Klasterizacija podataka korisnika kreditnih kartica

Uvod

Klasterizacija je forma nenadgledanog učenja u kojem grupišemo podatke koje imamo, u cilju izdvajanja grupa koje sadrže neku zajedničku osobinu. Jedan od jednostavnijih formi klasterizacije je K Means algoritam, koji radi tako što nasumično inicijalizuje centre klastera i „grabi“ sve podatke koji su blizu. Zatim izračunava novi centar na osnovu svih novododatih podataka i ponavlja ovaj proces dok se centri više ne pomeraju. Algoritam zahteva parametar koji predstavlja broj klastera, do kojeg se može doći na nekoliko načina. Način na koji smo odredili broj klastera je opisan u sekciji 2.

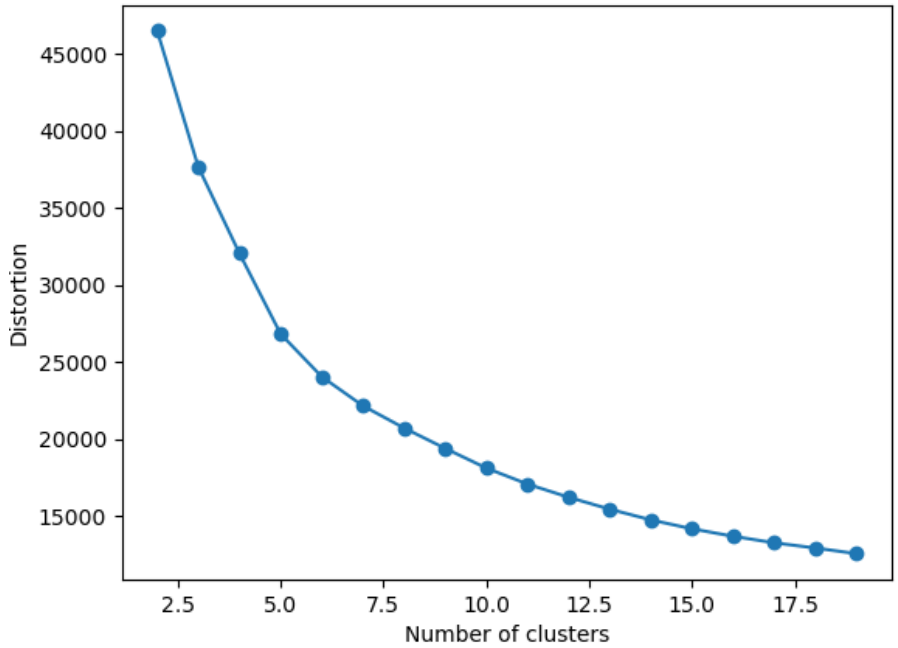
1. Pretprocesiranje

Pre primene algoritma, potrebno je odraditi pretprocesiranje, tj. pripremu podataka za algoritam. Prvo smo proverili da li ima podataka koji nedostaju. U kolonama minimum payments i credit limit smo imali ovakve podatke, te smo njih dopunili sa medijanom iz njihovih respektivnih kolona. Zatim smo primenili normalizaciju, tj. svođenje podataka na slične raspone u cilju podataka koji su centrirani oko 0 i imaju manju standardnu devijaciju. Takođe, sačuvali smo sve originalne podatke da bismo prilikom tumačenja podataka imali realne vrednosti, zbog lakšeg izvođenja zaključaka, a pretprocesirane podatke smo koristili prilikom primene K Means algoritma.

Pošto su nam podaci vrlo visoke dimenzionalnosti (za svaku osobu postoje 17 značajnih parametara), gledali smo da izdvojimo atribute koji će najviše (i najbolje) opisati izdvojene grupe. Na početku smo pokušavali sa svim podacima da klasterujemo, ali nismo dobijali lepo definisane klastere. Zbog ovoga smo ostavili samo podatke koji će nam najviše pomoći da objasnimo obrasce ponašanja neke grupe. Naša razmišljanja i odabiri atributa su opisani u poglavlju 3.

2. Određivanje broja klastera

Za određivanje broja klastera smo koristili Elbow method, koji se svodi na primenu K Means algoritma za svaki broj u nekom rasponu, i zatim se proverava vrednost od *sum-of-squares* grešaka. Broj klastera za koji greška naglo prestaje da se smanjuje predstavlja optimalan broj klastera. Primenom ovog algoritma i gledajući na plot (slika 1), nismo dobili jasnu granicu, te smo iskoristili implementaciju formule sa ovog sajta. Inicijalno smo dobili 8 za broj klastera prilikom korišćenja nenormalizovanih podataka, a 9 prilikom korišćenja normalizovanih. Nakon što smo dopunili nedostajuće vrednosti (opisano u pogavlju 1) i posmatrali samo selekciju kolona (a ne sve) dobili smo 7 za broj klastera, što smo koristili za sve naše testove.



Slika 1. Vizuelizacija elbow algoritma za odabir broja klastera. Ne vidi se jasna granica te smo matematički izračunali gde nam se najviše „isplati“ da stanemo.

3. Posmatranje određenih kolona za klastere

Kako su podaci veoma visoke dimenzionalnosti, odlučili smo da koristimo samo deo kolona prilikom pravljenja klastera. Krenuli smo od najopštijih kolona – balance (stanje na računu, važno zbog opšteg uvida u finansijsko stanje osobe), purchases (koliko troše na kupovinu), credit limit (koliko najviše mogu da se zaduže), payments (koliko su uplatili novca na račun) i minimum payments (koliko novca su najmanje uplatili), kako bismo grubo videli kakve klastere dobijamo. Rezultati su sledeći (vrednosti su u šablonu – prosečna vrednost (standardna devijacija)):

Klaster	Balance	Purchases	Credit limit	Payments	Min. payment	Broj
Klaster 0	6058 (2145)	978 (1472)	9080 (3315)	2520 (2045)	2100 (1287)	947
Klaster 1	813 (851)	471 (628)	2260 (1151)	868 (948)	474 (737)	5510
Klaster 2	1112 (1129)	1380 (1367)	7455 (2816)	1994 (1752)	423 (480	2047
Klaster 3	3511 (2641)	870 (1732)	3737 (3023)	1656 (3307)	15955 (6153)	73
Klaster 4	5351 (4235)	27085 (8459)	5626 (7500)	27158 (9708)	3201 (4916)	25
Klaster 5	3135 (2726)	5504 (3853)	9324 (4262)	10088 (5921)	1395 (1759)	341
Klaster 6	6674 (2595)	1714 (2797)	6457 (2622)	2169 (1914)	52620(13197)	7
Svi podaci	1564 (2081)	1003 (2136)	4494 (3638)	1733 (2895)	844 (2332)	8950

Iz ovoga smo videli da nismo baš najbolje odabrali kolone – broj podataka u nekim klasterima je vrlo mali, što smo očekivali, ali ne u ovolikoj meri (imamo klaster sa 7 ljudi i jedino što ih izdvaja je jako velik minimum payments), a dva klastera zajedno uključuju preko 80% podataka. Zbog velikog raspona minimum payments kolone i ne toliko velikog semantičkog značaja u tumačenju osobina ljudi iz klastera, odlučili smo da izbacimo ovu kolonu.

Umesto nje, ubacili smo tri nove kolone – purchase frequency (koliko često osoba kupuje generalno), purchase installments frequency (koliko često kupuje na rate) i prc full payments (koliki procenat nije otplaćen od purchases, tj. koliko novca osoba duguje banci). Sa ovim novim kolonama, dobili smo sledeće rezultate:

Kl.	Balance	Purchases	Purchase frequency	P. inst. frequency	Credit limit	Payments	Prc full payments	Broj
0	4015 (2623)	2894 (2016)	92% (12%)	76% (26%)	8926 (3449)	3103 (2164)	2% (7%)	732
1	1018 (905)	268 (490)	15% (18%)	5% (10%)	2970 (2078)	965 (1262)	4% (10%)	3712
2	810 (826)	969 (869)	86% (16%)	73% (27%)	3024 (2030)	1072 (1023)	7% (12%)	2230
3	3520 (3502)	6589 (4429)	75% (36%)	5% (40%)	10379 (4450)	13224 (6550)	35% (37%)	174
4	5351 (4235)	27085 (8459)	89% (23%)	70% (37%)	16360 (5626)	27159 (9708)	49% (41%)	25
5	136 (165)	1329 (1381)	77% (27%)	58% (36%)	4855 (3728)	1541 (1532)	81% (19%)	1182
6	5103 (2404)	343 (636)	16% (22%)	5% (12%)	8898 (3187)	2748 (2639)	1% (5%)	895
Svi	1564 (2081)	1003 (2136)	49% (40%)	36% (39%)	4494 (3638)	1733 (2895)	15% (29%)	8950

Ovog puta smo dobili mnogo bolje rezultate – imamo dva veća klastera koji su jasno razvojeni (klasteri 1 i 2, koji imaju sličan balance, credit limit i payments, ali zato se jasno razlikuju u purchases, purchase frequency i purchase installments frequency, što nas navodi da zaključimo da klaster 2 uključuje ljude koji često uzimaju na rate i redovno otplaćuju), tri srednja klastera (klasteri 0, 5 i 6, svaki od kojih ima svoje značajne karakteristike koje će biti diskutovane u sledećoj sekciji) i dva klastera sa malim brojem ljudi, ali koji jasno opisuju ljude koji mnogo troše. Detaljna analiza ovih klastera, uz naše pretpostavke i zaključke sledi u narednom poglavlju.

4. Analiza rezultata klasterovanja

Klaster 0 – Buržuji: „Ima se, može se!“

Prvi iz grupe srednje velikih klastera, sa poprilično visokim stanjem na računu (3. po redu, oko 2.5 puta više od proseka), ali i velikom količinom novca potrošenog na kupovine (skoro 3 puta više od proseka). Takođe, imaju najveću frekvenciju kupovine i kupovine na rate, kao i dosta visoke uplate (skoro 2 puta više od proseka). Zbog visokog stanja i uplata, credit limit im je isto veoma visok (3. po redu, oko 2 puta viši od proseka). Međutim, skoro sve što su kupili su i otplatili, što signalizira da su ovo najodgovorniji i najredovniji potrošači. Jedina loša strana je što ih ima relativno malo.

Klaster 1 – Prosečni Srbin: „Pokrij se kol’ko imaš!“

Grupa koja ima najveći broj potrošača, imaju balans koji je ispod proseka, purchases koji su drastično ispod proseka (4 puta manje), frekvenciju kupovine i kupovine na rate koji su takođe veoma niski, i payments koji su takođe ispod proseka. Dakle, ova grupa ljudi ne troši mnogo, relativno često uzima na rate (33% kupovina), ima srazmerno mali credit limit sa njihovim balansom i uplatama, a i ne zadužuju se mnogo. Činjenica da je u ovom klasteru skoro 50% od posmatranih podataka pokazuje da je ovo prosečan korisnik kreditne kartice.

Klaster 2 – Vrsni ekonomisti: „Ma to kad uzmeš na rate izađe te džabe...“

Druga najveća grupa potrošača, koji po mnogim osobinama liče prethodnom klasteru (sličan balans, limit i payments), sa nekoliko ključnih razlika – oni jako često kupuju (čak 86%, 3. po redu od svih), i najčešće na rate (73%, drugi po redu), sa tim da se ne zadužuju mnogo, tj. redovno otplaćuju račune. Možemo zaključiti da novac koji uplate ide na otplatu rata, te je ovo glavna karakteristika ove grupe.

Klaster 3 – Šopingholičari: „Plaćam karticom...“

Drugi najmanji klaster po broju ljudi i po količini novca potrošenoj na kupovine. Dakle, ovo su srednji potrošači, koji jako često kupuju (6.5 puta veći od prosečnog), ali zato jako retko uzimaju na rate (samo 5%, a prosek je 36%). Imaju drugi najveći credit limit (2 puta veći od prosečnog) zbog visokog stanja na računu i velike količine uplaćenog novca (8 puta veći od prosečnog). Imaju oko 35% procenta neplaćenih dugova, što je dosta visoko (2 puta više od proseka), ali i dalje manje od najvećih potrošača, koji su naša sledeća grupa.

Klaster 4 – Bahati: „Malo Havaji, skupe stvari, skup ferari!“

Klaster koji ima najmanji broj ljudi, ali zato su prvi u gotovo svakoj drugoj kategoriji – balance im je oko 3.5 puta veći od prosečnog, purchases oko 6.5 puta, purchase frequency skoro 2 puta, purchase installments oko 2 puta veće, credit limit 4 puta veći, payments čak 16 puta veći od prosečnog, a prc full payments 3 puta veći. Dakle, ovo su najveći potrošači, jako mnogo troše i jako često kupuju. Imaju visok balans, velike uplate i credit limit, ali zato imaju i veliki dug koji moraju da izmire.

Klaster 5 – Utopisti: „Sve će se rešiti kad dobijem na sedmicu na Loto-u...“

Najveći po broju ljudi od srednje-velikih klastera, ovde spadaju ljudi sa najmanjim stanjem na računu (10 puta manje stanje od prosečnog), ali koji i dalje jako često kupuju i troše na kupovinu, više nego što su u mogućnosti da plate. Oni imaju najveći procenat neotplaćenih dugova, te bi se mogli klasifikovati kao high-risk korisnici.

Klaster 6 – Štediše: „Valja imati za crne dane!“

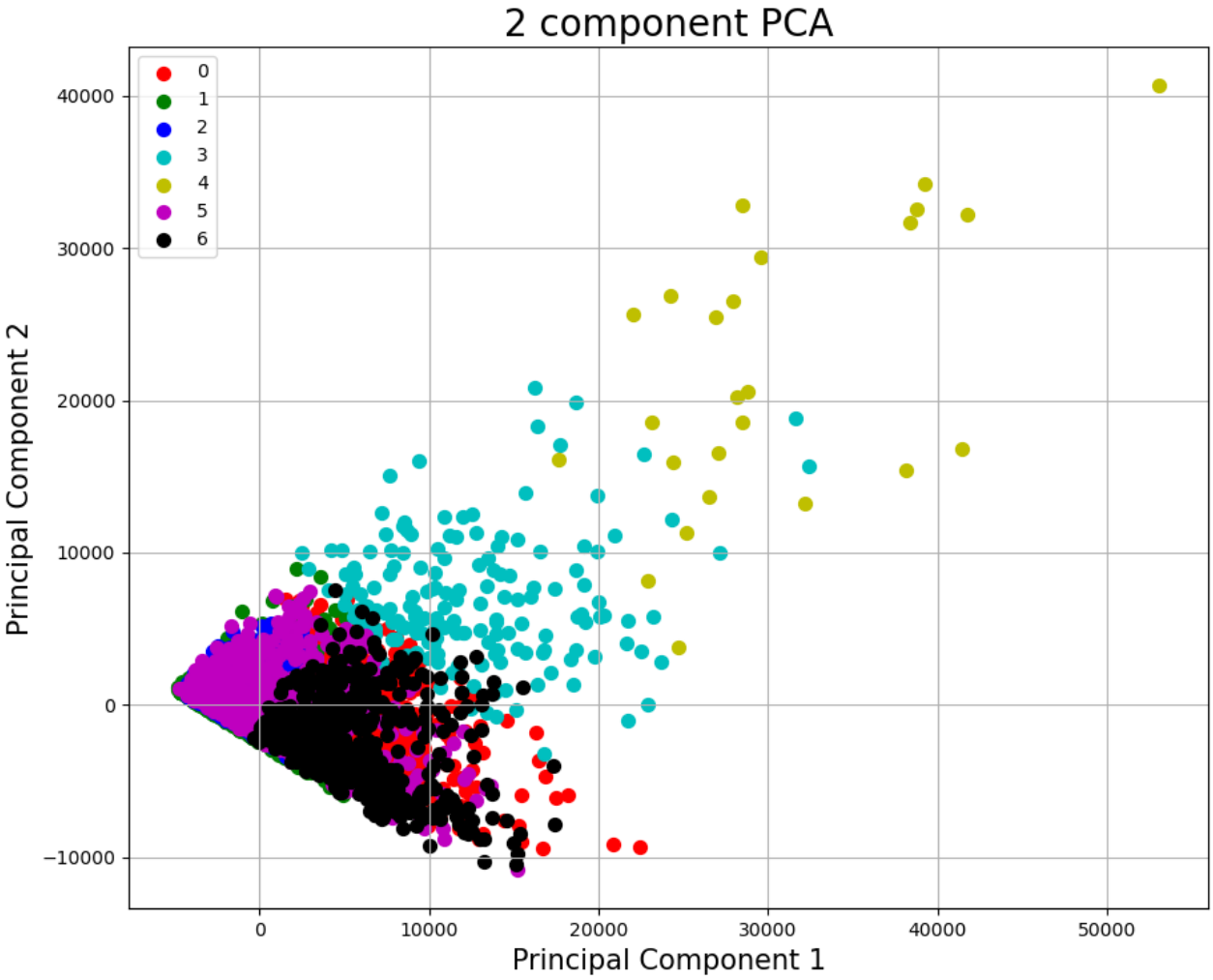
Poslednji od 3 srednje-velikih klastera, ovde spadaju ljudi koji imaju veoma veliko stanje na računu, ali zato skoro ništa ne troše (stanje je oko 3 puta veće od proseka, a novac potrošen na kupovinu je 3 puta manji od proseka). Zbog velikog stanja na računu imaju 2 puta veći od prosečnog credit limita, i uplate isto pokazuju ovo (malo manje od 2 puta veće od proseka). Frekvencija kupovine je niska, kao i frekvencija kupovine na rate i procenat koji duguju. Dakle, ovo su ljudi koji štede novac, a kada kupuju, ne troše mnogo novca i relativno često kupuju na rate (33% vremena uzimaju na rate).

5. Vizueliacija klastera u 2D prostoru

Pokušali smo da koristimo dva algoritma, PCA i TSNE, kako bismo prikazali klastere u 2D prostoru, u cilju vizuelizacije naših rezultata. Naši podaci imaju 17 dimenzija, te je potrebno smanjiti tu dimenzionalnost na 2.

PCA algoritam

PCA algoritam gleda koji podaci objašnjavaju najveću količinu varijanse u podacima.

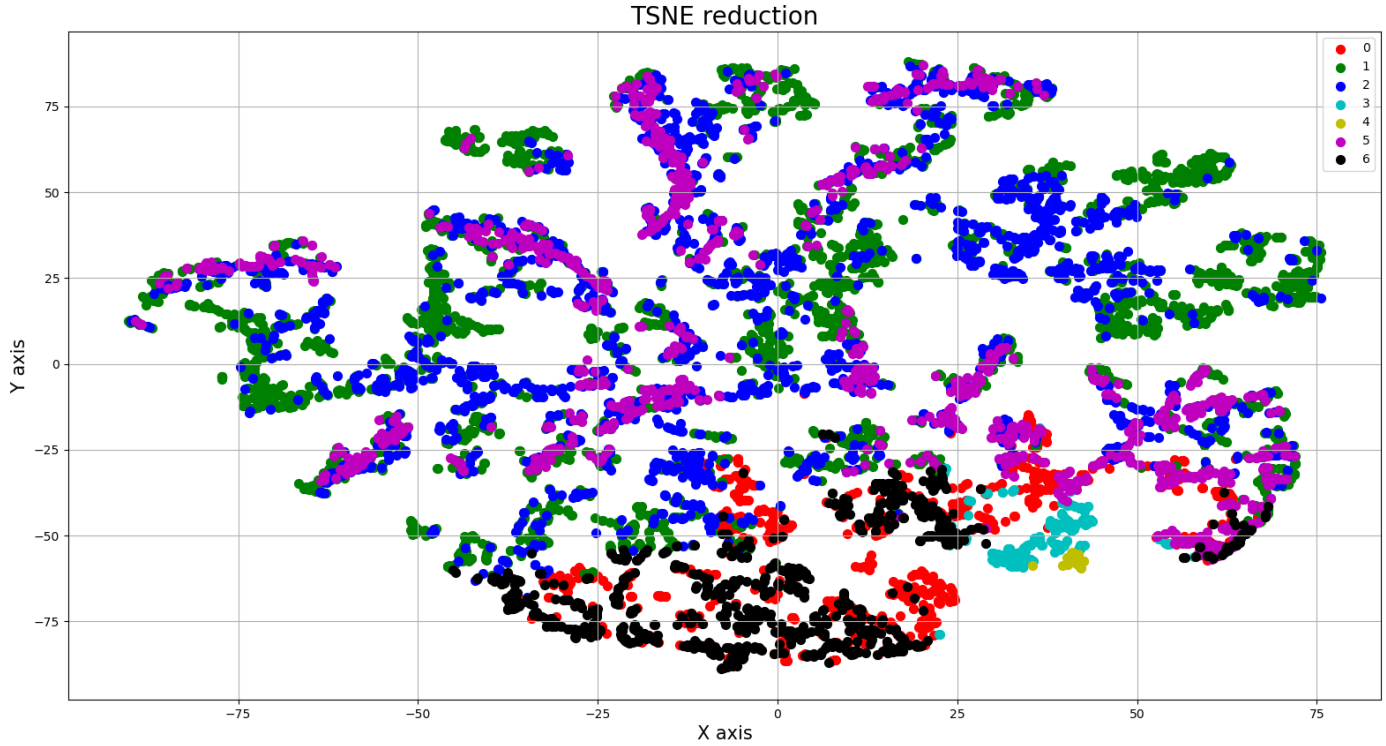


Slika 2. Rezultati PCA vizuelizacije u 2D prostoru.

Vidimo da algoritam nije uspeo baš najlepše da vizualizuje podatke, verovatno zbog velikih varijansi u određenim klasterima, a u isto vreme zbog velike sličnosti u značajnom broju osobina. I dalje možemo videti da su najveći outlier-i iz klastera koji mnogo troše, dok su ostali slični, te možemo zaključiti da su purchases i payments kolone koje su prikazane na x i y osi ovog grafa.

TSNE algoritam

Sledeće smo pokušali TSNE, koji gleda različitosti u svakoj od dimenzija i time postiže bolji rezultat. Međutim, njega tumačimo samo vizuelno. Dobijen grafik je sledeći:



Slika 3. Rezultati TSNE algoritma za vizuelizaciju u 2D prostoru.

Rezultati su ovog puta mnogo bolji i pregledniji – gledajući donji deo, vidimo da su klasteri 0 i 6 dosta blizu, što im i realno odgovara, jer je najveća razlika između njih zapravo koliko često kupuju. Desno od njih vidimo klastere 3 i 4, koji su pomalo izdvojeni od svih ostalih (to su naši veliki potrošači). Ostatak podataka su iz klastera 1, 2 i 5, što su zapravo najveći klasteri, te imaju i veliki raspon između njihovih podataka.

Pošto ne postoje jasno definisane granice između klastera, pretpostavljamo da nije trebalo ni da očekujemo takav rezultat (mnogo osobina su deljene iz grupe u grupu i vrednosti imaju veoma velike raspone). Podaci iz ovih klastera su generalno iz sličnih opsega, što možemo da vidimo i na slici, ali i dalje imaju dovoljno unikatnih osobina da smatramo da je algoritam uspešno odradio klasterizaciju, a ova reprezentacija nam potvrđuje to.

Box plotovi i pair plotovi

Box plotovi koji ilustruju podatke iz tabele sa rezultatima se nalaze u folderu *pokusaj 2* i predstavljaju vizuelni prikaz podataka koje smo diskutovali ovde. Tamo se nalazi i pair plot graf, koji predstavlja zavisnosti svake kolone sa svakom drugom, i može biti koristan da se vidi otprilike kako je prošla klasterizacija (da li su se podaci razdvojili ili nisu). Ove plotove nismo stavili u dokument jer smo smatrali da su ovi koje smo stavili najreprezentativniji, ali slobodno možete baciti pogled.

Zaključak

Ova tematika je vrlo interesantna i korisna za izučavanje i mislimo da smo postigli dobre rezultate za naše prvo susretanje sa klaster-izazovima. Primene ovih metoda su široke – konkretno za naš problem, banke bi mogle da koriste klasterizaciju nad sličnim podacima kao što smo i mi koristili u cilju odabira ciljnih grupa za povlastice ili da procene koliko je rizično nekome dati kredit i slično.