

Projektaufgabe: Introduction to Data Science

In Gruppen von 3-4 Studierenden solltet Ihr einen vollständigen Data Science Workflow durchführen - von der Datenexploration bis zur kritischen Bewertung der Ergebnisse. Jede Gruppe wählt einen Datensatz und bearbeitet die untenstehenden Aufgaben.

Überlegt euch eine sinnvolle Aufteilung innerhalb des Teams. Neben dem finalen Notebook formuliert bitte jedes Team-Mitglied ein bis maximal zwei Seiten, was die eigene Rolle innerhalb des Teams war, was die Hauptaufgaben waren und was die größten Herausforderungen und Learnings waren. Dies kann ggf. als Grundlage für individuelle Noten dienen.

Der Bewertungsbogen unten dient als Orientierungshilfe bei der Abgabe und deren Bewertung. Als optional gekennzeichnete und besprochene Aspekte, sind optional und können durch andere Aspekte ausgeglichen werden. Verwendet für die Abgabe gerne KI-Tools in sinnvollem Umfang, achtet aber auf eigenes Verständnis und kritisches Hinterfragen – lasst lieber eine Aufgabe als unbeantwortet, als nur mit KI beantwortet aber ohne Verständnis. Verwendet hauptsächlich die in der Vorlesung behandelten Tools und Frameworks.

Beispiel-Datensätze (wählt pro Gruppe einen aus oder findet einen eigenen)

1. [Heart Disease UCI](#) - Medizinische Daten zur Vorhersage von Herzerkrankungen
2. [Bank Marketing](#) - Kundendaten einer Bank für Marketingkampagnen
3. [Air Quality](#) - Luftqualitätsmessungen mit verschiedenen Sensoren
4. [Online Retail](#) - Transaktionsdaten eines Online-Händlers
5. Oder einen eigenen Datensatz aus den vorgestellten Quellen

Teil 1: Business Understanding & Datenexploration

1. **Business Context**
 - Definiert ein konkretes (Business-)Problem für euren Datensatz
 - Identifiziert 2-3 konkrete (Geschäfts-)Fragen, die Ihr beantworten möchtet
 - Diskutiert kurz, warum Data Science für dieses Problem geeignet ist
2. **Explorative Datenanalyse**
 - Erstellt mindestens 5 verschiedene Visualisierungen (unterschiedliche Typen!)
 - Identifiziert Muster, Anomalien und interessante Zusammenhänge
 - Diskutiert potentielle Verzerrungen (Bias) in den Daten
3. **Statistische Analys**
 - Berechnet ein paar relevante Statistiken
 - Führt Korrelationsanalysen durch
 -

Teil 2: Datenaufbereitung & Feature Engineering

1. **Datenqualität**
 - Analysiert und behandelt fehlende Werte
 - Identifiziert und behandelt Ausreißer
2. **Feature Engineering (10 Punkte)**
 - Erstellt wenn notwendig neue Features basierend auf Domänenwissen
 - Wendet wenn notwendig Normalisierung und/oder auch One-Hot Encoding an
3. **Dokumentation (5 Punkte)**
 - Erstellt ein Data Dictionary für alle Features

- Dokumentiert die Transformationsschritte möglichst nachvollziehbar
-

Teil 3: Modellierung & Evaluation

1. Modellauswahl

- Implementiert wenn möglich ein Modell
- Begründet die Auswahl basierend auf den Datencharakteristika

2. Evaluation

- Verwendet angemessene Metriken zur Evaluierung dieses Modells (ggf. beispielhaft)
- Diskutiert Overfitting/Underfitting anhand eurer Ergebnisse (ggf. beispielhaft)

Teil 4: Kritische Reflexion & Geschäftsempfehlungen

1. Geschäftsempfehlungen

- Übersetzt eure technischen Ergebnisse in Business-Sprache
- Gebt konkrete, umsetzbare Empfehlungen
- Diskutiert Grenzen eurer Analyse

2. Weiterführende Fragen

- Formuliert kritische Fragen, die eure Analyse aufwirft
- Schlägt nächste Schritte für eine erweiterte Analyse vor

Abgabeformat

1. Jupyter Notebook

- Strukturiertes, gut dokumentiertes Notebook
- Code und Markdown-Zellen
- Reproduzierbar (requirements.txt beifügen)

2. Präsentation

- 15 Minuten Präsentation + 5 Minuten Diskussion
- Max. 12 Folien
- Fokus auf Business-Insights, nicht technische Details
- Jedes Gruppenmitglied präsentiert

3. Eigner Kurzbericht

- Beschreibung der eigenen Rolle und Aufgabenbereich(e)
- Reflektion der größten Herausforderungen
- Analyse der eigenen Learnings aus dem Projekt und der Vorlesung allgemein

Abgabetermine

- **Notebook & Repository:** Am Tag vor der Präsentation um 20.00 Uhr.
- **Präsentationen und Kurzbericht:** Wird gemeinsam in der ersten Vorlesung definiert.

Viel Spaß und Erfolg bei dem Projekt!

Bewertungsbogen: Introduction to Data Science

Gruppennummer:
Gruppenmitglieder:
Gewählter Datensatz:

Bewertungsaspekt	Erreichte Punkte	Kommentar	Hinweise zur Bewertung:
Konkretes Problem definiert und Business-Relevanz erklärt	/5		<ul style="list-style-type: none"> Ist das Problem klar und nachvollziehbar formuliert? Sind die Geschäftsfragen konkret und sinnvoll? Wird die Eignung von Data Science begründet?
Qualität und Vielfalt der Visualisierungen	/7		<ul style="list-style-type: none"> Werden verschiedene Visualisierungstypen sinnvoll eingesetzt? Sind die Visualisierungen aussagekräftig und gut beschriftet? Werden interessante Erkenntnisse abgeleitet und kritisch hinterfragt?
Identifikation von Mustern, Anomalien und Zusammenhängen	/5		
Diskussion von Bias und Datenverzerrungen	/3		
Relevante statistische Kennzahlen berechnet und interpretiert	/7		<ul style="list-style-type: none"> Sind die gewählten Statistiken für den Datensatz sinnvoll? Werden Korrelationen korrekt interpretiert? Wird zwischen Korrelation und Kausalität unterschieden?
Korrelationsanalysen durchgeführt und diskutiert	/5		
Umgang mit fehlenden Werten	/5		<ul style="list-style-type: none"> Werden fehlende Werte systematisch analysiert? Sind die gewählten Strategien zur Behandlung begründet? Wird mit Ausreißern angemessen umgegangen?
Identifikation und Behandlung von Ausreißern	/5		

Erstellung und Begründung neuer Features (optional)	/5		<ul style="list-style-type: none"> • Werden Features basierend auf Domänenwissen erstellt? • Sind Transformationen korrekt angewendet und begründet? • Wird die Notwendigkeit der Transformationen erklärt?
Anwendung von Transformationen (Normalisierung, Encoding, etc.)	/3		
Data Dictionary und nachvollziehbare Dokumentation	/5		<ul style="list-style-type: none"> • Sind alle Features dokumentiert? • Sind die Transformationsschritte nachvollziehbar? • Ist der Code gut kommentiert?
Implementierung und Begründung des Modells (optional)	/10		<ul style="list-style-type: none"> • Ist die Modellwahl für den Datensatz und die Fragestellung geeignet? • Wird die Auswahl nachvollziehbar begründet? • Ist die Implementierung korrekt?
Angemessene Metriken zur Evaluierung	/5		<ul style="list-style-type: none"> • Sind die gewählten Metriken für das Problem geeignet? • Werden die Ergebnisse kritisch interpretiert? • Wird auf mögliche Probleme (Overfitting/Underfitting) eingegangen?
Diskussion von Overfitting/Underfitting	/5		
Übersetzung technischer Ergebnisse in Business-Sprache	/4		<ul style="list-style-type: none"> • Sind die Empfehlungen für Nicht-Techniker verständlich? • Sind die Empfehlungen praktisch umsetzbar? • Werden Limitationen ehrlich adressiert?
Konkrete, umsetzbare Empfehlungen	/4		
Diskussion der Grenzen der Analyse	/2		
Formulierung kritischer Fragen	/5		<ul style="list-style-type: none"> • Werden durchdachte, kritische Fragen gestellt? • Zeigen die Fragen tiefes Verständnis des Problems?
Vorschlag sinnvoller nächster Schritte	/5		

			<ul style="list-style-type: none"> Sind die vorgeschlagenen nächsten Schritte realistisch?
--	--	--	---

Abgabe

Kategorie	Maximale Punkte	Erreichte Punkte	Prozent
Teil 1: Business Understanding & Datenexploration	20	_____	____%
Teil 2: Datenaufbereitung & Feature Engineering	30	_____	____%
Teil 3: Modellierung & Evaluation	10	_____	____%
Teil 4: Kritische Reflexion & Geschäftsempfehlungen	20	_____	____%
Jupyter Notebook & Bericht Gesamt	80	_____	____%

Präsentation

Kriterium	Erreichte Punkte
Struktur, Zeitmanagement und Foliengestaltung	___/5
Fokus auf Business-Insights (nicht nur technische Details)	___/5
Qualität der Visualisierungen und Storytelling	___/5
Gleichmäßige Beteiligung aller Gruppenmitglieder	___/5
Präsentation Gesamt:	___/20

Zusätzliche Kommentare (Stärken/ Verbesserungspotential/ Besondere Anmerkungen):

Prüfer: Joel Weiß & Samuel Schlenker

Datum:

Unterschrift: