TP
MIAGE M2

Marketing + CRM

Fichier : **TP_03_MIAGE_M2_CRM_Nov2016**.docx

Date **: novembre 2016**

Rédacteur : DCN

Etudiant(s) auteur(s) de la réponse

-
-
-
-

Diffusion

- Etudiants M2 MIAGE

SOMMAIRE

Exercice n°1 : L'âge probable des clients 2
Exercice n°2 : Match-codage et dédoublonnage 4
Exercice n°3 : Signature et dédoublonnage de fichiers 6

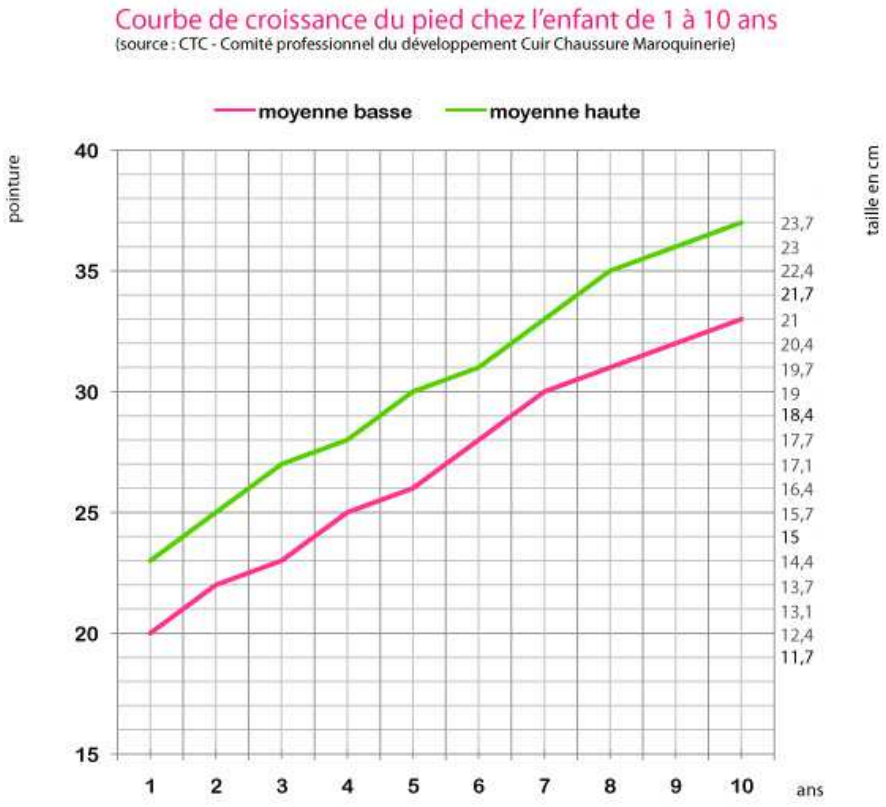
Didier Certain Professeur associé
☎02 99 84 74 57 ☎06 08 63 63 28
e-mail : didier.certain@univ-rennes1.fr

Exercice n°1 : L'âge probable des clients

L'âge est un critère de segmentation fréquemment utilisé car, dans de nombreuses situations, c'est un marqueur significatif de de comportement. Pas toujours, mais souvent.

Lorsqu'on connait la date de naissance des prospects ou des clients, on peut utiliser cette variable pour calculer l'âge des individus. Puis on segmente ensuite sur la variable « **Âge** ».

Lorsqu'on ne connait pas la date de naissance des prospects ou des clients, il faut essayer de trouver une méthode pour corréler les informations disponibles sur eux et leur âge. Par exemple, on peut déduire l'âge probable d'un enfant en fonction de la taille de son pied comme le montre le diagramme suivant :



Corrélation taille du pied / âge d'un enfant

Quand on ne connaît pas la date de naissance d'un client B2C, quand on ne connaît pas son âge et quand on ne connaît pas la taille de ses chaussures, il faut essayer d'estimer son âge en fonction d'autres caractéristiques personnelles.

L'expérience montre qu'il existe des corrélations statistiques entre l'âge d'un individu et certaines de ses données personnelles. Il est donc tentant de s'appuyer sur ces corrélations statistiques pour en déduire, de manière probabiliste, l'âge d'individus présents dans un fichier B2C.

Par exemple, il existe une corrélation possible (avec une marge d'erreur importante) entre le prénom usuel du client B2C et son âge. Nous allons tenter, dans cet exercice, de nous appuyer sur cette corrélation.

Pour vous en convaincre, vous lirez l'article sur les prénoms paru la revue Octant publiée par l'INSEE **(INSEE_octant99_art10_2004.pdf)** et vous en tirerez les conclusions pour déterminer l'âge probable des clients B2C présents dans le fichier Excel (**TP001_fichier_clients_sept-2016.xls** / **onglet B2C**) :

- Prendre le fichier EXCEL ("TP001_fichier_clients_sept-2016.xls")
- Cliquer sur l'onglet "B2C" ➔ Fichier client B2C

- ❑ Pour chaque client, déterminer son âge minimum (colonne "A1")
- ❑ Pour chaque client, déterminer son âge probable (colonne "A2")
- ❑ Pour chaque client, déterminer son âge maximum (colonne "A3")

L'idée de ce cadrage est de faire en sorte que :

- o Probabilité [âge réel est supérieur à A3] < 10%.
- o Probabilité | âge réel est inférieur à A1] <10%.
- o Probabilité [âge réel est inférieur à A2] = 50%.
- o Probabilité [âge réel est supérieur à A2] = 50%.

Exercice n°2 : Match-codage et dédoublonnage

- Regarder le fichier image « **mailing_avec_doublon.jpg** ».
- Commenter.

- Prendre le fichier EXCEL (**TP002_fichier clients V2016_a-dedoublonner.xls**)
- On cherche à "dédoublonner" les adresses de ce petit fichier client B2C afin d'éviter d'envoyer le même courrier deux fois au même prospect.

- Un contrôle visuel est possible sur un petit fichier client.
- Mais quand la taille du fichier augmente, il faut changer de tactique

❑ Trier le fichier sur les noms de client (par ordre ascendant).
❑ Exercer un contrôle visuel pour mettre en évidence les lignes consécutives identiques.

❑ Trier le fichier sur les noms de ville (par ordre ascendant).
❑ Exercer un contrôle visuel pour mettre en évidence les lignes consécutives identiques.

- La question devient plus compliquée quand des erreurs (ou des variations) de saisie génèrent des doublons indétectables en contrôle visuel. Pour essayer de détecter des doublons, on va utiliser la technique de la signature (également appelée « empreinte »).

❑ Construire par concaténation, pour chaque client, un champ "**match-code1**" de la manière suivante :
  - o 2 premiers caractères du nom
  - o 3 premiers caractères de l'adresse
  - o 2 premiers caractères du code postal
  - o 2 premiers caractères de la ville

❑ Trier le fichier sur les champs "**match-code1**" (par ordre ascendant)
❑ Exercer un contrôle visuel pour mettre en évidence les lignes possédant un "**match-code1**" identique.
❑ Créer un champ mettant en évidence automatiquement les **match-codes consécutifs et identiques**.

❑ Construire par concaténation, pour chaque client, un champ "**match-code2**" de la manière suivante :
  - o 3 premiers caractères du nom
  - o 3 derniers caractères de l'adresse
  - o 2 premiers caractères du code postal
  - o 3 premiers caractères de la ville

❑ Trier le fichier sur les champs "**match-code2**" (par ordre ascendant)
❑ Exercer un contrôle visuel pour mettre en évidence les lignes possédant un "**match-code2**" identique.
❑ Créer un champ mettant en évidence automatiquement les **match-codes consécutifs et identiques**.

- Construire par concaténation, pour chaque client, un champ "**match-code3**" de la manière suivante :
    - 4 premiers caractères du nom
    - 3 premiers caractères de l'adresse
    - 3 derniers caractères de l'adresse
    - 2 premiers caractères du code postal
    - 4 premiers caractères de la ville
- Trier le fichier sur les champs "**match-code3**" (par ordre ascendant)
- Exercer un contrôle visuel pour mettre en évidence les lignes possédant un "**match-code3**" identique.
- Créer un champ mettant en évidence automatiquement les **match-codes consécutifs et identiques**.

Comparer ces trois essais de match-codage.

On pourra consulter le document « **filtrage_doublons-V4.pdf** » pour trouver quelques bases de réflexion sur la méthode de calcul de signature (ou d'empreinte) qui est sous-jacente au match-codage.

Exercice n°3 : Signature et dédoublonnage de fichiers

Le dossier **Share_etu/general/Stages_old** contient quelques centaines de fichiers de texte. (URL = https://etudiant.istic.univ-rennes1.fr/current/general/Stages_old)

On recherche les fichiers en doublon afin de les éliminer du dossier. Pour cela, on va appliquer la méthode du calcul de signature des fichiers qui est expliquée précédemment dans le document « **filtrage_doublons-V4.pdf** ».

1. On commence par définir un vocabulaire de référence (adéquat avec la sémantique des fichiers).

2. Ensuite, sur la base de ce vocabulaire de référence, on va calculer la signature de chaque fichier. Soit N le nombre de fichiers. On va donc calculer N signatures.

3. Puis, pour chaque fichier, on calcule la distance entre sa signature et celle des autres fichiers. A priori, si N est le nombre de fichiers, il faut calculer N x (N-1) / 2 distances.

4. Si la distance entre les deux signatures est nulle, on présume qu'il s'agit d'un doublon. On note les noms des deux fichiers présumés être des doublons afin de les vérifier manuellement.

5. Existe-t-il une méthode plus rapide ?