DSCB420 – Datenanalyse und Business Intelligence 2

Hochschule Karlsruhe
University of
Applied Sciences

Aufgabenstellung Projektarbeit

Sommersemester 2025

Inhalt

1 Idee der Projektarbeit (wichtig, bitte lesen!)

In diesem Modul sollen Ihre Kenntnisse und Kompetenzen aus den verschiedenen Modulen des Studiengangs „Data Science“ **integriert und ausgebaut** werden: Sie werden die Kompetenz aufbauen, verschiedenen analytische Komponenten (bspw. ETL-Prozesse, OLAP-Modellierung, explorative Analysen mit pandas, Datentransformationen, Feature Engineering, Machine Learning etc.) in einer Gesamtarchitektur aufzubauen.

Dafür sind die Hintergründe entsprechender Systeme zu durchdringen und deren Anwendung zu üben („Data Engineering“ und Veranstaltungen im Teil „Küppers“). Weiterhin sollen Konzepte insbesondere im Kontext der Datentransformation und Datenvorverarbeitung erlernt werden, um Datenströme und Analyseprozesse sinnvoll entsprechend der Anforderungen in verschiedenen Praxis-Anwendungsfällen anwenden zu können.

In der Projektarbeit haben Sie neben der Bearbeitung von **vorgegebenen Fragestellungen (im Folgenden Teil 1)** die Aufgabe, aus den gegebenen Datenquellen weitere **interessante Anwendungsfälle (im Folgenden Teil 2)** zu erarbeiten. Es wird ein Datensatz vorgegeben, den Sie ggf. anreichern können und Ihren Interessen entsprechend auswerten sollen. Sie sind in Teil 2 also sehr frei in der Gestaltung des Anwendungsfalls und dürfen auch Daten (teilweise oder falls nötig komplett) selbst generieren. Insgesamt sollte die Gesamtkomplexität in dem „freien Teil“ jedoch angemessen gehalten werden.

Die im Rahmen des Projekts prototypisch aufzusetzenden Systeme und Methoden sind als eine Vorstufe zum operativen Betrieb zu sehen, d.h. Sie müssen nicht jedes einzelne Detail „durchautomatisieren“, aber die verwendeten Tools müssen dazu grundsätzlich in der Lage sein. Die Umsetzung wird **lokal** erfolgen. Es sind dabei insbesondere die anzuwendenden analytischen Methoden und deren Integration in einen sinnvollen Datenfluss von der Datenquelle bis zur Senke (bspw. Dashboard, DWH-Tabellen, Prozesse, einzelne Stakeholder usw.) festzuhalten. Die „Konsumenten“ (Stakeholder und Prozesse) sowie deren Anforderungen an eine Versorgung mit Daten sind festzuhalten.

Die Architektur ist **prototypisch in Python mit dem Framework „prefect“ umzusetzen**. Dies wird das Spezifizieren von ETL-Pipelines, Erzeugen von Datenbanken, sowie Implementierung komplexerer Analyse- und Transformationsschritte in diesem Framework bedingen.

Abschließend sollen Sie Ihre Ergebnisse „reflektieren“ und die eingesetzten Methoden sowie die Architektur **evaluieren**. Insbesondere folgende Fragen sind zu beantworten: Welche Herausforderungen sind aufgetreten? Wie konnten diese gelöst werden oder waren alternative Lösungsansätze nötig?

2 Gegebener Datensatz

Folgende gegebene Daten eines Online-Lieferdienstes für Lebensmittel sind aufzunehmen:

(A) orders:
Enthält eine Zeile pro Bestellung mit folgenden Spalten:

user_id:	Eindeutige ID für jeden Kunden
order_id:	Eindeutige ID für jede Bestellung
order_date:	Zeitstempel der Bestellung

(B) **order_products_denormalized:**
Enthält eine Zeile pro Bestellposition (=Artikel), d.h. pro Bestellung existieren so viele Zeilen wie Artikel in der Bestellung enthalten waren.
Enthaltene Spalten:

order_id:	
product_id:	
product_name:	
add_to_cart_order:	Als wievieltes Produkt wurde dieses Produkt bei der Bestellung in den Warenkorb gelegt?
aisle_id, aisle_name:	Beschreibt den Gang, in dem das Produkt im Supermarkt platziert ist (1:N-Beziehung zu product_id)
department_id, department_name:	Produktgruppe, der das Produkt zugerechnet wird. (1:N-Beziehung zu product_id)

(C) tips_public:
Enthält eine Zeile pro Bestellung mit folgenden Spalten:

order_id:	
tip:	Wurde dem Lieferboten Trinkgeld gegeben? (yes/no)

Die letzte Bestellung jedes users fehlt. Für diese ist Ihnen nicht bekannt, ob Trinkgeld gegeben wurde. Sie dienen als Testdatensatz.

3 Aufgabenstellung

Bitte setzen Sie folgende Aufgaben im Rahmen der Projektarbeit um. Die Aufgaben lassen teilweise Interpretationsspielraum und erfordern eine sinnvolle Auslegung. Sprechen Sie in diesem Zusammenhang auftretende Fragen in der Vorlesung an, dann wird dort geklärt, welchen Ermessens-Spielraum Sie haben und welche Untersuchungen mindestens erwartet werden. **Diese Absprachen in der Vorlesung sind auch Teil der Aufgabenstellung – es ist Ihre Verantwortung, diese auch mitzubekommen und zu berücksichtigen.**

(1) **Abruf, Aufbereitung und Kombination** der genannten Datenquellen
Transformieren Sie die Daten nach eigenem Ermessen, wie es Ihnen für die nachfolgenden Aufgaben zweckmäßig erscheint

(2) Beantwortung **vorgegebener analytischer Fragen (Teil 1)**
`Beschränken Sie sich für die folgenden Fragen zunächst auf die unter 2. (A) – (C) spezifizierten Daten. Falls Sie Erweiterungen für dringend geboten halten, sprechen Sie sie bitte mit den Aufgabenstellern ab.`
Die Trinkgeldgabe oder nicht-Gabe bei den Bestellungen jedes Users kann als univariate Zeitreihe betrachtet werden. Untersuchen Sie folgende Fragen und machen Sie das Ergebnis jeder Teilaufgabe auch durch geeignete Visualisierungen oder Kennzahlen nachvollziehbar und quantitativ greifbar:
   a. Mit welcher Genauigkeit lässt sich die Trinkgeldgabe bei einer Bestellung aus der Information vorhersagen, ob der User bei der vorhergehenden Bestellung Trinkgeld gegeben hat? (AR(1)-Modell)
   b. Mit welcher Genauigkeit lässt sich die Trinkgeldgabe vorhersagen, wenn man zusätzlich zu (a) die Information als Input benutzt, ob bei der vor-vorhergehenden Bestellung eines Users Trinkgeld gegeben wurde? (AR(2)-Modell)
   c. Untersuchen Sie die Autokorrelationen und partiellen Autokorrelationen der Zeitreihe und schätzen Sie daraus ab, welchen Wert für *n* man verwenden sollte, wenn man ein AR(n)-Modell für diese Zeitreihe benutzt. Welche Genauigkeit erreicht dieses optimale AR(n)-Modell?
   d. Hat die Zeitreihe Periodizität(en)?
   e. Gibt es einen Trend?
   f. Erweitern Sie Ihr Prognosemodell aus (c) passend zu Ihren Erkenntnissen aus (d) und (e) um eine geeignete Behandlung von Perioditizät und Trend, so weit diese vorliegen.
   g. Erstellen Sie mit Hilfe des Prognosemodells aus (f) für die order_ids aus der Datei tip_testdaten_template.csv jeweils eine Vorhersage, ob Trinkgeld gegeben wurde, und geben Sie diese als tip_prediction2.csv ab.

(3) Definition **weiterer offener Aufgaben (Teil 2)**
`Für diesen Aufgabenteil steht Ihnen frei, beliebige weitere Datenquellen zuzuziehen.`
   a. Identifizieren Sie weitere Einflussgrößen auf die Trinkgeldwahrscheinlichkeit einer Bestellung und verbessern Sie damit Ihre Prognosen aus 2. (Feature Engineering) Liefern Sie sowohl Prognosen für die Bestellungen aus tip_testdaten_template (Abgabe als Datei tip_prediction3), als auch Erkenntnisgewinn, indem Sie interpretieren, welchen weiteren Einflussgrößen Sie erkannt haben und in welcher Weise sie die Trinkgeldwahrscheinlichkeit beeinflussen.

(4) Spezifikation **eines integrierten Anwendungsfalls** (Kombination von Teilen 1 und 2)
   a. Definition (fiktiver) **Geschäftsprozesse oder eines Geschäftsmodells**. Der Prozess oder das Geschäftsmodell sollen auf Basis der genannten Daten, Fragestellungen und analytischen Prozesse gestaltet werden (von „einfachen“ Reports bis hin zu optimierten/automatisierten Prozessen ist alles denkbar)
   b. Festlegung von **Stakeholdern und Konsumenten der Daten** (Festlegung verschiedener Anspruchsgruppen, bspw. Kapazitätsplanung eines Supermarktbetreibers, Marketing-Abteilung, Einkäufer mit Reporting-Anforderungen bzw. Self-Service BI-Nutzer, automatisierbare Prozesse, etc.)
   c. Spezifikation der **analytischen Anforderungen** (notwendige analytische Berechnungen, Berechnungen, Modelle, Transformationsschritte etc.) – **WICHTIG**:

Aus Teil 1 dürfen die gegebenen Fragestellungen (teilweise) verwendet und mit den eigenen Ideen aus Teil 2 integriert werden.

d. Zusammenführung der genannten Aspekte in einer **Modellierung des Datenflusses**
e. **WICHTIG**: Aus den behandelten Bereichen im Teil „Küppers“ sind in dem Datenfluss und der Architektur alle Aspekte aus den Vorlesungen zu „Data Warehousing / OLAP“ (durch ein entsprechendes Datenmodell und Ladeprozesse) abzudecken und in prefect umzusetzen. Sie sollten hier zusätzlich Algorithmik aus dem Teil „Hofmann“ in Ihre prefect-Pipeline(s) einfügen, sofern dies für die von Ihnen definierten Prozesse sinnvoll erscheint.
f. **WICHTIG**: Halten Sie alle Punkte einfach, um den Modulrahmen nicht zu sprengen!

(5) **Umsetzung der Architektur**

a. Ggf. Entwicklung von Jupyter-Notebooks / lokalem Python- / SQL-Code zur Umsetzung der methodischen / analytischen Anforderungen
b. Festlegung der zu verwenden Architekturkomponenten (insb. Datenbanken bspw. sqlite, ETL-Tools, Analytics-Tools, etc.)
c. Deployment des Codes in prefect (workflow und task Definitionen) und dbt (SQL-zu-SQL).
d. Ggf. Entwicklung von Simulatoren für die Datenquelle(n) (bspw. Verkaufsevents)

4 Organisation und einzureichende Dokumente

4.1 Ablauf

Die Übung ist durch Projektgruppen bestehend aus **4 Studierenden** (im Ausnahmefall auch 3 oder 5, bitte vorher abklären) umzusetzen.

- **Gruppencalls mit Betreuern** (pro Gruppe 15min, Zeitslots werden noch bekannt gegeben):
  Kalenderwoche 21 (19.5.-23.5.) „Meilenstein 1st Draft“
  Erwartet werden fertige Ergebnisse zu (1), Teilergebnisse zu (2) und ersten Ideen zu (3)-(5)
  Kalenderwoche 25 (16.6.-20.6.) „Statuscall Anwendungsfall und weitere Ideen“
  Finalisierung weiterer Datenquellen, konkreter Datenfluss, analytischer Aufgaben und eingesetzten Methoden, sowie des prefect-Architekturentwurfs (d.h. (3)-(5))

  WICHTIG: Der in diesen Calls gezeigte Fortschritt der Projektbearbeitung geht mit in die Benotung ein.
- **Individueller Test**: **Dienstag, 24.06.2025**, ab 11:30 Uhr in E102 (60 Minuten)
- **Abgabe Projektbericht und „Deliverables“: Montag, 30.06.2025, 20 Uhr**
  Über ILIAS muss der Projektbericht von der Gruppe als PDF-Datei bis zur Deadline eingereicht werden (Abschnitt „Übungsaufgaben“ □ „Baustein Übung“ – dort können Sie die Datei mit dem Projektbericht je Gruppe hochladen). Darüber hinaus müssen Sie in einer zip-Datei sämtlichen Code einreichen. Bitte fügen Sie im Anhang des Projektberichts eine entsprechende Ordner-Übersicht mit Erklärung der einzelnen Dateien ein. Cloud-Pipelines sollten Sie mit Screenshots dokumentieren und im Anhang abbilden.
- **Präsentation: Dienstag, 01.07.2025, ab 09:50 Uhr** in 2 Vorlesungsblöcken (pro Gruppe 15 Minuten Präsentation, ca. 5 Minuten Diskussion). Achtung: die Präsentationen sind an dem Tag bis 09:50 Uhr in ILIAS als PDF einzureichen.

4.2 Einzureichende Dokumente und Prüfungsleistung

4.2.1 Projektbericht

- Erstellen Sie eine Titelseite Bezeichnung „DSCB420 – Projektbericht SoSe2025“ sowie die Matrikelnummern der Gruppenteilnehmer (keine Namen!)
- **12-14 Seiten (A4) inkl. Abbildungen**, die Titelseite zählt nicht dazu, verlagern Sie Details in den Anhang, beschränken Sie sich auf Kernaussagen und achten Sie auf Systematik
  - Schriftart Arial, Schriftgrad 11, Zeilenabstand „mehrfach 1,2“
  - Seitenränder 2,5cm
  - Abstand zwischen Absätzen 3 Punkte (Einstellung unter „Absatz“ in Word)
  - Abstand nach Überschriften 6 Punkte (es sollten Kapitel / Überschriften zur Strukturierung verwendet werden)
- Erstellen Sie einen Executive Summary (1/2 Seite, zählt nicht zu den 12-14 Seiten).
- Erstellen Sie ein Inhalts- und Abbildungs- sowie Tabellenverzeichnis (zählt nicht zu den 12-14 Seiten).
- Achten Sie auf Systematik im Argumentationsaufbau, präzise Formulierungen, ein durchgängiges Abbildungsdesign mit hoher Abbildungsqualität, sorgfältige Prüfungen der formellen Aspekte, eine zielgruppengerechte Aufbereitung der Inhalte, etc.
- Reichen Sie den Projektbericht als PDF-Datei ein.

4.2.2 Präsentation

Alle Gruppen müssen die Ergebnisse präsentieren und bei sämtlichen Präsentationen besteht Anwesenheitspflicht. Es müssen alle Teilnehmer der jeweiligen Gruppe präsentieren und Fragen beantworten können.

Die finale Präsentation ist eine Art „Pitch“: Nehmen Sie an, Sie müssen Ihren selbst definierten „fiktiven Auftraggebern“ aus dem Anwendungsfall Ihre Ergebnisse vermitteln.

Formelle Anforderungen an die Präsentation

- Zeitlicher Rahmen der Präsentation: 15 Minuten Präsentation, ca. 5 Minuten Diskussion
    - Genaues Timing ist ein Bewertungskriterium!

Wichtige Hinweise zu den Bewertungskriterien: Achten Sie insbesondere auf

- eine saubere Struktur und Systematik in der Präsentation (unterstützt bspw. durch eine Agenda),
- übersichtliche und leicht nachvollziehbare Folien (es gibt keine Formatvorlage, hier können Sie selbst kreativ sein),
- eine zielgruppengerechte Aufbereitung der Inhalte (Zielgruppe: „fiktive Auftraggeber aus Ihrem Anwendungsfall sowie den Pflichtfragen“),
- eine Darstellungsform, die auf eine Präsentation zugeschnitten ist (nicht 1:1 Kopie aus dem Projektbericht!) und klares Foliendesign,
- lesbare Abbildungen, Tabellen, etc.,
- einen flüssigen Vortragsstil ohne ablesen (üben!), sowie
- Nutzung „interaktiver“ Inhalte – sofern möglich und angebracht. Beispiele:
    - Nutzung von Jupyter-Notebooks, um die Elemente der „methodischen Komponenten“ vorzustellen
    - Darstellung der ETL-Pipelines (ggf. als Screencast mit „Zeitraffer“) in prefect
    - Animationen können bei der systematischen Herleitung komplexer Inhalte helfen (aber nicht übertreiben!)
- Live-Vorstellung von Artefakten und ggf. „Dashboards“ für die Stakeholder, die das „Ende Ihrer Pipeline“ darstellen (nicht zu viel Aufwand in Dashboards stecken!)

4.2.3 Code und weitere Artefakte

Reichen Sie zusätzlich – abhängig von Ihrem Anwendungsfall und der Architektur – die erstellten Artefakte ein (vgl. „Ablauf“). Diese können sein:

- Prefect ETL Pipeline – dokumentiert über Code bzw. Screenshots im Anhang
- pandas Code (Prototyping)
- Sämtlicher Code zu angewandten Methoden, Simulationen etc.

4.2.4 Sonstige Hinweise

Zur Individualisierung der Leistung findet ein Test statt (siehe oben).

Es gibt keine weitere Prüfungsleistung / Klausur.