DSCB310 – Datenanalyse und Business Intelligence 1

Projektaufgabe

Hochschule Karlsruhe
University of
Applied Sciences

Wintersemester 2023/24

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 Anwendungsfall ziehen und Hypothesen über den Betrachtungsgegenstand aufstellen und verifizieren zu können.

Die Analyse bedingt im Kontext dieser Veranstaltung vornehmlich drei Aspekte: (1) **Vorverarbeitung** der Daten (Bereinigung; Generieren neuer Features, die z.B. aus der Vorgeschichte eines Spiels abgeleitet werden), (2) **Analyse von Zusammenhängen** (und ggf. Überprüfung der Aussagekraft mit statistischen oder anderen Methoden) 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 Übung werden Sie Daten aus Tennisturnieren untersuchen. 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.

Stakeholder der Analyse soll die **sportliche Leitung des deutschen Tennisbundes** sowie die **Trainer deutscher Tennisspieler** sein, die an den im Datensatz erfassten Turnieren teilnehmen. Ziel ist es, Erkenntnisse zu gewinnen, wovon der **Erfolg eines Spielers generell und in einem konkreten Match** abhängt. Erkenntnisse hierzu können einerseits für die Talentsuche nützlich sein und andererseits lassen sich daraus möglicherweise Schlussfolgerungen über Fördermaßnahmen und Training der Spieler ziehen – allgemein, individualisiert für bestimmte Spieler, oder spezifisch für die Vorbereitung auf einzelne Matches.

Im Vorfeld wurden bereits konkrete Hypothesen und Fragen formuliert, die Sie überprüfen sollen:

1. Gibt es „Angstgegner"-Nationen, gegen die die deutschen Spieler besonders schlecht abschneiden? Lässt sich das einfach durch die generelle Spielstärke der Spieler dieser Nationen erklären, oder gibt es Nationen, gegen die gerade deutsche Spieler schlecht abschneiden? Gibt es analog „Lieblingsgegner"-Nationen?
2. Gibt es „Angstgegner" einzelner deutscher Spieler, d.h. Paarungen, bei denen der deutsche Spieler schlechter abschneidet, als es aufgrund der allgemeinen Spielstärke beider Spieler zu erwarten ist? Gibt es analog „Lieblingsgegner" einzelner deutscher Spieler?
3. Gibt es Turniere, bei denen die Deutschen besonders gut oder besonders schlecht abschneiden? Wenn ja, ist ein regionales Muster oder eine andere Ursache zu erkennen?
4. Wie entwickelt sich der sportliche Erfolg mit dem Alter? Gilt das für die Spielstatistiken gleichermaßen wie für die Platzierung auf der Rangliste?
5. Sinkt die Wahrscheinlichkeit zu siegen, wenn das *vorhergehende* Spiel eines Spielers besonders lange gedauert hat?
6. Gibt es einen Zusammenhang dazwischen, wie man zuletzt gegen einen bestimmten Gegner gespielt hat, und wie das nächste Spiel gegen diesen Gegner ausgehen wird?
7. Wie gut lässt sich das Ergebnis eines Matches vorhersagen? Welches sind die wichtigsten Features für eine solche Prognose?

2 Organisation und einzureichende Dokumente

Die Übung ist durch Projektgruppen bestehend aus 3-4 Studierenden umzusetzen. Die Gruppen müssen in ILIAS registriert sein (vgl. Hinweise in den Vorlesungen).

- **Start Übung 1**: 10.11.2023 (Bekanntgabe der Daten)

- **0. Meilenstein bis Freitag 24.11.2023**:
  Tragen Sie Ihre Teamzusammensetzung in ILIAS ein. (Siehe Ordner „Projektaufgabe“)
- **1. Meilenstein**: Durchsprache Ihrer Vorgehensplanung und bisherigen Ergebnisse:
  Vereinbaren Sie einen Feedback-Termin im Zeitraum **4.12. bis 21.12.** (Link zur Terminvereinbarung in Kürze in ILIAS unter „Projektaufgabe“.
- **Abgabe der Ergebnisse (Detailhinweise s. u.): Mittwoch 10.01.2023, 8 Uhr**
  Das Jupyter Notebook sowie der Power BI-Report der Gruppe sind fristgerecht in ILIAS hochzuladen (zip-Datei).
- **Präsentation und Diskussion (Detailhinweise s.u.):** 12.01.2023, 08:00-11:30 Uhr

3 Datensatz

Der gegebene Datensatz steht in ILIAS zur Verfügung. Dort ist auf die Website von Kaggle verlinkt, wo auch eine kurze Dokumentation der Daten zu finden ist.

Der Datensatz beschreibt die Ergebnisse aller Spiele aller wichtigen Tennisturniere seit 1968 und beinhaltet neben den Spielergebnissen auch Informationen zu den Rahmenbedingungen (z.B. Untergrund), Metriken zum Spielverlauf (z.B. Anzahl Asse) und zu den Spielern (Weltranglistenpunkte und -Platzierung, Alter, Größe, …)

Der Datensatz verfügt über eine relationale Struktur (Matches, Spieler) sowie über eine zeitliche Struktur (Matches, Weltranglistenplatzierung). Die Daten müssen geeignet transformiert werden, um Zusammenhänge auch über die relationale und zeitliche Struktur hinweg zu untersuchen.

4 Aufgabenstellung (Jupyter & Power BI)

Ihre Aufgabe ist eine zielgruppengerechte Analyse des Datensatzes, d.h. die *Sportliche Leitung des Deutschen Tennisbundes* bzw. die *Trainer der Spieler* sollen anhand Ihrer Auswertung in der Lage sein, Zusammenhänge besser zu durchdringen und Rückschlüsse für die Auswahl bzw. das Training von Spielern zu ziehen.

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.

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, die Sie in Erwägung ziehen können (nicht abschließende Aufzählung!):

- Verknüpfen der Spieler- und der Match-Tabelle
- Änderung der Datenstruktur im Hinblick auf die Aufteilung in „Winner_“- und „loser_“-Spalten
- Anreichern der Tabelle um zeitliche Features, die die Attribute eines Matches um solche erweitern, die relevante Aspekte der Vorgeschichte beinhalten
- Wo Ihnen es sinnvoll erscheint, Umwandlung numerischer in diskrete Attribute (bspw. mittels „pd.cut“), um einen leichten Überblick zu erhalten
- 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.

Erwartet wird, dass Sie die in Data Analytics behandelte **Analysemethoden** beherrschen, und anwenden, wo dies zielführend ist, um aus den Daten Erkenntnisse zu gewinnen. Hinterfragen Sie mögliche Schlussfolgerungen aus Ihren Beobachtungen und überprüfen sowie belegen Sie diese so weit wie möglich. Dokumentieren Sie nicht nur die Schlussfolgerungen, sondern auch die Belege und Argumente, die diese stützen. Bei Ergebnissen, bei denen es nicht offensichtlich ist, weisen Sie auch deren **Signifikanz** nach, oder prüfen Sie auf anderem Weg, ob Sie wirklich einen systematischen Zusammenhang entdeckt haben, oder es sich einfach um Zufall handeln kann.
Beispiele für Analysemethoden:

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

Auch wenn das durch Datenanalyse nur schwer nachzuweisen ist, ist das Ziel hinter den zu prüfenden Hypothesen eigentlich, **kausale Zusammenhänge zu verstehen**. Begnügen Sie sich also nicht damit, nur die zur jeweiligen Hypothese gehörende statistische Abhängigkeit zu untersuchen, sondern **prüfen Sie z.B. auch, ob es naheliegende andere erklärende Variablen gibt**.

Die obigen Fragestellungen sind nicht als präzise mathematische Vorgaben gemeint. **Legen Sie sie fachlich sinnvoll so aus**, dass die Antwort das liefert, was dem Fragesteller vermutlich am meisten hilft. **Denken Sie mit!**

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.

Setzen Sie weiterhin ein Business Dashboard **prototypisch mit PowerBI** um. Wählen Sie dazu zunächst einen der beiden oben genannten Stakeholder als Nutzer dieses Dashboards aus. Definieren Sie stichpunktartig, welchen Informationsbedarf Sie für den Stakeholder durch das Dashboard decken wollen. Notieren Sie den gewählten Stakeholder und die Informationsbedarfs-Definition bitte im Notebook ganz unten in einer Markdown-Zelle. Laden Sie die Rohdaten in Power BI und erzeugen Sie auf maximal 2 Seiten entsprechende Visualisierungen, Tabellen, Filter, passende Kennzahlen und/oder Hierarchien etc. Sie dürfen dabei natürlich auf den Erkenntnissen aus den o.g. Analysen sowie entsprechenden Visualisierungen in Python aufbauen. Achten Sie in diesem Teil bitte stark auf die Übersichtlichkeit!

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“, „Analyse“ 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 Projektaufgabe WS2023/24“ sowie die Matrikelnummern der Gruppenteilnehmer 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).
- Zielgruppen der Präsentation sind die genannten Stakeholder.
- Achten Sie dabei darauf, dass es sich bei der Präsentation um einen „Pitch“ handeln soll und diese somit auf das „Verkaufen Ihrer Ergebnisse“ fokussiert.
- Weitere Details (Zeitrahmen etc.) folgen.