Football Strategy

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Mirko Brkušanin
1087/2016

jun 2017.

1 Podaci

Podaci su preuzeti sa adrese: https://www.crowdflower.com/wp-content/uploads/2016/03/Football-Scenarios-DFE-832307.csv. Jedan red tabele sadrži informacije o zadataoj situaciji (vreme, rezultat, pozicija na terenu...) kao i odluku koja je proglašena kao najbolja od strane sudija zajedno sa pouzdanošću same odluke. Pored ovih postoje i dodatni podaci koji nama neće biti od značaja i u nastavku će biti ignorisani.

Primer jedne situacije može biti: *"It is first down and 10. The ball is on your opponent's 20 yardline. There is 3 seconds left in the second quarter. You are down by 3 points."*. Iz ovoga izvlačimo naredne podatke (redom kojim su navedeni):

down	yards	yardline	time	quarter	score
1	10	20	3	2	-3

- *down* $\in \{1, 2, 3, 4\}$. Daun može imati samo vrednosti iz ovog skupa. Prvi daun prevodimo u broj 1, drugi u 2, itd.
- *yards* $\in \{1, 3, 8, 10\}$. Označava broj jardi koji moramo da osvojimo. Može biti bilo koji broj između 1 i 100 (veličina terena) ali u dobijenim podacima se javljaju samo navedene 4 vrednosti. Oznaka *inches* u podacima je drugi naziv za jedan jard.
- *yardline* $\in \{1, 5, 20, 40, 55, 80\}$. Označava poziciju na terenu. Kao i *yards* može biti između 1 i 100 ali se u dobijenim podacima javljaju samo navedene vrednosti. Navedeni su u jednom od dva oblika:
  - *The ball is on your opponent's X yardline.* Što znači da imamo X jardi do protivničke *end zone*. Koristićemo broj X.
  - *The ball is on your Y yardline.* Što znači da imamo 100-Y jardi do protivničke *end zone*. Koristićemo broj 100-Y.
- *time* $\in \{3, 120, 420\}$. Preostalo vreme do kraja četvrtine. Javlja se samo u 3 oblika u podacima:
  - *There is 3 seconds left....* Prevodimo u 3.
  - *There is 2 minutes left....* Prevodimo u 120.
  - *There is 7 minutes left....* Prevodimo u 420.
- *quarter* $\in \{1, 2, 3, 4\}$. Trenutna četvrtina.
- *score* $\in \{-20, -10, -7, -3, 3, 7\}$. Razlika trenutnog rezultata. Broj je pozitivan ukoliko vodimo, inače je negativan. U podacima se javljaju samo navedene vrednosti.

Pored ovih koristićemo još i kolonu sa donetom odlukom koja će biti nazvana *decision* i može imati jednu od 5 vrednosti: *punt*, *kick a field goal*, *run*, *pass*, *kneel down*; kao i kolonu koja označava pouzdanost odluke sa nazivom *confidence* $\in [0, 1]$.

Datoteka **Football-Scenarios.csv** sadrži obrađene podatke. Redovi sa nepotpunim vrednostima su izbrisani. Ono što smo na kraju dobili su 6 kolona sa celobrojnim vrednostima koje u potpunosti opisuju zadatu situaciju (označeni prefiksom s i brojem u tabeli), jednu kolonu sa decimalnim brojevima (od 0 do 1) i jednu sa nominalnim podacima (5 različitih vrednosti).

2 Primenjene metode

Nad podacima su primenjene razne tehnike za klasifikaciju. Na osnovu 6 atributa koji opisuju situaciju pokušaćemo da odredimo koju je akciju najbolje preduzeti. Svaku od akcija ćemo posmatrati kao jednu klasu koju želimo da predvidimo.

Primenjene su sledeće tehnike: Stabla odlučivanja, naivni Bajesov klasifikator, K najbližih suseda i neuronske mreže. Korišćen je alat KNIME.

Sa podešavanjem parametara sve četiri navedene metode uspevaju da dostigu preciznos oko 0.9. Neuronske mreže se najbolje ponašaju za mali broj slojeva $[2, 4]$ i veći broj neurona $[8, 12]$ (videti: FS_NN_02.knwf). K najbližih suseda se pokazuje kao najlošija tehnika, a preciznost slabo varira sa promenom broja suseda (FS_KNN_02.knwf, FS_CV_KNN.knwf). Naivni Bajesov klasifikator takođe ima malo lošiju preciznost koja često zavisi od izbora trening skupa (FS_NB_01.knwf).

Kao najpreciznija se pokazala metoda sa stablom odlučivanja koja je takođe i najlakša za tumačenje upravo zbog postojanja stabala (FS_DTL_02.knwf). Što se parametara tiče mera kvaliteta za podelu nema značajnih razlika između Ginijevog koeficijenta sa MDL odsecanjima i bez. Parametri koji čine razliku su izbor atributa za podelu u korenu gde se preporučuje atribut *time*. Još jedan parametar koji dosta utiče na preciznost je izbor minimalnog broja instanci po čvoru. Iako preciznost varira za broj minimalnih instanci od 2 do 20 ipak za vrednost 1 je u proseku preko 0.95 na test skupu (FS_DTL_04.knwf). Međutim ovo može izazvati preprilagođavanje nad datim podacima jer iako su atributi celobrojnog tipa oni uzimaju samo mali broj različitih vrednosti (u najboljem slučaju samo 6).

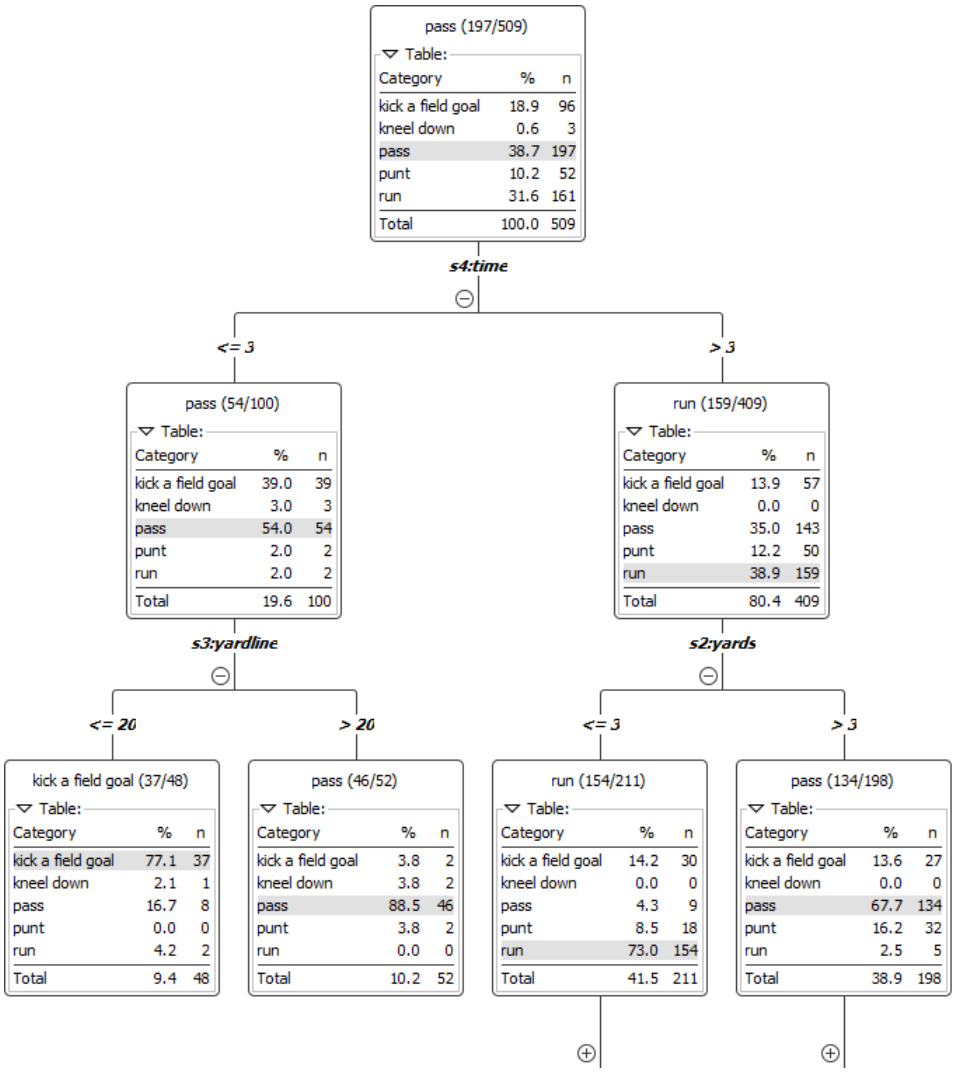
Prilikom primene ovih metoda dodatno su izostavljene i instance kod kojih je atribut *confidence* različit od 1 što nas dovodi od 2582 do 728 različitih unosa.

3 Rezultati

U nastavku slede rezultati i zaključci dobijeni primenom stabla odlučivanja nad datim skupom podataka. Jedini parametar koji se menja je izbor atributa za podelu u korenu (root split column). Svi ostali atributi ostaju nepromenjeni. Minimalni broj instanci po čvoru je 10.

3.1 Koji atribut treba prvo razmatrati prilikom donošenja odluke?

Kao što je već spomenuto izbor vremena kao podele u korenu daje najveću preciznost za stablo odlučivanja. Ovo se može videti na slici 1 ili u primeru FS_DTL_02.knwf gde se za zadate parametre 100 puta računa stablo i pronalazi prosek dobijenih preciznosti. Promena atributa *Root split column* daje manju preciznost. Ostali parametri su zanemarljivi pošto nemaju značajan uticaj izuzev minimalnog broja instanci po čvoru o kojem je već diskutovano.



Slika 1: Stablo odlučivanja sa *time* u korenu

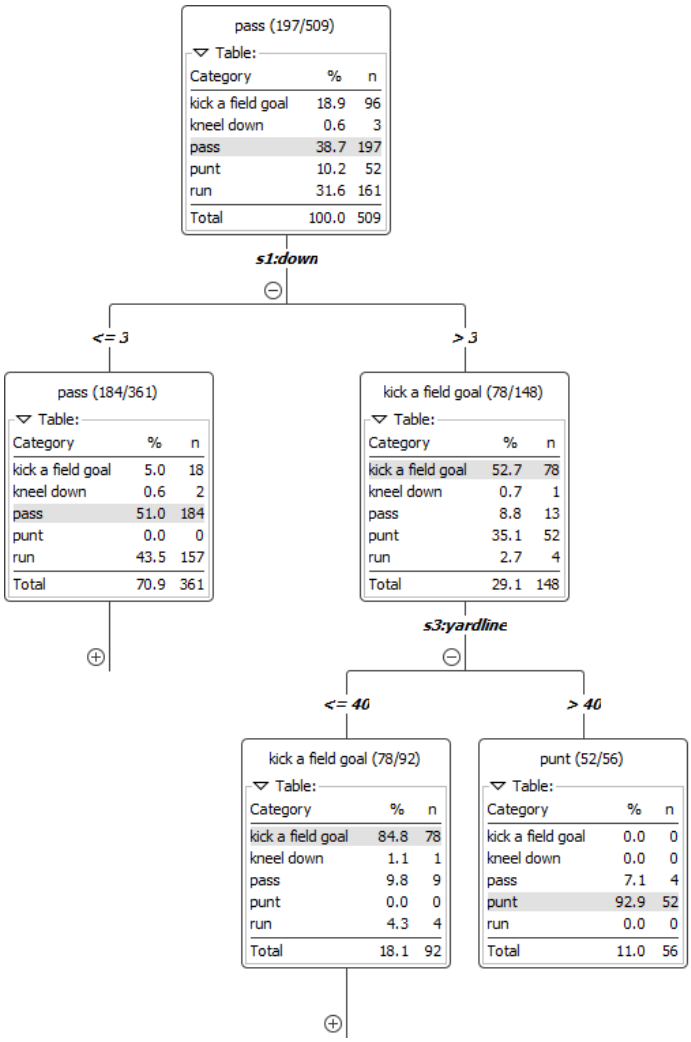
Rezultat koji stablo daje je sledeći: Prvo se vrši podela po vremenu i to tako da je jedna grana <= 3. U toj grani su skoro u potpunosti eliminisane akcije *punt* i *run*. Ove odluke su logične. U slučaju nedostatka vremena trčanje nije validna opcija, a ispucavanja (punt) nam ne donosi nikakvu korist jer se bliži kraj četvrtine i protivnik u svakom slučaju neće imati vremena za bilo kakvu kontra akciju. Odluku koju nam stablo preporučuje u ovoj situaciji zavisi od blizine protivničkog gola. Šut ako smo na 20 jardi ili manje ili pâs ako smo dalje od 20.

Druga grana koja potiče iz korena zahteva više razmatranja. Ali prva naredna podela značajno deli opcije između trčanja (run) koje se preporučuje ako

nam je ostao mali broj jardi i pâsa (pass) koji se preporučuje kada imamo više jardi da pređemo.

3.2 Kada je najbolje šutirati na gol ili ispucati loptu? (nasuprot tračanju i pâsu)

Prateći prethodni rezultat, sledeći po redu atribut koji ima najveću uticaj na preciznost je broj napada ili daun (down). Podela u korenu deli daun na prva tri i četvrti. Ukoliko smo na nekom od prva tri akcije (klase) koje preovladaju su pâs (pass) i trčanje (run). Dok za četvrti odnosno poslednji daun je obrnuta situacija i preovladaju šut (kick a field goal) i ispucavanje (punt). Delimično stablo se može videti na slici 2, a celokupno u primeru FS_DTL_02_01.knwf.



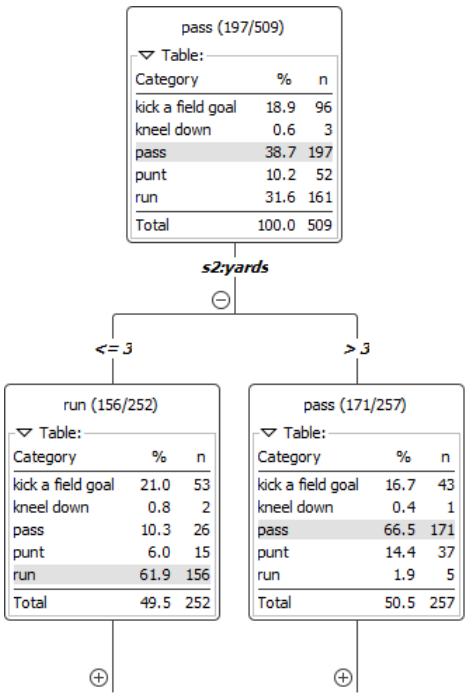
Slika 2: Stablo odlučivanja sa *down* u korenu

3.2.1 Podpitanje: Ispucavanje ili šut?

Odluka između ispucavanja ili šuta na gol se svodi na našu udaljenost od gola (slika 2). Za udaljenost veću od 40 jardi se ne preporučuje šut na gol već se ide na bezbednu opciju ispucavanja lopte. Za manje od 40 jardi šut je uvek bolja opcija. Jedini izuzetak (deo koji nije na slici) je kada smo u poslednjoj četvrtini i gubimo više od 3 razlike kada se preporučuje pâs jer nam 3 poena od gola nemaju veliki značaj.

3.3 Pâs ili trčanje?

Korena odluka koja nam nudi najveću podela između pâsa i tračanja je broj jardi koji moramo da pređemo u trenutnom napadu (slika 3). Međutim nešto precizniju odluku možemo da dobijemo ako prvo posmatramo vreme kao na slici 1 gde se preporučuje pâs kao bolji od trčanja u nedostatku vremena. Ukoliko je vreme ipak dostupno onda možemo doneti odluku na osnovu jardi gde se opet za 3 ili manje jardi preporučuje trčanje, a za više pâs.



Slika 3: Stablo odlučivanja sa *yards* u korenu

3.4 Kneel down. Retka opcija

Kneel down je opcija koja se vrlo retko koristi. Takođe je vrlo retka i u podacima koje obrađujemo. U sređenim podacima koji imaju 2582 instance javlja se samo 85 puta, a kada se ograničimo na 728 instanci koji imaju pouzdanost jednaku 1 onda samo 5 puta. Ovaj mali broj će biti zanemaren u stablima

odlučivanja pogotovo pošto smo koristili podrazumevanu vrednost od minimum 10 instanci po čvoru. Čak iako se pojavi čvor sa 10 instanci gde su baš svih 5 *kneel down* instanci a preostalih 5 takođe pripadaju istoj klasi i dalje nemamo garanciju da će konačna odluka za taj čvor biti *kneel down*. Ovo je još manje verovatno ako uzmemo u obzir da se tih 5 instanci dele na test i trening skup.

Pošto smo je ignorisali u prethodnim rezultatima i zaključcima sada ćemo navesti kada se ona koristi. Pošto imamo samo 5 instaci možemo ih sve navesti u tabli (1) i ručno razmatrati.

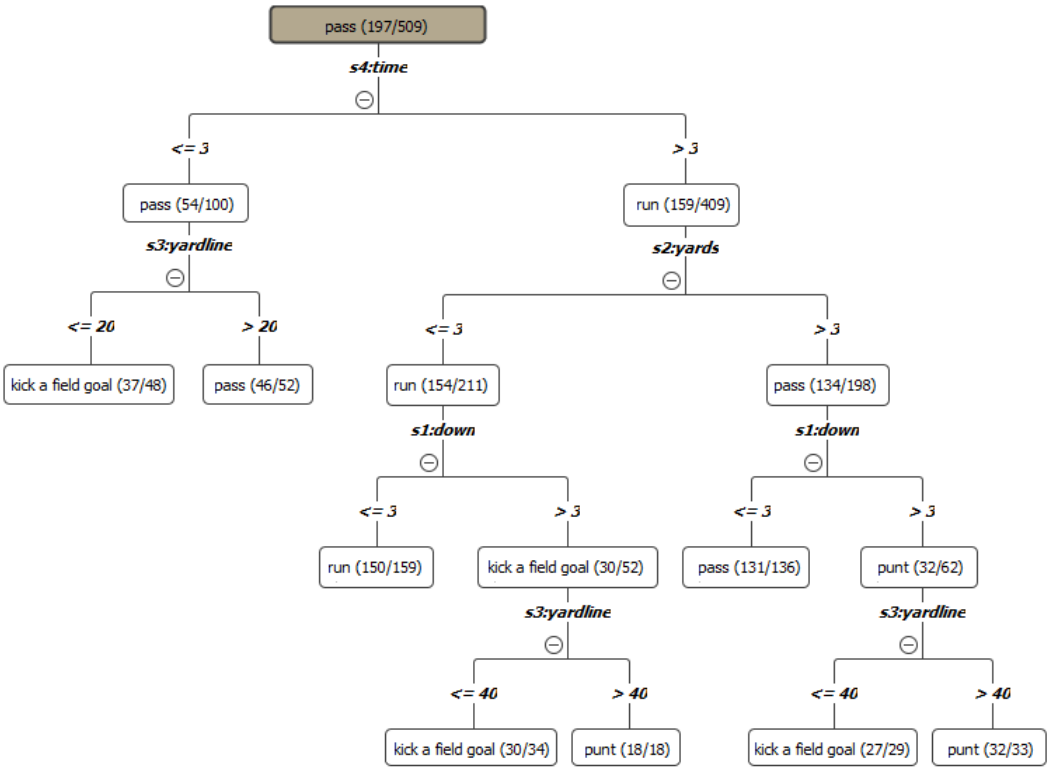
Tabela 1: Sve *kneel down* instance sa *confidence*=1

decision	confidence	down	yards	yardline	time	quarter	score
kneel down	1	3	1	55	3	4	3
kneel down	1	3	8	80	3	4	3
kneel down	1	3	8	80	3	4	7
kneel down	1	3	10	20	3	4	3
kneel down	1	4	1	40	3	4	7

Kneel down je akcija koja se koristi kada želimo da potrošimo vreme. U svim instancama ostalo je samo 3 sekunde do kraja utakmice i mi imamo prednost što predstavlja idealnu situaciju za *kneel down*.

4 Najbolje stablo

Za kraj nudimo primer stabla koji daje najbolju preciznost. U pitanju je stablo koje koristi vreme u korenu (slika 4).



Slika 4: Puno stablo odlučivanja sa *time* u korenu

A Dodatak

Korišćeni prevodi:

- down - daun / napad
- yardline - udaljenost od (protivničkog) gola
- kick a field goal - šut na gol
- punt - ispucavanje (lopte)
- pass - pâs (dobacivanje lopte)
- run - trčanje
- kneel down - *nije prevedeno*