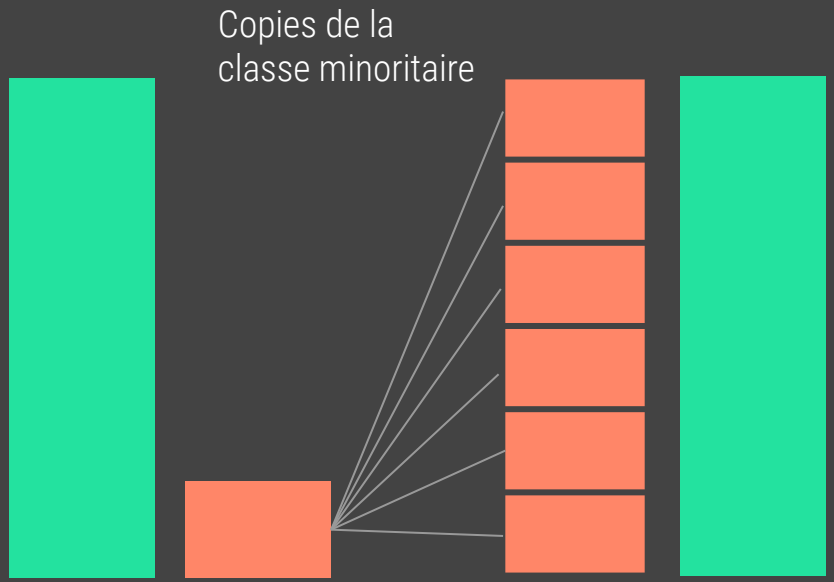
Etape 2 - Projet DS

Trieu Xoung YU
François LE GAC
Mercredi 26 Février

M2 ISIFAR
Promotion 2019 - 2020
Université PARIS DIDEROT

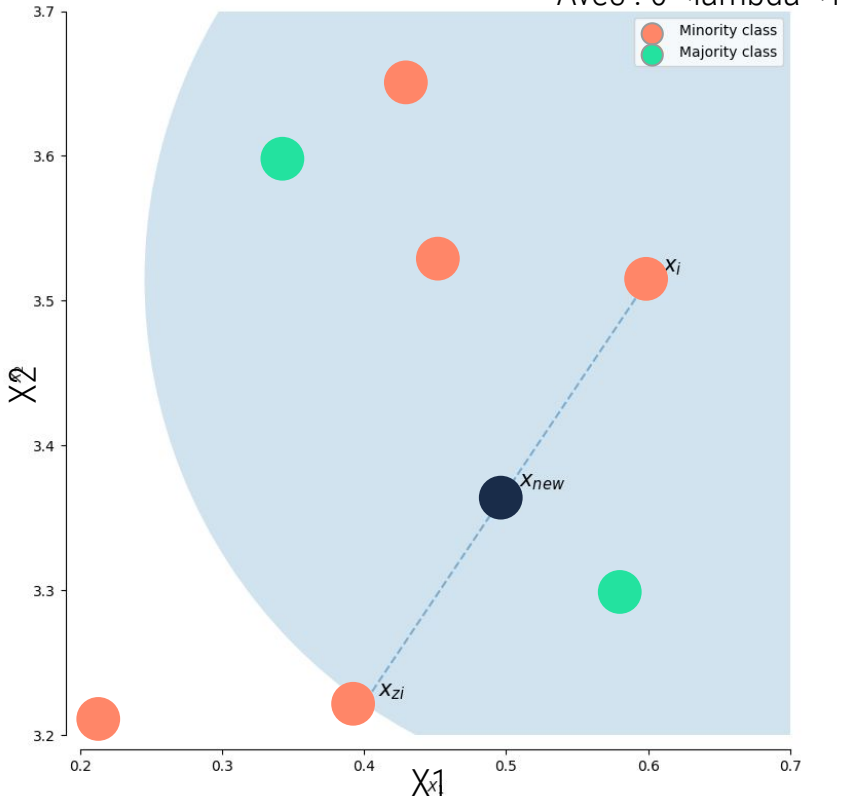
OVERSAMPLING



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

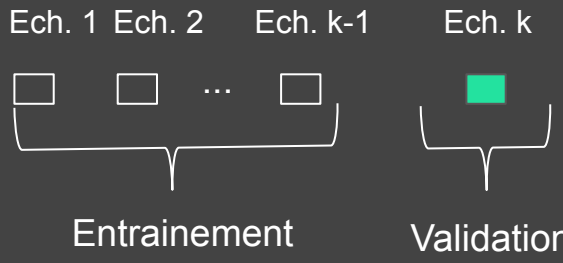
$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, x_i^\top w + b)$$

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

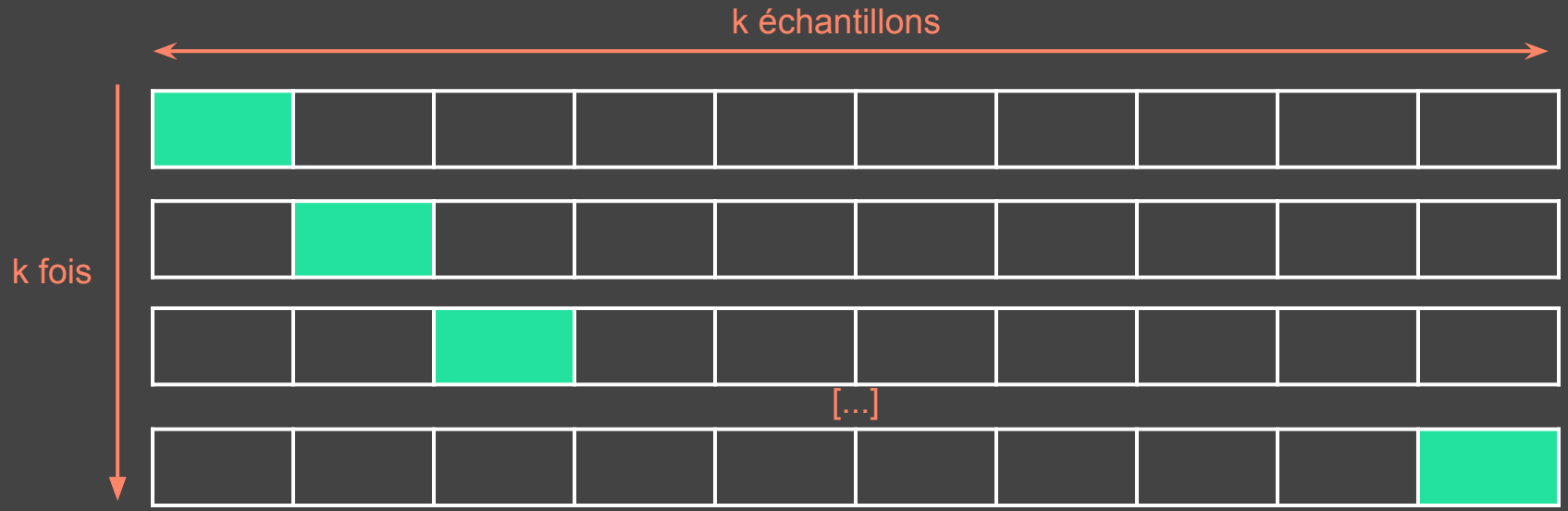
⟺ A minimiser $argmin(-\log(L(w, b)) + C * \|w\|)$

Moins log-vraisemblance **La pénalité: l1/l2**

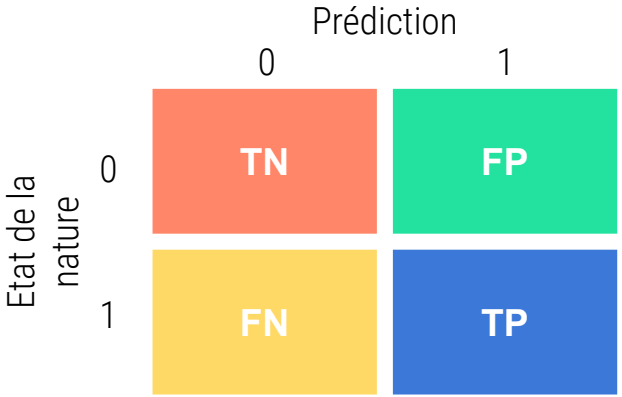
Validation croisée (& stratifiée) k-fold



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

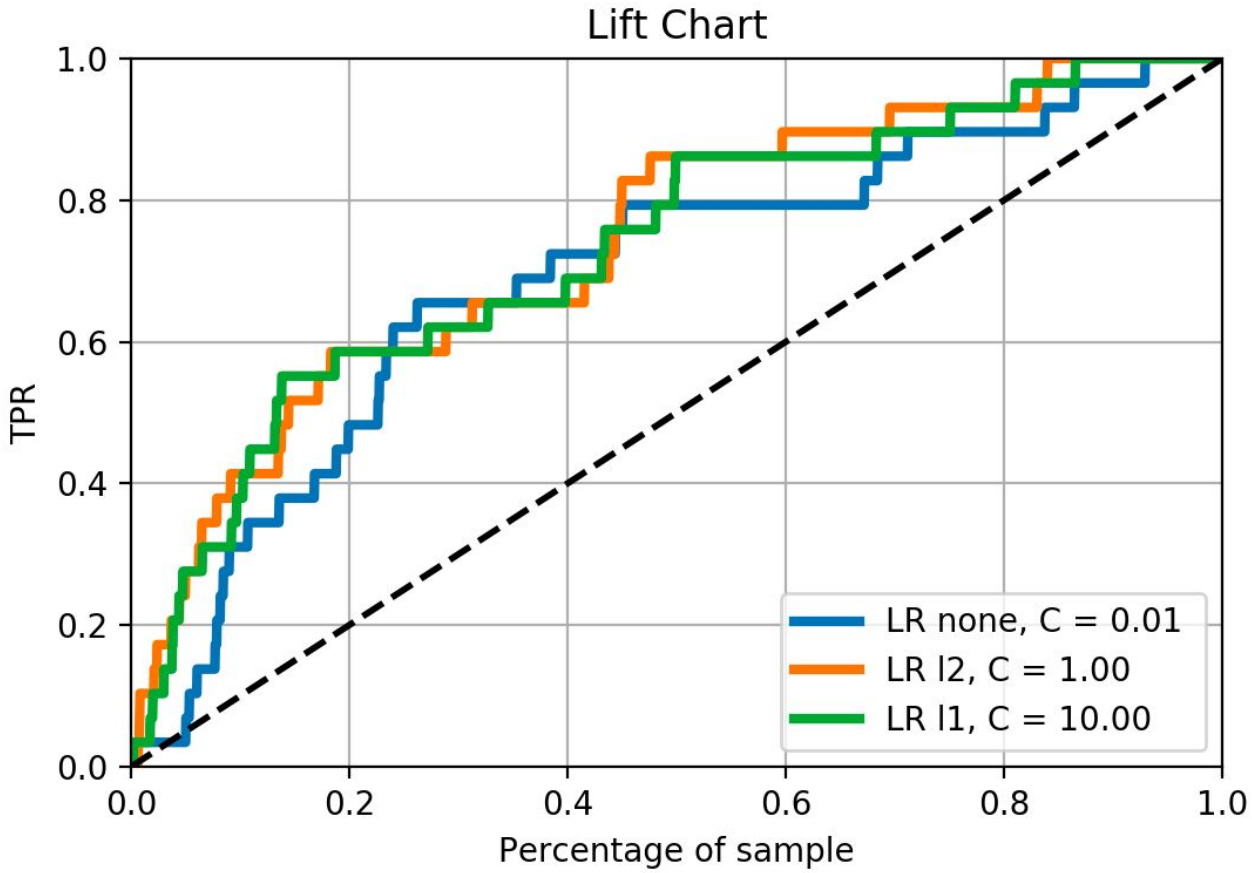
Recherche des meilleurs paramètres

Tableau de meilleurs résultats de GridSearchCV

Ranking	solver	penalty	C	precision	recall	f1
1	liblinear	l1	10	76.3%	84.08%	79.98%
4	newton-cg	l2	1	75.75%	82.17%	78.81%
6	newton-cg	none	0.01	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}



Scores calculés sur X_test

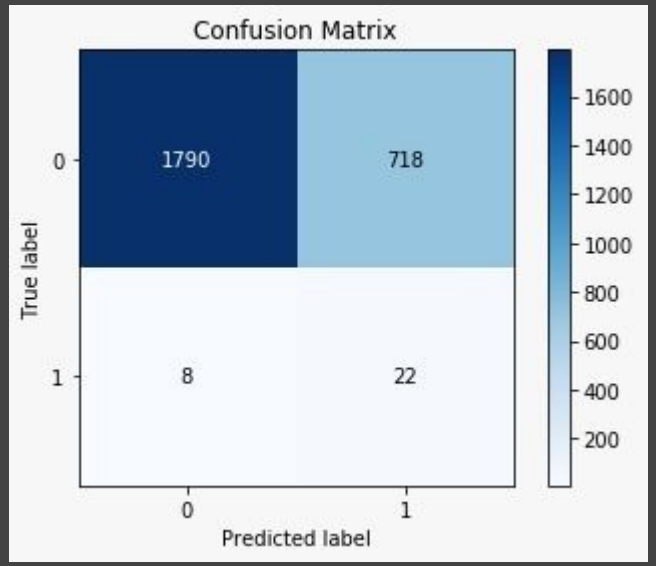
Tableau du lift cumulé

CLF : liblinear, l1 et C=10

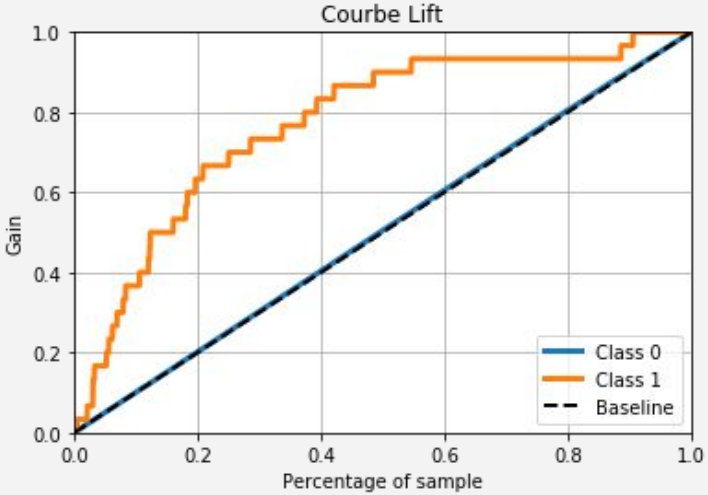
alpha	effectif	nb de défaut	alpha lift	effectif cumulé	nb cumulé de positif	alpha lift cumulé
0.1	254	12	4,13	254	12	4,13
0.2	254	5	1,72	508	17	2,93
0.3	254	1	0,34	762	18	2,07
0.4	254	2	0,69	1016	20	1,72
0.5	254	5	1,72	1270	25	1,72
0.6	254	0	0,00	1524	25	1,44
0.7	254	1	0,34	1778	26	1,28
0.8	254	1	0,34	2032	27	1,16
0.9	254	2	0,69	2286	29	1,11
1.0	254	0	0,00	2540	29	1,00

La performance

Meilleur clf : liblinear, l1 et C=10



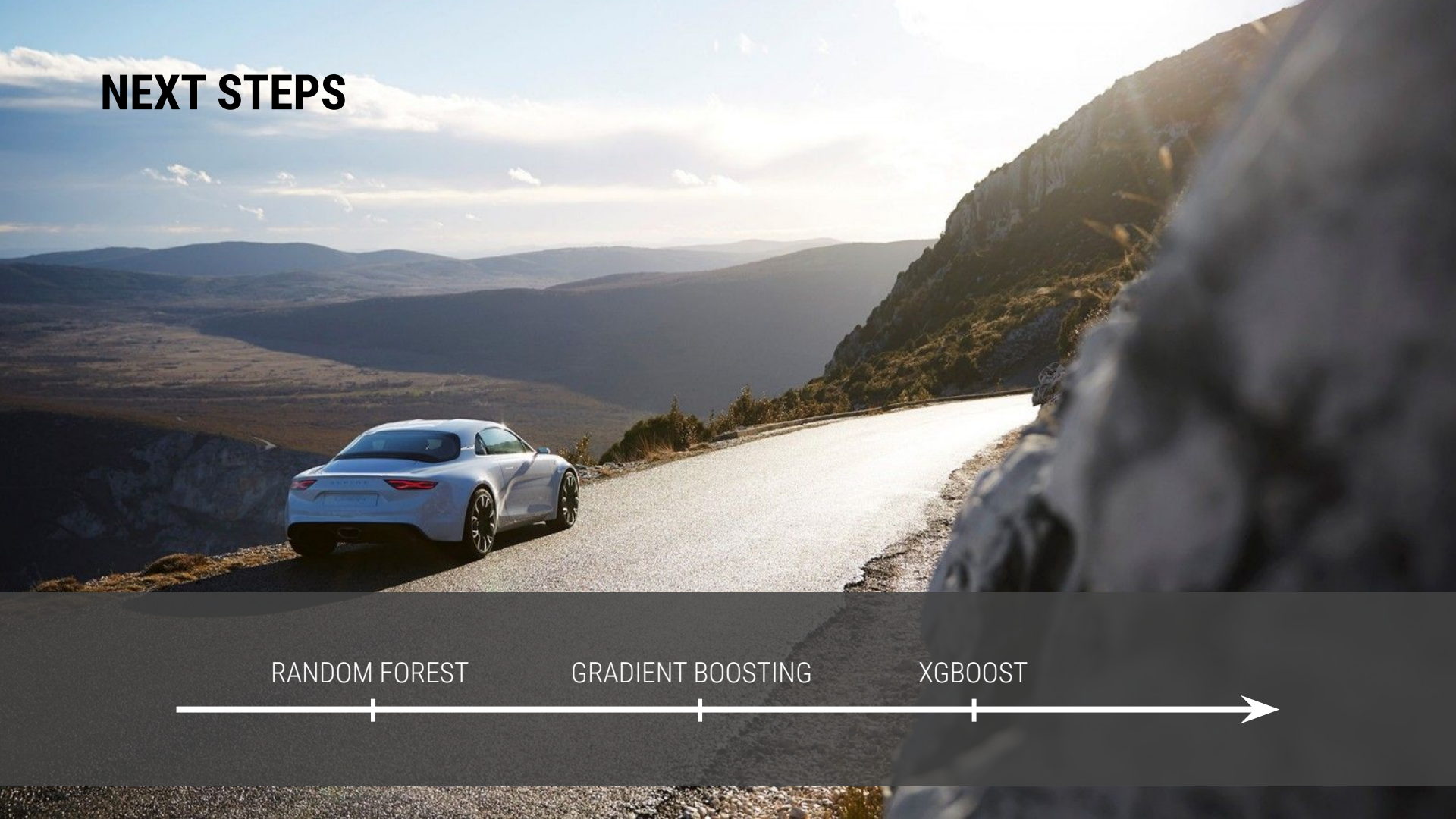
Accuracy : 71%
Precision : 3%
Recall : 73%
f1 : 2.33%



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

1. **Situation de familiale** (autre, celibataire): 0.505, 0.356
2. **Ancienneté à l'emploi** (0 à 5 ans, plus de 20 ans): 0.445, -0.376
3. **Mode de logement** (chez les parents): 0.285

NEXT STEPS



ANNEXE 1
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