Cours : Data Mining

Enseignant : Professeur. Kilian Stoffel

Assistant : Iulian Ciorascu

Switzernet

Élaboré par :

Bouzerda Feriel- Hadjira

Louati Mortadha

Maâtallah Amine

Table des matières

Introduction générale :

Dans le cadre du cours de Data Mining, nous avons été amenés à appliquer à une base de données clients d'une entreprise que nous avons choisi trois techniques en utilisant l'outil Weka. Le but de notre travail est de pourvoir analyser le comportement des clients et de pouvoir mieux cibler les bons clients lors de prochaines campagnes publicitaires.

Avant de commencer à appliquer des techniques, il faut tout d'abord, préparer les données, c'est-à-dire prendre la base de données clients dans son format initial, analyser tous les attributs présents, ensuite choisir tous ceux qui peuvent apporter une information pertinente aux questions qu'on peut se poser sur les clients et ensuite la « nettoyer ».

Ensuite, nous avons appliqué la méthode d'*arbre de décision*, qui consiste à classifier les données dans différentes catégories. Familièrement, on dit qu'on essaie de « faire parler les données ».

La deuxième technique appliquée est le *clustering* qui a pour but de faire ressortir plus de patterns et plus de règles qu'avec la première méthode en définissant des groups de clients qui ont le même comportement.

Pour finir, nous avons appliqué la technique de *règles d'association* pour trouver plus de patterns et essayer d'affirmer ou infirmer les hypothèses de départ.

Pour chaque technique, nous avons interprété les résultats et plus important encore, nous avons essayé de donner des recommandations à l'entreprise par rapport à la gestion de leurs clients.

1. Première phase : Préparation des données

1.1 Introduction :

Nous avons choisi comme entreprise Switzernet. Cette entreprise propose un abonnement de téléphonie VOIP pour des particuliers ou des entreprises. Elle offre des appels gratuits vers plusieurs pays (CH, DE, UK, FR (réseau fixe), USA, Singapour et Hawaï (fixe et mobile)) ainsi que des tarifs très avantageux vers plusieurs destinations. Chaque client se voit attribuer un numéro de téléphone suisse (avec indicatif vaudois ou genevois), ce qui lui permet d'être joignable partout où il se trouve, si il a accès à Internet bien sûr.

En 2003, Switzernet s'est lancée dans ce marché de la téléphonie via Internet en se fixant comme client cible les PME. Depuis 2005, ils proposent des abonnements pour les particuliers à 9.- par mois. En 2006, ils ont mis en vente des téléphones VOIP pré configuré avec à l'intérieur le contrat Switzernet chez Darty et MediaMarkt à Lausanne et Genève ainsi que des adaptateurs qui se connectent à des téléphones traditionnels.

Aujourd'hui, ils ont plus 10 points de vente et quelques milliers d'abonnés. Les clients peuvent également commander et s'inscrire directement sur leur site Internet : http://switzernet.com

D'après une interview de la directrice Sona Gabrielyan, le principal défi de Switzernet est de convaincre les clients potentiels que la technologie VOIP leur permet de réaliser des grandes économies mais aussi les rassurer en disant que la téléphonie via Internet n'est pas très différente de la téléphonie traditionnelle.

Dans cette première phase, nous avons pris les données en état brut et nous avons analysé et choisi les attributs importants tout en « nettoyant » les données pour les rendre propre.

1.2 Forme des données :

Le fichier qu'on nous a fourni concerne des données générales pour chaque numéro Switzernet, et vient d'une base de données dans un serveur Linux. Elle contient des informations sur les clients (nom, prénom, adresse, numéro de téléphone attribué, type de contrat, type de revendeur, type d'équipement et nom du SIP serveur utilisé)

Nous avons trouvé dans le premier fichier 31 attributs que nous avons détaillés ci-dessous (tableau 1).

Nom	Catégorie	Remarque
Date	Date	Date de signature du contrat avec le jour le mois et l'année.
Month	Date	Date un peu plus générale uniquement le mois et l'année.
Contrat	Variable catégorielle	Nous indique normalement le type de contrat (« Prepaid », « Commercial », « Particulier »). Cependant nous trouvons diverses informations sur la façon dont a été signé le contrat comme le fait qu'il a été envoyé par fax ou par mail.
Compagny	Variable catégorielle	Indique le nom des entreprises qui ont signé chez Switzernet. Donc des clients qui ont un contrat de type « Commercial ».
First Name	Variable catégorielle	Nom du client.
Last Name	Variable catégorielle	Prénom du client.
Adresse	Variable catégorielle	Adresse du client
C.P	Variable numérique	Code postal du client
City	Variable catégorielle	Ville où est situé le client
Private Phone	Variable numérique	Numéro de téléphone privé du client.
Prof Phone	Variable numérique	Numéro de téléphone professionnel du client.
Mobile	Variable numérique	Numéro de téléphone portable du client.
Fax	Variable numérique	Numéro de Fax du client.
E-mail	Variable catégorielle	Email du client.
E-mail 2	Variable catégorielle	Deuxième E-mail du client.
Carte ID	Variable Binaire	Cette variable indique si une photocopie d'une pièce d'identité a été fournie ou pas.
Revendeur	Variable catégorielle	Switzernet a un système d'affiliation et a donc diffèrent vendeur en Suisse romande. On indique donc dans ce champ le revendeur qui a effectué la transaction.
SIP	Numérique	Numéro de téléphone international du client.

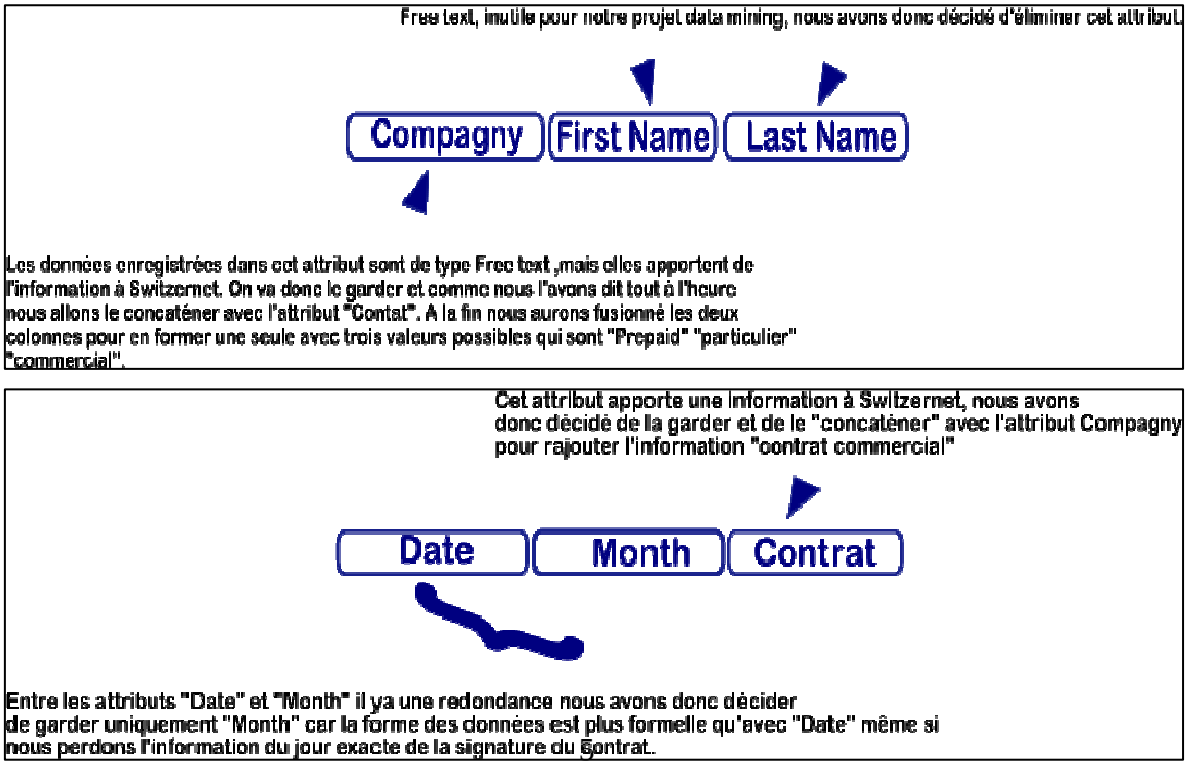
Phone	Numérique	Numéro national du client.
International	Numérique	Numéro international du client avec le petit « + ».
Short SIP	Numérique	Les 4 derniers chiffres du numéro de téléphone.
Voip Password	Variable catégorielle	Password du client pour s'authentifier sur l'application Voip.
Web login	Variable catégorielle	Login du client pour s'identifier sur son compte sur internet.
Web password	Variable catégorielle	Password du client pour s'authentifier sur son compte sur internet.
Model	Variable catégorielle	Model de l'appareil choisit par le client.
MAC	Variable catégorielle	Adresse MAC de l'appareil vendu au client.
Price	Monetary amounts	Prix de l'appareil en question.
Monthly	Monetary amounts	Facture mensuelle de l'abonnement du client.
Remarque	Variable catégorielle	Diverses remarques de Switzernet.
SIP Server	Variable catégorielle	Serveur auquel est rattaché le numéro Switzernet en question.
Number Price	Monetary amounts	Prix du numéro choisit par le client.

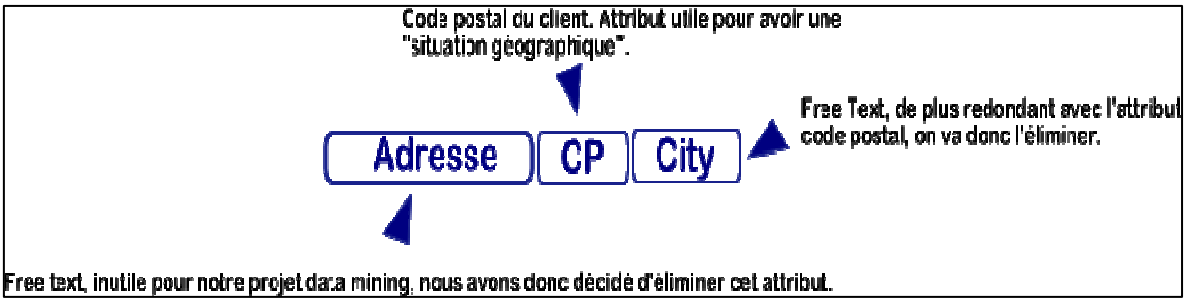
Tableau 1

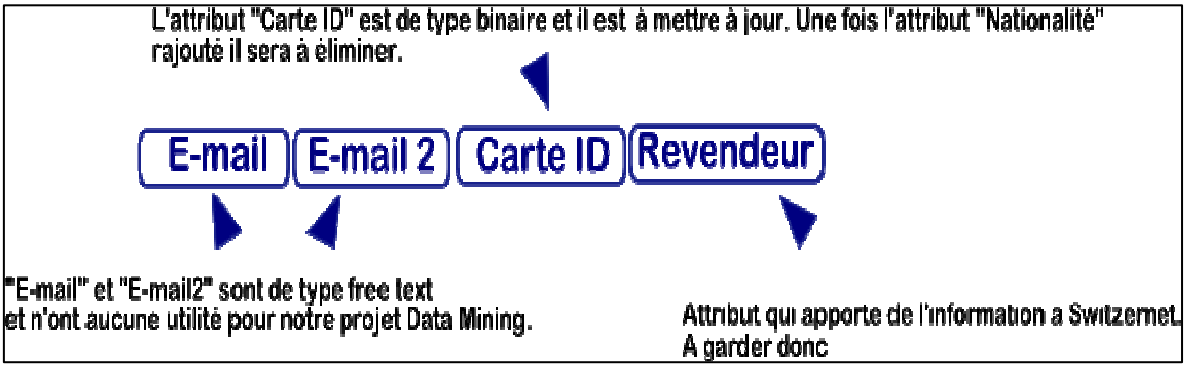
Pour Switzernet, un client = un numéro même si une même personne prend plusieurs numéro de téléphone. Par conséquent, nous avons l'égalité suivante :

Une ligne de la base de donnée = Un numéro de téléphone Switzernet

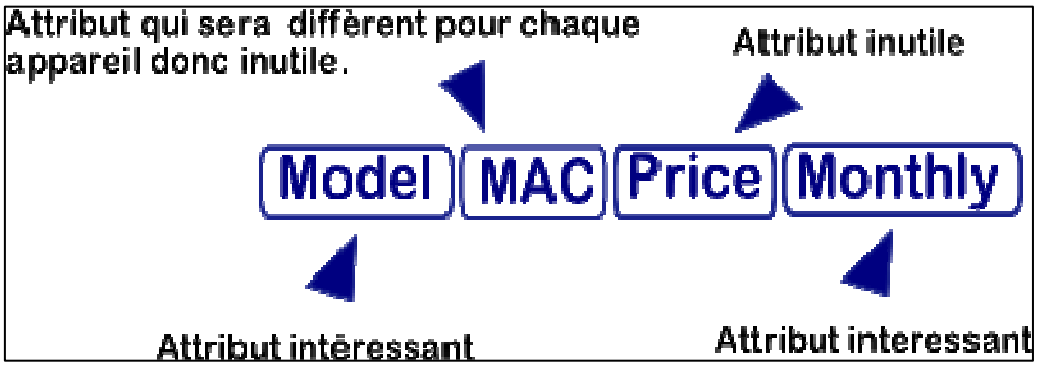
Voici avec plus de détail les choix effectués sur chaque attribut :



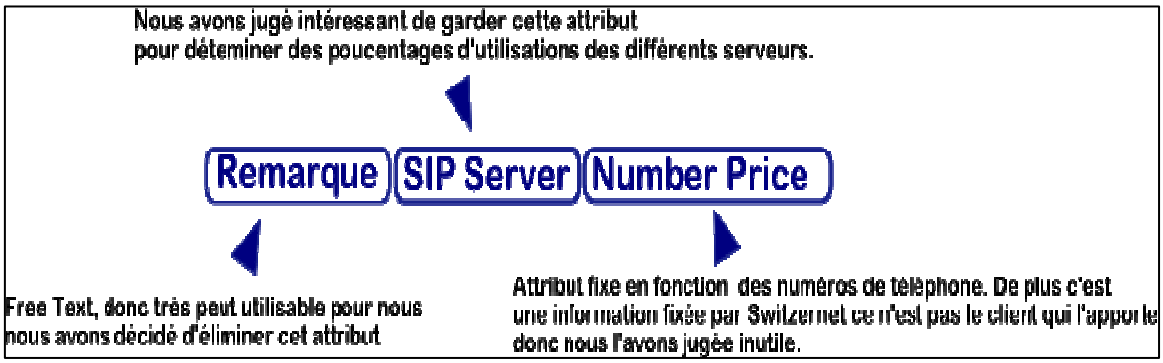

Code postal du client. Attribut utile pour avoir une "situation géographique".
Adresse
CP
City
Free Text, de plus redondant avec l'attribut code postal, on va donc l'éliminer.
Free text, inutile pour notre projet data mining, nous avons donc décidé d'éliminer cet attribut.


L'attribut "Carte ID" est de type binaire et il est à mettre à jour. Une fois l'attribut "Nationalité" rajouté il sera à éliminer.
E-mail
E-mail 2
Carte ID
Revendeur
"E-mail" et "E-mail2" sont de type free text et n'ont aucune utilité pour notre projet Data Mining.
Attribut qui apporte de l'information a Switzernet. A garder donc


SIP
Phone
International
Short SIP
Ces quatre attributs donnent exactement la même information: l'identidiant du produit, plus précisement, du numéro de téléphone (qui est en fait le numéro de téléphone lui même). On va donc uniquement garder l'attribut "SIP".
Dans ce projet Data Mining on a "une ligne = un numéro de téléphone"


Attribut qui sera différent pour chaque appareil donc inutile.
Attribut inutile
Model
MAC
Price
Monthly
Attribut intéressant
Attribut interessant


Private Phone
Professional Phone
Mobile
Fax
Tous ces attributs sont inutiles pour notre projet Data Mining, pour la simple raison qu'on ne trouvera jamais deux client avec le même numéro de fax ou de mobile.


Nous avons jugé intéressant de garder cette attribut pour déteminer des poucentages d'utilisations des différents serveurs.
Remarque
SIP Server
Number Price
Free Text, donc très peut utilisable pour nous nous avons décidé d'éliminer cet attribut
Attribut fixe en fonction des numéros de téléphone. De plus c'est une information fixée par Swilzernet ce n'est pas le client qui l'apporte donc nous l'avons jugée inutile.

1.3 Analyse des données :

Après avoir décidé quel champ utile garder pour analyser et préparer nos données, nous avons choisi de prendre seulement 1000 clients parmi les 2000 que compte Switzernet. De plus, nous avons seulement choisi ceux du canton de Vaud (Numéro commençant avec 021).

Pour rendre les données « propres » et lisibles par Weka, il a fallu remplacer les valeurs manquantes par des *Null* et aussi uniformiser les valeurs de certains champs, par exemple :

- Il a fallu remettre le champ « Date » sous le même format (mois, année) écrit de la même manière (en anglais et tous commençant par une majuscule).

- Concernant le champ Model qui décrit les produits additionnels que Switzernet vend, il y en a 11 types qui avaient la plupart une orthographe différente.

Lorsque nous avons pu exporter le fichier Excel (transformé en csv), nous avons pu observer quelques histogrammes intéressants :

- La figure 1 nous montre quel est le mois où il y a eu le plus de souscription d'abonnement. Ici, c'est septembre et octobre 2007. On peut se demander s'il y a un rapport avec la rentrée universitaire puisque Switzernet est très présente à l'EPFL et a mis des affiches publicitaires dans le TSOL.

- La figure 2 nous montre quel est le revendeur le plus efficace. Il faut d'abord remarquer que la majorité des souscriptions des abonnements se fait directement chez Switzernet et que le système de parrainage est assez bien utilisé. Les revendeurs principaux sont les points de vente Darty et puis VoIP Shop Nyon.

- La figure 3 nous montre quel est le serveur le plus utilisé et par conséquent celui dont il faut le plus prendre soin. Ici, c'est le serveur 66.234.138.73, ensuite sip.youroute.net et enfin sip.switzernet.com. Il est assez étonnant de trouver le serveur principal de Switzernet en troisième position.

- La figure 4 montre quels sont les appareils les plus vendus et pour lesquels il faudrait faire plus de publicité, comme HandyTone 286 (adaptateur), le logiciel ainsi que le BudgeTone 100(adaptateur).

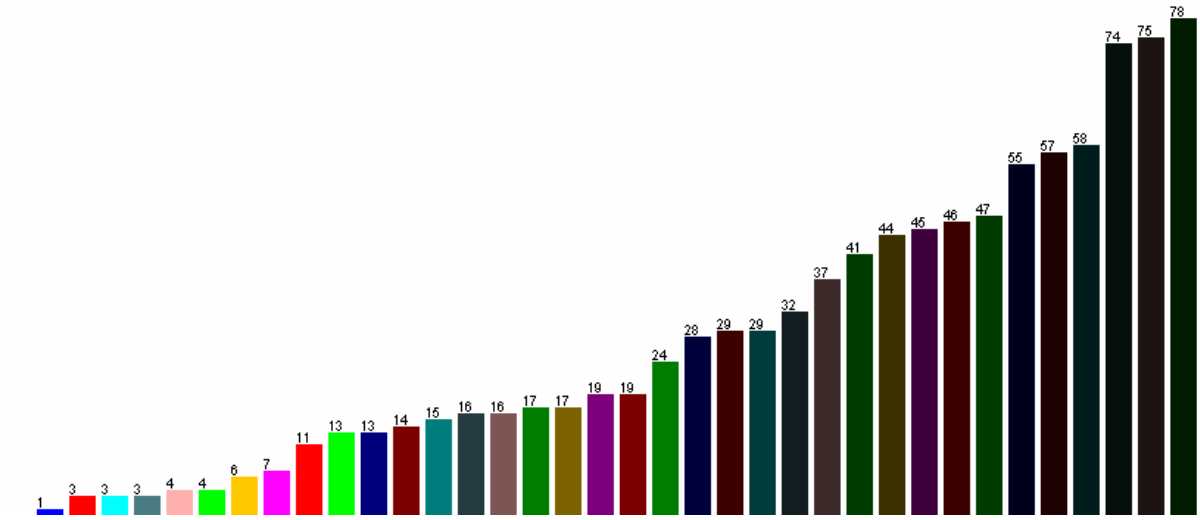


Fig. 1 Souscription d'abonnement par mois.

Fig 2. Principaux revendeurs

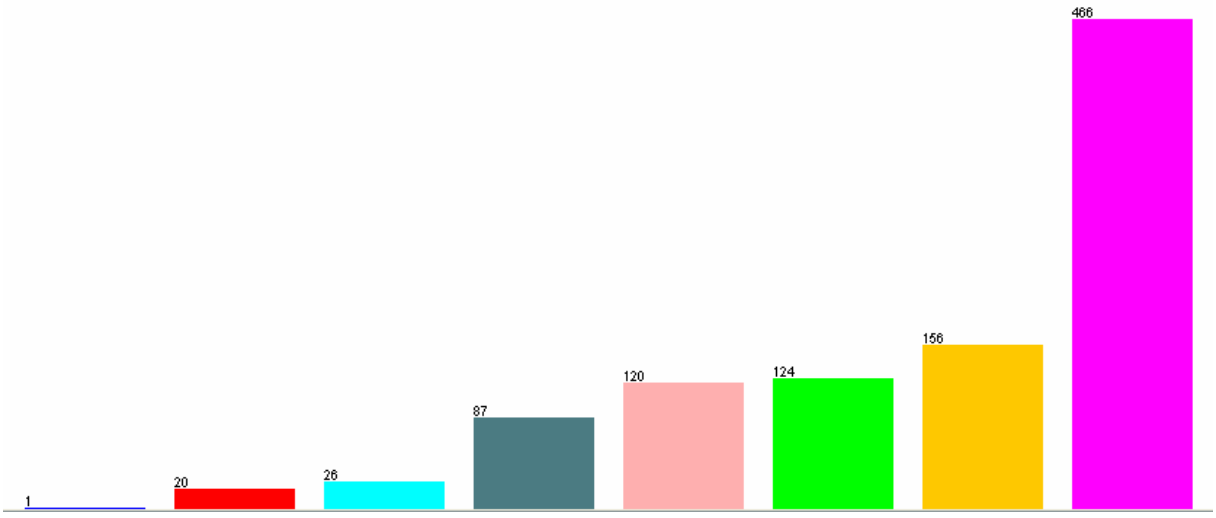


Fig. 3 Serveurs SIP

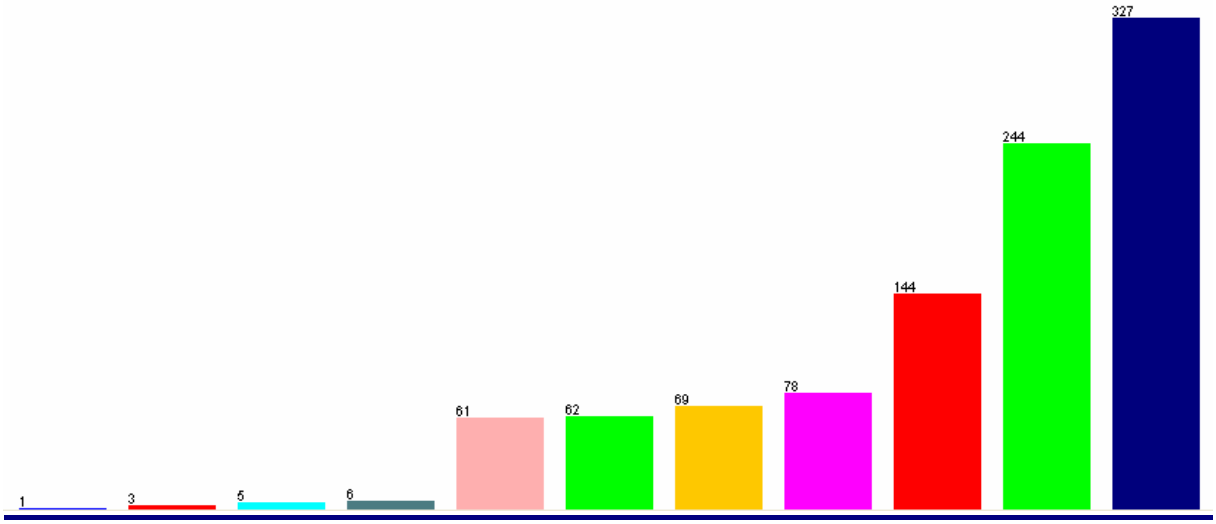


Fig.4 Appareils les plus vendus

1.4 Conclusion :

Cette première étape nous a permis de préparer les données en ayant une base de données « épurées » avec les attributs qui nous intéresse et surtout pas de données manquantes et des orthographes uniformisées.

Nous nous sommes posé d'autres questions que nous aimerions aborder lors de la prochaine étape comme :

- Quel est le profil de client qu'il faudra cibler lors d'une prochaine campagne publicitaire ?
- Quels sont les bons et les mauvais clients ?

Nous espérons vraiment pouvoir apporter des solutions à ces questions grâce à la prochaine étape.

2. Deuxième Phase : Arbres de Décision

2.1 Introduction :

Dans cette deuxième partie, nous nous sommes intéressés aux arbres de décisions qui vont nous permettre de faire des classifications de nos clients.

Nous aurions aimé répondre aux questions que nous nous sommes posées pendant la première phase, c'est-à-dire quels sont les bons et mauvais clients et quelles personnes cibler pour une campagne publicitaire.

Il faut néanmoins noter qu'avec la base de données que nous avons à disposition, nous n'avons vraiment pas assez d'attributs pertinents pour pouvoir répondre à nos problèmes, nous avons donc ajouté trois attributs et nous avons appliqué l'algorithme (J48) sur Weka. Nous avons obtenu deux arbres que nous présentons plus loin.

2.2 Analyse des données :

Après l'analyse des données, nous avons réalisé qu'il fallait ajouter des attributs qui permettraient de réaliser des bonnes catégorisations de classe. Nous avons donc ajouté les attributs suivants :

- Trimestre : nous avions un champ « Month » qui nous donnait les dates de signature de contrat. Nous avons décidé de les regrouper en trois périodes de l'année, c'est-à-dire :
    1. Janvier → Avril.
    2. Mai → Août.
    3. Septembre → Décembre.
- GlobalRevendeur : cet attribut regroupe tous les revendeurs de l'entreprise en quatre catégories : Parrainage, revendeur en magasin, publicité par des flyers et enfin aucun revendeur, c'est-à-dire souscriptions des clients sans intermédiaire.
- TypeEquipement : catégorise tous les types d'appareil vendus en téléphone ou adaptateur IP.

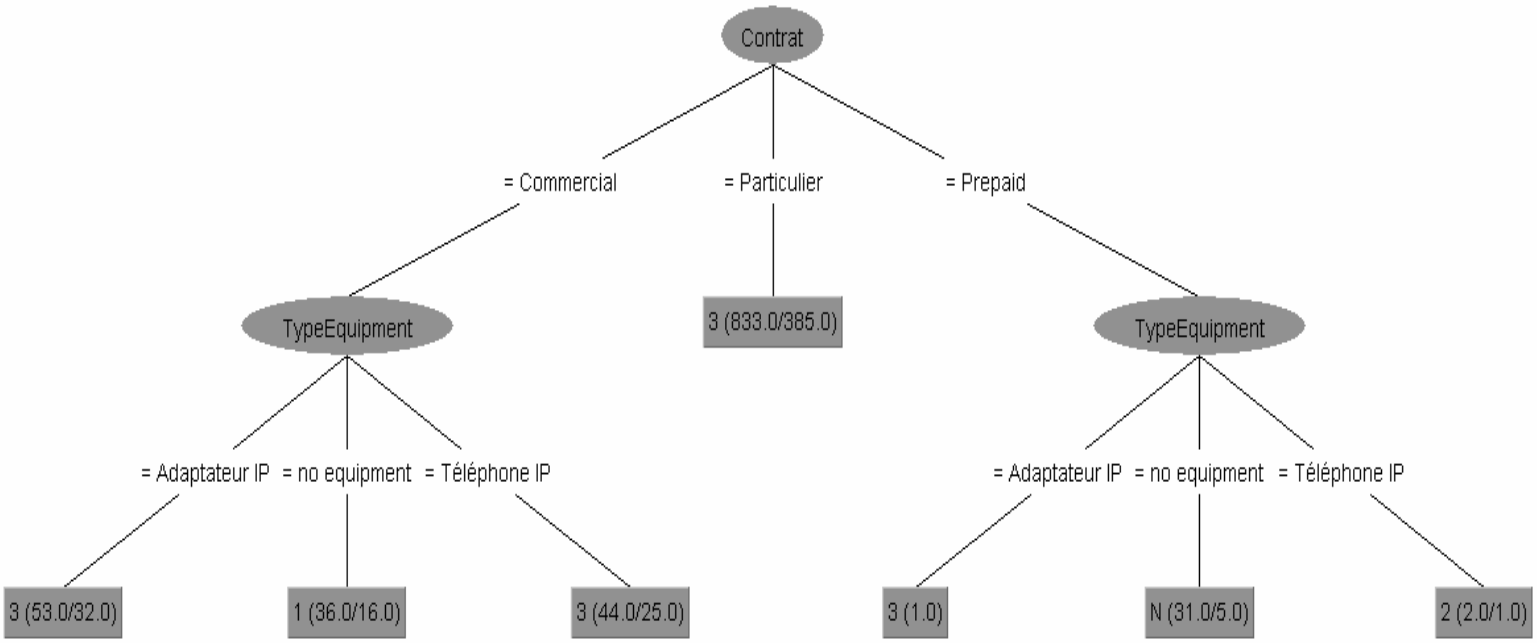
2.3 Premier problème :

Nous avons voulu savoir comment se faisait la répartition de la signature des différents contrats tout le long de l'année. Pour rappel, Switzernet propose trois types de contrats : Commercial, Prepaid et Commercial.

Nous avons donc gardé comme champ le contrat, le type d'équipement et le trimestre dans l'onglet « preprocess » dans Weka et ensuite nous avons appliqué l'algorithme J48 dans l'onglet « classify ».

2.3.1 Résultats obtenus et interprétations:

Tree View



```
Instances:    1000
Attributes:   3
              Trimestre
              Contrat
              TypeEquipment
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Contrat = Commercial
|   TypeEquipment = Adaptateur IP: 3 (53.0/32.0)
|   TypeEquipment = no equipment: 1 (36.0/16.0)
|   TypeEquipment = TÈlÈphone IP: 3 (44.0/25.0)
Contrat = Particulier: 3 (833.0/385.0)
Contrat = Prepaid
|   TypeEquipment = Adaptateur IP: 3 (1.0)
|   TypeEquipment = no equipment: N (31.0/5.0)
|   TypeEquipment = TÈlÈphone IP: 2 (2.0/1.0)



Number of Leaves:     7

Size of the tree:         10

=== Stratified cross-validation ===
=== Summary ===


Correctly Classified Instances         526               52.6    %
Incorrectly Classified Instances       474               47.4    %

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
  26   0 284   0 |   a = 1
  11   0 12 8   2 |   b = 2
  21   0 474   6 |   c = 3
   4   0  18  26 |   d = N
```

Nous constatons que l'algorithme arrive à classifier 52,6% des instances correctement donc à peu près une instance sur deux. Cependant cela reste le meilleur résultat obtenu. En effet, nous avons rajouté d'autres attributs pour essayer de classifier les instances de façon plus pure et malgré cela le résultat reste le même, nous avons toujours à peu près 53% de classification correcte sauf que l'arbre est moins bien élagué que celui que nous avons choisi.

La période N°3 enregistre le plus grand nombre de souscriptions pour le contrat de type « particulier ». Il y a probablement un lien avec la rentrée scolaire et universitaire. Il pourrait également avoir un rapport avec le fait que les gens rentrent de vacances et changent leurs abonnements, chose qu'ils ont préféré laisser pour après les vacances. Sur les 833 recensés, il y en a 448, soit plus de la moitié. Il faudrait donc concentrer la campagne publicitaire concernant le contrat « particulier » pendant cette période.

- En ce qui concerne le contrat « prepaid », nous avons pu constater que l'algorithme passe par deux étapes pour effectuer une classification. Pour cela, il utilise l'attribut contrat et l'attribut type d'équipement.

- Il en est de même pour le contrat de type « commercial ».

2.3.2 Recommandations :

On suggère à Switzernet les éléments suivants :

- Pour les contrats « particuliers », il serait judicieux de mettre en place une campagne publicitaire ciblée durant la 3ème période de l'année.
- Pour les contrats « commercial », les deux règles suivantes seraient intéressantes :
    - Coupler l'offre contrat commercial + téléphone IP ; ou
    - Coupler l'offre contrat commercial + adaptateur IP. Ceci ne s'applique que pour la 3ème période.
- Par contre, durant la 1ére période, cette offre « couplée » ne serait pas judicieuse. En effet, on constate que 55% des abonnés « commercial » et qui n'ont pas pris d'équipements ont signé durant cette période. À part si Switzernet veut bien sûr inverser la tendance.

2.4 Deuxième problème :

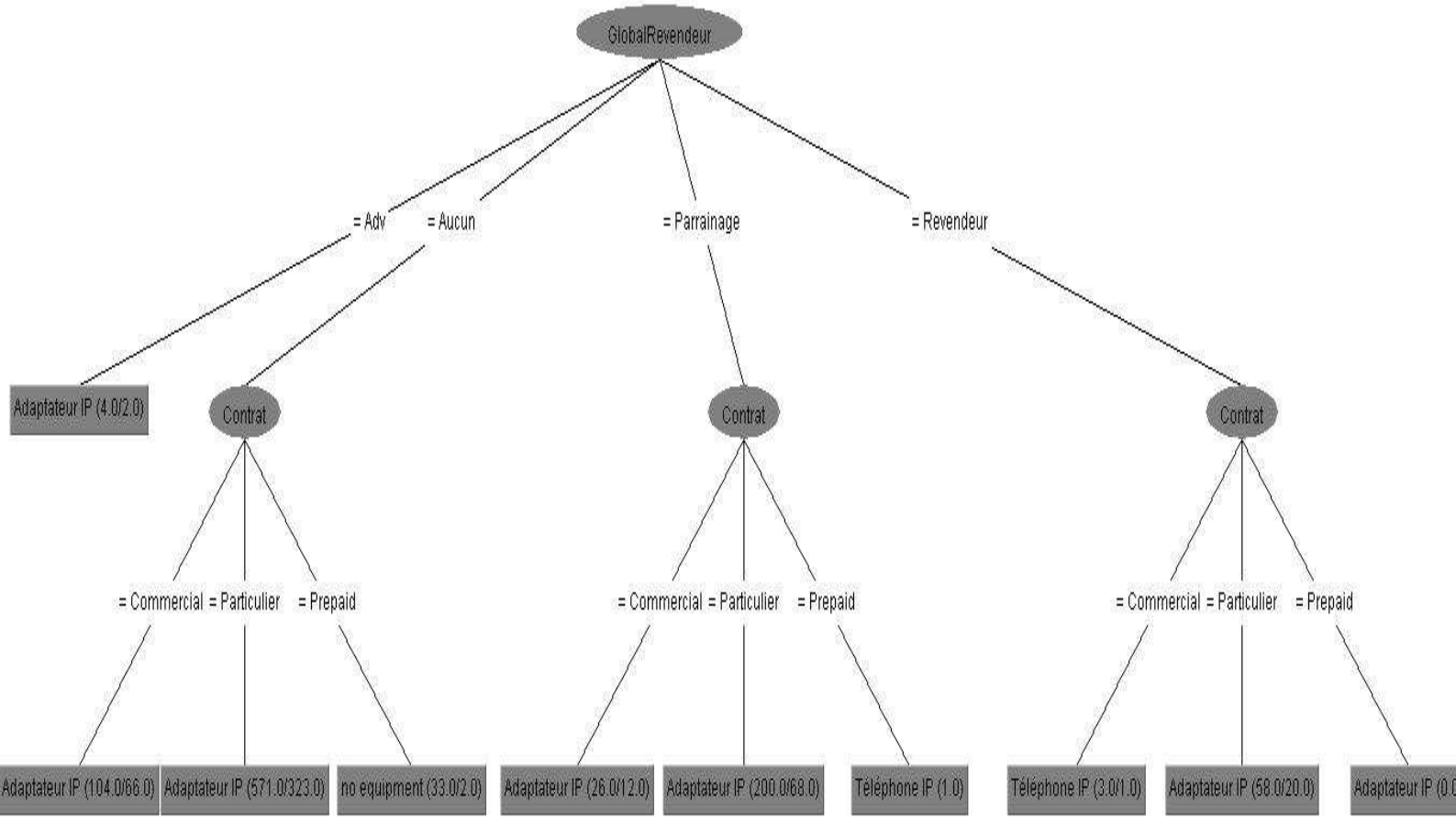
Nous avons voulu savoir qu'elle était le canal de distribution le plus efficace quant à la promotion des propositions de valeurs de Switzernet, et s'il y avait des problèmes, quels serait les meilleurs arrangements à faire.

Voici les différents canaux de distribution de Switzernet :

1. Le bouche-à-oreille: parrainage
2. Les vendeurs
3. La publicité
4. Autre: Site web

Nous avons utilisé les champs: GlobalRevendeur, type de contrat et type d'équipement et nous avons choisi l'algorithme J48 pour obtenir le résultat suivant.

2.4.1 Résultats obtenus et interprétations:



On constate que le canal de distribution le plus efficace est le quatrième, c'est-à-dire le site web de Switzernet en général. Arrive en deuxième position le premier canal de distribution. Et enfin pour finir le deuxième et troisième.

2.4.2 Recommandations :

Suite à ces résultats, il semblerait que le site Internet de Switzernet explique bien les différentes offres que propose cette dernière et que sa position sur la toile lui permet de figurer sur les pages clés concernant la téléphonie VoIP ce qui lui permet de toucher un maximum de clients potentiels. On peut donc dire qu'il est indispensable de continuer dans cette démarche et même de concentrer une importante partie du budget publicitaire à ce niveau-là.

Même constat pour le parrainage qui donne des résultats très satisfaisants et qui indique qu'il faut continuer et faire évoluer ce principe.

Par contre le canal de distribution « publicité » présente des résultats très médiocres malgré un financement important de la part de Switzernet. Il faudrait donc remettre en question cette façon de faire et peut être se focaliser pour le moment sur les trois autres canaux de distribution.

Pour finir en ce qui concerne le canal de distribution « revendeurs » son utilisation reste mitigée malgré un très grand potentiel. En effet, on voit qu'il apporte des clients à Switzernet mais pas autant que les autres. Ceci est certainement dû au fait que les vendeurs sont peu qualifiés pour expliquer en détail les offres que propose Switzernet.

Nous tenons à remarquer que cet arbre ne nous apporte pas beaucoup d'information puisque nous avions déjà cette information en analysant les données en « preprocess » sur Weka mais malheureusement nous n'avons pas des données qui nous permettent de faire de meilleurs arbres.

2.5 Conclusion :

Nous avons pu répondre aux deux questions que nous nous sommes posées au départ mais il est évident que l'information importante est quand faut-il cibler les futurs clients par compagne publicitaire et nous en avons conclu qu'il faudrait le faire la dernière période de l'année.

Nous tenons à vous faire remarquer qu'il n'a pas été facile d'obtenir de résultats plus élaborés et ce pour plusieurs raisons :

- Les données que nous avions n'étaient pas assez pertinentes pour pouvoir en sortir des règles assez élaborées.
- Pour créer des arbres de décision qui donnent un résultat, il a fallu catégoriser plusieurs champs qui ne donnent aucune valeur dans un arbre de décision, comme par exemple catégoriser les revendeurs, les trimestres ainsi que les types d'équipement.
- Nous avons également essayé d'obtenir un arbre en catégorisant le code postal par canton mais malheureusement sans succès.

3. Troisième Phase : Clustering

3.1 Introduction :

Après l'utilisation des arbres comme première méthode de classification, nous nous sommes intéressé de plus prés aux méthodes de « clustering » notamment à l'algorithme « SimpleKMeans », dans l'attente de faire ressortir plus de patterns et plus de règles d'associations qu'avec les arbres de décisions. Cette fois, nous avons procédé différemment : sans questions au préalable nous avons tout simplement testé nos données avec l'algorithme en jouant avec les deux variables « nombre de cluster » et « seed », et à notre grande surprise il y a eu plus de résultats qu'avec la technique précédente. Nous allons donc présenter un par un les résultats obtenus et les commenter.

3.2 Premier résultat : Confirmation

Fort de nos résultats obtenus dans la partie précédente nous avons voulu, dans un premier temps, avoir une confirmation de ces derniers avec cette nouvelle technique (voir figure 1 pour les résultats).
Rappelons tout d'abord que nous avons divisé l'année en trois périodes :

- Période 1 : Janvier → Avril
- Période 2 : Mai → Août
- Période 3 : Septembre → Décembre

Et tout comme avec les arbres de décision, on arrive à confirmer le fait que les particuliers ont tendance à signer un contrat avec Switzernet durant la 3ème période. Dans notre cas, ceci est représenté par le cluster numéro 1 avec 50% des instances.
Autre remarque intéressante avec le résultat obtenu par clustering, c'est que la comparaison entre les différents clusters joue sur le même dénominateur commun, qui sont les attributs « type de contrat : particulier » et « global revendeur : Aucun », et au numérateur on a l'attribut « Trimestre » ce qui n'était pas le cas avec l'arbre de décision. On se retrouve donc avec une répartition sur toute l'année des signatures de contrats de type « particulier » beaucoup plus précise qu'avec les arbres de décision.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -S 100
Relation:     clients_Switzernet-weka.filters.unsupervised.attribute.Remove-R2,5-6,8-10,12-
weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-
R3,5
Instances:    1000
Attributes:   3
              Trimestre
              Contrat
              GlobalRevendeur
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 507.0

Cluster centroids:

Cluster 0
        Mean/Mode:  2 Particulier Aucun
        Std Devs:   N/A   N/A   N/A
Cluster 1
        Mean/Mode:  3 Particulier Aucun
        Std Devs:   N/A   N/A   N/A
Cluster 2
        Mean/Mode:  1 Particulier Aucun
        Std Devs:   N/A   N/A   N/A

Clustered Instances

0      189 ( 19%)
1      501 ( 50%)
2      310 ( 31%)
```

Figure 1

3.3 Deuxième résultat : Efficacité sur la façon d'atteindre les clients :

Commençons encore par un petit rappel, histoire de ne pas être perdu lors de l'interprétation des résultats. On a trois types de contrats, et trois types d'appareils.
Type de contrat :

- Particulier
- Commercial
- Prepaid

Type d'appareil:

- Aucun (seulement logiciel)
- AdaptateurIP
- TéléphoneIP

En jouant uniquement avec ces deux attributs, nous avons obtenu le clustering en figure 2. Les résultats sont une sorte de mapping sur l'efficacité des différents canaux de distributions (d'après notre interprétation) même ci cet attribut n'apparaît pas.

Il y a principalement 3 canaux de distribution :

- « Aucun » qui fait référence au site Web de Switzernet.
- « Parrainage » où là c'est les anciens clients qui proposent les produits.
- « Revendeur » les différents partenaires de Switzernet comme Media-Markt.

Le canal de distribution "Aucun" présente toutes les combinaisons possibles avec l'attribut « Type d'équipement », c'est-à-dire :

{Aucun, Téléphone IP} {Aucun, no equipement} {Aucun, Adaptateur IP}

Ce qui veut dire (toujours selon notre interprétation) que le site Web fournit toutes les informations et les choix possibles nécessaires, pour que le client fasse son choix ce qui est très positive.

Mêmes constatations pour le canal de distribution « Parrainage ». Les clients déjà chez Switzernet savent plutôt bien expliquer ce que propose cette dernière. Les trois différentes combinaisons possible semblent être bien représentées (voir figure 2), donc l'initiative du parrainage apporte ses fruits.

Par contre au niveau du canal de distribution « Revendeur » le constat n'est pas aussi positif. Même si, sur les trois combinaisons possibles il en apparaît deux, pour ces dernières les chiffres obtenus ne sont pas glorieux. En effet, on a 2% des instances pour la combinaison {Revendeur, Téléphone IP} et 4% pour la combinaison {Revendeur, Adaptateur IP}. Switzernet est donc en droit de se poser des questions sur l'efficacité de ce canal de distribution.

Voici quelques questions que nous avons été amenés à nous poser avec ces résultats :

- Pourquoi la combinaison {Revendeur, no eqquipment} n'apparaît pas alors qu'elle est tout à fait possible ?
- Les salariés des différents partenaires sont-il capables de bien expliquer aux clients potentiels les différentes offres que proposent Switzernet ?
- Peut être faut-il former ces salariés d'une différente façon ou tout simplement plus régulièrement pour atteindre des chiffres plus élevés ?

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 8 -S 235
Instances:    1000
Attributes:   2
              GlobalRevendeur
              TypeEquipment
Test mode:    evaluate on training data

Number of iterations: 2
Within cluster sum of squared errors: 5.0

Cluster centroids:

Cluster 0
        Mean/Mode:  Aucun Téléphone IP
        Std Devs:   N/A   N/A
Cluster 1
        Mean/Mode:  Aucun no equipment
        Std Devs:   N/A   N/A
Cluster 2
        Mean/Mode:  Parrainage no equipment
        Std Devs:   N/A   N/A
Cluster 3
        Mean/Mode:  Parrainage Adaptateur IP
        Std Devs:   N/A   N/A
Cluster 4
        Mean/Mode:  Parrainage Téléphone IP
        Std Devs:   N/A   N/A
Cluster 5
        Mean/Mode:  Aucun Adaptateur IP
        Std Devs:   N/A   N/A
Cluster 6
        Mean/Mode:  Revendeur Téléphone IP
        Std Devs:   N/A   N/A
Cluster 7
        Mean/Mode:  Revendeur Adaptateur IP
        Std Devs:   N/A   N/A

Clustered Instances

0      196 ( 20%)
1      228 ( 23%)
2       16 (  2%)
3      148 ( 15%)
4       65 (  7%)
5      287 ( 29%)
6       21 (  2%)
7       39 (  4%)
```

Figure 2

3.4 Troisième résultat : Efficacité des revendeurs sur Vaud et Genève :

Nous avons constaté dans nos données que les clients ont majoritairement une adresse dans le canton de Vaud à 54% et à 16% dans le canton de Genève. Nous nous sommes posés la question suivante : quel type de revendeur utilise les clients dans chaque canton et surtout est-ce que les revendeurs sont bien représentés dans les différents cantons ?

Nous avons appliqué l'algorithme avec plusieurs attributs et plusieurs combinaisons (num-clusters et seed). Nous avons finalement trouvé 4 groupes en utilisant seulement deux attributs : GlobalRevendeur qui comme expliqué plus haut est composé de {Aucun, Parrainage, Revendeur et Flyer (Publicité)}. Les revendeurs sont représentés par Void Shop Nyon (VD) et les points de vente suivants : Agepoly-EPFL (VD), Darty à Crissier, Etoy et Villeneuve (VD), MediaMarkt Crissier (VD) et Meyrin (GE), New Telecom (VD), Steg à Ecublens (VD) et à Sion (VS) et enfin Tele Croset (VD).

Sur l'échantillon de 1000 clients, voici le tableau qui donne le compte pour chaque revendeur : (Tableau tiré de Weka)

Label	Count
N	708
Agepoly	5
DARTY	26
Flyer	4
new telecom	4
Parrainage	227
Steg Ecublens	12
Telecroset	3
Voip Shop Nyon	11

Tableau 1

Il faut noter que la majorité des clients n'utilise aucun revendeur (71%), c'est-à-dire qu'ils souscrivent directement sur le site de Switzernet. Le deuxième moyen de vente est le parrainage (23%). Il est donc assez difficile d'obtenir des groupes bien représentatifs des clients qui utilisent des revendeurs. Néanmoins, nous avons voulu savoir lesquels de ces points de vente sont les plus utilisés dans les deux cantons représentatifs, c'est-à-dire Vaud et Genève.

Les résultats que nous avons obtenus sont illustrés dans la figure 3. Les clusters 0, 2 et 3 représentent le canton de Vaud à 58%, et comme nous l'avons constaté, les revendeurs ne sont présents qu'à 7%, le parrainage à 19% et le site web est donc le meilleur moyen de vente pour le canton de Vaud avec 32%.En ce qui concerne le canton de Genève, là encore c'est au site que revient le meilleur score.

Ce que nous avons voulu montrer ici est que le seul revendeur se trouvant dans le canton de Genève (Steg) n'est représenté aucune fois sur les 1000 clients. Il est clair que l'absence ou presque de points de vente à Genève est un point assez négatif de la stratégie de vente de Switzernet.

Si cette dernière veut être plus présente dans la suisse romande, elle devrait peut - être renforcer sa présence dans le canton de Genève par plus de points de vente ou par une meilleure publicité. Il est à noter également que pour le canton de Vaud la publicité ne donne pas vraiment de bon retour, ce dont Switzernet est assez consciente et essaie d'améliorer.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 4 -S 250
Relation: clients_Switzernet-weka.filters.unsupervised.attribute.Remove-R1-2,4,6-7,9-11,13-14-weka.filters.unsupervised.attribute.Remove-R1weka.filters.unsupervised.attribute.Remove-R3
Instances: 1000
Attributes: 2
Canton
GlobalRevendeur
Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 340.0

Cluster centroids:

Cluster 0
Mean/Mode: Vaud Revendeur
Std Devs: N/A N/A
Cluster 1
Mean/Mode: Genève Aucun
Std Devs: N/A N/A
Cluster 2
Mean/Mode: Vaud Parrainage
Std Devs: N/A N/A
Cluster 3
Mean/Mode: Vaud Aucun
Std Devs: N/A N/A

Clustered Instances

0 65 (7%)
1 419 (42%)
2 194 (19%)
3 322 (32%)

Figure 3

3.5 Conclusion :

En appliquant l'algorithme « SimpleKMeans », nous avons pu obtenir trois clustering intéressants :

- Le premier nous a permis de confirmer ce que nous avions trouvé avec les arbres de décision c'est-à-dire que les particuliers ont tendance à signer un contrat avec Switzernet durant la 3ème période, ce qui correspond à la période de la rentrée scolaire/universitaire.
- Le deuxième clustering nous a permis de mieux répartir les différents canaux de distribution de Switzernet. Nous avons pu voir que le site Web représente le meilleur canal, ensuite le parrainage et enfin les revendeurs. Pour améliorer ce dernier canal, nous proposons à Switzernet de mieux former et informer les revendeurs ou alors de mettre de temps en temps des personnes qui connaissent bien les produits Switzernet pour faire un peu de publicité chez les revendeurs.
- Le troisième clustering nous a permis de réaliser que la présence de Switzernet dans le canton de Genève est très faible et pour cette raison, nous pensons que les clients ont plus de peine à devenir clients chez eux par rapport au canton de Vaud. Là encore, nous proposons à Switzernet de renforcer cette présence à Genève en acquérant plus de points de vente ou alors en faisant plus de publicité dans ce canton.

Nous espérons pouvoir apporter plus de réponses lors de la prochaine phase qui consistera à trouver des règles d'association.

4. Quatrième Phase : Règles d'association

4.1 Introduction :

La dernière technique que nous allons appliquer est de trouver des patterns par des règles d'association. Nous avons utilisé l'algorithme « A priori » sur Weka. Pour pouvoir estimer si les règles que nous avons trouvées sont bonnes, nous avons choisi comme métrique le Lift qui nous donne combien la règle trouvée est meilleure que l'intuition ou la règle par défaut.
Nous avons un choisi un Lift entre 1.2 et 1.7 car dans notre cas, c'est ce qui peut nous donner les meilleurs résultats.

4.2 Résultats :

Nous avons choisi les attributs suivants comme points d'entrée à l'algorithme :

- Trimestre : {1,2,3,Null}
- Contrat : { Commercial, Particulier, Prepaid}
- Canton : 19 cantons dont les plus représentés Vaud (538), ensuite Genève (162).
- GlobalRevendeur : {Aucun, Parrainage, Revendeur, Publicité}
- TypeEquipement : {Adaptateur IP, Téléphone IP, no equipement}

Parmi plusieurs règles que nous avons obtenues voici les règles qui nous ont le plus intéressées :

4.2.1 Règle 1 :

1. Canton=Vaud TypeEquipment=Adaptateur IP 280 ==> GlobalRevendeur=Parrainage 106 conf:(0.38) < lift:(1.67)> lev:(0.04) [42] conv:(1.24)

La première règle que nous avons obtenu et qui a le lift maximum de 1.67 nous confirme l'efficacité du canal de distribution parrainage. D'après elle, si le client habite au canton de Vaud et qu'il choisit un adaptateur IP, c'est qu'il a contracté le contrat avec Switzernet par parrainage. Cette règle est très intéressante car d'après nos précédentes constatations, le canal de distribution le plus efficace est le site web et en deuxième lieu le parrainage. Il faut encore noter que l'indice de confiance n'est pas très élevé (0.38)

4.2.2 Règle 2 :

2. GlobalRevendeur=Parrainage 227 ==> Canton=Vaud TypeEquipment=Adaptateur IP 106 conf:(0.47) < lift:(1.67)> lev:(0.04) [42] conv:(1.34)

Si l'on prend la règle dans le sens opposé, c'est-à-dire que lorsqu'un client est parrainé, c'est qu'il vient plutôt du canton de Vaud et qu'il achète plutôt un adaptateur IP. Cette règle ne nous donne pas plus d'information que la précédente car elle confirme toujours l'efficacité du parrainage. Il est intéressant de noter que l'indice de confiance est plus élevé ici (0.47).

4.2.3 Règle 3 :

4. Canton=Vaud GlobalRevendeur=Parrainage 157 ==> TypeEquipment=Adaptateur IP 106 conf:(0.68) < lift:(1.42)> lev:(0.03) [31] conv:(1.59)

Avec un lift plus bas mais un indice de confiance plus élevé que la première règle (0.68), nous avons ici une autre combinaison des trois attributs Canton, GlobalRevendeur et TypeEquipment, c'est-à-dire que si le client vient du canton de Vaud et qu'il s'est souscrit par parrainage c'est qu'il a acheté un Adaptateur IP.

4.2.4 Règle 4 :

6. Contrat=Particulier GlobalRevendeur=Parrainage 200 ==> TypeEquipment=Adaptateur IP 132 conf:(0.66) < lift:(1.39)> lev:(0.04) [37] conv:(1.52)

L'attribut Contrat apparaît dans cette règle et c'est pour cette raison qu'elle nous a intéressés. Si un client a pris un contrat de type Particulier et qu'il a été parrainé, il va plutôt acheter un adaptateur IP. Nous ne trouvons pas ici d'explication à ce phénomène car comment peut-on savoir si un client va choisir un adaptateur IP plutôt qu'un téléphone IP, mis à part le prix.

4.2.5 Règle 5 :

8. GlobalRevendeur=Parrainage 227 ==> Contrat=Particulier TypeEquipment=Adaptateur IP 132 conf:(0.58) < lift:(1.38)> lev:(0.04) [36] conv:(1.37)

Une autre combinaison des trois attributs GlobalRevendeur, Contrat et TypeEquipment nous permet de confirmer là encore l'efficacité du parrainage par cette règle qui dit que lorsqu'un client a été parrainé, il va plutôt choisir un contrat de type particulier et plutôt un adaptateur IP. Il n'y pas vraiment d'explication logique ou stratégique à ce phénomène mais il est intéressant de voir qu'il a un bon niveau de confiance (0.58)

4.2.6 Règle 6 :

10. GlobalRevendeur=Parrainage 227 ==> TypeEquipment=Adaptateur IP 146 conf:(0.64) < lift:(1.36)> lev:(0.04) [38] conv:(1.46)

Cette règle est assez surprenante car elle ne prend que deux attributs : Si un client a été parrainé, il va plutôt choisir un adaptateur IP. Nous savions déjà par nos précédentes techniques appliquées que l'adaptateur IP est mieux vendu que le téléphone IP mais ce qui nous a interpellé dans cette règle est que cela est plutôt lié au canal de distribution Parrainage qu'au canal de distribution site internet.

4.2.7 Règle 7 :

15. TypeEquipment=no equipment 244 ==> GlobalRevendeur=Aucun 227 conf:(0.93) < lift:(1.31)> lev:(0.05) [54] conv:(3.96)

Cette règle fait également intervenir que deux attributs mais ce qui est intéressant c'est que les clients qui ne prennent aucun équipement, c'est-à-dire qu'ils téléchargent le logiciel gratuitement, souscrivent via le cite web. On peut confirmer que le site web est le meilleur canal de distribution lorsqu'il s'agit de clients qui ne prennent aucun équipement, ce qui est assez logique car les clients ne vont pas se déplacer chez les revendeurs ou points de vente que pour faire un contrat mais plutôt pour acheter un appareil.

4.2.8 Règle 8:

21. Contrat=Particulier TypeEquipment=no equipment 177 ==> GlobalRevendeur=Aucun 161 conf:(0.91) < lift:(1.28)> lev:(0.04) [35] conv:(3.04)

Cette règle nous confirme que les clients qui prennent un contrat de type particulier et aucun appareil le font plutôt via le site web .Là encore aucune surprise. Ce qui est assez surprenant c'est l'indice de confiance qui est très élevé (0.91).

4.3 Commentaires :

- **Non adaptation de nos données par rapport à la technique des règles d'associations**

Précédemment, nous vous avons montré toutes les règles qui nous ont semblé les plus intéressantes et comme vous avez pu le constater ce sont des confirmations des règles trouvées avec la technique des arbres et de clustering. Nous essayons donc dans ce paragraphe de justifier, pourquoi nos données ne sont pas adaptées à la technique des règles d'associations.

1er point : On associe de manière générale les règles d'association au « caddie de la ménagère » dans un supermarché ou dans un E-shop. Or, ce qu'offre Switzernet ce sont des abonnements de téléphonie VoIP pour joindre les quatre coins de la planète à des prix réduits. Peut-on réellement dire que si un particulier s'abonne chez Switzernet et qu'il est du canton de Vaud, alors le type d'appareil qu'il va prendre sera un téléphone IP ? Switzernet peut elle conseiller, à un client qui s'est inscrit via son site Internet, de prendre un certain type d'abonnement et un certain type d'équipement parce que le voisin le plus proche de celui-ci a fait de même ? A notre avis les réponses à ces questions sont NON d'où le fait qu'aucune règle d'association intéressante ne nous est parue.

Caddie de la ménagère.

2ème point : Il semblerait aussi, que la technique des règles d'associations soit très sensible au fait que les groupes d'individus doivent être proportionnellement représentés (voir le début du cours). Car, comme on peut le constater dans les règles d'associations obtenues, le type de contrat « Particulier » est largement représenté et nous n'avons pas trouvé des règles associant les types de contrats « commercial » et « prepaid » et ceci s'explique par le fait que sur les 1000 itérations que nous avons obtenus de Switzernet 833 ont un type de contrat « Particulier » 133 ont un type de contrat « Commercial » et 34 ont un type de contrat « Prepaid ».

4.4 Conclusion :

Pour réellement voir ce que pourrait nous offrir les règles s'association, il nous faudrait des données avec des groupes beaucoup mieux proportionnés. Il nous faudrait donc plus de données, mais pour une question de temps nous allons rester sur cette insatisfaction et nous contenter des règles données par les techniques des arbres et de clustering.

Conclusion générale :

Dans le présent rapport, nous avons voulu analyser une base de données clients en appliquant différentes techniques de Data Mining vues en cours dans le but de les interpréter et de partager nos résultats avec l'entreprise qui nous l'a fournie.

La première phase a été la plus importante car une bonne préparation des données permet une bonne analyse et surtout permet à l'outil de bien utiliser les données. Dans cette phase, le plus difficile a été de choisir les attributs qui allaient nous servir pour la suite car il fallait avoir une bonne idée de ce que nous voulions obtenir comme résultat pour pouvoir juger de la pertinence des attributs. Côté technique, il faut noter que plus nous avons de données et plus c'est difficile de travailler avec la base de données car tout dépend comment cette dernière a été remplie. Nous avons rencontré différents problèmes d'uniformisation et de standardisation. Pour n'en citer que quelques uns, Il a fallu tout d'abord regarder toutes les valeurs des attributs qui pouvaient avoir différentes orthographes pour le même mot, ou encore différents formats pour les champs date ou code postal. Néanmoins, cette phase nous a permis de nous familiariser avec les données et avec l'outil que nous allions utiliser tout au long du projet.

La deuxième phase consistait à générer des arbres de décision en soumettant nos données à un algorithme de l'outil Weka. Les résultats que nous avons obtenus sont assez intéressants car ils nous ont permis de suggérer à Switzernet de cibler leur campagne publicitaire durant la troisième période de l'année et de faire des couplements d'offres pour les contrats de type commercial avec les différents appareils. Le deuxième résultat obtenu concerne plutôt les différents canaux de distribution avec le site Internet qui obtient la première place et il est donc indispensable de continuer à le maintenir voire à l'améliorer en essayant de mettre le maximum d'informations pour attirer les nouveaux clients. Le parrainage a également un bon retour mais il y a malheureusement un effort considérable à fournir du côté publicité. Les revendeurs offrent également un très bon potentiel qui, à notre humble avis, n'est pas assez exploité, il faudrait penser à mieux informer, voire à former des personnes qui connaîtrait bien les atouts de Switzernet pour pouvoir mieux les partager. Côté technique, il a fallu mieux catégoriser les attributs pour pouvoir obtenir des arbres avec des feuilles de plus en plus pure, ce qui ne donne pas toujours de bons résultats.

Dans la troisième phase, nous avons utilisé la technique du clustering. Nous avons pu confirmer notre hypothèse sur la tendance des signatures de contrat plutôt vers la troisième période. Nous avons également pu faire une analyse plus approfondie des canaux de distributions. Pour finir, nous avons pu faire une comparaison sur la présence des points de vente entre Genève et Vaud qui est clairement accès dans le canton de Vaud. Nous conseillons donc une acquisition de plus de points de vente à Genève.

Malheureusement, la quatrième phase où nous avons utilisé la technique de règle d'association n'a pas pu nous apporter de résultats pertinents et ceci est dû au manque de richesse de produits qu'offre Switzernet car elle ne propose que trois types de contrats et que deux types d'appareil. Nous n'avons donc pas pu apporter de nouveaux résultats ou améliorer les précédents.

Pour résumer, les arbres de décision nous ont permis d'apporter les premiers résultats qui nous permis de bien comprendre le comportement de certains clients de celui des canaux de distribution de l'entreprise. Le clustering nous a aidé à approfondir nos résultas avec plus de précision. Quant aux règles d'association, nous n'avons trouvé que des règles triviales et inexplicables mais aucune actionnables qui aurait pu nous aider à fournir plus de recommandations à l'entreprise.

Pour finir, nous dirons que ce projet nous a vraiment aidés à mettre en pratique les connaissances que nous avons acquises en cours car nous avons pu réaliser les obstacles techniques ou d'ordre de l'interprétation Business auxquels on peut être confronté lorsqu'on soumet une base de données à un outil. Nous avons également pu apprendre comment utiliser un outil simple de Data Mining mais il est clair que le travail le plus intéressant est de pouvoir interpréter les résultats et de pouvoir donner des recommandations à l'entreprise.

Nous tenons à remercier le professeur ainsi que son assistant pour l'aide précieuse qu'ils nous apportée et pour le temps qu'ils nous sont consacré. Nous tenons également à remercier chaleureusement la directrice de l'entreprise qui nous a fait confiance en nous fournissant la base de données clients et nous espérerons vraiment avoir pu apporter quelques éléments de réponses par rapport aux questions que nous nous sommes posées au départ.