Klasifikacija podataka prikupljenih u okviru eksperimenta o kratkim sastancima na slepo

Darinka Zobenica

Sažetak

U ovom radu prikazana je klasifikacija podataka prikupljenim u okviru eksperimenta vršenog u Kolumbiji u svrhu naučnog rada o razlikama između polova prilikom biranja partnera.[2] Ipak, eksperiment je prikupio mnogo drugih zanimljivih podataka o ispitanicima, te će u ovom radu biti predstavljeni rezultati dobijeni klasifikacijom o uticaju etničke pripadnosti, novčanih prihoda, zajedničkih interesovanja, i razlike u godinama, na izbor partnera.

Treba uzeti u obzir da su ispitanici iz SAD i da ne mora da znači da su rezultati relevantni na globalnom nivou, iako se daljim ispitivanjem može utvrditi da jesu.

Sadržaj

1 Priprema podataka

Kako korišćeni skup podataka sadrži 190 kolona i 8378 podataka, prvo sam izbacila kolone koje neće biti korišćene. Sačuvane relevantne kolone su:

- wave - Pošto je eksperiment ponavljan sa različitim ispitanicima tokom tri nedelje, u koloni wave nalaze se podaci o tome kada se određeni susret desio. Ova kolona ima celobrojne vrednosti od 1 do 21.
- age, age_o - Podaci o uzrastu ispitanika i njihovog partnera za dati sastanak, iskazani kao ceo broj (godina).
- race, race_o - Podaci o izraženoj etničkoj pripadnosti ispitanika i njihovog partnera za dati sastanak, gde brojevi sa sledećeg spiska odgovaraju brojevima u podacima:
  1. Negroidna rasa
  2. Evropeidna rasa
  3. Hispano i Latinoamerikanci
  4. Mongoloidna rasa
  5. Američki starosedeoci
  6. Ostali
- samerace - Oznaka da li su ispitanik i trenutni partner iste rase ili ne.
- imprace - Koliko je ispitanik označio da mu je rasa bitna u izboru partnera.
- int_corr - Korelacija između interesovanja ispitanika i trenutnog partnera, izvedena na osnovu popunjenog upitnika.

Napravila sam dve nove kolone na osnovu već postojećih:

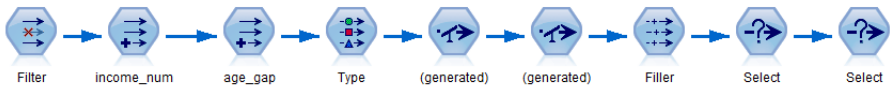
- age_gap - Brojčani podatak dobijen kao apsolutna razlika kolona age i age_o, koje se odnose na godine prvog i drugog ispitanika, redom.
- income_num - Brojčani podatak dobijen konverzijom niske iz kolone income u broj.

Primetila sam da nekoliko kolona koje sadrže kategoričke podatke nemaju jednaku količinu podataka za svaku od kategorija. Kako balansiranje podataka može dovesti do pristrasnosti modela ukoliko nema dovoljno podataka, odlučila sam da izbalansiram samo rasu. Pre balansiranja modeli su favorizovali belce kao najčešću etničku grupu među ispitanicima, a činjenica da je balansiranje poboljšalo preciznost krajnjeg rezultata govori o tome da u ovom slučaju do pristrasnosti nije došlo. Isti postupak nije primenjen na stavke kao što su starost i primanja, zbog manjka podataka za ekstremne vrednosti, a njihove istovremene relevantnosti za rezultate.

Tokom preprocesiranja, izbačeni su podaci za wave=12, jer su uslovi eksperimenta tog dana bili specifični (ispitanici su imali limitiran broj ljudi koje su mogli da odaberu da su im se svideli).

Za kraj, izbačeni su svi podaci za koje postoji nedostajuća vrednost u bilo kojoj od ovih kolona. Početni skup podataka je konzistentno sadržao isključivo sistemske nedostajuće vrednosti i dozvoljene vrednosti, tako da ništa drugo nije bilo potrebno uraditi.

Na sledećoj slici prikazani su korišćeni čvorovi:



Slika 1: Preprocesiranje

2 Klasifikacija

2.1 Algoritmi

2.1.1 C&R Tree

Algoritam C&R Tree je klasifikacija[4] i prediktivna metoda bazirana na strukturi stabla. Slično kao C5.0 algoritam koji će takođe biti korišćen, ova metoda koristi rekurzivno particionisanje da podeli trening podatke u segmente sa sličnim rezultatima. Proces počinje pregledom ulaznih vrednosti u cilju pronalaženja najbolje podele, a kvalitet podele se meri redukcijom mere nečistoće nakon podele. Podela definiše dve podgrupe, od kojih se svaka deli u još dve podgrupe, i tako dalje, dok god se ne ispuni jedan od kriterijuma za zaustavljanje. Sve podele su binarne.

Ovaj model je jako dobar za analizu nedostajućih podataka, kao i za korišćenje na velikom broju kolona. Nije potrebno mnogo vremena za trening. Uz to, dobijeni modeli su lakši za razumevanje od nekih dobijenih drugim algoritmima. Za razliku od C5.0, algoritam može da radi sa kontinualnim podacima kao rezultujućim vrednostima, ne samo kategoričkim.

2.1.2 CHAID

CHAID algoritam[5] je metod izrade stabla odlučivanja koji koristi hi-kvadrat statistiku za određivanje optimalnih podela. CHAID meri koje od ulaznih vrednosti su statistički značajne za određivanje ciljne vrednosti, i bira najznačajniju. Ako ima više od dve kategorije ulaznih vrednosti, one koje su jednako prediktivne se spajaju (inkrementalno redom one koje imaju najmanju razliku). Proces se zaustavlja kada razlika između kategorija dostigne predefinisanu vrednost.

CHAID radi za sve tipove ulaznih i izlaznih vrednosti, i pravi proizvoljna stabla (ne isključivo binarna).

2.1.3 C5.0

C5.0 algoritam[3] može da pravi stablo odlučivanja ili pravila pridruživanja. Ovaj model deli uzorak na osnovu atributa koji ima najveću informacionu dobit. Svaki poduzorak definisan prvom podelom se onda ponovo deli i proces se ponavlja dok god poduzorci i dalje mogu da se dele. Na kraju se vrši isecanje tako da se smanji greška zbog preprilagođavanja.

Ovaj algoritam radi samo za kategoričke ciljne vrednosti.

Prednosti algoritma su iste kao C&R Tree.

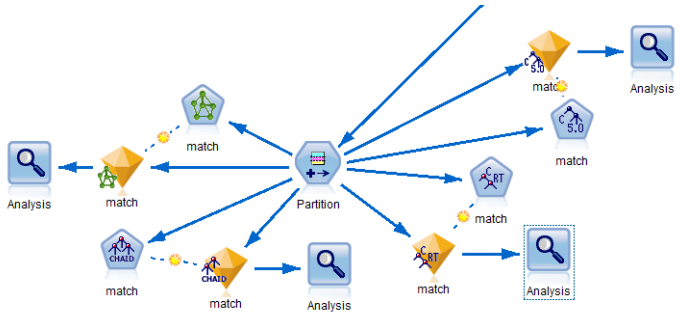
2.1.4 Neuronska mreža

Neuronska mreža[1] je pojednostavljen model načina na koji ljudski mozak procesira informacije. Radi tako što simulira veliki broj povezanih "neurona", podeljenih u ulazni, izlazni, i skriveni sloj. Mreža dodeljuje nasumične težine vezama između neurona, i onda ih optimizuje na osnovu podataka. Optimizacija radi tako što mreža generiše predikciju na osnovu podataka, a onda u zavisnosti od toga da li je predikcija tačna ili ne popravlja odgovarajuće težine grana grafa. Algoritam se zaustavlja kad se ispuni neki od kriterijuma za zaustavljanje.

2.2 Rezultati

Od svih ispitanika, 83.53% se nisu svideli međusobno, dok preostalih 16.47% jesu. Ovo znači da bi naši modeli imali visoku preciznost kad bi uvek pogađali da se par neće svideti jedan drugom, ali nama je cilj da upravo što češće pogodi te slučajeve, čak i po cenu nekog nivoa greške kod onih koji se ne svide jedni drugima.

Na sledećoj slici prikazan je deo toka podataka u kome se vrši klasifikacija:

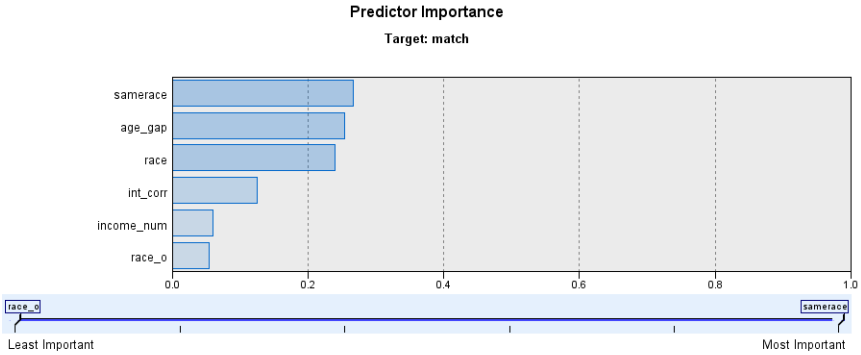


Slika 2: Klasifikacija SPSS

2.2.1 C&R Tree

C&R Tree je procenio da etnička pripadnost i razlike u godinama imaju najveći uticaj na to da li će se dvoje svideti jedno drugom na kratkom sastanku, dok su interesovanja i novčana primanja manje bitni.

Najbolje rezultate dobila sam korišćenjem Twoing algoritma.



Slika 3: Uticaj atributa C&R Tree

Klasifikacija je izvršena sa preciznošću od 81.6% na test skupu, a samo u 33.78% slučajeva pogađa kada će se dvoje svideti jedno drugom. Zaključujemo da je klasifikacija loše izvršena. twoig

Results for output field match

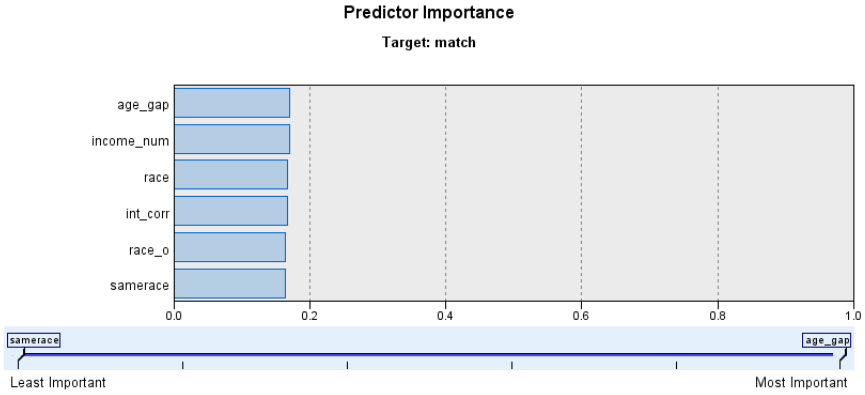
Comparing $R-match with match

'Partition'	1_Training		2_Testing	
Correct	15,005	81.24%	6,444	81.6%
Wrong	3,464	18.76%	1,453	18.4%
Total	18,469		7,897	

Slika 4: Analiza modela C&R Tree

2.2.2 CHAID

CHAID algoritam je procenio da su svi atributi jednako važni za predikciju. Model na test podacima ima preciznost od 82.72%, i pritom u samo 41% slučajeva pogađa kada će se dvoje svideti jedno drugom. Algoritam je procenio sve atribute kao skoro jednako bitne.



Slika 5: Uticaj atributa CHAID

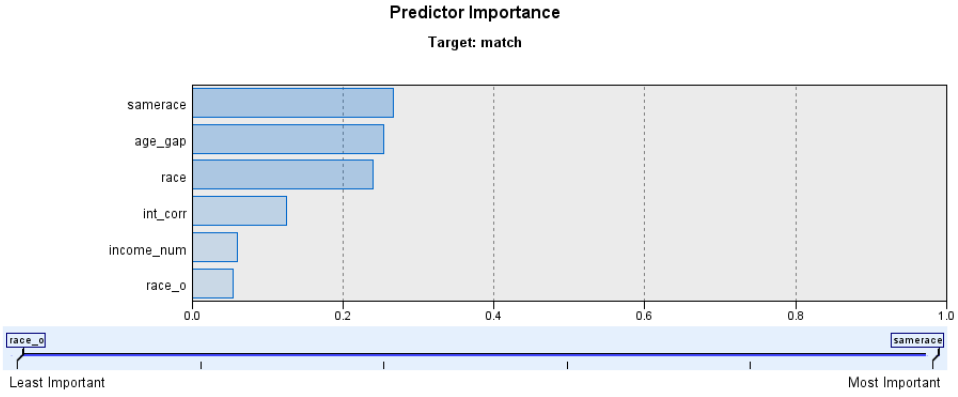
Results for output field match

Comparing $R-match with match

'Partition'	1_Training		2_Testing	
Correct	15,114	81.78%	6,536	82.72%
Wrong	3,368	18.22%	1,365	17.28%
Total	18,482		7,901	

Slika 6: Analiza modela CHAID

2.2.3 C5.0



Slika 7: Uticaj atributa C5.0

Algoritam je procenio da su rasa i razlika u godinama gotovo jednako bitni za izbor partnera, pri čemu je da li su partneri iste rase malo bitniji faktor. Iz dobijenog stabla odlučivanja se može videti da se partneri iste rase češće biraju međusobno, posebno u određenim demografijama. U poređenju s tim, novčana primanja i interesovanja se nisu pokazala kao toliko bitan faktor.

Results for output field match

Comparing $C-match with match

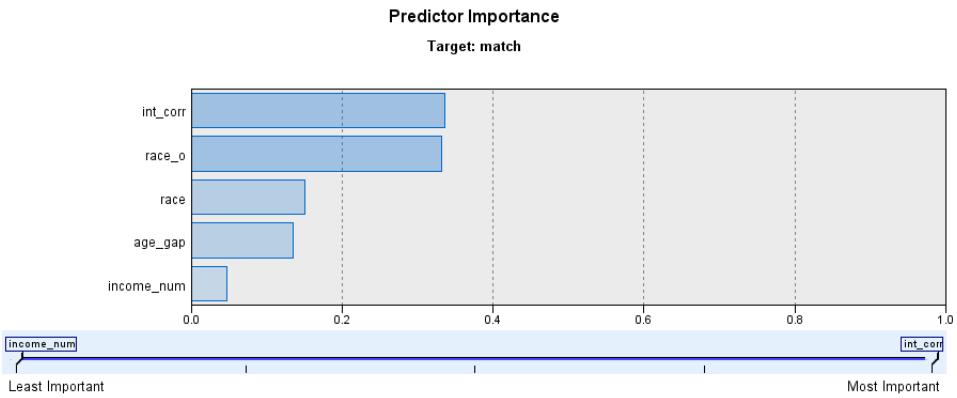
'Partition'	1_Training		2_Testing	
Correct	18,573	99.97%	7,933	99.96%
Wrong	6	0.03%	3	0.04%
Total	18,579		7,936	

Slika 8: Analiza modela C5.0

Algoritam C5.0 pokazao je najbolje rezultate, sa čak 99.96% pogođenih test slučajeva. Prilikom testiranja algoritam je pogrešio samo u 5 primera, od kojih je samo jedan u slučaju kada su se partneri jedan drugom svideli. Ovakve rezultate postigla sam korišćenjem optimizacija boosting i global pruning.

2.2.4 Neuronska mreža

Nakon C5.0, neuronska mreža je imala najbolje rezultate, sa 86.83% pogodaka i čak 57.33% pogodaka u slučaju kada su se partneri svideli jedno drugom. Kao najbitnije faktore procenila je zajednička interesovanja. Dalji trening nije popravljao dobijene rezultate.



Slika 9: Uticaj atributa Neuronska mreža

Results for output field match

Comparing $N-match with match

'Partition'	1_Training		2_Testing	
Correct	16,180	87.23%	6,881	86.83%
Wrong	2,369	12.77%	1,044	13.17%
Total	18,549		7,925	

Slika 10: Analiza modela Neuronska mreža

3 Zaključak

Najefikasnijim se pokazao algoritam C5.0 sa čak 99.96% pogođenih rezultata. Ostali algoritmi su se znatno gore pokazali, pri čemu je C&R Tree bio najgori.

Na osnovu rezultata C5.0 algoritma možemo zaključiti da su za ispitanike u Kolumbiji od testiranih faktora najbitniji godine i etnička pripadnost, dok su zarada i zajednička interesovanja manje bitni.

Literatura

[1] *SPSS Online Documentation - Neural Network Node.*

[2] Fisman, Iyengar, Kamenica, and Simonson. Gender differences in mate selection: Evidence from a speed dating experiment. *Quarterly Journal of Economics*, 2006.

[3] IBM. *SPSS Online Documentation - C5.0 Node.*

[4] IBM. *SPSS Online Documentation - C&R Tree Node.*

[5] IBM. *SPSS Online Documentation - CHAID Node.*