

Big Data I - présentation de projet

Construction d'un modèle de prédiction

Rémy DECOCQ Sam BOOSKO Dimitri WAELEKENS

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	...	y
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 :
 - 1 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 :

- 1 seulement **3** observations de **default** valent **yes** sur **30436**.
- 2 le nombre de **1** n'est pas proportionnel nombre de **0**
- 3 il y a **1316** observations où

$$pdays = 999 \wedge previous \geq 1$$

dont **108** se rapportent à **y = 1**

yes	no	unknown
3	7532	22901

FIGURE – table of people\$default

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

age	job	marital	...	y
56	housemaid	married	...	0
57	services	married	...	0
37	services	married	...	0
...
35	blue-collar	single	...	1
61	retired	married	...	0

y = 0

y = 1

age	job	marital	...	y
56	housemaid	married	...	0
57	services	married	...	0
37	services	married	...	0
40	admin.	married	...	0
56	services	married	...	0
45	services	married	...	0
59	admin.	married	...	0
41	blue-collar	married	...	0
24	technician	single	...	0
25	services	single	...	0
41	blue-collar	married	...	0
66	unemployed	single	...	0
88	retired	divorced	...	0
63	retired	married	...	0
38	management	married	...	0
...
57	entrepreneur	married	...	0
61	retired	married	...	0

30.436 - k =
28.094 observations

age	job	marital	...	y
41	blue-collar	divorced	...	1
49	entrepreneur	married	...	1
49	technician	married	...	1
41	technician	married	...	1
...
25	unknown	single	...	1
56	entrepreneur	married	...	1
35	blue-collar	single	...	1

 k observations

$k - t = v$ observations pour le training

$k / \text{prop_ok} = t$ observations pour la validation

$t * \text{balance}$ observations pour le training

v observations pour la validation

Analyse et preprocessing du dataset

Preprocessing du dataset

- Catégorisation des valeurs des variables **job** et **pdays**.

$$\blacksquare \text{ jobs : } \begin{matrix} \text{poor} = \left\{ \begin{array}{l} \text{blue - collar} \\ \text{housemaid} \\ \text{services} \\ \text{technician} \\ \text{unknown} \end{array} \right. \end{matrix}, \begin{matrix} \text{middle} = \left\{ \begin{array}{l} \text{admin.} \\ \text{entrepreneur} \\ \text{management} \\ \text{self - employed} \\ \text{unemployed} \end{array} \right. \end{matrix},$$

$$\text{good} = \left\{ \begin{array}{l} \text{retired} \\ \text{student} \end{array} \right.$$

- **pdays** : 999 = never, > 5 = late, ≤ 5 = recent

Sélection des prédicteurs

- Analyse de la signification des variables.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			12647	11397		
job	2	83.49	12645	11314	< 2.2e-16	***
marital	3	18.01	12642	11296	0.0004371	***
contact	1	108.98	12641	11187	< 2.2e-16	***
month	8	475.58	12633	10711	< 2.2e-16	***
day_of_week	4	17.93	12629	10693	0.0012730	**
campaign	1	9.31	12628	10684	0.0022823	**
pdays	1	24.83	12627	10659	6.269e-07	***
previous	1	19.05	12626	10640	1.271e-05	***
default	2	3.37	12624	10637	0.1858031	
loan	2	0.06	12622	10637	0.9694916	
housing	1	0.18	12621	10637	0.6728589	

 signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- 3 variables ne sont pas significatives :

- 1 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 WAELEKENS

Faculté des Sciences
Université de Mons



20 mai 2019