



Journal de l'enseignement des statistiques

ISSN : (imprimé) 1069-1898 (en ligne) Page d'accueil de la revue : <https://amstat.tandfonline.com/loi/ujse20>

« Ce prêt doit-il être approuvé ou refusé ? » : un grand ensemble de données avec des directives pour l'attribution des cours

Min Li, Amy Mickel et Stanley Taylor


Pour citer cet article : Min Li, Amy Mickel et Stanley Taylor (2018) « Ce prêt doit-il être approuvé ou refusé ? » : Un grand ensemble de données avec des directives pour les devoirs en classe, Journal of Statistics Education, 26 : 1, 55-66, DOI : [10.1080/10691898.2018.1434342](https://doi.org/10.1080/10691898.2018.1434342)

Pour accéder à cet article : <https://doi.org/10.1080/10691898.2018.1434342>



© Min Li, Amy Mickel et Stanley Taylor©
Min Li, Amy Mickel et Stanley Taylor



Voir le matériel supplémentaire 



Publié en ligne : 05 avril 2018.



Soumettez votre article à cette revue



Nombre de vues de l'article : 5351



Afficher les données Crossmark 

« Ce prêt doit-il être approuvé ou refusé ? » : un grand ensemble de données avec des directives pour l'attribution des cours

Min Li, Amy Mickel et Stanley Taylor

Faculté d'administration des affaires, Université d'État de Californie, Sacramento, Californie

ABSTRAIT

Dans cet article, nous présentons un vaste et riche ensemble de données de la Small Business Administration (SBA) des États-Unis et un exercice d'accompagnement conçu pour enseigner les statistiques en tant que processus d'investigation de la prise de décision. Des directives pour l'exercice intitulé « Ce prêt doit-il être approuvé ou refusé ? », ainsi qu'un sous-ensemble de l'ensemble de données plus vaste, sont fournies. Pour cet exercice d'étude de cas, les étudiants assument le rôle d'agent de crédit dans une banque et sont invités à approuver ou à refuser un prêt en évaluant son risque de défaut à l'aide d'une régression logistique. Étant donné que cet exercice est conçu pour des cours d'introduction aux statistiques commerciales, des méthodes supplémentaires pour des cours d'analyse de données plus avancés sont également suggérées.

MOTS CLÉS

Étude de cas; Classification;
Règle de décision; Régression
logistique; Données réelles;
Indicateur de risque

1. Introduction

Dans les Lignes directrices pour l'évaluation et l'enseignement de l'enseignement des statistiques (GAISE) de l'American Statistical Association (ASA)

Selon le rapport du Collège (GAISE College Report ASA Revision Committee 2016), les recommandations suivantes ont été formulées pour enseigner les statistiques d'introduction :

- (a) Enseigner la pensée statistique. Enseigner les statistiques comme un processus d'investigation de résolution de problèmes et de prise de décision. Donnez aux élèves l'expérience de la pensée multivariable. (b) Concentrez-vous sur la compréhension conceptuelle. (c) Intégrez des données réelles avec un contexte et un objectif. (d) Favorisez l'apprentissage actif. (e) Utilisez la technologie pour explorer les concepts et analyser les données. (f) Utilisez des évaluations pour améliorer l'apprentissage.

Dans cet article, nous prenons en compte ces recommandations en fournissant un ensemble de données riche et volumineux qui constitue en soi une contribution significative, car il peut être utilisé par les éducateurs pour créer des opportunités d'apprentissage conformes aux recommandations GAISE 2016. En conjonction avec l'ensemble de données, un ensemble de lignes directrices pour une étude de cas conçue en tenant compte des recommandations susmentionnées est également décrit.

L'ensemble de données accompagnant cet article est un ensemble de données réel de la Small Business Administration (SBA) des États-Unis. L'étude de cas intitulée « Ce prêt doit-il être approuvé ou refusé ? » est conçue pour enseigner la pensée statistique en se concentrant sur la façon d'utiliser des données réelles pour prendre des décisions éclairées dans un but particulier. Pour cette tâche, les étudiants assument le rôle d'un agent de crédit qui décide d'approuver ou non un prêt à une petite entreprise.

En analysant des données réelles, les étudiants expérimentent les statistiques comme un processus d'investigation de prise de décision, car l'étudiant est

L'étudiant doit répondre à la question suivante : En tant que représentant de la banque, dois-je accorder un prêt à une petite entreprise en particulier (société X) ? Pourquoi ou pourquoi pas ? L'étudiant prend cette décision en évaluant le risque d'un prêt.

L'évaluation est réalisée en estimant la probabilité de défaut du prêt en analysant cet ensemble de données historiques, puis en classant le prêt dans l'une des deux catégories suivantes : (a) risque plus élevé, susceptible de faire défaut sur le prêt (c'est-à-dire, être radié/ne pas payer en totalité) ou (b) risque plus faible, susceptible de rembourser le prêt en totalité.

Le processus de prise de cette décision exige que les élèves comprennent conceptuellement les concepts statistiques et comment les appliquer.

Nous avons utilisé une version adaptée de cette étude de cas dans les cours d'analyse des données pour les étudiants en commerce de premier et de deuxième cycle. Ces cours couvrent des sujets allant de la régression et de l'analyse de la variance dans le cours de premier cycle à l'exploration de données dans le cours de deuxième cycle. Pour tous les cours, la régression logistique est incluse dans le devoir tandis que les réseaux neuronaux et les machines à vecteurs de support (SVM) ne sont introduits que dans le cours de deuxième cycle.

Pour les deux cours, nous présentons d'abord ce travail sous forme de devoir interactif en classe. Nous passons deux ou trois périodes de cours de 75 minutes dans des laboratoires informatiques pour guider les étudiants à travers des étapes spécifiques sur la façon d'analyser ce grand ensemble de données afin de les aider à éclairer leurs processus de prise de décision. Pour favoriser un environnement d'apprentissage actif, nous encourageons la discussion et les questions pendant ces périodes de cours et divisons généralement les étudiants en groupes pour discuter de certaines étapes, puis leur demandons de présenter leurs idées et leur raisonnement. Pour évaluer la pensée statistique des étudiants, on leur présente un cas similaire et on leur demande de rédiger un rapport décrivant leurs décisions de prêt et la justification de ces décisions.

Cette tâche est idéale pour les cours d'analyse de données pour plusieurs raisons.

(a) L'étude de cas intègre l'ensemble du GAISE 2016 recommandations.

(b) Le sujet lui-même capte l'intérêt des étudiants, car il s'agit d'une application de données réelles liées à la situation financière réelle décisions.

(c) Les étudiants sont exposés à la gestion d'un grand ensemble de données et comprendre comment les données historiques peuvent être utilisées pour faire des décisions éclairées.

(d) La pensée critique est encouragée ; l'analyse, la synthèse et des compétences de prise de décision sont utilisées.

(e) Les étudiants sont initiés à la régression logistique et à d'autres méthodes de classification plus avancées.

(f) L'importance d'identifier des explications raisonnables

Les variables (par exemple, les indicateurs de risque de défaut de prêt) à incorporer dans les modèles statistiques donnent lieu à des discussions animées et engageantes.

De plus, les instructeurs en statistiques commerciales ont signalé que l'utilisation d'études de cas a entraîné une motivation et une participation accrues des étudiants, une sensibilisation accrue des étudiants de la pertinence des statistiques pour la prise de décision commerciale, et des expériences de classe plus positives pour l'instructeur (par exemple, Bryant 1999; Nolan et Speed 1999; Parr et Smith 1998; Smith et Bryant 2009). Nous avons connu des avantages similaires avec cette étude de cas.

2. Contexte et description des ensembles de données

La SBA américaine a été fondée en 1953 sur le principe de promotion et d'assistance aux petites entreprises sur le marché du crédit américain. (Présentation et historique de la SBA, Small Business Administration des États-Unis) (2015)). Les petites entreprises ont été une source principale d'emplois création aux États-Unis ; favorisant ainsi les petites entreprises

La formation et la croissance ont des avantages sociaux en créant des opportunités d'emploi et en réduisant le chômage. L'une des façons dont la SBA aide ces petites entreprises se font par le biais d'une garantie de prêt programme qui vise à encourager les banques à accorder des prêts aux petites entreprises. La SBA agit un peu comme un fournisseur d'assurance réduire le risque pour une banque en assumant une partie du risque en garantissant une partie du prêt. Dans le cas où un le prêt est en défaut, la SBA couvre alors le montant qu'elle garanti.

Il existe de nombreuses histoires de réussite de start-ups recevoir des garanties de prêt SBA telles que FedEx et Apple Ordinateur. Cependant, il y a aussi eu des histoires de petites entreprises et/ou start-ups qui ont fait défaut sur leurs Prêts garantis par la SBA. Le taux de défaut sur ces prêts est une source de controverse depuis des décennies.

les économistes estiment que les marchés du crédit fonctionnent efficacement sans la participation du gouvernement. Les partisans des prêts garantis par la SBA soutiennent que les avantages sociaux de l'emploi création par ces petites entreprises bénéficiant de prêts garantis par le gouvernement dépassent de loin les coûts encourus prêts en défaut.

Étant donné que les prêts SBA ne garantissent qu'une partie du prêt total En revanche, les banques subiront des pertes si une petite entreprise fait défaut sur son prêt garanti par la SBA. Par conséquent, les banques sont toujours confrontées

Tableau 1(a). Description de 27 variables dans les deux ensembles de données.

Nom de la variable	Type de données	Description de la variable
LoanNr_ChkDgt	Texte	Identifiant – Clé primaire
Nom	Texte	Nom de l'emprunteur
Ville	Texte	Ville emprunteuse
État	Texte	État emprunteur
.....	Texte	Code postal de l'emprunteur
Banque	Texte	Nom de la banque
Relevé bancaire	Texte	État bancaire
SCIAN	Texte	Classification des industries nord-américaines code système
Date d'approbation	Date/Heure	Date d'émission de l'engagement de la SBA
ApprobationFY	Texte	Exercice financier d'engagement
Terme	Nombre	Durée du prêt en mois
Pas d'Emp	Nombre	Nombre d'employés de l'entreprise
NouveauExiste	Texte	1 D Entreprise existante, 2 D Nouvelle entreprise
Créer un emploi	Nombre	Nombre d'emplois créés
Emploi conservé	Nombre	Nombre d'emplois conservés
Code de franchise	Texte	Code de franchise, (00000 ou 00001) D Non franchise
UrbainRural	Texte	1 D Urbain, 2 D rural, 0 D indéfini
Ligne de révisionCr	Texte	Ligne de crédit renouvelable : YD Oui, ND Non
Faible Doc	Texte	Programme de prêt LowDoc : YD Oui, ND Non
Date de dissolution du chantage		Date/Heure La date à laquelle un prêt est déclaré en défaut
Date de déboursement	Date/Heure	Date de déboursement
Déboursement Brut	Devise	Montant déboursé
Solde brut	Devise	Montant brut en souffrance
Statut MIS	Texte	Statut du prêt radié D CHGOFF, payé en entier D PIF
ChgOffPrinGr	Devise	Montant radié
GrAppv	Devise	Montant brut du prêt approuvé par la banque
SBA_Appv	Devise	Montant garanti de la SBA approuvé prêt

avec un choix difficile quant à savoir s'ils doivent accorder une telle prêt en raison du risque élevé de défaut. Une façon d'informer leur prise de décision se fait par l'analyse des données historiques pertinentes des données telles que les ensembles de données fournis ici.

Deux ensembles de données sont fournis : (a) Ensemble de données « National SBA » (nommé SBAnational.csv) de la SBA américaine qui comprend données historiques de 1987 à 2014 (899 164 observations)¹ et (b) ensemble de données « SBA Case » (nommé SBACase.csv) qui est utilisé dans la tâche décrite dans cet article (2102 observations). L'ensemble de données « SBA Case » est un sous-ensemble de la « SBA nationale ».

Le nom de la variable, le type de données et une brève description de chaque variable est fournie pour les 27 variables dans les deux ensembles de données (voir le tableau 1(a)). Pour l'ensemble de données « SBA Case », une variable supplémentaire huit variables ont été générées par les auteurs dans le cadre de la affectation (voir tableau 1(b)) et décrite dans les sections 4.1.4, 4.1.5, 4.1.6, 4.1.7 et 4.3.1. Pour la plupart des variables, la description est évidente. Les variables nécessitant une explication plus approfondie incluent : NAICS, NewExist, LowDoc et MIS_Status et sont décrits ci-dessous.

SCIAN (Système de classification des industries de l'Amérique du Nord) : Il s'agit d'un système de classification hiérarchique de 2 à 6 chiffres utilisé par les agences statistiques fédérales dans la classification des établissements commerciaux pour la collecte, l'analyse et la présentation des

¹ Veuillez noter que l'ensemble de données que nous fournissons ici est limité aux prêts provenant dans les 50 États-Unis et à Washington DC (les territoires américains ont été exclus) et pour lesquels le résultat (payé en totalité ou radié/défaut) est connu ; afin pour enseigner la régression logistique, une variable dépendante binaire est nécessaire.

² Le code SAS utilisé pour créer le sous-ensemble de données se trouve dans le « SBA » qui l'accompagne. Fichier de documentation des données « Cas ».

Tableau 1(b). Description de 8 variables supplémentaires dans l'ensemble de données de cas SBA.

Nom de la variable	Type de données	Description de la variable
Nouveau	Noméro D1 si NewExistD2 (Nouvelle Entreprise), D0 si NewExistD1 (Entreprise existante)	
Partie	Nombre	Proportion du montant brut garanti par la SBA
Immobilier	Noméro D1 si le prêt est adossé à un bien immobilier, D0 sinon	
Récession	Noméro D1 si le prêt est actif pendant la Grande Récession, D0 sinon	
Choisi	Noméro D1 si les données sont sélectionnées comme données de formation pour construire un modèle pour l'affectation, D0 si les données sont sélectionnées comme données de test pour valider le modèle	
Délai par défaut	Noméro D1 si MIS_StatusDCHGOFF, D0 si MIS_StatusDPIF	
	Nombre	Variable supplémentaire générée lors de la création « Récession » dans la section 4.1.6
xx	Nombre	Variable supplémentaire générée lors de la création « Récession » dans la section 4.1.6

données statistiques décrivant l'économie américaine. Les deux premiers chiffres de la classification SCIAN représentent le secteur économique. Le tableau 2 montre les secteurs à 2 chiffres et une description correspondante pour chaque secteur.

Note pédagogique : Le tableau des codes SCIAN à deux chiffres publié par le Bureau de recensement des États-Unis (<http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chartD2012>) fusionne quelques secteurs (voir Fabrication, Commerce de détail, transport et entreposage). Pour être cohérent avec la publication du US Census Bureau, nous faisons également la même chose fusions. Toutefois, les instructeurs peuvent souhaiter examiner les cas individuels secteurs de la fabrication, du commerce de détail, des transports et Entreposage.

NewExist (1 D Entreprise existante, 2 D Nouvelle entreprise) : Ceci indique si l'entreprise est une entreprise existante (existante depuis plus de 2 ans) ou une nouvelle entreprise (existante depuis inférieur ou égal à 2 ans).

LowDoc (YD Oui, ND Non) : Afin de traiter davantage de prêts efficacement, un programme de « prêt LowDoc » a été mis en œuvre où Les prêts de moins de 150 000 \$ peuvent être traités à l'aide d'une demande d'une page. « Oui » indique les prêts avec une demande d'une page, et « Non » indique les prêts avec plus d'informations jointes au application. Dans cet ensemble de données, 87,31 % sont codés comme N (Non) et 12,31 % pour Y (Oui) pour un total de 99,62 %. Il convient de noter que

Tableau 2. Description des deux premiers chiffres du SCIAN.

Secteur	Description
11	Agriculture, sylviculture, pêche et chasse
21	Exploitation minière, exploitation en carrière et extraction de pétrole et de gaz
22	Utilitaires
23	Construction
31-33	Fabrication
42	Commerce de gros
44-45	Commerce de détail
48-49	Transport et entreposage
51	Information
52	Finances et assurances
53	Immobilier et location et leasing
54	Services professionnels, scientifiques et techniques
55	Gestion des sociétés et des entreprises
56	Gestion administrative et de soutien et des déchets et services d'assainissement
61	Services éducatifs
62	Soins de santé et assistance sociale
71	Arts, divertissement et loisirs
72	Hébergement et restauration
81	Autres services (à l'exception de l'administration publique)
92	Administration publique

0,38 % ont d'autres valeurs (0, 1, A, C, R, S) ; il s'agit de saisie de données erreurs. Il y a également 2582 valeurs manquantes pour cette variable, exclus lors du calcul de ces proportions. Nous avons choisi de laisser ces entrées « telles quelles » pour offrir aux étudiants la possibilité pour apprendre à traiter des ensembles de données contenant de telles erreurs.

MIS_Status : Cette variable indique le statut du prêt : en défaut/facturé (CHGOFF) ou ont été avec succès payé en totalité (PIF).

3. Considérations préalables à la création d'une mission

Avant l'attribution de l'étude de cas, il est suggéré que les éducateurs envisagent : (a) d'élaborer des objectifs d'apprentissage pour les mission; (b) en utilisant des progiciels d'analyse statistique qui sont facilement accessibles aux étudiants pour analyse ; (c) déterminer une période de temps à inclure dans les analyses ; et (d) décider comment intégrer l'étude de cas dans une classe et façons d'évaluer l'apprentissage.

3.1. Objectifs d'apprentissage

- Il s'agit sans doute de l'étape la plus importante avant l'affectation création. Une compréhension et une explication claires de ce que mission est conçue pour enseigner est nécessaire. Pour le « Devrait Ce prêt doit-il être approuvé ou refusé ? » mission, nous voulons que notre les étudiants à :
- 1. Analyser un grand ensemble de données pour promouvoir la pensée statistique ;
 - 2. Identifier les variables explicatives qui peuvent être de bons « prédicteurs » ou indicateurs de risque du niveau de risque associé avec un prêt;
 - 3. Parcourez les étapes de construction du modèle et validation;
 - 4. Appliquer la régression logistique (et d'autres méthodes plus avancées) méthodes pour les étudiants diplômés) pour classer un prêt en fonction sur le risque de défaut prévu ; et
 - 5. Prendre une décision fondée sur un scénario, éclairée par des analyses de données (par exemple, financer ou non le prêt).

3.2. Logiciels d'analyse statistique

Les ensembles de données sont préparés pour analyse dans la plupart des logiciels d'analyse statistique disponibles. Il est suggéré aux enseignants choisir un progiciel auquel les étudiants peuvent facilement accéder et Nous utilisons les produits Microsoft Excel, R et SAS (JMP, University Edition) car ils sont facilement accessibles à nos étudiants gratuitement.

Pour nos étudiants, nous exportons les données dans les formats suivants : données permanentes SAS (.sas7bdat) et valeurs séparées par des virgules (.csv). Nous avons nos étudiants de premier cycle utilisent JMP pour ouvrir le fichier de données SAS afin d'effectuer des opérations logistiques régression et autres analyses. L'interface conviviale de JMP, par pointer-cliquer, est parfaite pour nos données de premier cycle cours d'analyse. Nos étudiants en MBA utilisent R pour ouvrez le fichier de données à valeurs séparées par des virgules et exécutez des analyses qui incluent la régression logistique, les réseaux neuronaux, et les SVM.

3.3. Période de temps

Les enseignants peuvent également vouloir réfléchir à la période à inclure dans les analyses. Par exemple, dans notre travail, l'accent est mis sur les taux de défaut des prêts dont la date de décaissement est antérieure à 2010.³ Nous avons choisi cette période pour deux raisons.

Nous souhaitons tenir compte des variations dues à la Grande Récession (décembre 2007 à juin 2009)⁴ ; nous avons donc besoin des prêts décaissés avant, pendant et après cette période. Deuxièmement, nous limitons la période aux prêts en excluant ceux décaissés après 2010, car la durée d'un prêt est souvent de 5 ans ou plus.⁵

Nous pensons que l'inclusion des prêts dont les dates de décaissement sont postérieures à 2010 donnerait plus de poids aux prêts radiés qu'aux prêts remboursés en totalité. Plus précisément, les prêts radiés le seront avant la date d'échéance du prêt, tandis que les prêts qui seront probablement remboursés en totalité le seront à la date d'échéance du prêt (qui s'étendrait au-delà de l'ensemble de données se terminant en 2014). Étant donné que cet ensemble de données a été limité aux prêts dont le résultat est connu, il y a plus de chances que les prêts radiés avant la date d'échéance soient inclus dans l'ensemble de données, tandis que ceux qui pourraient être remboursés en totalité ont été exclus. Il est important de garder à l'esprit que toute restriction temporelle sur les prêts inclus dans les analyses de données pourrait introduire un biais de sélection, en particulier vers la fin de la période. Cela peut avoir un impact sur les performances de tout modèle prédictif basé sur ces données.

3.4. Format de l'étude de cas

Ce devoir peut être adapté aux cours en classe, hybrides et en ligne. Bien que nous décrivions comment ce devoir a été appliqué dans nos cours en classe, nous encourageons les instructeurs à adapter les devoirs pour répondre aux besoins des étudiants et aux différents modes de prestation.

Pour les cours de premier et de deuxième cycle, nous présentons d'abord ce cours sous forme de devoir interactif en classe. Nous consacrons deux ou trois périodes de cours de 75 minutes à guider les étudiants à travers les différentes étapes décrites ci-dessous. Nous encourageons la discussion et les questions pendant ces périodes de cours. Pour promouvoir l'apprentissage actif, nous divisons les étudiants en groupes pour discuter de certaines étapes, puis nous leur demandons de présenter leurs idées et leur raisonnement. En tant qu'instructeurs, nous facilitons une discussion en classe plus large après ces présentations pour nous assurer que les étudiants comprennent les différentes étapes.

Pour évaluer l'apprentissage des étudiants, nous élaborons un travail d'étude de cas noté similaire à celui présenté en classe. Pour les étudiants de premier cycle, nous les laissons réaliser le travail en groupes de trois personnes. Pour les cours de deuxième cycle, les étudiants doivent réaliser le travail individuellement.

4. Lignes directrices pour « Ce prêt doit-il être approuvé ou « Refusé ? » Étude de cas

Cette section est organisée autour des étapes impliquées dans le processus d'investigation consistant à analyser ces données pour prendre une décision.

prendre une décision éclairée quant à l'approbation ou au refus d'un prêt, l'un des principaux objectifs d'apprentissage de cette mission.

Les étudiants sont guidés à travers :

Étape 1 : Identifier les indicateurs de risque potentiel ;

Étape 2 : Comprendre l'étude de cas ; Étape 3 :

Construire le modèle, créer des règles de décision et valider le modèle de régression logistique ; et

Étape 4 : Utiliser le modèle pour prendre des décisions.

4.1. Étape 1 : Identification des variables explicatives (indicateurs ou prédicteurs) du risque potentiel

Au cours de la première période de cours, nous fournissons aux étudiants l'ensemble de données « National SBA », un aperçu du contexte de la SBA et le devoir avec ses objectifs d'apprentissage. Étant donné que les modèles économiques doivent être basés sur une théorie économique solide, nous engageons les étudiants dans une discussion qui les oblige à identifier les variables explicatives qui, selon eux, seraient de bons indicateurs ou prédicteurs du risque potentiel d'un prêt : probabilité de défaut (risque plus élevé) par rapport au remboursement intégral (risque plus faible).

Pour atteindre l'objectif d'apprentissage suivant, à savoir identifier les variables explicatives qui peuvent être de bons prédicteurs ou indicateurs de risque du niveau de risque associé à un prêt, nous encourageons les étudiants à considérer les taux de défaut pour un groupe représenté par le pourcentage de prêts classés comme étant en défaut. Pour un groupe particulier de prêts, le taux de défaut est déterminé en utilisant la variable « MIS_Status » et en calculant le pourcentage du nombre total de prêts (CHGOFF C PIF) classés comme étant en défaut (CHGOFF).

Note pédagogique : Nous divisons les élèves en groupes pour une discussion et leur demandons de fournir une justification écrite pour chaque variable afin de savoir si elle serait un bon indicateur de risque et de les présenter brièvement à la classe. Cette activité renforce l'importance d'avoir une théorie solide lors de la construction de modèles et favorise l'apprentissage actif.

Il existe un certain nombre de variables qui apparaissent systématiquement comme des indicateurs de risque pouvant expliquer la variation des taux de défaut de prêt. Sept variables, ainsi que certaines analyses exploratoires, sont abordées ci-dessous, notamment la localisation (État), le secteur d'activité, le décaissement brut, les entreprises nouvelles ou établies, les prêts garantis par l'immobilier, la récession économique et la part garantie par la SBA des prêts approuvés. Pour un certain nombre de ces indicateurs, des variables fictives sont créées à des fins d'analyse et sont abordées dans l'enseignement

Remarques.

4.1.1. Localisation (État)

La localisation par État (représentée par « État » dans le [tableau 1\(a\)](#)) est un prédicteur possible que les étudiants identifient dans leurs discussions.

Ils reconnaissent que les 50 États et Washington DC ont des environnements économiques différents dans lesquels ils opèrent, ce qui entraîne des taux de défaut différents. Nous présentons cette carte thermique ([Figure 1](#)) en classe pour étayer cette discussion.

Note pédagogique : Les étudiants sont encouragés à explorer les raisons des différences dans les taux de défaut selon les États. Par exemple, pendant la Grande Récession, la Floride a connu une baisse importante des prix de l'immobilier, ce qui pourrait contribuer à des taux de défaut élevés ; des États comme le Wyoming et le Dakota du Nord avaient des économies plus fortes (en raison de leur dépendance aux minéraux et au pétrole), ce qui peut expliquer leurs taux de défaut plus faibles. Étant donné que nous opérons en Californie, la Californie

³ « DisbursementDate » est la variable utilisée pour déterminer cette classification.

⁴ Les dates telles que déclarées par le Bureau national de recherche économique (voir http://money.cnn.com/2010/09/20/news/economy/recession_over/)

⁵ La répartition des durées des prêts est telle que le mode est de 7 ans (27% des prêts ont une durée de 7 ans) et 73% ont une durée supérieure à 5 ans. Pour les prêts distribués à partir de 2010, 66% ont une durée supérieure à 5 ans.

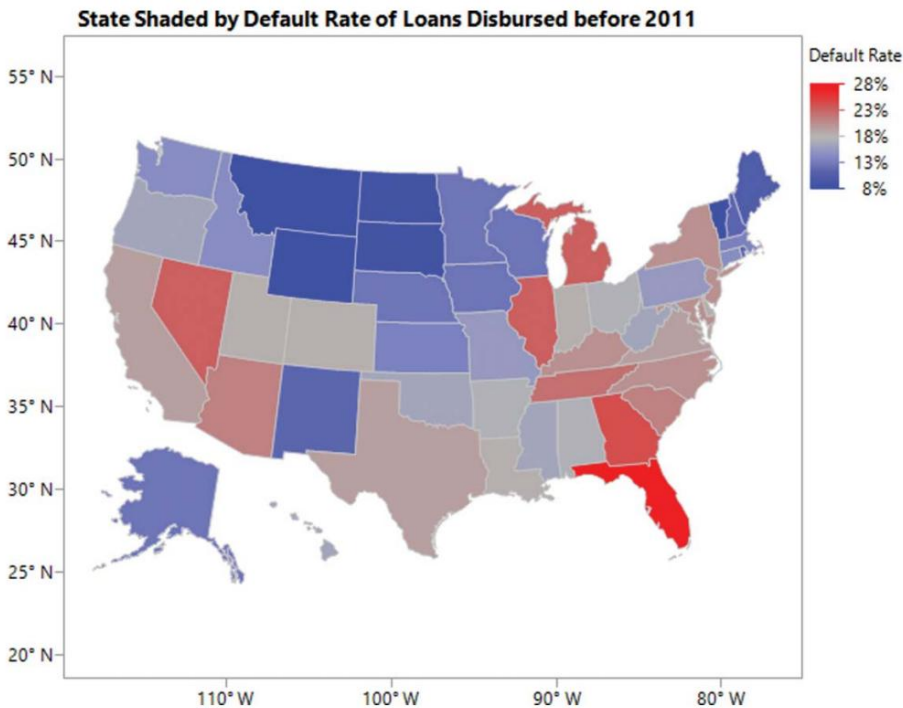


Figure 1. Carte thermique, taux de défaut par état (la figure 1 a été créée à l'aide de JMP).

par rapport aux autres États est mis en évidence, car il « ramène à la maison » discussion. Les instructeurs peuvent choisir de se concentrer sur les états d'intérêt à leurs étudiants.

4.1.2. Industrie

Comme indiqué dans le tableau 3, l'industrie (les deux premiers chiffres du SCIAN codes) est un autre indicateur de risque que les étudiants prennent en compte en raison de la variation importante des taux de défaut. À un moment donné à l'extrémité du spectre se trouvent les industries ayant de faibles taux de défaut (8%–10%), tels que : l'exploitation minière, l'exploration pétrolière et gazière (21), agriculture (11), sociétés de portefeuille de valeurs mobilières (55) et médecins et dentistes (62). À l'autre extrémité de la le spectre sont des industries avec des taux de défaut plus élevés (28%–29 %), comme les institutions financières comme les coopératives de crédit (52) et agences immobilières (53).

La variation des taux de défaut de l'industrie est souvent due à la nature cyclique de la demande de produits ou de services. par exemple, le secteur de la construction (23) connaît une expansion et une contraction spectaculaires au cours d'un cycle économique, tandis que le secteur des services médicaux (62) a tendance à être beaucoup plus stable ; par conséquent, les revenus et le résultat net sont beaucoup moins volatils pour les services médicaux que pour la construction. De plus, contrairement la construction et les services médicaux ont des exigences de licence qui créent des obstacles que les nouvelles entreprises doivent surmonter. Comme il n'est pas facile d'accéder aux services médicaux, ceux qui y accèdent cette industrie est très sérieuse dans sa nouvelle aventure et Cela contribue encore davantage à la baisse du secteur médical risque de crédit.

Comme la construction, une autre industrie qui a une plus grande le taux par défaut est l'hébergement à l'hôtel et le service de restauration industrie (c'est-à-dire l'hôtellerie) (72). Au fil du temps, les prêts hôteliers les défauts de paiement ont tendance à être élevés car les hôtels construisent souvent trop de nouveaux logements unités lorsque les taux d'occupation sont élevés et ils peuvent alors sont confrontés à de faibles taux d'occupation pour diverses raisons inattendues.

En ce qui concerne la restauration, le succès de tout nouveau restaurant est hautement imprévisible, et le succès continu de Les restaurants existants sont souvent menacés par de nouvelles entreprises.

Note pédagogique : Dans nos cours, nous avons tendance à utiliser les codes à deux chiffres montré dans le tableau 3. Cependant, on peut demander à leurs étudiants d'utiliser chiffres supplémentaires dans l'analyse. Par exemple, les médecins sont codés comme 6211 et les dentistes sont 6212. Le lien suivant fournit les schéma de codage plus détaillé que ceux fournis dans le tableau 3 : <http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chartD2012>. Ces définitions ainsi que les codes détaillés fournis dans le La variable NAICS permettra aux étudiants d'analyser des données plus spécifiques industries.

Tableau 3. Taux de défaut de l'industrie (deux premiers chiffres des codes SCIAN).

Code à 2 chiffres	Description	Taux de défaut (%)
21	Exploitation minière, carrières, pétrole et gaz extraction	8
11	Agriculture, sylviculture, pêche et chasse	9
55	Gestion des sociétés et des entreprises	10
62	Soins de santé et assistance sociale	10
22	Utilitaires	14
92	Administration publique	15
54	Professionnel, scientifique et technique services	19
42	Commerce de gros	19
31–33	Fabrication	19, 16, 14
81	Autres services (sauf publics) administration)	20
71	Arts, divertissement et loisirs	21
72	Hébergement et restauration	22
44–45	Commerce de détail	22, 23
23	Construction	23
56	Administratif/support & déchets Service de gestion/remédiation	24
61	Services éducatifs	24
51	Information	25
48–49	Transport et entreposage	27, 23
52	Finances et assurances	28
53	Immobilier et location et leasing	29

4.1.3. Déboursement brut

Décaissement brut (représenté par « Décaissement Brut » dans (l'ensemble de données) est un autre indicateur de risque que de nombreux étudiants identifient comme une variable clé à prendre en compte. La raison derrière la sélection « Décaissement brut » signifie que plus le montant du prêt est important, plus il est probable que l'entreprise sous-jacente sera établie et en expansion (c'est-à-dire acheter des actifs qui ont une certaine valeur de revente), augmentant la probabilité de rembourser le prêt. Cette logique est confirmé par l'examen des quartiles présentés dans le [tableau 4](#).

4.1.4. Entreprises nouvelles et entreprises établies

Le fait qu'une entreprise soit nouvelle ou établie (représentée par « NewEx-ist » dans l'ensemble de données) est un autre indicateur de risque potentiel que les étudiants identifier. Par conséquent, une variable fictive a été créée pour la logistique régression : « Nouveau » D 1 si l'entreprise a moins ou égal à 2 ans ancien et « Nouveau » D 0 si l'entreprise a plus de 2 ans.

La plupart des étudiants affirment que les nouvelles entreprises échouent à un taux plus élevé que les entreprises établies. Les entreprises déjà établies ont fait leurs preuves et demandent un prêt pour développer ce qu'ils font déjà avec succès. Considérant que les nouveaux les entreprises n'anticipent parfois pas les obstacles qu'elles peuvent rencontrer. Les investisseurs sont confrontés à de tels défis et peuvent être incapables de les surmonter avec succès, ce qui peut les conduire à ne pas rembourser leur prêt.

Cependant, lorsque les taux de défaut des prêts aux nouvelles entreprises (moins ou égal à 2 ans) et entreprise établie (plus (de plus de 2 ans) dans cet ensemble de données sont comparés, il existe une Il n'y a pas de différence notable entre les deux. Le taux de défaut des nouvelles entreprises est de 18,98 %, et celui des entreprises établies est de 17,36 %.

4.1.5. Prêts garantis par des biens immobiliers

La question de savoir si un prêt est garanti par un bien immobilier (possession d'un terrain) est un autre indicateur de risque qui est discuté. La justification de cela L'indicateur est que la valeur du terrain est souvent suffisamment élevée pour couvrir le montant du principal impayé, réduisant ainsi la probabilité de défaut.

Étant donné que la durée du prêt est fonction de la durée de vie prévue des actifs, les prêts garantis par des biens immobiliers auront des conditions 20 ans ou plus (240 mois) et sont les seuls prêts accordés pour une durée aussi longue, alors que les prêts non garantis par des fonds réels Les biens immobiliers auront une durée inférieure à 20 ans (< 240 mois). Par conséquent, les auteurs ont créé une variable fictive, « Immobilier », où « Immobilier » D 1 si « Durée » 240 mois et « Immobilier » D 0 si « Durée » < 240 mois.

Comme le montre le [tableau 5](#), les prêts garantis par des biens immobiliers ont un taux de défaut nettement inférieur (1,64 %) à celui des prêts non garantis par des biens immobiliers. immobilier (21,16%).

4.1.6. Récession économique

Un indicateur de risque qui apparaît constamment dans les discussions est la façon dont l'économie peut avoir un impact sur les taux de défaut. Les prêts aux petites entreprises sont

Tableau 4. Quartiles de décaissement brut.

Quartiles	MÉMOIRE	PIF
100 % maximum	4 362 157 \$	11 446 325 \$
75 % quartile	140 796 \$	255 000 \$
50 % médiane	61 962,5 \$	100 000 \$
25 % quartile	27 767 \$	49 034 \$
Minimum	4 000 \$	4000 \$

Tableau 5. Prêts garantis par des biens immobiliers.

	Défaut	Payé en totalité
Prêts garantis par l'immobilier (Durée 240 mois)	2472 (1,64 %)	147 868 (98,36 %)
Prêts non garantis par des biens immobiliers (Durée < 240 mois)	153 876 (21,16 %)	573 212 (78,84 %)

affecté par l'économie en général, et davantage de petites entreprises les prêts ont tendance à faire défaut juste avant et pendant une crise économique récession. Par conséquent, les auteurs ont créé une variable fictive, « Récession », où « Récession » D 1 si les prêts étaient actifs6 pendant la Grande Récession (décembre 2007 à juin 2009) et « Récession » D 0 pour toutes les autres périodes.

Comme l'illustre un graphique à barres empilées ([figure 2](#)), les prêts actifs pendant la Grande Récession ont un taux de défaut plus élevé (31,21 %) que les prêts qui n'étaient pas actifs pendant la récession (16,63 %).

4.1.7. Part garantie par la SBA du prêt approuvé

La partie qui est le pourcentage du prêt garanti par la SBA (représentée par « Partie » dans l'ensemble de données) est une valeur finale indicateur de risque qui est abordé dans nos cours. C'est l'un des variables que les auteurs ont générées en calculant le rapport entre les montant du prêt garanti par la SBA et le montant brut approuvé par la banque (SBA_Appv/GrAppv). Le [figure 3](#) montre la répartition des parts pour les prêts entièrement remboursés et en défaut prêts décaissés de 2002 à 2010. Ces deux boxplots montrent que les prêts qui sont remboursés en totalité ont généralement un taux légèrement plus élevé Pourcentage garanti par la SBA, comme indiqué par la moyenne la plus élevée partie destinée aux prêts entièrement remboursés.

Il convient de noter que la médiane n'est pas affichée dans les boîtes à moustaches pour les prêts en défaut, car 54 % de ces prêts ont la moitié du montant du prêt garanti par la SBA (portion D 0,5). en conséquence, il n'y a aucune différence entre 1%, 5%, 10%, 25% et 50% percentiles (tous ces percentiles sont égaux à 0,5).

Note pédagogique : En plus des variables de l'ensemble de données, nous demandons à nos étudiants s'il existe d'autres variables qui peuvent être significatives et devraient être prises en considération. Les étudiants sont généralement incapables de proposer des sources spécifiques de variation. Cependant, il convient de noter que l'ensemble de données n'inclut aucun élément qui représente directement risque de crédit. Au cours des dernières années, la SBA a collecté et évalué la cote de crédit Fair Isaac (FICO) des garants et des emprunteurs. Si un emprunteur ou un garant n'est pas une personne, alors un Dun et Le score Bradstreet est obtenu. De nombreuses institutions financières s'appuient désormais sur les cotes de crédit lors de l'octroi de petits prêts. Malheureusement, cela l'ensemble de données n'inclut pas ces informations.

4.2. Étape 2 : Comprendre l'étude de cas et l'ensemble de données

Après avoir identifié les indicateurs de risque potentiel, une étude de cas, où l'étudiant assume le rôle d'un agent de prêt qui est tenu de déterminer s'il faut approuver des prêts à deux petites entreprises, est présenté. Nous soulignons le fait que les banques tentent de

⁶ Les prêts codés comme « RécessionD1 » incluent ceux qui étaient actifs pendant au moins au moins un mois pendant la période de la Grande Récession. Cela a été calculé par ajouter la durée du prêt en jours à la date de décaissement du prêt. Le codage dans SAS pour cela est : RecessionD0 ; daystermDTerm 30 ; xxDDisbursementDateCdaysterm ; si xx ge '1DEC20070 d AND xx le '30JUN20090 d then RécessionD1.

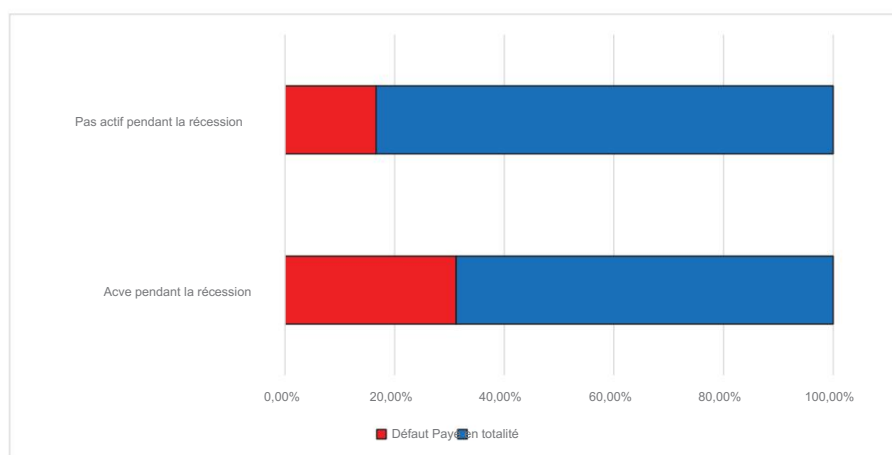


Figure 2. Statut des prêts actifs ou non actifs pendant la Grande Récession.

minimiser le risque de défaut de paiement (radiation) et approuver uniquement les prêts susceptibles d'être remboursés intégralement ultérieurement.

Note pédagogique : Pour tenir compte de deux des indicateurs de risque, l'État et le secteur, nous limitons l'étude de cas à un État et à un secteur (code secteur à deux chiffres). Nous suggérons aux enseignants d'envisager de faire de même pour trois raisons : (a) cela crée un scénario de prise de décision plus réaliste ; (b) l'inclusion de 50 États (plus Washington DC) et de 20 classifications industrielles (NAICS à 2 chiffres) entraînerait un grand nombre de variables binaires et pourrait créer des problèmes d'estimation ; et (c) l'ensemble de données extrait de l'ensemble de données plus vaste est plus facile à gérer pour les étudiants. Nous décrivons ce processus et sa justification aux étudiants en classe.

Pour nos cours, nous avons choisi de limiter l'étude de cas à l'État de Californie et au code à deux chiffres 53 : Immobilier et location et leasing. Nous extrayons les données pertinentes de l'ensemble de données plus vaste, « National SBA », qui produit un échantillon de 2102 observations et est inclus dans le document en tant que données « SBA Case ». Nous fournissons cet ensemble de données aux étudiants pour qu'ils l'analysent

dans leur rôle d'agents de crédit lorsqu'ils doivent décider d'approuver ou de refuser deux demandes de prêt.

Note pédagogique : nous limitons le scénario de la mission à la Californie, car c'est là que nous nous trouvons. Les instructeurs peuvent choisir de se concentrer sur les États qui intéressent leurs étudiants. Pour le code de l'industrie, on peut utiliser n'importe quel code à deux chiffres ou sélectionner un code utilisant plus de deux chiffres.

Étude de cas basée en Californie : Vous, un agent de crédit de Bank of America, avez reçu deux demandes de prêt de deux petites entreprises : Carmichael Realty (une agence immobilière commerciale) et SV Consulting (un cabinet de conseil en immobilier). Les informations pertinentes sur la demande sont résumées ci-dessous (voir le tableau 6). En tant qu'agent de crédit, vous devez déterminer si vous devez accorder ou refuser ces deux demandes de prêt et fournir une explication du « pourquoi ou du pourquoi pas ». Pour prendre cette décision, vous devrez évaluer le risque du prêt en calculant la probabilité estimée de défaut à l'aide d'une régression logistique. Vous souhaitez ensuite classer ce prêt

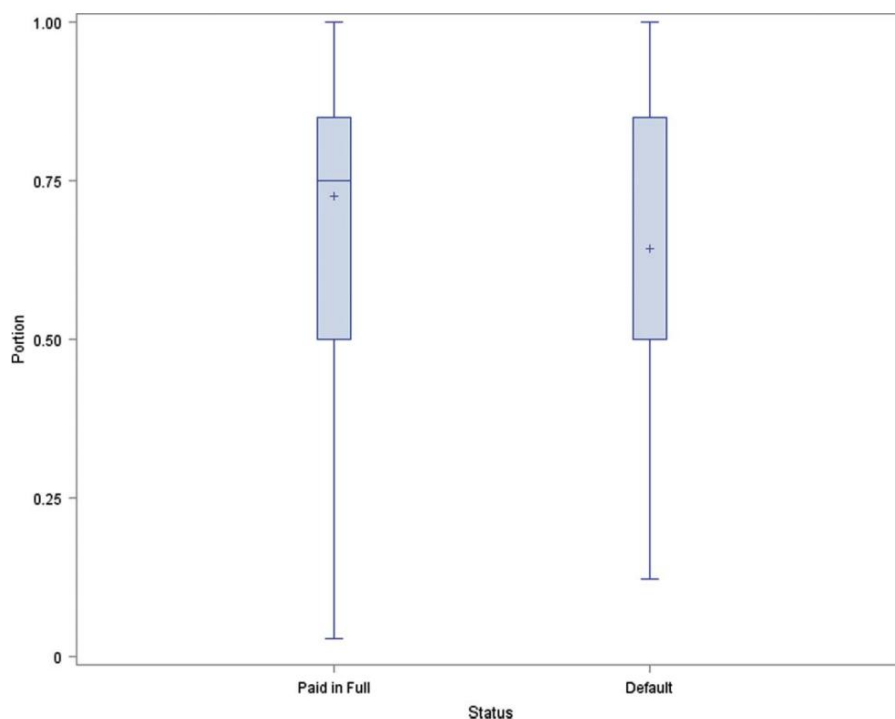


Figure 3. Parts garanties par la SBA pour les prêts entièrement remboursés et les prêts en défaut.

Tableau 6. Étude de cas basée en Californie : informations pour deux demandes de prêt.

Prêt	Nom	Ville	Date	Montant du prêt demandé	Portion SBA garantie	Sécurisé par l'immobilier ?
1 2	Immobilier Carmichael	Carmichael, Californie	Actuel (pas de récession)	1 000 000 \$	750 000 \$	Oui
	SV Conseil	San Leandro, Californie	Actuel (pas de récession)	100 000 \$	40 000 \$	Non

soit comme : « risque plus élevé – plus de probabilité de défaut » ou « risque plus faible – « vous êtes plus susceptible de payer en totalité » au moment de prendre votre décision.

Note pédagogique : Nous demandons aux étudiants de fournir un résumé écrit de la décision commerciale en question et les limites potentielles de la ensemble de données. Nous nous concentrons spécifiquement sur le cadre temporel et le biais de sélection, comme indiqué dans la section 3.3.

4.3. Étape 3 : Construction du modèle, choix d'une règle de décision, et validation du modèle de régression logistique

Nous guidons nos étudiants à travers le processus de construction d'un modèle de régression logistique pour estimer la probabilité de défaut de l' diverses demandes de prêt. Pour atteindre l'objectif d'apprentissage, comprendre les étapes de la construction et de la validation d'un modèle, nous parcourons les étudiants à travers un modèle itératif de construction en trois phases processus de spécification, d'estimation et d'évaluation, puis valider le modèle.

Pour construire le modèle de régression logistique pour l'étude de cas basée en Californie, nous avons sélectionné au hasard la moitié des données à analyser. nos données « d'entraînement » (1051 des 2102 observations originales). l'ensemble de données « SBA Case », la variable « Selected » indique laquelle les observations sont les données « d'entraînement » et lesquelles sont les données « de test » données (1 données de formation D à utiliser pour construire le modèle, 0 données de test D pour valider le modèle).

Note pédagogique : Il existe un certain nombre de techniques de classification possibles qui peut être utilisé pour modéliser ces données. Étant donné que notre commerce de premier cycle Le cours de statistiques est un cours de service pour les domaines fonctionnels de l'entreprise et un prérequis pour un certain nombre de cours tels que la finance et le marketing, les objectifs d'apprentissage de ce cours sont alignés sur les objectifs globaux de notre collège. objectifs d'apprentissage et les objectifs d'autres cours (qui incluent une compréhension de la régression logistique). Par conséquent, dans cet article, nous présentons notre couverture de la régression logistique de base pour nos étudiants de premier cycle en commerce. Les étudiants des cours de statistiques plus avancés peuvent être capable d'explorer les interactions dans la régression logistique, les covariables dépendantes du temps, ainsi que des méthodes de classification plus avancées.

4.3.1. Spécification et estimation du modèle

Lorsqu'il s'agit d'une réponse binaire, comme c'est le cas ici, la régression logistique est un choix de modèle populaire pour décrire la relation entre la réponse binaire et les variables explicatives.

(prédicteurs). Les modèles de régression logistique enregistrent les probabilités sous forme de courbes linéaires combinaison de variables explicatives (prédicteurs) :

$$P = \frac{e^{b_0 + b_1X_1 + b_2X_2 + \dots + b_KX_K}}{1 + e^{b_0 + b_1X_1 + b_2X_2 + \dots + b_KX_K}}$$

La probabilité d'intérêt P peut alors être obtenue comme

$$PD = \frac{e^{b_0 + b_1X_1 + b_2X_2 + \dots + b_KX_K}}{1 + e^{b_0 + b_1X_1 + b_2X_2 + \dots + b_KX_K}}$$

D

$$= \frac{1}{1 + e^{-b_0 - b_1X_1 - b_2X_2 - \dots - b_KX_K}}$$

où b0 C b1 X1 C b2 X2 C $\phi\phi\phi$ bKXK représente les coefficients et variables explicatives de la ligne généralisée

structure du modèle de régression auriculaire. La probabilité d'intérêt P peut être prédit avec les coefficients estimés.

Lors de la construction du modèle, nous indiquons aux étudiants que la variable dépendante est une variable binaire. Dans notre analyse, la La variable dépendante binaire est « Default » qui est une variable fictive créée à partir de la variable « MIS_Status ». La valeur de « Par défaut » D 1 si MIS_Status D CHGOFF, et « Par défaut » D 0 si MIS_Status D PIF. Par conséquent, le modèle de régression logistique pour ce Le scénario prédit la probabilité d'un défaut de prêt.

Nous soulignons pourquoi le modèle de régression logistique est utilisé, plutôt que la régression linéaire ordinaire, en discutant de la hypothèses de régression linéaire ordinaire et violation de certaines de ces hypothèses avaient été réalisées à l'aide d'une régression linéaire ordinaire. appliquée à cet ensemble de données. Puisque nous avons affaire à une résultat ici (c'est-à-dire par défaut ou non) plutôt qu'une valeur quantitative 1. La régression par les moindres carrés ordinaires n'est pas appropriée. Au lieu de cela, nous utilisons la régression logistique pour prédire les rapports de cotes et probabilités.

Pour les variables explicatives possibles, nous revisitons les résultats de l'étape 1 où sept variables sont identifiées comme potentielles indicateurs de risque. Étant donné que « l'emplacement (état) » et « l'industrie » sont déjà pris en compte en limitant les analyses à un seul état et une industrie, il y a cinq variables qui devraient être prises en compte pour inclusion dans le modèle en tant que variables explicatives : Récession économique (« Récession »), Nouvelles affaires (« Nouvelles »), Prêts Adossé à l'immobilier (« RealEstate »), décaissement brut (« Décaissement brut ») et la part garantie de la SBA Prêt approuvé (« Portion »).

Pour illustrer le processus de construction du modèle, nous présentons aux étudiants deux versions différentes du modèle en utilisant les données de formation : (a) modèle initial avec cinq variables explicatives (Tableau 7(a)), y compris le test du rapport de vraisemblance pour l'effet partiel obtenu à partir d'une analyse de type III du PROC GENMOD de SAS (Tableau 7(b)) ⁷; et (b) modèle re-spécifié avec trois explications variables (tableau 8). Une fois le modèle initial produit, une discussion sur les variables significatives et les valeurs p s'ensuit. les étudiants déterminent que les indicateurs de risque « Nouveau » et « Décaissement brut » ne sont pas statistiquement significatifs et qu'ils suggérer de re-spécifier le modèle sans ces variables. Étant donné l'objectif est la prédiction, le modèle final avec les trois variables explicatives « Immobilier », « Portion » et « Récession » sera utilisé pour classer les prêts dans l'étude de cas en utilisant la décision règles décrites dans la section 4.3.2.

Il convient de mentionner que : (a) les auteurs ont confirmé avec un employé de la SBA avec plus de 30 ans d'expérience qui fait sens économique de supprimer « Nouveau » et « Décaissement brut » de le modèle et (b) il n'y a presque aucune différence dans les taux de mauvaise classification calculés pour les données de test, avec ou sans deux variables « Nouveau » et « Décaissement brut ». Alors que le

⁷ L'analyse de type III teste la signification de chaque effet partiel et la signification de un effet avec tous les autres effets du modèle.

Tableau 7(a). Étude de cas de la Californie : modèle de régression logistique initial avec cinq variables explicatives.

Paramètre	DF	Estimation	Erreur standard	Chi carré de Wald	Pr > ChiSq
Intercepter	1	1,3537	0,3229	17,5729	<0,0001
Nouveau	1	¿0,0772	0,2101	0,1349	0,7134
Immobilier	1	¿2,0331	0,3636	31,2663	<0,0001
Déboursement Brut	1	¿3,37E-7	3,52E-7	0,9173	0,3382
Partie	1	¿2,8298	0,5594	25,5909	<0,0001
Récession	1	0,4971	0,2413	4,2441	0,0394

le modèle ne correspond pas aux données aussi bien qu'on pourrait l'espérer, il donne performance prédictive raisonnable qui est illustrée dans le section de validation (4.3.3) où les « données de test » sont appliquées à le modèle.

Comme l'a déclaré George Box, « Essentiellement, tous les modèles sont faux, mais certains sont utiles » (Box et Draper 1987, p. 424). Et, Seymour Geisser (1993) a affirmé qu'un modèle est utile tant que car il offre de bonnes performances prédictives.

Note pédagogique : La variable « Sélectionné » indique lequel des cas il s'agit de données d'entraînement par rapport à des données de test (1 pour l'entraînement et 0 pour les tests). Cet échantillon aléatoire a été tiré dans SAS à l'aide de SURVEYSELECT procédure : PROC SURVEYSELECT OUTALL OUT D dataca53 METHODE D SRS TAILLE D'ECHANTILLON D 1051 GRAINE D 18467;

En plus de la discussion sur les valeurs p, comment interpréter les estimations des paramètres du modèle en mettant l'accent sur les cotes de défaut est décrit. Par exemple, étant donné que l'immobilier est une variable fictive, nous pouvons interpréter ce coefficient comme : « Étant donné la même partie soutenue par la SBA et considérations économiques (récession ou non), le rapport de cotes estimé de défaut (soutenu par des fonds réels immobilier vs. non adossé à un bien immobilier) est $e_{¿2,1282 D 0,12}$. Ainsi, les chances de défaut lorsqu'elles sont garanties par des biens immobiliers ne sont que de 12 % de défaut lorsqu'il n'est pas adossé à un bien immobilier. Par conséquent, comme prévu, le risque de défaut est moindre lorsque le prêt est adossé à des actifs réels. domaine.

Nous avons pris en compte d'autres variables explicatives et interactions entre les variables fictives « Immobilier » et « Récession » et la variable explicative continue « Portion ». Bien qu'aucune des variables explicatives supplémentaires significatives sont apparues, il y a il y avait deux effets d'interaction significatifs : « Partie immobilière » et « Part de récession » ; cela suggère que la « Part » avait une influence supplémentaire si le prêt impliquait des biens immobiliers ou s'il était survenu pendant la récession. Étant donné que l'interaction dans la régression logistique est un concept complexe à conceptualiser dans ces introductions cours, nous avons décidé de ne pas inclure une discussion sur ces effets d'interaction dans cet article.

4.3.2. Choix d'une règle de décision

Ensuite, les étudiants sont guidés tout au long du processus de choix d'un règle de décision. Nous discutons de la manière dont la probabilité estimée de

le défaut d'un prêt particulier doit être comparé à un seuil probabilité lors de la prise d'une décision, suivie d'une discussion quant à ce que pourrait être une probabilité de coupure appropriée. Les étudiants suggèrent souvent 0,5 comme seuil, un choix évident pour beaucoup car cela équivaut aux probabilités (imputées ou payées en totalité) de 1.

Nous demandons aux étudiants de calculer le taux d'erreur de classification en utilisant différents niveaux de probabilité de coupure. Les résultats sont présentés dans Figure 4.

Le niveau de probabilité de coupure entraînant le taux de classification erroné le plus bas commence autour de 0,5. Le taux de classification erronée commence d'augmenter autour d'un niveau de probabilité de coupure de 0,6. Par conséquent, un un niveau de probabilité de coupure de 0,5 est un bon choix.

Des règles de décisions sont ensuite adoptées :

- (i) classer la demande de prêt dans la catégorie de risque inférieur et approuver le prêt lorsque la probabilité estimée de par défaut 0,5, ou
- (ii) classer la demande de prêt dans la catégorie de risque plus élevé et refuser le prêt lorsque la probabilité de défaut est estimée >0,5.

Note pédagogique : Dans la section 3.3, le potentiel de biais de sélection dû la période de temps utilisée dans les analyses a été discutée. Il devrait être a noté qu'il existe ici une autre source importante de biais de sélection Cependant, il existe une inadéquation critique entre les données utilisées pour construire le modèle prédictif et les prêts qui seront évalués en utilisant le modèle. On peut supposer que seuls les prêts perçus avoir un risque tolérablement faible si jamais ils avaient été approuvés en premier lieu. Cela signifie que tous les prêts représentés dans les données auraient été perçus comme présentant un risque « faible » par quelqu'un. Ceux considérés comme présentant un risque plus élevé risque (et donc, n'ont pas été approuvés) n'apparaissent pas dans les données du tout. Par conséquent, le taux de défaut dans l'échantillon sera probablement plus faible que le taux de défaut réel de toutes les demandes de prêt qui ont été soumises en premier lieu.

4.3.3. Validation et classification erronée

Nous validons le modèle final en l'appliquant à l'autre moitié de les données (les données de « test » qui incluent les 1051 restantes observations pour l'exemple basé en Californie) et évaluer ses performances en calculant le taux d'erreur de classification. Pour ce faire, les étudiants utilisent le modèle de régression logistique final pour générer le probabilité estimée du taux de défaut pour chacun des prêts du

Tableau 7(b). Analyse de type III.

Source	DF	Chi-carré	Pr > ChiSq
Nouveau	1	0,14	0,7130
Immobilier	1	39,96	<0,0001
Déboursement Brut	1	0,97	0,3258
Partie	1	27,41	<0,0001
Récession	1	4,27	0,0389

Tableau 8. Étude de cas basée en Californie : modèle re-spécifié avec trois explications variables.

Paramètre	DF	Estimation	Erreur standard	Wald	Chi-carré	Pr > ChiSq
Intercepter	1	1,3931	1	0,3216	18,7670	<0,0001
Immobilier	1	¿2,1282	1	0,3450	38,0529	<0,0001
Partie	1	¿2,9875	1	0,5393	30,6898	<0,0001
Récession	1	0,4971	0,2412	0,2412	4,3679	0,0366

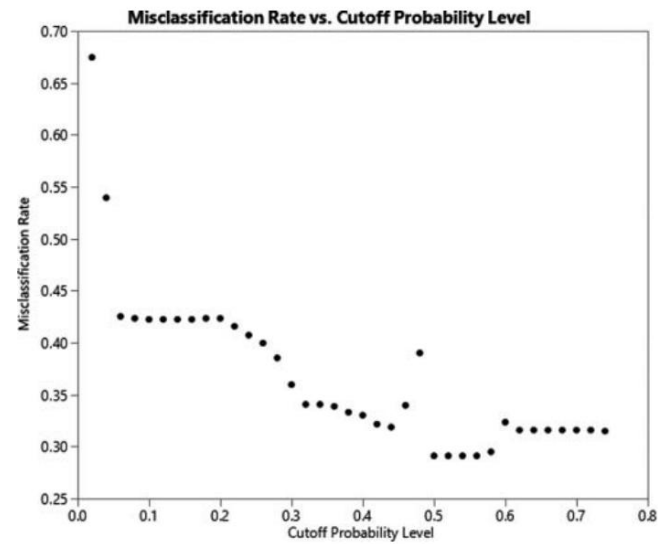


Figure 4. Taux d'erreur de classification par rapport au niveau de probabilité limite.

Échantillon de données « test ». Ensuite, les étudiants sont invités à classer les prêts dans les données de test comme « à risque élevé » ou « à risque faible » en utilisant les règles de décision de la section 4.3.2.

Étant donné que les résultats réels des prêts dans les données de test sont connus (MIS_Status de radié ou payé en totalité), le taux de classification erronée peut être déterminé pour le scénario basé en Californie. Dans le tableau 9, les colonnes représentent la réalité de savoir si un prêt a été radié ou payé en totalité, et les lignes représentent la classification du prêt selon la règle de décision (risque plus élevé ou risque plus faible). Comme indiqué ci-dessous, où le nombre de classifications erronées est représenté en caractères gras, 324 prêts ont été classés à tort comme « risque faible » et 14 prêts ont été classés à tort comme « risque plus élevé ». Le taux global de classification erronée est de 32,16 % ((324 C 14)/1051).

Note pédagogique : En classe, nous discutons de la manière dont ce processus fait partie de l'évaluation des performances prédictives d'un modèle et du fait que ce modèle particulier donne des performances prédictives raisonnables.

Note pédagogique : Étant donné que deux types d'erreurs différents peuvent être commises, à savoir la classification erronée d'un prêt comme « à risque élevé » ou comme « à risque faible », nous encourageons les étudiants à discuter des conséquences de l'un ou l'autre type d'erreur et à déterminer si le fait de traiter les deux types d'erreurs de la même manière est une décision commerciale judicieuse. Les discussions tournent généralement autour du fait que la banque perdra à la fois le capital et les intérêts si un prêt est classé à tort comme « à risque faible » et est ensuite radié, tandis que la banque n'encourra qu'un coût d'opportunité correspondant au montant des intérêts si un prêt est classé à tort comme « à risque élevé ».

Note pédagogique : Dans notre cours de troisième cycle, nous couvrons également la courbe ROC (caractéristique de fonctionnement du récepteur) pour décrire la précision de la classification en

Tableau 9. Scénario basé en Californie : Classification des prêts.

Classification	État de nature : Réalité		
	Prêts radiés	Prêts remboursés en totalité	Total
Risque plus élevé (plus de chances d'être radié)	31	14	45
Risque plus faible (plus de chances d'être payé en totalité)	324	682	1006
Total	355	696	1051

demander aux étudiants de regarder un court didacticiel vidéo sur <http://www.data school.io/roc-curves-and-auc-explained/>. La courbe ROC représente le taux de vrais positifs (sur l'axe des Y) par rapport au taux de faux positifs (sur l'axe des X) pour chaque seuil de probabilité de classification possible, tandis que le taux de mauvaise classification ne concerne qu'un seul seuil de probabilité. L'aire sous la courbe (AUC) est la proportion de la case (l'aire de cette case est 1) sous cette courbe ROC. L'AUC est la plus élevée (supérieure à 0,75) pour le modèle le plus parcimonieux avec trois variables explicatives dans le tableau 8, ce qui indique une performance de classification acceptable (voir Hosmer et Lemeshow 2000, p. 162).

4.4. Étape 4 : Utilisation du modèle pour prendre des décisions

Pour atteindre les objectifs d'apprentissage, apprendre à appliquer la régression logistique pour classer un prêt en fonction de la probabilité de défaut et expérimenter le processus d'investigation pour prendre une décision basée sur un scénario informé par les analyses de données, la dernière étape de ce devoir consiste à demander aux étudiants de répondre à la question initiale d'approuver ou de refuser un ou plusieurs prêts en utilisant : (a) le modèle de régression logistique final généré pour déterminer la probabilité estimée de défaut d'un prêt spécifique et (b) les règles de décision pour classer le prêt. Pour l'exemple basé en Californie, le modèle final avec les indicateurs de risque du tableau 8 est utilisé pour estimer la probabilité de défaut pour les deux demandes de prêt ; la probabilité estimée de défaut pour Carmichael Realty (prêt 1) est de 0,05 et SV Consulting (prêt 2) est de 0,55. En appliquant les règles de décision et la probabilité de coupure de 0,5 de la section 4.3, le prêt 1 est classé comme « à faible risque » et doit être approuvé, et le prêt 2 est classé comme « à risque plus élevé » et doit être refusé (voir tableau 10).

5. Évaluation de l'apprentissage, exploration de méthodes de classification plus avancées et remarques finales

5.1. Évaluation des apprentissages

Comme mentionné précédemment, nous évaluons l'apprentissage des étudiants en élaborant une étude de cas similaire à celle présentée en classe et en l'attribuant aux étudiants pour une note sous forme de lettre. Pour les étudiants de premier cycle, nous les laissons compléter le devoir noté en groupes de trois personnes. Pour les cours de deuxième cycle, les étudiants sont tenus de compléter le devoir individuellement.

Pour les devoirs notés, les étudiants doivent soumettre un rapport expliquant toutes les étapes qu'ils ont suivies (qui doivent refléter les étapes décrites ci-dessus) et une recommandation finale quant à l'approbation ou au refus du ou des prêts. Nous suggérons que le rapport fasse trois pages, plus des tableaux, des figures et des graphiques qui aideraient à illustrer et à étayer leur recommandation. Nous accordons aux étudiants deux semaines pour terminer le devoir après les séances en classe. Pour évaluer leur apprentissage, nous utilisons la grille de notation présentée dans le tableau 11.

5.2. Méthodes avancées de classification pour les diplômés Étudiants

Bien que nous nous soyons concentrés sur la régression logistique dans le cadre de la tâche « Ce prêt doit-il être approuvé ou refusé ? », d'autres méthodes de classification telles que les réseaux neuronaux (voir Odom et Sharda 1990 ; Tam et Kiang 1992 ; Lacher et al. 1995 ; Zhang et al. 1999) et les SVM (voir Chen et al. 2010 ; Kim et Sohn 2010) pourraient être utilisées .

Tableau 10. Résumé du scénario basé en Californie.

Prêt	Nom	Date	Montant du prêt demandé	Portion SBA garantie	Sécurisé par l'immobilier ?	Probabilité estimée de défaut	Approuver?
1 2	Immobilier Carmichael	Actuel (pas de récession)	1 000 000 \$	750 000 \$	Oui	0,05	Oui
	SV Conseil	Actuel (pas de récession)	100 000 \$	40 000 \$	Non	0,55	Non

enseigné en utilisant cet ensemble de données dans des cours d'analyse de données de niveau supérieur plus avancés.

Dans notre cours d'exploration de données de troisième cycle, nous insistons auprès des étudiants sur le fait que les modèles paramétriques traditionnels tels que la régression logistique reposent sur des hypothèses strictes. Lorsque ces hypothèses ne sont pas vérifiées, les méthodes de classification non paramétrique non linéaires telles que les réseaux neuronaux et les SVM constituent de puissantes alternatives. Les réseaux neuronaux (feed-forward) sont des modèles de régression non linéaire flexibles avec de nombreux paramètres, reliant les entrées (variables explicatives ou prédicteurs) aux sorties (la variable dépendante) via des couches cachées entre les entrées et les sorties. La « fonction d'activation » des unités de la couche cachée est généralement la fonction logistique (voir Venables et Ripley 2002, sec. 8.10). La régression logistique est équivalente au réseau neuronal sans nœud caché (Zhang et al. 1999), et il est naturel de comparer les résultats du réseau neuronal à ceux de la régression logistique. Si l'objectif d'apprentissage d'une tâche est de séparer les prêts des prêts susceptibles de faire défaut sans avoir besoin de la probabilité de défaut prédite, alors les réseaux neuronaux et les SVM sont de bons choix.

Note pédagogique : Nous commençons notre introduction aux réseaux neuronaux en montrant comment appliquer la fonction de réseaux neuronaux « nnet » dans R aux données d'entraînement et de test que les étudiants avaient déjà utilisées pour la régression logistique. Nos étudiants ont essayé deux configurations de réseaux neuronaux : aucune couche cachée et une couche cachée de 5 unités. Nous passons ensuite à la discussion de certains aspects théoriques des réseaux neuronaux.

Note pédagogique : Venables et Ripley (2002) fournissent une courte introduction très lisible avec des instructions claires pour utiliser le package de réseaux neuronaux « nnet » dans R. Nous demandons également aux étudiants de consulter la mise à jour

documentation pour « nnet » avec des exemples sur <http://cran.r-project.org/web/packages/nnet/nnet.pdf>. L'ajustement d'un tel modèle de réseau neuronal peut être facilement réalisé par nos étudiants diplômés avec quelques lignes de code en R.

Avec la même tâche décrite ci-dessus, les étudiants diplômés ont pu facilement adapter le modèle de réseaux neuronaux avec les mêmes variables explicatives et données de formation en utilisant R et obtenir un taux d'erreur de classification légèrement inférieur de 31,97 % ((324 C 12)/1051) pour les données de test. Le code R est : #Données sur les réseaux neuronaux <- read.csv(fichier D « C:/SBACase.csv », en-tête D TRUE, sep D « , ») summary(données) attach(données) x1 D Immobilier x2 D (Portion-mean(Portion))/sqrt(var(Portion)) x3 D Récession y D as.factor(MIS_Status) dat D data.frame(x1,x2,x3,y) library(nnet) train D (Selected>0) nnfit D nnet(y ~., données D dat[train,], taille D 5, sauter D VRAI, rang D 0,02, décroissance D 1e-3, maxit D 10000) résumé(nnfit) test D dat[!train,] model_pred D prédire(nnfit, test, type D « classe ») table(model_pred, test\$y) Une autre méthode de classification populaire pour ce problème de classification binaire (payé en totalité ou imputé) est celle des SVM. SVM est une extension du classificateur à vecteur de support, qui est étroitement lié

Tableau 11. Grille d'évaluation pour le devoir.

Pas (poids)	Ne répond pas aux attentes	S'approche des attentes	Conforme aux attentes	Dépasse les attentes
1) Identifier les indicateurs de risque potentiel (30%)	Identifie les indicateurs de risque potentiel qui n'ont pas de sens	Identifie les indicateurs de risque potentiel, mais ne fournit pas de justification raisonnable	Identifie les indicateurs de risque potentiel fournit une justification raisonnable expliquant pourquoi les variables doivent être prises en compte ou non, et fournit des preuves à l'appui (analyses).	Identifie les indicateurs de risque potentiel et
2) Comprendre l'étude de cas (10%)	Comprend l'étude de cas, un synopsis inexact et/ou déroutant de la décision commerciale en question, mais n'inclut pas la discussion des limites de l'ensemble de données	Comprend l'étude de cas en fournissant	Comprend l'étude de cas en fournissant un synopsis de la décision commerciale en question et inclut une discussion sur limitations liées soit au délai soit au biais de sélection	Comprend l'étude de cas en fournissant un synopsis de la décision commerciale en question et inclut une discussion sur les limites liées au délai et au biais de sélection
3) Construction du modèle, création des règles de décision et validation du modèle de régression logistique (50 %)	Construit un modèle qui n'a pas de sens	Crée un modèle qui fait ne crée pas de règle de décision ou de décision appropriée, mais ne valide pas adéquatement le modèle	Crée un modèle qui a du sens, mais qui	Crée un modèle qui a du sens, crée une règle de décision appropriée et valide adéquatement le modèle
4) Utiliser le modèle pour prendre des décisions (10%)	Dérivé d'une inexactitude estime la probabilité de défaut pour les deux prêts et prend de mauvaises décisions pour les deux prêts, en utilisant leur modèle	Dériver l'estimation correcte probabilité de défaut pour l'un ou les deux prêts, mais prend de mauvaises décisions pour les deux prêts, en utilisant leur modèle	Dériver l'estimation correcte probabilité de défaut pour les deux prêts et prend une bonne décision pour un prêt (mais pas pour l'autre), en utilisant leur modèle	Dériver l'estimation correcte probabilité de défaut pour les deux prêts et prend une bonne décision pour les deux prêts, en utilisant leur modèle



lié à la régression logistique (voir James et al. 2013, chap. 9.5).

Il est donc naturel de demander aux étudiants de comparer la SVM avec la régression logistique. En classe, les étudiants peuvent facilement adapter la SVM aux données à l'aide de la fonction « SVM » de la bibliothèque R e1071. On a constaté que les erreurs de classification étaient plus élevées que celles de la régression logistique ou des réseaux neuronaux.

5.3. Remarques finales

En conclusion, cet ensemble de données riche offre aux enseignants la possibilité de créer des tâches utiles pour enseigner une gamme de concepts statistiques et mettre en évidence la manière dont les données peuvent être utilisées pour éclairer les décisions commerciales réelles. Conformément aux recommandations GAISE de 2016, l'étude de cas « Ce prêt doit-il être approuvé ou refusé ? » est un excellent exemple de la manière de promouvoir l'apprentissage actif et d'enseigner la pensée statistique dans un contexte commercial à l'aide de données réelles.

Nous encourageons les autres à réfléchir à des façons créatives d'intégrer les données dans des tâches alternatives. Nous espérons que les enseignants partageront leurs tâches avec la communauté de l'enseignement statistique afin d'améliorer l'efficacité de l'enseignement dans le domaine des statistiques.

Matériel supplémentaire

Des informations complémentaires à cet article sont accessibles sur le [site Web de l'éditeur](#). Cela comprend les fichiers de données « National SBA » et « SBA Case » et leur documentation correspondante.

Références

Bryant, PG (1999), « Discussion, débat et désaccord : enseignement de la régression multiple par discussion de cas », Bulletin de l'Institut international de statistique, actes de l'Institut international de statistique, vol. 58, livre 2, Helsinki : Institut international, pp. 215–218.

Box, GEP, et Draper, NR (1987), Construction de modèles empiriques et surfaces de réponse, New York : Wiley, p. 424.

Chen, S., H€ardle, WK, et Moro, RA (2010), « Modélisation du risque de défaut avec des machines à vecteurs de support », Quantitative Finance, 11, 135–154.

Rapport du Collège GAISE, Comité de révision de l'ASA (2016), « Lignes directrices pour l'évaluation et l'enseignement de l'enseignement des statistiques, rapport du Collège 2016 », disponible sur <http://www.amstat.org/education/gaise>.

Geisser, S. (1993), Inférence prédictive : une introduction, New York : Chapman & Hall.

Hosmer, DW, et Lemeshow, S. (2000), Applied Logistic Regression (2e éd.), New York : Wiley.

James, G., Witten, D., Hastie, T., et Tibshirani, R. (2013), Une introduction tion à l'apprentissage statistique, New York : Springer.

Kim, HS, et Sohn, SY (2010), « Machines à vecteurs de support pour la prédiction des défauts des PME en fonction du crédit technologique », Revue européenne de recherche opérationnelle, 201, 838–846.

Lacher, RC, Coats, PK, Sharma, SC, et Fant, LF (1995), « Un réseau neuronal pour classer la santé financière d'une entreprise », European Journal of Operational Research, 85, 53–65.

Nolan, D., et Speed, TP (1999), « Enseigner la théorie statistique par le biais d'applications », The American Statistician, 53, 370–375.

Odom, MD, et Sharda R. (1990), « Un modèle de réseau neuronal pour la prédiction des faillites », Actes de la conférence internationale IEEE sur les réseaux neuronaux, II, 163–168.

Parr, WC, et Smith, MA (1998), « Développement de cours de statistiques commerciales basés sur des cas », The American Statistician, 52, 330–337.

Smith, M., et Bryant, P. (2009), « Gestion des discussions de cas dans les cours d'introduction aux statistiques commerciales : approches pratiques pour les instructeurs », The American Statistician, 63, 348–355.

Tam, KY, et Kiang, MY (1992), « Applications managériales des réseaux neuronaux : le cas des prévisions de faillite bancaire », Management Science, 38, 926–947.

US Small Business Administration (2015), Historique récupéré le 22 août 2015 à partir de <https://www.sba.gov/about-sba/what-we-do/history>.

Venables, WN, et Ripley, BD (2002), Statistiques appliquées modernes avec S (4e éd.), New York : Springer.

Zhang, G., Hu, MY, Patuwo, BE, et Indro, DC (1999), « Réseaux neuronaux artificiels dans la prévision des faillites : cadre général et analyse de validation croisée », Revue européenne de recherche opérationnelle, 116, 16–32.