BARBIER CONSULTING

Détection des faux billets

Xavier BARBIER - Avril 2020 - @xavbarbier

MISSION

Office Central pour la Répression du Faux Monnayage (OCRFM)

Jeu de billets

Détection faux billets

Python / Notebook Jupyter

MISSION

Questions :

1. Caractéristiques géométriques pour différencier billets vraix/faux ?
2. Quelle modélisation permet le mieux d’identifier les billets ?
3. Est-il possible de proposer un outil pour différencier les billets vraix/faux ?

NETTOYAGE :

- Vérification valeurs manquantes

  -> aucune valeurs manquantes

DONNÉES

	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
2	True	171.83	103.76	103.76	4.40	2.88	113.84
3	True	171.80	103.78	103.65	3.73	3.12	113.63
4	True	172.05	103.70	103.75	5.04	2.27	113.55

170 observations

- 1 variable qualitative
- 6 variables quantitatives

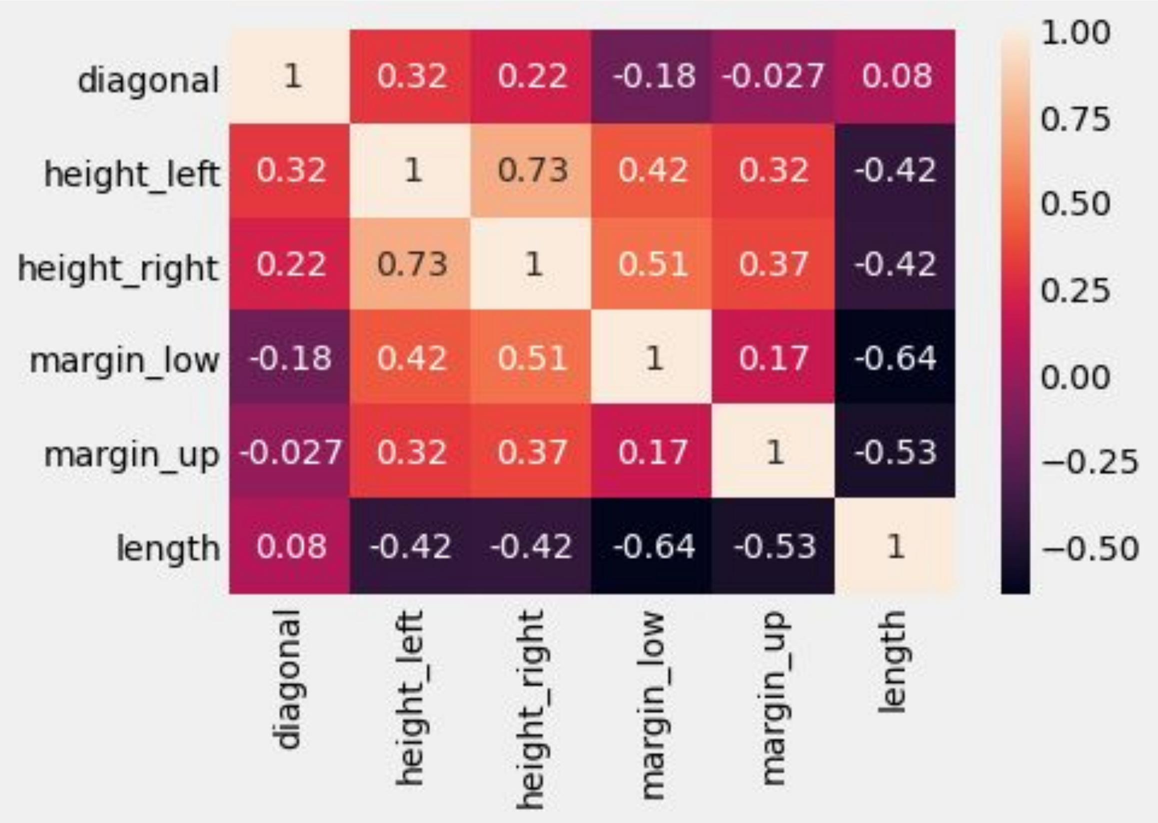
ANALYSE DESCRIPTIVE

TENDANCES CENTRALES

	diagonal	height_left	height_right	margin_low	margin_up	length
count	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000
mean	171.940588	104.066353	103.928118	4.612118	3.170412	112.570412
std	0.305768	0.298185	0.330980	0.702103	0.236361	0.924448
min	171.040000	103.230000	103.140000	3.540000	2.270000	109.970000
25%	171.730000	103.842500	103.690000	4.050000	3.012500	111.855000
50%	171.945000	104.055000	103.950000	4.450000	3.170000	112.845000
75%	172.137500	104.287500	104.170000	5.127500	3.330000	113.287500
max	173.010000	104.860000	104.950000	6.280000	3.680000	113.980000

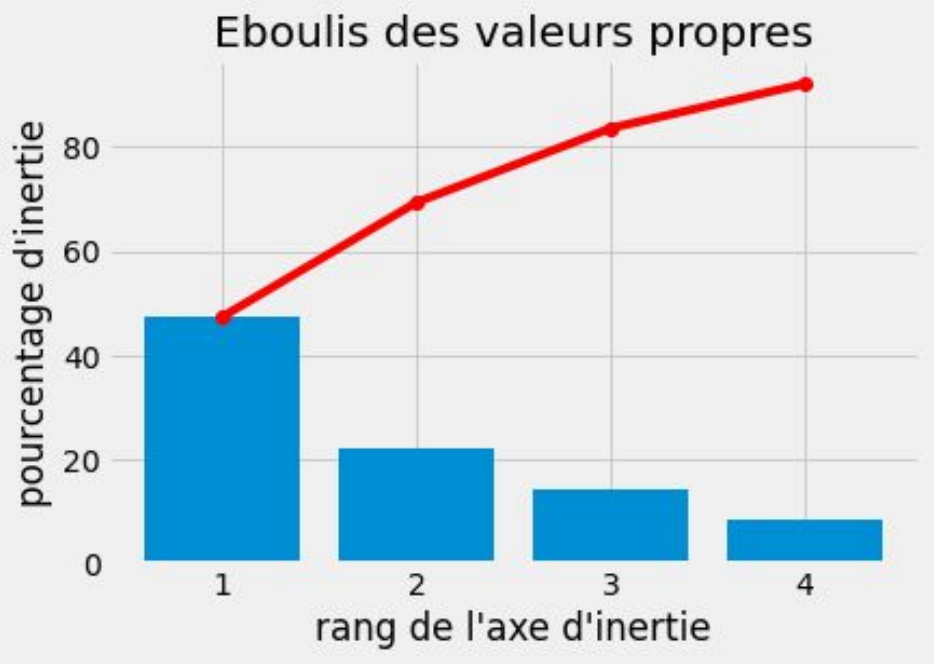
MATRICE PAR PAIRES

MATRICE CORRELATIONS



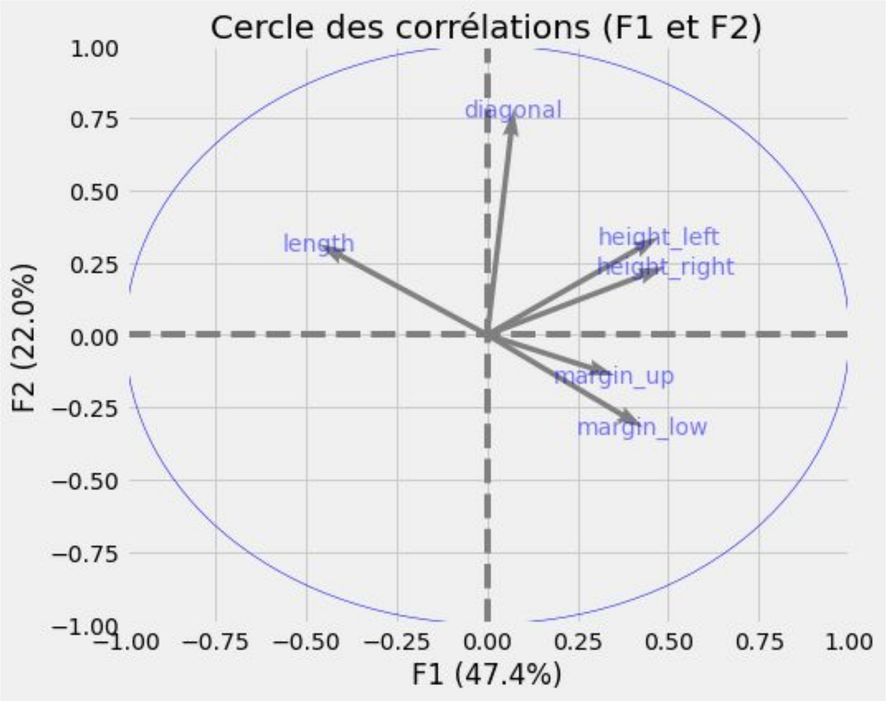
ACP

EBOULIS VALEURS PROPRES



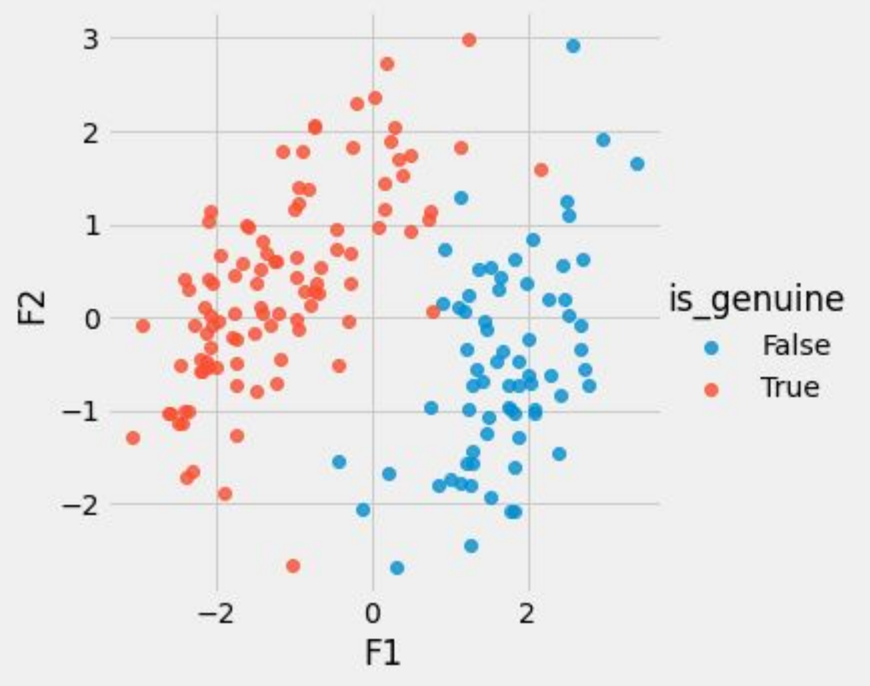
Hypothèse : conserver composante 1 (méthode coude - critère de Kaiser, 100/p, soit 50%)

CERCLE CORRÉLATION 1er PLAN



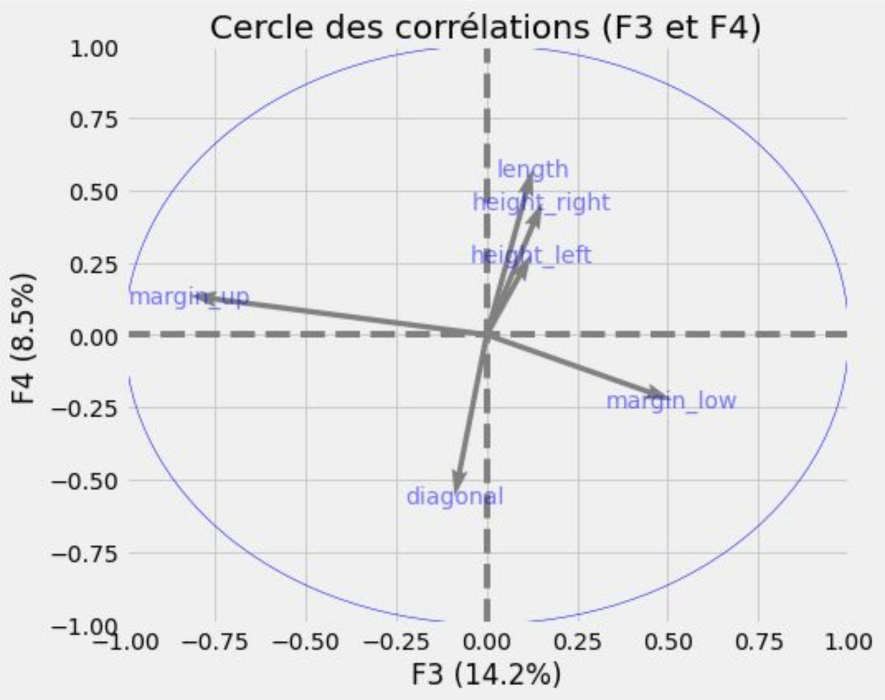
- Corrélation négative pour “lenght” avec 4 autres variables

PROJECTION INDIVIDUS 1er PLAN



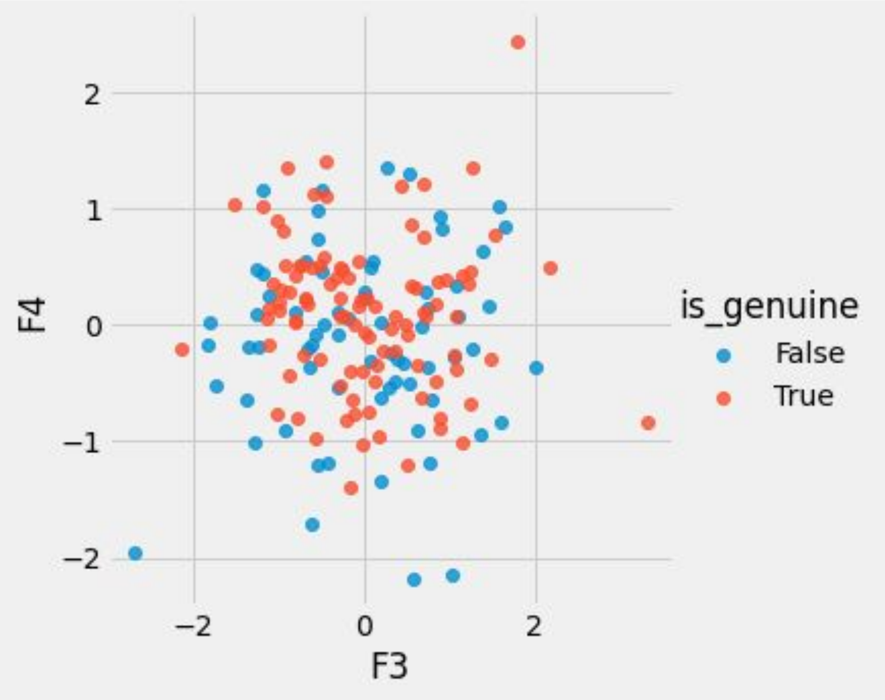
- bonne séparation billets sur 1er plan
- F1 & F2

CERCLE CORRÉLATION 2ème PLAN



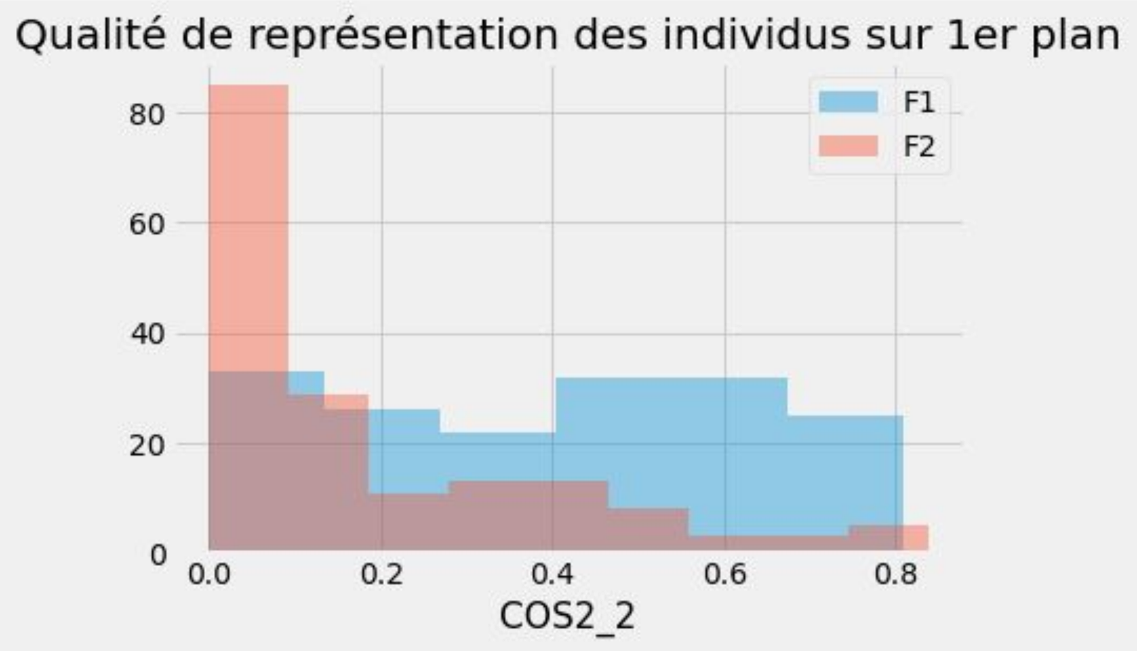
- Corrélation négative “margin”
- Corrélation négative “height”+”length” avec “diagonal”

PROJECTION INDIVIDUS 2ème PLAN



- Mauvaise séparation billets sur 2ème plan

QUALITÉ DE REPRÉSENTATION DES INDIVIDUS 1ER PLAN



- Meilleures représentation sur F1 que sur F2

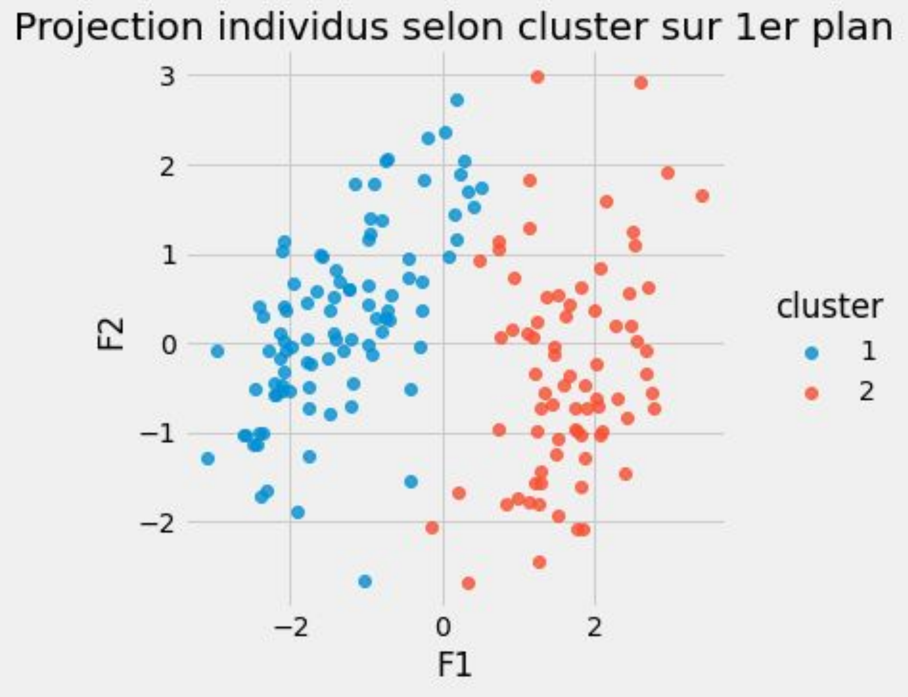
Caractéristiques géométriques pour différencier billets vrai/faux ?

Probable

- F1 et F2 / F1 ou F2 ?
- Variables initiales ?

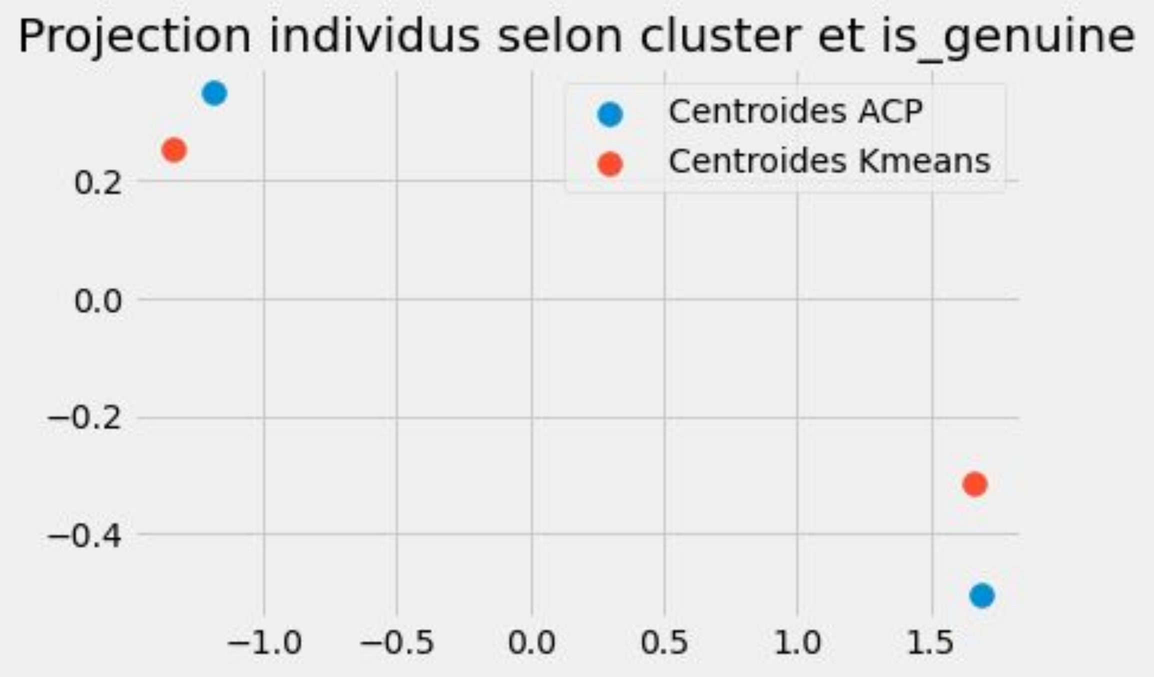
CLASSIFICATION

Kmeans



- bonne séparation clusters sur 1er plan

Centroïdes ACP et Kmeans



- Centroïdes assez similaires

Caractéristiques géométriques pour différencier billets vrai/faux ?

OUI

- F1 et F2 / F1 ou F2 ?
- Variables initiales ?

MODÉLISATION

PRÉSENTATION

Régression logistique

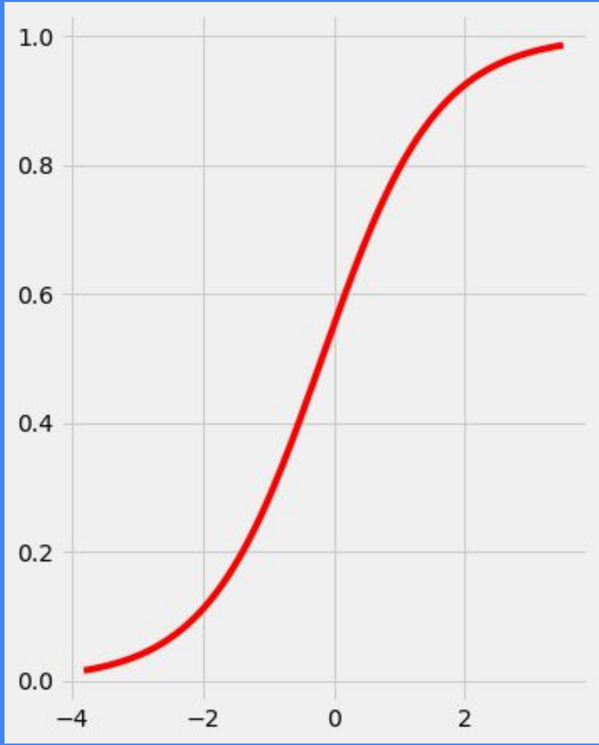
Statsmodel

Is_genuine

diagonal | height_left | height_right | margin_low | margin_up | length

F1 | F2

Modèle 1



Is_genuine

diagonal | height_left | height_right

margin_up | margin_low | length

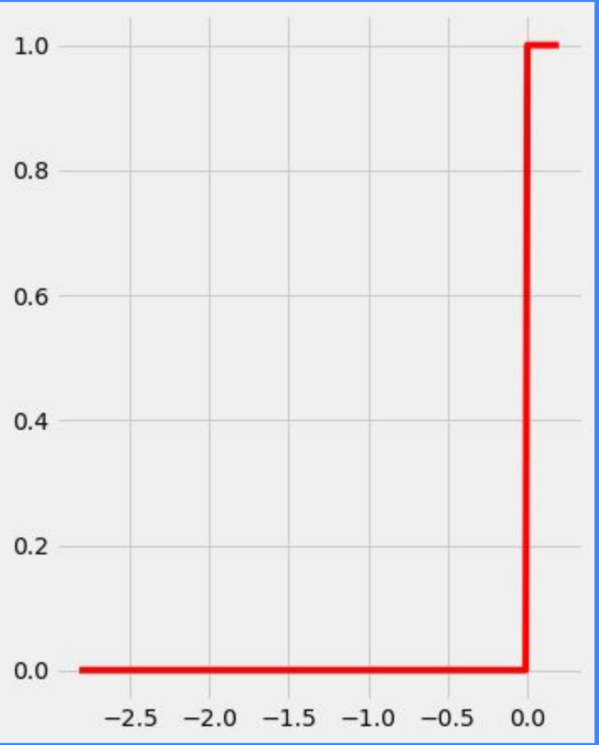
```
                           Logit Regression Results
==============================================================================
Dep. Variable:             is_genuine   No. Observations:                  170
Model:                          Logit   Df Residuals:                      164
Method:                           MLE   Df Model:                            5
Date:                Mon, 20 Apr 2020   Pseudo R-squ.:                  0.9999
Time:                        12:58:12   Log-Likelihood:             -0.0086318
converged:                       True   LL-Null:                       -115.17
Covariance Type:            nonrobust   LLR p-value:                 9.084e-48
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
diagonal        -1.5840     46.270     -0.034      0.973     -92.271      89.103
height_left      0.1350     51.880      0.003      0.998    -101.547     101.817
height_right     4.2889     69.600      0.062      0.951    -132.124     140.702
margin_low     -37.5951     76.858     -0.489      0.625    -188.234     113.044
margin_up      -17.0606     44.873     -0.380      0.704    -105.010      70.889
length          14.6691     40.654      0.361      0.718     -65.012      94.350
================================================================================

Possibly complete quasi-separation: A fraction 0.96 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Modèle 2 (backward)

Is_genuine

margin_low

length




```
                           Logit Regression Results
==============================================================================
Dep. Variable:             is_genuine   No. Observations:                  170
Model:                          Logit   Df Residuals:                      168
Method:                           MLE   Df Model:                            1
Date:                Mon, 20 Apr 2020   Pseudo R-squ.:                  0.9613
Time:                        12:58:12   Log-Likelihood:                -4.4520
converged:                       True   LL-Null:                       -115.17
Covariance Type:            nonrobust   LLR p-value:                 4.379e-50
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
margin_low    -9.5509      3.700     -2.582      0.010     -16.802      -2.300
length         8.1480      3.041      2.680      0.007       2.189      14.107
==============================================================================

Possibly complete quasi-separation: A fraction 0.75 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Is_genuine

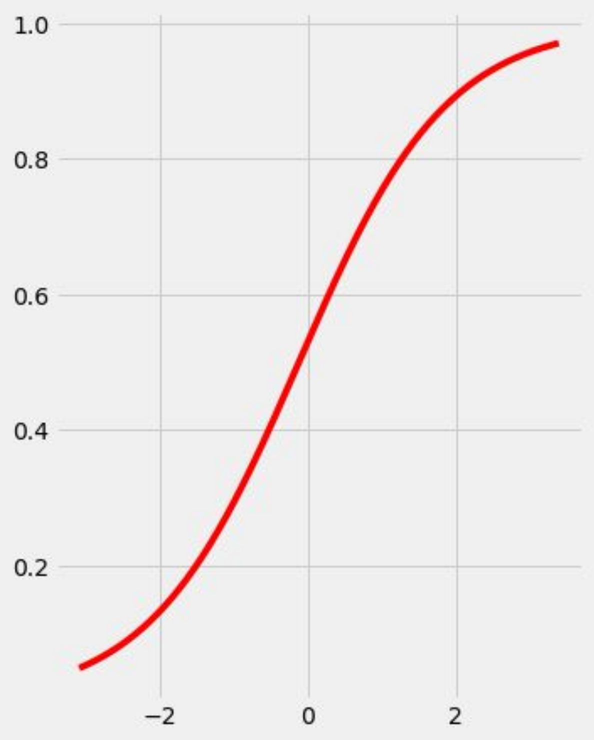
F1

F2

Modèle 3


```
                           Logit Regression Results
==============================================================================
Dep. Variable:             is_genuine   No. Observations:                  170
Model:                          Logit   Df Residuals:                      168
Method:                           MLE   Df Model:                            1
Date:                Mon, 20 Apr 2020   Pseudo R-squ.:                  0.8703
Time:                        12:58:13   Log-Likelihood:                -14.943
converged:                       True   LL-Null:                       -115.17
Covariance Type:            nonrobust   LLR p-value:                 1.656e-45
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
F1            -3.3735      0.685     -4.922      0.000      -4.717      -2.030
F2             2.4916      0.613      4.063      0.000       1.290       3.694
==============================================================================

Possibly complete quasi-separation: A fraction 0.11 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Modèle 4



Is_genuine

F1

```
                           Logit Regression Results
==============================================================================
Dep. Variable:             is_genuine   No. Observations:                  170
Model:                          Logit   Df Residuals:                      169
Method:                           MLE   Df Model:                            0
Date:                Mon, 20 Apr 2020   Pseudo R-squ.:                  0.6585
Time:                        12:58:13   Log-Likelihood:                -39.335
converged:                       True   LL-Null:                       -115.17
Covariance Type:            nonrobust   LLR p-value:                       nan
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
F1            -2.1334      0.313     -6.824      0.000      -2.746      -1.521
==============================================================================
```

CROSS VALIDATION

PRÉSENTATION

Séparation 70/30

Stratification

Accuracy

Sensitivity

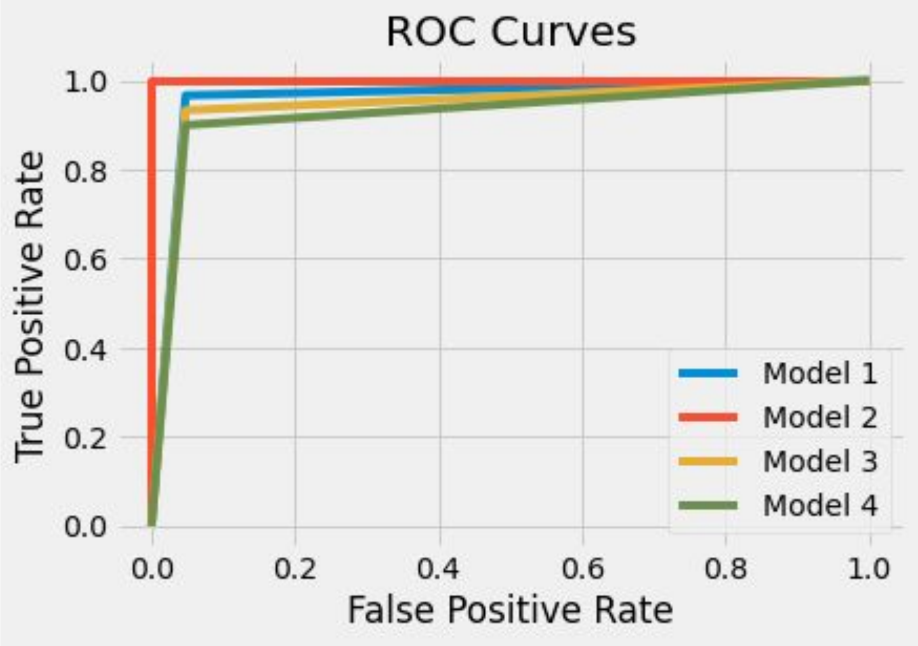
Specificity

Precision

AUC

Performances

	model1	model 2	model 3	model 4
accuracy	0.960784	1.0	0.941176	0.921569
sensitivity	0.952381	1.0	0.952381	0.952381
specificity	0.966667	1.0	0.933333	0.900000
precision	0.966667	1.0	0.965517	0.964286
auc	0.959524	1.0	0.942857	0.926190



Quelle modélisation permet le mieux d’identifier les billets ?

margin_low | length

OUTIL POUR DIFFÉRENCIER LES BILLETS VRAIX/FAUX ?

	diagonal	height_left	height_right	margin_low	margin_up	length	id
0	171.76	104.01	103.54	5.21	3.30	111.42	A_1
1	171.87	104.17	104.13	6.00	3.31	112.09	A_2
2	172.00	104.58	104.29	4.99	3.39	111.57	A_3
3	172.49	104.55	104.34	4.44	3.03	113.20	A_4
4	171.65	103.63	103.56	3.77	3.16	113.33	A_5

	id	proba	is_genuine
0	A_1	0.025227	False
1	A_2	0.015513	False
2	A_3	0.084346	False
3	A_4	0.993360	True
4	A_5	0.999580	True

MERCI !

@xavbarbier

https://www.linkedin.com/in/barbierxavier/

https://github.com/xavierbarbier/

contact@xavierbarbier.com