Organisation nationale de lutte contre le faux-monnayage (ONCFM)
Système de détection des faux billets

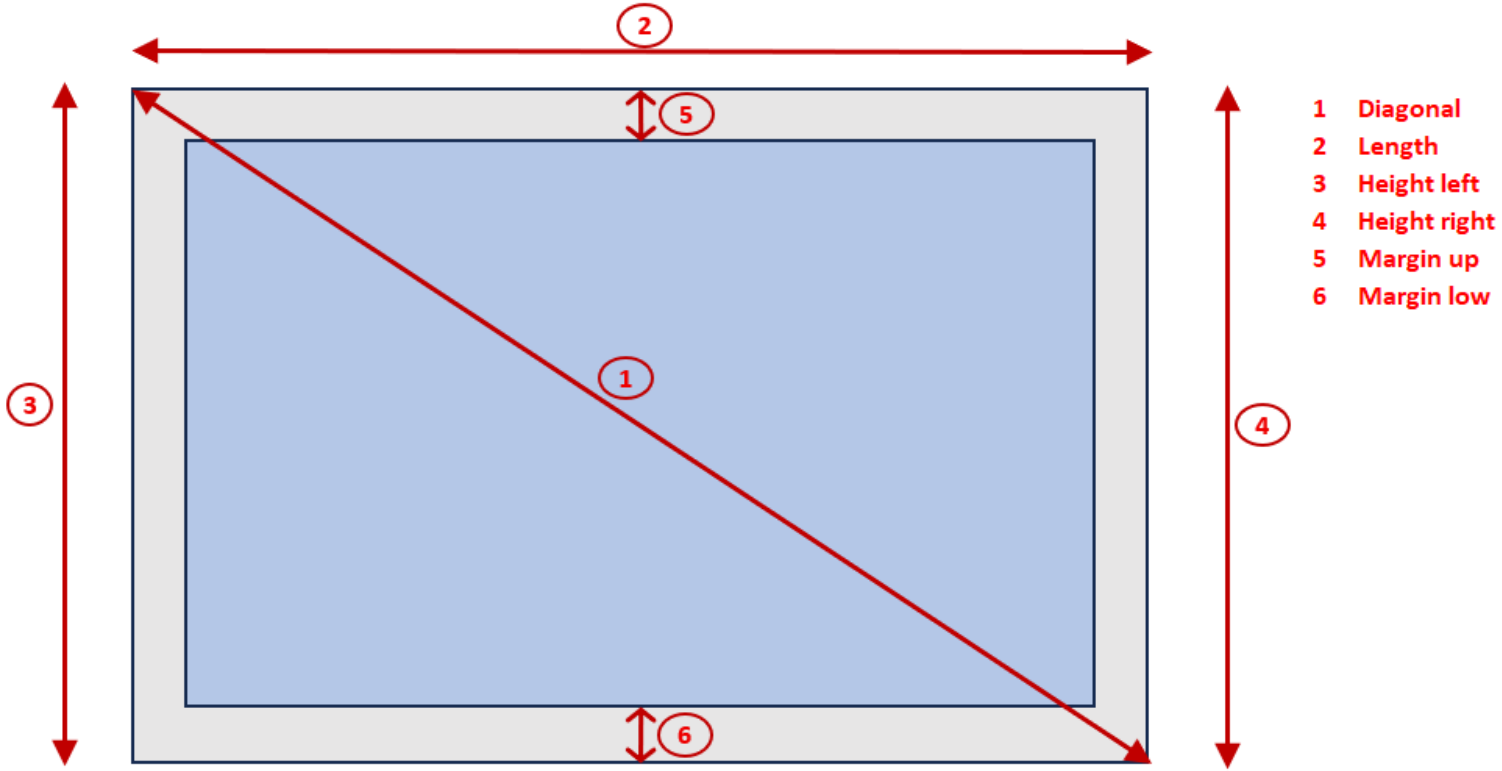
Arnaud Golliot
Senior Data Consultant
1ier septembre 2025

Sommaire

1 – Objectifs de l’étude

1. Objectifs de l'étude

Sur la base des 6 mesures ci-dessous, implanter un système automatique de détection des faux billets



2 – Imputation des valeurs manquantes

2. Imputation des valeurs manquantes [1/2]

OLS Regression Results

Dep. Variable:	margin_low	**R-squared:**	0.993
Model:	OLS	**Adj. R-squared:**	0.992
Method:	Least Squares	**F-statistic:**	2718.
Date:	Fri, 29 Aug 2025	**Prob (F-statistic):**	2.91e-104
Time:	11:09:45	**Log-Likelihood:**	205.27
No. Observations:	105	**AIC:**	-398.5
Df Residuals:	99	**BIC:**	-382.6
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
diagonal	-0.1323	0.013	-10.181	0.000	-0.158	-0.107
height_left	0.1632	0.014	12.069	0.000	0.136	0.190
height_right	0.2788	0.014	19.630	0.000	0.251	0.307
margin_up	0.2717	0.019	14.498	0.000	0.235	0.309
length	-0.4051	0.005	-73.691	0.000	-0.416	-0.394
intercept	26.0736	3.091	8.434	0.000	19.940	32.207

Omnibus:	7.176	**Durbin-Watson:**	2.105
Prob(Omnibus):	0.028	**Jarque-Bera (JB):**	3.580
Skew:	0.207	**Prob(JB):**	0.167
Kurtosis:	2.195	**Cond. No.**	2.27e+05

OLS LINEAR REGRESSION	
Count(X)	1463.0
Count(y)	1463.0
Count(X_train)	105.0
Count(y_train)	105.0
Count(X_test)	293.0
Count(y_test)	293.0
Count(y_pred)	293.0
R2 (First iteration)	0.47
R2 (Last iteration)	0.99
Shapiro pvalue	0.08
Précision	92.02

Annexe : processus itératif d'optimisation du modèle (obtenir pour le test de Shapiro sur les résidus standardisés une Pvalue > 0.05)

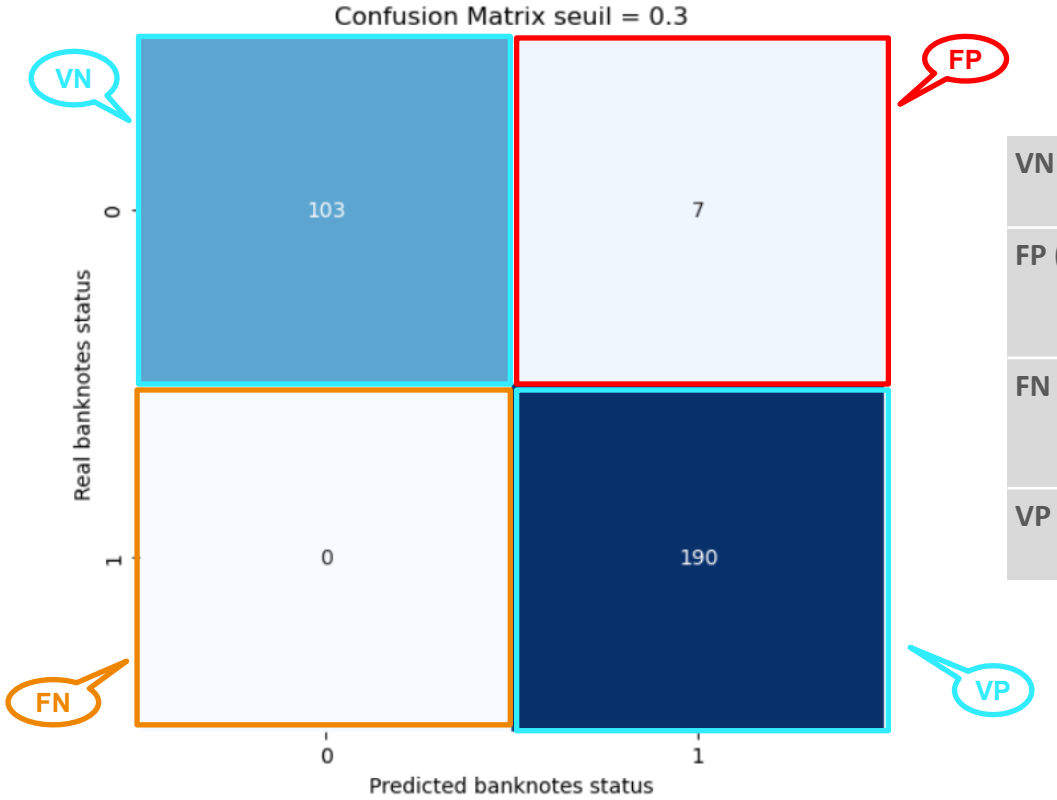
2. Imputation des valeurs manquantes [2/2]

	predicted_margin_low
72	4.31
99	4.40
151	4.42
197	4.35
241	4.65
251	3.80
284	4.19
334	4.14
410	4.12
413	4.15
445	4.15
481	3.78
505	4.06
611	4.30
654	4.15
675	4.10
710	4.44
739	4.47
742	4.34

	predicted_margin_low (ols standard)
780	4.07
798	3.62
844	4.39
845	4.09
871	4.26
895	3.90
919	3.74
945	4.23
946	4.74
981	4.13
1076	5.06
1121	4.81
1176	5.07
1303	5.02
1315	4.78
1347	5.73
1435	5.18
1438	5.14

3 – Modèles prédictifs

3.1 Modèles prédictifs : quelques notions préalables



VN (Vrais Négatifs)	Nombre de billets dont le statut réel et prédit est faux
FP (Faux Positifs)	Nombre de billets dont le statut prédit est vrai alors que le statut réel est faux ***A minimiser impérativement***
FN (Faux Négatifs)	Nombre de billets dont le statut prédit est faux alors que le statut réel est vrai ***A minimiser idéalement***
VP (Vrais positifs)	Nombre de billets dont le statut réel et prédit est vrai

3.1 Modèles prédictifs : vérification de non colinéarité des variables prédicatives

Calcul du variance_inflation_factor (VIF) pour chacune de variable prédicative des données d'entraînement (X_train).

Un score supérieur à 1 pour toutes les variables prédicatives confirme qu'elles sont toutes indépendantes les unes des autres

	Attribute	VIF Scores
0	diagonal	169608.194308
1	height_left	115336.336244
2	height_right	104539.727245
3	margin_low	90.755053
4	margin_up	263.860126
5	length	31241.821380

3.2 – Modèles prédictifs : Régression logistique MLE (Maximum Likehood Estimation)

3.2 Modèles prédictifs : régression logistique MLE [1/3]

Model:	Logit	Method:	MLE
Dependent Variable:	is_genuine	Pseudo R-squared:	0.952
Date:	2025-08-31 20:39	AIC:	86.6362
No. Observations:	1200	BIC:	122.2667
Df Model:	6	Log-Likelihood:	-36.318
Df Residuals:	1193	LL-Null:	-756.70
Converged:	1.0000	LLR p-value:	3.6192e-308
No. Iterations:	13.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
diagonal	-0.3386	1.1429	-0.2962	0.7671	-2.5787	1.9016
height_left	-1.8980	1.2709	-1.4935	0.1353	-4.3888	0.5929
height_right	-2.2359	1.0916	-2.0482	0.0405	-4.3754	-0.0963
margin_low	-5.3575	0.9827	-5.4519	0.0000	-7.2836	-3.4315
margin_up	-8.7177	2.1010	-4.1494	0.0000	-12.8355	-4.5999
length	5.6808	0.9094	6.2465	0.0000	3.8984	7.4633
intercept	-98.4322	257.7129	-0.3819	0.7025	-603.5402	406.6758

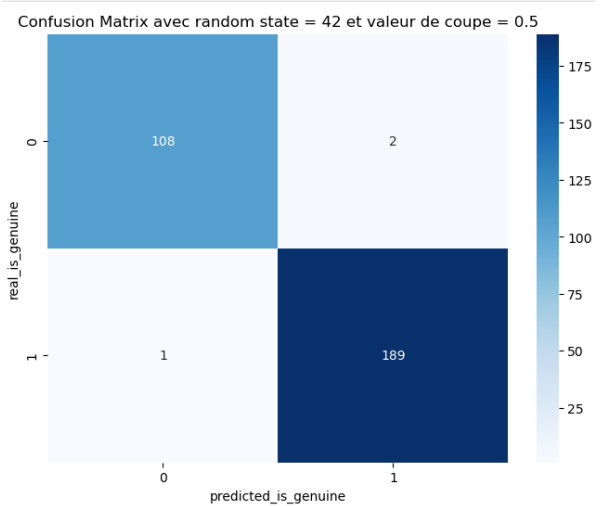
3.2 Modèles prédictifs : régression logistique MLE [2/3]

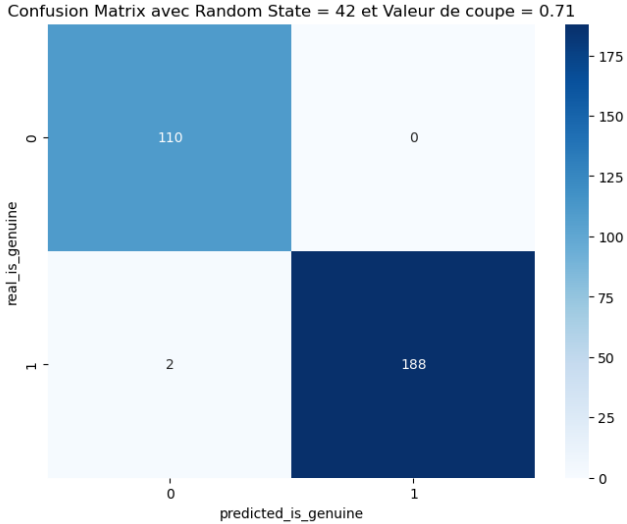
FIRST LOGIT REGRESSION MODEL	
Random State	42.0
Valeur de coupe	0.5
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.990000
RMSE	0.010000
MAPE	30023997515803.308594
Précision (directe)	-3002399751580231.000000
Précision (calculée)	-inf
Précision (ajustée)	99.00
VN	108.0
FP	2.0
FN	1.0
VP	189.0

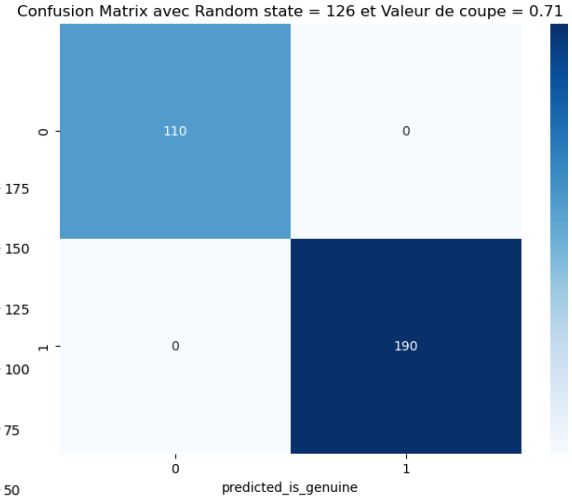
INTERMEDIATE LOGIT REGRESSION MODEL	
Random State	42.0
Valeur de coupe	0.7
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.993333
RMSE	0.006667
MAPE	0.006667
PRECISION (DIRECTE)	99.330000
PRECISION (CALCULEE)	98.947368
PRECISION (AJUSTEE)	99.33
VN	110.0
FP	0.0
FN	2.0
VN	188.0

OPTIMAL LOGIT REGRESSION MODEL	
Random state value	126.0
Valeur de coupe	0.71
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.993333
RMSE	0.006667
MAPE	0.006667
Précision (directe)	99.33
Précision (calculée)	100.00
Précision (ajustée)	100.0
VN	110.0
FP	0.0
FN	0.0
VP	190.0

3.2 Modèles prédictifs : régression logistique MLE [3/3]

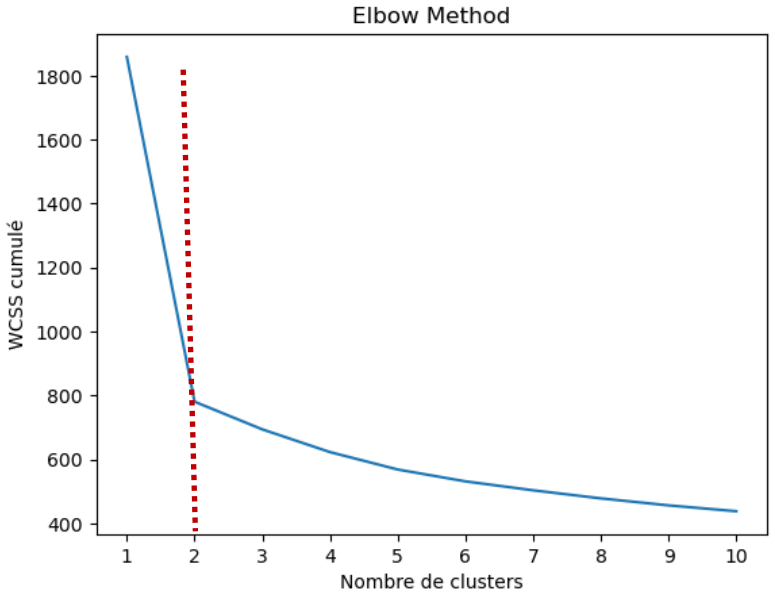


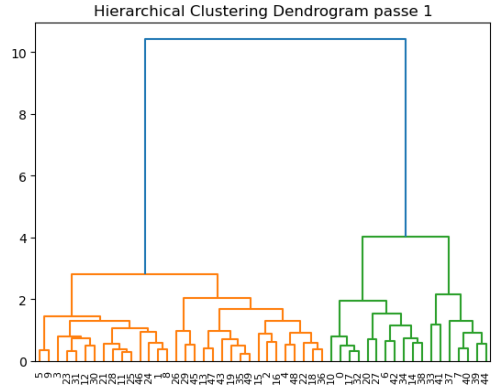




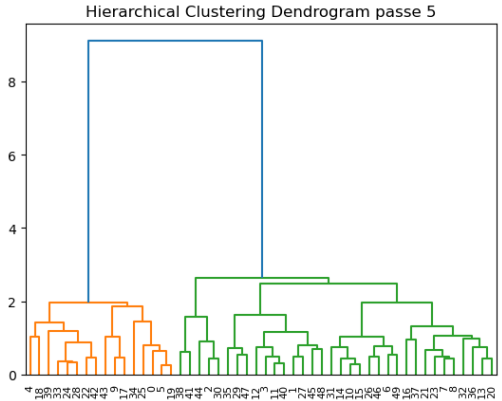
3.3 – Modèles prédictifs : Kmeans Clustering

3.3 Modèles prédictifs : Kmeans Clustering [1/3]





...



STEP	NB CLUSTERS
1	2.0
2	2.0
3	2.0
4	2.0
5	2.0
6	2.0
7	2.0
8	2.0
9	2.0
10	2.0

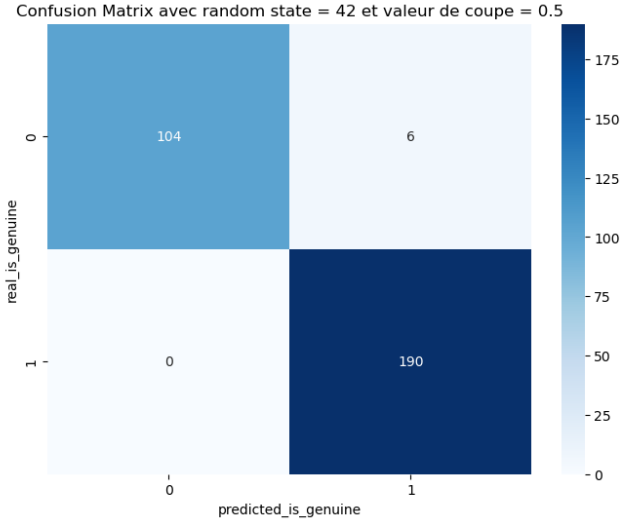
3.3 Modèles prédictifs : Kmeans Clustering [2/3]

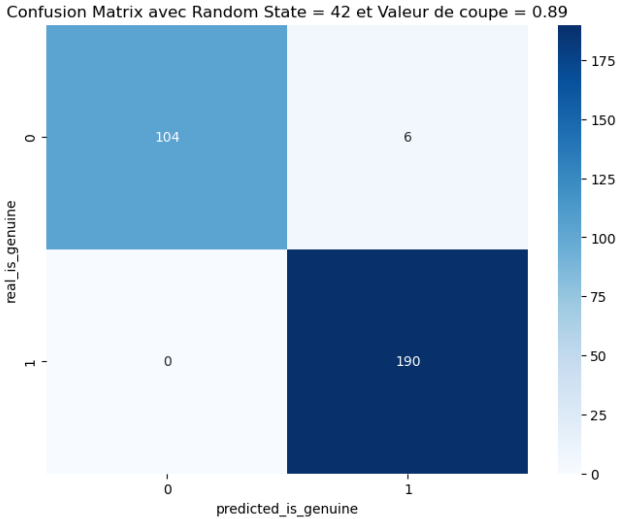
FIRST KMEANS CLUSTERING MODEL	
Random State	42.0
Valeur de coupe	0.5
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.980000
RMSE	0.020000
MAPE	90071992547409.921875
Précision (directe)	-9007199254740892.000000
Précision (calculée)	-inf
Précision (ajustée)	98.00
VN	104.0
FP	6.0
FN	0.0
VP	190.0

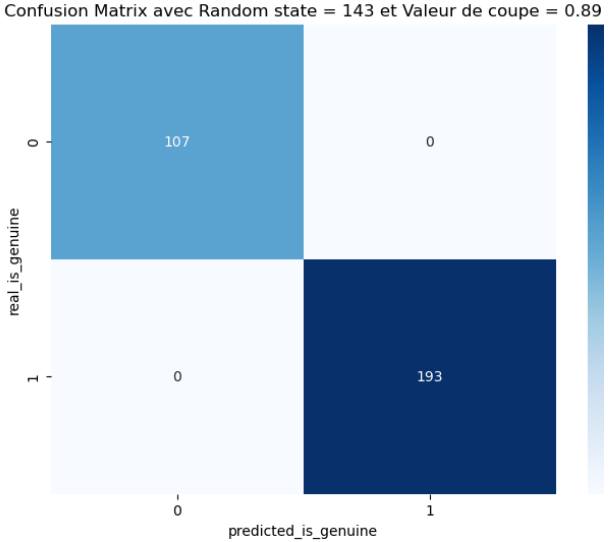
INTERMEDIATE KMEANS CLUSTERING MODEL	
Random State	42.0
Valeur de coupe	0.89
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.980000
RMSE	0.020000
MAPE	90071992547409.921875
Précision (directe)	-9007199254740892.000000
Précision (calculée)	-inf
Précision (ajustée)	98.00
VN	104.0
FP	6.0
FN	0.0
VN	190.0

FINAL KMEANS CLUSTERING MODEL	
Random state value	143.0
Valeur de coupe	0.89
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.980000
RMSE	0.020000
MAPE	90071992547409.921875
Précision (directe)	-9007199254740892.00
Précision (calculée)	100.00
Précision (ajustée)	100.0
VN	107.0
FP	0.0
FN	0.0
VP	193.0

3.3 Modèles prédictifs : Kmeans Clustering [3/3]

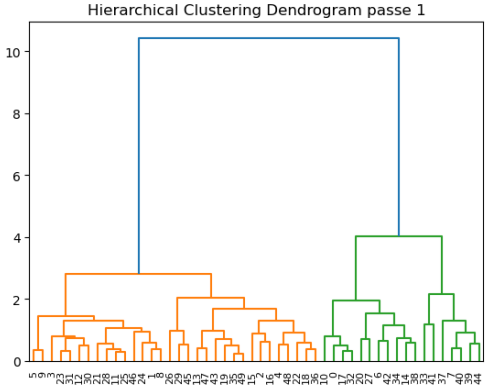




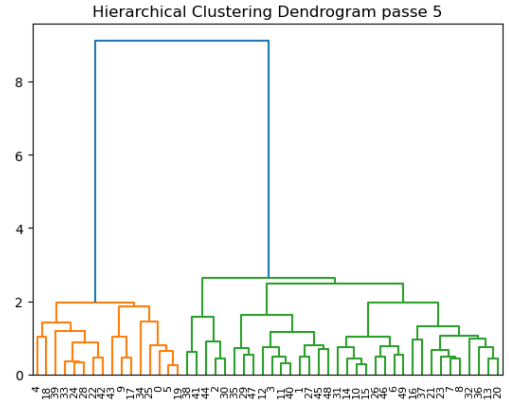


3.4 – Modèles prédictifs : Classification KNN

3.4 Modèles prédictifs : Classification KNN [1/3]



...



STEP	NB CLUSTERS
1	2.0
2	2.0
3	2.0
4	2.0
5	2.0
6	2.0
7	2.0
8	2.0
9	2.0
10	2.0

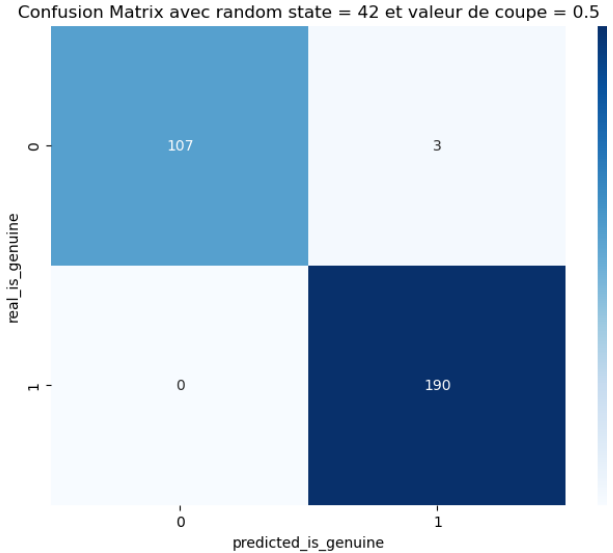
3.4 Modèles prédictifs : Classification KNN [2/3]

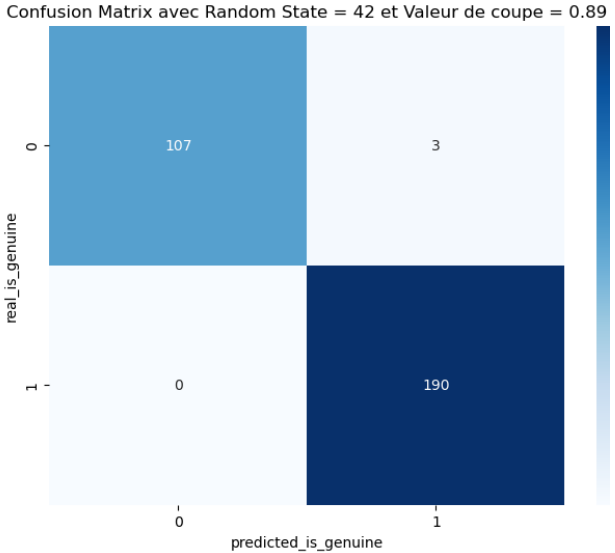
	INITIAL KNN MODEL
Random state value	42.0
Valeur de coupe	0.50
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.990000
RMSE	0.010000
MAPE	45035996273704.960938
Précision (directe)	-4503599627370396.00
Précision (calculée)	-inf
Précision (ajustée)	99.0
VN	107.0
FP	3.0
FN	0.0
VP	190.0

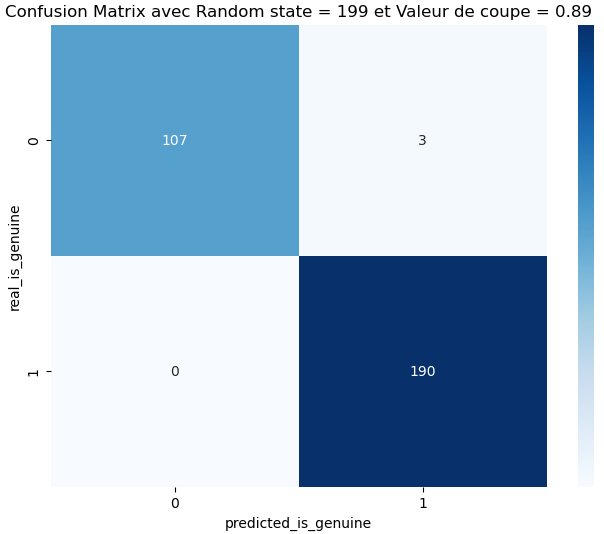
	INTERMEDIATE KNN MODEL
Random State	42.0
Valeur de coupe	0.89
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.540000
RMSE	0.460000
MAPE	1140911905600525.750000
Précision (directe)	-1140911905600524 64.000000
Précision (calculée)	-inf
Précision (ajustée)	98.33
VN	107.0
FP	3.0
FN	0.0
VN	190.0

	FINAL KNN MODEL
Random state value	199.0
Valeur de coupe	0.89
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.990000
RMSE	0.010000
MAPE	45035996273704.960938
Précision (directe)	-4503599627370396.00
Précision (calculée)	-inf
Précision (ajustée)	98.3
VN	107.0
FP	3.0
FN	0.0
VP	190.0

3.4 Modèles prédictifs : Classification KNN [3/3]

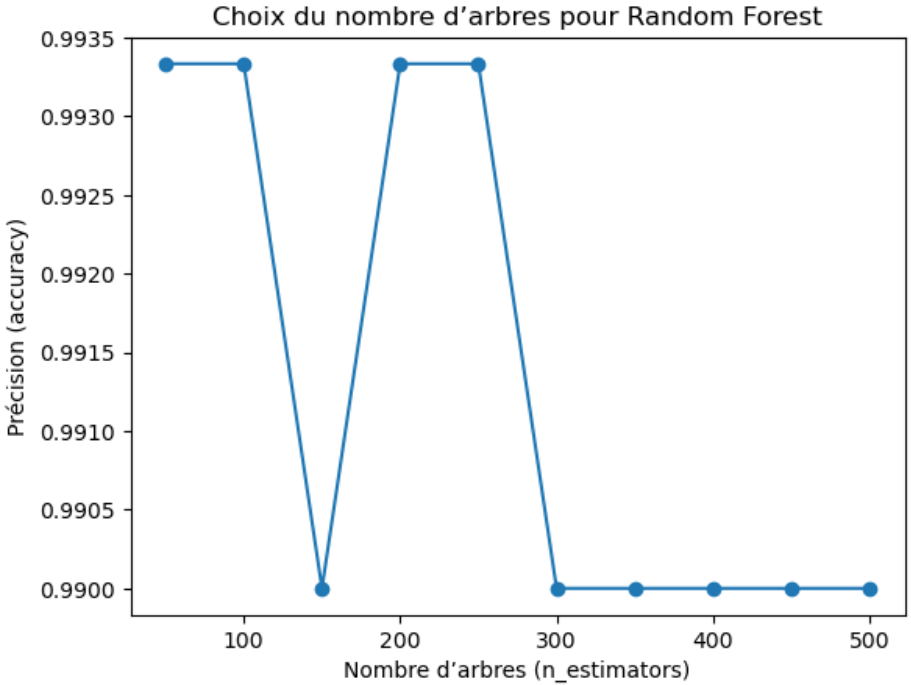






3.5 – Modèles prédictifs : Random Forest

3.5 Modèles prédictifs : Random Forest [1/3]




```
RandomForestClassifier
RandomForestClassifier(n_estimators=250, random_state=42)
```

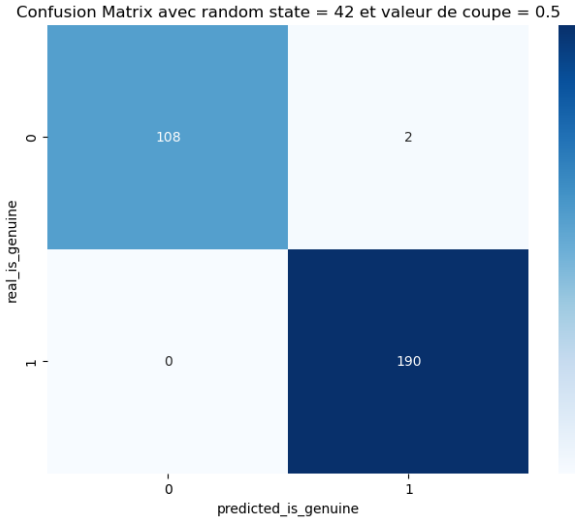
3.5 Modèles prédictifs : Random Forest [2/3]

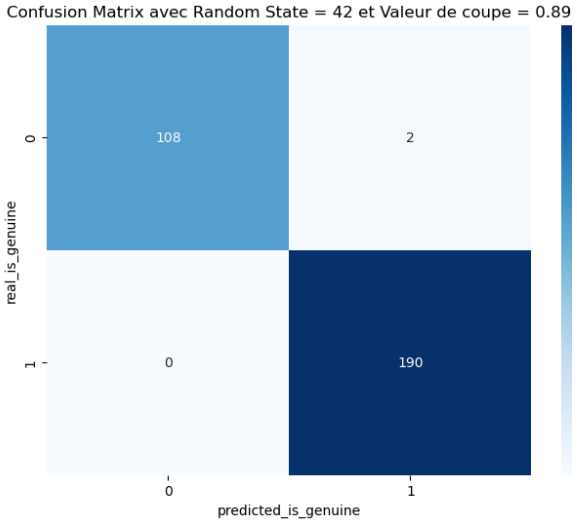
	INITIAL RANDOM FOREST
Random state value	42.0
Valeur de coupe	0.50
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.993333
RMSE	0.006667
MAPE	30023997515803.308594
Précision (directe)	-3002399751580231.00
Précision (calculée)	-inf
Précision (ajustée)	99.3
VN	108.0
FP	2.0
FN	0.0
VP	190.0

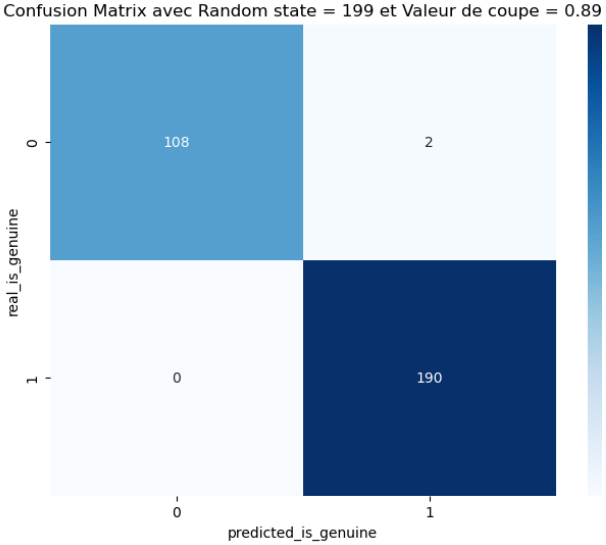
	INTERMEDIATE RANDOM FOREST
Random State	42.0
Valeur de coupe	0.9
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.993333
RMSE	0.006667
MAPE	30023997515803.308594
Précision (directe)	-3002399751580231.000000
Précision (calculée)	-inf
Précision (ajustée)	99.33
VN	108.0
FP	2.0
FN	0.0
VN	190.0

	FINAL RANDOM FOREST
Random state value	199.0
Valeur de coupe	0.89
count(X)	1500.0
count(y)	1500.0
count(X_train)	1200.0
count(y_train)	1200.0
count(X_test)	300.0
count(y_test)	300.0
count(y_pred)	300.0
Accuracy	0.993333
RMSE	0.006667
MAPE	30023997515803.308594
Précision (directe)	-3002399751580231.00
Précision (calculée)	-inf
Précision (ajustée)	98.7
VN	108.0
FP	2.0
FN	0.0
VP	190.0

3.5 Modèles prédictifs : Random Forest [3/3]







4 – Conclusions de l’étude

4. Conclusions de l'étude

Modèle	Modèle initial							Modèle optimisé (ou dégradé)							Commentaire
	Paramètres	R2 ou Accuracy	Précision	VN	FP	FN	VP	Paramètres	R2 ou Accuracy	Précision	VN	FP	FN	VP	
Régression logistique MLE	Cut : 0.5 Rstate : 42	0.99	99.0%	108	2	1	189	Cut : 0.71 Rstate : 126	0,993	99,33%	110	0	0	190	Un modèle optimisé avec effet report cohérent de FP sur VN (2) et de FN sur VP (1)
Kmeans	Cut : 0.5 Rstate : 42	0,98	98%	104	6	0	190	Cut : 0.89 Rstate : 143	0,98	100%	107	0	0	193	Un effet report incohérent de FP sur VN et de FN sur VP
KNN	Cut : 0.5 Rstate : 42	0.99	98.3%	107	3	0	190	Cut : 0.89 Rstate : 199	0,99	99,3%	107	3	0	190	Les optimisations sont sans en effet sur ce modèle
Random Forest	Cut : 0.5 Rstate : 42	0.993	99.3%	108	2	0	190	Cut : 0.89 Rstate : 199	0,993	98.7%	108	2	0	190	Les optimisations sont sans en effet sur ce modèle

5 – Limites de l’étude

5. Limites de l’étude [1/2]

Taille du jeu de données d’entraînement et de test : Disposer d’un jeu de données d’entraînement et de test plus grand (15 000 billets au lieu de 1500 par exemple) de manière à vérifier la robustesse du modèle.

Appliquer le principe de parcimonie au modèle de régression logistique : Expliquer la variable "cible" ("is_genuine") avec un minimum de variables prédicatives (comment ? voir page suivante)

Affermir la validité du modèle de régression logistique au moyen de l’homoscédasticité : seule la vérification de normalité des résidus standardisés a été appliquée (Pvalue > 0.05 sur le test de Shapiro)

Confronter le modèle de régression logistique MLE (fonction Python Logit de la librairie StatsModels) avec le modèle de régression logistique OLS (fonction Python LogisticRegression de la librairie Sklearn)

5. Limites de l'étude [2/2]

Model:	Logit	Method:	MLE
Dependent Variable:	is_genuine	Pseudo R-squared:	0.952
Date:	2025-08-31 20:39	AIC:	86.6362
No. Observations:	1200	BIC:	122.2667
Df Model:	6	Log-Likelihood:	-36.318
Df Residuals:	1193	LL-Null:	-756.70
Converged:	1.0000	LLR p-value:	3.6192e-308
No. Iterations:	13.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
diagonal	-0.3386	1.1429	-0.2962	0.7671	-2.5787	1.9016
height_left	-1.8980	1.2709	-1.4935	0.1353	-4.3888	0.5929
height_right	-2.2359	1.0916	-2.0482	0.0405	-4.3754	-0.0963
margin_low	-5.3575	0.9827	-5.4519	0.0000	-7.2836	-3.4315
margin_up	-8.7177	2.1010	-4.1494	0.0000	-12.8355	-4.5999
length	5.6808	0.9094	6.2465	0.0000	3.8984	7.4633
intercept	-98.4322	257.7129	-0.3819	0.7025	-603.5402	406.6758

Principe de parcimonie : Expliquer la variable "cible" ("is_genuine") avec un minimum de variables prédicatives

Les résultats obtenus montrent que la variable "diagonal" principalement, et la variable "height_left" en seconde intention, contribuent à l'explication de la variable "is_genuine" de façon marginale.

On pourrait donc potentiellement s'affranchir de ces deux variables dans notre modèle

5 – Annexes

Annexe 5.1 : processus itératif de régression linéaire OLS

_distance	count(X_train)	count(y_train)	count(reduced_X_train)	count(reduced_y_train)	updated_ols_r2	updated_shapiro_pvalue	count(y_test)	count(y_pred)	precision
0.600000	1170	1170	1170	1170	0.473171	3.641788e-10	293	293	92.008314
0.500000	1170	1170	1170	1170	0.473171	3.641788e-10	293	293	92.008314
0.004000	1170	1170	1114	1114	0.534528	9.043207e-06	293	293	92.051856
0.003500	1170	1170	1098	1098	0.548862	2.198142e-05	293	293	92.053575
0.003000	1170	1170	1083	1083	0.566115	5.359341e-06	293	293	92.052002
0.002500	1170	1170	1057	1057	0.573017	2.829902e-05	293	293	92.065505
0.002000	1170	1170	1019	1019	0.590218	1.432584e-05	293	293	92.080038
0.001500	1170	1170	968	968	0.618124	1.097475e-05	293	293	92.093481
0.001000	1170	1170	883	883	0.670412	4.065772e-06	293	293	92.082485
0.000500	1170	1170	719	719	0.770866	4.064763e-05	293	293	92.070960
0.000200	1170	1170	518	518	0.857727	1.750023e-05	293	293	92.073097
0.000010	1170	1170	109	109	0.992247	2.173495e-02	293	293	92.027668
0.000009	1170	1170	105	105	0.992768	8.310307e-02	293	293	92.016637

OLS LINEAR REGRESSION	
Count(X)	1463.0
Count(y)	1463.0
Count(X_train)	105.0
Count(y_train)	105.0
Count(X_test)	293.0
Count(y_test)	293.0
Count(y_pred)	293.0
R2 (First iteration)	0.47
R2 (Last iteration)	0.99
Shapiro pvalue	0.08
Précision	92.02