Model de analiza si inferenta a Matricei Recomandarilor

History			
NR	Data	Autor	Ver
1	19.01.2017	A.I. DAMIAN	Draft 1

Intregul model inferential si predictiv poate fi explicat intuitiv prin urmatorul exemplu concret. In vederea construirii exemplului vom pleca de la mai multe ipoteze-pasi dupa cum urmeaza:

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 atribute 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 Produs	Nume	ID Producator	UM	PU	Marja	Categorie Generala	Clasificare	Este Analgezic	Este Sampon	Este Cosmetic	Este OTC	Este Occitane	Este GNC	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17321	Nurofen	100	Pastila	2.15	0.3	Analgezice	Farmaceutice	1	0	0	0	0	0	...

34	Sampon AAA	203	Cutie	59.99	0.4	Sampoane	Cosmetice	0	1	1	0	1	0	...
32543	Vitamina C	455	Blister	20.00	0.25	Supliment e 1	OTC	0	0	0	1	0	1	...
47	Balsam B	203	Cutie	49.99	0.4	Sampoane	Cosmetice	0	1	1	0	1	0	...
1801	Sare	122	Punga	3.44	0.15	Supliment e 2	OTC	0	0	0	1	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Pe langa atributele

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:

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

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 **<u>vectorul de comportament</u>** 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 atribute de produs (ce cele 6 din exemplul naiv de mai sus) pentru fiecare client vom putea construi un model individual de forma

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

Unde

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

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 $numarul\ total\ de\ atribute\ (coloane)\ construite.$

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 confom urmatoarei formule generalizate:

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

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 Θ 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 aveam:

CLIENT 32245529	
ID Produs	T
...	...
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:

CLIENT 32245529	
ID Produs	T

...	...
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>=0.5)

Continuand exemplul de mai sus ideea predictiei scorului T pentru un produs necumparat sau nou functioneaza in felul urmator: 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:

$$\theta^{(Client_{32245529})} = \left(\theta^{(Client_{32245529})}_{EsteAnagezic} = 0.001, \theta^{(Client_{32245529})}_{EsteSampon} = 0.3, \theta^{(Client_{32245529})}_{EsteOccitane} = 0.41, \dots, \theta^{(Client_i)}_{N}\right)$$

Iar produsul ***47***:

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

Deci in urma produsului scalar inmultirile dintre ponderile $\theta^{(Client_{32245529})}_{EsteSampon} = 0.8, \theta^{(Client_{32245529})}_{EsteOccitane} = 0.91$ cu atributele 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 $\theta^{(Client_{32245529})}_{EsteAnagezic} = 0.001$ determina faptul ca acest client nu obisnuieste sa cumpere ***analgezice*** (plecand de la premiza ca valoarile ponderii sunt pe intervalul 0-1)

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

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

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

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:**

$$\boldsymbol{E^{(C_i)}}\left(\widehat{\boldsymbol{Y}}, \boldsymbol{Y} \,\middle|\, \boldsymbol{h(\theta)}\right) = \sum(\widehat{\boldsymbol{y}}_{\boldsymbol{k}} - \boldsymbol{y}_{\boldsymbol{k}})^2 + \sum \boldsymbol{\theta^2}$$

Unde $\widehat{\boldsymbol{Y}} = h\left(X \mid \theta^{(C_i)}\right)$ **este vectorul de predictii realizate de modelul pentru clientul** $\boldsymbol{C_i}$ **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** $\sum \boldsymbol{\theta^2}$ **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:**

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, \boldsymbol{E^{(C_i)}}\left(\widehat{\boldsymbol{Y}}, \boldsymbol{Y} \,\middle|\, \boldsymbol{h(\theta)}\right)$$

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

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

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)**