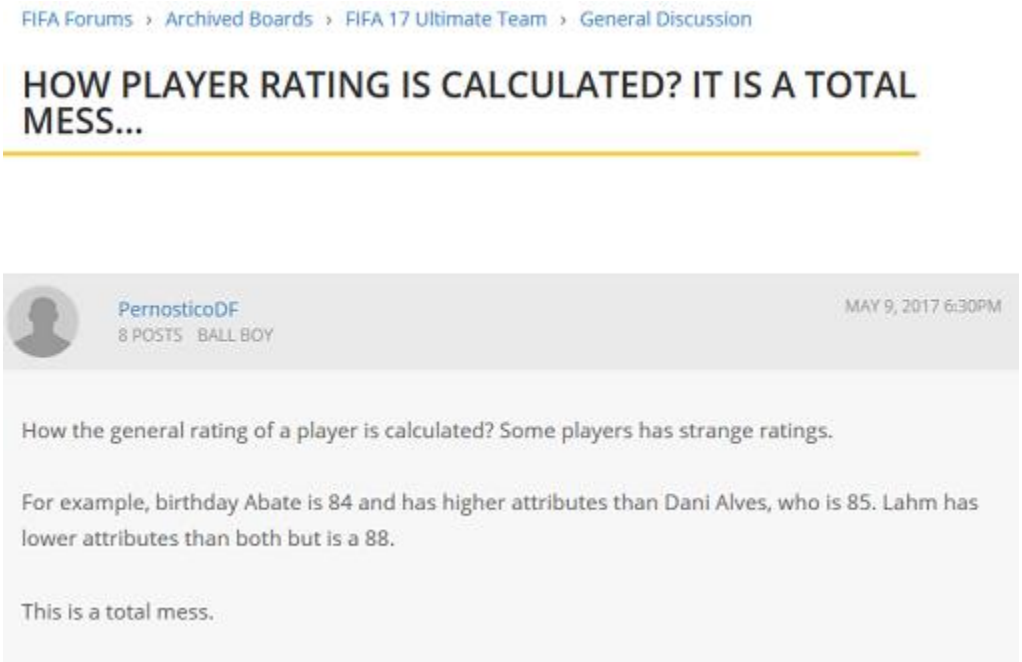
MONEYBALL

Projektni prijedlog iz kolegija Strojno učenje, ak.god. 2018./2019.
Prirodoslovno matematički fakultet Sveučilišta u Zagrebu
Profesor: dr.sc. Tomislav Šmuc
Asistenti: Tomislav Lipić, Matija Piškorec
Studenti: Ana Dugandžić, Ivan Zvonimir Kos, Lucija Marinčić, Lovro Sindičić
Travanj, 2019

1. Opis problema

Nogomet je već dugi niz godina najpopularniji sport na svijetu. FIFA kao najviša svjetska nogometna organizacija organizira svjetska prvenstva u nogometu. Po tome je napravljena računalna igrica FIFA, kojoj svake godine izlazi nova verzija. Uz igricu FIFA 2019, na KAGGLE-u je izašla i prikladna baza podataka za tu igricu, s podacima o nogometnim igračima. Među podacima o njima stoji i procjena njihove općenite vrijednosti, odnosno njihov *overall rating.* Problem koji mi pokušavamo riješiti je kako predvidjeti *overall rating* nogometnog igrača pomoću drugih podataka koje dobijemo o njemu.



Nismo jedini koji su se ovo zapitali.

Naša baza podataka sadrži 18000+ primjera, koje čine nogometni igrači, s 89 različitih značajki, kao što su *id* igrača, ime, nacionalnost, zastava, potencijal, novčana vrijednost, ime kluba, snaga udarca, brzina udarca i još mnogo drugih značajki vezanih na vještine

koje su tražene u nogometu. Ovim radom ćemo prikazati kako bismo mogli iz tih podataka izračunati *overall rating* nogometnog igrača.

data.csv (8.72 MB)

	ID	Name	Nationa...	Overall
1	158023	L. Messi	Argentina	94
2	20801	Cristiano Ronaldo	Portugal	94
3	190871	Neymar Jr	Brazil	92
4	193080	De Gea	Spain	91
5	192985	K. De Bruyne	Belgium	91
6	183277	E. Hazard	Belgium	91
7	177003	L. Modrić	Croatia	91
8	176580	L. Suárez	Uruguay	91

2. Cilj i hipoteza

U igrici svaki nogometaš dobije svoj tzv. *overall rating*, broj od 1 do 100, koji označava njegovu cjelokupnu vrijednost. Igrač se može osloniti na taj broj da bi recimo vidio gdje se Luka Modrić nalazi među top 100 nogometaša. Koliko je *overall rating* bitan u igrici je pitanje kojim se nećemo baviti, ali ćemo zato pokušati ustanoviti na koji način se preko ostalih podataka određuje *overall rating*.

Naša hipoteza je da postoji poveznica između ostalih podataka i *overall rating*-a, i da postoji način kako se ta vrijednost računa.

3. Pregled dosadašnjih istraživanja

Posljednjih desetak godina proveden je veći broj istraživanja u kojima se analizirala situacijska učinkovitost nogometaša i ekipa na utakmicama u igrici. Kao primjerom dosadašnjih istraživanja poslužili smo se također stranicom www.kaggle.com. Prije godinu dana objavljeni su podaci iz računalne igrice FIFA 2018. Podaci su sadržavali iste značajke. Korisnici KAGGLE-a su mogli s tom bazom podataka napraviti bilo kakvu analizu. Na skupu podataka mogle bi se analizirati performanse igrača na temelju mnogih atributa.

Primjerice, jedna osoba se bavila analizom najboljeg igrača grupiranih po godinama. Budući da su se neki igrači podudarali u godinama, pa čak i u *overall*-u, bilo je potrebno gledati još neke značajnije atribute po kojima je na temelju svih prosječnih vrijednosti određeno koji su bili najbolji igrači za 2018 godinu. Drugi primjer koji je sličan našem problemu je analiziranje igrice da bi se pronašla najbolja ekipa. U cjelini riječ je dobrom

projektu što pokazuju i pozitivni komentari, ali problem vrijedi pogledati iz malo više kuteva i probati nekolicinu različitih metoda.

4. Materijali, metodologija i plan istraživanja

Nadziranim učenjem ćemo naučiti računalo da na bazi preostalih značajki, tj. atributa, koje nogometni igrač ima, procjeni broj za njegov *overall rating.* Iz skupa od 18000 primjera ćemo najprije izdvojiti 10% primjera, koje ćemo ostaviti kao krajnji testni skup u lipnju. Na njemu ćemo obaviti evaluaciju modela. Preostale primjere ćemo podijeliti na skup za **trening**, skup za **validaciju** i skup za **testiranje**. Podjelu u skupove ćemo obaviti nasumičnim biranjem primjera. Valja napomenuti da se u tih 18000 primjera nalaze svi igrači, uključujući golmane, koji neke značajke ne dijele s ostalim igračima. Zbog toga ćemo izdvojiti golmane kao zaseban skup.

Budući da se bavimo procjenom jedne značajke na bazi preostalih značajki, zaključujemo da je u pitanju regresijski problem. Prema tome, najjednostavniji pristup rješavanju problema bi bio da koristimo model **linearne regresije**. Osim što je jednostavan pristup, naša analiza podataka upućuje na to da zaista postoji linearna korelacija. Za optimizaciju parametra koristit ćemo **gradijentni spust**. Pomoću njega ćemo procijeniti koji atributi su nam u stvari bitni za računanje *overall rating*-a. Zdravom logikom se može zaključiti da nešto poput nacionalnosti nije bitno, ali teško je procijeniti koliko je bitna brzina trčanja u odnosu na jačinu udarca nogom, itd. To ostavljamo algoritmu da odluči.
Za provjeru učinkovitosti modela koristit ćemo ***cross-validation*** algoritam za ispitivanje primjera. Točnije, koristit ćemo **k-fold** cross validation. Na kraju, što se tiče **evaluacije greške**, točnost linearne regresije ćemo ispitati pomoću standardne kvadratne pogreške, odnosno ***mean squared error***.

Drugi model koji možemo koristiti je ***K Nearest Neighbors***, odnosno možemo koristiti **k-NN algoritam** za predviđanje vrijednosti. Budući da nemamo premalo podataka, algoritam bi trebao biti dovoljno učinkovit, a budući da nemamo previše podataka, algoritam bi trebao biti dovoljno brz. Da bi izbjegli *overfitting*, koeficijent *k* pronaći ćemo pomoću ***elbow* metode**.

Još jedan način na koji možemo razmišljati kako bismo riješili traženi problem je ***Learning to rank*** ili ***machine-learned ranking*** (MLR).

Na kraju možemo koristiti model **stabla odlučivanja.** Iako se ono u većini slučajeva koristi za klasifikacijske probleme, postoje i regresijska stabla za slučaj poput našeg.

5. Očekivani rezultat projekta

Očekujemo da će se formula za računanje overall ratinga dobiti pomoću linearne regresije i da rješavanje problema neće biti previše problematično.

6. Literatura

[1] https://hr.wikipedia.org/wiki/FIFA ,travanj 2019.
[2] https://www.kaggle.com/karangadiya/fifa19, travanj 2019.
[3] Zaki & Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms (ch 3);
[4] https://www.kaggle.com/laowingkin/fifa-18-find-the-best-squad, travanj 2019.
[5] Andrew Ng, Machine learning, Coursera, travanj 2019.
[6]https://fifaforums.easports.com/en/discussion/277545/how-player-rating-is-calculated-it-is-a-total-mess?fbclid=IwAR1sylPJv6sbCHUrHY0TnmxfZ9BRqbb9ifZYtCQuiyB4uETJMCtA7vG6JHg, travanj 2019.