Détection automatique de faux billets

Nicolas Pautet

ONCFM

Septembre 2025

Contexte : objectif de l’étude

Mettre en place un algorithme de classification de billets

Source de données :

- Relevés de mesures géométriques sur échantillon de billets
- 1500 billets, 1000 authentiques et 500 faux

Objectif - détecter automatiquement les faux billets :

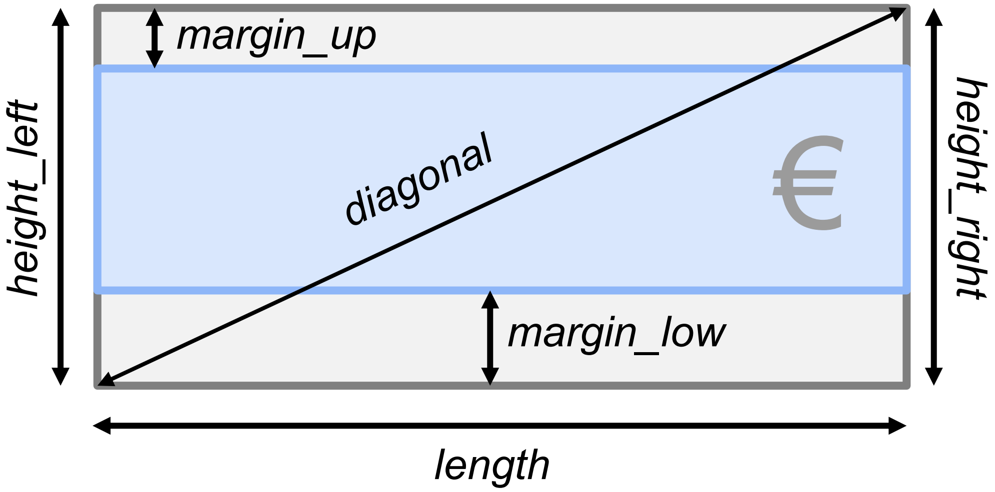
- Tester et évaluer la performance de modèles prédictifs
- Sélectionner et appliquer le modèle le plus performant

Données d'entraînement

1500 billets, 7 variables :

- 6 cotes en mm
- *is_genuine* : booléen, indique si le billet est authentique ou non
→ variable à prédire

Position des cotes relevées sur les billets



Données manquantes

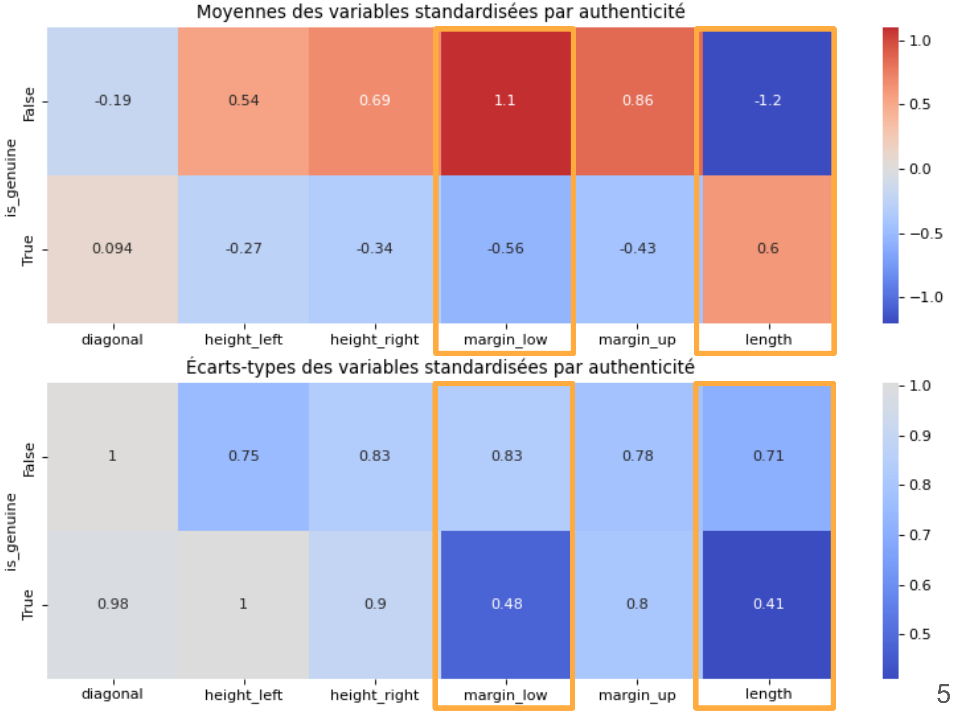
/!\ 37 billets (29 vrais et 8 faux) → 2,5% de l'échantillon pour lesquels on n'a pas la mesure *margin_low*

Ces valeurs manquantes seront imputées par régression linéaire à partir des autres variables

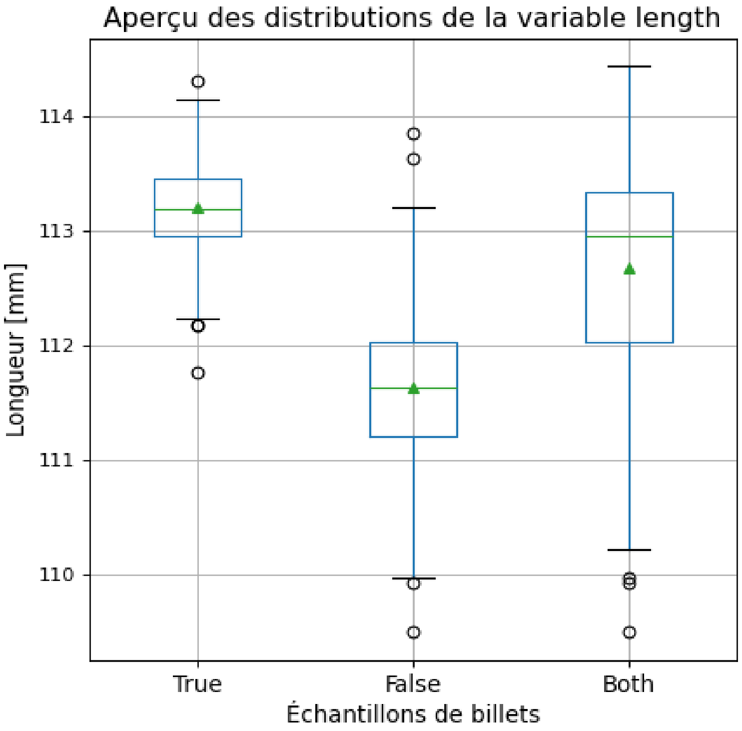
Analyses univariées

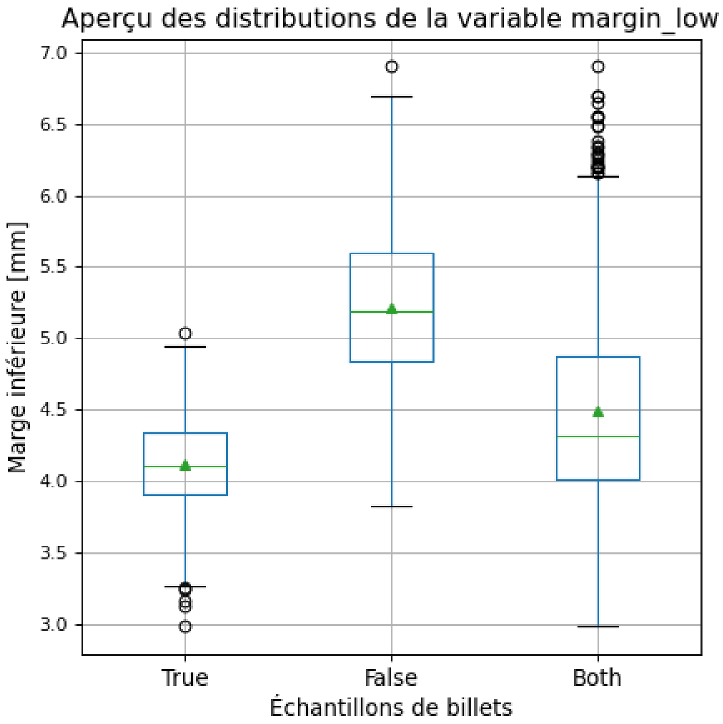
Écarts de cotes significatifs entre vrais et faux billets pour variables *length* et *margin_low*

Variance plus faible pour ces 2 variables sur les vrais billets



Analyses univariées

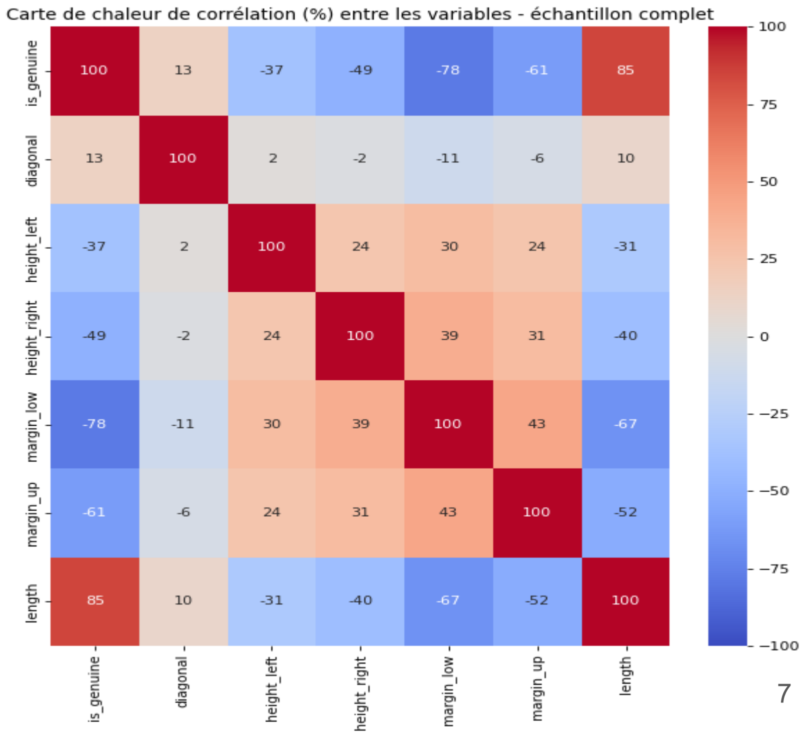




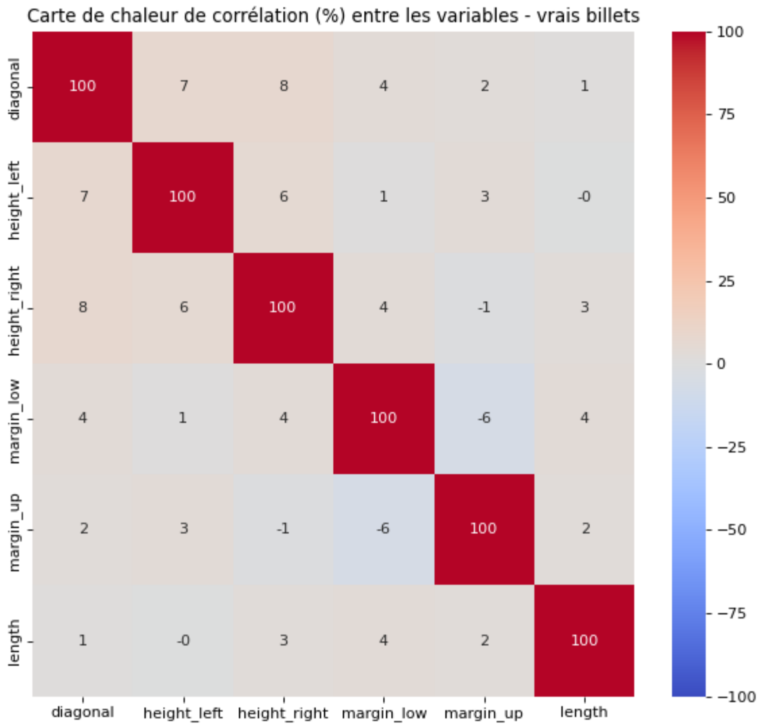
Analyse bivariée sur l’échantillon complet

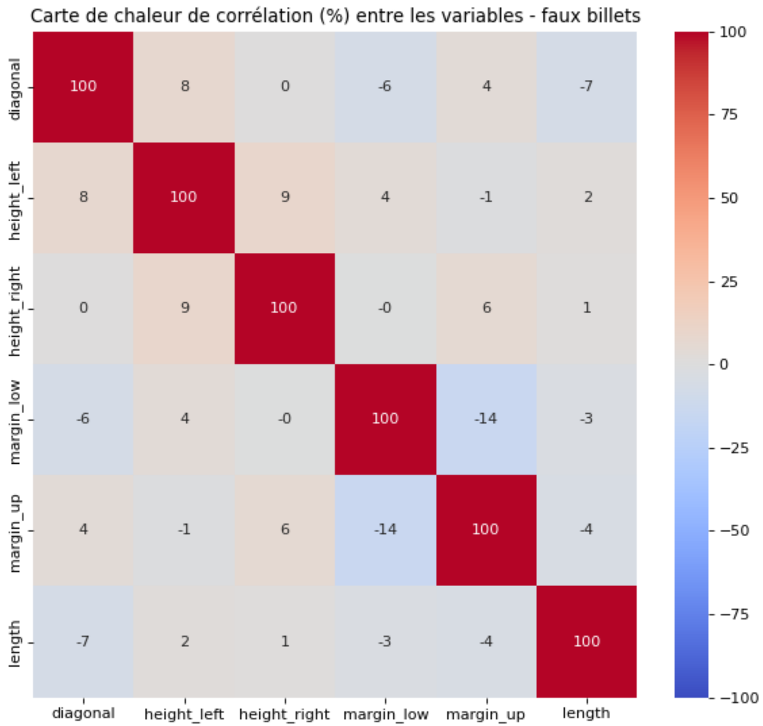
La variable *is_genuine* a un poids considérable relativement à l’information contenue dans les données

Une fois les billets séparés selon leur authenticité, les variables semblent décorrélées



Analyse bivariée selon l’authenticité





Imputation par régression linéaire

D'après ce qui précède, on devrait traiter séparément les vrais billets des faux billets quant à l'imputation de valeurs manquantes pour la variable *margin_low*

Pour chaque authenticité de billets, on utilise une approche incrémentale de comparaison de modèles, afin d'établir des modèles de régression simples et statistiquement pertinents

Démarche de comparaison de modèles

1. On détermine l'ordre dans lequel devraient être ajoutées une par une les variables prédictives, en fonction du score R^2 des régressions associées, cela donne une suite ordonnée de modèles à tester

2. On effectue un test F comparant le modèle à n prédicteurs vs. le modèle à n+1 prédicteurs, et on suit le résultat du test (rejeter le modèle n au profit du modèle n+1)

3. On peut s'arrêter dès qu'un modèle n'est pas rejeté par un test F, et on le compare à un modèle avec tous les prédicteurs possibles

Application aux faux billets

- Prédiction constante vs. *margin_low ~ margin_up*
  $p \sim 10^{-303}$ → modèle compact rejeté
- *margin_low ~ margin_up* vs. *margin_low ~ {margin_up + diagonal}*
  $p \approx 0{,}184$ → modèle compact acceptable
- *margin_low ~ margin_up* vs. *margin_low ~ {tous les prédicteurs}*
  $p \approx 0{,}486$ → modèle retenu validé

Le modèle retenu est : $margin_{low} = -1{,}100 * margin_{up} + 6{,}690$
pour lequel on a $R^2 = 0{,}02$

Application aux vrais billets

- Prédiction constante vs. *margin_low ~ margin_up*
  $p \approx 0{,}060$ → modèle compact acceptable
- *margin_low ~ margin_up* vs. *margin_low ~ {margin_up + length}*
  $p \approx 0{,}210$ → modèle compact acceptable
- *margin_low ~ margin_up* vs. *margin_low ~ {tous les prédicteurs}*
  $p \approx 0{,}385$ → modèle retenu validé

Le modèle retenu est : $margin_{low} = -0{,}104 * margin_{up} + 4{,}434$
pour lequel on a $R^2 = 0{,}004$

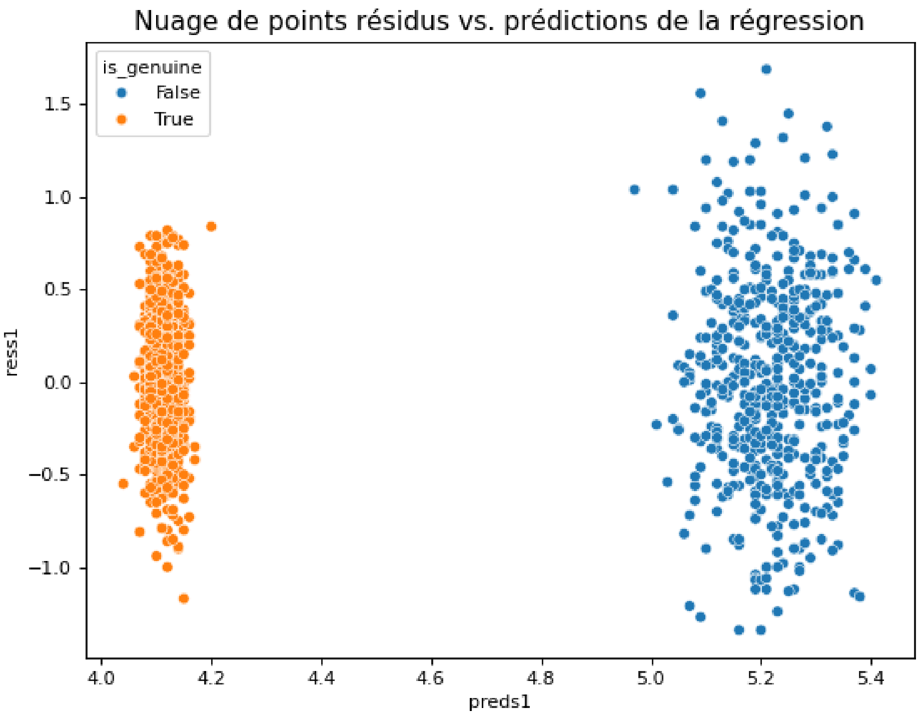
Performances de la régression

Comparaison des prédictions $\hat{y}_k$ avec les valeurs y_k connues :

$$MAE = \frac{1}{N} * \sum |\hat{y}_k - y_k| = 0{,}315\ mm$$

$$RMSE = \sqrt{\frac{1}{N} * \sum (\hat{y}_k - y_k)^2} = 0{,}410\ mm$$

$$MAPE = \frac{1}{N} * \sum \frac{|\hat{y}_k - y_k|}{|y_k|} = 7{,}0\ \%$$



Modèles prédictifs

Variable prédite : *is_genuine*

Le billet est-il authentique ?

Prédicteurs :

Les mesures géométriques

Métrique pour évaluer et classer les modèles

Rappel de l'objectif : **détecter en priorité les faux billets**

Vrai billet = Négatif   Faux billet = Positif

La métrique $Rappel = \frac{VP}{VP+FN}$ (à maximiser) apparaît pertinente

Problème : prédire tout le temps que le billet est faux est optimal

On va plutôt utiliser un F-score (à maximiser) :

$$F_{score} = (1 + rc)\frac{Précision * Rappel}{rc * Précision + Rappel} \quad \text{où :} \quad \mathrm{rc} = \frac{coût\ traitement\ FN}{coût\ traitement\ FP} = 20$$

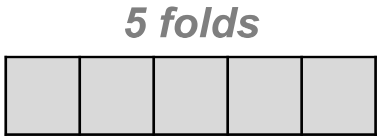
Modèles testés

- K-Means avec 2 clusters (k = 2)
- Régression logistique avec seuil de décision à déterminer
- K-NN avec k (nombres de voisins) à déterminer
- Forêt aléatoire avec n (nombre d’arbres) à déterminer

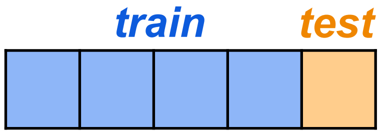
Principe de la validation croisée

Les données de test sur lesquelles on fait des prédictions doivent être indépendantes des données d'entraînement

Échantillon complet découpé en 5 *folds* de taille égale :



On fait tour à tour des prédictions sur chaque *fold* en utilisant les données des autres *folds* pour l'entraînement



On compare les valeurs prédites avec les valeurs réelles (connues)

Performances des modèles testés

Résultats obtenus par validation croisée :

Nom	Valeur paramètre	F_score	Rappel	Exactitude	Précision
K-Means	2	0,976	0,976	0,986	0,982
Régression logistique	0,85	0,994	0,996	0,983	0,956
KNN	2	0,985	0,986	0,984	0,966
Forêt aléatoire	2	0,986	0,990	0,966	0,916

Visualisation des résultats

→ Matrice de confusion

Performances de l'algorithme de régression logistique avec seuil de décision à 0,85 en validation croisée :

<u>Rappel</u> : Positif = Faux billet Négatif = Vrai billet		Valeur prédite	
		Positif	Négatif
Valeur réelle	Positif	498	2
	Négatif	23	977

Observations et conclusion

Modèle retenu : régression logistique avec seuil de décision à 0,85

Les modèles présentés sont conservateurs car ils sont évalués en priorité sur leur capacité à détecter les faux billets quitte à ne pas prédire correctement des vrais billets (faux positifs)

Selon leur probabilité prédite (par exemple si $0{,}5 \leq p \leq 0{,}85$), les billets testés ne seront pas toujours faux mais seront au moins suspects

Merci pour votre attention

Avez-vous des questions ?