Projet d'étude "Crédit à la consommation"

Ekuba Kabuiku Bénicia

25 Mars 2019

Notre base des données "Crédit-Consommation" contient 41 variables et 10000 observations.
Parmi ces variables,certaines ont été enlevées car elles sont des facteurs mais aussi redondantes.
Et nous avons également additionné certaines variables qui représentaient chacune d'elle la même chose
(" $recovery, sub_g rade$ ").
Ainsi notre base des données finale ne contient plus que 34 variables avec des données manquantes soit 14.

Nous avons premièrement réalisé un découpage $\frac{2}{3}$ / $\frac{1}{3}$ pour nos échantillons test($\frac{1}{3}$) et apprentissage($\frac{2}{3}$) ; puis deuxièment une imputation par knn en utilisant le package "VIM" pour enlever nos données manquantes.

LDA au seuil de 5%

		Négatif	Positif	Total
Réalité	Négatif	2578	93	2671
	Positif	482	180	662
	Total	3060	273	3333

LDA au seuil de 2%

		Négatif	Positif	Total
Réalité	Négatif	2025	646	2671
	Positif	218	444	662
	Total	2243	1090	3333

Le taux d'erreur global est de 0.2592 pour la LDA.

```
Prior probabilities of groups:
        N         Y
0.8107095 0.1892905
```

On en conclut que 81% d'observations ne sont pas BadLoan alors que 18% d'observations

Prévision QDA

		Négatif	Positif	Total
Réalité	Négatif	2009	662	2671
	Positif	219	443	662
	Total	2228	1105	3333

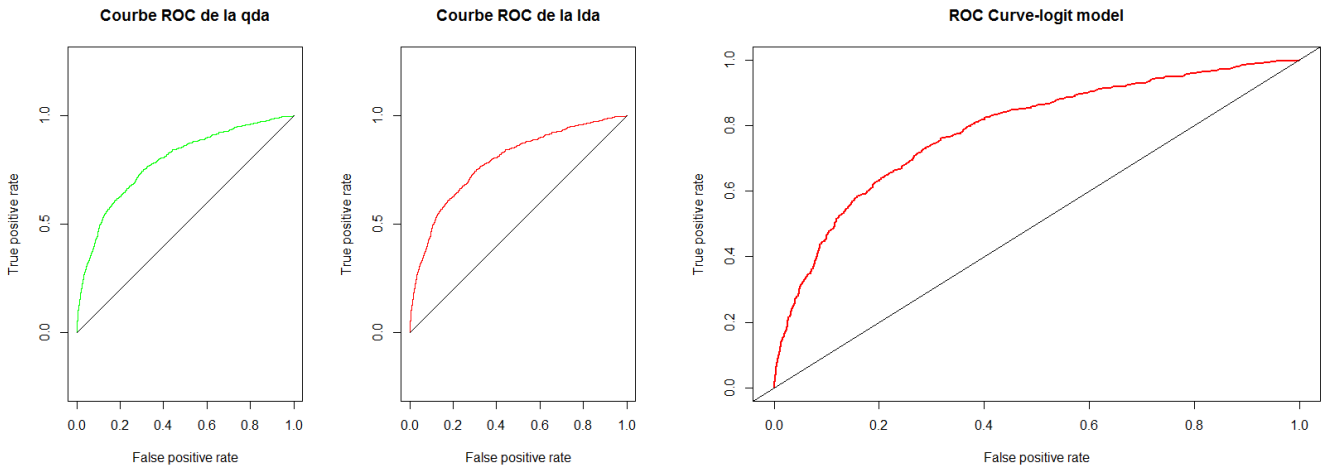
le taux d'erreur global est de 0.2643 pour la QDA

		Prévision Logit		
		Négatif	Positif	Total
Réalité	Négatif	1984	687	2671
	Positif	205	457	662
	Total	2189	1144	3333

Le taux d'erreur global est de 0.2676 pour le Logit

TABLE 1 – Modèle de régression logistique

	Dependent variable :
	BadLoan
Dim.1	-0.204^{***}
	(0.017)
Dim.2	-0.219^{***}
	(0.021)
Dim.3	0.302^{***}
	(0.023)
Dim.4	0.635^{***}
	(0.026)
Dim.5	0.527^{***}
	(0.029)
Constant	-1.908^{***}
	(0.043)
Observations	6,667
Log Likelihood	−2,526.545
Akaike Inf. Crit.	5,065.089
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01



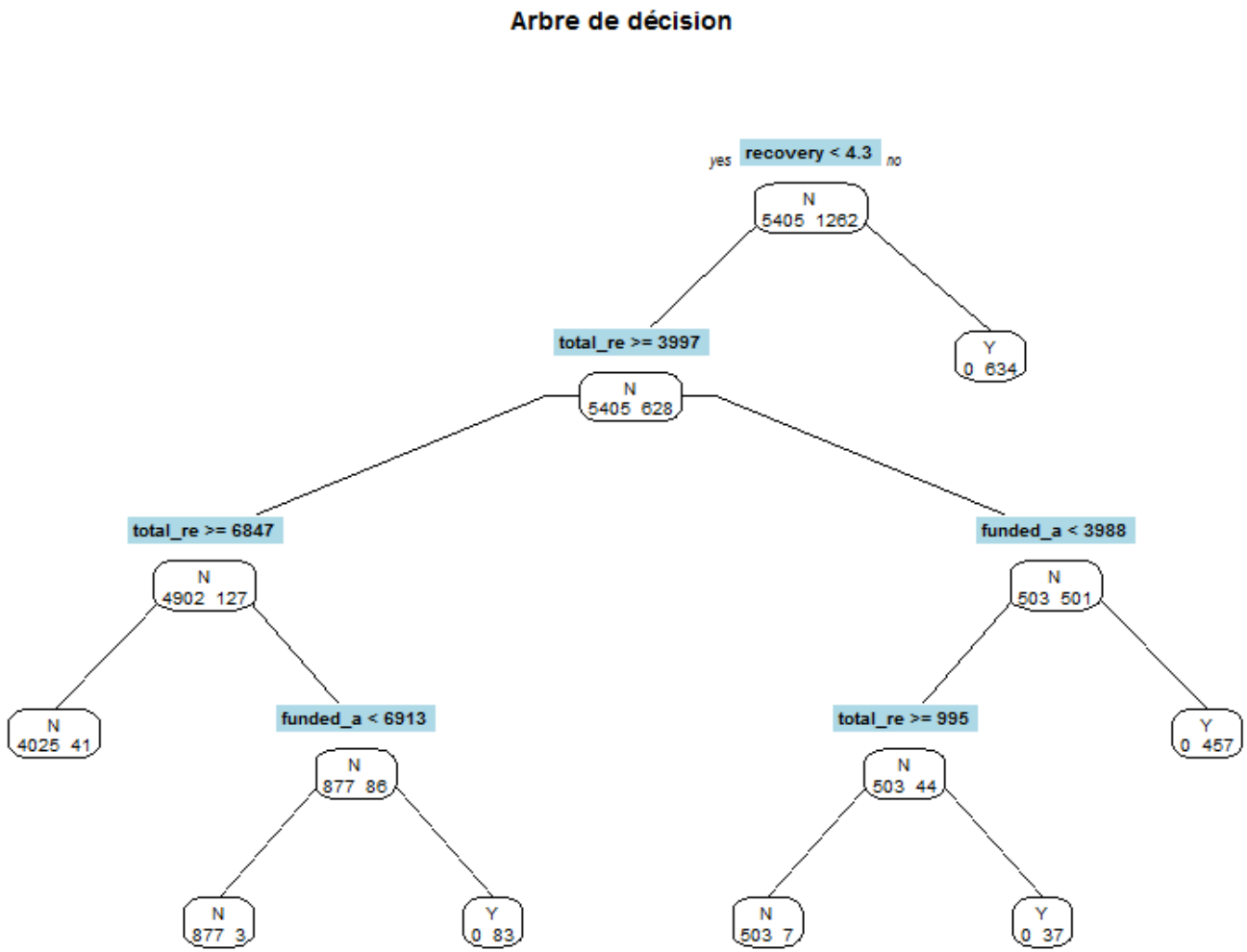
L'aire sous la courbe ROC de la LDA est de 0.7868, de 0.7860 pour la QDA et de 0.7862 pour le logit.

		Prévision Knn		
		Négatif	Positif	Total
Réalité	Négatif	2493	178	2671
	Positif	356	306	662
	Total	2849	484	3333

Le taux d'erreur global est de 0.1602 pour la knn.

		Prévision Arbre		
		Négatif	Positif	Total
Réalité	Négatif	2671	40	2711
	Positif	0	622	622
	Total	2671	662	3333

Le taux d'erreur global est de 0.012 pour l'arbre de décision.


Arbre de décision
yes
recovery < 4.3
no
N
5405 1262
total_re >= 3997
Y
0 634
N
5405 628
total_re >= 6847
funded_a < 3988
N
4902 127
N
503 501
N
4025 41
funded_a < 6913
total_re >= 995
Y
0 457
N
877 86
N
503 44
N
877 3
Y
0 83
N
503 7
Y
0 37

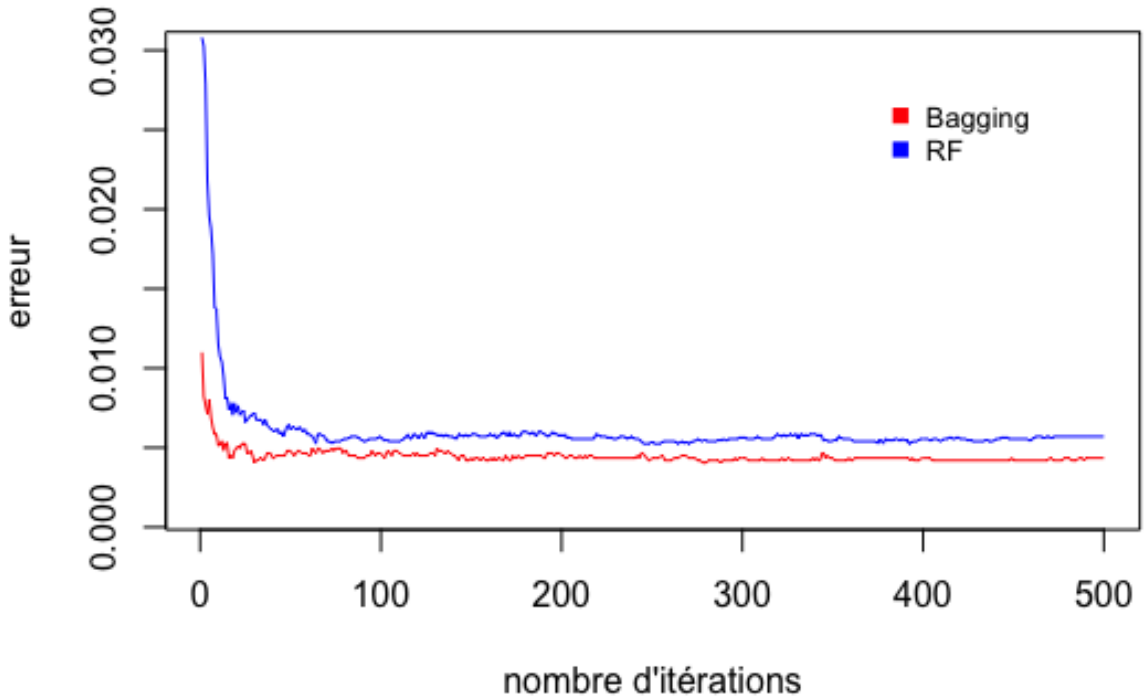
Prévision RF

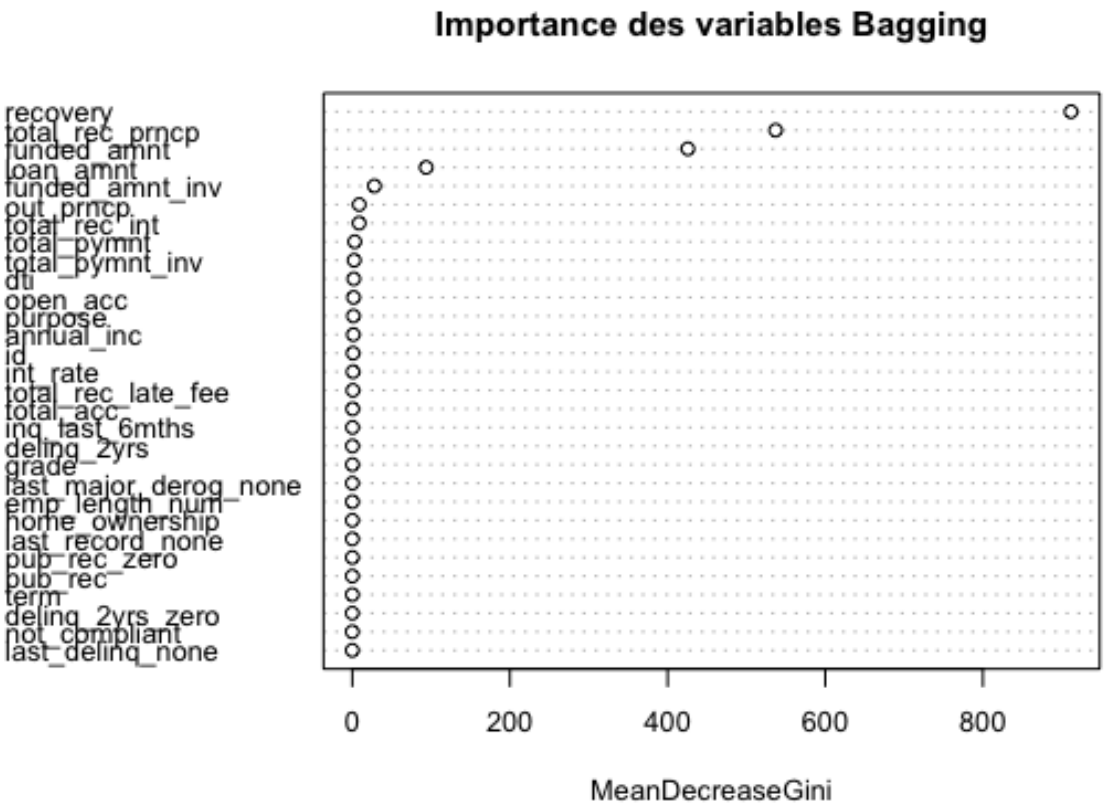
		Négatif	Positif	Total
Réalité	Négatif	2667	14	2681
	Positif	4	648	652
	Total	2671	662	3333

Prévision Bagging

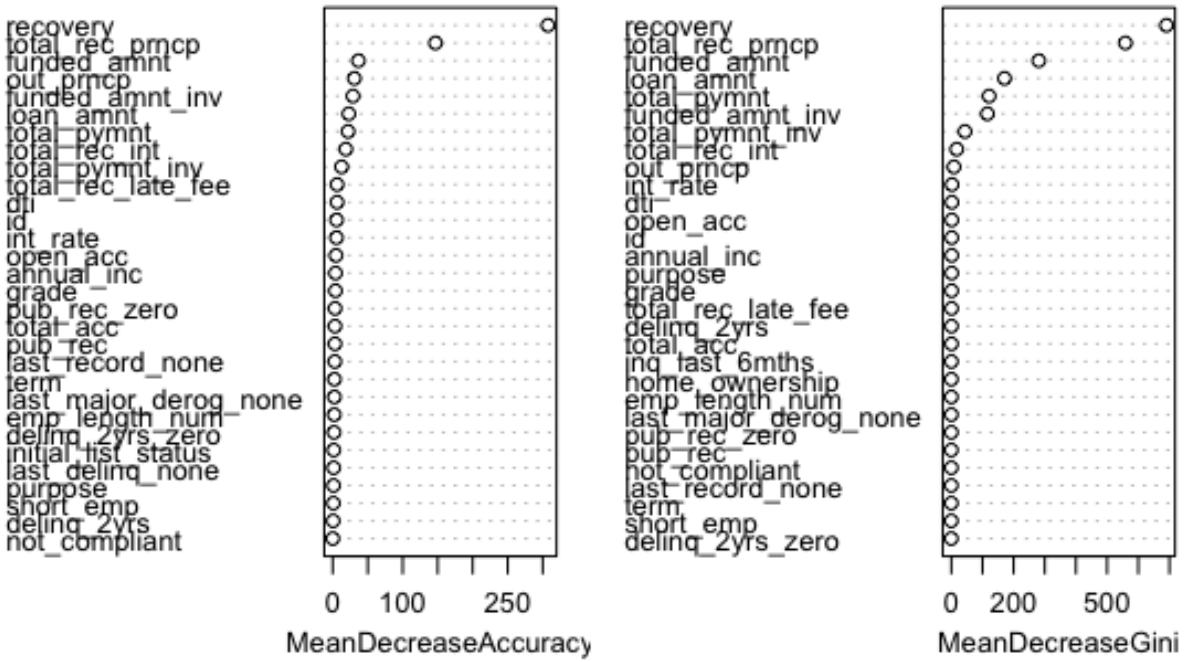
		Négatif	Positif	Total
Réalité	Négatif	2665	651	3316
	Positif	6	11	17
	Total	2671	662	3333

Le taux d'erreur global est de 0.0054 pour la forêt aléatoire et de 0.0051 pour le Bagging.


erreur
0.000
0.010
0.020
0.030
0
100
200
300
400
500
nombre d'itérations
Bagging
RF


Importance des variables Bagging
recovery
total_rec_prncp
funded_amnt
loan_amnt
funded_amnt_inv
out_prncp
total_rec_int
total_pymnt
total_pymnt_inv
dti
open_acc
purpose
annual_inc
id
int_rate
total_rec_late_fee
total_acc
inq_last_6mths
delinq_2yrs
grade
last_major_derog_none
emp_length_num
home_ownership
last_record_none
pub_rec_zero
pub_rec
term
delinq_2yrs_zero
not_compliant
last_delinq_none
0
200
400
600
800
MeanDecreaseGini

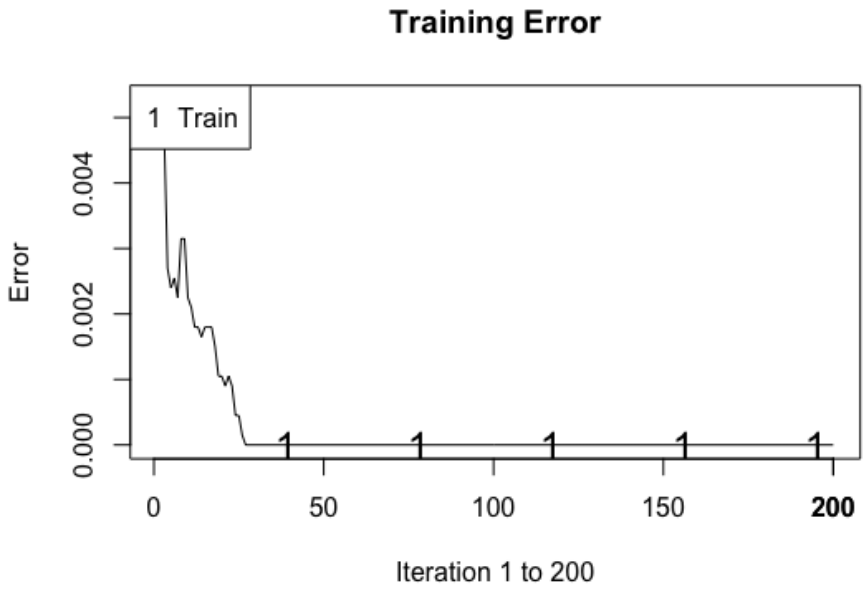
Importance des variables Random Forest

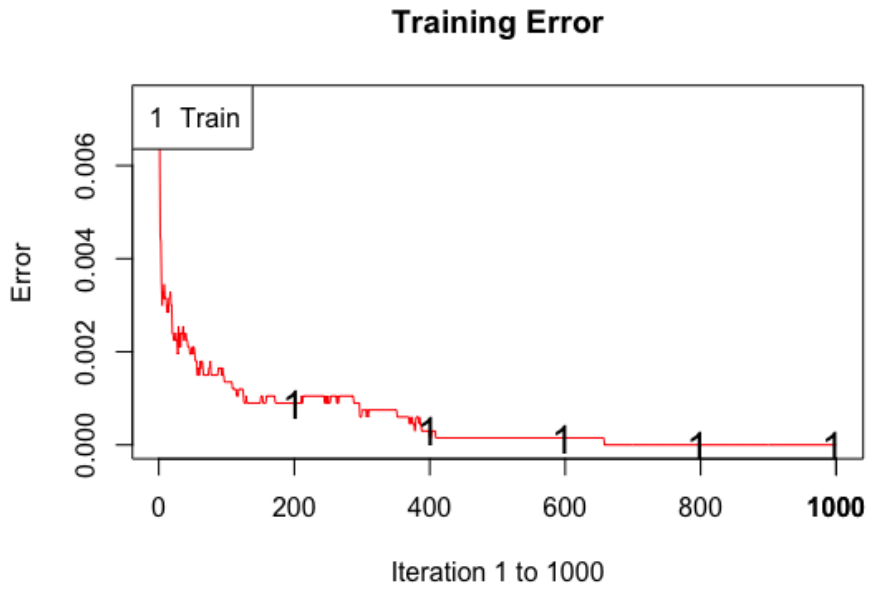


Prévision Boosting

		Négatif	Positif	Total
Réalité	Négatif	2671	649	3320
	Positif	0	13	13
	Total	2671	662	3333

Le taux d'erreur global est de 0.0039 pour le Boosting.





Après notre étude et au vu des résultats,la lda a été retenue comme le meilleur modèle parmi les paramétriques ; et le Boosting a été retenu comme le meilleur modèle parmi les non paramétriques.