Construction d’une grille de score

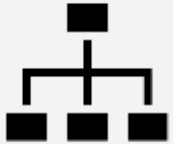
À partir de la base clients de LendingClub

BERRADA Abdoul Aziz – CAILLOSSE Morgane – HAMON Hugo – NIANG Mamad

Introduction

Objectif

Déterminer les clients à qui on souhaite prêter notre argent, pour ce faire, il s'agit de déterminer **leur probabilité de défaut bancaire**.



Moyen

Construction d'un modèle de credit scoring dans l'optique de **segmenter** la population en **classes de risques homogènes**.

Définition : grille de score

Une grille de score est un outil permettant de **noter** un individu en lui attribuant des points à partir de ses caractéristiques.

Des statistiques au métier

Points

→ La grille de score est calibrée sur **1000 points**

→ Plus un individu a de points moins il est **risqué**

Son rôle dans l'accord du prêt

1. Demande de prêt
2. Calcul des points à partir de la grille de score
3. Placement de l'individu dans une classe de risque
4.
   - Classe **peu risquée** → accord du prêt
   - Classe **très risquée** → refus d'accord du prêt
   - Classe **intermédiaire** → analyse de sa situation

→ Passerelle entre les statistiques et le métier

→ Gain de temps

→ Gain d'argent

Présentation des données

La période considérée

On considère uniquement les données à partir de 2010 :
→ La **crise des Subprimes** a entrainé des défauts à des profils n'étant pas supposés faire défaut.

La cible

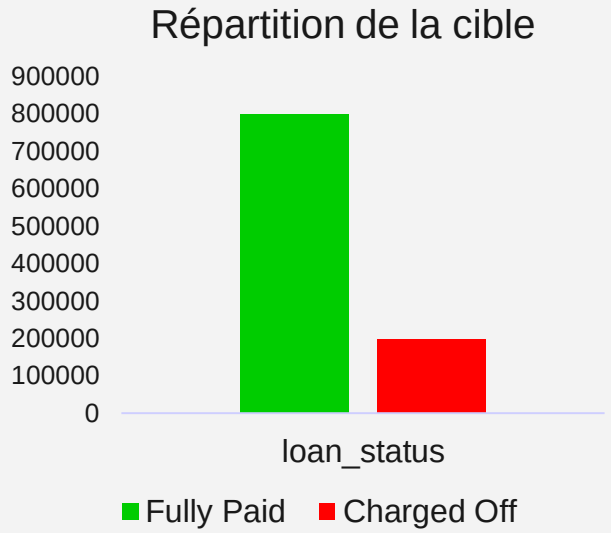
Les prêts entièrement payés et en défauts.

Caractérisation de la cible

Défaut = Default + Does not meet the credit policy. Status : Charged Off + Charged Off

Non défaut = Fully paid + Does not meet the credit policy. Status : Fully paid

Moins de 20% de défaut.



Choix des variables

Premier tri

Suppression des variables avec plus de 40% de valeurs manquantes, des variables floues, non pertinentes et celles dont on ne dispose pas au moment de l'octroi du crédit.

→ On conserve **33 variables.**

Valeurs manquantes

→ **Séparation** du train en 2 selon la modalité de la cible

Variables continues :

→ On **impute** la **médiane** si la distribution est asymétrique ;

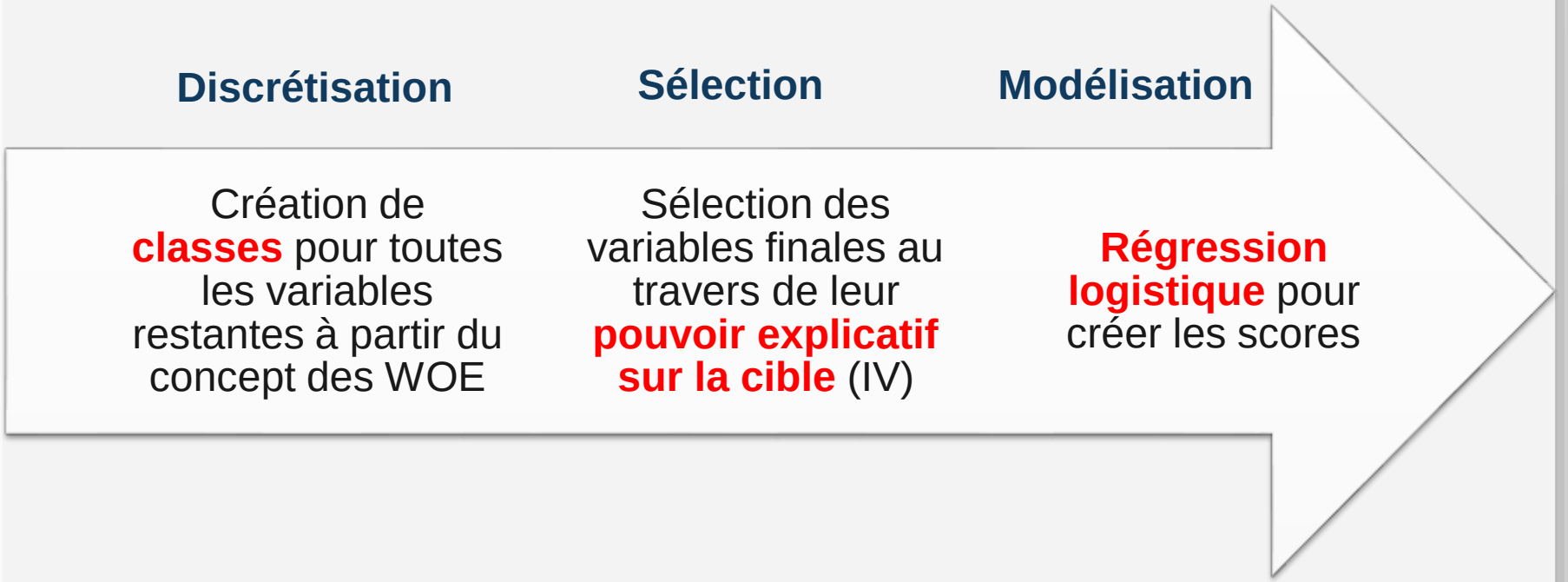
→ Sinon, on impute la **moyenne**

Variables catégorielles :

→ On impute la valeur **modale** si moins de 10% de valeurs manquantes ;

→ Sinon on crée une **nouvelle catégorie**

Choix des variables finales et modèle



Grille de score : exemple de résultats

Ratio d'endettement		Montant du prêt	
Classes	Points	Classes	Points
[-inf, 9[	24	[-inf, 4000[	18
[9, 12[	18	[4000,10000[	12
[12, 15[	12	[10000, 11000[	2
[15, 18[	5	[11000, 15000[	-4
[18, 21[	0	[15000, 16000[	-2
[21, 25[	-8	[16000, 20000[	-11
[25, 30[	-17	[20000, 29000[	-7
[30, inf[	-31	[29000, inf[	-13

Transformation des coefficients de la régression logistique

→ **17 variables** construisent la grille de score.

→ Plus l'individu a **un montant d'emprunt élevé** plus il est **risqué**.

→ Plus l'individu a **un ratio d'endettement élevé** plus il est **risqué**.

Classes de risque

Très risqué

Peu risqué

Classes de risque	Effectif	Probabilité de défaut
[550,650[	6072	0.66
[650,700[	39656	0.49
[700,750[	122026	0.35
[750,800[	240145	0.23
[800,850[	237419	0.13
[850,900[	125102	0.07
[900,1100]	65217	0.02

Avis d'un **spécialiste métier** pour décider de l'accord du prêt.

Challenger le modèle

Modèles\Métriques	F1-score	AUC	Recall	Accuracy
Régression logistique	✓	✓		✓
Decision Tree				
Random Forest			✓	
Gradient Boosting	✓			

→ La régression logistique apparaît comme étant le meilleur modèle car il permet d'arbitrer entre **interprétabilité** et **performance**.

Conclusion

Un pouvoir prédictif

Permet d'obtenir de bonnes performances de prédictions.

Un modèle simple

Permet d'attribuer des classes de risques aux individus simplement et rapidement.

Un modèle pratique

Peut être utilisé par des experts métiers **sans connaissance** particulière en Data Science.

Résultat du modèle

AUC : 0.682

Matrice de confusion :

118634	58962
19772	28271

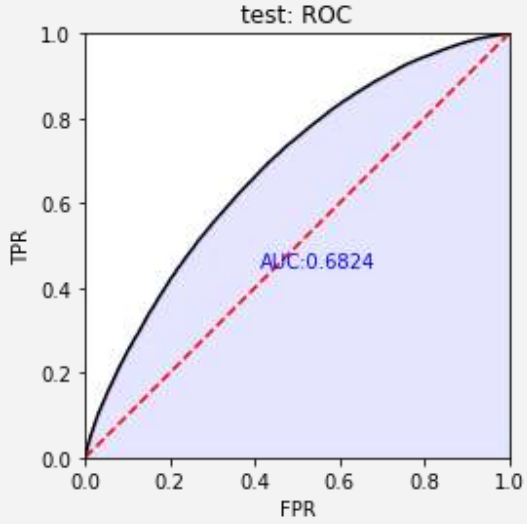
Classification \ report :

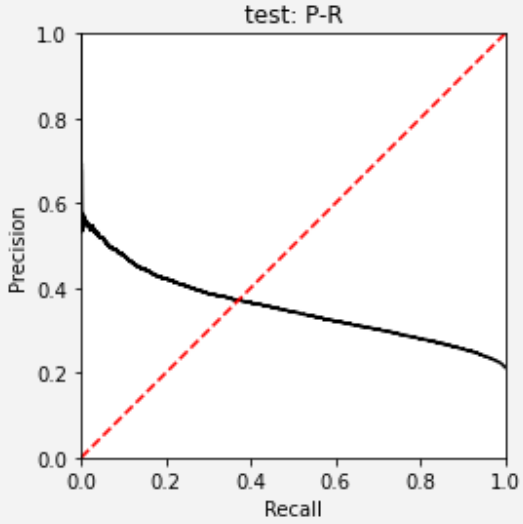
	precision	recall	f1-score	support
0	0.86	0.67	0.75	177596
1	0.32	0.59	0.42	48043
accuracy			0.65	225639
macro avg	0.59	0.63	0.58	225639
weighted avg	0.74	0.65	0.68	225639

Precision-Recall et Roc Curve

→ Une courbe ROC (Receiver Operating Characteristic) représente les **performances** d'un modèle de classification pour tous les seuils de classification.

→ Cette courbe trace le taux de **vrais positifs** en fonction du taux de **faux positifs**.





Point to Double the Odds

- **Base 1000**
- **Odds0 : Target odds = p/(1-p) = 1/19**
- **PD0=50**

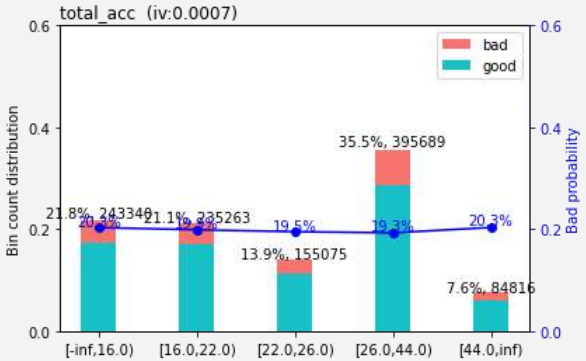
Nous attribuons une certaine signification à 1000 par exemple, on considère que 1000 points la probabilité de défaut est de 1/19. Typiquement, un saut de 50 points signifie un doublement des odds values, par exemple 1050 signifie que la probabilité de défaut est de 1/38.

WOE et IV

Weight Of Evidence

→ Formule : **WOE = ln(**$\frac{\%\ défaut}{\%\ non-défaut}$**)**

→ Outil de **discrétisation** : une variable doit présenter des modalités avec des probabilités de défaut différentes pour que chaque classe explique le risque de façon complémentaire.

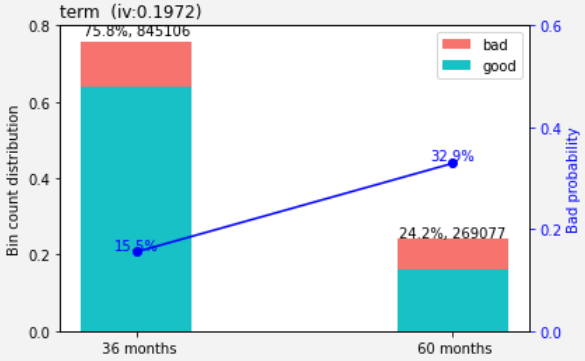


Mauvaise variable

Information Value

→ Donne une information sur le **pourvoir explicatif** d'une variable et de ses modalités au regard de la cible.

→ **Critère de sélection** des variables :
IV > 0.02



Bonne variable