

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Data mining u sportu

Alen Murtić

Voditelj: *doc. dr. sc. Damir Pintar*

Zagreb, travanj, 2017.

Sadržaj

1. Uvod	1
2. Potreba i korist data mininga u sportu	2
2.1. Izvorišne ideje	2
2.2. Potrebe za deskriptivnom analizom	3
2.2.1. Raspodjela resursa	3
2.2.2. Analiza kvalitete i stila vlastite momčadi	3
2.2.3. Pronalaženje optimalne strategije	3
2.2.4. Analiza protivnika	3
2.2.5. Procjena slučajnosti	3
2.3. Potrebe za prediktivnom analizom	3
2.3.1. Predviđanje rezultata utakmica	3
2.3.2. Kvaliteta igrača kroz godine	3
2.3.3. Rizik ozljeda	3
2.3.4. Simulacija taktike	3
3. Postojeća rješenja	4
3.1. Košarka	4
3.2. Nogomet	4
3.3. Američki nogomet	4
3.4. Ostali timski sportovi	4
4. Nove mogućnosti	5
5. Zaključak	6
6. Sažetak	7

1. Uvod

Proteklih 120 godina sport je polako postajao sve važniji element u svakodnevnom životu "običnih" ljudi. Iz krpenih lopti i dresova razvila se milijunska industrija koja pazi na svaki detalj kako bi sportaši postigli najveće moguće rezultate. Jedan od najvažnijih aspekata momčadskih sportova je donošenje odluka - treneri, sportski direktori, vlasnici i ostali djelatnici upravljaju velikim rizicima, a često nemaju potpune ili kvalitetne informacije. Što je pomalo otužna činjenica u trenutnom informacijskom vremenu kada postoji toliko dostupnih podataka o sportu. Zato je stvorena posebna disciplina analize podataka u cilju poboljšanja kvalitete odluka koje sportski djelatnici moraju donositi.

Posebni dio ekosustava modernog sporta je klađenje. Sportsko klađenje mnogima je jedini razlog zašto prate sport. Kladionice također zarađuju milijune te su često veliki sponzori sportkih momčadi, npr. glavni sponzor 10 od 20 trenutnih članova engleske Premier lige su tvrtke koje nude igre na sreću. Svaki sportski fan koji se razumije u matematiku jednom je želio napraviti model predviđanja utakmica, no kladionice ne dopuštaju najuspješnijim igračima da se klade nakon određenog dobitka. Kladionice su uvijek na dobitku, a to postižu pažljivim promjenama koeficijenata uz pomoć naprednih algoritama. Očito je one imaju veliku potrebu za analizom podataka kod promjene koeficijenata, detekcije najboljih igrača te detekcije namještenih sportskih dvoboja.

Data mining (rudarenje podataka) je vrsta analize podataka čiji je cilj pronalaženje relevantnih informacija u podacima te iskorištenje istih za poboljšanje procesa donošenja odluka. Postoje dvije kategorije *data mininga*: deskriptivna (otkrivajuća, opisna) te prediktivna (predviđajuća) analiza. Deskriptivna analiza u nestrukturiranim, prevelikim ili nejasnim podacima nalazi "skrivena" informacije te ih kranjem korisniku predstavlja u razumljivom obliku. Prediktivna pokušava naći pravilnosti u podacima te ih iskoristiti za predviđanje budućih. Ne postoji neka čvrsta granica između *data mininga* i ostalih vrsta analiza podataka, no njegova najčvršća značajka je traženje i dubinska analiza.

Ovaj seminarski rad opisuje potrebu i korist *data mininga* u sportu, postojeća rješenja te moguće nadogradnje. Fokusiran je na 3 vrlo različita timska sporta: košarku, nogomet i američki nogomet kako bi čitatelju prikazao što širi opus problema i ideja *data mininga* u sportu.

2. Potreba i korist data mininga u sportu

2.1. Izvorišne ideje

Većina ideja statističke analize u sportu dolazi iz baseballa, a za to postoje jasni razlozi:

1. baseball je sport s dugom tradicijom i velikom popularnošću, uvijek je bio profitabilan pa je privlačio mnogo novih ideja i pametnih ljudi. Bitna je i činjenica da je američka profesionalna liga MLB nastala spajanjem dvaju liga AL (American league) i NL (National league) koji su unutar MLB-a i dan danas donekle neovisni. Zbog toga su postojala dva "tržišta" za promociju sportske statistike.
2. legende sporta vrlo su važan element razgovora o baseballu, a budući da on ima turbulentnu i rasističku prošlost u SAD-u, postoji potreba za revizijom nekih dijelova iste.
3. njime dominira statistika, pogledom na detaljne rezultate od prije 100-injak godina može se razumjeti mnogo više o utakmicama nego kod ikojeg drugog sporta.

U engleskom jeziku postoji izraz sabermetrics koji označava korištenje statistike u cilju razumijevanja baseballa (danas se koristi i u kontekstu ostalih sportova), a potječe od organizacije SABR (Society for American Baseball Research) nastale 1971. s ciljem elemenata tog sporta. Specifičnost organizacije je u tome da je sačinjena od novinara, bivših igrača, navijača, statističara i ostalih dionika sporta kako bi pružila što širi i točniji pogled na teme koje ju zanimaju.

Pravi rast data mininga u baseballu počinje kasnih 1990-ih i ranih 2000-ih, a najviše ga simboliziraju dvojica menadžera Billy Beane i Theo Epstein. Billy Beane je 1998. postao generalni menadžer (GM) Oakland Athleticsa, momčadi s prilično ograničenim budžetom u odnosu na protivnike. Beane je osmislio princip nazvan Moneyball, ideju kupovine podcijenjenih igrača u cilju optimizacije potrošnje novca. Igrače je evaluirao analitičkim metodama zasnovanim na matematici. Beane je godinama imao značajno bolje rezultate od mnogih skupljih momčadi, ali Oakland nije uspio osvojiti ligu. Theo Epstein je 2002. sa samo 28 godina postao GM Boston Red Soxa, kluba koje je 84 bio bez trofeja prvaka. Samo 2 godine kasnije uspjeli su osvojiti naslov te to ponoviti 2007. 2016. je uspio sličan pothvat s Chicago Cubsima. Njegov uspjeh je analitičkom pristupu u sportu dao velik zamah.

2.2. Potrebe za deskriptivnom analizom

2.2.1. Raspodjela resursa

- minute, novci

2.2.2. Analiza kvalitete i stila vlastite momčadi

2.2.3. Pronalaženje optimalne strategije

2.2.4. Analiza protivnika

- dug i naporan proces skratiti što je više moguće

2.2.5. Procjena slučajnosti

Donošenje odluka u svakom sportu, a posebno nogometu, je često temeljeno na pitanju: učiniti promjenu ili ne? Treba li klub dati otkaz treneru, treba li trener zamijeniti napadača koji ne postiže golove nekoliko utakmica, promijeniti taktiku ili ne i sl.? Analitički pristup može dati odgovor koja je vjerojatnost da se dogodio negativan ishod utakmice uz igru koju je momčad pružila. Nekad je lako shvatiti da je protivnički vratar imao odličan dan i da je zato naš tim izgubio, ali u pravilu je teško objektivno procijeniti je li nešto slučajnost ili realnost, pogotovo ako se klub nalazi u zoni ispadanja. Npr. u sezoni 2011.-12. Wigan FC je imao katastrofalne rezultate prvih 29 kola engleske Premier lige, iako su očito igrali kvalitetno. No, klub nije dao otkaz treneru i sezonu je završio s 9 pobjeda iz 11 utakmica i spasio se od ispadanja. Takva strpljivost je rijetka, a dubinska analiza rezultata može dati odgovor isplati li se čekati.

2.3. Potrebe za prediktivnom analizom

2.3.1. Predviđanje rezultata utakmica

- kladionice
- klubovi: planiranje sezone

2.3.2. Kvaliteta igrača kroz godine

2.3.3. Rizik ozljeda

2.3.4. Simulacija taktike

3. Postojeća rješenja

3.1. Košarka

3.2. Nogomet

3.3. Američki nogomet

3.4. Ostali timski sportovi

4. Nove mogućnosti

5. Zaključak

6. Sažetak