

Documentation Technique

Général
MAI 2024

Pour produire SocialValues, CannabisInsights Powered by Vividata, CommunityHealth, MoneyMatters Powered by Canadian Financial Monitor et d'autres bases de données comportementales, nous utilisons plusieurs types de cadres de modélisation : coefficients PRIZM, K plus proches voisins (ou K-NN, de l'anglais K-Nearest Neighbours) ou fusion.

Cadre de modélisation par produit

Product	Coefficient	K-NN	Fusion+KNN	Autres
SocialValues	X			
CannabisInsights	X			
MoneyMatters	X			
ShopperChoice	X			
VaccineInsights	X			
Opticks Mobile		X		
Opticks Social		X		
Opticks eShopper		X		
Opticks Numeris		X		
Opticks Media		X		
Opticks Vividata		X		
Opticks Vividata (Genre)		X		
CommunityHealth			X	
GreenLiving			X	
GivingBack			X	
CommunityLife			X	
AutoView				X

Pour nos ensembles de données basés sur les coefficients, nous assignons à chaque enregistrement de sondages un segment PRIZM®. Selon le sondage, nous pouvons ajouter PRIZM® au niveau de l'aire de diffusion (AD) ou du code postal. Nous resumons le sondage codé par PRIZM aux 67 segments. Au niveau PRIZM, nous calculons des coefficients empiriques, qui peuvent être soit un taux, soit une moyenne. Les taux sont le nombre total de personnes qui pratiquent l'activité divisé par l'univers total de l'activité. Les moyennes sont le nombre total d'unités consommées/dépensées/détenues divisé par le nombre total de personnes qui consomment/dépensent/détiennent. Les coefficients empiriques sont appliqués à la géographie en utilisant les affectations RTAUDL PRIZM et en multipliant le coefficient de chaque variable par nos univers DemoStats de l'année en cours. Les univers communs pour nos sondages sont : les ménages, la population des ménages de 12 ans et plus et la population des ménages de 15 ans et plus.

La méthodologie du K plus proches voisins (K-NN) permet d'apparier la collection d'enregistrements provenant des sondages à notre fichier de codes postaux, également appelé données cibles. À partir de la collection d'enregistrements des sondages, nous calculons des coefficients moyens pondérés. Cela signifie que chaque code postal géographique possède un ensemble unique de coefficients. Les coefficients sont comparés à nos univers DemoStats de l'année en cours et nous utilisons 50 variables pour "appairer" les enregistrements du sondage à notre fichier de codes postaux. Ces 50 variables proviennent de notre base de données DemoStats et d'Equifax Neighbourhood View™, et sont ajoutées aux données des sondages au niveau des codes postaux. L'utilisation de ces 50 variables nous permet de calculer les différences multidimensionnelles entre chaque sondage et le code postal cible. Les codes postaux du sondage présentant la plus petite différence par rapport au code postal cible sont utilisés pour générer les coefficients pour le code postal cible. Il convient de noter que nous limitons les

correspondances entre les codes postaux des sondages et les codes postaux cibles par région géographique (Atlantique, Québec, Ontario, Prairies, C.- B./Territoires).

La fusion est une technique similaire au K-NN. Nous utilisons cette technique comme étape de prétraitement pour K-NN lorsque nous travaillons avec des fichiers de microdonnées à grande diffusion (FMGD). Les données FMGD sont accompagnées de peu, voire d'aucun, indicateur géographique. La technique de fusion est utilisée pour appairer les enregistrements des FMGD à un autre ensemble de données, connu sous le nom d'ensemble de données source. L'ensemble de données source comportera des codes postaux comme identifiant géographique. Nous générons ces correspondances sur la base de variables disponibles dans les données FMGD et dans notre ensemble de données source. Cette approche nous permet d'ajouter les codes postaux aux enregistrements FMGD. Remarque : le code postal ajouté est l'un des nombreux codes postaux possibles. Une fois que les codes postaux ajoutés aux données FMGD, nous pouvons appliquer la méthodologie K-NN décrite ci-dessus.

Les ensembles de données AutoView sont des comptages agrégés d'immatriculations de véhicules. Experian fournit des comptages au niveau AD par classe de véhicule et année de modèle. Nous répercutons ces données à tous les autres niveaux géographiques.

Résumé du cadre de modélisation

Coefficient

- Modélisé au niveau PRIZM
- Coefficients égaux à la propension empirique par segment
- Les coefficients PRIZM sont déployés au niveau géographique
- Toutes les zones géographiques ayant la même affectation de segment PRIZM ont la même propension à être affectées par un segment PRIZM

K-NN et/ou fusion (Modélisé)

- Les données sont modélisées directement au niveau du code postal
- Les modèles sont limités par la région
- En théorie, chaque code postal obtient un ensemble unique de propensions

Autres

- Chiffres agrégés par niveau géographique