Rezultatele experimentelor de clusterizare

in faza 1 a proiectului Hyperloop

Contents

Exemplul 1: Analiza Customer Revenue & Margin pentru clientii judetului Prahova in anul 2016

Pe axa verticala avem marja realizata iar pe axa orizontala avem cifra totala de afaceri. Algoritmul bazat pe kMeans a determinat 4 segmente optime pentru care a propus 4 categorii valorice de clienti (rosu: clientii cei mai valorosi, albastru: clientii buni, verde: clienti medii si mov clienti cei mai putin importanti dpdv al cifrei de afaceri si a profitului net generat). De mentionat ca pentru fiecare client atat cifra de afaceri cat si profitul net au fost “comprimate” sub forma unui indicator cu valori intre 0 si 1 (prin scaderea minimului si impartirea la maxim). De remarcat “linia” de la mijlocul graficului reprezentata de clientii care au facut marja negativa (profitul total generat din tranzactii este negativ) care se intinde de la clientii cu cifra de afaceri mica pana la clinetii cu cifra de afaceri mare.

Datorita clientilor “exceptie” (outliers) graficul arata cel putin ciudat (este vorba de clinetii din categoria de top care au facut o cifra f mare afaceri si au generat un profit f mare comparativ cu restul clientilor).

In urma aplicarii unei logaritmari naturale a celor doua valori (cifra si profit) s-a eliminat impactul negativ al clientilor “exceptie” asupra analizei datelor si acum se pot vizualiza in forma de mai jos:

Practic prin aplicarea logaritmarii naturale discrepantele intre clientii f buni si cei mai putin buni au fost reduse fara a se distruge impartirea logica a acestora in functie de performanta. De remarcat “linia” de la mijlocul graficului reprezentata de clientii care au facut marja negativa (profitul total generat din tranzactii este negativ) care se intinde de la clientii cu cifra de afaceri mica pana la clinetii cu cifra de afaceri mare insa este putin “ridicata” fata de graficul nelogaritmat (evident datorita logaritmarii).

O alta varianta mai logica (ClusterID: 20170208160241 in [HyperloopCluster] pe TESTEBI-2012) desi vizual arata mai putin intuitiv decat imaginea de mai sus este data de liniarizarea exclusiva pentru marja absoluta dupa care scalarea min-max. In aceasta varianta "segmentul" de clienti cu marje negative (linia) nu mai este distribuit pe mai multe clustere ci este inclus in worst-customers-segment: (obs: legenda din pacate a for suprapusa peste scatter-plot. deasemenea ordinea label-urilor pe legenda este putin anapoda insa numele si culorile sunt relevate: rosu pentru cei mai buni, albastru pentru cei buni, verde medii si mov pentru worst case):

Exemplul 2: Analiza pe Recenta, Frecventa si Masa monetara a clientilor pe judetul Prahova inanul 2016

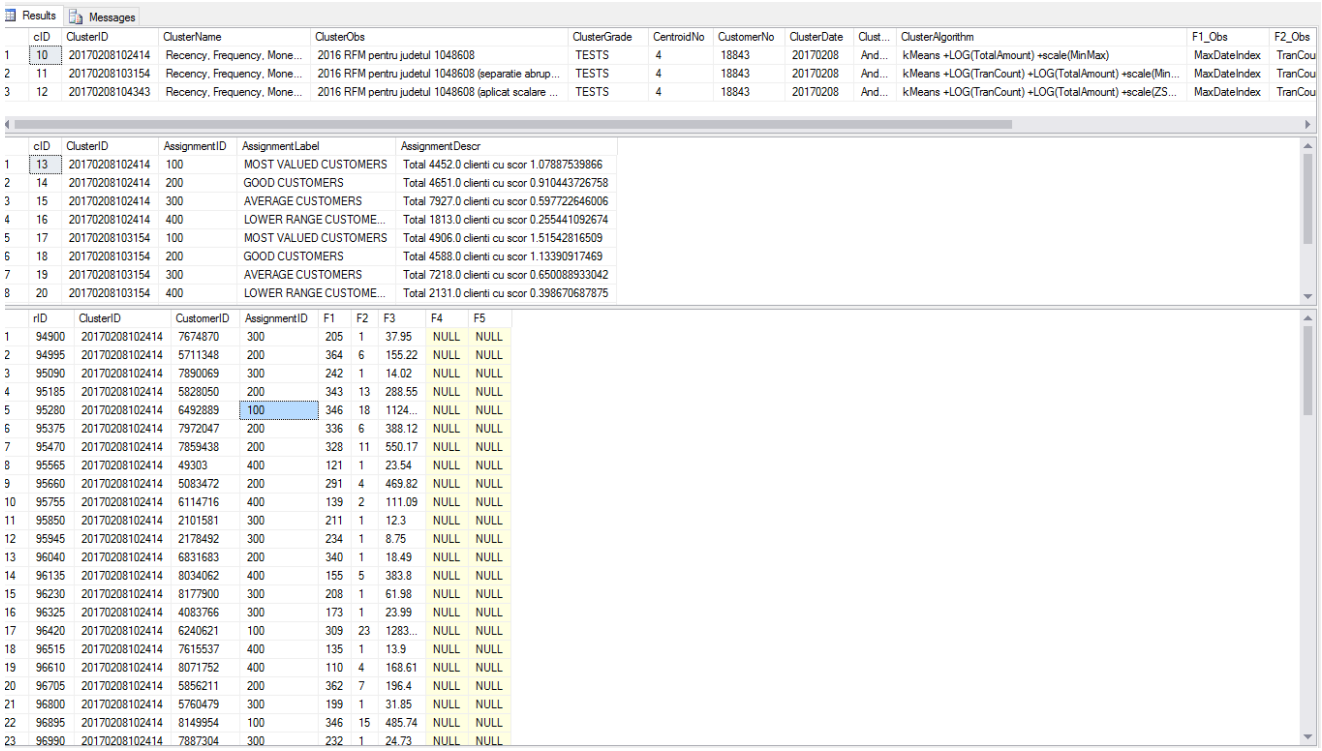
Graficul de mai jos este un exemplu de vizualizare 2D al unei segmentari 3D (bazata pe 3 atribute conform descrierii de mai sus) in care avem pe orizontala recenta iar pe verticala frecventa.

Datele au fost normalizate pentru a incapea atat plajele de valori ale cifrei de afaceri cat si frecventele si recenta in acelasi interval 0-1. Recenta este data de un index de recenta in care 1 reprezinta cea mai buna recenta pe intervalul respectiv de timp iar 0 reprezinta cea mai proasta recenta sau practic cumpararea doar in prima zi a intervalului analizat. Frecventa este data to de un index calculat pentru care 1 reprezinta cele mai multe tranzactii din populatie iar 0 cele mai putine. De mentionat ca dimensiunea atributului masa monetara este "in spate" 3D – acest grafic este o proiectie in 2D.

Dupa cum se observa frecventa a fost log-scalata (logaritmata natural) pentru a accentua diferentele intre clienti cu frecventa mare (zona de sus) si cei cu frecvente f mici (zona de jos) precum si a scapa de influenta nefasta a outlier-ilor in analiza segmentarii. (OBS: Legenda nu este ordonata cum trebuie dar din label/explicatii se poate intelege).

Disponibilitate datelor – importarea in excel

Atat pentru exemplu 1 cat si pentru exemplul 2 datele pot fi cu usurinta importate in Excel intr-o forma total accesibila cu multiple explicatii in limbaj natural. Mai jos este data o descriere a structurii si semanticii datelor:



Results Messages

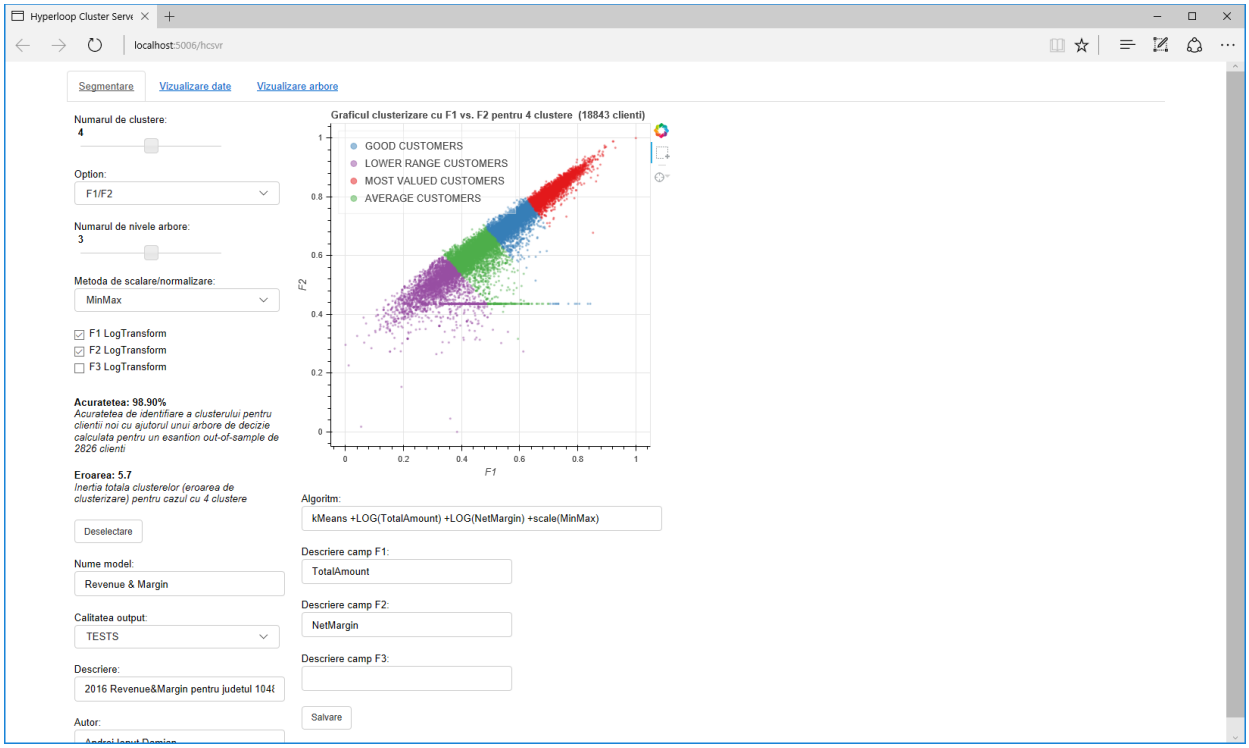
	cID	ClusterID	ClusterName	ClusterObs	ClusterGrade	CentroidNo	CustomerNo	ClusterDate	Clust...	ClusterAlgorithm	F1_Obs	F2_Obs
1	10	20170208102414	Recency, Frequency, Mone...	2016 RFM pentru judetul 1048608	TESTS	4	18843	20170208	And...	kMeans +LOG(TotalAmount) +scale(MinMax)	MaxDateIndex	TranCou
2	11	20170208103154	Recency, Frequency, Mone...	2016 RFM pentru judetul 1048608 (separatie abrup...	TESTS	4	18843	20170208	And...	kMeans +LOG(TranCount) +LOG(TotalAmount) +scale(Min...	MaxDateIndex	TranCou
3	12	20170208104343	Recency, Frequency, Mone...	2016 RFM pentru judetul 1048608 (aplicat scalare ...	TESTS	4	18843	20170208	And...	kMeans +LOG(TranCount) +LOG(TotalAmount) +scale(ZS...	MaxDateIndex	TranCou

	cID	ClusterID	AssignmentID	AssignmentLabel	AssignmentDescr
1	13	20170208102414	100	MOST VALUED CUSTOMERS	Total 4452.0 clienti cu scor 1.07887539866
2	14	20170208102414	200	GOOD CUSTOMERS	Total 4651.0 clienti cu scor 0.910443726758
3	15	20170208102414	300	AVERAGE CUSTOMERS	Total 7927.0 clienti cu scor 0.597722646006
4	16	20170208102414	400	LOWER RANGE CUSTOME...	Total 1813.0 clienti cu scor 0.255441092674
5	17	20170208103154	100	MOST VALUED CUSTOMERS	Total 4906.0 clienti cu scor 1.51542816509
6	18	20170208103154	200	GOOD CUSTOMERS	Total 4588.0 clienti cu scor 1.13390917469
7	19	20170208103154	300	AVERAGE CUSTOMERS	Total 7218.0 clienti cu scor 0.650088933042
8	20	20170208103154	400	LOWER RANGE CUSTOME...	Total 2131.0 clienti cu scor 0.398670687875

	rID	ClusterID	CustomerID	AssignmentID	F1	F2	F3	F4	F5
1	94900	20170208102414	7674870	300	205	1	37.95	NULL	NULL
2	94995	20170208102414	5711348	200	364	6	155.22	NULL	NULL
3	95090	20170208102414	7890069	300	242	1	14.02	NULL	NULL
4	95185	20170208102414	5828050	200	343	13	288.55	NULL	NULL
5	95280	20170208102414	6492889	100	346	18	1124...	NULL	NULL
6	95375	20170208102414	7972047	200	336	6	388.12	NULL	NULL
7	95470	20170208102414	7859438	200	328	11	550.17	NULL	NULL
8	95565	20170208102414	49303	400	121	1	23.54	NULL	NULL
9	95660	20170208102414	5083472	200	291	4	469.82	NULL	NULL
10	95755	20170208102414	6114716	400	139	2	111.09	NULL	NULL
11	95850	20170208102414	2101581	300	211	1	12.3	NULL	NULL
12	95945	20170208102414	2178492	300	234	1	8.75	NULL	NULL
13	96040	20170208102414	6831683	200	340	1	18.49	NULL	NULL
14	96135	20170208102414	8034062	400	155	5	383.8	NULL	NULL
15	96230	20170208102414	8177900	300	208	1	61.98	NULL	NULL
16	96325	20170208102414	4083766	300	173	1	23.99	NULL	NULL
17	96420	20170208102414	6240621	100	309	23	1283...	NULL	NULL
18	96515	20170208102414	7615537	400	135	1	13.9	NULL	NULL
19	96610	20170208102414	8071752	400	110	4	168.61	NULL	NULL
20	96705	20170208102414	5856211	200	362	7	196.4	NULL	NULL
21	96800	20170208102414	5760479	300	199	1	31.85	NULL	NULL
22	96895	20170208102414	8149954	100	346	15	485.74	NULL	NULL
23	96990	20170208102414	7887304	300	232	1	24.73	NULL	NULL

1. In partea de sus este prezentata tabela care descrie fiecare clusterizare realizata si salvata in baza de date cu explicatii extinse.
2. In partea de mijloc sunt prezentate – pentru fiecare clusterizare – care sunt nivelele de segmentare si cum a ajuns algoritmul la concluzia ca pentru un anumit clientii sunt la un anumit nivel valoric in sens de marketig/vanzari. De exemplu segmentul 100 al clusterizarii 20170208102414 este format din cei mai buni clienti ai analizei deoarece cei 4452 au o medie maxima a indexului de recenta, frecventa si masa monetara realizata pe intervalul analizat (1.0788) fata de celelalte categorii de clineti (0.9 pentru cei 4651 de clienti buni, 0.59 pentru cei 7927 de clienti medii si respectiv 0.25 pentru cei 2513 clienti “slabi”)
3. In partea de jos a imaginii sunt prezentate efectiv “asocierile” realizate de catre algoritmii de machine learning pentru fiecare client in cadrul fiecarei clusterizari fiind incluse si atributele care au generat respetiva segmentare (F1, F2, F3, F4, F5). In cazul de fata datele vizualizabile sunt din cadrul unei clusterizari (denumita 20170208102414) pentru care s-au folosit trei atribute de segmentare si practic F1 reprezinta recenta, F2 reprezinta frecventa si F3 reprezinta masa monetara tranzactionata de catre client.

Deasemena a fost finalizata o aplicatie web-based cu componenta de machine learning care trage automat din mssql datele, face inferente dupa care incarca rezultatele. Mai jos un exemplu al utilizarii ei pentru generarea clusterizarii pe revenue/netmargin.


Hyperloop Cluster Serv
localhost:5006/hcsvr
Segmentare
Vizualizare date
Vizualizare arbore
Numarul de clustere:
4
Option:
F1/F2
Numarul de nivele arbore:
3
Metoda de scalare/normalizare:
MinMax
F1 LogTransform
F2 LogTransform
F3 LogTransform
Acuratetea: 98.90%
Acuratetea de identifiare a clusterului pentru clientii noi cu ajutorul unui arbore de decizie calculata pentru un esantion out-of-sample de 2826 clienti
Eroarea: 5.7
Inertia totala clusterelor (eroarea de clusterizare) pentru cazul cu 4 clustere
Deselectare
Nume model:
Revenue & Margin
Calitatea output:
TESTS
Descriere:
2016 Revenue&Margin pentru judetul 104
Autor:
Graficul clusterizare cu F1 vs. F2 pentru 4 clustere (18843 clienti)
GOOD CUSTOMERS
LOWER RANGE CUSTOMERS
MOST VALUED CUSTOMERS
AVERAGE CUSTOMERS
F2
F1
Algoritm:
kMeans +LOG(TotalAmount) +LOG(NetMargin) +scale(MinMax)
Descriere camp F1:
TotalAmount
Descriere camp F2:
NetMargin
Descriere camp F3:
Salvare