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

Il manquait 37 valeurs dans la colonne margin_low. Pour prédire les valeurs de margin_low j'ai effectué un test du R2.
Le R2 permet de montrer si le modèle explique bien les données. Après avoir fait ce test je m'aperçois que la valeur diagonal n'apporte pas grand chose dans la prédiction de margin_low. Je décide donc de faire les prédictions de valeurs manquantes sans cette donnée. Finalement après avoir testé le modèle les valeurs manquantes sont créées.

Test du R2

Le R^2 mesure la qualité du modèle :

- Il est compris entre 0 et 1.
- Plus il est proche de 1, mieux le modèle explique les résultats.
- Ici $R^2 = 0.477$, ça veut dire que le modèle explique environ 47.7% de ce qui se passe dans les données.

Donc c'est un modèle moyennement bon.

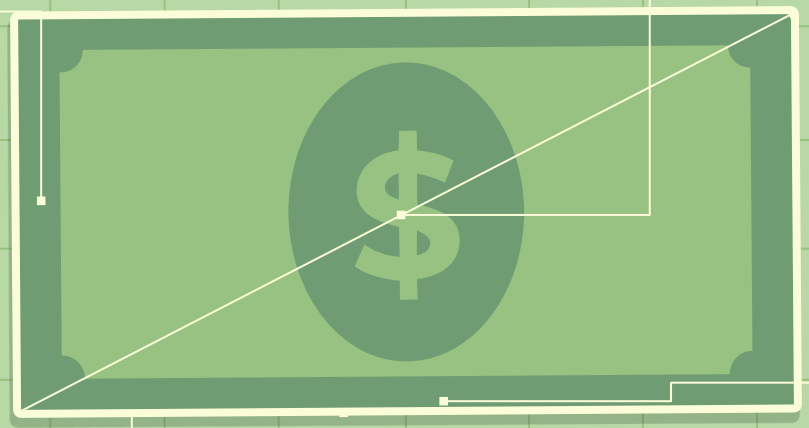
length est la variable la plus utile.

diagonal ne sert pas à grand-chose, donc je décide de faire sans.

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.

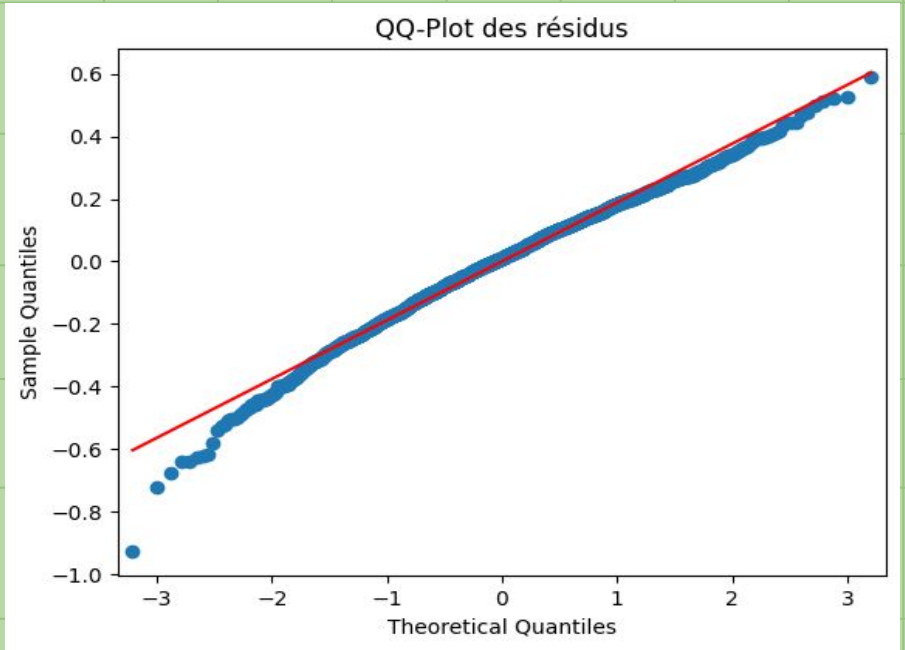
Dans notre analyse :
✔ On a vérifié la normalité des résidus à l'aide du QQ plot.
✔ On a contrôlé l'homoscédasticité (variance constante des résidus) avec le graphique des résidus vs valeurs ajustées.

Pourquoi c'est important ?
➢ Si les hypothèses sont respectées, alors :
- Les estimations sont fiables
- Le modèle est statistiquement solide

En bref :
Les erreurs doivent ressembler à du hasard, pas à un motif.
Si la variance reste stable et les erreurs sont normales, on peut faire confiance aux résultats du modèle.

Le QQ plot

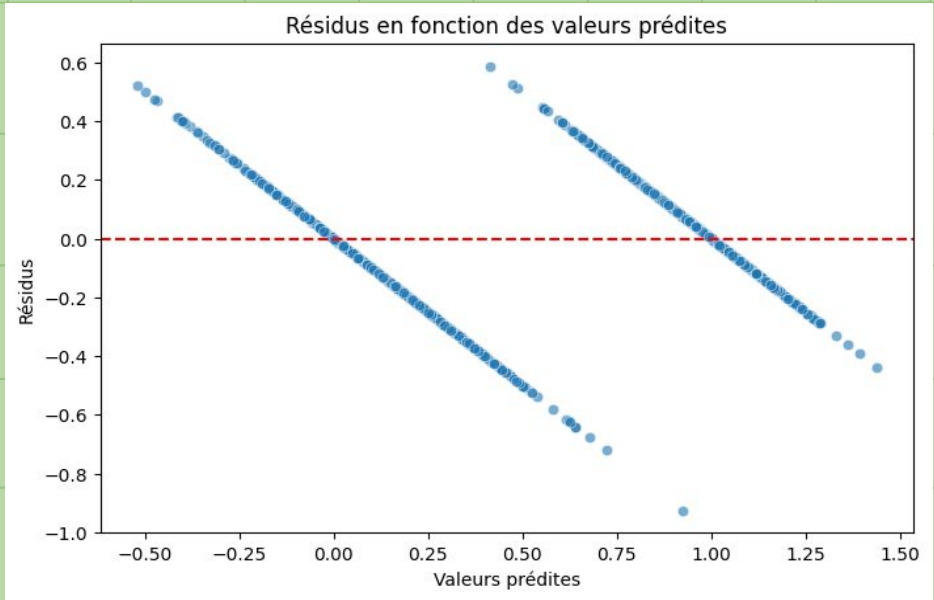
- Objectif : S'assurer que les erreurs (résidus) du modèle suivent une loi normale.
- Résultat : Les points du graphique suivent globalement la diagonale.
- Conclusion : L'hypothèse de normalité est globalement respectée, ce qui valide l'utilisation des tests statistiques du modèle.



Un QQ plot sert à vérifier si des données (comme les erreurs d'un modèle) suivent une courbe normale . Il compare les données qu'on a avec ce qu'on attendrait si elles étaient normales. Si les points du graphique forment une ligne droite, c'est bon signe. Ça veut dire que les erreurs sont "normales".

L’homoscédasticité

- 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.



Ce graphique permet de voir si les erreurs du modèle restent stables. On regarde si leur “taille” change selon les prédictions. Si les points sont un peu partout sans forme spéciale, c’est que le modèle est bon. Ça veut dire que les erreurs sont à peu près pareilles tout le temps. C’est ce qu’on veut.

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).
➢ Simplifier les données sans trop perdre d'information grâce à l'ACP.
➢ Séparer les données pour tester le modèle dans des conditions réalistes.
Dans notre analyse :
✔ On a séparé les données en un jeu d'entraînement et un jeu de test (train/test split)
✔ On a appliqué une ACP (Analyse en Composantes Principales) pour réduire la dimension du jeu de données
✔ On a comparé les performances du modèle sur les deux jeux de données
Pourquoi c'est important ?
➢ Si le modèle fonctionne bien sur les jeux de données :
  • Il est robuste
  • Il peut faire de bonnes prédictions sur de nouvelles données
➢ Si l'ACP donne de bons résultats :
  • Le modèle est plus simple, plus rapide, et plus stable
  • Il utilise uniquement l'information la plus utile des données
En bref :
On teste le modèle dans des conditions réalistes, avec des données nouvelles, et on l'aide à se concentrer sur l'essentiel grâce à l'ACP

L'ACP

Comment ça marche :

- L'ACP prend toutes les variables et cherche à en créer quelques nouvelles (appelées *composantes principales*) qui résument au mieux l'information.
- Ces nouvelles variables sont faites en combinant les anciennes, sans redondance.
- Elles sont classées : la 1re résume le plus de choses, la 2e un peu moins, etc.

Pourquoi c'est bien :

- Ça réduit le nombre de variables, donc le modèle est plus simple et plus rapide.
- Ça élimine les doublons d'information (ex : deux variables qui disent presque la même chose).
- Ça aide à éviter le sur-apprentissage, surtout avec beaucoup de variables.
- Ça peut améliorer la performance du modèle sur les données de test.

L'ACP simplifie le jeu de données sans trop perdre d'info, pour aider le modèle à mieux généraliser.
C'est comme résumer un long texte en gardant les idées principales.

Le train test

Comment ça marche :
On coupe le jeu de données en deux parties :

- Le jeu d'entraînement (train) sert à construire et ajuster le modèle.
- Le jeu de test sert à vérifier si le modèle fonctionne bien sur des données qu'il ne connaît pas.

L'idée, c'est de reproduire une situation réelle : on entraîne avec ce qu'on connaît, puis on teste avec du nouveau.

Pourquoi c'est bien :

- Ça permet de voir si le modèle généralise bien, c'est-à-dire s'il peut faire de bonnes prédictions ailleurs.
- Ça aide à détecter le sur-apprentissage : un modèle trop bon en entraînement mais mauvais en test a probablement mémorisé au lieu d'apprendre.
- C'est une étape clé pour valider la performance réelle du modèle.
- C'est simple à mettre en place, mais essentiel pour avoir un modèle fiable.

En bref :
Le train/test split, c'est comme s'entraîner à la maison (train) puis passer un vrai examen (test). Ça montre si le modèle est vraiment prêt.

04. Choix des modèles

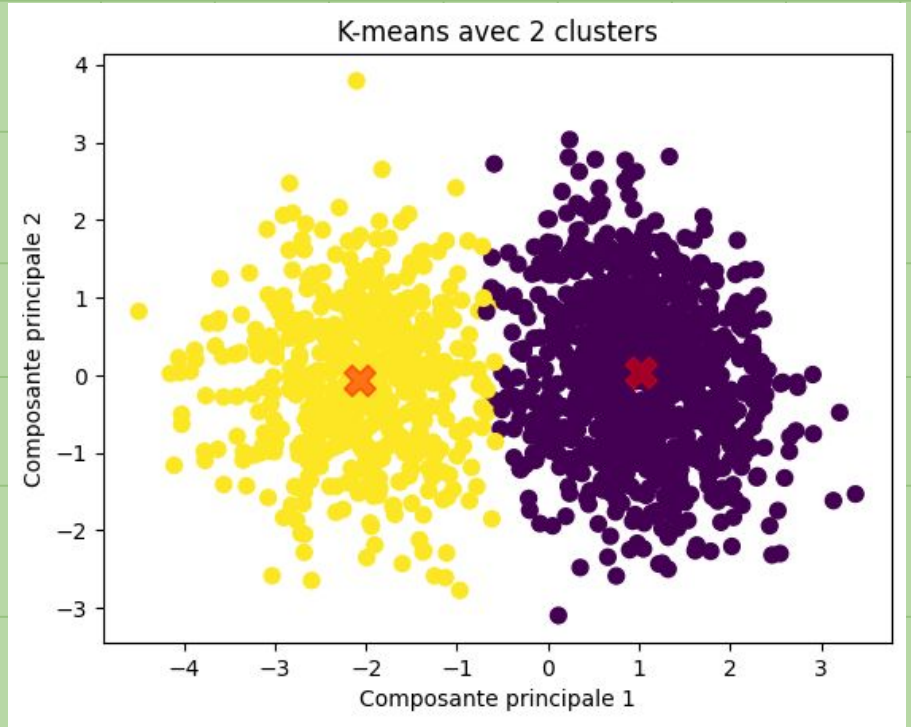
But principal :
➢ Combiner deux approches complémentaires :
• K-Means pour détecter des groupes cachés dans les données sans connaître les étiquettes (clustering non supervisé)
• Random Forest pour prédire de manière fiable si un billet est faux ou vrai (classification supervisée)

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



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

Accuracy: 0.9833333333333333
Rapport de classification :

	Precision	recall	f1-score	support
False	0.97	0.98	0.98	110
True	0.99	0.98	0.99	190
Accuracy			0.98	300
macro avg	0.98	0.98	0.98	300
weighted avg	0.98	0.98	0.98	300

Précision (accuracy) : C'est la proportion des prédictions positives correctes parmi toutes les prédictions positives. Par exemple, quand le modèle prédit "True", il le fait correctement dans 99% des cas.
Recall : C'est la proportion des vrais positifs correctement identifiés par rapport au nombre total de vrais positifs. Le modèle détecte 98% des billets vrais comme étant "True".
F1-score : C'est la moyenne harmonique de la précision et du rappel. Il donne une bonne idée de la performance générale du modèle, et ici il est très élevé (99% pour les billets vrais, 98% pour les faux)

05. L’application

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


MERCI !