AVIS DE SOUTENANCE

THESE DE DOCTORAT

Présentée par

Mme : SALMA EL HAJJAMI

Discipline **:** Informatique

Spécialité : Informatique

Sujet de la thèse : Elaboration d'une technique d'apprentissage automatique pour la détection de fraude par carte de crédit : Application sur les données massives déséquilibrées.

Formation Doctorale : Sciences de l'ingénieur Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le samedi 26 juin 2021 à 11h au Centre de conférences devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Nour El Houda CHAOUI	PES	Ecole Nationale des Sciences Appliquées de Fès	Président
Ahmed ZELLOU	PES	École Nationale Supérieure d'Informatique et d'Analyse des Systèmes Rabat	Rapporteur
El Habib NFAOUI	PES	Faculté des Sciences Dhar El Mehraz de Fès	Rapporteur
Youssouf EL ALLIOUI	PH	Faculté Polydisciplinaire de Khouribga	Rapporteur
Aziza EL OUAAZIZI	PH	Faculté Polydisciplinaire de Taza	Examinateur
Mohamed BENSLIMANE	PH	Ecole Supérieure de Technologie de Fès	Examinateur
Khalil MAALMI	PES	Ecole Supérieure de Technologie de Fès	Examinateur
Mostafa HARTI Mohammed BERRADA	PES PH	Faculté des Sciences Dhar El Mehraz de Fès Ecole Nationale des Sciences Appliquées de Fès	Directeurs de thèse

Laboratoire d'accueil : LIASSE.

Etablissement : Ecole Nationale des Sciences Appliquées de Fès

Titre de la thèse : Elaboration d'une technique d'apprentissage automatique pour la détection de fraude par carte de crédit : Application sur les données massives déséquilibrées.

Nom du candidat : Salma EL HAJJAMI

Spécialité : Informatique

Résumé de la thèse

A l'ère numérique d'aujourd'hui, les cartes de crédit sont utilisées dans pratiquement toutes les transactions. La facilité de paiement en ligne au cours de la dernière d´décennie, a ouvert une nouvelle passerelle pour les opportunités, les entreprises, l'abaissement des frontières géographiques, mais aussi l'augmentation des transactions frauduleuses. Les transactions frauduleuses peuvent ˆêtre effectuées par le biais de diverses tentatives telles que le vol de cartes de d´débit et de crédit, l'utilisation abusive d'autres cartes et comptes en ligne et le piratage du nom d'utilisateur et de son mot de passe. Alors que le pic du graphique du commerce ´électronique ne cesse d'atteindre de nouveaux sommets chaque année, il est ´également devenu la toile de fond des fraudeurs qui abusent fortement de la transparence des m´méthodes de paiement en ligne et des transactions par carte de crédit. Par conséquent, la d´détection de fraudes est un aspect essentiel de l'identification des transactions frauduleuses qui ont ´été ´émises illégalement au nom du propriétaire l´légitime de la carte et de la correction des pertes. La d´détection de fraude représente le d´défi de trouver des anomalies dans les données de transactions. En général, la d´détection d'anomalies (AD) est le problème de la distinction entre les instances de données normales avec des motifs ou signatures bien d´définis et ceux qui ne sont pas conformes aux profils attendus. Le comportement frauduleux peut se manifester par des modèles inhabituels dans les transactions financières.

Les progrès récents des approches d'intelligence artificielle, comme l'apprentissage automatique ou Machine Learning (ML), font une percée importante dans ce domaine grâce à leur travail en temps réel, leur vitesse et leur efficacité. La d´détection des fraudes qui était autrefois un domaine qui n´nécessitait beaucoup de travail manuel, est maintenant l'un des nombreux à avoir connu les progrès grâce à la capacité des algorithmes de ML à apprendre des modèles de fraude historiques et à les reconnaitre dans les transactions futures. Néanmoins, la conception d'algorithmes de d´détection de fraude est sujette `a divers problèmes, qui ont un impact considérable sur la précision du système. Le travail de recherche présente dans cette thèse concerne l'élaboration d'une technique `a base d'apprentissage automatique pour la d´détection d'anomalies dans les transactions bancaires. Nous traitons un problème majeur rencontré par les m´méthodes d'apprentissage dans la d´détection de fraudes qui concerne la distribution asymétrique des classes, également connue sous le nom de déséquilibre des classes, qui se produit lorsqu'une des classes a plus d'instances qu'une autre classe. C'est particulièrement le cas du domaine de d´détection de fraudes, en effet nous rencontrons souvent des transactions frauduleuses minoritaires qui sont mal représentées lors de la phase d'apprentissage, ce qui perturbe les algorithmes ML qui présentent une grande précision sur les instances normales mais une faible précision sur les instances frauduleuses. Ce problème est plus vulnérable lorsque nous le considérons dans le contexte des données massives "Big Data". Afin d'apporter des éléments de réponse pour cette problématique, nous proposons différentes contributions. Nous avons fourni comme première contribution, une étude expérimentale de l'influence du déséquilibre de classes sur les performances de prédiction des

algorithmes d'apprentissage automatique pour comprendre les véritables faiblesses. Nous avons montré l'impact du déséquilibre des données sur la performance de ces approches. Ainsi, à partir des résultats obtenus, nous notons l'existence d'un ensemble de paramètres adaptés aux modèles supervisés et d'autres aux modèles non supervisés. Dans une seconde contribution, nous avons proposé une m´méthode qui agit au niveau des données d'apprentissage appelée : One Side Behavioral Noise Reduction (OSBNR). C'est une technique d'échantillonnage qui permet de rééquilibrer les données en sous-échantillonnant la classe majoritaire. OSBNR est basée sur deux ´étapes. Tout d'abord, une analyse de cluster est appliquée à des groupes d'instances similaires de la classe minoritaire en plusieurs clusters de comportement. Deuxièmement, nous s´sélectionnons et éliminons les instances de la classe majoritaire, considérées comme du bruit comportemental, qui se chevauchent avec les clusters de comportement de la classe minoritaire. Enfin, nous avons proposé une m´méthodologie générale basée sur la formalisation, la mise en œuvre et la conception technique d'une plateforme de d´détection de fraude en mesure de satisfaire les conditions de travail dans le monde réel. La partie formalisation se compose de quatre couches : le stockage distribué, le traitement des données, la construction du modèle et enfin l'évaluation du modèle. La partie implémentation utilise le système de traitement de données distribué Spark. En particulier, nous nous sommes basée sur son cadre dédié au ML, appelé MLlib. La partie conception technique de la plateforme s'appuie sur des solutions open-source issues du monde du Big Data telles que Hadoop, Yarn, Livy.

Mots-clés : Détection de fraude, Détection d'anomalie, Apprentissage automatique, Déséquilibre de classe, Big Data, Méthodes d'échantillonnage, plates-formes Spark et Hadoop