Big Data I - présentation de projet Construction d'un modèle de prédiction

Rémy Decocq Sam Boosko Dimitri Waelkens

Faculté des Sciences Université de Mons





20 mai 2019

Sélection du type de modèle

Ensemble de données

 Les données se présentent sous la forme d'un tableau contenant des informations relatives à un individu.

	Age	job	marital	 У
1	56	housemaid	married	 0
2	57	services	married	 0
3	37	services	married	 0
30436	61	retired	married	 0

FIGURE - Dataset Dtrain.csv

Où la variable y vaut 1 si cet indivi a ouvert un compte en banque, 0 sinon.

Sélection du type de modèle

- Au vu des données et de la prédiction recherchée, nous pouvons en déduire qu'il s'agit d'un problème de classification.
- 2 méthodes de classifications vue au cours testées :
 - Logistic Regression
 - 2 Linear Discriminant Analysis (LDA)
- LDA est plus stable quand les prédicteurs suivent une distribution gaussienne. Or, les données n'ont pas cette distribution.
- Donnant de meilleurs résultats, la régression logistique a été retenue.

Analyse et preprocessing du dataset

Analyse du dataset

- L'analyse des données montre :
 - seulement **3** observations de **default** valent **yes** sur **30436**.
 - $oxed{2}$ le nombre de $oxed{1}$ n'est pas proportionnel nombre de $oxed{0}$
 - **3** il y a **1316** observations où

$$pdays = 999 \land previous \ge 1$$

dont 108 se rapportent à y = 1

yes	no	unknown
3	7532	22901

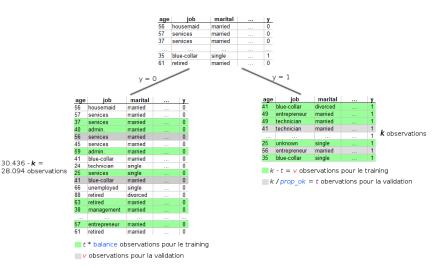
FIGURE -	table	of	people\$default
rigure -	Lable	Οı	peopleguerauit

1	0	total	
2342	28094	30436	

FIGURE – table of people\$y

Analyse et preprocessing du dataset

Sélection des observations pour le training et la validation



Analyse et preprocessing du dataset

Preprocessing du dataset

Catégorisation des valeurs des variables job et pdays.

pdays: 999 = never, > 5 = late, $\le 5 = \text{recent}$

Sélection des prédicteurs

Analyse de la signification des variables.

```
Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL
                             12647
                                        11397
iob
                  83.49
                             12645
                                        11314 < 2.2e-16
marital
                  18.01
                             12642
                                        11296 0.0004371
contact
                 108.98
                             12641
                                        11187 < 2.2e-16
month
                 475.58
                             12633
                                        10711 < 2.2e-16
day_of_week
                  17.93
                             12629
                                        10693 0.0012730
campaign
                   9.31
                             12628
                                        10684 0.0022823
pdays
                  24.83
                             12627
                                        10659 6.269e-07
previous
                                        10640 1.271e-05
                  19.05
                             12626
default
                  3.37
                            12624
                                        10637 0.1858031
loan
                   0.06
                             12622
                                        10637 0.9694916
housing
                   0.18
                             12621
                                        10637 0.6728589
                   '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Signif. codes:
```

- 3 variables ne sont pas significatives :
 - default
 - 2 loan
 - 3 housing

Résultats du modèle



Cross-Validation de la procédure



Conclusion

Big Data I - présentation de projet Construction d'un modèle de prédiction

Rémy Decocq Sam Boosko Dimitri Waelkens

Faculté des Sciences Université de Mons





20 mai 2019