

Fakultät Für Informatik und Mathematik
Stochastic Engineering in Business and Finance

Master-Thesis

Interpretation linearer Modelle mit SHAP

Interpretation of linear models with SHAP

Betreuer: Prof. Dr. Andreas Zielke

Eingereicht von:
Simon Symhoven, 49651418
Boschetsriederstraße 59A, D-81379 München
simon.symhoven@hm.edu

Eingereicht am:
München, den 4. November 2023

Abstract

Diese Masterarbeit beleuchtet die Interpretation linearer Modelle mit Shapley Additive exPlanations (SHAP), welches seine theoretische Basis in den Shapley-Werten der kooperativen Spieltheorie findet. Nach einer historischen Einordnung und Begriffsdefinition werden die Shapley-Werte formal hergeleitet und deren axiomatische Grundlagen beleuchtet. Der Übergang von Shapley-Werten zu SHAP wird zeigen wie Beiträge einzelner Merkmale zur Modellvorhersage beitragen. Am Beispiel einer ausgewählten Modellklasse und unter Verwendung des shap Python-Pakets wird die praktische Anwendbarkeit von SHAP auf einen konkreten Datensatz demonstriert. Die Arbeit schließt mit einer Diskussion über die Grenzen von SHAP und bietet einen Ausblick auf dessen Einsatzmöglichkeiten für transparente und nachvollziehbare Modellentscheidungen in der Datenwissenschaft.

Inhaltsverzeichnis

Abkürzungsverzeichnis	v
1 Einleitung	1
2 Hintergrund	3
2.1 Kooperative Spieltheorie	3
2.2 Formale Definition	3
2.3 Eigenschaften	4
3 Machine Learning und Shapley Values	7
3.1 Erwartete Auszahlung des Models	7
3.2 Axiome	7
3.3 Approximierung der Shapley Values	7
3.4 Causal Shapley Values	7
4 Praktische Anwendung	9
4.1 Vorstellung der Datensätze	9
4.1.1 Datensatz 1	9
4.1.2 Datensatz 2	9
4.2 Einleitung Python Paket SHAP	9
4.3 Model	9
5 Ausblick	11
6 Fazit	13
Literaturverzeichnis	15

Abbildungsverzeichnis	17
Tabellenverzeichnis	19
Quellcodeverzeichnis	21
Eidesstattliche Erklärung	23
Anhänge	25
Anhang A Quellcode	25

Abkürzungsverzeichnis

SHAP Shapley Additive exPlanations

1. Einleitung

In einer Zeit, in der datengetriebene Ansätze und automatisierte Modelle immer größere Relevanz erlangen, rückt die Notwendigkeit der Erklärbarkeit und Interpretierbarkeit von Modellen in den Vordergrund. Eines der vielversprechendsten Konzepte, das sich dieser Herausforderung annimmt, sind die sogenannten Shapley Values. Diese Masterarbeit erkundet die tiefgreifenden Konzepte der Shapley Values, ihre Anwendungen im Kontext von Machine Learning-Modellen und ihre praktische Umsetzung auf reale Datensätze.

Die Arbeit beginnt mit einer umfassenden Einführung in die Shapley Values und ihre historischen Wurzeln. Dabei wird insbesondere auf die kooperative Spieltheorie als Ursprung dieser Konzepte eingegangen. Anhand ausgewählter Literatur werden die theoretischen Grundlagen erörtert und der Forschungsstand auf diesem Gebiet aufgezeigt.

Im Anschluss daran wird die Brücke zur aktuellen Landschaft des maschinellen Lernens geschlagen. Es wird beleuchtet, wie die Shapley Values adaptiert werden können, um Einblicke in die Gewichtung von Merkmalen in komplexen Machine Learning-Modellen zu gewinnen. Dabei wird auf bestehende Methoden und Ansätze Bezug genommen und diskutiert, wie diese auf verschiedene Modelle angewendet werden können.

Ein zentraler Schwerpunkt der Arbeit liegt auf der praktischen Anwendung der Shapley Values. Ein realer Datensatz wird vorgestellt und die Methodik wird auf diesen angewendet, um die Wirksamkeit und Aussagekraft der Shapley Values in der Praxis zu evaluieren. Dies ermöglicht eine kritische Reflexion über die Stärken und Limitationen dieses Ansatzes im Kontext der Datenerklärung.

Abschließend werden die gewonnenen Erkenntnisse zusammengeführt und ein Ausblick auf zukünftige Entwicklungen und Forschungsrichtungen gegeben. Die Arbeit trägt somit dazu bei, das Verständnis für die Shapley Values als Instrument der Erklärbarkeit in komplexen Modellen zu vertiefen und ihre praktische Anwendbarkeit zu beleuchten.

2. Hintergrund

2.1. Kooperative Spieltheorie

Der Ursprung der Shapley Values liegt in der kooperativen Spieltheorie, einem fundamentalen Zweig der Spieltheorie. Dieser Bereich beschäftigt sich mit der Analyse von Situationen, in denen Akteure zusammenarbeiten, um gemeinsame Ziele zu erreichen. Zentrales Anliegen ist dabei die gerechte Verteilung der entstehenden Gewinne unter den Akteuren. Ein Schlüsselkonzept dieser Theorie ist die sogenannte "Charakteristische Funktion", welche die Bewertung der Gewinnverteilung einer Koalition von Akteuren ermöglicht.

Die Shapley Values, entwickelt von Lloyd Shapley in den 1950er Jahren, bieten einen methodischen Ansatz, um den individuellen Beitrag eines jeden Akteurs zur kooperativen Zusammenarbeit gerecht zu bewerten. Dies geschieht durch die Durchschnittsbewertung der Beiträge über sämtliche mögliche Koalitionen hinweg. Diese Methode erweist sich als äußerst nützlich, um eine gerechte und rationale Verteilung von Gewinnen in vielfältigen Szenarien zu ermöglichen, sei es in wirtschaftlichen Verhandlungen oder der Aufteilung von Ressourcen.

Das Verständnis der kooperativen Spieltheorie und ihrer Anwendung in Form der Shapley Values ermöglicht es, dieses theoretische Konzept auf den Bereich des maschinellen Lernens zu übertragen. In dieser Arbeit werden wir den Übergang von abstrakten Spieltheorie-Konzepten zu konkreten Anwendungen in der Welt der datengetriebenen Modelle erforschen.

Zur Erreichung dieses Ziels werden in den kommenden Abschnitten nicht nur die formalen Definitionen und Eigenschaften der Shapley Values erläutert, sondern auch ihre Adaption und Anwendung auf Machine Learning-Modelle in Betracht gezogen. Die Anwendbarkeit wird durch die praktische Anwendung auf einen realen Datensatz verdeutlicht.

2.2. Formale Definition

Sei $\mathcal{N} = \{1, \dots, n\}$ eine endliche Spielermenge mit $n := |\mathcal{N}|$ Elementen. Sei v die **Koalitionsfunktion**, die jeder Teilmenge von \mathcal{N} eine reelle Zahl zuweist

und insbesondere der leeren Koalition den Wert 0 gibt.

$$\begin{aligned} v &: \mathcal{P}(\mathcal{N}) \longrightarrow \mathbb{R} \\ &: v(\emptyset) \mapsto 0 \end{aligned}$$

Eine nicht leere Teilmenge der Spieler $\mathcal{S} \subseteq \mathcal{N}$ heißt Koalition. \mathcal{N} selbst bezeichnet die große Koalition. Den Ausdruck $v(\mathcal{S})$ nennt man den Wert der Koalition \mathcal{S} . Der Shapley-Wert ordnet nun jedem Spieler aus \mathcal{N} eine Auszahlung für das Spiel v zu.

Der marginale Beitrag eines Spieler $i \in \mathcal{N}$, also der Wertbeitrag eines Spielers zu einer Koalition $\mathcal{S} \subseteq \mathcal{N}$, durch seinen Beitritt, ist

$$v(\mathcal{S} \cup \{i\}) - v(\mathcal{S}).$$

Der Shapley-Wert eines Spielers i errechnet sich als das gewichtete Mittel der marginalen Beiträge zu allen möglichen Koalitionen:

$$\varphi_i(\mathcal{N}, v) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}} \underbrace{\frac{|\mathcal{S}|! \cdot (n - 1 - |\mathcal{S}|)!}{n!}}_{\text{Gewicht}} \underbrace{v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})}_{\text{marginaler Beitrag von Spieler } i \text{ zur Koalition } \mathcal{S}}.$$

2.3. Eigenschaften

Pareto-Effizienz Der Wert der großen Koalition wird an die Spieler verteilt:

$$\sum_{i \in \mathcal{N}} \varphi_i(\mathcal{N}, v) = v(\mathcal{N}).$$

Symmetrie Zwei Spieler i und j , die die gleichen marginalen Beiträgen zu jeder Koalition haben,

$$v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\})$$

erhalten das Gleiche:

$$\varphi_i(\mathcal{N}, v) = \varphi_j(\mathcal{N}, v).$$

Null-Spieler-Eigenschaft Ein Spieler der zu jeder Koalition nichts bzw. den Wert seiner Einer-Koalition beiträgt, erhält null bzw. den Wert seiner Einer-Koalition:

$$\varphi_i(\mathcal{N}, v) = 0,$$

bzw.

$$\varphi_i(\mathcal{N}, v) = v(\{i\}).$$

Additivität Wenn das Spiel in zwei unabhängige Spiele zerlegt werden kann, dann ist die Auszahlung jedes Spielers im zusammengesetzten Spiel die Summe der Auszahlungen in den aufgeteilten Spielen:

$$\varphi_i(\mathcal{N}, v + w) = \varphi_i(\mathcal{N}, v) + \varphi_i(\mathcal{N}, w).$$

3. Machine Learning und Shapley Values

3.1. Erwartete Auszahlung des Models

3.2. Axiome

3.3. Approximierung der Shapley Values

3.4. Causal Shapley Values

4. Praktische Anwendung

4.1. Vorstellung der Datensätze

4.1.1. Datensatz 1

4.1.2. Datensatz 2

4.2. Einleitung Python Paket SHAP

4.3. Model

5. Ausblick

6. Fazit

Literaturverzeichnis

- [1] Encarnación Algaba, Vito Fragnelli, and Joaquín Sánchez-Soriano. Hand-book of the shapley value. 2019.
- [2] Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.
- [3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [4] Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions - journal of computer-aided molecular design, May 2020.
- [5] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5572–5579. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Survey Track.
- [6] Joachim Schlosser. *Wissenschaftliche Arbeiten schreiben mit LATEX: Leit-faden für Einsteiger: Joachim Schlosser*. mitp, Heidelberg and München and Landsberg and Frechen and Hamburg, 4 edition, 2011.
- [7] L. S. Shapley. 17. *A Value for n -Person Games*, pages 307–318. Princeton University Press, Princeton, 1953.

Abbildungsverzeichnis

Tabellenverzeichnis

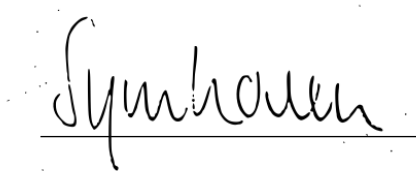
Quellcodeverzeichnis

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen übernommen wurden, sind als solche kenntlich gemacht. Alle Internetquellen sind der Arbeit beigefügt.

Des Weiteren versichere ich, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und dass die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

München, 4. November 2023

A handwritten signature in black ink, reading 'Symhoven', written over a horizontal line.

SIMON SYMHOVEN

A. Quellcode

Quellcode