INDUSTRIE- UND HANDELSKAMMER

ABSCHLUSSPRÜFUNG ZUM DATA ANALYST (IHK)

Online-Zertifikatslehrgang Data Analyst (IHK)

Abschlussbericht

Einkaufsabschlüsse im Online-Shop

Herr Alim Atca, MSc.	Autor
Herr Lewis Koua, MSc.	Autor
Herr Dominik Dietrich, MSc.	Autor
Frau Maddalena Zaffiro	Prüferin

Abgabefrist: 10.02.2026

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

1 Einleitung

1.1 Motivation und Zielsetzung des Projekts

1.2 Ausgangssituation und Hintergrund der Daten

Die vorliegenden Daten stammen aus dem Webtracking eines Online-Shops und erfassen das Verhalten von Besuchern während einzelner Website-Sessions. Ziel des Unternehmens ist es, besser zu verstehen, unter welchen Bedingungen ein Kaufabschluss erfolgt, um Marketing- und Shop-Maßnahmen gezielt zu optimieren.

Die Daten bilden einen realistischen E-Commerce-Prozess ab (vergleichbar mit Google Analytics) und enthalten Informationen zu:

- Nutzerverhalten (z. B. Produktseiten, Verweildauer),
- Einkaufssituation (Monat, Wochenende, Special Days),
- technischem Kontext (Browser, Betriebssystem),
- sowie dem Kaufabschluss (Revenue).

1.3 Geschäftlicher Hintergrund (Online-Shop & Kaufabschlüsse)

1.4 Projektabgrenzung und Fokus der Analyse

1.5 Hypothese H1: Zusammenhang zwischen Produktinteraktion und Kaufabschluss

2 Projektvorbereitung

2.1 Projektmanagement

2.1.1 Organisation und Kommunikation

2.1.2 Projektmanagementmethode und Arbeitsteilung

Die Projektplanung und -verwaltung wurde mithilfe eines Kanban-Boards (siehe **??**) und eines Gantt-Diagramms (siehe **??**) durchgeführt. Beide Visualisierungen befinden sich im Anhang.

2.2 Ist-Analyse

2.2.1 Geschäftsszenario

2.2.2 IT-Infrastruktur

3 Erhebung und Analyse

3.1 Bedarfsanalyse

3.2 Prozessanalyse (KNIME)

3.2.1 Datenaufbereitung und Datenqualität

3.2.2 Spaltenumbenennen

Die Spalten des ursprünglichen Datensatzes wurden vor der weiteren Verarbeitung umbenannt. Der Datensatz enthält zwar technisch korrekte Bezeichnungen, diese sind jedoch teilweise uneinheitlich und erschweren sowohl die Lesbarkeit als auch die spätere Dokumentation. Aus diesem Grund wurde eine einheitliche Benennung nach dem Schema `snake_case`[1] eingeführt. Dadurch lassen sich die Spalten leichter interpretieren, konsistent weiterverarbeiten und eindeutig im Bericht sowie im Dashboard referenzieren. Zusätzlich wurde darauf geachtet, dass ähnliche Inhalte ein gleiches Namensmuster erhalten, um Zusammenhänge direkt erkennbar zu machen. Beispiele:

- Spalten mit Seitenanzahlen enden auf `_page_count`
- Spalten mit Zeitangaben enden auf `_page_duration`
- Boolesche Merkmale erhalten das Suffix `_flag`

Diese Standardisierung reduziert Fehler im Workflow, erleichtert die Wartung und ist insbesondere für die Weiterverarbeitung in Power BI von Vorteil. Die vollständige Übersicht aller Umbenennungen ist in Tabelle **??** dargestellt:

Tabelle 3.1: Spaltenumbenennungen der CSV-Daten

Original CSV-Spalte	**Neuer Spaltenname**	**Begründung**
Administrative	`admin_page_count`	Einheitliche Benennung für Seitenanzahlen (_page_count), beschreibt die Anzahl administrativer Seitenaufrufe pro Session.
Administrative_Duration	`admin_page_duration`	Einheitliche Benennung für Zeitangaben (_page_duration), gibt die Verweildauer auf administrativen Seiten an.
Informational	`info_page_count`	Vereinheitlichung der Seitenkategorien, beschreibt die Anzahl besuchter Informationsseiten.

[1] https://developer.mozilla.org/de/docs/Glossary/Snake_case

Tabelle 3.1 – Fortsetzung

Original CSV-Spalte	**Neuer Spaltenname**	**Begründung**
Informational_Duration	`info_page_duration`	Klare Trennung zwischen Seitenanzahl und Verweildauer auf Informationsseiten.
ProductRelated	`product_page_count`	Zentrale Variable zur Überprüfung der Hypothese H1, da Produktseiten unmittelbar mit Kaufentscheidungen zusammenhängen.
ProductRelated_Duration	`product_page_duration`	Ergänzt product_page_count um die zeitliche Intensität der Produktbetrachtung.
BounceRates	`bounce_rate`	Vereinfachter Name für eine bekannte Google-Analytics-Metrik, beschreibt Absprungraten pro Seite.
ExitRates	`exit_rate`	Vereinfachter, einheitlicher Name für eine weitere Google-Analytics-Metrik, beschreibt Ausstiegsraten.
PageValues	`page_value`	Kürzere, besser interpretierbare Bezeichnung für den monetären Wert einer Seite vor einem Kauf.
SpecialDay	`special_day_score`	Verdeutlicht, dass es sich um einen normierten Score (0–1) handelt, der die Nähe zu Aktionstagen abbildet.
Month	`visit_month`	Präzisiert die Bedeutung als Monat der Session, nicht als abstrakter Kalendermonat.
OperatingSystems	`operating_system_id`	Technische Kontextvariable, numerische Kodierung zur Identifikation des Betriebssystems.

Tabelle 3.1 – Fortsetzung

Original CSV-Spalte	Neuer Spaltenname	Begründung
Browser	`browser_id`	Technische Kontextvariable, numerische Kodierung des verwendeten Browsers.
Region	`region_id`	Regionale Zuordnung des Nutzers, technisch als ID gespeichert.
TrafficType	`traffic_type_id`	Kennzeichnet den Ursprung des Traffics (z. B. Direkt, Referral, Kampagne).
VisitorType	`visitor_type`	Beibehaltung der semantischen Bedeutung ("New" / "Returning"), nur sprachlich vereinheitlicht.
Weekend	`weekend_flag`	Boolean-Indikator, der klar als Ja/Nein-Merkmal erkennbar ist.
Revenue	`revenue_flag`	Kennzeichnet, ob ein Kauf stattgefunden hat (TRUE/FALSE); Umbenennung macht die Rolle als Zielvariable eindeutig.

3.2.3 EDA

Die explorative Datenanalyse diente dazu, ein grundlegendes Verständnis der Daten zu gewinnen und die Hypothese H1 zu überprüfen. Untersucht wurden unter anderem:

- Verteilung der Zielvariable (Kauf vs. kein Kauf)
- Unterschiede im Nutzerverhalten zwischen Käufern und Nicht-Käufern
- Zusammenhang zwischen Produktseiteninteraktionen und Kaufabschluss

Neben den ursprünglichen Spalten wurden zusätzliche Merkmale berechnet, um das Nutzerverhalten auf Session-Ebene besser zu beschreiben. Ziel war es, robuste Kennzahlen zu erzeugen, die unabhängig von der absoluten Sessionlänge interpretiert werden können und eine bessere Vergleichbarkeit zwischen verschiedenen Besuchersessions

ermöglichen. Zunächst wurde die Gesamtanzahl aller Seiteninteraktionen pro Session berechnet (`total_page_count`). Diese Kennzahl beschreibt die allgemeine Aktivität eines Nutzers während einer Session und fasst alle besuchten Seitentypen zusammen.

$$\texttt{total\_page\_count} = \texttt{admin\_page\_count} + \texttt{info\_page\_count} + \texttt{product\_page\_count} \tag{3.1}$$

Ergänzend dazu wurde die gesamte Verweildauer innerhalb der erfassten Seitenkategorien bestimmt (`total_duration`). Diese Größe dient als Maß für die Dauer und Intensität der Session und berücksichtigt, wie viel Zeit ein Nutzer insgesamt im Shop verbracht hat.

$$\texttt{total\_duration} = \texttt{admin\_page\_duration} + \texttt{info\_page\_duration} + \texttt{product\_page\_duration} \tag{3.2}$$

Auf Basis dieser beiden Kennzahlen wurde die durchschnittliche Verweildauer pro Seiteninteraktion berechnet (`avg_time_per_page`). Diese Kennzahl ermöglicht es, zwischen kurzen, oberflächlichen Besuchen und intensiver Auseinandersetzung mit den Inhalten zu unterscheiden. Zur Vermeidung einer Division durch Null wurde der Nenner um eine Konstante ergänzt.

$$\texttt{avg\_time\_per\_page} = \frac{\texttt{total\_duration}}{\texttt{total\_page\_count} + 1} \tag{3.3}$$

Ein zentrales, hypothesengeleitetes Merkmal ist der Anteil der Produktseiten an allen Seiteninteraktionen (`product_share`). Diese Kennzahl beschreibt, wie stark sich ein Nutzer innerhalb einer Session auf Produktseiten konzentriert hat. Dadurch wird nicht nur die absolute Anzahl der Produktseiten berücksichtigt, sondern auch deren Bedeutung im Verhältnis zur gesamten Session.

$$\texttt{product\_share} = \frac{\texttt{product\_page\_count}}{\texttt{total\_page\_count} + 1} \tag{3.4}$$

Durch diese berechneten Merkmale wird das Nutzerverhalten differenzierter abgebildet. Insbesondere `product_share` stellt ein zentrales Merkmal zur Überprüfung der Hypothese dar, dass eine stärkere Fokussierung auf Produktseiten mit einer höheren Wahrscheinlichkeit für einen Kaufabschluss einhergeht.

3.2.4 Modellierung und Vorhersage

Für die Modellierung wurde bewusst ein begrenztes Feature-Set gewählt. Ziel war es nicht, möglichst viele Variablen zu verwenden, sondern diejenigen, die fachlich sinnvoll

begründet und gut interpretierbar sind.

Im finalen Feature-Set enthalten sind:

- `product_page_count` als zentrales Hypothesenmerkmal
- `product_share` als Verstärkung des Produktfokus
- `product_page_duration` als Maß für Engagement
- `page_value` als kaufnahe Metrik
- `total_duration` zur Abbildung der Sessionintensität
- kodierte Kontextvariablen wie Monat, Besuchertyp und Wochenende

Durch diese Auswahl bleibt das Modell übersichtlich, erklärbar und reduziert das Risiko von Überanpassung.

4 Visualisierung und Reporting (Power BI)

4.1 Datenübergabe von KNIME an Power BI

4.2 Definition der Business-KPIs

4.3 Dashboard-Design und Visualisierungen

4.4 Interpretation der Ergebnisse für Fachabteilungen

4.5 Management-Mehrwert durch datenbasierte Entscheidungen

5 Soll-Zustand

5.1 Zielbild der datengetriebenen Kaufvorhersage

5.2 Integration des Modells in Geschäftsprozesse

5.3 Nutzen für Marketing und Shop-Optimierung

5.4 Grenzen des Modells und Risiken

6 Ausblick

7 Anhang

7.1 Projektmanagement

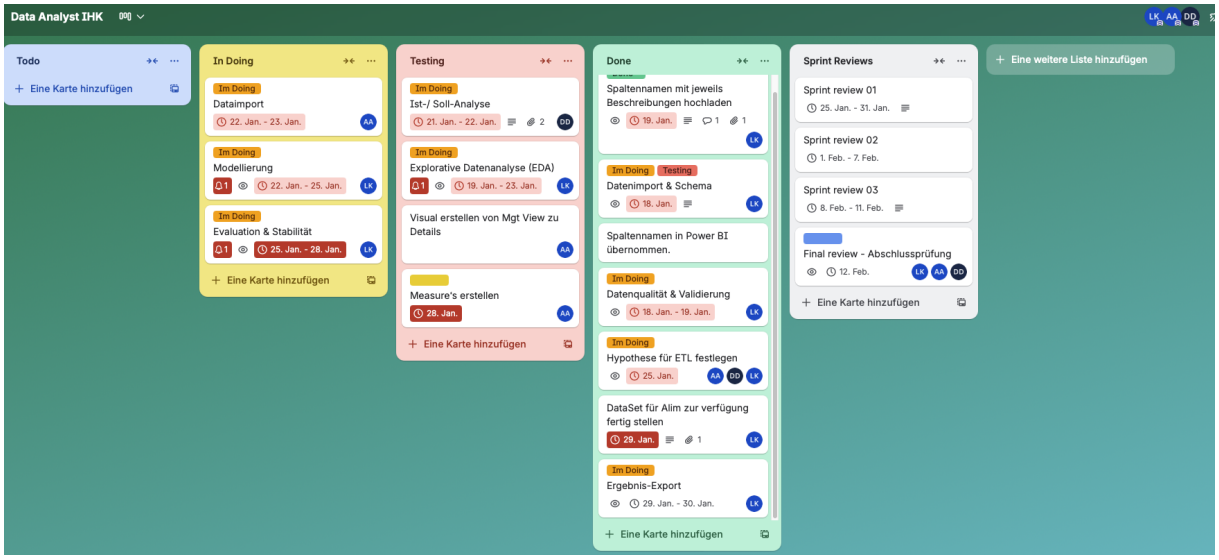

Abbildung 7.1: Kanban-Board mit Projektaufgaben und Sprint-Reviews

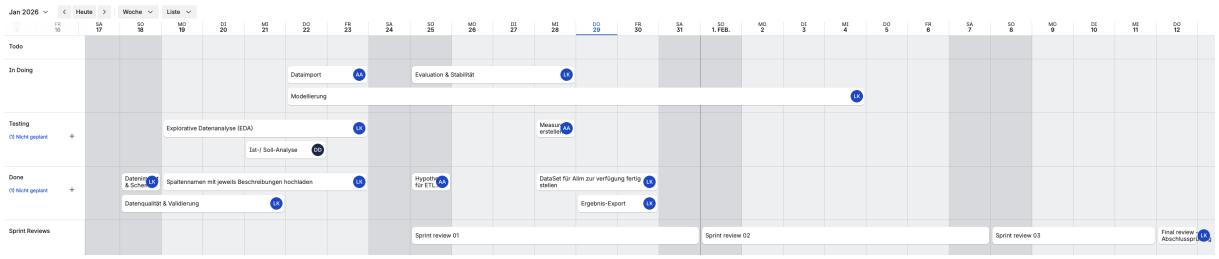

Abbildung 7.2: Gantt-Diagramm mit Projektzeitplan

Literaturverzeichnis

[1] Donald E. Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, 1984.

[2] Donald E. Knuth. *The TEX Book*. Addison-Wesley Professional, 1986.

[3] Leslie Lamport. *LATEX: a Document Preparation System*. Addison Wesley, Massachusetts, 2 edition, 1994.

[4] Michael Lesk and Brian Kernighan. Computer typesetting of technical journals on UNIX. In *Proceedings of American Federation of Information Processing Societies: 1977 National Computer Conference*, pages 879–888, Dallas, Texas, 1977.

[5] Frank Mittelbach, Michel Gossens, Johannes Braams, David Carlisle, and Chris Rowley. *The LATEX Companion*. Addison-Wesley Professional, 2 edition, 2004.