DSCB310 – Datenanalyse und Business Intelligence 1

Übungsblatt 1

Wintersemester 2022/23

Hochschule Karlsruhe
University of
Applied Sciences

Inhalt

1 Zweck und allgemeine Hinweise

Eine Basiskompetenz von Data Scientists ist die Analyse und Identifikation von Zusammenhängen und Abhängigkeiten in Daten, um Erkenntnisse aus den Daten im Hinblick auf den (betriebswirtschaftlichen) Anwendungsfall ziehen und Hypothesen über den Betrachtungsgegenstand aufstellen zu können.

Die Analyse bedingt im Kontext dieser Veranstaltung vornehmlich drei Aspekte: (1) **Vorverarbeitung** und Bereinigung der Daten (bspw. Ermittlung „neuer" Attribute wie „Wochentag aus einem Datum ableiten"), (2) Identifikation von Zusammenhängen unter Verwendung **statistischer Methoden** und Möglichkeiten der Datenaggregation und (3) **Visualisierung** der Erkenntnisse über die ermittelten Zusammenhänge sowie weiterer „interessanter" Inhalte aus den Daten (ggf. zusammengefasst als einzelne „Data Stories").

Bei allen Aspekten ist es wichtig, den „Anspruchsgruppen" (=Stakeholder) der Analyse zu kennen und zu berücksichtigen. In dieser ersten Übung werden Sie Daten eines E-Scooter-Verleihs analysieren, aufbereiten und visualisieren. Es handelt sich dabei um reale Daten. Ihr Ziel ist es, aus den Daten nutzenstiftende Informationen zu extrahieren und Ihre Analyse möglichst intuitiv nachvollziehbar und in angemessenem Detailgrad den Stakeholdern zur Verfügung zu stellen.

Die Stakeholder der Analyse sollen (1) die *Geschäftsführung* des E-Scooter-Verleihs sein, die Interesse an der Darstellung allgemeiner Erkenntnisse aus den Daten hat (bspw. Nutzungsentwicklung), um so das E-Scooter-Geschäftsfeld besser zu durchdringen.

Eine weitere Anspruchsgruppe ist (2) der *Abteilungsleiter der Disposition* des E-Scooter-Verleihs sein. Die Disposition kümmert sich darum, dass die sogenannten „Juicer"[1] eingeplant werden. Ihre Analysen sollen somit das operative Geschäft des Dispositionsleiters unterstützen. Wiederkehrende „Muster" in den Daten sind dabei von besonderem Interesse, um den betroffenen Geschäftsprozess dementsprechend optimieren zu können.

Nutzen Sie Ihre Analyseerkenntnisse kreativ, um Verbindungen zum „Business-Kontext" der beiden Anspruchsgruppen herzustellen und vermitteln Sie diese effektiv im Rahmen eines Jupyter-Notebooks.

2 Organisation und einzureichende Dokumente

Die Übung ist durch Projektgruppen bestehend aus 4 Studierenden umzusetzen. Die Gruppen müssen in ILIAS registriert sein (vgl. Hinweise in den Vorlesungen).

- **Start Übung 1**: 14.10.2022
- **Einreichung Jupyter Notebook (Detailhinweise s. u.): 25.11.2022, 8 Uhr**
  Über ILIAS muss das Notebook von der Gruppe bis zur Deadline eingereicht werden (Abschnitt „Übungsaufgaben" → „Baustein Übung" – dort können Sie die Datei mit dem Notebook je Gruppe hochladen)
- **Vorstellung Notebook (Detailhinweise s.u.):** 25.11.2022, 11:30-13 Uhr (Übung um 8:00 Uhr fällt an dem Tag aus) - pro Gruppe sind 10 Minuten Vorstellung und ca. 5 Minuten Diskussion eingeplant)

[1] „Juicer" sammeln die E-Scooter ein, laden sie selbständig auf und bringen die Scooter dann wieder zu vordefinierten Ausgangspunkten.

3 Datensatz

Der gegebene Datensatz steht in ILIAS zur Verfügung („escooter_history.parquet“). Pro Zeile ist eine einzelne E-Scooter-Ausleihe eines Nutzers hinterlegt (die einzelnen Nutzer sind nicht identifizierbar). Der Zeitstempel „datetime“ gibt den Start der Ausleihe an. Die Spalten „holiday“ und „workingday“ geben an, ob es sich um einen Feiertag gehandelt hat bzw. ob es ein Arbeitstag war.[2] In der Spalte „weather“ ist eine grobe Klassifikation des Wetters zu dem jeweiligen Zeitpunkt vorgegeben. Mittels „temp“ bzw. „atemp“ wird die Temperatur bzw. die gefühlte Temperatur beim Start der Ausleihe in der relevanten Region angegeben. Analog sind auch die Luftfeuchtigkeit und die Windgeschwindigkeit („humidity“ bzw. „windspeed“) vorgegeben. Über das Feld „registered_customer“ ist erkennbar, ob der E-Scooter spontan ohne vorherige Registrierung von einem Kunden ausgeliehen wurde, oder ob es sich um einen registrierten Kunden gehandelt hat.

	datetime	holiday	workingday	weather	temp	atemp	humidity	windspeed	registered_customer
0	2020-01-04 00:00:09	0.0	0.0	clear, few clouds	9.84	14.395	81.0	0.0000	True
1	2020-01-04 00:00:41	0.0	0.0	clear, few clouds	9.84	14.395	81.0	0.0000	True
2	2020-01-04 00:01:20	0.0	0.0	clear, few clouds	9.84	14.395	81.0	0.0000	True
3	2020-01-04 00:04:12	0.0	0.0	clear, few clouds	9.84	14.395	81.0	0.0000	True
4	2020-01-04 00:15:19	0.0	0.0	clear, few clouds	9.84	14.395	81.0	0.0000	True
...	...	...	...	...	...	...	...	...	...
9	2022-01-09 23:49:44	0.0	1.0	clear, few clouds	10.66	13.635	65.0	8.9981	True
9	2022-01-09 23:52:25	0.0	1.0	clear, few clouds	10.66	13.635	65.0	8.9981	True
9	2022-01-09 23:53:31	0.0	1.0	clear, few clouds	10.66	13.635	65.0	8.9981	True
9	2022-01-09 23:56:59	0.0	1.0	clear, few clouds	10.66	13.635	65.0	8.9981	True
9	2022-01-09 23:57:16	0.0	1.0	clear, few clouds	10.66	13.635	65.0	8.9981	True

3760822 rows × 9 columns

4 Aufgabenstellung

Ihre Aufgabe ist eine zielgruppengerechte Analyse des Datensatzes, d.h. die *Geschäftsführung* bzw. *Disposition* des Unternehmens soll anhand Ihrer Auswertung in der Lage sein, das Geschäftsfeld besser zu durchdringen und Rückschlüsse für die eigenen betrieblichen Handlungsbereiche aus den Daten zu ziehen.

Darüber hinaus hat ein Mitarbeiter der Disposition spezifische Fragen formuliert, die im Rahmen eines Projektes aufgekommen sind. Das Projekt bezieht sich dabei nur auf *nicht registrierte* User:

- Welche der Wetterattribute beeinflussen die Ausleihzahlen der nicht registrierten User am stärksten? Insbesondere: Ist *temp* oder *atemp* relevanter?
- Spielt an Feiertagen mit Windgeschwindigkeit>35 die Temperatur noch eine Rolle für die Ausleihzahlen der nicht registrierten User?

Bauen Sie systematisch ein Jupyter-Notebook mit allen Analyseschritten und -ergebnissen auf (nutzen Sie also insbesondere Markdown-Zellen und/oder Code-Kommentare). Die Stakeholder sollen sich in Kombination mit einer kurzen Erklärung Ihrerseits einen Überblick verschaffen und die für sie relevanten Informationen schnell auffinden können.

[2] Achtung: die Daten zu den Feiertagen sind nicht „sinnvoll“. Sie müssen diese im Hinblick auf den genauen Feiertag somit nicht interpretieren, sondern es reicht die Betrachtung „Feiertag vorhanden ja/nein“.

Für die Analyse sind einige **Vorverarbeitungsschritte** nötig. Nutzen Sie Ihre Kenntnisse in Python (bzw. den besprochenen Bibliotheken), um die Daten grundsätzlich zu verstehen und anzureichern. Beispiele für Vorverarbeitungs- und Analyseschritte (nicht abschließende Aufzählung!):

- Einfache Analysen anhand pandas-profiling oder „GroupBy“ und Kennzahlen aus der Deskriptiven Statistik, um sich mit dem Datensatz vertraut zu machen.
- Informationsextraktion aus dem Zeitstempel (Beispiele: Tages-/Wochen-/Monats-/Quartalsprofile)
- Wo Ihnen es sinnvoll erscheint, Umwandlung numerischer in nominal-skalierte Attribute (bspw. mittels „pd.cut“), um einen leichten Überblick zu erhalten.
- WICHTIG: Im zeitlichen Verlauf der Nachfrage sind Muster enthalten. Suchen Sie diese Muster und analysieren Sie sie im Hinblick auf die Effekte in der eScooter-Nachfrage so detailliert wie möglich (bspw. Dauer, Periodizität). Geben Sie eine Empfehlung, welches Attribut in die Quelldaten aufgenommen werden sollte.
- Bringen Sie ggf. weitere eigene Ideen ein, um aus den bekannten Informationen relevante Aspekte zu extrahieren und für weitere Analyseschritte zur Verfügung zu stellen.
- Identifizieren Sie etwaige Datenqualitätsprobleme, bewerten Sie diese im Hinblick auf die Kritikalität für die Stakeholder und schlagen Sie eine oder mehrere Lösungsmöglichkeiten zum Umgang mit diesen Problemen vor.

Empfehlung für das Notebook: Beschreiben Sie die umgesetzten Vorverarbeitungsschritte in Markdown-Zellen knapp (bspw. stichpunktartig) und nutzen Sie die Strukturierungsmöglichkeiten in Markdown-Zellen (Hierarchieebenen durch „#“). Gehen Sie in einem Abschnitt „Data Quality“ auf die identifizierten Datenqualitätsprobleme ein.

Wenden Sie in **Data Analytics behandelte Analysemethoden** an, um aus den Daten Erkenntnisse zu gewinnen. Hinterfragen Sie mögliche Schlussfolgerungen aus Ihren Beobachtungen und überprüfen sowie belegen Sie diese so weit wie möglich mit Daten. Dokumentieren Sie nicht nur die Schlussfolgerungen, sondern auch die Belege und Argumente, die diese stützen.
Beispiele für Analysemethoden:

- Vergleich bedingter relativer Häufigkeiten
- Untersuchung auf (bedingte) stochastische Unabhängigkeit
- Korrelationsanalyse
- Lineare Regression und Interpretation der Koeffizienten
- Prüfung auf Signifikanz

Empfehlung für das Notebook: Bereiten Sie Ihre Erkenntnisse jeweils zielgruppenbezogen in einzelnen Teilbereichen auf. Belegen Sie vorgenommene Schlussfolgerungen mit (aggregierten) Daten und Kennzahlen in angemessenem Detailgrad. Die Kenntnis der Stakeholder bzgl. gängiger Kennzahlen und ihrer Interpretation können Sie als bekannt voraussetzen.

Nutzen Sie die vorgestellten Möglichkeiten zur **Datenvisualisierung**, um zielgruppengerecht die von Ihnen ermittelten Erkenntnisse grafisch zu kommunizieren und damit insbesondere den Teil „Data Analytics“ visuell zu unterstützen.

- Stützen Sie Ihre statistischen Analysen und ggf. die Vorverarbeitungsschritte (s.o.) durch die Abbildungen und erstellen Sie ggf. weitere Visualisierungen, um den Anspruchsgruppen relevante Sachverhalte aus den Daten darzustellen.
- Verwenden Sie unterschiedliche Ansätze und Diagramme, um die jeweils hervorzuhebenden Aspekte zu visualisieren. Sie dürfen über die vorgestellten Diagrammtypen hinaus auch weitere Diagramme oder auch andere Frameworks verwenden.
- Ggf. bieten sich für einzelne Diagramme vorab auf die Visualisierung zugeschnittene Aggregationen des Original-DataFrames mittels „GroupBy“ an. Beispiel: um die Streuung in einem Tagesprofil darstellen zu können kann man zunächst einen DataFrame erstellen, der die Anzahl der Ausleihen pro Datum und Stunde aggregiert darstellt. Anschließend kann anhand

eines BoxPlots das Tagesprofil (also die Anzahl Ausleihen in den jeweiligen Stunden des Tages) geschickt dargestellt werden.
- Seien Sie vorbereitet auf Rückfragen wie „Warum hat sich in der Situation der gewählte Diagrammtyp angeboten? Gab es Alternativen?“.
- Achten Sie so weit wie möglich auf Konsistenz in der Darstellung aller Diagramme (bspw. Farbgebung, Schriftarten etc.).[3]
- Sie dürfen neben den Python-Tools weitere Programme verwenden, um Ihre Visualisierungen zu ergänzen (bspw. Excel/Visio/Powerpoint o.ä.). Siehe dazu die Hinweise unten.

Empfehlung für das Notebook: Führen Sie Ihre Visualisierungen auf und beschreiben Sie sehr knapp deren Interpretation sowie die Kernerkenntnisse in einer Markdown-Zelle bzw. auch dem Titel. Nutzen Sie einen angemessenen Detailgrad.

Die drei beschriebenen Aspekte „Vorverarbeitung“, „statistische Methoden“ und „Visualisierung“ müssen nicht sequentiell „hintereinander“ abgehandelt werden; Stattdessen wird ein systematischer Aufbau des Notebooks erwartet, mit dem Sie für die genannten Zielgruppen „Data Stories“ (also Erkenntnisse aus den Daten) berichten können.

Wichtig für alle Schritte: Relevant sind die in den Vorlesungen/Übungen vorgestellten Konzepte (statistische Methoden, Vorverarbeitungskonzepte, Visualisierungen) bis 1 Woche vor dem Abgabetermin sowie alle Inhalte aus den vorherigen Semestern (insbesondere in Bezug auf die Vorverarbeitungsschritte sowie Möglichkeiten zur Aggregation und Auswertung von Daten).

5 Anforderungen an das Jupyter Notebook

- Erstellen Sie eine Markdown-Zelle mit Inhalt „DSCB310 – Abgabe Übungsblatt 1 (WS202/23)“ sowie die Matrikelnummern der Gruppenteilnehmer (keine Namen!) und die Gruppennummer. In den folgenden Zellen sind Sie frei im Hinblick auf die Systematik und Gestaltung.
- Das Notebook sollte durch „Run All“ ausführbar sein (inkl. Einlesen der Quelldatei)
- Sie dürfen die Jupyter-Visualisierungen oder -Tabellen auch in einem anderen Tool wie Excel/Visio/PowerPoint oder ähnlichem nachbearbeiten, um Aspekte hervorzuheben (sofern die Methode der Hervorhebung in der Veranstaltung nicht behandelt wurde).
    - Bitte fügen Sie in dem Fall Screenshots in eine Markdown-Zelle ein und reichen Sie das Notebook zusammen mit den Bilddateien als Zip-Datei ein.
    - Kennzeichnen Sie entsprechende Abbildungen zusätzlich bitte mit „Erstellt in Jupyter-Notebook und nachbearbeitet mit <Tool-Name>“.
- Sprache im Notebook frei wählbar (Deutsch oder Englisch).
- Bitte laden Sie das Notebook bzw. ggf. die Zip-Datei in ILIAS hoch.

6 Anforderungen an die Präsentation

- Alle Gruppen müssen die Ergebnisse dieser Übung knapp präsentieren (Sprache: Deutsch).
- Bitte erstellen Sie keine PowerPoint (o.ä. Präsentationsfolien), sondern stellen Sie nur das Notebook vor (Live-Ausführung ist nicht notwendig, insb. bei langandauernden Operationen).
- Zeitlicher Rahmen: 10 Minuten Vorstellung und ca. 5 Minuten Diskussion
    - Genaues Timing ist ein Bewertungskriterium!
- Sie dürfen selbst entscheiden wer präsentiert (auch wie viele Studierende), jedoch muss im Laufe des Semesters jeder einmal präsentiert haben.

[3] Wenn Sie unterschiedliche Tools (bspw. seaborn und plotly) nutzen, dann sollten zumindest alle seaborn-Plots und alle plotly-Plots konsistent sein.