Problema Propusa:

* Vrem sa cream o formula de ranking generica (nu este personalizata in functie de utilizator) care sa determine ordinea optima a produselor in listarile de categorie, astfel incat profitul vanzarilor (volum\*margine) sa fie maxim. Evident, optimizarea se doreste a fi pe o perioada lunga.
* Intuim insa ca, pe langa vanzari, exista si alte caracteristici ale unui produs care ar putea influenta relevanta in listare.
* Intuim si ca aceste caracteristici si perioada pe care ele sunt evaluate ar putea sa difere de la o categorie la alta.

Abordarea problemei:

Partea I : Alegerea modului de calcul

1. Impartirea problemei in subprobleme:
   1. Determinarea perioadei relevante pentru calculul formulei de ranking la nivel de categorie (nivel correspondent unui listing de categorie)
   2. Determinarea modului de calcul cel mai potrivit pentru ranking la nivel de produs astfel incat sa optimizeze profitul
2. Determinarea perioadei relevante pentru calculul formulei de ranking la nivel de produs:
   1. Stabilirea unor factori pentru determinarea perioadei
      1. Variabilitatea vanzarilor in decursul acelei perioade masurata prin coeficientul de variatie (deviatia standard a vanzarilor pe perioada respectiva / media vanzarilor pe perioada respective)
      2. Frecventa de cumparare a produsului respectiv = 365 zile (1 an) / numarul de bucati vandute din acel produs intr-un an => 1 unitate din produsul X se vinde la Y zile
   2. Extragerea unui esantion relevant pentru acest calcul – sufficient de mare si divers pentru a fi reprezentativ pentru populatia de produse eMAG, sufficient de mic pentru a putea evalua rapid diverse formule pe el (timp de rulare <=5 min in R/Python pe un laptop obisnuit = procesor i5, 8G RAM)
      1. Sunt ~1400 de categorii
      2. La nivelul fiecarei categorii extrag random 10% din produse
      3. Ultimele 60 de zile maxim per produs
         1. Suficient de mare cat sa existe variatie in vanzari per produs si sa apara variatie de pret datorita promotiilor / campaniilor
         2. Suficient de mica incat sa evite schimbari mari din cauza schimburilor de anotimpuri
      4. Tabelul ar arata asa: IdCategorie,IdProdus, Data, Frecventa de cumparare a produsului in ultimul an,Valoarea Vanzarilor in acea Zi
      5. Estimez ca ar ajunge la o dimensiune de 1.15-1.5 milioane randuri = nivel optim pentru resursele unui laptop + capacitate de generalizare pentru un algoritm / ecuatie statistica
   3. Utilizarea factorilor intr-o anumita logica pentru a calcula perioada relevanta la nivel de produs
      1. Perioada sa fie cat de cat stabila in ceea ce priveste variatia vanzarilor:
         1. Ipoteza este ca in decurs de 2 luni apar diverse evenimente care influenteaza fluctuatia vanzarilor. Testez aceasta ipoteza prin a ma uita la cum se schimba coeficientul de variatie pe vanzari la 60,45,30 de zile. Daca diferenta este de +/-3 procente per acea categorie aplic logica de mai jos. Daca nu folosesc 45 de zile ca perioada de analiza pentru acea categorie.
         2. La nivel de produs elimin pe rand cea mai veche zi pana cand coeficientul de variatie <=35% sau raman mai putin de 30 de zile. Un numar
         3. Dupa acest pas, daca perioada la care am ajuns este mai mica decat frecventa de cumparare a produsului, inlocuiesc acea perioada cu frecventa lui de cumparare. Logica este ca daca de exemplu din punct de vedere variabilitate o perioada de 15 zile este ok, dar produsul este cumparat 1 data la 20 de zile este ilogic sa fac vreo previziune pe valoarea vanzarilor pentru ca nu voi avea vanzari.
         4. La nivel de categorie voi alege cuantila de 5% ( 95% din valori sunt mai mari decat acest nivel) ca perioada de calcul pentru categoria respectiva. In acest fel exista o valoare uniforma pentru toate produsele din acea categorie.
   4. Ca output la finalul acestei etape am pentru fiecare categorie perioada pe care sa o iau in considerare mai departe pentru a maximiza profitul
3. Determinarea modului de calcul cel mai potrivit pentru ranking la nivel de produs astfel incat sa optimizeze profitul
   1. Logica este ca pe baza unor anumiti factori sa fac un Forecast de care ar fi Profitul generat de un anumit produs pe perioada calculata anterior. Pentru asta ar putea fi 2 abordari: bazata pe ecuatii matematice (closed form solution): regresie liniara / arbore de regresie sau bazata pe algoritmi (machine learning): ensemble model (random forest / xgboost), K Nearest Neighbour etc. Avantajul primei variante este simplitatea si capacitatea de a fi explicat. Avantajul celei de-a 2-a variante este o putere de predictie mai mare, acuratete mai buna in general. Voi imbina cele 2 – in prima faza o ecuatie simpla pentru a vedea utilitatea factorilor pe care i-am ales in determinarea profitului, iar pentru productie voi alege un algoritm mai complex si mai puternic pe baza factorilor cei mai relevanti.
   2. Colectarea unui model de date posibil relevant – care sunt factorii care ar putea avea o influenta asupra ranking-ului?
      1. Mai jos este perspectiva mea. Mai departe o voi folosi pe aceasta in gandire. In mod normal as discuta cu cineva care se ocupa de partea de vanzari pentru a afla din experienta lor care ar putea fi lucrurile ce influenteaza profitul.
      2. Profit = Margine\*Volum
         1. Marginea este influentata de pret, discount, competitivitatea pretului
         2. Volumul este influentat de satisfactia clientului, calitatea produsului (produsul este returnat / ajunge in service de putine ori), interactiunea cu acel produs (click-uri etc.), vizibilitatea produsului (pozitia in listing), opinia altori client despre acel produs.
         3. Tabelul de la care as porni ar fi cam asa:

Tabelul 1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IdCategorie | IdProdus | Data | Pret | Discount% | Indicele de competitivitate al pretului | NPS pe produs (daca n-a existat nicio comanda in acea zi, ramane camp gol) | Numarul de retururi in acea zi | Numarul de retururi in service in acea zi | Click-uri din listing pe produs in acea zi | Afisari ale produsului in listing in acea zi | Pozitia in listing | Numarul de review-uri | Nota medie la review-uri pe care o are in acea zi | Profit (Margine \* Volum) |

* + 1. Mai departe as grupa la nivel de Categorie, Produs, si perioada de analiza pe care am stabilit-o anterior. In loc de data va veni numarul de zile pentru care fac forecast-ul.
    2. Urmatorul pas este analiza exploratorie a datelor – grafice de corelatie intre variabila de raspuns (profit) si fiecare din variabilele predictoare (toate mai putin, IdCategorie, IdProdus, Nr. Zile Forecast), si intre variabilele predictoare intre ele. + matricea de corelatie. Scopul este sa evaluez ochiometric care variabile influenteaza profitul si care nu si sa le elimin pe cele care se coreleaza mult intre ele (evitarea multicolinearilitatii care duce la erori / greseli in modelare predictiva). In acest sens daca 2 variabile au indice de corelatie >=0.75 o voi pastra pe cea care se coreleaza mai puternic cu profitul.
    3. Urmatorul pas este reducerea numarului de factori prin metode analitice – Principal Component Analysis (determina care din variabile si combinatii lineare ale lor explica cel mai mult din variabilitatea in profit) si Random Forest Feature Importance (rezultat similar cu PCA, metoda este diferita). Dupa acest pas elimin variabilele cele mai putin importante, avand grija sa pastrez cele care ofera cea mai mare simplitate cu cea mai mare putere de predictie modelului.
    4. In scopul de a construe un model cat mai generalizabil voi imparti datele asa: 90% pentru modelare & testare, 10% pentru validare. Datele de validare le voi extrage intr-un set separat si vor fi folosit pentru evaluarea finala a modelului predictiv. Din restul de 90% de date, fac impartirea random: 80% din date sunt pentru training, 20% din date sunt pentru testare. Aplic modelul predictiv asa cum este descris mai jos si testez dupa indicatorii de mai jos. Impartirea random, modelarea, testarea, calculul indicatorilor de evaluare o fac de 10 ori.
    5. Urmeaza modelarea efectiva - in prima faza 2 modele simple regresie liniara si arbore de regresie, apoi 2 modele mai complexe – random forest si xgboost. Voi evalua puterea predictiva a fiecarui model folosind 2 indicatori: Root Mean Squared Error (Se face suma patratelor erorilor, se imparte la numarul de observatii, se face radical de ordin 2) si Explained Variance Score Function (proportia de variatie in variabila raspuns explicata de model).
    6. Alegerea finala a modelului predictiv o voi face dupa urmatoarele criterii, dupa ce aplic modelul pe restul de 10% date de validare ramase.
       1. Simplitatea modelului – o regresie liniara este mai usor de explicat decat 500 de arbori de regresie combinati pe baza zonelor in care arborii precedenti au prezis in mod gresit (xgboost)
       2. RMSE cat mai mic
       3. Explained Variation cat mai mare
    7. Mai departe ca output de folosit pentru Web-Developeri ar fi un tabel de forma de mai jos, din care ar ordona produsele in listing dupa profit. Logica este ca un profit previzionat mai mare va insemna si in realitate un profit mai mare daca acel produs este pus mai in fata in listing.

Tabelul 2

|  |  |  |  |
| --- | --- | --- | --- |
| IdCategorie | IdProdus | Perioada Relevanta | Profit Forecastata |

Aceasta parte ar dura in jur de 60-80 ore de munca din partea de Data Scientist

Partea a II-a: Operationalizarea/automatizarea calculului pentru a fi folosit de catre web-developeri

1. Pentru operationalizare ar trebui o infrastructura de masini care sa:
   1. Stocheze toate datele de input pentru model, grupate dupa IdCategorie, IdProdus, Perioada Relevanta (Tabelul 1) – probabil ceva cu SQL Server / Hadoop
   2. Sa ruleze modelul predictiv – probabil ceva cu R / Python
   3. Sa stocheze tabela de ouput (Tabelul 2) pentru a fi folosita mai departe de catre Web Developeri
2. Frecventa de rulare a script-ului de calcul ar fi saptamanala

Partea a III-a: Evaluarea rezultatelor:

1. Pentru evaluarea utilitatii ar fi nevoie de un A/B test – 50% din useri interactioneaza cu listing-ul current, 50% din useri interactioneaza cu listing-ul cu model predictiv pe profit. Metricile de mai jos ar trebui calculate pe fiecare din variante si sa determine
   1. CTR-ul din listing – pagini de produs vizualizate din listing / afisari de listing
   2. Rata de conversie din listing – comenzi din listing / vizite ce contin listing-ul
   3. Profitul generat
2. Ar fi de determinat cu Product Owner-ul / userii de business care este metrica cea mai important iar pe aceasta sa fac o testare de ipoteze statistice pentru a verifica daca diferentele observate sunt semnificative statistic (se vor sustine pe termen lung, mai mari decat variatia naturala a indicatorului respectiv)