

Cahier des charges

Algorithme de détection automatique de
faux billets



**Organisation nationale de lutte contre le
faux-monnayage (ONCFM)**

Sommaire

Contexte du projet	3
Objectifs	3
Modèle de données	3
Dimensions géométriques	3
Algorithme	4
Langage	4
Fonctionnement général	4
Livrables attendus	4
Calendrier	5
Contexte et objectif	6
Périmètre et fonctionnement	6

Contexte du projet

L'Organisation nationale de lutte contre le faux-monnayage, ou ONCFM, est une organisation publique ayant pour objectif de mettre en place des méthodes d'identification des contrefaçons des billets en euros. Dans le cadre de cette lutte, nous souhaitons mettre en place un algorithme qui soit capable de différencier automatiquement les vrais des faux billets.

Objectifs

Lorsqu'un billet arrive, nous avons une machine qui consigne l'ensemble de ses caractéristiques géométriques. Au travers de nos années de lutte, nous avons observé des différences de dimensions entre les vrais et les faux billets. Ces différences sont difficilement notables à l'œil nu, mais une machine devrait sans problème arriver à les différencier.

Ainsi, il faudrait construire un algorithme qui, à partir des caractéristiques géométriques d'un billet, serait capable de définir si ce dernier est un vrai ou un faux billet.

Modèle de données

Dimensions géométriques

Nous disposons actuellement de six informations géométriques sur un billet :

- `length` : la longueur du billet (en mm) ;
- `height_left` : la hauteur du billet (mesurée sur le côté gauche, en mm) ;
- `height_right` : la hauteur du billet (mesurée sur le côté droit, en mm) ;
- `margin_up` : la marge entre le bord supérieur du billet et l'image de celui-ci (en mm) ;
- `margin_low` : la marge entre le bord inférieur du billet et l'image de celui-ci (en mm) ;
- `diagonal` : la diagonale du billet (en mm).

Ces informations sont celles avec lesquelles l'algorithme devra opérer.

Fichier pour paramétrisation

Nous vous laissons un fichier d'exemple contenant 1 500 billets, que vous pouvez utiliser comme bon vous semble pour paramétrer votre algorithme. Parmi ces 1 500 billets, 1 000 sont vrais et 500 sont faux ; une colonne a été ajoutée pour vous préciser la nature du billet.

Dans un premier temps, nous souhaitons avoir une analyse descriptive des données, notamment la répartition des dimensions des billets, le nombre de vrais / faux billets, etc.

Algorithme

Langage

L'algorithme peut être écrit soit en Python, soit en R : le choix est laissé libre au consultant chargé de sa réalisation.

Fonctionnement général

Comme vu précédemment, nous avons à notre disposition six données géométriques pour chaque billet. L'algorithme devra donc être capable de prendre en entrée un fichier contenant les dimensions de plusieurs billets, et de déterminer le type de chacun d'entre eux, à partir des seules dimensions. Nous fournissons à ce sujet le format type de nos fichiers de billets avec lequel l'algorithme sera censé fonctionner, au sein d'un fichier nommé `billets_production.csv`.

Nous aimerions pouvoir mettre en concurrence deux méthodes de prédiction :

- une régression logistique classique ;
- un k-means, duquel seront utilisés les centroïdes pour réaliser la prédiction.

Cet algorithme se devra d'être naturellement le plus performant possible pour identifier un maximum de faux billets au sein de la masse de billets analysés chaque jour.

Pour une évaluation optimale des modèles, nous souhaitons avoir une analyse des nombres de faux positifs et faux négatifs via une matrice de confusion.

Livrables attendus

L'algorithme devra être livré au sein d'un notebook Python ou R. Il devra être fonctionnel avec le format de données ne comportant que les dimensions géométriques.

Calendrier

Le délai prévu pour l'expérimentation et la production de l'algorithme est d'un mois.