Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Anže Kravanja

Napovedovanje zmagovalcev košarkarskih tekem

diplomsko delo

univerzitetni študij računalništva in informatike

Mentor: prof. dr. Igor Kononenko

Ljubljana 2012

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

[1] V dogovoru z mentorjem lahko kandidat diplomsko delo s pripadajočo izvorno kodo izda tudi pod katero izmed alternativnih licenc, ki ponuja določen del pravic vsem: npr. Creative Commons, GNU GPL.

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01883/2012
Datum: 03.12.2012

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ANŽE KRAVANJA**

Naslov: **NAPOVEDOVANJE ZMAGOVALCEV KOŠARKARSKIH TEKEM**
PREDICTING THE WINNERS OF BASKETBALL GAMES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Napovedovanje izida košarkarskih tekem je zelo težak problem, ki ga najbolje rešujejo z uporabo verjetnosti, ki izhajajo iz kvot stavnih borz. Naloga je uporabiti različne algoritme strojnega učenja za napovedovanje zmagovalcev košarkarskih tekem iz podatkov o preteklih odigranih tekmah. Pri tem naj kandidat poskuša sestaviti in uporabiti čimveč čimbolj informativnih atributov. Razvito metodologijo naj pretestira na realni bazi podatkov in rezultate primerja z dvema referenčnima modeloma, na podlagi deleža zmag v odigranih tekmah in na podlagi verjetnosti stavnih borz.

Mentor:

prof. dr. Igor Kononenko

Dekan:

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Anže Kravanja, z vpisno številko **63080010**, sem avtor diplomskega dela z naslovom:

Napovedovanje zmagovalcev košarkarskih tekem

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Igorja Kononenka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 18. januar 2013 Podpis avtorja:

Dejanu, Niki, Petru, Andreju, Nini, Matevžu in Luki hvala za nepozabna štiri leta, ki sem jih preživel z vami.

A journey of a thousand miles begins with a single step.

Kazalo

Povzetek

Cilj diplomske naloge je bil razviti napovedni model za napovedovanje košarkarskih tekem v ligi NBA. Problema smo se lotili z uporabo tehnik podatkovnega rudarjenja, ki obsega uporabo algoritmov strojnega učenja, umetne inteligence in statistike. V nalogi je predstavljen celoten potek dela, od začetnih obdelav t. i. podatkov "play-by-play", pa vse do testiranja in izbire končnega napovednega modela. Raziskali smo kakovost posameznih značilk, njihovo povezanost z razredom in pomoč pri napovedovanju. Pri testiranju smo preverili napovedi različnih algoritmov in izluščili tiste, ki so napovedovali najbolje, z njihovim združevanjem pa smo dobili končni model. Rezultate smo primerjali tudi s stavnimi kvotami, ki veljajo za najboljše napovedi. Ob koncu smo raziskali še težave, ki pestijo zgrajeni napovedni model, ponudili pa smo tudi možne izboljšave in predloge za nadaljnje delo.

Ključne besede

podatkovno rudarjenje, strojno učenje, košarka, napovedovanje zmagovalcev, športno napovedovanje

Abstract

The main goal of this thesis was to develop a prediction model for predicting basketball matches in the NBA. The problem was solved by using the techniques found in data mining, which include using machine learning algorithms, artificial intelligence and statistics. We present the whole path of development of such model from the initial parsing of so called "play-by-play" data, to the testing phase and choosing the final prediction model. We explored the qualitiy of attributes used, their connection to the class variable and help in prediction. While in testing phase, we tried different machine learning algorithms and we kept those whose predictions were promising. We combined the best algorithms in an ensemble to get even better results. We compared our results with betting odds, which give best predictions. At the end, we look deeper into the problems of our prediction model and gave suggestions about possible improvements and future work.

Keywords

data mining, machine learning, basketball, winner predictions, sports forcasting

Poglavje 1

Uvod

1.1 Motivacija

Šport ni imel nikoli veliko skupnega z računalništvom, vendar časi se spreminjajo in z naraščanjem prakse zbiranja podatkov o tekmah, se je pojavila tudi potreba po vse hitrejši obdelavi teh podatkov. To pa je kot nalašč za računalnik, ki lahko te podatke učinkovito ter v kratkem času obdela in vrne zanesljive informacije, za katere bi človek, da jih izlušči in uredi, potreboval precej več časa. Sprva je bil računalnik namenjen le preprostim operacijam, kot sta seštevanje in povprečenje. Potem so prišli statistiki, ki so se zanimali za določen šport in so v množicah podatkov iskali različne indekse, s katerimi bi ocenili igralce in ekipe. Kar nekaj časa je trajalo, da so se statistiki uveljavili v svetu športa, vendar jim je to vsekakor uspelo. Praktično na katerokoli športno spletno stran gremo, vidimo, da se v športnih analizah tekem skoraj vedno pojavljajo raznorazni indeksi in faktorji, ki so kombinacije vsega mogočega, predvsem z namenom, da bi več različnih podatkov o tekmi združili v eno samo številko, ki naj bi nam povedala, kateri igralec je boljši od drugega. Danes so statistični podatki o odigranih tekmah na voljo domala vsem in tudi trenerji vrhunskih ekip jih uporabljajo za pripravo na naslednje tekme. Tako se poraja vprašanje, če si vrhunski trenerji, ki so v določenem športu že mnogo let, pomagajo z golimi številkami ali so te številke same po

sebi dovolj informativne, da bi lahko iz njih napovedali zmagovalca tekme brez prisotnosti človeškega faktorja. Odgovor je jasen. So! Vendar, kako dobro lahko zmagovalce napovemo?

1.2 Cilj diplomske naloge

V diplomski nalogi smo poskušali napovedati zmagovalce košarkarskih tekem. Košarka je šport, kjer tekom posamezne tekme lahko beležimo ogromno podatkov, zato je zelo primerna za napovedovanje. Iz teh podatkov si namreč želimo pridobiti čim več različnih značilk, ki opisujejo tekmo, izmed vseh pa lahko nato izberemo tiste najboljše. Napovedovanje je narejeno s pomočjo tehnik odkrivanja znanj iz podatkov (ang. data mining), ki združuje različna področja računalništva, kot so strojno učenje, umetna inteligenca in statistika. Pod drobnogled smo vzeli ameriško profesionalno košarkarsko ligo NBA. Ta liga je bila izbrana iz dveh preprostih razlogov. Prvi je ta, da je to najbolj gledana košarkarska liga na svetu, drugi pa, da v njej vsaka ekipa odigra vsaj 82 tekem na sezono, kar je v primerjavi z ostalimi ligami precej več. Več tekem seveda pomeni več podatkov, več podatkov pa lahko vodi k boljši napovedi. Cilj je torej čim bolje napovedati zmagovalca na posamezni tekmi lige NBA. V diplomski nalogi smo poizkušali ugotoviti, do katere mere lahko naše napovedi približamo napovedim stavnih hiš, kjer stavne kvote določajo skupine analitikov, ki se z ligo NBA ukvarjajo vsak dan. Zanima nas torej ali lahko računalnik napove zmagovalce košarkarskih tekem v ligi NBA bolje od človeka.

1.3 Kratek opis vsebine

V 2 poglavju smo na kratko opisali pretekla dela na področju športnega napovedovanja v košarki, predstavljena pa so tudi orodja, s pomočjo katerih je bila diplomska naloga narejena.

3 poglavje predstavlja začetno točko gradnje vsakega napovednega mo-

dela. Razložili smo, katere podatke smo uporabili in kako smo jih pripravili za nadaljnjo obdelavo.

Grajenje učnih primerov smo opisali v 4 poglavju. Razložili smo vse uporabljene značilke in kako do njih priti. Zadnji del poglavja je posvečen gradnji učnih primerov, torej kako smo dobili predstavitev ene tekme, ki jo razumejo učni algoritmi.

V 5 poglavju smo naprej predstavili način testiranja, s katerim smo dobili najboljši model, čemur sledi podrobnejši opis dobljenega modela in rezultati testiranja. Za konec poglavja pa smo naredili še primerjavo svojih napovedi z napovedmi stavnic.

V sklepnih ugotovitvah (6 poglavje) smo poskušali oceniti narejeno in predlagati, kaj bi se še dalo izboljšati.

Poglavje 2

Pretekla dela in uporabljena orodja

2.1 Pretekla dela

V preteklosti so se, precej več kot računalničarji, s košarkarskimi podatki ukvarjali statistiki. Večina moštev NBA ima v svojih pisarnah zaposlene statistike, ki poskušajo v poplavi podatkov najti najmanjše detajle, ki bi njihovi ekipi pomagale do zmage. Trend statistikov v športu se je znatno povečal po uspehu ekipe Oakaland Athletics, ki sicer spada med bejzbolska moštva. Tam so prvi začeli z izčrpno statistično analizo igralcev in prišli do ugotovitve, da morajo ekipo sestaviti tako, da bodo optimizirali določene v bejzbolu pomembne statistične elemente. Kljub temu, da so imeli leta 2002, ko so s tem projektom začeli, tretji najmanjši znesek namenjen plačam igralcev v celotni ligi, so uspeli priti globoko v izločilne boje, kjer jih nihče ni pričakoval. Kmalu za tem uspehom so vse ekipe v ameriški bejzbolski ligi (MLB) v svoj način sestave ekipe uvedle ta sistem. Zgodba je lepo prikazana v filmu *Moneyball (2011)*.

Poskus napovedovanja košarkarskih tekem je že bil predmet preučevanj. Večina literature na to temo v povezavi s podatkovnim rudarjenjem, pa je bila napisana bolj za hobi. S tem ne mislimo, da avtorji niso izbrali pravih

metod ali niso imeli dovolj znanja, vendar malo je literature, v kateri bi se avtorji dejansko poglobili v problem in ga nato poskušali rešiti. Namesto tega najdemo veliko "zabavnih poskusov" napovedovanja. V [7] je omenjeno, da imajo do zdaj razviti modeli mnoge pomanjkljivosti. Dodaja tudi, da naj bi bila zgornja meja napovedne točnosti okoli 70%; okoli tega odstotka se gibljejo tudi napovedi iz stavnih hiš.

Boljši članki na to temo izhajajo iz poskusa simuliranja in preko tega napovedi zmagovalca tekme. Shirley [6] je uporabil Markovske verige za določanje prehodov ekipe iz stanja v stanje. Podobno in še malce naprednejšo uporabo Markovskih verig sta pokazala Štrumbelj in Vračar [5], kjer sta s poskusom simuliranja tekme napovedala zmagovalca. Njun model je imel točnost blizu "magičnih" 70%. Najdemo lahko tudi napovedi s sistemom ELO, ki je namenjen za rangiranje igralcev šaha. Dobro prilagojen sistem ELO lahko vrača solidne napovedi [5].

Napovedovanje kot je predstavljeno v tej nalogi, torej z uporabo podatkovnega rudarjenja, do zdaj ni bilo ravno močno zastopano v raziskavah; tudi zaradi tega smo se odločili, da raziščemo še ta način in ugotovimo ali je lahko uspešen.

2.2 Uporabljena orodja

Vsa koda uporabljena za izdelavo te diplomske naloge od začetka do konca je bila napisna v programskem jeziku Python. Python je bil izbran zaradi njegove preprostosti dela z matematičnimi objekti, kot so vektorji in matrike, v veliki meri pa tudi zaradi prejšnjih pozitivnih izkušenj s tem jezikom in odlično podprtjostjo jezika s knjižnicami za strojno učenje.

Večino postopkov in algoritmov smo napisali sami s pomočjo Pythonovih nepogrešljivih matematičnih knjižnic Numpy in Scipy. Pomagali pa smo si tudi z algoritmi iz knjižnic za strojno učenje. V največjo pomoč nam je bila knjižnica *scikit-learn*, uporabljena pa je bila tudi knjižnica *Orange* `orange.biolab.si`], ki je bila izdelana na naši fakulteti. Orange nam je

najboj pomagal pri delu z značilkami, saj ima odlične metode za shranjevanje učnih podatkov, njihovo izbiranje in transformiranje. Iz scikit-kearn-a pa smo vzeli večino algoritmov, ki jih sami nismo napisali, predvsem zaradi njihove kompleksnosti in hitrosti scikit-ove implementacije. Večkrat smo uporabili tudi knjižnico matplotlib za izris grafov in histogramov, ki jih lahko najdete na naslednjih straneh.

scikit-learn - `scikit-learn.org` Je odprtokodna knjižnica namenjena strojnemu učenju za programski jezik Python. Vključuje večino popularnih algoritmov za klasifikacijo, regresijo in deljenje (ang. clustering).Trenutna stabilna različica nosi oznako 0.12.

Orange - `orange.biolab.si` Podobno kot scikit-learn je tudi Orange odprtokodna knjižnica namenjena strojnemu učenju za programski jezik Python. Poleg klasičnih algoritmov, ki jih najdemo tudi drugje, ima Orange tudi veliko podporo za vizualizacjo podatkov, premore pa tudi grafični vmesnik, preko katerega lahko upravljamo vizualizacije in preprostejša opravila v strojnem učenju. Kot že omenjeno, je Orange nastal na naši fakulteti pod vodstvom Laboratorija za bioinformatiko. Trenutna verzija je 2.5.

Poglavje 3

Opis podatkov

3.1 Struktura podatkov

V diplomski nalogi smo uporabili podatke tipa play-by-play. Za razliko od klasične oblike agregiranih podatkov v tabeli (ang. box score), ki jih najdemo na večini športnih strani, podatki play-by-play ne združujejo podatkov, ampak se preprosto vsak dogodek, ki se na tekmi zgodi zabeleži v vnaprej določeni obliki. Taki podatki so pri tej diplomski nalogi bolj zaželeni, saj ne vidimo le stanja po koncu tekme, ampak lahko iz takih podatkov izluščimo tudi dogajanje med tekmo, ki je lahko še bolj pomembno, kot pa stanje po tekmi.

Podatke za to diplomsko nalogo smo dobili iz spletne strani `http://www.basketballgeek.com/data/`. Na voljo so nam tri sezone rednega dela lige NBA(sezona 2007/08, 2008/09 in 2009/10), kar pomeni 3573 tekem, s pomočjo katerih smo razvijali model. Prednost takih podatkov so tudi izpisane koordinate ob izvedenih metih (glej sliki 3.1 in 3.2 ter tabelo 3.1).

Vse odigrane tekme so podane v tekstovnih datotekah v formatu csv. Vsaka datoteka opisuje potek ene tekme. Iz imena datoteke lahko razberemo datum tekme in ime domače in gostujoče ekipe. Vsak dogodek na tekmi je podan v svoji vrstici. Prva vrstica v datoteki predstavlja glavo, ki vsebuje imena podatkov, ki so zapisani v tem stolpcem. Pri določenih dogodkih so

PHOENIX SUNS (7-11)

			FIELD GOALS				REBOUNDS									
	POS	MIN	FGM-A	3PM-A	FTM-A	+/-	OFF	DEF	TOT	AST	PF	ST	TO	BS	BA	PTS
M. Beasley	F	21:10	4-10	1-1	0-0	-20	1	3	4	2	0	2	2	0	2	9
M. Morris	F	18:21	2-6	1-4	1-1	-7	1	2	3	3	4	0	3	0	0	6
M. Gortat	C	41:23	8-11	0-0	2-4	-1	1	9	10	1	3	0	4	0	0	18
S. Brown	G	35:28	6-16	1-6	4-5	-8	3	4	7	3	3	1	1	0	2	17
G. Dragic	G	26:28	4-9	1-4	0-1	-11	1	3	4	5	2	0	2	0	0	9
J. Dudley		24:32	3-7	2-4	2-3	+11	0	1	1	2	2	1	0	0	0	10
P. Tucker		25:42	5-6	0-1	0-0	0	1	4	5	2	2	1	0	0	0	10
L. Scola		21:48	3-8	0-0	3-3	+1	2	6	8	0	2	0	3	0	1	9
S. Telfair		25:08	4-7	2-3	1-2	0	1	2	3	3	5	0	2	0	0	11
D. Garrett	DNP - COACH'S DECISION															
W. Johnson	DNP - COACH'S DECISION															
J. O'Neal	DND - STRAINED RIGHT QUAD															
L. Zeller	DNP - COACH'S DECISION															
Totals		**240**	**39-80**	**8-23**	**13-19**		**11**	**34**	**45**	**21**	**23**	**5**	**17**	**0**	**5**	**99**
			48.8%	34.8%	68.4%		TEAM REBS: 7						TOTAL TO: 17			

Slika 3.1: Prikaz podatkov o tekmi s pomočjo tabele (ang. box score) za eno od udeleženih ekip.

PHOENIX SUNS (7-11)		NEW YORK KNICKS (12-4)
START OF 1ST QUARTER		
(12:00) JUMP BALL CHANDLER VS GORTAT (MORRIS GAINS POSSESSION)		
Beasley Driving Jump shot: Missed	11:42	
	11:41	Thomas Rebound (Off:0 Def:1)
	11:23	Anthony Jump Shot: Missed
Dragic Rebound (Off:0 Def:1)	11:22	
Dragic Turnover : Traveling (1 TO)	11:15	
	11:01 [NYK 2-0]	**Felton Jump Shot: Made (2 PTS) Assist: Anthony (1 AST)**
Beasley Jump Shot: Made (2 PTS) Assist: Morris (1 AST)	10:41 [PHX 2-2]	
	10:29 [NYK 4-2]	**Anthony Jump Shot: Made (2 PTS)**
Gortat Turnaround Fadeaway shot: Made (2 PTS)	10:13 [PHX 4-4]	
Morris Foul: Shooting (1 PF) (2 FTA)	10:04	
	10:04 [NYK 5-4]	**Chandler Free Throw 1 of 2 (1 PTS)**
	10:04 [NYK 6-4]	**Chandler Free Throw 2 of 2 (2 PTS)**
Brown Foul: Offensive Charge (1 PF)	09:51	
Brown Turnover : Foul (1 TO)	09:51	

Slika 3.2: Primer play-by-play podatkov z začetka tekme. Vsak dogodek se zabeleži, razvidno je tudi, kateri ekipi dogodek pripada in kdaj se je zgodil.

time	team	etype	assist	player	points	reason	result	steal	type	x	y
11:31	CLE	shot	Shaquille O'Neal	Anderson Varejao	2		made		step back jump	21	15
11:12	BOS	shot	Ray Allen				missed		jump	4	14
11:10	CLE	rebound	Shaquille O'Neal						def		
11:03	CLE	shot		LeBron James	2		made		jump	12	25
10:46	BOS	shot		Kevin Garnett	2		made		running hook	21	14
10:26	CLE	shot	Mo Williams	Anderson Varejao	2		made		jump	22	14
10:13	BOS	turnover		Ray Allen		bad pass		Anthony Parker			

Tabela 3.1: Nekaj začenih vrstic tekme med Clevelandom in Bostonom. Prikazana je struktura podatkov. Zaradi lepšega prikaza so nekateri stolpci izpuščeni.

lahko nekateri stolpci tudi prazni. Vseh stolpcev je 32.

a1,..,a5 Imena in priimki gostujoče peterke, ki je v tistem trenutku na igrišču.

h1,..,h2 Imena in priimki domače peterke, ki je v tem trenutku na igrišču.

period Pove nam v kateri četrtini se je zgodil dogodek. Običajno ima vrednosti od 1 do 4, v primeru podaljška lahko 4 plus številka podaljška.

time Preostali čas do konca četrtine. Zapisan je v obliki MM:SS.

team Okrajšava ekipe, ki je izvedla akcijo opisano v vrstici. Izjemen primer je, če žoga ni v posesti nobene ekipe, takrat je atribut team enak 'OFF'.

etype Opis dogodka na tekmi (*timeout, shot, foul, ejection, jump ball, rebound, sub, free throw, violation, turnover*)

assist Ime igralca, ki je asistiral. Ime igralca je napisano le ob dogodku *shot*.

away Ime gostujočega igralca, ki skače za žogo. Napisano le ob dogodku *jump ball*

block Ime igralca, ki je izvedel blokado. Napisano le ob dogodku *shot*

entered Ime igralca, ki vstopa v igro. Napisano le ob dogodku *sub*

home Ime domačega igralca, ki skače za žogo. Napisano le ob dogodku *jump ball*

left Ime igralca, ki odhaja iz igre. Napisano le ob dogodku *sub*

num Številka prostega meta. Napisano le ob dogodku *free throw*

opponent Ime igralca nad katerim je storjen prekršek. Napisano le ob dogodku *foul*

outof Dosojeno število prostih metov. Napisano le ob dogodku *free throw*

player Ime igralca, ki je izvedel akcijo oziroma mu trenutni dogodek pripada.

points Število doseženih točk ob dogodku.

possession Ime igralca, ki je ujel žogo ob dogodku *jump ball*

reason Vzrok dogodka. Ob dogodkih *free throw, turnover,...* Primer *foul, bad pass.*

result Izid meta na koš. Dve možnosti *missed, made.*

steal Ime igralca, ki je ukradel žogo. Ob dogodkih *lost ball, bad pass.*

type Dodatni opis akcije oziroma dogodka. Primer *running layup, hook, slam dunk, layup, jump ball, kicked ball, 3pt,...*

x,y Koordinate na igrišču s katerih je igralec sprožil met. Napisano le ob dogodku *shot*

3.2 Priprava podatkov

Neobdelani, surovi podatki nam v kasnejših fazah ne koristijo, zato je treba iz njih izluščiti atribute, ki nas zanimajo. Vsaka tekma je bila do sedaj predstavljena kot datoteka z vrsticami, ki predstavljajo dogodke na tekmi. Za učinkovitejše nadaljnje delo je treba omenjen zapis strniti, in sicer tako,

da bo dostop do podatkov o vsaki tekmi hiter in preprost. V duhu objektnega programiranja smo podatke pripravili na tak način, da je vsaka tekma predstavljena z razredom. Razred pa seveda vsebuje vse atribute, ki smo jih izluščili iz vrstic z dogodki.

Vsaka tekma je vsebovala naslednje atribute:

Ukradene žoge Predstavlja število ukradenih žog celotne ekipe na celi tekmi.

Podaje Predstavlja število podaj celotne ekipe na celi tekmi.

Izgubljene žoge Predstavlja število izgubljenih žog celotne ekipe na celi tekmi.

Blokade Predstavlja število blokad celotne ekipe na celi tekmi.

Prekrški Predstavlja število prekrškov celotne ekipe na celi tekmi.

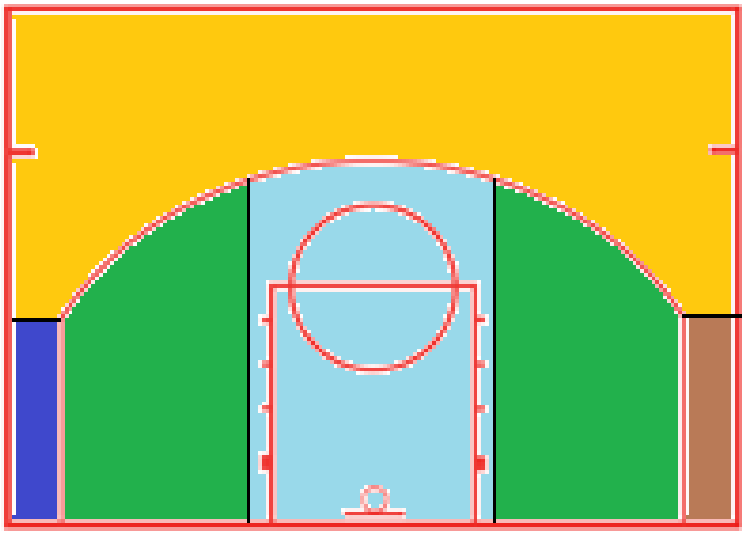
Zadeti meti in poskušani meti Pod to kategorijo štejemo več atributov. Beleži se število zadetih in poskušanih prostih metov, število zadetih in poskušanih metov za dve točki in število zadetih in poskušanih metov za tri točke.

Skoki Tudi tu se beleži več atributov. Število skokov v napadu, obrambi in poseben primer skoka imenovan ekipni skok (ang. team rebound); za vse tri velja, da predstavljajo število skokov celotne ekipe na celi tekmi.

Sprememba rezultata To je poseben atribut, ki tekmo razdeli na intervale po šest minut. V vsakem takem obdobju beležimo število doseženih točk celotne ekipe.

Posesti Beleži število posesti ekipe na celi tekmi. Ena posest predstavlja trenutek, ko ekipa dobi žogo v svojo posest do trenutka, ko posest žoge izgubi oziroma posest nad žogo prevzame druga ekipa.

Meti To je skupek atributov, ki beleži uspešnost meta ekipe iz različnih položajev na igrišču. Igrišče je razdeljeno kot prikazuje slika 3.3 na 6 delov. Za vsak del imamo podatek, kakšna je bila uspešnost meta.



Slika 3.3: Prikaz razdelitve igrišča na 6 delov. Trije deli beležijo uspešnost meta za dve točki, ostali trije pa met za tri točke.

Dosežene točke po tem, ko je nasprotna ekipa izgubila žogo Predstavlja število točk, ki jih ekipa doseže, po tem ko je nasprotna ekipa izgubila žogo, bodisi zaradi lastne napake (slaba podaja) ali pa ji je druga ekipa žogo ukradla.

Bonus Predstavlja štiri atribute za vsako četrtino. Beleži koliko časa je ekipa potrebovala, da je dosegla bonus. Ekipa je v bonusu, ko naredi 4 prekrške ali več.

Minute igranja Beleži koliko minut so skupaj odigrali igralci, ki so tekmo začeli v prvi peterki in koliko minut so igrali igralci, ki so tekmo začeli na klopi.

Pomembno je dodati, da se vsak zgoraj opisani atribut beleži dvakrat. Enkrat za domačo ekipo drugič pa za gostujočo ekipo. Beseda atribut se bo vedno nanašala na zgoraj opisane oznake, beseda značilka pa bo vedno pomenila skupek enega ali več atributov povprečenih ali s kako drugo metodo pridobljenih števil. Značilka (ang. feature) je že del učnega modela

oziroma učnih podatkov, s pomočjo katerih se algoritmi učijo. Kar je bilo v tem poglavju opisano, so bili atributi posamezne tekme, le-ti pa so kasneje uporabljeni za gradnjo značilk.

Iz analize v tabeli 3.2 lahko predpostavljamo, kateri atributi bodo gradili boljše značilke. Vzeli smo vse odigrane tekme in preverili, kakšen je odstotek zmag ekip, ki imajo določen atribut večji od nasprotne ekipe. Izkaže se, da ekipe, ki imajo skupni odstotek meta boljši od nasprotne ekipe zmagajo kar v 76.5% primerov. To ni presenetljivo, saj je bistvo košarke zadeti koš, če ti to uspe bolj natančno kot nasprotni ekipi, imaš jasno precej visoke možnosti za zmago. Nasploh lahko vidimo, da so vsi atributi povezani z metom blizu vrha tabele. Vsak trener ve, kako pomembni so skoki v obrambi. Ekipa si znatno poveča možnosti zmage, če uspe biti boljša v skoku v obrambi. Manj pomembni elementi košarkarske igre pa očitno ležijo v skoku v napadu, blokadah in ukradenih žogah.

Atribut	Odstotek zmag
Odstotek vseh metov	76.5%
Odstotek meta za dve točki	71.7%
Skok v obrambi	70.9%
Podaje	68.7%
Odstotek meta iz rakete	67%
Odstotek meta za tri točke	65.9%
Prekrški (manj)	65%
Izgubljene žoge (manj)	62%
Odstotek zadetih prostih metov	56%
Ukradene žoge	55.2%
Blokade	53.8%
Ekipa, ki kasneje izpolne bonus v 4. četrtini	51%
Skok v napadu	44.9%

Tabela 3.2: Odstotek zmag ekip, ki so na odigrani tekmi imele določen atribut večji od nasprotne ekipe razen pri atributih, ki imajo v oklepaju napisano manj. Ti atributi so negativni, kar pomeni, da je za ekipo slabše, če je število atributa večje. Analiza obsega tekme vseh treh sezon.

Poglavje 4

Grajenje učnih primerov

4.1 Opis značilk

Korak opisan v tem poglavju, je izjemnega pomena. Brez dobrih informativnih značilk napovedni model ne more dobro napovedati zmag, če pa uspemo najti dobre značilke, se lahko nadejamo boljšega napovedovanja. Namen iskanja značilk je, da poskušamo z njimi čim bolje opisati kakovost določene ekipe. Za to potrebujemo veliko značilk, pridobljenih iz različnih elementov in pogledov košarkarske igre. Take značilke bodo bolj koristne pri ločevanju med dobrimi in slabimi ekipami. Kadar napovedujemo tekmo, ne vemo vnaprej, koliko skokov, asistenc ali izgubljenih žog bo ekipa imela, saj tekma še ni bila odigrana. Značilke zaradi tega gradimo na podatkih že odigranih tekem, s pomočjo katerih dobimo o ekipi določeno sliko. V vsaki tekmi sodelujeta dve ekipi, katerih značilke združimo, primerjava le-teh pa nam pomaga pri napovedovanju izida tekme oziroma nam pokaže potencialnega zmagovalca.

V diplomski nalogi smo preverjali mnogo različnih značilk. Vse niso bile dobre. Take značilke smo izločevali, saj napovedovanju ne bi koristile oziroma bi jim celo škodile. Izbiro značilk (ang. feature selection) smo delali s pomočjo vzvratnega odstranjevanja (ang. backward elimination). Pri tem načinu iz množice vseh značilk vedno odstranimo tisto, ki ima najslabši rezultat izbrane mere. Nato postopek ponovimo in ga ponavljamo, dokler vre-

dnost izbrane mere pri najslabšem atributu ne presega neke vnaprej določene vrednosti. Za mere kakovosti značilk smo izbrali Pearsonovo korelacijo, informacijski prispevek in naprednejšo mero ReliefF [1], ki ne gleda le posameznega atributa, ampak tudi odvisnosti med njimi. Na koncu je bilo izbranih 39 značilk za vsako ekipo, ki so na naslednjih straneh podrobneje opisane.

Značilke so razdeljene na zunanje in notranje. Delitev smo naredili zato, da bo njihova predstavitev lažja in bolj smiselna in se ne navezuje na nobenega od izrazov v literaturi.

4.1.1 Zunanje značilke

Zunanje značilke so tiste, ki opisujejo ekipo "od zunaj". S tem je mišljeno, da ne gledamo njihovih iger in dosežkov znotraj tekme, ampak le to, kar se, da videti na površju. Primera takih značilk sta odstotek zmag ekipe in število zaporednih zmag ali porazov pred tekmo, ki jo napovedujemo.

Odstotek zmag ekipe Vsaka ekipa je do nekega trenutka odigrala že nekaj tekem. Odstotek zmag je število zmag ekipe na teh tekmah ulomljeno z vsemi odigranimi tekmami.

Število domačih zmag / število gostujočih zmag Značilka je zelo podobna prvi, le da tu upoštevamo le domače tekme. V primeru, da ekipa naslednjo tekmo gostuje, pa upoštevamo le gostujoče tekme. Tako dobimo odstotek zmag, ki jih ima ekipa na domačem ali gostujočem igrišču.

Število zaporednih zmag ali porazov Pove koliko zaporednih zmag ali koliko zaporednih porazov je ekipa imela do tega trenutka.

Napadalni rating (ang. offensive rating) Nekoliko kompleksnejša značilka, ki je sestavljena iz dveh atributov tekme. Izračuna se na sledeči način:

$$\text{Napadalni rating} = \frac{\text{Število doseženih točk na tekmi}}{\text{Število posesti na tekmi}} \times 100$$

Izračun nam pove, koliko točk bo ekipa v povprečju dosegla na 100 posesti žoge. Razlaga je preprosta; ekipa, ki je sposobna svojo posest bolje izkoristiti, to je doseči več točk, bo bolj uspešna.

Obrambni rating (ang. defensive rating) Podobno kot napadalni rating, tudi ta značilka združuje dva atributa. Poleg posesti še število doseženih točk nasprotnika.

$$\text{Obrambni rating} = \frac{\text{Število prejetih točk na tekmi}}{\text{Število posesti nasprotne ekipe na tekmi}} \times 100$$

Izračun nam pove, koliko točk ekipa v povprečju prejme na 100 posesti nasprotne ekipe. To si, podobno kot prej, lahko razlagamo na način, da ekipa, ki prejme povprečno manj točk na posest nasprotne ekipe, igra boljšo obrambo in je zaradi tega lahko uspešnejša.

Ratings Percentage Index (RPI) Indeks je namenjen ocenjevanju moči ekipe v nekem času. Sestavljen je iz treh uteženih komponent. Prva je klasični odstotek zmag ekipe na preteklih tekmah s tem dodatkom, da je domača zmaga utežena z utežjo 0.6, gostujoča zmaga pa z 1.4. S tem nagradimo ekipe, ki uspejo zmagati tudi v gosteh, kjer je običajno težje zmagovati. Druga komponenta je povprečen odstotek zmag brez uteži ekip proti katerim je ekipa, ki jo ocenjujemo, igrala. Tretja komponenta pa je povprečni odstotek zmag ekip v medsebojnih igrah, s katerimi je igrala ekipa, ki jo ocenjujemo.

$$\begin{aligned} RPI = \quad & 0.25 \times UtezenOdstotekZmag \\ & +0.5 \times OdstotekZmagNasprotnikov \\ & +0.25 \times NasprotnikovOdstotekZmagNasprotnikov \end{aligned}$$

RPI je na splošno zelo dobra značilka, saj nasprotno od klasičnega odstotka zmag, nagradi gostujoče zmage, upošteva pa tudi kakovost ekip, s katerimi so bile odigrane tekme.

4.1.2 Notranje značilke

Z notranjimi značilkami smo poskušali opisati notranje delovanje ekipe in jo oceniti glede na dosežke na tekmi. Ekipo opisujejo oziroma ocenjujejo s stališča, dobrega delovanja ekipe kot celote. Zato so vse notranje značilke izdelane iz atributov tekme, ki jih ekipa doseže kot skupina. Boljša, kot je skupina, boljše bodo značilke v primerjavi s slabšimi ekipami.

Štirje dejavniki

Štirje dejavniki predstavljajo štiri elemente košarkarske igre, ki prinašajo zmago. Večina trenerjev stalno opozarja igralce na te elemente igre. Trenerji so do teh ugotovitev prišli v času vodenja mnogoterih tekem in preko nabranih izkušenj skozi leta dela v košarki. Statistik Dean Oliver, ki je kot analitik delal tudi za nekatera moštva NBA, je v svoji knjigi *Basketball on paper* [3] s statističnimi metodami pokazal, da mit o štirih dejavnikih drži.

Vsaka ekipa, ki si želi povečati možnosti za zmago, mora poskrbeti za te štiri elemente igre. Prvi in najpomembnejši del je met na koš. Z večjo uspešnostjo, kot lahko ekipa doseže koše, več ima možnosti za zmago. Popolnoma logična predpostavka. Drugi element so izgubljene žoge. Če si hoče ekipa omogočiti dobro organiziran napad za dosego koša, žog ne sme po nepotrebnem izgubljati, saj se v tem primeru napad oziroma posest konča brez meta na koš. Brez meta pa je možnost zadetega koša enaka nič. Tretji element so skoki. V to so všteti tako obrambni kot napadalni skoki. Spet je razlaga dokaj enostavna in intuitivna. S skokom v obrambi ekipa pridobi posest žoge in ima posledično možnost dosega koša. V primeru napadalnega skoka sicer ekipa ne pridobi nove posesti, ker jo je imela že prej, vendar pa jo s tem obdrži oziroma podaljša, kar ji zopet nudi novo možnost dosege koša. Četrti dejavnik pa predstavlja uspešnost ekipe pri prihajanju do prostih metov. Bolj uspešne ekipe bodo uspele izsiliti več prekrškov in imele posledično več možnosti za doseganje "lahkih" košev iz prostih metov, ki pa jih je treba tudi učinkovito zadeti, zato lahko rečemo, da se v četrtem faktorju skrivata pravzaprav dve značilki. Druga meri uspešnost meta iz črte prostih metov.

Učinkoviti odstotek meta (ang. effective field goal percentage) - eFG%

$$\text{eFG\%} = \frac{FGM + 0.5 \times 3PM}{FGA}$$

FGM predstavlja število zadetih metov za dve točki, 3PM pa število zadetih metov za tri točke. V imenovalcu so sešteti vsi poskušani meti za dve in tri točke. Učinkoviti odstotek meta upošteva to, da je met za tri težje zadeti, zato zgrešen met za tri celotnega odstotka ne zniža veliko. eFG% se meri tako v napadu kot v obrambi. S to razliko, da si ekipe v napadu želijo imeti čim višji eFG%, v obrambi pa poskušajo nasprotno ekipo prisiliti v težke mete, zato je za obrambni eFG% bolje, da je manjši.

Izgubljene žoge na 100 posesti - TOpPOS

$$\text{TOpPOS} = \frac{TO}{POS}$$

Značilka meri izgubljene žoge na eno posest. Manj žog kot ekipa zapravi bolje izkoristi posest, v obrambi pa v več napak kot je uspelo prisiliti nasprotnika, težje bodo ti prišli do koša oziroma več priložnosti dobi druga ekipa.

Odstotek pridobljenih skokov - DREB% in OREB%

$$\text{DREB\%} = \frac{DREB_t}{DREB_t + DREB_o}$$

Oznaka "t" pomeni ekipo, ki ji značilko računamo, oznaka "o" pa pomeni nasprotno ekipo. Prikazana je formula za izračun obrambnega odstotka pridobljenih skokov, formula za napadalni odstotek je ista le številke za napadalni skok je potrebno vstaviti. Tak način izračuna skokov nam pove več kot, da bi napadalne ali obrambne skoke povprečili. Preko te značilke tudi vemo, ali je ekipa v povprečju v skoku boljša od druge po koncu tekme. Če je značilka nad 0.5, potem vemo, da skok v povprečju dobi.

Prosti meti - FTR in FT%

$$\text{FTR} = \frac{FTM}{FGA}$$

FTR (ang. free throw rate) beleži odstotek števila zadetih prostih metov (FTM) in števila poskušanih vseh metov za dve in tri točke. Kot že prej omenjeno, ekipa s prostimi meti dobi priložnost doseči lahke koše, zato je pomembno, da ji uspe do prostih metov tudi priti. Spet seveda za obrambo velja ravno obratna logika.

$$\text{FT\%} = \frac{FTM}{FTA}$$

Ker FTR ne meri uspešnosti izvajanja prostih metov, je pametno dodati tudi to značilko, s katero dobimo vpogled, kako dobra je posamezna ekipa pri izvajanju prostih metov. V model smo dodali še eno značilko, povezano s prostimi meti, to je FTA, torej le povprečno število prostih metov, ki jih ekipa meče na tekmi. Značilka je sicer zelo preprosta in se ni izkazala za najbolj informativno, vendar je vseeno izboljšala napoved.

Ostale

Odstotek asistenc - AST%

$$\text{AST\%} = \frac{AST_t}{AST_t + AST_o}$$

Oznaka "t" pomeni ekipo, ki ji značilko računamo, oznaka "o" pa pomeni nasprotno ekipo. Tako kot pri skokih tudi tu ne vzamemo le števila asistenc, ampak odstotek asistenc ekipe na tekmi.

Odstotek blokad - BLK%

$$\text{BLK\%} = \frac{BLK_t}{BLK_t + BLK_o}$$

Oznaka "t" pomeni ekipo, ki ji značilko računamo, oznaka "o" pa pomeni nasprotno ekipo. Meri odstotek blokad ekipe na tekmi.

Odstotek ukradenih žog - STL%

$$\text{STL\%} = \frac{STL_t}{STL_t + STL_o}$$

Oznaka "t" pomeni ekipo, ki ji značilko računamo, oznaka "o" pa pomeni nasprotno ekipo. Meri odstotek ukradenih žog ekipe na tekmi.

Prekrški - PF_t**,** PF_o Število prekrškov opazovane ekipe in število prekrškov na tekmi nasprotne ekipe.

Odstotek meta v raketi - PIP%

$$\text{PIP\%} = \frac{\text{Zadeti meti v raketi}}{\text{Vsi poskušani meti v raketi}}$$

Meti v raketi so pomembni, ker so najbližje košu. Bližje kot si, lažje je koš dati. Če ekipi uspe priti do takih metov, bo imela večjo možnost zmage oziroma dosege koša. Spet obratna logika velja za obrambo. Tej značilki smo dodali še eno, ki pa pove le število točk, doseženih v raketi.

Odstotek meta za tri točke - 3PS%

$$\text{PS\%} = \frac{\text{Zadeti meti za tri}}{\text{Vsi poskušani meti za tri}}$$

Met za tri točke, kot že samo ime pove, prinese tri točke. Težje ga je zadeti, vendar taki meti prinesejo več točk. Visok odsotek meta za tri točke je zato vsekakor pomemben. Vsaka ekipa poskuša uspešnost tega meta v obrambi seveda omejiti.

Hitrost izpolnitve bonusa Pod tem naslovom se skrivajo štiri značilke. Kot vemo, je košarkarska tekma razdeljena na štiri dele. Te značilke pa zabeležijo, koliko sekund pred koncem četrtine je ekipa prišla v bonus. Če ekipa pride v bonus, to pomeni, da vsak naslednji prekršek pomeni izvajanje prostih metov, kar pomeni dosego lahkih košev.

Gibanje rezultata Tudi pod tem naslovom se skriva več značilk, kar 8. Tekmo smo razdelili na osem delov, to pomeni intervale po 6 minut. V vsakem takem intervalu zabeležimo razliko med doseženimi in prejetimi točkami v času pretečenih šestih minut.

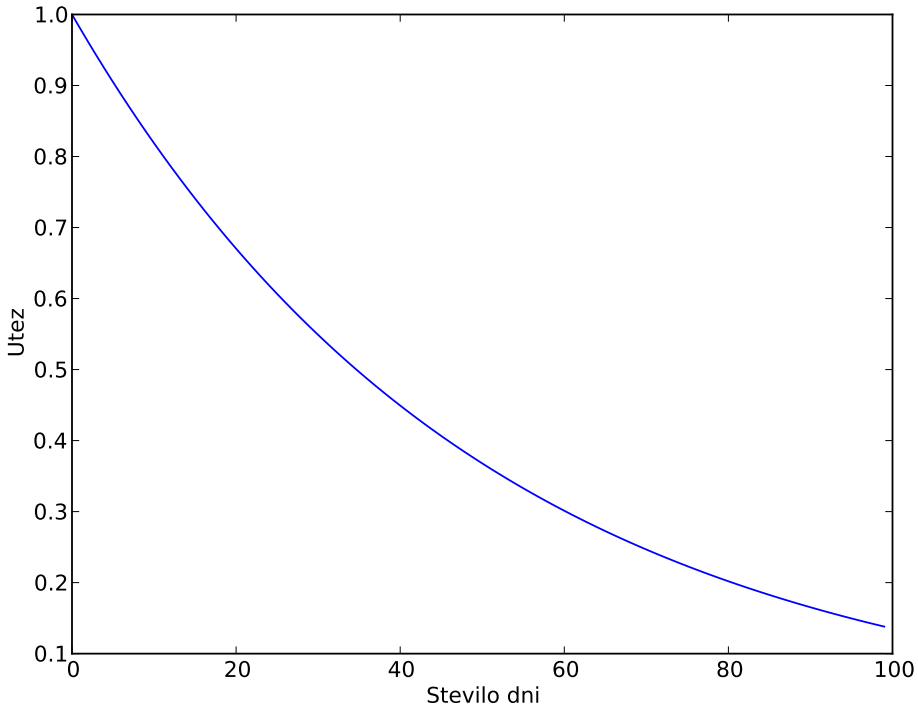
4.2 Način gradnje značilk za posamezen učni primer

Učni primer je vrstica z značilkami, ki predstavljajo stanje obeh ekip pred tekmo. Na konec te vrstice dodamo še razred. V našem primeru smo razred izbrali tako, da je razred 1 pomenil zmago domače ekipe, razred 0 pa zmago gostujoče ekipe. Problem gradnje učnih primerov pri napovedovanju košarkarskih tekem je, da se moramo odločiti, katere pretekle tekme bomo vzeli in iz njih izračunavali značilke.

Ob tem takoj pomislimo na dve teoriji. Prva, kjer v izračun vzamemo vse odigrane tekme v sezoni ter druga, pri kateri vzamemo v izračun značilk samo zadnjih nekaj tekem s predvidevanjem, da nam le-te pokažejo realno stanje ekipe. Težava pri obeh teorijah pa se pojavi, kadar napovedujemo prvo tekmo sezone ene ali druge ekipe. Nesmiselno bi bilo izračunavati značilke na podlagi tekem iz prejšnjih sezon, saj sta bili ekipi, kateri ocenjujemo, lahko sestavljeni popolnoma drugače. V tem primeru je potrebno določiti število tekem, ki jih ne bomo napovedovali, saj bi bile napovedi slabe, ker nimamo zadostnega vzorca tekem, s katerim bi lahko dobili kakovosten opis ekipe.

Odločili smo se, da za zamik vzamemo prvih 12 tekem. Torej na začetku sezone se vedno počaka 12 tekem, da se nabere začetno znanje o ekipah, nato pa začnemo graditi značilke. 12 tekem v ligi NBA predstavlja okoli 3 tedne, v celotni sezoni pa je, kot že omenjeno 82 tekem.

Ostane torej le še prvi opisani problem; ali upoštevati vse ali le določeno število odigranih tekem. Vsaka teorija ima svoje prednosti. Pri prvi dobimo zelo nespreminjajoče značilke, več dobrih ali slabih tekem le malo zboljša ali poslabša značilke ekipe, po drugi strani pa je sezona v ligi tako dolga, da je nemogoče držati visoko raven forme celo leto. Pod predpostavko, da se forma stalno spreminja, bi bilo bolje vzeti le nekaj zadnjih tekem, s čimer bi zajeli trenutno formo ekipe. Srednja pot se večkrat izkaže za najboljšo, zato smo se odločili, da problem rešimo na tak način, da vedno vzamemo vse do sedaj odigrane tekme sezone, nato pa jih utežimo glede na število dni od tekme za



Slika 4.1: Prikaz funkcije uteži. Pomembnosti tekem padajo z večanjem števila dni.

katero računamo značilke.

$$w = \exp\left(-\frac{\text{Število dni}}{2 \times \theta^2}\right) \tag{4.1}$$

Uteži tekem smo računali na podlagi formule 4.1. θ predstavlja razpon oziroma naklon funkcije. Večja kot je θ, bolj linearno izgleda funkcija.

Pojavila se je še ena predpostavka. Sklepali smo, da so morebitne že odigrane tekme med dvema ekipama, za katere računamo značilke bolj informativne kot tiste z drugimi. Zato teh tekem nismo utežili glede na število dni, ampak je bila njihova utež vedno 1.

Na tak način smo za vsako od treh sezon najprej počakali 12 začetnih tekem, nato pa začeli z gradnjo značilk na zgoraj opisani način. Vsakemu

Značilke	Pearsonova korelacija z razredom
Odstotek zmag domače ekipe	0.309
Odstotek zmag domače ekipe na domačem parketu	0.295
Odstotek zmag gostujoče ekipe	0.266
RPI_h	0.254
$OffRTG_h$	0.254
Odstotek zmag gostujoče ekipe na gostovanjih	0.241
RPI_o	0.237

Tabela 4.1: Povezanost najboljših osmih zunanjih značilk z razredom.

učnemu primeru smo kot razred na koncu dodali 1 ali 0, glede na to, kako se je tekma, za katero smo gradili učni primer, končala. Dobili smo 3003 učne primere, na katerih smo opravili testiranja.

4.3 Analiza značilk

Ustaljena praksa v podatkovnem rudarjenju je, da se, ko imaš na razpolago učno množico, lotiš njenega raziskovanja. Ta del je pomemben, saj s tem dobimo občutek, s kakšnimi podatki imamo opravka. Prvi logičen korak pri tem je, da preverimo, kako so značilke povezane z razredom. To nam da vpogled v značilke in njihovo kakovost.

Zaradi preglednosti smo zopet ločili značilke na notranje in zunanje. Tabela 4.1 prikazuje povezanost najboljših zunanjih značilk. Pričakovano so na vrhu odstotki zmag. Nekako intuitivno je razmišljanje, da ima ekipa, ki je v primerjavi z nasprotno ekipo do sedaj nanizala več zmag, tudi na naslednji tekmi več možnosti za zmago.

Tudi iz tabele 4.2 vidimo, da so najboljše značilke tiste, povezane z metom in skokom. Preseneča pa dejstvo, da se je odstotek asistenc domače ekipe pojavil na prvem mestu. Vidimo lahko tudi to, da so atributi, kot sta domači eFG_o in domači PIP%$_o$ bolj povezana z razredom, kot njuni nasprotni

Značilke	Pearsonova korelacija z razredom
AST%$_h$	0.241
Domači eFG_o	0.226
DREB%	0.222
Domači eFG	0.216
Gostujoči eFG_o	0.204
Gostujoči DREB%	0.195
Gostujoči eFG	0.177
Domači PIP%$_o$	0.173
Gostujoči PIP%$_o$	0.161

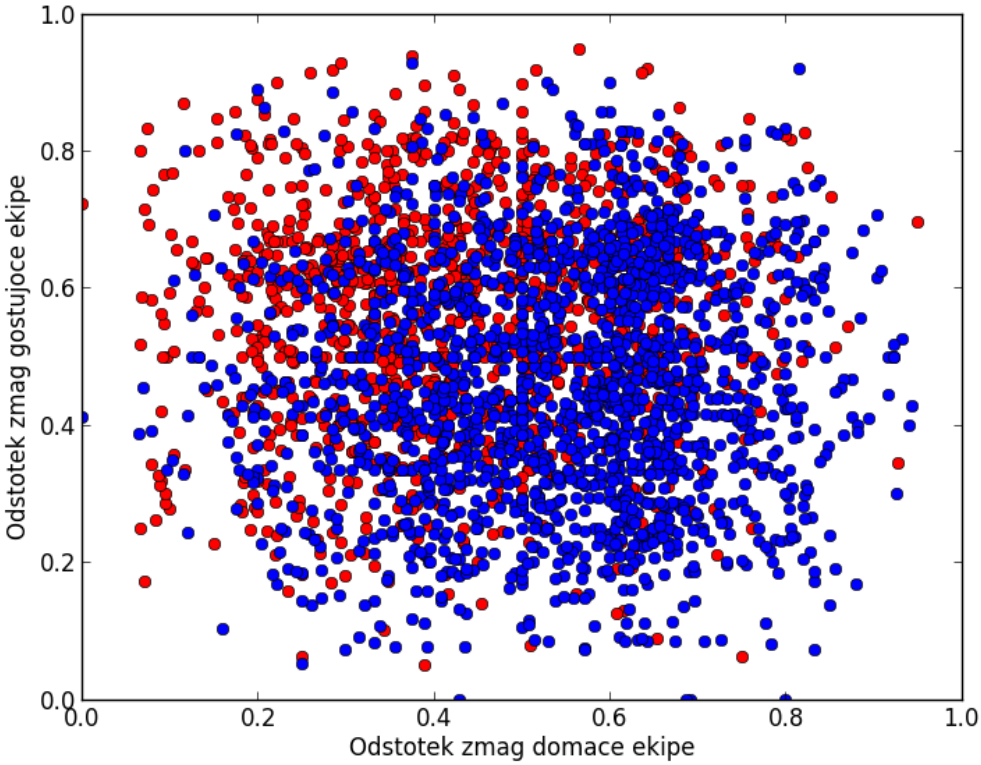
Tabela 4.2: Povezanost najboljših devetih notranjih značilk z razredom.

značilki. To bi si lahko razlagali s tem, da je bolj pomembno omejiti nasprotnikov met, kot pa izboljšati svojega. Z drugimi besedami; boljša obramba je za zmago ekipe pomembnejša kot boljši napad.

Na splošno so zunanje značilke bolj povezane z razredom, kar je tudi pričakovano, saj je pri notranjih značilkah večkrat mnogo odvisno od dnevne forme, kar lahko nekoliko pokvari dejansko stanje.

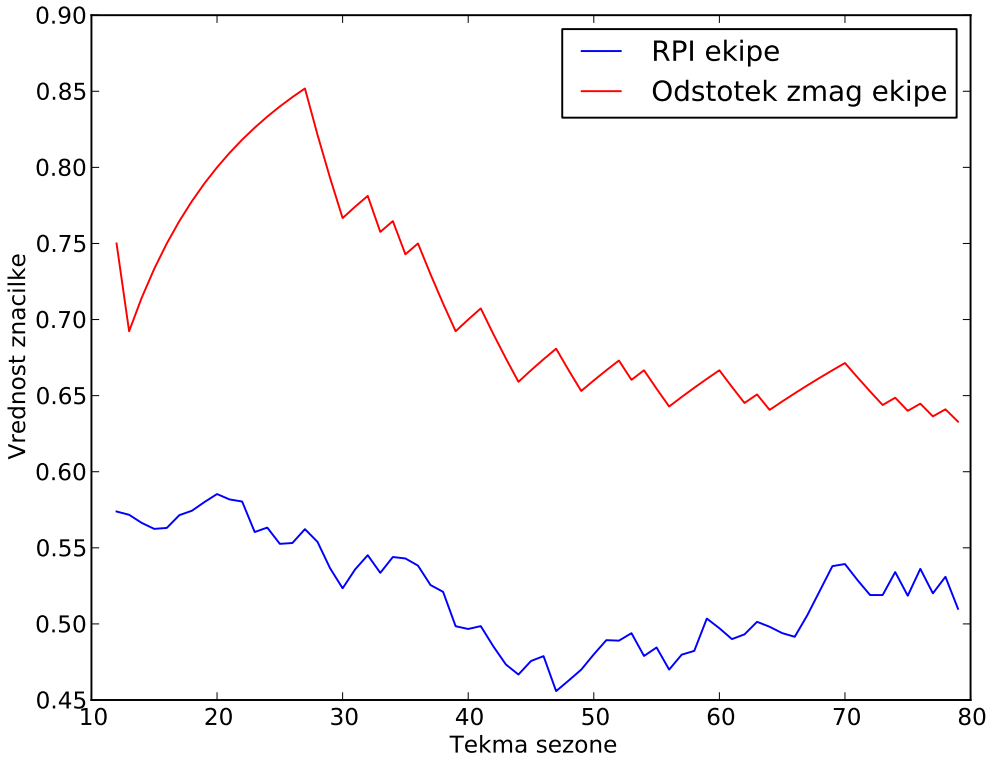
Slika 4.2 prikazuje zmage domače in gostujoče ekipe glede na odstotek zmag pred tekmo. Razvidno je, da sta zgornji levi kot in spodnji desni kot bolj "čista", saj je razlika v odstotku zmag med ekipama velika. Včasih pa se tudi v ta kota prikrade kakšna pika drugačne barve, to so tekme, kjer je favorit izgubil. Bolj ko pogledamo proti sredini slike, več je mešanja barv. Teh tekmem se samo s tema dvema značilkama ne bo dalo dobro napovedati. Tu bodo razliko delale druge značilke.

Kot je bilo že prej omenjeno, je sezona v ligi NBA zelo dolga, zato je visoko formo težko zadržati celotno sezono, seveda to velja tudi obratno. Ekipam se zgodi, da padejo v krizo, ki pa seveda ne traja celo sezono, ampak le določen del le-te. RPI je značilka, ki poskuša oceniti trenutno formo ekipe, saj gleda zmage in moč nasprotnikov le nekaj tekem nazaj. Iz grafa 4.3 je razvidno,



Slika 4.2: Modre pike označujejo domačo zmago, rdeče pa gostujočo zmago.

da je ekipa Bostona sezono začela izvrstno, vsaj sodeč po odstotku zmag. RPI pa kaže drugačno zgodbo; te zmage so bile najverjetneje dosežene proti lažjim nasprotnikom, zato iz krivulje RPI-ja ni razvidno, da bi se pretirano dvignila. Nekje na sredi sezone RPI doseže najnižjo točko, potem pa se začne dvigati vse do konca sezone, kljub temu da krivulja odstotka zmag ne narašča, temveč celo rahlo pada. To bi si lahko razlagali tako, da je imel Boston drugi del sezone težji razpored tekem; to pomeni, da je igral proti močnejšim nasprotnikom. Njegov odstotek zmag je vseskozi rahlo padal. Zaradi težjega razporeda so igrali proti več kakovostnejšim ekipam in jih nekaj uspeli tudi premagati. RPI je upošteval težji razpred oziroma boljše ekipe, zato krivulja ne pada, ampak se rahlo dviga.



Slika 4.3: Prikaz gibanja značilk za ekipo Boston Celtics v sezoni 2009/10.

Relativne značilke

Značilke so definirane tako, da za eno značilko domače ekipe, vedno obstaja tudi enaka značilka nasprotne ekipe. To lahko imenujemo par značilk. Par značilk bi torej bili značilki odstotek števila zmag domače ekipe in odstotek zmag nasprotne ekipe. Tak par lahko združimo v eno značilko in tako dobimo eno relativno značilko. Združimo jih lahko s preprostim deljenjem med seboj, ali s kako drugo metodo. Relativne značilke imajo potencial, da so bolj informativne kot par značilk, hkrati pa s tem število značilk razpolovimo. Manjša dimenzija učne množice lahko olajša učenje učnim algoritmom. V tej diplomski nalogi smo analizirali povezanost relativnih značilk pridobljenih

Značilke	Pearsonova korelacija z razredom
RPI	0.345
$OffRTG_h$	0.320
$DefRTG$	0.307
$\frac{\text{WIN\%}_\text{h}}{\text{WIN\%}_\text{o}}$	0.028
Razmerje števila zaporednih zmag ali porazov	0.019
$\frac{\text{WIN\%}_{\text{Domače ekipe doma}}}{\text{WIN\%}_{\text{Gostujoče ekipe v gosteh}}}$	0.007

Tabela 4.3: Povezanost zunanjih relativnih značilk z razredom. Prikazane so relativne značilke pridobljene z deljenjem parov značilk.

Značilke	Pearsonova korelacija z razredom
$\frac{\text{WIN\%}_\text{h}}{\text{WIN\%}_\text{o}}$	0.399
$\frac{\text{WIN\%}_{\text{Domače ekipe doma}}}{\text{WIN\%}_{\text{Gostujoče ekipe v gosteh}}}$	0.379
RPI	0.349
$OffRTG_h$	0.322
$DefRTG$	0.306
Razmerje števila zaporednih zmag ali porazov	0.195

Tabela 4.4: Povezanost zunanjih relativnih značilk z razredom. Prikazane so relativne značilke pridobljene z odštevanjem parov značilk.

z deljenjem in z odštevanjem para značilk.

V tabelah 4.3 in 4.5 so prikazani rezultati relativnih značilk pridobljenih z deljenjem. Primerjava s tabelama 4.1 in 4.2 pokaže, da so relativne značilke bolj povezane z razredom, kot najboljša značilka iz para značilk. Izjema so relativne značilke povezane z zmagami, katerih povezanost močno pade.

Bolje od relativnih značilk pridobljenih z deljenjem, so se izkazale relativne značilke pridobljene z odštevanjem. Rezultati so prikazani v tabelah 4.4 in 4.6. Značilke povezane z zmagami z odštevanjem niso izgubile na povezanosti, ampak so na njej pridobile.

Značilke	Pearsonova korelacija z razredom
AST%	0.297
eFG_o	0.304
DREB%	0.289
eFG	0.278
$\text{PIP\%}_o$	0.255
PIP%	0.237
3PS%	0.197
$\text{3PS\%}_0$	0.168

Tabela 4.5: Povezanost najboljših osmih notranjih relativnih značilk z razredom. Prikazane so relativne značilke pridobljene z deljenjem parov značilk.

Značilke	Pearsonova korelacija z razredom
AST%	0.297
eFG_o	0.304
DREB%	0.291
eFG	0.279
$\text{PIP\%}_o$	0.253
PIP%	0.238
3PS%	0.194
$\text{3PS\%}_0$	0.167

Tabela 4.6: Povezanost najboljših osmih notranjih relativnih značilk z razredom. Prikazane so relativne značilke pridobljene z deljenjem parov značilk.

Poglavje 5

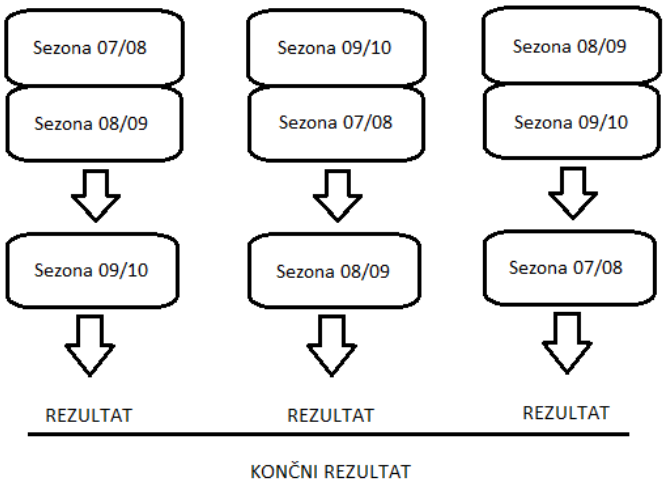
Gradnja in testiranje modela

5.1 Način testiranja

Izbira načina testiranja je pomembna, saj želimo na podlagi testnih rezultatov izbrati najprimernejši učni model, ki bo dobro napovedoval tudi na še ne videnih podatkih. Ker so tekme organizirane v sezone, nam to onemogoča klasično prečno preverjanje, saj bi tako prišlo do mešanja tekem iz različnih sezon. To sicer ni tako zelo narobe, vendar smo sklenili, da bi bilo bolje izbrati kakšen drug način. Obdržali smo idejo prečnega preverjanja s to razliko, da smo podatke razdelili na sezone. Na sliki 5.1 je prikazan diagram preverjanja. Ker so značilke vedno grajene znotraj ene sezone, učni algoritem "ne ve", katero sezono napoveduje in je zanj vseeno iz katerih sezon se uči. Tako se vedno učimo iz dveh sezon na eni pa napovedujemo. Rezultat napovedovanja vseh treh sezon se povpreči in tako dobimo enoten rezultat za celotno preverjanje.

Odločiti smo se morali tudi o tem, kaj sploh bomo merili oziroma katere mere bomo poskušali minimizirati. Na misel hitro pade napovedna točnost. Preprosto preštejemo vse pravilno napovedane tekme in to število delimo z številom vseh tekem, kot je prikazano v 5.1.

$$\text{Napovedna točnost} = \frac{\text{Število pravilnih napovedi}}{\text{Število vseh tekem}} \tag{5.1}$$



Slika 5.1: Diagram izvedbe testiranja. Rezultati treh testov se povprečijo in tako dobimo končni rezultat.

Napovedna točnost je sicer dobra mera, vendar za primer vzemimo naslednji scenarij. Tekma se je končala z zmago domače ekipe, eden od učnih algoritmov je napovedal 60% verjetnost zmage domače ekipe, drugi je prav tako napovedal zmago domače ekipe s 70%. Oba algoritma sta napovedala pravilno in sta iz stališča napovedne točnosti enaka, vendar je jasno, da je algoritem, ki je zmago domače ekipe napovedal s 70%, bolj prepričan v svojo napoved. Zato smo morali poiskati mero, ki bo merila tudi to, da učni algoritem s čim večjo verjetnostjo napove določeno napoved. Hkrati pa tudi kaznuje napovedi, ki so napačne oziroma preveč oddaljene od pravilne napovedi. Taka mera je "Brier score", njena formula pa je prikazana v 5.2.

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}\left(\text{Napoved}_i - \text{Razred}_i\right)^2 \tag{5.2}$$

Zmago domače ekipe označimo z razredom 1, z razredom 0 pa zmago

gostujoče ekipe. Vedno napovedujemo verjetnost zmage domače ekipe, to je lahko številka med 0 in 1.

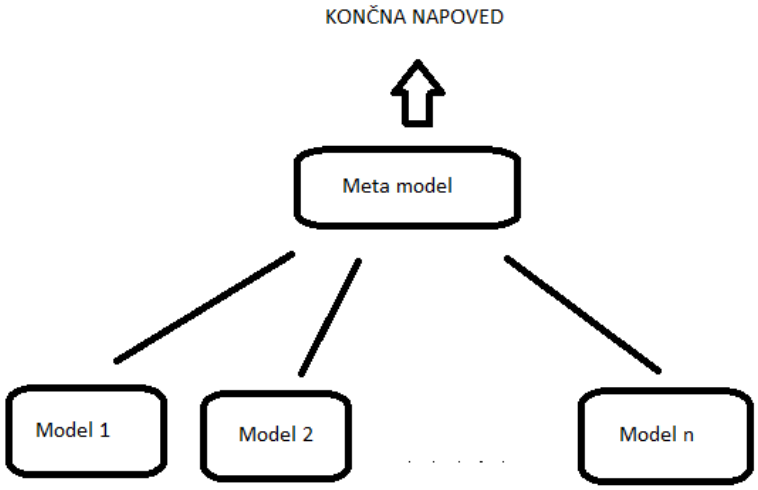
Preverjanje bi lahko naredili še malenkost bolj realno. Še vedno bi imeli podatke razdeljene na tri dele, ki predstavljajo tri sezone, prav tako bi se vedno učili iz dveh sezon in napovedovali na tretji. Razlika bi bila v tem, da bi vsako že napovedano tekmo dali v učno množico. Učni algoritem bi se tako učil iz dveh sezon vključno z dodatno tekmo, ki smo jo na novo prestavili v učno množico. To bi morda prineslo celo nekoliko boljše rezultate, saj bi se učni algoritem učil iz vedno več podatkov, vendar bi bil ta način računsko zelo zahteven in posledično počasen, zato smo se raje odločili, da bomo ostali na zgoraj opisanem postopku.

5.2 Opis modela in rezultati testiranj

Preden smo šli v podrobno testiranje raznih učnih algoritmov, smo morali določiti meje (ang. benchmark), ki jih želimo premagati. Te meje morajo temeljiti na preprostih predpostavkah, če bo učni algoritem slabši od mejnega, potem je jasno, da ne dela dobro oziroma je slab. Najbolj enostaven možni algoritem je, da za zmagovalca vedno napovemo domačo ekipo s tako verjetnostjo, kot so domače ekipe zmagovale v učni množici. Temu se reče večinski klasifikator (ang. majority). Napove razred, ki je najbolj pogost v učni množici. Z večinskim klasifikatorjem je meja postavljena izredno nizko, zato smo se odločili še za eno mejo. Tej bomo rekli "Višji odstotek zmag". Ekipa, ki ima pred tekmo višji odstotek zmag, je napovedana kot zmagovalna. Ta meja bi vsaj po logiki morala biti precej višja od večinskega klasifikatorja.

Testirali smo veliko učnih algoritmov, za katere smo menili, da se bodo na tej učni množici odrezali dobro. Izoblikovala se je množica treh algoritmov, ki so bili po rezultatih testiranja zelo dobri. To so bili Support Vector Machines (v nadaljevanju SVM), nevronska mreža in logistična regresija. Ob treh tako različnih algoritmih, ki pa napovedujejo približno enako dobro, na pamet hitro pade to, da bi jih nekako združili in tako dobili še boljše napovedi.

Odločili smo se za združevanje z meta - učenjem (ang. stacking), vendar se je izkazalo za neuspešno. Napovedi algoritmov so bile med seboj preveč podobne, zato ni bilo mogoče ničesar pridobiti.



Slika 5.2: Model združevanja z meta - učenjem (ang. stacking). Za meta algoritem se običajno vzame logistično regresijo. V našem primeru smo vzeli nevronske mreže.

Ideja je bila, da učno množico nekoliko spremenimo, ji odvzamemo nekaj značilk, v upanju, da dobimo malce drugačne napovedi, a vseeno take z dobrimi rezultati. Sledili smo načelu preprostosti in iz učne množice odstranili vse notranje značilke, tako da smo ostali le z dvanajstimi zunanjimi značilkami. Na tej množici smo pognali logistično regresijo in bili smo presenečeni nad rezultatom. Ta model je bil celo najboljši na testiranjih, če gledamo le Brier score. Tudi v napovedni točnosti pa ni zaostajal veliko za SVM-jem, ki je bil od prej omenjenih treh najboljši.

V duhu tega odkritja smo se spomnili tudi na že prej analizirane relativne značilke, ki so imele dobro povezanost z razredom in jih je bilo zato

Model	Brier score	Točnost
SVM	0.1978	0.7001
SVM (+ najboljše rel. značilke)	0.1982	0.6991
SVM (-)	0.2004	0.6940
SVM (/)	0.2392	0.6045
Logistična regresija (+ najboljše rel. značilke)	0.1976	0.6973
Logistična regresija	0.1977	0.6950
Logistična regresija (-)	0.1979	0.6941
Logistična regresija(/)	0.1982	0.6930

Tabela 5.1: Primerjava rezultatov algoritmov z uporabo relativnih značilk in brez.

na nek način smiselno vključiti v učno množico. Nove teste smo opravili na tri načine. Prvi je, da smo obstoječi učni množici dodali najboljše relativne značilke. Pri drugem smo učno množico skrčili na le relativne značilke pridobljene z deljenjem, pri tretjem pa smo storili enako kot pri drugem, le da smo vzeli relativne značilke pridobljene z odštevanjem. Pri skrčeni množici le zunanjih značilk, ki se je dobro izkazala za logistično regresijo, smo naredili podobno, le da smo pri delu z relativnimi značilkami upoštevali zgolj zunanje relativne značilke. Rezultati so prikazani v tabeli 5.1. Vidimo, da relativne značilke SVM-ju niso pomagale. Logistični regresiji pa je dodaja najboljših zunanjih relativnih značilk nekoliko pomagala.

To je znova dalo upanje, da bo združevanje teh dveh modelov prineslo nekaj boljšega. Pri združevanju z meta-učenjem smo preizkusili dva različna načina. Pri prvem smo vzeli SVM na običajni učni množici in logistično regresijo z le zunanjimi atributi. Pri drugem pa SVM na enak način kot prej, logistični regresiji pa smo v učno množico dodali še najboljše zunanje relativne značilke. Prvi način se je za odtenek bolje obnesel. Drugi način je imel sicer rahlo boljši Brier score od prvega, vendar je "preveč" zaostal v

napovedni točnosti. Zato smo za končni model vzeli prvi način.

Končni model je torej sestavljen iz treh delov. Prvi del so napovedi učnega algoritma SVM na celotni učni množici, to pomeni, da so uporabljene vse značilke, ki so opisane v poglavju 4. Drugi del so napovedi logistične regresije na skrčeni učni množici, ki uporablja le zunanje značilke. Tretji del pa je združitev prejšnjih dveh delov v novo boljšo napoved. To smo storili z meta učenjem. Za meta algoritem pa smo uporabili nevronske mreže. Rezultati so predstavljeni v tabeli 5.2, kjer vidimo, da smo z združevanjem uspeli izboljšati rezultata posameznih algoritmov.

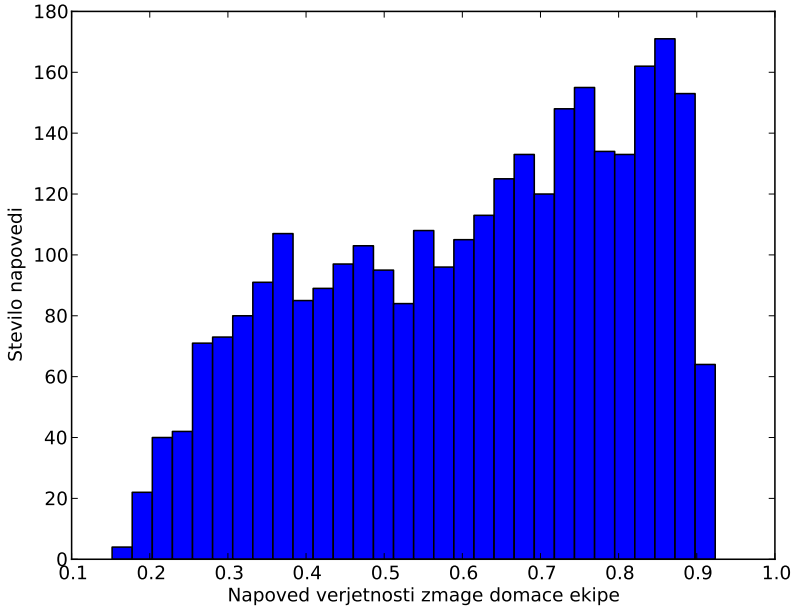
Model	Brier score	Točnost
Stack (brez relativnih značilk)	0.1976	0.7024
Stack (z relativnimi značilkami)	0.1974	0.6975
Logistična regresija (+ najboljše rel. značilke)	0.1976	0.6973
SVM	0.1978	0.7001
Logistična regresija	0.1977	0.6950
Višji odstotek zmag	0.2164	0.6597
Domača ekipa	0.2391	0.6045

Tabela 5.2: Rezultati mej in najboljših modelov.

Analiza napovedi

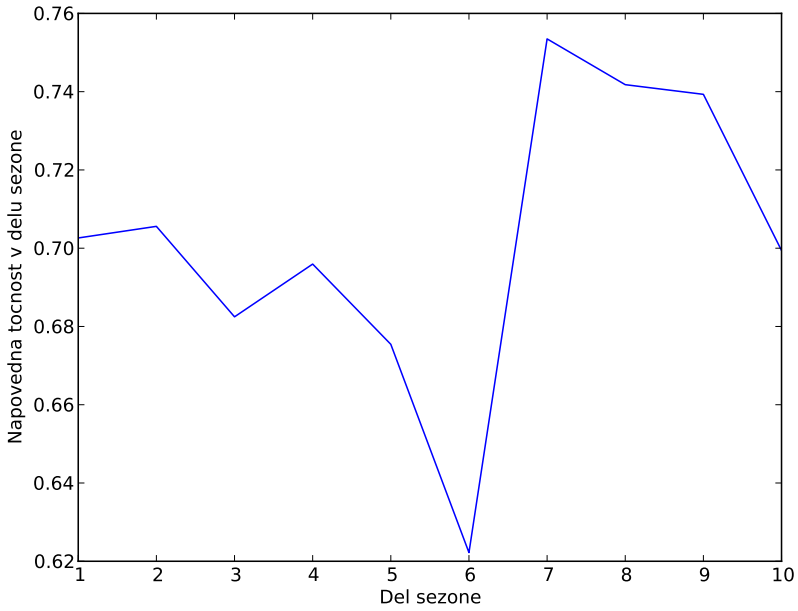
Iz slike 5.3 je razvidno, da so napovedi precej pristranske, saj zgrajeni model dosti raje napoveduje domačo zmago oziroma daje prednost domači ekipi. Kar je do neke mere sicer smiselno, če vemo, da domače ekipe zmagajo na približno 60% tekem. Podrobnejši pregled napovedi nam kaže, da je povprečna napoved zmag domače ekipe (kjer je domača ekipa dejansko zmagala) 67.7%. Povprečna napoved zmag gostujoče ekipe (kjer je gostujoča ekipa dejansko zmagala) pa 50.8%. Zato lahko sklepamo, da na tekmah, kjer igrata dve podobno dobri ekipi, model da prednost domači ekipi.

Zelo zanimiva je naslednja ugotovitev. Nekako naravno se nam zdi, da bi



Slika 5.3: Histogram napovedi.

morale biti napovedi boljše, bolj ko se bliža sezona koncu. Ekipe so odigrale že ogromno tekem, zato dobimo boljše opise ekip z značilkami. Pričakujemo, da se bo napovedna točnost linearno povečevala proti koncu sezone. Na sliki 5.4 pa je razvidno, da se zgodi nekaj zelo zanimivega. Sezono smo razdelili na 10 enakih delov in v vsakem delu izračunali napovedno točnost. Kar dobimo, vsekakor ni linearna krivulja oziroma niti blizu tega. Nekje malo čez polovico sezone napovedna točnost močno pade. Kar pa je še bolj čudno, v naslednjem delu se dvigne na najvišjo raven. Kaj točno se zgodi v teh dveh delih sezone, je dokaj težko razložiti. Morda se prava razmerja med ekipami vzpostavijo šele po drugem delu sezone in je v prvem delu več negotovosti; več zmag ekip, ki niso favoriti. Morda na to vpliva tudi vikend vseh zvezd (ang. All-Stars weekend), kjer se za nekaj dni liga prekine in se odigra nekaj ekshibicijskih tekem, na katerih so zbrani vsi najboljši igralci lige.



Slika 5.4: Sezona razdeljena na deset delov. V vsakem delu je izračunana napovedna točnost.

Omenjeno je bilo, da smo pri združevanju preizkusili dva različna načina. Pri enem smo pri logistični regresiji uporabili relativne značilke, pri drugem pa ne. Kot končni model smo nato izbrali prvi način, ki ni uporabljal relativnih značilk, to pa zato, ker je bil v izbranih merah skupno rahlo boljši. Z besedama boljši in slabši je včasih dobro biti bolj zadržan. Na voljo so nam le tri sezone in lahko se zgodi, da je rahlo boljši model boljši le zaradi naključja.

Odločili smo se preveriti, ali sta si združena modela v napovedih statistično res različna. Izbrani končni model je res boljši od drugega, če je razlika med njunima napovedima statistično značilna. Vzeli smo napovedi za vse tri sezone skupaj, ki smo jih dobili iz testiranj. Rezultata obeh modelov na testiranjih sta predstavljena v prvi in drugi vrstici tabele 5.2. Izvedli smo McNemaryev statistični test, ki med seboj primerja le klasifikacije tekem, to-

rej vektorje ničel in enic. Hipoteze, da se napovedi ne razlikujeta značilno, ni bilo mogoče zavreči. Za stopnjo zaupanja smo uporabili α=0.05, ugotovljena p-vrednost pa je bila bila 0.24. To pomeni, da je razlika med rezultati na testiranjih najverjetneje rezultat naključja.

Zanimalo nas je tudi, ali se združena napoved značilno razlikuje od posameznih napovedi SVM-ja in logistične regresije. Pri tem testu smo uporabili rezultate iz prve, četrte in pete vrstice v tabeli 5.2. Najprej smo uporabili Friedmanov test, ki vzame vektorje napovedanih verjetnosti, le da pri tem testu lahko primerjamo več vektorjev hkrati. Test je pokazal, da se trije primerjani vektorji (združena napoved, SVM napoved in napoved logistične regresije) med seboj statistično značilno razlikujejo. Za stopnjo zaupanja smo zopet izbrali α=0.05, ugotovljena p-vrednost pa je bila 2.17e-135. Nato je bilo potrebno ugotoviti, katere so tiste napovedi, ki se razlikujejo. Tu pride v poštev Nemenyijev test, ki je pokazal, da se vse tri napovedi med seboj statistično značilno razlikujejo, kar pomeni, da je združevanje dveh modelov obrodilo sadove in je združena napoved tudi dejansko boljša od dveh posameznih. P-vrednosti med modeli so prikazane v tabeli 5.3.

Model	Log. reg.	SVM	Stack
Log. reg.	-	0	3.57e-14
SVM	0	-	0
Stack	3.57e-14	0	-

Tabela 5.3: P-vrednosti med posameznimi napovednimi modeli. Vse napovedi se med seboj statistično značilno razlikujejo.

5.3 Napovedni model in stavnice

Naš napovedni model je brez težav presegel mejne rezultate. To ni bilo ravno težko, saj je bila že naša začetna predpostavka, da tekme lahko napovemo bolje od obeh mej. Pravi pokazatelj kakovosti modela so kvote stavnic. Te

	Sezona 07/08		Sezona 08/09		Sezona 09/10	
	Brier score	Točnost	Brier score	Točnost	Brier score	Točnost
Betfair	0.1900	0.7042	0.1867	0.7106	/	/
Napovedni model	0.1963	0.7102	0.1922	0.7103	0.2042	0.6848
Bovada	0.2277	0.6946	0.2258	0.7016	0.2285	0.6941

Tabela 5.4: Primerjava napovednega modela s stavnicami.

veljajo za najboljše napovedi, ki jih lahko dobimo. Stavna hiša Betfair ima prav posebno vrsto stav. Kvote niso določene vnaprej, kot je to običaj pri normalnih stavnicah, ampak lahko vsak sam podpre ali stavi proti dogodku s svojo lastno kvoto. Temu se reče izmenjevalne kvote (ang. exchange odds). Tako se kvote s časom spreminjajo. V [5] je bilo pokazano, da so kvote, ki so bile vplačane v intervalu 5 minut do začetka tekme, najboljša napoved za tekme lige NBA. Zgodovinske podatke stavnice Betfair so na voljo na `http://data.betfair.com/`.

Poleg Betfair-ovih kvot, smo vzeli tudi kvote navadne stavnice, ki svoje kvote določa vnaprej in s časom ostajajo iste. Te smo pridobili na `http://www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm`. To so kvote manj znane športne stavnice "Bovada". Izkazalo se je, da je z razlogom manj znana, vsaj sodeč po napovedih za ligo NBA.

Tabela 5.4 prikazuje primerjavo stavnic z našim modelom. Vidimo, da ima Betfair precej boljši Brier score od ostalih dveh. Toda ni vse tako slabo, naš model je v napovedni točnosti v sezoni 2007/08 uspel celo preseči Betfair, v sezoni 2008/09 pa za njim zaostal le za malenkost. Stavne kvote stavnice Bovada so se izkazale za zelo slabe, vendar so izpostavile že znano slabost našega modela. Sezona 2009/10 je bila zelo nepredvidljiva, domače ekipe so v povprečju zmagovale manj, med sezono pa se je zgodilo tudi nekaj menjav na trenerskih mestih, bilo je tudi nekaj poškodb ključnih igralcev in tudi menjave igralcev med ekipami. Vse to je botrovalo temu, da je bila ta sezona nepredvidljiva s stališča napovedovanja. Naš model je v tej sezoni

napovedoval občutno slabše, kot v dveh prejšnjih sezonah.

Poglavje 6

Sklepne ugotovitve

6.1 Ugotovitve

Glavni cilj je bil pokazati, da se z načinom opisanim v tej diplomski nalogi da učinkovito napovedovati košarkarske tekme. Uspelo se nam je zelo približati napovedim stavnice Betfair, ki veljajo za najboljše, oziroma jih celo preseči v eni od sezon. Pokazali smo tudi, da obstajajo stavnice z zelo slabo postavljenimi kvotami, ki jih naš model zlahka preseže, tako v Brier score-u kot v napovedni točnosti. Izkazalo se je, da je Brier score stavnice Betfair, kljub temu da smo se ji z napovedno točnostjo zelo približali, še vendno precej boljši. Tu bi glavne zasluge lahko pripisali človeškemu faktorju in poznavanju stanja v ekipi, kar je težko zapisati s številkami ali najti prave značilke.

V okviru te diplomske naloge sicer nismo preizkušali ali bi se dalo dejansko zaslužiti na stavnicah z uporabo tega modela. Mislimo pa, da bi s pravo stavno strategijo nedvomno lahko zaslužili, saj je napovedna točnost zelo solidna.

6.2 Izboljšave in nadaljnje delo

Napovedi se vsekakor da še izboljšati. Bodisi da nadgradimo obstoječ model ali pa naredimo novega in ga združimo s tem opisanim v diplomski nalogi.

Značilke izbrane za ta model gledajo na ekipo kot celoto. Izluščijo skoke, odstotek metov, asistence, ..., ki jih doseže celotna ekipa. Nimamo vpogleda v posamezna igralna mesta, ne vemo, kdo prispeva skoke, asistence, kdo največ meče. Možna izboljšava bi bil v tem primeru nov model z drugimi značilkami, ki bi ekipo ocenjevale kot skupek petih igralnih mest.Vsako igralno mesto bi imelo svoje značilke. Tak model bi dajal neke svoje verjetnosti, ki bi jih lahko združili z verjetnostmi že obstoječega modela in morda tako dobili boljše napovedi. Ena od možnosti je tudi, da bi nove značilke le dodali že obstoječim, tudi to bi morda prineslo boljše rezultate. Verjetno pa bi se dalo ekipo ravno tako pogledati še iz kakšnega drugega zornega kota.

Po zgledu aplikacij za finančno napovedovanje, bi lahko tudi tu uporabili tekstovno rudarjenje (ang. text mining). Z izluščenjem ključnih besed iz povzetkov tekem ali analiz tekem, bi lahko dobili boljši vpogled v stanje ekipe, ki ga nobena statistika ne more dobiti. Iz besedila bi ustvarili nove značilke, ki bi jih dodali obstoječim. Ta način bi bilo zelo zanimivo preizkusiti, saj ga, kot že omenjeno, finančne aplikacije zelo uspešno izkoriščajo. Vključili bi lahko tudi omembe ekip ali igralcev na socilanih omrežjih oziroma tekstovno rudarili tudi socialna omrežja, ki so danes zelo pogosto uporabljena za izražanje pripadnosti ekipam.

Kot slabost se je izkazalo tudi to, da je model preveč pristranski in daje preveliko prednost domačim ekipam. To bi bilo treba omejiti. Lahko bi podrobneje analizirali tekme, kjer so gostujoče ekipe zmagale in iz njih poskušali izluščiti razloge, zakaj se je to zgodilo. Potem bi lahko na podlagi zaključkov izoblikovali značilke, ki bi uravnotežile model. S tem bi morda lahko omejili tudi slabšanje napovedi v primeru nepredvidljive sezone. Taka sezona zmanjša napovedno točnost tudi stavnicam, a vendar ne v taki meri, kot smo to zaznali pri našem modelu.

Znano je, da verjetnosti, ki jih daje SVM niso uravnotežene oziroma ka-

librirane. Splačalo bi se uporabiti sigmoidno kalibracijo, ki verjetnosti uravnoteži po sigmoidni krivulji. Take verjetnosti bi lahko pomagale pri vsaj delni odpravi pristranskosti in boljši napovedi.

V poglavju 6, kjer je opisan način testiranja, smo omenili tudi nekoliko izboljšano in realnejšo različico testiranja. Vedno napovedujemo eno sezono iz dveh pa se sistem uči. Bolje bi bilo, da bi se vsaka napovedana tekma, potem tudi vključila v učno množico. Učna množica bi tako postopoma rasla in sistem bi imel na voljo vedno več učnih primerov za učenje. Zaradi večjega števila učnih primerov, zlasti v drugem delu napovedovanja sezone, bi bile napovedi po vsej verjetnosti nekoliko boljše. Bilo pa je že omenjeno, da bi bil tak način testiranja, precej bolj računsko zahteven, saj bi se sistem moral ob vsaki napovedi ponovno učiti.

Literatura

[1] M. Robnik-Šikonja, I. Kononenko (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF , *Machine Learning Journal*. 53, 23-69. Dostopno na: `http://lkm.fri.uni-lj.si/rmarko/papers/robnik03-mlj.pdf`

[2] W. Frank (2011). Data Mining, Morgan Kaufman.

[3] D. Oliver (2004). Basketball on paper. Potomac Books.

[4] J. Kubatko, D. Oliver, K. Pelton, D. T. Rosenbaum (2007), A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3), 1-22.

[5] E. Štrumbelj, P. Vračar (2012). Simulating a basketball match with a homogeneous Markov model and forcasting the outcome, *International Journal of Forecasting*, 28, 532-542.

[6] K. Shirley (September 2007). A Markov model for basketbal. Predstavitev na New England Symposium for Statistics in Sports v Bostonu. Dostopno na: `http://www.amstat.org/Chapters/boston/nessis07/presentation_material/Kenny_Shirley.pdf`

[7] H. O. Stekler, D. Sendor, R. Verlander (2009), Issues in sports forecasting , *International Journal of Forecasting*. 26, 606-621.