Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/13BE33Nisic

PREDIKCIJA BROJA INDEKSNIH POENA IGRAČA U ABA LIGI SA FOKUSOM NA PRIKUPLJANJU I EKSPLORATIVNOJ ANALIZI PODATAKA

PREDICTION OF THE NUMBER OF INDEX POINTS OF PLAYERS IN THE ABA LEAGUE WITH A FOCUS ON DATA COLLECTION AND EXPLORATORY ANALYSIS

Miloš Nišić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad se bavi predikcijom broja indeksnih poena koje igrač ostvari na košarkaškoj utakmici. Fokus rada je na prikupljanju i eksplorativnoj analizi podataka. Prikupljanje podataka je vršeno sa sajta eurobasket.com pomoću tehnika web-scrapinga. Nakon sređivanja skupa podataka, ekstrakcije obeležja i eksplorativne analize podataka vršena je predikcija pomoću tri različita regresora: Lasso, Random Forest i LightGBM. Optimizacijom hiperparametara implementacija ovih algoritama došlo se do modela pomoću kojih je vršena predikcija broja indeksnih poena. Najbolje rezultate među njima pokazao je model LASSO regresije sa srednjom apsolutnom greškom MAE = 5.617. Izneti su predlozi za poboljšanje skupa podataka, a samim tim i za dalji razvoj ovog rešenja.*

Ključne reči: *Lasso regresija, Random Forest regresija, LightGBM regresor, predikcija indeksa korisnosti*

Abstract – *This paper presents the machine learning approach to automatically predict the number of index points that a player achieves in a basketball game. The focus of the work is on data collection and exploratory analysis. Data was collected from eurobasket.com using web-scraping techniques. After cleaning the data set, feature extraction and exploratory data analysis were performed. The prediction was made using three different regressors: Lasso, Random Forest, and LightGBM By optimizing the hyperparameters of these algorithms' implementations, we came up with models that were used to predict the number of index points. The LASSO regression model achieved the best results with the mean absolute error MAE = 5.617. The paper proposes further improvements of the data set as future work.*

Keywords: *Lasso regression, Random Forest regression, LightGBM regressor, Personal index rating predictions*

1. UVOD

Kada je sport nastajao, njegov osnovni cilj bilo je takmičenje i zabava. Međutim, vremenom, sport je prerastao u unosan biznis i veliku industriju. Jedan od najzastupljenijih sportova danas, u koji se ulažu enormne sume novca, jeste košarka. Prosečna vrednost timova u NBA (engl. *National Basketball Association*) ligi je 2020. godine bila preko dve milijarde dolara [1]. Kao dodatni vid zabave za ljubitelje sporta osmišljene su razne igre na sreću, kao i sve popularnije fantazi (engl. *fantasy*) igre.

ABA liga, kao najveće i najpoznatije regionalno sportsko takmičenje, je takođe pokrenula svoju fantazi igru. S obzirom na to nastala je ideja primene mašinskog učenja u svrhe predikcije broja indeksnih poena koje će igrač osvojiti na utakmici čime bi se olakšalo takmičenje u fantazi ligi.

Kao osnovni parametar prilikom bodovanja učinka igrača za ovu fantazi igru koristi se broj indeksnih poena koje igrač ostvari na utakmici. U ovom radu vršena je predikcija ovog košarkaškog statističkog parametra. U ovom radu je prikazan postupak prikupljanja i eksplorativne analize podataka neophodnih za treniranje modela mašinskog učenja. Prikupljanje podataka je vršeno sa sajta eurobasket.com pomoću tehnika web-scrapinga. Nakon sređivanja skupa podataka, ekstrakcije obeležja i eksplorativne analize podataka vršena je optimizacija različitih regresionih modela. Kao najbolji model se pokazao model LASSO regresije koji je postigao srednju apsoulutnu grešku od 5.617.

2. PRETHODNA REŠENJA

Još od devedesetih godina prošlog veka ljudi su počeli ozbiljnije da se bave analizom statističkih podataka u košarci. U literaturi postoji mnogo radova koji se bave košarkaškom statistikom. Mnoge knjige, radovi i naučni članci u časopisima fokusirani su na uticaje različitih parametara na individualnu i timsku statistiku [2, 3, 4, 5]. Takođe, prethodna istraživanja bave se predikcijama mnogih drugih stvari koje se tiču sporta, kao što su različite mere učinka igrača na utakmici ili pak čitavoj karijeri [6, 7], a takođe i predikcijama uspešnosti timova [8] i ishoda utakmica [9, 10].

Problem rešavan u ovom radu je da se na osnovu statističkih podataka iz prošlosti pronađu ili kreiraju parametri koji mogu pomoći u predviđanju igračevog učinka. Postoji dosta sličnih radova na ovu temu, što u košarci [6, 7, 10], što u nekim drugim sportovima [11, 12], ali nijedan koji se konkretno bavi *ABA* ligom.

Različitim merama učinka igrača najviše se bave autori rada [3], ali više informativno, dajući formule za svaku od

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, docent.

navedenih metrika i objašnjavajući parametre koji se koriste u tim formulama. Oni se takođe bave i različitim parametrima koji utiču na predikciju, između ostalih ofanzivnim i defanzivnim rejtingom, efektivnim i pravim procentom šuta, tempom kojim igraju timovi i procentom uspešnosti skokova.

Rad koji se najviše bavi obeležjima i na koji se najviše referenciraju drugi radovi je [7]. U tom radu predikuje se broj poena, za koje kažu da su tradicionalno dominanta mera učinka igrača na utakmici, ali ne i najtačnija, a takođe se pominju i druge mere koje su pominjane u radu [3] i koje su prednosti, odnosno mane jednih i drugih.

3. METODOLOGIJA I ALATI

U ovom poglavlju predstavljena je arhitektura rešenja, kako se došlo do skupa podataka, kako je on izgledao, a predstavljena je i izgradnja modela, kao i evaluacija.

3.1. Arhitektura rešenja

Arhitektura rešenja sastoji se iz četiri dela i čine je:

1. Kreiranje skupa podataka
2. Ekstrakcija obeležja
3. Izgradnja modela
4. Evaluacija modela

Ulaz sistema predstavljaju podaci sa utakmica koji su prikupljeni sa *eurbasket.com* sajta, odnosno obeležja koja su dobijena iz prikupljenih podataka i čine ih prosečne vrednosti ostvarenih učinaka koje su igrač, njegova i protivnička ekipa ostvarile na prethodne 3 i na prethodnih 5 utakmica, dok izlaz sistema predstavlja predikciju indeksa korisnosti koji je igrač ostvario na utakmici.

3.2. Kreiranje skupa podataka

Kreiranje skupa podataka je vršeno sa sajta eurobasket.*com* tehnikama *web-scraping*-a. Preuzeti podaci su sređivani, odnosno čišćeni od nedostajućih vrednosti, pogrešno unetih podataka, itd. U ovom koraku vršeno je oblikovanje skupa podataka, odnosno organizovanje u oblik pogodan za ekstrakciju obeležja i kasnije korišćenje u modelima.

Za prikupljanje podataka korišćena je *BeautifulSoup* biblioteka [14] i *Selenium Web Driver* [15]. *Selenium Web Driver* se koristio za automatski pristup stranicama sa podacima, nakon čega je *BeautifulSoup* parsirao *Hypertext Markup Language* (*HTML*) dokument.

3.3. Ekstrakcija obeležja

U delu ekstrakcije obeležja vršena je transformacija "sirovih" podataka u podatke koji se mogu analizirati i koristiti tako da generišu validna i upotrebljiva znanja. Ovaj korak čine prvenstveno razumevanje i istraživanje podataka.

Na osnovu postojećih obeležja kreirana su nova obeležja, koja predstavljaju naprednu košarkašku statistiku i koja su upotrebljavana u drugim radovima.

Za predikciju indeksa korisnosti igrača na nekoj utakmici ne mogu se koristiti podaci sa te utakmice. Zbog toga sledeći problem bio je dobiti istorijski relevantne podatke za utakmicu za koju se predviđa indeks korisnosti.

Zbog toga su za svaku kategoriju kreirane dve nove koje su se odnosile na ostvaren prosek u određenoj statističkoj kategoriji na prethodne 3 i na prethodnih 5 utakmica. Na osnovu ovako dobijenih obeležja, kao i drugih koja su se ticala karakteristika igrača, vršila se predikcija.

Nakon kreiranja obeležja vršena je eksplorativna analiza podataka uz pomoć matrica korelacije i toplotnih mapa. Za pregled najuticajnih obeležja na predikciju indeksa korisnosti, pre nego što nam sami modeli pokažu, korišćna je i *SelectKBest*[1] klasa koju nudi *Scikit-learn* [16] biblioteka.

3.4. Izgradnja modela

Bez kvalitetnih podataka nema ni kvalitetnih rešenja, međutim, odabir modela predstavlja takođe jako bitan deo ovog sistema. Ovaj segment rada predstavlja odabir algoritma mašinskog učenja, kao i odabir hiperparametara koji se koriste tokom treniranja nekih od modela. U ovom radu razmatrani su i upotrebljeni modeli *LASSO* regresije, *Random Forest* regresije, kao i *Light Gradient Boosting Machine* (*LightGBM*) model.

Optimizacija hiperparametara modela predstavlja problem pronalaska optimalnih parametara za algoritme učenja. Od modela do modela zavisi koliko oni iziskuju ovu operaciju. *LASSO* regresija i *Random Forest* algoritam sadrže mnogo manje parametara koje je potrebno podesiti od *LightGBM* modela.

Za *LASSO* regresiju odabir *alfa* vrednosti urađen je odabirom najboljeg rezultata prilikom unakrsne validacije, pri čemu je opseg vrednosti automatski odredio sam algoritam, odnosno njegova implementacija iz *Scikit-learn* biblioteke.

Za pronalaženje hiperparametara za *Random Forest* algoritam korišćena je nasumična pretraga, uz pomoć *RandomizedSearchCV* , što takođe predstavlja *Scikit-learn* implementaciju ovog načina pretrage.

Hiperparametri i opsezi vrednost iz kojih su birani mogu da se vide u tabeli 1.

[1] *SelectKBest* rangira obeležja koristeći prosleđenu funkciju mere, u ovom slučaju f_regression [17], a zatim uklanja sve osim prvih k najbolje rangiranih obeležja. f_regression je linearni model za ispitivanje individualnog efekta svakog od mnogih regresora. Ovo je funkcija bodovanja koja se koristi u postupku odabira obeležja.

Tabela 1: *Hiperparametri i njihov opseg za Random Forest algoritam*

Hiperparametri	**Opseg vrednosti**
n_estimators	10 vrednosti iz opsega [1, 2000]
max_features	['*auto*', '*sqrt*']
max_depth	11 vrednosti iz opsega [10, 110]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
bootstrap	[*True, False*]

Za *LightGBM* algoritam korišćena je Bajesova optimizacija hiperparametara uz pomoć *HyperOpt* biblioteke [18] . Hiperparametri koji su korišćeni i opsezi vrednost iz kojih su birani mogu da se vide u tabeli 2.

Tabela 2: *Hiperparametri i njihov opseg za* LightGBM *algoritam.*

Hiperparametri	**Opseg vrednosti**
n_estimators	10 vrednosti iz opsega [1, 2000]
max_depth	11 vrednosti iz opsega [10, 110]
num_leaves	6 vrednosti iz opsega [100, 1300]
subsample	5 vrednosti iz opsega [0.1, 1]
colsample_bytree	5 vrednosti iz opsega [0.5, 1]
learning_rate	1 vrednosti iz opsega [0.1, 1]
reg_alpha	6 vrednosti iz opsega [0.1, 0.6]
reg_lambda	6 vrednosti iz opsega [0.1, 0.6]
min_child_samples	4 vrednosti iz opsega [0.1, 100]
verbose	-1

Nakon optimizacije hiperparametara korišćenih modela, vršena je evaluacija na prethodno izdvojenom test skupu i njihovo poređenje, kao i poređenje sa rezultatima sličnih radova.

4. REZULTATI I DISKUSIJA

Konačan skup podataka *ABA 2010-2019* sadrži 23 677 instanci. Za svrhe primene algoritama mašinskog učenja, podeljen je na trening i test skup, tako što trening skup čini 80%, a test skup 20% od ukupnog broja instanci. Odnosno, trening skup sadrži 18 940 instanci, dok test skup ima ukupno 4 736 instanci.

Za evaluaciju modela korišćena je srednja apsolutna greška (engl. *MAE – Mean Absolute Error*). Ova greška predstavlja apsolutnu razliku između ciljne vrednosti i vrednosti predviđene modelom.

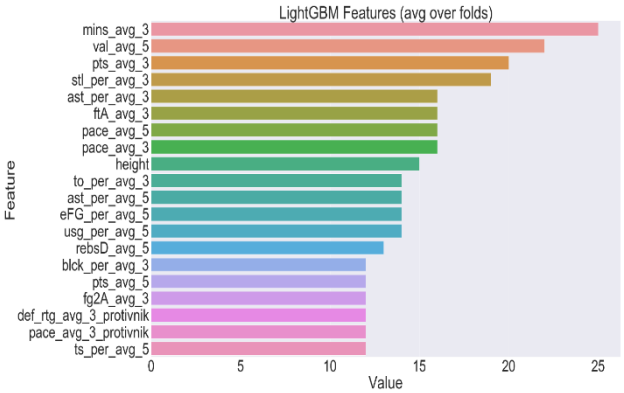
Najbolje rezultate među korišćenim algoritmima pokazao je model *LASSO* regresije sa srednjom apsolutnom greškom *MAE* = 5.617. Rezultati svih modela mogu da se vide u tabeli 3.

Kao najbitnija obeležja koja isticala su se prosečan broj indeksnih poena na prethodnih 3 i 5 utakmica, prosečan broj poena na prethodih 3 i 5 utakmica, prosečan broj odigranih minuta na prethodnih 3 i 5 utakmica, kao i procenat upotrebe na prethodnih 3 i 5 utakmica.

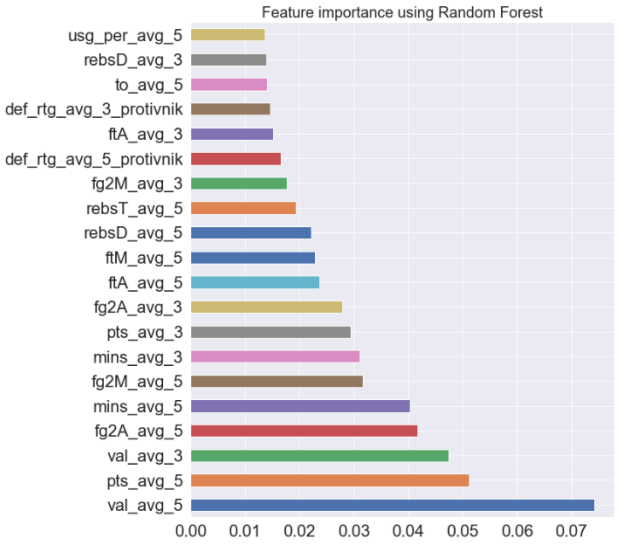
Na slici 1 mogu da se vide najbitnija obeležja koja je izdovjio *LightGBM* model, a na slici 2 ona koja je izdvojio model *Random Forest.*

Tabela 3: *Poređenje rezultata upotrebljenih modela.*

Model	*MAE* trening	*MAE* test	Opis
Linearna regresija/*Lasso*		5,617	Sa laso selekcijom obeležja
LightGBM	5,72	5,669	Sa optimalnim hiperparametrima
Random Forest		5,882	Sa standardno podešenih hiperparametrima
Random Forest	5,708	5,625	Sa optimalnim hiperparametrima



Slika 1: 20 najuticajnih obeležja koje je odabrao LightGBM model.



Slika 2: Najuticajnija obeležja koja je odabrao Random Forest algoritam.

U radu [13] najbolji rezultat je dobijen uz pomoć neuronskih mreža, pri čemu je srednja apsolutna greška iznosila *MAE* = 6.8805.

5. ZAKLJUČAK

Zadatak ovog rada je predikcija indeksa korisnosti košarkaša za *ABA* ligu. Indeks korisnosti predstavlja jednu od mera učinka igrača na utakmici i kao takva služi za bodovanje u fantazi igri koju organizuje *ABA* liga. Glavni problem u radu bilo je prikupljanje podataka, kreiranje odgovarajućeg skupa podataka, njihovo analiziranje, kao i kreiranje naprednih košarkaških statističkih kategorija, sve u cilju bolje predikcije broja indeksnih poena.

Rezultati koji su ostvareni su sasvim zadovoljavajući. Međutim, za dalji rad i unapređenje do sada urađenog najbitnije bi bilo proširenje skupa podataka *ABA 2010-2019*, kao i pronalazak i kreiranje dodatnih obeležja, odnosno njihovo detaljno analiziranje kakav uticaj imaju na igru, odnosno na kreiranje indeksa korisnosti. Da bi se došlo do ovoga potrebno je i da se sama liga više pozabavi domenom statistike.

6. LITERATURA

[1] NBA Team Values 2020, https://www.forbes.com/sites/kurtbadenhausen/2020/02/11/nba-team-values-2020-lakers-and-warriors-join-knicks-in-rarefied-4-billion-club/#2d8633322032

[2] Oliver, Dean. Basketball on paper: rules and tools for performance analysis. Potomac Books, Inc., 2004.

[3] Kubatko, Justin, et al. "A starting point for analyzing basketball statistics." Journal of Quantitative Analysis in Sports 3.3 (2007).

[4] Page, Garritt L., and Fernando A. Quintana. "Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates." Bayesian Analysis 10.2 (2015): 379-410.

[5] South, Charles, et al. "A Starting Point for Navigating the World of Daily Fantasy Basketball." The American Statistician 73.2 (2019): 179-185.

[6] FiveThirtyEight's Career-Arc Regression Model Estimator with Local Optimization, https://projects.fivethirtyeight.com/carmelo/

[7] Casals, Martí, and A. Jose Martinez. "Modelling player performance in basketball through mixed models." International Journal of Performance Analysis in Sport 13.1 (2013): 64-82.

[8] Cai W, Yu D, Wu Z, Du X, Zhou T. A hybrid ensemble learning framework for basketball outcomes prediction. Physica A: Statistical Mechanics and its Applications. 2019 Aug 15;528:121461.

[9] Thabtah F, Zhang L, Abdelhamid N. NBA game result prediction using feature analysis and machine learning. Annals of Data Science. 2019 Mar 7;6(1):103-16.

[10] Shreyas S. Shivakumar. "Learning to Turn Fantasy Basketball Into Real Money." https://shreyasskandan.github.io/Old_Website/files/report-ChanHuShivakumar.pdf

[11] Porter, Jack W. "Predictive Analytics for Fantasy Football: Predicting Player Performance Across the NFL." (2018).

[12] Lutz, Roman. "Fantasy football prediction." arXiv preprint arXiv:1505.06918 (2015).

[13] Kengo Arao, https://github.com/KengoA/fantasy-basketball/blob/master/report.pdf

[14] Beautiful soup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[15] Selenium Web Driver, https://www.selenium.dev/documentation/en/

[16] Scikit-learn, https://scikit-learn.org/

[17] f_regression, https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html#sklearn.feature_selection.f_regression

[18] Hyperopt, https://github.com/hyperopt/hyperopt

Kratka biografija

Miloš Nišić rođen je 1994. godine u Novom Sadu. Osnovnu školu "Desanka Maksimović" je završio u Futogu 2009. godine. Gimnaziju "Isidora Sekulić" u Novom Sadu je završio 2013. godine. Iste godine je upisao Fakultet tehničkih nauka, odsek Računarstvo i automatika. Zvanje diplomirani inženjer elektrotehnike i računarstva je stekao 2018. godine. Master akademske studije je upisao 2018. godine na istom studijskom programu na Fakultetu tehničkih nauka.

kontakt: milosnisic94@gmail.com