Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004:007
DOI: https://doi.org/10.24867/03OO02Mijatovic

MODEL ZA PREDIKCIJU PERFORMANSI SPORTISTA BAZIRAN NA TEHNIKAMA MAŠINSKOG UČENJA

ATHLETE'S PERFORMANCE PREDICTION MODEL BASED ON MACHINE LEARNING TECHNIQUES

Dejan Mijatović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – INFORMACIONI SISTEMI

Kratak sadržaj – *Tema ovog istraživanja jeste izrada modela za predviđanje performansa sportista korišćenjem tehnika mašinskog učenja. Za potrebe modelovanja, velike količine podataka su prikupljene i obrađene. Model je na kraju evaluiran i testiran u praksi.*

Ključne reči: *mašinsko učenje, predikcija, sportski rezultati, NBA*

Abstract – *The purpose of this research is to train model for athlete performance prediction, based on machine learning techniques. In order to train the model, large amount of data was collected, preprocessed and the fed to the model. Model was eventually evaluated and tested.*

Keywords: *machine learning, prediction, data science, sports prediction*

1. UVOD

Svedoci smo vremena u kome se velike promene dešavaju veoma brzo, toliko da neretko nismo u stanju da ih propratimo i automatski se prilagodimo na njih.
Tehnologija se rapidno razvija i napreduje, te se neretko susretnemo sa činjenicom da ono što je do juče smatrano naučnom fantastikom, danas postaje svakodnevnica. Pojavom interneta, naš stil života je promenjen iz korena. Informacije se prenose velikim brzinama, bez geografskih barijera, i danas su svima dostupne.
Znanje koje se može steći korišćenjem informacija i izvora na internetu je lako dostupno i besplatno, i škole i univerziteti nemaju više monopol nad obrazovanjem ljudi. Samim tim je mnogo veći broj ljudi aktivno uključen u razvoj i napredak tehnologija [1-2].
Prema dosadašnjim istraživanjima [3], mi imamo tu priliku da budemo deo vremena i dostignuća koja imaju kapacitet da promene život kakav poznajemo iz korena. Veštačka inteligencija je ono što bi lako nazvali trenutno „gorućom" temom na polju računarskih nauka i gotovo je sigurno da nam najveća otkrića iz ove oblasti tek predstoje.
Pravci u kojima će se ova disciplina razvijati, njene mogućnosti i potencijal da doprinese razvoju civilizacije za sada je još uvek nepoznata.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Srđan Sladojević, docent

Danas, softverski proizvodi zasnovani na veštačkoj inteligenciji i mašinskom učenju su dominantno zastupljeni u procesima proizvodnje, a sve više i u oblastima komunikacija i drugim disciplinama gde je potrebno razmišljanje, tj. gde sama automatizacija procesa ne igra najznačajniju ulogu, već je mašinu potrebno naučiti da prepoznaje šablone i smisleno na njih reaguje.
Veštačka inteligencija svakako nije nov pojam i naučnici o njoj diskutuju već unazad nekoliko desetina godina. Iako je mnogo radova napisano i mnogo diskusija vođeno o njenoj teorijskoj osnovi uz razvoj mnogobrojnih teorija i modela, sama primena u praksi koja bi omogućila rešavanje konkretnih problema i praktičnu upotrebu veštačke inteligencije je došla nešto kasnije usled nedostatka resursa koji su tek nedavno dovoljno uznapredovali da bi podržali masovniju primenu tehnologije na tako visokom nivou, i osposobili mašine da uče.
Ovaj rad ima za cilj da testira i pokaže primenu mašinskog učenja za zadatke prognoziranja događaja u budućnosti, oslanjajući se na informacije o prethodnim događajima od značaja za predviđanje sportskih rezultata igrača. Podaci koje smo koristili su iz domena košarke, tačnije analiza koja je napravljena u sklopu ovog rada obuhvata sve košarkaške utakmice u američkoj NBA ligi i statistike igrača na svakoj od utakmica. Svrha rada je da pokaže kako se mašinsko učenje i veštačka inteligencija danas mogu koristiti za razvijanje modela koji imaju sposobnost da predviđaju performanse igrača na sledećoj utakmici na osnovu informacija o njihovim uspesima iz prethodnih utakmica čiji su učesnici bili.

2. KORIŠĆENE TEHNOLOGIJE

Oblast veštačke inteligencije i mašinskog učenja je relativno nova, i kao takva razvija se veoma brzo. U cilju praćenja zahteva korisnika i pospešivanja novih, boljih rešenja, u ovoj oblasti koristi se više jezika i softverskih pristupa, kako bi se stvorila mogućnost da se sve tehnološki zahtevne ideje vremenom i praktično realizuju. Prilikom izrade ovog rada i ispitivanja mogućnosti treniranja datog modela u svrhu prognostike sportskih rezultata korišćene su tehnologije i programi koji su u oblasti mašinskog učenja veoma zastupljeni:
Python programski jezik [4] i njegove biblioteke i alati se koriste kroz sve segmente istraživanja. *Jupyter notebook,* kao interaktivno Python okruženje koje se izvršava u internet pretraživaču, ubrzava ceo eksperimentalni proces istraživanja. Za prikupljanje podataka sa interneta se koriste *BeautifulSoup* i *requests* biblioteke. *Numpy i Pandas,* popularni tandem za rad sa podacima, se koristi

za *preprocessing* fazu istraživanja. U fazi tumačenja podataka i vizuelizacije se koristi *matplotlib* biblioteka, a uz *scikit-learn* [2] alata ćemo trenirati modele mašinskog učenja.

3. METODOLOGIJA

Cilj istraživanja je bilo kreiranje modela za predviđanje učinka igrača na košarkaškoj utakmici. Model je kreiran i razvijen sa primarnim ciljem da predviđa broj poena igrača na narednoj utakmici, na način da prognozira da li će igrač postići više ili manje poena od zadate granice. Zadata granica je uglavnom igračev prosek, uz manje korekcije kako bi odgovaralo trenutnoj formi igrača.

Za rešenje našeg problema, kao i za većinu problema koji se danas rešavaju u okviru domena mašinskog učenja su nam bitni podaci. Podaci su nam dostupni na internetu, ali nisu lako dostupni za preuzimanje pa moramo za ovu svrhu napraviti skripte koje će automatski potreban sadržaj sajta preuzeti, obraditi i pretvoriti u nama potrebni format podataka.

Podatke je potom potrebno pospremiti na čvrsti disk kako bi nam bili stalno dostupni. U cilju obezbeđivanja pouzdane i kontinuirane dostupnosti podataka, potrebna je relaciona baza podataka, izmodelovana i kreirana na lokalnom disku.

Podaci koje smo na ovaj način prikupili i sačuvali, iako struktuirani, su i dalje sirovi. Podatke je potrebno prečistiti i obraditi kako bi bili kao takvi spremni za treniranje modela. Ovo je ujedno i najkompleksniji i vremenski najzahtevniji proces i potrebno mu je posvetiti mnogo pažnje, kako bi kreirali svrsishodnu bazu podataka koja se može koristiti za pozdano i precizno prognoziranje.

Treniranje modela se vrši tako što se obrađeni podaci serviraju algoritmima mašinskog učenja. Kako smo za treniranje koristimo *supervised* algoritme, bitno je da svaki podatak bude i obeležen (eng. *labelled*).

a) Prikupljanje podataka

Podaci o rezultatima i statistikama sportskih događaja su u većini slučajeva široko rasprostranjeni na internetu, i mogu se relativno lako pronaći kao javno dostupni podaci. Za svrhe prognoziranja i kreiranja pouzdanog modela predviđanja kao i treniranja istog od izuzetnog je značaja i pitanje koliko su ti podaci detaljni, jesu li u dovoljnoj meri potpuni, kao i to da li poseduju sva obeležja koja su nama potrebna.

Kao polazna tačka za prikupljanje podataka potrebnih u ovom radu korišćen je zvanični internet portal NBA lige. Uz zvanični sajt, korišćen je i portal poznate američke novinske agencije, specijalizovane za izveštavanje o sportskim događajima, ESPN, kao i portal sa sportskim rezultatima Flashscore kako bi kreirali bazu podataka koja sadrži potpune informacije koje su nam potrebne za razvoj i treniranje modela predviđanja sportskih rezultata košarkaša.

player_id	3992
game_id	4.009e+08
wl	L
side	away
min	37
fgm-fga	4-16
fg%	.250
3pm-3pa	1-5
3p%	.200
ftm-fta	4-7
ft%	.571
reb	7
ast	13
blk	0
stl	3
pf	1
to	6
pts	13
nick	james-harden
name_player	James Harden
Born	Aug 26, 1989 in Bellflower, CA (Age: 28)
Drafted	2009: 1st Rnd, 3rd by OKC
College	Arizona State
Num	#13
Pos	SG
Experience	8 years
Current_team	Houston Rockets
Extra_data	6' 5", 220 lbs
date	Wed 11/16
home_id	OKC
away_id	HOU
h_score	105
a_score	103
season	2017
type	REGULAR
link_home	http://www.espn.com/nba/teams/roster?team=Okc
name_home	Oklahoma City Thunder
link_away	http://www.espn.com/nba/teams/roster?team=Hou
name_away	Houston Rockets

SLIKA 1 - ESPN FORMAT PODATAKA

Na slici 1 vidimo primer jedne statistike igrača sa utakmice, dobijene sa ESPN sajta. Podaci u ovom stanju su sirovi i kao takve ih moramo obraditi.

Kako nijedan od ovih portala pojedinačno ne poseduje sve potrebne podatke, preuzeti su podaci sa sva tri zajedno. Na ovaj način ujedno je izvršena I provera preklapanja podataka u korišćenim bazama, te kreirana jedinstvena baza podataka koja sadrži sve elemente neophodne za razvoj našeg modela.

Prikupljeni su podaci sa 19848 utakmica za 20 sezona od 1999. godine po 2018. godine. U podacima se nalazi statistika za 490 igrača i ukupan broj statistika igrača na utakmicama koji je prikupljen je 191331. Jedna sezona u NBA ligi se sastoji od predsezone, regularnog dela i doigravanja.

Kako se s pojavom tehnologija vodi opširnija i preciznija statistika, tako su i podaci u ovom skupu raspoređeni nejednako, odnosno sa protekom vremena, informacije koje možemo naći su se umnožavale, i postale detaljnije. Tako za poslednjih nekoliko sezona imamo mnogo preciznije i kompletnije podatke nego za početne sezone u ovom skupu podataka, počevši od 1999. godine

b) Faze modelovanja

Tradicionalni CRISP-DM model ima šest faza [5]:

1. Razumevanje domena
2. Razumevanje podataka
3. Obrada podataka
4. Treniranje modela
5. Evaluacija modela
6. Puštanje modela u rad

1. *Razumevanje domena*

Istraživanje i određivanje domena je prvi korak razvoja modela. U ovom radu, domen je *NBA* košarka, a problem koji pokušavamo da rešimo je predviđanje učinka igrača po utakmici, tačnije, da li će postići više ili manje poena od svog proseka.

Takođe, veoma je bitno je da razumemo sve specifične karakteristike samog sporta. Ovo podrazumeva da razumemo pravila sporta, kako se igra i koji faktori mogu da utiču na sam ishod meča, a zatim i na učinak igrača pojedinačno. Do ovih saznanja možemo doći na osnovu ličnog poznavanja oblasti, izučavanje dostupne literature ili konsultovanjem ljudi koji su stručni u tome.

Za rešenje problema, bitno je da znamo sve faktore koji su relevantni i od uticaja za krajnji ishod, a koje imamo među prikupljenim podacima i kao takve ih možemo ispitati.

2. *Razumevanje podataka*

Podaci se prvo moraju dobro razumeti kako bi se izveo zaključak koje su korekcije potrebne nad njima. Ovo podrazumeva da smo najpre dobro upoznati sa domenom koji istražujemo, kako bismo mogli razumeti i podatke, te njihove nedostatke i prednosti, kao i njihovu upotrebljivost i pouzdanost u konkretnom slučaju za koji su namenjeni. Potrebno je da razumemo format podataka, način na koji su zapisani i šta svako od obeležja predstavlja.

Granularitet podataka treba biti uzet u obzir. Na našem primeru vidimo da je bitno razlikovati da li su podaci vezani za sezonu ili pojedinačnu utakmicu, kao i da je statitstika predstavljena na nivou tima ili pojedinačnog igrača.

S obzirom na to da je naš problem u domenu klasifikacije, tačnije ima dva moguća ishoda, manje ili više poena od proseka, moramo odrediti šta nam predstavlja *class* obeležje. To obeležje, kao takvo, nemamo u trenutno dostupnim podacima i potrebno je da ga sami kreiramo. Broj poena za svaku utakmicu nam je već dostupan i možemo ga iskoristiti kako bismo izračunali prosek poena po utakmici, a kasnije i da uporedimo taj prosek sa brojem poena po jednoj utakmici i obeležimo sa više ili manje.

Drugi pristup je bio računanje koliko broj poena odstupa od proseka igrača i u kom smeru, što je iz domena regresije.

3. *Obrada podataka*

Najbitniji deo treniranja modela mašinskog učenja jeste obrada podataka, metoda koja se još popularno naziva tzv. *Preprocesing.* U ovom koraku računamo i prethodno pomenuto *class* obleležje, više ili manje poena. Podaci su ulazni faktor modela i od njih zavisi koliko dobro će naša mašina naučiti. Koliko su nam dobri podaci, toliko je dobar i predviđanja.

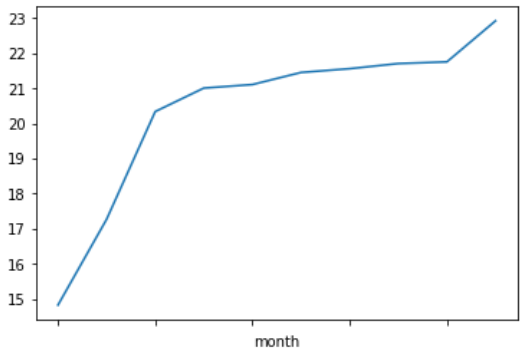
Vrednosti je potrebno prekontrolisati da nemamo *null* vrednosti, neodgovarajući format podataka, neispravne, netačne ili nekompletne vrednosti. Parametre (eng. *features*) je potrebno filtrirati, odnosno, odrediti koji su faktori bitni i koji nam povećavaju preciznost krajnjeg modela. Potrebno je odrediti koje parametre već imamo među podacima, a koje je potrebno da sami uvedemo i način za njihovo računanje.

Sledeći primer prikazuje jedan od načina određivanja parametara

month	
6	22.923670
4	21.755118
5	21.707233
2	21.560923
3	21.453663
12	21.108962
1	21.008574
11	20.338827
10	17.270211
9	14.826800

SLIKA 2 - PROSEK POENA PO MESECU U SEZONI

Na slici 2 je prikazana tabela u kojoj se nalaze proseci poena igrača na utakmici po mesecima u kojima se igra. Prvi mesec gledajući od gore ka dole bi bio mesec jun i prosek za taj mesec je 22.92 poena po meču. Tabela pokazuje jasnu korelaciju proseka poena i meseca.



SLIKA 3 - GRAFIKON PROSEKA POENA PO MESECU U SEZONI

Na slici 3 je to još bolje prikazano gde se jasno vidi trend rasta proseka poena igrača kako se zona odmiče. Pošto je septembar početak sezone, tačnije predsezone, a jun mesec je finalni deo sezone, kada se igraju doigravanja i odlučuje se pitanje pobednika lige, vidi se da timovi i sami igrači tempiraju formu i igraju sve bolje kako sezona odmiče. Na osnovu ovoga možemo zaključiti i da je period sezone jedan od bitnih faktora koji utiče na učinak igrača i kao takav se mora naći kao jedan od *feature*-a pripremljenih podataka.

4. *Treniranje modela*

Pripremljeni podaci sadrže 79661 uređenu torku sa 32 obeležja. Bitno je napomenuti da su prilikom filtriranja odbačeni svi nepotpuni podaci, kao i podaci za igrače sa zanemarljivim brojem utakmica.

Na slici 4 vidimo sve parametre podataka koje serviramo algoritmu za treniranje modela.

```
df.columns
```

```
Index(['name_player', 'season', 'game_id', 'index', 'opp', 'win_opp',
       'win_opp_ly', 'team', 'win', 'win_ly', 'ou_ratio', 'date', 'rest_days',
       'dayofweek', 'timestamp', 'wl', 'margin', 'side', 'min', 'pts', 'espn',
       'avgpts', 'ma5pts', 'ma10pts', 'ma20pts', 'avgc', 'avgespn', 'ma5espn',
       'ma10espn', 'ma20espn', 'ou', 'ou%'],
```

SLIKA 4 - LISTA KONAČNIH PARAMETARA

Pripremljene podatke delimo u dve grupe. Grupa za treniranje i grupa za testiranje. Uobičajeno je da se koristi odnos 7 prema 3 u korist modela za treniranje, za podelu podataka. Podaci se prethodno izmešaju nasumično i rasporede u dve grupe. Skup podataka za treniranje se servira algoritmu mašinskog učenja, a skup podataka za testiranje se koristi za evaluaciju modela, tačnije određivanje procenta uspešnosti [6].

Kako naš problem možemo posmatrati i kao klasifikaciju i kao regresiju, dostupne su nam obe grupe algoritama mašinskog učenja. Klasifikacioni algoritmi kao *class* obeležje primaju vrednosti više ili manje, dok algoritmi regresije kao ulaz primaju kontinuiranu numeričku vrednost koja se računa razlikom poena sa jedne utakmice i proseka poena igrača.

5. Evaluacija modela

Istreniran model je potrebno testirati kako bi se odredila preciznost predviđanja i potvrdio izbor algoritma.

Za evaluaciju regresivnih modela korišćena je *Mean Squared Error* metrika, dok je za klasifikacione modele preciznost određena odnosom tačnih pretpostavki i ukupnog broja pretpostavki, *Classification Accuracy* [7].

Rezultati evaluacije su pokazali da bi regresivni modeli mogli biti uspešniji u praksi i zbog toga je sa njima nastavljeno dalje testiranje.

6. Puštanje modela u rad

Najbolji način da se odredi usprešnost modela je da se isproba u praksi i testira nad podacima dobijenim u realnom vremenu.

Za ovu potrebu je kreiran Python program koji ima za cilj da automatizuje proces testiranja modela nad „živim“ podacima. Pošto se utakmice igraju svaki dan, jednom dnevno preuzme „sveže“ podatke sa interneta, obradi ih i pruži modelu kako bi od njega dobio predviđanje. Predviđanje od jučerašnjeg dana uporedi sa pravom vrednošću i odredi da li je bio uspešan ili ne. Na osnovu kratkog dvomesečnog testiranja, uspešnost modela je 62%, za šta se može reći da je dobar rezultat, daleko iznad nasumičnog.

4. ZAKLJUČAK

Mašinsko učenje danas ima veoma široku primenu. Teško je pronaći domen u kome ono nije zastupljeno, pogotovo kada govorimo o tehnologiji i njenom razvoju. Čitavo polje veštačke inteligencije ima toliki potencijal da unapredi svaki segment života, a mi smo tek zagrebali po površini i ostaje nam da vidimo u kom pravcu će se dalje usavršavati.

U ovom radu, cilj je bio da pokažemo moć mašinskog učenja kada je reč o predviđanju sportskih rezultata košarkaša. Model je bio razvijen i treniran tako da prognozira rezultate igrača na osnovu njihove igre iz prošlosti, koja je u model unešena u formi baze podataka. Kao rezultat istraživanja, dobili smo model za predviđanje učinka košarkaša po utakmici čiji je procenat uspešnosti 62%, što možemo protumačiti kao dobar rezultat, s obzirom na dostupne podatke.
NBA timovi koriste slične tehnike za analizu igre i svakako poseduju neizmerno veću i opširniju količinu podataka. Podaci kojima raspolažu su toliko detaljni, da za svaki sekund utakmice, mogu izvući poziciju i rastojanje za svakog od igrača na terenu [8].
Najzad, važno je napomenuti da je za dobijanje dobrog i preciznog modela kvalitet podataka od krucijalnog značaja. Ukoliko dobri i pouzdani podaci nisu dostupni i ne predstavljaju polaznu tačku, samom obradom podataka nismo u mogućnosti da dođemo do prihatljivog ishoda.

5. LITERATURA

[1] Justin M. Weinhardt, Traci Sitzmann, Revolutionizing training and education? Three questions regarding massive open online courses (MOOCs), Human Resource Management Review, 2018
[2] Alison George, Free online MIT courses are an education revolution, New Scientist, Volume 219, Issue 2925, 2013
[3] Spyros Makridakis, The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms, Futures, Volume 90, 2017
[4] http://python.org
[5] C. Shearer, The CRISP-DM model: the new blueprint for data mining, J. Data Warehousing 5 (4) (2000)
[6] Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: Machine learning in python*. Birmingham ; Mumbai: Packard publishing limited.
[7] http://scikit-learn.org/stable/documentation
[8] Yuanhao (Stanley) Yang (2015) Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics

Kratka biografija:

Dejan Mijatović rođen je u Osijeku 1991. god. Završio je gimnaziju Jovan Jovanović Zmaj u Novom Sadu, kao i osnovne akademske studije na Fakultetu Tehničkih nauka. Master rad na Fakultetu tehničkih nauka iz oblasti Inženjerstvo informacionih sistema odbranio je 2018.god.
kontakt: dejan.mijatovic@uns.ac.rs