1 Objectif

Etudier l'influence des différentes techniques de traitement des données manquantes sur le comportement de la régression logistique.

L'appréhension des données manquantes est un problème difficile. Non pas à cause de sa gestion informatique qui est relativement simple, il suffit de signaler les valeurs manquantes par un code spécifique, mais plutôt à cause des conséquences de leur traitement (suppression des lignes ou des colonnes du fichier ; ou remplacement par une valeur calculée à partir de observations disponibles, on parle alors d'imputation) sur les caractéristiques des modèles élaborés.

Nous en avions parlé dans un précédent document. Il s'agissait alors d'étudier l'impact des différentes techniques de traitement de valeurs manquantes sur les arbres de décision construits avec la méthode C4.5 (Quinlan, 1993)¹ dans le logiciel SIPINA. Aujourd'hui, nous réitérons l'analyse en étudiant leur influence sur les résultats de la régression logistique. Nous utiliserons principalement le logiciel **R**, avec la procédure **glm(.)**. Par la suite, nous examinerons le comportement des outils proposés dans des logiciels tels qu'Orange, Knime et RapidMiner placés dans un contexte identique.

Notons, concernant les logiciels, que Tanagra ne gère pas les valeurs manquantes. C'est un choix délibéré. Je ne souhaitais pas que les étudiants, qui sont le principal public de Tanagra, évacuent d'un clic la gestion de données manquantes parce qu'un traitement par défaut est proposé. Réfléchir en amont au problème et proposer une technique en adéquation avec les caractéristiques de l'étude que l'on mène me paraît pédagogiquement incontournable.

Justement, de quelle nature sont ces caractéristiques qui pèsent sur le traitement des valeurs manquantes ? Tout d'abord, il y a **le processus à l'origine de leur formation**². Il peut être complètement aléatoire (**MCAR** : missing completely at random). Dans ce cas, la survenance d'une observation manquante sur une variable ne dépend ni des valeurs de la variable, ni des valeurs d'autres variables de la base. Il peut être aléatoire (**MAR** : missing at random). Dans ce cas, sa survenance ne dépend pas des valeurs de la variable après avoir contrôlé le rôle des autres variables de la base. Par exemple, les cadres sont moins enclins à déclarer leur salaire dans une enquête. Mais dans cette catégorie de personnes, la probabilité d'apparition de valeur manquante ne dépend pas de la valeur du salaire. Il peut être non aléatoire (**MNAR** : missing not at random). Par exemple, les personnes qui ont un salaire élevé préfèrent éluder la question lors d'une enquête. Le problème est alors particulièrement difficile. Il faut passer par un traitement spécifique.

De toutes ces configurations, le cas MCAR est le plus facile à gérer. Nous obtenons des estimations non biaisées des paramètres, même lorsque nous utilisons des techniques très frustes comme la suppression des observations comportant au moins une valeur manquante.

¹ <u>http://tutoriels-data-mining.blogspot.com/2009/10/sipina-traitement-des-donnees.html</u>

² http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

L'analyse statistique appiquée sur les données est une autre pièce importante du puzzle. Clairement, l'influence des valeurs manquantes et de leur traitement n'est pas la même dans une analyse prédictive (la régression logistique par exemple), dans un clustering (ex. une classification ascendante hiérarchique) ou dans une analyse factorielle (ex. une analyse en composante principale). La distinction est d'autant plus importante que certaines méthodes intègrent nativement un procédé d'appréhension des valeurs manquantes (ex. l'algorithme NIPALS³ pour l'analyse en composantes principales et la régression PLS).

Enfin, le critère d'évaluation du traitement des données manquantes est la dernière pièce du puzzle. L'objectif n'est pas de retrouver les hypothétiques « vraies » valeurs des observations nonrenseignées pour chaque variable. Mais plutôt de proposer des valeurs de remplacement (dans le cas de l'imputation) qui ne dénaturent pas les résultats de l'étude que nous sommes en train de mener. On met souvent en avant le biais et la variance des paramètres estimés des modèles en statistique. Mais dans le cadre spécifique du data mining prédictif, on peut aussi légitimement se poser la question de la performance. Est-ce que la technique de traitement des données manquante permet de produire le modèle le plus efficace possible en prédiction ?

Dans ce tutoriel, nous nous plaçons dans la configuration suivante : (1) les valeurs manquantes sont MCAR, nous avons écrit un programme qui retire de manière complètement aléatoire les valeurs dans l'échantillon d'apprentissage ; (2) nous appliquons la régression logistique sur les données d'apprentissage post-traitées ; (3) nous évaluons les différentes techniques de traitement des données manquantes en observant le taux de bon classement (ou taux de succès) du modèle sur un échantillon test à part qui, lui, ne comporte aucune valeur manquante.

Dans un premier temps, nous mènerons l'expérimentation avec <u>R</u>. Nous opposerons la suppression de ligne (listwise deletion) à l'imputation univariée (remplacement par la moyenne pour les variables quantitatives, par le mode pour les qualitatives). Nous constaterons que cette dernière est une approche très viable dans le cadre MCAR. Dans un deuxième temps, nous étudierons les options proposées dans les logiciels <u>Orange</u>, <u>Knime</u> et <u>RapidMiner</u>. Nous constaterons que malgré leur sophistication, ils ne font pas mieux que l'imputation univariée. Mais rappelons-le encore une fois, nous nous plaçons dans un cadre très particulier (MCAR + régression logistique + taux d'erreur en test). D'autres études seraient le bienvenu (*avis aux amateurs*) dans des configurations différentes.

2 Données

Données disponibles. Nous utilisons les données GERMAN CREDIT DATA dans notre expérimentation⁴. Nous cherchons à prédire la classe d'un individu (bon ou mauvais client) lors d'une demande de crédit. Nous avons scindé aléatoirement le fichier en deux feuilles dans un classeur Excel (**credit-german-md-simulation.xls**) : CREDIT-GERMAN-TRAIN-FULL, réservé à l'apprentissage, comporte 300 observations ; CREDIT-GERMAN-TEST, réservé au test, 700 observations. Pour ce

³ http://en.wikipedia.org/wiki/Non-linear_iterative_partial_least_squares

⁴ http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

second échantillon, nous avons créé une fois pour toutes un fichier au format texte avec séparateur tabulation CREDIT-GERMAN-TEST.TXT.

Il n'y a aucune valeur manquante dans ces deux échantillons.

	credit-german-md-simulation.xls [Mode de compatibilité] - Microsoft Excel													
	6	Accueil	Insertion	Mise en p	age	Formule	es Données	Révision	Affichag	e	Développeur	Complémen	ts 🕜 – 🖻	X
	Vis Ba	ual Macros	Enregistrer un Utiliser les réfi Sécurité des m	ne macro érences rela nacros	tives	Insérer	Mode réation	iétés liser le code ter la boîte de c	lialogue	Sour	Ce Propriét	és du mappage tension er les données	Importer	
۲			Code	0			Contro	nes				AIVIL		
		L27	- (°	Ĵx	car									×
	4	D	E	F		G	Н		J		K	L	M	
	1	purpose	credit_amoun	savings_s	tatu e	mploymen	t_installment_co	personal_stat	other_pa	rties	residence_sir	property_mag	age	ot
	2	radio/tv	5943	no known s	avin <	1	1	female div/dep/r	none		1	car	44	no
	3	furniture/equipm	3650	<100	<	1	1	female div/dep/r	none		4	car	22	2 no
1	4	furniture/equipm	2069	no known s	avin 1-	<=X<4	2	male mar/wid	none		1	car	26	6 no
	5	business	1568	100<=X<50	0 1.	<=X<4	3	female div/dep/r	none		4	life insurance	24	no
	6	radio/tv	1533	<100	<	1	4	female div/dep/r	none		3	car	38	3 sti
	7	business	1941	>=1000	1.	<=X<4	4	male single	none		2	life insurance	35	5 no
	8	radio/tv	1364	<100	4-	<=X<7	3	male single	none		4	real estate	59) no
	9	used car	2993	<100	1.	<=X<4	3	male single	none		2	real estate	28	8 sti
1	10	new car	5045	no known s	avin >:	=7	1	female div/dep/r	none		4	car	59) no 🔻
ŀ		🕨 🕨 🤇 credit	-german-md	🦯 credit-	germa	in-train-full	credit-gern	nan-test 🦯 🐮	14				1	
F	Prêt	: 2			1							100 % 🕞 —		+ .::

Génération d'un échantillon d'apprentissage comportant des valeurs manquantes. Nous utilisons une macro VBA (visual basic pour applications) pour générer un fichier d'apprentissage intégrant des valeurs manquantes dans la première feuille du classeur (CREDIT-GERMAN-MD)⁵. Pour ce faire, nous actionnons le menu DEVELOPPEUR / MACRO d'EXCEL⁶ et dans la boîte de paramétrage qui apparaît, nous sélectionnons la macro STARTDATAGENERATOR.

	(≌ •) ∓ credit	erman me simula	tion.xls [Mode de compatibilité] - Microsoft Excel	
Accueil	Lisertion Mise en page	Formules	Macri Renzel	- = ×
Visual Macros	 Enregistrer une macro Utiliser les références relatives Sécurité des macros Code 	Insérer Mode • Création	Nom de la macro : StartDataGenerator StartDataGenerator Exécuter restricted a pas détaille	orter orter
Afficher I Affiche 1 pur 2 radic 3 furn	es macros (Alt+F8) la liste des macros que vous pou r, créer ou supprimer. uyez sur F1 pour obtenir de l'aid	e.	Modifier Créer Supprimer	• • • • • • • • • • • • • • • • • • •
4 furniture/equipr	r 2069 no known savin 1568 100<=X<500	1<=X<4 1<=X<4		26 no 24 no
6 radio/tv 7 business	1533 <100 1941 >=1000	<1 1<=X<4	Macros dans : Tous les classeurs ouverts	38 str 35 no
9 used car 10 new car	2993 <100 5045 no known savin	4<=X <br 1<=X<4 >=7		28 str 59 no 🔻
II I I I I Cred Prêt 200	it-german-md 🖉 credit-gern	nan-train-full 📜 c	Annuler	→ I

Le programme est démarré, une seconde boîte de paramétrage, propre à la macro, permet de spécifier la proportion de valeurs manquantes qu'il faut générer, en pourcentage des valeurs disponibles. Par exemple, si nous mettons 0.05 (5 %), le programme créera un fichier d'apprentissage

⁵ Attention, il faut veiller à ce que les macros soient activées lorsque vous importez le fichier dans Excel !

⁶ Pour Excel 2003 et antérieures, il faudrait chercher dans le menu OUTILS / MACROS / MACROS.

avec 0.05 * (300 lignes * 20 variables) = 300 cellules **avec l'indicateur « ? » pour signaler l'absence de valeur**. La colonne de la variable à prédire CLASSE n'est pas concernée par cette opération.

Missing values generator	×
Missing values proportion : .05 START Cancel	

Les données résultantes sont intégrées dans la première feuille du classeur (CREDIT-GERMAN-MD).

0		€	crea	dit-german-md.	txt en a de c	ompatibilité] -	Microsoft Excel)		, D X
ſ	Accueil	Insertion	Mise en page	Formules	Données	Révision Af	fichage Dév	veloppeur C	ompléments	0 - 🗖 X
Pre	Coller 🛷	Bitstream Ver		= = = = = = = = = = = = ∞ • t = ∞••	Standard	Style	Insérer ▼ Supprimer ▼ Format ▼ Cellules	Σ × A ▼ Z Trier e Z × filtrer Éc	t Rechercher et sélectionner •	
	C3	- (*	f _x ?							*
	А	В	С	D	E	F	G	Н		J
1	checking_stat	duration	credit_history	purpose	credit_amoun	savings_statu	employment	installment_c	personal_stat	other_partie
2	<0	12	critical/other exi	i new car	3499	<100	1<=X<4	3	female div/dep/r	co applicant
3	<0	6	?	furniture/equipm	1872	<100	?	4	male single	none
4	<0	6	critical/other exi	new car	1361	<100	<1	2	male single	none
5	no checking	15	existing paid	used car	3029	<100	4<=X<7	2	male single	none
6	>=200	18	all paid	radio/tv	1445	no known savin	4<=X<7	4	male single	none
7	no checking	18	all paid	new car	6458	<100	>=7	2	male single	none
8	no checking	48	critical/other exi	business	7629	no known savin	>=7	4	male div/sep	none
9	0<=X<200	30	critical/other exi	new car	4249	<100	unemployed	4	male mar/wid	none
10	no checking	18	delayed previou	business	2169	<100	1<=X<4	4	male mar/wid	none
11	no checking	18	critical/other exi	used car	?	no known savin	unemployed	2	male single	none
12	no checking	?	all paid	used car	7485	no known savin	unemployed	4	female div/dep/r	none
13	>=200	10	?	radio/tv	1347	no known savin	4<=X<7	4	?	none
14	no checking	12	critical/other exi	new car	926	<100	unemployed	1	female div/dep/r	none
15	no checking	24	existing paid	new car	1249	<100	<1	4	male mar/wid	none
16	0<=X<200	13	existing paid	radio/tv	2101	<100	<1	2	female div/dep/r	guarantor
17	<0	18	critical/other exi	new car	5302	<100	>=7	2	male single	none 👻
14	Credit	-german-md	credit-gerr	man-train-full	🦯 credit-germ	nan-test 📈 🛙 🕨				▶ 1
Prê	t 🛅							100 %	. 🖃 — – 🗘	(+) .:i

Dans le même temps, le fichier d'apprentissage au format texte « CREDIT-GERMAN-MD.TXT » est automatiquement créé dans notre répertoire de travail (le répertoire du fichier XLS).

00				- • ×
Comparison of the state of t	_orange 🕨 german	▼ + ₇	Rechercher dans : german	ىر
Organiser 🔻 Inclure dans la bibliothèque 🔻	tager avec 🔻 Diaporama Graver	Nouveau dossier		▼ □ ②
🔶 Favoris	^ Nom	Modifié le	Туре	Taille
💻 Bureau	🖉 credit_imputation.r	29/11/2011 11:06	Tinn-R	3 Ko
📃 Emplacements récents	credit-german-md.ows	30/11/2011 07:07	Fichier OWS	21 Ko
〕 Téléchargements	credit-german-md.txt	30/11/2011 22:46	Document texte	40 Ko
	credit-german-md-0_00.txt	29/11/2011 10:36	Document texte	41 Ko
🧮 Bureau	credit-german-md-0_50.txt	29/11/2011 10:37	Document texte	41 Ko
🥽 Bibliothèques	📄 credit-german-md-1_00.txt	29/11/2011 10:38	Document texte	41 Ko
Documents	credit-german-md-2_00.txt	29/11/2011 10:38	Document texte	40 Ko
📔 Images	📄 credit-german-md-5_00.txt	29/11/2011 10:39	Document texte	40 Ko
👌 Musique	credit-german-md-10_00.txt	29/11/2011 10:39	Document texte	38 Ko
📑 Vidéos	📄 credit-german-md-20_00.txt	29/11/2011 11:06	Document texte	36 Ko
🜏 Groupe résidentiel	🕲 credit-german-md-simulation	.xls 03/11/2011 15:14	Feuille Microsoft	357 Ko
🔒 Maison	📋 credit-german-test.txt	03/11/2011 15:14	Document texte	95 Ko
🖳 Ordinateur	尾 process - rapidminer.png	30/11/2011 19:39	Image PNG	53 Ko
📬 Réseau	results_experiments.xls	30/11/2011 18:58	Feuille Microsoft	9 Ko
📴 Panneau de configuration	-			
14 élément(s)				

Pour les férus de programmation, voici le code VBA associé au générateur de données.

```
Private Sub cmdBOk Click()
'retrieve the proportion from a dialogbox
Dim proportion As Double
proportion = Val(tbProportion.Text)
'checking the range of the value (< 0.5 max.)
If (proportion < 0\#) Or (proportion >= 0.5) Then
   proportion = 0.01
End If
'copying the full training set to the first sheet
Sheets("credit-german-train-full").Select
Range("A2:U301").Select
Selection.Copy
Sheets("credit-german-md").Select
Range("A2").Select
ActiveSheet.Paste
'number of cells to clear
Dim nbMd As Long
nbMd = Int(6000# * proportion)
'counter for the number of missing values
Dim counter As Long
counter = 1
'coordinates
Dim i As Long, j As Long
'initialize the random number generator
Randomize (100)
Do While (counter <= nbMd)
   'row
    i = 2 + Int(300 \# * Rnd)
    'column
    j = 1 + Int(20\# * Rnd)
    'checking if the cell is already empty
```

```
If (Cells(i, j).Value <> "?") Then
        Cells(i, j).Value = "?"
        counter = counter + 1
    End If
Loop
'set the text file name
Dim nomFichier As String
nomFichier = "\credit-german-md.txt"
nomFichier = ThisWorkbook.Path + nomFichier
'checking if the file already exists
Dim fs As Variant
Set fs = CreateObject("Scripting.FileSystemObject")
If (fs.FileExists(nomFichier) = True) Then
    fs.deletefile (nomFichier)
End If
'save the sheet into a text file
ThisWorkbook.SaveAs fileName:=nomFichier, FileFormat:=xlTextWindows
'close the form (dialogbox for the proportion setting)
formMD.Hide
End Sub
```

Notons une information très importante : nous avons inséré 300 valeurs manquantes dans notre ensemble d'apprentissage. Cela ne veut pas dire pour autant qu'il y a 300 observations avec valeurs manquantes. En effet, dans certains cas, certaines observations peuvent comporter deux, voire plus, cellules vides (ex. la ligne n°13 dans notre copie d'écran ci-dessus). Dans les traitements qui suivront, nous compterons les lignes complètes, c.-à-d. ne comportant aucune valeur manquante, dans notre base d'apprentissage.

3 Traitement des valeurs manquantes avec R

3.1 Architecture du script

Nous souhaitons opposer deux techniques de traitement des données manquantes : la suppression d'observations comportant au moins une valeur manquante (listwise deletion) et l'imputation univariée (remplacement par la moyenne pour les variables quantitatives, par le mode pour les qualitatives). Pour évaluer la qualité de la stratégie, nous appliquons les modèles élaborés sur le même échantillon test exempt de valeurs manquantes⁷. La meilleure stratégie est celle qui induit le modèle le plus performant en prédiction. Nous nous basons sur le taux de succès.

Décomposons notre programme R pour comprendre notre démarche.

Chargement des données. Dans un premier temps, nous chargeons l'échantillon d'apprentissage (comportant éventuellement des valeurs manquantes, sauf sur la variable à prédire) et de test (exempt de valeurs manquantes).

⁷ L'application d'un modèle sur des individus partiellement décrits est un autre type de problème qui n'entre pas dans le cadre de notre étude. Il n'en est pas moins intéressant. Peut être le sujet d'un futur tutoriel ?

```
#chargement du fichier test
setwd("... votre répertoire ...")
data.test <- read.table(file="credit-german-test.txt",dec=".",sep="\t",header=T)
#chargement des données d'apprentissage avec les valeurs manquantes
data.train <- read.table(file="credit-german-md.txt",dec=".",sep="\t",header=T,na.strings="?")
summary(data.train)</pre>
```

Suppression des lignes. La première stratégie consiste à supprimer du fichier les lignes comportant au moins une valeur manquante. Elle est un peu brutale, on peut vider très rapidement le fichier avec ça. Mais elle a le mérite de produire des estimations non biaisées des paramètres de la régression dans le contexte MCAR.

```
#première solution : listwise deletion
data.train.omitted <- na.omit(data.train)
summary(data.train.omitted)
print(paste('Remaining observations : ',as.character(nrow(data.train.omitted))))
model.omitted <- glm(classe ~ ., family = binomial, data = data.train.omitted)</pre>
```

na.omit(.) supprime de la base les observations présentant au moins une valeur manquante. Nous profitons de ce traitement pour afficher le nombre d'observations complètes dans le fichier d'apprentissage. **model.omitted** est le modèle obtenu à l'issue de ce pré-traitement.

Remplacement univarié. Pour l'imputation univariée, nous utilisons le programme suivant pour générer le modèle prédictif **model.imputed**.

```
#seconde solution : imputation univariée (moyenne, mode)
data.train.imputed <- as.data.frame(lapply(data.train,traiter_missing_data))
summary(data.train.imputed)
model.imputed <- glm(classe ~ ., family = binomial, data = data.train.imputed)</pre>
```

La fonction **traiter_missing_data(x)** fait la distinction entre les variables quantitatives et qualitatives (le type factor de R).

```
#traiter les données manquantes
traiter_missing_data <- function(x) {
    if (is.factor(x) == T) {
        return(traiter.discrete(x))
    } else {
        return(traiter.numeric(x))
    }
}</pre>
```

Ainsi, pour le remplacement par la moyenne, nous utiliserons la fonction :

```
#variable numérique, remplacement des NA par la moyenne
traiter.numeric <- function(x) {
    y <- x
    z <- na.omit(y)
    if (length(z) < length(y)) {
        m <- mean(z)
        y[is.na(y)] <- m
    }
}</pre>
```

return(y)

L'idée est simple. Nous calculons la moyenne sur les données disponibles. Puis, nous remplaçons les valeurs manquantes par cette moyenne.

De la même manière, nous utilisons le mode pour les variables qualitatives.

```
#variable catégorielle, remplacement des NA par le mode
traiter.discrete <- function(x) {
    y <- x
    z <- na.omit(y)
    if (length(z) < length(y)) {
        frequence <- table(z)
        m <- which.max(frequence)
        y[is.na(y)] <- levels(x)[m]
    }
    return(y)
}</pre>
```

Evaluation des modèles. Dernière étape de notre expérimentation, nous devons évaluer les modèles. Nous utilisons une fonction spécifique pour cela, elle prend en entrée l'échantillon test et le modèle. Elle se charge d'effectuer la prédiction, de construire la matrice de confusion et de calculer le taux de bon classement (= 1 - taux d'erreur)⁸.

```
#construction matrice de confusion à partir du modèle
#new.dataset représente l'échantillon test
#model est le modèle
pred_and_confusion_matrix <- function(new.dataset,model){
    #probabilité prédité par le modèle
    data.pred.prob <- predict(model,newdata = new.dataset)
    #affectation
    data.pred.class <- ifelse(data.pred.prob > 0.5,"B","A")
    data.pred.class <- as.factor(data.pred.class)
    #matrice de confusion (classe observée vs. prédite)
    mc <- table(new.dataset$classe,data.pred.class)
    print(mc)
    #taux de bon classement
    ca <- (mc[1,1]+mc[2,2])/sum(mc)
    print(ca)
}
```

3.2 Organisation de l'expérimentation

L'idée est d'observer l'impact de la stratégie de traitement à mesure que la proportion de valeurs manquantes augmente. Nous essaierons les pourcentages suivants : **0.5%**, **1%**, **2%**, **5%**, **10%**, **20%**.

⁸ Nous utilisons le taux de bon classement (ou taux de succès) et non pas le taux d'erreur pour être en adéquation avec les valeurs fournies par Orange dans la suite de ce tutoriel.

Le résultat de référence est la performance du modèle lorsque l'échantillon d'apprentissage est complet (0% de données manquantes). Bien évidemment, dans ce cas, les deux stratégies devraient fournir le même résultat puisqu'aucun traitement n'est effectué.

L'exécution du programme fournit les sorties suivantes.



300 observations sont disponibles après suppression des lignes comportant des valeurs manquantes dans la première stratégie. C'est tout à fait normal puisqu'il n'y en a pas (de valeurs manquantes). Le taux de succès en test est de 73%. Nonobstant les fluctuations d'échantillonnage, nous ne devrions pas faire mieux lorsque nous supprimerons des valeurs dans le fichier d'apprentissage.

3.3 Résultats de l'expérimentation

Les résultats sont recensés dans le tableau suivant.

GERMAN D	ATASET	Accı	uracy rate
% missing	# complete obs.	Listwise Del.	Univ. Imputation
0,00%	300	0,7300	0,7300
0,50%	272	0,7100	0,7257
1,00%	246	0,7129	0,7286
2,00%	201	0,7214	0,7186
5,00%	111	0,6729	0,7114
10,00%	40	ERR	0,7086
20,00%	4	ERR	0,7214

Plusieurs informations attirent notre attention :

- La présence de valeurs manquantes dégrade le modèle. Le contraire aurait été étonnant.
- Tant que la proportion des valeurs manquantes reste faible, jusqu'à 2%, ce qui correspond à 201 observations complètes sur 300, les deux stratégies se valent.
- Lorsque cette proportion augmente (à partir de 5%), la « listwise deletion » dégrade fortement l'apprentissage au point de la rendre impossible pour 10% de valeurs manquantes. En effet, il ne reste que 40 observations après suppression des lignes. R n'a pas voulu lancer la régression.
- A contrario, l'imputation univariée se maintient remarquablement bien. Dans la configuration MCAR, la stratégie tient la route même lorsqu'il y a 20% de valeurs manquantes dans l'échantillon d'apprentissage. A la réflexion ce n'est pas très étonnant. Les « ? » étant bien répartis dans l'ensemble des colonnes, il y a finalement peu d'observations manquantes pour chaque variable. Les estimations des valeurs de remplacement sont viables. Un traitement rapide sous Excel nous a permis de constater qu'il y a en moyenne 240 observations disponibles dans chaque colonne.

Encore une fois, nous nous plaçons dans un contexte « de laboratoire » très singulier dans ce tutoriel. Notre principal mérite est d'avoir, autant que faire se peut, délimité le champ de validité de notre expérimentation. Des études complémentaires sont nécessaires pour évaluer le comportement des différentes approches dans d'autres contextes. A ce sujet, je conseille la lecture des excellents ouvrages d'Allison (2001) et de Little & Rubin (2002) référencés en bibliographie.

4 Traitement des données manquantes dans d'autres logiciels

Pour être honnête, l'idée de ce tutoriel m'est venue en découvrant l'outil très élaboré de gestion de données manquantes dans le logiciel Orange. J'ai un peu regardé ce qu'il en était dans les autres logiciels puis, de fil en aiguille, j'ai monté l'expérimentation décrite dans la section précédente pour donner un tour plus scientifique à tout cela.

Nous réalisons un tour d'horizon dans cette section en utilisant le fichier avec 5% de valeurs manquantes (111 observations complètes sur les 300).

4.1 Traitement des données manquantes avec Orange

4.1.1 Définition du schéma d'analyse.

Imputation des données manquantes lors de l'apprentissage. Au lancement du logiciel, nous retrouvons l'espace de travail qui permet de définir des schémas d'analyse. Nous plaçons le composant FILE (onglet DATA) que nous paramétrons pour charger le fichier CREDIT-GERMAN-MD-5_00.TXT (avec 5.00% de valeurs manquantes). Nous fixons l'option DON'T KNOW à « ? » pour spécifier le code représentant les valeurs manquantes.

Image Canvas		x
File View Options Widget Help	 File Data File Credit-german-md-5_00.txt Info 300 example(s), 20 attribute(s), 0 meta attribute(s). Classification; Discrete class with 2 value(s). Advanced settings Missing Value Symbols Don't care: Don't know: ? New Attributes Create a new attribute when existing attribute(s) Have mismatching order of values Have no common values with the new (recommended) Miss some values of the new attribute Always create a new attribute 	Prepr
	Keport	

Première vérification à faire, nous visualisons les données avec l'outil DATA TABLE (onglet DATA).

🥲 Orange Canvas						
File View Options Widget Help	🔚 Data Table					
	Info	cre	dit-german-md-5_00	(Examples)		
Data Visualize Classify Regression	300 examples,		checking_status	duration	credit_history	рі 🔺
- 2 - D -	109 (05.0 %) with hissing values.	1	<0	12	critical/other ex	new ca
	20 attributes,	2	<0	6	? 🔶	furnitu
Data Table	no meta attributes.	3	<0	6	critical/other ex	new ca
•	Discrete class with 2 values.	4	no checking	15	existing paid	used c
	Colliner	5	>=200	18	all paid	radio/1
	Settings Show meta attributes Show attribute labels (if any) Resize columns: Desize 0 Desize 0	6	no checking	18	all paid	new ca
		7	no checking	48	critical/other ex	busine
		8	0<=X<200	30	critical/other ex	new ca
		9	no checking	18	delayed previo	busine
	Restore Order of Examples	10	no checking	18	critical/other ex	used c
0 0	Colors	11	no checking	? 🔶	all paid	used c
Examples Data Table	Visualize continuous values	12	>=200	10	? 🔶	radio/1
	Color by class value	13	no checking	12	critical/other ex	new ca
Ela	Set colors	14	no checking	24	existing paid	new ca
1110		15	0<=X<200	13	existing paid	radio/1
	Selection	16	<0	18	critical/other ex	new ca
	Send selections	17	no checking	18	critical/other ex	radio/1
	Commit on any change	18	no checking	9	existing paid	radio/1
		19	<0	18	existing paid	new ca 🔻
	Report	< □				•

Effectivement, Orange a réalisé correctement le chargement. Il nous annonce que 189 observations contiennent *au moins* une valeur manquante (189 + 111 obs. complètes = 300, le compte y est).

Nous plaçons l'outil IMPUTE (onglet DATA) dans le schéma. Nous actionnons le menu OPEN pour accéder à la boîte de paramétrage. Plusieurs options sont possibles pour traiter globalement l'ensemble des variables. Le fichier d'aide du logiciel fournit les indications suivantes :

- DON'T IMPUTE. Orange laisse le soin à l'algorithme d'apprentissage (LOGISTIC REGRESSION) en aval de réaliser à la volée l'imputation des valeurs manquantes. Nous y reviendrons plus loin.
- AVERAGE /MOST FREQUENT. C'est l'imputation univariée que nous avons mis en œuvre à l'aide de R dans la section précédente.
- MODEL-BASED IMPUTER. Orange utilise l'algorithme du plus proche voisin pour prédire la valeur manquante d'une variable à partir des autres variables prédictives de la base. Si l'idée semble séduisante, nous nous appuyons sur les liens entre les variables pour produire la meilleure imputation possible, le temps de calcul peut être rapidement prohibitif sur les grandes bases. On parle aussi d'approche « hot deck » dans la littérature.
- RANDOM VALUES. Orange essaie de reconstituer la distribution des données pour chaque variable. Il utilise alors une valeur extraite, en respectant la distribution approximée, de la plage des valeurs possibles.
- REMOVE EXAMPLES WITH MISSING VALUES. Il s'agit de la stratégie « listwise deletion ».

Nous choisissons l'option MODEL BASED IMPUTER.



N.B. : Il est possible de définir une stratégie spécifique pour chaque variable. Nous pouvons même définir la valeur d'imputation de manière ad hoc (INDIVIDUAL ATTRIBUTE SETTNGS = VALUE).

Il nous faut maintenant placer le composant LOGISTIC REGRESSION (onglet CLASSIFY). Nous ne procédons pas à une sélection de variables.



Notons que le composant « régression logistique » dispose d'un mécanisme interne d'imputation des valeurs manquantes (IMPUTATION OF UNKNOWN VALUES). Plusieurs solutions sont proposées : une très classique imputation univariée (moyenne, minimum, maximum) ; ou une imputation multivariée où l'on cherche à prédire les valeurs manquantes d'une variable prédictive à partir des autres variables présentes dans la base. Dans ce cas, un arbre de décision (resp. un arbre de régression) est utilisé pour imputer les valeurs d'une variable qualitative (resp. quantitative). Cette fonctionnalité est activée si nous choisissons l'option DON'T IMPUTE lors du paramétrage du composant IMPUTE.

Nous ne l'effectuons pas dans ce didacticiel mais il est possible de visualiser le modèle en utilisant l'outil NOMOGRAM par exemple. Nous détaillons son utilisation et la lecture des résultats dans notre didacticiel consacré au « classifieur bayésien naïf »⁹.

Evaluation du modèle sur l'échantillon test. Le modèle est directement construit lorsque nous fermons la boîte de paramétrage. Il ne nous reste plus qu'à évaluer ses performances sur l'échantillon test. Nous insérons une seconde fois le composant FILE. Nous sélectionnons le fichier CREDIT-GERMAN-TEST.TXT.

⁹ http://tutoriels-data-mining.blogspot.com/2010/03/le-classifieur-bayesien-naif-revisite.html

or Orange Canvas	
<u>File View Options Widget Help</u>	File (2)
	Data File
Data Visualize Classify Regression	credit-german-test.txt 🔻 🕕 🐼 Reload
File Paint Info Save Data Table	Info 700 example(s), 20 attribute(s), 0 meta attribute(s). Classification; Discrete class with 2 value(s).
Examples Data Table Examples	Advanced settings Missing Value Symbols Don't care: Don't know:
Impute Log	New Attributes E Create a new attribute when existing attribute(s) Have mismatching order of values Usave no common values with the new (recommended)
File (2)	Miss some values of the new attribute Always create a new attribute Report
	· · · · · · · · · · · · · · · · · · ·

Puis nous plaçons le composant TEST LEARNER (onglet EVALUATE). Nous lui connectons successivement : la régression logistique [LOGISTIC REGRESSION], le fichier d'apprentissage imputé [IMPUTE]¹⁰ et le fichier test [FILE (2)].



¹⁰ Je ne comprends pas très bien pourquoi il faut connecter de nouveau ce composant au TEST LEARNER puisqu'il est déjà relié à la régression logistique. Mais, l'évaluation n'est pas possible sans cette opération.

Attention pour cette dernière connexion, il faut spécifier explicitement (en double-cliquant sur le lien) qu'il s'agit de l'échantillon test (SEPARATE TEST DATA).

Il ne nous reste plus qu'à cliquer sur le menu OPEN de TEST LEARNER. Dans la fenêtre de visualisation qui apparaît, nous indiquons d'une part que l'évaluation doit être réalisée à l'aide de l'échantillon test (SAMPLING – TEST ON TEST DATA), d'autre part que seul le taux de bon classement (CLASSIFICATION ACCURACY) nous intéresse dans notre étude. Nous obtenons ainsi un taux en test de 72.43%.

TestLearners		
Sampling	Evaluation Results	
Cross-validation	Method CA	
Number of folds: 5	1 Logistic regression 0.7243 🔶	
Leave-one-out		
Random sampling		
Repeat train/test: 10 🚔		
Relative training set size:		
57%		
Test on train data		
Itest on test data		
Apply on any change		
Apply		
Performance scores		
Classification accuracy		
Sensitivity		
Specificity Area under BOC curve		
Information score		
Target dass		
good 🔹		
Report		

4.1.2 Comparaison des techniques d'imputation

Voyons ce qu'il en est des positions relatives des techniques. Les résultats sont directement comparables parce que nous utilisons exactement les mêmes échantillons d'apprentissage et de test. Orange propose une fonctionnalité très pratique. Il suffit de laisser ouvert la fenêtre TEST LEARNERS et de modifier la technique d'imputation dans la fenêtre de IMPUTE pour voir les résultats directement mis à jour (il faut que l'option SEND AUTOMATICALLY soit cochée dans la boîte de dialogue IMPUTE).

	Timpute	TestLearners
Credit-german-md - Orange Canvas	Default imputation method Don't Impute Average/Most frequent Model-based imputer Random values	Samping Evaluation Results Cross-validation Number of folds: 5 © Leave-one-out
Data Visualize Classify Regression Evaluate Paint Info Save Data Select R tributes R		 Random sampling Repeat train/test: 10 Relative training set size:
Examples Examples Examples Examples Impute	employment installment_commitment personal_status residence_since property_magnitude age other_payment_plans Set All to Default dass Imputation Impute dass values Send data and imputer	Apply Performance scores Classification accuracy Sensitivity Specificity Area under ROC curve Information score F-measure * * #
File	Apply Send automatically Report	Target dass good Report

Nous obtenons le tableau suivant à l'issue de l'expérimentation :

Approach	Class. Accuracy
Don't impute (Internal method of Logistic Reg.)	0,7357
Average / Most Frequent	0,7271
Model-based imputer	0,7243
Random Values	0,7357
Remove Examples with missing values	0,6614

Mise à part la suppression des lignes, toutes les méthodes se valent. Très étrangement, dans certains cas le taux de bon classement (73,57%) est meilleur que celui de glm(.) de R construit à partir du fichier de données complet (73%), sans observations manquantes.

A mon sens il ne faut pas trop s'attarder sur le résultat brut. D'une part parce que nous ne savons pas comment fonctionne exactement la régression logistique d'Orange¹¹. D'autre part, parce que nous nous situons dans le cadre très spécifique d'une expérimentation artificielle. Après coup on peut se dire que dans notre configuration MCAR, très extrême il faut le reconnaître, les valeurs étant retirées complètement aléatoirement de l'échantillon d'apprentissage, toute technique d'imputation un tant soit peu raisonnable s'avère viable. Finalement, le grand perdant de ce tutoriel est la suppression de lignes. Elle est très pénalisante dès que la proportion de valeurs manquantes augmente sensiblement (à partir de 5% sur notre fichier).

4.2 Traitement des données manquantes avec KNIME

KNIME propose le composant MISSING VALUES pour le traitement de données manquantes. Nous reproduisons dans cette section le schéma d'analyse que nous avons défini sous Orange.

¹¹ Curieusement, le taux de succès en test du modèle élaboré à partir du fichier d'apprentissage complet est de 72.14%, j'avoue que ce résultat assez singulier m'a laissé un peu perplexe...

Construction du modèle sur l'échantillon d'apprentissage traité par suppression de lignes. Nous importons le fichier CREDIT-GERMAN-MD-5_00.TXT à l'aide du composant FILE READER (branche IO / READ du dépôt des méthodes). Les valeurs manquantes sont symbolisées par le caractère « ? ».

									6	- • ×
File Edit View Search Run Node Help		,								
📑 🕶 🔚 🔞 🛷 🕶 🖢 🖛 🍫	Q - 💛 V) 100% • ••	🔬 Dialog -	2:1 - File Reade	r					
🖄 Workflow Projects 📃 🗆	🛕 0: Tree Again	🔺 *2: Missing	File	/	\frown					
			Settings	El Wariahlan	Memory Deligy					
	File Reader	ī	Contrast AC	TT data fila la sati	Heritory Policy					
A Tree Again			CLACE AS		on: press ente	er toupdate preview)				
		Configure 🖊	vali	d URL: 3/md e	operiments with	n orange/german/cre	dit-german-md-	5 00.txt 👻	Browse	1
	1 📈 🖸	Execute					-	-		'
		Execute and Ope	-Bacic Sett	inge		Dracarija licar cat	tings for new lo	cation		
	Node1	Cancel	Dasic Set	1193			-L.			
	•	Reset	read r	OW IDS		Column delimiter:	ab> 🔻		Advanced.	
🔷 Node Repository		E d'A Ne de Neuer	🔽 read o	olumn headers:		ignore spaces an	d tabs			
- ☆ ↔ ▽	-	- Edit Node Name				Java-style comm	ents	Single line	comment:	
A 01		New Workflow A								
Bead		Collapse into Me								
File Reader		Expand Meta No	Preview							
ARFF Reader		Show Flow Varia		Clic	k column heade	er to change column p	roperties (* = n	ame/type user :	settings)	
CSV Reader			Row ID	S checkin	- duration	S credit history	S purpose	+ credit	S savings	S employ.
Table Deader	4	Cut	Row0	<0	12	critical/other exist	new car	3499	<100	1<=X<4
Table Keader DNANI Reader		Сору	Row1	<0	6	?	furniture/eq	1872	<100	? =
Model Reader	1	Paste	Row2	<0	6	critical/other exist	new car	1361	<100	<1
Write E	🗉 Console 🖇 🎸	Undo	Row3	no checking	15	existing paid	used car	3029	<100	4<=X<7
i Other		Redo	Row4	>=200	18	all paid	new car	6458	<100	>=7
Cache	********	Delete	Row6	no checking	48	critical/other exist	business	7629	no known sa	>=7
🗧 Database	*** Welco	Delete	Row7	0<=X<200	30	critical/other exist	new car	4249	<100	unemployed
🗟 Data Manipulation	*** Cor 🖸	0 File Table	Row8	no checking	18	delayed previously	business	2169	<100	1<=X<4
🔍 Data Views	********	*********	Row9	no checking	18	critical/other exist	used car	?	no known sa	unemployed +
Σ Statistics	Log file is	located at: C		•						
🌍 Mining	WARN Fil	e Reader								
12 Meta										
Flow Control										
and have							0	К	Apply	Cancel

Nous lançons l'importation effective du fichier en cliquant sur le menu EXECUTE.

Pour vérifier l'importation, nous visualisons les données à l'aide du composant INTERACTIVE TABLE (DATA VIEWS). Nous lui connections le composant précédent et nous cliquons sur le menu EXECUTE AND OPEN VIEWS.

File Reader	Interactive Tabl	e	\frown				
Node 1	Node 2	Table View - 2	2 - Interactive	Table			- 0 X
	F	lie Hilite Navig	ation view C				
	F	low ID	S checkin	duration	S credit_history	S purpose	credit
		Row0	<0	12	critical/other exist	new car	3499
		Row1	<0	6	?	furniture/eq	1872 =
		Row2	<0	6	critical/other exist	new car	1361
		Row3	no checking	15	existing paid	used car	3029
		Row4	>=200	18	all paid	radio/tv	1445
		Row5	no checking	18	all paid	new car	6458
		Row6	no checking	48	critical/other exist	business	7629
		Row7	0<=X<200	30	critical/other exist	new car	4249
		Row8	no checking	18	delayed previously	business	2169
		Row9	no checking	18	critical/other exist	used car	?
		Row 10	no checking	?	all paid	used car	7485
		Row11	>=200	10	?	radio/tv	1347
		Row12	no checking	12	critical/other exist	new car	926
		Row13	no checking	24	existing paid	new car	1249
		Row14	0<=X<200	13	existing paid	radio/tv	2101
		Row15	<0	18	critical/other exist	new car	5302
		Row 16	no checking	18	critical/other exist	radio/tv	2238
		Row17	no checking	9	existing paid	radio/tv	2697
		Row 18	<0	18	existing paid	new car	2249
		Row19	no checkina	12	existing paid	radio/tv	1262
		Row20	no checking	24	all paid	business	1559 +
		-	<u>الا</u>	·	+ ··		4

	🗻 Dialog - 2:3 - Missing Value	
	File	Nous plaçons
	Default Individual Flow Variables Memory Policy	maintenant le
	Integer Columns	composant MISSING
	Do Nothing Remove Pow	VALUE (DATA
		ΜΑΝΙΡΗΙΑΤΙΟΝ /
	Mean Most Frequent	
	⑦ Fix Value:	COLUMIN /
Interactive Table	Double Columns	TRANSFORM). Nous le
>	D Double	paramétrons (menu
	Do Nothing	CONFIGURE) de la
File Reader / Node 2 Missing Value	Min Max	manière à supprimer les
	© Fix Value: 0	lignos (PENAOVE POWA)
	String Columns	
Node 1 Node 3	S String	tant pour les variables
	O Nothing Remove Row	numériques (entières -
	Most Frequent	INTEGER ou réelles
	○ Fix Value:	DOUBLE) que pour les
	Unknown Columns	variables qualitatives
	? Unknown	(STRINC) Il c'agit de la
	Do Nothing	(STRING). IT Sagit de la
		stratégie « listwise
		deletion ».
		De nouveau nour voir
	UK Apply Cancel	l'état de l'ensemble de
		données, nous utilisons

le composant INTERACTIVE TABLE. Nous constatons que les lignes ROWO, ROW1, ROW2, ROW9, ROW10, etc. ont été exclues de l'ensemble d'apprentissage.



Il reste maintenant à brancher le composant d'apprentissage. La régression logistique n'est pas disponible dans Knime (**04/12/2011** : *En réalité, elle y est. Voir page 21*). Nous lui substituons le modèle bayésien naïf NAIVE BAYES (modèle d'indépendance conditionnelle, MINING / BAYES)¹². Il induit aussi un séparateur linéaire dans l'espace de représentation. Les performances en prédiction devraient être similaires. Nous désignons la variable à prédire CLASSE.



Nous actionnons le menu EXECUTE AND OPEN VIEWS. Knime fournit les distributions conditionnelles (moyenne et variances conditionnelles pour les variables quantitatives, avec l'hypothèse d'une distribution gaussienne ; le tableau de contingence pour les variables qualitatives).

A Naive Bayes Learner View - 0:5 - Naive Bayes Learner										
Class counts for classe										
Class:			bad			good				
Count:		35				76				
Total count: 111										
Gaussian distribution for age per class value COITITITUOUS COSCIPIOI bad good										
Count:			35		76					
Mean:			33.42857			38.07895				
Std. Deviation:			11.53584			11.65248				
Rate:		35/111				76/111				
P(checking_status class=?)	Dis	cret	e desc	ript	or					
Class/checking_status	0=X2	:00	0	>	=200	no checking				
bad	12		12 3		3	8				
good	23	23 9			5	39				
Rate:	329	32% 19% 7% 42%								
•										

¹² http://eric.univ-lyon2.fr/~ricco/cours/slides/naive_bayes_classifier.pdf

Evaluation sur l'échantillon test. Nous devons maintenant appliquer ce modèle sur l'échantillon test (credit-german-test.txt). Nous chargeons ce dernier à l'aide d'un second FILE READER.

Puis nous calculons les prédictions à l'aide du composant NAIVE BAYES PREDICTOR auquel nous relions : le modèle appris (NAIVE BAYES LEARNER) ; les données sur lesquelles doivent être calculées les prédictions (FILE READER – échantillon test).



Passons à la matrice de confusion. Nous utilisons l'outil SCORER (MINING / SCORING / SCORER). Nous le paramétrons (menu CONFIGURE) de manière à opposer la classe observée (FIRST ROW) et la classe prédite par le modèle [SECOND COLUMN – WINNER (NAIVE BAYES)].



Nous actionnons le menu EXECUTE AND OPEN VIEWS. Nous constatons un taux de succès de 70.286%.



Le classifieur bayésien naïf souffre moins de la suppression de lignes que la régression logistique. En effet, le modèle élaboré à partir des données complètes (sans valeurs manquantes) propose un taux d'erreur de 72%. Cette caractéristique est vraisemblablement due au faible nombre de paramètres estimés dans le modèle d'indépendance conditionnelle. Sa dépendance aux données d'apprentissage – sa variance - est plus faible. Il est (un peu) moins sensible à la réduction de l'échantillon.

Remarque : Il y a des limites à tout bien sûr. Lorsque nous passons à 10% de valeurs manquantes, l'apprentissage reste encore possible contrairement à la régression logistique. Mais le très faible nombre d'observations exploitées (40 lignes sur les 300 initialement disponibles) se répercute sur le taux de succès : 63,429%.

Imputation univariée. Lorsque nous sélectionnons l'imputation univariée (MEAN pour les colonnes INTEGER et DOUBLE / MOST FREQUENT pour les STRING) dans le composant IMPUTE, nous obtenons un taux de succès de 71.571% sur l'échantillon test. Encore une fois, dans un cadre très similaire (classifieur bayésien naïf à la place de la régression logistique), l'imputation univariée surpasse très facilement la suppression de lignes dans le cadre des données manquantes MCAR.

🛆 Confusion Matrix - 0:8 - Sco 📼 📼 💌									
File Hilite									
classe \ Wi	bad	good							
bad	103	107							
good	92	398							
Correct class	Wrong classifi Error: 28.4	ed: 199 29 %							

La régression logistique sous Knime. Correction, 04/12/2011. Très peu de temps après la publication de ce tutoriel, Loïc Lucel – qu'il en soit très chaleureusement remercié – m'a indiqué que la régression logistique était en réalité présente dans le logiciel Knime. Il fallait chercher dans la branche STATISTICS / REGRESSION du « Node Repository ». J'ai donc reproduit l'analyse en substituant les outils de la régression logistique (LEANER et PREDICTOR) à ceux du classifieur bayésien naïf. Nous avons le diagramme suivant :



Les variables prédictives qualitatives sont automatiquement recodées. Les sorties de la régression logistique respectent les standards du domaine¹³. Nous disposons du coefficient calculé, de l'estimation de son écart-type, de la statistique de test de significativité et de la p-value associée.

🔺 Logistic Regression Result View - 0:9 - Logistic Regression (Learner)											
File											
Stati	stics on Logistic Regressi	on									
Logit	Variable 🤇	Coeff.	Std. Err.	z-score	P> z						
bad	checking_status=0	1.2391	0.4972	2.4919	0.0127						
	checking_status=>=200	-0.3502	0.9284	-0.3772	0.706						
	checking_status=no checking	-1.3266	0.5226	-2.5383	0.0111	ĺ					
	duration	0.006	0.0192	0.3108	0.756						
	credit_history=critical/other existing	-2.8089	0.9262	-3.0328	0.0024						
	credit_history=delayed previously	-2.415	1.008	-2.3959	0.0166						
	credit_history=existing paid	-2.6995	0.8692	-3.1059	0.0019						
	credit_history=no credits/all paid	-0.1139	1.3018	-0.0875	0.9303						
	purpose=domestic appliance	0.8377	1.6268	0.5149	0.6066						
	purpose=education	1.6245	1.0658	1.5241	0.1275						
	purpose=furniture/equipment	-0.3342	0.8601	-0.3886	0.6976						
	purpose=new car	0.7304	0.7587	0.9627	0.3357						
	nurnass-other	0 7004	1 612	0.4240	0 6641						

¹³ Décidément, j'apprécie beaucoup ce logiciel. Tout comme Orange, il sait résister à la tentation de la course aux armements qui consiste à programmer tout un tas de techniques d'apprentissage dont on a du mal parfois à en saisir la teneur réelle. Et les outils, lorsqu'ils sont intégrés dans le logiciel, sont souvent d'excellente facture.

Test sample



Node 10

Node 8

Nous avons exécuté le « workflow » avec la base d'apprentissage comportant 5% de données manquantes. En utilisant **l'imputation univariée**, nous obtenons un taux de succès de **72.286%**.

Avec la stratégie « listwise deletion » (Remove row), il passe à 67.286%.



4.3 Traitement des données manquantes avec RapidMiner

Deux outils sont nécessaires pour détecter les valeurs manquantes dans RapidMiner. Par la suite, l'imputation se fait naturellement avec un composant dédié. Outre la suppression de lignes, le logiciel propose l'imputation univariée (moyenne /mode). Les techniques plus sophistiquées ne sont pas disponibles.

Détection des valeurs manquantes. Après avoir démarré RapidMiner, nous créons un nouveau « Process ». Nous utilisons le composant READ CSV (branche IMPORT / DATA dans OPERATORS) pour importer les données. Son paramétrage peut être compliqué, mieux vaut passer par le wizard (bouton IMPORTATION CONFIGURATION WIZARD) pour paramétrer la lecture du fichier.

🚯 Missing values* – RapidMiner@Maison-PC			
<u>File Edit Process Tools View H</u> elp			
🖹 📦 🔒 🕞 🗊 🔊 🔌	🕨 📗 🛐 🛒 🜒		
Overview 🗙 💱 🖨 🗐	C ² Process X E XML X	að - 10 🗐 🔊 -	
C Operators			Read CSV
🛷 👻 [Filter] 🚳 🎾 👫			🎢 🎢 Import Configuration Wizard
Control (35) C	Read CSV		csv file
⊡ 🔄 Import (26) ⊡ 🔄 🔄 Data (18)	fii 🌛 out D		column separators
Read CSV			T trim lines
– 🥙 Read Excel – 🌉 Read Excel with Format			🕜 Help 🗶 📓 Comment 🗶
– 🧳 Read XML – 🌙 Read Access			
- 🏕 Read AML - 🏕 Read XRFF - 🏕 Read Database			A Read CSV
– 🌛 Stream Database – 🌛 Read SPSS			Synopsis
- A Read Stata	Problems Cog Cog Cog		This operator can read csv files.
– 🥔 Read Sparse – 🎒 Read DBase	No problems found		Description
– 🛃 Read C4.5	Message	Fixes Location	This operator can read csv files, where all values of an example are
- A Read Biblex			writen into one line and separated by an constant separator. The separator might be specified in the column separators parameter.
📗 🔤 🕹 Read URL			The default will split the line on each comma, semicolon and blank.
H G Attributes (2)			Arbitrary regular expressions are usable as separator. Empty values
Results (1)			nunte the values (including the column senarators) with a double
0			

Le typage des variables est important. Dans notre cas, toutes les variables numériques doivent être spécifiées REAL, les autres sont POLYNOMINAL (variables catégorielles à plus de 2 modalités) ou BINOMINAL (variables catégorielles à 2 modalités, c'est le cas de la CLASSE). De plus, nous devons indiquer au logiciel la variable à prédire CLASSE (LABEL).

🚯 Data	a import wizard	d - Step 4 of 4											x
2	This wizard guides you to import your data. Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.												
	, <u>R</u> eload data	<u>G</u> ue	ss value types	s 🔽 Previe	w uses only fi	rst 100 rows.	<u>D</u> ate format			•			
	\checkmark	\checkmark	✓	\checkmark	\checkmark	\checkmark	\checkmark	V	\checkmark	\checkmark	\checkmark	V	
nal_st	other_partie	residence_s	property_ma	age	other_paym	housing	existing_cre	job	num_depen	own_telepha	foreign_worl	classe	
o 🔻	polyno 🔻	real 💌	polyno 🔻	real 🔻	polyno 🔻	polyno 🔻	real 💌	polyno 🔻	real 💌	polyno 🔻	polyno 🔻	binomi	-
ute 🔻	attribute 💌	attribute 💌	attribute 🔻	attribute 💌	attribute 🔻	attribute 💌	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	[label]	•
e div/di	co applicant	2	real estate	29	none	?	2	skilled	1	none	yes	bad	
single	none	4	no known pr	36	none	for free	3	high qualif/s	1	yes	yes	good	
	-		-	·					555				
📀 0 e	rrors.									🖌 Ignore	errors 📃 S	how only <u>e</u>	errors
	Row,	Column			Error			Original value			Message		
								(Previous I	<u>N</u> ext	Finish Einish	X <u>C</u> a	ncel

RapidMiner se sert de cette première étape pour détecter les valeurs manquantes. Pour toutes celles indiquées REAL, le caractère « ? » est incohérent avec la définition de la variable. Le logiciel en déduit qu'il s'agit d'une valeur manquante. Nous le constatons en exécutant le PROCESS (menu PROCESS / RUN) (*Remarque* : il est très vraisemblable que RapidMiner vous demande de sauvegarder la description des traitements au préalable). Dans l'onglet des résultats EXAMPLE SET (READ CSV), il recense le nombre d'observations manquantes pour chaque variable REAL (ex. 17 pour DURATION, 20 pour CREDIT_AMOUNT, etc.).

🥸 Missing values – Rap	pidMiner@Maison-PC			100			x			
<u>File Edit</u> Process										
P 📦 🖩 🖬 🔊 \land 🕨 📗 🖏 🐺 👀										
🛛 🐺 Result Overview	/ 🛛 🗍 📑 ExampleSe	t (Read CSV) 🔪								
💿 Meta Data View 🔘) Data View 🔵 Plot View	w 🔘 Annotations				ته 🗳	•			
ExampleSet (300 exan	nples, 1 special attribute	, 20 regular attributes)			1	- 🏥 👻				
Role	Name	Туре	Statistics	Range	Missin	gs				
label	classe	binominal	mode = good (210), le	bad (90), good (210)	0					
regular	checking_status	polynominal	mode = no checking (1	<0 (85), no checking (1	0					
regular	duration	real	avg = 22.512 +/- 13.23	[4.000 ; 60.000]	17 🔶 🗕					
regular	credit_history	polynominal	mode = existing paid (critical/other existing (0					
regular	purpose	polynominal	mode = new car (84), I	new car (84), furniture/	0					
regular	credit_amount	real	avg = 3709.750 +/- 328	[276.000 ; 15857.000]	20					
regular	savings_status	polynominal	mode = <100 (179), le	<100 (179), no known	0					
regular	employment	polynominal	mode = 1<=X<4 (93), I	1<=X<4 (93), ? (12), <1	0					
regular	installment_commitm	real	avg = 3.045 +/- 1.098	[1.000 ; 4.000]	10 🔶 💳					
regular	personal_status	polynominal	mode = male single (1	female div/dep/mar (92	0					
🔒 Log 💥 💱 🖨										
🔒 🥔 🤻						nc	nitor			
Dec 3, 2011 5:49:19 AM CONFIG: Loading perspectives. Dec 3, 2011 5:49:20 AM INFO: Checking for updates. Dec 3, 2011 5:49:21 AM INFO: No updates since Wed Nov 30 08:59:11 CET 2011. Dec 3, 2011 5:49:24 AM INFO: Decoupling process from location //NewLocalRepository/Missing values. Process is now associated with file //NewLocalRepository/Missing values. Dec 3, 2011 5:52:27 AM WARNING: Password in XML file looks like unencrypted plain text. Dec 3, 2011 5:52:27 AM WARNING: Password in XML file looks like unencrypted plain text. Dec 3, 2011 5:53:02 AM INFO: Saved process definition at //NewLocalRepository/Missing values'. Dec 3, 2011 5:58:17 AM INFO: Reading example set										

Revenons à la fenêtre de définition des traitements. Nous insérons le composant DECLARE MISSING VALUES (DATA TRANSFORMATION / VALUE MODIFICATION) dans l'espace de travail. Nous lui connectons READ CSV. Nous le paramétrons de manière à ce qu'il ne traite que les variables nominales, en lui indiquant que le code « ? » désigne les valeurs manquantes.



Eile Edit Process Tools View Help									
🗋 😋 📰 🔂 🔊 🔺 🕨 🔠 🖏 🔯									
🔀 Result Overview 🕱 🔲 ExampleSet (Declare Missing Value) 🕺									
Meta Data View Data View Plot View Annotations	🖫 🦂 🚽								
ExampleSet (300 examples, 1 special attribute, 20 regular attributes)	. •								
Role Name Type Statistics Range 💎 Mis	sings								
label classe binominal mode = good (210), le: bad (90), good (210) 0									
regular checking_status polynominal mode = no checking (1 <0 (85), no checking (1 14									
regular duration real avg = 22.512 +/- 13.23 [4.000; 60.000] 17									
regular credit_history polynominal mode = existing paid (critical/other existing (19									
regular purpose polynominal mode = new car (84), I new car (84), furniture/i 18									
regular credit_amount real avg = 3709.750 +/- 328 [276.000; 15857.000] 20									
regular savings_status polynominal mode = <100 (179), le: <100 (179), no known : 11									
regular employment polynominal mode = 1<=X<4 (93), li 1<=X<4 (93), ? (0), <1 (12									
regular installment_commitme real avg = 3.045 +/- 1.098 [1.000 ; 4.000] 10									
🔂 Log 🕺 📲 🖅									
	initor								
Dec 3, 2011 6:11:38 AM INFO: Process //NewLocalRepository/Missing values starts									
Dec 3, 2011 6:11:38 AM INFO: Loading initial data.									
Dec 3, 2011 611/38 AM INFO: Saving results.									
Dec 3, 2011 0.11.30 Aw INFO. Process invewLocalRepository wissing values infished successfully after 0.5 Dec 3, 2011 6:31:43 AM INFO: No filename given for result file, using stdout for logging results!									
Dec 3, 2011 6:31:43 AM INFO: Process //NewLocalRepository/Missing values starts									
Dec 3, 2011 6:31:43 AM INFO: Loading initial data.									

En sélectionnant l'option DATA VIEW dans l'onglet des résultats, nous pouvons visualiser les données en les filtrant de différentes manières. Nous n'affichons que les lignes comportant au moins une valeur manquante (MISSING_ATTRIBUTES). RapidMiner nous indique qu'ils sont au nombre de 189¹⁵.

🚯 Missing va	🐼 Missing values - RapidMiner@Maison-PC										
<u>F</u> ile <u>E</u> dit	Process To	ols ⊻iew <u>H</u>	leip								
P 🔄											
🛛 🛒 Resul	🔀 Result Overview 🕱 🖉 ExampleSet (Declare Missing Value) 🕱										
🔘 Meta Data	a View 💿 Da	ta View 🔵 Pl	ot View 🔘 A	nnotations.						i.	ے 😓
ExampleSet	(300 example	s, 1 special at	tribute, 20 reç	jular attributes)			iew Filter (189	9 / 300): [missing_attributes	•
Row No.	classe	checking_st	duration	credit_history	purpose	credit_amo	savings_st	employment	installm	ent personal_st	other
1	bad	<0	12	critical/other	new car	3499	<100	1<=X<4	3	female div/di	co ap
2	good	<0	6	? 🔶	furniture/equ	1872	<100	? 🔶	4	male single	none
3	good	<0	6	critical/other	new car	1361	<100	<1	2	male single	none
4	good	no checking	18	critical/other	used car	?	no known sa	unemployed	2	male single	none
5	bad	no checking	?	all paid	used car	7485	no known sa	unemployed	4	female div/d	none
6	good	>=200	10	?	radio/tv	1347	no known sa	4<=X<7	4	?	none
7	good	0<=X<200	13	existing paid	radio/tv	2101	<100	<1	2	female div/d	guar
8	good	no checking	18	critical/other	radio/tv	2238	<100	1<=X<4	2	female div/d	none 🤜
<											
🗟 Log	X 🚼 🗢 🖻										
R <i>Q</i>	۵									Mo	nitor 🖹
Dec 3, 2011	6:11:38 AM IN	IFO: Process /	/NewLocalRe	pository/Missi	ng values star	rts					
Dec 3, 2011	6:11:38 AM IN	/FO: Loading i/	nitial data.								
Dec 3, 2011	6:11:38 AM IN	IFO: Saving res	sults.				6.0 A 0 -				
Dec 3, 2011	6:31:43 AM IN	JEO: No filenar	me diven for r	pository/wissi esult file_usin/	ng values finis 1 stdout for loc	sned success ading results)	runy atter o s				
Dec 3, 2011	6:31:43 AM IN	IFO: Process /	/NewLocalRe	pository/Missi	ng values star	rts					
Dec 3, 2011	6:31:43 AM IN	IFO: Loading i	nitial data.								

¹⁴ Ca paraît un peu compliqué tout ça. La difficulté vient de la nécessité de faire comprendre à RapidMiner le code des valeurs manquantes dans un fichier texte. Lorsque les données proviennent d'une base, le processus est nettement simplifié comme le montre ce tutoriel : <u>http://www.youtube.com/watch?v=0IVZmAk0pI4</u>
¹⁵ Dans le fichier à 5% de valeurs manquantes, il y a 111 observations complètes dixit le traitement dans R. Ainsi, 111 + 189 = 300 observations, la taille initiale de notre échantillon d'apprentissage.

Imputation et construction du modèle. Nous utilisons REPLACE MISSING VALUES pour l'imputation¹⁶ (DATA TRANSFORMATION / DATA CLEANSING). Nous lui connectons l'opérateur précédent.



AVERAGE est le type de remplacement proposé par défaut. En réalité, cette option inclus le remplacement par le mode pour les variables nominales.

Nous pouvons maintenant introduire l'outil d'apprentissage supervisé. La régression logistique proposée dans RapidMiner ne correspond pas à la méthode que l'on retrouve usuellement dans les logiciels de statistique. Il m'a paru préférable d'utiliser le classifieur bayésien naïf encore une fois.



Nous lançons l'exécution. Dans le tableau décrivant les distributions conditionnelles, nous constatons que les modalités « ? » des variables ne correspond à aucune observation maintenant. Les imputations ont bien été effectuées.

¹⁶ Un outil IMPUTE MISSING VALUES existe. Mais il semble encore non stabilisé. Son utilisation n'est pas recommandée dixit la documentation du logiciel.

🚯 Missing va	alues – RapidN	liner@Maisor	n-PC		• X
<u>F</u> ile <u>E</u> dit	<u>P</u> rocess <u>T</u> o	ols <u>V</u> iew <u>I</u>	<u>H</u> elp Nouve	elle capture Ctrl+N	
i 🕒 🕥		I	A		
🛛 🐺 Resul	t Overview 🛛	🍸 💡 Sim	pleDistribution	(Naive Bayes) 🔀 🔪	
🔵 Text View	O Plot View	🤊 💽 Distribu	tion Table) 🔘	Annotations	🖫 🤳 🚽
Attribute	Parameter	bad	good		
checking_st:	value=<0	0.489	0.195		
checking_st:	value=no chi	0.167	0.505		
checking_st;	value=>=200	0.033	0.062		
checking_st:	value=0<=X∘	0.311	0.238		
checking_st:	value=?	0.000	0.000 🧹		
checking_st:	value=unknc	0.000	0.000		
duration	mean	25.345	21.298		
duration	standard dev	13.375	12.462		
credit_histor	value=critica	0.178	0.329		
credit_histor	value=?	0.000	0.000 🧲		
credit_histor	value=existir	0.533	0.533		
Log	x 23 - 0				
🔲 🚽 🍭	•				onitor
Dec 3, 2011	6:10:58 AM IN	FO: Loading	initial data.		
Dec 3, 2011	6:10:58 AM S	EVERE: Proc	ess failed: one	erator cannot be executed (Extraction of nominal example value for	

Evaluation sur l'échantillon test. Pour l'évaluation, nous devons appliquer ce modèle sur l'échantillon test et confronter la prédiction avec les valeurs observées de la variable cible. Nous insérons une nouvelle fois le composant READ CSV pour charger l'échantillon test (CREDIT-GERMAN-TEST.TXT). Nous le connectons au composant APPLY MODEL (MODELING / MODEL APPLICATION) qui permet d'appliquer le classifieur sur un échantillon.



Enfin, l'opérateur PERFORMANCE CLASSIFICATION (EVALUATION / PERFORMANCE MEASUREMENT / CLASSIFICATION AND REGRESSION) permet de mesurer le taux de succès (ACCURACY).

😥 Missing values - RapidMiner@Maison-PC	
Eile Edit Process Iools View Help	
🗋 📦 🖩 🖬 🔊 🔌 🕨 🔳 🔳 🛐 🕏	
Poverview 🛛 💱 🗇 🔟 🔗 Process 🗶 🕞 XML 🗶	🛛 🔂 Parameters 💥 💱 🖛 🔟
(+ + + +) → + (+ + + + + + + + + + + + + + + +	að + 🛊 🔲 🔯 🚸 + 🚨 👦 👳 🛼 +
	Performance (Performance (Classifica
Main Process	main criterion first
Declare Missi Replace Missi	accuracy 🦛 💶
era era era of p of p	Classification error
inp D	🗌 карра
	weighted mean recall
	weighted mean precision
S Centermance	
E Strate (1)	spearman rho
Example Control Contro	kendali tau
E a Results (1) Read CSV (2)	lodel Performance
E S Evaluation (17)	mod oper % exa
E Serformance Measurement (17)	e relative error
- % Performance (Classification)	relative error lenient
- % Performance (Regression)	
- % Performance (Costs)	
K Performance (Support Vector Co	normalized absolute error
Comparison of the state of	Tixes Location not mean squared error
E 🗐 Clustering (5)	
- 7 Cluster Courti Performance - 7 Cluster Distance Performance	root relative squared error
Cluster Density Performance	squared error
0	

Nous exécutons le diagramme. Nous obtenons un taux de succès de 74.86% sur l'échantillon de 700 observations.

🥸 Missing values – Rapidl	Miner@Maison-PC	Annes					
Eile Edit Process Tools View Help							
🖹 👒 📰 🛃 🔊 \land 🕨 📗 🖏 🐺 💿							
🔀 Result Overview 🕱 🥂 况 PerformanceVector (Performance) 💈							
💿 Table / Plot View 🔿 Text View 🔿 Annotations 🛛 🕞 🚽							
Criterion Selector 💿 Multiclass Classification Performance 🔿 Annotations 🕞 🚛							
accuracy O Table View O Plot View							
accuracy: 74.86%							
		true bad	true good	class precision			
	pred. bad	91	57	61.49%			
	pred. good	119	433	78.44%			
	class recall	43.33%	88.37%				
Monitor 2							
Dec 3, 2011 7:03:38 AM INFO: Saving results. Dec 3, 2011 7:03:38 AM INFO: Process //NewLocalRepository/Missing values finished successfully after 0 s							

5 Expérimentation sur la base WAVE à 2 classes

Pour donner une plus grande assise à notre étude, nous avons réitéré l'expérimentation sous R sur la base WAVE¹⁷ réduite à 2 classes (21 variables prédictives quantitatives). Nous disposons de 500 observations en apprentissage et, surtout, de 32867 observations en test. L'estimation du taux de succès n'en sera que plus précise.

¹⁷ http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29

Nous obtenons le tableau suivant :

WA	WAVE DATASET		uracy rate
% missing	# complete obs.	Listwise Del.	Univ. Imputation
0,00%	500	0,9150	0,9150
0,50%	451	0,9147	0,9150
1,00%	405	0,9082	0,9162
2,00%	331	0,9068	0,9160
5,00%	177	0,8731	0,9174
10,00%	72	0,7847	0,9192
20,00%	6	ERR	0,9188

Les résultats corroborent ceux obtenus sur la base GERMAN. Lorsque la proportion de valeur manquante augmente, l'imputation univariée tient parfaitement son rôle alors que la suppression de lignes dégrade fortement la construction du modèle prédictif. Assez curieusement, on aurait même dit que l'introduction de valeurs manquantes pour les remplacer par la moyenne ou le mode améliorerait la qualité des modèles. Les différences sont infimes, gardons-nous bien de ce genre de conclusion. Mais j'avoue que ces résultats m'intriguent quelque peu, et ce n'est pas faute d'avoir vérifié l'exactitude des taux de succès recueillis¹⁸.

6 Conclusion

On ne le répétera jamais assez, le traitement des données manquantes est un problème difficile. Il faut faire des choix en fonction de facteurs que l'on ne maîtrise pas toujours très bien (le processus de formation des valeurs manquantes notamment). Dans ce tutoriel, nous avons essayé de montrer les solutions proposées par plusieurs logiciels. Nous constatons que dans le cadre des données manquantes totalement MCAR, l'imputation univariée (moyenne / mode) convient très bien dans la construction de modèles prédictifs linéaires, dans la mesure où notre principal critère d'évaluation est l'efficacité en prédiction.

7 Bibliographie

Allison, P.D. (2001), « Missing Data ». Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA : Sage.

Little, R.J.A., Rubin, D.B. (2002), « Statistical Analysis with Missing Data », 2nd Edition, New York : John Wiley.

¹⁸ De toute manière, les données réellement utilisées et les scripts sont accessibles sur notre site de tutoriels. R, tout comme Orange, Knime et RapidMiner, sont eux-mêmes librement téléchargeables sur leurs sites de distribution respectifs. Tout le monde peut reproduire à l'identique les résultats décrits dans ce document. C'est une règle que l'on devrait voir instaurer pour les publications dites « scientifiques ».