



Institut National de Statistique
et d'Economie Appliquée



Centre des Etudes Doctorales
Sciences, Ingénierie
et Développement Durable

Avis de soutenance de thèse de Doctorat

Monsieur Monir EL ANNAS

Soutiendra publiquement sa thèse de Doctorat
le lundi 20 février 2023 à 10h
à la salle de conférence de l'INSEA

Intitulé de la thèse

« **Développement d'un modèle d'apprentissage
semi-supervisé pour la notation de crédit** »

Devant le jury composé de :

Président :

Pr. Professeur ZOGLAT Abdelhak, PES, Faculté des Sciences Med V - Rabat

Directeur de thèse :

Pr. Mohamed OUZINEB, PES, INSEA

Co-Directeur de thèse :

Pr. Badreddine BENYACOUB, PA, INSEA (invité)

Membres du jury :

Pr. Mohammed EL HAJ TIRARI, PES, INSEA

Pr. Ahmed EL GHINI, PES, Faculté des Sciences Juridiques, Economiques et Sociales de
Rabat

Pr. Rachid SAADAN, PES, EHTP

Pr. Imad EL HARRAKI, PH, ENSMR

Résumé

Dans ce projet de thèse, nous nous intéressons à l'application des modèles d'apprentissage automatique en crédit scoring s'inscrivant dans le cadre des approches dédiées à l'évaluation des risques-clients. Les modèles classiques de scoring ne tiennent pas en compte d'informations des demandes de clients rejetées dès le départ, ce qui est susceptible d'engendrer un biais de sélection. Par ailleurs, les dossiers refusés des demandes de prêt sont très majoritaires et leurs stratus de remboursement (bon/mauvais) sont inconnus. Ainsi, pour remédier à ce problème, nous utilisons les modèles d'apprentissage semi-supervisés pour traiter le problème d'inférence des demandes rejetées permettant leur réintégration dans le processus d'élaboration des modèles de crédit scoring.

Plus spécifiquement, nous développons dans ce travail un modèle semi-supervisé basé sur les chaînes de Markov cachées (HMM). A cet effet, le problème de l'inférence de demandes rejetées est modélisé par la structure discrète du modèle HMM, présentant un cadre probabiliste rigoureux pour l'élaboration d'estimateurs performants pour les données manquantes.

L'approche HMM semi-supervisée proposée, notée par SSHMM, permet de remédier au problème de biais de sélection et tenir en compte d'autant d'informations dans le processus d'apprentissage. La plupart des travaux proposés récemment dans la littérature utilisent les données de la plateforme Lending Club pour valider et tester leurs approches semi-supervisées qui ont été développées, y compris notre modèle SSHMM. Il s'agit d'une plateforme en ligne de prêt pair-à-pair la plus utilisée aux États-Unis, qui permet de créer de la valeur en connectant des emprunteurs et des prêteurs et de réduire les coûts de transaction et de recherche d'une contrepartie.

Abstract

In this thesis project, we are interested in the application of machine learning models in credit scoring within the framework of approaches dedicated to customer risk assessment. Classical credit scoring models do not take into account information about rejected applications, which may lead to a selection bias. In contrast, rejected applications represent the majority of loan applications and their repayment status (good/bad) is unknown. Thus, to address this problem, we use semi-supervised learning models for inferring rejected applications, allowing them to be reintegrated into the credit scoring model building process.

More specifically, we develop in this work a semi-supervised model based on hidden Markov models (HMM). To this aim, the problem of inferring rejected applications is modelled by the discrete structure of the HMM model, presenting a rigorous probabilistic framework for developing efficient estimators for missing data. The proposed semi-supervised HMM approach, denoted by SSHMM, addresses the selection bias problem and takes into account as much information in the learning process. Most recent proposed works in the literature use data from the Lending Club platform to validate and test their developed semi-supervised approaches, including our SSHMM model. Lending Club is the most widely used online peer-to-peer lending platform in the US, which creates value by connecting borrowers and lenders and reduces transaction and matching costs.

ملخص

في مشروع الأطروحة هذا، أولينا اهتمامنا لتطبيق نماذج التعلم الأوتوماتيكي في التصنيف الائتماني والذي يندرج في إطار المقاربات المخصصة لتقييم مخاطر الزبناء. إن نماذج التنقيط التقليدية لا تأخذ في الاعتبار المعلومات الواردة من طلبات الزبناء المرفوضة منذ البداية، والتي من المحتمل أن تولد تحيزًا في الاختيار. بالإضافة إلى ذلك، فإن ملفات طلبات القروض المرفوضة تُشكل أغلبية كبيرة ووضعية السداد غير معروفة (جيدة / سيئة). ومن أجل معالجة هذه المشكلة، استخدمنا نماذج التعلم شبه الخاضعة للإشراف للتعامل مع مشكلة الاستدلال على الطلبات المرفوضة مما مكننا من إعادة إدماجها في عملية تطوير نماذج التقييم الائتماني. وبصفة مُحددة، طورنا في هذا العمل نموذجًا شبه خاضع للإشراف مبني على سلاسل ماركوف المخفية (HMM). وتحقيقًا لهذه الغاية، تم القيام بنموذج مشكلة الاستدلال من الطلبات المرفوضة من خلال البنية المفصلة لنموذج HMM (مما يقدم إطارًا احتماليًا صارمًا لصياغة مُقدِّرات ناجعة للبيانات المفقودة. إن نهج HMM شبه الخاضع للإشراف، الذي أشار إليه SSHMM، يجعل من الممكن معالجة مشكلة التحيز في الاختيار ومراعاة أكبر قدر من المعلومات في عملية التعلم. تستخدم معظم الأعمال المقترحة مؤخرًا في الأدبيات بيانات من منصة Lending Club للتحقق من صحة واختبار مناهجها شبه الخاضعة للإشراف التي تم تطويرها، بما في ذلك نموذج SSHMM الخاص بنا. إن الأمر يتعلق بمنصة الإقراض من نظير إلى نظير عبر الإنترنت والتي تعتبر الأكثر استخدامًا في الولايات المتحدة، مما يساعد على خلق قيمة من خلال ربط المقترضين والمقرضين ويقلل من تكاليف المعاملات والبحث عن مقابل.