

## **Jim Godfrey (suite)**

### **Famille :**

Je suis présentement célibataire, et ma seule proche parente est ma mère.

### **Animaux :**

J'ai trois chiens répondant aux noms de Chico, Honey et Harley. J'ai aussi un perroquet youyou dénommé Kenya et un perroquet gris d'Afrique appelé Zulu.

### **Sports/Passe-temps :**

Je fais partie d'une société de reconstitution médiévale où j'étudie les techniques de combat à armes lourdes. Je fais aussi de la plongée et je joue au paint-ball.

### **La fin de semaine idéale :**

Pour moi, ce serait de partir en jet privé vers Cozumel, au Mexique, pour deux jours de plongée sur d'incroyables récifs!

### **Mets préférés :**

J'aime TOUS les mets indiens et thaïlandais...oh, et les sushis aussi!

### **Si je pouvais faire autre chose (que consultant SAS), je voudrais ...**

...entraîner des dauphins! J'ai fait mes études de premier cycle en biologie marine, et j'ai toujours voulu travailler avec des dauphins.

### **Lorsque je ne suis pas occupé par des projets de consultation SAS, j'aime ...**

Poursuivre ma formation SAS en suivant des cours offerts par le service de formation de SAS et assister à des conférences sur l'exploration de données comme la série M de conférences commanditée par SAS et qui a lieu chaque année à Las Vegas.

### **Une chose apprise sur le terrain en travaillant avec SAS qui pourrait être utile à d'autres utilisateurs SAS, c'est ...**

...l'intérêt d'appliquer le suréchantillonnage aux problèmes de modélisation prédictive binaire, en particulier lorsque le nombre global d'enregistrements de modélisation est très élevé et que le nombre de cas binaires d'intérêt sont rares. Je vais illustrer le concept de suréchantillonnage par l'exemple qui suit :

Prenons un problème de modélisation prédictive où vous essayez de prédire, disons, la probabilité qu'un client de la banque qui ne possède pas de carte de crédit choisisse un produit de carte de crédit. De plus, disons qu'il y a deux millions de clients potentiels pour se procurer une carte de crédit dans votre population de modélisation, et que

seulement 2 000 de ces clients sont devenus titulaires d'une carte de crédit dans la fenêtre cible définie (habituellement un mois).

Maintenant, disons que dans cet ensemble de données de 2 millions de clients, dont 2 000 avec présence de l'événement d'intérêt, il y a 1 500 variables d'entrée potentielles qui incluent des caractéristiques des clients, comme des données démographiques et des comportements passés à l'égard d'autres produits bancaires (historiques de transaction et de soldes de comptes).

Il est maintenant possible d'utiliser cet ensemble de données dans son format brut pour la modélisation prédictive, pour en tirer éventuellement des résultats de modélisation plutôt intéressants. Il s'agit toutefois d'un très gros ensemble de données, et son utilisation dans son format brut pour la modélisation exigerait d'énormes capacités de traitement, d'espace et de temps. De plus, l'événement cible est très rare, puisque seulement 0,1 pour cent des clients présentent l'événement d'intérêt! Plusieurs techniques de modélisation prédictive font appel aux mesures de précision prédictive comme objectif de modélisation par défaut. Par conséquent, un modèle serait précis à 99,9 pour cent en prédisant simplement qu'aucun client ne prendra de carte de crédit. Il s'agit d'un résultat très précis et totalement inutile! Ne serait-ce pas préférable que l'événement cible ne soit pas si rare et que la taille des données ne soit pas aussi grande ? La solution à ce problème, c'est le suréchantillonnage !

Le suréchantillonnage implique la création d'un échantillon à partir d'une population de modélisation où la densité des classes d'événements est enrichie. Pour ce faire, il faut sélectionner tous vos enregistrements avec événement et un sous-ensemble aléatoire de vos enregistrements sans événement, enrichissant ainsi votre population avec les cas d'événement. Dans le cas de l'exemple précédent, vous sélectionneriez d'abord la totalité des 2 000 enregistrements où l'événement d'intérêt est présent (c.-à-d. que le client est devenu détenteur d'une carte de crédit) et un sous-ensemble aléatoire d'enregistrements sans événement (c.-à-d. des clients qui ne sont pas détenteurs d'une carte de crédit). Ainsi, si nous voulions créer une population suréchantillonnée avec 50 pour cent d'enregistrements avec événement, nous choisirions les 2 000 enregistrements avec événement et 2 000 enregistrements parmi les 1 998 000 enregistrements sans événement restants, créant ainsi un ensemble de modélisation comptant 50 pour cent d'événements et 50 pour cent de non-événements. Nous pourrions aussi créer un ensemble de données de modélisation contenant un tiers d'enregistrements avec événement et deux tiers d'enregistrements sans événement en sélectionnant la totalité des 2 000 enregistrements avec événement et 4 000 enregistrements sans événement. En faisant cela, nous avons 1) réduit la taille des données, passant d'un déconcertant 2 millions d'enregistrements à un échantillon de 4 000 observations, que l'on peut facilement traiter sur un ordinateur personnel, et 2) enrichi la population d'événements cible, ce qui facilite la détection des cas d'événement et de non-événement.

Des études de simulation ont montré que les modèles construits au moyen des populations suréchantillonnées donnaient en moyenne des résultats tout aussi bons que les modèles construits avec un ensemble complet de données. Par conséquent, vous

pouvez obtenir à peu près les mêmes résultats en utilisant beaucoup moins de ressources informatiques. Toutefois, les modèles que vous construirez au moyen des ensembles de données suréchantillonnés donneront des résultats applicables à la population suréchantillonnée et NON à la population originale dont le taux d'événement est de 0,1 pour cent. Heureusement, cette erreur de précision du modèle se corrige facilement au moyen de la formule suivante toute simple :

$$p_1 = \frac{\tilde{p}_1(\pi_1/\rho_1)}{\tilde{p}_0(\pi_0/\rho_0) + \tilde{p}_1(\pi_1/\rho_1)}$$

Où  $\tilde{p}_1$  est la probabilité prédictive d'un événement dans le modèle construit à partir d'une population suréchantillonnée,  $\tilde{p}_0$  est la probabilité prédite d'un non-événement dans le modèle construit à partir de la population suréchantillonnée,  $\pi_1$  est la probabilité de l'événement dans la population réelle (dans ce cas  $2\,000/2\,000\,000 = 0,001$ ),  $\pi_0$  est la probabilité de non-événement dans la population réelle (dans ce cas  $1 - 0,001 = 0,999$ ),  $\rho_1$  est la probabilité de l'événement dans la population SURÉCHANTILLONÉE (qui serait de 0,33333 dans le cas d'un suréchantillon de 1/3, ou 0,5 pour un suréchantillon de 50/50), et finalement  $\rho_0$  est la probabilité de non-événement dans la population SURÉCHANTILLONÉE ( $1 - \rho_1$ ).