



Déetectez des faux billets

Glossaire

diagonal : diagonale du billet (en mm)

height_left : hauteur du billet (mesurée sur le côté gauche, en mm)

height_right : hauteur du billet (mesurée sur le côté droit, en mm)

margin_low : marge entre le bord inférieur du billet et l'image de celui-ci (en mm)

margin_up : marge entre le bord supérieur du billet et l'image de celui-ci (en mm)

length : longueur du billet (en mm)

Is_genuine : authenticité du billet

ACP : Analyse par Composantes principales
Objectif : rechercher la projection pour laquelle l'inertie des points est maximale.

KMeans: algorithme de clustering non supervisé

Composantes principales ou **facteurs** : nouvelles variables formées par combinaison linéaire des anciennes variables (variables synthétiques)

Bon partitionnement : homogénéité intra-classe (les individus se ressemblent) et hétérogénéité inter-classe (les groupes diffèrent)



Mission : créer un algorithme de détection de faux billets

“

Votre société de consulting informatique vous propose une nouvelle mission au ministère de l'Intérieur, dans le cadre de la lutte contre la criminalité organisée, à l'Office central pour la répression du faux monnayage.

Votre mission si vous l'acceptez : créer un algorithme de détection de faux billets.

Votre mission

Sommaire

- **Importation et lecture** du fichier
- **Analyse et description** des données
- **ACP** de l'échantillon
- **Partitionnement** par KMeans et **visualisation** par ACP
- **Modélisation** par régression **logistique**

Mission 0

Analyse des données

Importation et lecture du fichier csv

Données

Contient les **caractéristiques géométriques** du billet de banque.

Nom : notes.csv

Source : OpenClassrooms
<https://openclassrooms.com/fr/paths/65/projects/147/assignment>

170 Lignes - 7 colonnes

Caractéristiques

6 variables numériques
(même unité mm, type float)

1 variable qualitative (type booléen)

Pas de valeurs nulles

Pas de valeurs dupliquées

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
0	True	171.81	104.86	104.95	4.52	2.89	112.83
1	True	171.67	103.74	103.70	4.01	2.87	113.29

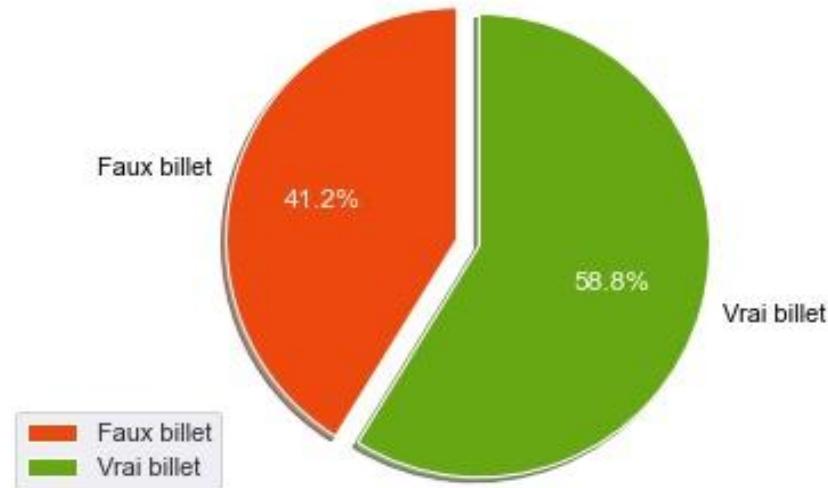
Majorité de vrais billets

170 billets au total

100 billets vrais

70 billets faux

Répartition billets selon leur authenticité

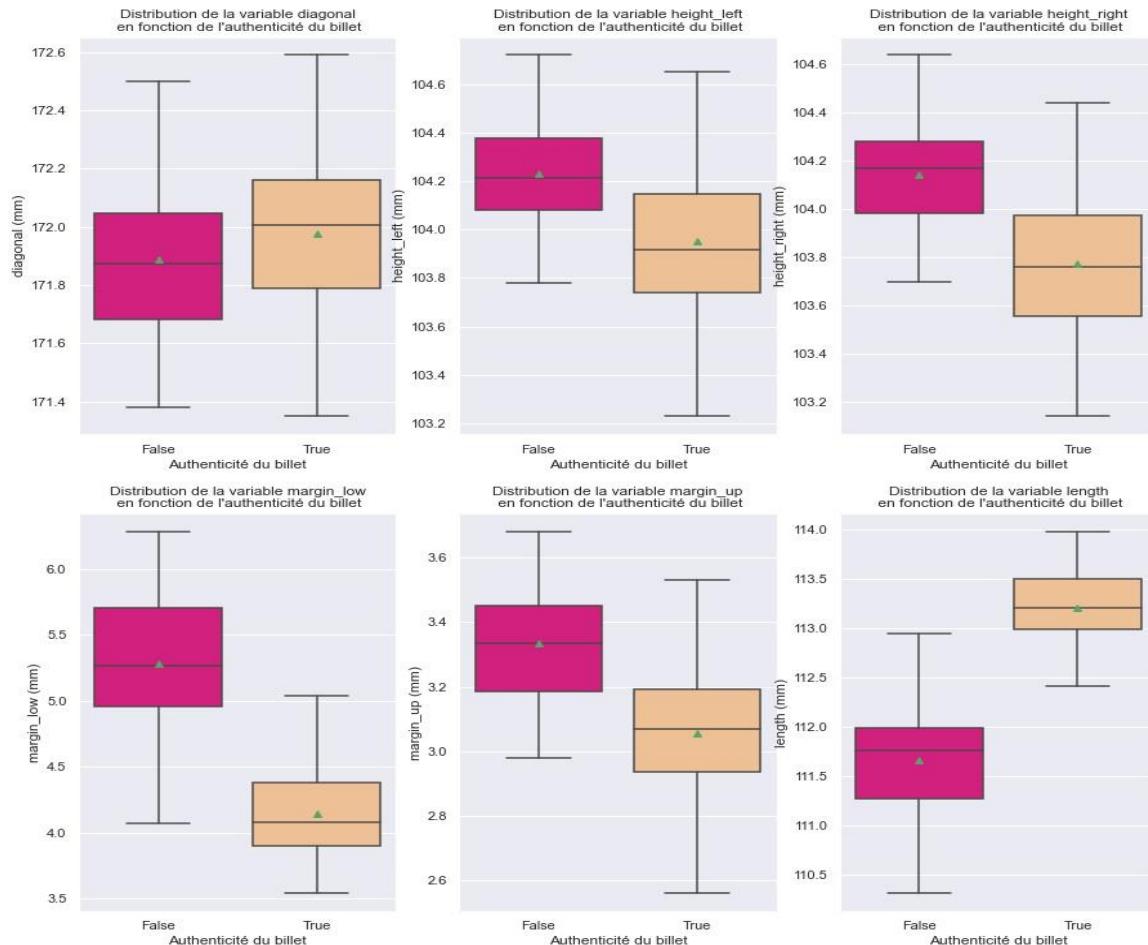


Des métriques différentes selon l'authenticité du billet

Billet vrai

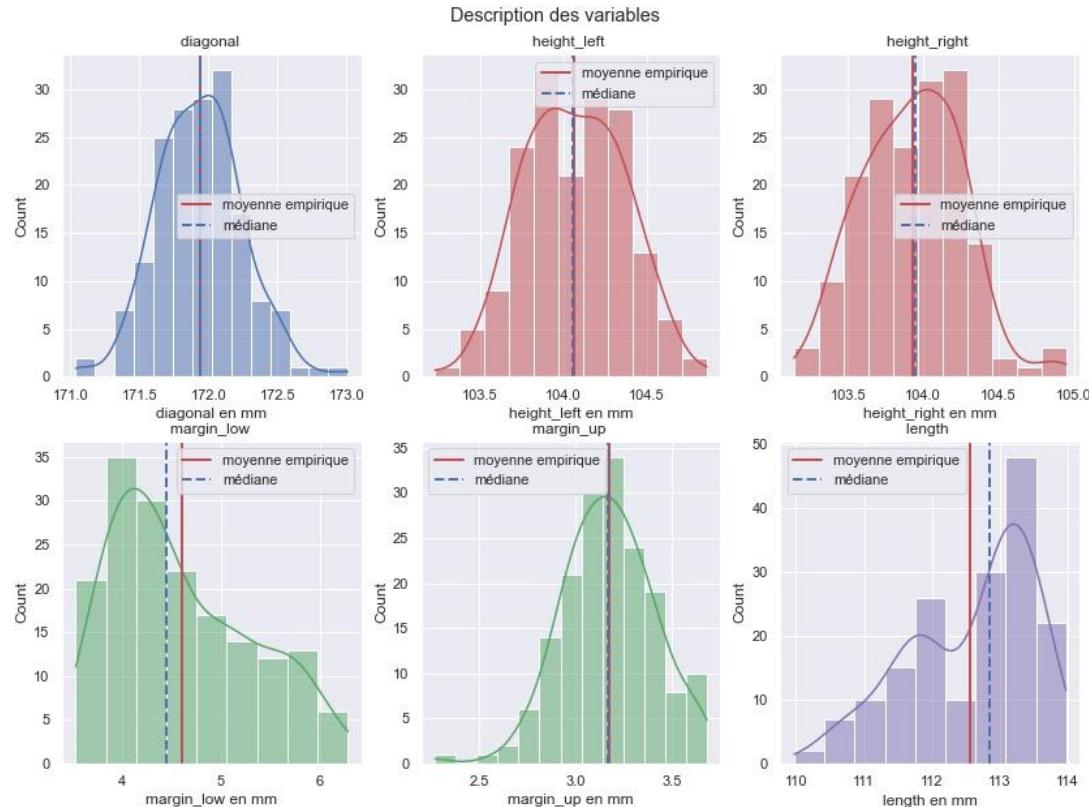
- étendue des valeurs plus grande
- diagonale plus grande
- hauteur plus faible
- marge plus petite
- longueur plus grande

Répartition des métriques des billets en fonction de leur authenticité (True) ou non (False)



Des variables gaussiennes ?

Application du test de Kolmogorov-Smirnov



Les métriques en chiffre

	Médiane	Moyenne	Ecart type	Kurtosis	Skewness
diagonal	171.9	171.9	0.09	0.58	0.19
height_left	104.05	104.06	0.08	-0.46	0.028
height_right	103.95	103.92	0.10	-0.004	0.16
margin_low	4.45	4.61	0.49	-0.74	0.58
margin_up	3.17	3.17	0.05	0.54	-0.20
length	112.84	112.57	0.84	-0.53	-0.65

Test de Kolmogorov-Smirnov

Test d'hypothèse

H0 : la variable suit une loi normale

H1 : la variable ne suit pas une loi normale

Seuil alpha : 0.05

H0 non rejetée au seuil alpha

Les variables suivantes suivent une **loi normale**

H0 rejetée au seuil alpha

Les variables suivantes ne suivent **pas une loi normale**

Margin_low : p_value : 0.007

Length : p_value : 0.02

Diagonal : p_value : 0.92

height_left : p_value : 0.83

height_right : p_value : 0.82

margin_up : p_value : 0.38



Les groupes False et True sont-ils distincts ?

**Test d'égalité des variances
(Levene car plus robuste
aux écarts non gaussiens)**

**HO non rejetée, les
variances des 2 groupes
sont égales** pour les variables

diagonal : p_value : 0.62

height_right : p_value : 0.17

margin_up : p_value : 0.98

Test d'hypothèse

HO : les variances sont
égales

H1 : les variances ne sont
pas égales
seuil alpha : 0.05

**HO rejetée, les variances
des 2 groupes sont inégales**
pour la variable

height_left : p_value : 0.03

Poursuite de l'étude avec un test de Student sur les variables dont les variances sont égales

Test d'hypothèse

H_0 : les moyennes sont égales

H_1 : les moyennes ne sont pas égales

seuil alpha : 0.05

H_0 rejetée, les moyennes des 2 groupes sont inégales pour les variables

height_right : p_value : 2.42e-16

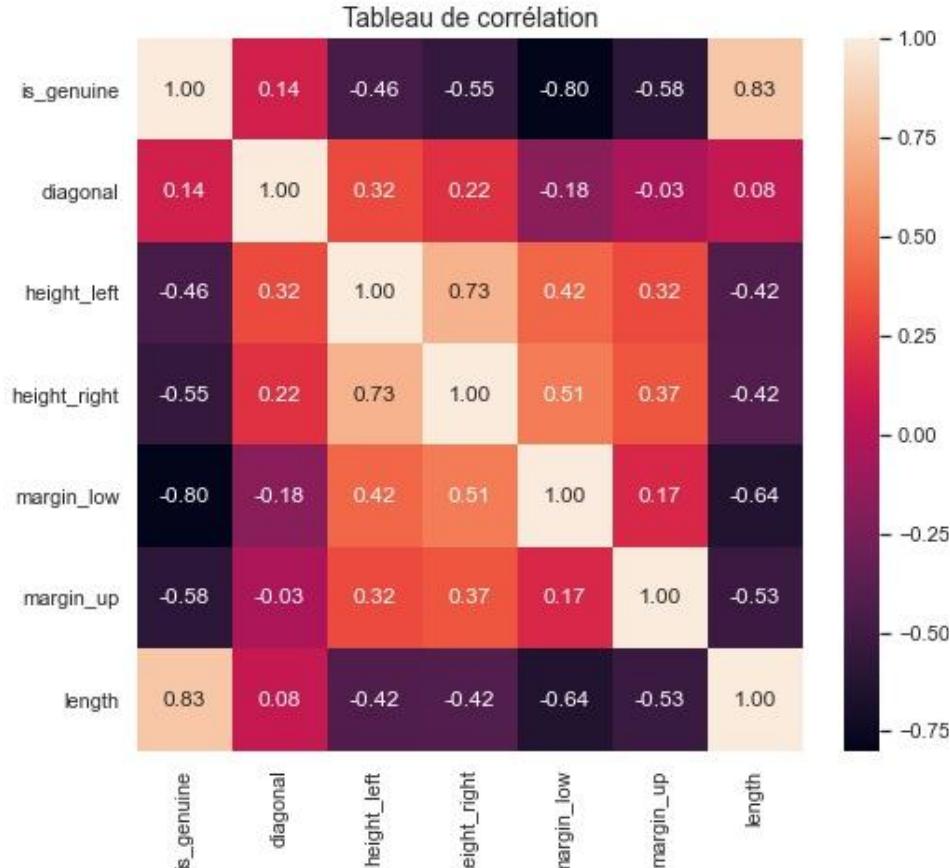
margin_up : p_value : 4.25e-13

H_0 non rejetée, les moyennes des 2 groupes sont égales pour la variable
diagonal : p_value : 0.12

Seule la variable **diagonal** ne permet pas de conclure à la **distinction** des échantillons True et False

Des variables corrélées ?

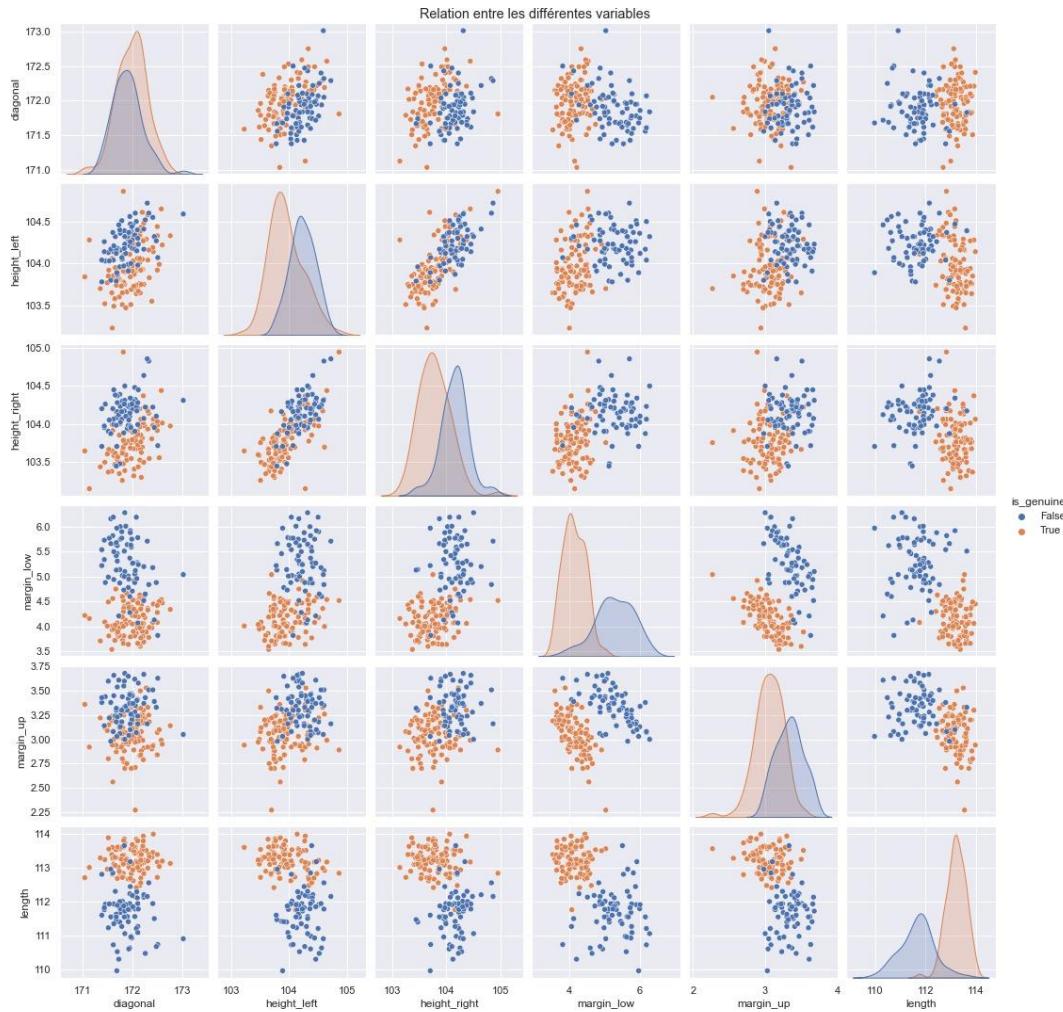
Authenticité des billets corrélée avec **length** (0.83) et anti corrélée **margin_low** (- 0.80)



Vrai billet VS faux billet

Différentiation par
margin_low
length

La variable **diagonal**
est la variable la **moins**
significative





L'authenticité corrélée à certaines métriques

**η^2 = rapport de corrélation
= variation interclasse /
variation totale**

η^2 diagonal : 0.02
 η^2 height_left : 0.21
 η^2 height_right : 0.30
 η^2 margin_low : 0.64
 η^2 margin_up : 0.34
 η^2 length : 0.68

Forte corrélation
authenticité - length
authenticité - margin_low

Absence de corrélation
authenticité - diagonal

Mission 1

Analyse en Composantes Principales ACP

Analyse en Composantes Principales (ACP)

Permet d'étudier

La **variabilité** des billets

La **liaison** des variables entre elles afin de créer des **variables synthétiques** calculées à partir des variables initiales corrélées entre elles

Permet de

Réduire les dimensions du dataset

Repérer les outliers

Comment ?

Projection des données (centrées - réduites) sur les composantes principales tout en minimisant la distance de projection

Conservation d'un maximum de variance (et ainsi limiter la perte d'information)

Composantes principales

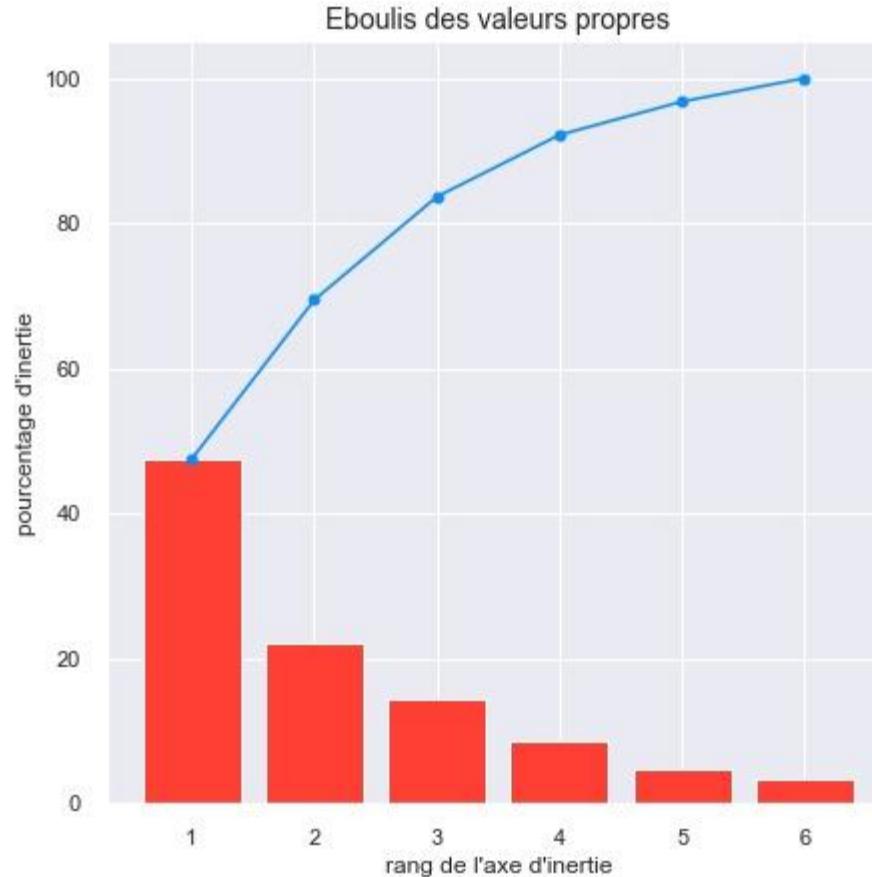
Vecteurs propres de la matrice de covariance - combinaison linéaire des autres variables



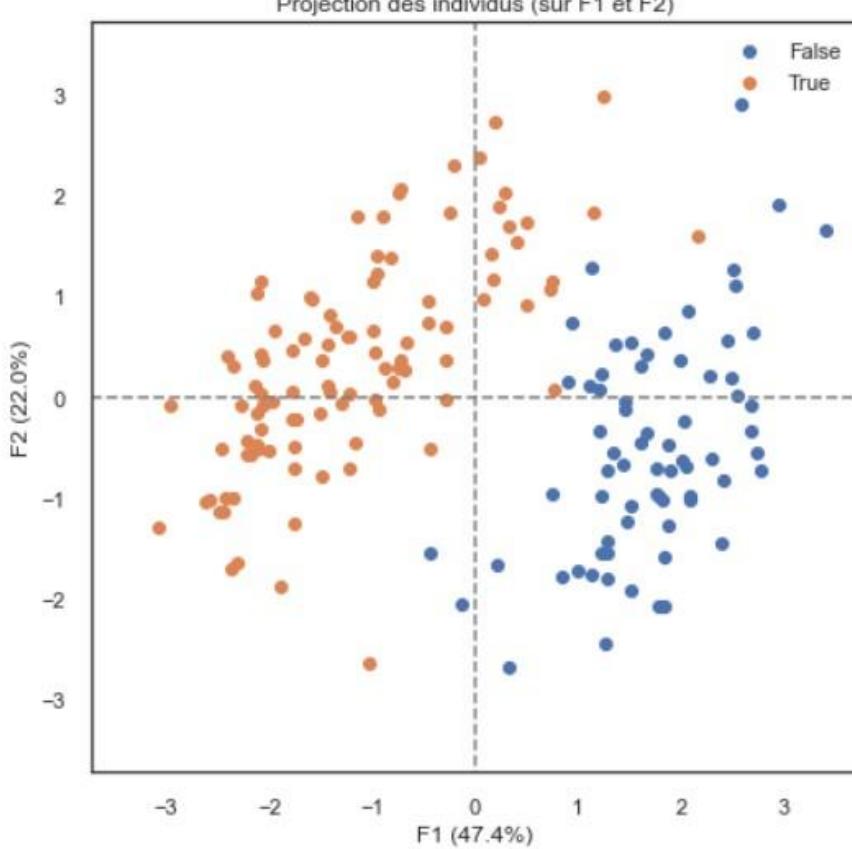
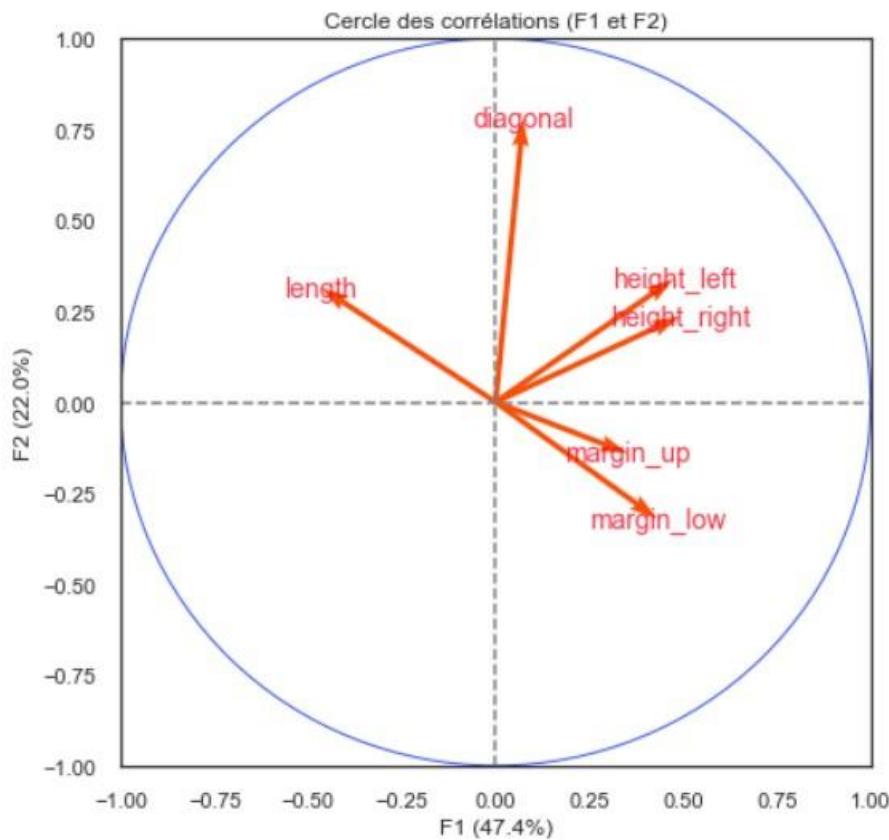
Eboulis des valeurs propres

Décrit le pourcentage d'inertie totale associé à chaque axe

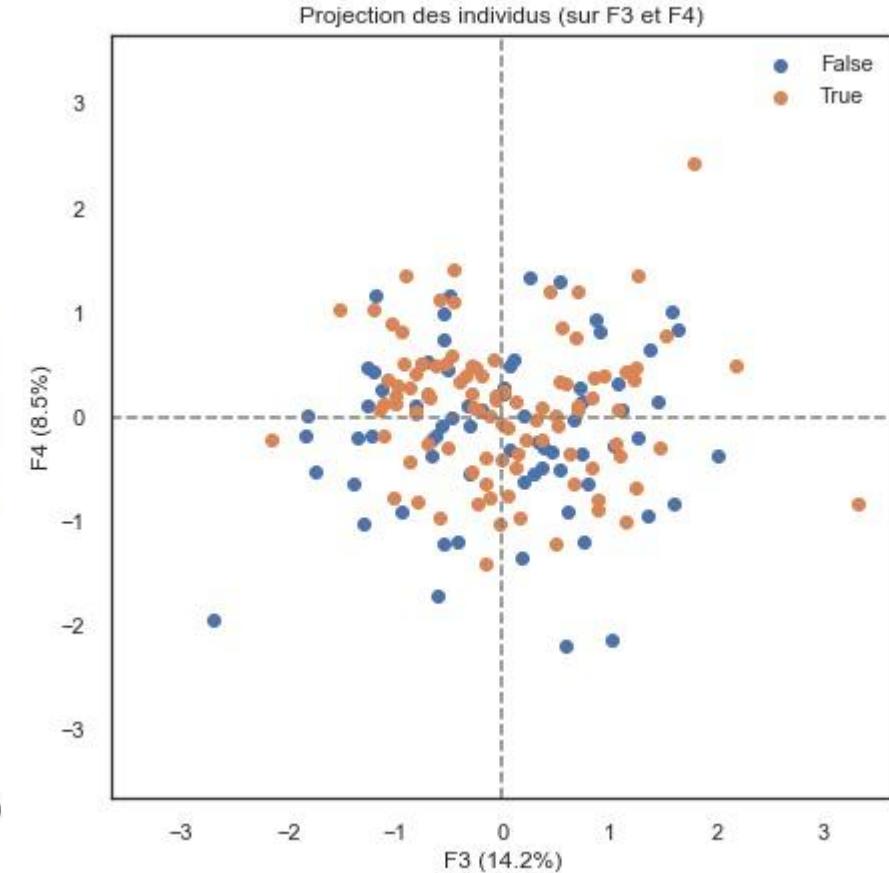
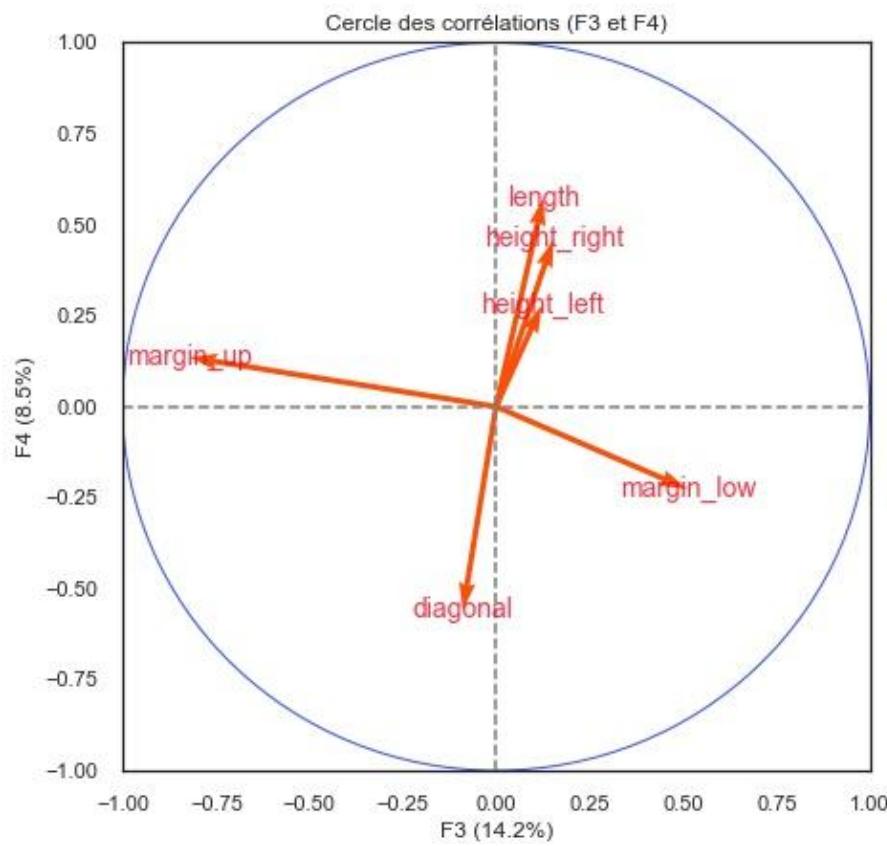
Les 4 premières composantes suffisent à expliquer 92% de l'inertie totale



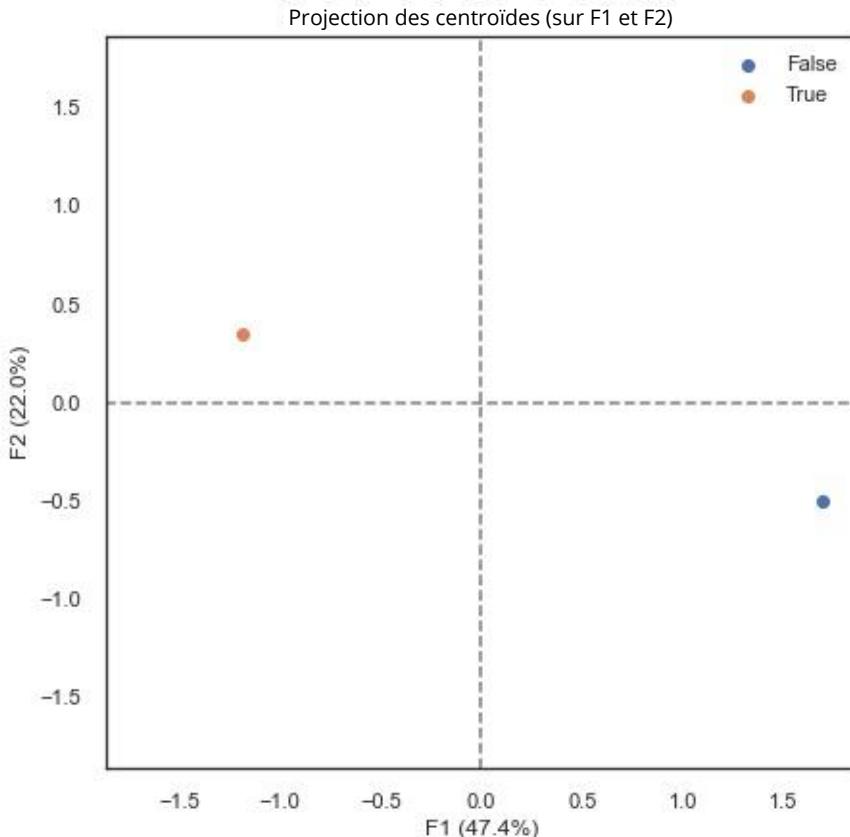
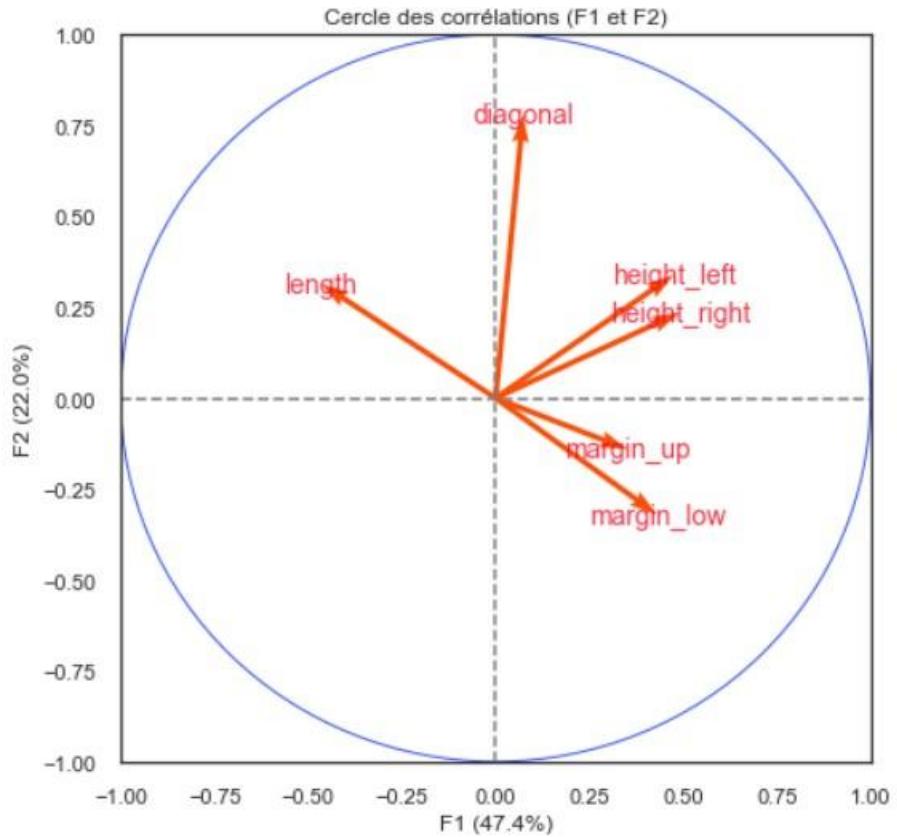
Cercle de corrélation et projection des individus sur F1 et F2



Cercle de corrélation et projection des individus sur F3 et F4



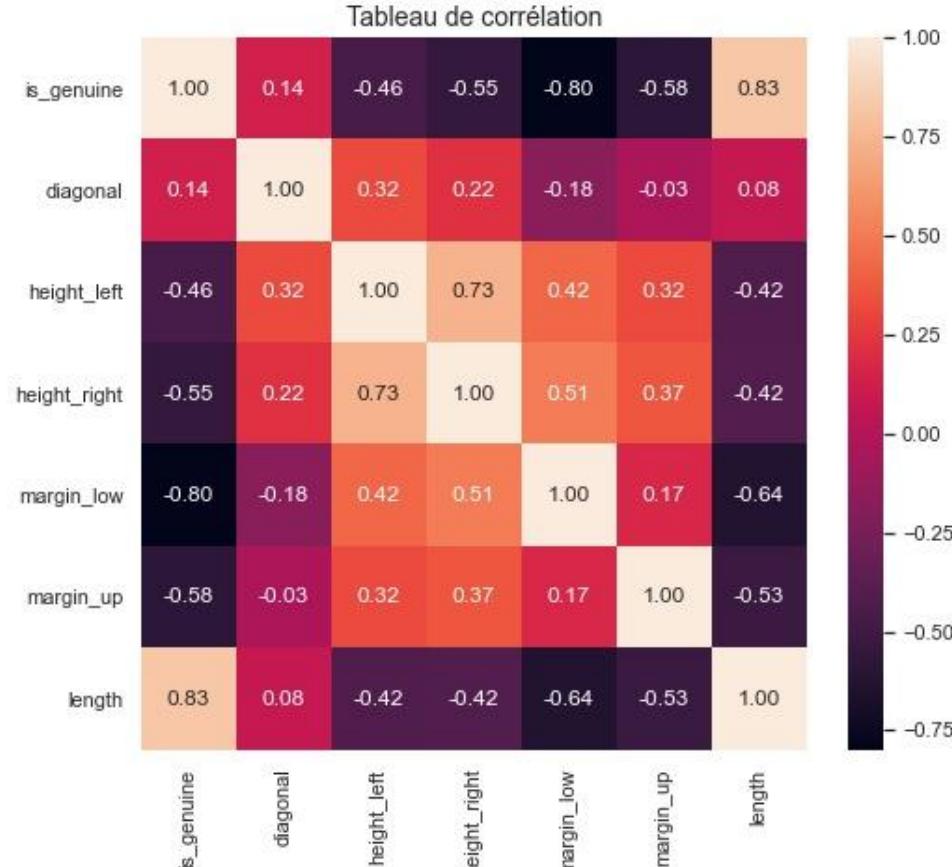
Cercle de corrélation et projection des centroïdes sur F1 et F2



Corrélation entre les variables

Grande corrélation avec length et is_genuine (0.83)

corrélation avec height_right et height_left (0.73)



Interprétation ACP

Variables les plus corrélées à F1

height_left

height_right

margin_low

margin_up

Length

Variable la plus corrélée à F2

diagonal

F1

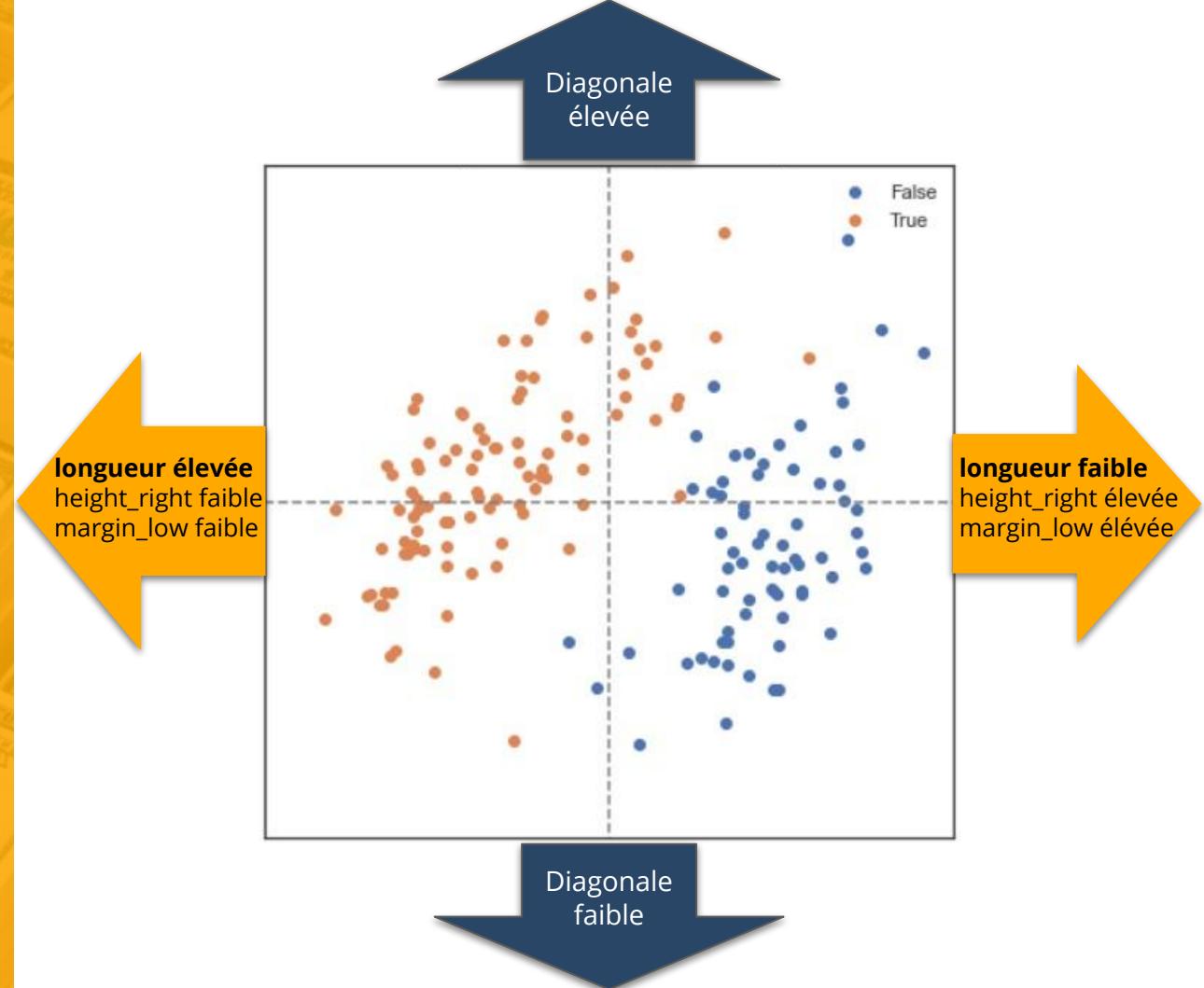
+0.07 * Diagonal
+0.47 * height_left
+0.49 * height_right
+0.43 * margin_low
+0.35 * margin_up
-0.46 * length

F2

+0.77 * diagonal
+0.33 * height_left
+0.23 * height_right
-0.32 * margin_low
-0.14 * margin_up
+0.31 * length

Pour résumer

L'authenticité des billets se caractérise bien avec la longueur. La diagonale n'est pas significative.



Qualité de représentation des individus sur les axes

Carré des distances à l'origine des individus

Correspond également à leur contribution dans l'inertie totale

Détermination du \cos^2

Calcul du \cos^2 de l'angle du vecteur (centre du nuage et l'individu) et vecteur (centre du nuage et projection sur l'axe de l'individu)

Individu bien représenté

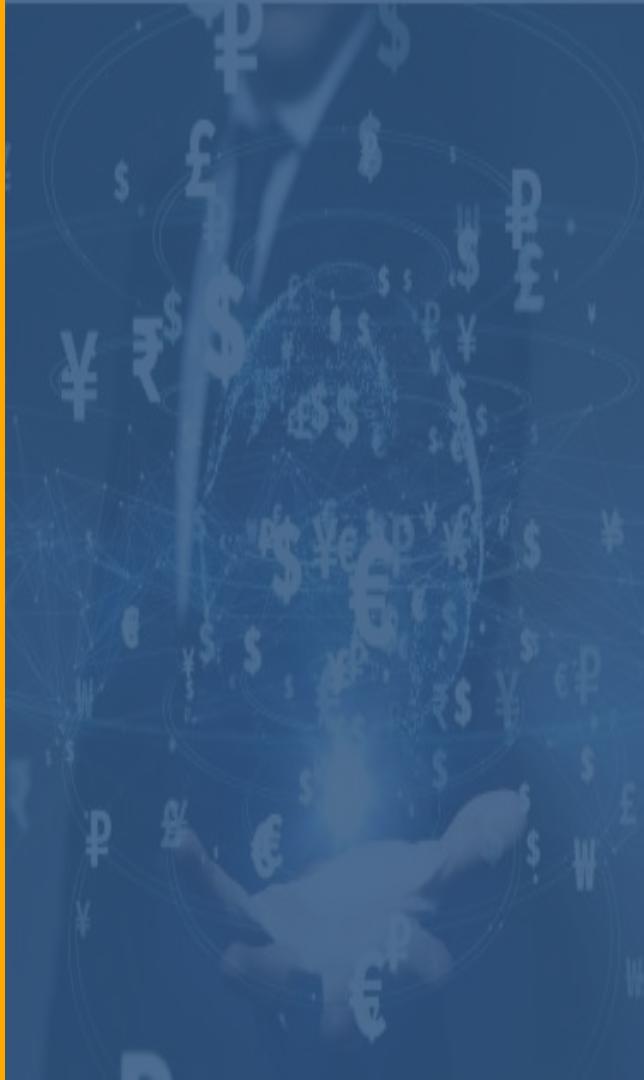
\cos^2 proche de 1 donc l'angle est proche de 0

Individu bien projeté = bien interprétable

Individu mal représenté

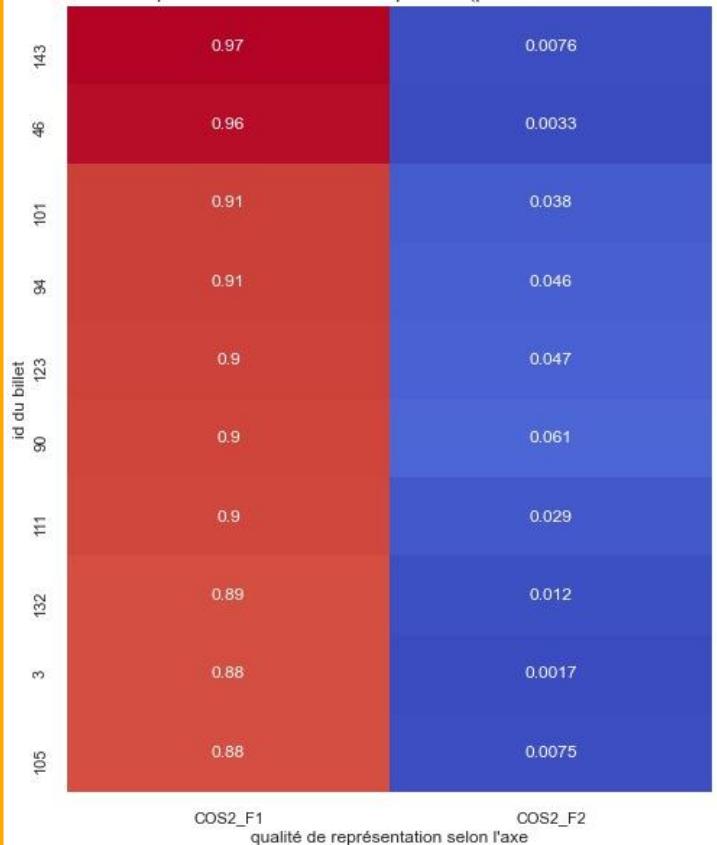
\cos^2 proche de 0 donc l'angle est proche de 90°

Loin du plan de projection

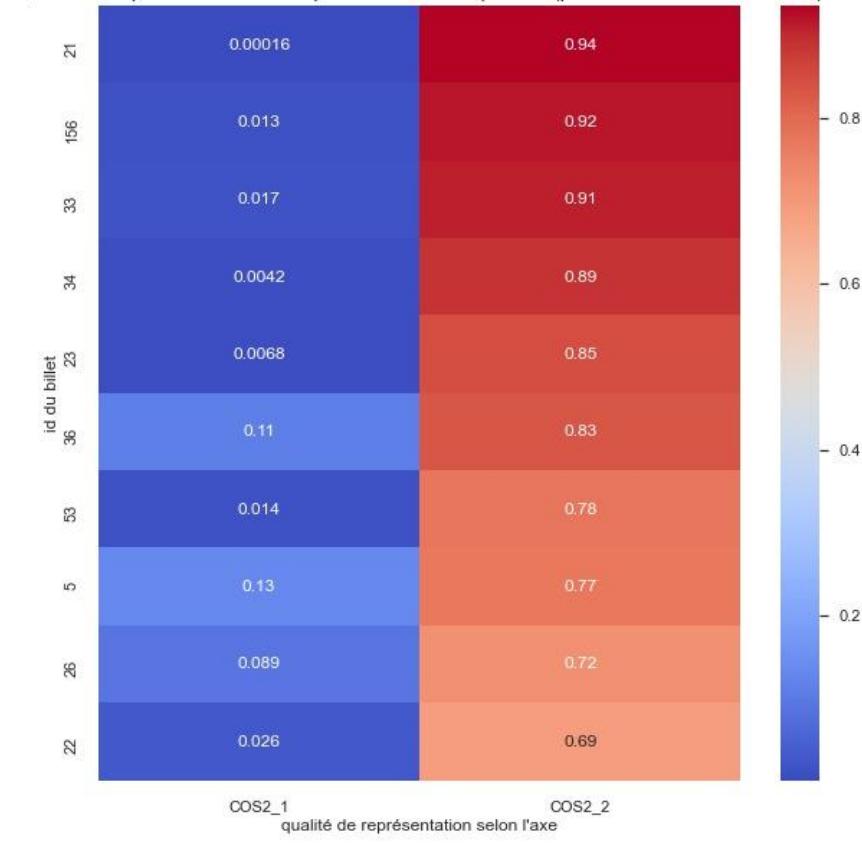


Les 10 meilleures représentations d'individus en fonction de l'axe

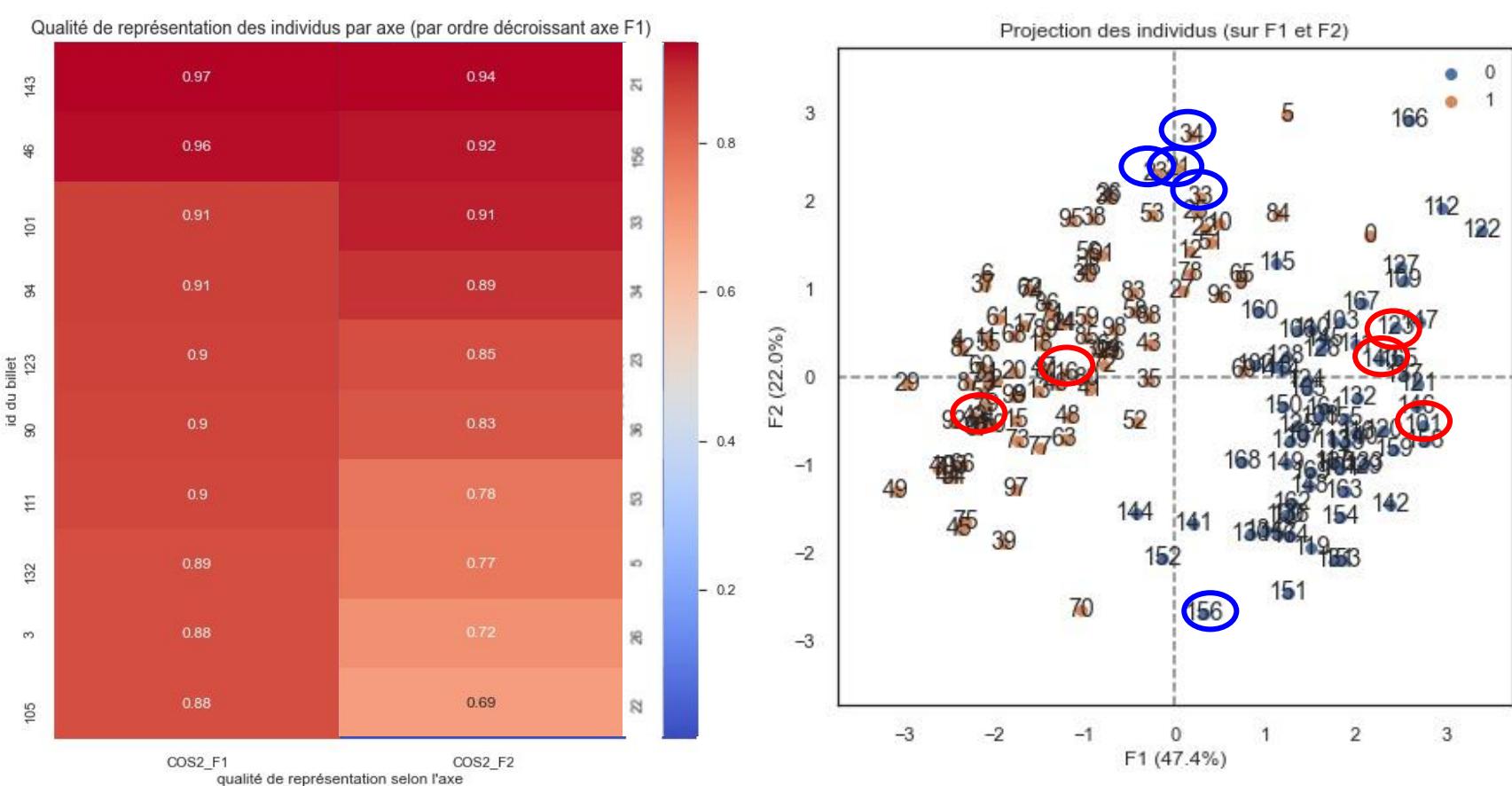
Qualité de représentation des individus par axe (par ordre décroissant axe F1)



Qualité de représentation des 10 premiers individus par axe (par ordre décroissant axe F2)



Les 10 meilleures représentations d'individus en fonction de les axes F1 et F2



Contribution des individus aux axes

Mesure

L'influence de l'individu sur le calcul de certaines composantes

Permet de

Repérer si un individu est particulier

Détermination simplifiée du CTR

Inertie projeté de l'individu sur l'axe / inertie projetée totale de tous les individus sur l'axe

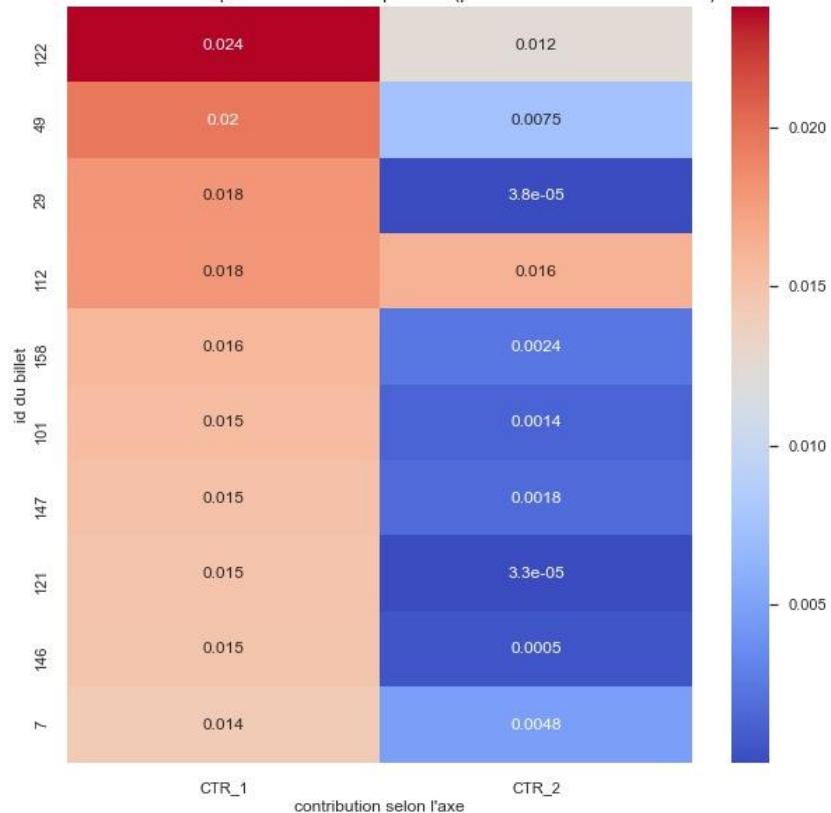
Permet

L'interprétation des axes si les individus sont identifiés

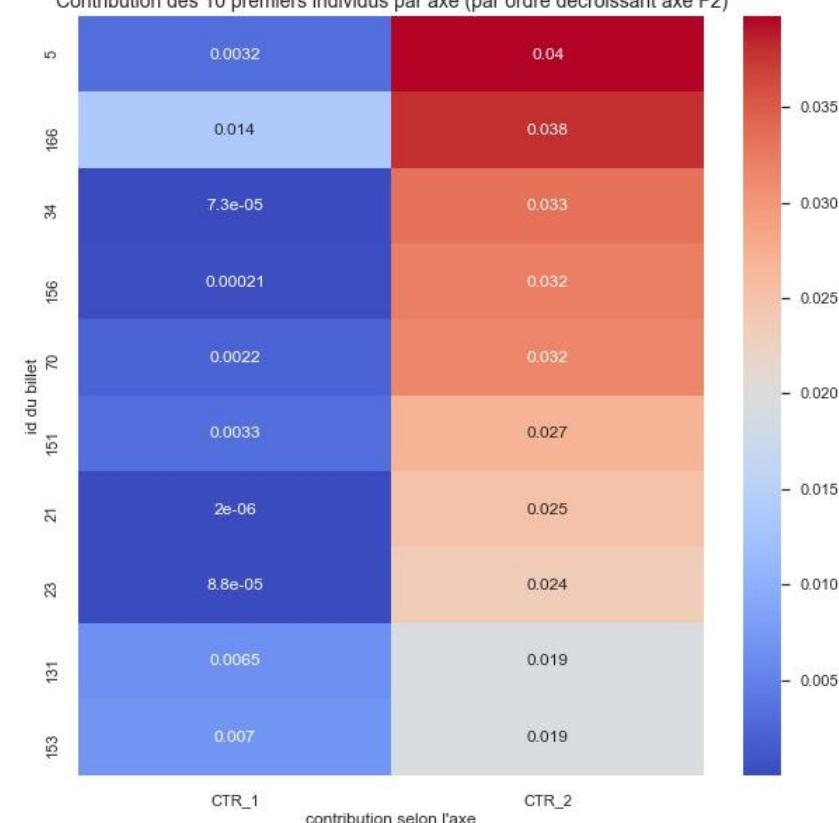


Les 10 meilleures contributions d'individus en fonction de l'axe

Contribution des 10 premiers individus par axe (par ordre décroissant axe F1)



Contribution des 10 premiers individus par axe (par ordre décroissant axe F2)



Mission 2

KMeans et ACP

A photograph of a person from the chest up, wearing a virtual reality headset. A large, semi-transparent yellow grid is overlaid on the image, covering most of the background. The person appears to be looking slightly to the right.

KMeans

Algorithme non supervisé de clustering

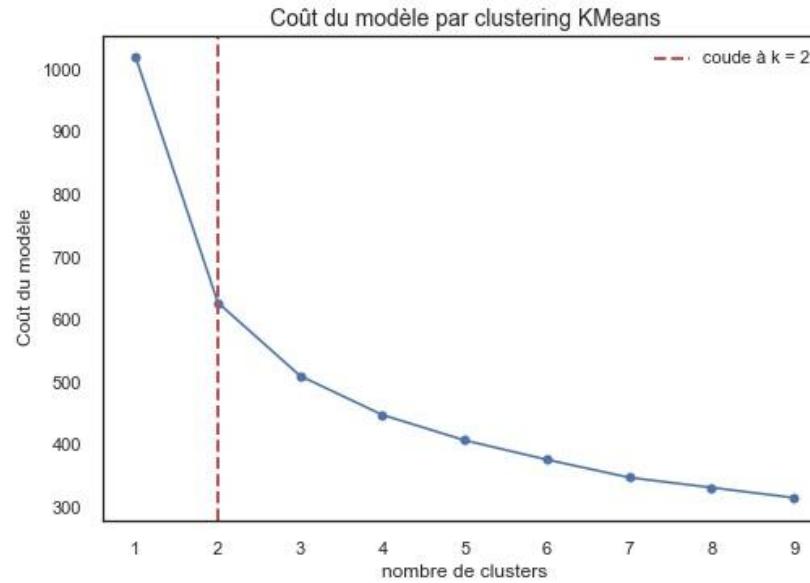
Clustering par KMeans

Nombre de clusters : 2

groupe vrais billets : 94

Groupe faux billets : 76

Il existe des vrais billets dans groupe de faux et vice versa



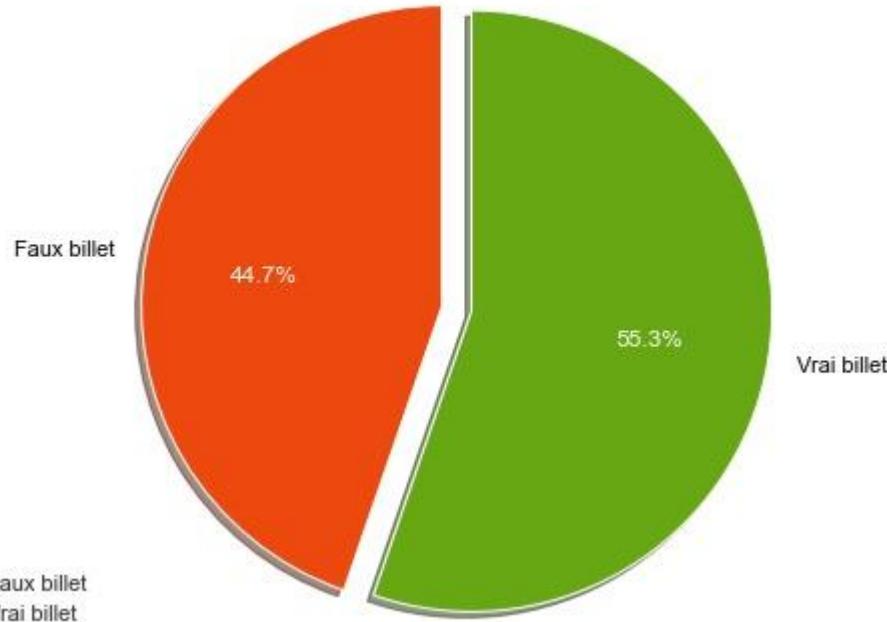
Des billets mal répartis

1 billet faux classé vrai

7 billets vrais classés faux

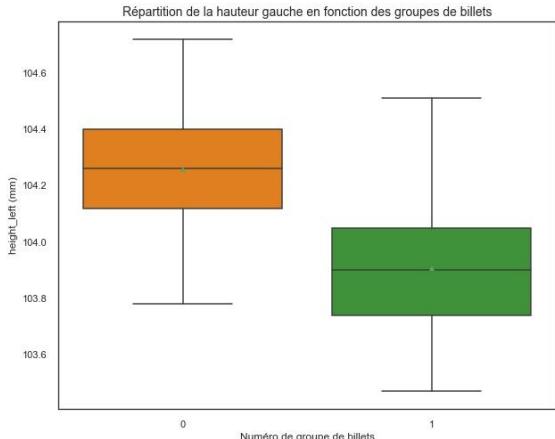
Taux d'erreur après clustering : 4.7 %

Répartition billets selon leur authenticité (après clustering)



Les 2 groupes sont-ils distincts ?

Variable gaussienne height_left



Test de Levène

H0 : les variances sont égales

H1 : les variances ne sont pas égales

Seuil alpha : 0.05

p_value = 0.3

H0 ne peut pas être rejetée au seuil alpha choisi

Les variances des deux groupes sont égales

Test de Student

H0 : les moyennes sont égales

H1 : les moyennes ne sont pas égales

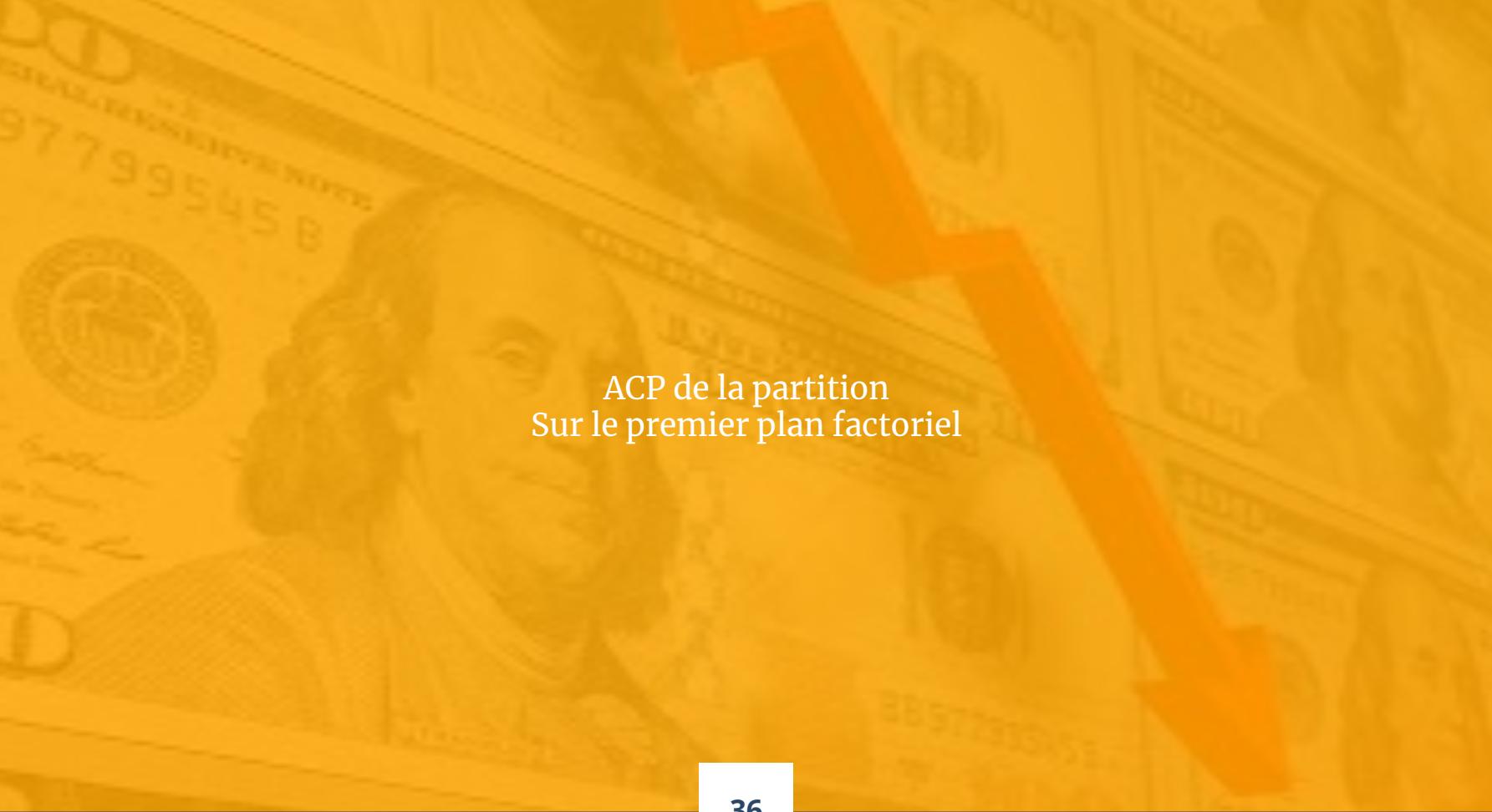
Seuil alpha : 0.05

p_value = 4.9 e-16

H0 peut être rejetée au seuil alpha choisi

Les moyennes des deux groupes sont différentes

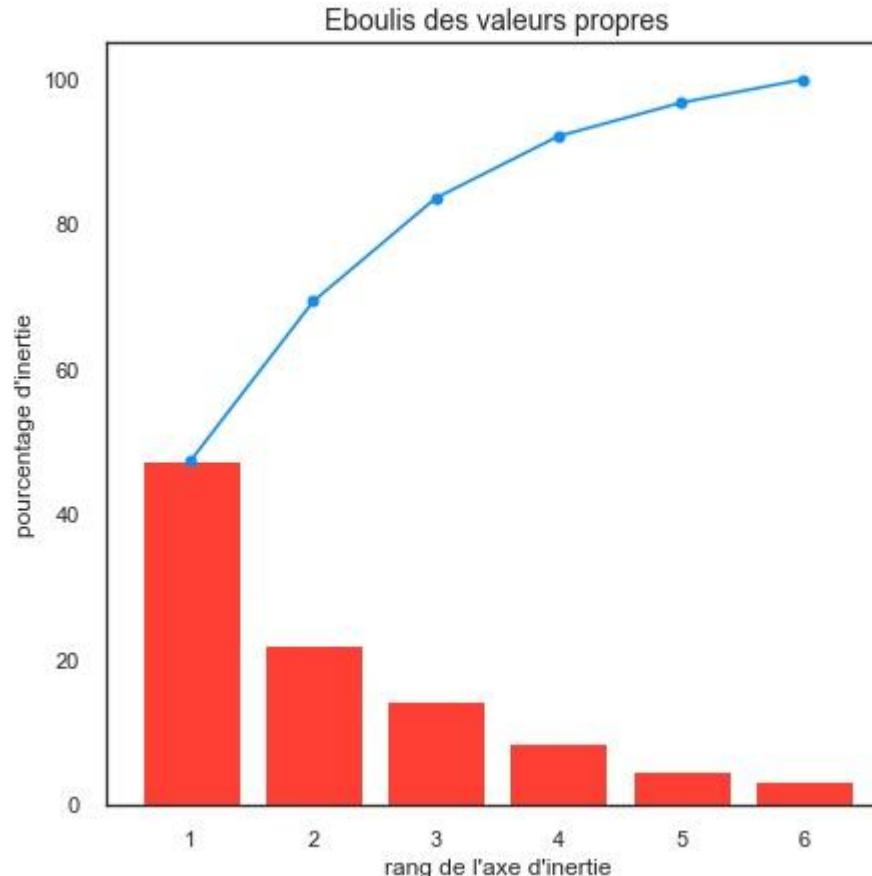
Les groupes sont distincts



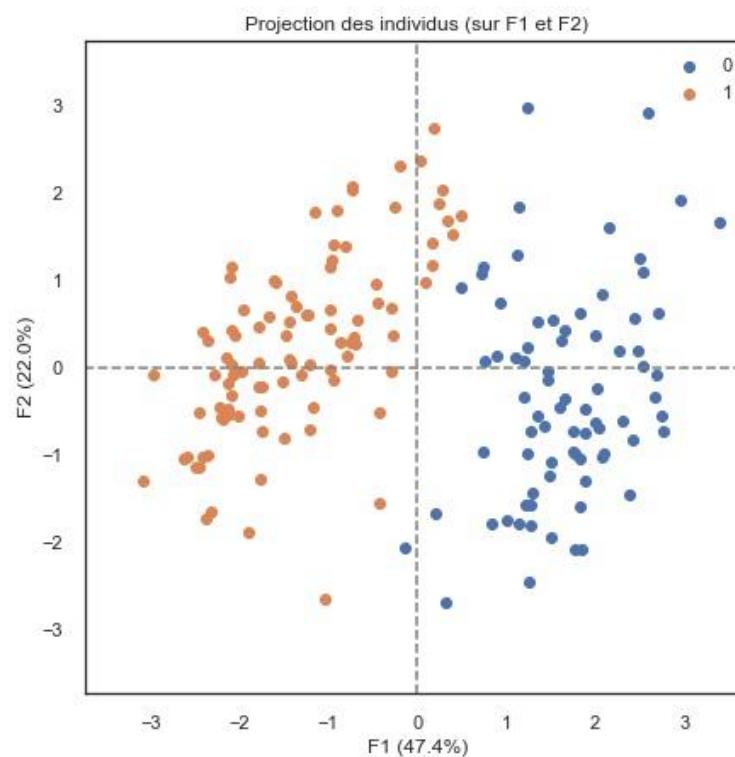
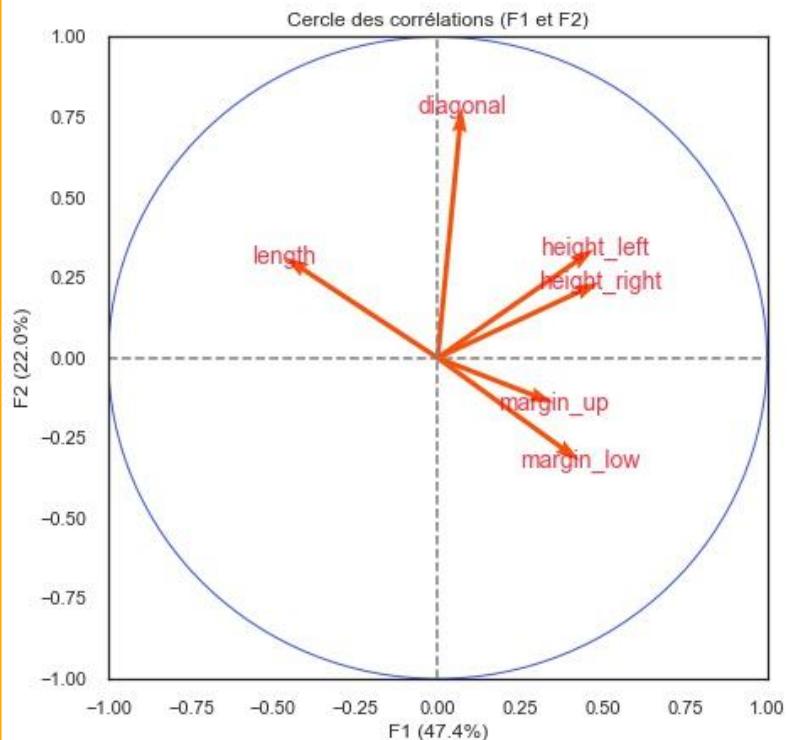
ACP de la partition
Sur le premier plan factoriel

Eboulis des valeurs propres

Les 2 premières composantes suffisent à expliquer 69.4 % de l'inertie totale



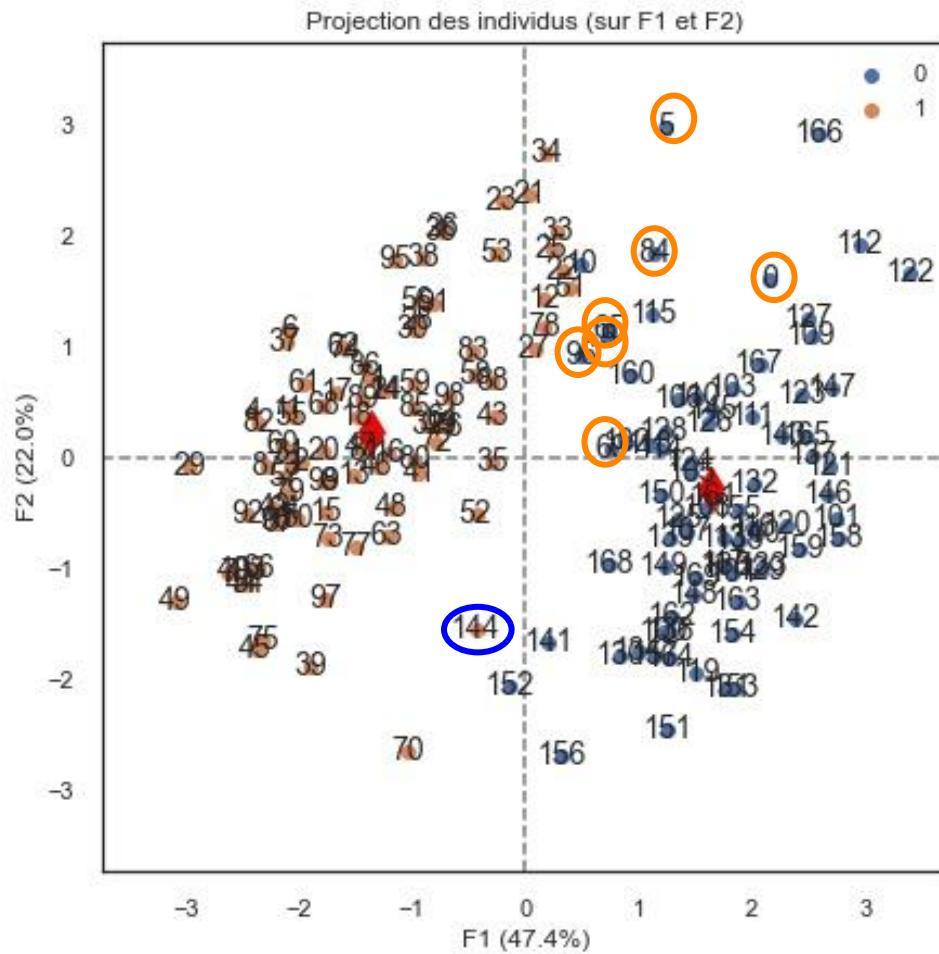
Cercle de corrélation et projection des individus sur F1 et F2



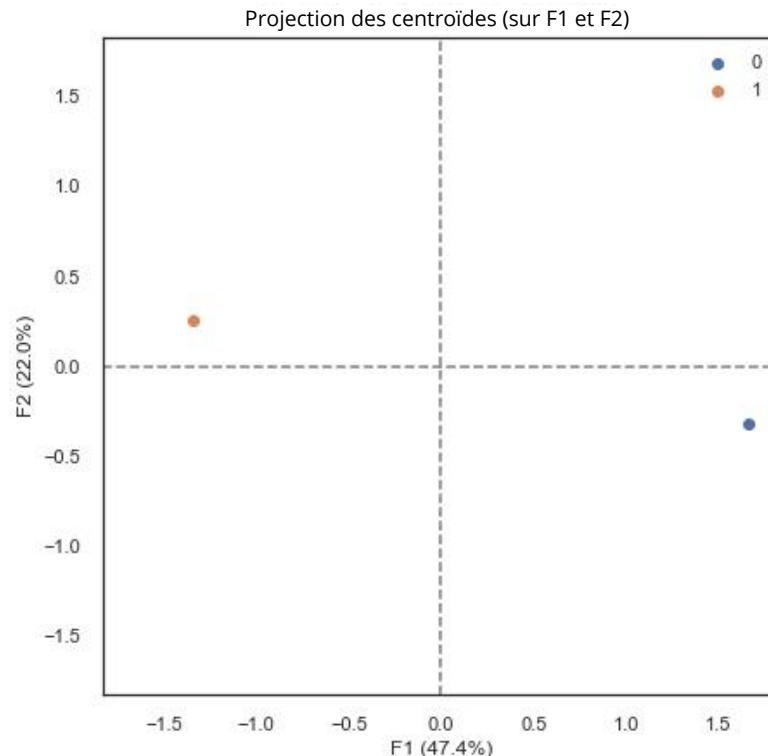
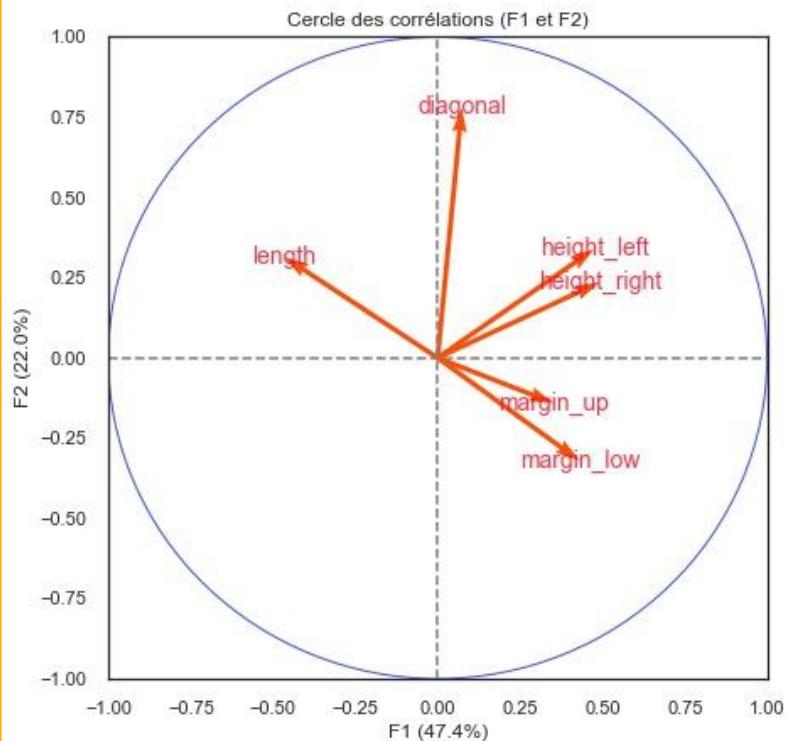
8 individus mal classés

Billet 144 : billet faux
classé vrai

Billets vrais classés
dans les faux
0, 5, 9, 65, 69, 84, 96



Cercle de corrélation et projection des centroïdes sur F1 et F2





Interprétation ACP

Variables les plus corrélées à F1

height_left

height_right

margin_low

margin_up

Length

Variable la plus corrélée à F2

diagonal

F1

+0.07 * Diagonal

+0.47 * height_left

+0.49 * height_right

+0.43 * margin_low

+0.35 * margin_up

-0.46 * length

F2

+0.77 * diagonal

+0.33 * height_left

+0.23 * height_right

-0.32 * margin_low

-0.14 * margin_up

+0.31 * length

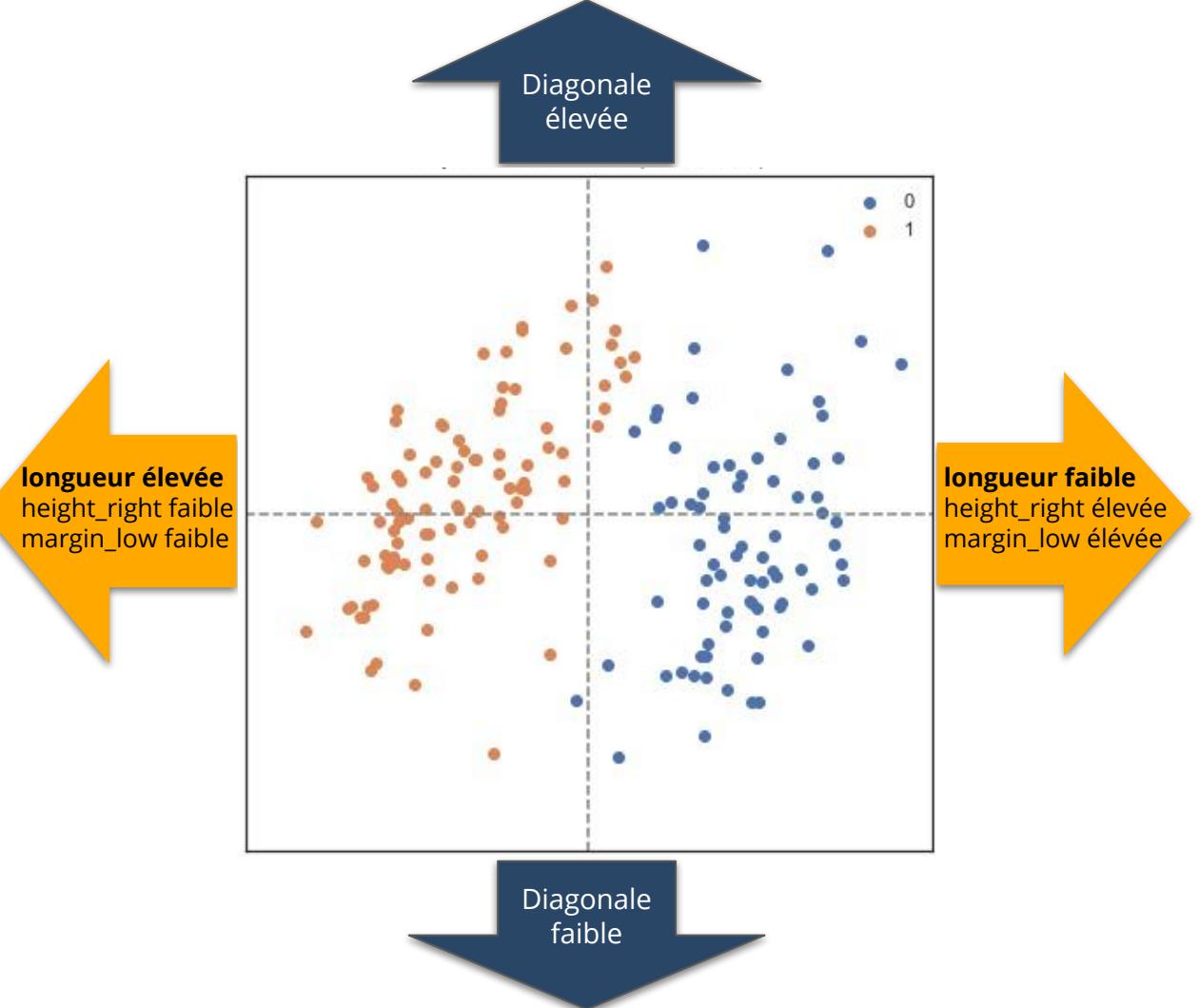
Pour résumer

Billet vrai :

Length plus grande

Margin_low plus faible

Height plus faible



Mission 3

Modélisation des données à l'aide d'une
régression logistique

Régression logistique (classification supervisée)

Qu'est ce que c'est ?

Modèle statistique permettant d'étudier les relations entre des variables explicatives (catégorielles ou continues) et une variable dépendante qualitative (type binaire)

Permet de

Prédire la **probabilité** qu'un évènement survienne

Survenu de l'évènement

Quand la probabilité est supérieure à un seuil (seuil de classification 0.5)

Fonction logistique (logit)

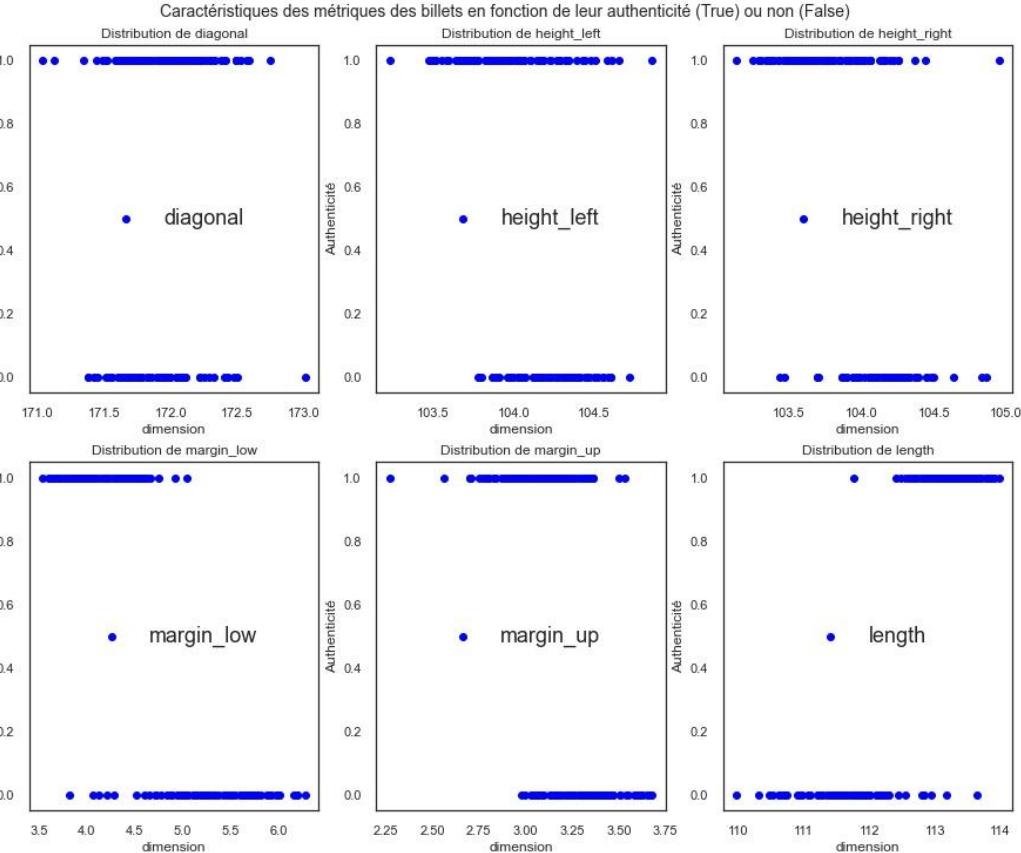
Fonction reliant la probabilité aux variables explicatives



Caractéristiques des métriques en fonction de l'authenticité

Variables significatives
margin_low - length

Variable non significative
Diagonal



Création du modèle logistique

Variables retenues

Is_genuine - margin_low - length
- height_right

Représentent un résumé de la
métrique d'un billet

Standardisation des données

Uniquement sur les données
d'entraînement

Données test

Normalisées en utilisant la
moyenne et l'écart type calculés
sur les données d'entraînement

Echantillons

80 % de données d'entraînement
20 % de données test



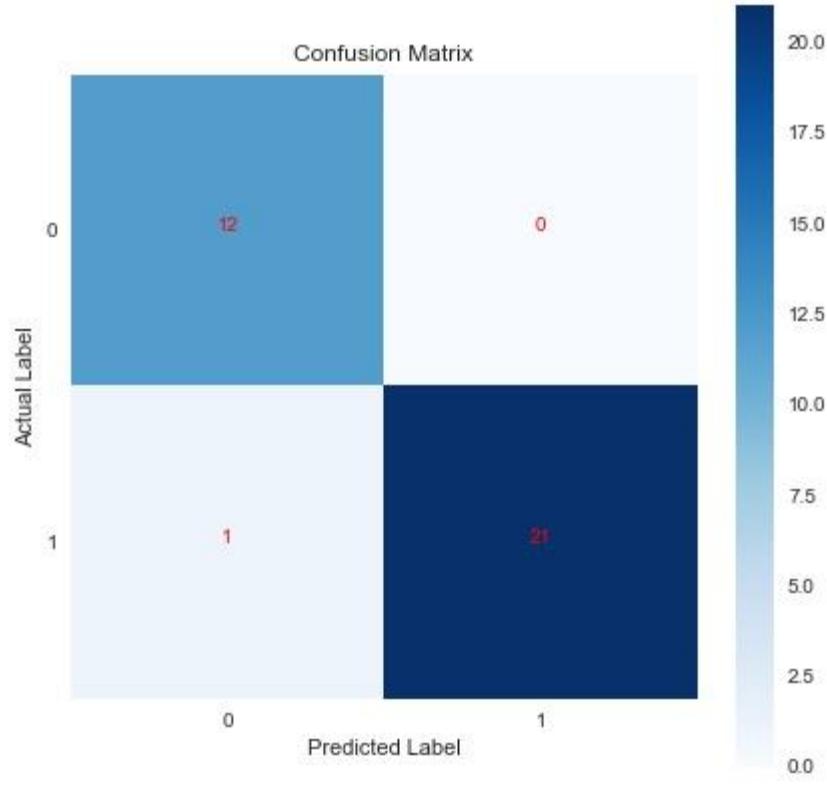
Matrice de confusion

1 billet vrai prédit en faux
0 billet faux prédit en vrai

Sensibilité (taux de positifs classés positifs)
 $VP / (VP+FN) = 0.95$

Spécificité (taux de négatifs classés négatifs)
 $VN / (VN + FP) = 1$

Modèle **fidèle à**
97,8 % pour les données entraînement
97.1 % pour les données test



Influence des variables

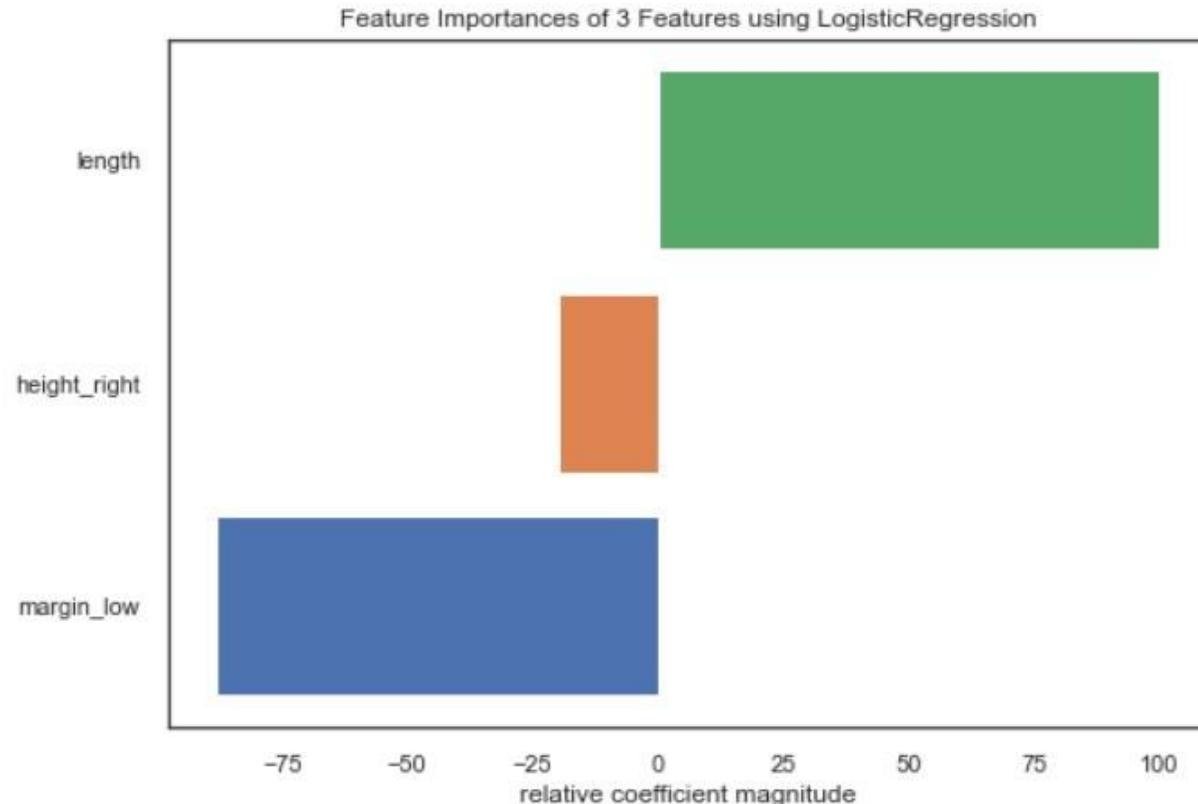
length et margin_low
sont les variables les plus
influentes du modèle

Score :

Length : 2.52

Margin_low : -2.22

Height_right : -0.50



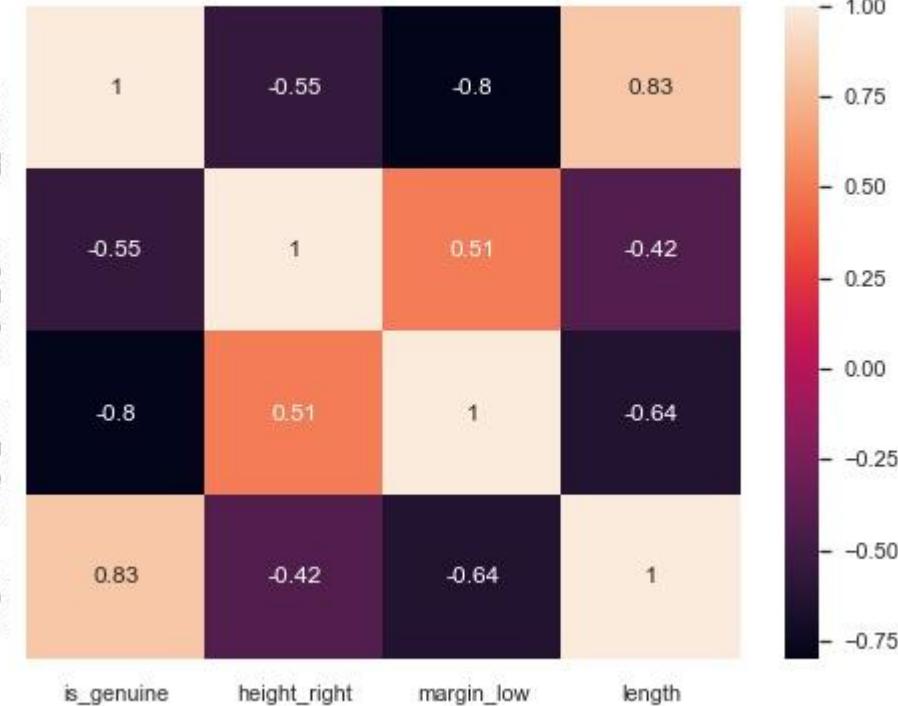
Corrélation des variables

is_genuine est bien corrélée avec les variables choisies

Corrélation de margin_low et length

$R^2 = -0.64$

Tableau de corrélation des variables les plus influentes

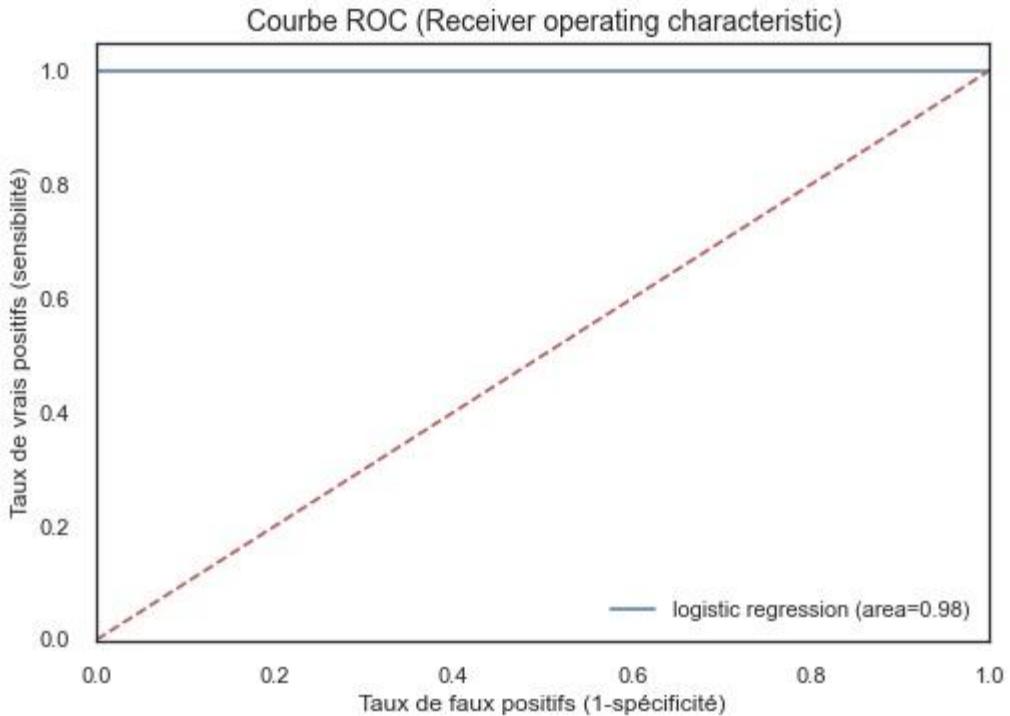


Courbe de ROC (Receiver Operating Characteristic)

AUC : mesure de la qualité de la classification

AUC = 0.5 : dans le pire des cas

AUC = 1 : dans le meilleur des cas



A man with glasses and a white shirt is sitting at a desk, looking down at a computer screen. He is in an office environment with other desks and equipment visible in the background.

Merci!

Des questions ?

Contactez - moi

www.linkedin.com/in/isabelle-barbier