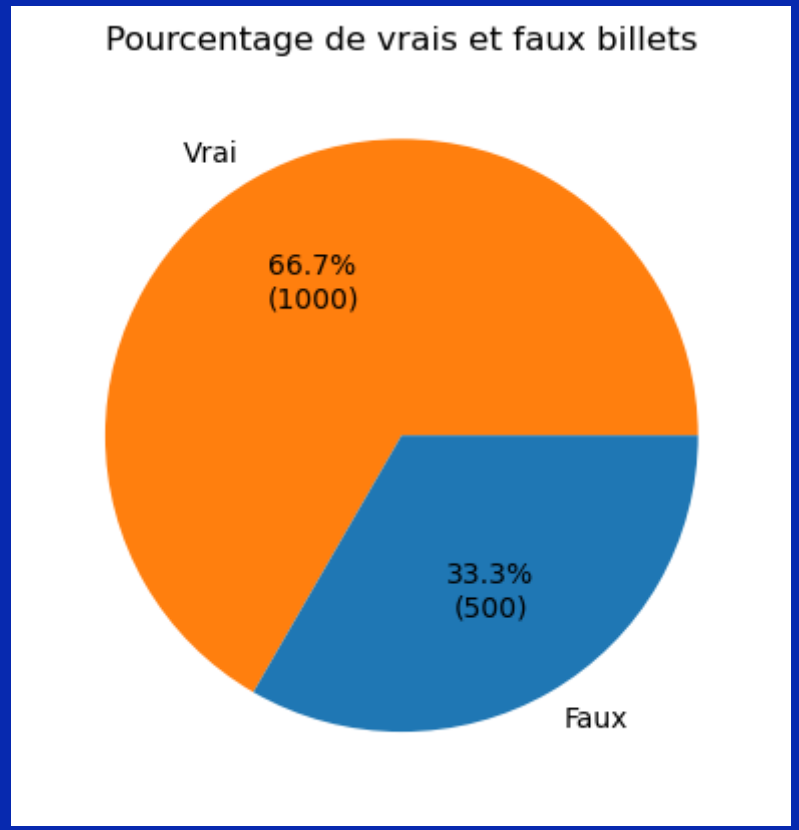
DÉTECTION AUTOMATIQUE DE FAUX BILLETS

ONCFM

Contexte

- L'Organisation nationale de lutte contre le faux-monnayage – association publique
- Il y a des différences de dimensions entre les faux et les vrais billets.
- Objectif du projet - créer un algorithme capable de différencier automatiquement les vrais et les faux billets en utilisant les dimensions des billets.

ONCFM



Le Fichier

1500 billets

- 1000 vrais, 500 faux

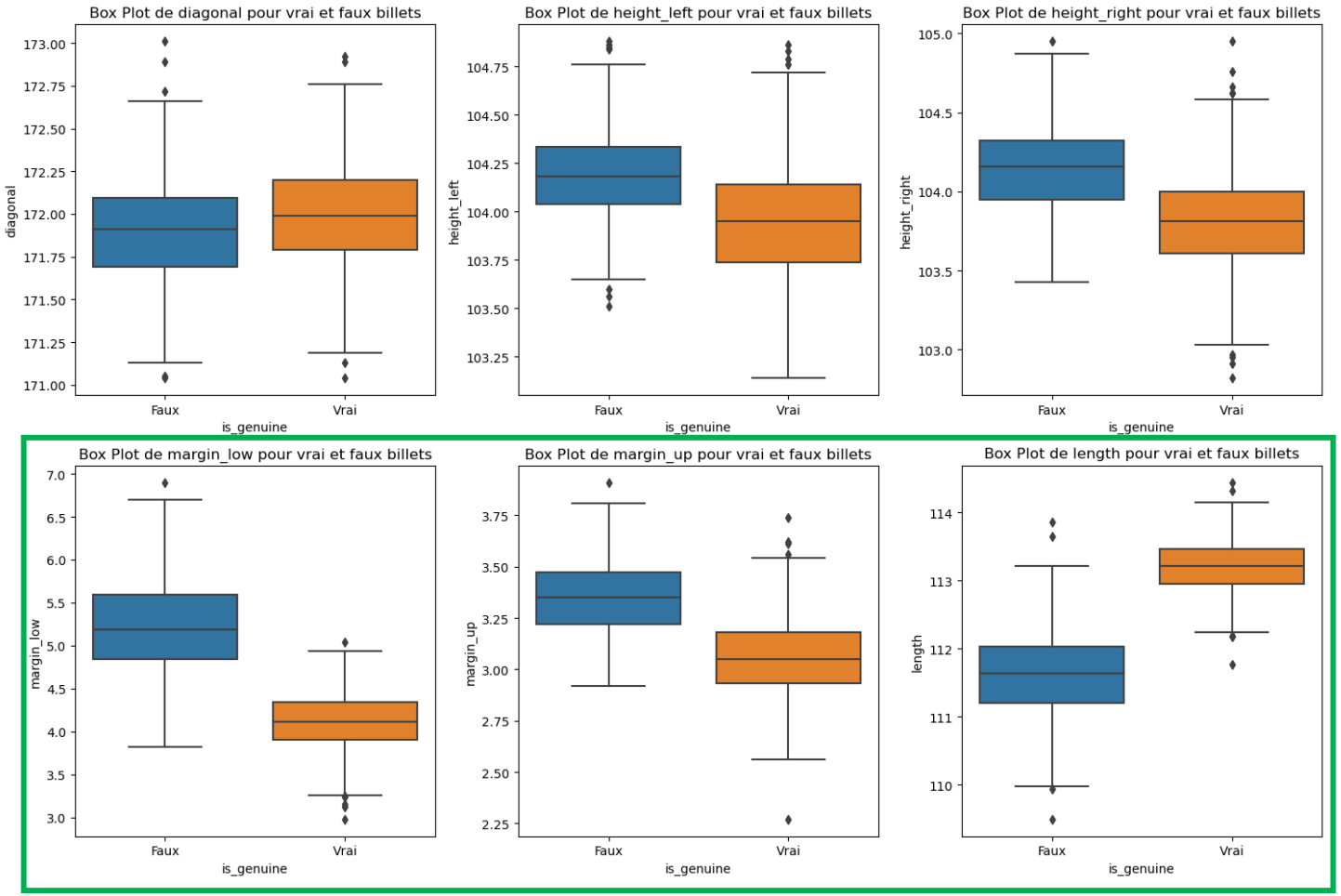
7 colonnes

- is_genuine : si le billet est vrai ou faux
- length : la longueur du billet
- height_left : la hauteur du billet sur le côté gauche
- height_right : la hauteur du billet sur le côté droit
- margin_up : la marge entre le bord supérieur du billet et l'image de celui-ci
- margin_low : la marge entre le bord inférieur du billet et l'image de celui-ci
- diagonal : la diagonale du billet

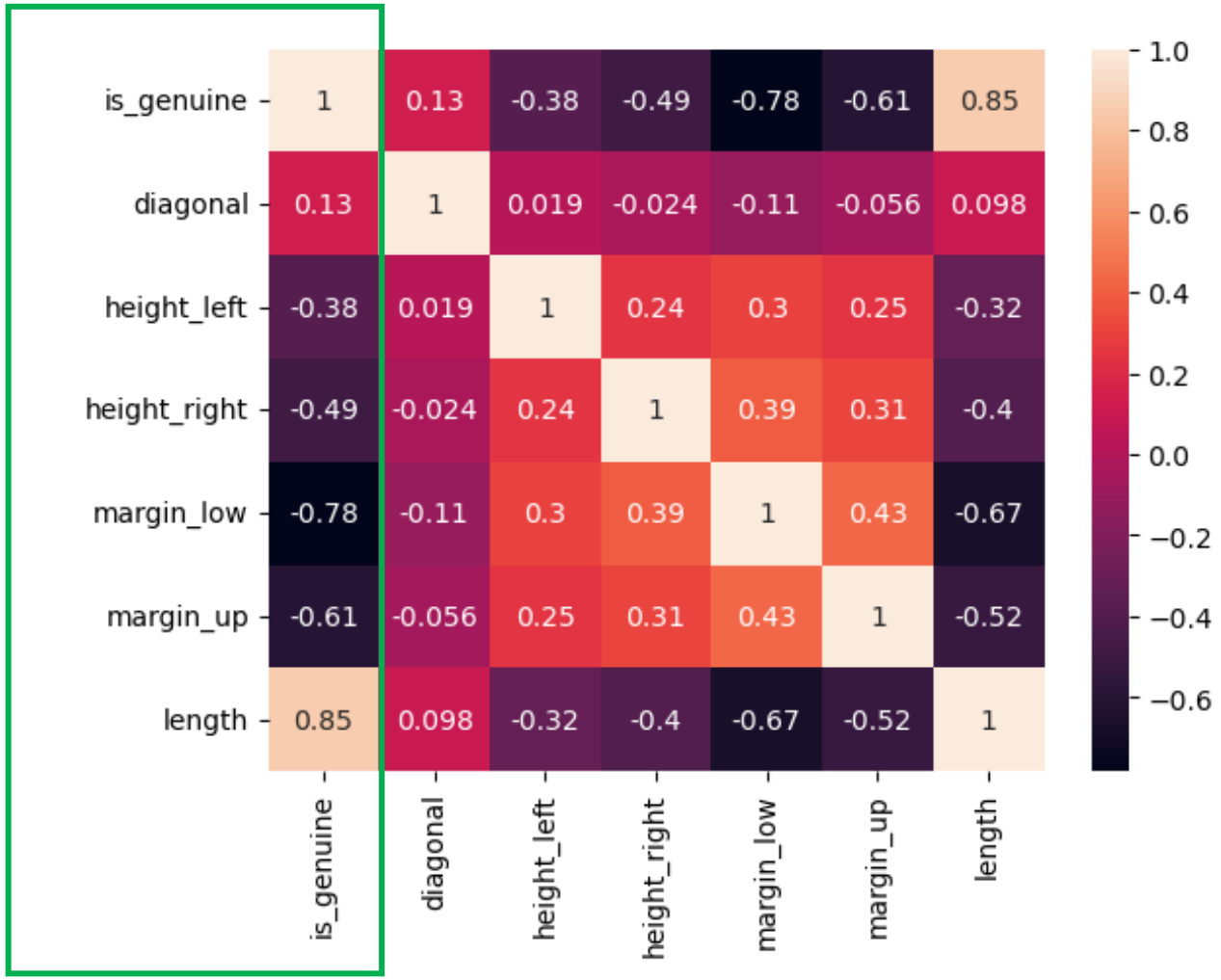
ONCFM

Analyse Descriptive

- Les vrais et faux billets pour « length », « margin_up » et « margin_low » sont assez différenciés.
- Les vrais et faux billets pour « diagonal », « height_left », et « height_right » sont assez similaires.



ONCFM



ONCFM

Analyse Descriptive

Variables les plus pertinentes pour « is_genuine » :

- « length »
- « margin_low »
- « margin_up »

Variable la moins pertinente pour « is_genuine » :

- « diagonal »

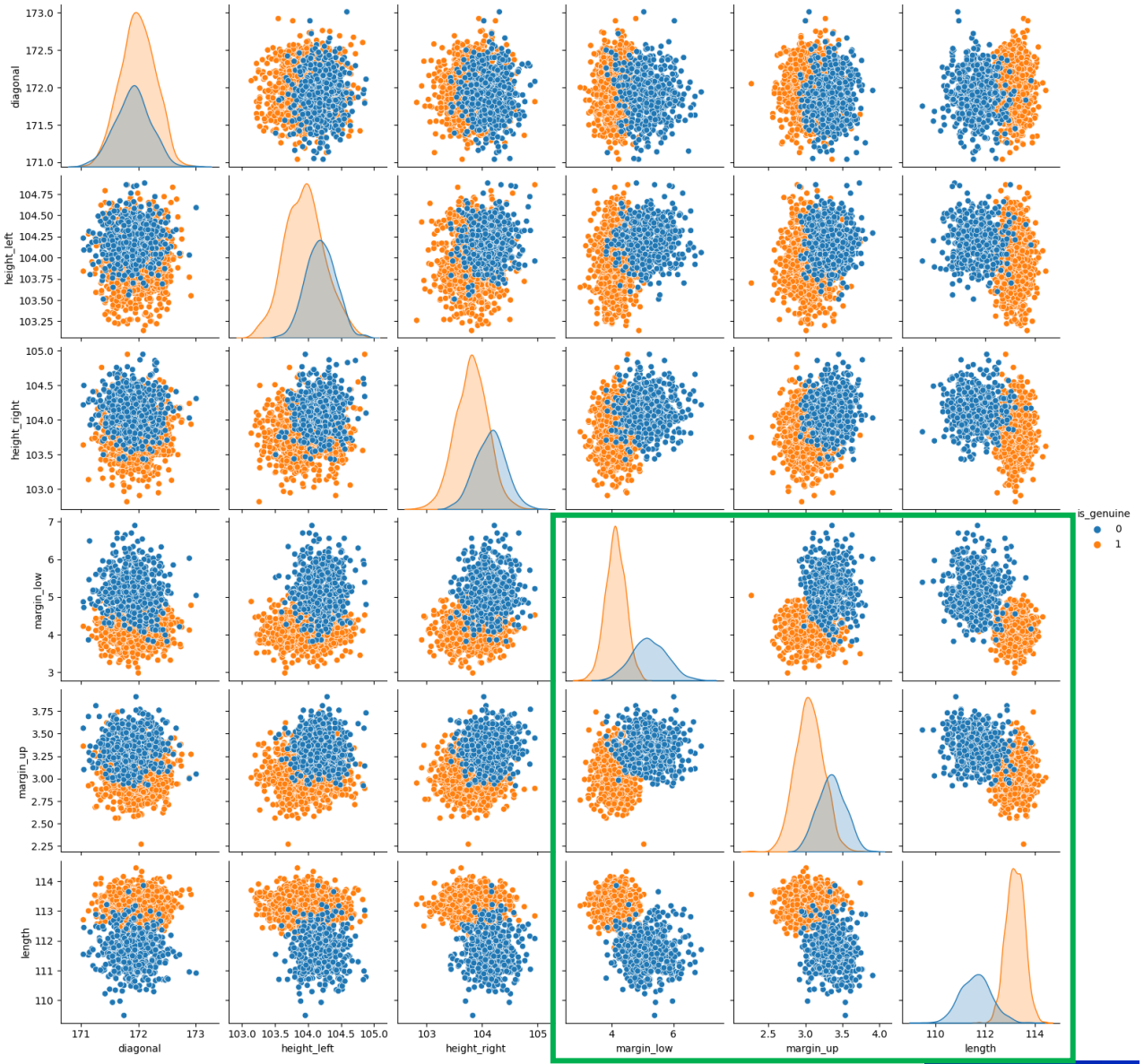
Analyse Descriptive

Graphiques de densité noyau

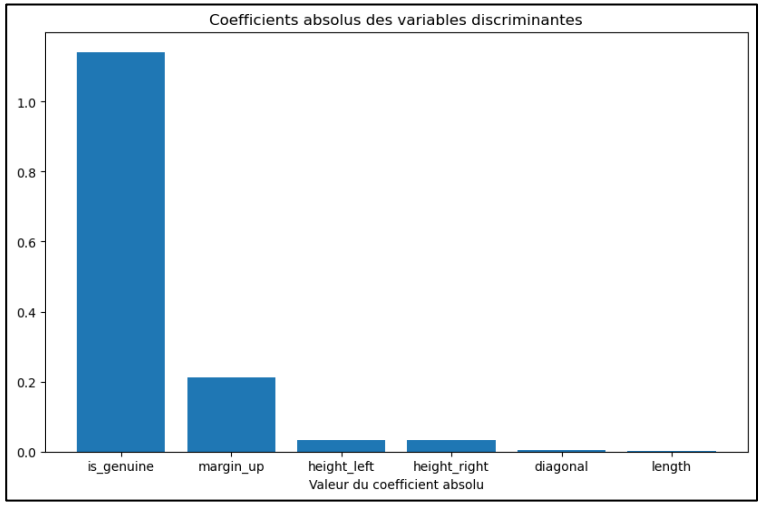
- Les vrais et les faux billets sont bien différenciés par les variables « length », et « margin_low »
- L'inverse est vrai pour « diagonal », « height_left » et « height_right ».

Scatterplots

- Les scatterplots qui contiennent « length », « margin_low » et « margin_up » ont des clusters assez bien séparés.
- Utiles pour déterminer quels billets sont réels et lesquels sont faux.



ONCFM



OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	1174.
Date:	Thu, 08 Feb 2024	Prob (F-statistic):	1.24e-304
Time:	17:48:57	Log-Likelihood:	-774.73
No. Observations:	1463	AIC:	1555.
Df Residuals:	1460	BIC:	1571.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9263	0.198	30.003	0.000	5.539	6.314
margin_up	-0.2119	0.059	-3.612	0.000	-0.327	-0.097
is_genuine	-1.1632	0.029	-40.477	0.000	-1.220	-1.107

Omnibus:	22.365	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.106
Skew:	0.057	Prob(JB):	3.22e-09
Kurtosis:	3.793	Cond. No.	65.0

Régression Linéaire

Bar plot

- « is_genuine » influence beaucoup la variable « margin_low »
- « margin_up » a aussi une légère influence.

Modèle

- Le meilleur modèle utilise seulement « margin_up et » « is_genuine »
- $R^2 = 0,617$
- Statistiquement significatif $(1,24^{e}\text{-}403<0,05)$
- Coefficients – « margin_up» = -0,2119, is_genuine = -1,1632

ONCFM

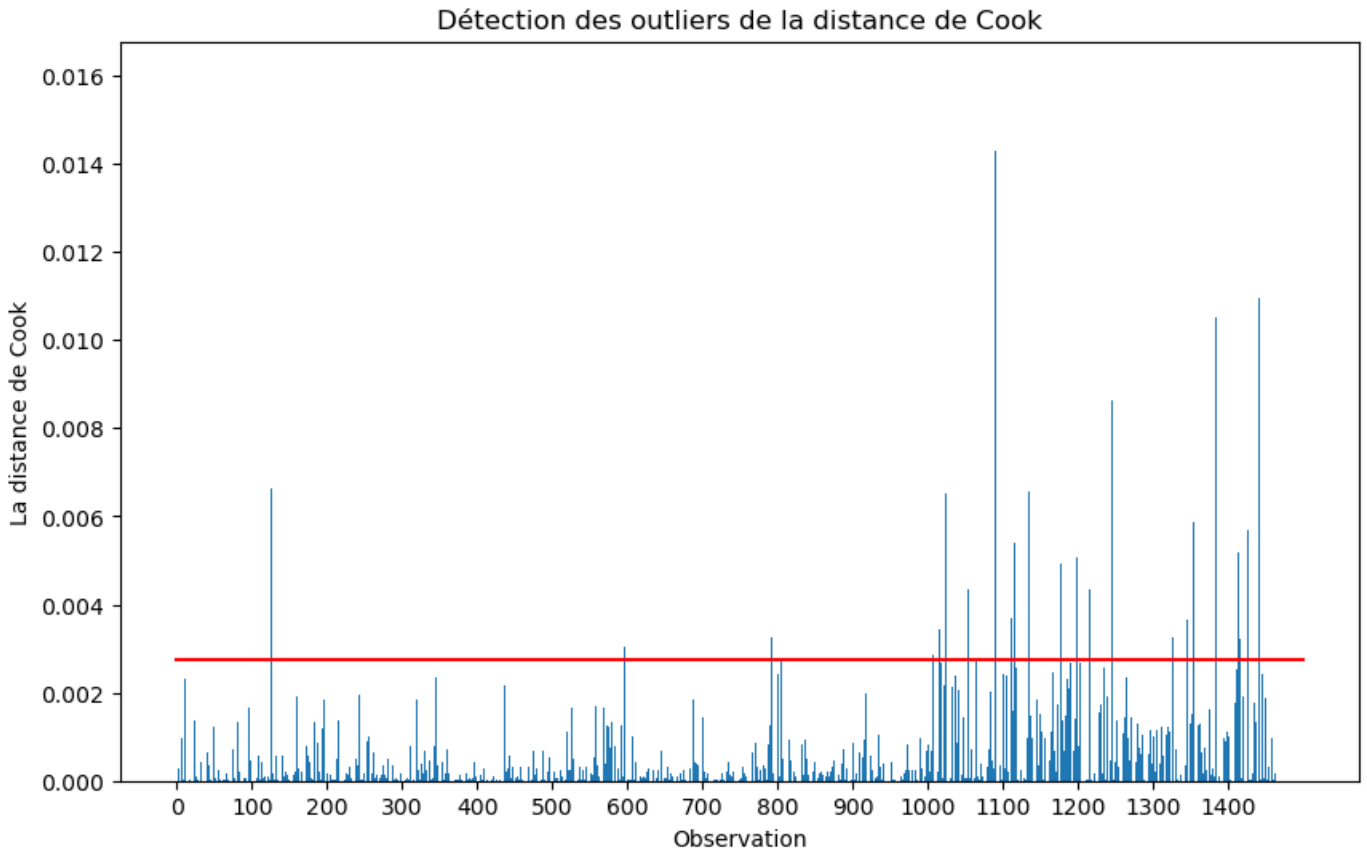
Régression Linéaire

Outliers de la distance de Cook

- 79 faux billets et 14 vrais billets sont des outliers

Processus

- Valeurs aberrantes pour les faux billets ne sont pas très surprenantes
- Mais les différences relatives entre les outliers des vrais billets et les autres vrais billets sont plus surprenantes.
- Donc, après avoir examiné les différences, j'ai supprimé les 14 vrais billets qui étaient des outliers.



93 outliers

ONCFM

Régression Linéaire

Avant

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.617
Model:	OLS	Adj. R-squared:	0.616
Method:	Least Squares	F-statistic:	1174.
Date:	Thu, 08 Feb 2024	Prob (F-statistic):	1.24e-304
Time:	17:48:57	Log-Likelihood:	-774.73
No. Observations:	1463	AIC:	1555.
Df Residuals:	1460	BIC:	1571.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.9263	0.198	30.003	0.000	5.539	6.314
margin_up	-0.2119	0.059	-3.612	0.000	-0.327	-0.097
is_genuine	-1.1632	0.029	-40.477	0.000	-1.220	-1.107

Omnibus:	22.365	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.106
Skew:	0.057	Prob(JB):	3.22e-09
Kurtosis:	3.793	Cond. No.	65.0

Après

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.627
Method:	Least Squares	F-statistic:	1249.
Date:	Sat, 10 Feb 2024	Prob (F-statistic):	1.00e-318
Time:	16:42:32	Log-Likelihood:	-749.04
No. Observations:	1486	AIC:	1504.
Df Residuals:	1483	BIC:	1520.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.0124	0.194	30.969	0.000	5.632	6.393
margin_up	-0.2377	0.058	-4.119	0.000	-0.351	-0.124
is_genuine	-1.1671	0.028	-41.835	0.000	-1.222	-1.112

Omnibus:	30.919	Durbin-Watson:	2.046
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59.806
Skew:	0.087	Prob(JB):	1.03e-13
Kurtosis:	3.967	Cond. No.	66.0

ONCFM

Régression Linéaire

Normalité des résidus

- En termes de statistique, le modèle est bon.
- Mais la valeur p < 0,05 remet en question la normalité des résidus.
- Les résidus ne sont pas très différents d'une distribution symétrique et l'échantillon a plus de 30 individus
- Donc, les résultats obtenus par le modèle ne sont pas absurdes

Conclusion

- Je vais utiliser ce modèle pour imputer les valeurs manquantes

Pas de problème avec la colinéarité = Validé

VIF pour les coefficients = [1,6202, 1,6202]
(Inférieure à 10)

L'homoscédasticité = Pas validé

Le test de p-valeur de Breusch Pagan: 1,9624e-39

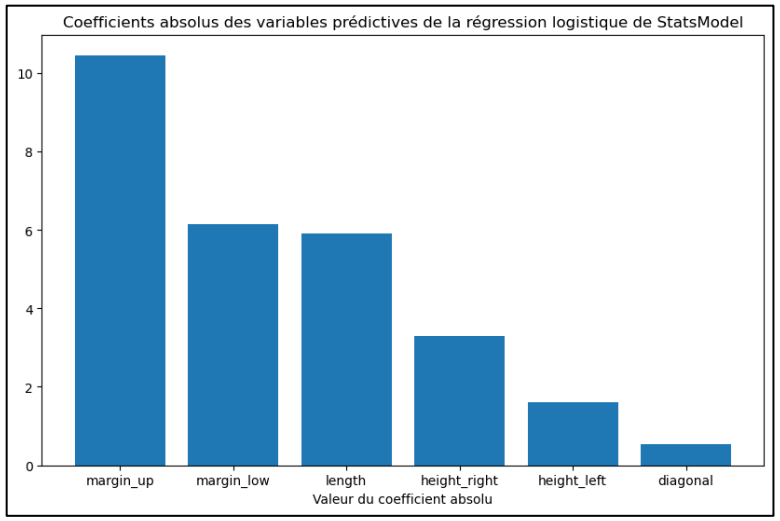
La normalité des résidus = Pas validé

Statistique = 0,9928
p-valeur = 1,7644e-06

ONCFM

CRÉATION DES ALGORITHMES

ONCFM




```
is_genuine ~ margin_up + height_right + length + margin_low + 1
Optimization terminated successfully.
         Current function value: 0.026782
         Iterations 13
is the final model !

                           Logit Regression Results
==============================================================================
Dep. Variable:              is_genuine   No. Observations:                 1500
Model:                          Logit   Df Residuals:                     1495
Method:                           MLE   Df Model:                            4
Date:                Thu, 08 Feb 2024   Pseudo R-squ.:                 0.9579
Time:                        17:49:00   Log-Likelihood:                -40.173
converged:                       True   LL-Null:                       -954.77
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -282.4740    139.419     -2.026      0.043    -555.731      -9.217
margin_up     -10.4098      2.197     -4.738      0.000     -14.716      -6.103
height_right   -3.3512      1.123     -2.984      0.003      -5.553      -1.150
length          6.1592      0.889      6.931      0.000       4.418       7.901
margin_low     -6.3058      0.963     -6.550      0.000      -8.193      -4.419
==============================================================================
```

Statsmodels : Régression Logistique

Bar plot

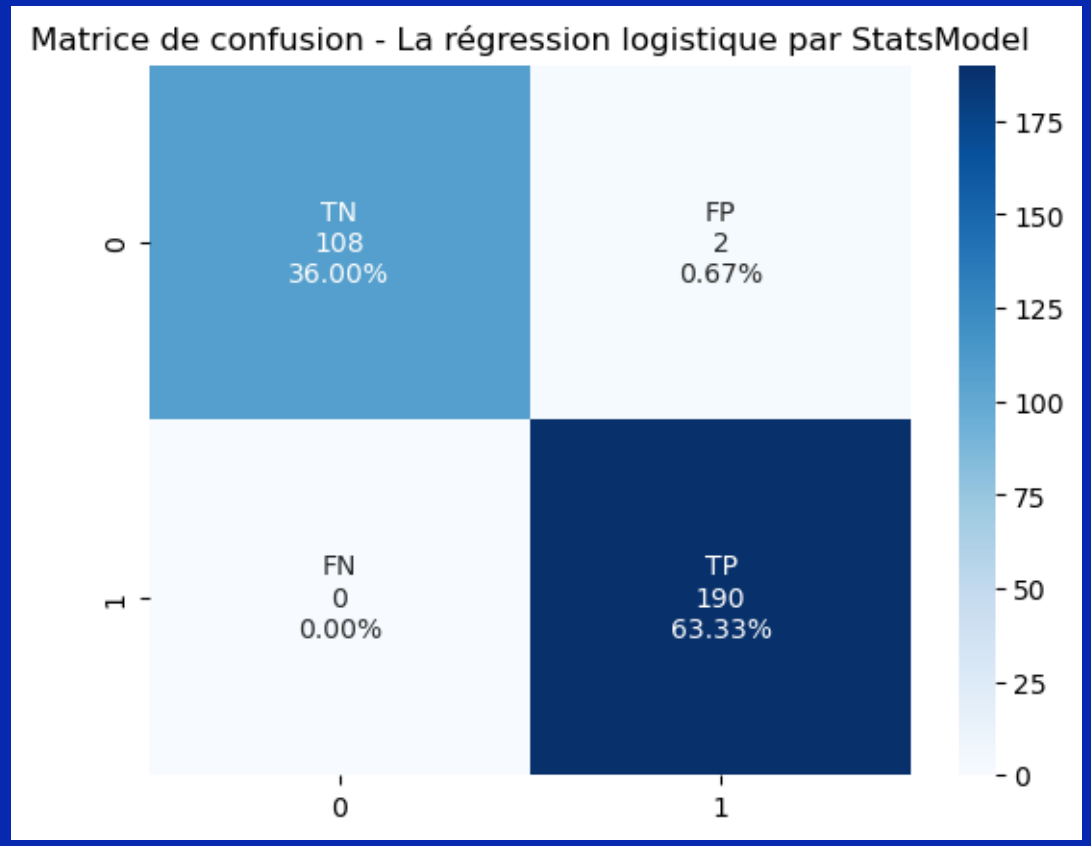
- « margin_up », « margin_low » et « length » ont une influence relativement forte sur « is_genuine »

Modèle

- Le meilleur modèle utilise « margin_up », « margin_low », « length » et « height_right».
- Pseudo $R^2 = 0{,}9579$
- Statistiquement significatif (0,00<0,05)
- Coefficients – « margin_up » = -10,41, « height_right » = -3,35, « length » = 6,16, « margin_low » = -6,31

ONCFM

Statsmodels – Régression Logistique



Accuracy : 99,33%

Précision : 98,96%

Recall : 100%

ROC-AUC : 0,9909

Statsmodels : Régression Logistique

Augmenter le seuil = Réduire les faux positifs

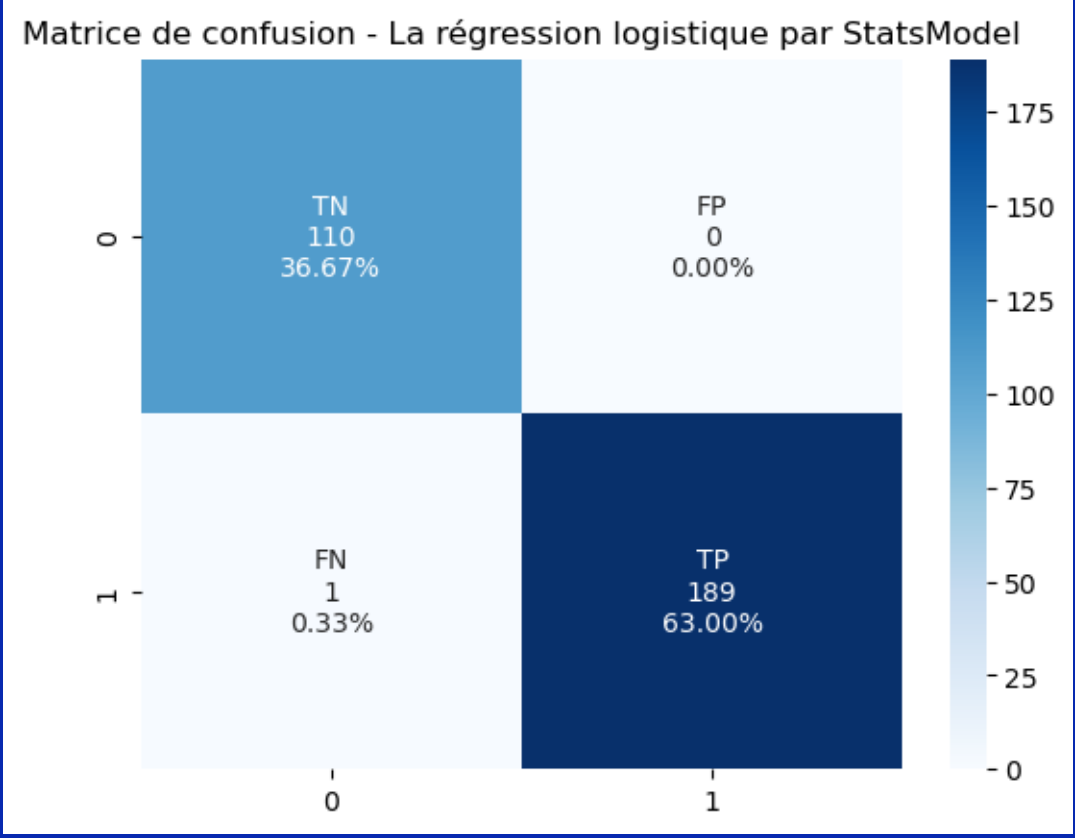
Abaisser le seuil = Réduire les faux négatifs

Seuil de 0,7

- Pas de faux positifs
- MAIS 1 faux négatif

```
Seuil 0.3 - Nombre de True Negative = 106
Seuil 0.3 - Nombre de True Positive = 190
Seuil 0.3 - Nombre de False Negative = 0
Seuil 0.3 - Nombre de False Positive = 4
-------------------------------------------------
Seuil 0.4 - Nombre de True Negative = 106
Seuil 0.4 - Nombre de True Positive = 190
Seuil 0.4 - Nombre de False Negative = 0
Seuil 0.4 - Nombre de False Positive = 4
-------------------------------------------------
Seuil 0.5 - Nombre de True Negative = 108
Seuil 0.5 - Nombre de True Positive = 190
Seuil 0.5 - Nombre de False Negative = 0
Seuil 0.5 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.6 - Nombre de True Negative = 108
Seuil 0.6 - Nombre de True Positive = 189
Seuil 0.6 - Nombre de False Negative = 1
Seuil 0.6 - Nombre de False Positive = 2
-------------------------------------------------
Seuil 0.7 - Nombre de True Negative = 110
Seuil 0.7 - Nombre de True Positive = 189
Seuil 0.7 - Nombre de False Negative = 1
Seuil 0.7 - Nombre de False Positive = 0
-------------------------------------------------
Seuil 0.8 - Nombre de True Negative = 110
Seuil 0.8 - Nombre de True Positive = 187
Seuil 0.8 - Nombre de False Negative = 3
Seuil 0.8 - Nombre de False Positive = 0
-------------------------------------------------
```

Statsmodels – Régression Logistique

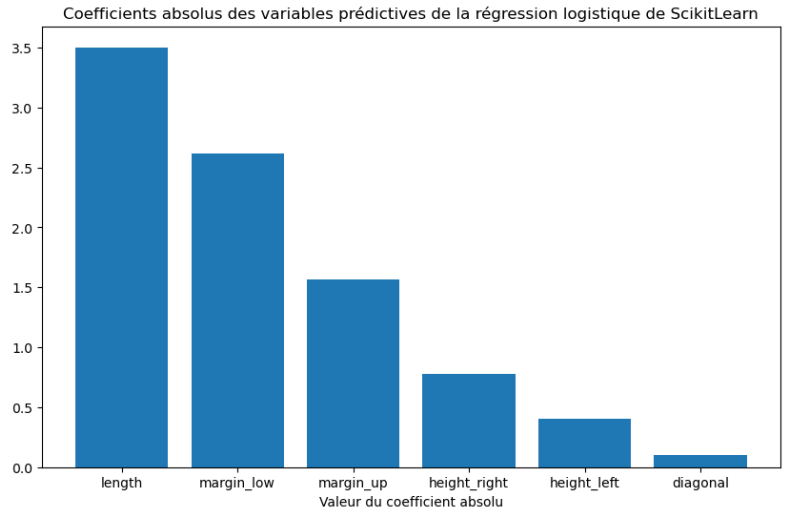


Accuracy : 99,67%

Précision : 100%

Recall : 99,47%

ROC-AUC : 0,9974



	Accuracy	Précision	Recall	Score de ROC-AUC
Length, margin_low, margin_up, height_right, height_left, diagonal	0,9933	0,9896	1	0,9909
Length, margin_low, margin_up, height_right, height_left	0,9933	0,9896	1	0,9909
Length, margin_low, margin_up, height_right	0,99	0,9845	1	0,9864
Length, margin_low, margin_up	0,9933	0,9896	1	0,9909
Length, margin_low	0,9867	0,9794	1	0,9818
Length	0,9533	0,94	1	0,9402

SciKitLearn : Régression Logistique

Bar plot

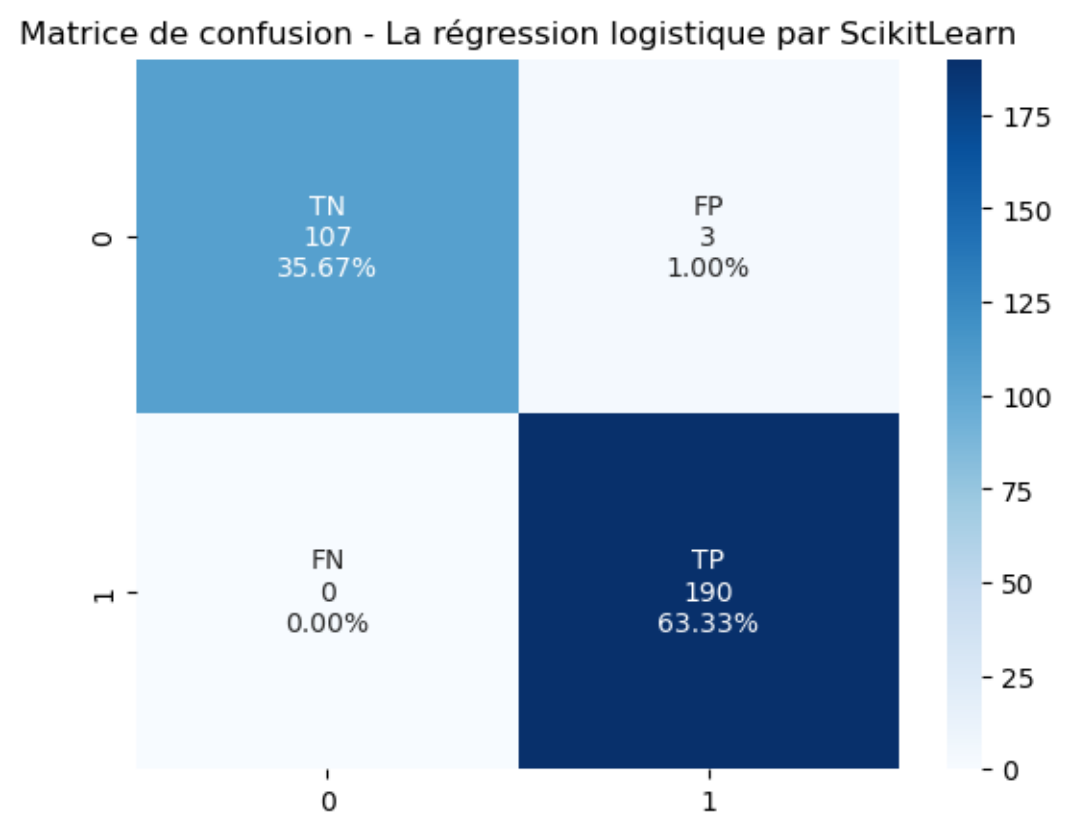
- « length », « margin_low » et « margin_up » sont les variables avec la plus forte influence sur « is_genuine »

Modèle

- Le meilleur modèle utilise « length », « margin_low » et « margin_up », et « height_right».
- Moins de variables simplifie le modèle et réduit les risques de overfitting.

ONCFM

SciKitLearn – Régression Logistique



Accuracy : 99%

Précision : 98,45%

Recall : 100%

ROC-AUC : 0,9864

Statsmodels : Régression Logistique

Augmenter le seuil = Réduire les faux positifs

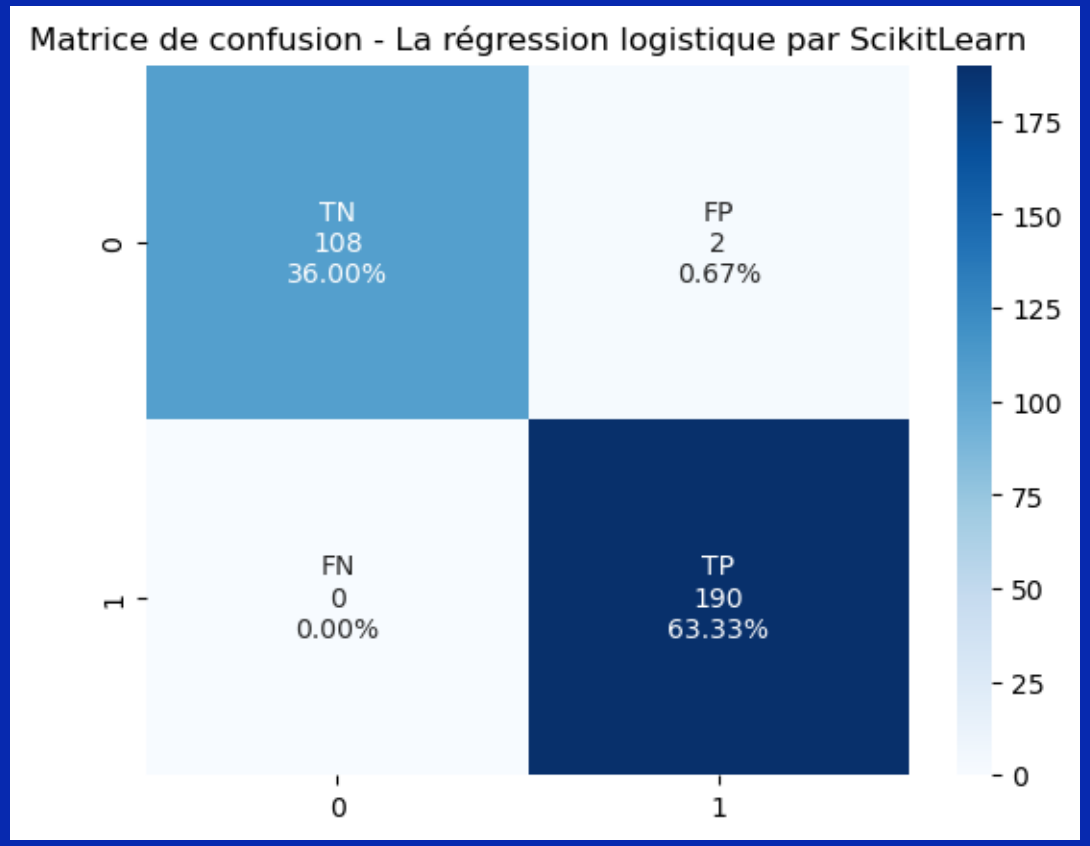
Abaisser le seuil = Réduire les faux négatifs

Seuil de 0,6 ou 0,7

- Pas de faux négatif
- 2 faux positif

```
Seuil 0.3 - Nombre de True Negative = 102
Seuil 0.3 - Nombre de True Positive = 190
Seuil 0.3 - Nombre de False Negative = 0
Seuil 0.3 - Nombre de False Positive = 8
-----------------------------------------------
Seuil 0.4 - Nombre de True Negative = 105
Seuil 0.4 - Nombre de True Positive = 190
Seuil 0.4 - Nombre de False Negative = 0
Seuil 0.4 - Nombre de False Positive = 5
-----------------------------------------------
Seuil 0.5 - Nombre de True Negative = 107
Seuil 0.5 - Nombre de True Positive = 190
Seuil 0.5 - Nombre de False Negative = 0
Seuil 0.5 - Nombre de False Positive = 3
-----------------------------------------------
Seuil 0.6 - Nombre de True Negative = 108
Seuil 0.6 - Nombre de True Positive = 190
Seuil 0.6 - Nombre de False Negative = 0
Seuil 0.6 - Nombre de False Positive = 2
-----------------------------------------------
Seuil 0.7 - Nombre de True Negative = 108
Seuil 0.7 - Nombre de True Positive = 190
Seuil 0.7 - Nombre de False Negative = 0
Seuil 0.7 - Nombre de False Positive = 2
-----------------------------------------------
Seuil 0.8 - Nombre de True Negative = 109
Seuil 0.8 - Nombre de True Positive = 187
Seuil 0.8 - Nombre de False Negative = 3
Seuil 0.8 - Nombre de False Positive = 1
-----------------------------------------------
```

SciKitLearn – Régression Logistique

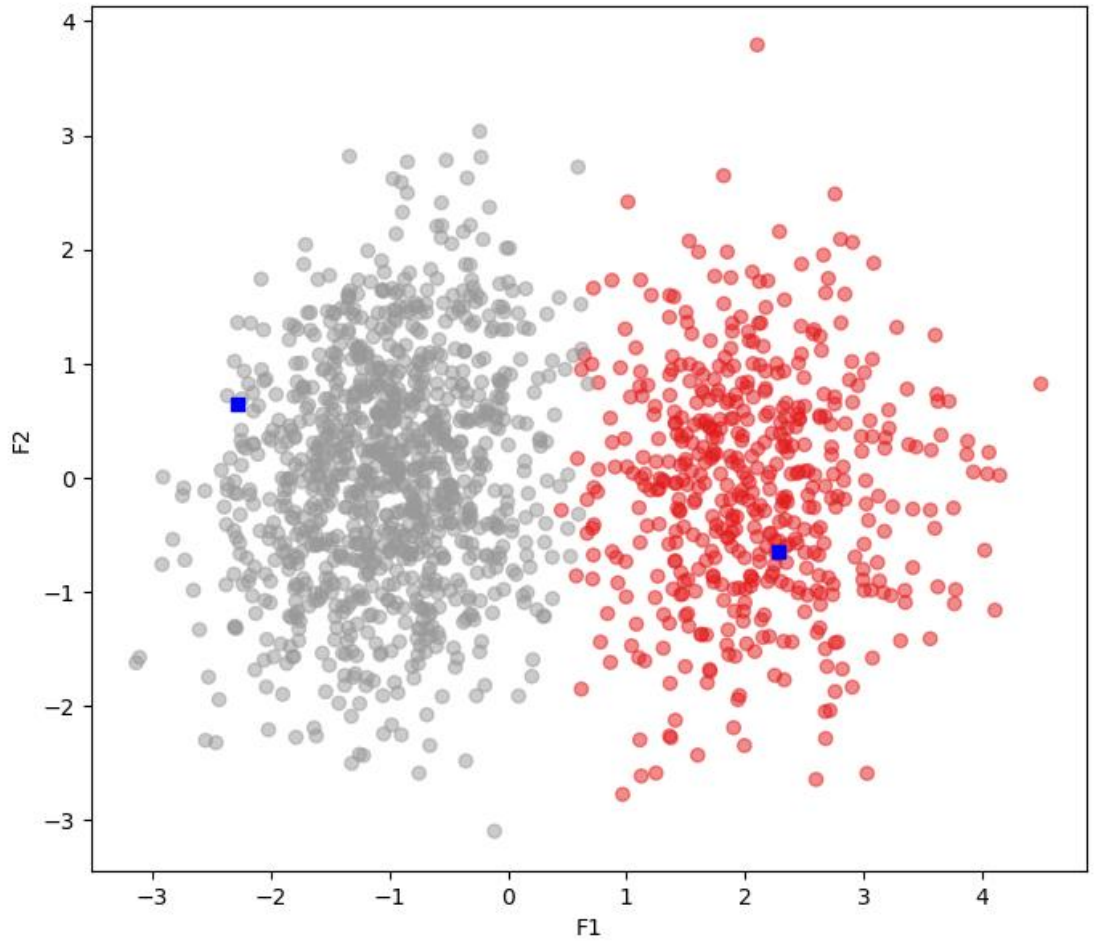


Accuracy : 99,33%

Précision : 98,96%

Recall : 100%

ROC-AUC : 0,9909



K-means

Nombre de clusters

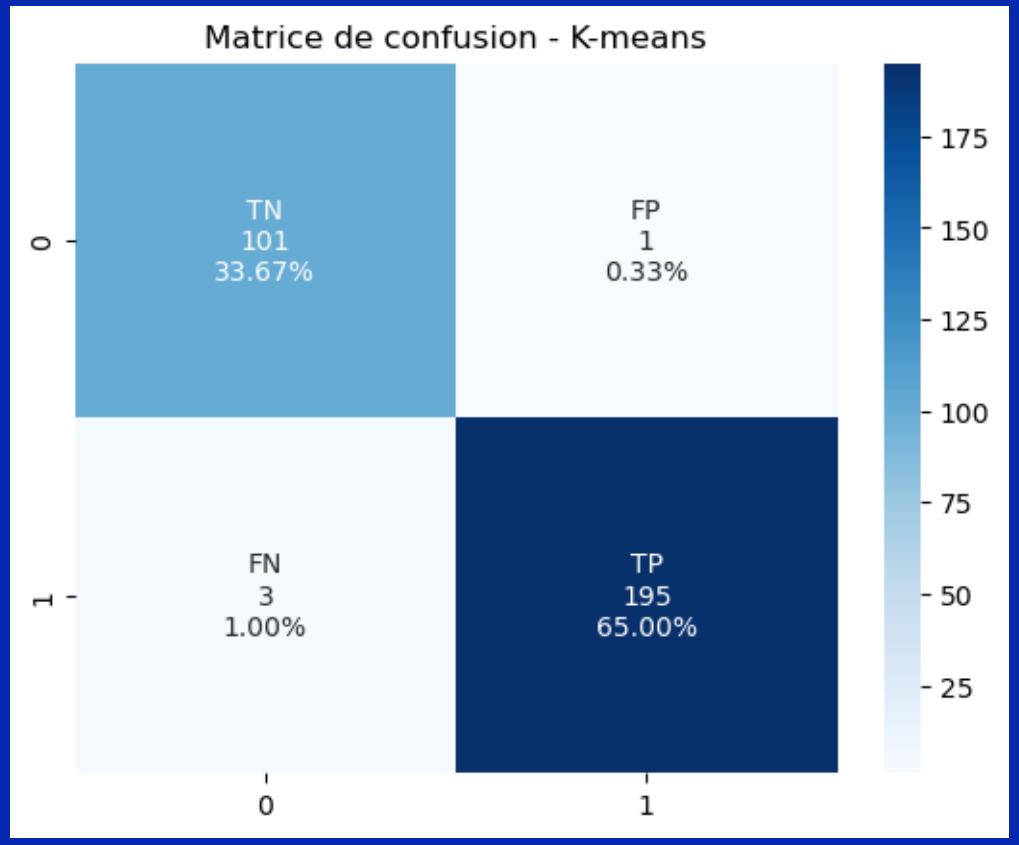
- Pas besoin de la méthode du coude : 2 clusters, les vrais et les faux billets

K-means Plot

- 2 clusters bien définis

ONCFM

K-means



Accuracy : 98,67%

Précision : 99,49%

Recall : 98,48%

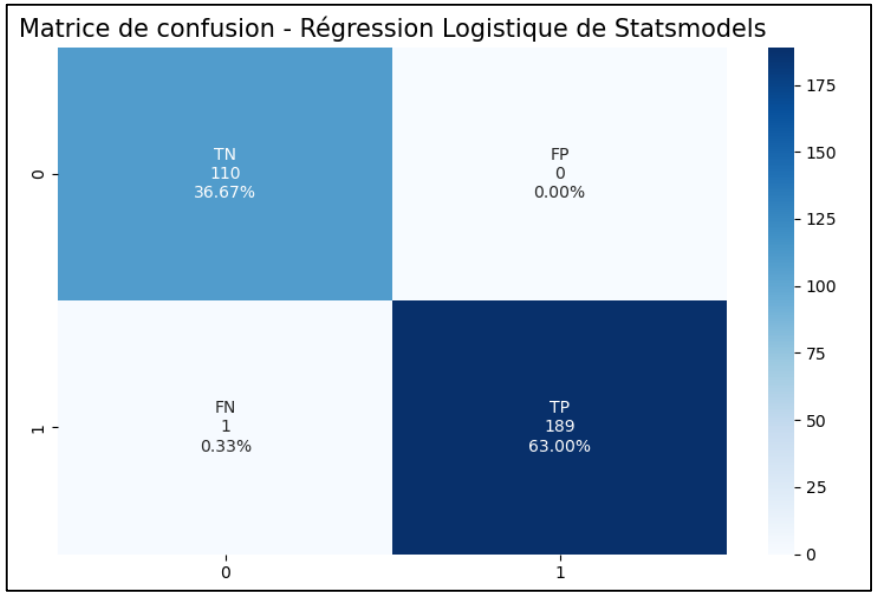
ROC-AUC : 0,9875

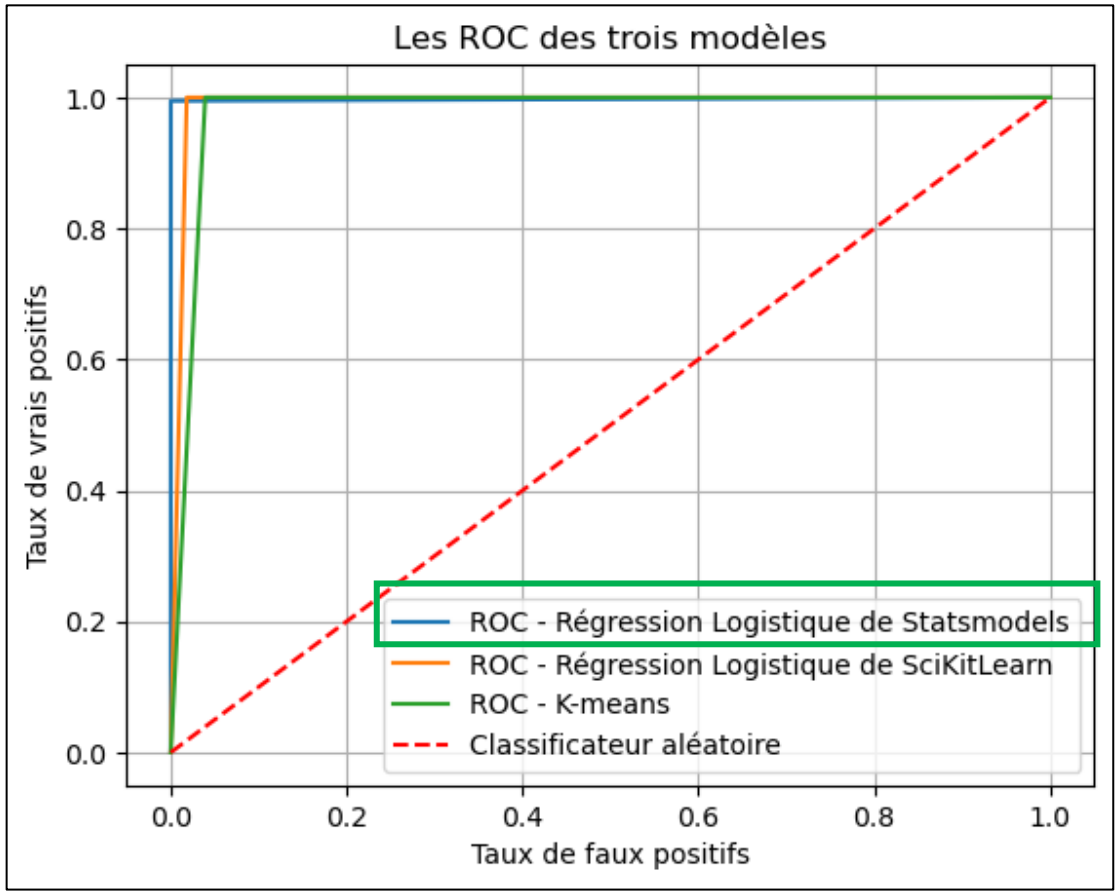
Conclusion

Meilleur algorithme :

Régression logistique de Statsmodels

- Le meilleur score d'accuracy
- Le meilleur score de précision
- Le meilleur score de ROC-AUC





	Accuracy	Précision	Recall	Score de ROC-AUC
Régression logistique de Statsmodels	0,9967	1	0,9947	0,9974
Régression logistique de SciKitLearn	0,9933	0,9896	1	0,9909
K-means	0,9867	0,9802	1	0,9804