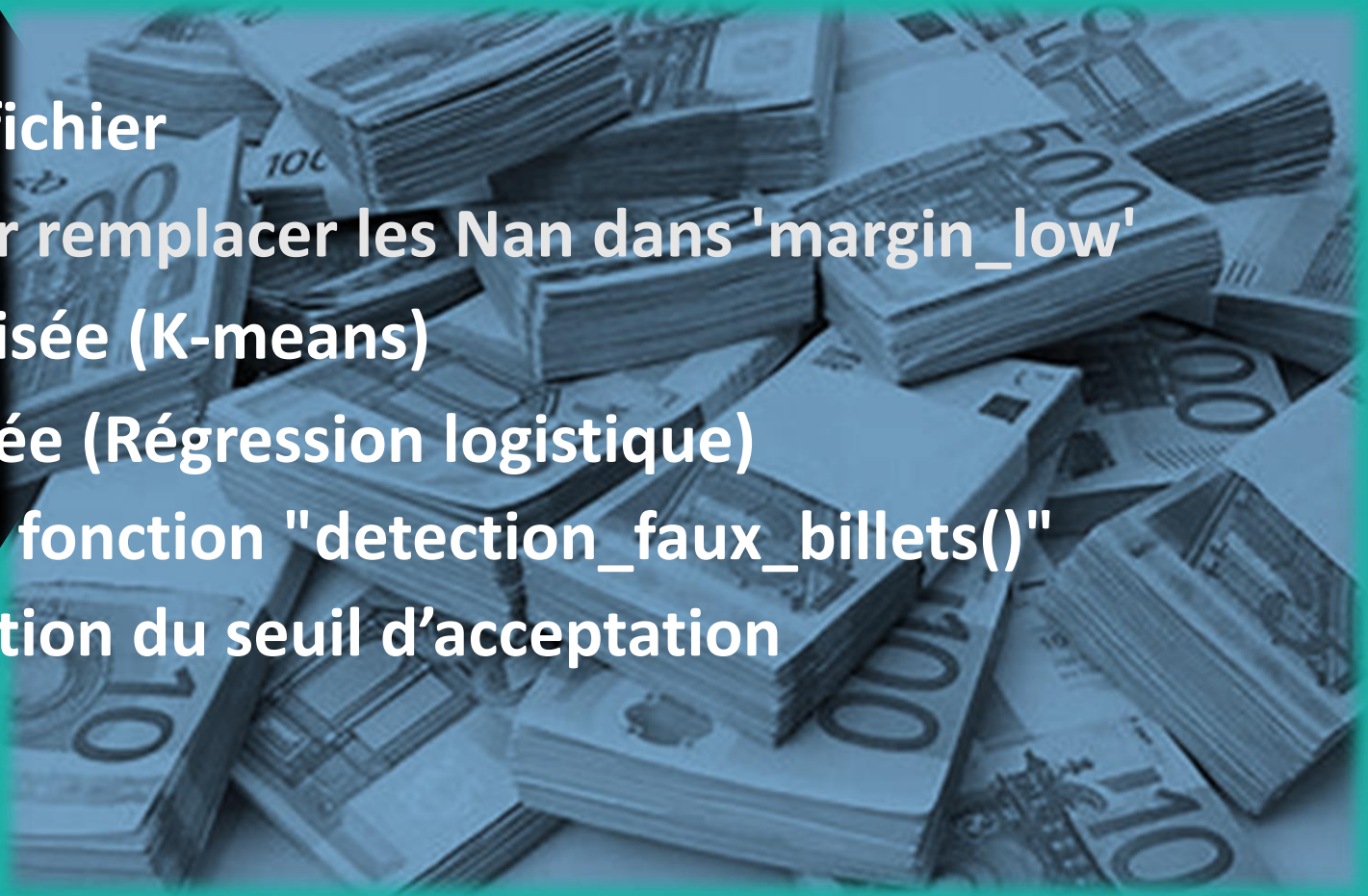
Détectez des faux billets avec Python

ONCFM

- Contexte
- Analyse préparatoire du fichier
- Régression linéaire pour remplacer les Nan dans 'margin_low'
- Méthode non supervisée (K-means)
- Méthode supervisée (Régression logistique)
- Résultats avec la fonction "detection_faux_billets()"
- Outil de définition du seuil d'acceptation



Stéphane Srsa

Contexte

L'Organisation nationale de lutte contre le faux-monnayage, ou ONCFM, est une organisation publique ayant pour objectif de mettre en place des méthodes d'identification des contrefaçons des billets en euros.

La mission

Mise en place d'un algorithme capable de différencier automatiquement les vrais des faux billets pour l'ONCFM.

Dimensions à prendre en compte pour créer le modèle :

- **length** : la longueur du billet (en mm)
- **height_left** : la hauteur du billet (mesurée sur le côté gauche, en mm)
- **height_right** : la hauteur du billet (mesurée sur le côté droit, en mm)
- **margin_up** : la marge entre le bord supérieur du billet et l'image de celui-ci (en mm)
- **margin_low** : la marge entre le bord inférieur du billet et l'image de celui-ci (en mm)
- **diagonal** : la diagonale du billet (en mm)

Analyse préparatoire du fichier

ONCFM

Information fichier :

‘is_genuine’ différencie vrais et faux billets

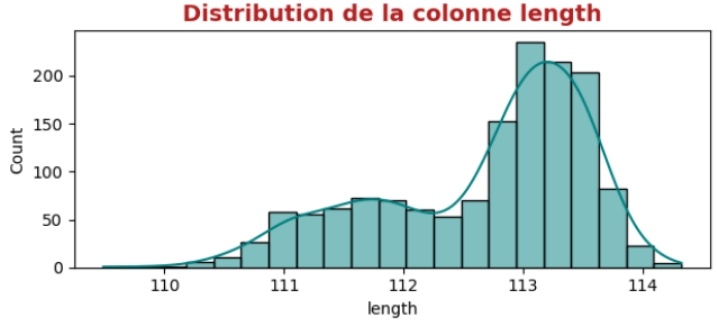
Il manque 37 valeurs pour ‘margin_low’.

Statistiques descriptives du dataframe

	Dtype	Non-Null Count	Valeurs unique	moyennes	medianes	ecart_types	min	max	Valeurs manquantes	Valeurs manquantes%
is_genuine	bool	1500	2	0.667	1.00	0.472	False	True	0	0.00
diagonal	float64	1500	159	171.958	171.96	0.305	171.04	173.01	0	0.00
height_left	float64	1500	155	104.030	104.04	0.299	103.14	104.88	0	0.00
height_right	float64	1500	170	103.920	103.92	0.326	102.82	104.95	0	0.00
margin_low	float64	1463	285	4.486	4.31	0.664	2.98	6.9	37	2.47
margin_up	float64	1500	123	3.151	3.14	0.232	2.27	3.91	0	0.00
length	float64	1500	336	112.678	112.96	0.873	109.49	114.44	0	0.00

Analyses univariés :

Exemple avec la colonne ‘lenght’.



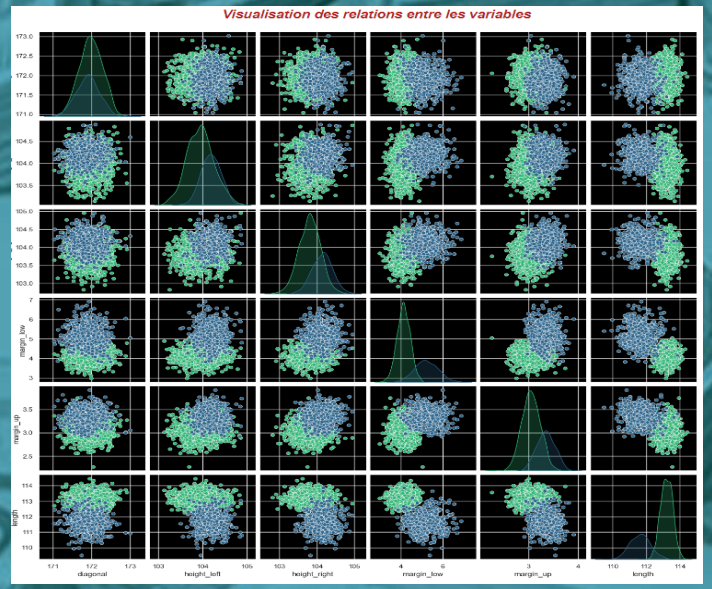
length : 3 Outliers

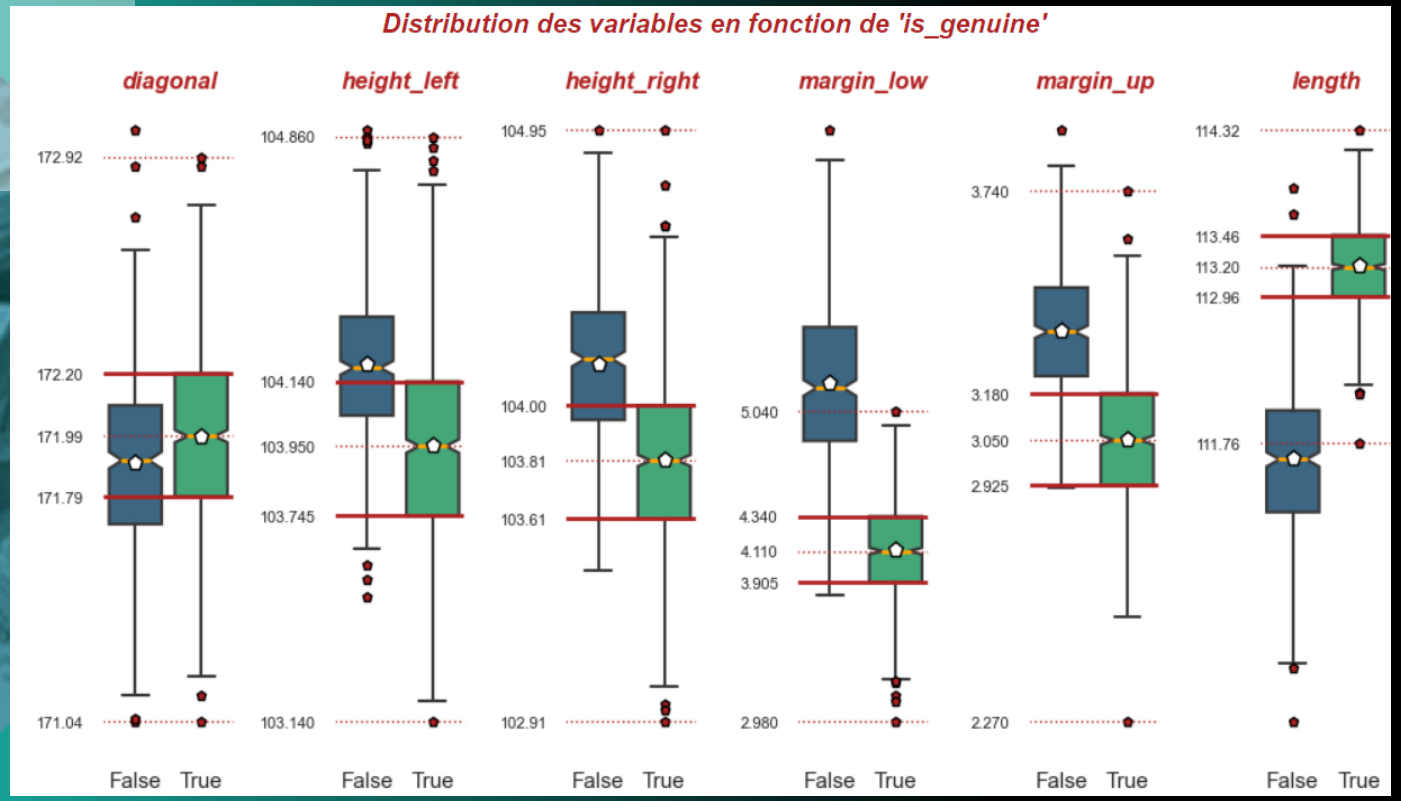
Analyse préparatoire du fichier

Analyses bivariés :

- sns.pairplot(DF_F, hue="is_genuine", palette="viridis")
- Boxplot en fonction de ‘is_genuine’.

On peut déjà distinguer les variables importantes différenciant un billet faux d’un vrai.

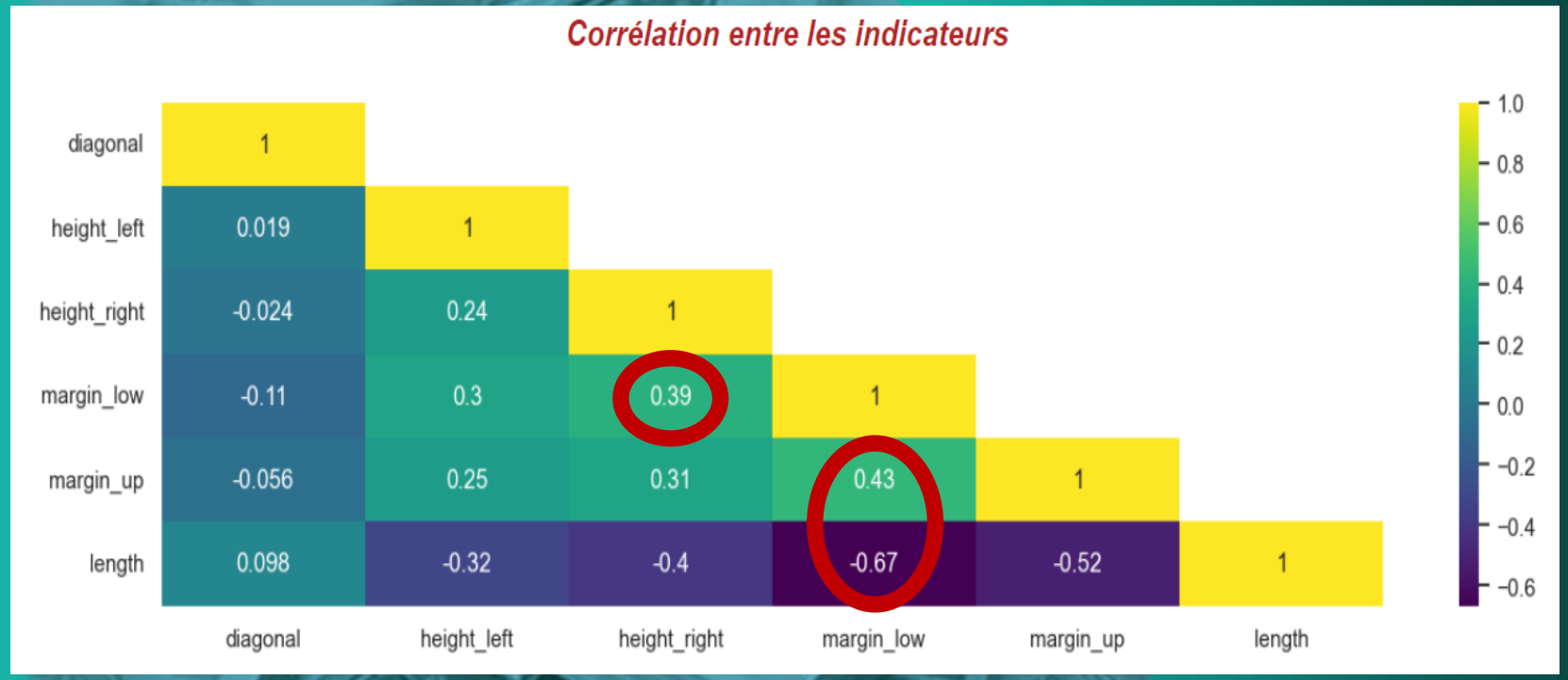




Pour vérifier 'is_genuine'
'margin_low', 'margin_up' et 'length' sont les plus prédictives

Analyse préparatoire du fichier

Table des corrélations :



Prédire ‘margin_low’:

Les variables les plus prédictives sont : ‘height_right’ - ‘margin_up’ - ‘length’

Régression linéaire pour remplacer les Nan dans 'margin_low'

ONCFM

1 - Optimisation du Random_state :

- Vérification du seed optimal avec toutes les variables prédictives :
random_seeds = [n for n in range(0,200,1)]

- les valeurs de performances sont optimisées avec un seed à 61

On fixe random_state à 61
Cette valeur n'est valable que dans le cadre de notre dataset

2 - Choix du pré-traitement :

- Réalisation : Scaler + modèle

- Aucun pré-traitement ne sort du lot

On conserve le DF sans traitement

margin_low ~ diagonal + height_left + height_right + margin_up + length

Normalizer()

R^2 : 0.476 --- RMSE: 0.2513 --- MAE : 0.3863 --- MAPE: 0.0849

PowerTransformer()

R^2 : 0.46 --- RMSE: 0.2656 --- MAE : 0.4022 --- MAPE: 0.0886

StandardScaler()

R^2 : 0.478 --- RMSE: 0.252 --- MAE : 0.3873 --- MAPE: 0.0851

MinMaxScaler()

R^2 : 0.478 --- RMSE: 0.252 --- MAE : 0.3873 --- MAPE: 0.0851

Transformation logarithmique

R^2 : 0.478 --- RMSE: 0.2518 --- MAE : 0.387 --- MAPE: 0.085

Sans traitement

R^2 : 0.478 --- RMSE: 0.252 --- MAE : 0.3873 --- MAPE: 0.0851

Régression linéaire pour remplacer les Nan dans 'margin_low'

ONCFM

3- Variables prédictives à utiliser :

- Valeur cible : ‘margin_low’

margin_low ~ 'toutes variables' a le MAPE le plus bas
Nous avons vu que 'diagonal' et 'height_left' ne sont que peu prédictives
----
margin_low ~ height_right + margin_up + length est validé avec MAPE = 7.115 %

Résultats avec Statsmodels :

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.464
Model:	OLS	Adj. R-squared:	0.463
Method:	Least Squares	F-statistic:	337.1
Date:	Mon, 22 Jan 2024	Prob (F-statistic):	1.38e-157
Time:	13:01:35	Log-Likelihood:	-828.97
No. Observations:	1170	AIC:	1666.
Df Residuals:	1166	BIC:	1686.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	22.8377	6.157	3.709	0.000	10.758	34.917
height_right	0.2718	0.049	5.590	0.000	0.176	0.367
margin_up	0.3075	0.074	4.155	0.000	0.162	0.453
length	-0.4222	0.020	-20.999	0.000	-0.462	-0.383

Omnibus:	49.349	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59.018
Skew:	0.454	Prob(JB):	1.53e-13
Kurtosis:	3.622	Cond. No.	6.56e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.56e+04. This might indicate that there are strong multicollinearity or other numerical problems.

RMSE: 0.4512

MAE: 0.3293

MAPE: 7.1151 %

Résultats avec Sklearn :

----------Variables prédictives pour 'margin_low' ----------
margin_low ~ diagonal + height_left + height_right + margin_up + length
R^2 : 0.471 --- RMSE: 0.1981 --- MAE : 0.3251 --- MAPE: 7.03 %

----------Variables prédictives pour 'margin_low' ----------
margin_low ~ height_left + height_right + margin_up + length
R^2 : 0.469 --- RMSE: 0.1998 --- MAE : 0.3257 --- MAPE: 7.037 %

----------Variables prédictives pour 'margin_low' ----------
margin_low ~ height_right + margin_up + length
R^2 : 0.464 --- RMSE: 0.2036 --- MAE : 0.3293 --- MAPE: 7.115 %

----------Variables prédictives pour 'margin_low' ----------
margin_low ~ margin_up + length
R^2 : 0.45 --- RMSE: 0.2101 --- MAE : 0.3355 --- MAPE: 7.271 %

----------Variables prédictives pour 'margin_low' ----------
margin_low ~ length
R^2 : 0.44 --- RMSE: 0.2123 --- MAE : 0.3435 --- MAPE: 7.466 %

Métriques disponibles dans le notebook

Régression linéaire pour remplacer les Nan dans 'margin_low'

ONCFM

Modèle retenu :

----------Variables prédictives pour 'margin_low' ----------

margin_low ~ height_right + margin_up + length

R^2 : 0.464 --- RMSE: 0.2036 --- MAE : 0.3293 --- MAPE: 7.115 %

4 - Comparaison des modèles avec et sans Intercept:

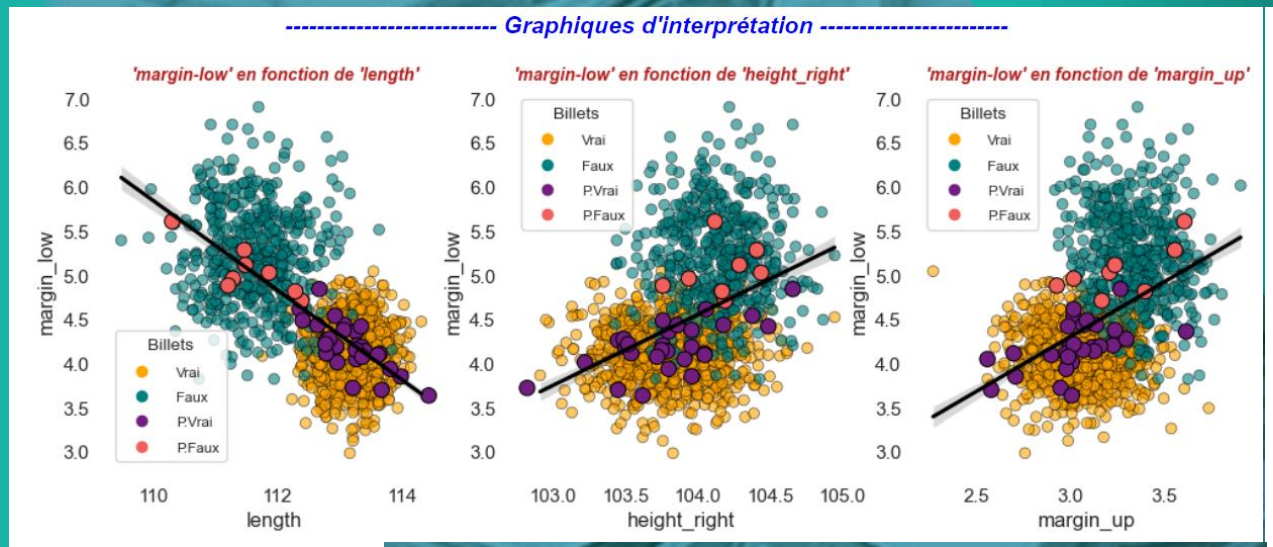
- Le coefficient de détermination R^2 est extrêmement différent suivant le modèle.
- Les autres mesures sont quasi identiques.
- Comparons les 2 modèles graphiquement.

Résultats avec intercept	*Résultats sans intercept*
R-squared (R^2): 0.464	*R-squared (R^2): 0.98810*
MAE: 0.32930	*MAE: 0.33046*
RMSE: 0.45122	*RMSE: 0.45481*
MAPE: 7.11513%	*MAPE: 7.12140%*

Régression linéaire pour remplacer les Nan dans 'margin_low'

ONCFM

Modèle avec Intercept :



Résultats avec intercept

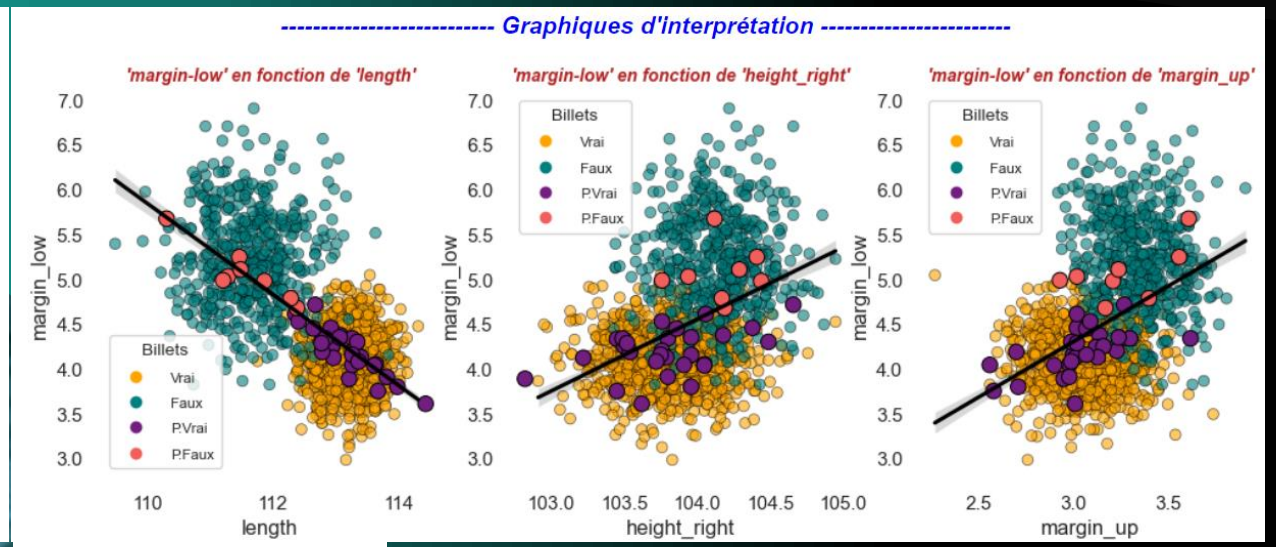
R-squared (R^2): 0.464

MAE: 0.32930

RMSE: 0.45122

MAPE: 7.11513%

Modèle sans Intercept :



Résultats sans intercept

R-squared (R^2): 0.98810

MAE: 0.33046

RMSE: 0.45481

MAPE: 7.12140%

R^2 à 0,988 indique que le modèle capte une grande partie de la variance des données.
Cependant, l'omission de l'intercept peut introduire des biais dans l'estimation des coefficients.

Vérifions via les métriques que ce modèle est robuste face au modèle avec intercept

Régression linéaire pour remplacer les Nan dans 'margin_low

ONCFM

AIC (Akaike Information Criterion) & BIC (Bayesian Information Criterion) :

Mesures utilisées en statistiques pour évaluer la qualité d'un modèle statistique en tenant compte de la trade-off entre la complexité du modèle et sa capacité à expliquer les données.

Avec Intercept:

AIC : 1665.931 / BIC : 1686.19

Sans Intercept:

AIC : 1677.658 / BIC : 1692.852

Les différences entre les modèles ne sont pas suffisantes pour déterminer le meilleur compromis entre ajustement et complexité

Vérifier la colinéarité des variables :

Pour les deux modèles, pas de problème de colinéarité

Facteur d'inflation de la variance avec Intercept

	Variables	VIF
0	height_right	1.207999
1	margin_up	1.396446
2	length	1.519821

Coefficients inférieurs à 10 (Avec Intercept et Sans Intercept(data scalées))

Facteur d'inflation de la variance sans Intercept

	Variables	VIF
0	height_right	17037.112090
1	margin_up	261.096721
2	length	14793.687996

On scale les données pour minimiser les différences de tailles des variables

Facteur d'inflation de la variance sans Intercept (data scalées)

	Variables	VIF
0	height_right	1.208128
1	margin_up	1.396593
2	length	1.519928

Régression linéaire pour remplacer les Nan dans 'margin_low

ONCFM

Tester l'homoscédasticité :

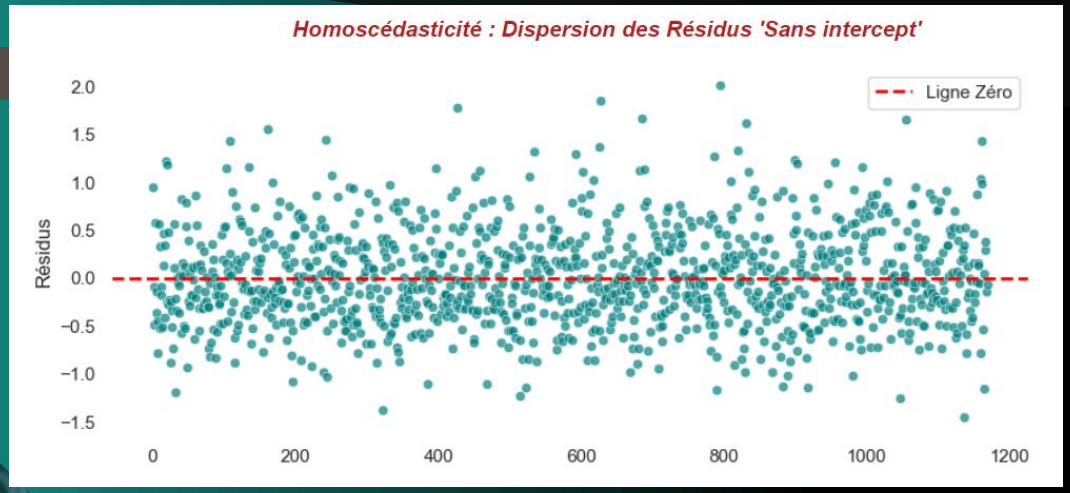
p-valeur < à 5% pour les deux modèles

L'hypothèse $H0$ est rejetée selon laquelle les variances sont constantes (hypothèse d'homoscédasticité)

Tests de Breusch Pagan

P-value sans intercept: 3.894809736584079e-14

P-value avec intercept: 8.873495179862817e-15



Tester la normalité des résidus :

p-valeur < à 5% pour les deux modèles, l'hypothèse $H0$ est rejetée.

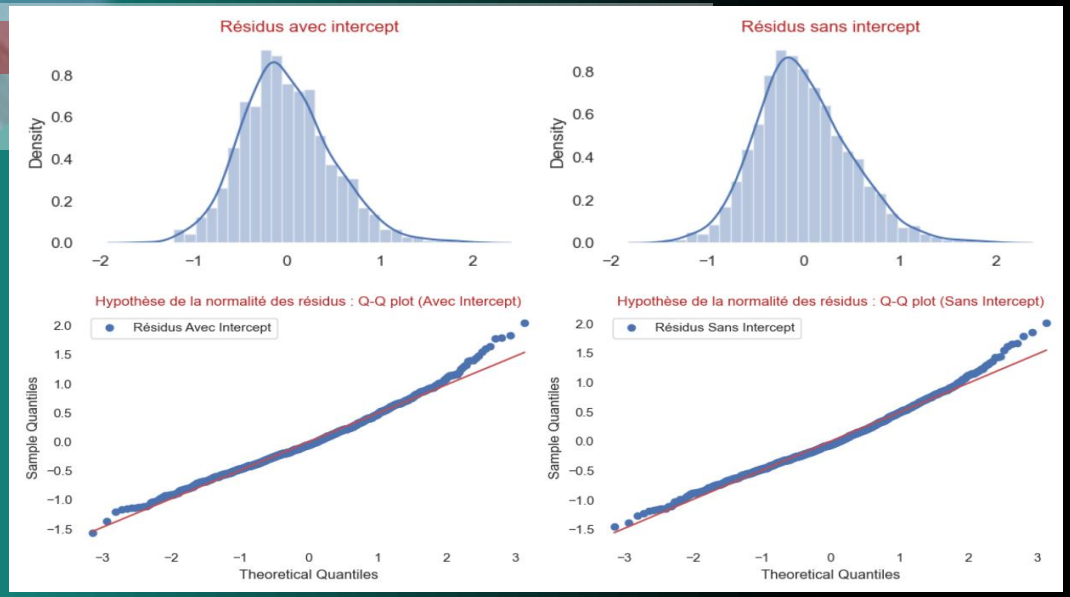
Normalité des résidus

Modèle avec intercept

```
ShapiroResult(statistic=0.9879472851753235, pvalue=3.12514139011455e-08)
```

Modèle sans intercept

```
ShapiroResult(statistic=0.9878649115562439, pvalue=2.842064006358669e-08)
```



Régression linéaire pour remplacer les Nan dans 'margin_low'

ONCFM

Modèle sans Intercept :

	height_right	margin_up	length	margin_low	margin_low_Predict	% erreur
1301	104.19	3.25	111.99	5.72	4.882714	14.64
1028	103.93	3.49	111.34	5.38	5.093796	5.32
281	104.21	3.07	113.01	4.18	4.447908	6.41
665	104.44	2.99	113.16	4.54	4.465495	1.64
555	104.10	2.70	113.99	4.47	3.906402	12.61

Le coefficient de détermination R^2 est de 0.988

Vérification erreur moyenne : 7.122 %

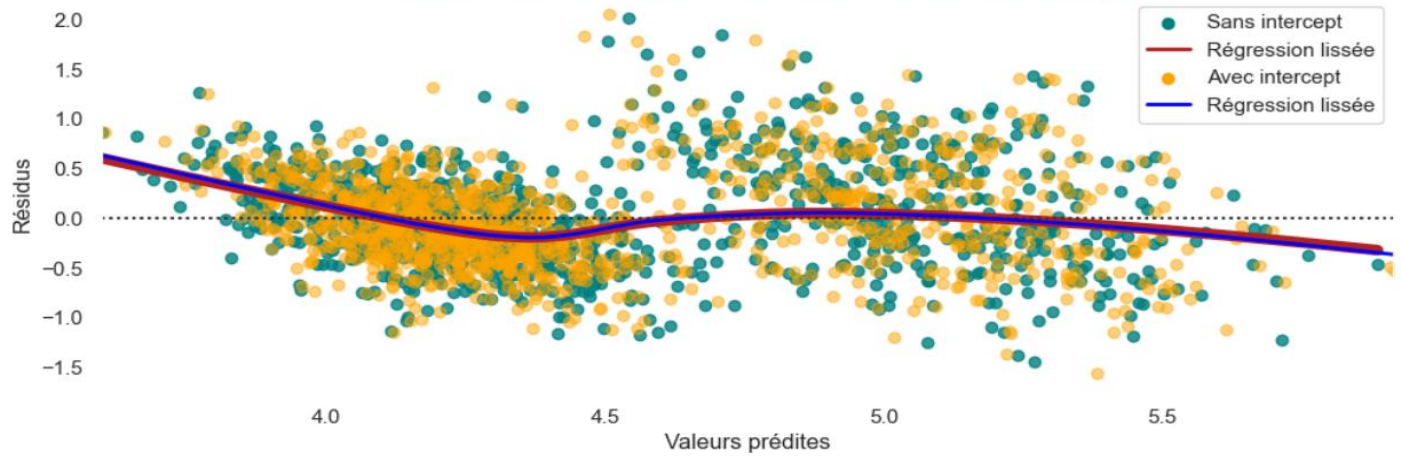
Modèle avec Intercept :

	height_right	margin_up	length	margin_low	margin_low_Predict	% erreur
1301	104.19	3.25	111.99	5.72	4.867881	14.90
1028	103.93	3.49	111.34	5.38	5.145473	4.36
281	104.21	3.07	113.01	4.18	4.387276	4.96
665	104.44	2.99	113.16	4.54	4.361852	3.92
555	104.10	2.70	113.99	4.47	3.829804	14.32

Le coefficient de détermination R^2 est de 0.464

Vérification erreur moyenne : 7.115 %

Graphique des résidus pour une régression linéaire multiple

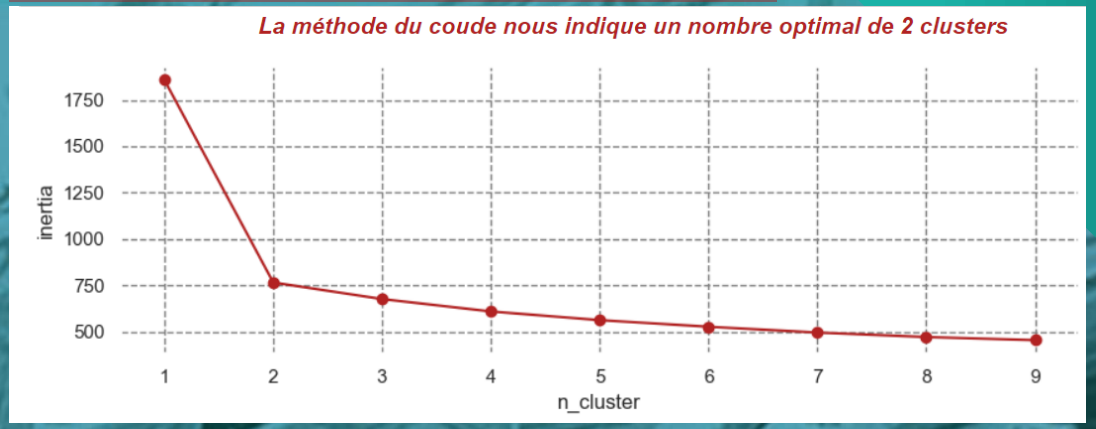


Pas de différences notables entre les deux modèles

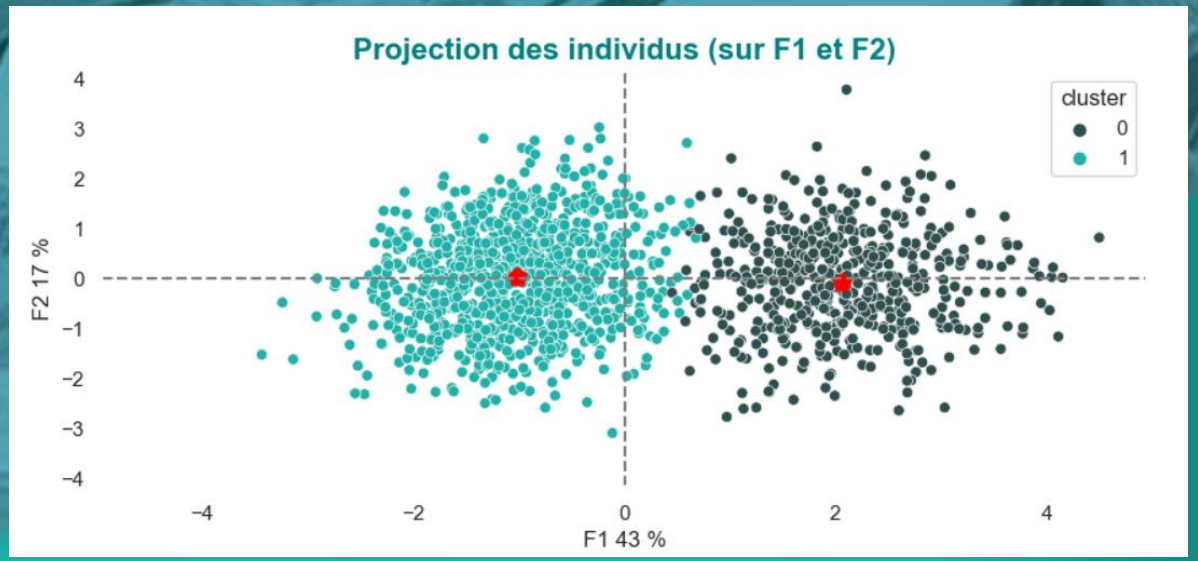
Avec une erreur moyenne équivalente et un R^2 à 0.988, le premier modèle est validé

Méthode non supervisée (K-means) – Méthode du Coude & PCA

Nombre de cluster :



Projection sur le premier plan factoriel :

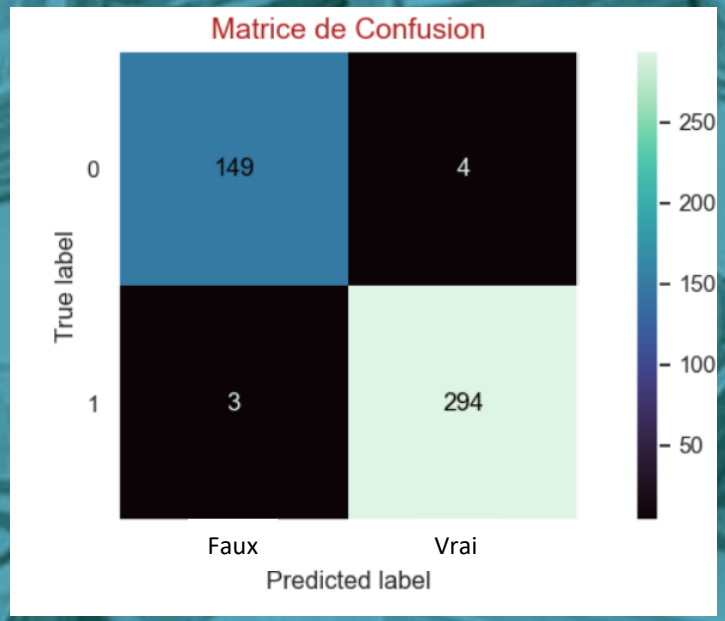


Méthode non supervisée (K-means)

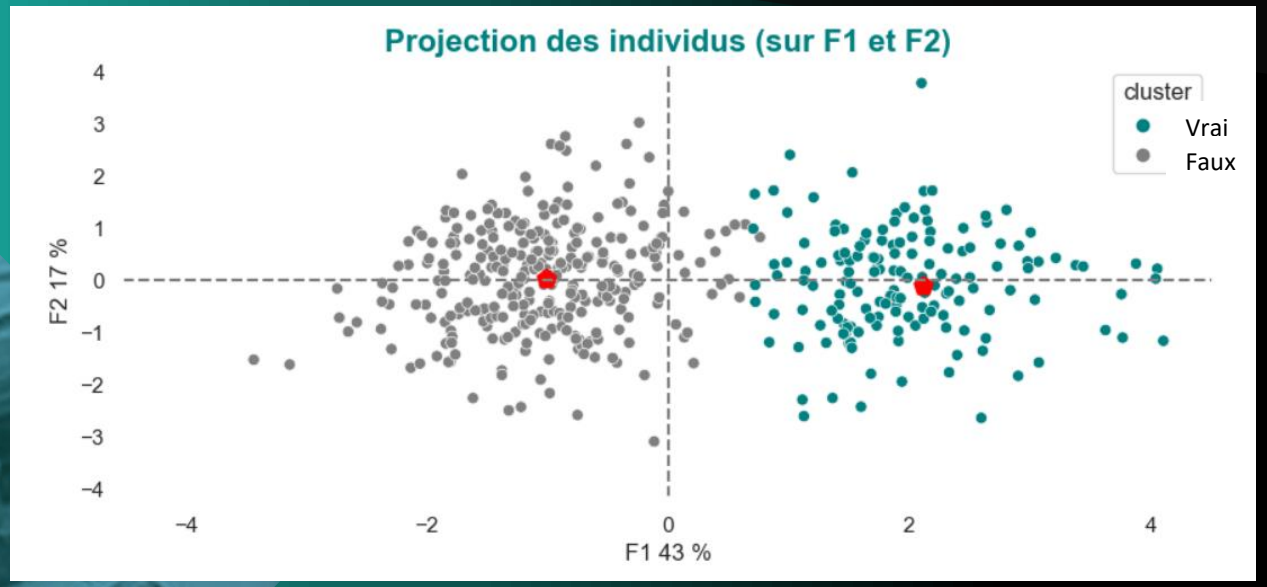
Résultats sur le set de test :

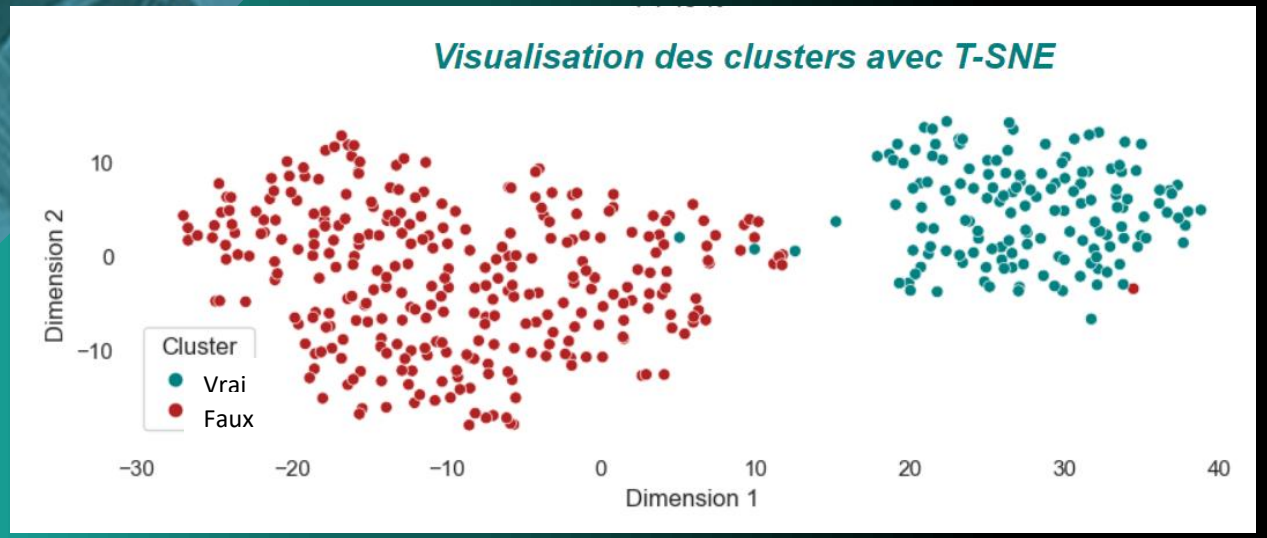
Affectation des clusters au set test (30 % du DataFrame)

	diagonal	height_left	height_right	margin_low	margin_up	length	cluster
1212	-0.191548	1.003691	1.473628	0.328799	0.684084	-0.640159	0
954	-0.617647	-1.334617	-1.014709	-1.461078	0.727236	0.104878	1
1207	-0.814308	0.302198	0.674903	0.541158	0.986152	-1.144493	0



F1-score sur l'ensemble de test: 98.4432 %





Méthode non supervisée (K-means)

Métriques d'évaluation de classifications non supervisées

Calcul du score de Silhouette :

- Score de silhouette = 0.3506
- Ce score indique une certaine cohésion intra-cluster et une séparation inter-cluster raisonnable.
- D'autres métriques sont nécessaires pour avoir une meilleure compréhension de la qualité du clustering

Score => entre -1(Mal placé) et 1 (Bien regroupé)

Résultats

Score de silhouette : 0.3506

Davies-Bouldin Index : 1.1811

Calinski-Harabasz Index : 255.6412

Adjusted Rand Index (ARI) : 0.938

Calcul du Davies-Bouldin Index :

- Davies-Bouldin Index : 1.1811
- Les clusters sont relativement bien séparés avec une certaine compacité à l'intérieur de chaque cluster.
- Avec un risque de superposition ou une dispersion, utilisons d'autres métriques pour évaluer le modèle.

Plus le DBI est bas, meilleure est la partition des données

Méthode non supervisée (K-means)

Métriques d'évaluation de classifications non supervisées

Calcul du Calinski-Harabasz Index :

- Calinski-Harabasz Index : 255.64

- Cette valeur, relativement élevée, indique une bonne séparation entre les clusters mais n'est pas assez élevée pour suggérer une séparation des clusters et une compacité intra-cluster parfaitement définis.

Plus le score CH est élevé, meilleure est la séparation entre les clusters et une plus grande homogénéité intra-cluster

Calcul de l'ARI (Adjusted Rand Index) :

- Adjusted Rand Index (ARI) : 0.938

- L'ARI suggère une forte similarité entre les clusters prédits et les clusters réels.

- La performance de l'algorithme de clustering non supervisée (K-means) est très bonne pour détecter des faux billets.

Un ARI à 0.938 indique une correspondance très forte entre les deux ensembles d'étiquettes.
Le CHI indique une bonne séparation entre les clusters dans l'algorithme de clustering

Pour ce projet, il est intéressant d'obtenir de bons résultats avec le K-mean (Algorithme non supervisé)

Régression logistique

Variables prédictives après vérifications : ‘margin_low’, ‘margin_up,’ ‘height_right’ et ‘length

Régression logistique avec Statsmodels

Logit Regression Results

Dep. Variable:	is_genuine	**No. Observations:**	1200
Model:	Logit	**Df Residuals:**	1195
Method:	MLE	**Df Model:**	4
Date:	Mon, 22 Jan 2024	**Pseudo R-squ.:**	0.9492
Time:	13:54:33	**Log-Likelihood:**	-38.843
converged:	True	**LL-Null:**	-765.20
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-283.1799	157.034	-1.803	0.071	-590.961	24.601
height_right	-2.6311	1.226	-2.147	0.032	-5.033	-0.229
margin_low	-5.7816	0.979	-5.906	0.000	-7.700	-3.863
margin_up	-9.4716	2.089	-4.533	0.000	-13.567	-5.376
length	5.4540	0.799	6.829	0.000	3.889	7.019

Accuracy : 0.9933

Précision : 0.9902

Recall : 1.0

Le score f1 du modèle sur les données Test est de : 99.5074%

Fonction backward_selected_log

```
_____________________________________
is_genuine ~ height_left + height_right + length + margin_low + diagonal + margin_up + 1
Optimization terminated successfully.
         Current function value: 0.028632
         Iterations 13
remove diagonal (p-value: 0.922 )


_____________________________________
is_genuine ~ height_left + height_right + length + margin_low + margin_up + 1
Optimization terminated successfully.
         Current function value: 0.028635
         Iterations 12
remove height_left (p-value: 0.105 )


_____________________________________
is_genuine ~ height_right + length + margin_low + margin_up + 1
Optimization terminated successfully.
         Current function value: 0.029539
         Iterations 12
is the final model!
```

Performance du modèle LogisticRegression() de Sklearn

Accuracy : 0.9933

Précision : 0.9902

Recall : 1.0

Le score f1 du modèle sur les données Test est de : 99.5074%

Le coefficient de détermination R^2 est : 0.9697

Régression logistique

Variables prédictives après vérifications : ‘margin_low’, ‘margin_up,’ ‘height_right’ et ‘length

Régression logistique avec Statsmodels :

Billets considérés comme Vrais: 204 - Billets considérés comme Faux : 96

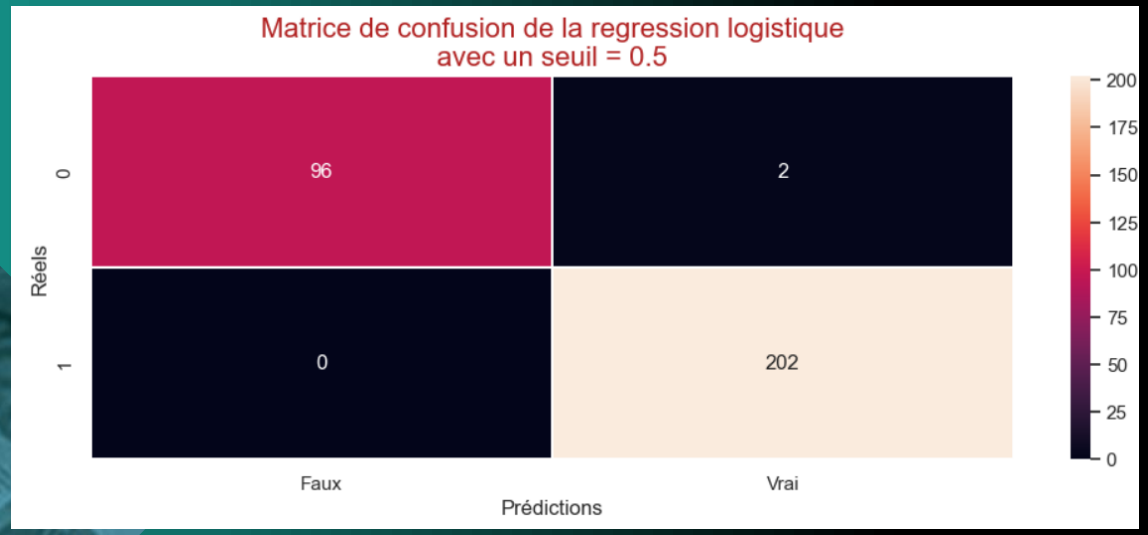
Après avoir croiser les données réelles de celles prédites, on observe :

Faux billets (classe 0): il existe 96 faux billets et 2 faux billets identifiés comme vrais

Vrais billets (classe 1): il existe 202 vrais billets et 0 vrais billets identifiés comme faux

Métriques d’évaluation :

Accuracy: 0.9933	*Recall: 1.0*
Précision: 0.9902	*F1_score: 99.5074 %*



Régression logistique avec Sklearn :

Billets considérés comme Vrais: 204 - Billets considérés comme Faux : 96

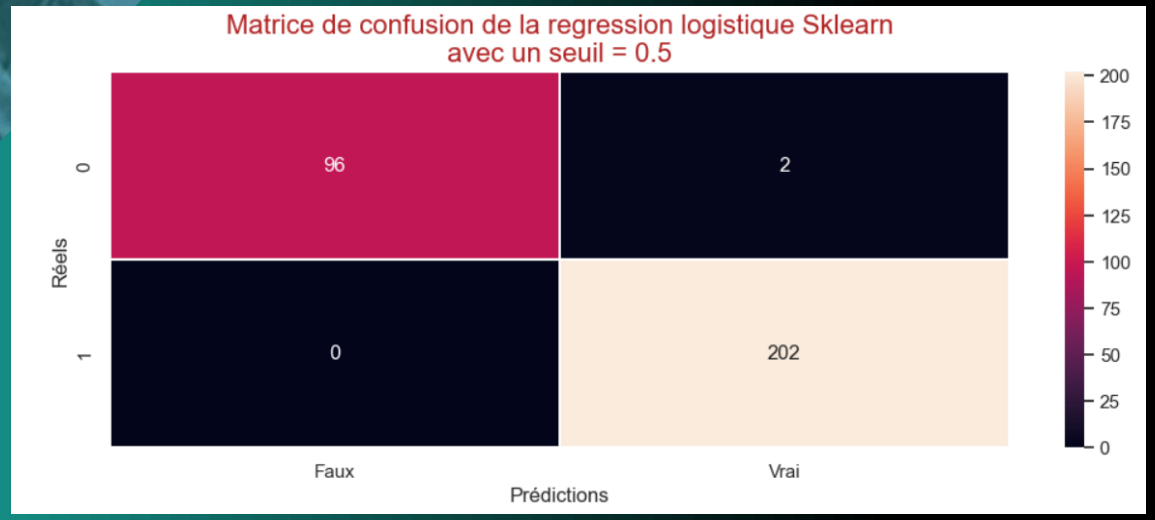
Après avoir croiser les données réelles de celles prédites, on observe :

Faux billets (classe 0): il existe 96 faux billets et 2 faux billets identifiés comme vrais

Vrais billets (classe 1): il existe 202 vrais billets et 0 vrais billets identifiés comme faux

Métriques d’évaluation :

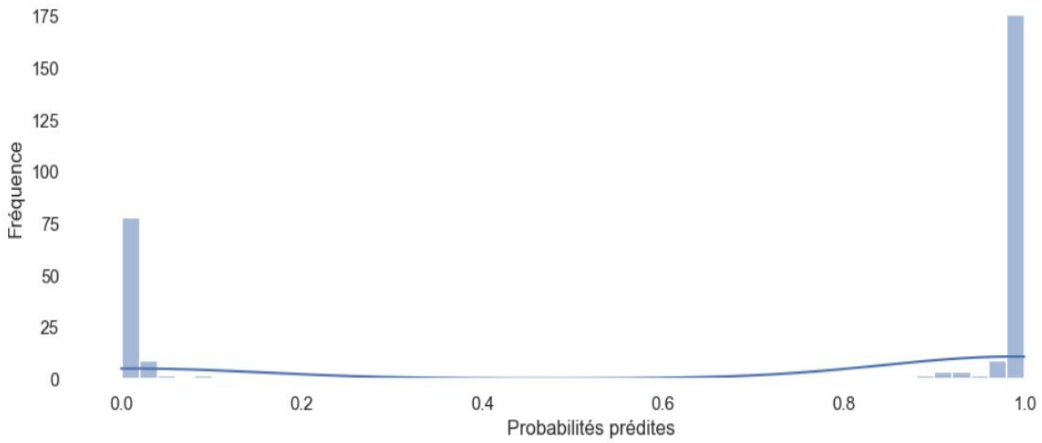
L'accuracy : 0.9933	*Recall : 1.0*
Précision : 0.9902	*F1_score : 99.5074 %*



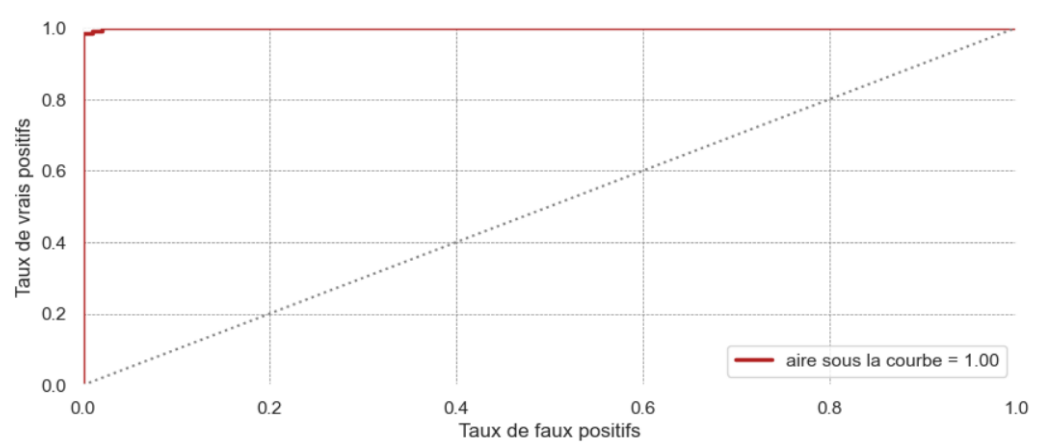
Régression logistique

Régression logistique avec Sklearn

Probabilités des prédictions :

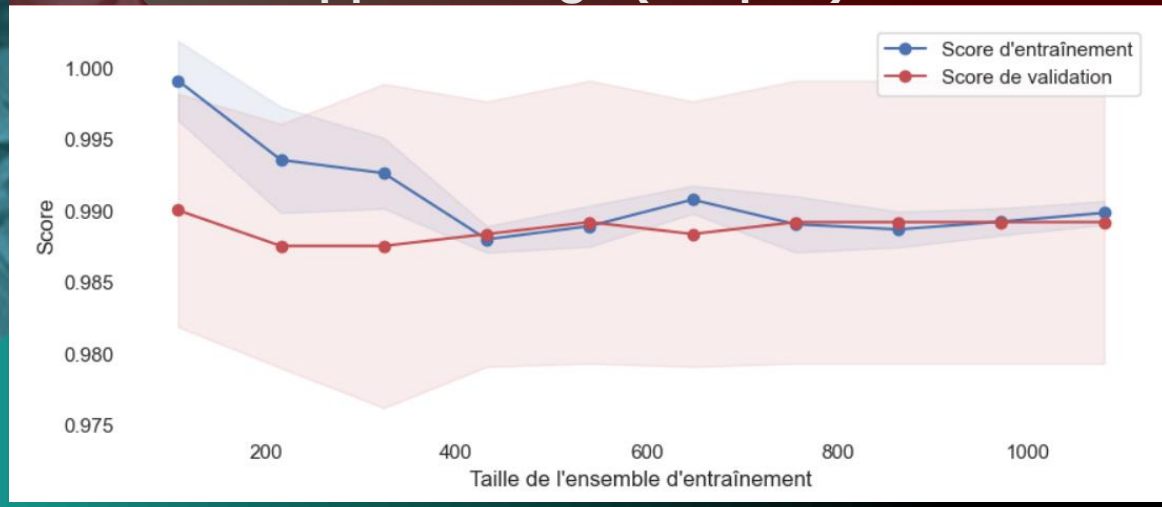


Courbe R.O.C:



Capacité du modèle à discriminer entre les classes - ROC-AUC: 0.999747

Courbes d'apprentissage (10 splits) :



Choix de l'algorithme

ONCFM

Algorithme de régression logistique

F1_score sur le set Test: 99.5074 %

Nombre de prédictions correctes sur le nombre total d'échantillons - Accuracy : 0.9933

Précision des prédictions positives - Precision: 0.9902

Capacité du modèle à capturer tous les exemples positifs - Recall: 1.0

```
LogisticRegression
LogisticRegression()
```

Algorithme K_means

F1-score sur l'ensemble de test: 98.4432 %

Précision sur l'ensemble de test: 0.9844

Rappel sur l'ensemble de test: 0.9844

```
KMeans
KMeans(n_clusters=2, n_init='auto')
```

Avec un meilleur score, la régression logistique sera ulisée pour identifier les billets

Résultats avec la fonction "detection_faux_billets()"

ONCFM

Outil de détection de faux billets

Aperçue du fichier test

	diagonal	height_left	height_right	margin_low	margin_up	length	id
0	172.09	103.95	103.73	4.39	3.09	113.19	B_1
1	171.52	104.17	104.03	5.27	3.16	111.82	B_2
2	171.78	103.80	103.75	3.81	3.24	113.39	B_3
3	172.02	104.08	103.99	5.57	3.30	111.10	B_4
4	171.79	104.34	104.37	5.00	3.07	111.87	B_5

On ne conserve que les colonnes prédictives

	diagonal	height_left	height_right	margin_low	margin_up	length
0	172.09	103.95	103.73	4.39	3.09	113.19
1	171.52	104.17	104.03	5.27	3.16	111.82
2	171.78	103.80	103.75	3.81	3.24	113.39

Le fichier test contient 5 billets à vérifier

Suppression des lignes contenant des valeurs non autorisées

le fichiers ne contient aucune valeur nulle

le fichiers ne contient aucune valeur manquante

le fichiers ne contient aucune valeur texte

Vérification présence de dimensions abérrantes

- 'diagonal' est Ok

- 'height_left' est Ok

- 'height_right' est Ok

- 'margin_low' est Ok

- 'margin_up' est Ok

- 'length' est Ok

Toutes les lignes aberrantes ont été supprimées du dataset

Le dataset contient 5 lignes avant détection des vrais billets

---

Imputation de 'is_genuine' suivant le modèle

	is_genuine_Predict	% Probablement VRAI	diagonal	height_left	height_right	margin_low	margin_up	length	id
0	1	99.930000	172.090000	103.950000	103.730000	4.390000	3.090000	113.190000	B_1
1	0	0.060000	171.520000	104.170000	104.030000	5.270000	3.160000	111.820000	B_2
2	1	100.000000	171.780000	103.800000	103.750000	3.810000	3.240000	113.390000	B_3
3	0	0.000000	172.020000	104.080000	103.990000	5.570000	3.300000	111.100000	B_4
4	0	0.150000	171.790000	104.340000	104.370000	5.000000	3.070000	111.870000	B_5

Contenu du fichier test : 2 Vrais billets et 3 Faux billets

Lignes du fichier mises à l'écart

Aucune ligne n'a été mise de coté

Outil de définition du seuil d'acceptation