**INDUSTRIE- UND HANDELSKAMMER**

ABSCHLUSSPRÜFUNG ZUM DATA ANALYST (IHK)


## Online-Zertifikatslehrgang Data Analyst (IHK)

---

Abschlussbericht

# Einkaufsabschlüsse im Online-Shop

**Herr Alim Atca, B.A.** Autor
**Herr Lewis Koua, MSc.** Autor
**Herr Dominik Dietrich** Autor

Abgabefrist: 10.02.2026

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# 1 Einleitung und Ist-Analyse

## 1.1 Ausgangssituation

In unserem Online-Shop sehen wir aktuell eine deutliche Lücke zwischen den Besucherzahlen und den tatsächlichen Abschlüssen. Viele Nutzer investieren Zeit in die Produktsuche und füllen ihren Warenkorb, brechen den Prozess dann aber ohne ersichtlichen Grund ab.

Wir möchten herausfinden, welche Verhaltensmuster statistisch gesehen zu einem Kauf führen und wo wir potenzielle Käufer verlieren. Ziel ist es, eine Prognose darüber abzuleiten, unter welchen Bedingungen ein Kaufabschluss wahrscheinlicher ist.

Dabei entsteht eine große Menge an Verhaltens- und Session-Daten. Diese Daten bilden die Grundlage, um Muster zu erkennen und erfolgreiche Käufe von Abbrüchen zu unterscheiden.

Um die Analyse überschaubar zu halten, werden die Hypothesen auf ausgewählte Dimensionen beschränkt. Dazu gehören Besuchertyp (Visitor Type), die Anzahl der aufgerufenen Seiten (Page Count), technischer Kontext (Browser, Betriebssystem), Kaufabschluss (Revenue), die Verweildauer (Page Duration) sowie die Einkaufssituation (Monat, Wochenende, Special Days).

## 1.2 Erläuterung der Datenquellen und Prozesse

Die verwendeten Daten stammen aus dem aufgezeichneten Kundenverhalten einer Online-Shop-Website. Die Datenbasis bildet der Datensatz „Online Shoppers Purchasing Intention“[1] aus dem UCI Machine Learning Repository. Dabei wurden Informationen zum Verhalten der Nutzer sowie zur jeweiligen Einkaufssituation erfasst.

- **Anzahl Sessions:** 12.330
- **Merkmale in 18 Bereichen:**
  - Nutzerverhalten: Seitenanzahl, Verweildauer
  - Engagement: BounceRate, ExitRate, PageValue
  - Zeit & Kontext: Monat, SpecialDay, Wochenende
  - Technik & Segment: Browser, OS, VisitorType, TrafficType
  - Zielvariable: Revenue (Kaufabschluss ja/nein)

Die Spalten des ursprünglichen Datensatzes wurden vor der weiteren Verarbeitung umbenannt; die Struktur der Quelldaten aus [1] hat uns dabei als Grundlage für die Spaltenumbenennungen geholfen. Eine einheitliche Benennung nach dem Schema `snake_case`[1] erleichtert die Lesbarkeit, die konsistente Weiterverarbeitung und die Referenzierung im Bericht sowie im Dashboard. Die vollständige Übersicht aller Umbenennungen ist im Anhang in Tabelle 8.3 dargestellt.

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

# 2 Backlog und Projektziele

## 2.1 Sammlung von User Storys und Potenzialen

Auf Basis der Daten wurden erste User Storys erstellt, die uns helfen, Fragestellungen aus Sicht des Unternehmens zu formulieren. Die Projektplanung und -verwaltung wurde mithilfe eines Kanban-Boards (siehe Abb. 8.1) und eines Gantt-Diagramms (siehe Abb. 8.2) durchgeführt; beide Visualisierungen befinden sich in Kapitel 8.

- **US 1 – Produktinteraktion beeinflusst Kaufabschluss positiv**
  Als Online-Marketing-Manager: Möchte ich vorhersagen, ob eine Session zu einem Kauf führt, um Marketingmaßnahmen gezielt auszurichten.
- **US 2 – Returning Visitors kaufen häufiger als neue Besucher**
  Als Marketing Manager: Möchte ich zwischen neuen und wiederkehrenden Besuchern unterscheiden können, um gezielte Marketingmaßnahmen für Returning Visitors umzusetzen.
- **US 3 – Je länger sich der Besucher auf der Seite aufhält, desto wahrscheinlicher der Kaufabschluss**
  Als Data Analyst: Möchte ich Muster im Nutzerverhalten erkennen, um wichtige Einflussfaktoren des Kaufabschlusses zu identifizieren.
- **US 4 – Besuche an Wochenenden führen zu höherer Kaufwahrscheinlichkeit**
  Als Produktmanager: Möchte ich ein ML-Modell, das die Kaufwahrscheinlichkeit prognostiziert, um Nutzer in Echtzeit zu klassifizieren.
- **US 5 – Returning Visitors haben eine höhere Conversion Rate als New Visitors**
  Als Marketing Manager: Möchte ich vorhersagen, ob eine Session zu einem Kauf führt, um Marketingmaßnahmen gezielt auszurichten.

## 2.2 Bewertung und Priorisierung nach Aufwand und Ertrag (Low Hanging Fruits)

Nach der Sammlung der verschiedenen Ansätze wurde die erste User Story priorisiert. Die Fragestellung liefert einen direkten Einblick in das Kaufverhalten der Nutzer; die Kennzahlen Page Count und Page Duration stehen in engem Zusammenhang mit dem Kaufabschluss und eignen sich gut, um das Kundeninteresse während einer Session

einzuschätzen. Die Priorisierung folgte dem Prinzip der Low Hanging Fruits: Es wurden Fragen ausgewählt, die mit geringem Aufwand untersucht werden können und einen hohen Mehrwert bieten.

## 2.3 Auswahl der erreichbaren Projektziele für den ersten Sprint

Für den ersten Sprint wurden Punkte ausgewählt, die mit geringem Aufwand umsetzbar sind und die vorhandenen Daten direkt nutzen. Ziel ist es, mögliche Zusammenhänge, ein besseres Verständnis und erste Ergebnisse bis zum Ende des ersten Sprints sichtbar zu machen – ohne dass dafür z. B. KNIME oder Power BI zwingend nötig sind. Konkret: Fokus auf Datenaufbereitung, EDA, Modellierung (Kaufvorhersage) sowie Visualisierung und Reporting in Power BI.

# 3 Projektausarbeitung

## 3.1 Formulierung von Hypothesen, die die Machbarkeit der Funktionen der User Storys untermauern

Für jede User Story wurden Hypothesen formuliert, um die Machbarkeit der geplanten Maßnahmen anhand der vorhandenen Daten zu untermauern.

**H1 – Produktinteraktion beeinflusst Kaufabschluss positiv:** Nutzer, die mehr produktbezogene Seiten ansehen oder mehr Zeit darauf verbringen, schließen Käufe ab. (Var: Product-related und Revenue.) Diese Hypothese untermauert die Machbarkeit der User Story zur Kaufvorhersage und wurde in der EDA (Kapitel 1) sowie in der Modellierung (Kapitel 6) überprüft.

**H2 – Returning Visitors kaufen häufiger als neue Besucher.** Var: Visitor Type und Conversion-Rate (Anzahl Revenue=True / Anzahl der Sessions).

**H3 – Je länger sich der Besucher auf der Seite aufhält, desto wahrscheinlicher der Kaufabschluss.** Var: Product-related-Duration und Revenue.

**H4 – Besuche an Wochenenden führen zu höherer Kaufwahrscheinlichkeit.** Var: Conversion-Rate und Weekend.

**H5 – Returning Visitors haben eine höhere Conversion Rate als New Visitors.** Messvariable: `visitor_type`, `revenue_flag` (relevant für Marketingmanager).

## 3.2 Zuordnung der Hypothesen zu User Storys und Potenzialen

Diese Hypothesen verdeutlichen, welche Zusammenhänge im Nutzerverhalten untersucht werden sollen. Sie zeigen, welche Annahmen den jeweiligen Fragestellungen zugrunde liegen und welches Optimierungspotenzial besteht – und machen auf einen Blick erkennbar, welche Maßnahme sinnvoll ist und welche Kundengruppe betroffen ist.

Tabelle 3.1: Zuordnung der User Storys zu Hypothesen und Potenzialen

| **User Story** | **Potenzial** | **Hypothese** |
|---|---|---|
| Als Online-Marketing-Manager: vorhersagen, ob eine Session zu einem Kauf führt, um Marketingmaßnahmen gezielt auszurichten. | Optimierung von Produktseiten, bessere Darstellung und Hervorhebung | Nutzer, die mehr auf produktbezogene Seiten interagieren oder mehr Zeit darauf verbringen, schließen Käufe ab. Var: Product-related und Revenue. |
| Als Marketing Manager: zwischen neuen und wiederkehrenden Besuchern unterscheiden, gezielte Maßnahmen für Returning Visitors. | Stärkung von Kundenbindung, gezieltes Retargeting und Newsletter-Maßnahmen | Returning Visitors haben eine höhere Conversion Rate als neue Besucher. Var: Visitor Type und Conversion-Rate. |
| Als Data Analyst: Muster im Nutzerverhalten erkennen, Einflussfaktoren des Kaufabschlusses identifizieren. | Verbesserung der Nutzerführung, Reduktion von Absprüngen, optimierte Seitenstruktur | Je länger sich der Besucher auf der Seite aufhält, desto wahrscheinlicher der Kaufabschluss. Var: Product-related-Duration und Revenue. |
| Als Produktmanager: ML-Modell zur Prognose der Kaufwahrscheinlichkeit, Nutzer in Echtzeit klassifizieren. | Zeitlich optimierte Marketingkampagnen und Angebote am Wochenende | Besuche an Wochenenden führen zu höherer Kaufwahrscheinlichkeit. Var: Conversion-Rate und Weekend. |
| Als Marketing Manager: vorhersagen, ob eine Session zu einem Kauf führt, Marketingmaßnahmen ausrichten. | Effizienter Einsatz des Marketingbudgets, Fokus auf profitable Zielgruppen | Returning Visitors haben eine höhere Conversion Rate als New Visitors. Messvariable: `visitor_type`, `revenue_flag`. |

# 4 Projektmetriken

## 4.1 Fortschritt und Erfolg des Projektes sowie Qualität der Analyse bzw. des Dashboards im Einsatz

Für die weitere Analyse werden zunächst relevante Attribute ausgewählt, die einen Einfluss auf den Kaufabschluss haben. Dazu zählen unter anderem das Nutzerverhalten auf Produktseiten, der Besuchertyp sowie zeitliche Merkmale.

Die Daten werden in KNIME aufbereitet. Dabei ist geplant, Ausreißer zu identifizieren und gegebenenfalls zu bereinigen, um verzerrte Ergebnisse zu vermeiden. Anschließend werden die Daten in Power BI visualisiert, um Zusammenhänge zu erkennen.

Relevante Attribute der Hypothese: Nutzerverhalten auf Produktseiten; Besuchertyp; zeitliche Merkmale (z. B. Wochenende, Monat).

Zusätzlich lassen sich Fortschritt und Erfolg messen durch: Erfüllung der Sprintziele (Backlog-Status), Qualität des Modells (z. B. Vorhersagegüte), Nutzbarkeit des Dashboards für Fachabteilungen sowie die Definition und Einhaltung von Business-KPIs (vgl. Kapitel 8). Die Qualität der Analyse zeigt sich in der Nachvollziehbarkeit der EDA, der Datenqualität und der Erklärbarkeit des Modells.

## 4.2 Definitions of done für die Funktionen und die Mehrwerte der ausgewählten User Storys

**User Story 1:** Als Betreiber eines Online-Shops möchte ich verstehen, wie sich die Anzahl der Seitenaufrufe und die Verweildauer auf Produktseiten auf den Kaufabschluss auswirken.

**Definition of Done:**

- Die relevanten Variablen (Page Count, Page Duration, Revenue) wurden eindeutig identifiziert.
- Es liegt eine nachvollziehbare Auswertung vor, die zeigt, ob und wie sich Seitenaufrufe und Verweildauer auf den Kaufabschluss auswirken.
- Die Ergebnisse sind so aufbereitet, dass sie auch für fachliche Stakeholder verständlich sind.

**Mehrwert:** Das Unternehmen erhält ein besseres Verständnis dafür, welche Nutzerinteraktionen kaufrelevant sind. Marketing- und Produktmaßnahmen können gezielter auf

kaufinteressierte Nutzer ausgerichtet werden. Erste Optimierungspotenziale im Online-Shop können identifiziert werden, ohne großen technischen Aufwand.

Zusätzlich gelten die übergreifenden DoD: Daten aufbereitet und dokumentiert; EDA durchgeführt und bewertet; Modell trainiert und evaluiert; Dashboard in Power BI erstellt und mit KNIME-Daten angebunden; Dokumentation und Anhang (Kanban, Gantt) vollständig.

# 5 Planung des ersten Sprints

## 5.1 Projektorganisation und Kommunikation

Die Zusammenarbeit im Projektteam wurde durch klar definierte Kommunikationswege und ein zentrales Projektmanagement unterstützt. Die Abstimmung erfolgte regelmäßig über Zoom und WhatsApp und ermöglichte eine flexible, ortsunabhängige Kommunikation. Die Besprechungen dienten der inhaltlichen Abstimmung, der Fortschrittskontrolle und gemeinsamen Entscheidungen.

Die Datenbasis wird wöchentlich aktualisiert, sodass Analysen und Auswertungen stets auf dem aktuellen Stand sind. Durch die regelmäßigen Updates können Hypothesen strukturiert überprüft werden.

Zur Planung, Koordination und Dokumentation der Aufgaben kam Trello als Projektmanagement-Tool zum Einsatz, da es eine übersichtliche Strukturierung der Aufgaben ermöglicht. Dort wurden alle User Stories, Arbeitspakete und Aufgaben erfasst, priorisiert und den jeweiligen Projektphasen zugeordnet. So konnten die einzelnen Aufgabenpakete klar definiert werden; der Fortschritt der Arbeit war jederzeit ersichtlich.

## 5.2 Formulierung der einzelnen Aufgabenpakete mit ihren berechneten oder geschätzten Aufwänden

Die Aufgabenpakete wurden im Kanban-Board (siehe Abb. 8.1) und im Gantt-Diagramm (siehe Abb. 8.2) in Kapitel 8 abgebildet. Die folgenden Tabellen fassen zentrale Pakete und geschätzte Aufwände zusammen.

Tabelle 5.1: Erstes Aufgabenpaket: Datenüberprüfung

| **Aufgabenpaket** | **Beschreibung** | **Geschätzter Aufwand** |
|---|---|---|
| Datenüberprüfung | Überprüfung der Vollständigkeit und Plausibilität der vorhandenen Kundendaten | ca. 1,5 Stunden |

Tabelle 5.2: Weitere Aufgabenpakete des ersten Sprints

| **Aufgabenpaket** | **Beschreibung** | **Geschätzter Aufwand** |
|---|---|---|
| Einfache Datenaufbereitung | Kategorisierung der Kundendaten nach Besuchertyp, Nutzungsverhalten und Kaufstatus | ca. 2 Stunden |
| Hypothesenprüfung | Erste Einschätzung, ob die Hypothesen plausibel sind | ca. 1,5 Stunden |
| Erste Visualisierung | Darstellung der Kennzahlen in Diagrammen, um Trends erkennbar zu machen | ca. 1 Stunde |
| Dokumentation der Ergebnisse | Zusammenfassung der Aufgaben, Beschreibung und Erkenntnisse | ca. 1,5 Stunden |

Weitere zentrale Pakete: Projektvorbereitung und Ist-Analyse, Datenaufbereitung in KNIME, Spaltenumbenennen, EDA, Modellierung (XGBoost), Datenübergabe KNIME–Power BI, Business-KPIs und Dashboard-Design, Interpretation und Management-Mehrwert. Die Aufwände wurden im Gantt-Diagramm geplant und in den Sprint-Reviews überprüft. Weitere Darstellungen (Kanban-Board, Gantt-Diagramm) siehe Kapitel 8.

# 6 Technische Beschreibung des ersten Sprints anhand des Machine Learning Canvas

Das Machine Learning Canvas (siehe Abb. 6.1) fasst Aufgaben, Datenquellen, Merkmale, Modellierung, Vorhersage und Evaluierung des ersten Sprints übersichtlich zusammen.

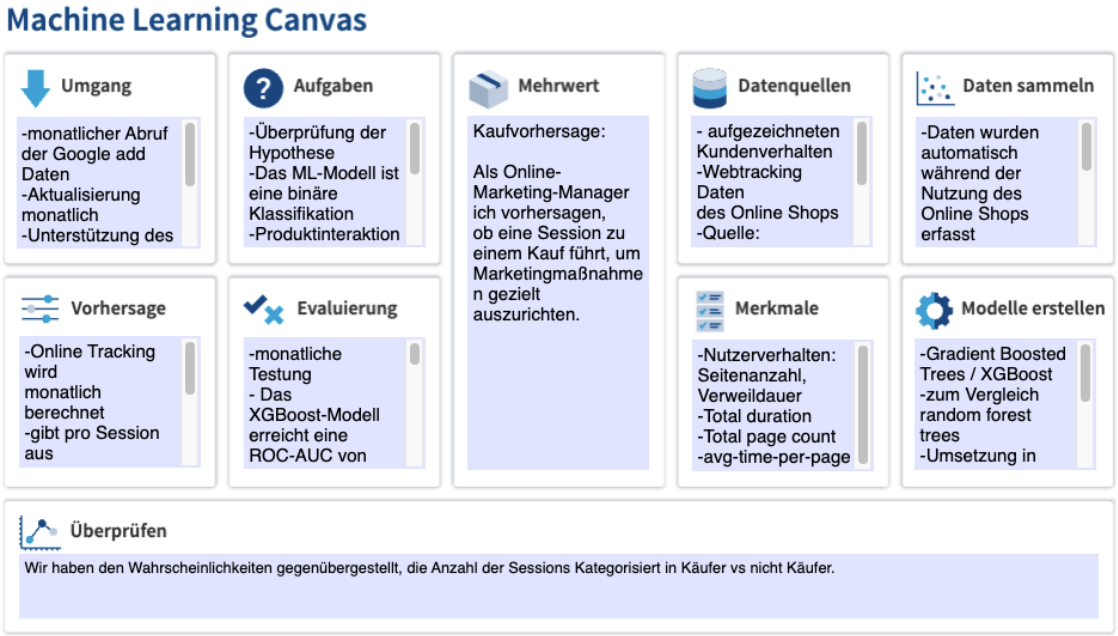



Abbildung 6.1: Machine Learning Canvas: Übersicht des ersten Sprints (Datenquellen, Aufgaben, Merkmale, Modell, Vorhersage, Evaluierung)

**Zeitliche Merkmale & Hypothese H1.** Hypothese H1 (Produktinteraktion beeinflusst Kaufabschluss positiv): Nutzer, die mehr auf produktbezogene Seiten interagieren oder mehr Zeit darauf verbringen, schließen Käufe ab. Variablen: Product-related und Revenue.

**Vorgehensweise:**

1. **Datenexploration:** Verteilung von Product-related und Revenue prüfen.
2. **Analyse:** Korrelation und ggf. logistische Regression, um den Einfluss der Produktinteraktion auf die Kaufwahrscheinlichkeit zu messen.
3. **Interpretation:** Je höher der Count an angesehenen Produktseiten, desto größer die Wahrscheinlichkeit für einen Kauf.

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.

**Modellierung.** Für die Modellierung wurde das XGBoost-Modell verwendet.

**Vorhersage.** Die Vorhersage wurde mithilfe des trainierten Modells durchgeführt (predict-Methode des Modells). Die Ausgabe kann in KNIME weiterverarbeitet und an Power BI übergeben werden (vgl. Kapitel 8: Datenübergabe von KNIME an Power BI).

**Modellbewertung.** Die Bewertung des Modells erfolgte u. a. anhand der ROC-Kurve (siehe Abb. 6.2). Das eingesetzte XGBoost-Modell erreicht eine AUC von 0,737 und damit eine solide Trennschärfe zwischen Kauf- und Nicht-Kauf-Sessions. Die Accuracy von 69 % ist aufgrund der unausgeglichenen Klassenverteilung nur eingeschränkt zur Bewertung der Modellqualität geeignet. Aussagekräftiger ist der Recall-Wert von etwa 62 %: Ein Großteil der tatsächlichen Kaufabschlüsse wird korrekt erkannt. Insgesamt liefert das Modell realistische und nachvollziehbare Ergebnisse, ohne Hinweise auf Data Leakage, und eignet sich als belastbare Grundlage für weiterführende Analysen.

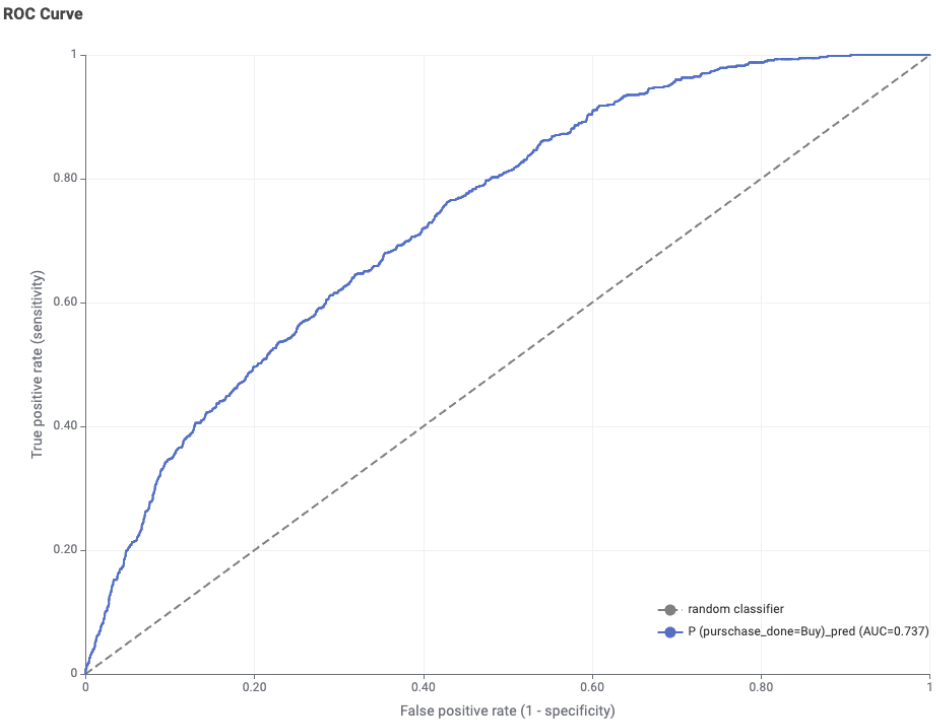



Abbildung 6.2: ROC-Kurve des XGBoost-Modells (AUC=0,737)

## 6.1 Modellvergleich: Baseline vs. XGBoost

Zur Einordnung der Modellleistung wurde eine einfache Baseline definiert, die standardmäßig keinen Kaufabschluss prognostiziert. Diese Baseline dient als Referenz zur Bewertung des Mehrwerts des Machine-Learning-Modells.

Im Vergleich zur Baseline zeigt das XGBoost-Modell eine deutlich bessere Trennfähigkeit, insbesondere gemessen an der ROC-AUC sowie dem Recall-Wert für Kaufabschlüsse. Dadurch wird deutlich, dass das ML-Modell über triviale Entscheidungsregeln hinausgeht und einen echten Mehrwert liefert.

# 7 Kurze Zusammenfassung der Ergebnisse

## 7.1 Beschreibung und Bewertung der Verifizierungen der Hypothesen

Hypothese H1 (Zusammenhang zwischen Produktinteraktion und Kaufabschluss) wurde in der EDA und in der Modellierung überprüft. Die berechneten Merkmale (`product_share`, `product_page_count`, `product_page_duration`) erwiesen sich als aussagekräftig; das XGBoost-Modell nutzt sie im finalen Feature-Set. Die Verifizierung stützt die Machbarkeit der Kaufvorhersage und den Einsatz im Dashboard.

## 7.2 Überprüfung und Erläuterung der Erfüllung der Definitions of done

Für die priorisierte User Story wurden im Vorfeld klare Definitions of Done festgelegt. Diese dienen dazu zu prüfen, ob die Ziele des Sprints erreicht wurden.

Die Definition of Done gilt als erfüllt, wenn:

- relevante Verhaltensmerkmale identifiziert wurden,
- ein nachvollziehbarer Zusammenhang zwischen Nutzerinteraktion und Kaufabschluss dargestellt werden kann,
- die Ergebnisse visuell aufbereitet und verständlich interpretiert wurden.

Diese Kriterien wurden im Rahmen des Projekts erfüllt.

Durch die Auswertung der Page Count und Page Duration konnte gezeigt werden, dass Käufer im Durchschnitt mehr Produktseiten aufrufen und sich länger auf diesen Seiten aufhalten als Nicht-Käufer. Die Visualisierungen machen diese Unterschiede sichtbar. Außerdem wurde überprüft, wie sich die Conversion Rate verändert, wenn Nutzer mehr Seiten aufrufen oder länger im Online-Shop bleiben. Dabei zeigt sich, dass Sitzungen mit hoher Interaktion eine deutlich höhere Kaufwahrscheinlichkeit aufweisen.

Die Definition of Done ist somit erreicht, da Erkenntnisse gewonnen wurden, die als Grundlage für weitere Analysen dienen können. Zusätzlich: Datenaufbereitung und Spaltenumbenennung dokumentiert (Tabelle im Bericht), EDA mit Ergebnisbeschreibung durchgeführt, Modell (XGBoost) trainiert und evaluiert, Vorhersage implementiert, Anbindung an Power BI und Dashboard-Themen geplant bzw. umgesetzt, Projektmanagement (Kanban, Gantt) dokumentiert.

## 7.3 Ausblick auf die nächsten Aufgaben bzw. Projektziele

Im bisherigen Projektverlauf konnte ein erster Mehrwert im Zusammenhang mit dem Kaufverhalten der Nutzer identifiziert werden. Insbesondere die Kennzahlen Page Count und Page Duration haben gezeigt, dass sie gut geeignet sind, das Interesse der Kunden zu bewerten.

In den nächsten Schritten soll geprüft werden, ob sich anhand dieser Kennzahlen zuverlässiger einschätzen lässt, ob ein Nutzer einen Kauf abschließt oder nicht. Ziel ist es, diese Erkenntnisse zu nutzen, um den Kaufprozess im Online-Shop weiter zu verbessern.

Weitere Sprints können die Verfeinerung des Modells, erweiterte KPIs im Dashboard, Integration in operative Geschäftsprozesse sowie Nutzen für Marketing und Shop-Optimierung vertiefen. Grenzen des Modells und Risiken (z. B. Datenqualität, Überanpassung) sind zu beobachten; das Zielbild der datengetriebenen Kaufvorhersage bleibt die langfristige Orientierung.

# 8 Dashboard & Workflow

## 8.1 Projektmanagement

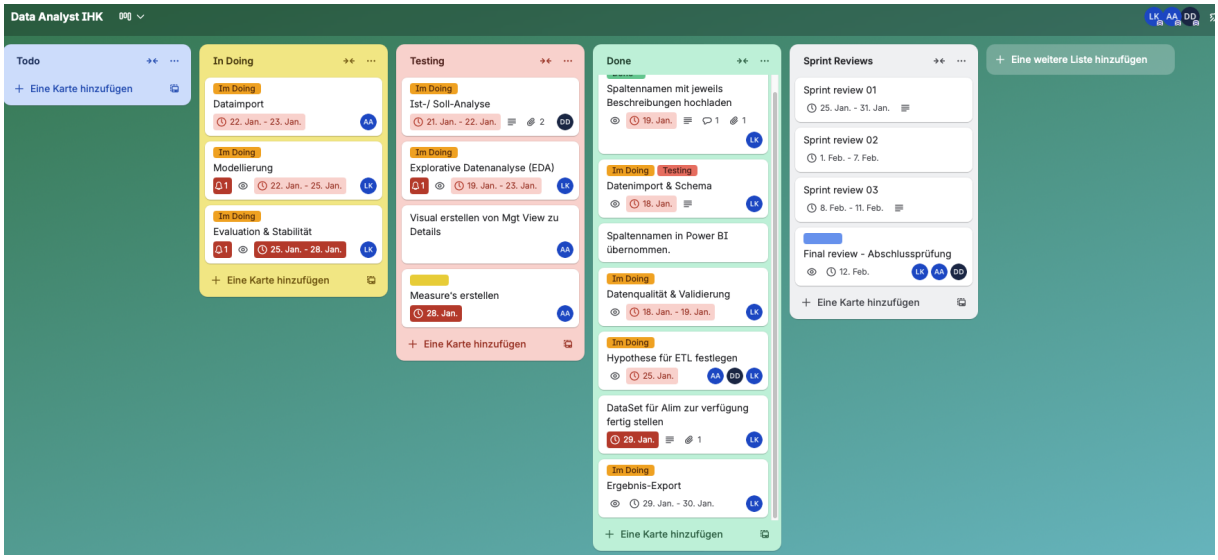



Abbildung 8.1: Kanban-Board mit Projektaufgaben und Sprint-Reviews

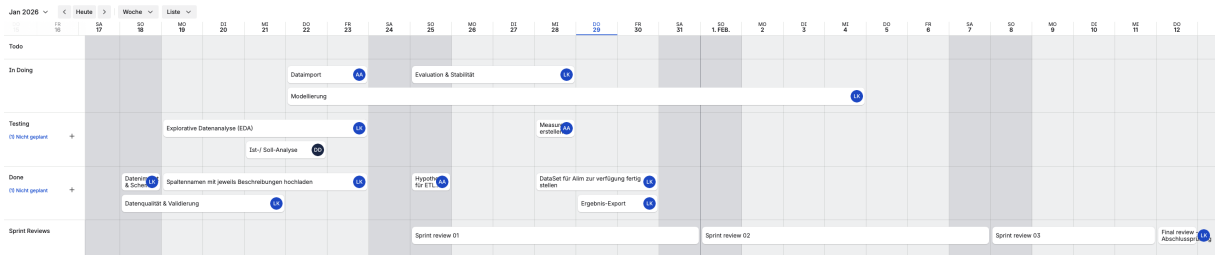



Abbildung 8.2: Gantt-Diagramm mit Projektzeitplan

## 8.2 Datenübergabe von KNIME an Power BI

Die aufbereiteten und modellierten Daten werden von KNIME exportiert und an Power BI übergeben (z. B. CSV-Export oder direkte Schnittstelle). So können die Business-KPIs und das Dashboard mit den gleichen Daten wie in der Analyse gefüllt werden. Abb. 8.3 zeigt einen Ausschnitt aus dem KNIME-Workflow bzw. der Ergebnisdarstellung.

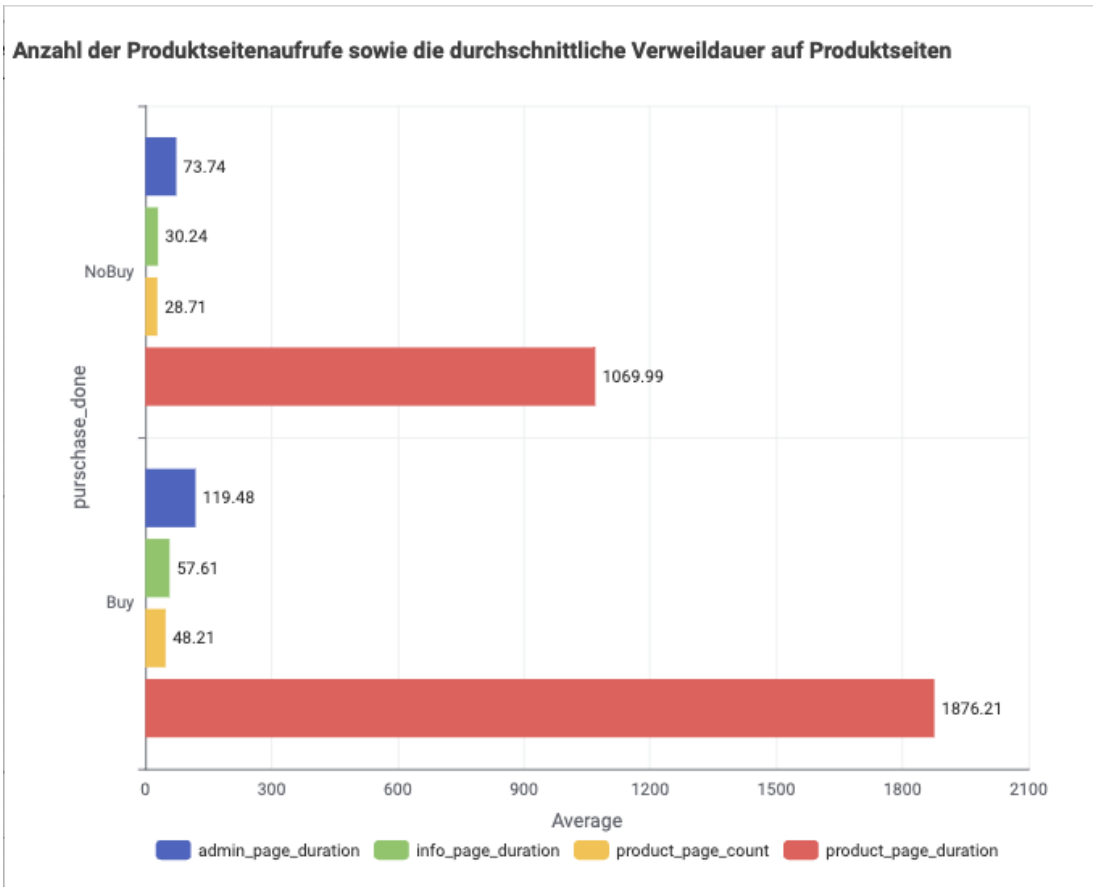



Abbildung 8.3: Ausschnitt aus KNIME-Workflow bzw. Ergebnisdarstellung (Daten aus CSV)

## 8.3 Definition der Business-KPIs

Die Business-KPIs wurden in Absprache mit den Fachabteilungen definiert und umfassen u. a. kaufrelevante Metriken (Conversion, Revenue), Nutzerverhalten (Produktseiten, Verweildauer) sowie Modelloutput (Vorhersage Kauf ja/nein). Sie werden im Dashboard abgebildet.

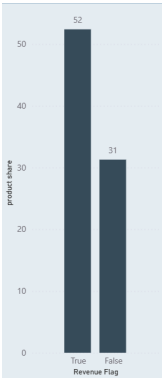
## 8.4 Dashboard-Design und Visualisierungen

Das Dashboard in Power BI nutzt die definierten KPIs und Visualisierungen (z. B. Verteilungen, Zeitverläufe, Segmentierungen nach Besuchertyp oder Monat), um die Ergebnisse der Analyse und der Vorhersage für Fachabteilungen zugänglich zu machen.

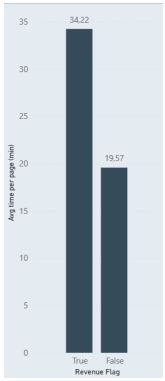
## 8.5 Auswertung im Dashboard: Käufer vs. Nicht-Käufer und Interaktionssegmente

Die beiden ersten Darstellungen (vgl. Abb. 8.4) vergleichen Käufer und Nicht-Käufer hinsichtlich der Seitenaufrufe und der Verweildauer. Es zeigt sich eine starke Korrelation: Käufer rufen im Mittel doppelt so viele Seiten auf wie Nicht-Käufer und verweilen länger auf der Website. Konkret liegen Käufer (Revenue = True) bei durchschnittlich 34,22 Minuten pro Seite, Nicht-Käufer (Revenue = False) bei 19,57 Minuten. Bei der Seitenanzahl (product share bzw. Anzahl aufgerufener Seiten) rufen Käufer im Mittel 52 Seiten auf, Nicht-Käufer 31 – also durchschnittlich 21 Seiten mehr bei Käufern. Die Nutzeraktivität ist damit stark mit dem Kaufverhalten verknüpft.

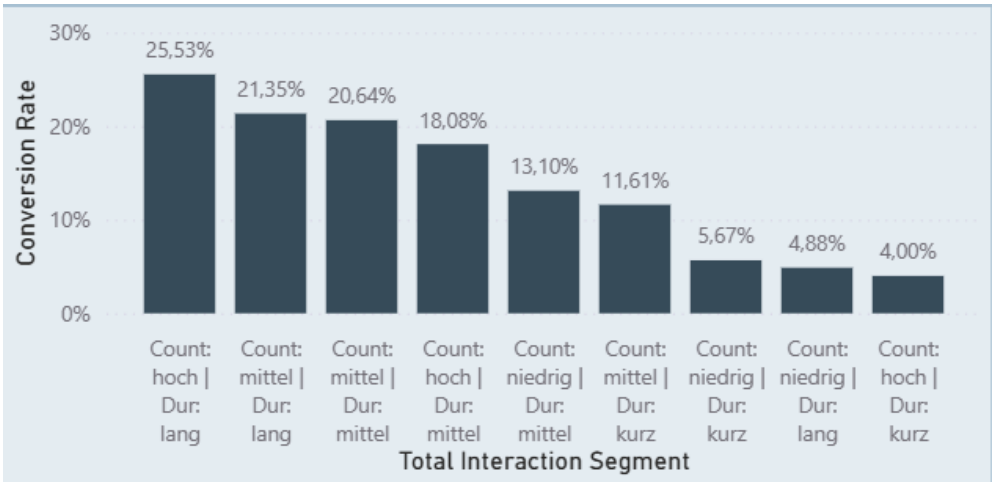
Die dritte Darstellung zeigt die Conversion Rate für unterschiedliche Interaktionssegmente. Ein Interaktionssegment setzt sich aus der Gesamtanzahl der pro Session aufgerufenen Seiten (Total Page Count) und der Gesamtdauer der Session (Total Duration) zusammen; beide Merkmale werden in die Kategorien hoch, mittel bzw. niedrig (Count) sowie lang, mittel bzw. kurz (Duration) unterteilt. Sehr aktive Nutzer weisen eine deutlich höhere Conversion Rate auf. Die höchste Rate weist das Segment *Count: hoch, Duration: lang* mit 25,53 % auf, gefolgt von *mittel/lang* (21,35 %) und *mittel/mittel* (20,64 %). Hohe Seitenanzahl bei mittlerer Dauer liegt bei 18,08 %, niedrige Count-Segmente bei kurzer oder mittlerer Dauer bei 11,61 % bzw. 13,10 %. Die niedrigsten Conversion Rates finden sich bei *hoch/kurz* (5,67 %), *mittel/kurz* (4,88 %) und *niedrig/lang* (4,00 %). Hohe Interaktion (viele Seiten, längere Verweildauer) geht mit deutlich höherer Conversion einher; kurze Sessions bei vielen Klicks bleiben vergleichsweise conversionschwach.




(a) Ø Seitenanzahl (Käufer vs. Nicht-Käufer)




(b) Ø Verweildauer pro Seite (Käufer vs. Nicht-Käufer)




(c) Conversion Rate nach Interaktionssegmenten

Abbildung 8.4: Dashboard-Auswertung: Käufer vs. Nicht-Käufer (Verweildauer, Seitenanzahl) und Conversion Rate nach Interaktionssegmenten.

**Sessions pro Prediction und bedingte Kaufquote.** Abbildung 8.5 zeigt die Kombination aus Balken- und Liniendiagramm zur Achse „Prediction (Container)“ (0 %, 20 %, 40 %, 60 %, 80 %). Die dunkelgrauen Balken (linke Achse „Sessions“) geben die Session-Anzahl pro Prediction-Bereich an: 0–10 %: 830, 10–20 %: 335, 20–30 %: 473, 30–40 %: 500, 50–60 %: 331, 60–70 %: 202, 70–80 %: 143, 80–90 %: 19 Sessions; die Anzahl ist in den niedrigen Prediction-Bereichen am höchsten und sinkt mit steigendem Prediction-Score. Die rosa Linie (rechte Achse „Relative Kaufquote“, 0–40 %) zeigt die bedingte Kaufquote: 1,45 % (0–10 %), 6,27 %, 12,31 %, 12,90 %, 18,40 %, 23,93 %, 27,79 %, 34,65 %, Maximum 39,86 % im Bereich 80–90 %, danach Abfall auf 31,58 %. Es besteht ein klarer Zusammenhang: Wo die Session-Zahl hoch ist, ist die relative Kaufquote niedrig; in den hohen Prediction-Bereichen (70–90 %) sind weniger Sessions, aber die Kaufquote ist deutlich

höher.

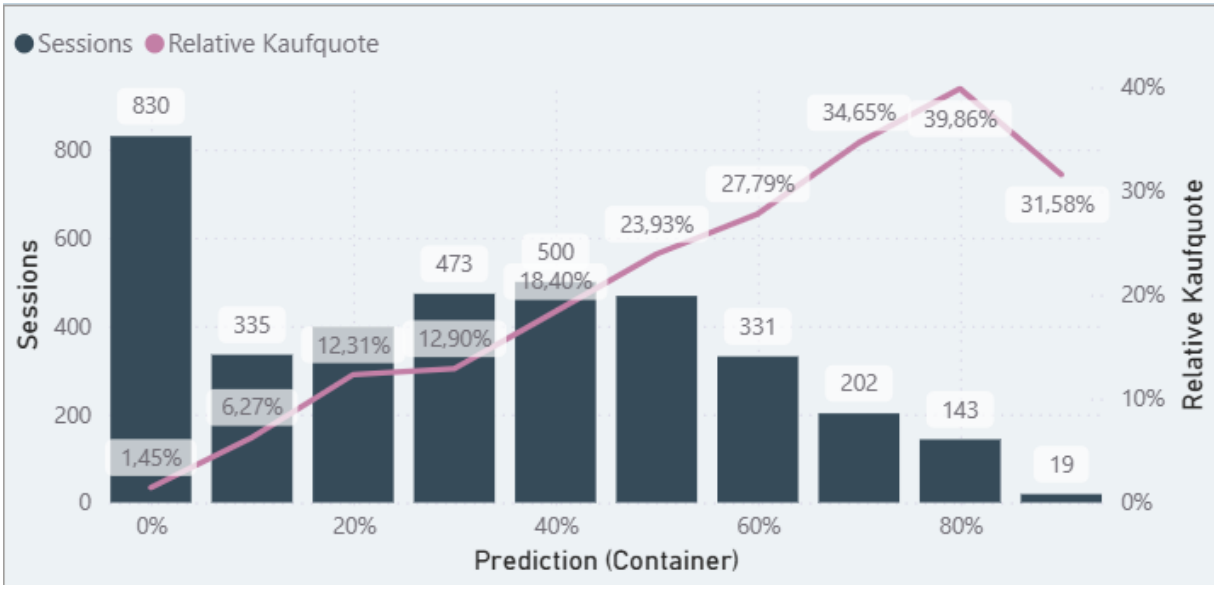



Abbildung 8.5: Sessions pro Prediction und bedingte Kaufquote.

**Segment-Definitionen im Dashboard.** Die Interaktionssegmente im Dashboard basieren auf zwei Unterteilungen. Tabelle 8.1 definiert das *Total Page Count*-Segment, Tabelle 8.2 das *Total Page Duration*-Segment.

Tabelle 8.1: Total Page Count Segment

| Kategorie | Total Page Count | Power BI Definition |
|---|---|---|
| Niedrig | 0–10 Seiten | `X <= 10` |
| Mittel | 11–40 Seiten | `X <= 40` |
| Hoch | ¿ 40 Seiten | ansonsten |

Das Total Page Duration Segment wurde ebenfalls in drei Gruppen eingeteilt:

Tabelle 8.2: Total Page Duration Segment

| Kategorie | Total Duration | Power BI Definition |
|---|---|---|
| Kurz | 0–10 Minuten | `X <= 10` |
| Mittel | 11–25 Minuten | `X <= 25` |
| Lang | ¿ 25 Minuten | ansonsten |

Das finale Segment ist eine textuelle Kombination aus dem Page-Count-Segment und dem Duration-Segment. Das Säulendiagramm stützt die Annahme, dass eine höhere Interaktionsintensität mit einer erhöhten Wahrscheinlichkeit einer Kaufentscheidung verbunden ist.

## 8.6 Interpretation der Ergebnisse für Fachabteilungen

Die Ergebnisse (EDA, Modell, Vorhersage) wurden so aufbereitet, dass Fachabteilungen sie interpretieren können: z. B. welche Faktoren den Kaufabschluss begünstigen, wie sich Nutzerverhalten unterscheidet und wie das Modell im Einsatz genutzt werden kann.

## 8.7 Management-Mehrwert durch datenbasierte Entscheidungen

Der Management-Mehrwert entsteht durch datenbasierte Entscheidungen: Nutzung der Kaufvorhersage und des Dashboards für Marketing- und Shop-Optimierung, Priorisierung von Maßnahmen und bessere Ressourcensteuerung.

## 8.8 Handlungsempfehlungen auf Basis hoher Produktinteraktion

Die Analyse zeigt, dass eine hohe Produktinteraktion (gemessen über Product Page Count und Product Page Duration) mit einer höheren Conversion Rate einhergeht. Produktinteraktion erweist sich damit als positiver Kaufindikator. Nachfolgend werden Handlungsempfehlungen für den Online-Shop formuliert, die diese Erkenntnis aufgreifen.

**1) Produktinteraktion gezielt fördern.** Der Online-Shop sollte Nutzer aktiv dazu anregen, sich intensiver mit Produkten auseinanderzusetzen. Konkrete Maßnahmen: ähnliche Produkte anzeigen (z. B. „Das könnte Sie auch interessieren“), Zubehör oder passende Ergänzungen vorschlagen, erweiterte Produktinformationen anbieten („Mehr Details anzeigen“) statt Informationen zu reduzieren. Begründung: Hohe Interaktion erhöht die Kaufwahrscheinlichkeit und ist Teil des Entscheidungsprozesses.

**2) Produktseiten als Entscheidungsraum gestalten.** Produktseiten sind nicht nur Informationsträger, sondern der Ort, an dem die Kaufentscheidung entsteht. Empfohlen werden: ausführliche Produktbeschreibungen, mehrere Bilder bzw. Anwendungsbeispiele,

klare Darstellung von Vorteilen und Nutzen sowie die sichtbare Integration von Bewertungen. Längere Verweildauer korreliert mit höherer Conversion; Information zahlt sich aus.

**3) Vergleichen erleichtern.** Da viele Produktseitenaufrufe mit hoher Conversion einhergehen, sollte das Vergleichen von Produkten erleichtert werden: z. B. Vergleichstabellen, Hervorhebung von Unterschieden zwischen Produkten, Kennzeichnungen wie „Unsere Empfehlung“ oder „Beliebteste Wahl“. Vergleich ist kein Abbruchsignal, sondern Teil einer aktiven Kaufentscheidung.

**4) Nutzer mit hoher Interaktion unterstützen.** Nutzer, die viele Produktseiten ansehen und viel Zeit investieren, sollten nicht gestört, sondern unterstützt werden. Konkret: auf aggressive Pop-ups verzichten, dezente Hilfsangebote anbieten (z. B. „Fragen zum Produkt?“), eine klare Navigation zurück zu bereits angesehenen Produkten bereitstellen. Diese Nutzergruppe weist die höchste Conversion Rate auf.

**5) Content-Qualität priorisieren.** Nicht ausschließlich auf „schnell zum Kauf“ optimieren, sondern auf qualitativ gute Entscheidungshilfen. Empfohlen werden: FAQ direkt auf der Produktseite, Hinweise zu Lieferung, Rückgabe und Garantie, kurze Entscheidungshilfen (z. B. „Für wen ist dieses Produkt geeignet?“). Die Daten zeigen: Kauf entsteht durch Auseinandersetzung mit dem Produkt, nicht durch Eile.

**6) Erfolgreiche Interaktionsmuster übertragen.** Seiten oder Produkte mit besonders hoher Interaktion und Conversion sollten gezielt analysiert und als Vorbild genutzt werden. Die Struktur, Inhalte und Darstellung dieser Seiten können auf andere Produkte übertragen und standardisiert werden. So wird sichtbar, was im Shop bereits funktioniert.

**Zusammenfassung:** Die Analyse zeigt, dass hohe Produktinteraktion (viele Produktseiten, längere Verweildauer) mit der höchsten Conversion Rate einhergeht. Ziel des Online-Shops sollte es sein, diese Interaktion gezielt zu fördern und optimal zu unterstützen. Drei zentrale Empfehlungen: Produktinteraktion fördern statt reduzieren; Produktseiten als Entscheidungshelfer ausbauen; Vergleich und Information bewusst ermöglichen.

## 8.9 Zielbild und Soll-Zustand (Ergänzung)

Zielbild: datengetriebene Kaufvorhersage im Einsatz. Das Modell soll in Geschäftsprozesse integriert werden (z. B. Personalisierung, Kampagnensteuerung). Nutzen für Marketing

und Shop-Optimierung sowie Grenzen des Modells und Risiken sind bei der weiteren Einführung zu berücksichtigen.

# Anhang

## Spaltenumbenennungen der CSV-Daten

Die Spalten des ursprünglichen Datensatzes wurden vor der weiteren Verarbeitung umbenannt; die Quelldaten aus dem Datensatz[1] haben uns bei den Spaltenumbenennungen als Orientierung gedient. Eine einheitliche Benennung nach dem Schema `snake_case`[2] erleichtert die Lesbarkeit, die konsistente Weiterverarbeitung und die Referenzierung im Bericht sowie im Dashboard. Die vollständige Übersicht aller Umbenennungen ist in Tabelle 8.3 dargestellt.

Tabelle 8.3: 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. |
| 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. |

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

**Tabelle 8.3 – Fortsetzung**

| Original CSV-Spalte | Neuer Spaltenname | Begründung |
|---|---|---|
| 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. |
| 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). |

**Tabelle 8.3 – Fortsetzung**

| Original CSV-Spalte | Neuer Spaltenname | Begründung |
|---|---|---|
| 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. |

# Literaturverzeichnis

[1] C. Sakar and Yomi Kastro. Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5F88Q.