Automatisation de la détection d'anomalies financières

Maha DAO, Cédric GRANDVAUX, Lucas HENRIO, Enzo MARTINEZ

Ecole Nationale de la Statistique et de l'Analyse de l'Information

Le point méthodologique

Mots-clés :

- **Score d'anomalie :** Compris entre 0 et 1, il indique le niveau d'anomalie du contrat. Plus le score est proche de 1, plus le contrat pourra être considéré comme étant anormal.
- **Seuil d'anomalie :** Score d'anomalie à partir duquel un contrat sera considéré comme étant anormal.
- **Variable pertinente :** Variable de la base de données utilisée dans l'implémentation des algortihmes.

Analyse des résultats

Les algorithmes divisent donc les individus statistiques en deux catégories : les contrats normaux et les contrats anormaux. De ce fait, les boîtes de dispersion constituent une méthode statistique pertinente pour comparer ces deux populations vis à vis des variables pertinentes.

Isolation Forest

Il s'agit d'un algorithme d'**apprentissage non supervisé** qui repose sur le calcul d'un **score d'anomalie** pour détecter des données atypiques d'un jeu de données. On peut représenter schématiquement son fonctionnement :

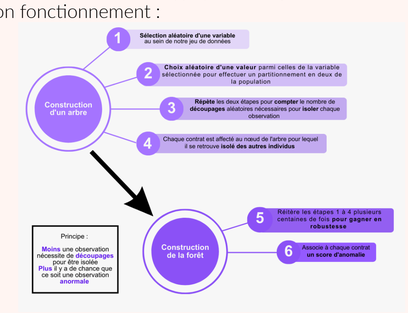

Figure 1. Principe de l'Isolation Forest

À l'aide du seuil d'anomalie préalablement fixé, l'algorithme catégorisera chaque contrat du jeu de données comme étant normal ou anormal.

Local Outlier Factor

C'est une méthode concurrente de détection d'anomalies à portée **locale**. Elle fait l'hypothèse que les valeurs abberrantes ont une **densité** plus faible que les valeurs dites normales et compare la densité d'un point avec celle de ses voisins. On peut synthétiser l'algorithme en quatre grandes étapes :

Figure 2. Etapes du Local Outlier Factor

Cet algorithme est plus **restrictif** dans son implémentation. En effet, à l'inverse de l'Isolation Forest, le LOF n'accepte **que les variables quantitatives** qui ne contiennent **pas de valeurs manquantes** (NA).

Classification And Regression Trees

Ces arbres de décision vont procéder à des **divisions** successives **binaires** de l'échantillon donné. Un noeud parent est ainsi partitionné en deux noeuds enfants. La division effectuée rend compte de la plus forte **diminution** de l'**héterogénéité** entre les **deux noeuds enfants**.

Objectif

Le pôle Modèles Portefeuilles fait partie du département Modélisation de la direction des risques du groupe bancaire BPCE. Le coeur du travail de ce pôle repose sur la projection des risques de crédit. La qualité et la fiabilité des données sont donc primordiales pour leurs missions.

Notre objectif était d'implémenter un algorithme non supervisé de détection d'observations financières anormales.

Contextualisation

Pour implémenter l'algorithme, on nous a partagé une base de données de contrats d'emprunts segmentée selon la norme comptable IFRS9.

La **norme IFRS9** distingue trois types de contrats :

1. Les contrats sains, S1
2. Les contrats sains mais ayant connu une dégradation, S2
3. Les contrats en défaut, S3

La base de données contenait plus de 900 000 observations auxquelles correspondaient une centaine de variables.

Ces trois types de contrats n'étant pas comparables entre-eux, nous avons implémenté les algorithmes de détection d'anomalies de manière séparée en créant un jeu de données pour chaque type de contrat. Les contrats de type S1 étant les plus nombreux (761 757 contrats), nos analyses se sont davantage focalisées sur ces derniers.

Afin de gagner en pertinence dans le processus de détection d'anomalies, nous avons effectué une liste de variables pertinentes. Les valeurs prises par ces variables apportent les informations nécéssaires pour juger de la qualité d'un contrat. Parmi elles, nous retrouvons :

- Les **pertes attendues** (*EL*)
- Les **probabilités de défaut** des clients (*PD*)
- L'**exposition au défaut** de chaque client (*EAD*)
- Les **pertes attendues en cas de défaut** pour chaque contrat (*LGD*)

Ces variables sont reliées entre elles selon la relation suivante :

$$EL = PD \times EAD \times LGD$$

Cette relation établie nous a permis d'identifier les cas où elle était significativement différente pour des contrats répondants aux mêmes critères qualitatifs.

Conclusion

→ Il s'agissait de mettre en pratique nos compétences en data science dans le cadre bancaire et financier
→ L'**Isolation Forest** permet de détecter des valeurs anormales à échelle **globale**, alors que le **Local Outlier Factor** agit à échelle **locale** et est plus **restrictif** dans son implémentation
→ Le **code** R a été **automatisé** pour être réutilisé
→ Dans notre cas, les observations identifiées comme **anormales** le sont en raison de **valeurs aberrantes**

Pour aller plus loin...

La base de données sur laquelle nous avons travaillé ne nous a pas permis d'établir des règles de prédiction : un futur projet pourrait s'interroger sur l'implémentation de l'algorithme *Partial Least Squares* par exemple.

Une autre piste peut concerner la comparaison des sorties de l'Isolation Forest et du Local Outlier Factor. Pour ce faire, il peut être intéressant de reprendre notre projet et d'approfondir plus particulièrement l'algorithme de réduction de dimension *UMAP - Uniform Manifold Approximation and Projection* à partir d'une base de données où il existerait un motif de séparation entre outliers et non outliers.

Sortie Isolation Forest pour les contrats S1

Parmi les 761 757 contrats S1 du jeu de données, l'algorithme a détecté **235** contrats anormaux.

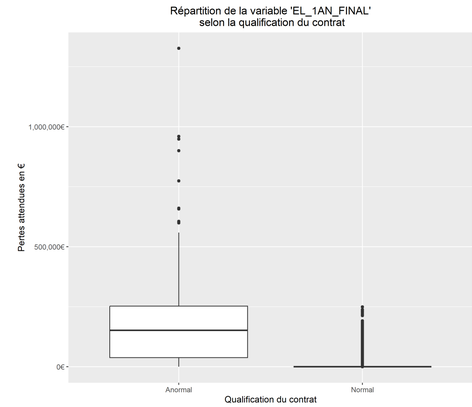

Figure 3. Boxplot pour la variable *EL_1AN_FINAL* pour les contrats S1 selon leur type de qualification

Les pertes attendues à un an (*EL_1AN_FINAL*) des contrats anormaux représentent des montants aberrants comparé à la population des contrats normaux.

Ce constat est applicable pour l'ensemble des variables pertinentes. Ainsi, **l'Isolation Forest isole principalement des contrats ayant des valeurs aberrantes pour au moins une des variables pertinentes.**

Le LOF, une approche concurrente pour la prédiction ?

Les deux méthodes s'opposent sur plusieurs aspects. Tout d'abord, la **volumétrie des anomalies** est **beaucoup plus importante** pour le Local Outlier Factor, qui procède à une **détection locale** alors que l'Isolation Forest opère au niveau global.

Dans une tentative de **prédiction** nous avons implémenté un arbre CART basé sur les détections du LOF pour les contrats S2 anormaux. Cette sortie doit toutefois être interprétée au regard de sa **matrice de confusion**.

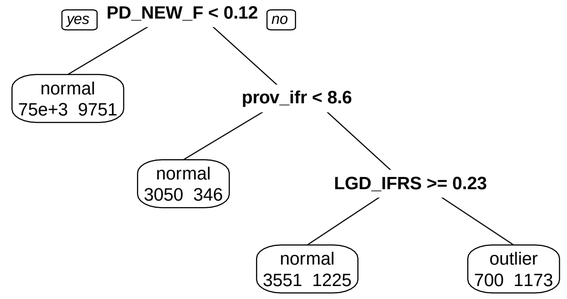

Figure 4. Arbre CART construit à partir des outliers S2 désignés par le LOF.

	normal	outlier
normal	9008	102
ourlier	1253	136

Table 1. Matrice de confusion de l'arbre CART pour le LOF