Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

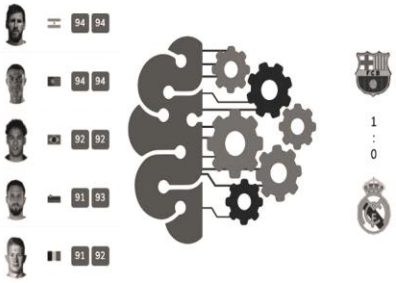
IDP Institut für Datenanalyse
und Prozessdesign

Vorhersage von Fussballresultaten mittels Maschinellen Lernens

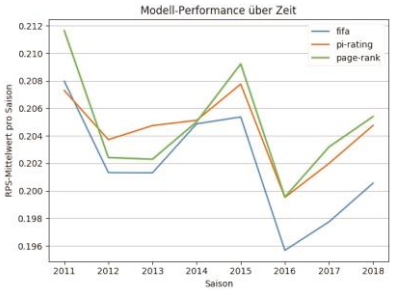
Diplomand
David Ljubas

Dozierende
Helmut Grabner
Martin Frey

Fussball ist nicht zuletzt darum weltweit populär, weil die Ergebnisse unvorhersehbar sind und es immer wieder Überraschungen gibt. Vorhersagen von zukünftigen Spielen sind auch für die Machine-Learning-Community herausfordernd. In einer Machine Learning Challenge 2018 haben verschiedene Teams versucht, basierend auf den Ergebnissen der vergangenen Spiele, Vorhersagen für zukünftige Fussballspiele zu machen. Am besten schnitt dabei ein Gradient Boosted Trees-Modell ab. Als wichtigste Features kristallisierten sich dabei das Pi-Rating und ein Page-Ranking heraus. Das Pi-Rating ist ein dynamisches Bewertungs-system basierend auf erhaltenen und erzielten Toren in den relevanten letzten Spielen. Beim Page-Ranking werden die Ergebnisse gegen andere, unterschiedlich starke Teams ausgewertet. Beide Ansätze machen dabei ihre Vorhersagen unabhängig davon, in welcher Besetzung ein Team antritt. Die Hypothese dieser Arbeit ist, dass sich die Vorhersagen durch den Einbezug der Mannschaftsaufstellung verbessern lassen. Zur Analyse wurden Modelle basierend auf dem Pi-Rating, dem Page-Ranking und dem Aufstellungsfeature verglichen. Zur Modellierung wurde jeweils ein Gradient Boosted Trees-Modell verwendet.
Um das Aufstellungsfeature zu generieren, mussten die Aufstellungsdaten zuerst mit einem eigens dafür entwickelten Webcrawler aus dem Web extrahiert werden. Um die Stärken der aufgestellten Spieler zu bewerten, wurde auf die Fifa-Ratings von EA Sports zurückgegriffen. Auch diese Daten mussten mittels Webcrawler aus dem Web gezogen und an-schliessend mit einem Fuzzy-Ansatz gematcht werden. Zur Modell-Performance-Evaluation wurde die Ranked-Probability-Score-Metrik (RPS) eingesetzt. Diese Metrik erlaubt den Vergleich zweier Wahrscheinlichkeitsverteilungen. Je tiefer der RPS-Wert, umso genauer die Prognosen.
Die analysierten Modelle wurden mit einer 3-fach wiederholten Time Series-Validation ausgewertet. Die Performance von Aufstellungsfeature, Pi-Rating und Page-Ranking ist sehr ähnlich. Betrachtet man die Per-formance aber einzeln über die Jahre hinweg, sieht man, dass das Aufstellungsfeature konstant leicht besser abschneidet (Abbildung 2). Ist man an den korrekten Prognosen interessiert (z.B. für Wetten), erzielt man auch mit dem Aufstellungsfeature mit +/- 7.6 Prozent richtigen Prognosen das beste Ergebnis.



Das vollständige End-to-End-System extrahiert die relevanten Daten aus dem Web und trainiert damit ein Vorhersage-Modell, das sie für die kommenden Spiele nutzt.



Ein Backtest, der die Leistung einzelner Modelle über die Zeit veranschaulicht (Tiefere RPS-Werte entsprechen einer besseren Vorhersagegenauigkeit).