ONCFM

Détection automatique de faux billets

Cruder Lea – Data Analyst

ONCFM

Table des matières

ONCFM

01

Contexte

Rappel du contexte et
de la mission

Organisation nationale de lutte contre le faux-monnayage (ONCFM)

ONCFM

- Organisation publique
- Objectif : mettre en place des méthodes d'identification des contrefaçons des billets en euros.
- Mission : mettre en place un algorithme qui soit capable de différencier automatiquement les vrais des faux billets à partir des dimensions géométriques

ONCFM

02

Méthodologie

Description des étapes de conception

ONCFM

Méthodologie suivie

1

Analyse de la mission en détails et des attendus

2

Analyse exploratoire

Prédiction des valeurs manquantes

3

Analyse en Composantes Principales

Conception des algorithmes

4

Optimisation de l'algorithme retenu

Test final

1

ONCFM

Préparation

Analyse

Analyse du cahier des charges et de la mission en détails

Planification

Récapitulatif et planification des attendus

Données

Importation des données

Vérification de l'intégrité

Exploration

Première analyse exploratoire des données afin de mieux les comprendre

1

Description des données

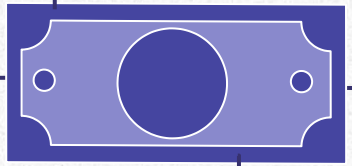
ONCFM

is_genuine
False / True

diagonal
Diagonale du billet en mm

height_left
Hauteur du billet à gauche en mm

margin_up
Hauteur entre bord supérieur et image du billet en mm



height_right
Hauteur du billet à gauche en mm

margin_low
Hauteur entre bord inférieur et image du billet en mm

length
Longueur du billet en mm

x 1500 billets

1

ONCFM

Description des données

Vrais billets

- 1000 exemplaires
- 29 valeurs manquantes sur margin_low

Faux billets

- 500 exemplaires
- 8 valeurs manquantes sur margin_low

2

Analyse exploratoire

ONCFM

Analyses univariées

Etude de chaque variable dans son ensemble et par type de billet

Analyses bivariées

Recherche et étude des liens potentiels entre les variables

Analyses multivariées

Recherche de multi colinéarité et de variables constantes

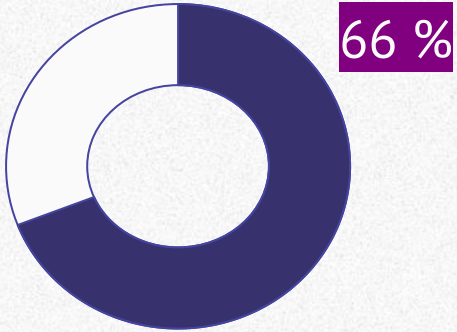
Prédiction des valeurs manquantes

Comparaison et choix du meilleur modèle pour prédire les valeurs manquantes

2

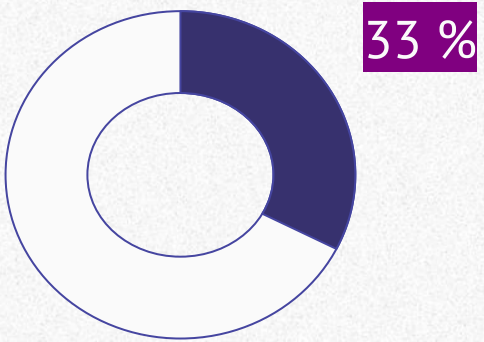
Variable : is_genuine

ONCFM



Vrais billets

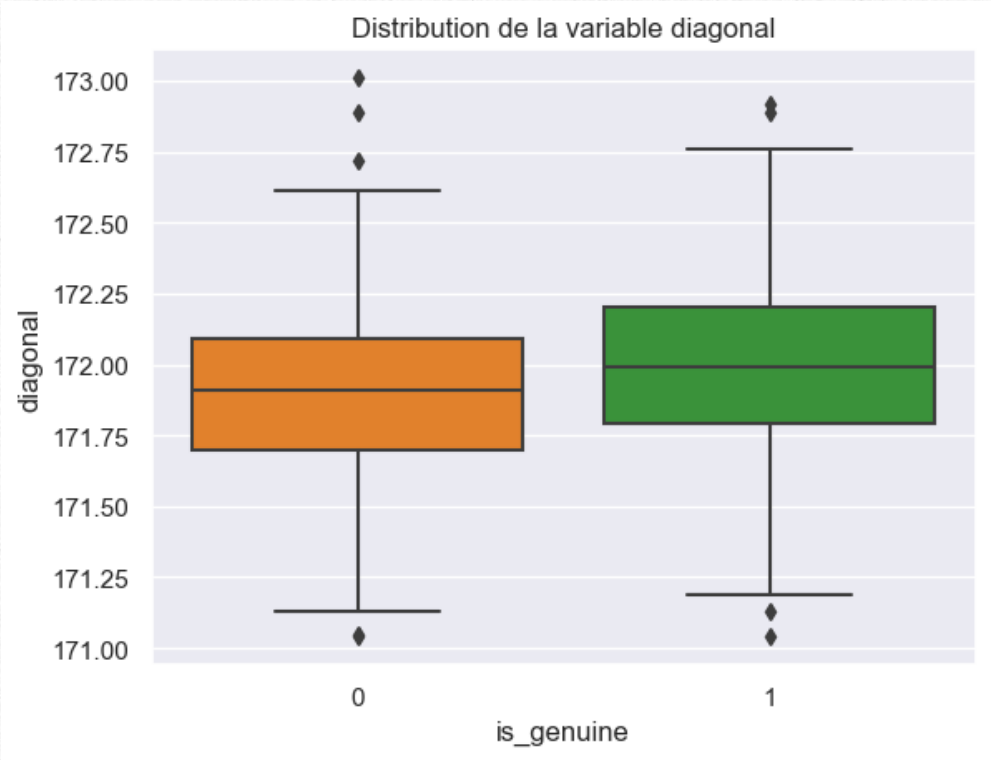
Remplacement par 1



Faux billets

Remplacement par 0

Variable : diagonal



- **Distribution normale**
- **9 outliers potentiels**

Vrais billets

Moyenne : 171,91 mm
Médiane : 171,99 mm

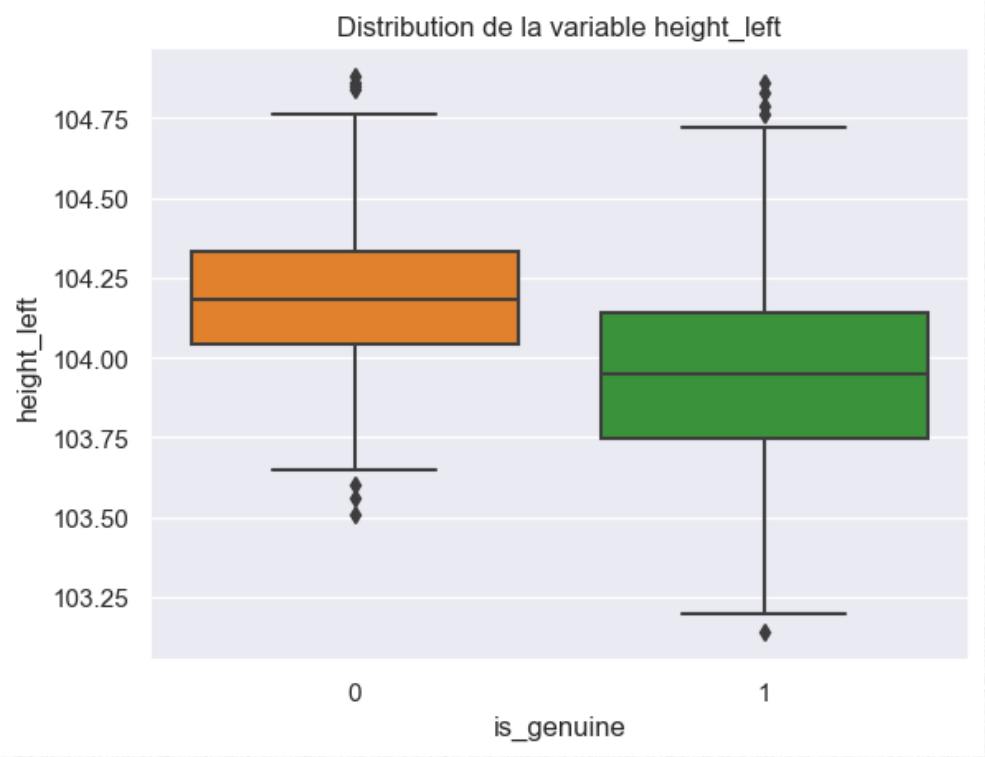
Faux billets

Moyenne : 171,99 mm
Médiane : 171,90 mm

2

Variable : height_left

ONCFM



- **Distribution normale**
- **12 outliers potentiels**

Vrais billets

Moyenne : 103,95 mm
Médiane : 103,95 mm

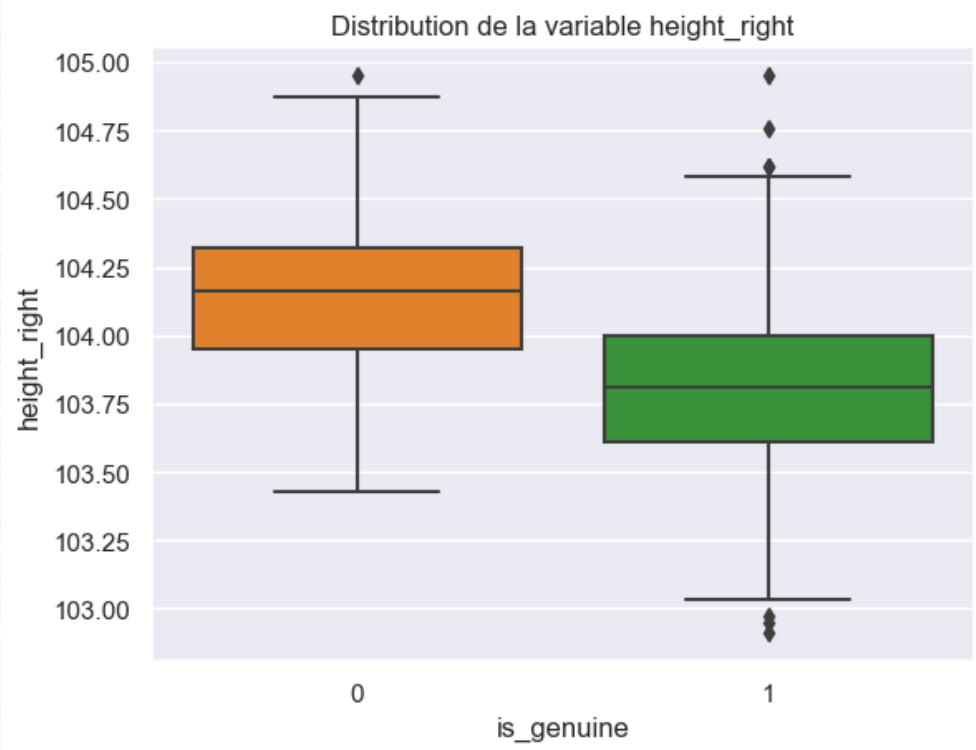
Faux billets

Moyenne : 104,18 mm
Médiane : 104,18 mm

2

Variable : height_right

ONCFM



- **Distribution normale**
- **7 outliers potentiels**

Vrais billets

Moyenne : 103,81 mm
Médiane : 103,81 mm

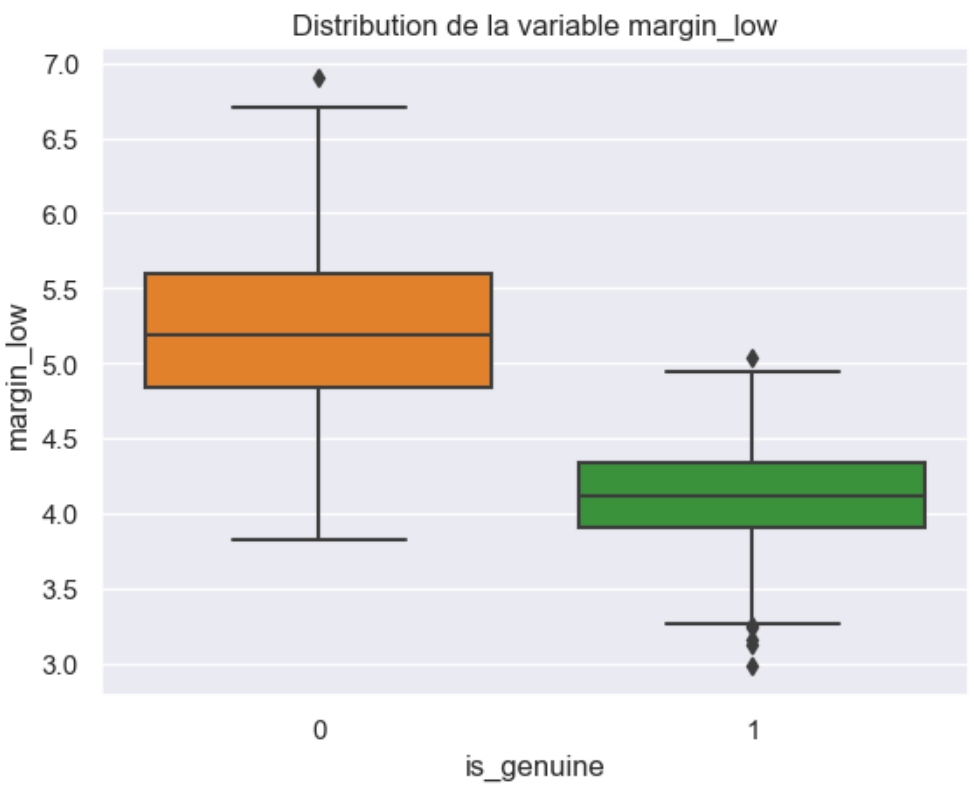
Faux billets

Moyenne : 104,14 mm
Médiane : 104,16 mm

2

Variable : margin_low

ONCFM



- **Défaut de normalité**
- **7 outliers potentiels**

Vrais billets

Moyenne : 4,11 mm
Médiane : 4,11 mm

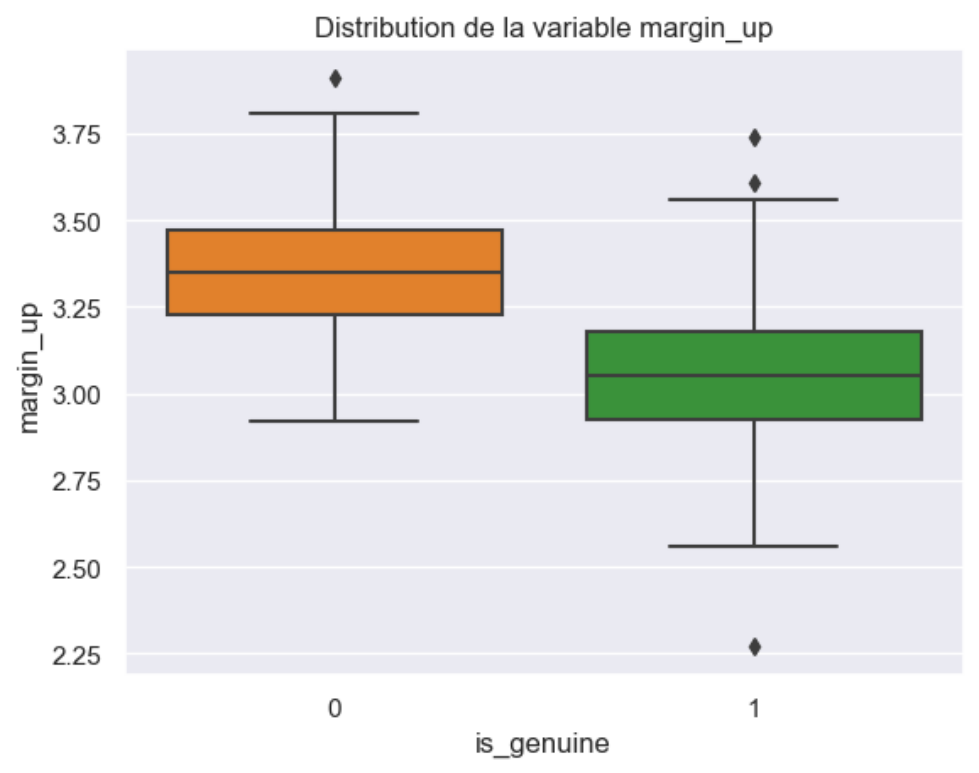
Faux billets

Moyenne : 5,21 mm
Médiane : 5,19 mm

2

Variable : margin_up

ONCFM



- **Défaut de normalité**
- **4 outliers potentiels**

Vrais billets

Moyenne : 3,05 mm
Médiane : 3,05 mm

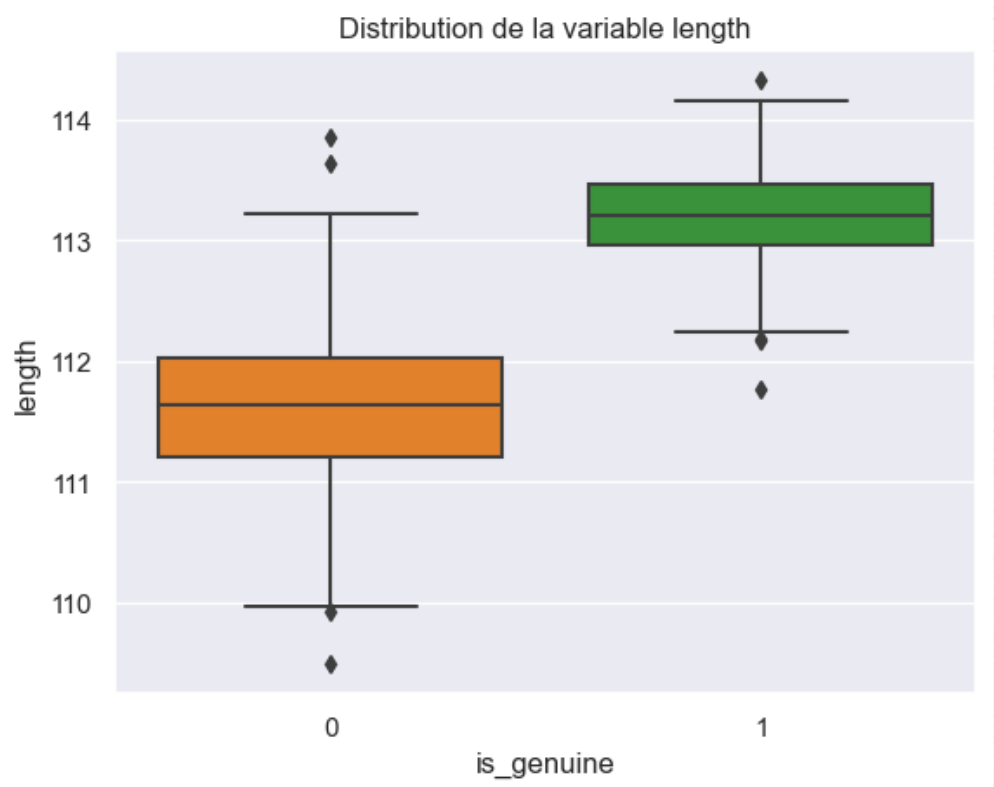
Faux billets

Moyenne : 3,35 mm
Médiane : 3,35 mm

2

Variable : length

ONCFM



- **Défaut de normalité**
- **8 outliers potentiels**

Vrais billets

Moyenne : 113,20 mm
Médiane : 113,20 mm

Faux billets

Moyenne : 111,63 mm
Médiane : 111,63 mm

Analyses bivariées

Après analyse visuelle, étude des liens potentiels et confirmation notamment :

is_genuine

- Lien confirmé avec **margin_low** : Stat H de Kruskal-Wallis = 828,74 / p-value <0,01
- Lien confirmé avec **length** : Stat H de Kruskal-Wallis = 916,9 / p-value <0,01
- Lien confirmé avec **height_right** : Stat H de Kruskal-Wallis = 355,69 / p-value <0,01
- Lien confirmé avec **height_left** : Stat H de Kruskal-Wallis = 221,40 / p-value <0,01

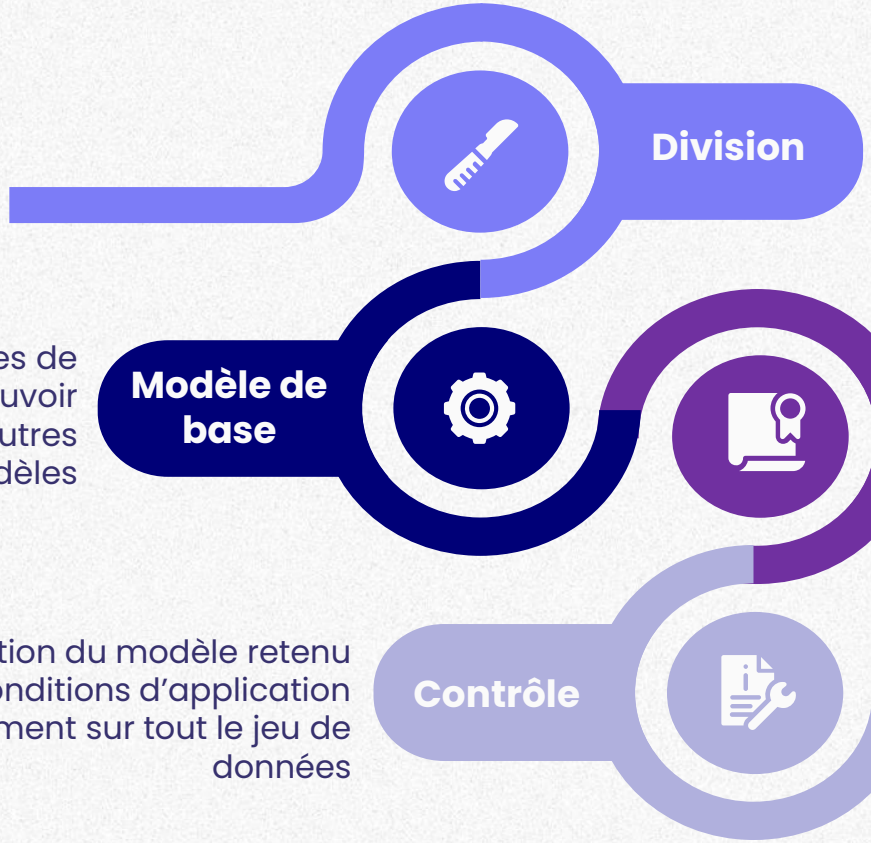
margin_low

- Corrélation confirmée avec **length** : Coeff. de Spearman = -0,59 / p-value <0,01
- Corrélation confirmée avec **height_right** : Coeff. de Spearman = 0,39 / p-value <0,01
- Corrélation confirmée avec **margin_up** : Coeff. de Spearman = 0,42 / p-value <0,01
- Corrélation confirmée avec **height_left** : Coeff. de Spearman = 0,29 / p-value <0,01

2

Conception d'un modèle

ONCFM



2

Choix du modèle final

ONCFM

Modèle de base

Dummy_median

Remplacement par la médiane (distribution non normale)

Scores obtenus :

Métrique	Train	Test
R2	-0.068643	-0.077451
MSE	0.475608	0.454152
RMSE	0.689643	0.673908

Régression linéaire simple

Variable prédictive : length

	Coef.	Std.Err.	t	P>|t|
const	60.705029	1.901646	31.922361	2.780726e-161
length	-0.498963	0.016877	-29.564881	7.173831e-144

Scores obtenus :

Métrique	Train	Test
R2	0.428035	0.513607
MSE	0.254557	0.205017
RMSE	0.504537	0.452788

Régression linéaire multiple

Variables prédictives :

	Coef.	Std.Err.	t	P>|t|
height_left	0.175760	0.039037	4.502434	7.391678e-06
height_right	0.283590	0.037374	7.587922	6.626067e-14
margin_up	0.294641	0.073321	4.018497	6.232521e-05
length	-0.392292	0.016678	-23.521624	1.968097e-100

Scores obtenus :

Métrique	Train	Test
R2	0.458700	0.539374
R2_adjusted	0.988241	
MSE	0.240909	0.194156
RMSE	0.490825	0.440632

2

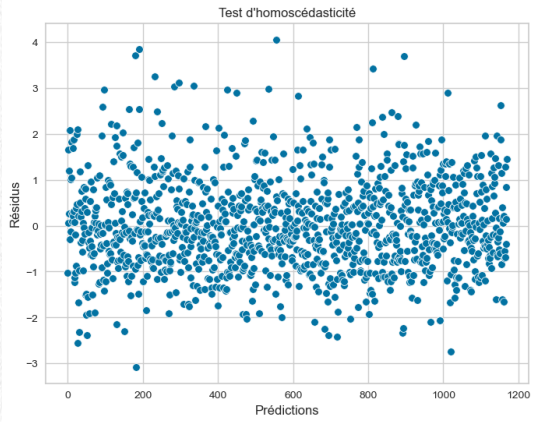
ONCFM

Conditions d'application

Linéarité

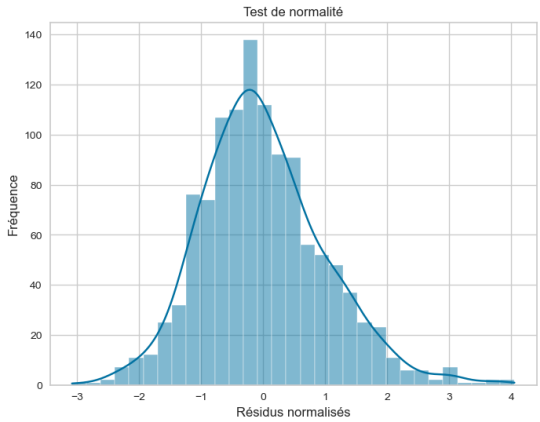
Les relations suivent une forme linéaire

Homoscédasticité



Variance presque constante des résidus
(Test de Breusch-Pagan ne passe pas)

Normalité



Distribution presque normale mais rejetée par les tests (Shapiro-Wilk)

2

ONCFM

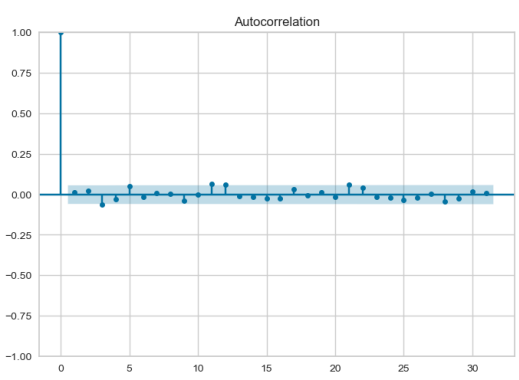
Conditions d'application

Absence de multicolinéarité

Variable	VIF
const	313283.808503
height_left	1.147526
height_right	1.259215
margin_low	1.903892
margin_up	1.419586
length	2.126807

Pas de VIF > 5 pour les variables sélectionnées
= pas de variables colinéaires

Absence d'autocorrélation



Les variables ne sont pas auto-corrélées
Confirmé par le test de Durbin-Watson

Significativité des coefficients

	Coef.	Std.Err.	t	P>|t|
height_left	0.175760	0.039037	4.502434	7.391678e-06
height_right	0.283590	0.037374	7.587922	6.626067e-14
margin_up	0.294641	0.073321	4.018497	6.232521e-05
length	-0.392292	0.016678	-23.521624	1.968097e-100

Les coefficients sont statistiquement différents de 0

Les p-values sont < 0,05

2

Prédiction et exportation

ONCFM

Préparation

Ré-entraînement du modèle final sur l'ensemble des données

Prédiction

Prédiction des valeurs manquantes sur margin_low

Vérification

Vérification de l'intégrité des données

Exportation

Exportation du fichier final

Exportation d'une version sans outliers pour tester l'impact

3

Analyse en Composantes Principales

ONCFM

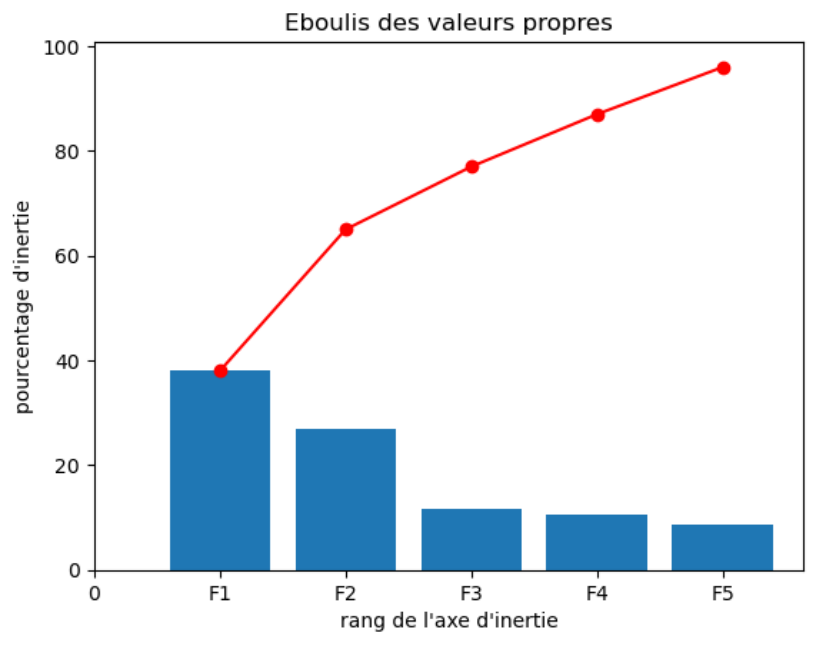
Calcul du nombre optimal de dimensions

- Graphique combinant le taux d'inertie et la variance expliquée

  But : en garder un maximum sans avoir trop de dimensions

- Nous permettra de visualiser nos clusters

- Choix de 3 axes : conserve 77% de la variance expliquée

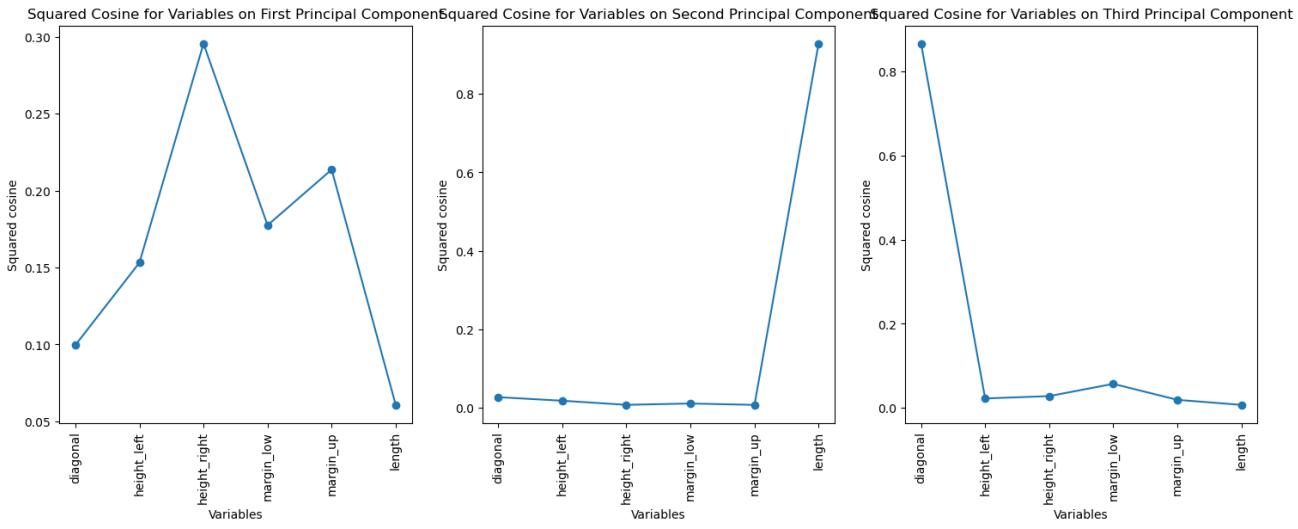


3

Analyse en Composantes Principales

ONCFM

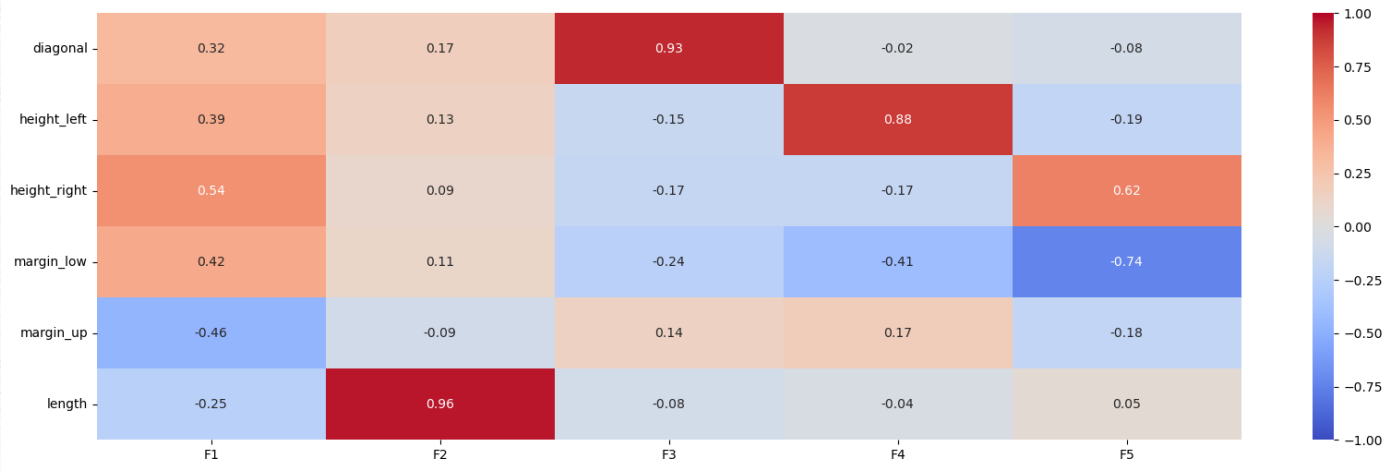
Comprendre et vérifier les axes



- Le premier axe capte les variables les plus liées à la véracité d'un billet
- Le deuxième axe correspond à la longueur du billet (très liée elle aussi)
- Le troisième axe représente la diagonale

Analyse en Composantes Principales

Calculer les valeurs sur les nouveaux axes



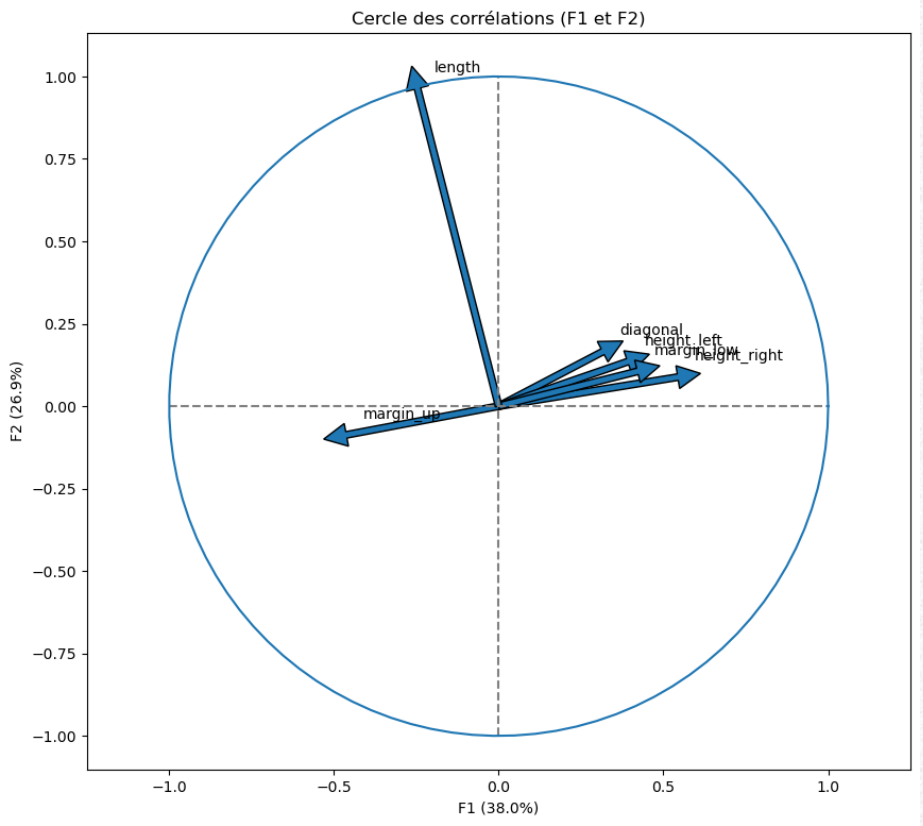
Le premier axe sera calculé sur chaque ligne en multipliant par les **coefficients** correspondants, puis en additionnant le tout

3 Analyse en Composantes Principales

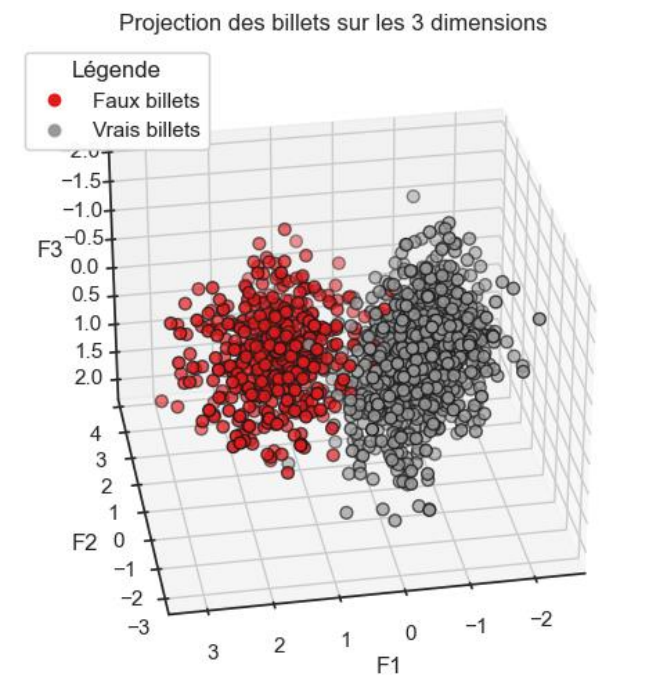
ONCFM

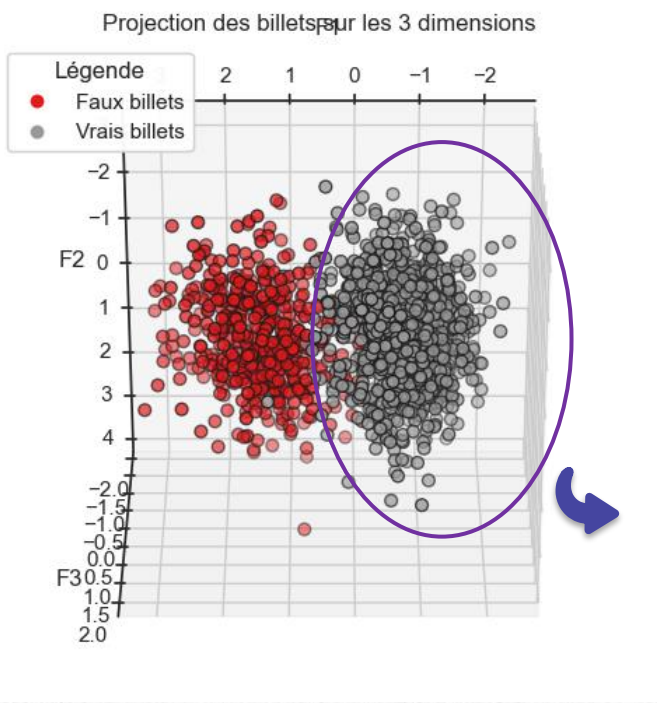
Cercle des corrélations

- margin_up est corrélée négativement à F1
- height_right, margin_down et height_left sont corrélées positivement à F1
- length est très liée à F2



Projection sur les axes





Les vrais billets ont presque tous des valeurs négatives sur F1

3

ONCFM

Algorithmes

Préparation

Découpage des différentes versions du dataset en Train et en Test pour validation croisée simple

Train

Entraînement des modèles

Calcul des métriques

Test

Test des modèles

Calcul des métriques et comparaison

Bilan sur l'impact des potentiels outliers

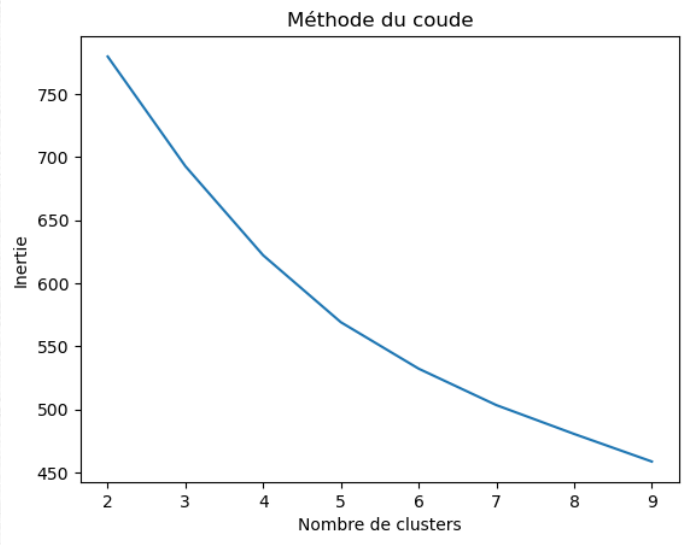
Comparaison

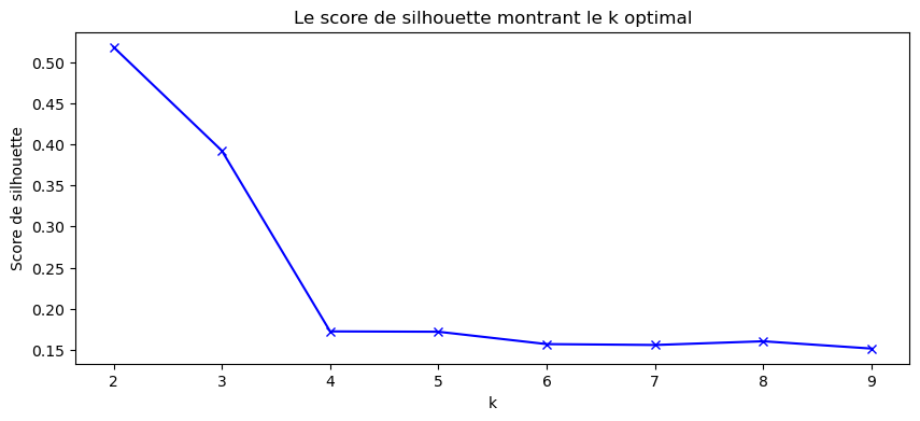
Choix du meilleur modèle en comparant les métriques et les scores obtenus

3 Méthode des K-means

ONCFM

Calcul du nombre optimal de clusters



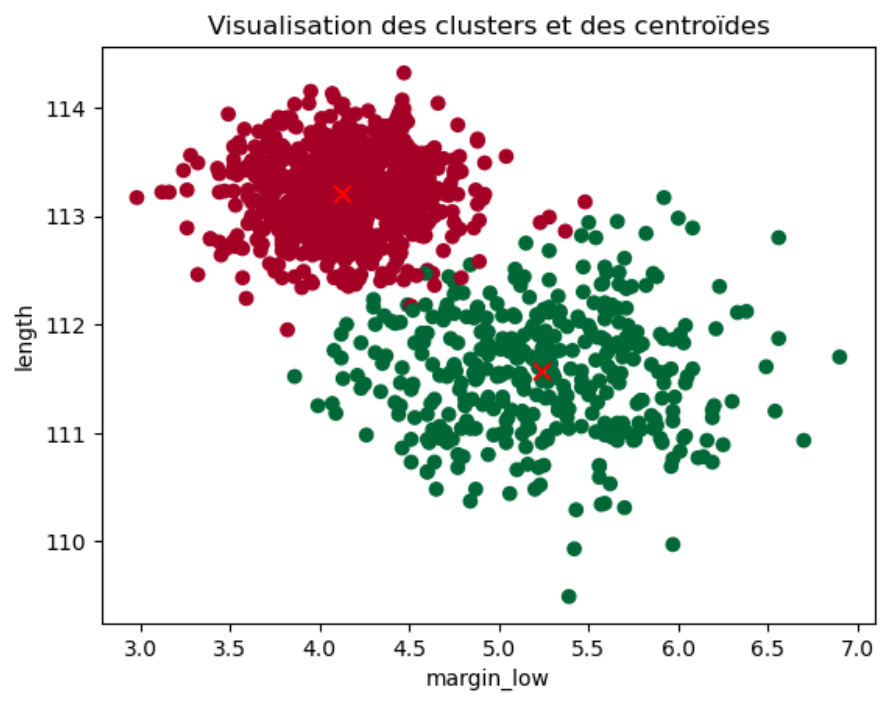


Ici 2 : on anticipe les classes de vrais / faux billets (à confirmer avec l'étude des centroïdes)

3 Vérification des centroïdes

ONCFM

Sur la base du Train (version non normalisée)



- Les centroïdes ont bien convergé et sont centrés
- Plusieurs tests pour confirmer l'inertie maximale
- Cluster 0 en rouge
  Cluster 1 en vert

Étude des centroïdes

ONCFM

Sur la base du Train (version non normalisée)

cluster_kmeans	diagonal	height_left	height_right	margin_low	margin_up	length	value_counts
0	171.982670	103.956359	103.816092	4.123944	3.059466	113.203859	824
1	171.900957	104.192473	104.147580	5.237287	3.351383	111.564096	376

- Des moyennes très proches de celles des types de billets
- Notre cluster 0 correspond aux vrais billets, et notre cluster 1 aux faux

Comparaison avec les moyennes des types de billets :

is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length	value_counts
0	171.903154	104.190026	104.146538	5.213462	3.353000	111.608154	390
1	171.983025	103.953457	103.810864	4.116173	3.053642	113.210988	810

- 14 erreurs de classification
- Un taux d'erreur de 1,5 % sur la base du Train

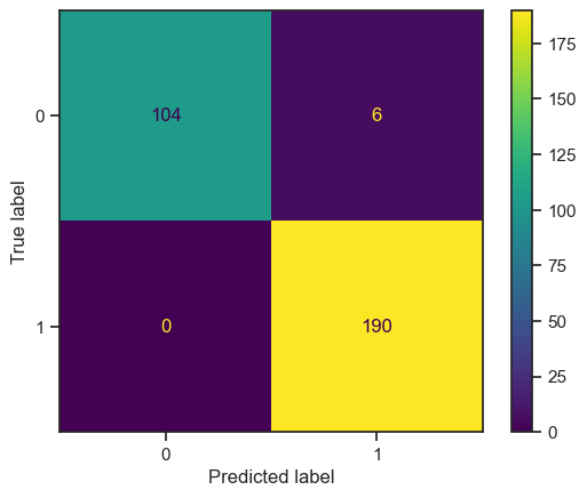
Rappel : is_genuine = 0 -> Faux billets
is_genuine = 1 -> Vrais billets

3 Performances du K-means

ONCFM

Sur la base du Test (version non normalisée)

Matrice de confusion :



- 6 faux positifs : notre modèle a prédit 6 faux billets comme étant vrais

Rapport de classification :

	precision	recall	f1-score	support
0	1.00	0.95	0.97	110
1	0.97	1.00	0.98	190
accuracy			0.98	300
macro avg	0.98	0.97	0.98	300
weighted avg	0.98	0.98	0.98	300

- Pour qu'un modèle soit performant, il doit identifier le plus possible de faux billets
- On veut donc la précision la plus proche de 1 possible (et pour le f1-score)
- Un taux d'erreur de classification de 2 % sur le Test

3

Régression logistique

Sur la base du Train (version non normalisée)

ONCFM

Premier essai avec toutes les variables

Logit Regression Results

Dep. Variable:	is_genuine	No. Observations:	1200
Model:	Logit	Df Residuals:	1193
Method:	MLE	Df Model:	6
Date:	Sat, 20 Apr 2024	Pseudo R-squ.:	0.9517
Time:	17:03:35	Log-Likelihood:	-36.563
converged:	True	LL-Null:	-756.70
Covariance Type:	nonrobust	LLR p-value:	4.619e-308

	coef	std err	z	P>|z|	[0.025	0.975]
const	-83.9258	257.673	-0.326	0.745	-588.957	421.105
diagonal	-0.3890	1.151	-0.338	0.735	-2.646	1.868
height_left	-1.8596	1.276	-1.458	0.145	-4.360	0.641
height_right	-2.2581	1.093	-2.066	0.039	-4.400	-0.116
margin_low	-5.3287	0.976	-5.460	0.000	-7.242	-3.416
margin_up	-8.8390	2.111	-4.186	0.000	-12.977	-4.701
length	5.6169	0.895	6.276	0.000	3.863	7.371

- On retirera par la suite la variable la moins significative et ainsi de suite
- Ces 4 variables sont significatives dans la configuration actuelle (coefficient statistiquement différent de 0 avec une p-value < 0,05)
- Pas de variables colinéaires

3 Régression logistique

Sur la base du Train (version non normalisée)

ONCFM

Troisième essai

Logit Regression Results

Dep. Variable:	is_genuine	No. Observations:	1200
Model:	Logit	Df Residuals:	1195
Method:	MLE	Df Model:	4
Date:	Sat, 20 Apr 2024	Pseudo R-squ.:	0.9511
Time:	17:03:36	Log-Likelihood:	-36.990
converged:	True	LL-Null:	-756.70
Covariance Type:	nonrobust	LLR p-value:	1.963e-310

	coef	std err	z	P>|z|	[0.025	0.975]
height_left	-2.4744	1.070	-2.314	0.021	-4.571	-0.378
height_right	-2.6582	0.979	-2.715	0.007	-4.577	-0.739
margin_low	-5.2354	0.911	-5.747	0.000	-7.021	-3.450
margin_up	-9.1009	2.101	-4.332	0.000	-13.219	-4.983
length	5.2192	0.724	7.205	0.000	3.799	6.639

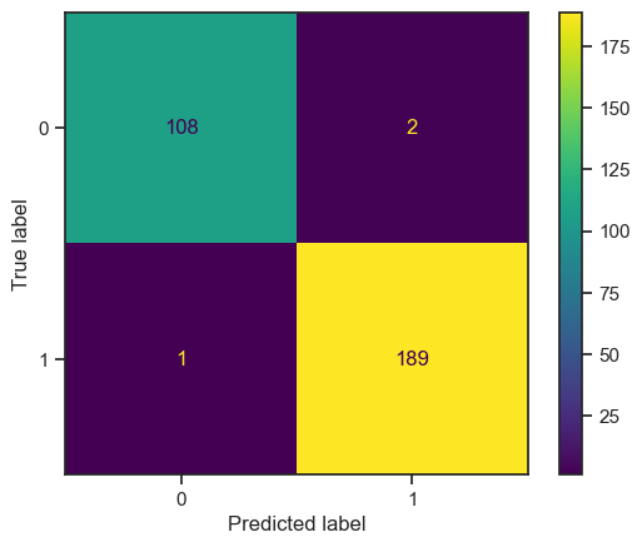
- Cette fois toutes les variables sont significatives
- Pas de variables colinéaires

3 Performances du modèle

ONCFM

Sur la base du Test (version non normalisée)

Matrice de confusion :



- 2 faux positifs pour 1 faux négatif

Rapport de classification :

	precision	recall	f1-score	support
0	0.99	0.98	0.99	110
1	0.99	0.99	0.99	190
accuracy			0.99	300
macro avg	0.99	0.99	0.99	300
weighted avg	0.99	0.99	0.99	300

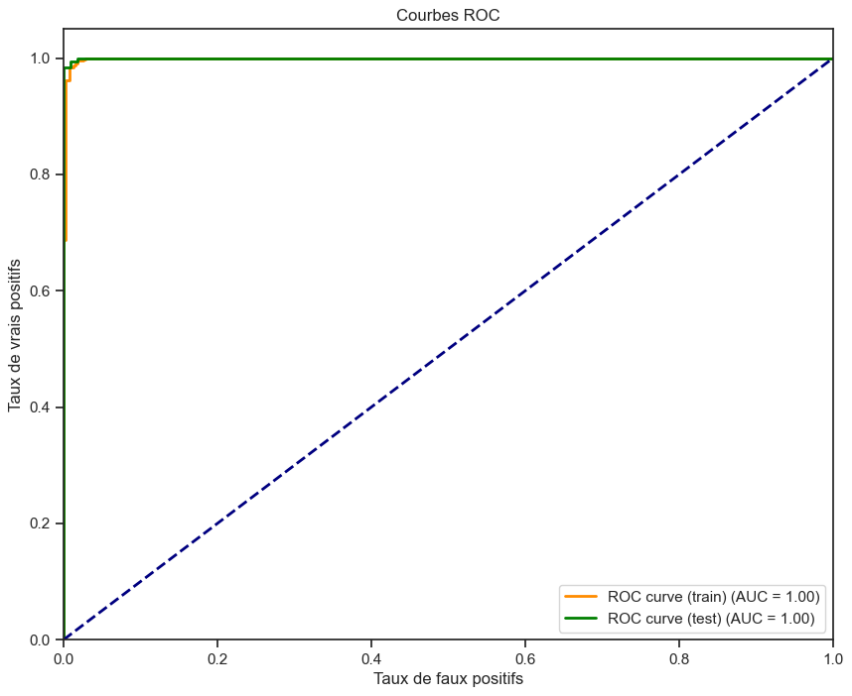
- De bons résultats sur les métriques
- Une précision et un F1-score très proches de 1
- Un taux d'erreur de classification de 1 % sur le Test

3

Performances du modèle

ONCFM

Courbe ROC :



- De bons résultats : les 2 courbes sont quasiment dans le coin supérieur gauche
- Une courbe de Test meilleure que celle d'entraînement
- Des performances qui peuvent encore être améliorées en calculant le seuil optimal

4

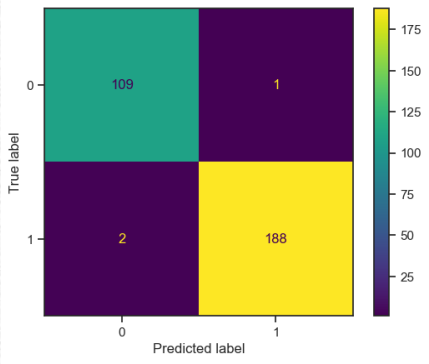
Optimisation du modèle

ONCFM

Recherche du seuil optimal

Approche basée sur la courbe ROC

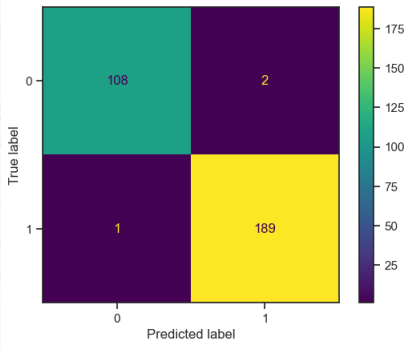
Seuil trouvé : 0,62



Mêmes résultats

Approche de la précision, du rappel et du F1-score

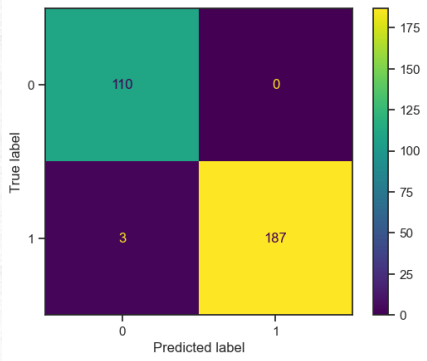
Seuil trouvé : 0,40



Résultats moins bons

Méthode de classification coût sensible

Seuil trouvé : 0,77

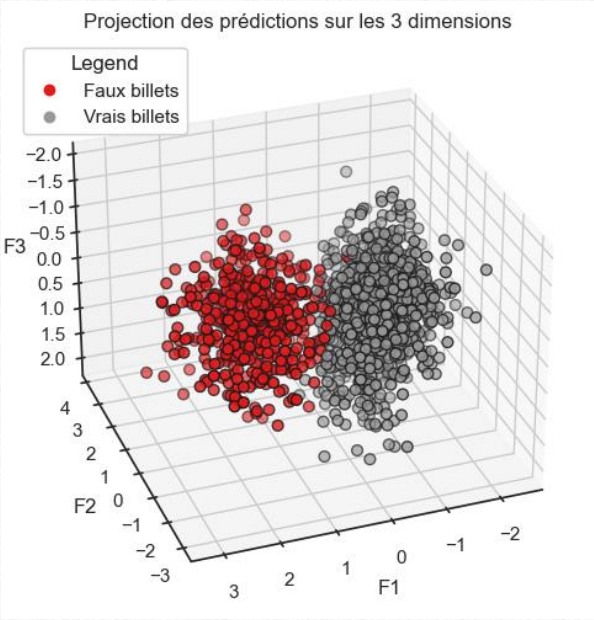


Meilleure prédiction

4 Test final : préparation

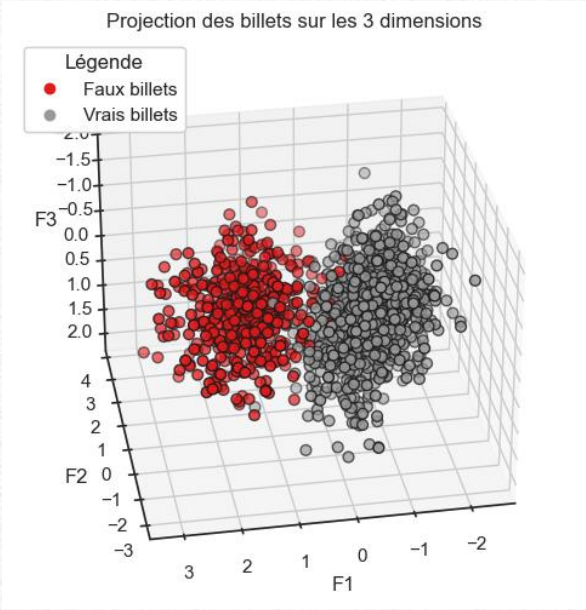
ONCFM

Ré-entraînement sur tout le dataset



6 faux positifs seulement sur les 1500 billets

Visualisation des vrais classes



4 Test final

ONCFM

Prédictions des 5 billets du fichier production :

	id	is_genuine	taux_proba
0	A_1	False	0.000229
1	A_2	False	0.000015
2	A_3	False	0.000023
3	A_4	True	0.980833
4	A_5	True	0.999987

Correspond à la probabilité qu'un billet soit vrai

ONCFM

03

Conclusion

Bilan et recommandations

Bilan et recommandations

ONCFM

Forces

- Modèle sachant réagir aux valeurs atypiques des vrais billets, et donc proche de la situation réelle
- Modèle performant pour la prédiction des faux billets : très peu de faux positifs sur l'entièreté du dataset

Améliorations possibles

- Ajouter l'info sur le type de billet (5€ / 50€...)
- Possibilité de créer une classe pour les prédictions proches du seuil

Recommandations

- Faire plusieurs tests en situation réelle avant déploiement
- Confirmer les valeurs atypiques

ONCFM

04

Démonstration

Démonstration de
l'algorithme final