Détection de faux billets

...

Août 2018

Antoine LEPAGE - yop1001@gmail.com

Démarche

- Mission 0 - Analyse des données :
  - (nettoyage des données)
  - comprendre les variables
  - comprendre ce qui différencie les vrais des faux billets

- Mission 1 - Analyse en composantes principales

- Mission 2 - Application d'un algorithme de classification
  - choix d'un algorithme de classification : KMeans
  - évaluation de la performance de l'algorithme

- Mission 3 - Modélisation des données avec une régression logistique
  - évaluation de la performance
  - prédiction de l'authenticité sur de nouveaux billets

Comprendre les données

...

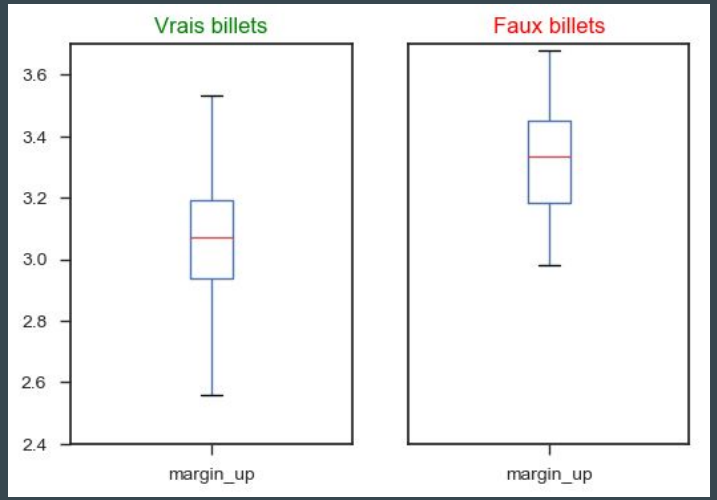
Mission 0

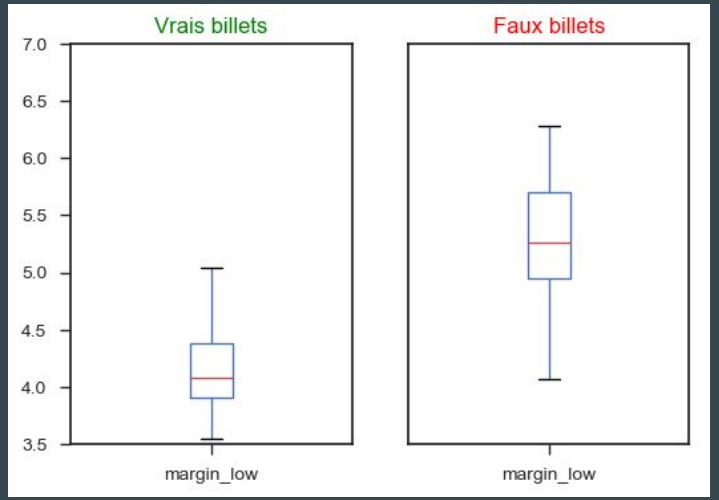
Variables

Fournis par la PJ

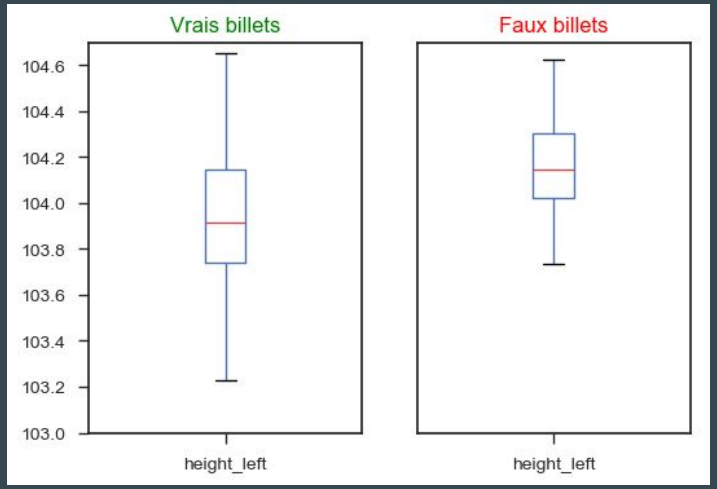
- 6 caractéristiques géométriques de 170 billets de banque
- information pour chacun s'il est *Vrai* ou *Faux* (*is_genuine*)

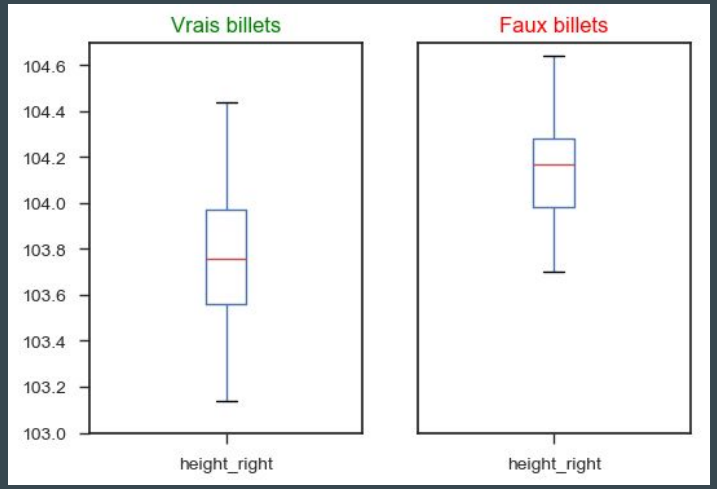
Analyse des variables



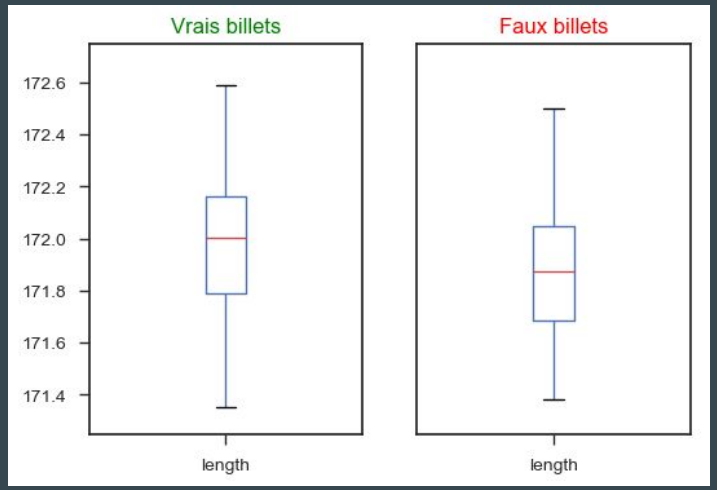


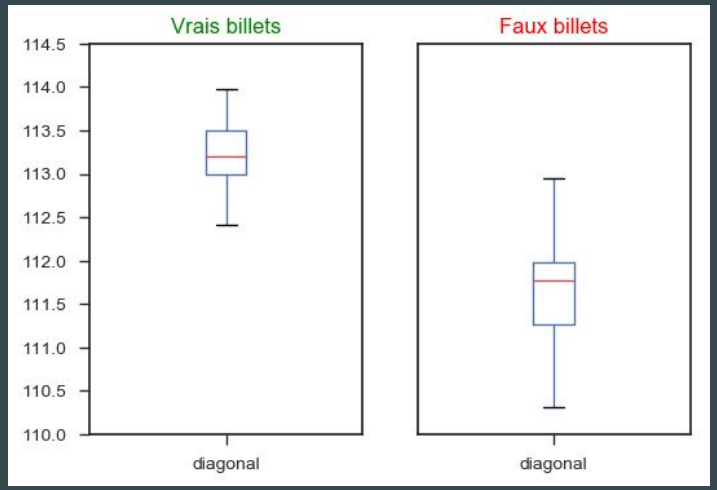
Analyse des variables





Analyse des variables





- La diagonale est la mesure qui discrimine le plus les vrais des faux billets

Analyse en Composantes Principales

...

Mission 1

Analyse en Composantes Principales (ACP)

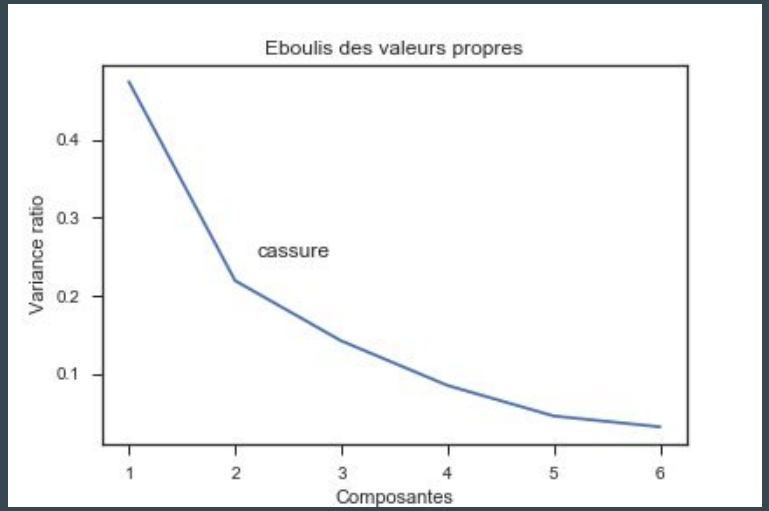
- => Permet d'analyser et de visualiser un jeu de données contenant des individus décrits par plusieurs variables quantitatives.

- ACP est utilisée pour extraire et de visualiser les informations importantes contenues dans une table de données multivariées. L'ACP synthétise cette information en seulement quelques nouvelles variables appelées composantes principales.

- L'information contenue dans un jeu de données correspond à la variance ou l'*inertie totale* qu'il contient. L'objectif de l'ACP est d'identifier les directions (i.e., *axes principaux* ou composantes principales) le long desquelles la variation des données est maximale.

Choix du nombre de composantes principales

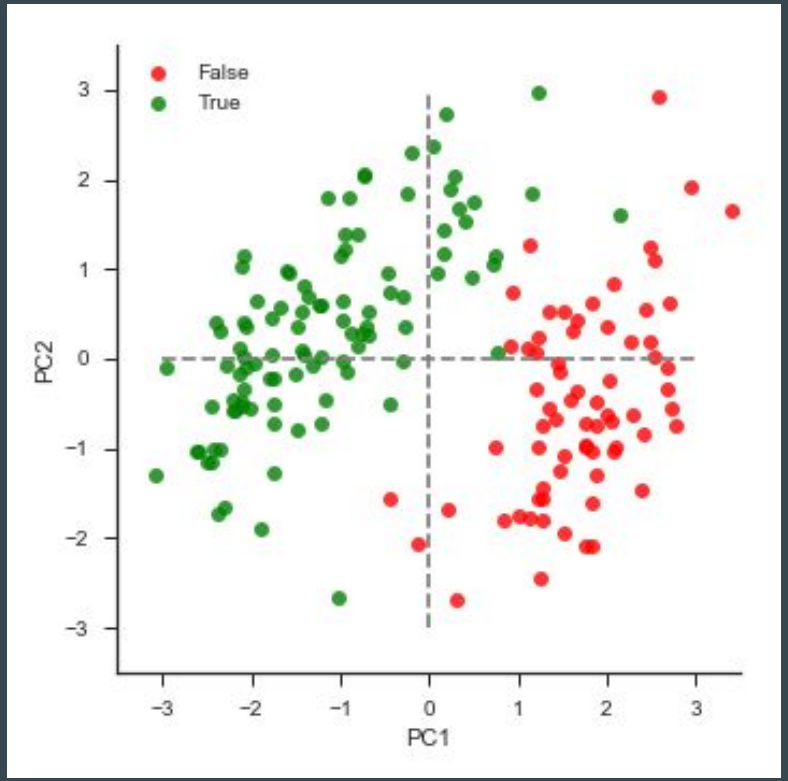
- Objectif : obtenir le maximum d'inertie conservée avec le minimum de composantes principales

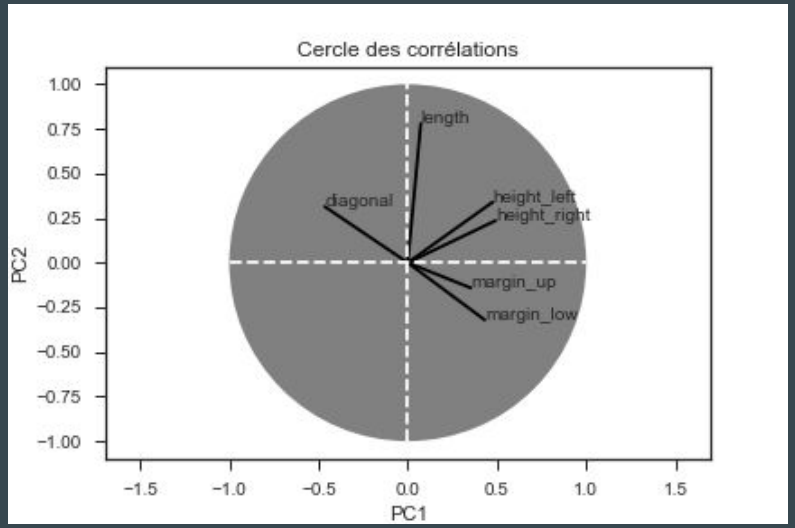


- Premier axe conserve 47% de l'inertie du nuage, 22% pour le second
- Chute importante dès le 3ie axe qui ne conserve que 14% de l'inertie

=> Nous retenons que les deux premiers axes (69% de la variance expliquée)

Visualisation sur 2 axes





Plus les variables sont proches du cercle + elles sont bien représentées

Les vrais billets se définissent principalement par une diagonale plus grande que les faux.

Qualité de représentation des individus sur les axes

- COS^2 permet de quantifier la qualité de la projection au niveau de chaque point
- = indice de non-déformation des distances vues sur le plan ACP

	COS2_1	COS2_2	COS2_sum
0	0.644437	0.355563	1.0
1	0.941505	0.058495	1.0
2	0.999406	0.000594	1.0
3	0.998132	0.001868	1.0
4	0.971425	0.028575	1.0

- Interprétation - exemple de l'individu 0
  Sur l'axe 1, l'individu 0 a un indice de non déformation de 0.64 et de 0.36 sur l'axe 2

- 103 billets sur 170 sont très bien réprésentés sur l’axe 1 (indice > 0.75) - VS 27 pour l’axe 2

Contributions des individus aux axes

- Elles permettent de déterminer les individus qui pèsent le plus dans la définition de chaque facteur.

	CTR_1	CTR_2
122	0.067637	0.016300
49	0.055856	0.009864
29	0.051497	0.000050
112	0.051102	0.021420
158	0.045083	0.003193

les 5 billets qui influent le plus dans la définition du premier facteur de l'ACP :

	CTR_1	CTR_2
5	0.008981	0.052350
166	0.039317	0.049966
34	0.000208	0.043873
156	0.000602	0.042505
70	0.006254	0.041520

les 5 billets qui influent le plus dans la définition du second facteur de l'ACP :

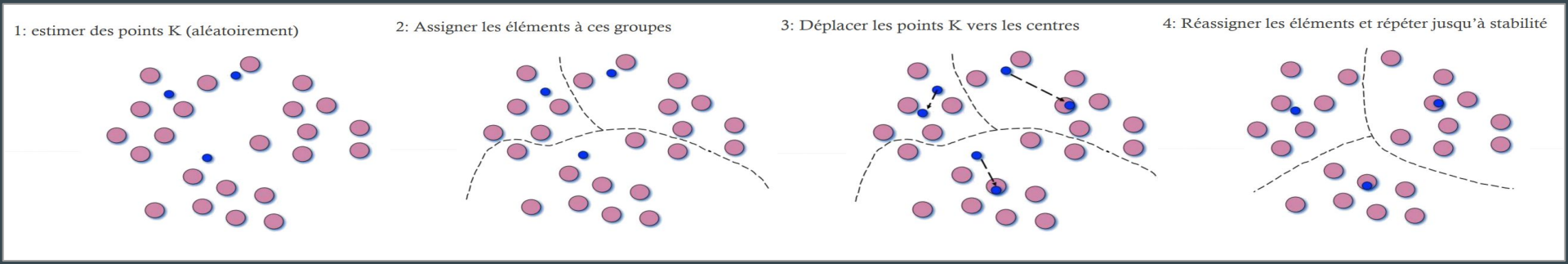
Classification - K-means

...

Mission 2

Algorithme K-means

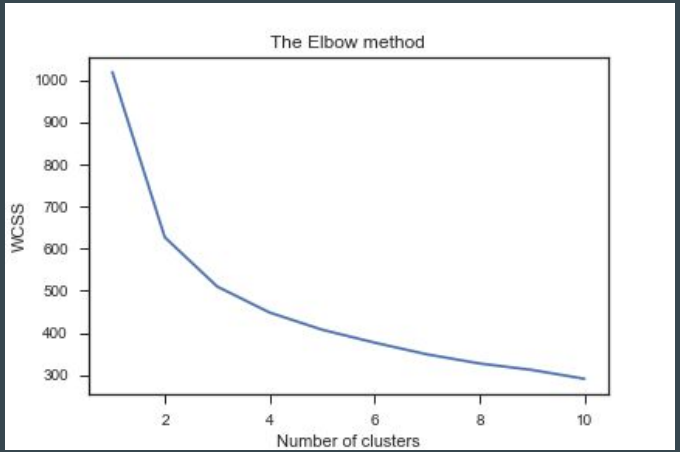
- Apprentissage non-supervisé. Va plutôt trouver des *patterns* dans les données.
- k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd.
- Le choix initial des centroïdes conditionne le résultat final.
- préconisé quand on connaît le nombre de groupe



Source : *http://www.ieee.ma/uaesb/pdf/Algo-k-Moyennes.pdf*

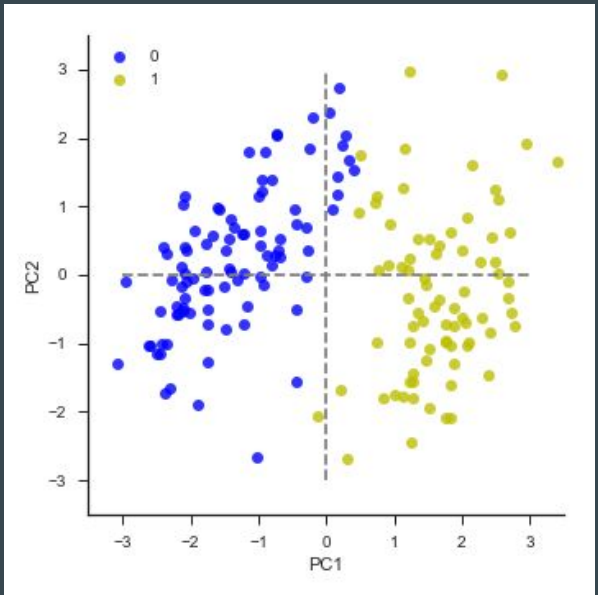
Définir le nombre de groupe

- Méthode “elbow method”
  - permet de déterminer le nombre de clusters en fonction de la variance expliquée



- Le coude peut être déterminé entre 2 et 3 clusters. Nous choisirons 2 clusters car notre problématique est de différencier les vrais des faux billets.

Visualisation de la classification dans le plan de l'ACP



is_genuine	False	True
Groupe		
0	1	92
1	69	8

- Le groupe 0 correspond aux “vrais billets” et le groupe 1 aux “faux billets” à quelques individus prêts.
- L'algorithme de classification donne un taux de précision de 80%

Modélisation - Régression logistique

...

Mission 3

Modélisation des données avec une régression logistique

- Elle permet de mesurer l'association entre la survenue d'un évènement (variable expliquée qualitative) et les facteurs susceptibles de l'influencer (variables explicatives).
    - variable expliquée qualitative : *is_genuine*
    - variables explicatives : *les différentes mesures du billet*

- Apprentissage supervisé (variable à expliquer est connue à l'avance)

- Technique prédictive

Fonctionnement

- Séparation des données :
  - Training set : 80%
  - Test set : 20%

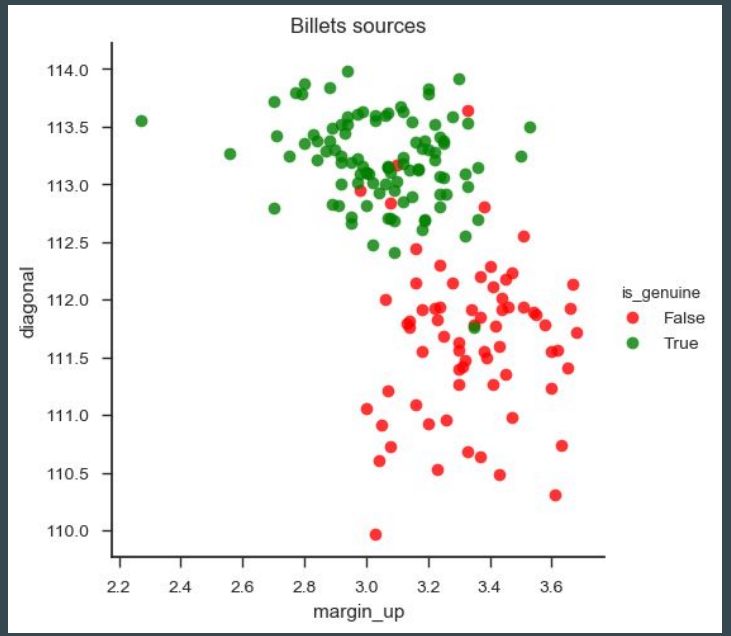
- Algorithme va modéliser la régression logistique sur le *training set* et tester les prédictions sur le *test set*

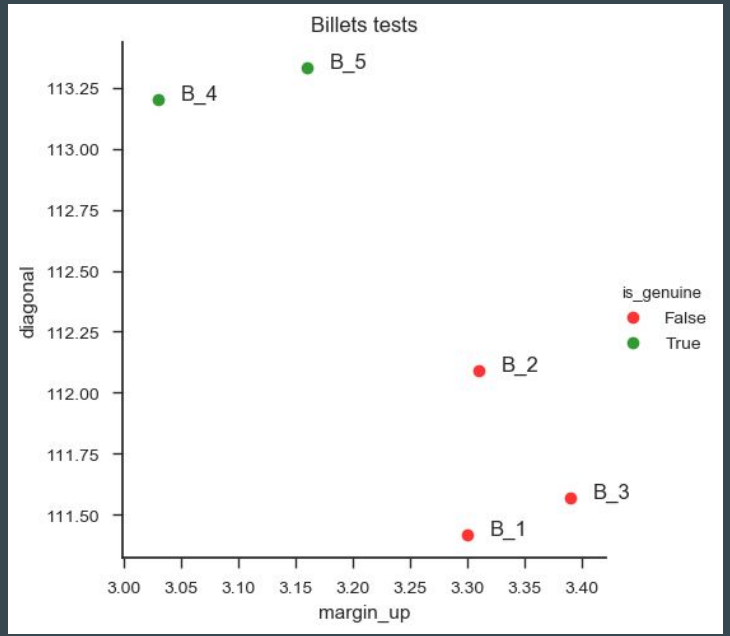
- Taux de précision : 100%

- Prédiction à partir d'un fichier d'évaluation

Prédictions de l'authenticité de 5 nouveaux billets

- données : *exemple.csv* donné dans le cours





id	False	True
B_1	0.959252	0.040748
B_2	0.990545	0.009455
B_3	0.971500	0.028500
B_4	0.123976	0.876024
B_5	0.003749	0.996251

Prédictions avec le fichier d’évaluation

- Montrer Exemple_out
- Montrer Proba