Olivier CHEVALLIER
ochevall@gmail.com

STA212 - Rapport Projet

Authentification de billets de banque

STA212 2021/2022

INTRODUCTION

L'objectif du projet est la mise en place d'un modèle de prédiction de faux billets à partir de 4 variables quantitatives dont la technique de recueil est décrite dans le premier chapitre. Les variables explicatives étant assez abstraites et peu interprétables, la priorité sera mise sur les performances en prédiction des modèles et non pas sur leurs possibilités d'interprétation.

Nous utiliserons pour cela 2 méthodes vues lors de l'enseignement sta212 : Régression logistique binaire et forêts aléatoires.

La limitation de cette étude est la méthodologie de recueil des faux billets : s'agit-il de billets issus de la collecte sur plusieurs années, de lot et de "producteurs" différents ? N'ayant pas plus d'informations sur cette collecte et sur l'origine de chaque faux billet, nous ferons l'hypothèse que les individus (les faux billets) sont indépendants et que le résultat du modèle prédictif obtenu est généralisable.

Le projet sera réalisé avec le logiciel R. Toutes les manipulations effectuées dans cette étude sont disponibles dans le fichier sta212.R joint à ce rapport.

DESCRIPTION DES DONNEES

Le jeu de données étudié dans ce rapport provient du Machine Learning repository de l'université de Californie à Irvine. Lien direct vers la page du jeu de données : https://archive-beta.ics.uci.edu/ml/datasets/banknote+authentication

Chaque individu représente un billet de banque (vrai ou faux). Le jeu de données en contient 1372 et il n'y a aucune donnée manquante. Les individus sont supposés être indépendants. Le fichier contient pour chaque individu 5 variables, 4 quantitatives et une qualitative binaire.

Afin d'obtenir les 4 premières variables du jeu de données, les billets ont été photographiés, puis une transformation en ondelletes (Wavelet transform) a été appliquée à chaque photo. Les valeurs des variables VARIANCE, SKEWNESS, CURTOSIS et ENTROPY correspondent pour chaque billet aux caractéristiques de l'ondelette obtenue à partir de la photographie digitale.

variance :	variance de l'ondelette obtenue après transformation de la photo d'un billet
skewness :	coefficient d'asymétrie de l'ondelette
curtosis :	coefficient d'aplatissement de l'ondelette
entropy :	entropie de l'ondelette

Il s'agit de variables quantitatives continues, à valeurs positives et négatives, observées. Ce sont les variables explicatives de notre modèle.

La cinquième et dernière variable FORGED est une variable qualitative binaire, imposée, et qui prend les valeurs suivantes :

0 s'il s'agit d'un vrai billet de banque
1 s'il s'agit d'un faux

Il s'agit de la variable expliquée du projet.

Les données sources sont disponibles dans le fichier data_banknote_authentication.txt.

METHODOLOGIE DU PROJET

RISQUE

Le risque défini pour ce projet est de 5%.

MESURE DE LA PERFORMANCE DES MODÈLES

Les 2 modalités de la variable cible étant relativement équilibrées au sein des différents jeux de données, nous utiliserons la métrique **précision** (appliquée aux prédictions sur le jeu de données de validation) afin de comparer les performances des modèles et sélectionner le modèle prédictif le plus efficace.

ECHANTILLONS

Pour les besoins du projet l'échantillon initial sera découpé en 3, en respectant la répartition initiale des modalités de la variable cible (44% de faux billets) :

- L'échantillon d'entrainement (50%) sera utilisé pour entrainer les modèles
- Celui de validation (25%) pour mesurer la performance des modèles
- Celui de test (25%) pour publication des résultats du modèle choisi

ANALYSE DES DONNEES

Dans cette partie du projet nous allons étudier en détail chacune des variables du projet dans le but de valider la pertinence d'un modèle prédictif, et d'identifier les éventuels pré-traitements nécessaires aux méthodes choisies. Le jeu de données ne comporte pas de valeurs manquantes.

Analyse univariée

VARIABLES EXPLICATIVES

VARIANCE	SKEWNESS	CURTOSIS	ENTROPY
MIN. :-7.0421	MIN. :-13.773	MIN. :-5.2861	MIN. :-8.5482
1ST QU.:-1.7730	1ST QU.: -1.708	1ST QU.:-1.5750	1ST QU.:-2.4135
MEDIAN : 0.4962	MEDIAN : 2.320	MEDIAN : 0.6166	MEDIAN :-0.5867
MEAN : 0.4337	MEAN : 1.922	MEAN : 1.3976	MEAN :-1.1917
3RD QU.: 2.8215	3RD QU.: 6.815	3RD QU.: 3.1793	3RD QU.: 0.3948
MAX. : 6.8248	MAX. : 12.952	MAX. :17.9274	MAX. : 2.4495

STATISTIQUES DESCRIPTIVES

Les 4 variables sont à valeurs positives et négatives, centrées au alentours de 1, et d'échelles à peu près semblables.

	VARIANCE	SKEWNESS	CURTOSIS	ENTROPY
ECART TYPE	2.842763	5.869047	4.31003	2.101013
VARIANCE	8.081299	34.445710	18.57636	4.414256

INDICATEURS DE DISPERSION

Les variances des 4 variables sont à peu près sur la même échelle, allant de 4 pour ENTROPY à 34 pour SKEWNESS.

	VARIANCE	SKEWNESS	CURTOSIS	ENTROPY
ASYMÉTRIE	-0.1492243	-0.3936725	1.087378	-1.021125
APLATISSEMENT	2.2467848	2.5600102	4.261481	3.491315

INDICATEURS DE FORME DES DISTRIBUTIONS

Les distributions des variables VARIANCE et SKEWNESS sont relativement centrées, par contre les distributions des variables CURTOSIS et ENTROPY sont asymétriques. Les méthodes de prédiction utilisées dans ce projet ne reposant pas sur des hypothèses normalité des distributions, cela ne posera pas de problème ici.

Distributions des variables

Fig.1 – Histogramme de la variable VARIANCE

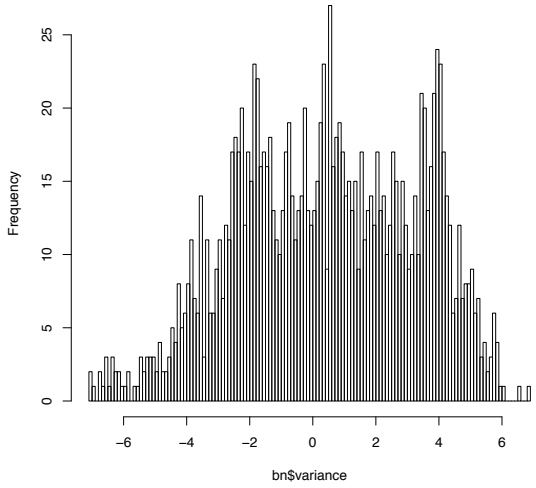


Fig.2 – Boxplot de la variable VARIANCE

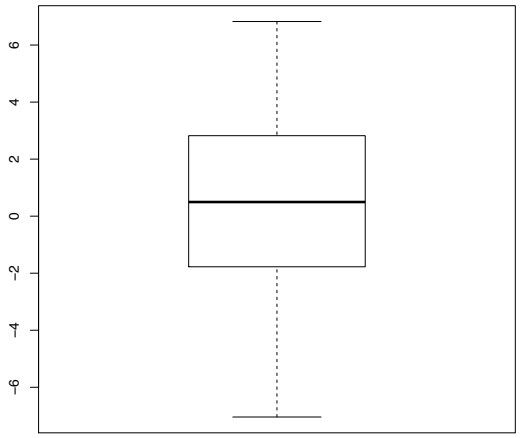


Fig.3 – Histogramme de la variable SKEWNESS

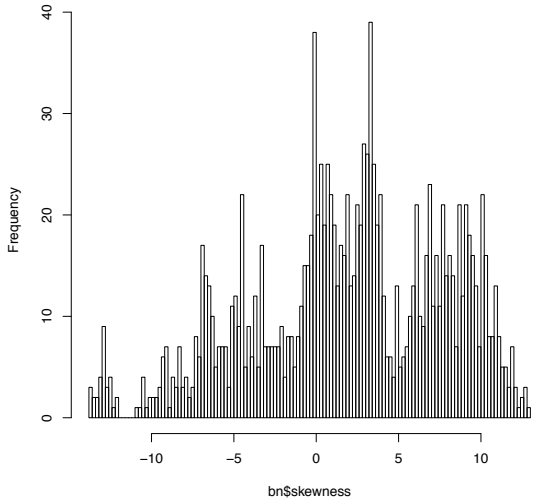


Fig.4 – Boxplot de la variable SKEWNESS

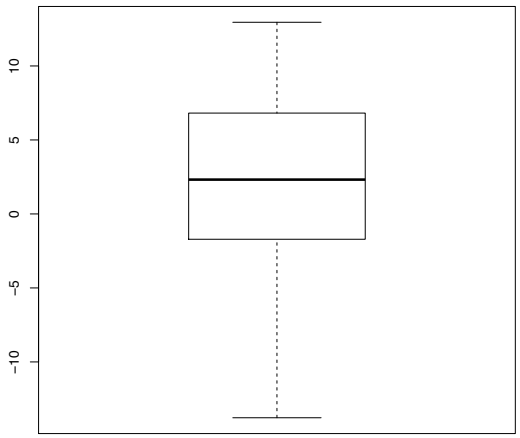


Fig.5 – Histogramme de la variable CURTOSIS

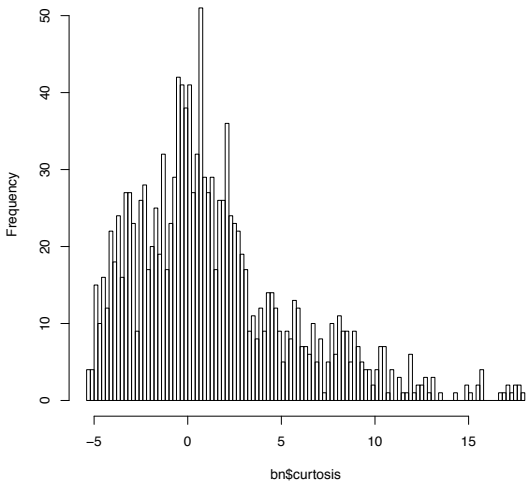


Fig.6 – Boxplot de la variable CURTOSIS

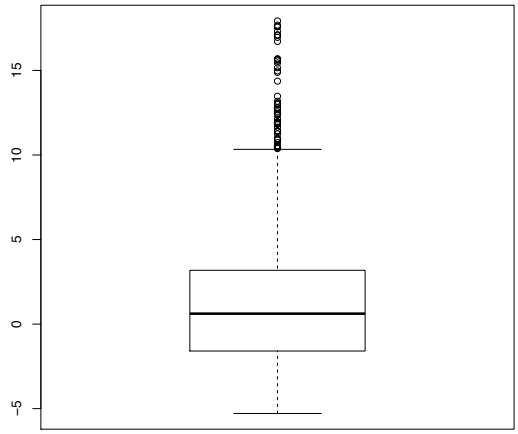


Fig.7 – Histogramme de la variable ENTROPY

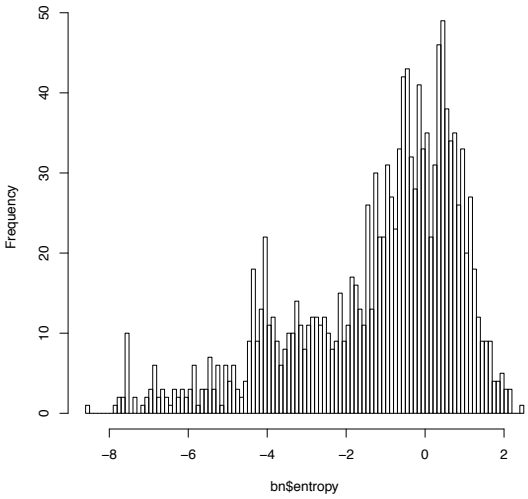
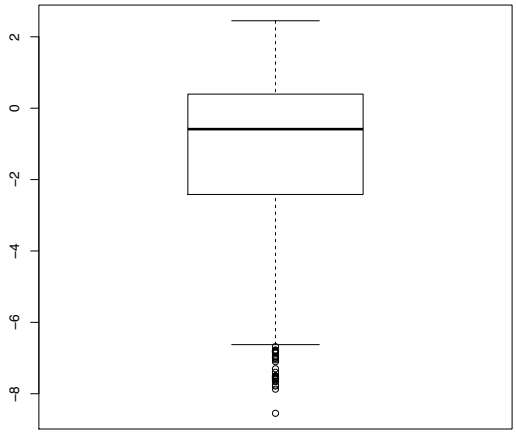


Fig.8 – Boxplot de la variable ENTROPY



Les variables CURTOSIS et ENTROPY contiennent un nombre élevé d'outliers, ce qui s'explique par leur distributions assez éloignées de la loi normale.

VARIABLE CIBLE

Le tableau de contingence de la variable FORGED sur l 'ensemble de l'échantillon nous permet de vérifier que les 2 modalités sont relativement équilibrées (44% de faux billets) :

	FORGED	
VALEUR	0	1
NB OBS.	762	610

Analyse bivariée

Dans cette partie nous étudierons les relations entre les variables 2 à 2, afin de d'identifier la présence de forte multicolinéarité, et d'étudier les liens éventuels entre chacune des variables explicatives et la variable cible.

CORRÉLATION ENTRE VARIABLES EXPLICATIVES

A la lecture de la matrice de corrélation des 4 variables quantitatives on voit que les variables SKEWNESS et CURTOSIS sont assez fortement négativement corrélées (-0.78). Les autres variables sont assez peu corrélées 2 à 2.

	VARIANCE	SKEWNESS	CURTOSIS	ENTROPY
VARIANCE	1.0000000	0.2640255	-0.3808500	0.2768167
SKEWNESS	0.2640255	1.0000000	-0.7868952	-0.5263208
CURTOSIS	-0.3808500	-0.7868952	1.0000000	0.3188409
ENTROPY	0.2768167	-0.5263208	0.3188409	1.0000000

MATRICE DE CORRELATIONS DES 4 VARIABLES EXPLICATIVES

VARIABLES EXPLICATIVES ET VARIABLE CIBLE

Fig.11 – Répartition de la variable VARIANCE

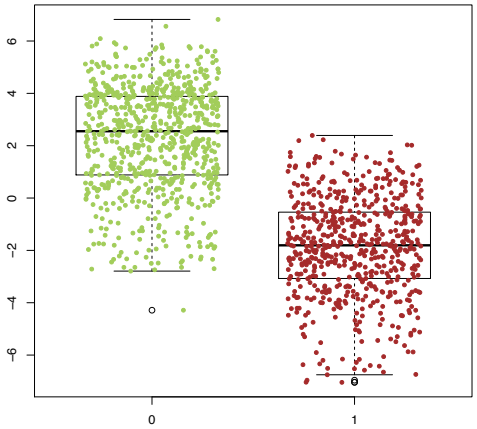
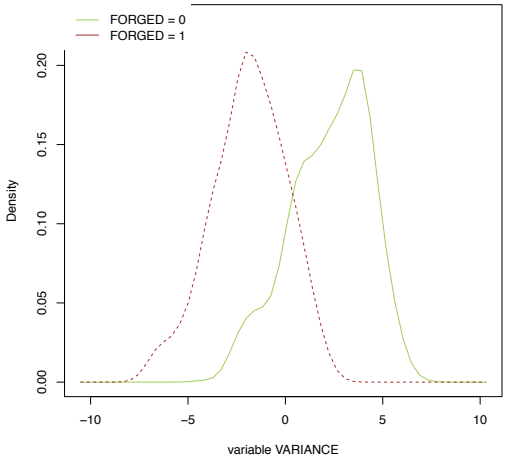


Fig.12 – Densité de probabilité de VARIANCE par modalité de FORGED



Variables VARIANCE et FORGED

En étudiant les 2 figures 11 (BOXPLOT) et 12 (distribution), on constate que les valeurs prises par la variable VARIANCE sont différentes qu'il s'agisse d'un vrai ou d'un faux billet.

Au dessus de 2, il s'agit principalement de vrais billets, et en dessous de -2 principalement des faux. Malgré un chevauchement entre les valeurs -2 et 2, on peut conclure de l'importance de la variable VARIANCE dans la mise en place du modèle prédictif.

Variables SKEWNESS et FORGED

De la même façon, sur les figures 13 et 14, on constate que les valeurs extrêmes positives de la variable SKEWNESS correspondent aux vrais billets de banque, alors que les valeurs extrêmes négatives correspondent elles à des faux billets. A un degré moindre que la variables VARIANCE, la variables SKEWNESS pourra être utile pour classifier les billets de banque.

Fig.13 – Répartition de la variable SKEWNESS

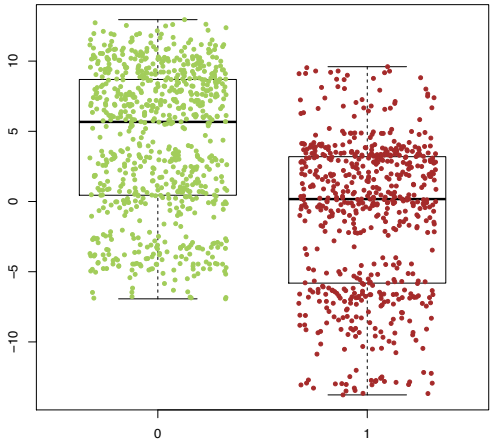


Fig.14 – Densité de probabilité de SKEWNESS par modalité de FORGED

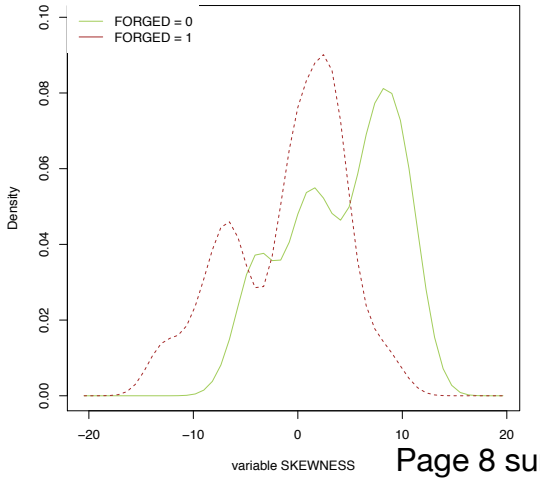
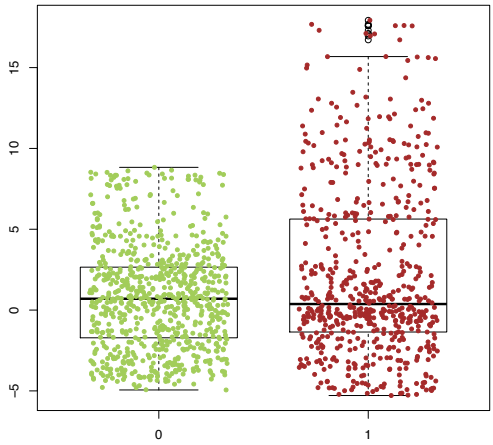


Fig.15 – Répartition de la variable CURTOSIS



Variables CURTOSIS et FORGED

Sur la figure 16 ci-contre, on peut voir que la distribution de la variable CURTOSIS est presque identique quelle que soit le type de billet, en dehors des valeurs extrêmes positives (> 10) composées uniquement de faux billets (comme on peut également le voir sur le boxplot figure 15).

Fig.16 – Densité de probabilité de CURTOSIS par modalité de FORGED

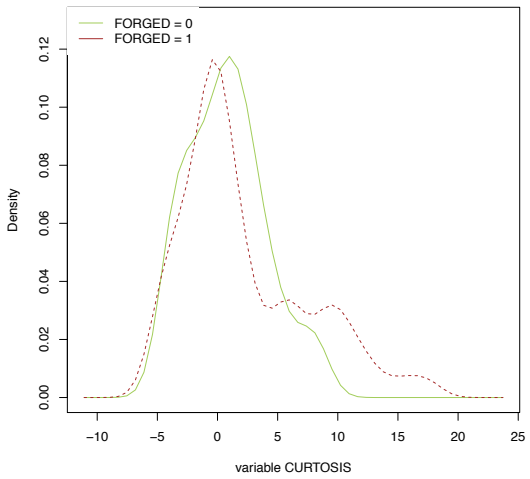


Fig.17 – Répartition de la variable ENTROPY

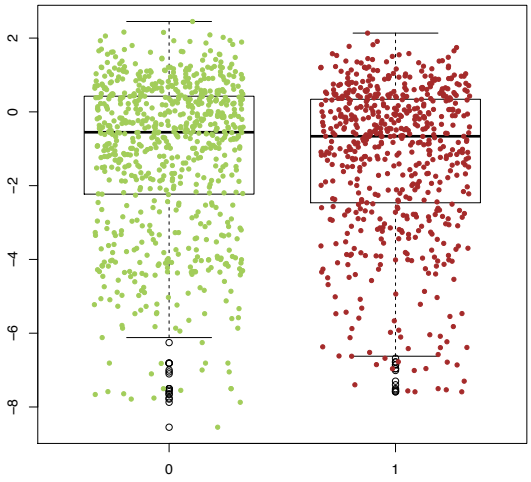
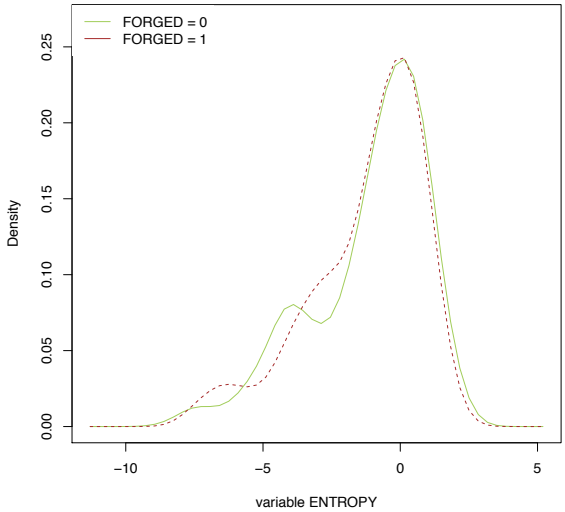


Fig.18 – Densité de probabilité de ENTROPY par modalité de FORGED



Variables ENTROPY et FORGED

Sur les figures 17 et 18, on voit bien que les distributions de la variable ENTROPY conditionnellement à FORGED sont identiques. On peut supposer du peu d'intérêt de la variable ENTROPY pour discriminer les billets de banque entre les vrais et les faux.

A l'issue de cette étude bivariée, on peut conclure que la variable VARIANCE semble être la plus importante pour identifier les faux billets, avec en complément les variables SKEWNESS et CURTOSIS. Dans cette vision bivariée, la variable ENTROPY n'a que très peu d'intérêt pour classifier les billets de banque.

Analyse multivariée

NUAGES DE POINTS

Fig. 20 – Nuage de points VARIANCE et SKEWNESS par modalité de FORGED

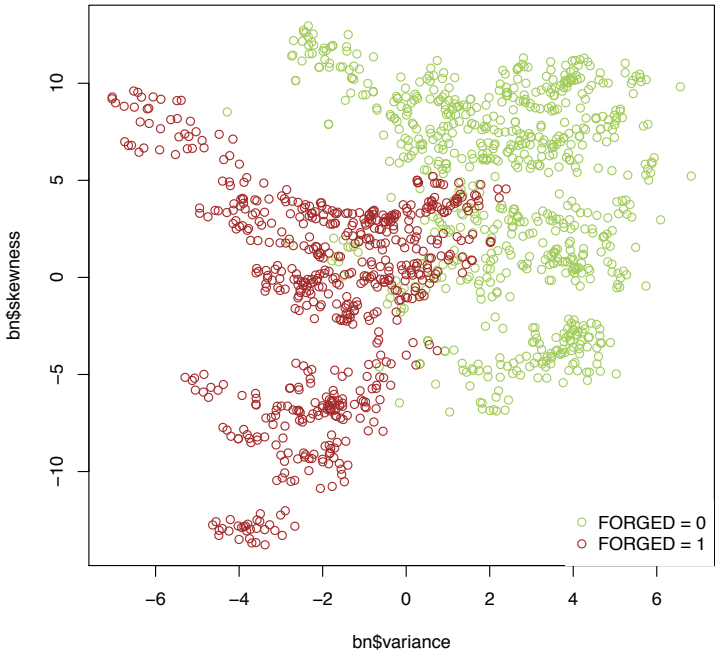


Fig. 21 – Nuage de points VARIANCE et CURTOSIS par modalité de FORGED

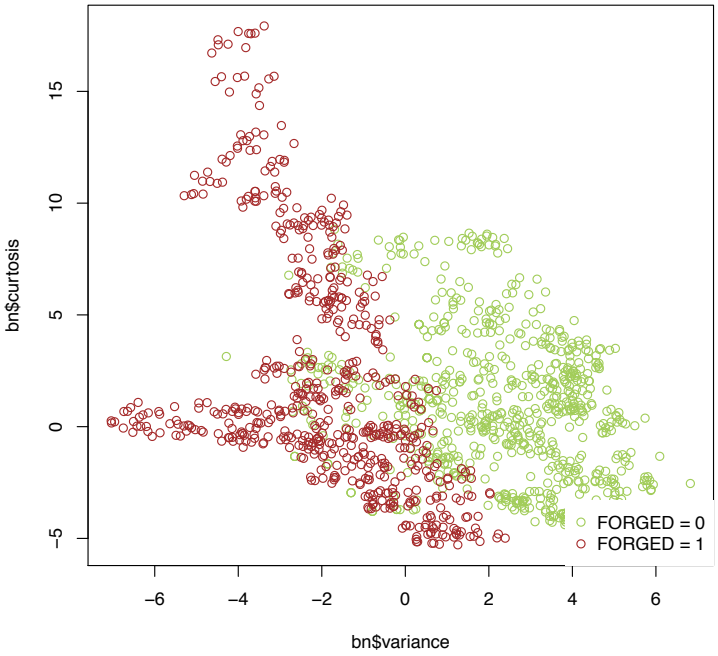
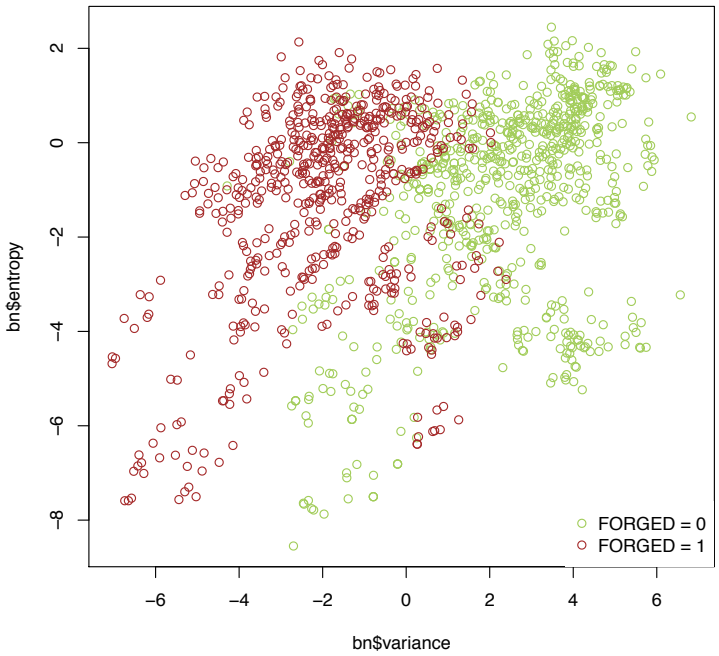
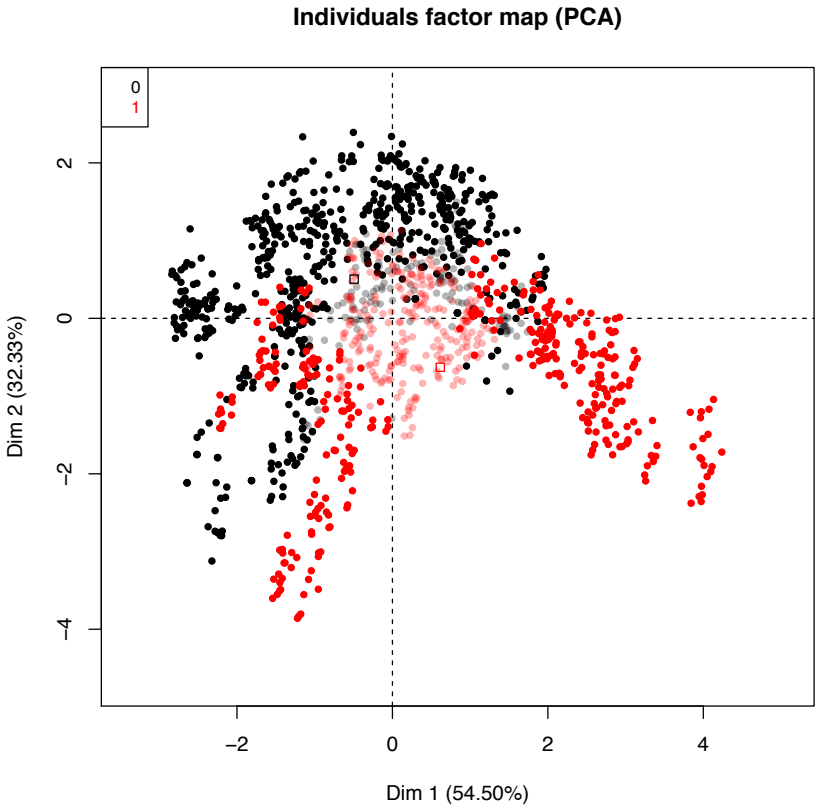


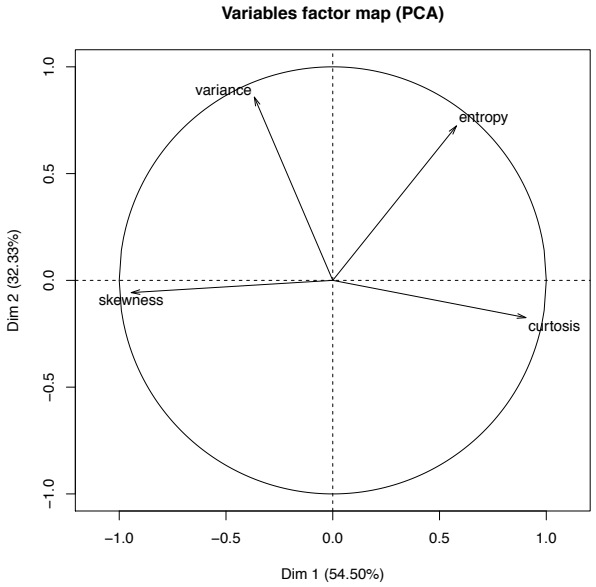
Fig. 22 – Nuage de points VARIANCE et ENTROPY par modalité de FORGED



L'étude des figures 20 à 22 représentant les 3 nuages de points de la variable VARIANCE (que l'on a précédemment identifié comme étant la plus interessante pour séparer les 2 modalités de FORGED) et d'une des 3 autres variables nous permet de constater que la meilleur séparation est obtenue avec le couple VARIANCE/SKEWNESS. La figure 22 montre que l'apport de la variable ENTROPY n'est pas flagrant.

ANALYSE EN COMPOSANTES PRINCIPALES





Après avoir réaliser une ACP centrée-réduite des 4 variables quantitatives (avec la variable cible en variable supplémentaire), étudions le cercle des variables ainsi que le plan factoriel principal, où les individus bien projetés apparaissent en noir et en rouge.

Sur le cercle des variables, on peut voir que le 1er axe est corrélé positivement à la variable SKEWNESS et négativement à CURTOSIS. Le deuxième axe représente principalement l'effet taille des variables VARIANCE et ENTROPY.

Sur le plan factoriel principal, contenant 86% de l'information initiale, on peut voir apparaitre 2 structures de faux billets. La première dans la partie droite du plan (SKEWNESS élevée et CURTOSIS faible), la seconde à gauche de l'axe 2 (VARIANCE et ENTROPY faibles à moyennes). Concernant les vrais billets (en noir pour les individus bien projetés), une structure apparait également (autour et à gauche de la droite de pente 1).

On peut voir quelques zones de chevauchement, mais on peut cependant conclure que l'information contenue dans les 4 variables quantitatives devrait pouvoir nous permettre d'obtenir un modèle de prédiction assez efficace.

Chacune des observations est modélisée par un couple (Xi, yi) où, pour l'observation i, Xi est le vecteur des 4 variables quantitatives (VARIANCE, SKEWNESS, CURTOSIS, ENTROPY) et yi l'indicatrice du fait que le billet soit un faux. La variable yi est une variable aléatoire suivant une loi de Bernoulli de paramètre πi représentant la probabilité que le billet soit un faux.

La fonction de lien utilisée sera le lien logistique ln (x / 1-x), le modèle sera donc de la forme :

Pour tout i,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_{var} \times VARIANCE_i + \beta_{ske} \times SKEWNESS_i + \beta_{cur} \times CURTOSIS_i + \beta_{ent} \times ENTROPY_i$$

Afin d'estimer les paramètres du modèle α, βvar, βske, βcur et βent, on utilisera le maximum de la vraisemblance des données à prédire.

Choix de modèle

La sélection des meilleurs modèles sera faite sur base des indices BIC et AIC, puis le meilleur modèle sera choisi après analyse et performances sur l'échantillon de validation.

Pour référence, ci-dessous la déviance du modèle nul :

```
NULL DEVIANCE: 942.561  ON 685  DEGREES OF FREEDOM
```

Compte tenu des conclusions de la partie précédente (importance de la variable VARIANCE), et du faible nombre de variables explicatives, nous pouvons nous permettre de tester l'ensemble des 8 modèles incluant la variable VARIANCE et de comparer leurs résultats.
Ci-dessous le tableau synthétisant les résultats obtenus :

	ITERATIONS	CONVERGENCE	LN.V	DEVIANCE	AIC	BIC	WARNING
FORGED ~.	12	1	-13.89084	27.78169	37.78169	60.43607	OUI
FORGED ~ VARIANCE	6	1	-227.86373	455.72747	459.72747	468.78922	NON
FORGED ~ VARIANCE + SKEWNESS	6	1	-175.43365	350.86731	356.86731	370.45994	NON
FORGED ~ VARIANCE + CURTOSIS	6	1	-210.47822	420.95643	426.95643	440.54907	NON
FORGED ~ VARIANCE + ENTROPY	6	1	-204.82919	409.65838	415.65838	429.25101	NON
FORGED ~ .-SKEWNESS	7	1	-153.03537	306.07074	314.07074	332.19425	NON
FORGED ~ .-ENTROPY	12	1	-14.46416	28.92832	36.92832	55.05183	OUI
FORGED ~ .-CURTOSIS	7	1	-175.21362	350.42724	358.42724	376.55075	NON

La colonne CONVERGENCE indique si l'algorithme a bien convergé (d'après la fonction glm). La colonne WARNING indique si l'algorithme a détecté une problématique de séparation/ quasi-séparation

Sur la base des indices AIC, BIC et de la déviance, nous sélectionnons les 2 modèles suivants pour analyse :

- Le modèle incluant les 4 variables explicatives
- Le modèle incluant 3 variables explicatives (VARIANCE, SKEWNESS, CURTOSIS)

Nous notons qu'il s'agit des 2 modèles pour lesquels il y a quasi-séparation.

Analyse et validation des modèles sélectionnés

MODÈLE À 4 VARIABLES EXPLICATIVES

Le modèle

```
DEVIANCE RESIDUALS:
     MIN        1Q    MEDIAN        3Q       MAX
-1.50671   0.00000   0.00000   0.00018   1.94596

COEFFICIENTS:
            ESTIMATE STD. ERROR Z VALUE PR(>|Z|)
(INTERCEPT)   7.2745     2.1491   3.385 0.000712 ***
VARIANCE     -7.9611     2.2427  -3.550 0.000386 ***
SKEWNESS     -3.9330     1.1316  -3.476 0.000510 ***
CURTOSIS     -5.0776     1.4473  -3.508 0.000451 ***
ENTROPY      -0.4683     0.4601  -1.018 0.308794
---
SIGNIF. CODES:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(DISPERSION PARAMETER FOR BINOMIAL FAMILY TAKEN TO BE 1)

    NULL DEVIANCE: 942.561  ON 685  DEGREES OF FREEDOM
RESIDUAL DEVIANCE:  27.782  ON 681  DEGREES OF FREEDOM
AIC: 37.782

NUMBER OF FISHER SCORING ITERATIONS: 12
```

Le coefficient pour la variable ENTROPY n’est pas statistiquement significatif.
Les estimations des coefficients ainsi que les erreurs standard sont élevées et suspectes (sauf pour la variable ENTROPY). Cela est du à la quasi-séparation détectée lors de la construction du modèle.

Coefficients et intervalles de confiance

```
                                        INTERCEPT                          BETA.VARIANCE                          BETA.SKEWNESS
ESTIMATION                       7.27446235535996                    -7.96105377066518                     -3.93299141311148
IC PROFILAGE 0.95 [ 4.13184431693969 , 12.9824323783488 ] [ -13.6342480030022 , -4.54950387953275 ] [ -6.8134623425526 , -2.22478551588629 ]
IC WALD 0.95      [ 3.06226042532364 , 11.4866642853963 ] [ -12.3566355987982 , -3.56547194253213 ] [ -6.1509473019983 , -1.71503552422467 ]

                                      BETA.CURTOSIS                         BETA.ENTROPY
ESTIMATION                       -5.07762983207253                    -0.468315647972522
IC PROFILAGE 0.95 [ -8.75108366839199 , -2.88998510476512 ] [ -1.54662215591558 , 0.371546823997076 ]
IC WALD 0.95       [ -7.91425397133775 , -2.2410056928073 ] [ -1.37018448799801 , 0.433553192052968 ]
```

Même constatation qu’au-dessus, de plus les intervalles de confiance pour l’estimation du coefficient beta pour la variable ENTROPY contient la valeur 0.

Influence des valeurs initiales

	ITÉRATIONS	LN.VRAISSEMBLANCE	COEF.ALPHA	COEF.BETA.VARIANCE	COEF.BETA.SKEWNESS	COEF.BETA.CURTOSIS	COEF.BETA.ENTROPY
PAR DEFAUT	12	-13.89084	7.274462	-7.961054	-3.932991	-5.07763	-0.4683156
REPRISE	1	-13.89084	7.274462	-7.961054	-3.932991	-5.07763	-0.4683156
SIMPLES (0,0,0,0,0)	12	-13.89084	7.274462	-7.961054	-3.932991	-5.07763	-0.4683156
HASARD (0.3,-0.7,0.2,-0.1,1,2)	13	-13.89084	7.274462	-7.961054	-3.932991	-5.07763	-0.4683156

On constate que les résultats sont identiques pour les différentes valeurs initiales utilisées, cela suggère qu’avec les valeurs initiales par défaut il y a bien convergence vers le maximum de vraisemblance.

Conclusion

Ces éléments nous permettent de conclure le modèle est bien valide, mais que statistiquement la variable ENTROPY n’apporte rien au modèle.

MODÈLE À 3 VARIABLES EXPLICATIVES (SANS ENTROPY)

Le modèle

```
DEVIANCE RESIDUALS:
     MIN        1Q    MEDIAN        3Q       MAX
-1.44947   0.00000   0.00000   0.00026   2.05583

COEFFICIENTS:
            ESTIMATE STD. ERROR Z VALUE PR(>|Z|)
(INTERCEPT)   7.4962     2.2648   3.310 0.000933 ***
VARIANCE     -7.5921     2.2102  -3.435 0.000592 ***
SKEWNESS     -3.4804     0.9655  -3.605 0.000312 ***
CURTOSIS     -4.6253     1.3169  -3.512 0.000444 ***
---
SIGNIF. CODES:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(DISPERSION PARAMETER FOR BINOMIAL FAMILY TAKEN TO BE 1)

    NULL DEVIANCE: 942.561  ON 685  DEGREES OF FREEDOM
RESIDUAL DEVIANCE:  28.928  ON 682  DEGREES OF FREEDOM
AIC: 36.928

NUMBER OF FISHER SCORING ITERATIONS: 12
```

On constate que tous les coefficients sont statistiquement significatifs, mais comme dans le cas du modèle précédent leurs valeurs sont extrêmement élevées et suspectes à cause de la problématique de quasi-séparation.

Coefficients et intervalles de confiance

```
                                          INTERCEPT                          BETA.VARIANCE                          BETA.SKEWNESS
ESTIMATION                          7.49621123919365                    -7.59207878992064                    -3.48037714861793
IC PROFILAGE 0.95 [ 4.25557935580412 , 13.4315970813716 ]  [ -13.2727501144298 , -4.3310787680984 ] [ -5.96928217222014 , -2.05952585673841 ]
IC WALD 0.95       [ 3.05727305791534 , 11.935149420472 ] [ -11.9239013132181 , -3.26025626662321 ] [ -5.37266703121099 , -1.58808726602486 ]

                                       BETA.CURTOSIS
ESTIMATION                          -4.62528662745859
IC PROFILAGE 0.95  [ -8.0272736365817 , -2.69266923036282 ]
IC WALD 0.95      [ -7.20628078173792 , -2.04429247317925 ]
```

On constate qu’aucun des intervalles de confiance ne contient la valeur 0, ce qui renforce le fait que ces coefficients sont statistiquement significatifs.

Tout comme les valeurs des coefficients, les intervalles de confiances à 95% sont énormes et exagérés. Il s’agit là encore d’un des impacts de la quasi-séparation. Ni les coefficients ni leurs IC ne sont exploitables pour interprétation, mais le modèle reste utilisable pour la prédiction.

Influence des valeurs initiales

	ITÉRATIONS	LN.VRAISSEMBLANCE	COEF.ALPHA	COEF.BETA.VARIANCE	COEF.BETA.SKEWNESS	COEF.BETA.CURTOSIS
PAR DEFAUT	12	-14.46416	7.496211	-7.592079	-3.480377	-4.625287
REPRISE	1	-14.46416	7.496211	-7.592079	-3.480377	-4.625287
SIMPLES (0,0,0,0)	12	-14.46416	7.496211	-7.592079	-3.480377	-4.625287
HASARD (0.3,-0.7,0.2,-0.1)	12	-14.46416	7.496211	-7.592079	-3.480377	-4.625287

On constate que les résultats sont identiques pour les différentes valeurs initiales utilisées, cela suggère qu’avec les valeurs initiales par défaut il y a bien convergence vers le maximum de vraisemblance.

Conclusion

Le modèle est valide, tous les coefficients sont statistiquement significatifs pour le risque 0.05, et aucun des intervalles de confiance ne contient la valeur 0. Cependant les valeurs des coefficients sont beaucoup trop grandes à cause de la quasi-séparation.

Performance prédictive des modèles

En appliquant les 2 modèles étudiés précédemment à l'échantillon de validation avec un seuil à 0.5, on voit que le modèle à 3 variables explicatives obtient de meilleurs performances que celui à 4 variables explicatives (Accuracy = 0,9884 contre 0,9826) :

```
MODELE : "FORGED ~ .-ENTROPY"
CONFUSION MATRIX AND STATISTICS

          REFERENCE
PREDICTION   0   1
         0 190   3
         1   1 150

               ACCURACY : 0.9884
                 95% CI : (0.9705, 0.9968)
    NO INFORMATION RATE : 0.5552
    P-VALUE [ACC > NIR] : <0.0000000000000002

                  KAPPA : 0.9764

 MCNEMAR'S TEST P-VALUE : 0.6171

            SENSITIVITY : 0.9804
            SPECIFICITY : 0.9948
         POS PRED VALUE : 0.9934
         NEG PRED VALUE : 0.9845
             PREVALENCE : 0.4448
         DETECTION RATE : 0.4360
   DETECTION PREVALENCE : 0.4390
      BALANCED ACCURACY : 0.9876

       'POSITIVE' CLASS : 1


MODELE : "FORGED ~ ."
CONFUSION MATRIX AND STATISTICS

          REFERENCE
PREDICTION   0   1
         0 190   5
         1   1 148

               ACCURACY : 0.9826
                 95% CI : (0.9624, 0.9936)
    NO INFORMATION RATE : 0.5552
    P-VALUE [ACC > NIR] : <0.0000000000000002

                  KAPPA : 0.9646

 MCNEMAR'S TEST P-VALUE : 0.2207

            SENSITIVITY : 0.9673
            SPECIFICITY : 0.9948
         POS PRED VALUE : 0.9933
         NEG PRED VALUE : 0.9744
             PREVALENCE : 0.4448
         DETECTION RATE : 0.4302
   DETECTION PREVALENCE : 0.4331
      BALANCED ACCURACY : 0.9810

       'POSITIVE' CLASS : 1
```

Compte tenu des conclusions de l'analyse des modèles et des performances obtenues lors de la prédiction sur l'échantillon de validation, le modèle de régression logistique sélectionné est celui à 3 variables explicatives (sans la variable ENTROPY).

Interprétation

Les valeurs des coefficients obtenus pour les 2 modèles étudiés (ainsi que les intervalles de confiance) sont extrêmement élevées et sujets à caution (-8 par exemple pour le coefficient beta pour la variable VARIANCE).

Dans le cas de séparation quais-complète, l'estimateur au sens du maximum de vraisemblance n'existe pas, et l'on obtient des pseudo-estimations classant correctement les individus. Dans ce cadre une interprétation de ces coefficients n'est pas réalisable.

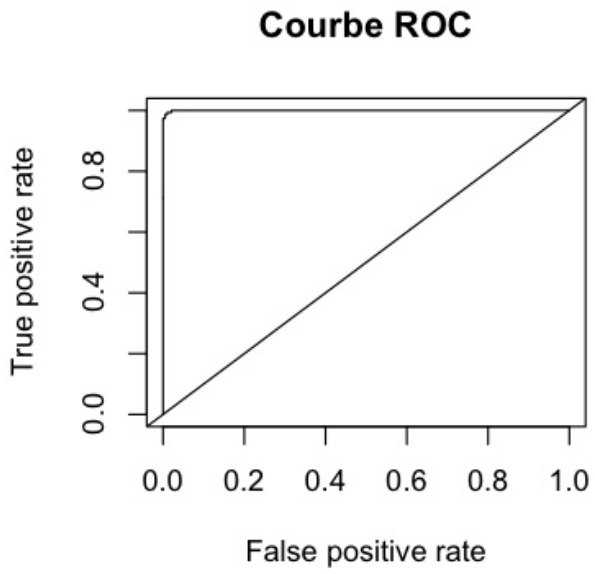
Les performances sur l'échantillon de validation des modèles ne présentant pas de quasi-séparation des 2 groupes étant moins bonnes, ces modèles ne seront pas étudiés plus profondément malgré des coefficients cette fois interprétables.

Choix du seuil et courbe ROC

Courbe ROC

Ci-dessous la courbe ROC obtenue sur les données de validation.

AUC : 0.9997262



La courbe ROC nous indique que le modèle classe presque parfaitement la variable FORGED, avec un AUC très proche de 1.

Choix du seuil

Ci-dessous les performances sur l 'échantillon de validation du modèle à 3 variables, pour les différentes valeur de seuil allant de 0 à 1 par pas de 0,05 :

	SEUIL	ACCURACY	IC.95.ACCURACY	SENSITIVITY	SPECIFICITY
0	0	0.444767441860465	[0.391475013835356 , 0.499018822490614]	1	0
0.05	0.05	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.993464052287582	0.989528795811518
0.1	0.1	0.988372093023256	[0.970497105564825 , 0.996822948212077]	0.986928104575163	0.989528795811518
0.15	0.15	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.2	0.2	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.25	0.25	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.3	0.3	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.35	0.35	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.4	0.4	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.45	0.45	0.991279069767442	[0.974726604340269 , 0.998197911001809]	0.986928104575163	0.994764397905759
0.5	0.5	0.988372093023256	[0.970497105564825 , 0.996822948212077]	0.980392156862745	0.994764397905759
0.55	0.55	0.988372093023256	[0.970497105564825 , 0.996822948212077]	0.980392156862745	0.994764397905759
0.6	0.6	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.973856209150327	0.994764397905759
0.65	0.65	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.973856209150327	0.994764397905759
0.7	0.7	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.973856209150327	0.994764397905759
0.75	0.75	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.973856209150327	0.994764397905759
0.8	0.8	0.988372093023256	[0.970497105564825 , 0.996822948212077]	0.973856209150327	1
0.85	0.85	0.988372093023256	[0.970497105564825 , 0.996822948212077]	0.973856209150327	1
0.9	0.9	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.967320261437909	1
0.95	0.95	0.98546511627907	[0.966408145990631 , 0.995264161816648]	0.967320261437909	1
1	1	0.555232558139535	[0.500981177509386 , 0.608524986164644]	0	1

TEST DES DIFFERENTS SEUILS SUR ECHANTILLON DE VALIDATION

Les meilleurs performances sont obtenues pour les valeurs de seuil de 0,05, et de 0,15 à 0,45 avec une précision de 0.991279069767442.

On pourra par exemple sélectionné le seuil de 0.45, qui est le plus proche du seuil classique de 0,5 ayant obtenu les meilleurs résultats. Ce qui nous donne comme performance pour le modèle de régression logistique :

Modèle : FORGED ~.-ENTROPY
Seuil : 0,45

Performances détaillées du modèle sur l'échantillon de validation

Précision	0.991279069767442
IC 95%	[0.974726604340269 , 0.998197911001809]
Sensibilité	0.986928104575163
Spécificité	0.994764397905759

FORETS ALEATOIRES

Cette partie du projet sera consacrée à la mise en place d'un modèle prédictif basé sur les forêts aléatoires. Nous utiliserons ici la fonction randomForest du package R du même nom.

Le principe des forêts aléatoires est basé sur le bagging d'arbres CART, auquel on ajoute une condition d'aléa sur les choix des variables à utiliser pour l'apprentissage. On s'affranchit ainsi des faiblesses des arbres de décision (instabilité, sur-apprentissage).

Remarque concernant la notion de seuil

Il est possible d'obtenir les probabilités d'appartenance à une classe obtenues par la méthode des forêts aléatoires avec le package randomForest en utilisant l'option type = "prob".
il s'agit de la proportion d'arbres ayant voté pour l'appartenance à la classe positive. Nous ne l'utiliserons pas dans ce projet, en nous contentant du seuil par défaut (majorité des votes).

Recherche et sélection du meilleur modèle

Les 2 principaux paramètres des forêts aléatoires sont le nombre d'arbres agrégés au sein de la forêt et le nombre de variables disponible pour la construction de chaque noeud (par défaut nombre de variable /2). Suite aux analyses préliminaires, se pose également la question de la pertinence d'inclure la variable ENTROPY dans le modèle.

Ici, compte tenu du faible nombre de variables, la méthode utilisée pour sélectionner les meilleurs paramètres sera de tester chaque combinaison sur l'échantillon de validation (plutôt que d'utiliser les résultats OOB de l'entrainement du modèle).

Nous allons donc tester, en mesurant la précision de la prédiction sur l'échantillon de validation :

- 2 modèles (avec ou sans la variable ENTROPY)
- Le nombre de variable pour la construction des noeud (de 2 à 3/4 selon le modèle)
- Le nombre de noeud (de 10 à 500 par pas de 10)

Ci-dessous les résultats obtenus.

Fig.50 - Modèle : forged ~. , mtry=2

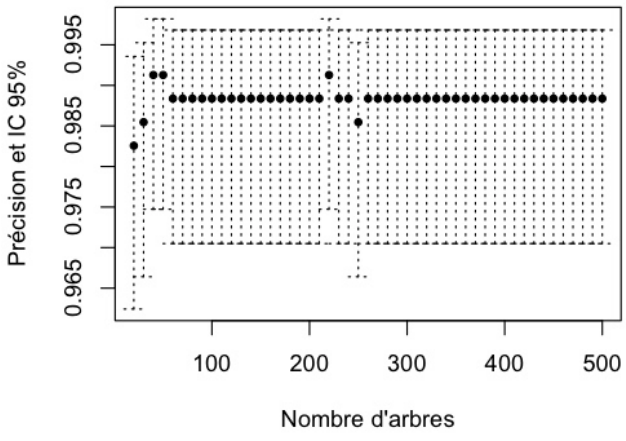


Fig.51 - Modèle : forged ~. , mtry=3

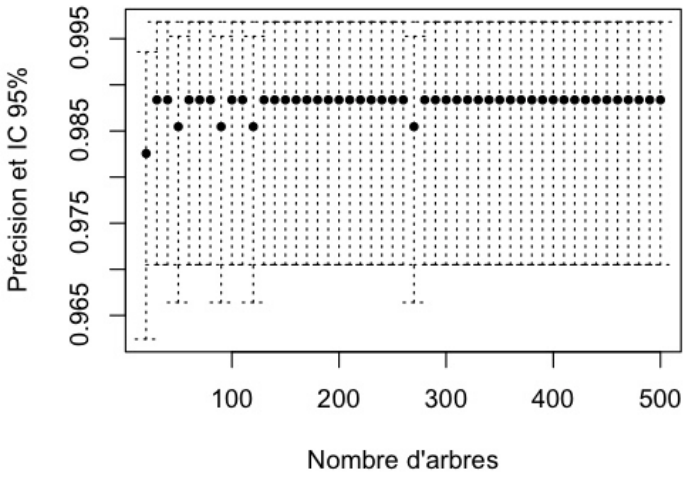


Fig.52 - Modèle : forged ~. , mtry=4

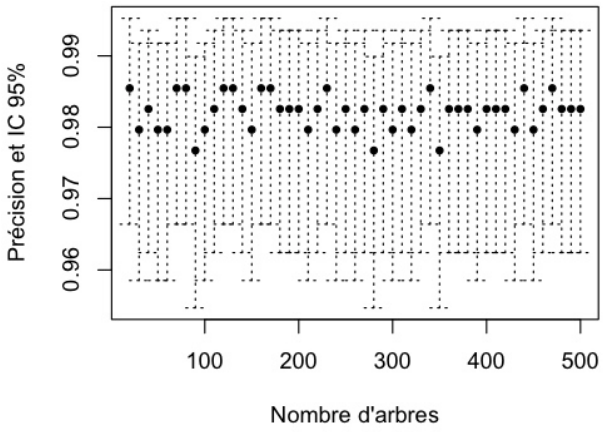


Fig.53 - Modèle : forged ~.-entropy , mtry=2

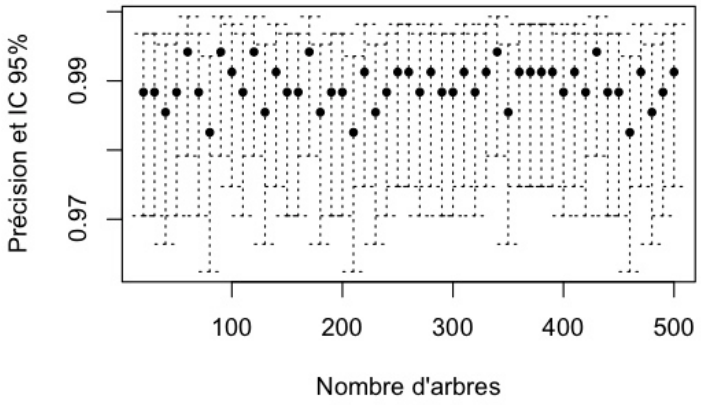
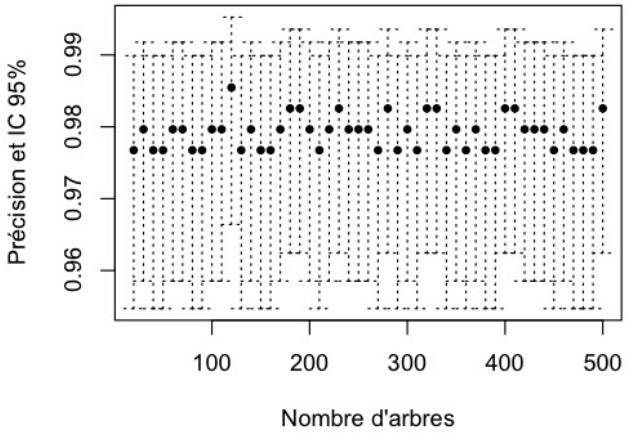


Fig.54 - Modèle : forged ~.-entropy , mtry=3



Ces représentations graphiques étant difficilement exploitables pour trouver la meilleure combinaison de paramètres, trions les résultats par précision descendante :

	MODELE	TREE	NVAR	ACCURACY	IC.95.ACCURACY
2	FORGED ~.	30	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
152	**FORGED ~.-ENTROPY**	**60**	**2**	**0.994186046511628**	**[0.97915688716803 , 0.999295126107053]**
155	FORGED ~.-ENTROPY	90	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
158	FORGED ~.-ENTROPY	120	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
163	FORGED ~.-ENTROPY	170	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
180	FORGED ~.-ENTROPY	340	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
189	FORGED ~.-ENTROPY	430	2	0.994186046511628	[0.97915688716803 , 0.999295126107053]
1	FORGED ~.	20	2	0.991279069767442	[0.974726604340269 , 0.998197911001809]
4	FORGED ~.	50	2	0.991279069767442	[0.974726604340269 , 0.998197911001809]
43	FORGED ~.	440	2	0.991279069767442	[0.974726604340269 , 0.998197911001809]

TABLEAU DES 10 MEILLEURS COMBINAISONS DE PARAMÈTRES (ACCURACY SUR ÉCHANTILLON VALIDATION)

Les 7 premières combinaisons obtiennent la même précision. En utilisant la règle privilégiant, à performance égale, les modèles les moins complexes, nous sélectionnons la combinaison suivante :

Modèle : FORGED ~.-ENTROPY
Nombre de variables pour chaque noeud : 2
Nombre d'arbres : 60

Performances détaillées du modèle sur l'échantillon de validation

Précision	0.994186046511628
IC 95%	[0.97915688716803 , 0.999295126107053]
Sensibilité	1
Spécificité	0.989528795811518
PPV	0.987096774193548
NPV	1

Interprétation

Le principal inconvénient des forêts aléatoires par rapport aux arbres de décision est la perte de lisibilité. Il s'agit de méthodes de type « boite grise », difficiles à interpréter. On peut cependant obtenir une idée de l'importance des variables dans la prédiction en étudiant les mesures suivantes (qui ont servi à la construction de la forêt ou en découlent) :

- Nombre d'apparition de la variable dans l'arbre
- Mean decreased accuracy
- Mean decreased GINI

Appliquées au modèle sélectionné précédemment :

	MEANDECREASEACCURACY	MEANDECREASEGINI
VARIANCE	0.3560560	203.28506
SKEWNESS	0.1975481	85.33804
CURTOSIS	0.1049816	50.12447

On constate que la variable la plus efficace (et la plus utilisée) est la variable VARIANCE, puis la variable SKEWNESS. La dernière variable du modèle, CUTOSIS, n'est jamais utilisée pour construire le premier noeud.

Variables utilisées pour le 1er noeud (nombre d'arbres) :

VARIANCE	SKEWNESS
42	18

CONCLUSION

Les performances des 2 modèles sont très proches, mais le modèle de forêts aléatoire ayant obtenu sur l'échantillon de validation des résultats un peu meilleurs sera celui sélectionné comme modèle final du projet. On évite également ainsi d'appliquer un modèle de régression logistique à des données présentant une quasi-séparation (où une analyse discriminante aurait été plus appropriée).

Modèle final du projet

Méthode : Random Forest
Modèle : FORGED ~.-ENTROPY
Nombre de variables pour chaque noeud : 2
Nombre d'arbres : 60

Performances détaillées du modèle sur l'échantillon de test

Précision	0,9912
IC 95%	[0.9746 , 0.9982]
Sensibilité	0.9803
Spécificité	1
PPV	1
NPV	0.9845

Les performance en prédiction du modèle retenu sont presque parfaites, avec une précision de 0.9912 (IC 95% 0.9746 à 0.9982).