Identification de faux billets

ONCFM

Introduction

Contexte : L'Organisation nationale de lutte contre le faux-monnayage (ONCFM) nous a sollicité afin de mettre en place une modélisation qui serait capable d'identifier automatiquement les vrais des faux billets, à partir de certaines dimensions des billets.

I. Analyse descriptive
II. Traitement des valeurs manquantes
    1. Preprocessing
    2. Régression linéaire multiple
III. Algorithmes
    1. Régression logistique
    2. Kmeans

Analyses descriptives

Présentation du jeu de données :

Le jeu de données est composé de 1500 individus et de 7 variables.

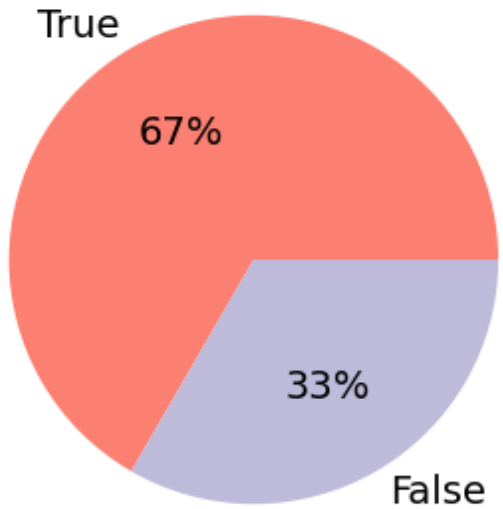
Ces variables sont constituées d'une variable binaire, 'is_genuine', qui nous renseigne sur l'authenticité des billets et de 6 variables quantitatives, précisant leurs dimensions (en mm) :

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

Vérification des données manquantes :

La variable **margin_low** est la seule variable ayant des données manquantes.

Répartition des billets vrais et faux



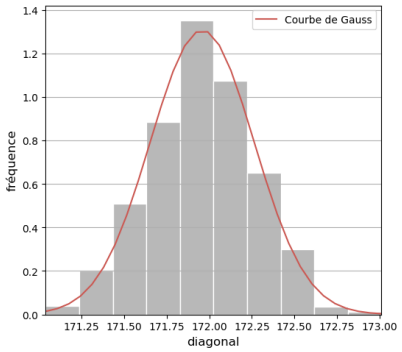
Dans notre jeu de données, 1/3 des billets sont faux.

Analyses descriptives

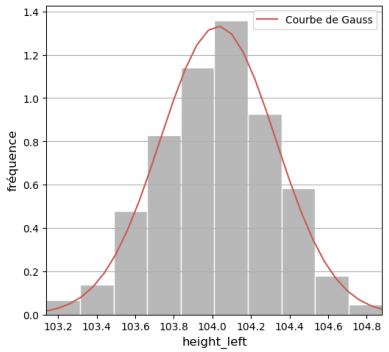
Vérification de la normalité des variables quantitatives

Distribution

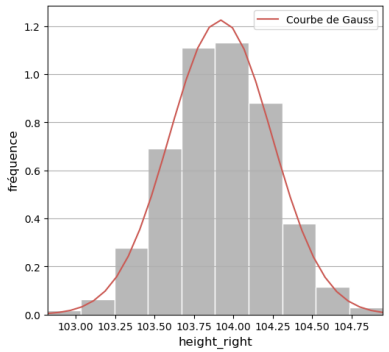
diagonal



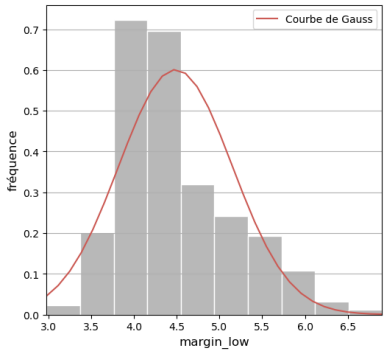
height_left



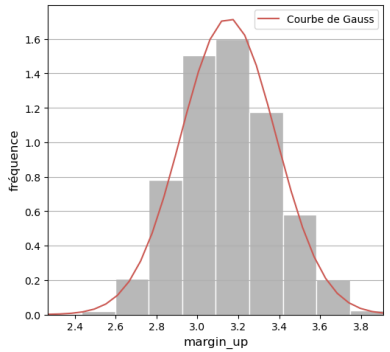
height_right



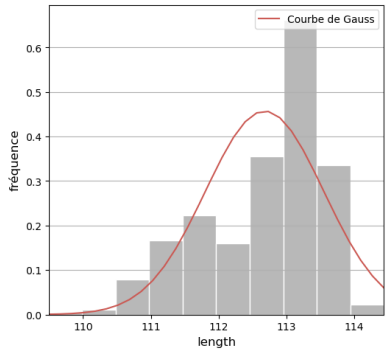
margin_low



margin_up



lenght



Test de Shapiro

Hypothèse : la distribution suit-elle une loi normale ?
H0 : la variable suit une loi normale
H1 : la variable ne suit pas une loi normale

Les variables **diagonal, height_left et height_right** ont une pvalue > 5%

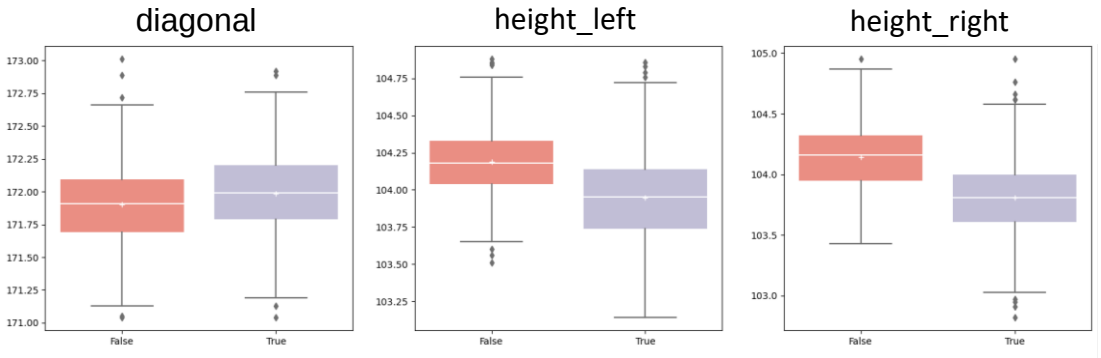
- Ces variables suivent une loi normale

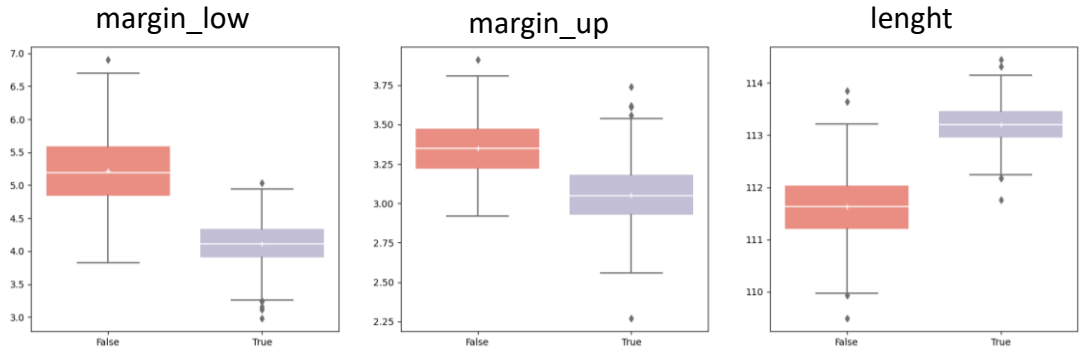
Les variables **margin_low, margin_up, lenght** ont une pvalue < 5%

- Ces variables ne suivent pas une loi normale

Analyses descriptives

- **Comparaison de la distribution des données**





Hypothèse : Y a-t-il une différence entre les billets faux et les vrais ?
H0 : Il n'y a pas de différence significative entre les 2 groupes
H1 : Il y a une différence significative entre les 2 groupes

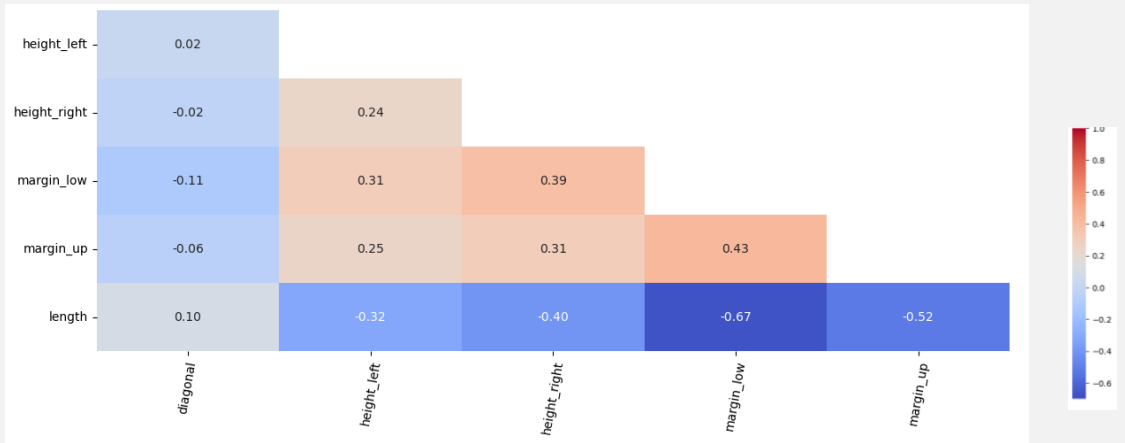
diagonal, height_left, height_right : Test t (Student) pvalue < 5%

- Il y a une différence significative entre les vrais et faux billets

margin_low, margin_up, lenght : *Test Wilcoxon-Mann-Whitney* pvalue < 5%

- Il y a une différence significative entre les vrais et faux billets

- **Matrice de corrélation**



Traitement des valeurs manquantes

➢ Il manque 37 valeurs dans la variable margin_low

- **Préparation des données (preprocessing)**
  - Création d'un jeu de données d'entrainement et d'un jeu de données test :
    - Le train set (jeu d'entrainement) est composé de :
      - 80% du jeu de données (sans les individus ayant une valeur manquante)
      - Soit 1170 lignes
    - le test set(jeu de test) est composé de :
      - 20% du jeu de données
      - Soit 293 lignes
  - Standardisation des données

- **Régression linéaire multiple**

Hypothèses d'application de la régression linéaire :

- Absence de multi-colinéarité entre les variables explicatives
- Homoscédasticité des résidus
- Distribution normale des résidus
- Indépendance des résidus

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.484
Model:	OLS	Adj. R-squared:	0.482
Method:	Least Squares	F-statistic:	218.8
Date:	Sat, 10 Dec 2022	Prob (F-statistic):	1.22e-164
Time:	22:27:51	Log-Likelihood:	-1272.6
No. Observations:	1170	AIC:	2557.
Df Residuals:	1164	BIC:	2588.
Df Model:	5		
Covariance Type:	nonrobust		

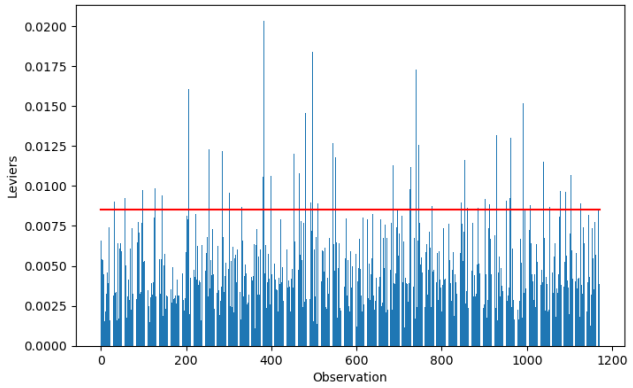
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.025e-15	0.021	-1.44e-13	1.000	-0.041	0.041
diagonal	-0.0463	0.021	-2.181	0.029	-0.088	-0.005
height_left	0.0860	0.023	3.822	0.000	0.042	0.130
height_right	0.1239	0.023	5.274	0.000	0.078	0.170
margin_up	0.0940	0.025	3.746	0.000	0.045	0.143
length	-0.5376	0.027	-20.124	0.000	-0.590	-0.485

Omnibus:	51.079	Durbin-Watson:	2.018
Prob(Omnibus):	0.000	Jarque-Bera (JB):	60.873
Skew:	0.467	Prob(JB):	6.05e-14
Kurtosis:	3.614	Cond. No.	2.16

Régression linéaire multiple

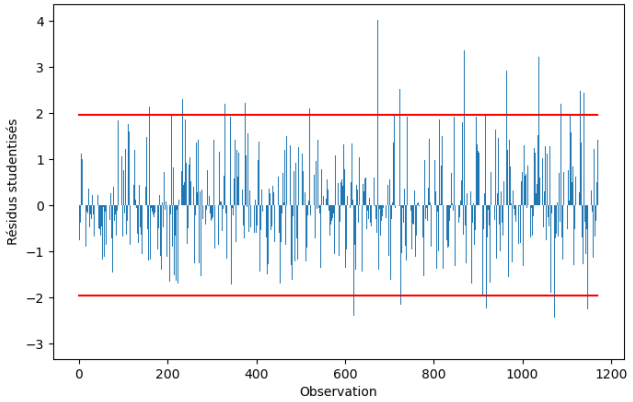
Analyses des valeurs atypiques et influentes

Atypicité sur les variables explicatives
(les leviers)



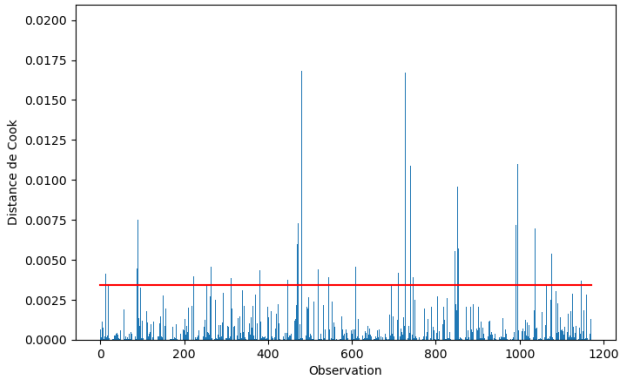
124 observations

Atypicité sur la variable à expliquer
(les résidus studentisés)



62 observations

Influence des observations
(distance de cook)



68 observations

➢ Eviction des valeurs atypiques et influentes

OLS Regression Results

Dep. Variable:	margin_low	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.546
Method:	Least Squares	F-statistic:	267.5
Date:	Mon, 26 Dec 2022	Prob (F-statistic):	2.11e-187
Time:	15:39:40	Log-Likelihood:	-1041.2
No. Observations:	1110	AIC:	2094.
Df Residuals:	1104	BIC:	2124.
Df Model:	5		
Covariance Type:	nonrobust		

➢ L'éviction des 60 observations atypiques et influentes a permis d'améliorer le coefficient de détermination $R2$ et le $R2$ ajusté qui valent maintenant 0,55.

Régression linéaire multiple

- **Vérification de la multi-colinéarité des variables**

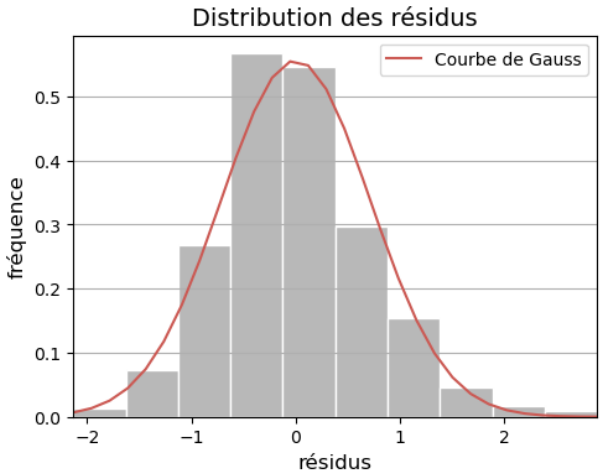
	VIF
diagonal	1.016418
height_left	1.143632
height_right	1.245960
margin_up	1.422592
length	1.611214

La non-colinéarité est vérifiée puisque les valeurs VIF sont inférieures à 2.

- **Indépendance des résidus**

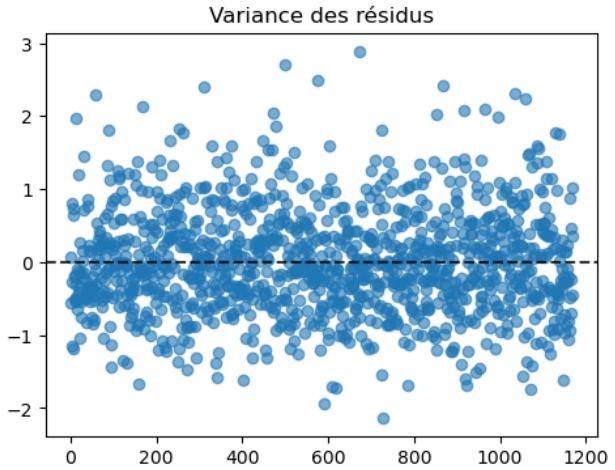
➢ Statistique de Durbin-Watson : 1.99

- **Vérification de la normalité des résidus**



Test de Shapiro
pvalue < 5%
➢ On rejette l'hypothèse nulle.
Les résidus ne suivent pas une loi normale.

- **Vérification de l'homoscédasticité**



Test Breusch Pagan
pvalue < 5%
➢ On rejette H0.
Les résidus ne sont pas homoscédastiques.

- **Evaluation du modèle**

L'écart quadratique moyen sur le train set est de 0.5 et sur le test set est de 0.51

➢ L'équivalence des écarts quadratiques moyens nous permet de valider la stabilité de ce modèle.

Régression logistique

- **Preprocessing**
  - Création d'un jeu de données d'entrainement 70% (1050 individus) et d'un jeu de données test 30% (450 individus)
  - Standardisation des données
- **Régression logistique**

hypothèses :

- La variable à expliquer doit être classée en 2 catégories.
- Les variables explicatives ne doivent pas avoir de multi-colinéarité.
- L'échantillon doit être de grande taille.

- **Vérification de la colinéarité des variables**

	VIF
diagonal	1.020868
height_left	1.158638
height_right	1.260927

	VIF
margin_low	2.029857
margin_up	1.365642
length	2.161742

- **Entrainement du modèle et évaluation de la performance**
  - Matrice de confusion

	Billet prédit Faux	Billet prédit Vrai
Billet Faux	133	4
Billet Vrai	1	312

  - Mesures

Précision : 0.987
Rappel : 0.997
Spécificité : 0.971
Accuracy score : 0.989

- **Sélection de variables significatives**

Variables significatives : 'margin_low', 'margin_up' et 'length'.

- **Réentrainement et évaluation de la performance**
  - Matrice de confusion

	Billet prédit Faux	Billet prédit Vrai
Billet Faux	133	4
Billet Vrai	0	313

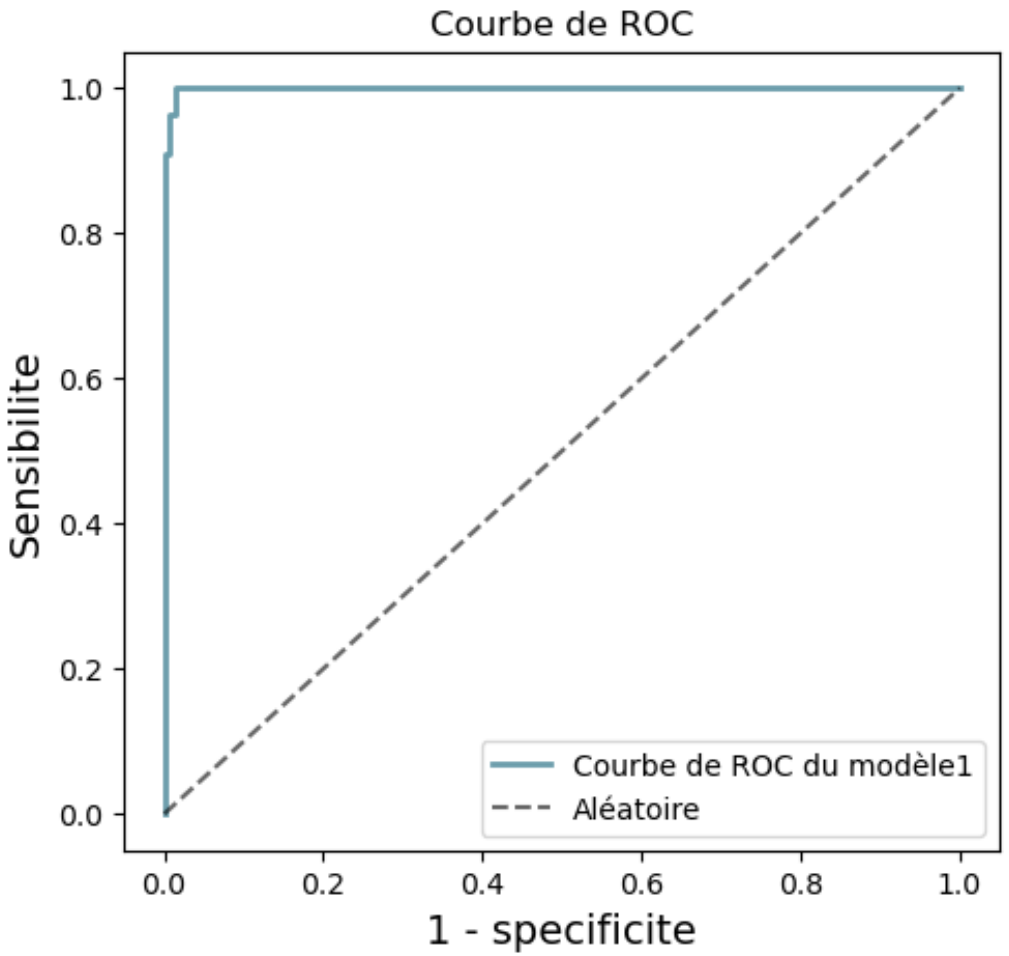
  - Mesures

Précision : 0.987
Rappel : 1
Spécificité : 0.971
Accuracy score : 0.991

➢ On obtient une performance légèrement meilleure avec seulement 3 variables

Régression logistique

Courbe de ROC



Calcul de l'aire sous la courbe

Air sous la courbe ROC : 0.999

- Le calcul de l'air sous la courbe ROC nous confirme que le modèle est performant.

Cross validation

Moyenne de précision : 0.99
Moyenne d'accuracy score : 0.989

Mesures du test set

Précision : 0.987
Accuracy score : 0.991

- La précision et le taux de bonnes prédictions sur le test set est équivalent aux moyennes de la cross validation, on peut donc en conclure qu'il n' y a pas surapprentissage sur ce modèle qui par ailleurs est performant

K-means

- **Choix du nombre de clusters : 2**
- **Entrainement de l'algorithme**
- **Evaluation**

Train set

	Billet prédit Faux	Billet prédit Vrai
Billet Faux	356	7
Billet Vrai	9	678

Précision : 0.99
Rappel : 0.987
Spécificité : 0.981
Accuracy score : 0.985

Test set

	Billet prédit Faux	Billet prédit Vrai
Billet Faux	131	6
Billet Vrai	2	311

Précision : 0.981
Rappel : 0.994
Spécificité : 0.956
Accuracy score : 0.982

➢ Les mesures de performance sur le test set sont équivalentes à celles du train set. Ainsi avec un taux de bonnes prédiction de 98%, ce modèle est performant.

Conclusion

Comparaison des 2 algorithmes

Regression logistique

- Précision : 0.987
- Rappel : 1.0
- Spécificité : 0.971
- Accuracy score : 0.991

K-means

- Précision : 0.981
- Rappel : 0.994
- Spécificité : 0.956
- Accuracy score : 0.982

- Les 2 algorithmes fournissent une très bonne performance. En les comparant, on peut observer que le taux de bonnes prédictions est meilleur avec la régression logistique, ainsi que la précision, qui pour rappel, est la mesure que l'on souhaite maximiser afin de mieux détecter les billets faux.

➢ C'est pourquoi la régression logistique est retenu.

✓ *Test de l'algorithme avec un jeu de données inconnu*

MERCI !