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 Mohamed BARHDADI

Doctorant au laboratoire de recherche

« Méthodes Appliquées en Statistique, Actuariat, Finance et Economie Qantitative »

(MASAFEQ)

Spécialité : ***Mathématiques Appliquées***

Soutiendra publiquement sa thèse de Doctorat

Le lundi 1er juillet 2024 à 10h

à la salle de conférence de l'INSEA

Intitulé de la thèse

« Développement et amélioration des modèles de classification de données complexes et de grande taille »

Devant le jury composé de :

Présidente :
Pr. Fatima Zahra MHADA, PES, ENSIAS-Rabat

Directeur de thèse :
Pr. Mohamed OUZINEB, PES, INSEA-Rabat

Co-Directeur de thèse :
Pr. Badreddine BENYACOUB, PH, INSEA-Rabat

Membres du jury :
Pr. Ali YAHYAOUI, PES, Faculté des sciences Dhar El Mahraz, Fès
Pr. Karim EL MOUAOUAKIL, PES, Université Sidi Mohamed Ben Abdellah, Fès
Pr. Monsieur Khalid EL YASSINI, PES, Université My Ismail, Meknès
Pr. Ahmed EL GHINI, PES, Université Mohamed V, Rabat
Pr. Abdelhadi SABRY, PH, Académie internationale Mohammed VI de l'aviation civile (AIAC)

INSEA
Institut National de
Statistique et d'Économie
Appliquée
CEDOC-SIDD

Les Résumés de la thèse (F1)

Date : le 07/05/2024

Réservé à l'administration

N°de thèse : .....

Nom : BARHDADI

Prénom : Mohamed

Résumé

Pendant de nombreuses années, pour résoudre un problème de classification, la pensée dominante était qu'il fallait recourir à des méthodes d'apprentissage statistiques. Toutefois, ces approches présentent des limites, notamment en ce qui concerne l'interprétation des résultats et la compréhension des décisions prises par les algorithmes. Ainsi, l'objectif principal de cette thèse est de proposer une alternative visant à pallier ces lacunes, en fournissant une méthode à la fois plus compréhensible et plus fiable. Cette alternative vise à offrir une interprétation claire des résultats, permettant aux utilisateurs de suivre le processus décisionnel et de comprendre les raisons sous-jacentes aux conclusions des modèles. Cette approche revêt une importance particulière dans des domaines sensibles tels que la finance, la santé ou le droit, où les décisions algorithmiques doivent être justifiables et explicables.

Dans ce travail, nous utilisons la programmation mathématique pour élaborer des modèles de classification binaire qui répondent à plusieurs critères clés. Ils doivent être interprétables par les humains, c'est-à-dire que les résultats et les processus de décision peuvent être compris sans difficulté. De plus, ces modèles doivent être suffisamment flexibles pour s'adapter à des ensembles de données massifs et complexes. Enfin, ils doivent être économes en ressources informatiques et rapides en termes de temps d'exécution, assurant ainsi une grande efficacité même dans des environnements à capacité limitée. Le développement de ce classificateur a suivi trois étapes principales. Premièrement, la stabilisation de la solution, qui consiste à choisir une fonction objectif appropriée. Deuxièmement, nous avons vérifié la signification statistique des résultats en utilisant des méthodes de ré-échantillonnage. Enfin, nous avons appliqué ce modèle de classification dans des contextes de données volumineux et fortement déséquilibrés. Pour mettre en œuvre ces étapes, nous avons appliqué notre méthode à deux domaines de grande importance : l'évaluation de la solvabilité pour les demandes de crédit et la prédiction de désabonnement.

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

INSEA
Institut National de
Statistique et d'Économie
Appliquée
CEDOC-SIDD

Abstract

For many years, the common approach was simple: "To solve a classification problem, use machine learning techniques." Yet, these methods have notable limitations, particularly regarding the interpretability of results and the ability to understand the logic behind the decisions that algorithms make. This thesis aims to address these challenges by offering an alternative solution that is not only easier to understand but also more reliable. This approach seeks to provide a clear explanation of the results, allowing users to track the decision-making process and understand the rationale behind the models' outputs. This clarity is especially crucial in sensitive sectors like finance, healthcare, and law, where algorithm-driven decisions must be transparent and defensible.

In this study, we apply mathematical programming to create binary classification models that satisfy several essential criteria. Firstly, these models must be easily interpretable, in other words, the results and the underlying decision-making processes are clear and understandable to humans. Secondly, the models need to be versatile enough to handle large, complex data sets without compromising performance. Finally, they should be designed to use resources efficiently and produce rapid results, even in settings with limited computational capacity. The creation of our classifier progressed through three crucial stages. Initially, we stabilized the solution by selecting an appropriate objective function. In the second stage, we validated the statistical significance of the results using resampling techniques. Finally, we applied our classification model to scenarios with extensive and highly imbalanced datasets. We assessed this approach in two primary applications: credit scoring for credit applications and churn prediction for subscription-based services.

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

INSEA
Institut National de
Statistique et d'Économie
Appliquée
CEDOC-SIDD

ملخص

لسنوات طويلة، ظل النهج الرائج لحل مشكلات التصنيف يعتمد ببساطة على استخدام تقنيات التعلّم الآلي، والتي على الرغم من فعاليتها، فهي تشكو من عيوب جوهرية، لاسيما فيما يتعلق بصعوبة تفسير النتائج والتحديات المرتبطة بفهم القرارات التي تنتج عن الخوارزميات. تهدف هذه الأطروحة إلى معالجة هذه التحديات بتقديم حل بديل يتسم بسهولة الفهم و ذي مستوى عالي من الموثقية و الإعتمادية. يسعى النهج الذي نقترحه إلى إلقاء الضوء على آليات عمل هذه النماذج و توضيح العمليات التي تقود إلى اتخاذ القرارات، مما يمَكِّن المستخدمين من متابعة هذه العمليات خطوة بخطوة لفهم كيفية الحصول على النتائج النهائية. هذه الشفافية في العمليات والنتائج تكتسي أهمية بالغة في مجالات عدة، خاصة الحساسة منها كالقطاع المالي والرعاية الصحية والقانون، حيث تتطلب القرارات التي تُتّخذ بواسطة هذه الخوارزميات درجة عالية من الوضوح وقابلية للتفسير و التبرير.

في هذا البحث، نستخدم البرمجة الرياضية لتطوير نماذج التصنيف الثنائي التي تستوفي معايير أساسية عدة. أولاً، يُشترط في هذه النماذج أن تكون ذات قابلية عالية للتفسير، بمعنى أن النتائج والآليات الأساسية لعملية اتخاذ القرار يجب أن تكون واضحة ومُستساغة للعقل البشري. ثانيًا، يتعيّن على النماذج أن تتّسم بالمرونة الكافية لمعالجة مجموعات البيانات الضخمة والمعقدة دون المساس بكفاءة الأداء. ثالثًا، ينبغي أن تُصمَّم هذه النماذج لاستخدام الموارد بفعالية وإنتاج نتائج سريعة، حتى في الحواسيب ذات القدرات الحسابيّة المحدودة. تطلّبت عملية بناء نموذج التصنيف ثلاث مراحل أساسية. في البداية، عزَّزنا استقرار الحل باختيار دالة الهدف الأنسب. في المرحلة الثانية، قمنا بالتحقق من كون النتائج المحصَّلة ذات دلالة من الناحية الإحصائية باستخدام تقنيات إعادة أخذ العينات. وأخيرا، طبقنا نموذج التصنيف في سياقات تشمل مجموعات شاسعة من البيانات غير المتّزنة. قمنا بتقييم هذه الطريقة عبر تطبيقها في مجالين أساسيين: تقييم الائتمان لطلبات القروض وتوقع انسحاب العملاء من الخدمات المعتمِدة على الاشتراكات.

***INSEA** B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
***Tél.** : +212 (0)5 37 77 48 59/60 **Fax** : + 212 (0) 537 77 94 57*