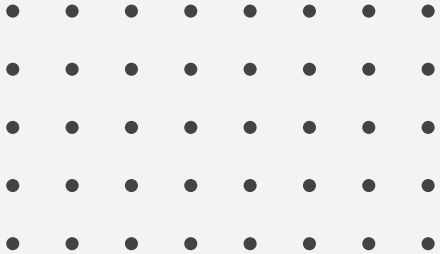


université PARIS DIDEROT PARIS 7

Projet Data Science

Trieu Xuong YU, François LE GAC
M2 ISIFAR
Promotion 2019 - 2020
Etape Finale : Mardi 31 Mars

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 (1/2)

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 (2/2)

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

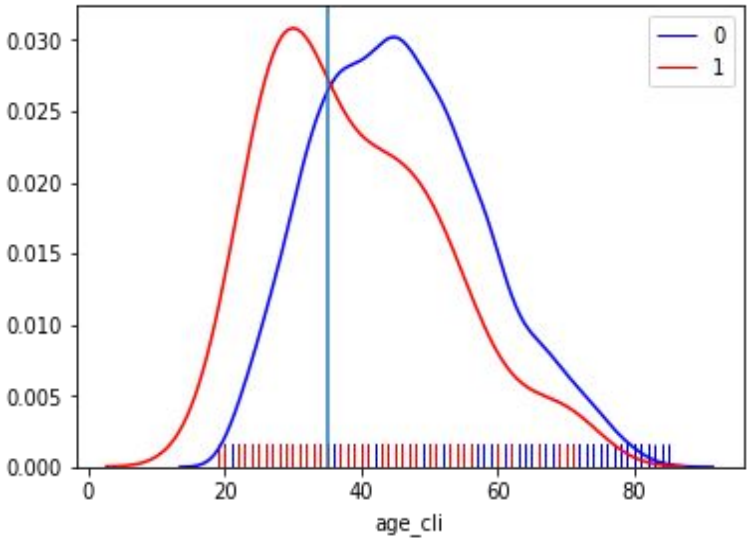
Tableau contingence mode de logement / situation familiale

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 du client

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

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

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

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

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

Contrats plus longs :
65% des inds. en défaut ont un contrat > 40 mois 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)

Cf Annexe : tableaux de lois conditionnelles pour plus de détails

Bilan

Idées feature engineering :

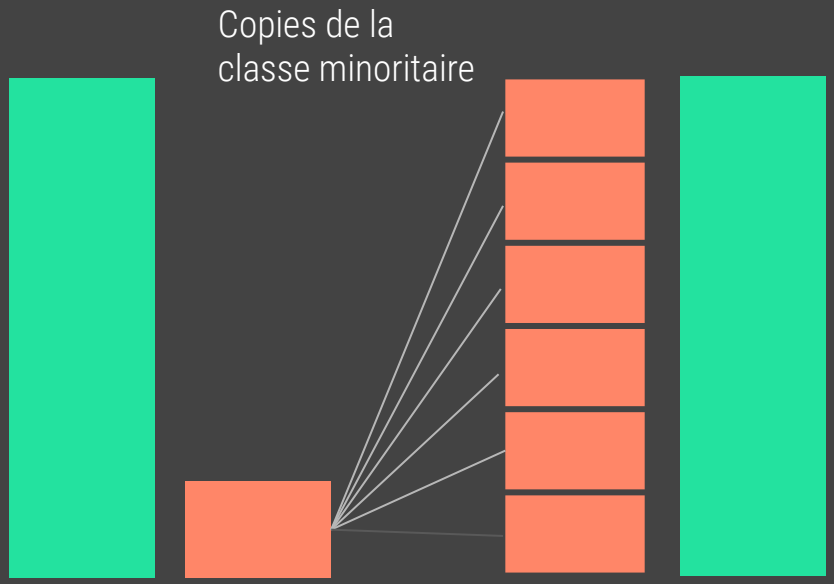
1. Une variable **‘duree gestion’** à partir de la variable ‘mois gestion’
2. Une variable ‘**prix total’** comme étant la somme du prix du véhicule, du montant de gestion et du montant de l’assurance.
3. Création de ratios

Dimension après transformation dummies :
8421 individus et 77 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

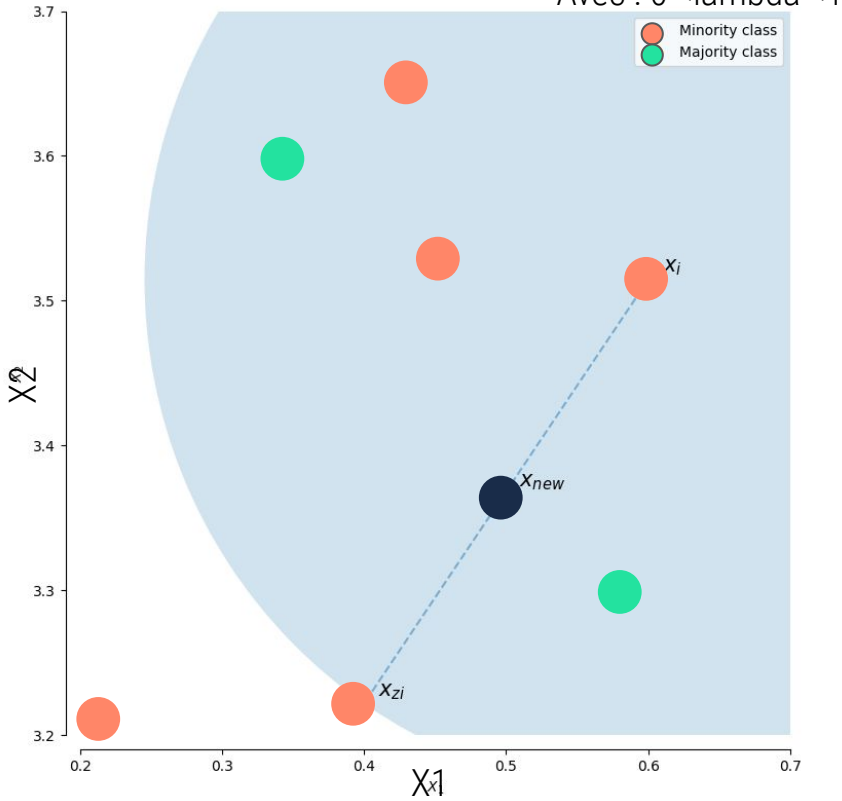
Oversampling



Smote

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i)$$

Avec : 0 <lambda <1



La régression logistique

- Modèle sur la loi Yi sachant Xi:

$\mathbb{P}(y_i = 1 \mid x_i) = \sigma(x_i^\top w + b)$ avec $\sigma(z) = \frac{1}{1+e^{-z}}$ la fonction sigmoid

- On trouve les meilleurs paramètres w et b en maximisant la vraisemblance

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w + b)$$

Où $\ell(y, y') = \log(1 + e^{-yy'})$ la perte de logistique

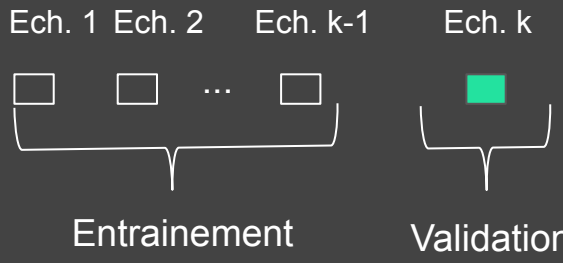
⟺ A minimiser $argmin(-\log(L(w, b)) + C * \|w\|)$

Moins log-vraisemblance

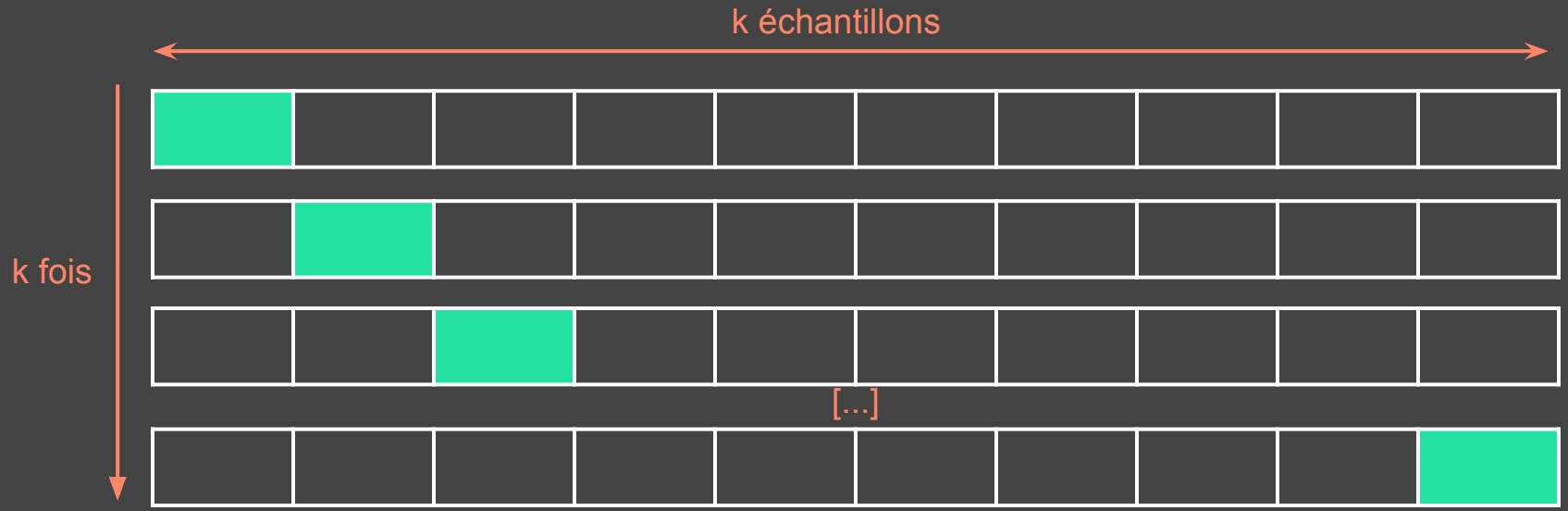
Paramètre à ajuster

La pénalité: l1/l2

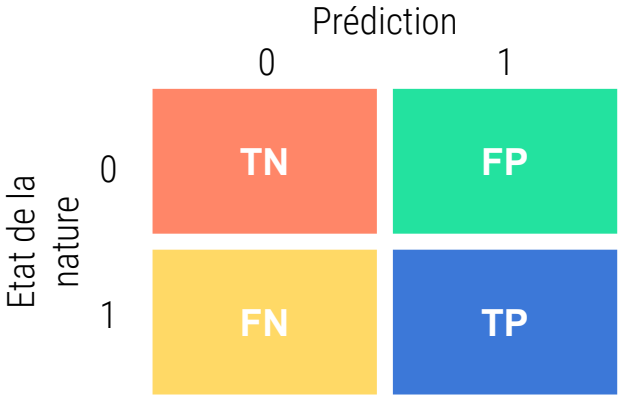
Validation croisée (& stratifiée) k-fold



Stratifiée : on conserve le même taux de ‘1’ dans chaque échantillon (~1%)



Scores



Accuracy = $\frac{TP + TN}{TP + FP + TN + FN}$

Recall* = $\frac{TP}{TP + FN}$

Precision = $\frac{TP}{TP + FP}$

**Recall = TPR = Sensitivity*

Score F1 = $\frac{2 \times precision \times recall}{precision + recall}$

Note : on utilisera aussi les courbes lift et ROC vues en cours

Recherche des meilleurs paramètres

Tableau de meilleurs résultats de GridSearchCV

Ranking	penalty	C	solver	precision	recall	f1
1	l1	0.01	liblinear	76.3%	84.08%	79.98%
2	l2	1	newton-cg	75.75%	82.17%	78.81%
3	none	0.01	newton-cg	75.7%	84.17%	78.78%

Paramètres utilisés :

1) Solver: ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’
2) Penalty: ‘none’, ‘l1’, ‘l2’, ‘elasticnet’
3) C: {0.001, 0.01, 0.1, 1, 10, 100}

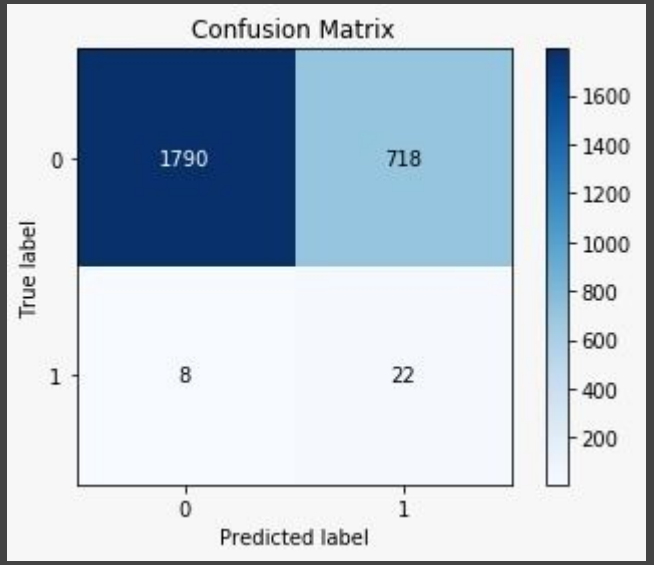
Tableau du lift cumulé

CLF : liblinear, l1 et C=0.01

alpha	effectif	nb de défaut	alpha lift	effectif cumulé	nb cumulé de positif	alpha lift cumulé
0.1	254	11	3.66	254	11	3.66
0.2	254	6	2.00	508	17	2.83
0.3	254	3	1.00	762	20	2.22
0.4	254	4	1.33	1016	24	2.00
0.5	254	4	1.33	1270	28	1.86
0.6	254	0	0,00	1524	28	1.55
0.7	254	0	0,00	1778	28	1.33
0.8	254	0	0,00	2032	28	1.16
0.9	254	2	0,66	2286	30	1,11
1.0	254	0	0,00	2540	30	1,00

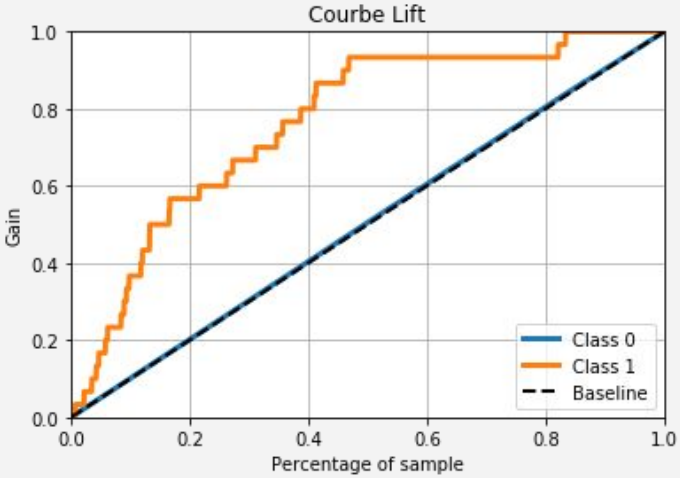
La performance

Meilleur clf : liblinear, l1 et C=0.01



Sur les données test :

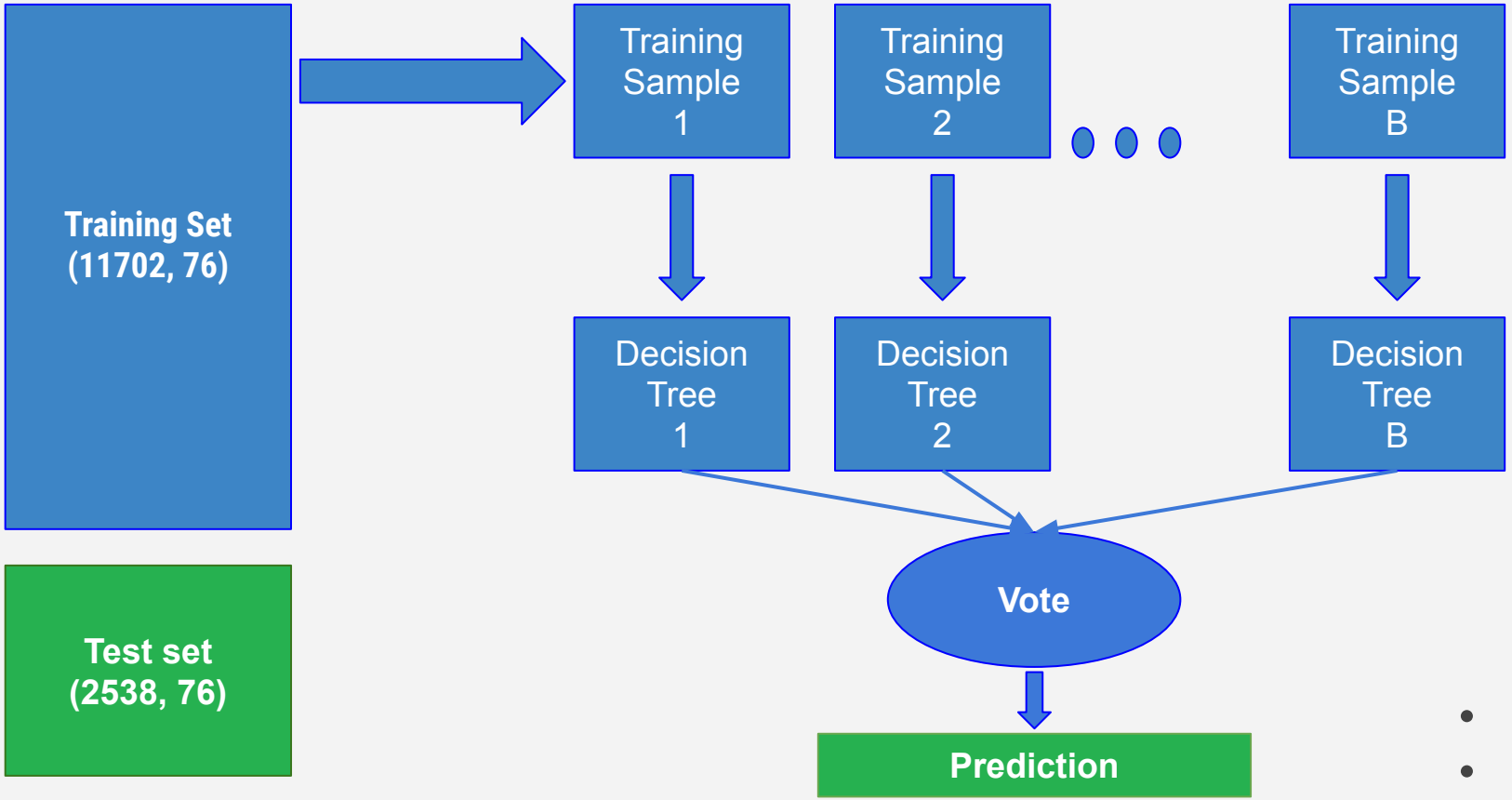
Accuracy : 71% **Recall :** 73%
Precision : 3% **f1 :** 2.33%



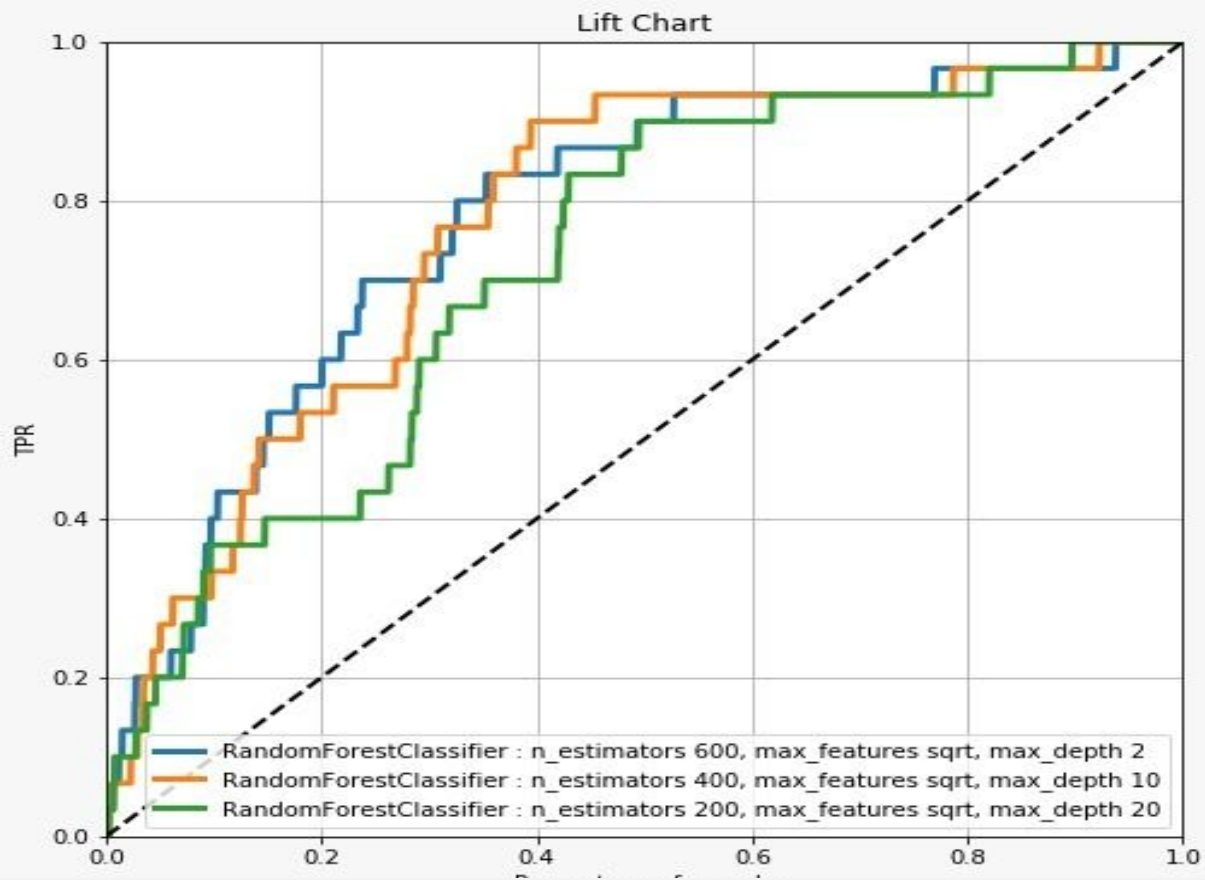
Basé les informations de notre modèle, les variables les plus importantes sont:

1. **Prix total/ Prix de véhicule** : -0.46, -0.23
2. **Situation de familiale** (celibataire): 0.24
3. **Mode de logement** (chez les parents): 0.2
4. **Age de client** (jeune): 0.15

Random Forest

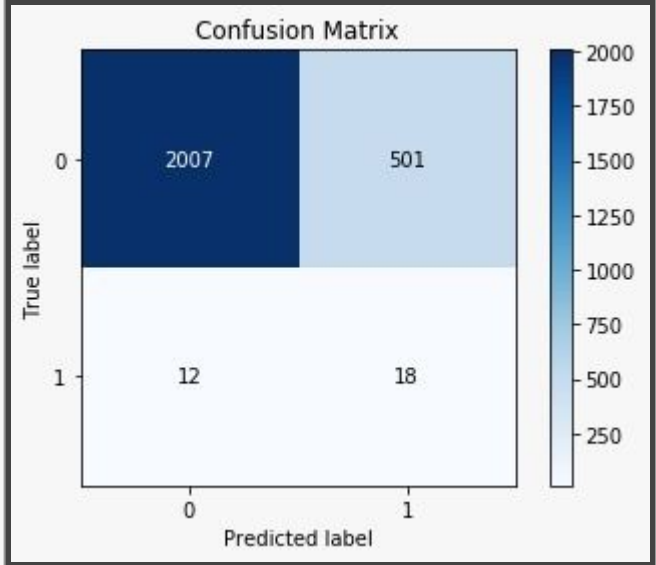


Hyperparameter Tuning: Random Forest



Meilleurs paramètres:

n_estimators: 600
max_features: sqrt
max_depth: 2

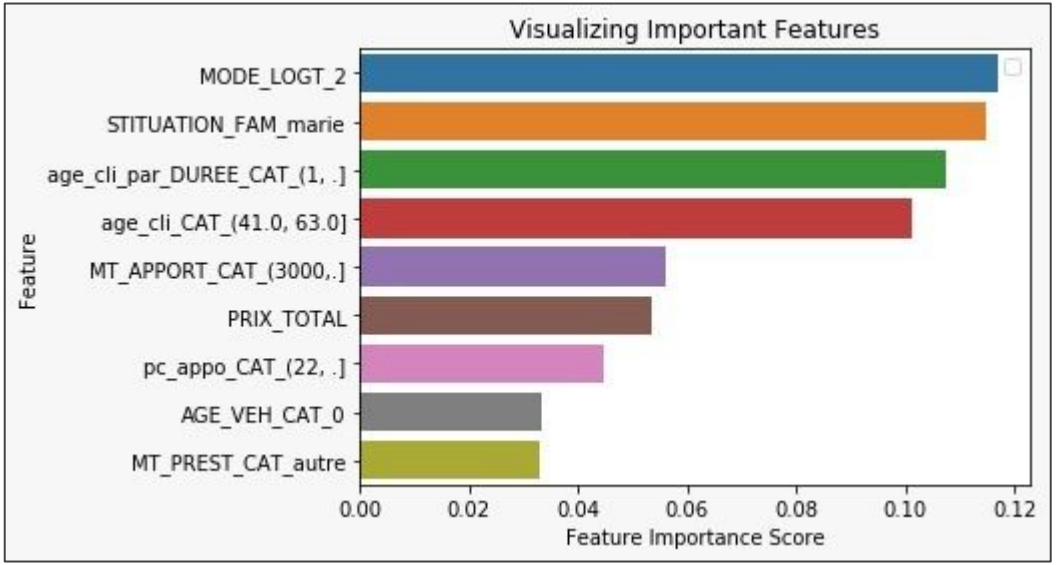


Importance des variables

Régression logistique

Variable	coefficient
Prix total	-0.46
Situation familiale: célibataire	0.24
Prix de véhicule	-0.23
Ancienneté rci	-0.21
Mode de logement: chez les parents	0.2
Age de client	0.15
Pourcentage d'apport	-0.14

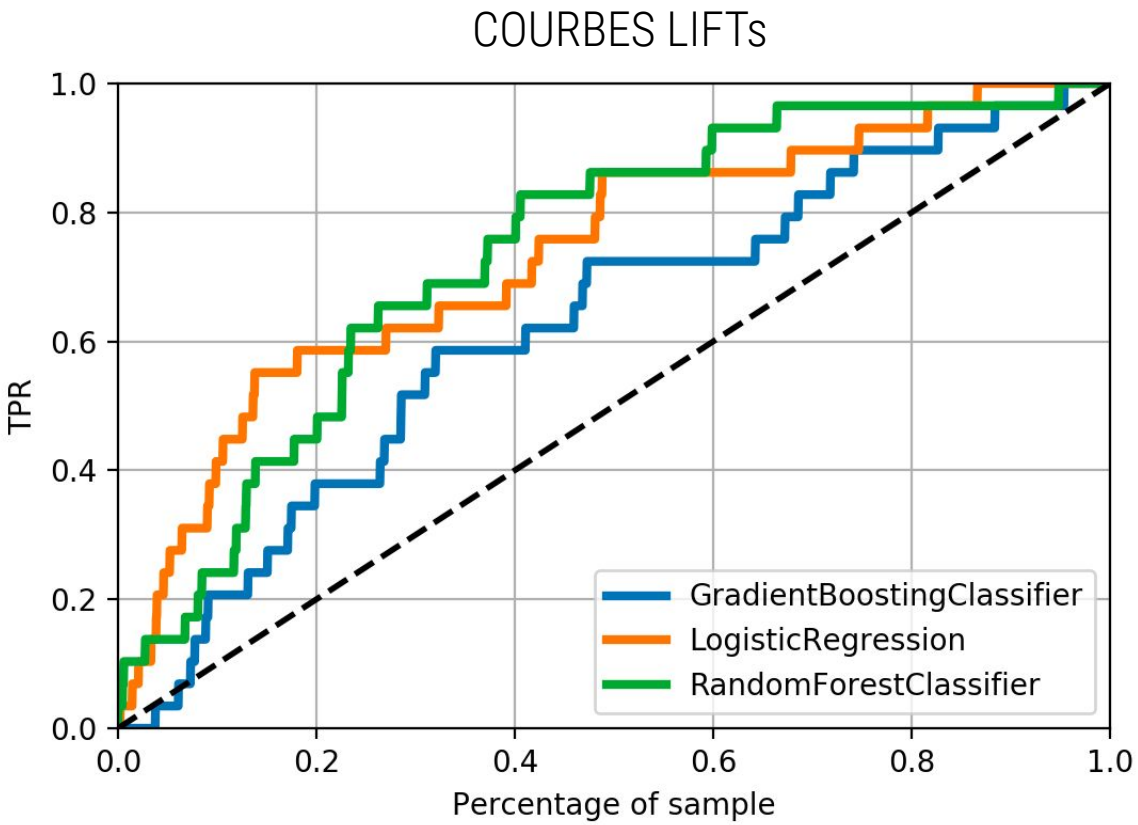
Random Forest



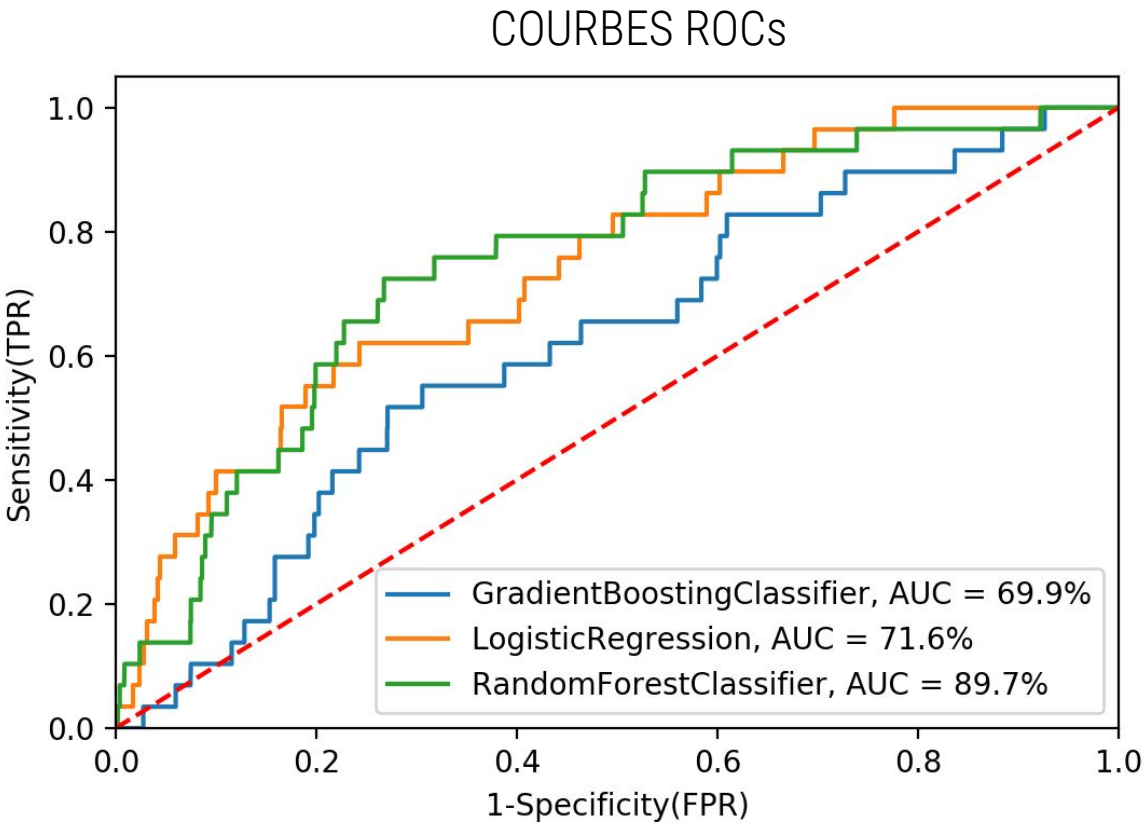
4 variables importantes qui apparaissent dans les 2 interprétations:

Situation familiale
Âge de client
Mode de logement
Pourcentage d'apport

Comparaison des modèles (1)



Comparaison des modèles (2)



Random Forest : tableau Lift

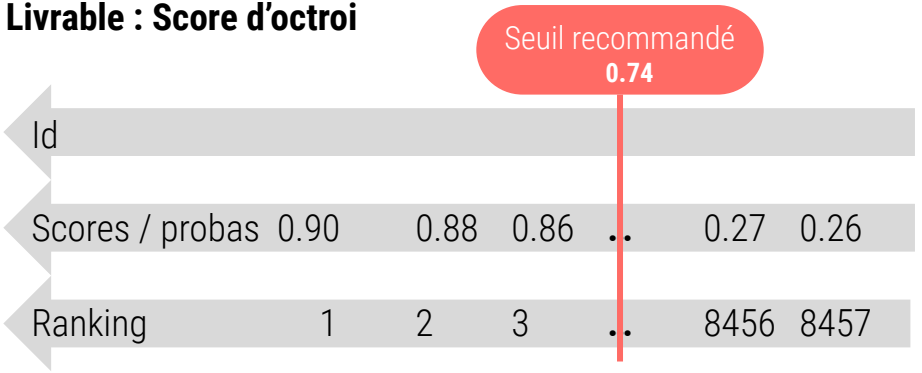
alpha	Nb de défauts	Alpha Lift	Effectif cumulé	Nb cumulé de défauts	Alpha Lift cumulé (Random. F)	Alpha Lift cumulé (Rég. Log)
0.1	12	4,00	254	12	**4,00**	**3.66**
0.2	5	1,67	508	17	2,83	2.83
0.3	4	1,33	762	21	2,33	2.22
0.4	4	1,33	1016	25	2,08	2.00
0.5	2	0,67	1270	27	1,80	1.86
0.6	1	0,33	1524	28	1,55	1.55
0.7	0	0,00	1778	28	1,33	1.33
0.8	1	0,33	2032	29	1,21	1.16
0.9	0	0,00	2286	29	1,07	1,11
1.0	1	0,33	2540	30	1,00	1,00

Décision et livrable

Aperçu du tableau lift : RF

alpha	Nb de défauts	Alpha Lift	Alpha Lift cumulé
0.1	12	4,00	4,00
0.2	5	1,67	2,83
0.3	4	1,33	2,33
0.4	4	1,33	2,08
0.5	2	0,67	1,80
0.6	1	0,33	1,55
0.7	0	0,00	1,33
0.8	1	0,33	1,21
0.9	0	0,00	1,07
1.0	1	0,33	1,00

Livrable : Score d'octroi



Aperçu

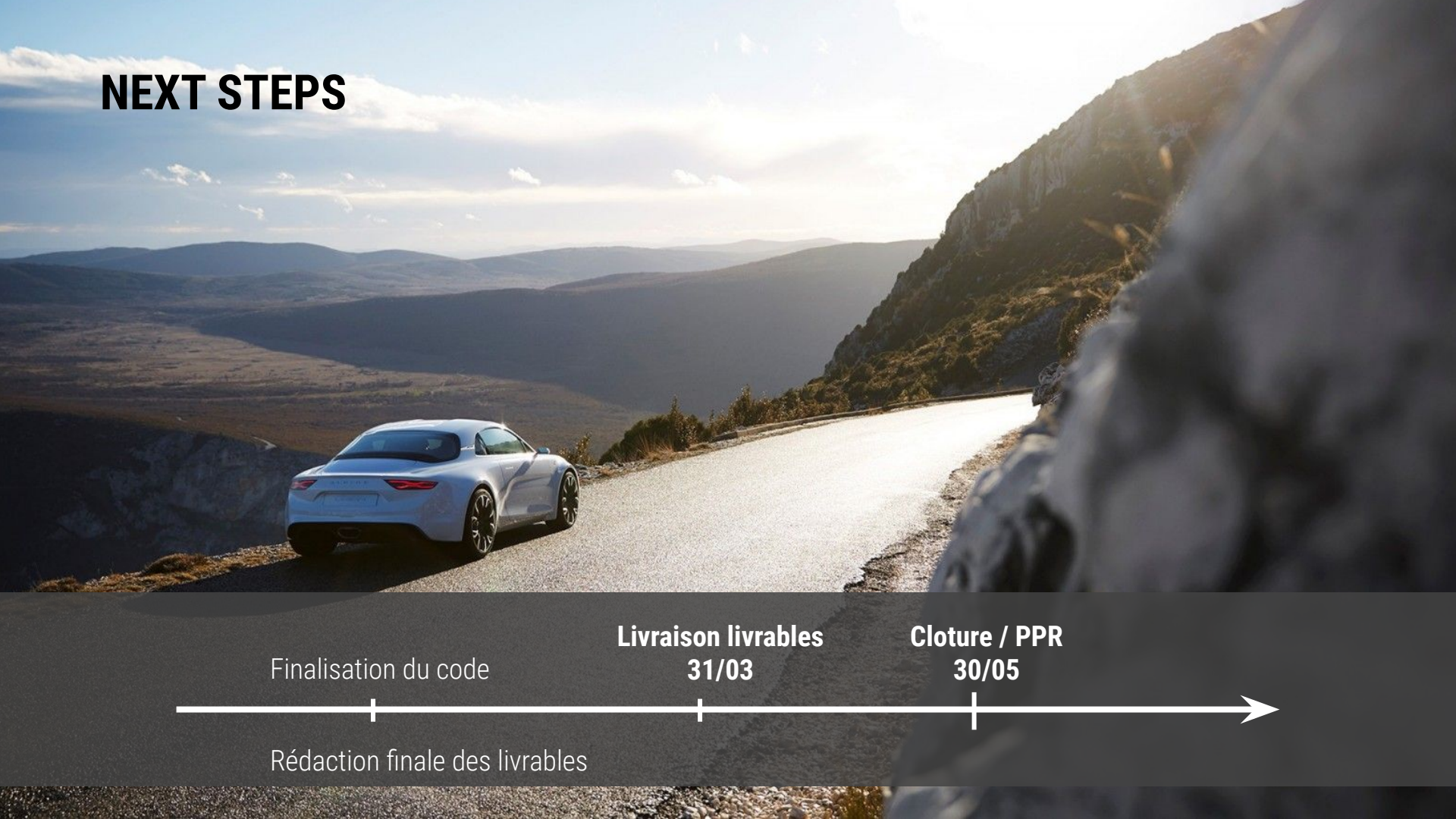
Client le plus favorable au crédit :

Score	Mode de Lgmt	Age	Situation Fam.	% apport
0.90	Propriétaire	59	Marié	60

Client le plus susceptible de faire défaut :

Score	Mode de Lgmt	Age	Situation Fam.	% apport
0.26	Chez les parents	22	Célibataire	15

NEXT STEPS


Finalisation du code
Livraison livrables
31/03
Cloture / PPR
30/05
Rédaction finale des livrables

Merci !