Predviđanje broja postiignutih poena na utkamici u Evroligi

1. Opis problema

Naš problem je da pomoću trenutno ofanzivne forme tima i defanzivne forme njenog sledećeg protivnika predvidimo koliko će poena dati. Samim tim ćemo moći i da ustanovimo koji aspekti forme(npr. preciznost sa trojke ili linije slobodnih bacanja utiču na broj postignutih poena).

2. Skup podataka

Skupove podataka koji će se koristiti su Euroleage & Eurocup Datasets(https://www.kaggle.com/datasets/babissamothrakis/euroleague-datasets) i ako bude bilo potrebe Euroleague Basketball Advnaced Stats (https://www.kaggle.com/datasets/avivshany/euroleague-basketball-advanced-stats).

Euroleage & Eurocup Datasets sadrži podatke od sezone 2007/2008 do danas i sadrže veoma puno različitih tipova podataka, među kojima su najbitniji ishodi mečeva kao i standardna statistika po svakom meču.

Euroleague Basketball Advnaced Stats sadrži naprednu statistiku mečeva od sezone 2016/2017 do danas za svaki meč.

3. Metodologija

Zavisna varijabla koju predviđamo je broj postignutih poena jedne ekipe na meču. Zavisne promenljive će nam biti sve ofanzivne statstike ekipe kojoj predviđamo broj poena sa poslednjih N utkamica, ali tako da je su iste statistike sa različitih utakmica distiktne(procenat šuta se prošle utakmice je sigurno bitniji nego procenat šuta utakmice 15 kola unazad) kao i defanzivne statistike protivničke ekipe sa poslednjih N utakmica.

Tokom pretprocesiranja za svaku utakmicu ćemo napravaiti njen „train vektor“ koji će se sastojati od svih podataka koji ćemo koristi da predikciju. Primer train vektora bi bio *[3p%_utkamica1, ft%_utakmica1, steals_utakmica1, 3p%_utkamica2, ft%_utakmica2, steals_utakmica2, … , 3p%_utkamicaN, ft%_utakmicaN, steals_utakmicaN].*

Potom cemo podeliti utakmice na ***train/test/val segmente*** u odnosu ***60/20/20***.

Kada to uradimo fitovaćemo model i evaluirati ga. I tako ćemo raditi sa raličitim brojem prethodnih utakmica, kombinacijama sa dodatnim naprednim statstikama i t-testovima uklanjati statistike koje us nam beskorisne.

4. Način evaluacije

Za evaluaciju modela koristićemo prilagođeni r^2 . Pre svega zato što koristimo prilagođeni, a ne standardni jer imamo više nezavisnih promenljivih. Pošto je košarka sport koji je sklon neočekivanim rezultatima, anomalije će imati manji uticaj na našu meru nego kad bismo koristili neku metriku poput SSE.

5. Tehnologija u izradi

Koristice se programski jezik Python i njegove biblioteke Pandas, Numpy, Matplotlib, Sklearn.