Hausarbeit Data Science 1

Fallstudie Bloomé Beauty

Prof. Dr. Christian Bockermann

Sommersemester 2026

Abgabemodalitäten

- Ausarbeitung der Hausarbeit auf 6 Seiten (Limit!)
- Abgabe der Hausarbeit bis 23.8.2026 23:59 Uhr (via Moodle Aufgabe im Kurs)
- Gruppenarbeit möglich, Gruppengröße bis 3 Personen
- Individuelle mündliche Prüfung, Terminvereinbarung nach Absprache (ab 1.9.2026 möglich)
- **Abgabe im LNCS Format als PDF Datei!** Eine entsprechende Formatvorlage für Word/LaTeX finden sie unter

`https://datascience.hs-bochum.de/templates/`

1 Use-Case – Die Ausgangslage

Bloomé Beauty ist eine Direct-to-Consumer-Marke für Hautpflege und Make-up. Verkauft wird ausschließlich über den eigenen Webshop; es gibt keinen stationären Handel und keine Marktplatz-Präsenz. Die Marke positioniert sich im Bereich „clean beauty“ und führt zwei Produktlinien: eine vegane, tierversuchsfreie Kernlinie („Clean Vegan“) und eine konventionelle Zweitlinie („Classic“).

Die wirtschaftliche Lage: Der Website-Traffic ist solide und wächst leicht, aber die Conversion Rate (Anteil der Besuche, die zu einem Kauf führen) stagniert seit drei Quartalen. Gleichzeitig wird das Marketingbudget breit über viele Kanäle gestreut, ohne dass klar ist, welche Maßnahmen tatsächlich zum Umsatz beitragen. Die Geschäftsführung formuliert ein einfaches Ziel: mehr Umsatz aus dem vorhandenen Traffic herausholen, statt nur teuer neuen Traffic einzukaufen.

Im Unternehmen werden drei Stellhebel diskutiert, an denen man ansetzen könnte:

- **Marketing:** Welche Influencer-Kooperationen werden eingekauft?
- **Personalisierung:** Lässt sich die Website abhängig vom (mutmaßlichen) Profil der Besucher anpassen?
- **Ansprache:** Mit welcher Tonalität werden Kunden angesprochen (z. B. jugendlich, hip, konservativ)?

Die Geschäftsführung hat das Data-Science-Team (= Sie) gebeten, einen dieser Hebel datengetrieben zu untermauern.

2 Ihr Auftrag

Wählen Sie **einen** der unten beschriebenen Anwendungsfälle (Abschnitt 3). Für den gewählten Fall bearbeiten Sie:

(a) **Analyse des Geschäftskontexts:** Welche konkrete betriebswirtschaftliche Entscheidung adressieren Sie, und welcher Teil davon ist überhaupt durch ein Modell beeinflussbar?

(b) **Datenauswahl:** Aus dem bereitgestellten Datensatz-Pool (Abschnitt 5) wählen Sie die relevanten Datensätze aus – und begründen ausdrücklich auch, welche Sie verwerfen und warum.

(c) **Präzise Definition der Lernaufgabe:** Typ, Zielvariable, Klassen, in Frage kommende Features, überwacht/unüberwacht.

(d) **Herkunft der Labels:** Woher kommt die Zielvariable? Für wie viele Datenpunkte liegt sie vor?

(e) **Verwertung im Unternehmen:** Wie genau wird das Modellergebnis in eine Handlung übersetzt?

(f) **Risiken:** Datenqualität, Verzerrung (Bias)

Hinweis: Der Datensatz-Pool (siehe Anhang) enthält bewusst mehr Datensätze, als Sie für eine einzelne Aufgabe brauchen. Die Auswahl ist Teil der Leistung.

3 Die Anwendungsfälle (Klassifikation)

Wählen Sie einen der folgenden Fälle.

Fall 1 – Ernährungstyp aus dem Verhalten erkennen (Personalisierung)

Bloomé möchte die Startseite und Produktempfehlungen abhängig davon ausspielen, ob ein Besucher sich vegan, vegetarisch oder mischköstlich ernährt – auch wenn der Besucher das nie angegeben hat. Überlegen Sie, ob und wie sich der Ernährungstyp aus dem beobachtbaren Verhalten ableiten lässt, welche Signale dafür in Frage kommen und woher überhaupt verlässliche Labels für ein Training stammen könnten.

Fall 2 – Kaufabsicht in der laufenden Sitzung (Echtzeit-Intervention)

Bloomé will Besuchern, bei denen ein Kauf wahrscheinlich ist, aber noch wackelt, gezielt eine Maßnahme anbieten (z. B. Rabatt-Pop-up, Gratisversand, Beratungs-Chat). Überlegen Sie, ob sich noch während einer laufenden Sitzung vorhersagen lässt, ob sie zu einem Kauf führt – und welche Konsequenzen es hat, dass die allermeisten Sitzungen ohne Kauf enden.

Fall 3 – Wiederkäufer nach dem Erstkauf erkennen (Kundenbindung)

Bloomé möchte Erstkäufer, die voraussichtlich treu werden, gezielt mit Bindungsmaßnahmen ansprechen – und das knappe Bindungsbudget nicht mit der Gießkanne verteilen. Überlegen Sie, wie sich direkt nach einem Erstkauf einschätzen lässt, ob dieser Kunde innerhalb der nächsten Monate erneut kaufen wird, und welche Informationen Sie dafür verwenden dürfen.

Fall 4 – Lohnende Influencer-Kooperationen vorab einschätzen (Marketing)

Bloomé erhält laufend Kooperationsangebote von Influencern und muss entscheiden, welche angenommen werden. Überlegen Sie, wie sich anhand der Merkmale eines Influencers vorab einschätzen lässt, ob eine Kooperation sich lohnen wird, und wie die Grenze „lohnt sich / lohnt sich nicht“ überhaupt sinnvoll festgelegt wird.

4 Abgabe – Leitfragen

Die folgenden Leitfragen sind als Orientierungshilfe gedacht. Es geht nicht darum, nur eine Ausarbeitung mit der Beantwortung genau dieser Fragen zu schreiben.

1. Welche betriebswirtschaftliche Entscheidung adressieren Sie? Was daran ist durch ein Modell beeinflussbar, was nicht?
2. Welche Datensätze aus dem Pool wählen Sie aus? Welche verwerfen Sie – mit Begründung?
3. Benennen Sie die Lernaufgabe präzise: Welche Zielvariable, welche Klassen, welche Features, überwacht oder unüberwacht?
4. Woher kommen die Labels? Für wie viele Datenpunkte liegen sie vor? Welche Probleme ergeben sich daraus?
5. Wie gut lässt sich mit Ihrem Modell die Zielvariable vorhersagen? Wie evaluieren Sie ihr Modell?
6. Angenommen, Ihr Modell funktioniert: Wie genau wird das Ergebnis im Unternehmen in eine konkrete Handlung übersetzt?
7. Welche Risiken sehen Sie – bei Datenqualität, Verzerrung (Bias) und rechtlich/ethisch?

5 Datensatz-Pool

Folgende Datensätze stehen zur Verfügung. Beschrieben sind jeweils Inhalt und Spalten. (Schlüssel: PK = Primärschlüssel, FK = Fremdschlüssel zur Verknüpfung.)

Sie finden sämtliche Datensätze zum Download unter

`https://data.hsbo.de/bloome-s26/`

`customers.csv` – **Kundenstammdaten**	
`customer_id`	(PK) eindeutige Kundennummer
`signup_date`	Datum der Registrierung
`birth_year`	Geburtsjahr (teilweise fehlend)
`region`	Bundesland/Region
`acquisition_channel`	Wie der Kunde gewonnen wurde (organic, paid_social, influencer, newsletter)
`newsletter_opt_in`	hat Newsletter abonniert (true/false)

`diet_survey.csv` – **Ergebnisse einer freiwilligen Kundenbefragung**	
`customer_id`	(FK) Verknüpfung zu customers
`declared_diet`	selbst angegebener Ernährungstyp (vegan, vegetarian, omnivore)
`survey_date`	Datum der Befragung

Hinweis: Liegt nur für einen kleinen Teil der Kunden vor (Teilnehmer der Befragung).

`web_sessions.csv` – **Besuchssitzungen auf der Website**	
`session_id`	(PK) eindeutige Sitzungsnummer
`customer_id`	(FK) zugehöriger Kunde
`session_start`	Zeitstempel des Sitzungsbeginns
`device_type`	mobile, desktop, tablet
`entry_source`	Einstiegsquelle (organic, paid, influencer_link, direct)
`duration_sec`	Dauer der Sitzung in Sekunden
`pages_viewed`	Anzahl betrachteter Seiten
`converted`	Sitzung endete mit Kauf (true/false)

`content_interactions.csv` – **Interaktion mit Inhalten innerhalb einer Sitzung**	
`session_id`	(FK) zugehörige Sitzung
`content_category`	Inhaltskategorie (skincare, makeup, animals_nature, travel, fitness, food_recipes, fashion, memes, sustainability)
`dwell_time_sec`	Verweildauer auf diesem Inhalt
`scroll_depth_pct`	wie weit gescrollt wurde (0–100)
`liked`	Inhalt wurde geliked (true/false)

Hinweis: Pro Sitzung mehrere Zeilen (eine je betrachteter Kategorie).

`orders.csv` – Bestellungen

`order_id`	(PK) Bestellnummer
`customer_id`	(FK) Besteller
`session_id`	(FK) auslösende Sitzung
`order_date`	Bestelldatum
`revenue_eur`	Bestellwert in Euro
`discount_pct`	gewährter Rabatt in Prozent
`campaign_id`	(FK) ggf. zugeordnete Kampagne (kann leer sein)

`order_items.csv` – Positionen je Bestellung

`order_id`	(FK) zugehörige Bestellung
`product_id`	(FK) gekauftes Produkt
`quantity`	Menge
`unit_price`	Stückpreis

`products.csv` – Produktkatalog

`product_id`	(PK) Produktnummer
`product_name`	Bezeichnung
`line`	Produktlinie (clean_vegan, classic)
`is_vegan`	vegan (true/false)
`category`	Produktkategorie (skincare, makeup, . . .)
`price_eur`	Listenpreis
`target_age_group`	anvisierte Altersgruppe

`influencer_campaigns.csv` – eingekaufte Influencer-Kooperationen

`campaign_id`	(PK) Kampagnennummer
`influencer_handle`	Name/Handle
`platform`	instagram, tiktok, youtube
`followers`	Followerzahl
`engagement_rate`	durchschn. Interaktionsrate (0–1)
`niche`	thematischer Schwerpunkt (beauty, lifestyle, fitness, sustainability, comedy)
`cost_eur`	Kosten der Kooperation
`promoted_line`	beworbene Produktlinie (clean_vegan, classic)
`start_date`	Startdatum

`campaign_attribution.csv` – Wirkung der Kampagnen

`campaign_id`	(FK) zugehörige Kampagne
`sessions_generated`	zugeführte Sitzungen
`conversions`	daraus resultierende Käufe
`attributed_revenue_eur`	zugeordneter Umsatz

`ab_test_tone.csv` – Ergebnisse eines A/B-Tests zur Ansprache

`customer_id`	(FK) getesteter Kunde

`assigned_tone`	ausgespielte Tonalität (hip, youthful, conservative)
`converted`	hat in der Testphase gekauft (true/false)
`test_period`	Zeitraum des Tests

`supplier_inventory.csv` – Lager- und Lieferantendaten

`product_id`	(FK) Produkt
`supplier`	Lieferant
`stock_units`	Lagerbestand
`lead_time_days`	Wiederbeschaffungszeit in Tagen
`restock_date`	nächster Wareneingang

`website_performance.csv` – technische Kennzahlen der Website

`date`	Datum
`avg_load_time_ms`	durchschnittliche Ladezeit in Millisekunden
`uptime_pct`	Verfügbarkeit in Prozent
`server_errors`	Anzahl Serverfehler