67 Daten- und modellgestützte Talentförderung von Nachwuchsfußballern**

Prof. Jan Mayer und Dr. Sascha Härtel von der TSG Hoffenheim sowie Oliver Wohak von der d-fine GmbH berichten von einem gemeinsamen Projekt, bei dem mathematische Modelle und Machine Learning Methoden auf physiologische und psychologische Leistungsdaten von Nachwuchsfußballern der TSG Hoffenheim angewandt wurden. In dem innovativen Projekt konnten erfolgreich datengestützte Stärken-Schwächen-Profile und Kompensationsmechanismen aufgezeigt, sowie individuelle Entwicklungsprognosen für den einzelnen Spieler entwickelt werden.

Mathematische Modelle unterstützen die individuelle Förderung von Jugendspielern im Fußball

Im Fußball stecken viele Daten. Mittlerweile werden an einem Bundesligaspieltag 3,6 Millionen Datenpunkte pro Spiel erhoben. Bevor ein Spieler es jedoch bis in die höchste deutsche Spielklasse schafft, durchläuft er in der Regel die Jugendakademien oder Nachwuchsleistungszentren. Und auch da – bereits ab den jüngsten Jahrgängen – gehört die Erfassung von Leistungsdaten dazu. Ein Vorreiter in der Nachwuchsleistungsförderung ist die TSG Hoffenheim. Neben Tracking Daten, die zu jedem Training und Spiel aktuelle Performance-Daten aufzeichnen, werden sehr umfangreich physiologische und psychologische Leistungstests durchgeführt. Mit Hilfe dieser interdisziplinären Daten soll ein möglichst ganzheitliches Bild über die einzelnen Spieler und Mannschaften entstehen. Die Bestimmung des Ist-Stands ermöglicht die Einschätzung der Grundvoraussetzungen (z.B. ist ein Spieler eher schnellkräftig oder ausdauerveranlagt; ist er im Kopf und/oder in den Beinen schnell) sowie einen Abgleich mit Soll-Werten (Benchmark-Werte Profis, positionsspezifische Anforderungsprofile).

Das aus den Testungen abgeleitete Stärken-Schwächen-Profil ermöglicht konkrete, individuelle Trainingsmaßnahmen und damit verbunden eine Optimierung der Leistungsfähigkeit. Mit Hilfe von Testwiederholungen kann die langfristige Entwicklung und damit die Wirksamkeit der gewählten Interventionen überprüft werden.

Die Verwertung von Leistungs- und Diagnostikdaten unterstützt die Talentförderung

Verglichen mit anderen Fußballvereinen hat die TSG schon sehr früh mit der strukturierten Datenerfassung begonnen. Somit liegen umfangreiche, mehrjährige Zeitreihen für verschiedene U-Mannschaften zu einem einheitlichen Satz von Kennzahlen vor, die sich sowohl zur explorativen Datenanalyse, als auch für die Entwicklung von Prognosemodellen nutzen lassen. Typische Fragestellungen in diesem Kontext sind: Wie verändern sich bestimmte Kennzahlen im Karriereverlauf? Wie entwickeln sich bestimmte Korrelationen im Zeitverlauf? Welche Prognosen lassen sich bzgl. des Übergangs in

den U-Mannschaften treffen? Alles mit dem Ziel Muster in den Daten zu erkennen, welche die individuelle Förderung der Spieler unterstützen und das Leistungspotential heben.

Der Blick zurück – erste explorative Datenanalysen zeigen Zusammenhänge auf

Einfachere Muster lassen sich mittels explorativer Analysen in den Daten erkennen. Dies sind zum Beispiel die zeitliche Entwicklung einzelner Attribute (univariate Analysen) oder auch Korrelationen verschiedener Attribute (multivariate Analysen). Korrelationsanalysen sind insbesondere dann hilfreich, wenn die „Ähnlichkeit“ von historischen Zeitreihen einzelner Attribute bewertet werden soll. Dabei gibt die Berechnung der Korrelation Erkenntnis über die Stärke des Zusammenhangs der Attribute sowie der Richtung: -1 ist maximal stark negativ korreliert und +1 maximal stark positiv korreliert. In Abbildung 96 sind die Korrelationswerte verschiedener Attribute aus den Leistungstests dargestellt.

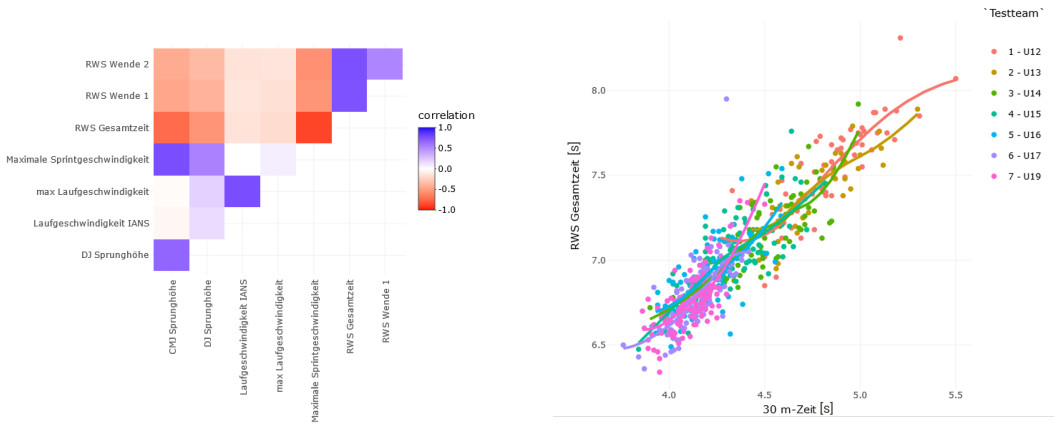

Abbildung 96: Korrelationsplot zu Schnelligkeits-, Ausdauer und Sprungkraftwerten (links) und bivariate Analyse zu Richtungswechselsprint (RWS) und 30 m-Sprintzeit rechts.

Zum Beispiel ist dort zu erkennen, dass die Attribute der Sprungkraft (konkret hier die verschiedenen Angaben zur Sprunghöhe DJ Sprunghöhe und CMJ Sprunghöhe) sehr mit den Sprintwerten korrelieren -– stark negativ mit den Richtungswechselsprintzeiten (RWS) und stark positiv mit der maximalen Sprintgeschwindigkeit. Daraus lässt sich ableiten, dass Spieler mit guten Schnelligkeitswerten typischerweise auch eine starke Sprungkraft haben.

Neben den Korrelationsplots können bivariate Analysen helfen, um Zusammenhänge zwischen zwei bestimmten Attributen näher zu untersuchen. In dem Punktwolkendiagramm in Abbildung 96 sind exemplarisch die beiden Attribute RWS Gesamtzeit und 30 m-Zeit gegenübergestellt. Durch die inhaltliche Nähe der beiden Attribute lässt sich

in der Analyse ein starker Zusammenhang erkennen. Spieler mit einer schnellen 30 m-Zeit absolvieren typischerweise auch den Richtungswechselsprint in einer kurzen Zeit. Allerdings ist ebenfalls zu erkennen, dass dieser Zusammenhang nicht perfekt ist. So kann im Einzelfall ein Spieler eine gute 30 m-Zeit aufweisen, während die Leistung beim Richtungswechseltest ausbaufähig ist. Zusammengefasst scheinen die Attribute 30 m-Zeit und RWS Gesamtzeit stark verwandt zu sein, jedoch noch spezifisch genug, um Stärken und Schwächen im jeweiligen Attribut getrennt identifizieren zu können. Darüber hinaus lässt sich aus den U-Mannschafts-spezifischen Trendlinien noch erkennen, dass die sich die Sprintzeiten insbesondere im Bereich der U12 bis U14 auf Grund der körperlichen Entwicklung von den anderen Mannschaften abheben und, dass die Unterschiede zwischen den Mannschaften unabhängig vom Alter von der U15 aufwärts sehr gering sind.

Der Blick in die Zukunft – Machine Learning Modelle ermöglichen Prognosen zur Spielerentwicklung

Neben den explorativen Datenanalysen ist es interessant zu untersuchen, inwiefern Prognosemodelle helfen können, Aussagen über zukünftige Entwicklungen der Spieler zu treffen. Während explorative Analysen die vorliegenden Daten lediglich beschreiben und nachträglich Zusammenhänge aufzeigen, ermöglichen Prognosemodelle eine Vorhersage der Spielerentwicklungen. In unserem Projekt erfolgte dies an Hand der Prognose der Übergangswahrscheinlichkeit in die nächst höhere U-Mannschaft, die für jeden Spieler ermittelt wurde. Dies ermöglicht somit das frühzeitige Eingreifen von Athletiktrainern oder Sportpsychologen, um die Weiterentwicklung der Spieler optimal zu unterstützen.

Hierfür bieten sich etablierte Machine Learning Methoden an -– also Mathematik — um sowohl lineare Abhängigkeiten (z.B. mittels linearer oder logistischer Regression) als auch komplexere Zusammenhänge (z.B. mittels Baum-basierten Modellen oder neuronalen Netzen) in den Daten zu erkennen und zu verwerten. Diese Methoden haben gemeinsam, dass in einem Prozess – der auch überwachtes Lernen (engl. „supervised learning“) genannt wird – in historischen Daten Gesetzmäßigkeiten erkannt werden, um basierend auf einem Satz an Eingabewerten (Modell Features) zugehörige Ergebnisse oder Ereignisse (Modell Label) möglichst gut und zuverlässig zu prognostizieren. In diesem Projekt wurde ein so genanntes Random Forest Modell (eine Kombination aus mehreren Baum-basierten Modellen) trainiert, um ausgehend von den physiologischen und psychologischen Testergebnissen der Spieler eine U-Mannschafts-spezifische Übergangswahrscheinlichkeit zu prognostizieren. Die Features im Modell sind somit z.B. die körperliche Aktivität, die Laufgeschwindigkeit an der individuellen anaeroben Schwelle (Grundlagenausdauer) oder auch der Geburtsmonat (spät im Jahr geborene Kinder fallen häufig aus der Förderung, da sie meist gegenüber den früh im Jahr geborenen weniger weit entwickelt sind), während der Label der (nicht) erfolgreiche Übergang in die nächsthöhere U-Mannschaft am Ende der Saison ist.

Das Modell wird zuerst mit den historischen Daten trainiert und anschließend validiert.

Dafür werden typischerweise die vorhandenen Daten in einen Trainingssatz (z.B. 80 % der Daten) und einen Validierungssatz (die restlichen 20 %) getrennt. Mithilfe des Validierungssatzes wird die Modellgüte getestet, zum Beispiel über die Anzahl der Richtig-Positiven und Falsch-Positiven Vorhersagen. Dabei spricht man dann von der ROC-Kurve (Receiver-Operating-Characteristics oder auf Deutsch Grenzwertoptimierungskurve) oder auch einem Gini-Koeffizient (der die Fläche unter der Kurve auf Werte zwischen 0 und 1 normalisiert). Es lässt sich somit mit einer gewissen Zuverlässigkeit vorhersagen, dass ein Mittelfeldspieler der U17 mit 16 Jahren, einer körperlichen Aktivität von 1 (Maximum), einer Laufgeschwindigkeit IANS (12,8 km/h) sowie einigen weiteren Attributen (Abbildung 97) mit einer Wahrscheinlichkeit von ca. 76 % den Sprung in die U19 schafft.

Abbildung 97: Darstellung der Attribute eines Spielers der U17 mit einer hohen Übergangswahrscheinlichkeit.

Modelle erklärbar zu machen liefert zusätzliche Erkenntnisse

Als Ergänzung zu den Expertenmeinungen der Trainer ist es hilfreich, einzelnen Spielern eine Übergangswahrscheinlichkeit zuordnen zu können und somit datengestützt Spieler mit einem hohen Potential bereits frühzeitig identifizieren zu können. Um darüber hinaus jedoch jeden Spieler individuell bestmöglich fördern zu können, müssen wir zusätzlich wissen, welche Eigenschaften bei den Spielern mit hohem Potential stark ausgeprägt sind und welche körperlichen oder mentalen Eigenschaften bei „schwächeren“ Spielern insbesondere noch trainiert werden sollten. Standardmäßig können Machine Learning Modelle dazu keine direkte Aussage treffen. Es gibt jedoch mathematische Verfahren, die es ermöglichen, den Anteil jedes Attributs an dem Prognoseergebnis zu quantifizieren. Dies wird dann auch „explainable AI“ (xAI) oder erklärbare künstliche Intelligenz genannt. Eines dieser Verfahren – das auch in diesem Projekt angewandt wurde – ist SHAP (Shapley Additive exPlanations). Es baut auf einer Idee auf, die ursprünglich aus der Spieltheorie kommt und versucht den marginalen Beitrag jedes Spielers in einem kooperativen Spiel mit messbarem Ausgang zu quantifizieren. Angewandt auf Machine Learning Methoden quantifiziert SHAP den marginalen Beitrag jedes Modell Features auf die Modellprognose. Dies erreicht man, indem man die

Ergebnisse vieler ähnlicher Modelle vergleicht. Diese Modelle unterscheiden sich nicht in ihrer Parametrisierung, decken jedoch alle möglichen Kombinationen an Feature-Zusammensetzungen ab. Durch den Vergleich der jeweiligen Modellprognosen kann zurückgerechnet werden, welchen Beitrag jedes einzelne Feature auf die Prognose des vollständigen Modells hat. Insbesondere ist es in unserem Modell somit möglich aufzuzeigen, welche Features die Übergangswahrscheinlichkeit bei einem bestimmten Spieler – im Vergleich zu dem Durchschnittsspieler – eher treiben bzw. drücken. Dies ist durch den „Force-Plot“ in Abbildung 98 dargestellt an dem erkennbar ist, dass der dargestellte Spieler insbesondere durch seine körperliche Aktivität und das gute Abschneiden im Determinationstest (Kognitionstest) heraussticht. Eine Eigenschaft, an der der Spieler noch arbeiten kann, um sich weiter zu verbessern, und wo er schlechter abschneidet als der Durchschnitt, ist seine Grundlagenausdauer – hier dargestellt als Laufgeschwindigkeit an der individuellen anaeroben Schwelle (IANS).

Abbildung 98: Force-Plot der anzeigt welche Attribute den dargestellten Spieler im Vergleich zu einem Durchschnittspieler (Basiswert) auszeichnen.

Aus dem xAI-Ansatz und der Darstellung mittels „Force-Plots“ lassen sich somit daten- und modellgestützte Stärken-Schwächen Profile für die Spieler ableiten – auch nochmal dargestellt in Abbildung 99. Darüber hinaus lassen sich weitere interessante Erkenntnisse ableiten, die von den Leistungsdiagnostikern und Sportpsychologen verwertet werden können. So wurden in dem Projekt zum Beispiel Kompensationseffekte analysiert. Diese geben an, ob die Stärken eines Spielers eher physiologischer oder psychologischer Art sind. Bei dem oben dargestellten Spieler der U17 beträgt das Verhältnis physiologischer zu psychologischer Stärken 28:72. Das bedeutet, dass der Spieler ausgeprägte mentale Fähigkeiten hat und seine athletischen Defizite darüber ausgleicht. Insbesondere lässt sich über die verschiedenen Jugendmannschaften hinweg erkennen, dass physiologische Eigenschaften in jüngeren Jahren für die Förderung wichtiger sind, im Alter für den Sprung in den Profi-Kader die psychologischen Kompetenzen jedoch umso wichtiger werden.

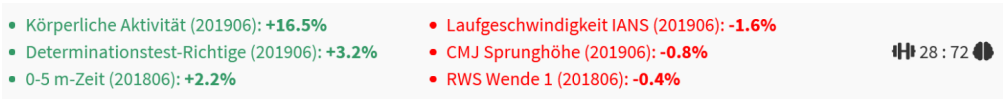

Abbildung 99: Das Stärken-Schwächen-Profil und der Kompensationseffekt (Hantel = physiologisch, Gehirn = psychologisch).

Daten und Modelle werden auch im Fußball immer wichtiger

Insgesamt hat das Projekt gezeigt, dass mathematische Methoden und Modelle helfen können vorhandene Daten im Fußball zu verwerten. Anhand historischer Daten zu Jugendspielern der TSG Hoffenheim wurde ein Machine Learning Modell implementiert, das die Übergangswahrscheinlichkeiten für einzelne Spieler ableitet und darüber hinaus die Definition von Stärken-Schwächen-Profilen sowie Kompensationseffekten ermöglicht. Weitergedacht können diese Art Daten im Fußball auch genutzt werden, um z.B. Verletzungsanfälligkeit oder Überbelastung zu identifizieren – zum Beispiel auch in Kombination mit den Trackingdaten aus Training und Spiel.