Klasifikacija podataka o kreditnim karticama

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Tamara Radovanović 430/2016

9. juni 2019

Sažetak

U okviru ovog rada biće prikazane metode za klasifikaciju koje su učene na kursu Istraživanje podataka i njihova primena na skupu podataka o kreditnim karicama.

Sadržaj

1 Uvod

Za istraživanje korišćen je skup podataka o klijentima banke. Ovaj skup podataka sadrži informacije o podrazumevanim plaćanjima, demografskim faktorima, kreditnim podacima, istoriji plaćaanja i računima klijenata kreditnih kartica na Tajvanu od aprila 2005. do septembra 2005. godine.

Baza sadrži 7000 instanci sa po 25 atributa. Ti atributi su:

- **ID:** identifikacija klijenta
- **LIMIT_BAL:** Količina dodeljenog kredita
- **EDUCATION:** edukacija (1=graduate school, 2=university, 3=high school, 4=other, 5=unknown)
- **SEX:** Pol (1=male, 2=female)
- **MARRIAGE:** Bračni status (1= married, 2=single, 3=other)
- **AGE:** Starost u godinama
- **PAY_0-PAY_6:** Status otplate u periodu od aprila do septembra 2005. godine
- **BILL_AMT1-BILL_AMT6:** Stanje računa od aprila do septambra 2005. godine
- **PAY_AMT1-PAY_AMT6:** Količina prethodne upalate od aprila do septambra 2005. godine
- **default.payment.next.month:** Redovna uplata (1=Yes, 0=No)

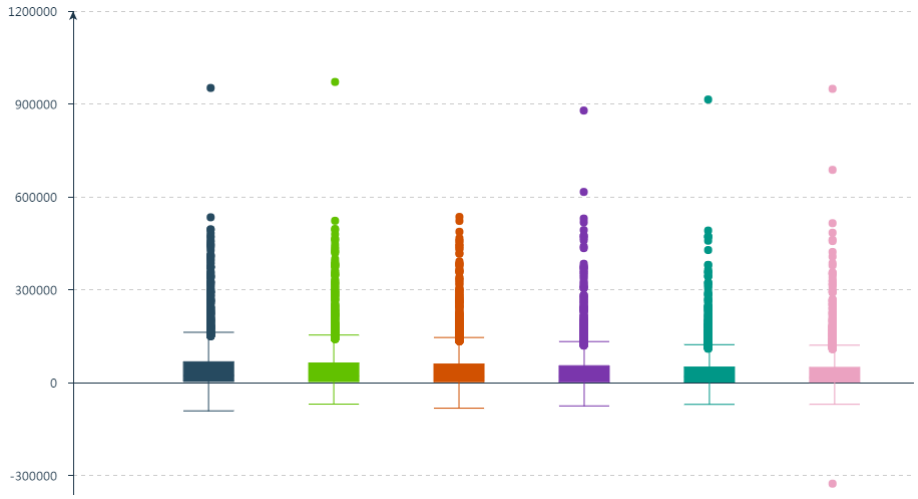
Nad ovim podacima su primenjeni algotirmi za klasterovanje. Prikazani su rezultati, njihovo međusobno uporođivanje.

2 Priprema podataka

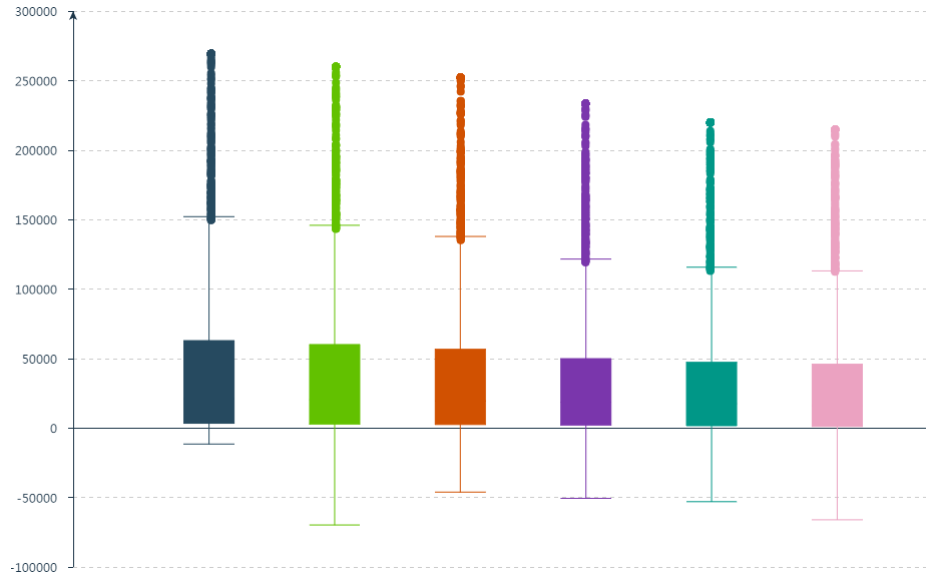
Podaci sadže informaciju o jednistenoj identifikaciji klijenta, koji ne utiču na verovatnoću da klijent vrati kredit. Iz tog razloga podaci o ID se ne uzimaju u razmatranje.

Podaci o BILL AMT i PAY AMT sadrže kontinualne podatke i za njih treba proveriti da li posotje ekstremne vrednosti i elementi van granice.Instance koje sadrže ekstremne vrednosti nisu uzete u razmatranje, a onima koje sadže elemente van granice su te vrednosti zamenjene najbližom vrednošću koja se ne smatra elementom van granice. Na slikama 1 i 2 se mogu videti vrednosti podataka pre i posle obrade ekstremnih vrednosti.

Atribut SEX sadrži vrednosti 1 i 2 i one su promenjene u "male"odnosno "female"radi bolje preglednosti. Kod EDUCATION vrednosti 1-3 su zamenjene redom sa "graduate school", "university"i "high school", a sve ostale sa "other". Kod atributa MARRIAGE urađeno je slično, vrednosti 1 i 2 su zamenjene sa "married"i "single", a ostale sa "other".



Slika 1: Vrednosti BILL_AMT atibuta pre uklanjanja ektremnih vrednosti



Slika 2: Vrednosti BILL_AMT atribura nakon uklanjanja ektremnih vrednosti

3 C5.0

Klasifikaicioni algoritam C5.0 pravi drvo odlučivanja na osnovu koga se određuje kojoj klasi instanca treba da pripada.

Primenom C5.0 algoritma se dobija model čija se uspešnost može videti u tebeli 1 kao i mtrica konfuzije u tabeli 2.

Tabela 1: Uspešnost C5.0. algoritme

	Trening		Test	
Tacni	12.278	82.5%	12.324	81.52%
Netacni	2.605	17.5%	2.793	18.48%

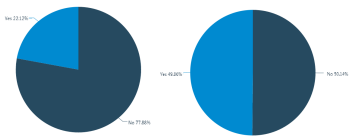
Tabela 2: Matrica konfuzije za C5.0 algoritam za test skup

	No	Yes
No	TN 11.255	FP 444
Yes	FN 2.349	TP 1.069

Iz matrice konfuzije možemo ustanoviti na koji nači model raporedjuje insnce po klasama:

- TPR= 0.312756 (stopa stvarno pozitvnih)
- TNR= 0.962048 (stopa stvarno negativnih)
- FPR= 0.037952 (stopa lažno pozitvnih)
- FNR= 0.687244 (stopa lažno negativnih)

Na osnovu stope stvarno negativnih i stvarno pozitivnih vidimo da je model skloniji da dobro klasifikuje instance iz "No"klase. Ovakvi rezultati su očekivani jer u korišćenom skupu preovlađuju instance koje pripadaju klasi "No", pa je potrebno uraditi balansiranje klasa. Na slici 3 se može videti odnos između klasa pre i posle balansiranja.



Slika 3: Odnos klasa pre i posle balansiranja

Za balansiranje klasa se mogu koristit dve tehnike. Prva je da se iz veće klasa izbaci deo instanci, tako da broj istanci ostane isti u obe klasa, dok druga podazumeva da se instance iz manje klase umnože nakon čega će klase biti balansirane. Primenjene su obe tehnike nad podacima i zatim algoritam C5.0, rezultati se mogu videti u tabelama 3 i 4. Matrica konfuzije koja je dobijen korišćenjem balansiranih podataka se može videti u tabeli 5, a mere za ocenu modela su sledeće:

- TPR= 0.822608 (stopa stvarno pozitvnih)
- TNR= 0.801415 (stopa stvarno negativnih)

- FPR= 0.198585 (stopa lažno pozitvnih)
- FNR= 0.177392 (stopa lažno negativnih)

Model napravljen pomoću blansiranih podataka nije nakoljen ni jednoj klasi i daje ujednačene razulate.

Tabela 3: C5.0 algoritam nad balansiranim podacima sa smanjenim brojem instanci "No"klase.

	Trening		Test	
Tacni	22,382	68.69%	6,447	69.12%
Netacni	12,204	31.31%	2,880	30.88%

Tabela 4: C5.0 algoritam nad balansiranim podacima sa povećanim brojem instanci "Yes"klase.

	Trening		Test	
Tacni	26,386	80,97%	7,593	81,41%
Netacni	6,201	19.03%	1.734	18.59%

Tabela 5: Matrica konfuzije za C5.0 algoritam nakon balansiranja za test skup

	No	Yes
No	TN 3.737	FP 926
Yes	FN 827	TP 3835

Nakon umnožavanja instanci klase "Yes"algoritam daje značajno bolje rezultate jer ne dolazi do gubitka informacija, pa će ova tehnika biti korišćena u ovom algoritmu, kao i u narednim algoritmima.

C5.0 algoritam pruža mogućnost boosting metode, kao i unkrsne validacije. Primenom ovih metoda uspešnost algoritma se povećava. Rezultati se mogu naći u tabeli 6. Preciznost algotitma kao i mera nečistoće računata preko Ginijevog indeksa se može videti u tablei 7.

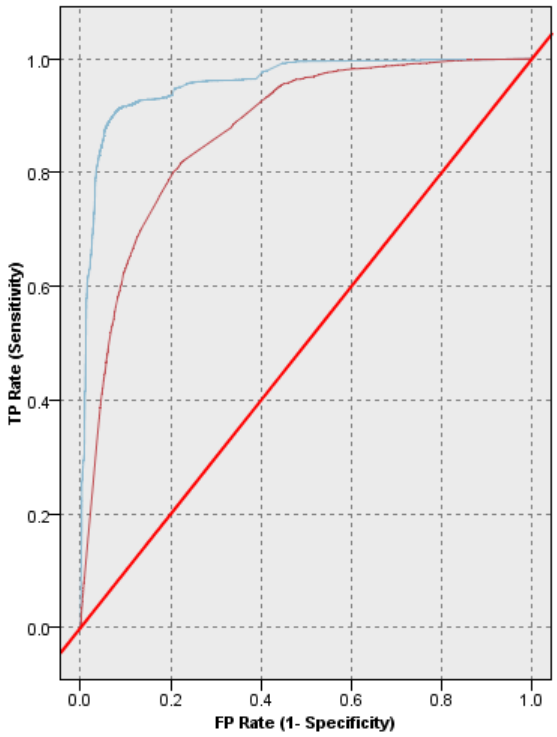
Tabela 6: C5.0 algoritam korišćenjem boosting metode i unakrsne validacija.

	Trening		Test	
Tacni	29,845	91,75%	8,528	91,67%
Netacni	2,682	8,25%	775	8,33%

Tabela 7: Preciznost algotitma C5.0 i mera nečistoće čvorova.

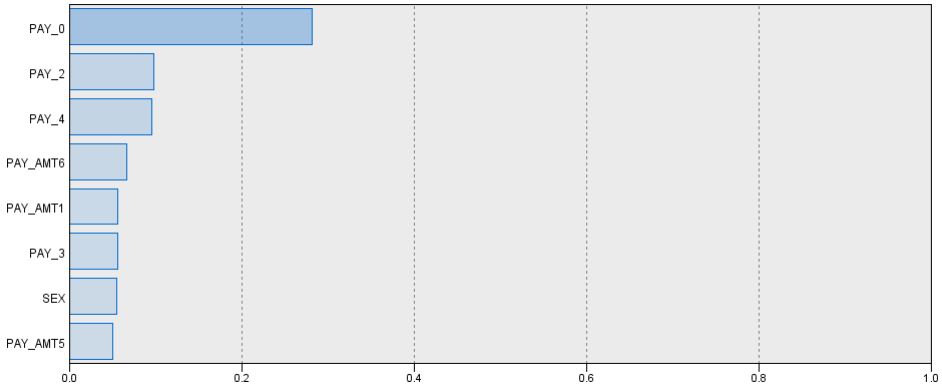
	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,97	0,94	0,97	0,94

Modeli se mogu vizuelno uporediti i korišćenjem ROC krive koja predstavlja grafički prikaz kompromisa izmedju TPR i FPR. Na grafiku 4 su prikazane ROC krive za C5.0 bez i sa korišćenja boosting metode.



Slika 4: ROC kriva za C5.0 algoritam sa i bez boosting metode

Nakon pravljenja modela može se videti koji atribut je u kojoj meri uticao na kasifikaciju. U ovim podacima najznačaniji atribut je PAY_0 što predstavlja staus otplate za prethodni mesec. Odnos atributa se može videti na slici 5.



Slika 5: Yastuljenost atributa u algoritmu C5.0

4 C&rt

Model kod ovog algoritma je kao i kod C5.0 drvo odlučivanja, ali za razliku od njega on pravi binarno drvo. U tabelama 8, 9 i 10 su prikazane uspešnosti modela dobijenih C&RT algoritmom bez dodtnih metoda, i sa boosting i bagging metodama.

Tabela 8: C&R tree

	Trening		Test	
Tacni	23.051	70.81%	9.915	70.58%
Netacni	9.502	29.19%	4.133	29.42%

Tabela 9: C&R tree sa korišćenjem boostin metode

	Trening		Test	
Tacni	22.960	70.56%	9.835	70.03%
Netacni	9.581	29.44%	4.209	29.97%

Tabela 10: C&R tree sa korišćenjem bagging metode

	Trening		Test	
Tacni	22.960	70.56%	9.835	70.03%
Netacni	9.581	29.44%	4.209	29.97%

U tabeli 11 možemo videi preciznost algoritma kao i meru nečistoće čvorova. Na osnovu dobijenih rezultata možemo da zaključimo da ovaj algoritam pravi lošiji model nad datim podacima.

Tabela 11: Preciznost algotitma C&RT i mera nečistoće čvorova.

	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,747	0,494	0,755	0,509

Kod ovih podataka se može pretpostaviti da je bitnije smanjiti mogućnost da se instance klase "No"klasifikuju kao instance klase "Yes". Da bi se ovo postiglo može se povećati cena tog promašaja. Nakon primene algoritma dobijaju se sledeći rezulati iz tebele 12 i matrica konfuzije 13. Odavde dobijamo sledeće rezultste:

- TPR= 0.380181 (stopa stvarno pozitvnih)
- TNR= 0.936788 (stopa stvarno negativnih)
- FPR= 0.063212 (stopa lažno pozitvnih)
- FNR= 0.619819 (stopa lažno negativnih)

Procenat uspesnosti modela se smanjio, ali je stopa lažno negativnih manja.

Algoritmi kod kojih je model drvo odlučivanja imaju pozitivnu stranu da rezultati mogu da budu lako interpretirani, odnosno nakon klasifikacije

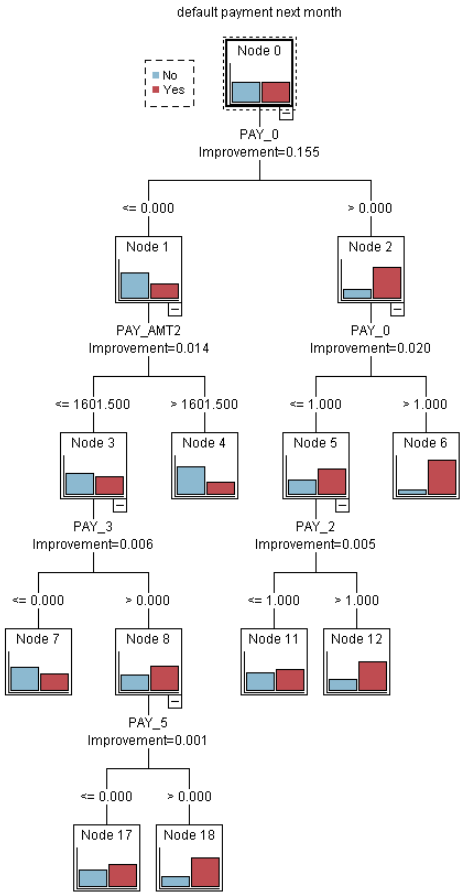
Tabela 12: C&R tree sa promenjenom cenom promašaja

	Trening		Test	
Tacni	20.653	65.84%	8.859	65.49%
Netacni	10.716	34.16%	4.669	34.51%

Tabela 13: Matrica konfuzije za C5.0 algoritam za test skup

	No	Yes
No	TN 6.254	FP 422
Yes	FN 4.247	TP 2.605

znamo zašto je instanca pripala baš toj klasi. Drvo odlučivanja koje se dobija korišćenjem C&RT algoritma se moze videti na slici 6.



Slika 6: Drvo odlučivanja dobijeno C&RT algoritmom

5 KNN

KNN je algoritam kod koga je klasifikacija instance zasnovana na sličnosti sa drugim instancama. Kako se koriste mere rastojanja, potrebno je da vrednosti svih atributa budu u istom opsegu da jedan atribut ne bi vršio preveliki uticaj u odnosu na ostale. Za predprocesiranje kontinualnih atributa koristi se normalizacija, dok se kategorički atributi transformišu u binarni vektor čija je dimenzija broj različitih klasa datog atributa. Ovo predprocesiranje mozemo da obavimo sami, a ukoliko ga ne uradimo SPSS modeler će to uraditi umsto nas.

Tabela 14: KNN algoritam bez dodatnih opcija.

	Trening		Test	
Tacni	26.618	85.85%	11,437	84.69%
Netacni	4.691	14.15%	2.067	15.31%

Tabela 15: Preciznost algotitma KNN i mera nečistoće čvorova

	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,931	0,862	0,931	0,862

Pravljenje modela pomoću KNN algoritma nad podacima velikih dimenzija može da oduzme dosta vremena, zato SPSS modeler omogućava uključivanje opcije za pravljenje modela sa redukovanim vremenom. Rezultati dobijeni primenom ove metodom se mogu videti u tebeli 16, a preciznost i mera nečistoće u 17.

Tabela 16: KNN algoritam sa smanjenim vremenom izvršavanja.

	Trening		Test	
Tacni	26,708	85.3%	11,478	84.98%
Netacni	4.604	14.7%	2.029	15.02%

Tabela 17: Preciznost algotitma KNN smanjenim vremenom izvršavanja i mera nečistoćce čvorova

	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,935	0,869	0,928	0,857

6 Neuralne mreže

Neuoronske mreze daju model koji koji simulira rad nervnog sisetam. Ovaj algoritam je dobar za široku upotrebu nad podacima bez dodatnih pretpostavki. Rezultati primene algoritama bez dodatnih opcija, kao i sa boosting metodom se mogu videtu u tabelama 18, 20 .

Tabela 18: Neuronske mreže bez dodatnih opcija.

	Trening		Test	
Tacni	22.239	70.8%	9,535	70.39%
Netacni	9,172	29,2%	4,011	29,61%

Tabela 19: Neuronske mreže bez dodatnih opcija. Preciznost i Gini

	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,789	0,578	0,789	0,577

Tabela 20: Neuronske mreže sa korišćenjem boosting metode.

	Trening		Test	
Tacni	22.501	71.73%	9,706	71.76%
Netacni	8.866	28.27%	3,819	28.24%

Tabela 21: Neuronske mreže sa korišćenjem boosting metode. Preciznost i Gini

	Trening		Test	
Model	Preciznost	Gin	Preciznost	Gini
	0,789	0,578	0,789	0,577

7 Zaključak

Cilj istraživanja je bio da se napravi model za klasifikaciju klijenata banke na osnovu toga da li će izmiriti svoje obaveze za sledeci mesec. U istraživanju je korišćeno nekoliko različitih načina za predprocesiranje i algoritama za pravljenje modela.

Ispitani algoritmi su C5.0 i C&RT koji prave drvo odlučivanja, KNN koji poredi sličnost sa ostalim instancama i neuronske mreže. Na osnovu svih ispitanih rezultata može se zaključiti da C5.0 algoritam sa boosting metodom daje najbolje rezultate, dok nam C&RT algoritam daje najjednostavniji model.