Projet N°10
Détectez des faux billet

Openclassroom

Sommaire

01

Le contexte

ONCFM

Organisation nationale de lutte contre le faux-monnayage

Notre but :

- Mettre en place des méthodes d'identification des contrefaçons des billets en euro

02

Les données

Aperçu et premières observations.

Les variables

LENGHT :
Longueur

MARGIN UP:
Marge haute

DIAGONAL :
diagonal

MARGIN LOW:
Marge basse

HEIGHT RIGHT :
Hauteur droite

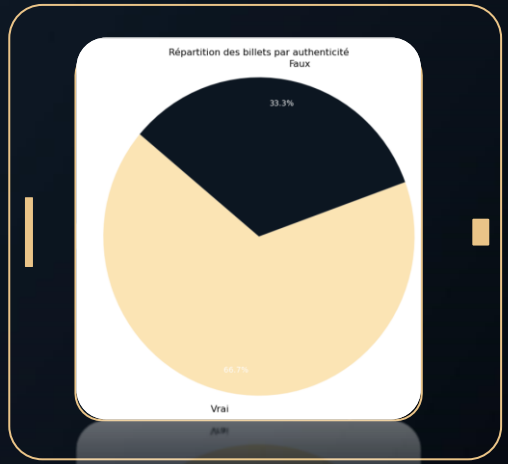
HEIGHT LEFT :
Hauteur gauche

Is Genuine

visualisation

- Nous avons :
  - 1000 vrai billets
  - 500 faux billets

Ce qui représente 66,7% de vrai VS 33,3 % de faux

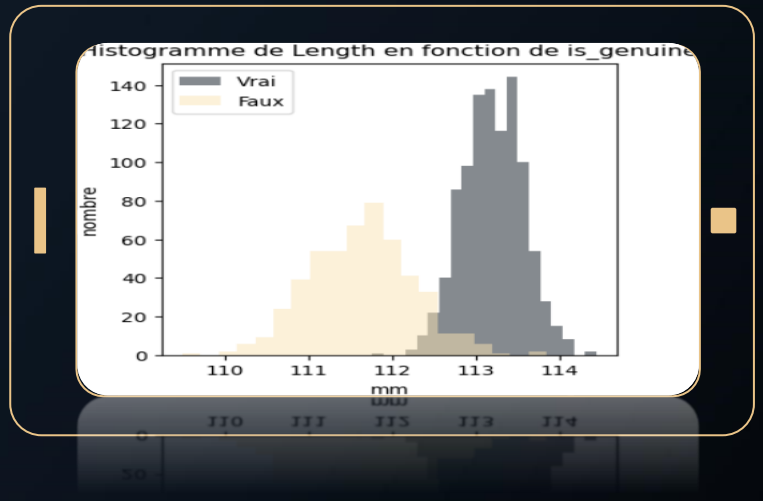


Length

visualisation

- Vrai billets :
  - Maximum : 114,4 mm
  - Moyenne : 113,2 mm
  - Minimum : 111,7 mm

- Faux billets :
  - Maximum : 113,8 mm
  - Moyenne : 111,6 mm
  - Minimum : 109,4 mm

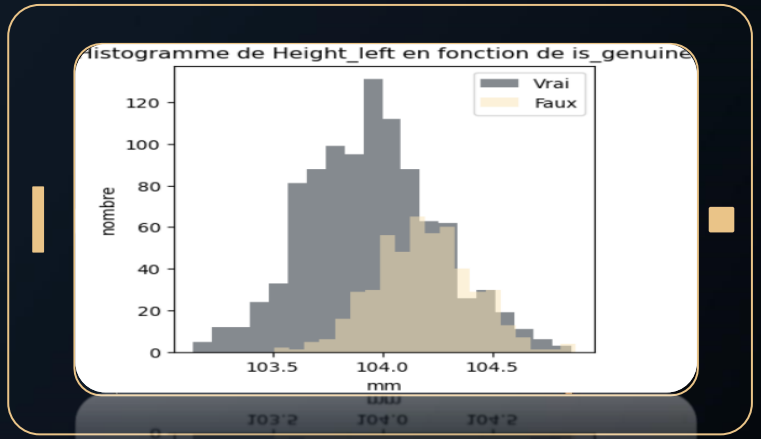


Height Left

visualisation

- Vrai billets :
  - Maximum : 104,8 mm
  - Moyenne : 103,9 mm
  - Minimum : 103,1 mm

- Faux billets :
  - Maximum : 104,8 mm
  - Moyenne : 104,1 mm
  - Minimum : 103,5 mm

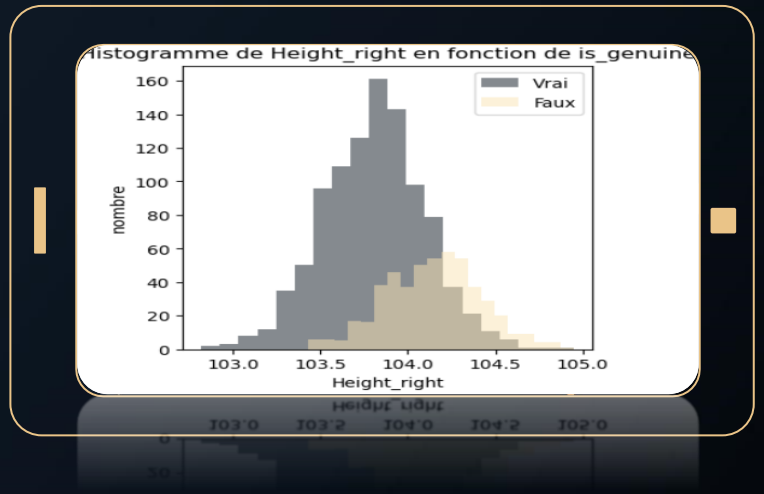


Height Right

visualisation

- Vrai billets :
  - Maximum : 104,9 mm
  - Moyenne : 103,8 mm
  - Minimum : 102,8 mm

- Faux billets :
  - Maximum : 104,9 mm
  - Moyenne : 104,1 mm
  - Minimum : 103,4 mm

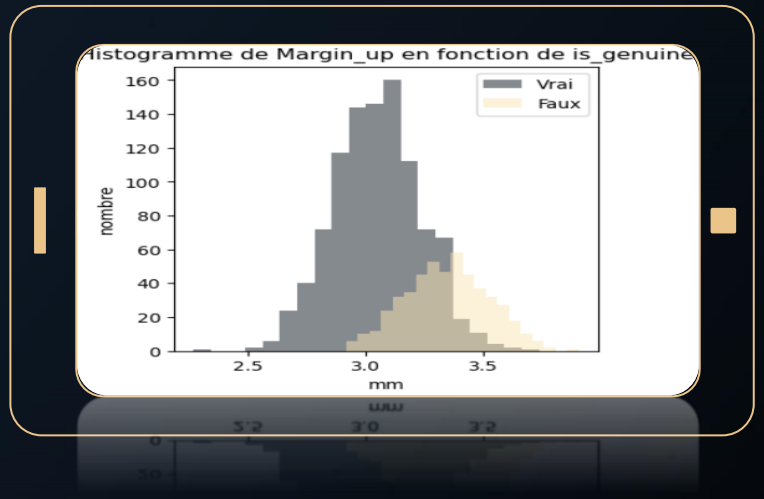


Margin Up

visualisation

- Vrai billets :
  - Maximum : 3,7 mm
  - Moyenne : 3 mm
  - Minimum : 2,3 mm

- Faux billets :
  - Maximum : 3,9 mm
  - Moyenne : 3,3 mm
  - Minimum : 2,9 mm

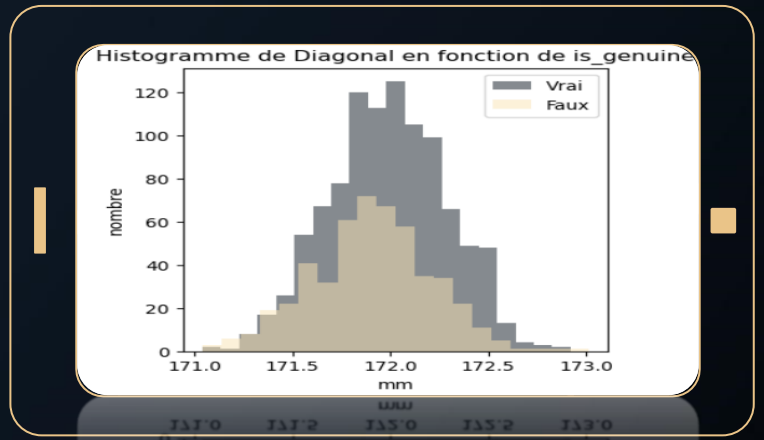


Diagonal

visualisation

- Vrai billets :
  - Maximum : 172,9 mm
  - Moyenne : 171,9 mm
  - Minimum : 171 mm

- Faux billets :
  - Maximum : 173 mm
  - Moyenne : 171,9 mm
  - Minimum : 171 mm



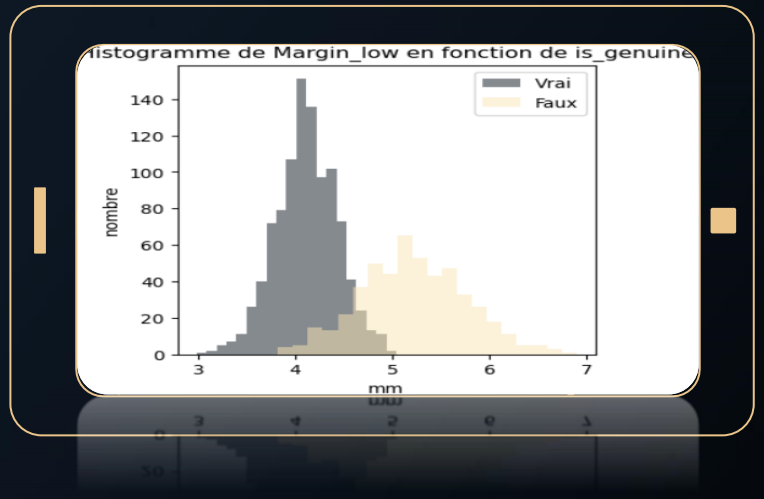
Margin Low

visualisation

- Vrai billets :
  - Maximum : 5 mm
  - Moyenne : 4,1 mm
  - Minimum : 3mm

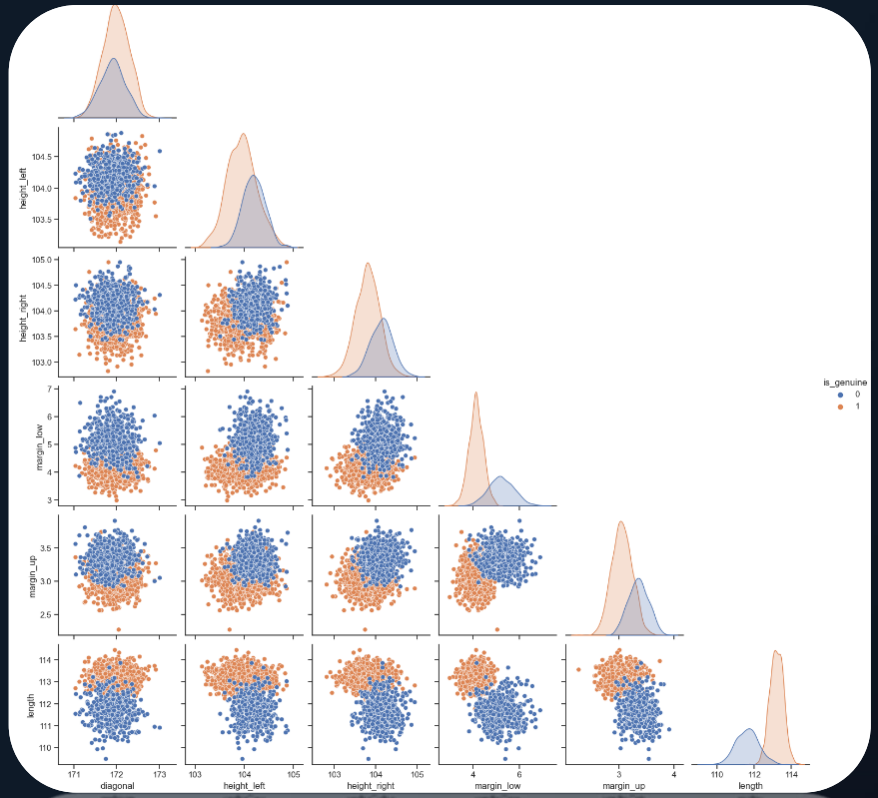
- Faux billets :
  - Maximum : 6,9 mm
  - Moyenne : 5,2 mm
  - Minimum : 3,8 mm

Nous avons 37 données manquante



Récapitulatif

visualisation



Forte dépendance des variables
Margin_low
Length

Ces variables joueront sans doute un rôle important plus tard

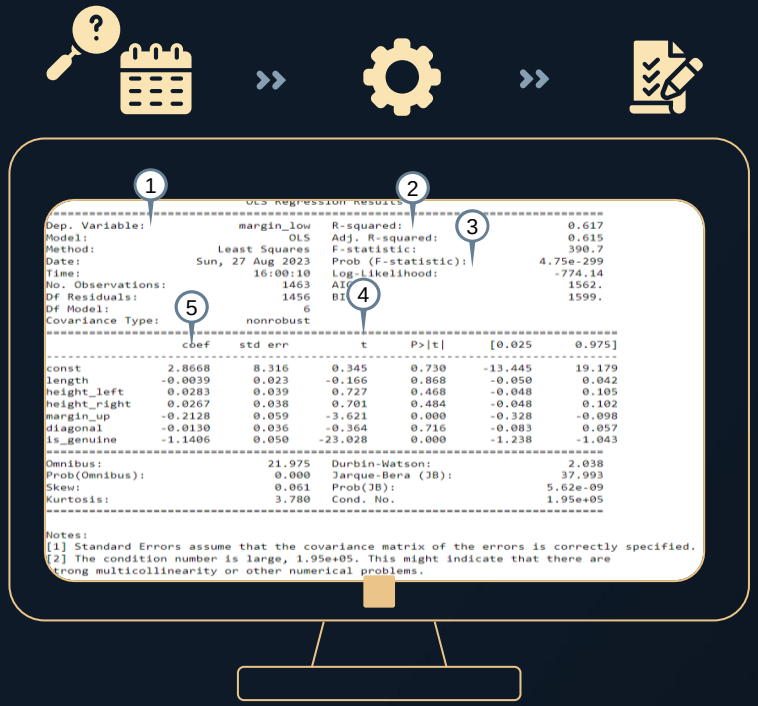
03

Prédiction

imputation des valeurs manquantes par régression linéaire

Régression linéaire

Explication



Plusieurs valeurs non significative P-value > 5 % nous allons les retirer.

1. La variable à expliquer : margin_low
2. R^2 l'estimation du modèle
   + élevé = meilleurs estimation du modèle
   Mieux vaut regarde le r^2 ajuster
3. La statistique F significativité du modelé
   Si f élever et prob faible = significatif. À travers ca P-value mesure la significative global du modèle. Son hypothèse nul est la nullité de l'ensemble des coef (B1=B2.. = 0) vs au moins des coef est différent de 0.
4. La statistique t mesure la significativité d'une variable.
   Seuil à 0,05
5. Coef : Le poids de chaque variable dans le modèle.
   Coef positif = augmenterons la variable et inversement avec le négatif

Test d’hypothèse

évalue la validité du modèle

Test de colinéarité

Le test de multicollinéarité vérifie si il y a une multicollinéarité entre les variables indépendantes.

Test de VIF:
Les coefficients sont inférieur à 5, il n'y a donc pas de problème de colinéarité

Test de normalité des résidus

Ce test vérifie si les résidus **(les erreurs entre les valeurs prédites par le modèle et les valeurs observées)** suivent une distribution normale.

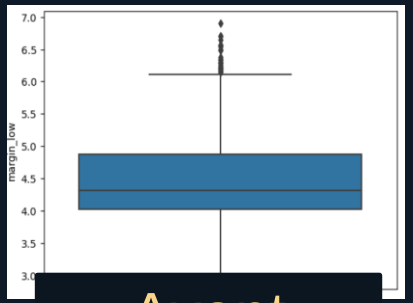
Shapiro test :
P-value < 5% et donc rejet de l'hypothèse H0.
Les résidus ne suivent pas une distribution normal

Test d'homoscédasticité

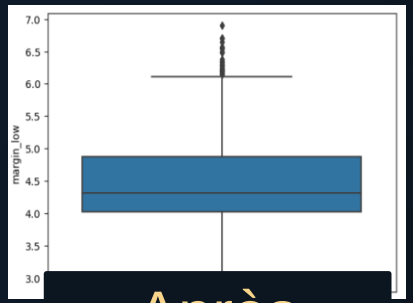
Le test d'homoscédasticité n'est pas respecter.

Verification

aprés imputation des valeurs manquantes



Avant



Après

04

Modélisation

par la méthode de la régression logistique et KMeans

Régression logistique

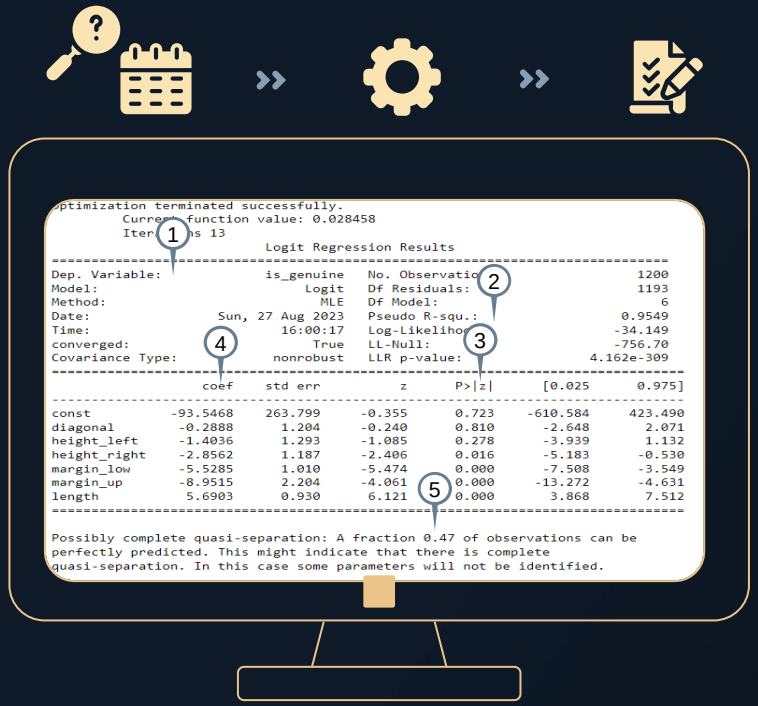
explication

La régression logistique est utilisée pour prédire une classe binaire (0 ou 1) ou variable qualitative en fonction de variables indépendantes.
Pour valider la modélisation, nous allons utiliser un schéma de validation simple en découpant notre dt initial en un jeu de données d'entrainement et de test (Train_test_ split) :

- Limiter le sur-apprentissage
- Mesurer la performance du modèle sur un jeu de donnée test

Regression logistique

Explication



Plusieurs valeurs non significative P-value > 5 % nous allons les retirer.

1. La variable expliquer : is_genuine
2. Pseudo R^2 l'estimation du modèle
   + élevé = meilleurs estimation du modèle
3. P-Value : aide à décider si une variable à un effet significatif sur la variable dépendante.

   Seuil à 0,05
4. Coef : Le coef de chaque variable dans le modèle.

   Coef positif augmente la probabilité que le billet soit vrai vs diminue le taux de vrai billet en cas de négatif
5. Message d'avertissement

   Signifie que 47% des prédictions sont prédite de façon quasi parfaite(0,99%)

Test d’hypothèse

évalue la qualité et la pertinence

Test de colinéarité

Le test de multicollinéarité vérifie la corrélation élevée entre les variables indépendantes.

Test de VIF:
les coefficients sont inférieur à 5, il n'y a donc pas de problème de colinéarité

La variable est Binaire/quali

Oui, is_genuine est bien qualitatif ou binaire (vrai/faux)

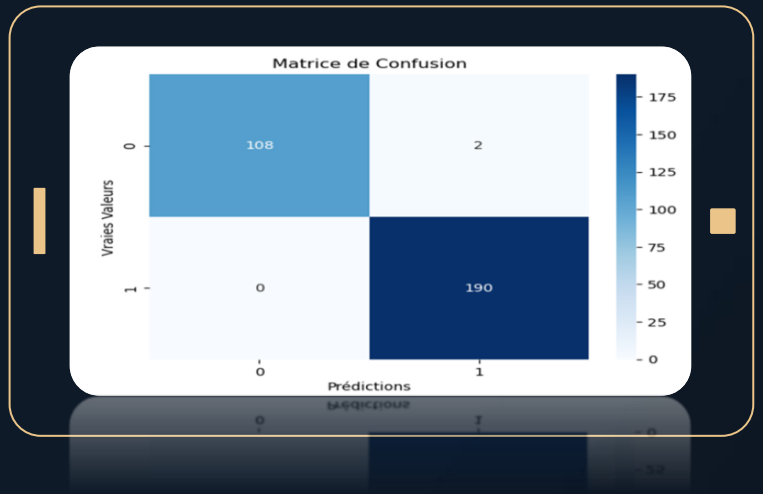
Taille et indépendance de l'échantillon

On souhait avoir plus de 500 observation indépendantes et plus de 20 observations pour l'outcome.

Les billets sont indépendants des autres et sont uniques

Matrice de confusion

visualisation des erreurs en test (20%)

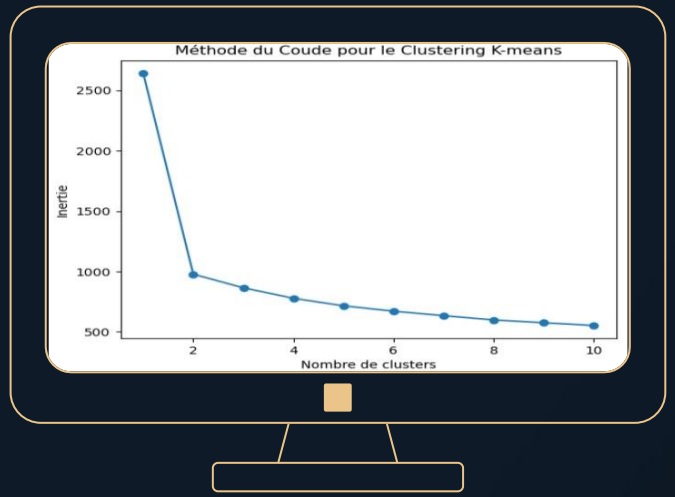


Nous permet de percevoir les erreurs du modèle.
2 faux qui ont était prédit vrai

Précision du modèle : 99%

La classification

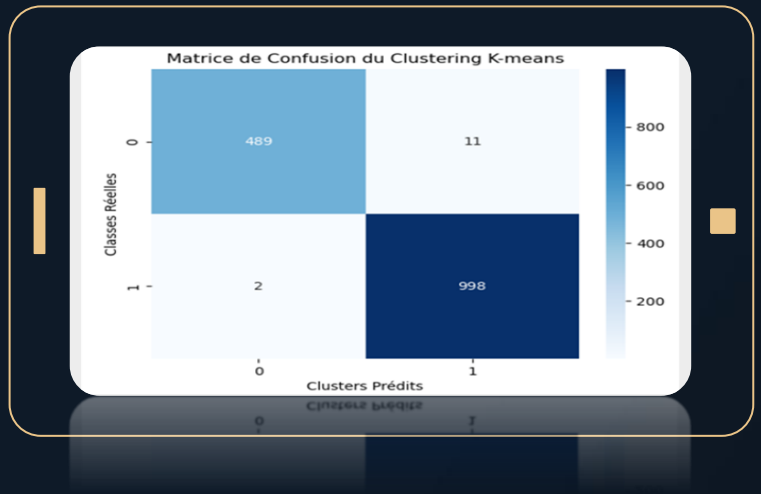
par la méthode du Kmeans (non-supervisé)



- Nombre de cluster optimal de 2
- Chaque billets sera classifié dans un groupe en les confrontant à la variable is_genuine

Matrice de confusion

visualisation des erreurs



Nous permet de percevoir les erreurs du modèle.
11 faux qui ont était prédit vrai
2 vrai qui ont était prédit faux

Précision du modèle : 99%

Il ne devrait pas trouver des groupes vrais faux aussi détaillé, à cause des données montré précédemment (groupe bien visible) méthode non superviser

05
Conclusion

R

La méthode de la régression logistique
99% de score
peu d'erreur dans la matrice de confusion

K

La méthode du KMeans:
99% de score
Quelque erreurs dans la matrice

Mais ne devrais pas pouvoir définir si parfaitement les groupes