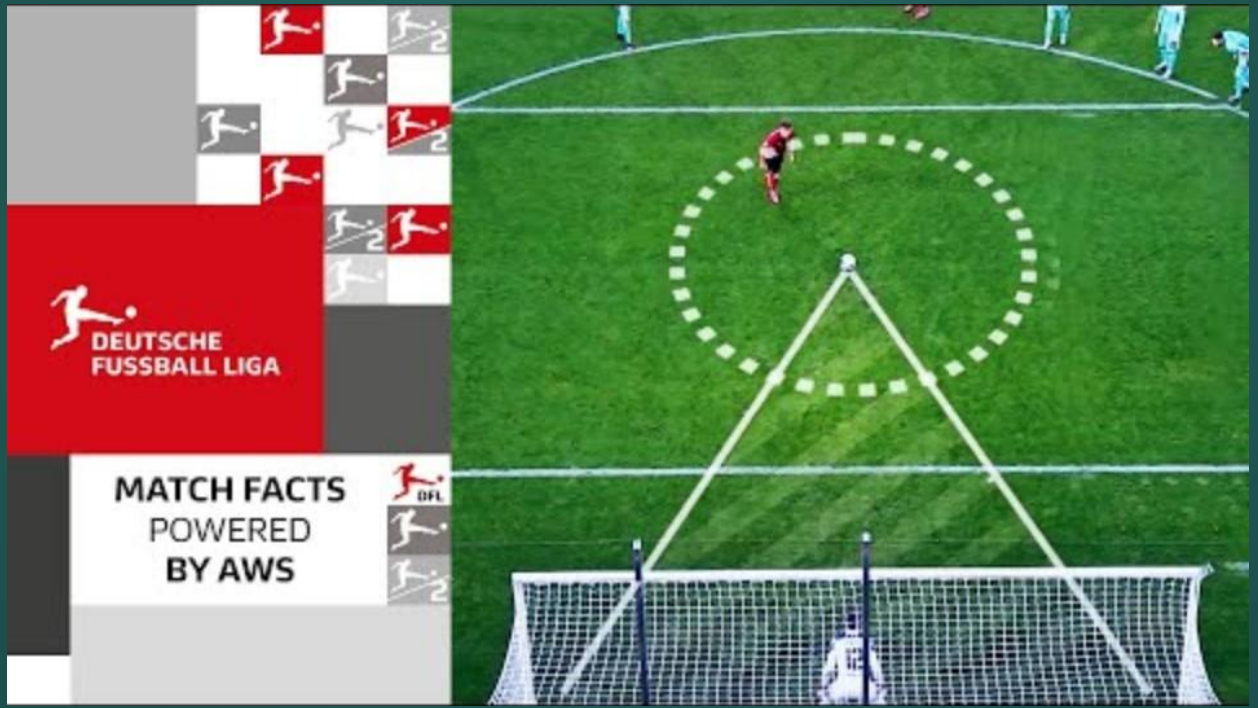
A MACHINE LEARNING APPROACH FOR PLAYER AND POSITION ADJUSTED EXPECTED GOALS

Autoren: James H. Hewitt, Oktay Karakuş (2023)
Seminar: Advanced Topics in Data Analysis and Deep Learning
Bearbeiter: Felix Narr

WIE VIELE TORE SIND AM 33. SPIELTAG IN DER BUNDESLIGA GEFALLEN?

WIE VIELE XG LIEFERTEN DIE SPIELE IN DER SUMME?

EINFÜHRENDES VIDEO (DFL)



GLIEDERUNG

1. Allgemeines
2. Entwicklung des xG-Modells
3. Ergebnisse
4. Code-Demo

1. ALLGEMEINES

WISSENSCHAFTLICHE PERSPEKTIVEN AUF XG

Lucey et al. (2015): Einführung von Spatio-temporalen Mustern zur Analyse von Schussqualität. Erkenntnis: Verteidigerpositionen müssen in xG integriert werden

Tippet (2018): "The Expected Goals Philosophy": Relevanz und Entwicklung von xG. Fehlende Verteidigerpositionierung, Angriffe ohne Torschuss fehlen

Fairchild et al. (2018): Kombination von Fraktalgeometrie und Machine Learning zur besseren Berechnung von xG → räumlich-zeitliche Bewegungen der Spieler entscheidend

Singh (2018) + Decroos et al. (2019): Einführung von expected threat (xT) und VAEP (Valuing Actions by Estimating Probabilities) → Parallel zur xG-Entwicklung

Tureen & Olthof (2022): Erstellung des „Estimated Player Impact" (EPI)-Maßes → Einfluss einzelner Spieler auf jede xG-Bewertung (Unterteilung in Positionsgruppen)

Cavus & Biecek (2022): Einführung von Explainable AI in xG-Modelle zur besseren Interpretierbarkeit.

Madrero (2020): Distanz zum Tor als wichtigstes Kriterium. Einführung eines Teambasierten xG-Modells → Logistische Regression, XGBoost and Neuronale Netze

Brechot & Flepp (2020): Nachweis, dass xG ein besseres Maß für (langfristige) Teamleistung ist als reine Tore (kurzfristig) → logistische Regression

RESULTAT

- **Viele verschiedene xG-Modelle**
- **Unterschiedliche statistische Verfahren**: Logistische Regression, Decision Trees, Random Forest, Ada-Boost, …
- **Unterschiedliche Features**: Verteidigerposition, Schusswinkel, Distanz zum Tor …

ZIELE DES PAPERS

- Entwicklung eines eigenen, zuverlässigen xG-Modells mit neuen Features
- Vergleich der Vorhersagegenauigkeit mit bestehenden Modellen (StatsBomb)
- Einführung von Positions- und Spieler-angepasstem xG zur besseren Bewertung von Schussqualitäten

DATENSATZ

Variables	Type	Values
Aerial shot	*Binary*	True & False
1st time	*Binary*	True & False
Open Goal	*Binary*	True & False
Pressure	*Binary*	True & False
1v1	*Binary*	True & False
Shot Technique	*Categorical*	Backheel, Diving Header, Half Volley, Lob, Normal, Overhead Kick, Volley
Pass From	*Categorical*	Corner, Counterattack, Free Kick, Goal Kick, Keeper, Kick off, Throw In, Regular Play
Shot Body Part	*Categorical*	Header, Left Foot, Right Foot, Other
Location	*Coordinates*	(x, y) coordinates — 120 yard ×80 yards
Goal	*Binary*	True & False (Target)

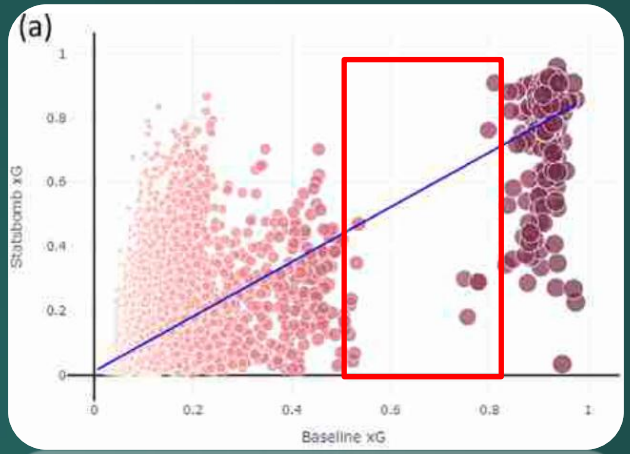
Player	Aerial shot	1st time	1v1	Open goal	Press.	Shot tech.	Pass from	Body part	Goal
Diego Garcia	1	0	0	0	1	Normal	Free Kick	Head	1
Antoine Griezmann	0	0	0	0	0	Normal	Regular Play	L. Foot	0
Alvaro Negredo	0	0	0	0	1	Normal	Corner	L. Foot	0
Lionel Messi	0	0	0	0	0	Normal	Free Kick	L. Foot	1
Xavier Hernandez	0	1	0	0	0	Normal	Regular Play	R. Foot	0
Gerard Pique	0	0	0	0	0	Normal	Corner	Head	0
Lionel Messi	0	1	0	0	0	Normal	Free Kick	L. Foot	1
Jordi Alba	0	1	0	0	0	Half Volley	Free Kick	L. Foot	0
Martin Odegaard	0	0	0	0	1	Normal	Regular Play	L. Foot	0
Lionel Messi	0	0	0	0	0	Normal	Free Kick	L. Foot	0

- Nur Schüsse aus dem Spiel, keine Elfmeter
- 15.575 Datensätze mit ursprünglich 95 Merkmalen
- Datenaufbereitung:
  - Schussinformationen mit fehlenden Werten entfernt
  - Numerische Merkmale skaliert
  - Kategoriale Merkmale mit Label-Encoding versehen
  - Boolesche Merkmale mit One-Hot-Encoding kodiert
  - Koordinaten in X und Y getrennt

2. ENTWICKLUNG DES XG-MODELLS

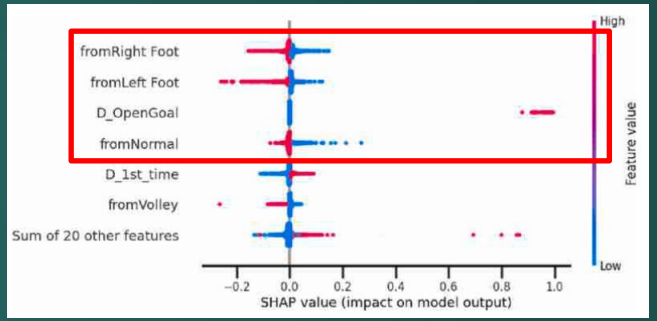
ENTWICKLUNG DES MODELLS (1)

- **Baseline-Model**: Basis für nachfolgende Verbesserungen
- **Logistische Regression** als statistisches Verfahren
- **Vergleich mit Branchenbenchmark StatsBomb** (Abb.)
  - Angepasste Regressionsgerade: $y = 0{,}01 + 0{,}85x$
  - Korrelation von **0,659**
  - Großer leerer Bereich ohne xG-Werte
  - Tendenzielle Unterschätzung der xG-Werte (Werte links)
- Vergleich der **akkumulierten Tore** mit den **akkumulierten xG**:
  - **Ziel:** Hohe Korrelation
  - **Ergebnis:** 1887 vs. 1866

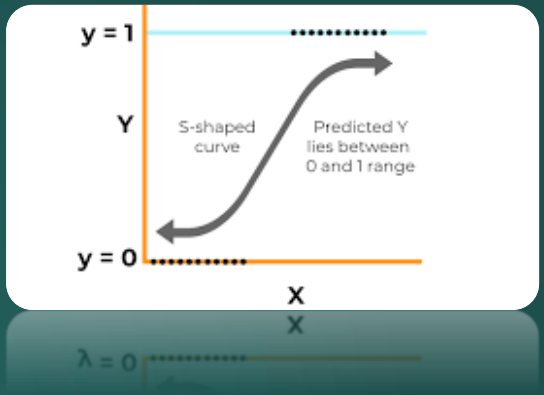


ENTWICKLUNG DES MODELLS (1)

- **SHAP** (SHapley Additive exPlanations) erklärt den Einfluss einzelner Merkmale auf eine Modellvorhersage → basiert auf Spieltheorie, berechnet den durchschnittlichen Beitrag jedes Features
- **Wichtigste Merkmale:** z. B. Open Goal
- SHAP-Werte insgesamt **gering** → Hinweis auf zu wenig Informationsgehalt im Modell
- **Fazit:** Weitere, aussagekräftigere Features nötig, um xG besser vorherzusagen



LOGISTISCHE REGRESSION

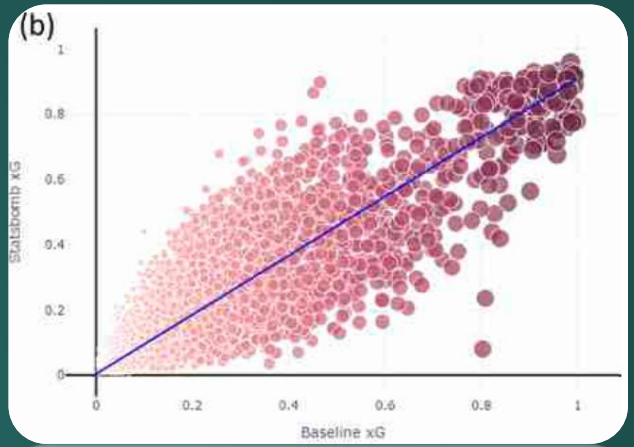


- **Zweck:** Modelliert die Wahrscheinlichkeit eines binären Ereignisses (z. B. Tor oder kein Tor).
- **Funktionsweise:** Verwendet eine S-Kurve (logistische Funktion), um eine lineare Kombination von Eingabewerten in eine Wahrscheinlichkeit (zwischen 0 und 1) zu verwandeln.
- **Formel:** $P\,(y{=}1) = \frac{1}{1+e^{-z}}$ → mit $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$
- **Ergebnis:** Zahl zwischen 0 und 1 → interpretiert als Wahrscheinlichkeit
- **Vorteil:** Einfach zu interpretieren und liefert Wahrscheinlichkeiten für die Entscheidungsfindung

- Distanz zum Tor: 12 Meter
- Winkel zum Tor: 30 Grad
- Mit dem Fuß geschossen: Ja = 1
- Fiktive Gewichtung der Parameter:
  - $z = \beta_0 + \beta_1 * \text{Distanz} + \beta_2 * \text{Winkel} + \beta_3 * \text{Fuß}$
  - $z = -1.5 + (-0.1*12) + (0.03*30) + (0.5*1)$
  - $z = -1.3$ (in Formel einsetzen)
- xG = $\frac{1}{1+e^{-(-1.3)}}$ → **0,214**

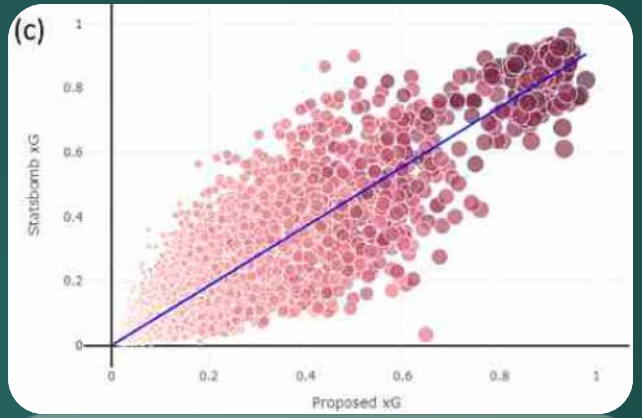
ENTWICKLUNG DES MODELLS (2)

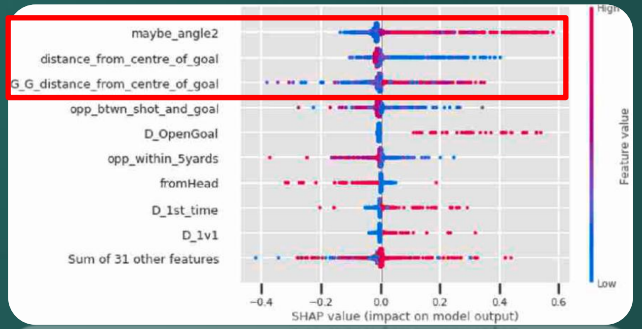
- **Überarbeitetes-Modell** → Erweiterung des Feature-Sets (insg. 40 Merkmale), u.a. Position des Torwarts, Druckradius gegnerischer Spieler, Gegnerische Spieler zwischen Ball & Tor
- **Logistische Regression** als statistisches Verfahren
- **Vergleich mit Branchenbenchmark StatsBomb** (Abb.)
  - angepasste Regressionsgerade: $y = 0.00 + 0.90x$
  - Korrelation von **0.887**
- Vergleich der **akkumulierten Tore** mit den **akkumulierten xG**:
  - **Ziel:** Hohe Korrelation zwischen den akkumulierten xG und den akkumulierten Toren
  - **Ergebnis:** 1887 vs. 1866
- **Fazit:** Stärkerer Zusammenhang mit Branchenstandard und realistischere xG-Verteilung



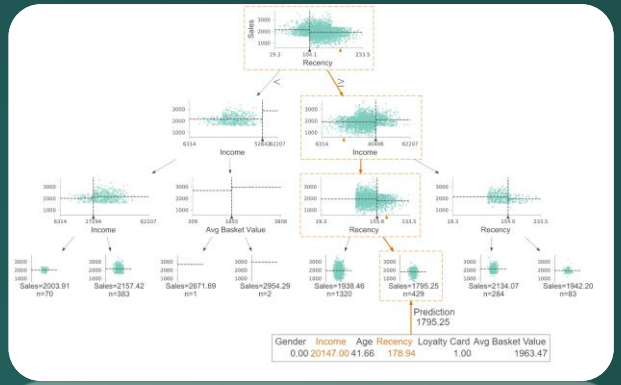
ENTWICKLUNG DES MODELLS (3)

- **Überarbeitetes-Modell**
- **ML-Entscheidungsbäume** als statistisches Verfahren i.V.m. **Gradient Boosting**
- **Vergleich mit Branchenbenchmark StatsBomb** (Abb.)
  - angepasste Regressionsgerade: $y = 0.00 + 0.92x$
  - Korrelation von 0.902
- Vergleich der **akkumulierten Tore** mit den **akkumulierten xG**:
  - **Ziel:** Hohe Korrelation zwischen den akkumulierten xG und den akkumulierten Toren
  - **Ergebnis:** 1887 vs. 1870
- SHAP mit deutlich höheren Werten, signifikante, neue Variablen
- **Fazit:** Noch besseres Modell





ML-ENTSCHEIDUNSBÄUME & GRADIENT BOOSTING



- Was sind **Entscheidungsbäume**?
  - Supervised Learning für Klassifikation & Regression
  - Entscheidungen anhand einfacher "Wenn-Dann"-Regeln
- Warum **Gradient (Tree) Boosting**?
  - Kombiniert viele schwache Modelle zu einem starken Gesamtmodell
  - Korrigiert Fehler schrittweise durch Gradient Descent
- Drei zentrale Bausteine:
  - Loss Function – misst den Vorhersagefehler
  - Weak Learner – meist flache Entscheidungsbäume
  - Additives Model – Fehler werden iterativ reduziert

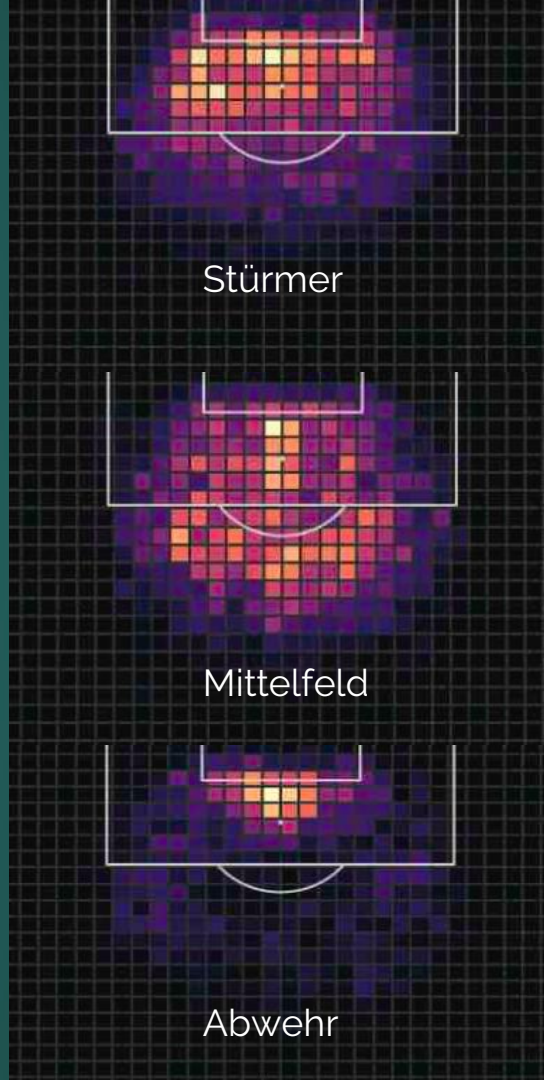
3. ERGEBNISSE

ERGEBNISSE (1) – POSITIONSANGEPASSTER XG

- **Stürmer**: 1276 Tore bei niedrigeren xG-Werten → Überperformance, hohe Abschlussqualität → Schüsse von Stürmern höher bewerten?!
- **Mittelfeldspieler**: 411 xG, 398 Tore → Unterperformance, schwächere Chancenverwertung
- **Verteidiger**: xG stets < tatsächliche Tore → Überperformance, effizient trotz seltener Chancen
- **Fazit**: Unterschiedliche Spielerrollen zeigen deutliche Abweichungen zwischen erwartetem und tatsächlichem Torerfolg – wichtig für Modellkalibrierung & Spielerbewertung → Daten individuell trainieren

Positional analysis of the data set.

Position	Total shots	Total goals	Shot/Goal	Baseline xG	Statsbomb xG	Proposed xG
Forward	8646	1276	6.77	1265.767	1154.624	1252.802
Midfield	4590	398	11.53	399.3236	391.2747	411.0013
Defender	2336	213	10.97	200.9173	205.6996	206.8463
Goalkeeper	2	0	0	0.096735	0.107875	0.105983

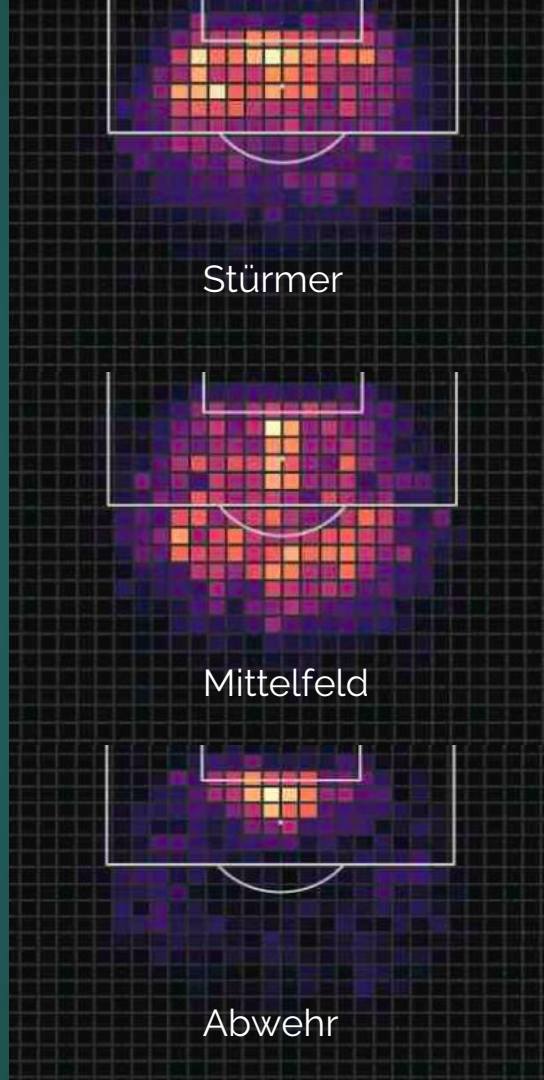


ERGEBNISSE (1) – POSITIONSANGEPASSTER XG

- Anpassung des Trainingsdatensatzes (3x)
- Modell wird für jede Position trainiert
  - Forward Adjusted xG
  - Midfield Adjusted xG
  - Defender Adjusted xG
- Ziel: Empirische Belege liefern, dass jede Position unterschiedliche Effizienzlevel über das gesamte Datenspektrum hinweg aufweist.

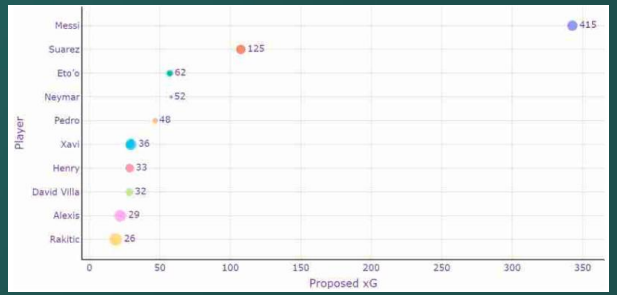
Positional Adjusted xG Values.

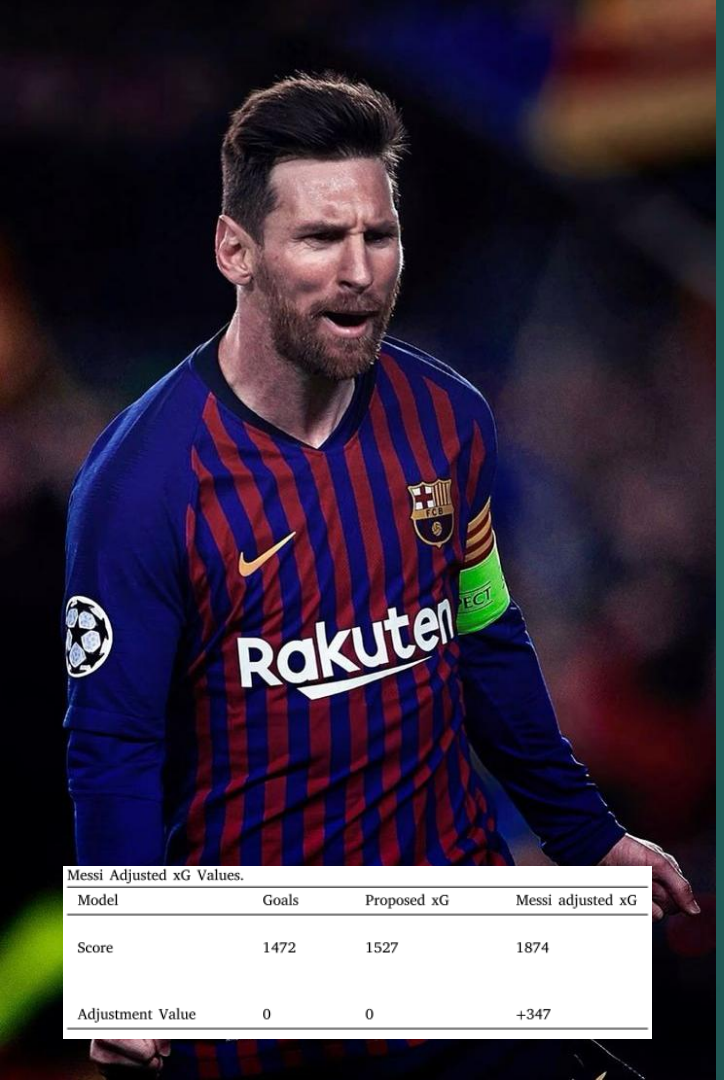
Model	xG	Forward xG	Midfield xG	Defender xG	Goals
Goals/xG	1870	1956	1728	1397	1887
Adjustment Value	0	+86	−142	−473	0



ERGEBNISSE (2) – SPIELERANGEPASSTER XG

- *„Wie viel besser performen Top-Stürmer wie Lionel Messi im Vergleich zu normalen Spielern und dem durchschnittlichen xG?“*
- Top-Stürmer: In der Regel Tore > xG
  - Messi: 415 Tore aus 342 xG
  - Suarez: 125 Tore aus 107 xG





Messi Adjusted xG Values.

Model	Goals	Proposed xG	Messi adjusted xG
Score	1472	1527	1874
Adjustment Value	0	0	+347

ERGEBNISSE (2) – SPIELERANGEPASSTER XG

- 2300 Messi-Schüsse als Trainingsdaten
- Vergleich der xG-Werte mit restlichem Datensatz (Nicht-Messi-Schüsse)

Player-specific results for Messi Adjusted xG.

Player	Goals	Proposed xG	Statsbomb xG	Messi adjusted xG	Over-performance?
Suarez	125	107.4513	108.2466	126.2198	No
Eto'o	62	57.1248	52.71371	63.02088	No
Neymar	52	58.12527	56.02049	66.74318	No
Pedro	48	46.75322	43.47725	53.60634	No
Xavi	36	29.35869	28.75191	41.78057	No
Henry	33	28.5776	28.18925	32.03867	**Yes**
David Villa	32	28.3766	29.0367	32.61256	No
Alexis	29	21.9309	21.33671	25.03286	**Yes**
Rakitic	26	18.57919	18.45674	26.59705	No

ERGEBNISSE (3) – BENCHMARKING

Game statistics for the 2018 Real Madrid vs Liverpool, Champions League final.

Team	Goals	Shots	Statsbomb	FBRef	infogol	Proposed
Liverpool	1	14	1.31442	**1.9**	**1.88**	1.61353
R. Madrid	**3**	14	**1.367858**	1.5	1.71	**1.816377**

- Anwendung des xG-Modells auf das Champions-League-Finalspiel Real Madrid vs. Liverpool (2017/18)
- Vergleich der kumulierten xG-Werte:
  - StatsBomb: 2,7 xG
  - FBRef: 3,4 xG
  - Infogol: 3,59 xG
  - Vorgeschlagenes Modell: 3,4 xG
- **Fazit**: Das Modell zeigt valide Ergebnisse im Vergleich zu anderen Quellen, Fallrückzieher von Bale ggf. etwas überbewertet

2018 Real Madrid vs Liverpool, Champions League final game goals.

Team	Player	Shot technique	Minute	Proposed xG	Statsbomb xG
R. Madrid	Karim Benzema	Volley	50	0.351569	0.517137
Liverpool	Sadio Mane	Volley	54	0.599426	0.548516
R. Madrid	Gareth Bale	Overhead Kick	63	0.131235	0.022605
R. Madrid	Gareth Bale	Normal	82	0.020848	0.013965

4. CODE-DEMO

FAZIT, LIMITATIONEN & AUSBLICK

Fazit

- Neues xG-Modell mit innovativen Features (z.B. Torwart-Position, Spieler-Druckradius)
- Positionsadaptiertes und Messi-adaptiertes xG liefern präzise Vorhersagen

Limitationen

- Begrenzte Daten und Ressourcen im Vergleich zu großen Anbietern
- Mögliche Überbewertung schwieriger Schüsse (z.B. Fallrückzieher).

Ausblick

- Weitere Anpassungen für verschiedene Ligen und Wettbewerbe
- Nutzung zusätzlicher Datenquellen wie StatsBomb 360 zur Modellverbesserung

"DAS RUNDE MUSS INS ECKIGE.„ SEPP HERBERGER