EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Wirtschafts- und Sozialwissenschaftliche Fakultät

LEHRSTUHL FÜR MARKETING
Prof. Dr. Dominik Papies

Telefon +49 7071 29-78202
Telefax +49 7071 29-5078
dominik.papies@uni-tuebingen.de
www.uni-tuebingen.de/wiwi/marketing

Modul „Data Science Project“ 2021/2022

im Rahmen des M.Sc. „Data Science in Business and Economics” (Stand: 5.10.2021)

Die Studierenden im M.Sc. Data Science in Business and Economics müssen ein praktisch orientiertes Modul “Data Science Project” belegen, das einen Umfang von 12 ECTS hat. Mit 12 ECTS hat das Projekt ein hohes Gewicht, was vor dem Hintergrund, dass eine eigenständige Arbeitsweise erwartet wird, zu rechtfertigen ist.

Die Grundidee dieses Moduls ist, dass Studierende **eigenständig** und **in Gruppen** den gesamten Prozess eines Data Science Projektes einmal von Beginn bis Ende konzipieren und programmieren. Ein Fokus soll dabei auf der **Automatisierung** liegen, d.h., dass die Beschaffung der Daten, das Strukturieren, Einlesen, Validieren, Modifizieren und Auswerten (weitestgehend) ohne Eingriff des Analysten oder Nutzers der Analyse erfolgt.

Ein Data Science Project soll die folgenden Komponenten umfassen.

1. **Datenakquise**: das Projekt soll die Daten beschaffen und einlesen, z. B. indem Daten von Webseiten ausgelesen werden („web scraping“), in dem eine oder mehrere APIs („application programming interface“) genutzt werden, oder indem Textdokumente ausgelesen werden. Möglich ist auch, dass Studierende einen neuen Datensatz semi-manuell selbst erstellen oder Studierende einen bestehenden Datensatz als Grundlage für das Projekt nehmen. Die Komplexität und der Aufwand der Datenakquise gehen mit in die Bewertung ein. Die Daten enthalten idealerweise Komponenten, die unstrukturiert sind, also z. B. Text oder Bilder.
2. **Datenaufbereitung**: die Daten sollen eingelesen und rechenbar gemacht werden. Das bedeutet unter anderem, dass die einzelnen Teil-Datensätze eingelesen und mit anderen Datensätzen gemerged werden, dass unstrukturierte Daten (z. B. Text) in rechenbare Formate umgewandelt werden, dass korrekte Identifier für cross-sectional units (z. B. Firmen, Marken, Produkte, Länder, Personen) und Zeitpunkte generiert werden.
3. **Datenvalidierung**: die Daten sollen validiert werden, u. a. indem deskriptive Statistiken und Abbildungen generiert werden, unplausible Beobachtungen und Ausreißer identifiziert werden. Nachdem Kriterien zur Validierung festgelegt sind, sollte auch dieser Prozess weitgehend ohne den Eingriff des Nutzers geschehen.
4. **Datenanalyse:** im Zentrum des Projekts soll eine relevante Fragestellung stehen, diese kann, muss aber nicht notwendigerweise ökonomischer Natur sein. Dieses kann z. B. ein Zusammenhang zwischen Variablen sein (korrelational oder kausal) oder es kann die Vorhersage ökonomischer Zustände betreffen (predictive analytics). Die entsprechende

Analyse soll mit adäquaten Methoden an dieser Stelle (traditionelle Statistik oder Maschinelles Lernen) eingebettet werden.

5. **Datenvisualisierung:** die Struktur und Verteilung der Daten, die Zusammenhänge und Ergebnisse der Datenanalyse sollen in ansprechender und informativer Art, dem State-of-the-Art der Datenvisualisierung entsprechend, dargestellt werden.
6. **Einbettung der Ergebnisse:** alle Ausgaben des Projektes (z. B. Tabellen, Grafiken, Ergebnisse), sollen in einer interaktiven Umgebung eingebettet werden, z. B. in Form einer Shiny-App. So soll der Nutzer / „Leser“ die Verteilung der Daten, die Variablen, und die Ergebnisse (z. B. Zusammenhänge, Vorhersagen) durchstöbern und erforschen können. Beispiele für vergleichbare Shiny-Apps finden sich unter anderem auf der Shiny-App Website von RStudio (https://shiny.rstudio.com/gallery/).
7. **Automatisierung:** das gesamte Projekt mit den genannten Schritten soll automatisiert ablaufen. Was ist darunter zu verstehen? Stellen wir uns einen beliebigen Datensatz vor. Diesen teilen wir nun in zwei Teile. Die oben genannten Schritte (1-6) werden nun programmiert (d.h., der Code wird erstellt) unter Nutzung einer Hälfte des Datensatzes. Wenn nun die zweite Hälfte des Datensatzes eingelesen wird, dann soll der gesamte Code ausgeführt und die Ergebnisse dargestellt werden können, ohne dass dafür Änderungen am Code vorgenommen werden müssen. Damit soll ein Workflow imitiert werden, der in der Praxis häufig relevant ist: eine Datenanalyse- und Visualisierungs*struktur* wird erstellt (d.h., programmiert und getestet), und dann kommen regelmäßig neue Daten rein, die ein Anwender, der den Code nicht erstellt hat, evaluieren und erforschen kann, ohne den Code „anfassen“ zu müssen. Wichtig ist hierbei, dass der Code vollständig und gut dokumentiert wird (Kommentare einbetten), damit externe Parteien den Code nachvollziehen und überprüfen können.

Was ist der Unterschied zwischen einer Masterarbeit und dem Data Science Project?

Bei einer Masterarbeit steht die Lösung eines relevanten Forschungsproblems im Fokus. Die Forschungsfrage muss in der relevanten Literatur verortet werden. Die Analysen werden im Text beschrieben, und die Ergebnisse werden in Form von statischen Tabellen oder Grafiken abgedruckt. Die Arbeit interpretiert dann die Ergebnisse und zieht Schlüsse. Es werden keine Anforderungen an Automatisierung gestellt, und der Leser „konsumiert“ den Text, ohne mit den Daten selber zu interagieren.

	Masterarbeit	**Data Science Project**
Forschungsproblem herausarbeiten	Hohes Gewicht	Geringes Gewicht
Literaturarbeit	Hohes Gewicht	Geringes Gewicht
Automatisierung	Geringes Gewicht	Sehr hohes Gewicht
Leser / Nutzer interagiert mit Daten	Geringes Gewicht	Sehr hohes Gewicht
Schlussfolgerung	Vor allem Autor	Eher Leser / Nutzer

Bewertung

Die Studierenden präsentieren (ggf. in ihren Gruppen) zwei Mal. In einer Zwischenpräsentation stellen sie den Plan und erste Schritte vor. In der Abschlusspräsentation zeigen sie das fertige Projekt, den zentralen Aufbau des Codes, und wie Nutzer mit dem Ergebnis interagieren können.

Bewertet wird das Projekt nach den folgenden Kriterien:

1. Komplexität des Gesamtprojektes
2. Qualität des Codes
3. Inhaltlicher Anspruch der untersuchten Fragestellung, Niveau der Umsetzung des inhaltlichen Problems, logische Konsistenz der Fragestellung und Umsetzung

Zur Bewertung erstellen die Studierenden die folgenden Komponenten:

1. Eine Präsentation, die die zentrale Fragestellung motiviert, die Relevanz herleitet, und ggf. Bezug auf die relevante Literatur herstellt.
2. Die Studierenden präsentieren im Rahmen der Abschlusspräsentation das fertige „Produkt“, z.B. eine Website (Shiny-App) oder eine vergleichbare Möglichkeit, als Nutzer mit den Daten zu interagieren.
3. Die Studierenden stellen den Code auf GitHub bereit. Der Code muss nachvollziehbar kommentiert sein, und es muss aus dem Code, der Benennung der Dateien, oder einer begleitenden Dokumentation klar hervorgehen, welche Dateien in welcher Reihenfolge ausgeführt werden müssen. Pfade müssen so gesetzt werden, dass keine manuelle Änderung von Pfadnamen notwendig ist, wenn der Code auf anderen Computern ausgeführt wird. Zudem stellen die Studierenden die im Projekt genutzten Daten bereit, im Regelfall durch ein zufälliges (kleines) Sample auf GitHub, dazu den vollständigen Datensatz als FileTransfer (z. B. Dropbox).
4. Studierende stellen das Interface (z. B. Shiny-App) den anderen Studierenden und den Lehrenden zur Verfügung, in dem dieses z. B. in der BWCloud gehostet wird.
5. Die Studierenden erstellen ein kurzes Video (ca. 3 Minuten), das das Projekt und das Ergebnis anschaulich für einen breiteren Kreis von Nutzern darstellt.

Themen und Projekteideen

Neben den unten genannten Themen ist es möglich und erwünscht, dass Studierende eigene Themenideen einbringen. Diese *können* sich an aktuellen Themen und Themen der vergangenen Jahre orientieren. Studierende sind eingeladen, ihre Themenvorstellungen inkl. Datenquellen und Analysemethoden in den Wochen vor dem Kick-off dem Modulverantwortlichen (Dominik Papies) vorzustellen. Dann wird auch besprochen, ob für das Thema eine zusätzliche Fachbetreuung notwendig ist.

1. **Gender gap in academic publishing**: Auf welchen Dimensionen unterscheidet sich das Publikationsverhalten von weiblichen und männlichen Autoren? Erstellen einer Datenbank mit Publikationen in Business & Economics, algorithmisches Bestimmen des Geschlechts der Autoren, Berechnen des Gendergaps, Analyse der Netzwerke von männlichen und weiblichen Autoren (Ko-Autorenschaften), Schätzen des Impacts von Publikationen mit weiblicher und männlicher Beteiligung.
2. **Soccer prediction App**: Sind die Vorhersagen von Sportergebnissen, die auf Basis von Machine Learning erstellt werden, besser als z.B. die Vorhersagen auf Basis von Wettquoten? Erstellen eines Dashboards, mit dem auf Basis eines prediction algorithms die Bundesliga-Ergebnisse des jeweils nächsten Spieltages vorhergesagt werden. Dabei sollen Spieler-Charakteristika sowie Ergebnisse der letzten Spiele mit einbezogen werden.
3. **Song lyrics**: Wie unterscheiden und entwickeln sich die Texte in der Popmusik? Gibt es strukturelle Unterschiede in Themen, Komplexität, Anzahl der Wörter, etc. zwischen Genres oder Künstlern? Aufbau einer Datenbank, Textanalyse, Darstellung über unterschiedliche Dimensionen.
4. **Benzinpreisvorhersagen**: Ist es möglich, mit hinreichend großer Genauigkeit, die Benzinpreise für bestimmte Tage und Uhrzeiten in der Zukunft für bestimmte Tankstellen vorherzusagen? Nutzung einer bestehenden Datenbank mit Benzinpreisen, Entwicklung eines Vorhersagetools, Ermittlung der Prognosegenauigkeit.
5. **Earnings conference calls & firm valuation**: Welche Rolle spielen Inhalte und Sentiment der Aussagen von Vorstandsmitgliedern (CEO und CFO) in earnings conference calls bei den direkten Investitionsentscheidungen von Investoren? Lassen sich auf Basis der Inhalte und des Sentiments Veränderungen in der Marktbewertung von Firmen vorhersagen? Aufbau einer Datenbank mit Transcripts von earnings conference calls, Extraktion wesentlicher Features mit language processing, Entwicklung eines Vorhersagemodells.
6. … weitere Themen werden im Verlauf der nächsten Wochen hinzugefügt.

Themen, die in vergangenen Jahren bearbeitet wurden:

1. **Corona-Datenbank**: Aufbau einer Datenbank, die auf möglichst niedrigem Aggregationsniveau (subnational) in möglichst vielen Ländern Informationen zu Infektionszahlen, Todeszahlen, Einführung

von NPI (nonpharmaceutical interventions wie z. B. Kontaktbeschränkungen, Maskenpflicht) und Aufhebung der NPI enthält.

2. **Immobilienpreise**: über API von Immobilienportalen großflächig Immobilienpreise für Kauf- und Mietimmobilien beschaffen, aufbereiten und visualisieren (u.a. auf interaktiven Landkarten). Durchführung von hedonischen Preisregressionen und Berechnung des Spreads Miete vs. Eigentum, über die Zeit aktualisieren.
3. **Gender gap in academic publishing**: über geeignete API Meta-Daten zu Publikationen in Business & Economics extrahieren, inkl. Autoreninformation. Anteil der Publikationen mit weiblichen Autoren berechnen und visualisieren, Darstellung über unterschiedliche Disziplinen.
4. **Sentiment and stock markets**: Erfassen des Social Media sentiments von Nutzern, Entwicklung über die Zeit, Berechnung des Zusammenhangs mit Stock Market Returns, Rolle der Corona-Pandemie.
5. **ESG-Dashboard**: Entwicklung eines Dashboards, mit dem Nutzer Merkmale von Aktienportfolios explorieren können, mit besonderem Fokus auf ESG-Merkmalen.

Zeitplan

Kick-off: 22.10.2021, 14 Uhr

Zwischenpräsentation: Letzte Vorlesungswoche vor Weihnachten

Abschlusspräsentation: 2. Februarwoche

Dazwischen sollen regelmäßige Treffen (ggf. Online) stattfinden, in denen Fragen diskutiert werden können.