

CASE 2026

STATISTIČKA ANALIZA DETERMINANTI UČINKA U NOGOMETU

Istraživanje korelacija i važnosti značajki primjenom strojnog učenja

Nikola Jakov Pavečić · Denis Romić · Bruno Polonijo

Sezona 2024./2025. | 1.982 igrača | Pet europskih liga

Uvod i Motivacija

Zašto statistička analiza u modernom nogometu?

Izazovi modernog nogometa

Sve veća taktička fluidnost i hibridne pozicije otežavaju tradicionalnu evaluaciju igrača

Subjektivna procjena skauta nije dovoljna u modernom profesionalnom nogometu

Pojava naprednih metrika (xG, progresivna nošenja) pruža novu dimenziju analize

Taksonomska dvosmislenost pozicija postaje statistički mjerljiv fenomen

Ciljevi istraživanja

01

Segmentirati igrače u homogene grupe primjenom K-means algoritma

02

Klasificirati primarnu poziciju igrača na temelju statističkih profila

03

Predvidjeti učinak igrača (gol/90 min) regresijskim modelima

Pregled Srodnih Radova

Od Moneyball-a do naprednih metrika i strojnog učenja u nogometu



Moneyball – Lewis

2003

Popularizacija kvantitativnog skautinga u baseballu. Pokrenula revoluciju u sportskoj analitici – data-driven pristup kao alternativa intuiciji.



Expected Goals (xG)

2014

Metrika koja mjeri vjerojatnost gola po udarca ovisno o poziciji, kutu i kontekstu. Postala standardni alat za procjenu napadačke učinkovitosti.



Bialkowski et al.

2014

Demonstrirali primjenu klasteriranja za analizu timskih formacija – prvi dokaz da ML može prepoznati taktičke uzorke iz statističkih podataka.



Pappalardo et al.

2019

Predviđanje učinka igrača nadgledanim metodama na temelju prostorno-vremenskih događajnih skupova podataka – referentna točka za metodologiju ovog rada.



Merlin et al.

2022

Statistički profili igrača u ligama visokog tempa ne odgovaraju uvijek nominalnim pozicijama – direktna motivacija za naš pristup klasifikacijom prema učinku.

Metodologija

CRISP-DM okvir | Podaci | Algoritmi

CRISP-DM

- 1 Razumijevanje problema
- 2 Razumijevanje podataka
- 3 Priprema podataka
- 4 Modeliranje
- 5 Evaluacija
- 6 Primjena

Skup podataka

- FBref.com / StatsBomb / Understat
2.854 igrača → 1.982 (≥ 450 min)
267 varijabli po igraču
5 liga, sezona 2024./2025.

Algoritmi

- K-means klasteriranje (K=4)
Decision Tree & Random Forest
Linear & RF Regresija
Python 3.10 | scikit-learn | pandas

Inženjering značajki

Normalizacija na 90 minuta:
goals_per_90, xG_per_90,
assists_per_90, shots_per_90

StandardScaler za K-means

Podjela podataka 80:20
(train/test)

5-struka unakrsna provjera
(5-fold cross-validation)

Ključne Metodološke Odluke

Zašto baš ovako? Obrazloženje odabira algoritama i parametara

Zašto K=4 klastera?

Metoda lakta pokazala je izravnjanje krivulje WCSS pri K=4. Silhouette koeficijent (0,271) potvrđuje prihvatljivu separaciju. Viši K rastakao bi klastere bez semantičkog smisla.

Zašto Random Forest, ne SVM ili Neural Network?

RF je robustan na prekomjerno prilagođavanje i ne zahtijeva skaliranje. Pruža interpretabilnost (feature importance) koja je ključna za praktičnu primjenu u skautingu. NN bi bio preteška arhitektura za 1.982 uzorka.

Zašto prag od 450 minuta?

Ekvivalent 5 punih utakmica osigurava statističku stabilnost naprednih metrika poput xG i PrgC. Igrači s manje minuta imaju ekstremne per-90 vrijednosti koje iskrivljuju modele.

Zašto class_weight='balanced'?

Skup je nebalansiran: daleko manje golmana nego braniča ili veznjaka. Bez balansiranja model ignorira manjinsku klasu. Ovaj parametar automatski težinski izjednačuje klase.

Skup podataka – Opisna statistika

n = 1.982 igrača | 5 europskih liga | Sezona 2024./2025.

Varijabla	Srednja vrijednost	Std. devijacija	Medijan	Maksimum
Minute igranja (90s)	15,2	8,9	13,5	41,8
Gol/90 min	0,12	0,19	0,05	1,43
xG/90 min	0,10	0,14	0,05	0,95
Asistencije/90 min	0,08	0,11	0,04	0,73
PrgC (Progr. nošenja)	3,21	3,44	2,20	22,4

Korelacija xG ↔ Stvarni golovi



Ključni nalaz EDA:

Distribucija ofenzivnih metrika asimetrična udesno – većina igrača su braniči i veznjaci koji rijetko postižu golove.

Rezultati – K-means Klasteriranje

Optimalni broj klastera: K=4 | Silhouette koeficijent = 0,271

Klaster 0	Klaster 1	Klaster 2	Klaster 3
Ofenzivni igrači Visoki xG/90 i goals/90; napadači i napadački krilni igrači xG/90: Visoki Gol/90: Visoki PrgC: Visoki Def. akcije: Niski	Defenzivni igrači Niski napadački stat; braniči i defenzivni vezni igrači xG/90: Niski Gol/90: Niski PrgC: Niski Def. akcije: Visoki	Vezni igrači (balansirani) Umjerene vrijednosti u svim kategorijama xG/90: Umjereni Gol/90: Niski PrgC: Umjereni Def. akcije: Umjereni	Kreativni vez./napadači Visok volumen progresivnih akcija; kreativni veznjaci i napadački bekovi xG/90: Umjereni Gol/90: Umjereni PrgC: Visoki Def. akcije: Niski

* PCA projekcija pokazuje umjerenu separaciju klastera što je očekivano zbog hibridnih pozicija modernog nogometa.

Rezultati – Klasifikacija Pozicija

Random Forest vs. Decision Tree | Točnost modela: 85,55%

85,55%

ukupna točnost
Random Forest modela

~75%

točnost
Decision Tree modela

⚠ Najčešće pogreške:

18× MF klasificiran kao DF | 5× MF klasificiran kao FW
→ Odražava statističku hibridnost modernih veznih igrača ("branič-veznjak")

Evaluacijske metrike po poziciji

Pozicija	Preciznos t	Odziv	F1	AUC
FW (Napadač)	0,88	0,85	0,87	0,990
MF (Vezni)	0,79	0,72	0,75	0,922
DF (Branič)	0,86	0,90	0,88	0,951
GK (Golman)	0,97	0,98	0,97	0,995

Rezultati – Predikcija Golova

Linearna regresija i Random Forest Regressor

$R^2 = 0,79$

koeficijent determinacije
(RF Regressor)

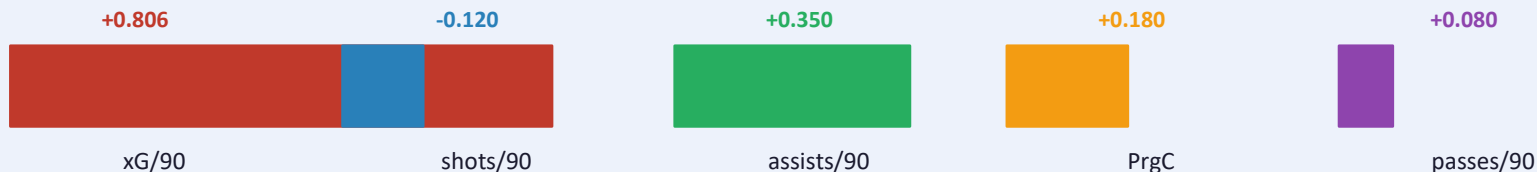
RMSE = 0,090

gol/90 min prosječna pogreška
(Linearna regresija)

xG koef. = 0,806

dominantni prediktor u
linearnoj regresiji

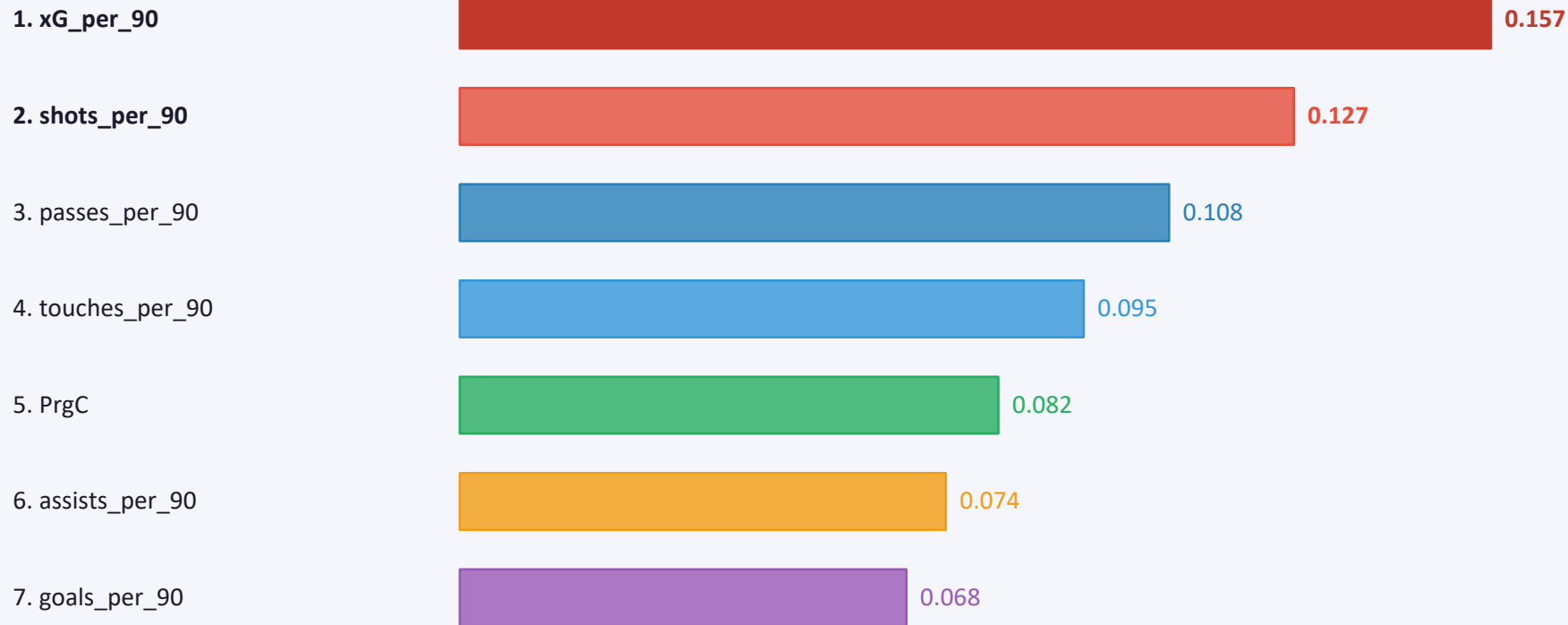
Koeficijenti Linearne Regresije – Relativna važnost prediktorskih varijabli



* Negativan koeficijent za shots/90 posljedica je multikolinearnosti – igrači s visokim volumenom udaraca često imaju niži xG.

Važnost Značajki (Feature Importance)

Random Forest model – rangiranje prediktora



xG/90 i shots/90 dominiraju zbog napadača | passes/90 i touches/90 razlikuju vezne igrače od ostalih pozicija

Ograničenja i Smjerovi Budućeg Istraživanja

Što model ne može – i kamo dalje?

Ograničenja

Samo 5 top liga

Podatci pokrivaju isključivo Premier League, La Ligu, Bundesligu, Serie A i Ligue 1. Generalizabilnost na niže rangiranim ligama nije testirana.

Jedna sezona (2024/25)

Model nije validiran na više sezona. Statistički profili igrača mogu varirati s godinom, ozljedama ili promjenom kluba.

Nema tracking podataka

Koriste se agregirani statistički podatci bez GPS/tracking informacija o trčanju, duelima i pozicijskom pokrivanju terena.

Budući rad

Proširenje na više liga i sezona

Uključiti Championship, Eredivisie i druge ligaše za bolju generalizabilnost modela.

Integracija tracking podataka

Dodavanje brzinskih zona, heatmapa i prostornih metrika bi znatno poboljšalo preciznost klasteriranja.

Personalizirani skauting alat

Razvoj interaktivne aplikacije koja bi na temelju modela sugerirala kompatibilne zamjene za svakog igrača.

Zaključci i Doprinos Istraživanja

1

xG kao dominantan prediktor:

xG_per_90 (koef. 0,806) potvrđuje se kao najsnažniji prediktor stvarnih golova, nadmašujući puki volumen udaraca.

2

K-means otkriva smislene profile:

4 klastera odgovaraju prepoznatljivim nogometnim profilima – napadači, braniči, veznjaci i kreativni igrači.

3

Hibridnost modernih veznih igrača:

18 pogrešnih klasifikacija MF→DF statistički dokazuje evoluciju uloge veznog igrača prema hibridnom 'braniču-vezniku'.

4

Visoka točnost klasifikacije:

Random Forest postiže 85,55% ukupne točnosti; golmani (GK) i napadači (FW) su najprepoznatljivije klase (AUC >0,99).

5

Kvantitativni okvir za skauting:

Predloženi model nudi objektivan alat za evaluaciju igrača koji nadopunjuje subjektivnu procjenu skauta.

Hvala na pažnji!

Pitanja i diskusija

Nikola Jakov Pavečić · Denis Romić · Bruno Polonijo

CASE 2026 | Analiza korelacija i važnosti značajki primjenom strojnog učenja

1.982

igrača

267

varijabli

85,55%

točnost

$R^2=0,79$

predikcija golova