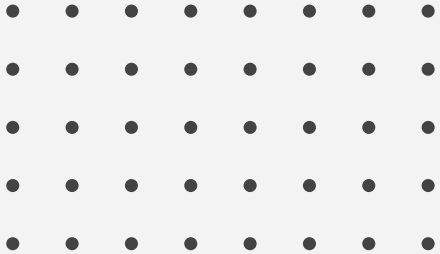


université PARIS DIDEROT PARIS 7

Projet Data Science

Trieu Xuong YU, François LE GAC
M2 ISIFAR
Promotion 2019 - 2020
Etape 1 : Mercredi 29 Janv.

Rappel du contexte

8457 individus, **18** variables explicatives, **1** variable cible

OBJECTIF :

Construire un score d'octroi de crédit pour les clients du groupe RCI

VARIABLE CIBLE :

Def12_31 **:** indique si la personne a fait défaut 1=Oui / 0=Non

	0	1
Effectif	8359	98
Pourcentage	99%	1%

=> On dispose d'un jeu de données **déséquilibré**

Description des variables

CONTRAT

pc_appo: Pourcentage d'apport. N

MT_APPORT : Montant de l'apport. N

MT_FINANCE : Montant restant dû après apport. N

MT_MENS: Montant de la mensualité. N

VR_BALLON: Montant ballon. N

DUREE_CONTRAT: Durée du contrat. N

MT_PREST: Montant des prestations. N

MT_ASSUR: Montant des assurances. N

CLIENT

age_cli: Age du client. N

ANC_EMPLOI : Ancienneté à l'emploi. N

STITUATION_FAM : Situation familiale. C (6 mods) (1=Marié, 2=Célib, 3=Divorcé, 4=Veuf, 5=Séparé, 11=Colloc)

MODE_LOGT : Mode de logement C (4 mods) (1=locataire, 2=proprio, 3=autre, 4=chez les parents)

anciennete_rci: Ancienneté relation rci. C (4 mods)

mois_gestion: Mois d'entrée en gestion. C

VEHICULE

PRIX_VEH : Prix du véhicule. N

AGE_VEH : Age du véhicule. N

MARQUE : Marque. C

VN_VO : Type de véhicule. C (2 mods : VN=véh. neuf /VO= véh. occasion)

N = Numérique, C = Catégorielle
(12 numériques, 6 catégorielles)

Qualité des données

Tableau des valeurs manquantes

	anciennete_rci	MT_ASSUR	AGE_VEH	VR_BALLON	MT_PREST	MODE_LOGT
Effectif	7271	6755	6003	5238	1280	108
Pourcentage	86%	79.9%	71%	61.9%	15.1%	1.3%

Prise de décisions :

1) Suppression de variable anciennete_rci
2) Remplacement NA par 0 pour les variables MT_ASSUR, VR_BALLON, AGE_VEH, MT_PREST

Qualité des données

Zoom sur le Mode de Logement

108 valeurs manquantes dont 10 qui font défaut (~10% des individus qui font défaut).

Mod_l \ Sit	Marié	Célib.	Divorcé	Veuf	Séparé	Coloc
Locataire	0.1	0.2	0.25	0.05	0.3	0.15
Proprio	**0.9**	0.4	**0.7**	**0.95**	**0.65**	**0.75**
Autre	0	0	0	0	0	0
Chez par.	0	0.4	0.05	0	0.05	0.1

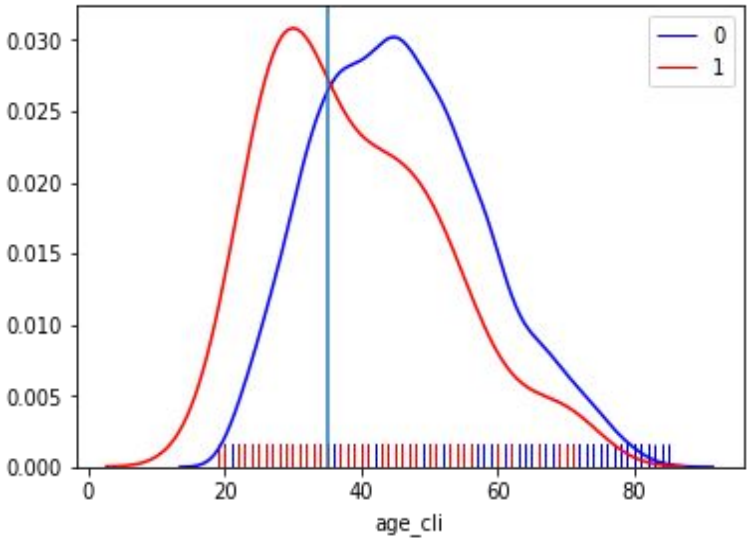
Tableau contingence mode de logement / situation familiale

On affecte la modalité propriétaire pour les colonnes en vert. **(72 obs)**
Pour les célibataires **(36 obs, dont 5 défauts),** on laisse les NAs.

Analyse d’impact âge

Âge du client vs cible

	Effectif	Pourcentage	Effectif	Pourcentage
cible	0		1	
age_cli_CAT (18,35]	1838	22%	47	49%
age_cli_CAT > 35	6488	78%	48	51%



Distributions conditionnelles : âge du client

Caractéristiques de l’individu en défaut

Jeunes :
(60% ont -40 ans pour les inds en défaut contre 40% pour les autres)

Faible expérience pro :
(30% ont entre 0 et 5 ans pour les inds. en défaut contre 15% pour les autres)

Faible apport :
(70% des inds. en défaut ont un apport <20% contre 50% pour les autres)

Célibataires :
(40% pour les inds. en défaut contre 20% chez les autres

Locataires ou chez les parents:
25% (1) contre 15% (0) pour les “locataires”. 25% (1) contre 10% (0) “chez les parents”.

Contrats plus longs :
65% des inds. en défaut ont un contrat > 40 mois contre 40% pour les autres

Cf les tableaux de lois conditionnelles en annexe pour plus de détails

Conclusion

Idées feature engineering :

1. Une variable **‘duree gestion’** à partir de la variable ‘mois gestion’
2. Une variable ‘**prix total’** comme étant la somme du prix du véhicule, du montant de gestion et du montant de l’assurance.

Dimension après transformation dummies :
8421 individus et 50 variables

Next steps :

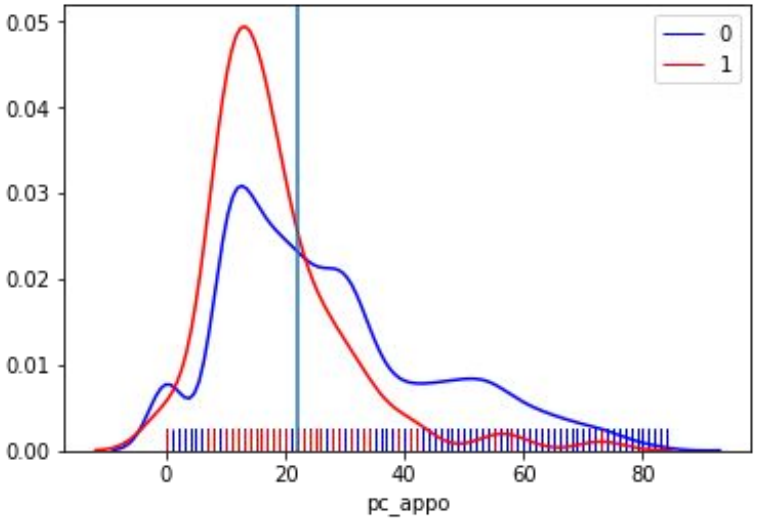
- Mise en place de techniques de rééchantillonnage pour lutter contre le déséquilibre du jeu de données
- Implémentation du modèle de régression logistique

Merci !

Annexe 1/3

Pourcentage de l’apport vs cible

	Effectif	Pourcentage	Effectif	Pourcentage
cible	0		1	
pc_appo_CAT (0,22]	4039	49%	71	75%
pc_appo_CAT (22,100]	4287	51%	24	25%

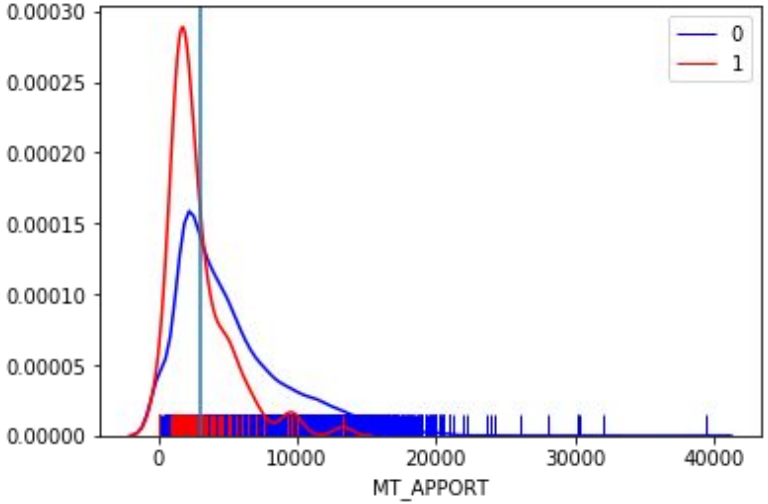


Distributions conditionnelles : pourcentage de l’apport

Annexe 2/3

Montant de l'apport vs cible

	Effectif	Pourcentage	Effectif	Pourcentage
cible	0		1	
MT_APPORT_CAT (0,3000]	3262	39%	64	67%
MT_APPORT_CAT > 3000	5064	61%	31	33%

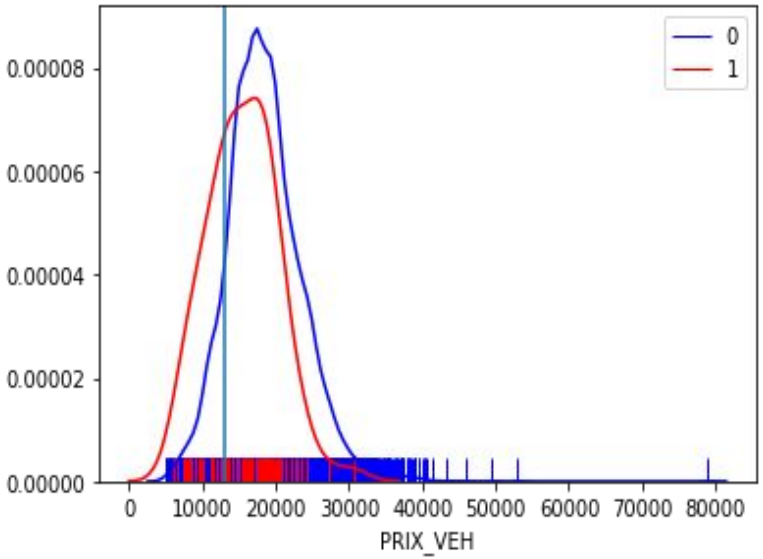


Distributions conditionnelles : montant apport

Annexe 3/3

Prix du véhicule vs cible

	Effectif	Pourcentage	Effectif	Pourcentage
cible	0		1	
PRIX_VEH_CAT (0,13000]	1003	12%	31	33%
pc_appo_CAT >13000	7323	88%	64	67%



Distributions conditionnelles : prix du véhicule