TP DE DATA MINING
1 : COMPREHENSION
ET PREPARATION DES DONNEES
AVEC SPSS CLEMENTINE

EPF – 4/ 5ème année - Option Ingénierie d'Affaires et de Projets - Finance

Bertrand LIAUDET

TP N°1 DE DATA MINING : LES DONNEES

Le premier TP a pour objectif de se familiariser avec le logiciel SPSS Clementine et de mettre en œuvre les phases de compréhension et de préparation des données d'un processus de data mining.

Un mode d'emploi succinct de SPSS Clementine est proposé dans un autre document.

Jeu de données d'attrition (churn)

1 Ouvrir un flux

2 Travailler sur le fichier d'attrition (Source/Délimité)

Combien d'attributs y a-t-il dans ce fichier. (Onglet Données). Triez-les par type (Tri de l'onglet stockage).

Observez les caractéristiques des attributs (Onglet Type). Lire les valeurs. Que constatez-vous ?

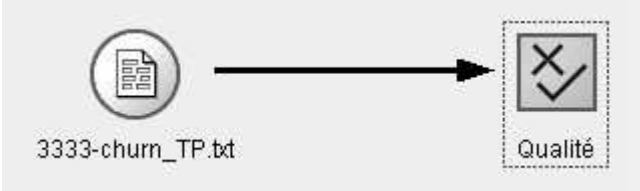
Forcez le type du n° de département : faites le passer de « intervalle » à « « ensemble ». Relire les valeurs.

Rappel : Le fichier de données d'origine doit être conservé intact.

Clementine

- Ouvrir un nouveau flux
- Ajouter le nœud « Source / Délimité » au flux.
- Double-cliquer sur ce nœud et associer à ce nœud le fichier texte à manipuler.
- **Avoir de l'information sur le type des données :** Sélectionner « Source / Délimité ». Double-cliquer. Choisir l'onglet « données » et regarder la colonne « stockage » : Clementine distingue entre « entier », « réel » « chaîne », et « date »
- **Avoir de l'information sur les valeurs possibles des données :** Sélectionner « Source / Délimité ». Double-cliquer. Choisir l'onglet « types », cliquer « effacer toutes les valeurs » puis « lire les valeurs » et regarder les colonnes « type » et « valeurs ». Clementine définit des types « intervalle », « booléens » et « ensemble », mais aussi « sans type ».
- **Forcer le type du n° de département :** le faire passer de « intervalle » à « ensemble ». Ainsi, on aura bien une variable catégorielle.

3 Observer la qualité des données (Sortie/Qualité)



Afficher le maximum de caractéristiques. Trier les résultats. Quels attributs posent problème ?

Clementine

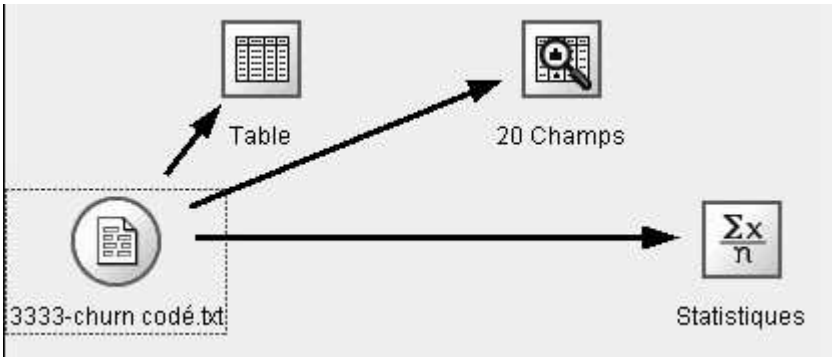
- Ajouter le nœud « Sortie / Qualité » au flux.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Qualité » .
- Double-cliquer sur le nœud « Sortie / Qualité ». Tout cocher.
- Exécuter.

4 Afficher le tableau des données (Sortie / Tableau)

Afficher le tableau des données.

Essayer de faire des tris. Que pouvez vous constater ?

Que pensez-vous de cet outil ?



Clementine

- Ajouter le nœud « Sortie / Table » au flux.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Table » (Sélectionner le nœud « Source / Délimité », bouton droit, connecter, cliquer sur « Sortie / Table »)
- **Afficher le tableau des données :** Sélectionner « Sortie / Table », bouton droit, exécuter.

5 Afficher l'audit de données (Sortie – Audit)

Combien d'attributs sont traités par l'audit ? Pourquoi ?

Trier les données par type.

Quelle est la durée de vie moyenne des contrats ?

Que signifie la moyenne des codes département ?

Expliquer la signification de chaque variable.

Trier les données par asymétrie. Que constatez-vous ?

Quelles conclusions pouvez-vous tirer de l'observation des histogrammes ?

Clementine

- Ajouter le nœud « Sortie / Audit données » au flux. L'audit de données traite les données continues et les données discrètes.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Audit données » (Sélectionner le nœud « Source / Délimité », bouton droit, connecter, cliquer sur « Audit données »)
- **Afficher l'Audit des données :** Sélectionner « Sortie / Audit données », bouton droit, exécuter.

RESULTATS

Les colonnes « Graphique », « Type », « Unique » sont double-cliquables pour obtenir des précisions.

Toutes les colonnes permettent de faire des tris en cliquant sur le nom de la colonne.

6 Statistiques détaillées (Sortie / Statistiques)

Afficher et analyser les statistiques détaillées.

Clementine

- Ajouter le nœud « Sortie / Statistiques » au flux. Les statistiques ne s'appliquent qu'aux données continues.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Statistiques » (Sélectionner le nœud « Source / Délimité », bouton droit, connecter, cliquer sur « Audit données »)
- Double-cliquer sur le nœud « Sortie / Statistiques », sélectionner les variables à examiner et les statistiques souhaitées (sélectionner les toutes).
- **Afficher le tableau des statistiques :** Sélectionner « Sortie / Statistiques», bouton droit, exécuter.

RESULTATS

Les calculs statistiques sont effectués sur 16 variables : les 16 variables continues. Les 4 variables discrètes n'ont pas été prises en compte. La variable numTél a déjà été éliminée.

Aide :

En cliquant sur le « ? » en haut à droite de la fenêtre, Clementine fournit une aide pour l'interprétation et le paramétrage du résultat.

7 Filtrer toutes les données hors norme (Sortie / Statistiques)

On a mis au jour des données hors norme.

Filtrer ces données et recommencer l'audit de données.

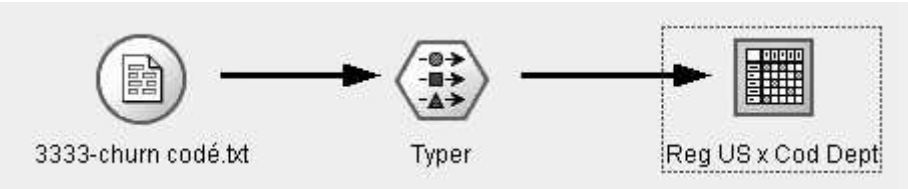
Clementine

- On peut produire un nœud « Sélection » à partir du graphique de l'histogramme ou des proportions.

9 Afficher la matrice croisée « département », « Reg US » (Sortie / Matrice)

Le n° de département est une variable catégorielle.

Quelles conclusion peut-on en tirer ?



A noter que le nœud « typer » est inutile car le type est déjà imposé dans le nœud « Source/Délimité ».

Faire les statistiques du numéro de téléphone :

- Ajouter le nœuds « Ops champs / Typer » au flux, « Graphique / proportions ».
- Connecter les nœuds « Source / Délimité », « Ops champs / Typer » et « Graphique / proportions »/
- Double-cliquer sur le nœud « Ops champs / Typer ». Passer le type de « numTel » à « ensemble » et la « direction » à « in » : constater qu'il y a maintenant 21 champs pour le nœud statistique.
- **Afficher le tableau de proportions :** trier par comptage.

10 Donner le nombre d'occurrences de chaque numéro de téléphone (Graphique / Proportion) (Champs / Typer)

On utilisera l'outil : de proportion et le nœud « typer ».

Quels sont les nombres d'occurrences min et max. Quelle conclusion peut-on en tirer ?

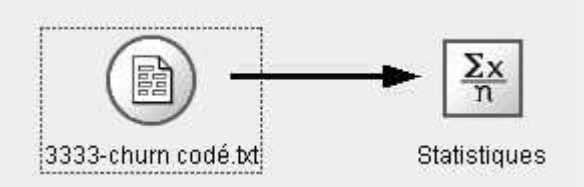
Faire les statistiques du numéro de téléphone :

- Ajouter le nœuds « Ops champs / Typer » au flux, « Graphique / proportions ».
- Connecter les nœuds « Source / Délimité », « Ops champs / Typer » et « Graphique / proportions »/
- Double-cliquer sur le nœud « Ops champs / Typer ». Passer le type de « numTel » à « ensemble » et la « direction » à « in » : constater qu'il y a maintenant 21 champs pour le nœud statistique.
- **Afficher le tableau de proportions :** trier par comptage.

11 Rechercher toutes les corrélations possibles entre les variables numériques (Sortie / Statistiques)

Quelles corrélations trouvez-vous ?

Montrez graphiquement les corrélations avec un nuage de point.



12 Filtrer le tableau de départ et faire un audit

Dans un nouveau flux, filtrer les attributs inutiles (on utilisera le nœud « typer »), filtrer les données hors norme et faire un audit.

13 Corrélation entre le Churn et l'option internationale (Graphique / Proportion)

Dans une nouvelle feuille de flux, afficher les proportions de churn selon la valeur de l'option internationale avec un graphique de proportion.

Quelles conclusions pouvez-vous en tirer ?

Clementine

- Ajouter le nœud « Graphique / Proportion » au flux.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Audit données ».
- Double-cliquer que le nœud « Graphique / Proportion ». Choisir le champs « international », superposer « churn », normaliser par couleur.
- Exécuter.

13 Corrélation entre le Churn et l'option messagerie (Graphique / Proportion)

Dans une nouvelle feuille de flux, afficher les proportions de churn selon la valeur de l'option messagerie avec un graphique de proportion.

Quelles conclusions pouvez-vous en tirer ?

13 Calcul du nombre de clients « récupérables » (Sortie / Matrice)

Afficher les données chiffrées des 2 analyses précédentes avec le nœud : « Sortie/Matrice ».

En ramenant le taux de churn de ceux qui ont pris l'option internationale au taux de churn de ceux qui ne l'ont pas pris, combien de clients pourrait-on conserver ?

En ramenant le taux de churn de ceux qui n'ont pas pris l'option mail au taux de churn de ceux qui l'ont pris, combien de clients pourrait-on conserver ?

Clementine

- Ajouter le nœud « Sortie / Matrice » au flux.
- Connecter le nœud « Source / Délimité » au nœud « Sortie / Matrice ».
- Double-cliquer que le nœud « Sortie / Matrice ». Choisir le champs « international » pour ligne, et le champs « churn » pour colonne.
- Exécuter. Dans « apparence », cliquer sur « inclure les totaux des lignes et des colonnes ».

13 Corrélation entre le Churn et le nombre d'appels au service client

Dans une nouvelle feuille de flux, afficher les proportions de churn selon le nombre d'appels au service client avec un graphique de proportion.

Quelles conclusions pouvez-vous en tirer ?

14 Corrélation du Churn et de la consommation de jour

Dans une nouvelle feuille de flux, afficher les proportions de churn selon la consommation par jour avec un histogramme.

Quelles conclusions pouvez-vous en tirer ?

15 Corrélation du Churn avec la consommation de jour et les appels au service client

Superposer le churn dans le nuage consommation jour et appels service client.

Quelles sont les zones remarquables ?

Quelles conclusions pouvez-vous tirer ?

16 Faites une synthèse de toute votre analyse.

Jeu de données de voitures

Traiter le problème de la compréhension et de la préparation des données.

Quelles sont vos conclusions. ?

Dictionnaire des variables

N°	Nom de la variable	Signification de la variable	Type de variable : Catégorielle ou autre	Autres caractéristiques
1	N°	Numéro du véhicule		
2	Consommation en mile par galon	Consommation en mile par gallon (4,546 litres = 1 gallon. 1 mile = 1609 mètres. 10 litres au 100 km = 1 gallon au 28,25 miles, soit 28,25 miles par gallon).		
3	Poids	Poids du véhicule		
4	Cylindres	Nombre de cylindres du moteur		
5	cm3	Volume du moteur		
6	Hp	Puissance du moteur		
7	Time-to-60	Temps pour atteindre les 60 miles /s		
8	Année	Année de sortie d'usine		
9	Origine	Origine du constructeur		

Autres jeux de données

Les employés et les départements pour refaire un peu de SQL !

Affichez le tableau des données pour Emp et Dept

Faites la jointure entre emp et dept. Affichez les résultats triés par nom et prénom d'employé en présentant les attributs dans le bon ordre.

Afficher le nombre d'employés et le salaire moyen par département.

En pratique :

Soit les deux tables suivantes :

Emp (numEmp, nom, fonction, suphié, date, salaire, prime, numDept)

Dept (numDept, nom, ville)

Clementine (exemple 01 – Emp-Dept)

Ouvrir un nouveau flux

Ajouter deux nœuds « Source / Délimité » au flux.

Double-cliquer sur les nœuds et associer le fichier texte à manipuler.

Ajouter le noeud « Ops sur lignes / fusionner ».

Connecter les nœuds « Source / Délimité » au nœud « Ops sur lignes / fusionner ».

Double-cliquer sur « Ops sur lignes / fusionner », sélectionner « clés » comme méthode de fusion, et déplacer la clé possible en clé pour fusion.

Ajouter le noeud « Sortie / table ».

Connecter le nœud « Ops sur lignes / fusionner » au nœud « Sortie / table ».

Les céréales

Traiter le problème de la compréhension et de la préparation des données.

Quelles sont vos conclusions. ?