Prédiction du risque de crédit bancaire sensible aux coûts financiers en intégrant des descripteurs extraits des graphes

Victor Nico DJIEMBOU TIENTCHEU[1] **and Armel Jacques NZEKON NZEKO'O**[1]

[1]Université de Yaoundé I, Cameroun

*E-mail : Victor nico.djiembou@facsciences-uy1.cm, Armel armel.nzekon@facsciences-uy1.cm

---

Résumé

Les prêts sont des opérations financières très importantes pour le développement et la croissance économique d'un pays, car ces derniers facilitent la création et la croissance des entreprises et donc l'emploi de plus de personnes tant par les entreprises privées, publiques ou parapubliques. Les non remboursements des prêts ont des coûts importants sur les institutions financières prêteuses, pouvant entraîner leur faillite et donc détruire tout le système de prêt et constituer par là, un frein au développement économique. Il est donc nécessaire de pouvoir prédire efficacement si un prêt sera remboursé ou non par l'emprunteur. A cet effet, la question de prédiction du risque de crédit est devenu un domaine majeur dans lequel des chercheurs en Apprentissage automatique proposent des modèles qui prédisent la classe d'un prêt à partir des attributs standards qui le décrivent dans l'institution préteuse. Ces attributs standards n'étant pas suffisants pour avoir les meilleures prédictions, ces dernières années, plusieurs travaux portent sur la création de nouveaux attributs descriptifs à mettre en entrée des modèles classiques de prédiction dans le but d'améliorer leur performance. C'est le cas des récents travaux sur l'extraction de nouveaux descripteurs des prêts modélisés par un graphe multicouches dans lesquels une seule application du PageRank personnalisé sur le graphe multicouches permet d'extraire les nouveaux descripteurs des différents prêts considérés. Les travaux actuels sur les graphes multicouches ont pour limites de ne pas être suffisamment personnalisés par prêt, de ne pas considérer les classes des prêts dans le processus de construction du graphe lors de l'apprentissage, de ne pas proposer de stratégie pour le choix des attributs à considérer comme couches du graphe construit et enfin de ne pas évaluer leur impact sur les coût financiers qui sont un aspect important pour les institution prêteuses. Dans ce mémoire, nous proposons d'intégrer les classes des prêts dans le processus de construction des graphes multicouches, et d'appliquer le PageRank personnalisé par prêt pour extraire les nouveaux descripteurs des ces graphes. Par ailleurs, nous proposons un protocole de sélection des attributs à considérer comme couches du graphe multicouches, et effectuons une évaluation des coûts financiers des modèles de prédiction du risque de crédit construits à partir des données enrichies par les nouveaux descripteurs. Des expérimentations sont menées sur 04 jeux de données, en considérant 06 modèles classiques de prédiction du risque de crédit et 03 métriques d'évaluation des performances des modèles, dont l'une sensible aux coûts financiers. Des valeurs de SHapley sont considérés pour évaluer l'importance des nouveaux descripteurs. Nous observons que le protocole de sélection des attributs à considérer comme couches du graphe multicouches permet d'avoir des performances très proches des meilleures et même les meilleures dans X% des cas. Par ailleurs, la considération des classes dans la construction du graphe multicouches et l'application du PageRank personnalisé par prêt permettent non seulement de diminuer significativement les pertes financières jusqu'à A%, B%, C% et D%, et les nouveaux descripteurs associés occupent W%, X%, Y% et Z% du Top-10 des meilleures valeurs de SHapley dans les 04 jeux de données considérés.

Mots-Clés
Prédiction du risque de crédit ; Graphe multicouches ; PageRank personnalisé ;
Apprentissage automatique ; SHapley.

I INTRODUCTION

1.1 Contexte d'application et d'intérêt

Les prêts sont des opérations financières très importantes pour le développement et la croissance économique d'un pays, car ces derniers facilitent la création et la croissance des entreprises et donc l'emploi de plus de personnes tant par les entreprises privées, publiques ou parapubliques. Que ce soit Tayo au Nigeria en 2020, Tazi.ai au Maroc en 2017 ou encore Kiro'o Games au Cameroun en 2013, tous ses entreprises de renommé actuel ont construit sur la base de prêts financiers. Elles sont aujourd'hui des piliers de l'économie de leur pays.

Le rapport annuel de la banque mondiale en 2023[**WorldBank2023**] montre que le volume total des prêts accordés à l'échelle mondiale a atteint un niveau record de 62 987 milliards de dollars, en hausse de 8,4% par rapport à 2022. Cette croissance est tirée par une demande accrue de financement de la part des entreprises et des ménages, ainsi que par des politiques monétaires accommodantes mises en place par de nombreuses banques centrales.

1.2 Transition vers la problématique

Cette hausse de volume de prêt est une opportunité pour la croissance économique mais un risque considérable lorsque nous savons que les non remboursements des prêts ont des coûts importants sur les institutions financières prêteuses, pouvant entraîner leur faillite et donc détruire tout le système de prêt et constituer par là, un frein au développement économique[**10.1093/oep/gpac013**].

La crise des subprimes en est un exemple pertinent qui a débuté aux États-Unis en 2007 et a conduit à une grave récession mondiale en 2008-2009. Ses origines remontent au début des années 2000, lorsque le marché hypothécaire américain a connu une bulle spéculative alimentée par une distribution massive de crédits immobiliers à haut risque, dits "subprimes", à des emprunteurs peu solvables[**CriseDesSubprimes2008**]. A cet effet, la question de prédiction du risque de crédit est devenu un domaine majeur dans lequel des chercheurs en apprentissage automatique proposent des modèles qui prédisent la classe d'un prêt à partir des attributs standards qui le décrivent dans l'institution préteuse.

Mais ses travaux se heurte aux faits que les données possède déjà très peu de prêt non remboursé et aussi, l'insuffisance des attributs standards pour former des modèles efficace. Alors, ces dernières années, plusieurs travaux se concentrent plutôt sur la création de nouveaux attributs descriptifs à mettre en entrée des modèles classiques de prédiction dans le but d'améliorer leur performance. C'est le cas des récents travaux sur l'extraction de nouveaux descripteurs des prêts modélisés par le graphe multicouches dans lesquels une seule application du PageRank personnalisé sur le graphe multicouches permet d'extraire les nouveaux descripteurs des différents prêts considérés.

De plus ces travaux actuels posent des limites, notamment le fait de ne pas être suffisamment personnalisés par prêt, de ne pas considérer les classes des prêts dans le processus de construction du graphes lors de l'apprentissage, de ne pas proposer de stratégie pour le choix des attributs

à considérer comme couches du graphe construit et enfin de ne pas évaluer leur impact sur les coûts financiers.

1.3 Problème

Il est donc question pour nous dans ce mémoire de trouver comment proposer à la fois une façon d'améliorer la personnalisation du PageRank, prendre en compte la décision de prêt dans la modélisation graphe multicouches, de sélectionner les attributs descriptives optimales pour la construction du graphe et d'évaluer les coûts financiers des modèles construits.

1.4 Objectif

Dans ce mémoire, nous travaillons sur la prédiction du risque de crédit bancaire en nous focalisant sur l'extraction d'attribut de graphe multicouches construit sur la base d'historique de crédit bancaire d'institution préteuse par exécution d'un algorithme de PageRank personnalisé. Par ailleurs, nous avons pour objectifs d'intégrer la classe dans le graphe multicouches construit lors de l'apprentissage. Dans la suite nous allons proposer un PageRank personnalisé par prêt pour extraire des attributs du graphe multicouhes. Enfin, Nous proposons une stratégie de sélection d'attributs à considérer comme couche dans la construction du graphe multicouches basé sur les coûts financiers qui va permettre d'estimer l'impacte des modèles de prédiction construits sur l'évaluation du risque de crédit par les institutions préteuses.

1.5 Contribution

Notre contribution pour ce mémoire se décrit comme suit :

- Proposition d'une métrique de coûts financiers pour les institutions préteuses.
- Proposition d'un cadre de modélisation de graphe multicouches qui intègre les classes de prêts ;
- Proposition d'un nouveaux PageRank personnalisation par prêts pour le graphe multicouches ;
- Proposition d'une stratégie de sélection d'attributs à considérer comme couche dans le graphe multicouches ;

1.6 Plan du mémoire

Le reste de ce mémoire se présentera comme suit, dans la section II nous présenterons l'état de l'art sur la prédiction du risque de crédit financier. La section III présente notre solution, un cadre d'extraction de nouveaux descripteurs par application du PageRank personnalisé par prêt sur des graphes multicouches construit en intégrant les classes. La section IV présente notre cadre expérimentale. Et enfin, en section V nous conclurons notre travail.

II PRÉDICTION DU RISQUE DE CRÉDIT

2.1 Définition du problème de prédiction du risque de crédit

Le problème de prédiction du risque de crédit consiste à estimer la probabilité qu'un emprunteur ne rembourse pas son prêt dans les délais impartis. C'est un enjeu majeur pour les institutions financières afin de minimiser leurs pertes et de maintenir la santé de leur portefeuille de crédits. Ce problème peut être formalisé mathématiquement comme suit :

Soit un ensemble d'emprunteurs $\mathcal{B} = b_1, b_2, ..., b_n$. Chaque emprunteur b_i est caractérisé par un vecteur de variables explicatives $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ qui décrivent son profil (revenus, historique de crédit, etc.).

L'objectif est de prédire la variable binaire $y_i \in 0, 1$ qui indique si l'emprunteur b_i rembourse son prêt ($y_i = 0$) ou non ($y_i = 1$) dans les délais. Formellement, on cherche à construire un modèle de prédiction $\hat{y}_i = f(\mathbf{x}_i)$ qui estime au mieux la probabilité de défaut $P(y_i = 1|\mathbf{x}_i)$ pour chaque emprunteur b_i.

2.2 Techniques classiques de prédiction du risque de crédit

Les premiers modèles de prédiction du risque de crédit étaient principalement basés sur des techniques statistiques telles que la régression logistique (LR). Ces modèles, comme l'ont montré [**khemais2016credit**], permettent d'obtenir des résultats satisfaisants en termes de prédiction, tout en étant relativement simples à mettre en œuvre. Cependant, ils reposent sur des hypothèses restrictives comme la linéarité des relations et l'indépendance des variables explicatives.

Pour pallier ces limitations, des approches plus avancées ont été développées, notamment les arbres de décision (DT) et les machines à vecteurs de support (SVM) [**dastile2020statistical**]. Ces modèles, plus flexibles, permettent de capturer des relations non linéaires et des interactions complexes entre les variables. Néanmoins, ils peuvent souffrir d'un manque de robustesse et nécessitent un réglage fin des paramètres.

Plus récemment, les techniques d'apprentissage automatique telles que les forêts aléatoires (RF) [**ziemba2020client**, **dastile2020statistical**] et le boosting de gradient extrême (XGB) [**dastile2020statistical**, **mushava2022novel**] ont démontré leur efficacité pour la prédiction du risque de crédit. Ces modèles de "boîte noire" sont capables de capturer des patterns complexes dans les données, tout en étant plus robustes que les approches précédentes. Cependant, leur manque de transparence peut rendre l'interprétation des résultats plus difficile.

Dans ses travaux, Les modèles de prédictions sont formé sur des données déséquilibrés car dans la réalité, les historiques des institutions financières prêteuses ne renseignent pas assez sur la descriptions des prêts qui étaient à défaut au vu du nombre réduit des cas existant. Des techniques d'augmentation de données par mélange synthétique des existantes comme SMOTE, ADASYN, BSMOTE, ROS, etc... [**barabasi2013network**, **lenka2022empirical**, **anis2017investigating**, **mushava2022novel**] permettent d'atténuer l'effet de ce déséquilibre. Toutefois, les modèles formés n'étant toujours pas très performance, de nouveaux travaux portent leur intérêt sur l'évaluation du risque de crédit sur des données autres que celles de décrivant les prêts par les institutions préteuses.

2.3 Nouveaux descripteurs pour la prédiction du risque de crédit

Présenter ici les autres natures de données alternatives utilisés pour la prédiction du risque de crédit (3 paragraphes chronologiques) et comment elles sont utilisées pour la prédiction du risque de crédit.

Présenter l'alternative portée sur la modélisations des prêts en graphes.

2.4 Évaluation de la pertinence des nouveaux descripteurs

2.5 Évaluation des coûts financiers des modèles de prédiction du risque de crédit

III EXTRACTION DE NOUVEAUX DESCRIPTEURS PAR APPLICATION DU PAGERANK PERSONNALISÉ PAR PRÊT SUR DES GRAPHES MULITCOUCHES CONSTRUIT EN INTÉGRANT LES CLASSES

3.1 Modèles d'apprentissage supervisé

3.1.1 Définition

3.1.2 Linear Discriminative Analysis (LDA)

3.1.3 Logistic Regression (LR)

3.1.4 Support Vector Machine (SVM)

3.1.5 Decision Tree (DT)

3.1.6 Random Forest (RF)

3.1.7 eXtreme Gradient Boosting (XGBoost)

3.2 Graphes multicouches biparti

3.3 Analyse des graphes

3.4 Évaluation des coûts financiers

3.5 Construction du graphes multicouches intégrant les classes

3.6 Proposition du PageRank Personnalisé par prêt

3.7 Protocole de sélection des attributs à considérer dans la construction du graphe

IV EXPÉRIMENTATIONS

4.1 Description du jeux de données

AFB, CREDIT RISK, GERMAN, JAPAN (nombre de ligne (exemple), nombre de colonnes, nombre de colonnes numériques, nombre d'attributs catégoriel, nombre d'exemples positif, nombre d'exemples négatifs)

4.2 Évaluation et paramétrage de modèles

4.2.1 Métriques d'évaluation

Acc + F1 + Cost + SHapLey

4.2.2 Approches de construction du graphes

Approches

- MlC
- MCA

4.2.3 Logiques de personnalisation personnalisation

Logiques

- GLO
- PER
- GAP

4.2.4 Configuration de descripteurs

Configs

- MX
- CX
- CY
- CXY

4.2.5 paramétrages des algorithmes

4.3 Résultats

SHAP + Tableaux

V CONCLUSION

5.1 rappel du problème abordé

Il était question pour nous dans ce mémoire de trouver comment proposer à la fois une façon d'améliorer la personnalisation du PageRank, prendre en compte la décision de prêt dans la modélisation graphe biparti multicouches, de sélectionner les attributs descriptives optimales pour la construction du graphe et mettre ce pieds un métrique de

5.2 Idée de solution

Prise en compte des décisions dans la modélisation graphe biparti multicouches et personnalisation du PageRank à un seul emprunteur pour une prédiction du risque de crédit sensible aux coûts financiers

5.3 Démarche

proposer un PageRank personnalisé porté sur un seul emprunteur à la fois

proposer un protocole qui va permettre d'identifier les relations les plus pertinentes à analyser

proposer une cadre de modélisation graphe biparti multicouches qui incorpore les informations de décision des prêts historique.

une métrique de coûts financiers

5.4 Les principaux résultats

5.5 Les perspectives

prendre en compte les données numérique dans la modélisation

proposer de nouvelles stratégies d'exploitation du graphes autres que le PageRank personnalisé.

RÉFÉRENCES

A ANNEXE 1

Dans Bibtex, comment écrire une citation d'un article de ARIMA (example : **arima**) sans rien oublier et dans le bon format ? Voir la structure dans les commentaires à la fin de *arima.tex*.

B REMERCIEMENTS

Nous tenons à remercier tous nos partenaires financiers : ANR ..., ERC ..., agences de financement, ...

C BIOGRAPHIE

Il est possible ici d'insérer de courtes biographies des auteurs.