Data Exploration Project

Daniel Wehner M.Sc., René Penkert

SAP SE / DHBW Mannheim

Sommersemester 2020

Agenda for this Unit

Planung und Organisatorisches

Grundsätzliches zum Projekt

- Name der Veranstaltung: *Data Exploration Project*
- Laut ⇒ Modulkatalog beträgt der Workload des Projekts **pro Person**:
  - Präsenzzeit: 27 Stunden
  - Selbststudium: 47 Stunden
- Definition des Projekts aus dem ⇒ Modulkatalog:
  „Anwendung von Methoden und Verfahren des maschinellen Lernens auf eine vorgegebene Datenbasis unter Laborbedingungen. Verwendung von üblichen Repositorien wie Hadoop/Spark/Flink/Mahout, Python-RASBT, R, etc. ***Ein besonderer Fokus soll auf einer ganzheitlichen wirtschaftsinformatischen Betrachtung liegen.*** *Es soll dabei neben der informatischen Betrachtung auch der betriebswirtschaftliche Nutzen, z. B. anhand eines Use Cases, betrachtet werden.“*

Bearbeitung

Grundsätzliches

- Das Projekt ist in Gruppen von **drei bis vier Studierenden** (nicht mehr und nicht weniger) zu bearbeiten.
- Die Organisation der Gruppen erfolgt selbständig durch die Studierenden. Bitte melden Sie sich, falls es Probleme bei der Gruppenfindung geben sollte.
- Geben Sie Ihrer Gruppe einen Namen!
- Jede Gruppe bearbeitet **ein anderes Thema**. Die Gruppen dürfen eigene Themen vorschlagen. *Siehe mögliche Themenvorschläge am Ende dieser Präsentation.*
- Bezüglich der zu verwendenden Technologien werden keine Einschränkungen gemacht, da je nach Projektthema andere Technologien sinnvoll sind.

Bei der Themenwahl ist darauf zu achten, den Umfang weder zu gering noch zu groß zu wählen! Bitte stimmen Sie daher das Thema mit uns ab!

Zwischenpräsentation

- Die Zwischenpräsentation beträgt **maximal 10 Minuten** und dient dem Zweck, die grobe Konzeption des Projektvorhabens darzulegen und zu präsentieren.
- Die Studierenden sind dazu angehalten, den anderen Gruppen Feedback zu geben, beziehungsweise der präsentierenden Gruppe Anregungen und Ideen mitzuteilen.
- Die Zwischenpräsentation geht **nicht** mit in die Endwertung ein.
- Von Seiten der Projektgruppen sind zu diesem Termin keine Arbeitsergebnisse oder Dokumente einzureichen.

Abgabe

- Neben dem erstellten ❶ **Quellcode** sind am Ende des Semesters eine ❷ **Abschlusspräsentation** sowie ein ❸ **Projektreport** anzufertigen.
- Die formalen Kriterien für die Abgabebestandteile werden auf den nachfolgenden Folien beschrieben.
- Sämtliche Dokumente sind **gezippt** als `*.pdf` Datei in Moodle einzureichen.
- Bitte beachten Sie folgende Namenskonventionen:

Abschlusspräsentation	`project_presentation_<group>.pdf`
Projektreport	`project_report_<group>.pdf`
Zip-Datei	`project_submission_<group>.zip`

Der Abgabetermin (und weitere wichtige Termine) können der ⇒ table 4 entnommen werden.

❶ Quellcode

- Der Quellcode ist auf einem **öffentlichen GitHub Repository** abzulegen.
- Dem Repository ist eine aussagekräftige `README.md` hinzuzufügen, welche mindestens Folgendes enthält:
  - Zielsetzung des Projekts
  - Eine Liste der Gruppenmitglieder
  - Eine kurze Beschreibung, wie der Quellcode ausgeführt werden muss (Installation/Dependencies/Packages/...)
- Auch die Qualität des Quellcodes geht mit in die Bewertung ein.

Kommentieren Sie den Quellcode ausreichend! Je leichter der Code für uns zu verstehen ist, desto besser ist es für Sie.

❷ Abschlusspräsentation

- Die Abschlusspräsentation weist einen zeitlichen Umfang von **20 Minuten** (+ max. 10 Minuten für Fragen) auf.
- Beachten Sie folgende Dinge für eine gute Präsentation:
  - Gehen Sie auf alle wichtigen Bestandteile Ihres Themas ein.
  - **Halten Sie unbedingt den zeitlichen Rahmen ein.** Der Spielraum betragt $\pm$ 3 Minuten.
  - Lesen Sie nicht ab, bauen Sie stattdessen Blickkontakt zum Publikum auf.
  - Gestalten Sie Ihren Vortrag lebendig (Modulation der Stimme, interessante Hinführung zum Thema, Fragen ins Publikum, ...).
  - Lassen Sie Raum für Fragen und Anmerkungen.
  - Die Folien sollten übersichtlich gestaltet sein, d. h. nicht zu viel Text und nutzen Sie Visualisierungen!

❸ Projektreport

- Der Projektreport ist gemäß den Regeln des wissenschaftlichen Arbeitens anzufertigen und weist einen Umfang von **minimal 3 und maximal 4 Seiten** (ohne Abbildungen und ohne Anhang) auf.
- Zum Zwecke der Vergleichbarkeit der Abgaben nutzen Sie bitte folgendes ⇒ LaTeX Template für die Erstellung Ihres Projektreports.
- Der Projektreport deckt mindestens folgende Bestandteile ab:
  - Thema und Motivation
  - Related Work (*welche wissenschaftlichen Publikationen gibt es zu diesem Thema bereits?*)
  - Verwendete Technologien und Bibliotheken (z. B. `scikit-learn`, `tensorflow`, ...)
  - Präsentation der Ergebnisse
  - Kritische Bewertung der Ergebnisse (*„lessons learned": Was hat (nicht) funktioniert und warum?*)
  - Anmerkungen zum Quellcode im Anhang (*wie ist der Code auszuführen und was gibt es zu beachten?*)

Bewertung des Projekts

- Die komplette Abgabe besteht aus dem Projektreport (`*.pdf`), der Abschlusspräsentation (`*.pdf`), sowie dem im Rahmen des Projekts erstellten Quellcode.
- **Nur rechtzeitig eingereichte Dokumente können bewertet werden!**
- Die einzelnen Bestandteile werden folgendermaßen gewichtet:
  - Quellcode und Ergebnisse (50 %)
  - Projektreport (30 %)
  - Abschlusspräsentation (20 %)
- Jeder Bestandteil wird mit **maximal 100 Punkten** bewertet (die Bewertungskriterien befinden sich auf den nachfolgenden Seiten).

Wichtig: Das Fehlen einer Teilabgabe führt zu erheblichem Punkteabzug (unter Umständen auch zum Nichtbestehen der gesamten Veranstaltung)!

Bewertungskriterien Quellcode

Nr.	Kriterium	Punktzahl
❶	Das Projekt ist auf GitHub veröffentlicht und enthält eine aussagekräftige `README.md`.	5
❷	Der Quellcode ist ausreichend kommentiert.	10
❸	Der Quellcode ist übersichtlich und ordentlich formatiert, intuitiv verständlich und folgt generell dem *„Clean Code“* Ansatz.	15
❹	Das gesamte Entwicklungsprojekt ist gut strukturiert und modular aufgebaut (z. B. Datenvorverarbeitung ist getrennt von Datenanalyse und Evaluation, etc.).	20
❺	Die Grundregeln des maschinellen Lernens werden berücksichtigt (z. B. korrektes Splitting, X-Val., Occam's razor, Hyper-Parameter Suche, etc.) und spiegeln sich im Code wider.	25
❻	Es wird eine geeignete Datenbasis ausgewählt und entsprechend dem vorliegenden Problem vorverarbeitet.	25
Summe		**100**

Table 1: Bewertungskriterien für den Quellcode

Bewertungskriterien Abschlusspräsentation

Nr.	Kriterium	Punktzahl
Mündlicher Vortrag		
❶	Der vorgegebene zeitliche Rahmen wird eingehalten (20 min $\pm$ 3 min).	10
❷	Die Vortragenden lesen nicht ab und sind imstande, frei zu sprechen (Blickkontakt).	10
❸	Die Geschwindigkeit des Vortrags ist dem Inhalt angemessen.	10
❹	Die Gruppenmitglieder können vom Publikum gestellte Fragen sicher beantworten.	20
❺	Der Vortrag geht auf alle relevanten Punkte ein.	15
Vortragsfolien		
❻	Die Folien liegen im `*.pdf`-Format vor.	5
❼	Die Gestaltung der Folien ist übersichtlich. Inhalte werden überwiegend visuell dargestellt.	15
❽	Alle wichtigen Aussagen sind in den Folien festgehalten.	15
Summe		**100**

Table 2: Bewertungskriterien für die Abschlusspräsentation

Bewertungskriterien Projektreport

Nr.	Kriterium	Punktzahl
❶	Der Projektreport liegt im richtigen Format vor (LaTeX-Vorlage und `*.pdf`-Format).	5
❷	Der Projektreport bedient sich einer sachgerechten Sprache und ist verständlich verfasst.	5
❸	Der Projektreport ist klar und übersichtlich strukturiert.	5
❹	Die Vorgaben bezüglich des Seitenumfangs werden eingehalten.	5
❺	Die Regeln des wissenschaftlichen Arbeitens werden berücksichtigt (z. B. Zitation).	25
❻	Inhaltlich werden alle geforderten Bereiche abgedeckt.	15
❼	Der wirtschaftliche Kontext wird ausreichend herausgestellt.	15
❽	Die erzielten Ergebnisse werden kritisch reflektiert und bewertet.	25
Summe		**100**

Table 3: Bewertungskriterien für den Projektreport

- Die Endnote für das Projekt berechnet sich somit folgendermaßen:

$$\text{Note} = \left\lceil \frac{5}{10} \cdot \Sigma_{\text{Code}} + \frac{3}{10} \cdot \Sigma_{\text{Report}} + \frac{2}{10} \cdot \Sigma_{\text{Präsentation}} \right\rceil \tag{1}$$

- Maximal zu erreichen sind 100 Punkte.
- Es gilt der offizielle Notenschlüssel der DHBW Mannheim.
- Die im Projekt erzielte Note wird mit der in der Klausur zur Vorlesung *„Applied Machine Learning Fundamentals“* erreichten Note verrechnet, um die Modulnote zu erhalten.

Das Ziel des Projekts ist es weniger, Ergebnisse zu erzielen, die dem „State of the Art“ entsprechen. Vielmehr steht eine korrekte (wissenschaftliche) Vorgehensweise im Vordergrund. Falls Sie keine guten Ergebnisse erzielen, sollten Sie jedoch darlegen, woran es gelegen haben könnte und diesbezüglich kritisch reflektieren!

Anwesenheitspflicht

- **Aufgrund der momentanen Situation wird die gesamte Veranstaltung ausschließlich virtuell stattfinden.**
- Mit wenigen Ausnahmen basiert die (virtuelle) Anwesenheit der einzelnen Projektgruppen an den Terminen **auf freiwilliger Basis**.
  - Es sollte jedoch regelmäßig Rücksprache bezüglich des Zwischenstands gehalten werden.
  - Bitte melden Sie sich **rechtzeitig** und **eigenverantwortlich**, falls von Ihrer Seite aus Diskussionsbedarf besteht. Nutzen Sie hierfür die Veranstaltungstermine, das Moodle-Forum, oder schreiben Sie uns eine E-Mail.
- **Anwesenheitspflicht** besteht an folgenden Terminen (siehe ⇒ table 4):
  - Einführung
  - Zwischenpräsentation
  - Finale Präsentation und Abgabe

Zeitlicher Ablauf des Projekts

Datum	von	bis	Bemerkung	Anwesenheitspflicht
08.05.2020	**3:30 pm**	4:30 pm	**Einführung**	ja
15.05.2020	3:00 pm	4:00 pm	*Abstimmungstermin*	nein
22.05.2020	3:00 pm	4:00 pm	*Abstimmungstermin*	nein
29.05.2020	**4:00 pm**	5:00 pm	*Abstimmungstermin*	nein
05.06.2020	3:00 pm	6:00 pm	**Zwischenpräsentation**	ja
12.06.2020	3:00 pm	4:00 pm	*Abstimmungstermin*	nein
19.06.2020	**4:30 pm**	5:30 pm	*Abstimmungstermin*	nein
26.06.2020	3:00 pm	4:00 pm	*Abstimmungstermin*	nein
03.07.2020	3:00 pm	4:00 pm	*Abstimmungstermin*	nein
10.07.2020	**3:30 pm**	4:30 pm	*Abstimmungstermin*	nein
17.07.2020	3:00 pm	4:00 pm	**Abgabe Moodle**	nein
20.07.2020	**9:00 am**	12:00 pm	**Finale Präsentation I**	ja
21.07.2020	**9:00 am**	12:00 pm	**Finale Präsentation II**	ja

Table 4: Alle wichtigen Termine der Veranstaltung

Themen

Eigene Themen

- Laut ⇒ Modulkatalog soll der Fokus auf einer *„ganzheitlichen wirtschaftsinformatischen Betrachtung“* liegen, und auch dem betriebswirtschaftlichen Aspekt Rechnung getragen werden.
- Es ist grundsätzlich erlaubt und auch erwünscht, **eigene Themenvorschläge** einzubringen.
- Eigene Themen müssen natürlich vorher genehmigt werden.
- Auf der nächsten Folie sind einige Projektvorschläge aufgelistet, falls einzelne Gruppen kein eigenes Thema finden sollten.

Themenvorschläge

- `Sentiment Analyse von Kundenrezensionen` (kommt das Produkt beim Kunden gut oder schlecht an?)
- `Vorhersage von Aktienkursen` (falls Sie reich werden möchten, ist das ein guter Anfang)
- `Recommender Systems` (z. B. *Collaborative Filtering*, siehe Netflix)
- `Baue deinen eigenen Chatbot` (z. B. zur automatischen Beantwortung von Kundenfragen)
- `Spracherkennung`
- `Analyse medizinischer Scans zur Krankheitsdiagnose`
- `CureMannheim` (autonomes Fahren, ⇒ Cure Mannheim e. V.)
- `Automatische Steuerung von Drohnen` (virtuell: ⇒ AirSim Simulation)
- `Erkennung von „Fake News“` (Falschmeldungen können wirtschaftliche Entscheidungen stark beeinflussen)
- `Vorhersage des Bitcoin Preises`

- `Aufdeckung von Kreditkartenbetrug`
- `Segmentierung von Kunden` (Unterteilen Sie Kunden bezüglich ihres Kaufverhaltens, Alters, Geschlechts, ...)
- `Identifikation von Emotionen in Texten / Audio`

Für zusätzliche Inspiration, siehe das ⇒ UCI Machine Learning Repository!

Thank you very much for the attention!

Topic: Data Exploration Project
Term: Sommersemester 2020

Contact:
Daniel Wehner M.Sc., René Penkert
SAP SE / DHBW Mannheim
daniel.wehner@sap.com

Do you have any questions?