

Création d'un algorithme détectant les faux billets

CLADIERE Nathan, projet 6 , Formation Data Analyst OC

Sommaire


- Contexte et objectifs
- Statistiques descriptives
- Analyse exploratoire (ACP)
- Classification non-supervisé : clustering avec K-means
- Classification supervisée: modélisation par régression logistique

Contexte et objectifs

- Data de billets vrai et faux
- Trouver un algorithme détectant les futurs faux billets:
 - Statistiques descriptives des billets
 - Analyse exploratoire de données (ACP)
 - Réalisation d'une classification non-supervisé (K-means)
 - Réalisation d'une classification supervisé (regression logistique)

Contexte et objectifs : données

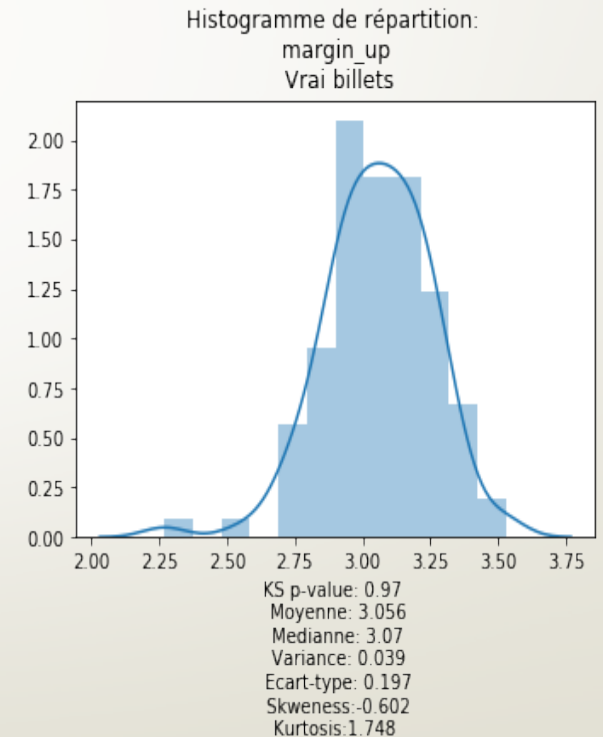
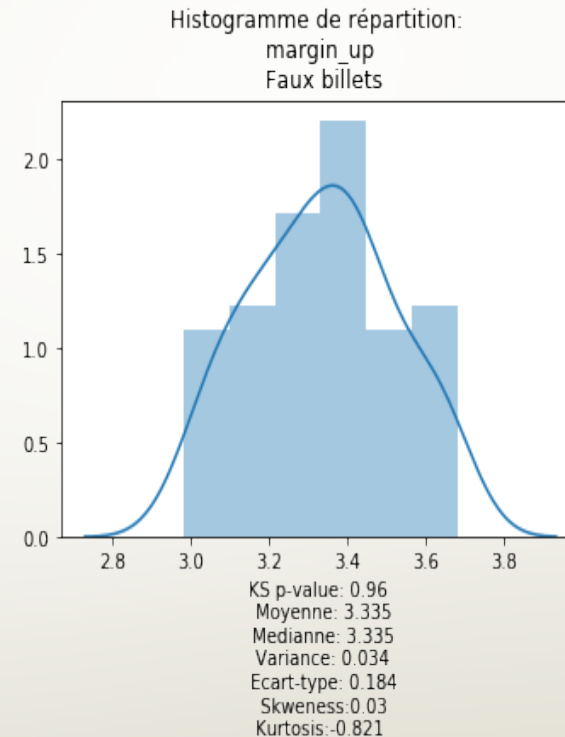
- Caractéristiques des billets
 - Longueur du billet (length)
 - Hauteur du billet à gauche (height_left)
 - Hauteur du billet à droite (height_right)
 - Marge bord supérieur/image du billet (margin_up)
 - Marge bord inférieur/image du billet (margin_low)
 - Diagonale (diagonal)
 - Billet vrai ou faux (is_genuine)
- Data prep: check données manquantes



M0: Statistiques descriptives des caractéristiques des billets

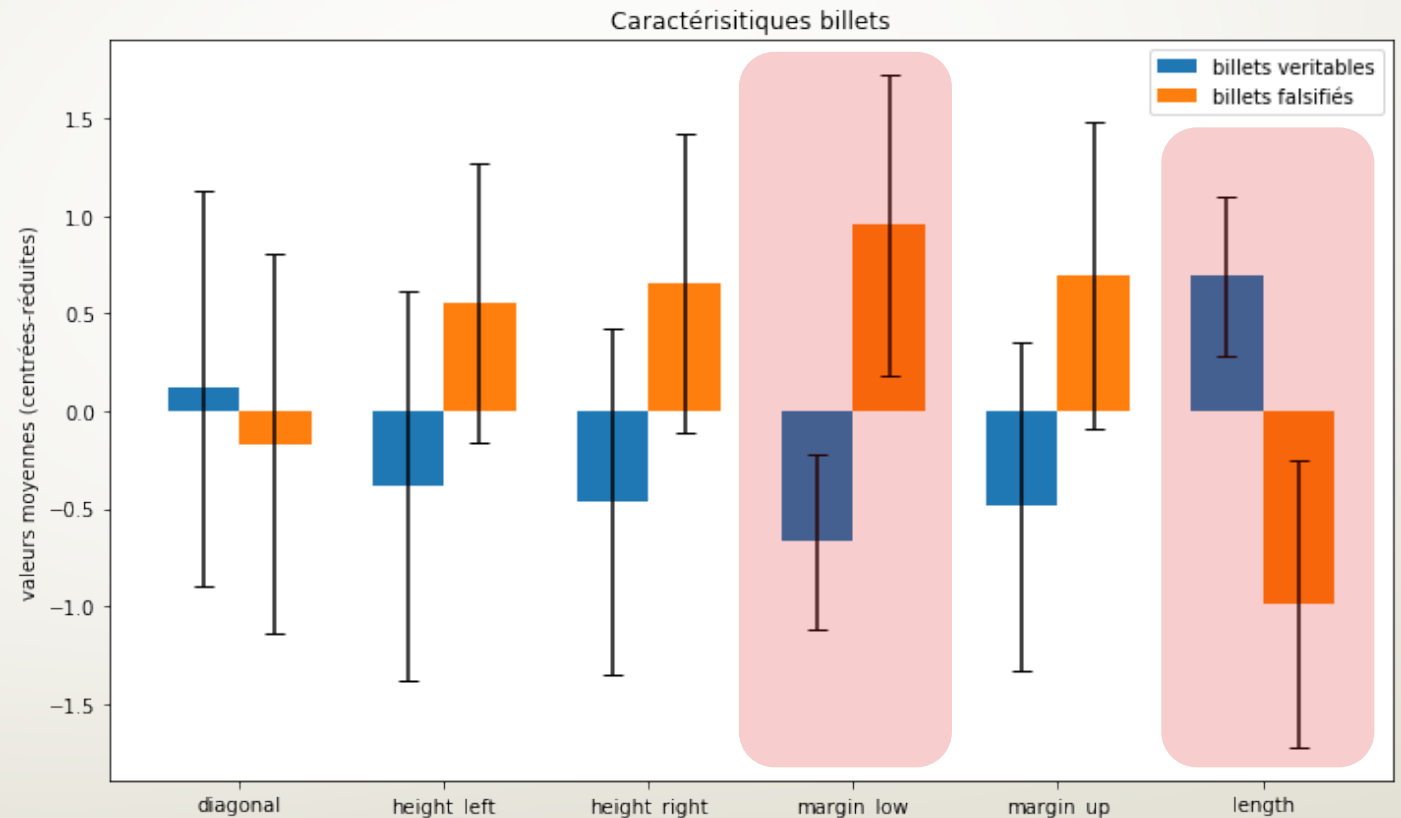
M0: lois de distribution et mesures de forme

- Normalité : Test Kolmogorov Smirnov, Pvalue $>0,05$
- Apparence de l'histogramme:
 - Kurtosis et Skewness varient beaucoup
 - Allure générale de l'histogramme de répartition : courbe Gaussienne



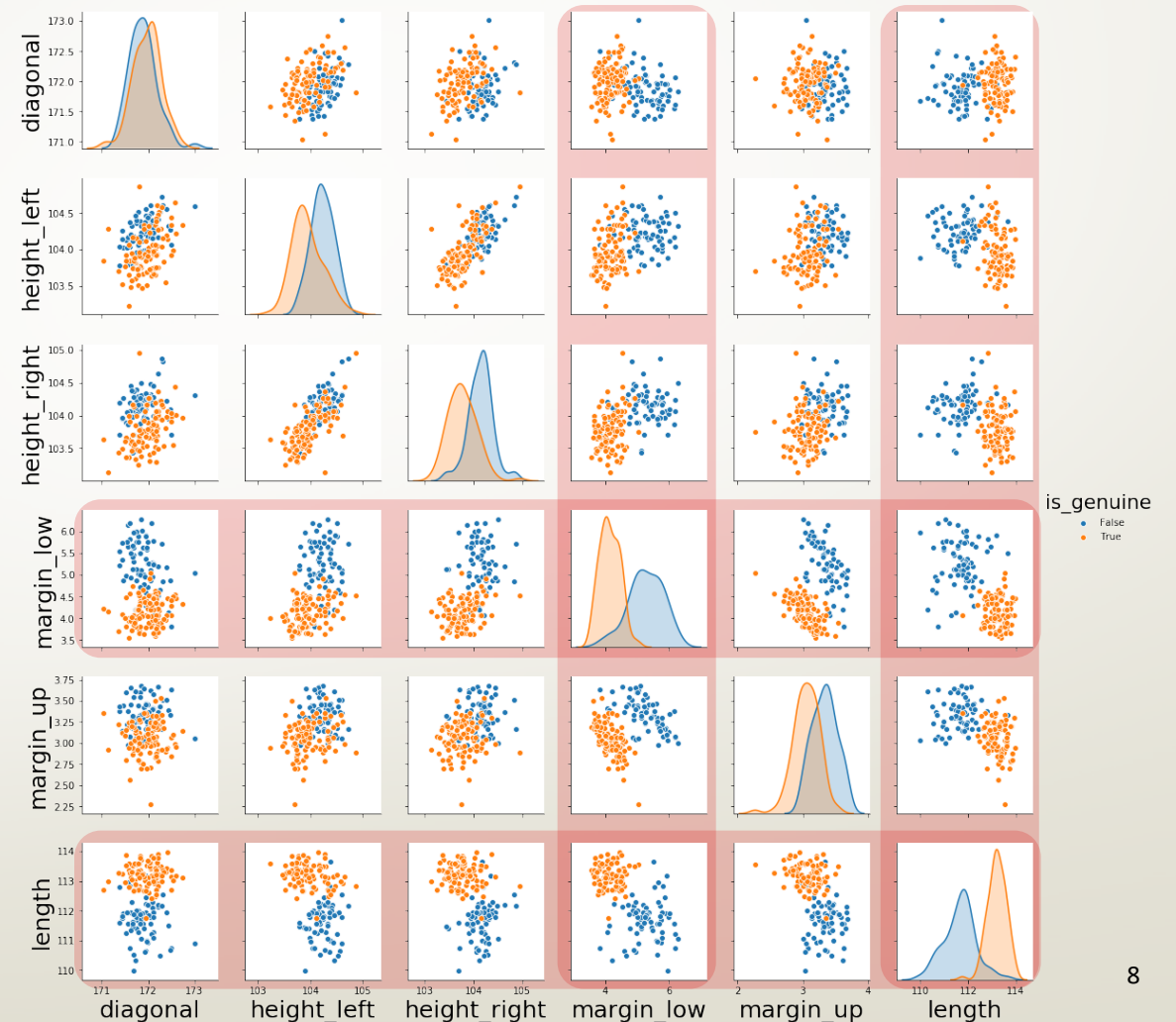
M0 : Moyennes et écart-types déjà révélateurs

- Mesure centrée-réduite
- Moyenne différente pour chaque caractéristiques
- Ecart-type moins grand :
 - marge_low
 - length



M0: analyses bivariés illustrant les différences

- Différence confuse
 - Diagonal
 - Heights
 - Marge_up
- Différence semble significative :
 - Margin_low
 - Length



M0: Conclusion

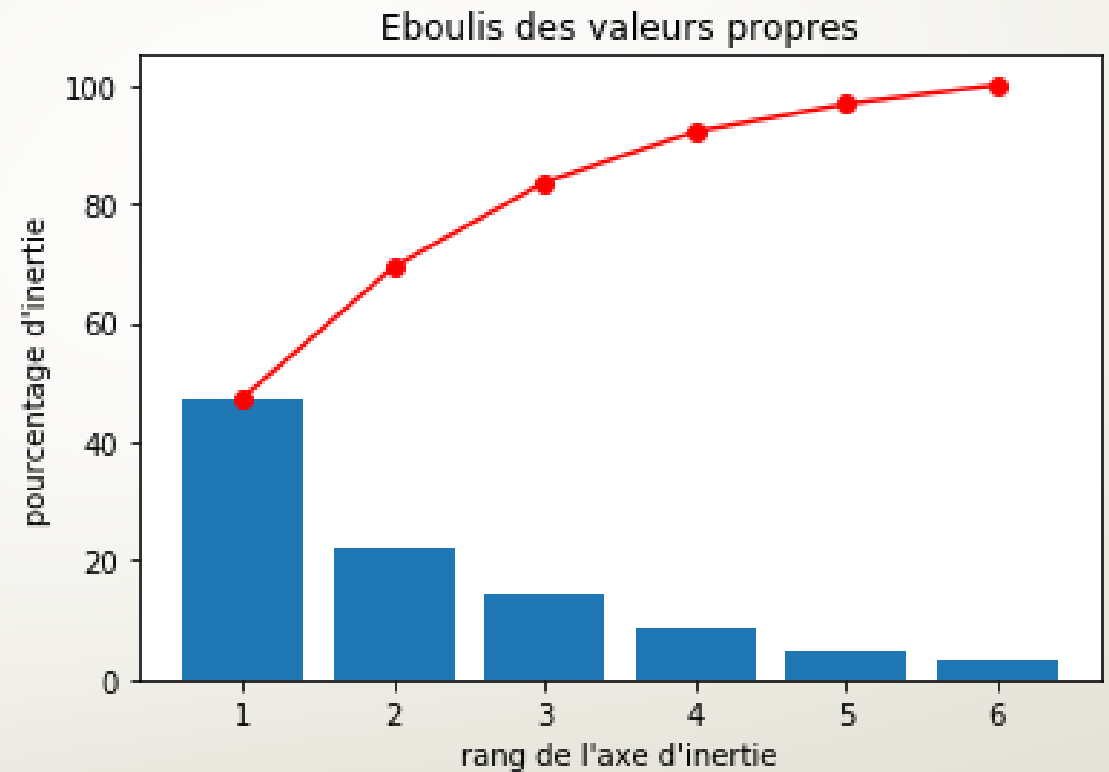
- Différences plus marqués sur ces deux caractéristiques :
 - Margin_low
 - Length
- Caractéristiques semblent illustrer la différences entre les billets
- Détermination de l'influence des caractéristiques via une ACP



M1 : Analyse en Composantes principales

M1: Détermination du nombre de composantes

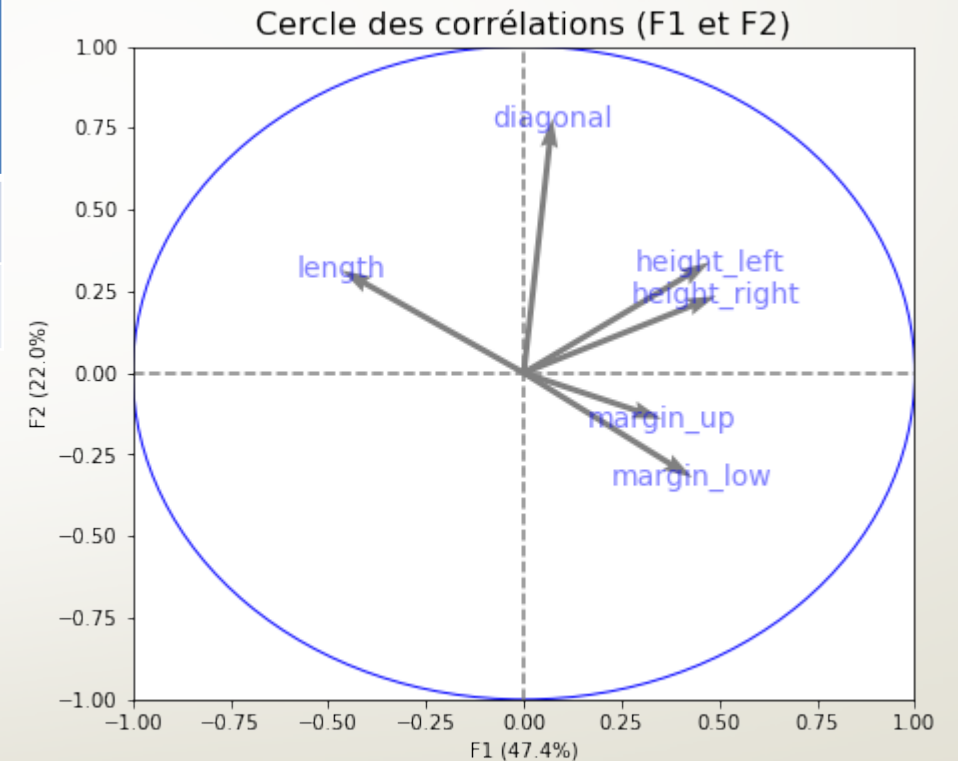
- Premier plan factoriels : 69 %
 - F1 : 47 %
 - F2 : 22 %
- Suffisant pour voir l'influence des différentes caractéristiques



M1: interprétation du cercle des corrélations

	diagonal	height_left	height_right	margin_low	margin_up	length
F1	0.07	0.48	0.49	0.43	0.35	-0.47
F2	0.78	0.34	0.24	-0.32	-0.14	0.31

- F1
 - Positivement : Hauteurs, Margins
 - Négativement : length
- F2: Diagonal



M1: Qualité et contribution des individus

Contributions: carrés des distance à l'origine

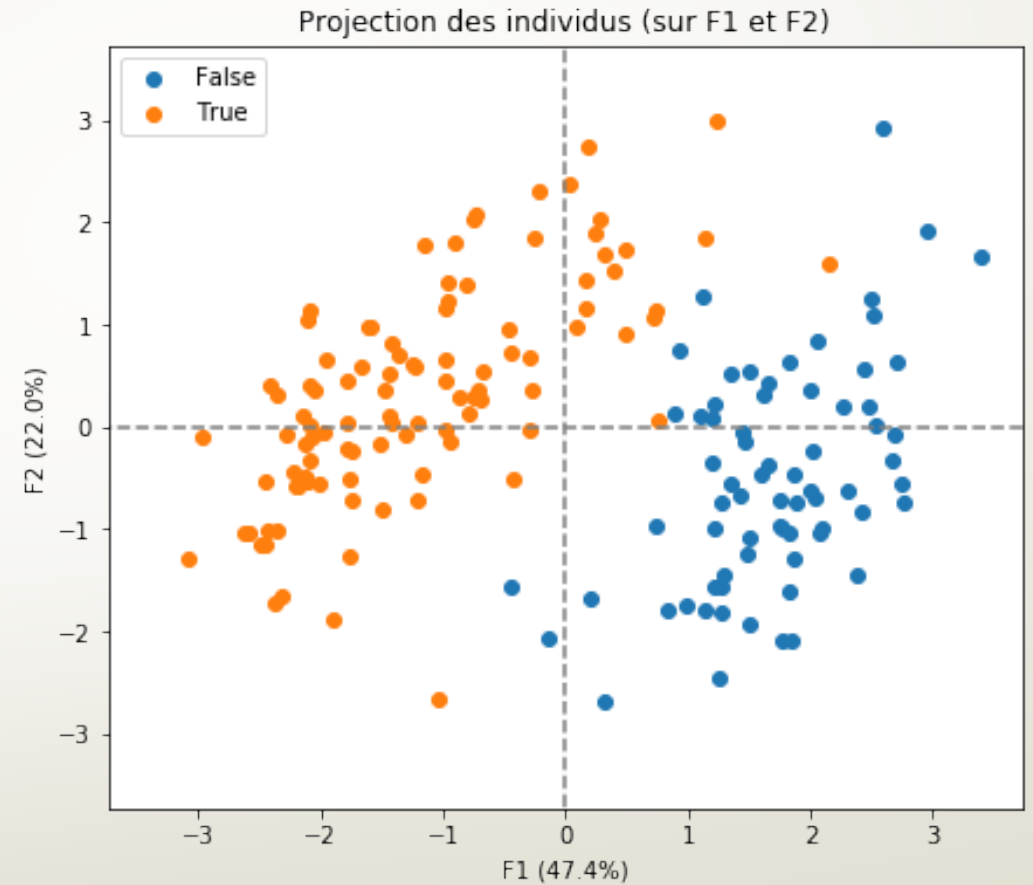
- Inertie du nuage de points provenant des vrais et faux (en moyenne)
 - Vrai billets : 5,3
 - Faux billets: 6,8
- Valable aussi pour les 20 valeurs les plus hautes (entre 20 et 9)
 - Nombre de billets vrai :10
 - Nombre de billets Faux

Qualités : \cos^2

- Bonne représentation sur l'axe F1 (en moyenne)
 - Vrais billets : 0,45
 - Faux billets: 0,49
- Représentation moins marquée sur F2(en moyenne)
 - Vrais billets : 0,25
 - Faux billets: 0,21


M1: Projections des individus

- Pas de différence sur la diagonale (F2)
- Influence sur F1 bien plus marqué
- Spécialement sur l'axe margin_low / length



M1: conclusion

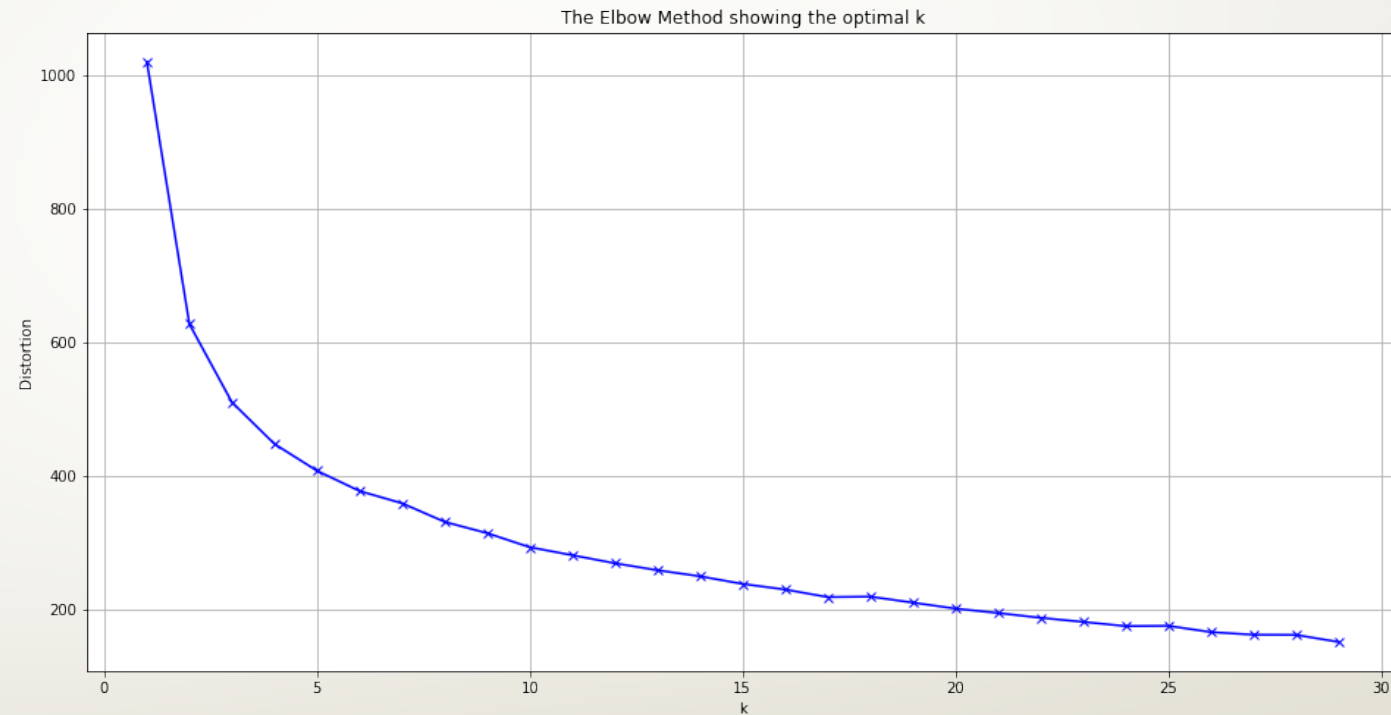
- Observation de caractéristiques plus influentes :
 - Margin_low
 - Length
- Quelques individus difficiles à caractériser avec des caractéristiques proches
- Est-ce qu'une classification non-supervisée classe bien les billets ?



M2 : Classification non-supervisé (K-means)

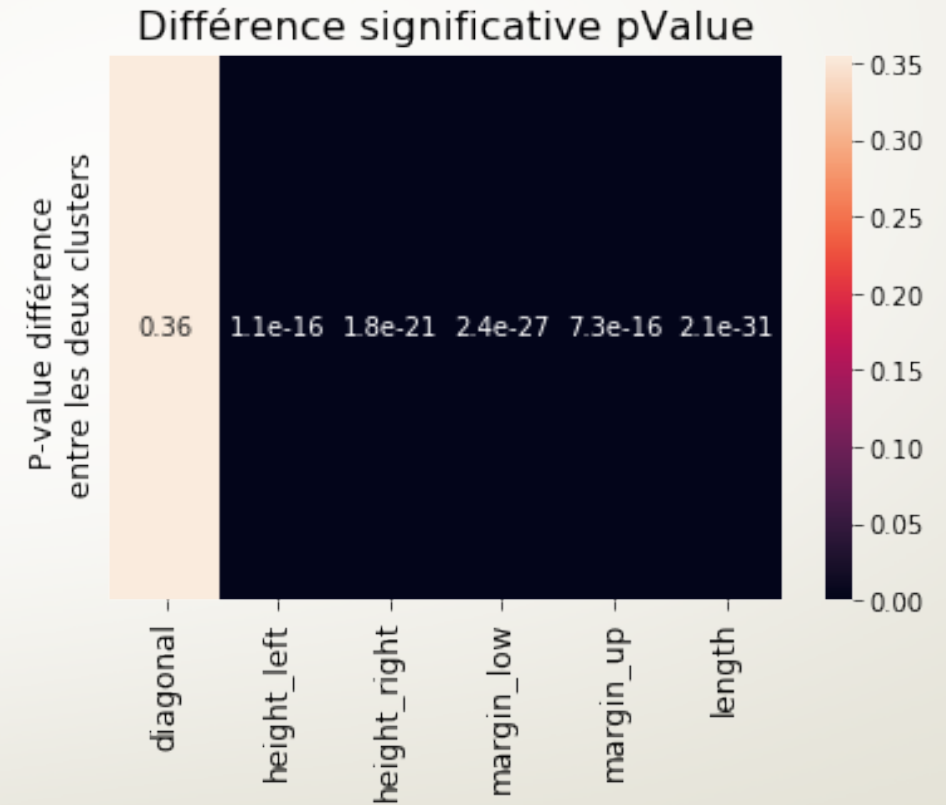
M2: choix du nombre de clusters

- Graphique de distorsion
méthode du coude : 5 clusters
- Contexte vrai /faux billets : 2 Clusters



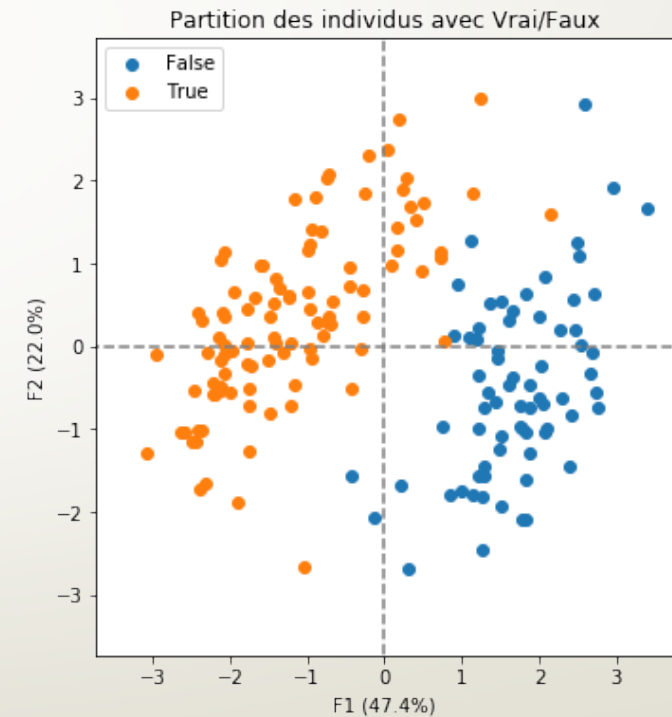
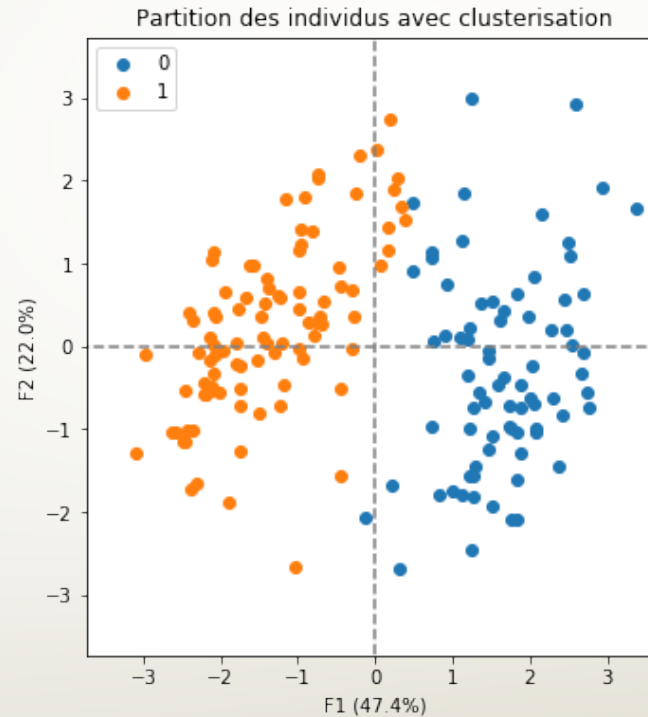
M2:différence significative entre clusters

- Méthode
 - Test d' égalité des variance (Fisher)
 - Tests statistiques (Welsh/Student)
- Clusters différents
 - Heights
 - Margins
 - Length
- Clusters sans différence
 - Diagonal
- Diagonal pas importante pour détecter des faux billets




M2: Comparaisons des cluster au vrai/faux sur le premier plan factoriel

- Numériquement assez proche
 - Vrai/Faux : 100/70
 - K-means : 93 /77
- Fiabilité du clustering
 - Taux d'erreur : 5,29%



M2: Conclusion

- Clusterisation non supervisée pas assez précise
 - Taux d'erreur trop important pour des billets ($>5\%$)
 - Différence entre cluster notable
- Test d'une classification supervisée pour prédire



M3: Classification supervisée régression logistique

M3: Méthode

- Utilisation de statsmodels.glm via python
- Procédure itérative descendante (backward)
 - Implémentation de tous les regressseurs
 - Retrait des paramètres non-significatifs
- Création d'un modèle avec les paramètres des différents regressseurs significatif
- Vérification du modèle
 - Matrice de de confusion
 - Sensibilité
 - Spécificité
 - Taux d'erreur
 - Précision
 - F1 score

M3: Procédure itérative (initialisation)

- Backward, problème avec 3 variables utilisées ensemble (séparation parfaite):
 - Margin_up
 - Margin_low
 - Length
- Analyses précédentes : margin_low et length plus importantes

M3: Procédure itérative (backward)

- Initialisation
- Suppression du regresseur le moins significatif
height_right : pValue = 0,785
- Fin de la procédure
- Tous les regresseurs sont significatifs
pValue = 0,014

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	is_genuineBol	No. Observations:	149			
Model:	GLM	Df Residuals:	143			
Model Family:	Binomial	Df Model:	5			
Link Function:	logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-4.1250			
Date:	Tue, 29 Sep 2020	Deviance:	8.2499			
Time:	18:10:25	Pearson chi2:	8.74			
No. Iterations:	12					
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]

Intercept	-1589.7988	1561.462	-1.018	0.309	-4650.208	1470.610
diagonal	3.5478	7.182	0.494	0.621	-10.529	17.625
height_left	-2.5560	8.917	-0.287	0.774	-20.033	14.921
height_right	1.4045	5.152	0.273	0.785	-8.693	11.502
margin_low	-14.8769	7.800	-1.907	0.056	-30.165	0.411
length	10.3834	5.761	1.802	0.071	-0.908	21.675

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	is_genuineBol	No. Observations:	149			
Model:	GLM	Df Residuals:	146			
Model Family:	Binomial	Df Model:	2			
Link Function:	logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-4.2763			
Date:	Tue, 29 Sep 2020	Deviance:	8.5525			
Time:	16:47:39	Pearson chi2:	10.0			
No. Iterations:	11					
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]

Intercept	-933.1189	379.932	-2.456	0.014	-1677.773	-188.465
margin_low	-13.2064	5.360	-2.464	0.014	-23.713	-2.700
length	8.8351	3.576	2.471	0.013	1.826	15.844

M3: Etablissement du modèle et vérification

- Modèle :
$$f(x) = \frac{e^{\beta_1 + \beta_2 \text{marginlow} + \beta_3 \text{length}}}{1 + e^{\beta_1 + \beta_2 \text{marginlow} + \beta_3 \text{length}}}$$
- Beta 1, 2 et 3 paramètres calculés par la régression logistique

M3: Vérification du modèle

- Le taux de spécificité:
 - $VN/(FP+VN) = 98,6 \%$
- Le taux d'erreur:
 - $(FN+FP)/\text{tot pop} = 1,18 \%$
- Le taux de sensibilité:
 - $VP/(VP+FN)=99,0 \%$
- Précision :
 - $VP/(VP+FP) = 99,0\%$
- F₁ score:
 - Moyenne harmonique de sensibilité et précision = 0,99

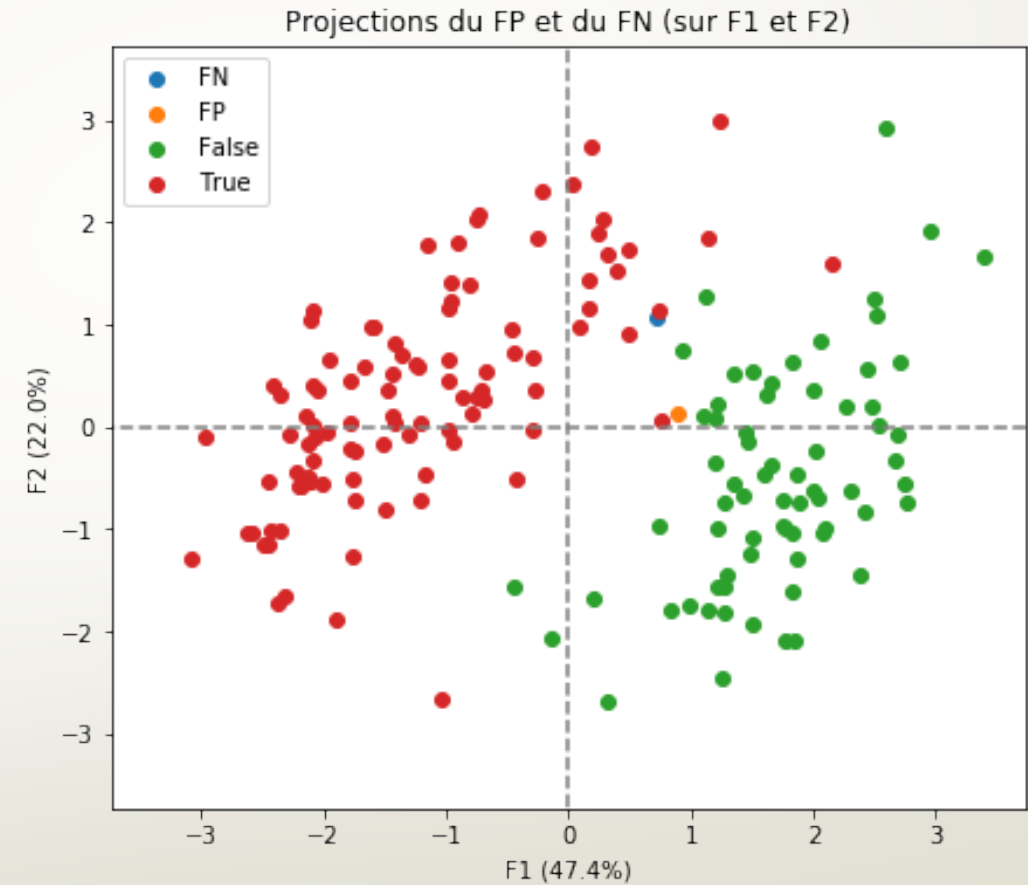
		Conditions	
		Positives	Négatives
Prédictions	Positives	VP : 99	FP: 1
	Négatives	FN : 1	VN:69

NB: vérification sur tout le data set car séparation groupe test groupe essai ressort trop facilement l'erreur parfaite vu dans l'initialisation

M3: illustration du FN et du FP

	Margin_low	Lenght	Di	Qualité F1	Qualité F2	Proba
FP	4.28	112.23	2,48	0,32	0,008	0,86
FN	4.63	112.47	2,37	0,22	0,47	0,35

- Billets mal représentés sur F1
- Peu d'influence sur l'inertie du nuage de points
- Différences sur les caractéristiques très faible
- Probabilité pas marquées



M3 Conclusion

- Régression logistique efficace
 - Précision : 99,0%
 - F1 score : 0,99
- Modèle prometteur
- A éprouver sur un nouveau jeu de billet