



Détection de faux billets de banque

Pour l'ONCFM

Thomas Fruchard Data Analyst

De l'analyse des faux billets à la mise en place de l'application

01

Gestion des données manquantes

02

Vérification des hypothèses

03

Tests algorithmiques

04

Choix des modèles

05

Mise en place de l'application

01. Gestion des données manquantes

Après avoir effectué un R^2 , le R^2 passe de 0.477 à 0.475, ce qui est pratiquement identique. Cela confirme que diagonal n'apporterait rien de significatif à la prédiction de `margin_low`, (cela ne nuit pas non plus).

Je décide de ne pas garder les lignes de billets avec les données manquantes.



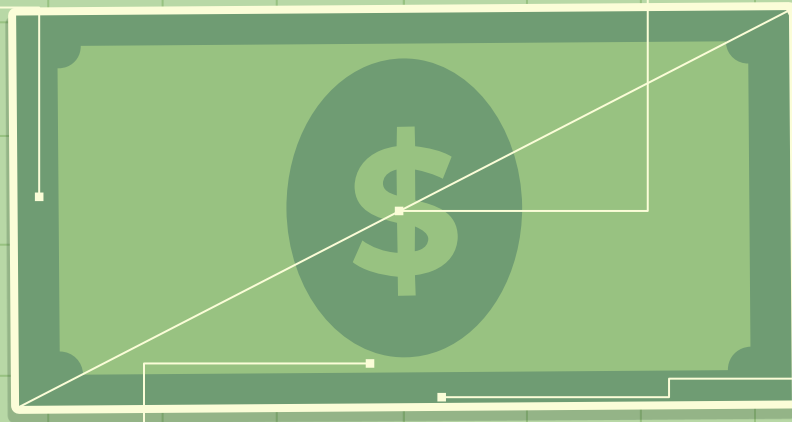
Les données

**Hauteur
droite et
gauche**

La diagonal

**Marge haute
et basse**

Longueur




02. Vérification des hypothèses

- But principal :
 - S'assurer que les conditions nécessaires pour utiliser et interpréter correctement le modèle de régression sont respectées.
- Dans notre analyse :
 - ✓ On a vérifié la normalité des résidus à l'aide du QQ plot.
 - ✓ On a contrôlé la variance constante des résidus avec le graphique des résidus vs valeurs ajustées.
- Pourquoi c'est important ?
 - Si les hypothèses sont respectées, alors :
 - Les estimations sont fiables
 - Les valeurs p et intervalles de confiance sont valides
 - Le modèle est statistiquement solide




Le QQ plot

(Insérer ici le QQ plot des résidus)

- Objectif : S'assurer que les erreurs (résidus) du modèle suivent une loi normale.
- Méthode utilisée : QQ plot (quantile-quantile plot)
- Résultat : Les points du graphique suivent globalement la diagonale.
-  Conclusion : L'hypothèse de normalité est globalement respectée, ce qui valide l'utilisation des tests statistiques du modèle.

L'homoscédasticité

- Objectif : Objectif : Vérifier que la variance des résidus reste constante.
- Méthode utilisée : Graphique des résidus vs valeurs ajustées.
- Résultat : Pas de motif clair (pas de cône ou de forme particulière).
-  Conclusion : L'hypothèse d'homoscédasticité est raisonnablement respectée.

(Insérer ici le graphique résidus vs valeurs ajustées)

03. Tests algorithmiques

But principal :

- Évaluer la performance réelle du modèle sur des données qu'il n'a jamais vues.
- Vérifier qu'il généralise bien et ne se contente pas de mémoriser l'entraînement (éviter le sur-apprentissage).

Dans notre analyse :

- ✓ On a séparé les données en jeu d'entraînement et de test
- ✓ On a comparé la performance sur les deux jeux avec des métriques comme le RMSE (Root Mean Square Error)

Pourquoi c'est important ?

- Si le modèle fonctionne bien sur les deux jeux de données :
 - Il est robuste
 - Il peut faire de bonnes prédictions sur de nouvelles données
- Si le modèle est très bon en entraînement mais mauvais en test :
 - Il est probablement sur-appris



04. Choix des modèles

But principal :

➤ Combiner deux approches complémentaires :

- K-Means pour détecter des groupes cachés dans les données sans connaître les étiquettes (clustering non supervisé)
- Random Forest pour prédire de manière fiable si un billet est faux ou vrai (classification supervisée)

Dans notre analyse :

- ✓ K-Means a permis d'identifier des regroupements naturels entre billets vrais et faux, ce qui aide à explorer et visualiser les données
- ✓ Random Forest a montré de très bonnes performances en test, avec une robustesse au sur-apprentissage et une bonne interprétabilité (importance des variables)

Pourquoi c'est important ?

- On utilise K-Means pour repérer des comportements suspects sans supervision
- On utilise Random Forest pour confirmer la détection avec un modèle fiable et précis
- Ensemble, ils offrent une solution complète : exploration + prédiction robuste

Le KMeans

- Objectif :
 - Regrouper les clients en profils similaires (clustering non supervisé)
- Méthode :
 - ✓ Choix du nombre optimal de clusters grâce à la méthode du coude (elbow method)
 - ✓ Utilisation de KMeans pour regrouper les individus selon leurs caractéristiques
- Résultat :
 - Les clusters obtenus permettent d'identifier des segments de clients avec des comportements ou besoins différents
 - Visualisation des clusters en 2D pour interprétation simple

(Insérer le graphique des clusters + inertie ici)

Le Random Forest

- Objectif :
 - Prédire une variable cible à partir de plusieurs variables explicatives
- Pourquoi Random Forest ?
 - ✓ C'est un modèle puissant et robuste
 - ✓ Il limite le sur-apprentissage grâce à l'agrégation de plusieurs arbres de décision
 - ✓ Il permet de connaître l'importance des variables
- Résultat :
 - Bonnes performances sur les données de test
 - Mise en évidence des variables les plus influentes dans la prédiction

(Insérer la matrice de confusion et/ou l'importance des variables)

05. L'application



Le mieux n'est pas d'en parler mais de la tester !



The image features a light green background with a subtle grid pattern. Scattered around the edges are several stylized green banknotes, each with a white dollar sign (\$) in the center. The banknotes are depicted with wavy, irregular edges, giving them a hand-drawn or cut-out appearance. They are positioned primarily along the top and bottom borders of the frame.

MERCI !