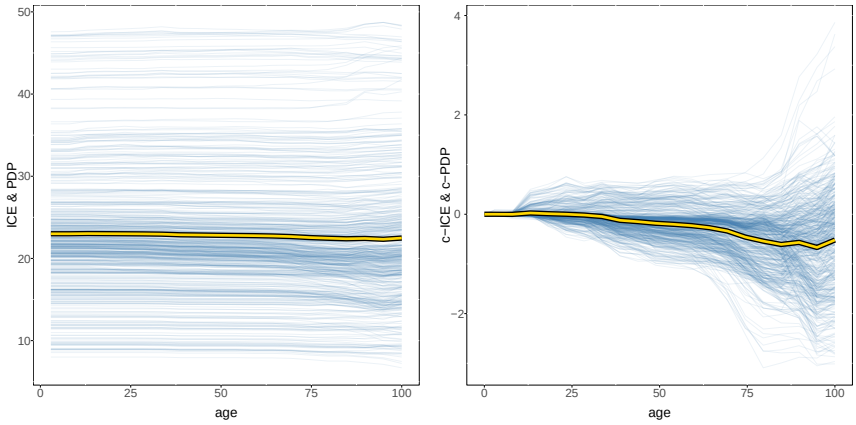
Modell 2.2.2.

$$Y = -5x_1 + 10x_1^3 - 0.5x_2 + \varepsilon$$
$$\varepsilon \overset{iid}{\sim} \mathcal{N}(0, 0.25), \quad x_1, x_2, \overset{iid}{\sim} \mathcal{U}(-1, 1), \quad N = 1000$$

Die Partial Dependence ist im Falle nichtvorhandener Interaktionseffekte ein repräsentatives Maß für den Variableneffekt von x_2 auf y, da die ICE-Kurven ähnliche Verläufe mit vertikalen Verschiebungen aufweisen.

2.2.2. Zentrierte ICE-Plots

Im vorliegenden Fall ist die Erkennung von Interaktionseffekten aufgrund der starken Divergenz der ICE-Kurven ein Leichtes. In vielen Anwendungsfällen weist die abhängige Variable eine große Spannweite auf. Das Erkennen der Kurvatur der ICE-Kurven wird in solchen Fällen erschwert. Als Beispiel soll der vorhergesagte *Median-Häuserpreis einer Wohngegend* im *Boston Housing* Datensatz dienen.



(a) (b)

Abbildung 2.2.4.: ICE sowie c-ICE der Variable *age*. Die Zentrierung am Stichprobenminimum erleichtert die Interpretation.

Die Variable *age* stellt den Häuseranteil einer Wohngegend (in Prozent) dar, die vor 1940 gebaut wurden. Die Spannweite des Prädiktors beträgt 97.1%. Abbildung 2.2.4 zeigt die geschätzten ICE-Kurven (Abb. 2.2.4 (a)). Der Prädiktor scheint keinen Einfluss auf die Zielvariable zu haben. Durch Zentrieren der Schätzungen am Minimum

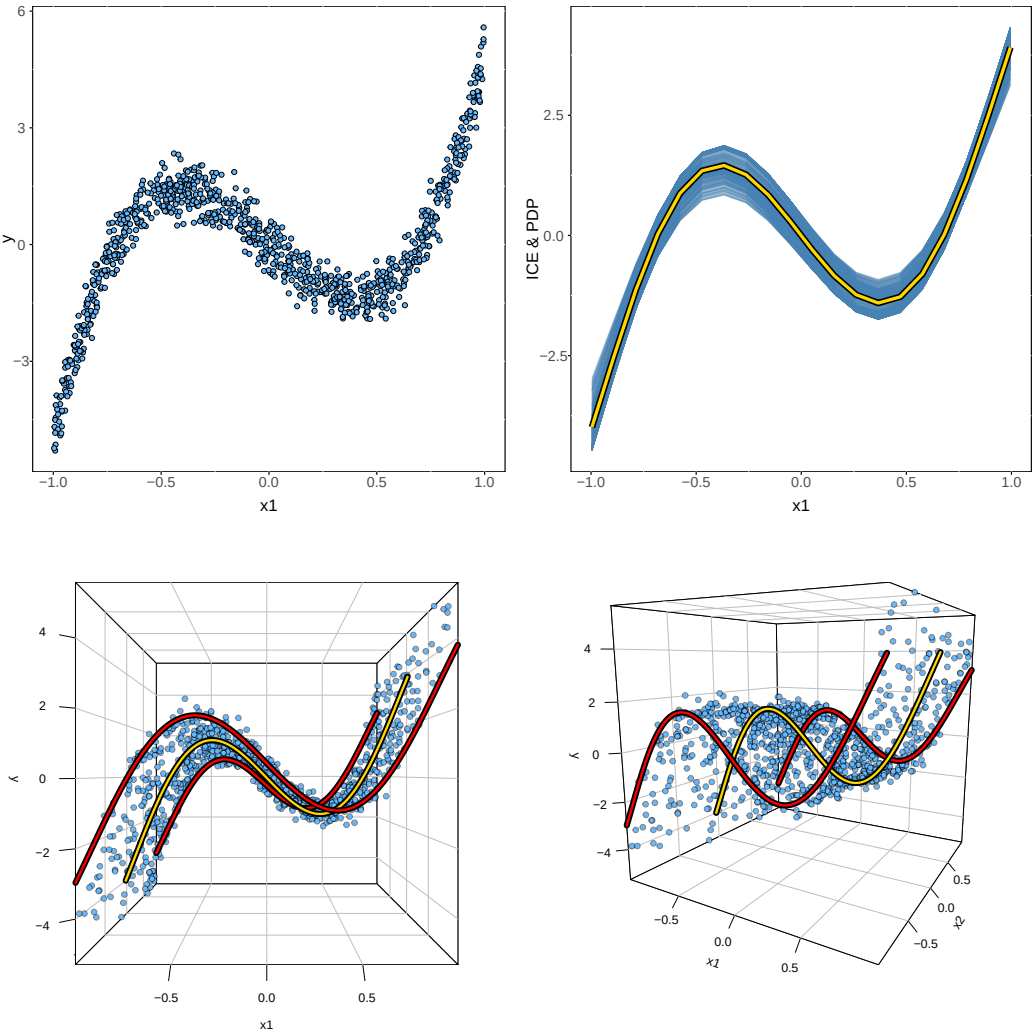


Abbildung 2.2.5.: Parallel verlaufende ICE-Kurven sind indikativ für eine Abwesenheit von Interaktionseffekten. In der dreidimensionalen Darstellung wird deutlich, dass keine Interaktionseffekte in den Daten existieren. Der zweidimensionale ICE-Plot ist ein verlässlicher Indikator. Verlaufen die ICE-Kurven parallel, ist die PD ein repräsentatives Aggregat des Prädiktoreffektes.

von *age* erhalten alle ICE-Kurven einen gemeinsamen Startpunkt, wodurch Divergenzen zwischen den Beobachtungen besser zu erkennen sind (Abb. 2.2.4 auf Seite 41 (b)). Da die Subtraktion einer Spalte eine monotone Transformation darstellt, verändert sich die Steigung der ICE-Kurven nicht. Der c-ICE-Plot dient der visuellen Erleichterung, die Änderungsraten der ICE-Kurven zu erkennen.

Definition 2.2.4 (c-ICE)**.** *Für jede ICE ist die korrespondierende c-ICE mit c^* als Zentrierungspunkt und $\widehat{f_{c-ICE}}^{(i)}(x_S, x_C^{(i)})[c^*]$ als Wert des Zentrierungspunktes gegeben durch*

$$\widehat{f_{c-ICE}}^{(i)}(x_S, x_C^{(i)}) = \widehat{f_{ICE}}^{(i)}(x_S, x_C^{(i)}) - \widehat{f_{ICE}}(x_S, x_C^{(i)})[c^*]$$

Die j-te Spalte einer ICE-Plot-Datenmatrix stellt die Vereinigung aller ICE-Schätzungen am j-ten Wert von x_S (auf der horizontalen Achse) dar. Goldstein et al. (2013) definieren den *Centered ICE Plot [c-ICE]* als ICE-Plot, dessen Werte spaltenweise um die Werte einer spezifizierten Spalte subtrahiert werden. Stellt die spezifizierte Spalte das Minimum von x_S dar, finden alle ICE-Kurven ihren Anfang am Ursprung.

Die Zentrierung der ICE-Kurven geschieht im Boston-Housing-Beispiel wie folgt. Die graue Spalte entspricht hierbei dem Zentrierungspunkt auf der horizontalen Achse. Als Zentrierungsspalte wird das Minimum der Verteilung des Prädiktors gewählt. Jede Reihe entspricht einer einzelnen ICE-Kurve.

Tabelle 2.2.2.: Subtraktion der Zentrierungsspalte der Variable age im Boston-Housing-Datensatz

2.9	6.5	6.8	...	100
27.65 - 27.65	24.70 - 27.65	27.65 - 27.65	...	28.20 - 27.65
30.86 - 30.86	30.89 - 30.86	24.70 - 30.86	...	24.53 - 30.86
39.92 - 39.92	39.92 - 39.92	30.89 - 39.92	...	32.00 - 39.92

Tabelle 2.2.3.: Datenstruktur der c-ICE der Variable age im Boston-Housing-Datensatz

2.9	6.5	6.8	...	100
0	-2.95	0	...	0.55
0	0.03	6.16	...	-6.33
0	0	-9.03	...	-7.92

Der Prädiktor weist bei manchen Gegenden einen negativen und bei anderen Gegenden einen positiven Effekt auf den Median-Häuserpreis auf. Dies ist wiederum ein Indiz für Interaktionseffekte des Prädiktors *age* mit den restlichen Prädiktoren.

2.2.3. Derivative ICE-Plots

Statt die Kurvatur der (zentrierten) ICE visuell zu beurteilen, können die Änderungsraten auch direkt berechnet und visualisiert werden. *Derivative ICE Plots [d-ICE]* stellen das Analogon zum ICE-Plot dar, das statt der vorhergesagten Response die lokale Änderungsrate wiedergibt. Goldstein et al. (2013) wenden vorangehend ein Smoothing-Verfahren auf die ICE-Schätzung an.

Definition 2.2.5 (d-ICE)**.** *Individual Conditional Expectation, deren Funktionsstellen nach x_S numerisch differenziert wurden. Der d-ICE-Plot gibt lokale Änderungsraten der ICE-Kurven an.*

Zur Illustration wird der d-ICE-Algorithmus auf das Boston Housing Beispiel angewandt. Wie vom c-ICE-Plot bereits suggeriert, zeigen sich auch beim d-ICE-Plot bei höheren Werten der Variable *age* Hinweise für Interaktionseffekte mit x_C. Die Änderungsraten steigen mit einem höheren Anteil an Häusern der Wohngegend, die vor 1940 gebaut wurden. Am oberen Ende der marginalen Verteilung von *age* weisen die d-ICE-Kurven einzelner Beobachtungen starke Divergenzen auf.

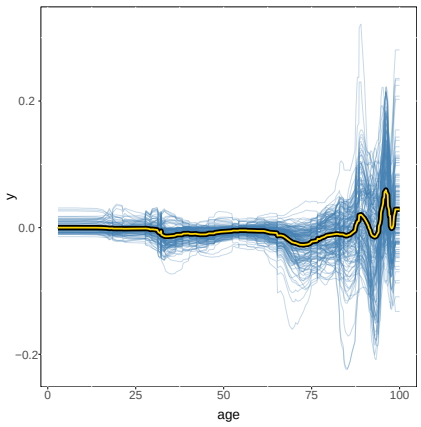


Abbildung 2.2.6.: Der d-ICE-Plot zeigt die lokalen Änderungsraten des ICE. Divergenzen sind ein Indiz für Interaktionseffekte.

Die d-ICE liefert über die lokale Änderungsrate wiederum eine Funktion marginaler Effekte. Anhand des vorherhigen Simulationsbeispiels 2.2.2 auf Seite 41 kann demonstriert werden, wie die d-ICE den wahren marginalen Effekt von x_1 auf y (in grün) approximiert. Die annähernd parallelen Verläufe der d-ICE-Kurven sind indikativ für das Nichtvorhandensein von Interaktionseffekten (Abb. 2.2.7).

Algorithmus 3: Berechnung der c-ICE

Data: Matrix X_{ICE} der geschätzten ICE (n Zeilen repräsentieren jeweils eine ICE), Beobachtungsindex $c^* \in 1, \ldots, N$ des Zentrierungspunktes

Result: Matrix X_{c-ICE} der zentrierten geschätzten ICE

Initialisiere;
Zentrierungsvektor = c^* -te Spalte von X_{ICE};
X_{c-ICE} = Matrix mit gleichen Dimensionen wie X_{ICE} ;
forall $j = 1, \ldots,$ *Spaltenanzahl von* X_{ICE} **do**
    `// Zentriere jede Spalte der ICE-Matrix`
    $X_{c-ICE}[\ , j] = X_{ICE}[\ , j] - X_{ICE}[\ , c^*]$
end
return X_{c-ICE}

Algorithmus 4: Berechnung der d-ICE

Data: Matrix X_{ICE} der geschätzten ICE (n Zeilen repräsentieren jeweils eine ICE), angepasstes Modell $\hat{f}$

Result: Matrix X_{d-ICE} der differenzierten geschätzten ICE

Initialisiere;
N = Zeilenzahl von X_{ICE} `// Anzahl der ICE-Kurven`
K = Spaltenzahl von X_{ICE} `// Anzahl der Knotenpunkte`
X_{d-ICE} = Matrix mit gleichen Dimensionen wie X_{ICE} ;
forall $i \in N$ **do**
    `// Durchlaufe jede einzelne ICE`
    forall $j \in K$ **do**
        `// Durchlaufe jeden einzelnen Knotenpunkt`
        $X_{d-ICE}[i, j] = \text{Gradient}_{x_S}\,[X_{ICE}[i, j]]$ `// Differenziere Knotenpunkt j von i-ter ICE numerisch bezüglich` x_S
    end
end
return X_{d-ICE}

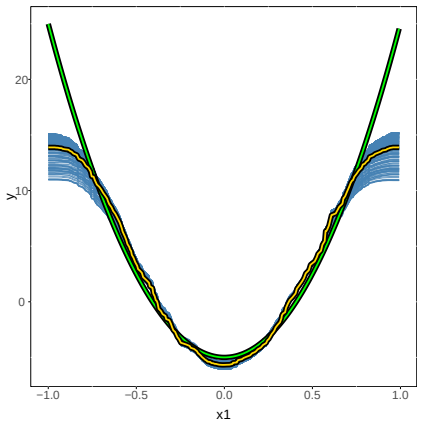


Abbildung 2.2.7.: Sind keine Interaktionseffekte vorhanden, verlaufen die d-ICE-Kurven annähernd parallel. Der d-ICE-Plot (gelb) approximiert im vorliegenden Beispiel den wahren marginalen Effekt (grün).

2.2.4. Relation der d-ICE zu marginalen Effekten

Durch Differenzierung der Zellelemente kann die Permutations-/Prädiktionsmatrix für ICE und PD in eine Permutations-/Prädiktions-/Differenzierungsmatrix (Tabelle 2.2.4 auf der nächsten Seite) überführt werden. Die Differenzierung der Partial Dependence am Punkt $x_S = k$ entspricht dem Durchschnitt aller d-ICE-Werte am Punkt $x_S = k$.

$$
\begin{aligned}
&\frac{\partial \widehat{f_{PD}}(x_S = k, x_C)}{\partial x_S} = \\
&= \frac{\partial \left[\frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S = k, x_C^{(i)}) \right]}{\partial x_S} = \\
&= \frac{1}{n} \frac{\partial \left[\sum_{i=1}^{n} \hat{f}(x_S = k, x_C^{(i)}) \right]}{\partial x_S} = \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{f}(x_S = k, x_C^{(i)})}{\partial x_S}
\end{aligned}
$$

Die überführte Matrix enthält die Werte der d-Varianten von ICE und PDP. Der differenzierte i-te Zeilendurchschnitt entspricht der Ableitung der Partial Dependence am Wert von x_S der i-ten Beobachtung. Die Vereinigung des j-ten Spaltenwerts repräsentiert die d-ICE-Kurve der j-ten Beobachtung. Die Diagonalelemente stellen die marginalen Effekte von x_S dar, d.h. die numerischen Differenzierungen an den tatsächlich beobachteten Wertkombinationen. Der AME von x_S auf die Prädiktion entspricht dem Durchschnitt der Diagonalelemente. Daraus folgt, dass der d-ICE-Plot einer spezifischen Grid-Struktur für marginale Effekte an repräsentativen

Tabelle 2.2.4.: Die Permutations-/Prädiktions-/Differenzierungsmatrix erzeugt AME, d-ICE und d-PD

$$
\begin{array}{cc}
 & \begin{array}{c} x_C^{(1)} \\ (B = B_1, C = C_1) \end{array} \\
x_S^{(1)} \quad (A = A_1) & \frac{\partial \hat{f}(A_1,B_1,C_1)}{\partial x_S} = \mathrm{ME}(x_s^{(1)}) \\
x_S^{(2)} \quad (A = A_2) & \frac{\partial \hat{f}(A_2,B_1,C_1)}{\partial x_S} \\
x_S^{(3)} \quad (A = A_1) & \frac{\partial \hat{f}(A_3,B_1,C_1)}{\partial x_S}
\end{array}
\qquad
\begin{array}{c} x_C^{(2)} \\ (B = B_2, C = C_2) \end{array}
\begin{array}{c} \\ \frac{\partial \hat{f}(A_1,B_2,C_2)}{\partial x_S} \\ \frac{\partial \hat{f}(A_2,B_2,C_2)}{\partial x_S} = \mathrm{ME}(x_s^{(2)}) \\ \frac{\partial \hat{f}(A_3,B_2,C_2)}{\partial x_S} \end{array}
\qquad
\begin{array}{c} x_C^{(3)} \\ (B = B_3, C = C_3) \end{array}
\begin{array}{c} \\ \frac{\partial \hat{f}(A_1,B_3,C_3)}{\partial x_S} \\ \frac{\partial \hat{f}(A_2,B_3,C_3)}{\partial x_S} \\ \frac{\partial \hat{f}(A_3,B_3,C_3)}{\partial x_S} = \mathrm{ME}(x_s^{(3)}) \end{array}
\Bigg)
$$

$$
\begin{array}{l}
\bigcup_{i \in [1,2,3]} \frac{\partial \hat{f}(x_S^{(i)}, x_C^{(1)})}{\partial x_S} \\
\quad = \text{d-ICE}(x_S^{(1)})
\end{array}
\qquad
\begin{array}{l}
\bigcup_{i \in [1,2,3]} \frac{\partial \hat{f}(x_S^{(i)}, x_C^{(2)})}{\partial x_S} \\
\quad = \text{d-ICE}(x_S^{(2)})
\end{array}
\qquad
\begin{array}{l}
\bigcup_{i \in [1,2,3]} \frac{\partial \hat{f}(x_S^{(i)}, x_C^{(3)})}{\partial x_S} \\
\quad = \text{d-ICE}(x_S^{(3)})
\end{array}
$$

$$
\begin{array}{l}
\frac{1}{3} * \sum_{i=1}^{3} \frac{\partial \hat{f}(x_S^{(1)}, x_C^{(i)})}{\partial x_S} = \text{d-PD}(x_S = x_S^{(1)}) \\
\frac{1}{3} * \sum_{i=1}^{3} \frac{\partial \hat{f}(x_S^{(2)}, x_C^{(i)})}{\partial x_S} = \text{d-PD}(x_S = x_S^{(2)}) \\
\frac{1}{3} * \sum_{i=1}^{3} \frac{\partial \hat{f}(x_S^{(3)}, x_C^{(i)})}{\partial x_S} = \text{d-PD}(x_S = x_S^{(3)})
\end{array}
$$

Werten [MER] gleicht, die auf den gesamten Vektor der nicht selektierten komplementären Prädiktoren bedingen.

2.2.5. Additive Unverzerrtheit der Partial Dependence

Die Partial Dependence ist bezüglich additiver Effekte erster Ordnung anderer Prädiktoren bis auf eine additive Konstante unverzerrt [(Apley, 2016) mit Verweis auf (Friedman, 2001)].

Sei eine Responsefunktion der Form $\hat{f}(x_1, x_2) = x_1 + x_2$ und N Beobachtungen gegeben. Die geschätzte Partial Dependence der Response von x_1 lautet:

$$\begin{aligned}
\mathbb{E}\left[\widehat{f_{PD}}(x_1)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_1, x_2^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N}(x_1 + x_2^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N}x_1 + \sum_{j=1}^{N}x_2^{(j)}\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[N\,x_1 + \sum_{j=1}^{N}x_2^{(j)}\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[N\,x_1\right] + \mathbb{E}\left[\sum_{j=1}^{N}x_2^{(j)}\right] \\
&= \frac{1}{N}\,N\,\mathbb{E}\left[x_1\right] + \mathbb{E}\left[\sum_{j=1}^{N}x_2^{(j)}\right] \\
&= x_1 + const.
\end{aligned}$$

Der Fall kann auf eine allgemeine Functional-ANOVA-Dekomposition erweitert werden. Sei nun angenommen, es existiere für jeden Prädiktor ausschließlich ein Effekt erster Ordnung. Wir können folgende Eigenschaft der PD ableiten.

Proposition 2.2.1 (Additive Unverzerrtheit der Partial Dependence)**.** *Die geschätzte Partial Dependence ist bezüglicher additiver Effekte erster Ordnung bis auf eine additive Konstante unverzerrt.*

$$\mathbb{E}\left[\widehat{f_{PD}}(x_S)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_S, x_C^{(j)})\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}(x_S + \sum_{p\in C} x_p^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N} x_S\right] + \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N}\sum_{p\in C}(x_p^{(j)})\right] \\
&= \frac{1}{N}\,N\,x_S + \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N}\sum_{p\in C}(x_p^{(j)})\right] \\
&= x_S + const.
\end{aligned}
$$

2.2.6. Multiplikative Unverzerrtheit der Partial Dependence

Sei nun eine Response der Form $\hat{f}(x_1, x_2) = x_1 x_2$ gegeben. Die geschätzte Partial Dependence der Response von x_1 lautet wie folgt und identifiziert x_1 bis auf eine multiplikative Konstante.

$$
\begin{aligned}
\mathbb{E}\left[\widehat{f_{PD}}(x_1)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_1, x_2^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[\sum_{j=1}^{N}(x_1\,x_2^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[x_1\sum_{j=1}^{N}(x_2^{(j)})\right] \\
&= x_1\,\frac{1}{N}\,\mathbb{E}\sum_{j=1}^{N}(x_2^{(j)}) \\
&= x_1 \times const.
\end{aligned}
$$

Proposition 2.2.2 (Multiplikative Unverzerrtheit der Partial Dependence)**.** *Die allgemeine Partial Dependence der Response von Prädiktor x_S identifiziert $v_S(x_S)$ im Falle multiplikativ verknüpfter Effekte zweiter Ordnung bis auf eine multiplikative Konstante.*

$$
\begin{aligned}
\mathbb{E}\left[\widehat{f_{PD}}(x_S)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\hat{f}(x_S, x_C^{(j)})\right] \\
&= \frac{1}{N}\,\mathbb{E}\left[\sum_{p\in C}\sum_{j=1}^{N}(x_S\,x_p^{(j)})\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N} \mathbb{E}\left[x_S \sum_{p \in C} \sum_{j=1}^{N} (x_p^{(j)}) \right] \\
&= x_S \frac{1}{N} \mathbb{E}\left[\sum_{p \in C} \sum_{j=1}^{N} (x_p^{(j)}) \right] \\
&= x_S \times const.
\end{aligned}
$$

Die Partial Dependence ist sowohl bei stochastischer Abhängigkeit, als auch bei stochastischer Unabhängigkeit von x_S mit Variablen in C bis auf eine multiplikative Konstante unverzerrt (Apley, 2016).

2.2.7. Laufzeitverhalten von ICE und Partial Dependence

Der maximale Rechenaufwand jeweils einer ICE-Kurve beträgt N Vorhersagen. Es können maximal N ICE-Kurven erzeugt werden. Die Partial Dependence mittelt an N Datenpunkten über jeweils N ICE-Werte. Ein zusätzlicher Datenpunkt hat für jede ICE-Kurve je eine zusätzliche Prädiktion zur Folge, sowie eine zusätzliche ICE-Kurve mit $N + 1$ Datenpunkten. Die Partial Dependence mittelt an einem zusätzlichen Datenpunkt über $N + 1$ neue ICE-Werte. Die Worst-Case-Laufzeit in Landau Notation beträgt:

$$f_{\text{ICE, PDP}} \in \mathcal{O}(N^2 + N) = \mathcal{O}(N^2)$$

Die zentrierten Varianten erfordern zusätzlich N^2 Subtraktionen:

$$f_{\text{c-ICE, c-PDP}} \in \mathcal{O}(2N^2 + N) = \mathcal{O}(N^2)$$

Für die d-Varianten fallen für N Prädiktionen zusätzlich N numerische Differenzierungen an. Die Worst-Case-Laufzeit der d-Varianten lautet folglich:

$$f_{\text{d-ICE, d-PDP}} \in \mathcal{O}(2N^2 + N) = \mathcal{O}(N^2)$$

2.2.8. Höherdimensionale Partial Dependence Plots

Die Partial Dependence kann für beliebig viele Prädiktoren x_S geschätzt werden. Für $|S| \geq 3$ ist die Partial Dependence zunehmend schwerer interpretierbar. Für zweielementige S kann die bivariate Partial Dependence als Heatmap oder als dreidimensionale Oberfläche visualisiert werden (Abb. 2.2.9).

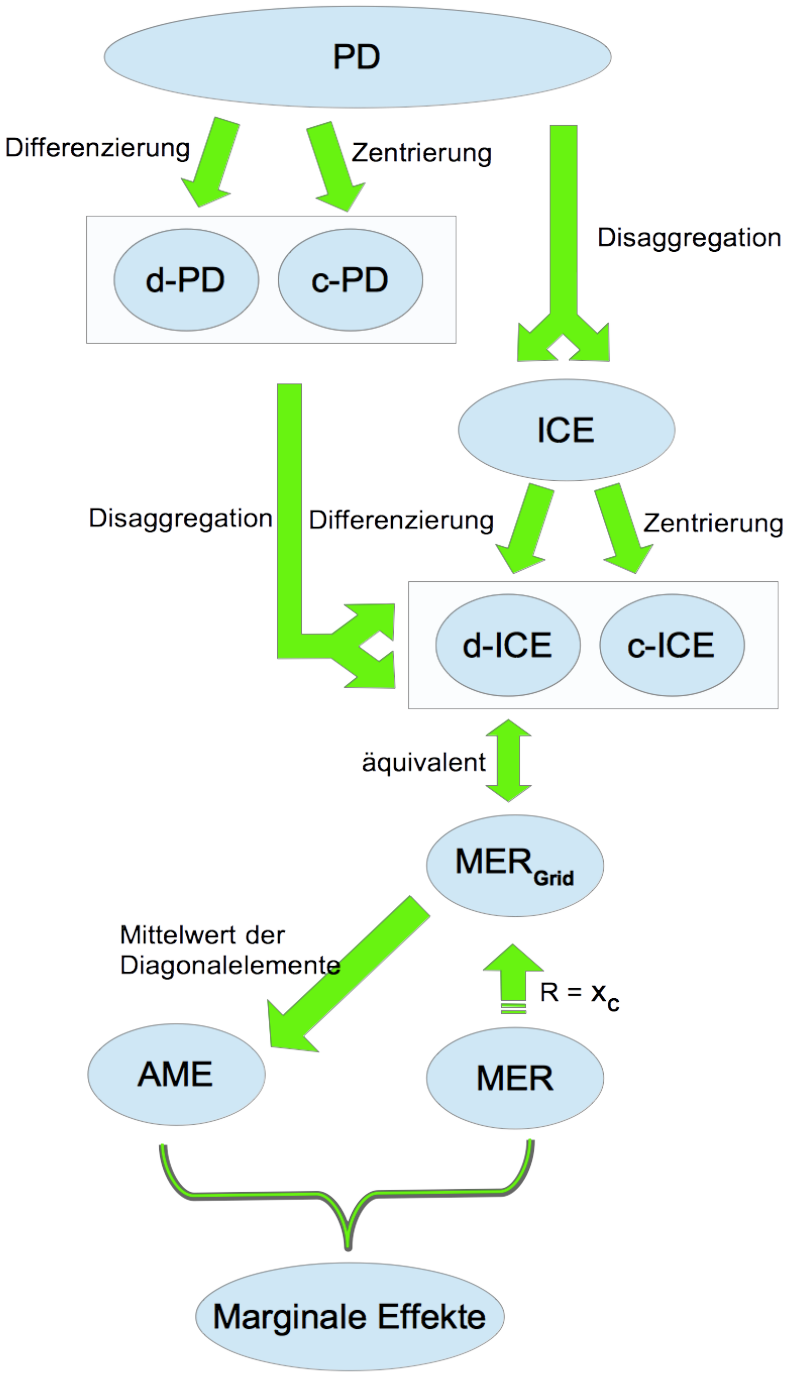


Abbildung 2.2.8.: Relationen zwischen marginalen Effekten, ICE und PD.

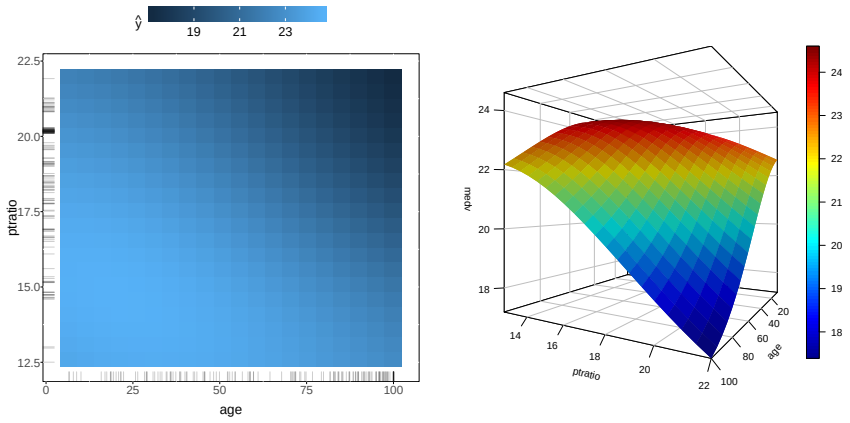


Abbildung 2.2.9.: Die bivariate Partial Dependence von *age* und *ptratio* im Boston-Housing-Datensatz ist indikativ für einen Interaktionseffekt.

2.2.9. Automatisierte Auswertungen der ICE

Goldstein et al. (2013) schlagen eine visuelle Beurteilung der ICE vor. Der Interpretationsprozess kann jedoch auch automatisiert werden. Black-Box-Prädiktionsfunktionen sind in Functional-ANOVA-Dekompositionen dreier Kategorien zu unterscheiden.

(i) Ausschließlich additive Effekte erster Ordnung.

(ii) Ausschließlich Interaktionseffekte.

(iii) Kombination aus Effekten erster Ordnung und Interaktionseffekten (der gängigste Fall).

Die drei Fälle resultieren in unterschiedlichen Charakteristika der ICE-Trajektorien. Vertikale Differenzen von ICE sind indikativ für additive Effekte erster Ordnung in der Responsefunktion (Goldstein et al. 2013). Fall (i) ist durch einen vertikalen Niveauunterschied der ICE-Kurven bei (annähernd) gleicher Form gegeben (Goldstein et al. 2013). Fall (ii) durch ein (annähernd) gleiches Niveau der ICE-Kurven, jedoch mit unterschiedlicher Form. Im gängigsten Fall (iii) besitzen die ICE-Kurven sowohl ein unterschiedliches Niveau, als auch unterschiedliche Trajektorien. Für automatisierte Auswertungen ist die Berechnung von Score-Metriken notwendig, die sensibel für die Fallunterscheidungen (i), (ii) und (iii) sind.

Es werden zwei Score-Metriken zur automatisierten Beurteilung von ICE-Kurven vorgeschlagen. Die erste Metrik beurteilt additiv verknüpfte Effekte anderer Prädiktorvariablen. Die zweite Metrik beurteilt Interaktionseffekte mit dem betrachteten Prädiktor.

Automatisierte Einschätzung der Additivität

Definition 2.2.6 (Kriterien für einen Additivitäts-Score der ICE)**.** *Eine Metrik zur Indikation von additiv verknüpften Effekten anderer Prädiktoren bei der Betrachtung der ICE sollte fünf Faktoren berücksichtigen.*

(1) Ausreichender Durchlauf der Stichprobenwerte von x_S.

(2) Berücksichtigung der Spannweite zwischen Maximum und Minimum der ICE-Kurven.

(3) Berücksichtigung der Varianz der ICE-Werte je Achsenabschnitt. Die ICE können um einen kleinen Bereich schwanken mit großen Ausreißern für die Maxima und Minima an den jeweiligen Achsenabschnitten. In derartigen Fällen sollte die Metrik eine geringere Indikation für additive Effekte erster Ordnung liefern, als wären die ICE (approximativ) gleichverteilt im Intervall zwischen Minimum und Maximum.

(4) Eine Normierung, um Ergebnisse auf verschiedenen Skalen vergleichbar zu machen.

(5) Die Berücksichtigung von unterschiedlichen Trajektorien aufgrund vorhandener Interaktionseffekte.

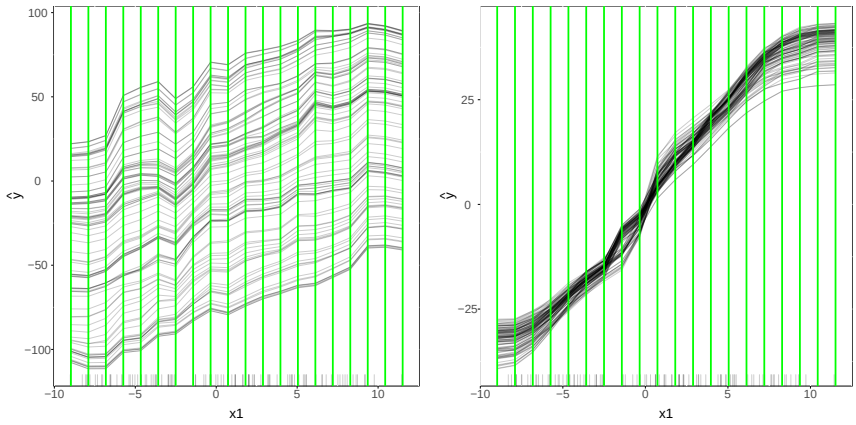
Die folgende Metrik zur Beurteilung der Höhendifferenzen von ICE-Kurven kann im Fall (i) als Indikator bzw. Score für die Additivitätsbeurteilung genutzt werden. Sie berücksichtigt die in Def. 2.2.6 genannten Kriterien (1) bis (4).

Definition 2.2.7 (Score-Metrik zur Beurteilung der Höhendifferenzen von ICE-Trajektorien)**.**

$$Score_{Additivität} = \frac{1}{N}\sum_{i=1}^{N}\left\{max\left[\cup\hat{f}_{ICE}(x_S^{(i)}, x_C)\right] - min\left[\cup\hat{f}_{ICE}(x_S^{(i)}, x_C)\right] \times \right.$$
$$\left. \times \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[\hat{f}_{ICE}(x_S^{(i)}, x_C^{(j)}) - \hat{f}_{PD}(x_S^{(i)})\right]^2}\right\}$$
$$Score_{Additivität,\ Normiert} = \frac{Score_{Additivität}}{max(Zielvariable) - min(Zielvariable)}$$

Der Algorithmus durchläuft eine Stichprobe der Verteilung von x_S. Die Verteilung von x_S stellt hierbei die horizontale Achse des Plots dar. Für jeden Wert wird die Differenz aus Maximum und Minimum der ICE-Kurven mit der Standardabweichung aller ICE am jeweiligen Achsenabschnitt multipliziert. Die Metrik je horizontalem

Achsenabschnitt wird anschließend gemittelt, um ein globales Maß zu erhalten. Zuletzt erfolgt eine Normierung, indem durch die Spannweite der Zielvariable dividiert wird.



(a) (b)

Abbildung 2.2.10.: ICE-Plot für einen additiven datengenerierenden Prozess mit $y = 5x_1 - 10x_2$ in (a) und einen nicht-additiven datengenerierenden Prozess mit $y = 5x_1$ in (b)

Je evaluiertem Achsenabschnitt ergeben sich folgende Score-Metriken für den additiven datengenerierenden Prozess.

```
[R]>  [1] 22.27493 23.81511 24.77575 29.63576 31.25170 28.34146 24.27066 23.73327 27.20126
[R]> [10] 27.74790 28.36175 27.88367 27.89475 27.63738 30.61439 28.55596 27.36420 24.07122
[R]> [19] 23.43217 22.98487
```

Der globale Score des additiven Modells lautet:

```
[R]> [1] 26.59241
```

Je evaluiertem Achsenabschnitt ergeben sich folgende Score-Metriken für den nicht-additiven datengenerierenden Prozess.

```
[R]>  [1] -0.14159282 -0.02008974  0.14102092 -0.37345372 -0.29038805  0.08435681 -0.18296849
[R]>  [8]  0.41592472  0.25031603 -0.05634181  0.21985064  0.14604284  0.13607287  0.14750871
[R]> [15]  0.53383085  0.26502921  0.15942636  0.19804038  0.19846455  0.85008952
```

Der globale Score des nicht-additiven Modells lautet:

```
[R]> [1] 0.134057
```

Die vorgeschlagene Metrik ist im simulierten Beispiel eine verlässliche Indikation für additive Effekte erster Ordnung, d.h. vertikale Unterschiede zwischen den ICE-Kurven. Treten Interaktionseffekte auf, ist der Score fälschlicherweise indikativ für Effekte erster Ordnung.

Algorithmus 5: Automatisierte Auswertung der Höhendifferenzen von ICE-Trajektorien

Data: Matrix X_{ICE} der geschätzten ICE
Result: Score zur Beurteilung additiver Haupteffekte anderer Variablen

Initialisiere;
K = Spaltenanzahl von X_{ICE} `// Anzahl an Achsenabschnitten`
scores = Vektor der Länge K ;
forall $j \in 1, \ldots, K$ **do**
    `// Wiederhole für jeden einzelnen Achsenabschnitt`
    $max = max(X[\,, j])$;
    `// Maximum der ICE-Kurven am Achsenabschnitt j`
    $min = min(X[\,, j])$;
    `// Minimum der ICE-Kurven am Achsenabschnitt j`
    $var = variance(X[\,, j])$;
    $sd = \sqrt{var}$;
    `// Standardabweichung der ICE-Kurven am Achsenabschnitt j`
    $scores[j] = (max - min) \times sd$ `// Score-Berechnung am Achsenabschnitt j`
end
score_total = mean(scores) ;
`// Bilde arithmetisches Mittel der Score-Berechnungen je Achsenabschnitt zum globalen Score`
score_total = $\frac{\text{score\_total}}{max(x_S) - min(x_S)}$;
`// Normiere den globalen Score`
return *score_total*

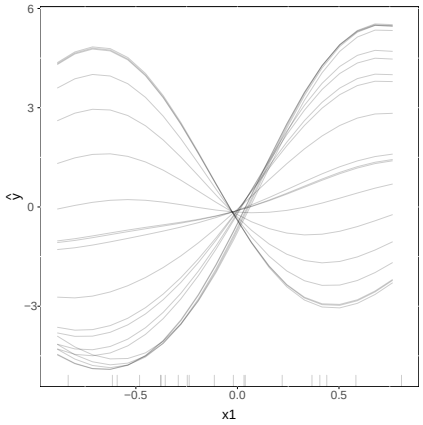


Abbildung 2.2.11.: Interaktionseffekte führen zu einer Fehleinschätzung durch die vorgeschlagene Score-Metrik.

Der Score für additive Effekte erster Ordnung lautet:

```
[R]> [1] 386.5431
```

Trotz ausschließlicher Interaktionseffekte ist der additive Score indikativ für Haupteffekte erster Ordnung. Der vorgeschlagene Score ist nur anwendbar, falls sich die Trajektorien der ICE-Kurven approximativ ähneln. Eine robuste Metrik muss ebenfalls Kriterium (5) der in Def. 2.2.6 auf Seite 53 genannten Kriterien berücksichtigen. Um anwendungsunabhängig eine Indikation für additive Effekte zu liefern, ist weiterhin eine Definition geeigneter Grenzwerte notwendig, die nicht durch die Skalierung der Variablen beeinflusst wird.

Automatisierte Einschätzung der Multiplikativität

Unterschiedliche Verläufe der ICE-Kurven sind immer indikativ für Interaktionseffekte. Die vorgeschlagene Score-Metrik kann in allen Fällen angewandt werden und basiert auf der *Fréchet-Distanz*.

Die *Fréchet-Distanz [FD]* ist ein Maß zur Beurteilung der Ähnlichkeit zweier Kurven. Die Idee der FD basiert auf der folgenden Vorstellung. Eine Person traversiert einen endlich gekurvten Pfad, während sie einen Hund an der Leine führt. Der Hund traversiert einen verschiedenen Pfad und variiert währenddessen seine Geschwindigkeit, um so viel Spannweite der Leine wie möglich zu halten. Die FD wird durch die Länge der kürzesten Leine beschrieben, die ausreicht, um Person und Hund ihre jeweiligen Pfade traversieren lassen zu können. Die FD ist symmetrisch bezüglich beider Kurven, d.h. sie ist identisch, falls der Hund die Person an der Leine führt (Alt und Godau, 1995). Abb. 2.2.12 illustriert das Gedankenexperiment.

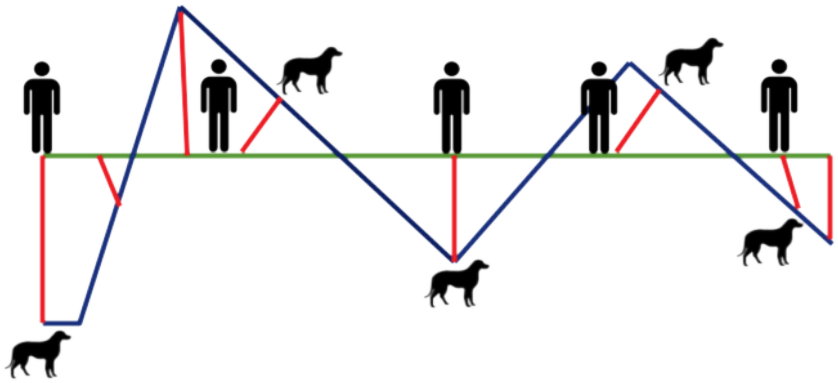


Abbildung 2.2.12.: Illustration des Gedankenexperiments zur Fréchet-Distanz.

Die vorgeschlagene Metrik zur Beurteilung von multiplikativ verknüpften Effekten basiert auf der paarweisen FD zwischen jeweils zwei ICE-Trajektorien. Die resultierenden Distanzen werden anschließend gemittelt. Da die FD symmetrisch ist, sind statt $\sum_{i=1}^{N}(N-1) = N \times (N-1)$ nur $\sum_{i=1}^{N}(N-i)$ Vergleiche notwendig. Die Anzahl der Vergleiche kann folgendermaßen umgeformt werden.

$$\begin{aligned}\sum_{i=1}^{N}(N-i) &= (N-1) + (N-2) + (N-3) + \cdots + (N-(N-1)) + (N-N) \\ &= N^2 - 1 - 2 - 3 - \cdots - (N-1) - N \\ &= N^2 - \sum_{i=1}^{N} i\end{aligned}$$

Seien beispielsweise 4 ICE-Kurven gegeben. Eine asymmetrische Distanz erfordert $3+3+3+3 = 4 \times 3 = 12$ Vergleiche zwischen allen Kurven. Für eine symmetrische Distanz hingegen müssen nur $3+2+1 = 6$ paarweise Distanzen berechnet werden. Um vertikale Unterschiede zwischen den ICE-Kurven ausschließlich auf unterschiedliche Trajektorien zurückzuführen, wird die c-ICE mit den jeweiligen Minima als Zentrierungspunkt gewählt.

Definition 2.2.8 (Score-Metrik zur Beurteilung der Ähnlichkeit von ICE-Trajektorien)**.**

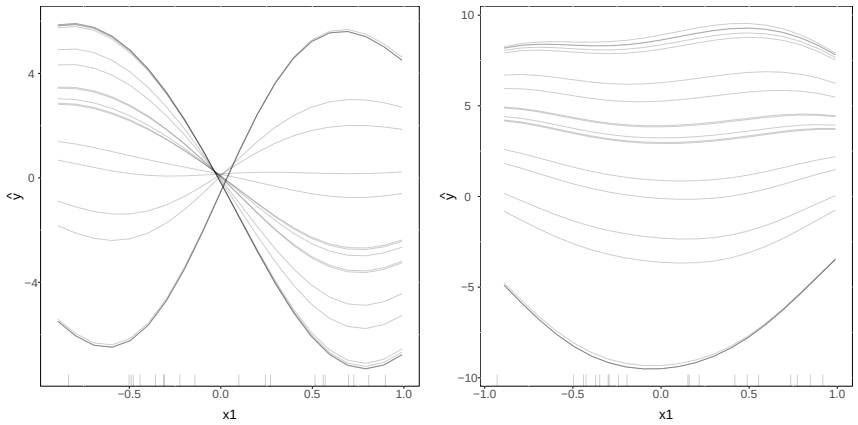
$$Score_{Multiplikativität} =$$

$$\frac{1}{N^2 - \sum\limits_{i=1}^{N} i} \left\{ \sum_{i<j} FD\left[\hat{f}_{ICE}^{(i)}(x_S, x_C^{(i)}) - \hat{f}_{ICE}^{(i)}(x_S^{(1)}, x_C^{(i)}), \hat{f}_{ICE}^{(j)}(x_S, x_C^{(j)}) - \hat{f}_{ICE}^{(j)}(x_S^{(1)}, x_C^{(j)}) \right] \right\}$$

Algorithmus 6: Automatisierte Auswertung der Ähnlichkeit von ICE-Trajektorien

Data: Matrix X_{ICE} der geschätzten ICE
Result: Score zur Beurteilung von Interaktionseffekten mit anderen Variablen

Initialisiere;
N = Zeilenanzahl von X_{ICE} `// Anzahl an ICE`
$X_{c-ICE} = center(X_{ICE},\ center = minimum)$;
`// Transformiere ICE in c-ICE mit jeweiligen Minima als Zentrierungspunkt`
distances = Liste der Länge N ;
forall $i \in 1, \ldots, N$ **do**
    `// Iteriere über alle zentrierten ICE`
    d = Vektor der Länge $(N - i)$ `// Vektor der Distanzen mit Kurve i`
    forall $j \in \{i + 1, \ldots, N\}$ **do**
        `// Iteriere über alle restlichen zentrierten ICE, die noch nicht mit i verglichen wurden`
        $d[j] = FrechetDist(X_{c-ICE}[i\ ,\ ], X_{c-ICE}[j\ ,\ ])$ `// Füge Frechet-Distanz zwischen i-ter und j-ter ICE an Vektorpunkt j ein`
    end
    distances[i] = d `// Füge den Vektor aller paarweiser Distanzen mit i-ter ICE an Listenpunkt i ein`
end
distances = unlist(distances) `// Hebe Listenstruktur auf`
score_total = mean(distances) ;
`// Bilde arithmetisches Mittel aller paarweisen Distanzen`
return *score_total*



(a) (b)

Abbildung 2.2.13.: ICE-Plot für Datenmodell mit Interaktionseffekt $y = 15x_1x_2$ in (a) und ohne Interaktionseffekt $y = x_1^2 - 10x_2$ in (b).

Der globale Score des multiplikativen Modells lautet:

```
[R]> [1] 9.945057
```

Der globale Score des nicht-multiplikativen Modells lautet:

```
[R]> [1] 1.936345
```

Die Auswertung muss nicht vollständig automatisiert werden. Eine Score-Metrik kann dem Anwender auch als Hilfestellung zur Interpretation angegeben werden. Dabei sind jedoch die Laufzeiten der Score-Algorithmen zu berücksichtigen.

```
[R]> [1] "Attention: multiplicativity score critical at 9.945."
[R]> [2] "Interaction effects between x1 and other predictors are likely."
[R]> [3] "Inspection of second order effects is recommended."
```

In der vorliegenden Simulation ist ein höherer Score indikativ für tatsächliche Interaktionseffekte. Es stellt sich die Frage nach Grenzwerten für den Score, d.h. welche Werte indikativ für eine Anwesenheit oder Abwesenheit von Interaktionseffekten sind. Diese unterscheiden sich je nach Anwendung, da die Skalierung der ICE auch den Distanzwert beeinflusst.

Des Weiteren ist die Fréchet-Distanz ein sehr rechenaufwändiges Verfahren. Einen Lösungsansatz, um den Rechenaufwand bzw. die Laufzeit zu senken, stellt ein Sampling der Beobachtungen, wie bei der PD dar. Darüber hinaus kann beispielsweise die *diskretisierte Fréchet-Distanz* (Eiter und Mannila, 1994) verwendet werden. Die dikretisierte FD betrachtet nur Positionen der Leine, bei denen die Endpunkte an Scheitelpunkten der Kurven liegen.

Auch andere Maße zur Beurteilung der Ähnlichkeit der Trajektorien können verwendet werden. Falls die verwendete Metrik die Distanz zwischen den Kurvenpunkten beurteilt, ist darauf zu achten, die vertikalen Unterschiede der Kurven vor der Distanzmessung zu nivellieren.

Goldstein et al. (2013) verwenden den d-ICE-Plot, um Regionen mit Interaktionseffekten zu finden. Der d-ICE-Plot besitzt die Eigenschaft, Höhenunterschiede zwischen den ICE-Kurven zu eliminieren. In Regionen ohne Interaktionen sind die Änderungsraten der ICE-Kurven nahezu identisch. Treten jedoch Interaktionseffekte auf, verändern sich die Trajektorien der ICE-Kurven und somit auch ihre ersten Ableitungen. Unähnliche d-ICE-Trajektorien sind daher indikativ für Interaktionseffekte. Statt der zentrierten ICE kann für die vorgeschlagene Metrik auch die d-ICE verwendet werden. Eine Einschätzung der Multiplikativität anhand der d-ICE ist mit doppeltem Rechenaufwand durch die anfallenden Differenzierungen verbunden.

Zur Verlässlichkeit von auf der ICE basierenden Metriken

Auf der ICE basierende Metriken sind nur für genau solche Effekte indikativ, die das zugrundeliegende Modell erfasst hat. Variierende Trajektorien sind zwar immer indikativ für Interaktionseffekte. Identisch verlaufende ICE-Trajektorien sind jedoch nur indikativ für eine Abwesenheit von Interaktionseffekten im vorliegenden Modell. Ein ausgelassener Prädiktor in den Trainingsdaten kann beispielsweise suggerieren, dass keine Interaktionseffekte existieren, obwohl dessen Inklusion zu variierenden Trajektorien führen würde.

Existieren jedoch Interaktionseffekte ohne den ausgelassenen Prädiktor, hat dessen Abwesenheit keinen Einfluss auf bereits bestehende Interaktionseffekte und somit auch keinen Einfluss auf die Score-Metrik. Die vorgeschlagenen Metriken sind daher nicht für Zielsetzungen geeignet, statistische Inferenz zu betreiben, sondern erlauben nur einen Einblick in die Funktionsweise des angepassten Modells.

2.2.10. Die Extrapolationseigenschaft von ICE und Partial Dependence

Die Monte-Carlo-Integration der Schätzung von ICE und Partial Dependence beruht auf einer angenommenen Gleichverteilung von x_C (Hooker, 2007). Problematisch wird dies, falls die Prädiktorvariablen voneinander abhängig sind. Die Integration über ein gleichverteiltes Wahrscheinlichkeitsmaß von x_C setzt in diesem Fall eine zu große Konzentration auf Regionen mit niedriger Wahrscheinlichkeitsmasse. Hooker (2007) & Apley (2016) raten dazu, die Vorhersage eines Machine-Learning-Modells in datenarmen Bereichen anzuzweifeln.

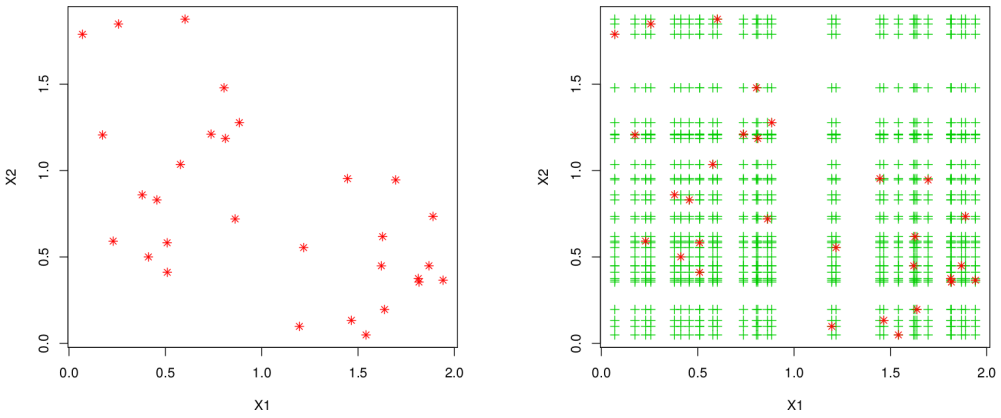


Abbildung 2.2.14.: Auszug aus Hooker und Friedman (2004); Beispielhafter Datensatz in rot (links), sowie darauf basierende Werte für die Berechnung der Partial Dependence in grün (rechts).

Abbildung 2.2.15 zeigt, ähnlich wie in Apley (2016), einen Zusammenhang zwischen zwei Prädiktoren x und y, die in die Funktion $\hat{f}(x, y)$ eingehen. Aufgrund der Multikollinearität zwischen den Prädiktoren unterscheidet sich die marginale Dichte $f(x_C) = f(y)$ an einer beliebigen Stelle $x = k$ deutlich von der bedingten Dichte $f(x_C|x_S = k) = f(y|x = k)$. Der Integrand der Partial Dependence (Def. 2.2.1 auf Seite 35) wird in Datenregionen evaluiert, die keine Daten enthalten.

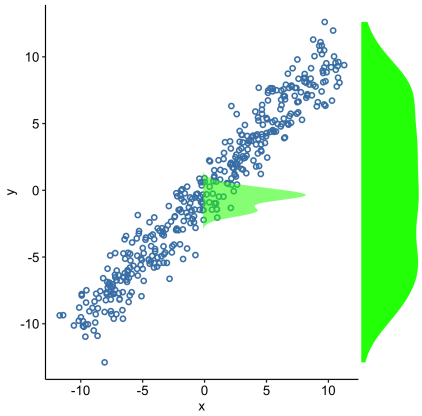


Abbildung 2.2.15.: Vergleich von bedingter Dichte $f(y|x = 0)$ und marginaler Dichte $f(y)$.

$$f_{PD} = \mathbb{E}_{x_c}\left[f(x_s, x_c)\right] = \int f(x_s, x_c) dP(x_c)$$

Das Integral ist der gewichtete Durchschnitt von $f(x, y)$, während y über die marginale Verteilung $f(y)$ variiert. Somit erfolgt die Integration über die gesamte vertikale

Achse, während die Trainingsdaten sich auf die bedingten Verteilungen $f(y|x = k)$ beschränken.

2.2.11. Der marginale Plot als Alternative zur Partial Dependence

Apley (2016) führt als Alternative zur Partial Dependence, die eine Extrapolation über die Trainingsdaten hinaus vermeidet, den *Marginalen Plot [M-Plot]* an. Statt auf die marginale Dichte der komplementären Variablen $f(x_C)$ zu bedingen, greift der M-Plot auf die bedingte Dichte $f(x_C|x_S = k)$ zurück. Für jeden bedingten Wert k werden die auf diesen Wert, oder auf ein kleines Intervall um diesen Wert fallenden Beobachtungen ermittelt. Anschließend wird die Vorhersage für das ermittelte *Subset* gemittelt.

Definition 2.2.9 (Marginaler Plot)**.**

$$f_M(x_S) = \mathbb{E}_{x_c}\left[f(x_S, x_C|x_S = k)\right] = \int f(x_s, x_c)p(x_c|x_S)dx_C$$

Definition 2.2.10 (Schätzung des marginalen Plots)**.**

$$\widehat{f_M(x_S)} = \frac{1}{n(x_S)} \sum_{i=1}^{N(x_S)} \hat{f}(x_S, x_C^{(i)})$$

mit $N(x_S) \subset \{1, 2, ..., n\}$ als Teilmenge der Beobachtungen i für die $x_S^{(i)}$ in eine angemessen kleines Intervall um $x_S = k$ fällt und $n(x_S)$ als die Anzahl der Beobachtungen im jeweiligen Intervall.

Wir simulieren Daten aus Modell 2.2.3 und visualisieren die Vorgangsweise der M-Plot-Schätzung in Abb. 2.2.16 auf der nächsten Seite.

Modell 2.2.3. *Wir entnehmen x_1 und x_2 jeweils der Sequenz $\{-10, -9.95, \ldots, 9.95, 10\}$ und addieren auf jeden Wert eine Ziehung einer standardnormalverteilten Zufallsvariable.*

$$y = \frac{3}{4}x_1\, x_2 + \varepsilon$$
$$\varepsilon \overset{iid}{\sim} \mathcal{N}(0, 3)$$

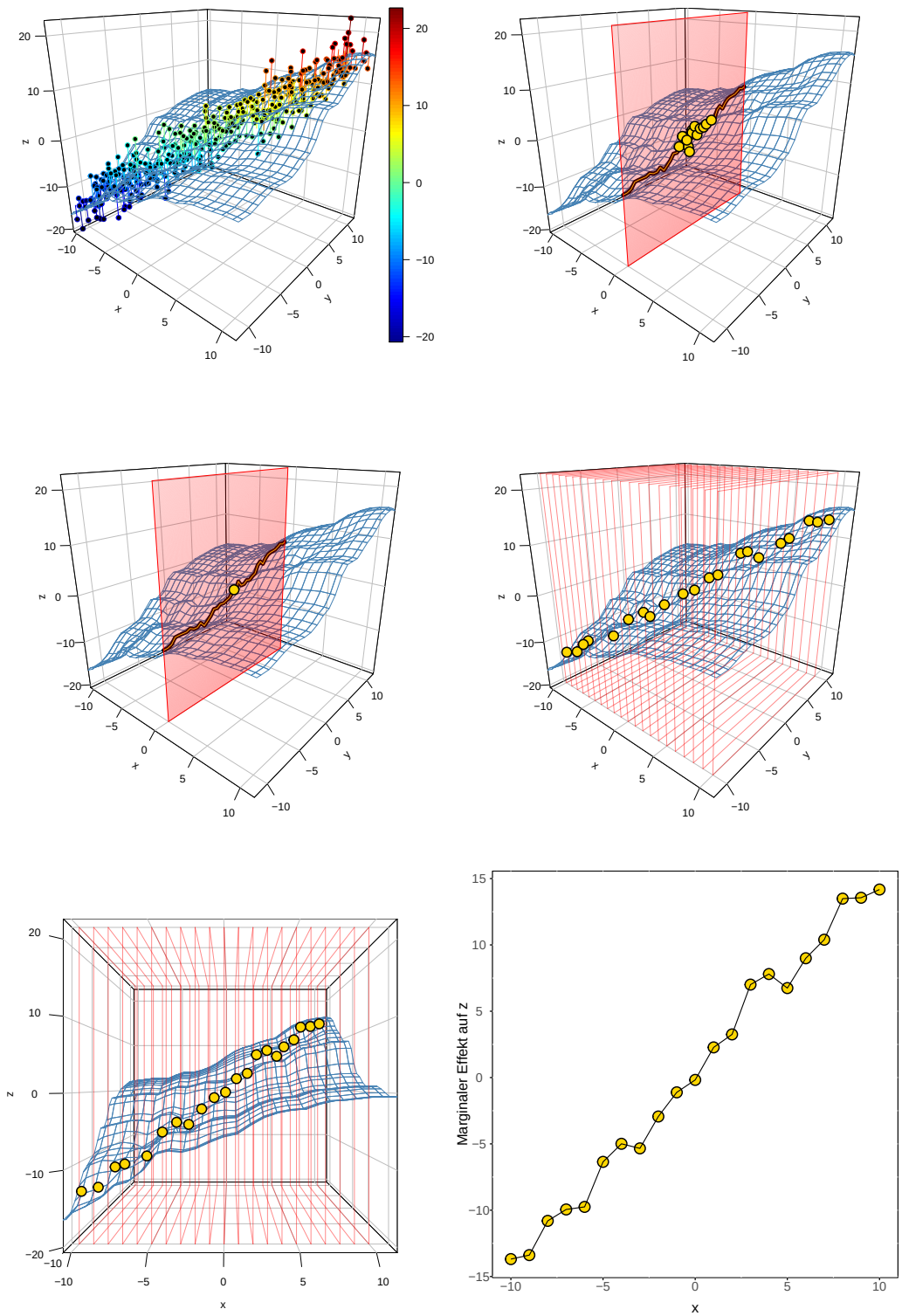


Abbildung 2.2.16.: Schätzung des marginalen Plots aus den Stichprobendaten. Für eine Auswahl bedingter Werte werden die auf einen jeweiligen Wert, oder in ein ausreichend kleines Intevall um den jeweiligen Wert fallenden Beobachtungen ausgewählt. Anschließend wird die Vorhersage für jede Teilmenge bedingter Beobachtungen gemittelt.

Apley (2016) kritisiert die Eigenschaft des M-Plots, über den Effekt von x_C auf die Zielvariable zu aggregieren. Falls Interaktionseffekte zwischen den Prädiktorvariablen existieren, sei der M-Plot verzerrt. Ein Phänomen, das dem *ausgelassenen Variablenproblem (Omitted Variable Bias)* in der Regression ähnelt (Apley, 2016). Als Alternative, die die Vorteile von Partial Dependence und M-Plot vereint, schlägt Apley (2016) *Accumulated Local Effects* vor.

2.3. Accumulated Local Effects

Accumulated Local Effects [ALE] nach Apley (2016) sind eine Alternative zu Partial Dependence und M-Plot, die die (graphische) Bestimmung von Prädiktoreffekten und deren Zerlegung in Haupt-/ und Interaktionseffekte ermöglichen. Die Idee des ALE ist es, die partielle Ableitung der Responsefunktion nach x_S wieder nach x_S zu integrieren, um somit einen akkumulierten partiellen Effekt von x_S auf die Zielvariable zu erhalten.

Definition 2.3.1 (Accumulated Local Effect erster Ordnung)**.** *Mit der partiellen Ableitung der Responsefunktion $\frac{\partial f(x_S,x_C)}{\partial x_S}$ nach x_S als der stetigen Funktion des lokalen Effektes von x_S auf $f(x_S, x_C)$ wird der ALE erster Ordnung definiert als*

$$f_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} \mathbb{E}\left[\frac{\partial f(z_S, x_C)}{\partial z_S}\right] dz_S - C$$
$$f_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} \int p(x_C|z_S) \frac{\partial f(z_S, x_C)}{\partial z_S} dx_C dz_S - C$$

c = const und wird zur Zentrierung des Plots benötigt (siehe Sektion 2.3.5 auf Seite 73).

Zur Illustration sei folgendes Datenmodell, sowie eine darauf trainierte SVM gegeben:

Modell 2.3.1.

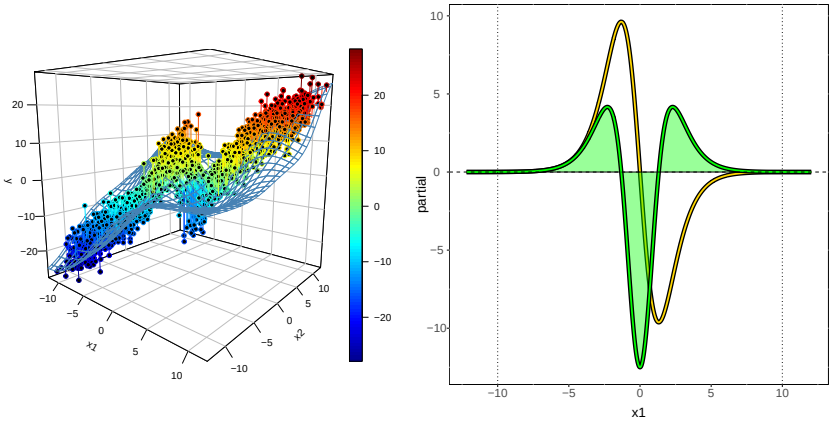
$$x_1, x_2 \overset{iid}{\sim} \{U(-10, 10) + N(0, 1)\}$$
$$y = f(x_1, x_2) = 100 \left[\frac{\partial^2 \left[\frac{1}{1+exp(-x_1)}\right]}{\partial x_1 \partial x_1}\right] + 2x_2 + \varepsilon \qquad \varepsilon \overset{iid}{\sim} N(0, 1)$$

Der Prädiktor x_1 geht über die zweite Ableitung der Sigmoid-Funktion in die Responsefunktion ein. Der lokale Effekt (in grün) von x_1 auf $f(x_1, x_2)$ lautet:

$$\frac{\partial\, g_1(x_1)}{\partial x_1} = \frac{\partial^3 \left[\frac{1}{1+exp(-x_1)}\right]}{\partial x_1 \partial x_1 \partial x_1}$$

Der ALE (in gelb) wiederum ist durch durch das Integral des lokalen Effektes gegeben:

$$\int \left[\frac{\partial^3 \left[\frac{1}{1+exp(-x_1)} \right]}{\partial x_1 \partial x_1 \partial x_1} \right] dx = \frac{\partial^2 \left[\frac{1}{1+exp(-x_1)} \right]}{\partial x_1 \partial x_1}$$



(a) (b)

Abbildung 2.3.1.: Der ALE (gelbe Linie) stellt das Integral der partiellen Ableitung nach x_S (grüne Linie) wiederum nach x_S (grüne Fläche) dar.

2.3.1. Schätzung von Accumulated Local Effects

Für die Schätzung des ALE erster Ordnung aus den Stichprobendaten sind drei Schritte notwendig.

(i) Der lokale Effekt ist im Allgemeinen nicht bekannt und muss intervallweise geschätzt werden. Die partielle Ableitung $\frac{\partial f(x_S, x_C)}{\partial x_S}$ wird durch ein numerisches Verfahren auf einer geeigneten Diskretisierung des Wertebereichs $[min(x_S), max(x_S)]$ in die Menge an Intervallgrenzen $\{z_0, z_1, \ldots, z_j, \ldots, z_k\}$ approximiert. Die Schätzung am Punkt $x_S = z_j$ repräsentiert die numerische Approximation im Intervall $[z_{j-1}, z_j]$.

(ii) Der Erwartungswert $\mathbb{E}\left[\frac{\partial f(x_S, x_C)}{\partial x_S} | x_S = z_j\right]$ des lokalen Effektes am Punkt z_j wird durch die Mittelung über alle geschätzten lokalen Effekte derjenigen Beobachtungen ersetzt, die in das Intervall $[z_{j-1}, z_j]$ fallen. Das Verfahren gleicht der Integration über die komplementären Prädiktoren x_C im M-Plot (siehe Sektion 2.2.11 auf Seite 62).

(iii) Anschließend ersetzt die Akkumulation aller realisierten lokalen Effekte auf die Zielvariable bis zum Wert $x_S = z_k$ das äußere Integral, d.h. die geschätzten lokalen Effekte in den Intervallen $\{[z_0, z_1], [z_1, z_2], \ldots, [z_{j-1}, z_j], \ldots, [z_{k-1}, z_k]\}$ werden kumulativ summiert.

Apley (2016) schlägt vor, den Wertebereich von x_S gemäß der Quantile in eine spezifizierte Anzahl von K Intervallen zu teilen. Bereiche mit einer hohen Konzentration an Beobachtungen erhalten somit eine kleinere Intervallschachtelung als Bereiche mit wenigen Beobachtungen. Der lokale Effekt je Intervall in (i) wird über die Bildung paarweiser finiter Differenzen approximiert. Für jede Beobachtung innerhalb des Intervalls $[z_{j-1}, z_j]$ wird eine paarweise Differenz der Form

$$\hat{f}(x_S^{(i)} = z_j, x_C^{(i)}) - \hat{f}(x_S^{(i)} = z_{j-1}, x_C^{(i)})$$

gebildet.

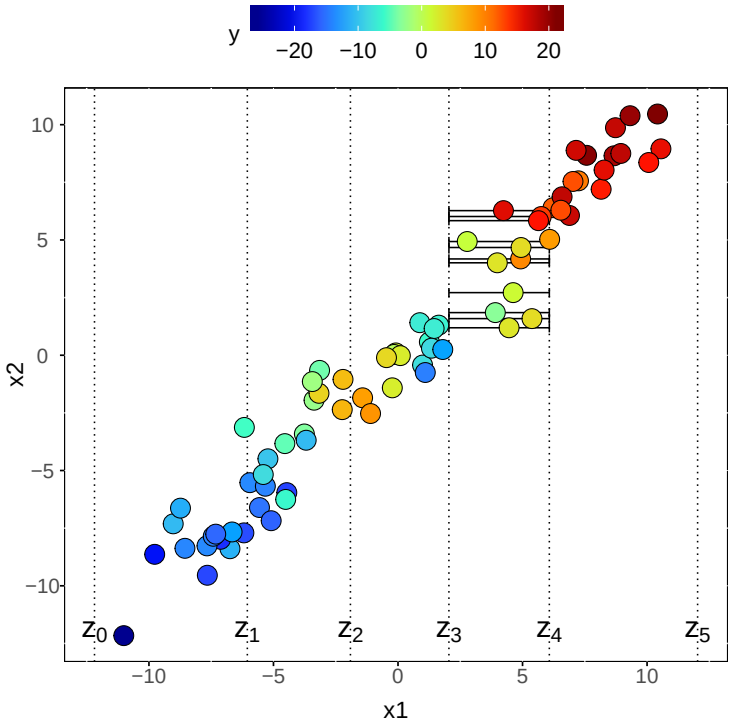


Abbildung 2.3.2.: Bildung paarweiser finiter Differenzen über die Intervallschachtelung von x_1. Je Beobachtung werden zwei neue Vektoren mit substituierten Werten von x_1 gebildet. Der Wert von x_1 wird jeweils durch die rechte bzw. linke Intervallgrenze substituiert. Anschließend wird die finite Differenz zwischen der Vorhersage mit rechts substituiertem Wert und links substituiertem Wert ermittelt. x_2 stellt hierbei eine Störvariable dar und bleibt je finiter Differenz konstant.

Der Durchschnitt (ii) aller finiten Differenzen je Intervall stellt den geschätzten intervallweiten lokalen Effekt dar. Anschließend werden die geschätzten durchschnittlichen finiten Differenzen (iii) zum ALE erster Ordnung kumulativ aufsummiert.

Algorithmus 7: Schätzung des ALE erster Ordnung

Data: Datenmatrix X, angepasstes Modell $\hat{f}$, selektierter Prädiktor x_S, Anzahl an Intervallen K
Result: Funktion $\widehat{f}_{ALE}(x_S)$

Initialisiere;
N = Zeilenanzahl von X `// Anzahl an Beobachtungen`
Berechne $K+1$ Quantile Z^* von x_S ;
ALE = Vektor mit Länge $(K+1)$;
forall $j \in 1 : (K+1)$ **do**
    `// Iteriere über jedes Quantil`
    Konstruiere Intervall $[z_{j-1}, z_j]$ aus den Werten von Z^*;
    Subset X_{Subset} von X $= X[x_S \geq z_{j-1} \wedge x_S \leq z_j, \quad]$ `// Teilmenge der Beobachtungen innerhalb des Intervalls`
    $X_{Subset,rechts} = X_{Subset}$;
    $X_{Subset,rechts}[\ , x_S] = z_j$;
    `// Setze Werte des selektierten Prädiktors aller Beobachtungen innerhalb des Intervalls auf die rechte Intervallgrenze`
    $X_{Subset,links} = X_{Subset}$;
    $X_{Subset,links}[\ , x_S] = z_{j-1}$;
    `// Setze Werte des selektierten Prädiktors aller Beobachtungen innerhalb des Intervalls auf die linke Intervallgrenze`
    FinDiffs $= \hat{f}(X_{Subset,rechts}) - \hat{f}(X_{Subset,links})$ `// Berechne geschätzte finite Differenzen innerhalb des Intervalls über die Differenz der Vorhersagen mit rechts und links substituierten Werten` x_S
    $LE = mean$(FinDiffs) `// geschätzter lokaler Effekt innerhalb des Intervalls ist durch die durchschnittliche finite Differenz der Vorhersage gegeben`
    $ALE[j] = ALE[j-1] + LE$ `// ALE am j-ten Quantil enstpricht ALE am (j-1)-ten Quantil zuzüglich des geschätzten lokalen Effekts in j-tem Quantil`
end
return Z^*, ALE

2.3.2. Additive Unverzerrtheit des ALE erster Ordnung

Der ALE erster Ordnung ist bezüglich additiv verknüpfter Effekte anderer Prädiktoren unverzerrt (Apley, 2016). Sei $\hat{f}(x_S, x_C)$ die Responsefunktion eines trainierten Supervised-Learning-Modells. Es existiere nur ein Haupteffekt des Prädiktors x_S auf $f(x_S, x_C)$ und Effekte erster sowie zweiter Ordnung der Prädiktoren in C, d.h. die Effektstruktur (Def. 2.0.5 auf Seite 15) lautet

$$f(x_S, x_C) = h_0 + g_S(x_S) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j), \quad i, j \in C$$

Proposition 2.3.1 (Additive Unverzerrtheit des ALE erster Ordnung)**.** *Treten keine Effekte höherer Ordnung des Prädiktors x_S auf, ist der geschätzte ALE erster Ordnung des Prädiktors x_S unverzerrt und blockiert Effekte erster sowie höherer Ordnung eventueller Störvariablen.*

$$\begin{aligned}
\mathbb{E}\left[\widehat{LE}(x_S = z_j)\right] &= \mathbb{E}\left[\hat{f}(x_S = z_j, x_C) - \hat{f}(x_S = z_{j-1}, x_C)\right] = \\
&= \mathbb{E}\left[\left(h_0 + g_S(j) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j)\right) - \right. \\
&\quad \left. - \left(h_0 + g_S(j-1) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j)\right)\right] = \\
&= [g_S(z_j) - g_S(z_{j-1})]
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\widehat{ALE}(x_S = z_k)\right] = \\
&= \mathbb{E}\left[\widehat{LE}(z_1) + \widehat{LE}(z_2) + \widehat{LE}(z_3) + \cdots + \widehat{LE}(z_k)\right] = \\
&= \mathbb{E}\left[\widehat{LE}(z_1)\right] + \mathbb{E}\left[\widehat{LE}(z_2)\right] + \mathbb{E}\left[\widehat{LE}(z_3)\right] + \cdots + \mathbb{E}\left[\widehat{LE}(z_k)\right] \\
&= g_S(z_1) - g_S(z_0) + g_S(z_2) - g_S(z_1) + g_S(z_3) - g_S(z_2)] + \cdots + g_S(z_k) - g_S(z_{k-1}) \\
&= g_S(z_k) - g_S(z_0) \\
&= \int_{z_0}^{z_k} g'(x_S) dx_S
\end{aligned}$$

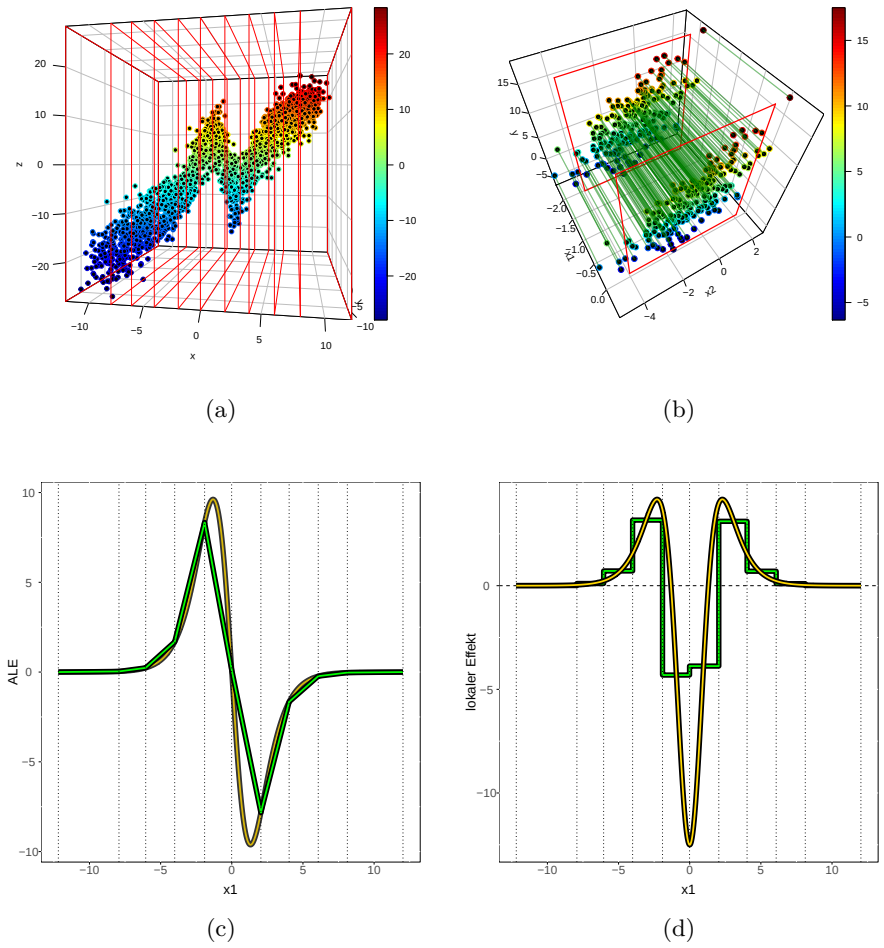


Abbildung 2.3.3.: Im mehrdimensionalen Fall additiver Haupteffekte blockiert der ALE den Effekt der Störvariablen. Die finiten Differenzen approximieren die partielle Ableitung, hier repräsentiert durch die dritte Ableitung der Sigmoid-Funktion. Der ALE entspricht der zweiten Ableitung der Sigmoid-Funktion.

Abbildung 2.3.3 zeigt die Schätzung für Datenmodell 2.3.1 auf Seite 65. Über die Bildung finiter Differenzen (b) wird der quadratische Effekt von y auf die Zielvariable herausgerechnet. Die finiten Differenzen approximieren den wahren lokalen Effekt (d). Die kumulierte Summe der finiten Differenzen approximiert das Integral des wahren lokalen Effektes von x auf $f(x, y)$ (c).

2.3.3. Multiplikative (Un-)Verzerrtheit des ALE erster Ordnung

Der ALE erster Ordnung ist bezüglich multplikativ verknüpfter Effekte mit x_S bis auf eine multiplikative Konstante unverzerrt, falls die Prädiktoren stochastisch unabhängig sind (Apley, 2016). Sei die Responsefunktion durch einen alleinigen Prädiktoreffekt zweiter Ordnung $f(x_1, x_2) = x_1 x_2$ gegeben. Der lokale Effekt von x_1 enstpricht der partiellen Ableitung $\frac{\partial x_1 x_2}{\partial x_1} = x_2$. Der ALE von x_1 wird durch drei Intervalle geschätzt:

$$\widehat{LE_1} = \frac{1}{n(z_1)} \sum_{i=1}^{N(z_1)} z_1 x_2^{(i)} - z_0 x_2^{(i)} = \frac{1}{n(z_1)} \sum_{i=1}^{N(z_1)} x_2^{(i)}(z_1 - z_0) = \overline{x_2}_{|x_1=z_1}(z_1 - z_0)$$

$$\widehat{LE_2} = \frac{1}{n(z_2)} \sum_{i=1}^{N(z_2)} z_2 x_2^{(i)} - z_1 x_2^{(i)} = \frac{1}{n(z_2)} \sum_{i=1}^{N(z_2)} x_2^{(i)}(z_2 - z_1) = \overline{x_2}_{|x_1=z_2}(z_2 - z_1)$$

$$\widehat{LE_3} = \frac{1}{n(z_3)} \sum_{i=1}^{N(z_3)} z_3 x_2^{(i)} - z_2 x_2^{(i)} = \frac{1}{n(z_3)} \sum_{i=1}^{N(z_3)} x_2^{(i)}(z_3 - z_2) = \overline{x_2}_{|x_1=z_3}(z_3 - z_2)$$

$$\begin{aligned}
&\mathbb{E}\left[\widehat{ALE}(x_1 = z_3)\right] \\
&= \mathbb{E}\left[\widehat{LE_1}\right] + \mathbb{E}\left[\widehat{LE_2}\right] + \mathbb{E}\left[\widehat{LE_3}\right] = \\
&= \mathbb{E}\left[\overline{x_2}|x_1 = z_1\right](z_1 - z_0) + \mathbb{E}\left[\overline{x_2}|x_1 = z_2\right](z_2 - z_1) + \mathbb{E}\left[\overline{x_2}|x_1 = z_3\right](z_3 - z_2) \\
&\overset{x_1 \perp\!\!\!\perp x_2}{=} \mathbb{E}\left[\overline{x_2}\right](z_1 - z_0 + z_2 - z_1 + z_3 - z_2) \\
&= (z_3 - z_0)\mathbb{E}\left[\overline{x_2}\right] \\
&= (z_3 - z_0) \times const.
\end{aligned}$$

Sind die Prädiktoren stochastisch unabhängig, hängt die Verteilung der Störvariable x_2 nicht von der Verteilung von x_1 ab. Es gilt: $\mathbb{E}\left[x_2|x_1\right] = \mathbb{E}\left[x_2\right]$. Somit ist der erwartete Wert der Störvariable in allen Intervallen identisch.

Sei nun der totale Prädiktoreffekt gegeben durch je einen zusätzlichen multiplikativ verknüpften Term (Interaktionseffekt) mit den jeweiligen anderen Störvariablen in C.

$$f(x_S, x_C) = h_0 + g_S(x_S) + \sum_{l \in C} g_{Sl}(x_S, x_l) + \sum_{l \in C} g_l(x_l) + \sum_{i \neq l} g_{il}(x_i, x_l)$$
$$i, l \in C,\ i \neq S$$

Proposition 2.3.2 (Multiplikative (Un-)Verzerrtheit des ALE erster Ordnung)**.** *Die paarweise finite Differenz im Intervall $[z_{j-1}, z_j]$ entspricht der paarweisen finiten Differenz des Haupteffektes von x_S zuzüglich paarweiser finiter Differenzen von Effekten höherer Ordnung mit x_S. Sind die Prädiktoren stochastisch abhängig, sind die paarweisen finiten Differenzen der Interaktionseffekte verzerrt (Apley, 2016).*

$$
\begin{aligned}
&\hat{f}(x_S = z_j, x_C) - \hat{f}(x_S = j-1, x_C) = \\
&= \left[h_0 + g_S(z_j) + \sum_{l \in C} g_{Sj}(z_j, x_l) + \sum_{l \in C} g_l(x_l) + \sum_{i \neq l} g_{il}(x_i, x_l) \right] - \\
&\quad - \left[h_0 + g_S(z_{j-1}) + \sum_{l \in C} g_{Sj}(z_{j-1}, x_l) + \sum_{l \in C} g_l(x_l) + \sum_{i \neq l} g_{il}(x_i, x_l) \right] = \\
&= g_S(z_j) - g_S(z_{j-1}) + \sum_{l \in C} \left[g_{Sl}(z_j, x_l) - g_{Sl}(z_{j-1}, x_l) \right]
\end{aligned}
$$

Apley (2016) führt an, dass eine multiplikative Unverzerrtheit bei stochastisch unabhängigen Prädiktoren wünschenswert sei. Bei stochastisch abhängigen Variablen sei es jedoch fragwürdig, inwiefern geschätzte Effekte erster Ordnung überhaupt eine Aussagekraft besitzen.

2.3.4. Verhältnis des ALE (erster Ordnung) zur numerischen Differenzierung

Die Effektidentifikation des ALE erster Ordnung weist die gleiche Struktur auf, wie die des zentralen Differenzenquotienten (Prop. 2.1.1 auf Seite 24). Der zentrale Differenzenquotient (Def. 2.1.6 auf Seite 19) im Intervall $[x_S - h, x_S + h]$ wurde definiert als:

$$
\frac{\partial f(x_S, x_C)}{\partial x_S} \approx \frac{f(x_S + h, x_C) - f(x_S - h, x_C)}{2h}, \quad h > 0
$$

Der Zähler stellt die finite Differenz des zentralen Differenzenquotienten dar. Demgegenüber lautet die finite Differenz als Approximation des lokalen Effektes im Intervall $[z_{j-1}, z_j]$:

$$
\hat{f}(x_S = z_j, x_C) - \hat{f}(x_S = z_{j-1}, x_C)
$$

Für $x_S + h = z_j$ und $x_S - h = z_{j-1}$ sind die finiten Differenzen identisch. Dies kann nur erfüllt sein, falls das Intervall symmetrisch um den betrachteten Wert liegt. Die

finite Differenz des ALE entspricht einer modifizierten finiten Differenz des zentralen Differenzenquotienten mit unterschiedlichen Vorwärts-/ und Rückwärtsdifferenzen:

$$\frac{\partial f(x_S, x_C)}{\partial x_S} \approx \frac{f(x_S + h_v, x_C) - f(x_S - h_r, x_C)}{h_v + h_r}, \quad h > 0$$

Abbildung 2.3.6 auf Seite 76 zeigt die Entwicklung bei einer iterativ feiner gewählten Intervallschachtelung. Zum Schluss existieren für N Beobachtungen N Intervalle (gemäß ihrer Quantile, siehe Sektion 2.3.6 auf der nächsten Seite). Für jeweils zwei Beobachtungen i, j mit $x_S^{(i)} - \mathrm{x}_S^{(j)} \approx 0$, d.h. die Werte von x_S liegen sehr nah aneinander, gilt:

$$h_v, h_r \ \to 0 \quad , x_S \to z_j \quad , x_S \to z_{j-1}$$

Der Nominator des Differenzenquotienten des marginalen Effektes entspricht in diesem Fall approximativ dem lokalen Effekt:

$$f(x_S + h_v, x_C) - f(x_S - h_r, x_C) \approx f(z_j, x_C) - f(z_{j-1}, x_C)$$

Die finiten Differenzen des ALE sind nur im Falle gleich breiter Intervalle eine Approximation der Funktion der ersten Ableitung. Variiert die Intervallbreite, ist für die Approximation der Ableitung eine Division der finiten Differenz durch die jeweilige Intervallbreite erforderlich. Für die anschließende Integration nach x_S reicht es jedoch, die finiten Differenzen aufzusummieren.

2.3.5. Zentrierung des ALE erster Ordnung

Die vertikale Zentrierung des Plots erfolgt durch $C = const$ (Def. 2.3.1 auf Seite 65). Apley (2016) schlägt vor, die Zentrierungskonstante so zu wählen, sodass $f_{S,ALE}(x_S)$ einen Mittelwert von 0 bezüglich der marginalen Verteilung von x_S besitzt:

$$\begin{aligned} f_{S,ALE}(x_S) &= \hat{f}_{S,ALE}(x_S) - \mathbb{E}\left[\hat{f}_{S,ALE}(x_S)\right] = \\ &= \hat{f}_{S,ALE}(x_S) - \int p_S(z_S)\, \hat{f}_{S,ALE}(z_S) dz \end{aligned}$$

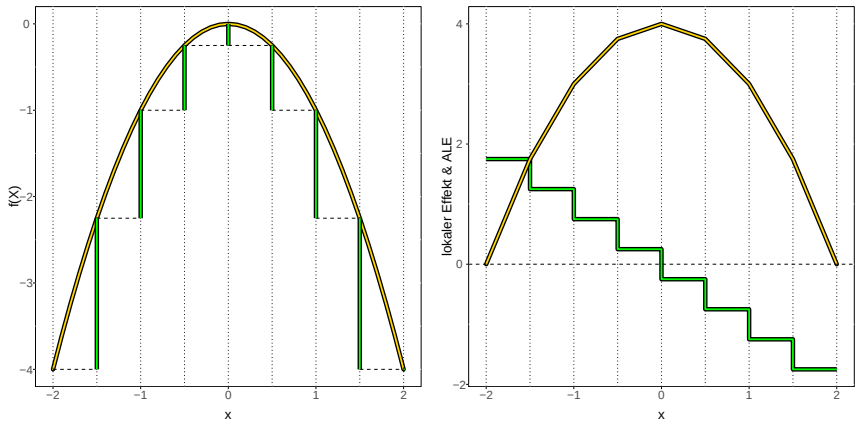
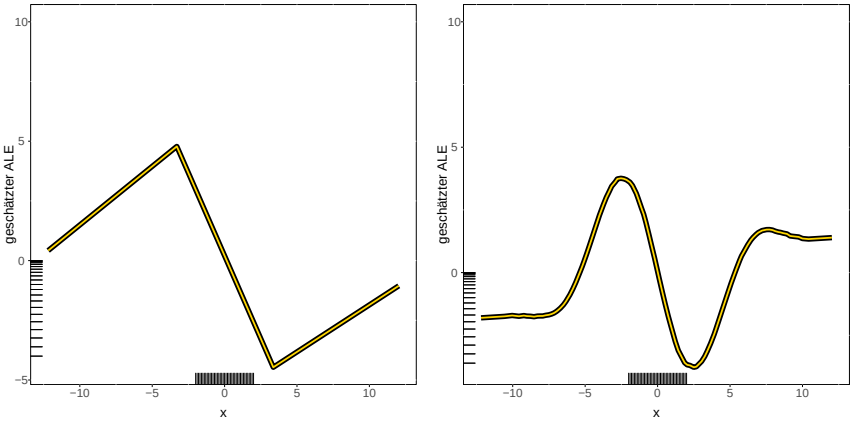


Abbildung 2.3.4.: Die Akkumulation paarweiser Differenzen approximiert im univariaten Fall die Responsefunktion bis auf eine additive Konstante. Diese kann über eine anschließende Zentrierung ausgeglichen werden.

2.3.6. Schätzgenauigkeit des lokalen Effektes



(a) (b)

Abbildung 2.3.5.: Schätzung des ALE für drei sowie 100 Intervalle. Mehr Intervalle liefern eine bessere Approximation an den wahren ALE.

Die Schätzgenauigkeit des ALE wird von zwei Faktoren determiniert:

(i) Über die gewählte Intervallschachtelung wird die partielle Ableitung $\frac{\partial \hat{f}(x_S, x_C)}{\partial x_S}$ approximiert. Stärkere Effektänderungen erfordern eine feinere Intervallschachtelung. Abbildung 2.3.5 zeigt eine Schätzung des ALE durch das *ALEPlot*-Paket mit jeweils drei sowie 100 Intervallen. Eine stärkere Intervallschachtelung liefert eine bessere Approximation an den wahren ALE. Apley (2016) betrachtet die Quantile von x_S und teilt Bereichen mit einer hohen Konzentration mit

beobachteten Werten eine feinere Intervallschachtelung zu als Bereichen mit einer geringen Datenkonzentration.

(ii) Die Approximation des angepassten Modells $\hat{f}(x_S, x_C)$ an das wahre Modell $f(x_S, x_C)$. Nur im Falle eines guten Modells stellt die Schätzung in (i) eine Approximation der partiellen Ableitung des wahren Modells $\frac{\partial f(x_S, x_C)}{\partial x_S}$ dar.

2.3.7. Wahl der Intervallschachtelung

Die Problematik der von Apley (2016) gewählten Intervallschachtelung wird in Abbildung 2.3.6 auf der nächsten Seite visualisiert. Links ist jeweils der wahre ALE in schwarzer Farbe und in grüner Farbe der geschätzte ALE zu sehen. Rechts sind die intervallweisen finiten Differenzen in grüner Farbe und der echte lokale Effekt in schwarzer Farbe gegeben. Der ALE, gegeben durch die zweite Ableitung der Sigmoid-Funktion, ist in Bereichen mit wenigen Beobachtungen stark gekrümmt. Folglich wird gemäß den Quantilen von x_S in diesen eine große Intervallbreite gewählt. Die finiten Differenzen können in den breiten Intervallen den lokalen Effekt nicht approximieren. Folglich ist der intervallweise geschätzte ALE divergent vom wahren ALE. Auch durch die Spezifizierung einer feineren Intervallschachtelung kann das Problem nicht gelöst werden.

Würde die Intervallschachtelung optimal gelöst, könnte das Supervised-Learning-Modell jedoch trotzdem nicht zuverlässig über die Trainingsdaten hinaus extrapolieren. Es kann nicht davon ausgegangen werden, dass finite Differenzen auf der Grundlage des Modells in datenarmen Regionen eine Approximation der wahren partiellen Ableitung sind. Hierzu wäre eine ausreichende Datenmenge im Trainingsprozess notwendig. Eine Indikation von Stellen mit Extrapolitionsgefahr ist bei der Evaluierung von Accumulated Local Effects wünschenswert. Einen solchen Indikator stellt beispielsweise ein *Teppich (Rug Plot)* der beobachteten Werte x_S auf der horizontalen Achse dar. Sind die Prädiktorvariablen voneinander abhängig, ist die marginale Verteilung von x_S nicht ausreichend, um Extrapolationsregionen zu erkennen. Weiterführende Ansätze zur Extrapolationsdetektion im Machine-Learning-Kontext werden beispielsweise in Hooker und Friedman (2004) beschrieben.

2.3.8. Laufzeitverhalten von Accumulated Local Effects

Für N Beobachtungen werden N finite Differenzen auf der Grundlage von jeweils zwei Prädiktionen berechnet. Der Rechenaufwand ist unabhängig von der gewählten Intervallschachtelung. Für jede Beobachtung wird in jedem Fall genau eine finite Differenz gebildet. Die Worst-Case Laufzeit des ALE entspricht (Apley, 2016):

$$f_{\text{ALE}} \in \mathcal{O}(N)$$

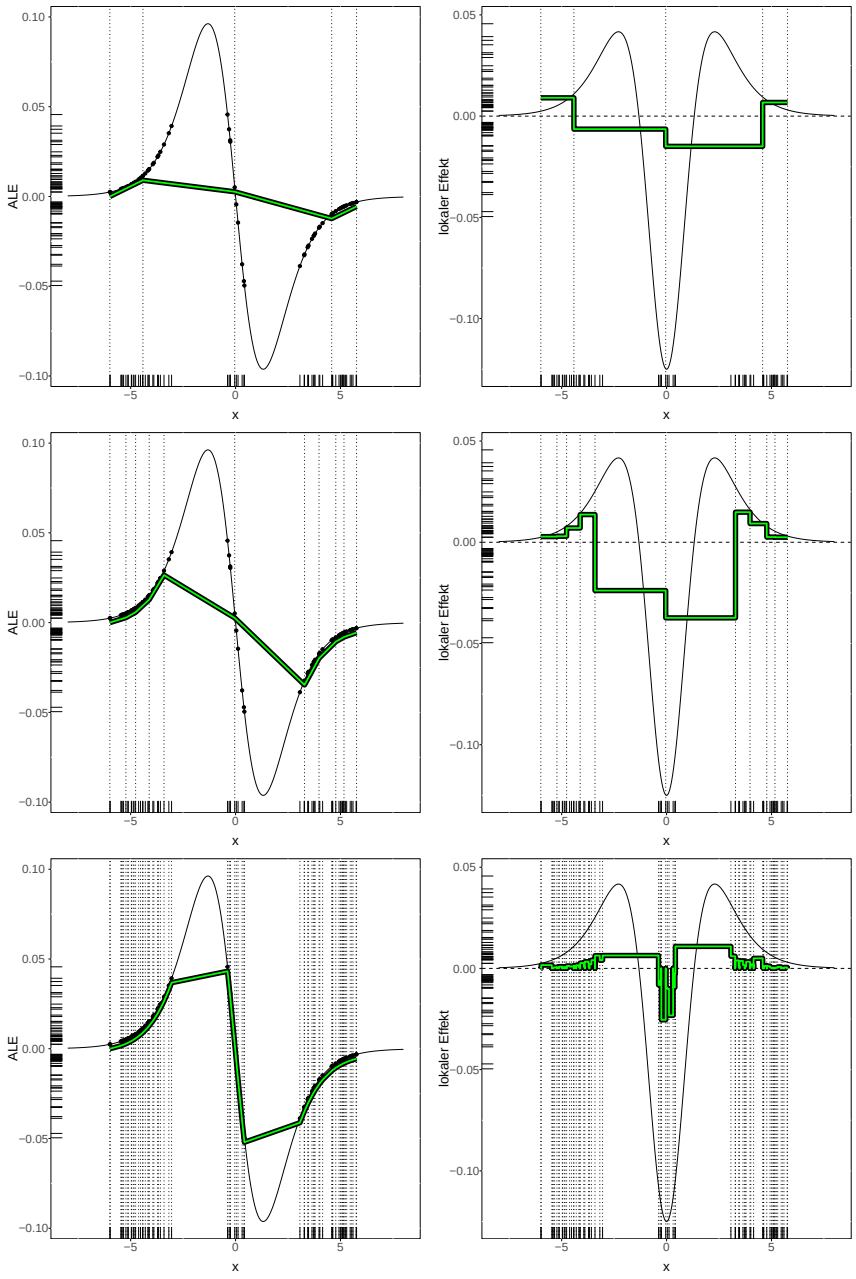


Abbildung 2.3.6.: Aufgrund der Quantile kann der wahre ALE auch durch eine höhere Intervallzahl nicht gänzlich approximiert werden (jeweils links). Die finiten Differenzen approximieren den lokalen Effekt sehr ungenau (jeweils rechts).

2.3.9. Accumulated Local Effects zweiter Ordnung

Im Falle existierender Interaktionseffekte und einer stochastischen Abhängigkeit zwischen den betroffenen Prädiktoren ist die Aussagekraft des ALE erster Ordnung fragwürdig (Apley, 2016). Der ALE zweiter Ordnung stellt die Erweiterung des ALE auf einen bivariaten Einfluss dar. S wird nun durch die beiden Prädiktorvariablen S_1 und S_2 repräsentiert.

Definition 2.3.2 (ALE zweiter Ordnung)**.** *Mit der partiellen Ableitung der Responsefunktion $\frac{\partial f(x_{S_1}, x_{S_2}, x_C)}{\partial x_{S_1} \partial x_{S_2}}$ nach x_{S_1} und x_{S_2} als der stetigen Funktion des bivariaten lokalen Effektes von x_{S_1} und x_{S_2} auf $f(x_{S_1}, x_{S_2}, x_C)$ wird der ALE zweiter Ordnung definiert als*

$$
\begin{aligned}
f_{\{S_1,S_2\},ALE}(x_{S_1,S_2}) = &\int_{z_{0,S_1}}^{x_{S_1}} \int_{z_{0,S_2}}^{x_{S_2}} \mathbb{E}\left[\frac{\partial f(z_{S_1}, z_{S_2}, x_C)}{\partial z_{S_1} \partial z_{S_2}}\right] dz_{S_1} dz_{S_2} \\
&- f_{S_1,ALE}(x_{S_1}) \\
&- f_{S_2,ALE}(x_{S_2}) \\
f_{\{S_1,S_2\},ALE}(x_{S_1,S_2}) = &\int_{z_{0,S_1}}^{x_{S_1}} \int_{z_{0,S_2}}^{x_{S_2}} \int p(x_C|z_{S_1}, z_{S_2}) \left[\frac{\partial f(z_{S_1}, z_{S_2}, x_C)}{\partial z_{S_1} \partial z_{S_2}}\right] dx_C dz_{S_1} dz_{S_2} \\
&- f_{S_1,ALE}(x_{S_1}) \\
&- f_{S_2,ALE}(x_{S_2})
\end{aligned}
$$

Der ALE zweiter Ordnung wird analog zum ALE erster Ordnung über die Bildung finiter Differenzen geschätzt. Diese werden nun auf einem zweidimensionalen Gitter gebildet (Abb. 2.3.7 auf der nächsten Seite).

Definition 2.3.3 (Schätzung des ALE zweiter Ordnung)**.** *Das Rechteck, das durch die Intervallgrenzen $x_{S_1} = [z_{j-1}, z_j]$ und $x_{S_2} = [z_{m-1}, z_m]$ festgelegt wird, sei mit den Indices $\{j, m\}$ bezeichnet. $N(z_j, z_m)$ bezeichne die Indices der Beobachtungen, die in das Rechteck fallen. $n(z_j, z_m)$ bezeichne die Anzahl der Beobachtungen im betroffenen Rechteck. Der geschätzte lokale Effekt innerhalb des Rechtecks lautet:*

$$
\begin{aligned}
\widehat{LE}(x_{S_1} = z_j, x_{S_2} = z_m) = &\frac{1}{n(z_j, z_m)} \sum_{i \in N(z_j, z_m)} \\
&- \Big[\Big(\hat{f}(x_{S_1} = z_j, x_{S_2} = z_m, x_C^{(i)}) - \hat{f}(x_{S_1} = z_{j-1}, x_{S_2} = z_m, x_C^{(i)})\Big) \\
&- \Big(\hat{f}(x_{S_1} = z_j, x_{S_2} = z_{m-1}, x_C^{(i)}) - \hat{f}(x_{S_1} = z_{j-1}, x_{S_2} = z_{m-1}, x_C^{(i)})\Big)\Big]
\end{aligned}
$$

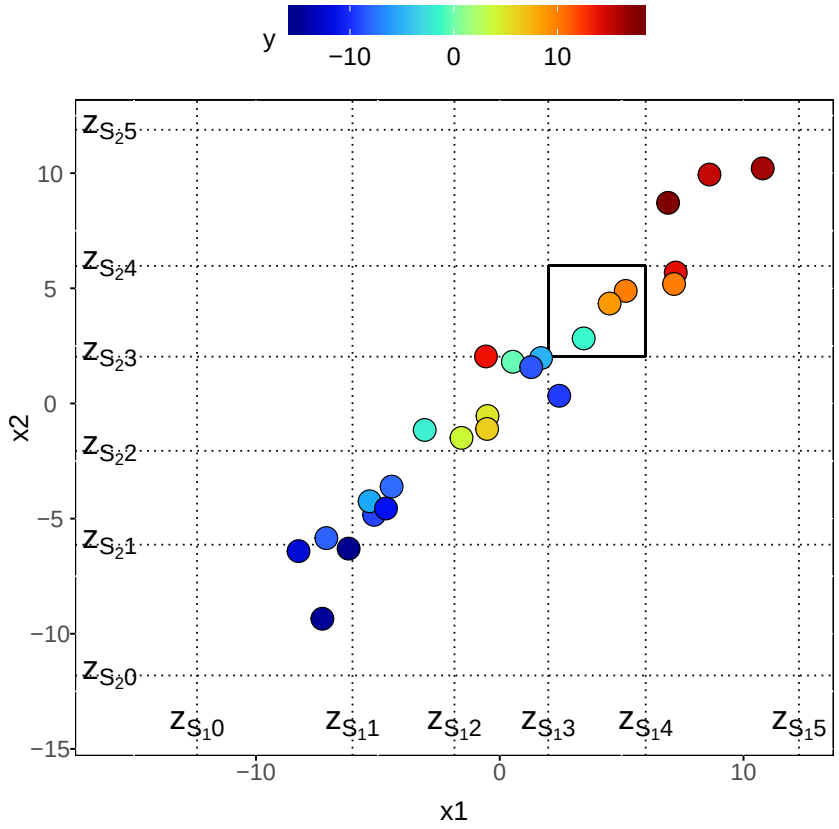


Abbildung 2.3.7.: Bildung finiter Differenzen über ein zweidimensionales Gitter.

Die finiten Differenzen zweiter Ordnung werden durch die finiten Differenzen aus ALE erster Ordnung an der oberen Intervallgrenze des Rechtecks und des ALE erster Ordnung an der unteren Intervallgrenze des Rechtecks repräsentiert. Die kumulativen bivariaten finiten Differenzen am Punkt $(x_{S_1} = j, x_{S_2} = m)$ *lauten:*

$$
\begin{aligned}
(*) &= \sum_{k=1}^{z_j, S_1} \sum_{l=1}^{z_m, S_2} \widehat{LE}(x_{S_1} = z_k, x_{S_2} = z_l) \\
&= \sum_{k=1}^{z_j, S_1} \sum_{l=1}^{z_m, S_2} \frac{1}{n(z_k, z_l)} \sum_{i \in N(z_k, z_l)} \\
&\quad - \Big[\Big(\hat{f}(x_{S_1} = z_k, x_{S_2} = z_l, x_C^{(i)}) - \hat{f}(x_{S_1} = z_{k-1}, x_{S_2} = z_l, x_C^{(i)})\Big) \\
&\quad - \Big(\hat{f}(x_{S_1} = z_k, x_{S_2} = z_{l-1}, x_C^{(i)}) - \hat{f}(x_{S_1} = z_{k-1}, x_{S_2} = z_{l-1}, x_C^{(i)})\Big)\Big]
\end{aligned}
$$

Die kumulativen bivariaten finiten Differenzen aus $(*)$ *werden anschließend von den akkumulierten lokalen Effekten erster Ordnung bereinigt:*

$$(*) - \widehat{ALE}(x_{S_1}) - \widehat{ALE}(x_{S_2})$$

Der ALE zweiter Ordnung kann im Gegensatz zur bivariaten PD ausschließlich in Kombination mit beiden ALE erster Ordnung interpretiert werden.

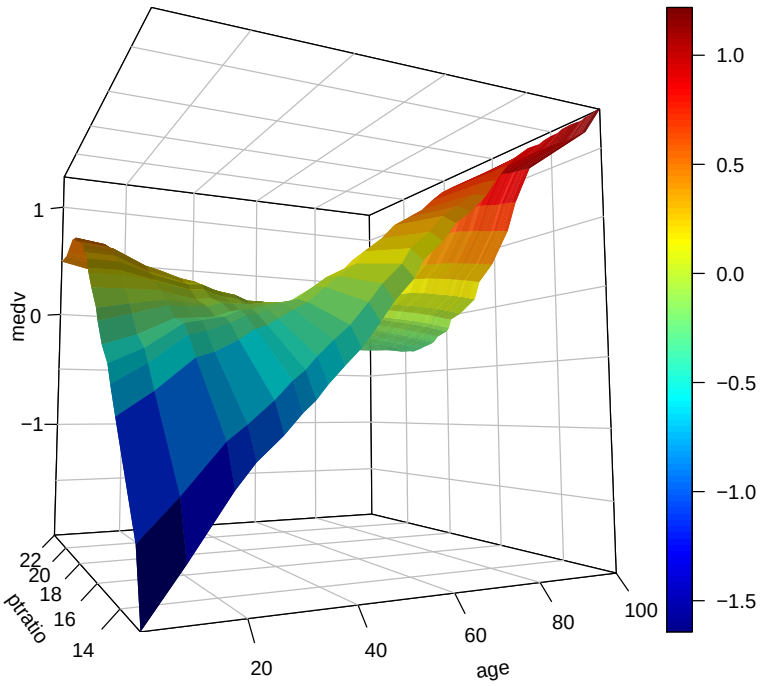


Abbildung 2.3.8.: Der ALE zweiter Ordnung ist indikativ für Interaktionseffekte zwischen *age* und *ptratio* im Boston-Housing-Datensatz.

KAPITEL 3.

Ein generalisiertes System zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen

„Particular facts are never scientific. Only generalization can establish science.“

Claude Bernard

Die in Kapitel 2 vorgestellten Verfahren verwenden eine gemeinsame Methodik. Aufgrund der Black-Box-Eigenschaft der betrachteten Supervised-Learning-Modelle ist eine interne Betrachtung der Funktionsweise nicht möglich. Zur Bestimmung von Prädiktoreffekten wird nicht versucht, die Black-Box „zu öffnen“. Stattdessen wird das Verhalten der Black-Box bei Veränderung der Eingangsgrößen betrachtet.

3.1. Die Konstruktion eines generalisierten Systems aus gemeinsamen Arbeitsschritten

Der Arbeitsablauf der vorgestellten Verfahren lässt sich auf eine Sequenz gemeinsamer Schritte reduzieren. Die gemeinsamen Verfahrensschritte bilden das generalisierte System zur Bestimmung von Prädiktoreffekten in (Black-Box-)Supervised-Learning-Modellen.

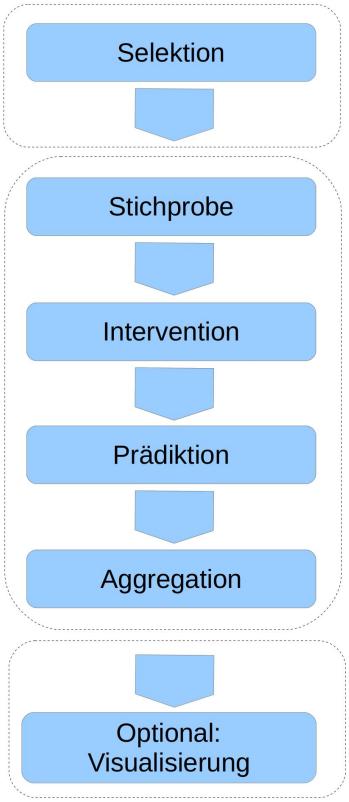


Abbildung 3.1.1.: Allgemeines System zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen. Selektions-/ und Visualisierungsschritt sind vom Kern des Systems abgekoppelt. Für die vorgestellten Verfahren wird der betrachtete Prädiktor als gegeben betrachtet. Die Visualisierung ist optional.

Zuerst wird ein Prädiktor S in einem *Selektionsschritt (Selection Step)* ausgewählt. Die Selektion eines relevanten Prädiktors ist vor allem bei einer großen Menge an Variablen problematisch. Da in der Regel die Datenmenge sehr groß ist, wird in einem *Stichprobenschritt (Sampling Step)* eine Stichprobe ausgewählt. In einem *Interventionsschritt (Intervention Step)* wird in die Stichprobe eingegriffen. Der Datensatz muss verändert werden, um das Verhalten der Black-Box zu analysieren. Entweder werden die Werte permutiert (ICE und PD), oder Werte werden direkt verändert (ME und ALE). Die intervenierten Werte werden im *Prädiktionsschritt (Prediction Step)* in die Responsefunktion eingesetzt und eine Vorhersage berechnet. Hier kommt der post-hoc-Aspekt der Thematik zum Tragen, da für den Prädiktionsschritt ein bereits angepasstes Modell notwendig ist. Die Werte können in einem weiterführenden *Aggregationsschritt (Aggregation Step)* aggregiert werden. Zuletzt kann das Ergebnis in einem optionalen *Visualisierungsschritt (Visualization Step)* graphisch dargestellt werden.

Für die Einbettung der Verfahren in das Rahmenkonzept werden diese in sequentielle Berechnungsschritte zerlegt (Abb. 3.1.2 auf Seite 85). Alle Verfahren benötigen zuerst eine Stichprobe aus den vorliegenden Daten. Anschließend folgen unterschiedliche Interventionen und Aggregationen.

3.1.1. Einbettung des (intervallweisen) AME in das generalisierte System

Für den intervallweisen AME ist die Intervallschachtelung des Intervalls $[min(x_S), max(x_S)]$ vorangestellt. Anschließend folgt die Berechnung marginaler Effekte auf den im jeweiligen Intervall befindlichen Werten. Der marginale Effekt besitzt zwei Komponenten, die finite Differenz (Zähler) und die Intervallbreite (Nenner). Für die finite Differenz des ME wird ein kleines und symmetrisches Intervall um den betrachteten Wert konstruiert. Die Bedingung $x_S^{(i)} - v_R = x_S^{(i)} + v_R \approx x_S^{(i)}$ repräsentiert die numerische Approximation der *infinitesimalen Änderung* innerhalb des Differentialquotienten.

Im Prädiktionsschritt wird die finite Differenz aus den Vorhersagen der Vorwärtsdifferenz und der Rückwärtsdifferenz gebildet. Da der ME ein Steigungsmaß darstellt, muss die finite Differenz der Vorhersage durch die Intervallbreite $2h$ dividiert werden. Die ME je Intervall werden zum intervallweiten AME aggregiert (gemittelt). Eine weitere Aggregation der intervallweiten AME führt zum globalen AME. Der intervallweite AME repräsentiert eine Approximation der partiellen Ableitung der Prädiktionsfunktion nach der selektierten Variable x_S.

3.1.2. Einbettung des ALE erster Ordnung in das generalisierte System

Das Vorgehen zur Berechnung des ALE erster Ordnung ähnelt dem des intervallbasierten AME. Im Interventionsschritt wird ebenfalls eine Intervallschachtelung vorgenommen. Anschließend wird in allen Intervallen die durchschnittliche finite Differenz geschätzt. Für jede Beobachtung, die in ein Intervall fällt, wird eine Vorwärts-/ und Rückwärtsdifferenz gebildet. Diese ist im Gegensatz zum ME nicht symmetrisch und kann auch größere Werte annehmen. Die Vorwärtsdifferenz wird in jedem Fall durch den absoluten Abstand des betrachteten Wertes zur rechten Intervallgrenze repräsentiert. Analog wird mit der Rückwärtsdifferenz und der linken Intervallgrenze verfahren.

Im Prädiktionsschritt wird nun wie bei der finiten Differenzenbildung der Vorhersage für den ME vorgegangen. Die geschätzten lokalen Effekte je Beobachtung werden zum intervallweisen lokalen Effekt aggregiert (gemittelt). Da das Ziel die anschließende Integration nach der gleichen Variable ist, ist im Gegensatz zum ME keine Division durch die Intervallbreite notwendig. Wird diese jedoch trotzdem vorgenommen, was von Apley (2016) nicht beabsichtigt wurde, erhalten wir wie beim intervallbasierten AME eine intervallweise Approximation der partiellen Ableitung nach der selektierten Variable. Da die zugrundeliegenden finiten Differenzen verschieden sind, sind die Approximationen nicht identisch.

Sei nun angenommen, wir haben die finiten Differenzen der Prädiktion (FD) durch die jeweiligen Intervallbreiten (IB) dividiert und eine Approximation der partiellen Ableitung erhalten. Zum ALE ist zusätzlich die Integration nach der selektierten Variable notwendig. Da wir die partielle Ableitung intervallweise approximiert hatten, wird unsere Approximation durch eine Treppenfunktion repräsentiert. Für die Bildung des Integrals einer Treppenfunktion wird folgendermaßen vorgegangen. Je Treppenstufe liegt der Flächeninhalt eines Rechtecks vor. Wir müssen daher jede Stufenhöhe mit der jeweiligen Stufenbreite (IB) multiplizieren und die Produkte aufsummieren. Die Stufenhöhe des Intervalls $[z_{j-1}, z_j]$ wird im vorliegenden Fall durch den Quotienten

$$\frac{FD_j}{IB_j}$$

repräsentiert. Zur Integration über ein Intervall j bilden wir das Produkt

$$\frac{FD_j}{IB_j} IB_j = FD_j$$

Zur Integration der partiellen Ableitung nach x_S wiederum nach x_S genügt es daher, die finiten Differenzen FD_j über alle j aufzusummieren.

3.1.3. Einbettung von ICE & PD in das generalisierte System

Für die ICE ist keine Intervallschachtelung notwendig. Je Beobachtung werden im Interventionsschritt die Werte des selektierten Prädiktors permutiert. Im Prädiktionsschritt wird jede einzelne Permutation in die Prädiktionsfunktion eingesetzt. Der Vektor von Vorhersagen für eine Beobachtung repräsentiert eine einzelne ICE. Diese können anschließend punktweise zur PD aggregiert (gemittelt) werden.

3.2. Erweiterung des generalisierten Systems auf die Feature-Importance

Definition 3.2.1 (Feature-Importance)**.** *Die Feature-Importance beschreibt, wie wichtig ein Prädiktor für die Vorhersagekraft (Predictive-Performance) eines Modells ist (Casalicchio, Molnar und Bischl, 2018).*

Somit ist die Bestimmung eines Prädiktoreffektes (Richtung und Magnitude) eng verbunden mit der Importance (Beitrag zur Vorhersagekraft) eines Prädiktors. Die Auswahl von Prädiktoren (*Feature-Selection*) spielt bereits beim Modellanpassungsprozess eine große Rolle. Wir betrachten hingegen ein bereits angepasstes Modell und stellen uns die Frage nach der Relevanz der verwendeten Variablen für die Modellgüte.

Im allgemeinen linearen Modell wird als natürliche Feature-Importance der Quotient aus absolutem Wert des jeweiligen Koeffizienten und dem korrespondierendem Standardfehler verwendet, was dem absoluten Wert der t-Statistik entspricht (Greenwell, Boehmke und McCarthy, 2018). Für Black-Box-Modelle existieren weder analytische Formen der Koeffizienten, noch deren Standardfehler. Sektion 3.2.1 stellt Verfahren zur Bestimmung der Feature-Importance vor, die Parallelen zum vorgeschlagenen generalisierten System zur Bestimmung von Feature-Effects aufweisen.

3.2.1. Verfahren zur Bestimmung der Feature-Importance für ein spezifisches Modell

Krümmung der ICE

Eine flach verlaufende Trajektorie einer ICE bedeutet, dass im vorliegenden Modell die Veränderung des betrachteten Prädiktors für die jeweilige Beobachtung keinen Einfluss auf die Vorhersage der Zielvariable hat. Verlaufen die ICE-Kurven annähernd parallel, kann auch die Partial Dependence betrachtet werden (Sektion 2.2.1 auf Seite 37). Greenwell, Boehmke und McCarthy (2018) schlagen ein Krümmungsmaß der PD vor.

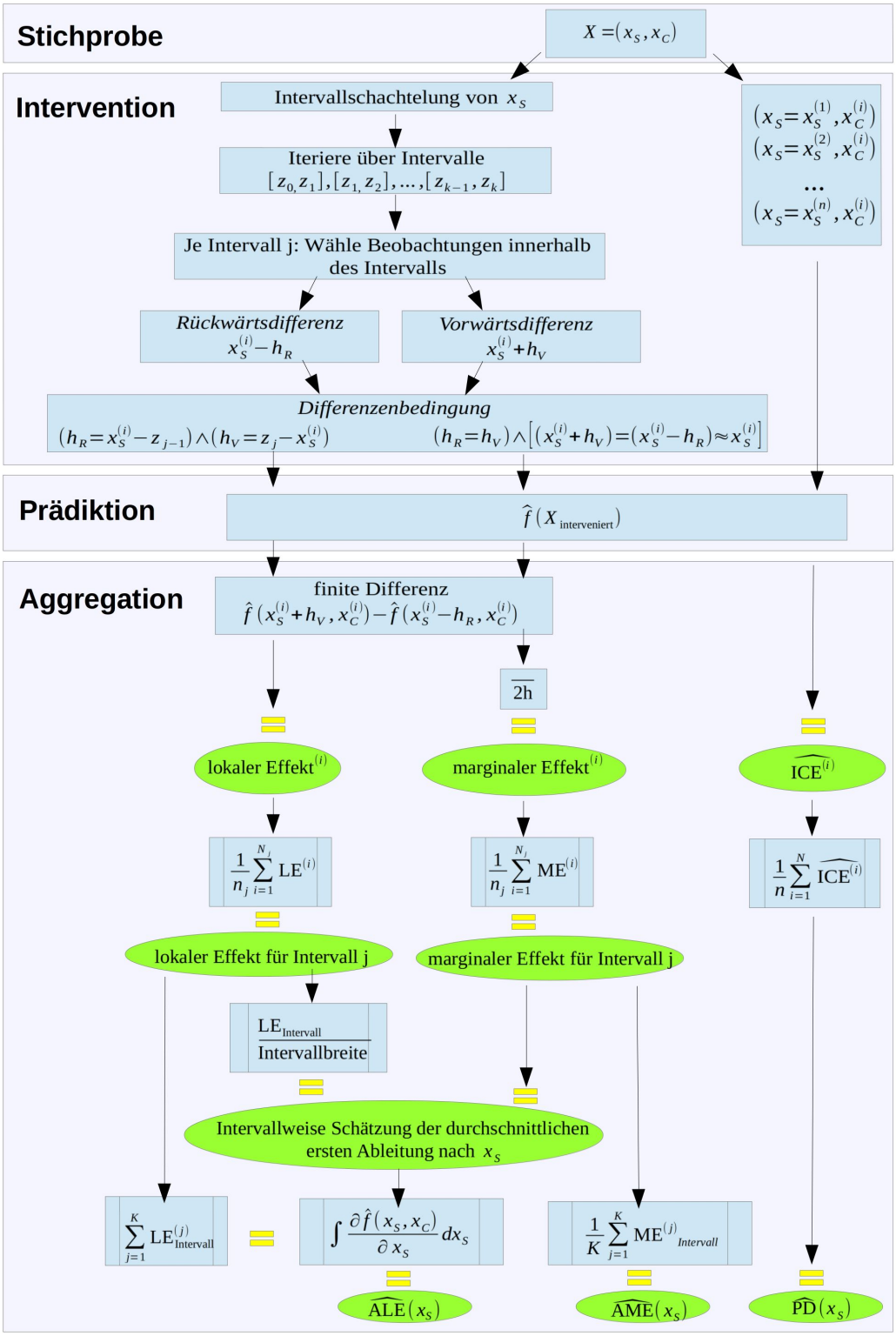


Abbildung 3.1.2.: Einbettung der vorgestellten Verfahren in das allgemeine Rahmenkonzept zur Bestimmung von Prädiktoreffekten. Intervallbasierte AME und der Quotient aus intervallweiten durchschnittlichen finiten Differenzen und der jeweiligen Intervallbreite sind zwei verschiedene Varianten, um die partielle Ableitung nach der selektierten Prädiktorvariable zu approximieren. Die beiden Verfahren sind nicht äquivalent.

Definition 3.2.2 (Krümmungsmaß der PD als Score der Feature-Importance)**.** *Wir notieren den durchschnittlichen Wert der PD von Variable x_j mit $\overline{\hat{f}_{PD}}(x_j) = \frac{1}{k}\sum_{i=1}^{k}\hat{f}_{PD}^{(j)}(x_j^{(i)})$. Die Importance von Variable x_j entspricht der geschätzten Standardabweichung ihrer PD-Funktion. Je flacher die PD verläuft, desto geringer die Standardabweichung und somit die Importance. Für kategoriale Variablen wird die Spannweite der PD durch 4 dividiert, was eine Schätzung der Standardabweichung für kleine bis mittlere Stichprobengrößen darstellt (Greenwell, Boehmke und McCarthy, 2018).*

$$Imp(x_j) = \begin{cases} \sqrt{\frac{1}{k-1}\sum\limits_{i=1}^{k}\left[\hat{f}_{PD}(x_j^{(i)}) - \overline{\hat{f}_{PD}}(x_j)\right]^2} & x_j \text{ kontinuierlich} \\ \frac{1}{4}\left[max_i\left\{\overline{f_{PD}}(x_j^{(i)})\right\} - min_i\left\{\overline{f_{PD}}(x_j^{(i)})\right\}\right] & x_j \text{ kategorial} \end{cases}$$

Die von Greenwell, Boehmke und McCarthy (2018) vorgeschlagene Metrik basiert ausschließlich auf der PD und ist im Falle von Interaktionseffekten kein zuverlässiges Maß der Feature-Importance. Ein geeigneteres Maß sollte die Standardabweichung der ICE betrachten.

Permutation Feature Importance, Individual Conditional Importance & Partial Importance

Die *Permutation Feature Importance [PFI]*, ursprünglich von Breiman (2001a) für Random Forests entwickelt, wurde von Fisher, Rudin und Dominici (2018) zu einem modellagnostischen Werkzeug erweitert, um die globale Feature-Importance eines Prädiktors für ein bereits angepasstes Modell zu bestimmen. Werden die Werte eines Prädiktors isoliert permutiert, wird der Zusammenhang zwischen dem Prädiktor und der Zielvariable durchbrochen. Falls der Prädiktor wichtig für die Vorhersagekraft des Modells war, resultiert die Permutation in einem erhöhten Verlust (Casalicchio, Molnar und Bischl, 2018). Die modellagnostische PFI einer Prädiktorvariablen x_S misst die Differenz des Verlustes auf Daten jeweils mit permutierten und nicht permutierten Werten für ein beliebiges Modell.

Casalicchio, Molnar und Bischl (2018) präsentieren die *Individual Conditional Importance [ICI]*, sowie die *Partial Importance [PI]* als Visualisierungsmöglichkeiten, die auf dem gleichen Prinzip wie ICE und PD basieren. Die ICI visualisiert den Einfluss eines Prädiktors auf die Modellgüte für eine individuelle Beobachtung, während die PI den durchschnittlichen Einfluss visualisiert. Die Schätzung der ICI entspricht der Vorgehensweise für die ICE, wenn im Prädiktionsschritt anstatt der

absoluten Vorhersage mit intervenierten Werten der Verlust zwischen Vorhersage mit beobachteten Werten und intervenierten Werten berechnet wird. Die j-te ICE-Kurve an der i-ten Beobachtung von x_S, notiert mit $\hat{f}^{(j)}_{ICE}(x^{(i)}_S)$, zeigt die Vorhersage der j-ten Beobachtung an, wobei der Wert von $x^{(j)}_S$ durch $x^{(i)}_S$ ersetzt wurde. Die j-te ICI-Kurve an der i-ten Beobachtung von x_S, notiert mit $\hat{f}^{(j)}_{ICI}(x^{(i)}_S)$, zeigt den Verlust für die j-te Beobachtung an, zwischen der Vorhersage mit beobachtetem Wert $x^{(j)}_S$ und der Vorhersage mit ersetztem Wert $x^{(i)}_S$. Die PI entspricht wiederum dem punktweisen Durchschnitt aller ICI je Achsenabschnitt, was der Monte-Carlo-Integration bei der Schätzung der PD entspricht.

In Abb. 3.2.1 auf Seite 91 sind die ICI und PI-Plots für die Variablen *LSTAT (Prozentualer Anteil der Bevölkerung mit niedrigem Status im betrachteten Distrikt)* und *RM (Durchschnittliche Anzahl der Räume je Wohnungseinheit im betrachteten Distrikt)* des *Boston-Housing-Datensatzes* gegeben. Die vertikale Achse zeigt die Differenz des MSE-Verlustes zwischen der Vorhersage mit beobachteten Werten und intervenierten Werten an. Je höher die Verluständerung durch die Intervention in die Daten, desto höher ist die der Variable zugewiesene Importance.

Der PI-Plot von *LSTAT* suggeriert, dass ab einem Anteil der Bevölkerung mit niedrigem Status von 10% oder weniger die globale Importance (für alle Beobachtungen des Datensatzes) von *LSTAT* steigt. Ab einem Anteil von 10% oder mehr hat die Intervention in die Werte von *LSTAT* keinen Einfluss auf die Performance. Der PI-Plot von *RM* sagt aus, dass ab einer durchschnittlichen Anzahl von 7 Räumen oder mehr die globale Importance von *RM* steigt. Darunter hat die Intervention in die Werte von *RM* keinen Einfluss auf die Performance. Das Integral der PI-Kurve stellt wiederum einen globalen Score (für alle Beobachtungen und alle Werte der Variable) der Importance dar und entspricht der PFI.

Die jeweiligen ICI-Plots dienen der Überprüfung, ob die Aggregation zur PI repräsentativ für die Importance der Prädiktorvariablen ist. Wir stellen fest, dass die Effekte sich unter den individuellen Beobachtungen unterscheiden, was für Interaktionseffekte spricht.

Shapley-Wert

Cohen, Ruppin und Dror (2005) stellten erstmals das Konzept des Shapley-Wertes aus der Spieltheorie als ein Verfahren zur Bestimmung der Feature-Importance vor. Casalicchio, Molnar und Bischl (2018) merken an, dass die Autoren sich auf die *Feature-Selection* konzentrieren, wobei das Modell mit verschiedenen Prädiktorvariablen neu angepasst wird. Dieses Vorgehen sei für die Betrachtung und Interpretation eines spezifischen Modells nicht zielführend. Casalicchio, Molnar und Bischl

(2018) schlagen die *Shapley Feature IMPortance [SFIMP]* vor, wobei über die restlichen Prädiktoren integriert wird, anstatt sie auszulassen. Die Autoren schätzen zwei Generalisierungsfehler. In der ersten Variante wird über alle komplementären Prädiktorvariablen x_C integriert, in der zweiten Variante wird über alle Prädiktorvariablen integriert. Die Differenz aus der ersten und zweiten Variante bildet die Grundlage des SFIMP.

Somit kann das Rahmenkonzept 3.1.1 auf Seite 81 zur Bestimmung von Prädiktoreffekten auf die Bestimmung der Feature-Importance erweitert werden, indem im Prädiktionsschritt statt der absoluten Vorhersage die Verlustfunktion betrachtet wird. Da für den Verlust auch die Vorhersage notwendig ist, lassen sich Feature-Effects und Feature-Importance mittels ICE & PD, sowie ICI & PI simultan berechnen. Daraus folgen Implikationen für die rechentechnische Umsetzung der Verfahren. Eine Implementierung der vorgestellten Verfahren kann auf gemeinsamen Datenstrukturen und Funktionen basieren, um die Erweiterung neuer Verfahren so einfach wie möglich zu gestalten.

3.2.2. Interaktion von Prädiktoren

Die univariate Betrachtung von Prädiktoreffekten und der Feature-Importance ist problematisch. Es ist möglich, dass zwei Prädiktorvariablen univariat betrachtet wenig oder keinen Einfluss auf die Zielvariable besitzen. Somit sind einerseits die Effekte erster Ordnung nullwertig. Andererseits sollte ein Maß wie die ICI oder der Shapley-Wert indikativ für eine geringe Importance sein. Werden nun jedoch beide Prädiktoren im Verbund betrachtet, kann der Interaktionseffekt erheblich sein.

Die auf der Partial Dependence basierende H-Statistik nach Friedman und Popescu (2008) ist ein Maß für die Importance eines Prädiktorenverbundes. Wir unterscheiden zwischen zwei relevanten Fällen: Die paarweise Interaktion zwischen zwei Prädiktoren und die Interaktion eines Prädiktors mit allen anderen Variablen (Molnar, 2018). Zunächst sei angenommen, zwei Prädiktoren x_j und x_k besäßen keinen Interaktionseffekt. Die PD kann folgendermaßen zerlegt werden, wobei $PD_{x_j,x_k}(x_j, x_k)$ die bivariate PD darstellt. Voraussetzung ist, dass die PD-Funktionen und das Modell $\hat{f}$ zentriert sind, so dass sie einen Mittelwert von 0 aufweisen (Friedman und Popescu, 2008).

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k)$$

In ähnlicher Art und Weise kann im Falle einer Abwesenheit von Interaktionseffekten die Responsefunktion $\hat{f}(x_S, x_C)$ in eine Summe von PD-Funktionen zerlegt werden,

wobei der erste Summand nur von x_j abhängt und der zweite Summand von allen restlichen Prädiktoren (Molnar, 2018).

$$\hat{f}(x_S, x_C) = PD_j(x_j) + PD_{-}j(x_{-j})$$

Im nächsten Schritt wird die Differenz der beobachteten und geschätzten PD mit der theoretischen Nicht-Interaktions-PD gebildet (Molnar, 2018). Die H-Statistik für die Interaktion zweier Prädiktoren (Def. 3.2.3) misst den Anteil der Varianz von $PD_{jk}(x_j, x_k)$, der nicht durch $PD_j(x_j) - PD_k(x_k)$ erklärt wird (Friedman und Popescu, 2008).

Die H-Statistik besitzt einen Wertebereich zwischen 0 und 1. Für beide Varianten ist sie nullwertig, falls keine Interaktionen auftreten. Sie beträgt 1, falls die gesamte Varianz von $PD_{jk}(x_j, x_k)$ bzw. von $\hat{f}(x_S, x_C)$ durch die Summe der PD-Funktionen erklärt wird (Molnar, 2018). Ein Wert von 1 der paarweisen H-Statistik bedeutet, dass beide univariaten PD-Funktionen konstant sind und die Vorhersage ausschließlich von der paarweisen Interaktion bestimmt wird (Molnar, 2018).

Definition 3.2.3 (H-Statistik für Interaktionseffekt zweier Prädiktoren)**.** *Die H-Statistik für den Interaktionseffekt zweier Prädiktoren x_j und x_k nach Friedman und Popescu (2008) ist gegeben durch*

$$H_{jk}^2 = \sum_{i=1}^{N} \frac{\left[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)})\right]^2}{PD_{jk}^2(x_j^{(i)}, x_k^{(i)})}$$

Analog wird die H-Statistik für den zweiten Fall definiert.

Definition 3.2.4 (H-Statistik für Interaktionseffekt eines Prädiktors mit allen restlichen Prädiktoren)**.** *Die H-Statistik für den Interaktionseffekt eines Prädiktors x_j mit allen restlichen Prädiktoren nach x_{-j} Friedman und Popescu (2008) ist gegeben durch*

$$H_j^2 = \sum_{i=1}^{N} \frac{\left[\hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)})\right]^2}{\hat{f}^2(x^{(i)})}$$

Die H-Statistik für die Interaktion eines Prädiktors mit allen anderen Variablen (Def. 3.2.4) ist von null verschieden, falls x_j mit mindestens einer Variablen interagiert (Friedman und Popescu, 2008) und wächst mit zunehmenden Interaktionseffekten. In diesem Fall sagt die H-Statistik jedoch nichts darüber aus, mit welcher Variable interagiert wird.

Friedman und Popescu (2008) schlagen einen statistischen Test vor, um zu prüfen, ob die H-Statistik signifikant verschieden von null ist. Getestet wird zur Nullhypothese, dass die H-Statistik nullwertig ist, d.h. dass keine Interaktionseffekte existieren. Molnar (2018) merkt an, dass die Interaktionsstatistik unter der Nullhypothese eine Anpassung des Modells erfordert, sodass darin keine Interaktionen auftreten. Dies sei nur für bestimmte Modellklassen möglich, weshalb der Test modellspezifisch sei und nicht modellagnostisch.

Die H-Statistik ist aufwändig zu berechnen, da sie über alle N Datenpunkte iteriert und an jedem Punkt die PD evaluiert, was wiederum eine Evaluation von N Datenpunkten darstellt (Molnar, 2018). Des Weiteren kann sie indikativ für *Spurious Interactions* sein (Friedman und Popescu, 2008). Sogar ein hoch akkurates Prädiktionsmodell kann substantielle Interaktionseffekte enthalten, die nicht im wahren zugrundeliegenden Modell enthalten sind (Friedman und Popescu, 2008). Laut Friedman und Popescu (2008) ist dies der Fall, falls die Prädiktorvariablen einen hohen Grad der Kollinearität aufweisen.

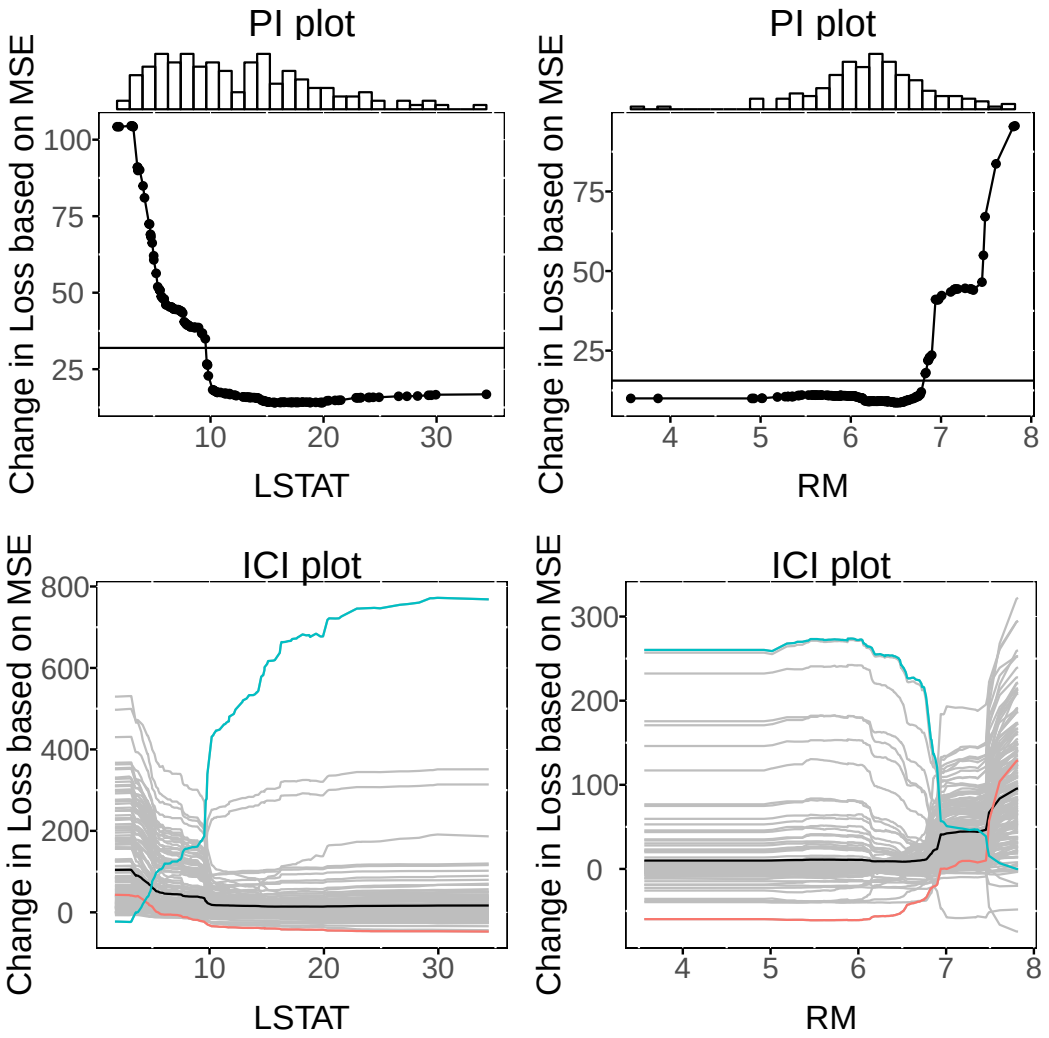


Abbildung 3.2.1.: Beispiel aus Casalicchio, Molnar und Bischl (2018) zur Visualisierung von Individual Conditional Importance und Partial Importance. Gezeigt sind jeweils ICI und PI der Prädiktorvariablen *LSTAT* und *RM* für einen Random Forest, trainiert auf dem Boston-Housing-Datensatz. Die horizontalen Linien in den PI-Plots repräsentieren die Werte der globalen PFI, d.h. das Integral der PI-Kurve. Die ICI mit jeweils größtem Integral ist in grüner Farbe gekennzeichnet. Die ICI mit jeweils kleinstem Integral ist in roter Farbe gekennzeichnet.

KAPITEL 4.

Demonstration

„Before you've practiced, the theory is useless. After you've practiced, the theory is obvious."

David Williams

Im Anschluss an die diskutierten theoretischen Überlegungen werden die vorgestellten Verfahren anhand eines gemeinsamen Datensatzes demonstriert. Wir verwenden den *Bike Sharing*-Datensatz (Fanaee-T und Gama, 2013).

4.1. Datenbeschreibung

Der Datensatz enthält die stündlichen und jährlichen Zählungen ausgeliehener Fahrräder im Zeitraum 2011 bis 2012 des *Capital-Bikeshare-Systems* in Washington D.C. Es wurden 731 Tage aufgezeichnet, sowie 17389 Stunden. Wir wollen die Anzahl ausgeliehener Fahrräder pro Tag über die Einflüsse der Umwelt vorhersagen. Es stehen folgende Prädiktorvariablen zur Auswahl:

- instant: Index für die Bestandsaufnahme
- dteday : Datum
- season: Frühling (1), Sommer (2), Herbst (3) und Winter (4)
- mnth : Monat (1 bis 12)
- hr : Stunde (0 bis 23)
- yr: Das Jahr (2011 oder 2012)
- holiday: Binäre Variable, die anzeigt, ob der jeweilige Tag ein Feiertag war (1) oder nicht (2)
- weekday: Wochentag
- workingday: Binäre Variable, die anzeigt, ob der jeweilige Tag ein Werktag (1) oder kein Werktag (0) war
- weathersit: Kategoriale Variable für die Wettersituation eines Tages:
  1. Clear, Few clouds, Partly cloudy,
  2. CloudyMist, Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist,
  3. Light Snow, Light Rain + Thunderstorm + Scattered clouds
  4. Light Rain + Scattered clouds , Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalisierte Temperatur in Grad Celsius. Die Werte werden über folgende Formel berechnet:

$$\frac{(t - t_{min})}{(t_{max} - t_{min})}, t_{min} = -8, t_{max} = +39 \quad \text{(nur in stündlicher Skala)}$$

- atemp: Normalisierte gefühlte Temperatur in Grad Celsius. Die Werte werden über folgende Formel berechnet:

$$\frac{(t - t_{min})}{(t_{max} - t_{min})}, t_{min} = -16, t_{max} = +50 \quad \text{(nur in stündlicher Skala)}$$

- hum: Normalisierte Humidität. Die Werte werden werden durch 100 dividiert (maximale Humidität)
- windspeed: Normalisierte Windstärke. Die Werte werden durch 67 dividiert (maximale Windstärke)
- casual: Anzahl an gelegentlichen Kunden
- registered: Anzahl an registrierten Kunden
- cnt: Anzahl aller ausgeliehenen Fahrräder

Aufgrund von ähnlichem Informationsgehalt von *atemp* und *temp*, *season* und *month*, sowie *holiday*, *weekday*, *workingday*, wird nur jeweils eine Variable in das Modell aufgenommen. Die Temperatur wird gemäß obenstehender Formel zurückgerechnet und zu Interpretationszwecken nicht-normalisiert aufgenommen. Darüber hinaus entfernen wir *casual* und *registered*, da die Anzahl registrierter Kunden und die Anzahl gemieteter Fahrräder in einem *Reverse-Causality*-Zusammenhang stehen könnten. Die Struktur des verwendeten Datensatzes lautet wie folgt. Wir verwenden *cnt* als Zielvariable.

```
[R]> 'data.frame': 731 obs. of  8 variables:
[R]>  $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
[R]>  $ yr        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
[R]>  $ workingday: Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 1 1 2 ...
[R]>  $ weathersit: Factor w/ 3 levels "1","2","3": 2 2 1 1 1 1 2 2 1 1 ...
[R]>  $ temp      : num  8.52 9.45 1.43 1.6 2.89 ...
[R]>  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
[R]>  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
[R]>  $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

4.2. Modellanpassung

Wir verwenden die SVM des *e1071*-Paketes, sowie *mlr* für die Modellanpassung und *iml* für die post-hoc Interpretation. Die Hyperparameterkonfiguration der SVM wird anhand einer 3-fachen Kreuzvalidierung und einer Random-Search der Kosten-/ und Gammaparameter optimiert. Abb. 4.2.1 auf der nächsten Seite zeigt die getesteten Hyperparameter und das gefundene Optimum.

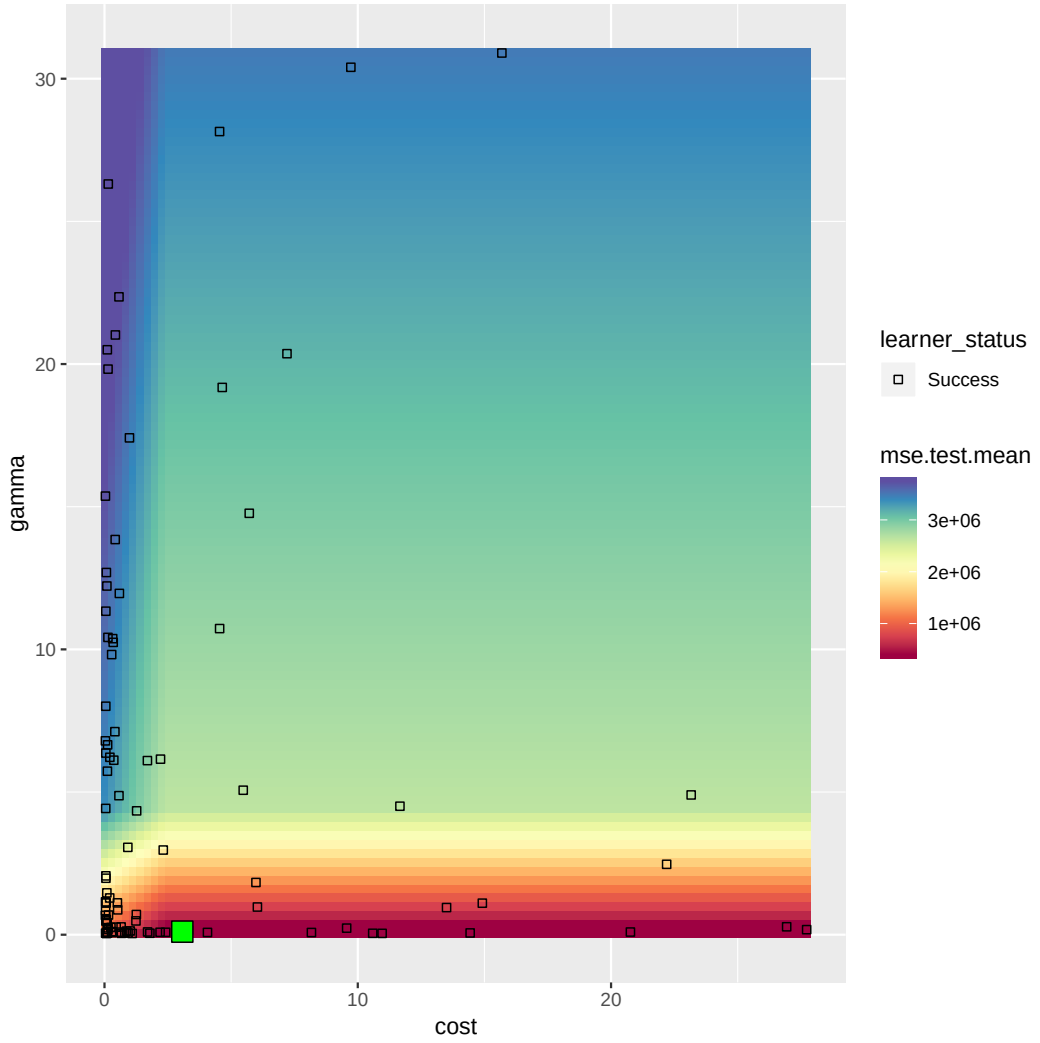


Abbildung 4.2.1.: Hyperparameterkonfigurationen, die im Zuge des Tuning-Prozesses getestet wurden. Die optimale Konfiguration ist vergrößert in grüner Farbe dargestellt.

Die gefundene Hyperparameterkonfiguration wird anschließend verwendet, um die SVM auf dem gesamten Datensatz zu trainieren:

```
[R]> Tune result:
[R]> Op. pars: cost=3.07; gamma=0.108
[R]> mse.test.mean=406666.7832791
[R]>
[R]> Call:
[R]> svm(formula = f, data = getTaskData(.task, .subset), cost = 3.07193487814972,
[R]>     gamma = 0.107773033766924)
[R]>
[R]>
[R]> Parameters:
[R]>    SVM-Type:  eps-regression
```

```
[R]>  SVM-Kernel:  radial
[R]>        cost:  3.071935
[R]>       gamma:  0.107773
[R]>     epsilon:  0.1
[R]>
[R]>
[R]> Number of Support Vectors:  493
```

4.3. Interpretation

Wir beginnen mit der Feature-Importance, um einen Überblick über den Beitrag jedes Prädiktors zur Modellgüte zu erhalten. Ein Überblick über die Feature-Importance aller Prädiktorvariablen ist bei der anschließenden Selektion eines Prädiktors zur Bestimmung eines Feature-Effects von Nutzen.

4.3.1. Feature-Importance

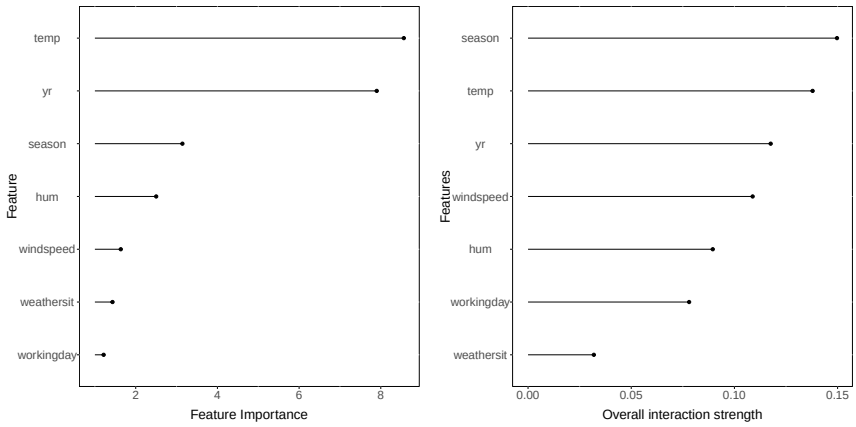
Für alle verwendeten Prädiktoren wird die Permutation-Feature-Importance geschätzt. Wir ziehen den *Mean-Square-Error* als Verlustfunktion heran. Die PFI (Abb. 4.3.1 auf der nächsten Seite (a)) dient als globales Maß zur Bestimmung der Feature-Importance.

```
[R]>      feature original.error permutation.error importance
[R]> 1       temp       287953.4         2466069.9   8.564127
[R]> 2         yr       287953.4         2275081.8   7.900867
[R]> 3     season       287953.4          905051.0   3.143046
[R]> 4        hum       287953.4          720115.0   2.500804
[R]> 5  windspeed       287953.4          470641.5   1.634436
[R]> 6 weathersit       287953.4          412205.6   1.431501
[R]> 7 workingday       287953.4          349751.0   1.214610
```

Darüber hinaus betrachten wir die H-Statistik (Kapitel 3, Def. 3.2.4 auf Seite 89), um Interaktionseffekte zwischen den Prädiktorvariablen einschätzen zu können. Zunächst wird für jeden Prädiktor die H-Statistik für die Interaktion mit allen restlichen Variablen geschätzt.

```
[R]> Interpretation method:  Interaction
[R]>
[R]>
[R]> Analysed predictor:
[R]> Prediction task: regression
[R]>
[R]>
[R]> Analysed data:
[R]> Sampling from data.frame with 731 rows and 7 columns.
[R]>
[R]> Head of results:
```

```
[R>     .feature .interaction
[R> 1     season   0.14970658
[R> 2         yr   0.11757526
[R> 3 workingday   0.07804823
[R> 4 weathersit   0.03189180
[R> 5       temp   0.13789380
[R> 6        hum   0.08951137
```



(a) (b)

Abbildung 4.3.1.: PFI und H-Statistik

In Abb. 4.3.1 (b) sehen wir, dass alle zur Verfügung stehenden Prädiktorvariablen Interaktionen untereinander aufweisen. Aufgrund der relativ geringen Datengröße können wir eine erschöpfende Berechnung aller paarweisen H-Statistiken durchführen. Die Schätzungen basieren auf unterschiedlichen Stichproben der Daten und sind somit zum Teil asymmetrisch. Die H-Statistik besitzt einen Wertebereich im Intervall $[0, 1]$. Die geschätzten H-Statistiken sind indikativ für schwache Interaktionseffekte zwischen den Prädiktorvariablen.

```
[R>                 .feature .interaction
[R> 1          hum:windspeed   0.31466264
[R> 2          windspeed:hum   0.26831505
[R> 3             season:hum   0.26195161
[R> 4             hum:season   0.19462609
[R> 5   windspeed:workingday   0.16117917
[R> 6  weathersit:workingday   0.14920527
[R> 7   workingday:windspeed   0.14831036
[R> 8      season:workingday   0.12518814
[R> 9            season:temp   0.11529106
[R> 10 workingday:weathersit   0.11005820
[R> 11           temp:season   0.10845219
[R> 12                hum:yr   0.10723302
[R> 13               yr:temp   0.10376922
[R> 14               temp:yr   0.09937009
[R> 15      windspeed:season   0.09486661
```

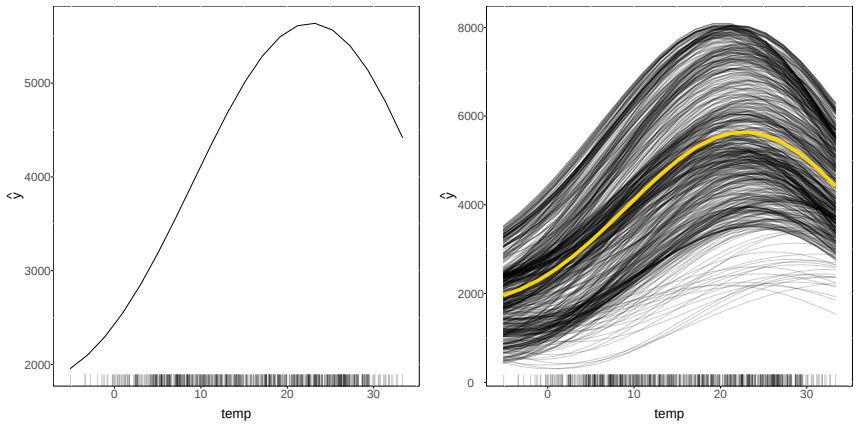
```
[R]> 16      season:windspeed   0.09227810
[R]> 17                yr:hum   0.08522654
[R]> 18     workingday:season   0.08095898
[R]> 19  windspeed:weathersit   0.07885410
[R]> 20       temp:workingday   0.07781736
[R]> 21  weathersit:windspeed   0.07497635
[R]> 22       workingday:temp   0.07155813
[R]> 23        weathersit:hum   0.06726529
[R]> 24        hum:weathersit   0.06268413
[R]> 25        windspeed:temp   0.06000066
[R]> 26        temp:windspeed   0.05849063
[R]> 27          yr:windspeed   0.05821403
[R]> 28          windspeed:yr   0.05531272
[R]> 29         yr:workingday   0.05404110
[R]> 30              hum:temp   0.05242782
[R]> 31       weathersit:temp   0.05151457
[R]> 32              temp:hum   0.05008835
[R]> 33     season:weathersit   0.04850348
[R]> 34            yr:season   0.04665760
[R]> 35         workingday:yr   0.04355320
[R]> 36         weathersit:yr   0.04205749
[R]> 37         yr:weathersit   0.04174178
[R]> 38             season:yr   0.03805152
[R]> 39       temp:weathersit   0.03764817
[R]> 40     weathersit:season   0.03664286
[R]> 41        workingday:hum   0.02528645
[R]> 42        hum:workingday   0.02445457
```

4.3.2. Feature-Effects

Wir beginnen mit der Betrachtung der absoluten Vorhersagen durch ICE & PD. Da die Auswahl an Prädiktorvariablen relativ gering ist (7 Prädiktoren), stehen beim Selektionsschritt Betrachtungen inhaltlicher Fragestellungen im Vordergrund. Ein Einfluss der Außentemperatur auf die Anzahl ausgeliehener Fahrräder erscheint sinnvoll.

Die PD ist indikativ für eine steigende Anzahl gemieteter Fahrräder bis zu einer Temperatur von 23 Grad Celsius. Darüber beginnt die Anzahl gemieteter Fahrräder im Mittel zu sinken. Wir betrachten den ICE-Plot, um die Repräsentativität der PD für den Prädiktoreffekt einzuschätzen. Die Spannweite der ICE-Werte ist zu groß, um Aussagen über die Ähnlichkeit der Trajektorien zu treffen. Wir zentrieren die ICE-Kurven am jeweiligen Anfangspunkt und betrachten zusätzlich den d-ICE-Plot, um etwaige Interaktionseffekte einschätzen zu könnnen. Abbildung 4.3.5 auf Seite 102 zeigt alle paarweisen Interaktionen mit *temp*, geschätzt mit der paarweisen H-Statistik.

Die zentrierten ICE-Kurven (c-ICE-Plot in Abb. 4.3.3 auf Seite 100 (a)) weisen leicht divergente Trajektorien auf. Die Änderungsraten der ICE-Kurven (d-ICE-Plot in Abb. 4.3.3 auf Seite 100 (b)) unterscheiden sich deutlich. Beide Varianten sind



(a) (b)

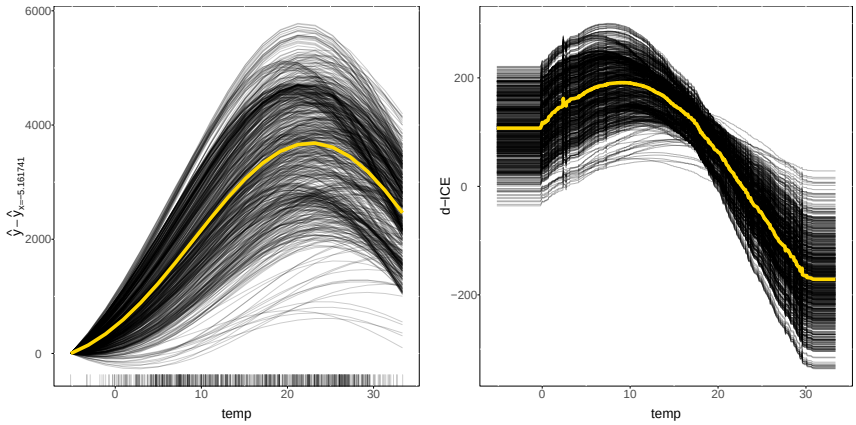
Abbildung 4.3.2.: Die Partial Dependence suggeriert einen parabolischen Zusammenhang zwischen der Umgebungstemperatur und der Anzahl gemieteter Fahrräder (a). Die Spannweite der Vorhersagen ist im Falle der ICE zu groß, um Aussagen bezüglich der Trajektorien zu treffen (b).

indikativ für Interaktionseffekte. Wir betrachten zusätzlich den ALE erster Ordnung in Abb. 4.3.4 auf Seite 101 (b). Univariate PD und ALE erster Ordnung zeigen einen ähnlichen geschätzten Effekt erster Ordnung.

Der Interaktionseffekt der Temperatur mit der Jahreszeit ist am stärksten ausgeprägt. Dabei ist zu beachten, dass die größte geschätzte H-Statistik etwa einen Wert von 0.1 aufweist. Bei einem Wertebereich der H-Statistik im Intervall $[0, 1]$ sind die Interaktionseffekte allesamt gering ausgeprägt. Der ALE zweiter Ordnung für die Interaktion mit einer kategorialen Variable wird durch den ALE erster Ordnung je Ausprägung repräsentiert (Abb. 4.3.6 auf Seite 103).

Die Effekte sind in allen Jahreszeiten ähnlich und unterscheiden sich hauptsächlich durch additive Konstanten bzw. vertikale Differenzen. Zusätzlich wird die Interaktion mit einer numerischen Variable visualisiert. Wir vermuten, dass Umgebungstemperatur und Humidität korrelieren und betrachten deshalb die bivariate PD (Abb. 4.3.7 auf Seite 104) und den ALE zweiter Ordnung von *temp* und *hum* (Abb. 4.3.8 auf Seite 105). Da der ALE zweiter Ordnung von Haupteffekten bereinigt ist, eignet sich die bivariate PD für die Einschätzung eines totalen Effektes beider Prädiktorvariablen auf die Vorhersage der Zielvariable.

Der ALE zweiter Ordnung (Abb. 4.3.8 auf Seite 105) ermöglicht die Einschätzung des Interaktionseffektes nach Abzug der beiden Haupteffekte. Im Gegensatz zur bivariaten PD kann dieser nicht ohne die Betrachtung der Haupteffekte interpretiert



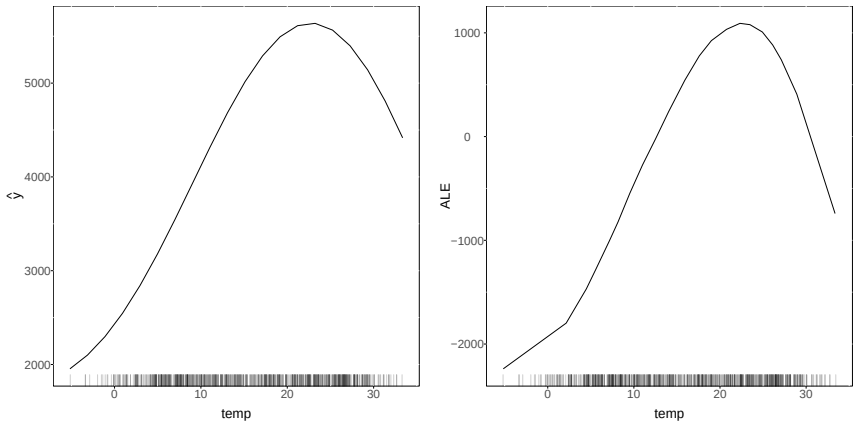
(a) (b)

Abbildung 4.3.3.: Über die Zentrierung der ICE sind Divergenzen zu erkennen (a). Der d-ICE-Plot zeigt unterschiedliche Änderungsraten (b). Beide Varianten sind indikativ für Interaktionseffekte.

werden. Die Luftfeuchtigkeit ist bis zu einem Wert von 0.5 mit einer steigenden Anzahl an gemieteten Fahrrädern verbunden. Ab einem Wert von 0.5 ist eine steigende Luftfeuchtigkeit mit einer sinkenden Anzahl an gemieteten Fahrrädern assoziiert. Der ALE erster Ordnung der Temperatur gleicht approximativ der univariaten geschätzten PD. Steigende Temperaturen sind ungefähr bis zu einem Wert von 23 Grad Celsius mit einer steigenden Anzahl an gemieteten Fahrrädern verbunden. Darüber sinkt die vorhergesagte Anzahl gemieteter Fahrräder. Zusätzlich zu den Haupteffekten können wir beobachten, dass die folgenden Interaktionen zu mehr gemieteten Fahrrädern führen. Eine geringe Temperatur mit einer geringen Humidität, eine geringe Temperatur mit einer hohen Humidität und eine hohe Temperatur mit hoher Humidität. Im mittleren Bereich der gemeinsamen Verteilung existiert kein Interaktionseffekt. Bei hohen Temperaturen und geringer Humidität ist ein negativer Interaktionseffekt auf die Anzahl gemieteter Fahrräder zu beobachten. Hierbei ist anzumerken, dass anhand der marginalen Verteilungen beider Prädiktorvariablen ersichtlich wird, dass sich an den Rändern nur wenige Datenpunkte befinden. Die Vorhersage von ML-Modellen in datenarmen Regionen ist unzuverlässig.

Zuletzt kann ein marginaler Effekt geschätzt werden. Aufgrund des approximativ quadratischen Effektes der Außentemperatur auf die Anzahl ausgeliehener Fahrräder ist die intervallbasierte Variante anzuraten.

```
[R]>     interval.left interval.right       AME
[R]> 0%      -5.161741       4.480000  110.0152
[R]> 10%      4.480000       7.159984  171.4168
```



(a) (b)

Abbildung 4.3.4.: Univariate PD (a) und ALE erster Ordnung (b). Geschätzt wird der Einfluss der absoluten Umgebungstemperatur auf die Anzahl gemieteter Fahrräder. Beide Verfahren liefern ähnliche Prädiktoreffektschätzungen.

```
[R]> 20%      7.159984       9.530416  192.4140
[R]> 30%      9.530416      12.520000  175.3035
[R]> 40%     12.520000      15.919984  166.2841
[R]> 50%     15.919984      19.000000  121.9017
[R]> 60%     19.000000      22.320016   49.8856
[R]> 70%     22.320016      24.919984  -27.3861
[R]> 80%     24.919984      27.120016 -124.7333
[R]> 90%     27.120016      33.360016 -214.3322
```

Der global geschätzte AME von *temp* auf die Anzahl gemieteter Fahrräder lautet:

```
[R]>       temp
[R]> 1 61.90845
```

Die vorgestellten Verfahren werden nun auf weitere Prädiktorvariablen angewandt. Ergänzend können auch andere Modelle zum Vergleich herangezogen werden.

Intervallbasierte AME sind nur für numerische Variablen möglich.

```
[R]> $temp
[R]>     left.boundary right.boundary       AME
[R]> 0%      -5.161741       4.480000  110.0152
[R]> 10%      4.480000       7.159984  171.4168
[R]> 20%      7.159984       9.530416  192.4140
[R]> 30%      9.530416      12.520000  175.3035
[R]> 40%     12.520000      15.919984  166.2841
[R]> 50%     15.919984      19.000000  121.9017
[R]> 60%     19.000000      22.320016   49.8856
[R]> 70%     22.320016      24.919984  -27.3861
```

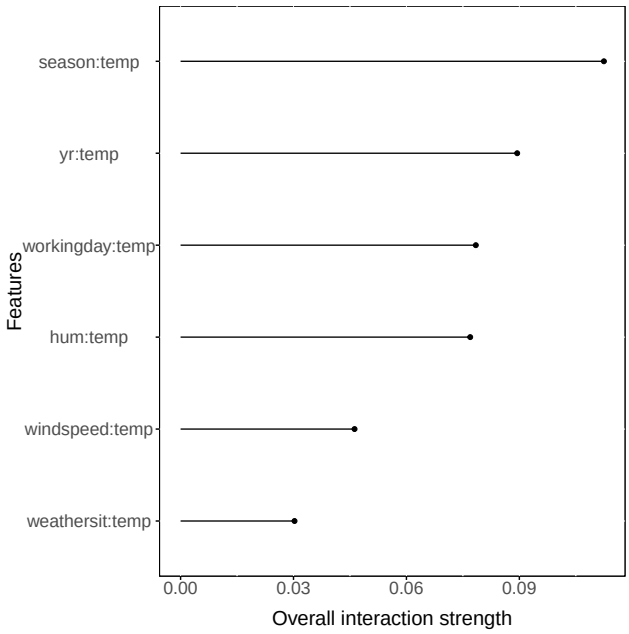


Abbildung 4.3.5.: Paarweise Interaktionen mit *temp*, geschätzt mit der paarweisen H-Statistik. Die geschätzten H-Statistiken sind indikativ für geringe Interaktionseffekte.

```
[R]> 80%     24.919984      27.120016 -124.7333
[R]> 90%     27.120016      33.360016 -214.3322
[R]>
[R]> $hum
[R]>     left.boundary right.boundary        AME
[R]> 0%       0.000000       0.450000   536.0596
[R]> 10%      0.450000       0.500000  -926.0460
[R]> 20%      0.500000       0.542083 -1074.8655
[R]> 30%      0.542083       0.585217 -1742.5340
[R]> 40%      0.585217       0.626667 -2007.2037
[R]> 50%      0.626667       0.668750 -2473.9285
[R]> 60%      0.668750       0.707500 -3101.6980
[R]> 70%      0.707500       0.752917 -3801.8093
[R]> 80%      0.752917       0.817500 -3967.3929
[R]> 90%      0.817500       0.972500 -4870.8229
[R]>
[R]> $windspeed
[R]>     left.boundary right.boundary        AME
[R]> 0%      0.0223917       0.100133  -969.1891
[R]> 10%     0.1001330       0.125248 -1858.6874
[R]> 20%     0.1252480       0.143042 -1818.0436
[R]> 30%     0.1430420       0.163554 -1407.9460
[R]> 40%     0.1635540       0.180975 -1150.9939
[R]> 50%     0.1809750       0.200258 -1710.9643
[R]> 60%     0.2002580       0.223267 -2845.2534
[R]> 70%     0.2232670       0.248309 -3300.6163
[R]> 80%     0.2483090       0.296029 -3899.9575
[R]> 90%     0.2960290       0.507463 -4704.3803
```

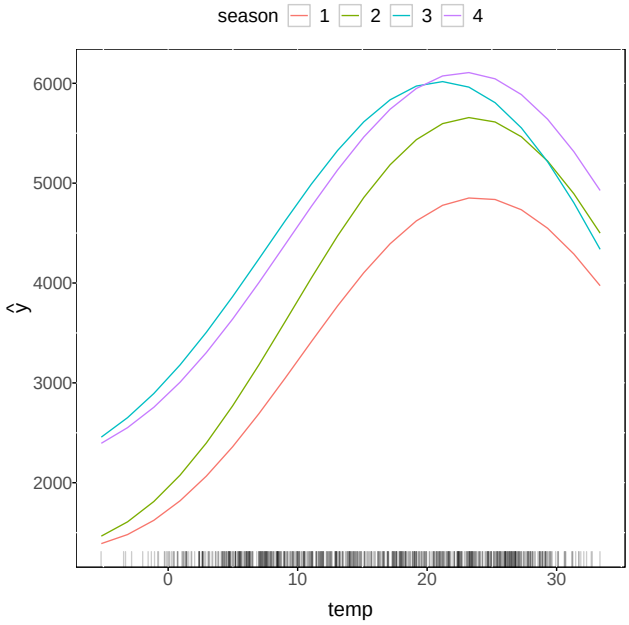


Abbildung 4.3.6.: PD je Faktorausprägung von *season*.

Die global geschätzten AME je numerischer Prädiktorvariable lauten:

```
[R]>        temp         hum   windspeed
[R]>    61.90845 -2344.37763 -2378.82543
```

Zusätzlich können wir einen AME je Faktorvariable ausgeben, der die durchschnittliche Änderung der Vorhersage im Vergleich zu einer Basiskategorie angibt. Wir erstellen für jede Faktorausprägung der betrachteten Variable einen intervenierten Datensatz, der die Werte der betrachteten Faktorvariable auf die jeweilige Ausprägung setzt. Anschließend wird die Differenz der Vorhersagen ermittelt und zum AME gemittelt. Wir wählen jeweils die erste Kategorie als Basiskategorie.

```
[R]> [1] "AME von season auf cnt mit Basiskategorie = 1"
[R]>         2         3         4
[R]>  627.7597 1263.1388 1249.9990
[R]> [1] "AME von yr auf cnt mit Basiskategorie = 0"
[R]>        1
[R]> 1941.635
[R]> [1] "AME von workingday auf cnt mit Basiskategorie = 0"
[R]>        1
[R]> 106.2257
[R]> [1] "AME von weathersit auf cnt mit Basiskategorie = 1"
[R]>         2         3
[R]> -236.5419 -787.1332
```

Die Interpretation des AME von kategorialen Variablen ist intuitiv. Wir stellen beispielsweise fest, dass im Vergleich zur Wetterkategorie 1 (Clear, Few clouds, Partly

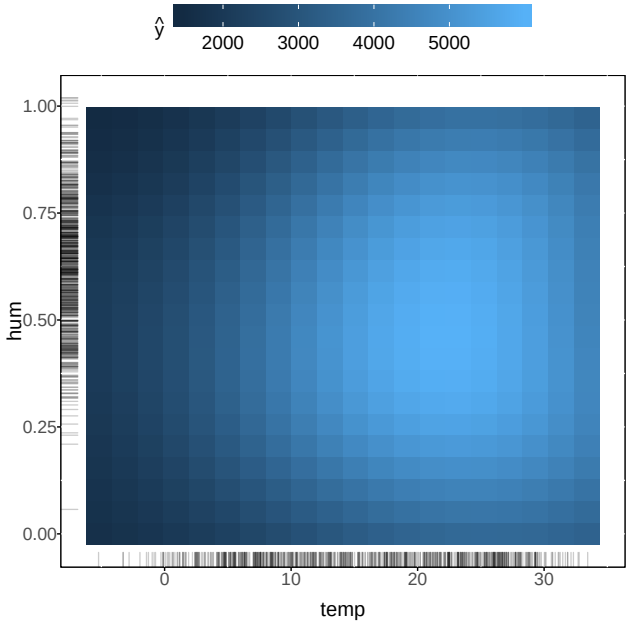


Abbildung 4.3.7.: Bivariate PD zwischen *temp* und *hum*. Mittlere Temperaturen und mittlere Humidität ist mit der höchsten Anzahl gemieteter Fahrräder verbunden.

cloudy) die Anzahl vorhergesagter gemieteter Fahrräder im vorliegenden Modell desto stärker sinkt, je schlechter die Wetterlage ist. So sinkt diese einmal, falls die Wetterlage aller Tage Kategorie 2 annehmen würde (CloudyMist, Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist). Sie sinkt noch einmal stärker, falls die Wetterlage aller Tage des Datensatzes Kategorie 3 annehmen würde (Light Snow, Light Rain + Thunderstorm + Scattered clouds).

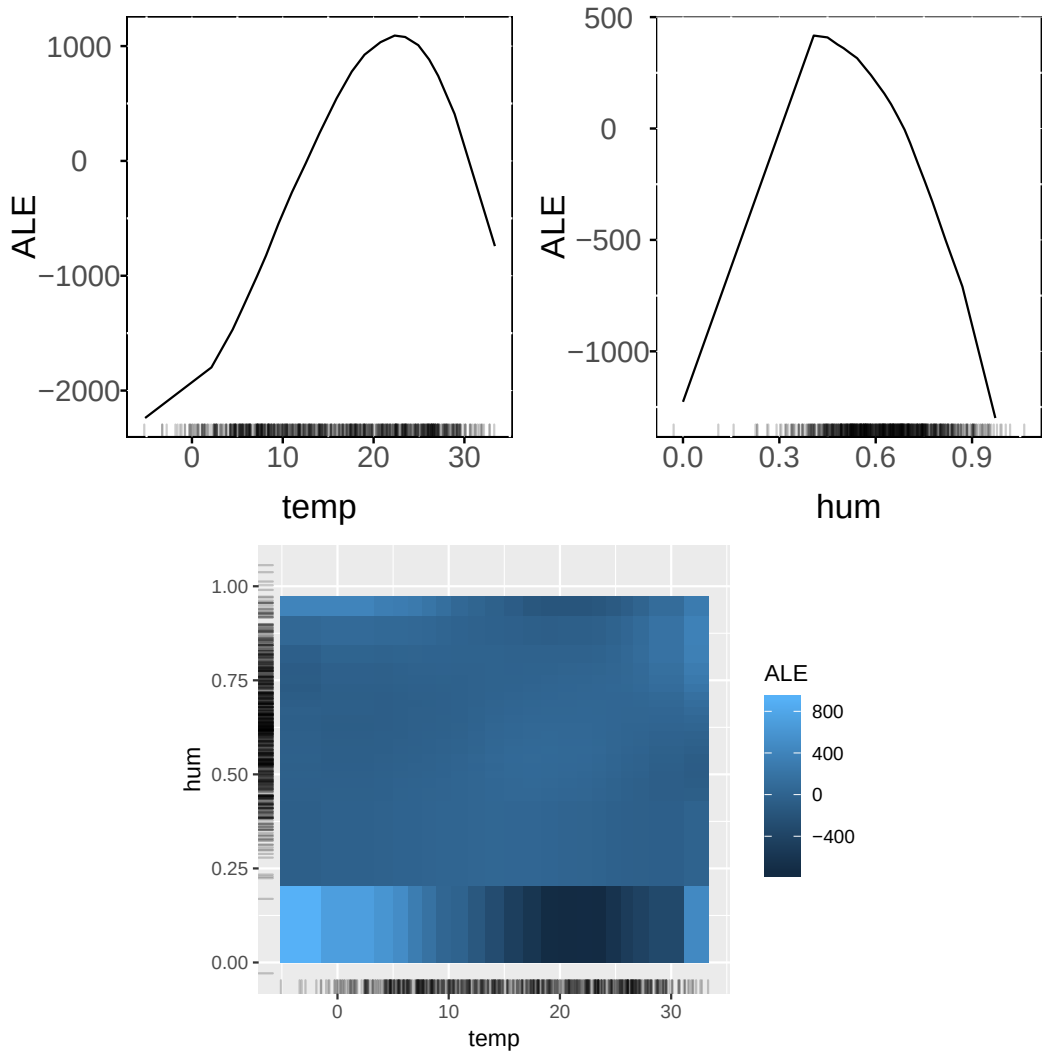


Abbildung 4.3.8.: ALE erster Ordnung für *temp* und *hum*, sowie ALE zweiter Ordnung für die Interaktion zwischen *temp* und *hum*. Der ALE zweiter Ordnung schätzt die Interaktion beider Variablen nach Abzug der Effekte erster Ordnung und kann nie in Isolation interpretiert werden.

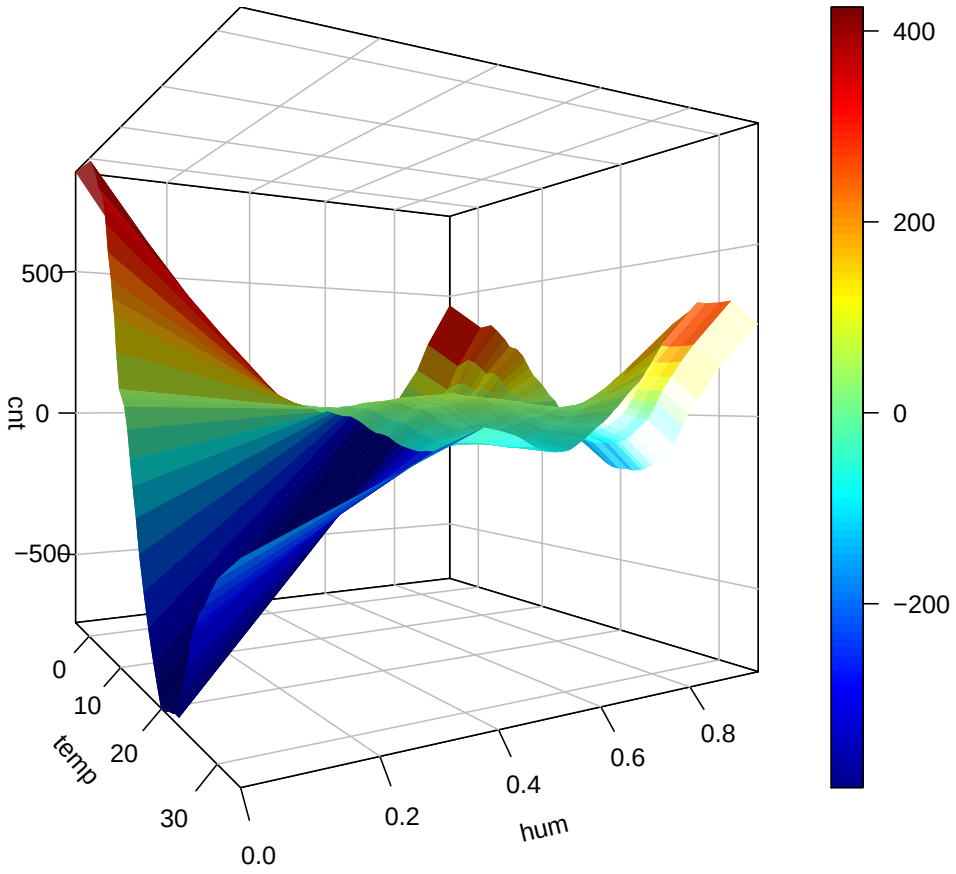


Abbildung 4.3.9.: Die Interpretation des ALE zweiter Ordnung kann durch dreidimensionale Darstellungen erleichtert werden.

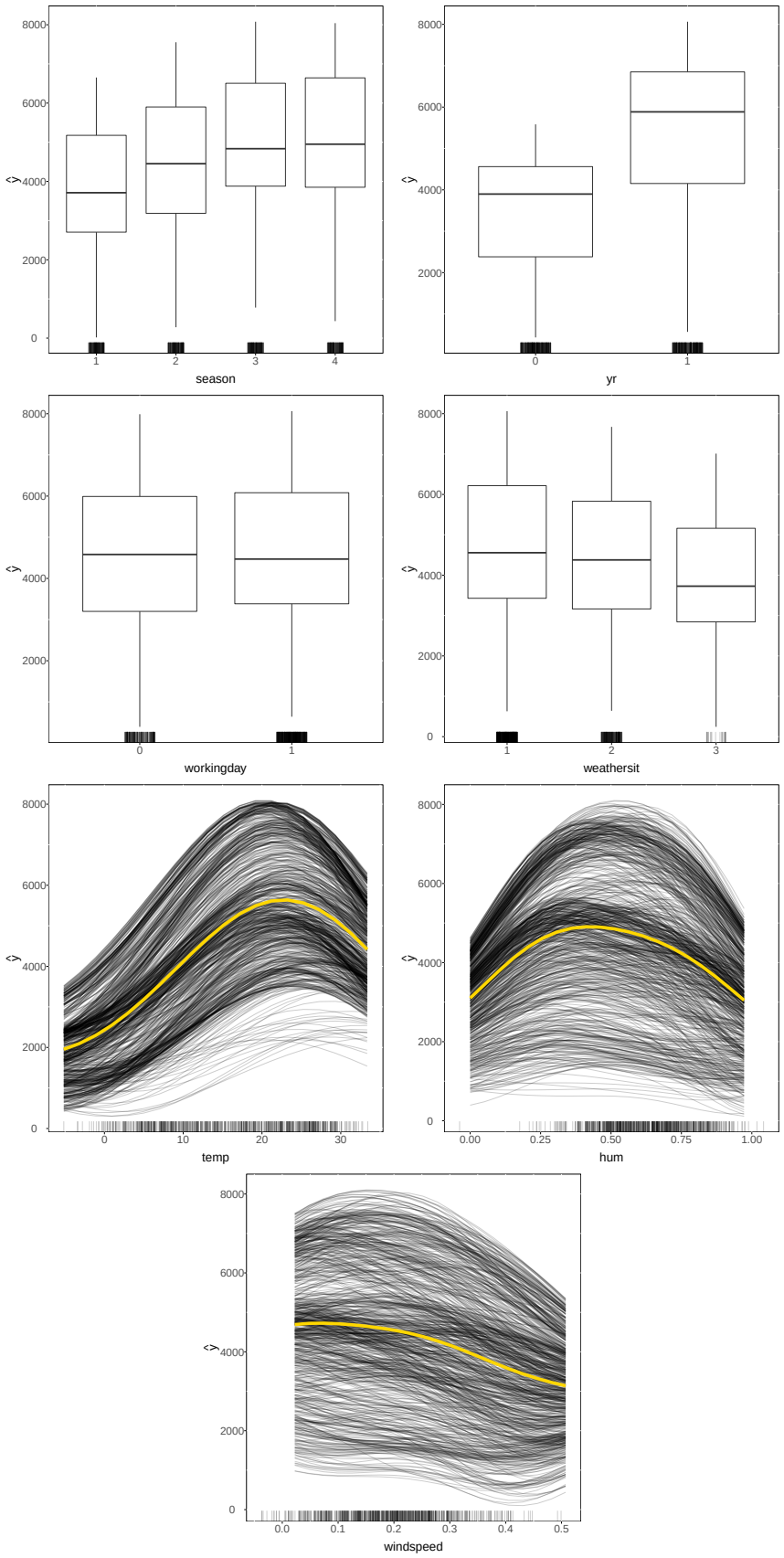


Abbildung 4.3.10.: ICE & PD aller verwendeter Prädiktorvariablen.

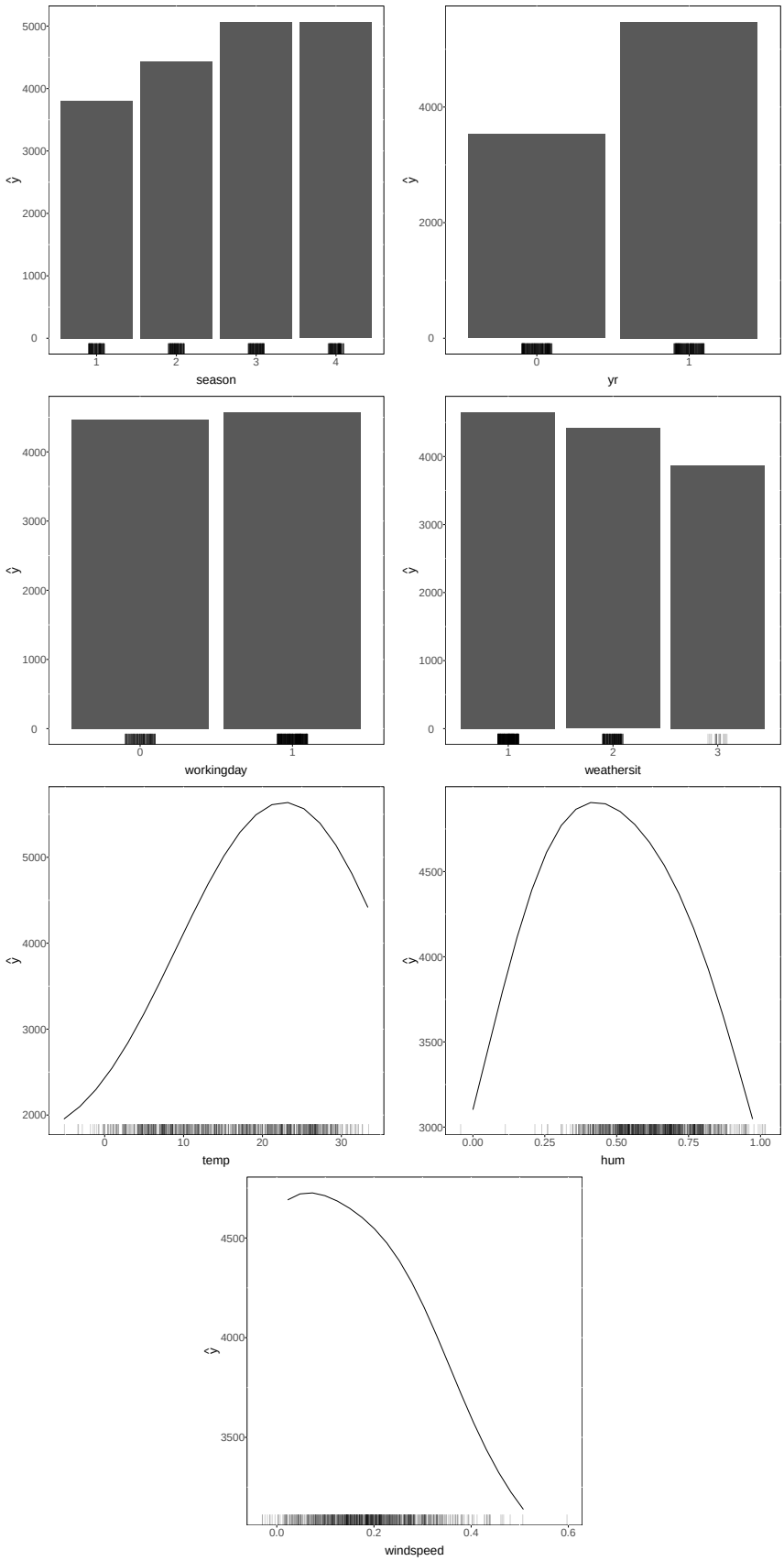


Abbildung 4.3.11.: PD aller verwendeter Prädiktorvariablen.

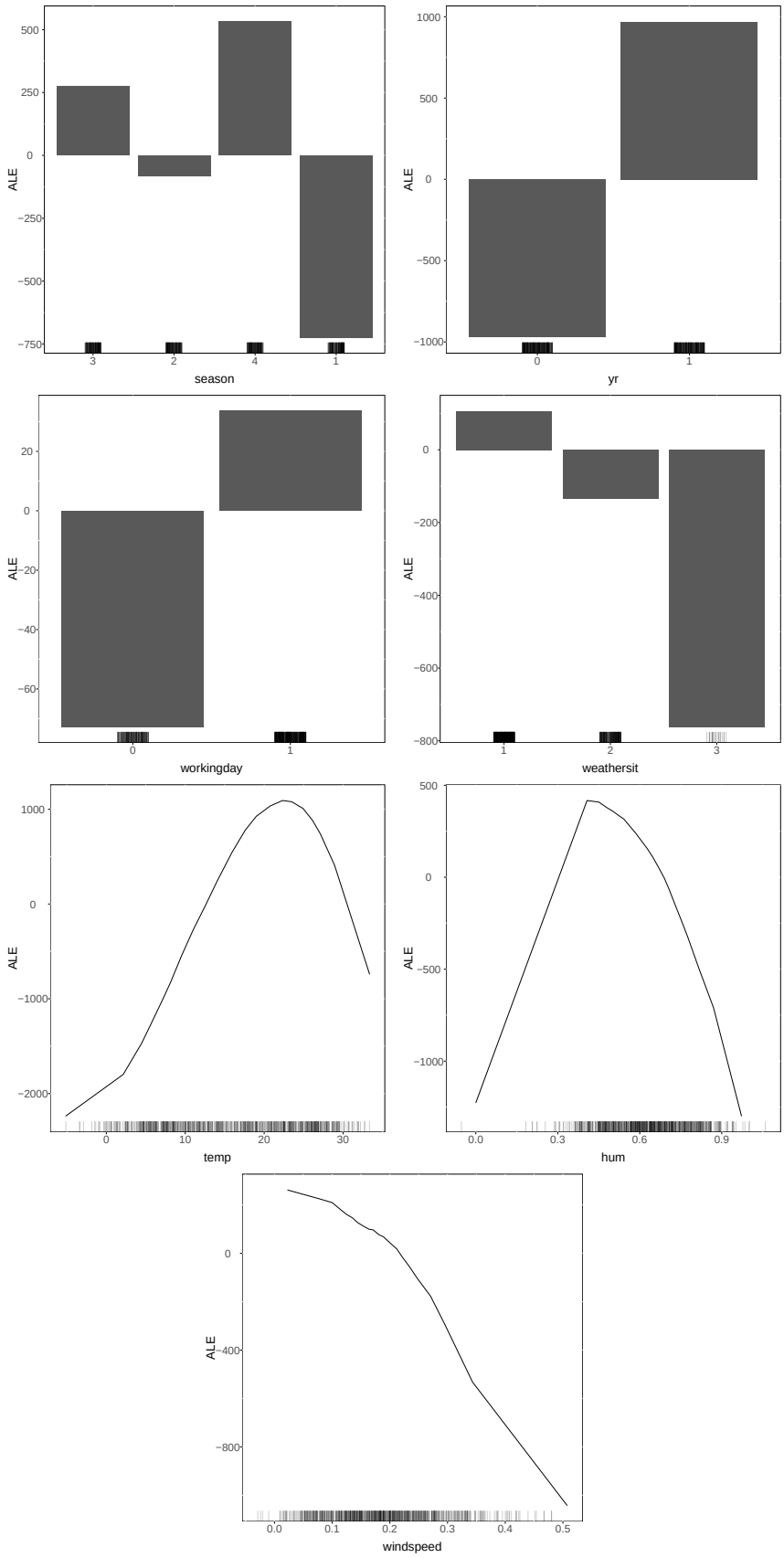


Abbildung 4.3.12.: ALE erster Ordnung aller verwendeter Prädiktorvariablen.

KAPITEL 5.

Konklusion

„Science is about the process. It's not about the conclusion."

Steven Novella

5.1. Zusammenfassung

Die vorliegende Arbeit hatte zum Ziel etablierte post-hoc modellagnostische Verfahren zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen vorzustellen und zu diskutieren. In einer konzeptionellen Einführung in das Interpretierbare Machine-Learning wurde dargelegt, weshalb Interpretierbarkeitsverfahren für ML-Modelle sinnvoll und notwendig sind. Für Interpretationszwecke musste bisher auf herkömmliche statistische Verfahren zurückgegriffen werden. ML-Modelle besitzen ein sehr viel höhereres prädiktives Potential zu Lasten ihrer Interpretierbarkeit. Post-hoc modellagnostische Interpretationsverfahren ermöglichen die Interpretation von Black-Box-Modellen jeder Art im Anschluss an den Modellanpassungsprozess. Neben der Bestimmung von Feature-Effekten ist die Bestimmung der Feature-Importance möglich. Lokale Effekte betrachten die Prädiktion eines einzelnen Datenpunktes, während bei der Bestimmung globaler Effekte eine Interpretation für die gesamten Daten vorgenommen wird.

Im Anschluss wurde eine Teilmenge der verfügbaren Verfahren im Detail vorgestellt. Über die Existenz der Functional-ANOVA-Dekomposition können Black-Box-Prädiktionsfunktionen in zueinander orthogonale Prädiktoreffekte unterschiedlicher Ordnung zerlegt werden. Die Schätzverfahren und Eigenschaften von marginalen Effekten, Individual Conditional Expectation & Partial Dependence, sowie Accumulated Local Effects wurden diskutiert und aus dem Blickwinkel der Functional-ANOVA-Dekomposition betrachtet.

Über die Zerlegung der vorgestellten Verfahren in gemeinsame Arbeitsschritte wurde ein generalisiertes System zur Bestimmung von Prädiktoreffekten in Supervised-Learning-Modellen entwickelt. Im Anschluss an die Auswahl einer Prädiktorvariablen erfolgt eine Stichprobenziehung. In einem Interventionsschritt wird der Datensatz verändert. Anschließend erfolgt eine Vorhersage anhand des intervenierten Datensatzes. In unterschiedlichen Aggregationsschritten werden die Prädiktionen weiterverarbeitet und können anschließend visualisiert werden.

Das generalisierte System kann auf die Feature-Importance erweitert werden. Wird im Prädiktionsschritt von ICE und PD statt der absoluten Vorhersage die Verlustfunktion herangezogen, erhalten wir ICI und PI als Maß für die lokale und globale Feature-Importance. Die Berechnung von Feature-Effect und Feature-Importance kann somit simultan geschehen. Dabei sollten ebenfalls Maße zur Interaktion von Prädiktoren herangezogen werden, wie die H-Statistik.

Im abschließenden vierten Kapitel wurde die vorgestellte Theorie auf den Bikesharing-Datensatz des Washington D.C. Capital-Bikeshare-Systems angewandt. Eine SVM sagt die Anzahl täglich gemieteter Fahrräder in Abhängigkeit der Wetter-

einflüsse vorher. Über die Bestimmung von Feature-Importance und Feature-Effects konnte ein detaillierter Einblick in die Arbeitsweise des Modells gewonnen werden.

5.2. Diskussion

Die vorgestellten Verfahren stellen eine Basis für die post-hoc modellagnostische Interpretation von ML-Modellen dar. Da diese verschiedene Stärken und Schwächen aufweisen, sollten sie grundsätzlich alle im Interpretationsprozess zur Anwendung kommen. (Intervallbasierte) AME eignen sich, um ein numerisches Maß für Prädiktoreffekte zu erhalten. ICE & PD, sowie ALE eignen sich vor allem für die visuelle Auswertung. Die PD kann aufgrund möglicher Interaktionseffekte nicht ohne Betrachtung der ICE interpretiert werden. Der ALE stellt darüber hinaus ein besseres Verfahren dar, falls die Prädiktorvariablen korreliert sind, was im Anwendungsfall sehr wahrscheinlich ist.

Zukünftig können Auswertungen über automatisierte Berechnungen von Score-Metriken erleichtert werden. Die entwickelten Score-Metriken sollten dabei skalenunabhängig sein. Die Interpretation muss nicht ausschießlich visuell oder automatisiert ausgeführt werden. So kann dem Anwender auch während der visuellen Betrachtung ein Score als Hilfestellung ausgegeben werden, falls der Score-Algorithmus nicht zu rechenintensiv ist.

Das generalisierte System zur Bestimmung von Feature-Effects und Feature-Importance stellt einen Leitfaden zur Entwicklung zukünftiger Verfahren dar. Software-Implementierungen können darauf aufbauen. Dabei ist zu beachten, dass der Rahmen verfügbarer Verfahren in der vorliegenden Arbeit nicht erschöpft wurde. Es existieren noch weitere vielversprechende Ansätze, wie beispielsweise die Verwendung von *Ersatz-Modellen (Surrogate Models)*.

5.2.1. Surrogate-Modelle

Ein globales Surrogate-Modell ist ein intrinsisch-interpretierbares Modell, das trainiert wurde, um die Vorhersagen eines Black-Box-Modells zu approximieren (Molnar, 2018). Das Surrogate-Modell kann anschließend anstatt der Black-Box interpretiert werden.

Local-Interpretable-Model-Agnostic-Explanations [LIME] (Tulio Ribeiro, Singh und Guestrin, 2016) stellen einen Surrogate-Ansatz für lokale Interpretationen dar. Im Interventionsschritt wird ausschließlich in die Eingangsgrößen einer einzelnen

Beobachtung interveniert. Im Anschluss an den Prädiktionssschritt werden die Vorhersagen der Zielvariable mit einem Maß ihrer Nähe zum beobachteten Zielvariablenwert gewichtet. Auf dem neuen, gewichteten Datensatz wird nun ein intrinsisch-interpretierbares Modell angepasst und anschließend interpretiert.

Der LIME-Ansatz zeigt vielversprechende Ergebnisse, auch bei der Verwendung von Text-Daten. So kann dieser beispielsweise bei der *Sentiment-Analyse* verwendet werden. Wird das Sentiment eines Satzes über ein Black-Box-Modell wie ein neuronales Netz klassifiziert, kann anhand von LIME geschätzt werden, wie viel einzelne Wörter zur Klassifikationsentscheidung beigetragen haben.

5.2.2. Machine-Learning & Kausale Inferenz

Darüber hinaus kann die Bestimmung von Prädiktoreffekten dazu verwendet werden, *kausale Effekte* in ML-Modellen zu bestimmen. Vielversprechnde Ansätze finden sich in Athey (2015), Wager und Athey (2015) oder in Ramachandra (2018). Kausale Inferenz mit der Hilfe von ML-Modellen besitzt das Potential, gesamte Wissenschaften zu transformieren, die sich zu einem großen Teil auf die Bestimmung kausaler Effekte in empirischen Daten konzentrieren, wie beispielsweise die Ökonometrie.

Die Ökonometrie stützt sich gegenwärtig auf herkömmliche Verfahren wie die logistische Regression, Diff-in-Diff-Ansätze, oder die Schätzung anhand von Instrumentalvariablen. Athey (2015) schlägt einen datenorientierten Ansatz vor, der die Daten in Subpopulationen partitioniert, die in der Größe des *Treatment-Effektes* variieren. Das *Causality Tree*-Verfahren verwendet eine erste Stichprobe, um die Partitionierung zu bestimmten und eine zweite Stichprobe, um den Treatment-Effekt zu schätzen. Der Ansatz basiert auf modifizierten Regressionsbäumen. Wager und Athey (2015) konstruieren weiterführend einen Random Forest, der zur kausalen Inferenz genutzt werden kann.

5.3. Ausblick

Molnar (2018) diskutiert mögliche Entwicklungen des Machine-Learnings, sowie des Interpretierbaren Machine-Learnings. Da die Entwicklung des Machine-Learnings wegweisend für die Entwicklung des Interpretierbaren Machine-Learnings ist, wird zunächst auf Ersteres eingegangen.

5.3.1. Machine-Learning

Die Entwicklung neuer Verfahren läuft mit enormer Geschwindigkeit voran, während diese im privaten Sektor nur langsam integriert werden. Es ist deshalb vorstellbar, dass die Domäne des Machine-Learnings in ihrer Vollumfänglichkeit nur langsam, jedoch kontinuierlich wachsen wird.

Derzeit ist eine zweite Entwicklung zu beobachten. Die Digitalisierung von Geschäftsmodellen auf einer globalen Skala führt zu einer zunehmenden Substitution von Arbeitskräften durch Maschinen. Die Entwicklung von intelligenten Maschinen auf der Basis von ML-Algorithmen wird diese Entwicklung beschleunigen. Somit wird die Integration von ML in reale Applikationen den Brennstoff für die Entwicklung der Digitalisierung darstellen.

Die bevorstehenden Entwicklungen werden nicht zuletzt aufgrund der Black-Box-Natur der zugrundeliegenden Algorithmen mit Skepsis betrachtet werden. Die Interpretierbarkeit von Machine-Learning-Modellen wird daher die Adaption erleichtern. Stellt Machine-Learning den Brennstoff der Digitalisierung dar, repräsentieren Interpretierbarkeitsverfahren den Katalysator (Molnar, 2018). Über die Möglichkeit zu erklären, wie ein Modell zu einer Entscheidung gelangt, wird die Hemmschwelle zum Einsatz von KI herabgesetzt.

5.3.2. Interpretierbares Machine-Learning

Die Fülle an verfügbaren Machine-Learning-Modellen erfordert Interpretationsverfahren, die vom zugrundeliegenden Modellanpassungsprozess bzw. Modell abgekoppelt sind (Molnar, 2018). Für die Skalierung von Systemen ist die Modularität der eingesetzten Verfahren maßgeblich (Molnar, 2018). So könne sowohl das zugrundeliegende Modell substituiert werden, als auch das Interpretationsverfahren. Es ist daher denkbar, dass der Großteil zukünftig entwickelter Verfahren unter dem Ziel der Skalierbarkeit post-hoc und modellagnostisch sein wird.

Während automatische Modellanpassungsverfahren bereits entwickelt und eingesetzt werden, ist dies für die Interpretation von Modellen noch nicht der Fall. Es ist bereits zum jetzigen Zeitpunkt möglich, alle vorgestellten Verfahren automatisiert zu berechnen. Die endgültige Interpretation kann jedoch noch nicht von Maschinen übernommen werden. Sobald robuste und effiziente Score-Algorithmen verfügbar sind, kann jedoch auch die Analyse der Interpretations-Outputs automatisiert werden. Daraus folgt, dass die Grenzen zwischen traditionellem ML und Interpretierbarem ML allmählich verschwinden und die beiden Disziplinen zum vollständig automatisierten Machine-Learning fusionieren werden.

Literaturverzeichnis

Alt, Helmut und Michael Godau (1995). „Computing the Fréchet distance between two polygonal curves“. In: *International Journal of Computational Geometry & Applications* 05.01n02, S. 75–91. DOI: 10.1142/S0218195995000064. eprint: https://doi.org/10.1142/S0218195995000064. URL: https://doi.org/10.1142/S0218195995000064.

Apley, D. W. (Dez. 2016). „Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models“. In: *ArXiv e-prints*. arXiv: 1612.08468 [stat.ME].

Athey, Susan (2015). „Machine Learning and Causal Inference for Policy Evaluation“. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: ACM, S. 5–6. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2785466. URL: http://doi.acm.org/10.1145/2783258.2785466.

Bartus, Tamás (2005). „Estimation of marginal effects using margeff“. In: *The Stata Journal* 5.3, S. 309 –329.

Best, Henning und Christof Wolf (2012). „Modellvergleich und Ergebnisinterpretation in Logit- und Probit-Regressionen“. In: *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 64.2, S. 377–395. ISSN: 1861-891X. DOI: 10.1007/s11577-012-0167-4. URL: https://doi.org/10.1007/s11577-012-0167-4.

Breiman, Leo (2001a). „Random Forests“. In: *Machine Learning* 45.1, S. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

— (Aug. 2001b). „Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)“. In: *Statist. Sci.* 16.3, S. 199–231. DOI: 10.1214/ss/1009213726. URL: https://doi.org/10.1214/ss/1009213726.

Casalicchio, G., C. Molnar und B. Bischl (Apr. 2018). „Visualizing the Feature Importance for Black Box Models“. In: *ArXiv e-prints*. arXiv: 1804.06620 [stat.ML].

Cohen, Shay, Eytan Ruppin und Gideon Dror (2005). „Feature Selection Based on the Shapley Value“. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., S. 665–670. URL: http://dl.acm.org/citation.cfm?id=1642293.1642400.

Doshi-Velez, F. und B. Kim (Feb. 2017). „Towards A Rigorous Science of Interpretable Machine Learning“. In: *ArXiv e-prints*. arXiv: 1702.08608 [stat.ML].

Eiter, Thomas und Heikki Mannila (1994). *Computing discrete Fréchet distance*. Techn. Ber. Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria.

Fanaee-T, Hadi und Joao Gama (2013). „Event labeling combining ensemble detectors and background knowledge“. In: *Progress in Artificial Intelligence*, S. 1–15. ISSN: 2192-6352. DOI: 10.1007/s13748-013-0040-3. URL: [WebLink].

Fisher, A., C. Rudin und F. Dominici (Jan. 2018). „Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the “Rashomon” Perspective“. In: *ArXiv e-prints*. arXiv: 1801.01489 [stat.ME].

Fornberg, Bengt (Dez. 1981). „Numerical Differentiation of Analytic Functions“. In: *ACM Trans. Math. Softw.* 7.4, S. 512–526. ISSN: 0098-3500. DOI: 10.1145/355972.355979. URL: http://doi.acm.org/10.1145/355972.355979.

Fornberg, Bengt und David M. Sloan (1994). „A review of pseudospectral methods for solving partial differential equations“. In: *Acta Numerica* 3, 203–267. DOI: 10.1017/S0962492900002440.

Friedman, Jerome H. (Okt. 2001). „Greedy function approximation: A gradient boosting machine.“ In: *Ann. Statist.* 29.5, S. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

Friedman, Jerome H. und Bogdan E. Popescu (Sep. 2008). „Predictive learning via rule ensembles“. In: *Ann. Appl. Stat.* 2.3, S. 916–954. DOI: 10.1214/07-AOAS148. URL: https://doi.org/10.1214/07-AOAS148.

Goldstein, A., A. Kapelner, J. Bleich und E. Pitkin (Sep. 2013). „Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation“. In: *ArXiv e-prints*. arXiv: 1309.6392 [stat.AP].

Greenwell, B. M., B. C. Boehmke und A. J. McCarthy (Mai 2018). „A Simple and Effective Model-Based Variable Importance Measure“. In: *ArXiv e-prints*. arXiv: 1805.04755 [stat.ML].

Hooker, Giles (2007). „Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables“. In: *Journal of Computational and Graphical Statistics* 16.3, S. 709–732. DOI: 10.1198/106186007X237892. eprint: https://doi.org/10.1198/106186007X237892. URL: https://doi.org/10.1198/106186007X237892.

Hooker, Giles und Jerome Friedman (2004). „Diagnostics and extrapolation in machine learning“. Diss. Stanford University.

Kleiber, Christian und Achim Zeileis (2008). *Applied Econometrics with R*. ISBN 978-0-387-77316-2. New York: Springer-Verlag, S. 126. URL: https://CRAN.R-project.org/package=AER.

Kohavi, Ron (1998). „Glossary of terms“. In: *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process* 30.271, S. 127–132. URL: https://ci.nii.ac.jp/naid/10018512237/en/.

Leeper, Thomas J. (2018). *margins: Marginal Effects for Model Objects*. R package version 0.3.23.

Lindfield, George R. und John E. T. Penny (1989). *Microcomputers in Numerical Analysis*. New York, NY, USA: Halsted Press. ISBN: 0-470-21415-5.

Markov, A. A. (1957). „Theory of Algorithms“. In: *Journal of Symbolic Logic* 22.1, S. 77–79.

Minsky, Marvin (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press. ISBN: 0262630222.

Molnar, Christoph (2018). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/. https://christophm.github.io/interpretable-ml-book/.

Moravec, Hans (1988). *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA, USA: Harvard University Press. ISBN: 0-674-57616-0.

Muehlenstaedt, Thomas, Olivier Roustant, Laurent Carraro und Sonja Kuhnt (2012). „Data-driven Kriging models based on FANOVA-decomposition“. In: *Statistics and Computing* 22.3, S. 723–738. ISSN: 1573-1375. DOI: 10.1007/s11222-011-9259-7. URL: https://doi.org/10.1007/s11222-011-9259-7.

Nelder, John und Robert Wedderburn (Aug. 1972). *Generalized Linear Models*. Wiley. ISBN: 0412317605.

Pinker, Steven (2003). *The Language Instinct: How the Mind Creates Language*. Penguin UK.

Ramachandra, V. (Feb. 2018). „Deep Learning for Causal Inference“. In: *ArXiv e-prints*. arXiv: 1803.00149.

Roosen, Charles Benjamin (1995). „Visualization and Exploration of High-dimensional Functions Using the Functional Anova Decomposition“. UMI Order No. GAX96-02949. Diss. Stanford, CA, USA: Stanford University.

Rosenblatt, F. (1958). „The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain“. In: *Psychological Review*, S. 65–386.

Squire, William und George Trapp (1998). „Using Complex Variables to Estimate Derivatives of Real Functions“. In: *SIAM Review* 40, S. 110–112.

Tulio Ribeiro, M., S. Singh und C. Guestrin (Feb. 2016). „“Why Should I Trust You?”: Explaining the Predictions of Any Classifier“. In: *ArXiv e-prints*. arXiv: 1602.04938.

Wager, S. und S. Athey (Okt. 2015). „Estimation and Inference of Heterogeneous Treatment Effects using Random Forests“. In: *ArXiv e-prints*. arXiv: 1510.04342 [stat.ME].

Zeiler, M. D und R. Fergus (Nov. 2013). „Visualizing and Understanding Convolutional Networks“. In: *ArXiv e-prints*. arXiv: 1311.2901 [cs.CV].

Abbildungsverzeichnis

Tabellenverzeichnis

Definitionen, Propositionen und Modelle

Appendix

Verwendete Software

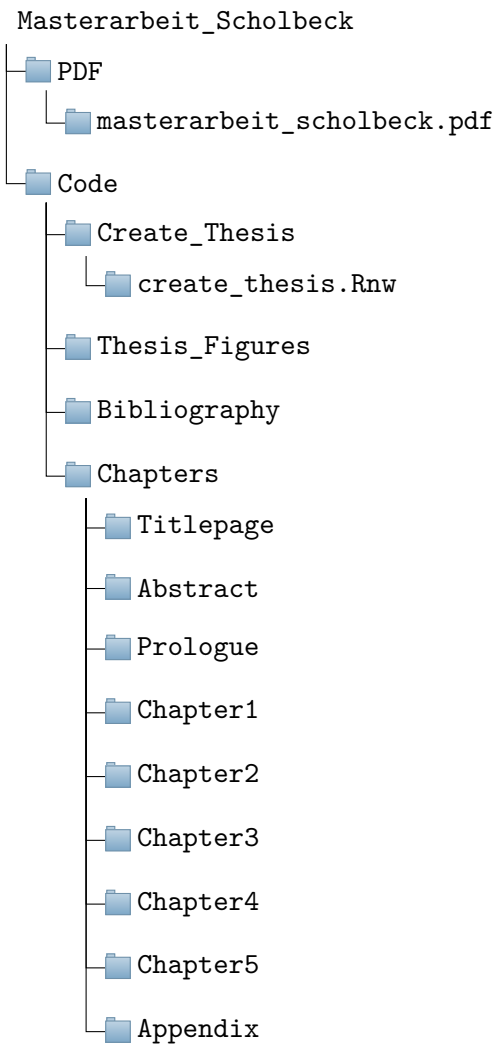
Die Generierung des vorliegenden Textmaterials erfolgte in der Zeichensetzungssprache *LaTex*. Die Umsetzung von Berechnungen und Visualisierungen erfolgte in der Programmiersprache *R*. Die folgenden Softwarepakete wurden für die prorgrammiertechnische Umsetzung in R verwendet:

- ALEPlot
- alphahull
- ame
- BBmisc
- data.table
- dplyr
- e1071
- ggplot2
- grDevices
- gridExtra
- ICEbox
- iml
- knitr
- MASS
- mlr
- numDeriv
- plot3D
- randomForest
- reshape2

- shape
- SimilarityMeasures
- tidyr

Elektronischer Anhang

Der beigefügte elektronische Anhang enthält sowohl die vorliegende Abschlussarbeit als PDF-Datei, als auch den erzeugenden Code. Die Dateien liegen in der folgenden Ordnerstruktur vor.



Eidesstattliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Diese Erklärung erstreckt sich auch auf in der Arbeit enthaltene Graphiken und Zeichnungen.