

# **MATRECOM - Model de analiza si inferenta a**

## **Matricei Recomandarilor si a Vectorilor de**

### **Comportament ai clientilor**

History			
NR	Data	Autor	Ver
1	19.01.2017	A.I. DAMIAN	Draft 1
2	30.01.2017	A.I. DAMIAN	Draft 2
3	01.02.2017	A.I. DAMIAN	Draft 3
4	16.02.2017	A.I. DAMIAN	Draft 4
5	21.02.2017	A.I. DAMIAN	Draft 5

## Contents

Analiza .....	2
Atribute generate de sistem .....	8
Considerente de implementare .....	11

## Analiza

Intregul model inferential si predictiv poate fi explicat intuitiv prin urmatorul exemplu concret.

În vederea construirii exemplului vom pleca de la mai multe ipoteze-pași după cum urmează:

- 1) Se presupune ca avem o baza de date a produselor farma in baza caruia sa putem construi un set de proprietati extinse ale fiecarui produs individual. Sa presupunem urmatoarea tabela snapshot ipotetic de cateva produse farma:

ID Produs	Nume	ID Producator	UM	PU	Marja	Categorie Generala	Clasificare
...	...	...	...	...	...	...	...
17321	Nurofen	100	Pastila	2.15	0.3	Analgezice	Farmaceutice
34	Sampon AAA	203	Cutie	59.99	0.4	Sampoane	Cosmetice
32543	Vitamina C	455	Blister	20.00	0.25	Suplimente 1	OTC
47	Balsam B	203	Cutie	49.99	0.4	Sampoane	Cosmetice
1801	Sare	122	Punga	3.44	0.15	Suplimente 2	OTC
...	...	...	...	...	...	...	...

Din tabela standard de mai sus putem deriva o serie de attribute calculabile prin SQL, care vor imbogati tabela si o vor duce la urmatoarea forma finala:

Coloane originale								Coloane construite						
1	2	3	4	5	6	7	8	1	2	3	4	5	6	.
ID Pro dus	Nume	ID Produc tor	UM	PU	Mar ja	Catego rie Genera la	Clasifi care	Este Analg ezic	Est e Sam pon	Este Cosm etic	Es te OT C	Este Occi tane	Es te GN C	. . . . .
...	...	...	...	...	...	...	...	...	...	...	.. .	...	.. .	. . . . .
173 21	Nuro fen	100	Past ila	2.1 5	0.3	Analge zice	Farmace utice	1	0	0	0	0	0	. . . . .
34	Samp on AAA	203	Cuti e	59. 99	0.4	Sampo ne	Cosmeti ce	0	1	1	0	1	0	. . . . .
325 43	Vita mina C	455	Blis ter	20. 00	0.2 5	Suplim ente 1	OTC	0	0	0	1	0	1	. . . . .

47	Bals am B	203	Cuti e	49. 99	0.4	Sampo a ne	Cosmeti ce	0	1	1	0	1	0	.
180 1	Sare	122	Pung a	3.4 4	0.1 5	Suplim ente 2	OTC	0	0	0	1	0	0	.
...	...	...	...	...	...	...	...	...	...	...	..	...	..	.
											.		.	.

Pe langa attributele

- 2) Plecand de la certitudinea ca pentru fiecare client posesor de card actualmente se poate determina numarul de unitati/tranzactii realizate pe un anumit produs se va putea genera o tabela care sa arate de forma urmatoare:

Tabelul aferent ID CLIENT 32245529 pe perioada 01.01.2016-01.01.2017								
ID Produs	Numar tranzactii normalizat (T)	Este Analgezic	Este Sampon	Este Cosmetic	Este OTC	Este Occitane	Este GNC	...
...	...	...	...	...	...	...	...	...
17321	0.5	1	0	0	0	0	0	...
34	0.9	0	1	1	0	1	0	...
32543	0.1	0	0	0	1	0	1	...
47	0	0	1	1	0	1	0	...
1801	0	0	0	0	1	0	0	...
...	...	...	...	...	...	...	...	...

Numarul de tranzactii standardizat este fie 0 in situatia in care acel produs nu a fost cumparat niciodata de catre clientul nostru sau:

$$T = \frac{A - \text{Min}(A)}{\text{Max}(A)}$$

Unde A este lista numarului de tranzactii efectuate pentru toate produsele cumparate de client. Practic T va avea aprox 1 pentru produsul cel mai cumparat, 0.5 pentru un produs

cumparat in medie si aproape de 0 pentru produsele cumparate o singura data in perioada de timp.

***Astfel, scopul intregului model si al Matricii de Recomandari este sa determinam coeficientul de tranzactii (sau potential de cumparare) T al fiecaruia din clienti pentru toate produsele – atat pentru cele pe care le-a cumparat DAR MAI ALES pentru cele pe care NU le-a cumparat niciodata (cele cu T=0)***

- 3) Deja la pasul 3 putem aplica un algoritm de machine learning de tip regresie care sa ne construiasca un model pentru a determina **vectorul de comportament** al clientului “**ID CLIENT 32245529**”. Acest vector de comportament va putea fi utilizat pentru determinarea “potentialului de cumparare” sau mai bine zis al predictiei de cumparare pentru un anumit produs de catre clientul dat.

Astfel plecand de la premiza ca vom construi un total de N attribute de produs (ce cele 6 din exemplul naiv de mai sus) pentru fiecare client vom putea construi un model individual de forma

$$\theta^{(Client_i)} = \left( \theta_{EsteAnagezic}^{(Client_i)}, \theta_{EsteSampon}^{(Client_i)}, \theta_{EsteOccitane}^{(Client_i)}, \dots, \theta_N^{(Client_i)} \right)$$

Unde

$i \in [1..M]$ ,  $M = \text{numarul total de clienti}$ ,  $N = \text{numarul total de attribute (coloane) construite}$ .

In continuare, reprezentand un produs sub forma vectoriala:

$$X^{(Produs_j)} = \left( x_1^{(Produs_j)}, x_2^{(Produs_j)}, x_3^{(Produs_j)}, \dots, x_N^{(Produs_j)} \right)$$

unde  $j \in [1..O]$ ,  $O$  fiind numarul total de produse iar  $N$  fiind ca si in cazul vectorului client *numarul total de attribute (coloane) construite*.

Vom putea realiza o predictie a potentialului de vanzare a unui produs  $k$  folosind produsul scalar al celor doi vectori reprezentati de un client  $i$  si un produs netranzactionat sau total nou  $k$  conform urmatoarei formule generalizate:

$$\text{Predictie Potential Cumparare } (C_i, P_k) = h(C_i, P_k) = \Theta^{C_i^T} X^{P_k} = \sum_{z=1}^N \theta_z^{C_i} x_z^{P_k}$$

Evident functia prezentata aici de ipoteza a predictiei este una liniara si poate fi inlocuita cu modele mai complexe bazate pe retele neurale artificiale in vederea determinarii unei aproximari mai reale a regresiei urmarite.

Concret pentru clientul nostru **32245529** putem sa realizam urmatoarele analize:

- a) Sa realizam inferenta cosului de cumparaturi (Market Basket)
- b) Sa determinam predictia de cumparare pentru produsul "Balsam" cu ID-ul 47 prezent in snapshot-ul nostru naiv
- c) Sa determinam predictia de cumparare pentru o intreaga lista de produse noi pe care dorim sa le lansam
- d) Sa determinam segmentul de comportament de cumparare al clientului realizand clusterizarea in functie de parametrii vectorului de comportament
- e) Modelul este self-explain-able in sensul in care pentru un anumit client datele din vectorul de comportament sunt extrem de usor de interpretat
- f) Probabil una din cele mai importante facilitati ale modelului consta in faptul ca orice tranzactie noua realizata de client modifica vectorul de comportament  $\Theta$  ducant la noi si imbunatatite inferente/predictii

Pentru market basket analiza datelor va genera in prima faza o tabela/matrice cu coeficientii T de tranzactii calculabili dupa care prin aplicarea modelului de regresie la nivel de client individual se vor putea determina coeficientii de tranzactii potentiale (predictiile) pentru produsele cumparate

Concret pentru exemplul nostru naiv in faza 1 vom avea:

<b>CLIENT 32245529</b>	
ID	T
Produs	
...	...
17321	0.5
34	0.9
32543	0.1
47	0
1801	0
....	...

Iar in urma aplicarii modelului de regresie vom obtine cel mai probabil:

<b>CLIENT 32245529</b>	
ID	T
Produs	
...	...
17321	0.5
34	0.9
32543	0.1
47	0.8
1801	0.05
....	...

Astfel Market Basket-ul final va fi MB = (MB Generat) + (MB Predictie) = (17321 : 0.5; 34 : 0.9; 47 : 0.8) plecand de la premiza ca vom selecta doar produsele cu T peste medie ( $T \geq 0.5$ )

Continuand exemplul de mai sus ideea predictiei scorului T pentru un produs necumparat sau nou functioneaza in felul urmatoar: fara a intra momentan in detaliile algoritmului (varianta simpla – nu varianta DNN) de regresie este evident ca vectorul de comportament de cumparare al clientului **32245529** arata de forma aproximativa:

$$\begin{aligned}\theta^{(Client_{32245529})} &= \left( \theta_{EsteAnagezic}^{(Client_{32245529})} = 0.001, \theta_{EsteSampon}^{(Client_{32245529})} = 0.3, \theta_{EsteOccitane}^{(Client_{32245529})} \right. \\ &= \left. 0.41, \dots, \theta_N^{(Client_i)} \right)\end{aligned}$$

Iar produsul **47**:

$$X^{(Produs_{47})} = \left( x_{EsteAnagezic}^{(Produs_{47})} = 0, x_{EsteSampon}^{(Produs_{47})} = 1, x_{EsteOccitane}^{(Produs_{47})} = 1, \dots, x_N^{(Produs_{47})} \right)$$

Deci in urma produsului scalar inmultirile dintre ponderile  $\theta_{EsteSampon}^{(Client_{32245529})} = 0.8, \theta_{EsteOccitane}^{(Client_{32245529})} = 0.91$  cu attributele produsului **47** vor genera un scor T destul de mare (0.8 in exemplul nostru naiv)

Analizand vectorul de client de mai sus este evident ca:

Valoarea  $\theta_{EsteAnagezic}^{(Client_{32245529})} = 0.001$  determina faptul ca acest client nu obisnuieste sa cumpere **analgizice** (plecand de la premiza ca valorile ponderii sunt pe intervalul 0-1)

Valoarea  $\theta_{EsteSampon}^{(Client_{32245529})} = 0.3$  determina faptul ca acest client destul de des cumpara des **sampoane** (plecand de la premiza ca valorile ponderii sunt pe intervalul 0-1)

Valoarea  $\theta_{EsteOccitane}^{(Client_{32245529})} = 0.41$  determina faptul ca acest client cumpara des marca **occitane** (plecand de la premiza ca valorile ponderii sunt pe intervalul 0-1)

*Astfel chiar si un non-statistician sau non-informatician poate intelege intuitiv comportamentul clientului*

## Atribute generate de sistem

In urma aplicarii modelului Matricii Recomandarilor sistemul va genera automatizat urmatoarele atribute la nivel de CLIENTID individual:

### 1. CLIENT\_<**BRAND**>\_PREFERENCE:

1.1. Serie de atribute care determina importanta unui anumit **brand** pentru respectivul PARTENER.

Atributele vor contine valori pe un interval bine determinat (de exemplu 0.00-0.99) care sa determine nivelul de preferinta al CLIENT-ului pentru

1.2. Minim 5 atribute de acest tip

1.3. Exemple:

1.3.1. "CLIENT\_APIVITA\_PREFERENCE" : 0.1 (interes redus pentru APIVITA)

1.3.2. "CLIENT\_OCCITANE\_PREFERENCE" : 0.33 (interes sub-mediocru pentru OCCITANE)

Name	Description	Notes
BP01_APIVITA	Client brand preference for APIVITA	Real value score [0.00-1.00]
BP02_OCCITANE	Client brand preference for L'OCCITANE	Real value score [0.00-1.00]
BP03_VICHY	Client brand preference for VICHY	Real value score [0.00-1.00]
BP04_DR_HART	Client brand preference for DR HART	Real value score [0.00-1.00]
BP05_NUROFEN	Client brand preference for NUROFEN	Real value score [0.00-1.00]
BP06_AVENE	Client brand preference for AVENE	Real value score [0.00-1.00]

1.4. Modele/functionality aplicabile:



1.4.1.Segmentare comportamentala

1.4.2.Determinare market-basket

1.4.3.Determinarea preferintelor de brand si asocierile aferente

1.4.4.Analiza churn si construire de model deep pentru churn prediction

1.4.5.Recomandare de cumparare (asocieri)

1.4.6.Campanii target

## 2. PARTNER\_<CATEG>\_PREFERENCE:

2.1. desemenea o serie de attribute care determina importanta unei anumite **categorii** de produse pentru respectivul partener. Valorile sunt similare cu cele ale preferintei partenerilor pentru brand-uri dupa cum urmeaza:

2.1.1.“CLIENT\_FRUMUSETE\_PREFERENCE” : 0.53 (interes peste medie pentru produsele de frumuseti)

2.1.2.“CLIENT\_ALIMENTATIE\_PREFERENCE” : 0.17 (interes foarte mic pentru produse de alimentatie)

Name	Description	Notes
CP01_BABY	Client preference for products tagged as BABY products (category)	Real value score [0.00-1.00]
CP02_COSMETICE	Client preference for products tagged as COSMETICE products (category)	Real value score [0.00-1.00]
CP03_DERMOCOSM	Client preference for products tagged as DERMOCOSMETICE products (category)	Real value score [0.00-1.00]
CP04_PHARMA	Client preference for products tagged as PHARMA products (category)	Real value score [0.00-1.00]
CP05_SPECIALE	Client preference for products tagged as SPECIALE products (category)	Real value score [0.00-1.00]

2.2. Minim 5 attribute de acest tip

2.3. Modele aplicabile:

- 2.3.1.Segmentare comportamentala
- 2.3.2.Determinare market-basket
- 2.3.3.Analiza churn si modele deep de churn prediction
- 2.3.4.Recomandare de cumparare (asocieri)
- 2.3.5.Campanii target

### 3. CLIENT\_<CATEG>\_PRICING:

3.1. Este un atribut care determina puterea de cumparare a CLIENT pentru o anumita categorie de produse, branduri, etc. Concret valoarea atributului determina o putere de cumparare:

3.1.1.“CLIENT\_FRUMUSETE\_PRICING” : 0.9 (cumpara produse scumpe de infrumusetare)

3.1.2.“CLIENT\_BACY\_PRICING” : 0.1 (cumpara produse ieftine pentru belelusi)

Name	Description	Notes
PR01_BABY	Price range score for BABY product category	Real value score [0.00-1.00]
PR02_COSMETICE	Price range score for COSMETICE product category	Real value score [0.00-1.00]
PR03_DERMOCOSM	Price range score for DERMOCOSMETICE product category	Real value score [0.00-1.00]
PR04_PHARMA	Price range score for PHARMA product category	Real value score [0.00-1.00]
PR05_SPECIALE	Price range score for SPECIALE product category	Real value score [0.00-1.00]
PR06_OTC	Price range score for OTC product category	Real value score [0.00-1.00]

3.2. Minim 10 attribute de acest tip la nivel de partener

3.3. Utilizabil la:

- 3.3.1.Segmentare comportamentala (behaviour based segmentation)
- 3.3.2.Modele deep de churn prediction

### 3.3.3.Recomandare de produse

## Metoda avansata de generare a atributelor de produs

In urma analizei bazelor de date s-a constat ca produsul sau produsele asociate unei anumite tranzactii pot apartine unor intrari diferite din nomenclatoarele de produse desi se refera la acelasi (sau aproape acelasi) produs. Mai mult, grupuri de produse care semantic sunt similare pot fi cu usurinta disociate datorita inconsistentei datelor de nomenclator.

Din fericire structura nomenclatorului de produse (care nu este de forma arborescenta ci sub forma unui graf ciclic) contine informatii de tip text limbaj natural de descriere a produselor respective. Testele initiale au fost facute prin analiza datelor pana la nivelul 4 de descriere (nume1,nume2, nume3, nume4) si s-au generat (prin hard-codare) coloane specifice fiecarui cuvant cheie (brand, categorie, etc). Ulterior s-a constat ca acesta metoda este atat ineficienta cat si inflexibila la modificarea structurii datelor putand fi aplicata pentru teste, segmentari fixe si inferente predefinite dar fara o reala aplicabilitate intr-un sistem de productie.

Reluand analiza am descoperit o metoda eficienta si eficace de analiza a produselor si de pregatire a structurii vectorului comportamental al clientului. Astfel, am ajuns la concluzia ca se poate aplica un model de preprocesare (analiza si hashing semantic) al produselor bazat pe descriere produselor regasita in campurile nume1,nume2, nume3, etc. Algoritmul multi-stage va construi o reprezentarea de tip “bag-of-words” a tuturor produselor posibile si ulterior va genera o structura cu dimensionalitate redusa prin aplicarea unui algoritm de machine learning de tip autoencoder adanc. Decizia de a folosi un autoencoder fata de un model bazat pe analiza componentelor principale (PCA) a fost luata pentru a putea “capta” non-liniaritati in structura produselor, tranzactiilor si implicit a comportamentului clientului. Ulterior reprezentarea generata de model se va aplica asupra produselor asociate unei tranzactii iar pe baza acestei reprezentari se va putea genera lista de attribute reale ale produsului respectiv.

Finalmente, lista de attribute va fi pastrata pentru fiecare produs si implicit va avea asociat vectorul de comportament al clientului format din ponderile fiecarui atribut de produs asa cum sunt descrise in sectiunea “Attribute generate de sistem” a documentului de prezentare a modelului de analiza a comportamentului clientului (MATRECOM).

De mentionat in final ca aceste atribute sunt realmente singure metode prin care se pot realiza inferente reale asupra segmentarii clientilor in baza comportamentului complet al acestora (plecand de la informatii elementare ca valori, tranzactii pana la informatii ca afinitate fata de branduri sau predispozitie fata de anumite nivele de preturi sau fata de anumite categorii). In afara de segmentare se pot analiza evolutiile/comportamentele periodice, predictia churn, predictii/sugestii de cumparare, recomandari, s.a.m.d.

## Considerente de implementare

- a) **Modelul prezentat pentru un client si construirea modelului regresiv pentru toate produsele se poate scala pentru aplicarea la scara mare pentru toti cei >2.000.000 de clienti si cele >30.000 produse pentru fiecare client in parte. Modelul a fost prezentat pentru un client din considerente de simplificare a analizei**
- b) **Modelul va fi implementat pe o structura de calcul masiv paralel capabila sa calculeze cele peste 2.000.000 de modele individuale (fiecare client/card) si sa le modifice o data cu aparitia de noi date. Ca volum de date: 2.000.000 modele a cate N (min 7-8 – in curs de definire) parametrii calculati in baza unui set de produse de peste 30.000 (rog corectie). Training-ul de “consolidare” va rula pe intreaga baza de date pe un masiv (matrice multi dimensionala) total de date teoretic de peste 200 GB. Update-urile se vor putea face cu resurse mult mai limitate si implicit se vor putea genera raspunsuri in timp real**
- c) **Dinamica modificarii celor peste 2.000.000 trebuie sa fie una ridicata pentru a putea introduce noi potentiale facilitati cum ar fi: propunerea unui cos de cumparaturi “instant” imediat dupa ce clientul a realizat o tranzactie si implicit modelul sau comportamental a fost proaspat updatat**
- d) **Din punct de vedere algoritmic sistemul se bazeaza pe optimizarea stohastica online prin gradienti. Astfel plecand de la premiza ca scopul principal este corectarea**

continua a erorilor de predictie pe care modelele le fac vom avea o functie de eroare de forma:

$$E^{(C_i)}(\hat{Y}, Y | h(\theta)) = \sum (\hat{y}_k - y_k)^2 + \sum \theta^2$$

Unde  $\hat{Y} = h(X | \theta^{(C_i)})$  este vectorul de predictii realizate de modelul pentru clientul  $C_i$  pentru toate produsele deja cumparate iar  $Y$  este adevarul constatat prin calcularea scorurilor coeficientilor de tranzactionare pentru toate produsele  $k$  (cu  $k$  de la 1 la nr de produse cumparate) iar  $\sum \theta^2$  este componenta de regularizare de tip regularizare elastic net (L2) care reduce riscul de over-fitting al datelor. Astfel problema noastra se poate reduce la rezolvarea urmatoarei ecuatii:

$$\Theta^* = \underset{\theta}{\operatorname{argmin}} E^{(C_i)}(\hat{Y}, Y | h(\theta))$$

Astfel la fiecare tranzactie noua pe care clientul o realizeaza se va calcula:

$$\theta^{(C_i)} = \theta^{(C_i)} + \alpha \nabla E^{(C_i)}(\theta) = \theta^{(C_i)} + \alpha \frac{\partial E}{\partial \theta}$$

obtinandu-se implicit un nou vector de comportament al clientului

e) Intregul system predictiv cu toate cele 2.000.000 de sub-modele poate fi programat si implementat pe solutia Microsoft Azure HDInsight sau poate rula in mediul propriu HTSS utilizand o masina cu resurse decente de memorie si GPU.

f) Sistemul va putea fi imbunatatit intr-o faza ulterioara cu adaugarea de noi ponderi in vectorul de comportament al clientului bazate pe corelatiile date de comportamentul similar de cumparare al altor clienti (practic sa adauga model de recomandari colaborative la modelul existent)

g) Un model alternativ va fi construit in fazele avansate ale proiectului bazat pe modele generative de tip RBM (Restricted Boltzmann Machine). Acest model generativ va analiza comportamentul clientilor fara a pleca de la premiza necesitatii unui "label-ing" al comportamentului (tranzactiilor) clientului.