Univerzitet u Beogradu
Elektrotehnički fakultet

Projektni rad

Softversko Inženjerstvo Velikih Baza Podataka

Studenti:

Sanja Vujovic 17/3397

Milan Bojovic 17/3358

Profesor:

Prof. dr. Miroslav Bojović

Asistent: Stefan Tubić

Beograd, Januar 2018.

Sadržaj

1.Uvod

Ovaj dokument je namenjen svima koji žele bolje da razumeju na koji način je izrađen projekat iz predmeta Softversko inženjerstvo velikih baza podataka.

2.Projekat

Projektni zadatak je preuzet sa veb sajta „**kaggle.com**“ i nosi naziv „**Kobe Bryant Shot Selection**“. U zadatku se radi o tome da se na osnovu karijere duge dvadeset godina otkrije koji šut bi bio postignut, a koji ne.

Od podataka je dostupan .csv fajl po nazivom „**data.csv**“ koji u sebi sadrži 30698 šuteva na koš ovog poznatog košarkaša. Od 30698 šuteva, treba napraviti predikciju za 5000 šuteva koji su izostavljeni, tačnije za njih nisu upisane vrednosti.

U .csv fajlu se nalazi 25 kolona u kojima su neke vrlo bitne, a neke nisu toliko bitne za projekat, svaka od kolona će biti objašnjena pojedinačno.

Neke od kolona možemo posmatrati kao numeričke, a neke kao nominalne atribute. Kao numeričke atribute posmatramo sve kolone sa brojevima, dok kao nominalne atribute posmatramo sve kolone sa svim ostalim vrednostima(tekst, tekst i brojevi, oznake itd.).

Kolona	Objašnjenje
action_type	Tip akcije.
combined_shot_type	Kombinovanje akcije.
game_event_id	ID dogadjaja.
game_id	ID utakmice.
lat	Geografska širina.
loc_x	Koordinata X.
loc_y	Koordinata Y.
lon	Geografska dužina.
minutes_remaining	Minuta ostalo do kraja.
period	Četvrtina.
playoffs	Playoff.
season	Sezona.

seconds_remaining	Sekunde do kraja napada.
shot_distance	Razdaljina šutiranja.
shot_made_flag	Koš postignut.
shot_type	Tip šuta.
shot_zone_area	Zona šuta.
shot_zone_basic	Pozicija šuta.
shot_zone_range	Oblast šuta(15-20ft ili 20-25ft).
team_id	ID tima za koji igra igrač.
team_name	Ime tima igrača.
game_date	Datum utakmice.
matchup	Gostovanja(LAL@POR, LAL@CHI).
opponent	Protivnik.
shot_id	ID šuta.

Imena kolona označena crvenom bojom su za nas u čitavom procesu predikcije nebitna,a razlozi su sledeći:

game_event_id	ID dogadjaja - nije posebno bitan u našoj predikciji
game_id	ID utakmice - takođe nije posebno bitan u našoj predikciji.
minutes_remaining	Minuta ostalo do kraja – uvek je bitno koliko je sekundi ostalo do kraja, i kako organizovati napad u tim sekundama.
team_id	ID tima za koji igra igrač – Kobi Brajant je ceo život igrao za tim Los Angeles Lakers.
team_name	Ime tima igrača – objašnjenje isto kao za team_id.
shot_id	ID šuta – bitno za submit na kaggle.com, za nas je apsolutno nebitno.
shot_zone_range	Oblast šuta(15-20ft ili 20-25ft) – analizom razdaljine šutiranja, može se napraviti custom razdaljina sa koje je igrač bio manje ili više uspešan.

Priprema podataka kao input parametara nije potrebna, jer se sve odvija kroz alat Excel u kome se nalaze I izvorni podaci. Na izvorni set podataka se može gledati kao na set podataka sa nedostajućim vrednostima – vrednostima za koje radimo predikciju.

Pre bilo kog vršenja predikcije, podaci su podeljeni u tri seta, jedan gde su samo postignuti koševi, drugi gde su samo promašeni koševi, treći gde su I postignuti I promašeni koševi zajedno. Na ovaj način iskorišćena je metoda stabla, "podeli I vladaj", gde su se podaci filtrirali, tj. radila se klasifikacija I odmah se znalo za koliko šuteva se radi predikcija.

U većini slučajeva, podaci su prikazivani kroz pivot tabele I pivot chart, kao I kroz običan chart. Korišćen je model linearnog prikaza, recimo u delu gde se prikazuju samo postignuti I samo promašeni šutevi. Za sve predikcije su korišćena pravila koja imaju smisla, recimo nema smisla koristiti pravilo da je igrac u timu Los Angeles Lakers I da je Team ID: 1610612747, jer se ti podaci nikada ne menjaju I uvek bi se dobio čitav set podataka na izlazu.

Fajl sa podacima je detaljno analiziran I nije zaključeno da nedostaju podaci, osim pomenutih pet hiljada šuteva za predikciju i da su svi podaci u istom formatu gledajući po kolonama, pa samim tim nema potrebe da se vrši prepravka ili dopuna podataka.

Predikcija je vršena u pet pokušaja uz dva dodatna seta obradjenih podataka koji su služili kao moralna prekretnica pri odlučivanju.

Prva predikcija je radjena na osnovu razdaljine šutiranja. Na osnovu ispitivanja podataka došlo se do zaključka da se svi šutevi preko 40 fita sigurno promaše, dok se šutevi ispod 40 fita sigurno pogode. Prilično crno-bela predikcija, ali zasigurno i nerealna jer zdrav razum ne dozvoljava da svaki šut ispod 40 fita bude sigurno pogodjen. Ovim metodom se došlo do rezulata da od 5000 šuteva, samo 11 bude promašeno, a 4989 bude pogodjeno, što je procentualno pogodjeno 99,78%, a promašeno 0,22%.

Druga predikcija se zasniva na većem razlagu razdaljine šuta, kao i saznanju da li se igra playoff ili ne. Na osnovu detaljnije analize razdaljine šuta, došlo se do zaključka da ukoliko se igra playoff svi šutevi preko 30 fita budu sigurno promašeni, a ispod 30 fita budu sigurno pogodjeni, ukoliko je playoff svi sutevi ispod 27 fita budu sigurno pogodjeni, a iznad 27 fita sigurno promaseni. Opet crno-bela predikcija. Ove infromacije donose zaključak da se od 5000 šuteva pogodi 4966, tj. 99,32%, a promaši 34, tj. 0,68%. Predikcija malo bolja od prethodne, ali za mene lično i dalje nerealna.

Za prvu I drugu predikciju se dosta koristio pomoćni set podataka koji je prikazivao položaj igrača na terenu pomoću lon/lat, loc_x/loc_y koordinata, drugim rečima ovde su se koristila pravila asocijacija podataka po principu:

- Za prvu predikciju
    - if(shootrange>40) then miss
    - if(shootrange<40) then score
- Za drugu predikciju

- if(shootrange>30) then miss
- if(shootrange<30) then score

U kasnijim predikcijama su se koristila pravila asocijacija sa izuzecima.

Treća predikcija se naslanja na drugu predikciju I na poziciju šutiranja "Back Court", gde se koristi sve iz druge predikcije I pomoćni set podataka prethodno opisan. Rezulat je apsolutno isti kao I u drugoj predikciji.

Četvrta predikcija je po svojim rezultatima najrealnija. Zasniva se na akciji koja se izvodila dok se šutiralo, kao I koliko vremena je ostalo do kraja. U početku su vršene analize vremana u razmacima od po pet sekundi, zatim od deset I na kraju od dvadeset sekundi, gde su se za skoro svaku utakmicu dobili identični rezultati I doslo se do zakljucka da vreme treba uzimati kao celokupnu jedinicu tj. čitavih šezdeset sekundi.

U ovoj predikciji se došlo do sledećih rezultata:

Alle Oop Dunk Shot – 94% pogodaka
Alle Oop Layup Shot – 71% pogodaka
Cutting Layup Shot – 66% pogodaka
Driving Bank Shot – 66% pogodaka
Driving Dunk Shot – 97% pogodaka
Driving Finger Roll Layup Shot – 88% pogodaka
Driving Finger Roll Shot – 85% pogodaka
Driving Floating Bank Jump Shot – 100% pogodaka
Driving Floating Jump Shot – 33% pogodaka
Driving Hook Shot – 61% pogodaka
Driving Jump shot - 43% pogodaka
Driving Layup Shot - 74% pogodaka
Driving Reverse Layup Shot - 74% pogodaka
Driving Slam Dunk Shot - 97% pogodaka
Dunk Shot - 77% pogodaka
Fadeaway Bank shot - 88% pogodaka
Fadeaway Jump Shot - 57% pogodaka
Finger Roll Layup Shot - 82% pogodaka
Finger Roll Shot - 46% pogodaka
Floating Jump shot - 72% pogodaka
Follow Up Dunk Shot - 90% pogodaka
Hook Bank Shot - 100% pogodaka
Hook Shot - 36% pogodaka
Jump Bank Shot - 77% pogodaka

Jump Hook Shot – 73 % pogodaka
Jump Shot - 32% pogodaka
Layup Shot - 38% pogodaka
Pullup Bank shot - 54% pogodaka
Pullup Jump shot – 72% pogodaka
Putback Dunk Shot - 66% pogodaka
Putback Layup Shot – 66 % pogodaka
Putback Slam Dunk Shot - 50% pogodaka
Reverse Dunk Shot - 91% pogodaka
Reverse Layup Shot - 63% pogodaka
Reverse Slam Dunk Shot - 100% pogodaka
Running Bank shot - 83% pogodaka
Running Dunk Shot - 88% pogodaka
Running Finger Roll Layup Shot - 60% pogodaka
Running Finger Roll Shot – 25% pogodaka
Running Hook Shot - 87% pogodaka
Running Jump Shot – 74% pogodaka
Running Layup Shot - 70% pogodaka
Running Pull-Up Jump Shot - 66% pogodaka
Running Reverse Layup Shot - 57% pogodaka
Running Slam Dunk Shot - 100% pogodaka
Running Tip Shot – 0% pogodaka
Slam Dunk Shot - 98% pogodaka
Step Back Jump shot - 63% pogodaka
Tip Layup Shot - 50% pogodaka
Tip Shot - 35% pogodaka
Turnaround Bank shot - 79% pogodaka
Turnaround Fadeaway shot - 58% pogodaka
Turnaround Finger Roll Shot - 100% pogodaka
Turnaround Hook Shot - 50% pogodaka
Turnaround Jump Shot - 59% pogodaka

Što nas na prvi pogled procentualno dovodi do dosta realnijih rezultata nego kod prve i druge predikcije. Analizom rezultata se dolazi do zaključka da je čak 55.72% procenta šuteva promašeno, odnosno 2786 šuteva, dok je 44.28% pogodjeno, što je 2214 šuteva.

Peta predikcija se zasniva na tome da li je playoff, ko je protivnik, koja je sezona i četvrtina.

Kada se podaci razlože i uradi detaljnija analiza, dobije se sledeća tabela:

Protivnik	Procentualno promašeno	Procentualno pogodjeno
ATL	54.79452055	45.20547945

BKN	60	40
BOS	58.8761175	41.1238825
CHA	56.4	43.6
CHI	56.97674419	43.02325581
CLE	56.0311284	43.9688716
DAL	54.5982575	45.4017425
DEN	54.21597633	45.78402367
DET	55.87734242	44.12265758
GSW	53.54330709	46.45669291
HOU	56.54038599	43.45961401
IND	59.90415335	40.09584665
LAC	53.91061453	46.08938547
MEM	54.99425947	45.00574053
MIA	57.05996132	42.94003868
MIL	58.97435897	41.02564103
MIN	55.53732568	44.46267432
NJN	56.39810427	43.60189573
NOH	54.94736842	45.05263158
NOP	59.23344948	40.76655052
NYK	52.29681979	47.70318021
OKC	58.11051693	41.88948307
ORL	56.45695364	43.54304636
PHI	55.05804312	44.94195688
PHX	53.5504886	46.4495114
POR	53.48297214	46.51702786
SAC	53.47172513	46.52827487
SAS	56.34920635	43.65079365
SEA	54.75504323	45.24495677

TOR	53.5971223	46.4028777
UTA	55.57350565	44.42649435
VAN	52.94117647	47.05882353
WAS	57.28542914	42.71457086
PROCENTUALNO UKUPNO	55.81037509	44.18962491

Što dovodi do zaključka da je promašeno 2790 šuteva, tačnije 55.8% i pogodjeno 2209 šuteva, odnosno 44.18%.

3.Zaključak

Gledajući sve statistike I dodatne setove podataka iz svih predikcija, za ovog igrača je apsolutno svejedno sa koje je daljine šutirao, koje je sezone igrao, protiv koga je igrao, sa koje pozicije je šutirao, jer su rezultati pogadjanja I mašenja šuteva uvek bili veoma slični, razlika u par procenata. Upravo ovo je još jedan od dokaza zašto je bio jedan od najboljih strelaca tima Los Angeles Lakers.

Svi dobijeni rezultati su u odbircima proveravani na veb stranici “basketball-reference.com” koja sadrži prave rezultate utakmica, grafikone i statistike, gde se zaključilo da dobijeni rezltati varijaju u meri +- 2% do 4%.

4.Literatura

[1] www.kaggle.com

[2] Data Mining – Ian H. Witten, Eibe Frank, Mark A. Hall.

[3] http://www.basketball-reference.com/ - za proveru dobijenih rezultata sa pravim rezultatima.