RAPPORT FINAL

PAREA 2015-0011

Persévérance et réussite scolaire par le forage de données d'éducation

Jonathan Guillemette, John Abbott College
Sameer Bhatnagar, Dawson College
Michael Dugdale, John Abbott College
Sahir Bhatnagar, McGill University
Nathaniel Lasry, John Abbott College

Déposé le 25 juillet 2018
Cette recherche a été effectuée à partir d'une subvention du Ministère de l'Éducation et de l'Enseignement supérieur. Les auteurs assument la responsabilité du contenu du rapport.

TABLE DES MATIÈRES

CHAPITRE 1

CONTEXTE ET PROBLÉMATIQUE

Bien que la société québécoise, dans les années 1960, faisait partie des sociétés les moins scolarisées (Donald, 1997), un demi-siècle plus tard, le taux de diplomation au Québec se trouve à être parmi les plus élevés au monde (Statistique Canada, 2016). Ces changements significatifs ont vu le jour en 1963 avec les recommandations de la *Commission royale d'enquête sur l'enseignement de la province de Québec*, connu sous le nom de *Commission Parent* du nom de son président, Monseigneur Alphonse-Marie Parent, alors recteur de l'Université Laval. S'appuyant sur les recommandations de la *Commission Parent,* le gouvernement provincial de l'époque met en place un système d'enseignement supérieur à deux niveaux successifs : 1) les collèges d'enseignement général et professionnel (cégeps) et 2) les universités.

La création du système collégial, système unique au Québec jusqu'à aujourd'hui, voit le jour avec les premiers cégeps fondés en 1967. Le nombre de cégeps a continué à croître de décennie en décennie, jusqu'à nos jours, de sorte qu'il existe aujourd'hui des cégeps dans chacune des régions administratives du Québec (Donald, 1997). Actuellement, le taux d'accès moyen au cégep est de plus de 60 % bien que le taux d'accès pour les garçons soit d'environ 15 % inférieur à celui des filles qui, lui, dépasse les 70 % (Kamazani, Uzenat et St-Onge, 2018). En effet, selon un rapport du MELS de 2014, le taux de diplomation au collégial a augmenté depuis la création des cégeps : de 22 % en 1976, il est passé à 34,4 % en 1986, à 39,4 % en 1996 et à 50 % en 2006 et en 2011. Cela dit, comme un verre à moitié plein peut aussi être perçu comme étant à moitié vide, que peut-on faire pour la moitié de la population qui n'obtient pas de diplôme collégial?

Plus d'un million d'élèves sont actuellement inscrits à la formation générale des jeunes (études primaires et secondaires) et se dirigent donc vers des études au collégial (D'Arisso, 2018). Que sait-on sur les facteurs qui peuvent prédire la réussite et la persévérance scolaires au collégial? Ce rapport présente trois années de recherches comportant des données colligées sur plus de 120 000

étudiants du collégial inscrits dans les cégeps Dawson, John Abbott et Vanier. L'objectif de cette recherche est simple : utiliser des méthodes d'analyses classiques et modernes pour trouver les facteurs qui prédisent le mieux la réussite et la persévérance scolaires au collégial.

PROBLÉMATIQUE

Le budget provincial de 2017-2018 du gouvernement du Québec (Secrétariat du Conseil du trésor, 2017) allouait 17,9 milliards de dollars à *l'éducation et l'enseignement supérieur*, ce qui correspond à 22,3 % du budget total de la province, soit la deuxième catégorie de dépense en importance après la santé à 36,8 milliards de dollars. Ces investissements gouvernementaux sont basés sur les données probantes démontrant que l'éducation postsecondaire est une source de bénéfices monétaires et sociaux pour les individus et la société en entier (McMahon, 2010). Or, pour maximiser ces investissements, les élèves inscrits doivent être en mesure de compléter avec succès leur formation.

La persévérance et la réussite scolaires (PRS) sont un sujet d'intérêt public ainsi qu'une priorité gouvernementale. En effet, un sondage révélait que « 97 % des Québécois estiment que le gouvernement devrait faire de la lutte au décrochage scolaire une priorité » et que « 93 % des Québécois jugent que le gouvernement devrait investir davantage en éducation » (Massé, 2009). Depuis un peu plus d'une décennie, des organismes subventionnaires comme le ministère de l'Éducation, du Loisir et du Sport, les Fonds de Recherche du Québec et certaines fondations privées se penchent sur la question de la persévérance et la réussite scolaires[1]. Ces études s'intéressent principalement à la PRS dans les écoles primaires et secondaires. Cependant, le passage au collégial est extrêmement difficile pour plusieurs étudiants. Ils se retrouvent exposés à un nouveau système d'éducation simultanément plus flexible et plus contraignant que le secondaire. Plusieurs ont du mal à s'adapter et à suivre le rythme collégial. Par conséquent, la réussite, au collégial, se situe fréquemment en deçà des attentes des étudiants, de leurs enseignants, des administrateurs et des organismes gouvernementaux. Depuis plus d'une décennie, près de 20 millions de dollars ont été

[1] Voir sur le site du MELS: http://www.mels.gouv.qc.ca/dossiers-thematiques/lutte-contre-le-decrochage-et-reussite-scolaire/programme-de-recherche-sur-la-perseverance-et-la-reussite-scolaires/

affectés au problème de la PRS[2]. Conséquemment, plusieurs facteurs influençant la PRS ont été identifiés. Ces facteurs étant maintenant connus, une approche de résolution du problème à la fine pointe de la technologie est de mise.

Ce que nous présentons dans ce projet de recherche est une vaste étude portant sur les données d'apprentissage provenant de plus de 122 000 étudiants du collégial. L'analyse comme telle est technique et comprend la construction de modèles analytiques qui utilisent les facteurs précédemment identifiés comme ayant un effet sur la PRS pour détecter le risque de décrochage avant qu'il se produise. Dans un deuxième temps, nous avons mis au point des modèles informatiques qui identifient le profil de l'étudiant à risque pour nous permettre de déterminer comment remédier aux difficultés propres à chaque étudiant. Dans un troisième temps, nous laissons de côté tous les savoirs et résultats préalables sur la PRS et utilisons des méthodes d'intelligence artificielle pour mettre au point des modèles capables de prédire la PRS à partir de nos données et non pas à partir de variables que nous estimons être pertinentes. En laissant les algorithmes d'apprentissage machine trouver les facteurs pertinents, nous sommes alors en mesure de produire des classificateurs puissants permettant d'identifier le plus tôt possible les étudiants les plus à risque. Quelle que soit l'approche préconisée, notre but, tout au long de cette étude, demeure constant : effectuer un forage de données en éducation pour minimiser les taux d'abandon et maximiser la réussite des étudiants du collégial.

RICHES EN DONNÉES, PAUVRES EN INFORMATION

Les salles de cours des cégeps se sont grandement modernisées. Tableaux tactiles et ordinateurs ornent la salle tandis que plusieurs étudiants suivent et interagissent avec leurs dispositifs mobiles. Nos étudiants sont plus connectés que jamais. Aujourd'hui, les enseignants font participer leurs étudiants à des activités informatisées et accumulent régulièrement un grand nombre de données. Les établissements d'enseignements ont accès à des données qui représentent le rendement scolaire de chaque élève au secondaire, leurs résultats aux examens d'admission, le programme

[2] Voir sur le site du FRQ : http://www.frqsc.gouv.qc.ca/fr/bourses-et-subventions/consulter-les-programmes-remplir-une-demande/bourse/developpement-d-un-reseau-de-recherche-sur-la-perseverance-et-la-reussite-scolaires--action-concertee-d2ajzm5q1416840286215

d'étude de l'élève, etc. Les établissements d'enseignement conservent aussi des données démographiques comme le sexe de l'élève, son âge et même son code postal, un indicateur du statut socioéconomique (Demissie, Hanley, Menzies, Joseph et Ernst, 2000; Deonandan, Ostbye, Tummon, Robertson et Campbell, 2000). De leur côté, un nombre grandissant d'enseignants sondent les élèves et colligent maintes formes de données (Burnstein et Lederman, 2001, 2003; Lasry, 2008; Lee, Ding, Reay et Bao, 2011) pour évaluer l'état des connaissances ou de la compréhension de leurs élèves. Les enseignants utilisent de plus en plus de systèmes de devoirs en ligne. Certains de ces systèmes sont maintenant fournis par les éditeurs de manuels scolaires (Pearson-MyLab, WileyPLUS, Norton-SmartWork). D'autres sont des entreprises commerciales (Blackboard, Edmodo ou Webassign) ou des logiciels libres (Moodle, WebWork, ou LON-CAPA). Les élèves qui utilisent ces logiciels pour leurs devoirs génèrent en quelques heures, souvent de façon hebdomadaire, une quantité considérable de données. Ces données peuvent correspondre à des notes obtenues sur des problèmes, mais aussi le temps pris pour les résoudre, le nombre d'essais, le nombre de demandes d'indices. La plupart de ces données ne sont jamais utilisées à leur plein potentiel. Comment analyser de façon sensée toutes les données colligées?

Le premier obstacle est l'analyse d'une quantité considérable et diverse de données. Un élève peut facilement créer plus de mille éléments de données par trimestre. Le second problème, plus important encore que le premier, survient lorsqu'on essaye de faire une gestion efficace et une analyse sensée de données qui sont très différentes. En effet, une augmentation du type de données se traduit par une explosion combinatoire des hypothèses probables. Il est difficile, voire impossible, d'effectuer une analyse exhaustive et sensée lorsque la quantité est grande et que le type de données est très varié. Au lieu d'émettre une hypothèse plausible parmi un océan d'hypothèses probables, le forage de données permet à un ou plusieurs modèles informatiques d'être créés pour « laisser parler les données » et découvrir les structures relationnelles sous-jacentes.

OBJECTIF GLOBAL : Prédicteurs de persévérance et de réussite

L'objectif premier de cette étude est de maximiser la persévérance et la réussite scolaires. Pour atteindre cet objectif, nous utilisons le forage de données, une méthode puissante et

relativement nouvelle pour atteindre notre objectif. L'accroissement de la puissance de calcul des ordinateurs actuels permet non seulement d'emmagasiner beaucoup de données, mais aussi d'analyser de grandes quantités de données variées. Le forage de données, connu sous le nom de « data-mining » en anglais, est le processus par lequel on découvre des structures émergentes dans de grands ensembles de données. Le forage de données est un domaine interdisciplinaire à l'intersection de « l'apprentissage machine » et des statistiques appliquées. On y utilise des algorithmes informatiques pour trouver des relations entre des variables qui, *a priori,* n'ont aucun rapport entre elles. Ces structures permettent de regrouper des points de données qui sont en apparence distincts ou bien de différencier des groupes semblables en apparence (Luan, 2002). En un mot, le forage de données permet de trouver des relations prédictives là où les humains n'en sont souvent pas capables. Les études précédentes sur la persévérance et la réussite scolaires sont en mesure de nous indiquer quels facteurs sont importants à examiner. En revanche, le forage nous permet d'utiliser ces facteurs pour déterminer quelles combinaisons de facteurs, ou quelles interactions entre des facteurs, permettent de prédire le décrochage scolaire ou d'optimiser la réussite scolaire.

Un des objectifs principaux du forage de données consiste à utiliser les données pour prédire les caractéristiques des données futures. Au cours des dernières années, le forage de données a grandement gagné en popularité dans le monde des affaires (UNESCO, 2012). Il permet aux gens d'affaires d'obtenir une image des ventes, par intervalle de temps et par secteur, et d'y découvrir des relations qui sont parfois inattendues (Long et Siemens, 2011). Un exemple mythique, mais fréquemment cité, est celui du forage de données fait par Walmart. La compagnie découvre une relation entre la vente de couches pour bébé les jeudis soirs et la vente de bière. Walmart décide de placer un étalage de bière haut de gamme près de l'étalage de couches de bébé pour stimuler leurs ventes et augmenter les revenus.

Certains diront que les applications pédagogiques du forage de données sont plus nobles. Étant à leurs débuts, ces applications restent largement inaccessibles à la plupart des enseignants (Romero, Ventura et Garcia, 2008). En partant des prémisses du forage de données, nous envisageons un système dont l'objectif est de trouver des relations inattendues dans l'apprentissage. Par exemple, une étude de San Pedro *et coll*. démontrait que l'usage d'un système de devoirs en

ligne au début du secondaire permettait de prédire le taux de présence des étudiants dans leurs cours au collégial (San Pedro, Baker, Bowers et Heffernan, 2013). Certains liens sont prévisibles, comme les liens entre la présence en classe et l'achèvement des devoirs sur la réussite d'un cours. Cependant, notre principal objectif est de trouver des relations inattendues dans le domaine de l'éducation qui sont équivalentes à celle de la bière et des couches pour bébé pour Wal-Mart. Il est important d'avoir en tête d'autres corrélations que celles entre les différents modes d'évaluation de la session.

Un de nos premiers objectifs avec le forage de données en éducation est de pouvoir dépister les étudiants à risque, et ce, dès que possible. Après une collecte initiale de données, un modèle informatique est créé. Les questions qui orienteront la création du modèle informatique sont de type : quelles variables permettent de prédire quels étudiants sont les plus susceptibles d'échouer ou d'abandonner un cours, de changer de programme ou encore de décrocher complètement?

Comment définir le décrochage?

Un des obstacles majeurs à la poursuite de nos objectifs est qu'il est actuellement difficile à définir le décrochage. Un étudiant inscrit lors d'une session, mais qui n'est plus inscrit la session suivante pourrait encore être aux études, mais avoir changé d'institution, de province ou bien avoir planifié un voyage de quelques mois avant de retourner aux études. De définir avec certitude le décrochage à partir de données colligées par les institutions académiques est difficile, voire impossible. C'est pourquoi la plupart des indicateurs actuels sont indirects et focalisent sur les taux de graduation, un paramètre plus facilement mesurable. L'objectif est donc de porter notre attention sur la maximisation des diplômés plus que la minimisation des décrocheurs puisque nous n'avons pas d'accès direct à ces données. Nos approches portent donc sur les prédicteurs de réussite, car celui-ci est facilement mesurable via les indicateurs de graduation.

RÉFÉRENCES

Arnold, K. E. et Pistilli, M. D. (2012). *Course Signals at Purdue: Using Learning Analytics to Increase Student Success.* Document présenté lors de l'International Conference on Learning Analytics and Knowledge, à Vancouver, en C.-B.

Baillargeon, G., Demers, M., Ducharme, P., Foucault, D., Lavigne, J., Lespérance, A., . . . Vigneault, A. (2001). Education Indicators, édition 2001. Québec, QC : ministère de l'Éducation, Gouvernement du Québec.

Burnstein, R. et Lederman, L. (2001). Using Wireless Keypads in Lecture Classes. *PHYSICS TEACHER, 39*(1), 8-13.

Burnstein, R. et Lederman, L. (2003). Comparison of Different Commercial Wireless Keypad Systems. *The Physics Teacher, 41*, 272.

Demissie, K., Hanley, J. A., Menzies, D., Joseph, L. et Ernst, P. (2000). Agreement in measuring socio-economic status: area-based versus individual measures. *Chronic Dis Can, 21*(1), 1-7.

Deonandan, R., Ostbye, T., Tummon, I., Robertson, J. et Campbell, K. (2000). A comparison of methods for measuring socio-economic status by occupation or postal area.

Donald, JG (1997). Higher education in Quebec: 1945-1995. Dans GA Jones (éd.), Higher Education in Canada: Different systems, different perspectives (p.161-188). New York, NY: Garland

Ewell, P. et Wellman, J. (2007). Enhancing student success in education. *National Postsecondary Education Cooperative (NPEC)*.

Hill, D., Rapoport, A., Lehming, R. et Bell, R. (2007). Changing U.S. Output of Scientific Articles: 1988-2003. Arlington, VA : National Science Foundation.

Kamanzi, P. C., Uzenat, M. et St-Onge, M. (2018). Évolution de l'enseignement supérieur : à la croisée de la démocratisation des études et de l'économie du savoir. Dans J. Masdonati et C. Montmarquette (dir.), « Le Québec économique. Éducation et capital humain. » (pp. 119-150) Québec : Presses de l'Université Laval.

Lasry, N. (2008). Clickers or Flashcards: Is There Really a Difference? *The Physics Teacher, 46*, 242.

Lee, A., Ding, L., Reay, N. W. et Bao, L. (2011). Single-concept clicker question sequences. *The Physics Teacher, 49*(6), 385-389.

Lehr, C. A., Hansen, A., Sinclair, M. F. et Christenson, S. L. (2003). Moving Beyond Dropout Towards School Completion: An Integrative Review of Data-Based Interventions. *School Psychology Review*.

Long, P. et Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *Educause Review, 46*.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education. Toronto, Canada: Annual Forum for the Association for Institutional Research.

Massé, D. (2009). Réussite éducative, abandon et décrochage : Nécessité d'une nouvelle stratégie pour le Québec? *La revue des Échanges, 98*, 4.

McMahon, W. (2010). The private and social benefits of higher education: The evidence, their value, and policy implications. Advancing Higher Education, 1-12.

Osborne, J. et Dillon, J. (2008). *Science education in Europe: Critical reflections* (Vol. 13): London: The Nuffield Foundation.

Romero, C., Ventura, S. et Garcia, E. (2008). Data mining in course management systems: Moodle case study & tutorial. *Computers and Education*(51), 368-384.

Rosenfield, S., Dedic, H., Dickie, L., Rosenfield, E., Aulls, M., Koestner, R., . . . Abrami, P. (2005). *Étude des facteurs aptes à influencer la réussite et la rétention dans les programmes de la science aux cégeps anglophones* : Fonds québécois de la recherche sur la société et la culture.

San Pedro, M. O. Z., Baker, R. S., Bowers, A. J. et Heffernan, N. T. (2013). *Predicting college enrollment from student interaction with an intelligent tutoring system in middle school.* Paper presented at the Proceedings of the 6th international conference on educational data mining.

Secrétariat du Conseil du trésor (2017). Budget des dépenses 2017-18. Plans annuels de gestions des dépenses des ministères et organismes.

Srikant, R., & Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements*: Springer.

Statistique Canada (2016). Indicateurs de l'éducation au Canada : Une perspective internationale, 2015. Ottawa, ON : Statistique Canada.

UNESCO. (2012). Learning Analytics. Moscow: Institute for Information Technologies in Education.

CHAPITRE 2

PROBLÈME D'ACCÈS ET DE COLLECTE DE DONNÉES

L'étude que nous présentons dans ce rapport consiste en plusieurs formes d'analyses de données nous permettant d'examiner les facteurs associés à la réussite et à la persévérance scolaires pour prédire certains risques d'échec ou de décrochage et de fournir des outils qui permettent d'intervenir en temps et lieu. Pour ce faire, notre première grande difficulté s'est avérée être, non pas la grande quantité de données à extraire et à préparer pour nos analyses, mais plutôt notre capacité à obtenir ces données des établissements d'enseignement impliqués. En effet, pour pouvoir obtenir des données historiques anonymisées de chacun des trois cégeps impliqués, nous devions passer par les comités d'éthique de chacun des trois cégeps. Bien que nous demandions des données historiques (aucun impact sur les étudiants actuels) anonymisées à l'aide de clés de cryptage unidirectionnelles (Hashing algorithm SH-256) qui nous empêchaient de connaître l'identité des étudiants, les comités d'éthique étaient néanmoins très prudents à cet égard.

La difficulté majeure se trouvait possiblement dans le fait que notre équipe était la première du genre à faire ce genre de requête. Les questions légales qui se posaient pour les comités d'éthiques sortaient donc de la norme et posaient alors des difficultés procédurales. Pour expliquer simplement, nous proposions d'obtenir des données anonymisées grâce à un processus irréversible d'anonymisation ('hashing algorithm'). L'anonymisation des noms et

prénoms ainsi que des codes postaux de chaque élève ne nous permettait pas de savoir qui étaient les sujets de notre étude. De plus, notre principal intérêt consistait à compiler une grande quantité de données plutôt que des données propres à une personne en soi. Pour venir à bout des longs va-et-vient entre les différents comités d'éthique des trois établissements d'enseignement, nous avons demandé à un professeur de droit à l'Université de Montréal de nous rédiger un avis juridique quant à l'admissibilité de notre requête.

Notre équipe est infiniment reconnaissante envers M^{e} Pierre Trudel, avocat et professeur de droit, pour l'avis juridique qu'il nous a rédigé. Cette lettre conclut que les données que nous demandions, étant anonymisées, ne pouvaient donc pas être sujettes à la *Loi sur les renseignements personnels* étant donné qu'aucune personne pouvant être identifié n'apparaissait dans les données. Étant donné l'importance de ce document et son impact sur notre capacité d'effectuer l'étude, nous concluons ce chapitre avec cette lettre qui a joué un rôle déterminant. Étant vraisemblablement le premier groupe de recherche à faire une demande inter-établissement de données de cette ampleur (données sur plus de 120 000 élèves), nous souhaitons ouvrir la voie et permettre aux chercheurs intéressés par les données anonymisées en éducation de comprendre les mécanismes légaux qui leur permettent de faire ce genre d'étude.

****

Le 15 février 2016

M. Nathaniel Lasry
Department of Physics
Room HS-218
John Abbott College
21275 Rue Lakeshore
Saint-Anne-de-Bellevue, QC, H9X 3L9

Objet : Le statut de "données personnelles" des renseignements anonymisés transmis par les collèges dans le cadre du projet « Persévérance et réussite scolaire par le forage de données d'éducation »

– N/D : T 348

Cher collègue,

La présente analyse répond à la question de savoir si les données qui seront transmises par les collèges ayant accepté de coopérer au projet de recherche « Persévérance et réussite scolaire par le forage de données d'éducation » sont des renseignements personnels au sens de la loi applicable au Québec.

En d'autres mots, il s'agit de déterminer si l'équipe engagée dans le projet de recherche décrit ci-après reçoit communication de renseignements personnels au sens de la législation régissant les collèges et autres organismes publics québécois, à savoir la *Loi sur l'accès aux documents des organismes publics et sur la protection des rensei-gnements personnels* ***[1]***.

Le projet de recherche

Le projet de recherche intitulé « Persévérance et réussite scolaire par le forage de données d'éducation » entend mettre à profit diverses méthodes de forage de données afin d'éclairer les problématiques de la persévérance et de la réussite scolaire des élèves de niveau collégial.

À cette fin, les chercheurs recevront des collèges ayant accepté de collaborer au projet plusieurs ensembles de données portant sur les élèves.

Le projet de recherche vise à mener des analyses par forage de données afin de mieux comprendre les multiples facteurs pouvant avoir une influence sur la persévérance et la réussite scolaire.

Le projet a pour « objectif intermédiaire (…) d'utiliser le forage de données, une méthodologie puissante et relativement nouvelle, pour atteindre (…) cet objectif.

Connu sous le nom de « data-mining » en anglais, le forage de données est un « processus par lequel on découvre des structures émergentes dans de grands ensembles de données ».

On y utilise des algorithmes informatiques recherchant des associations entre des variables qui, *a priori,* n'ont aucun rapport entre eux. Ces structures permettent de regrouper des points de données qui sont en apparence distincts ou bien de différencier des groupes en apparence similaires (Luan, "Data Mining and Knowledge Management in Higher Education, Toronto, Annual Forum for the Association for Institutional Research, 2002).

Le forage sera effectué afin de déterminer quelles combinaisons de facteurs, ou interactions entre facteurs permettent de prédire le décrochage ou permet d'optimiser la réussite scolaire.

Il est prévu que l'équipe de recherche ne sera en aucun moment en mesure d'identifier des élèves à partir des données reçues de la part des institutions collégiales participantes. La désidentification des données sera en quelque sorte inhérente aux requêtes de données auprès du serveur CLARA. En ayant recours à la technique du hachage unidirectionnel, la requête auprès du serveur est accomplie en effectuant une désidentification des données portant sur les élèves avant de livrer ces ensembles de données à l'équipe de recherche.

La notion de renseignement personnel

Dans la *Loi sur l'accès aux documents des organismes publics et sur la protection des rensei-gnements personnels*, la notion de renseignement personnel est extrêmement large. L'article 54 dispose que « dans un document sont personnels les renseignements qui concernent une personne physique et permettent de l'identifier ».

Selon l'article 53, les renseignements personnels sont confidentiels sauf situations exceptionnelles qui n'ont pas ici d'application.

L'article 59 prévoit qu'un « organisme public ne peut communiquer un renseignement personnel sans le consentement de la personne concernée ». Les exceptions à ce principe sont limitées aux cas où les renseignements sont nécessaires pour lutter contre le crime et les autres infractions aux lois [2], aux situations d'urgence mettant la sécurité ou la vie de la personne concernée en danger [3] et à des fins d'étude ou de recherche[4]. Les autres excep-tions importantes à la confidentialité des renseignements personnels sont celles ayant trait aux ententes de transfert de tels renseignements entre organismes.

En plus d'affirmer le caractère confidentiel des renseignements personnels, la loi en régle-mente la collecte. Celle-ci est interdite sauf si elle est nécessaire à l'exercice des attributions d'un organisme ou à la mise en œuvre d'un programme dont il a la gestion [5]. Elle est égale-ment permise si « cela est nécessaire à l'exercice des attributions ou à la mise en œuvre d'un programme

de l'organisme public avec lequel [l'organisme public] collabore pour la prestation de services ou pour la réalisation d'une mission commune » [6]. Ce principe suppose donc que les organismes évaluent les corrélations entre les impératifs des programmes qu'ils administrent et l'information demandée aux personnes.

En vertu de la L*oi concernant le cadre juridique des technologies de l'information*, les fichiers qui seront transmis à l'équipe de chercheurs sont des « documents ». Au sens de cette loi d'application générale, un document est constitué d'information portée par un support. L'information y est délimitée et structurée, de façon tangible ou logique selon le support qui la porte, et elle est intelligible sous forme de mots, de sons ou d'images. L'information peut être rendue au moyen de tout mode d'écriture, y compris d'un système de symboles transcriptibles sous l'une de ces formes ou en un autre système de symboles.

Au sens de la L*oi concernant le cadre juridique des technologies de l'information*, est assimilée au document toute banque de données dont les éléments structurants permettent la création de documents par la délimitation et la structuration de l'information qui y est inscrite. Un dossier peut être composé d'un ou de plusieurs documents. Les documents sur des supports faisant appel aux technologies de l'information visées au paragraphe 2° de l'article 1 sont qualifiés dans la présente loi de documents technologiques.

Le mécanisme utilisé par les collèges à l'égard des documents qui seront mis à la disposition de l'équipe de recherche suppose une anonymisation irréversible [7].

Par conséquent, les documents technologiques qui seront reçus par l'équipe de recherche et sur laquelle porteront les analyses ne sont pas des renseignements personnels. Certes, ils portent sur des faits et gestes relatifs à des personnes, mais ils ne permettent pas d'identifier une personne.

Par conséquent, dès lors que l'équipe de chercheurs ne reçoit pas de renseignements personnels. C'est-à-dire, de renseignements qui "concernent une personne physique et permettent de l'identifier », les interdictions de la Loi régissant les renseignements personnels détenus par les collèges ne reçoivent pas application.

Les collèges vont livrer des fichiers qui ne comporteront pas d'information permettant d'identifier les personnes (les élèves). D'autre part, une fois que l'équipe aura en sa possession les fichiers, il sera impossible de déduire l'identité d'un élève.

Dès lors qu'il est établi que les données qui sont transmises par les collèges ne permettent pas d'identifier une personne en particulier, celles-ci n'ont pas le caractère de renseignement personnel au sens de la loi.

La condition qui doit être remplie pour qu'on puisse considérer un renseignement comme étant un renseignement personnel est qu'il concerne une personne identifiable ou permette d'identifier une personne.
Même en Europe, là où les règles en matière de protection des données personnelles ont une application beaucoup plus étendue qu'au Québec, on reconnaît que l'anonymisation fait perdre le caractère de donnée personnelle à une information ou un document.

Dans un document rédigé par le Groupe « article 29 » réunissant les agences européennes de protection de données personnelles, il est expliqué que :
"Anonymous data" in the sense of the Directive can be defined as any information relating to a natural person where the person cannot be identified, whether by the data controller or by any other person, taking account of all the means likely reasonably to be used either by the controller or by any other person to identify that individual. "Anonymised data" would therefore be anonymous data that previously referred to an identifiable person, but where

that identification is no longer possible. Recital 26 also refers to this concept when it reads that "the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable". Again, the assessment of whether the data allow identification of an individual, and whether the information can be considered as anonymous or not depends on the circumstances, and a case-by-case analysis should be carried out with particular reference to the extent that the means are likely reasonably to be used for identification as described in Recital 26. This is particularly relevant in the case of statistical information, where despite the fact that the information may be presented asaggregated data, the original sample is not sufficiently large and other pieces of information may enable the identification of individuals **[8]**.

Par contre, un fichier ne peut être qualifié d'anonyme lorsqu'il est possible par un moyen ou un autre d'identifier une personne, lorsqu'un moyen de déduction logique permet de reconstituer une identité à partir de plusieurs renseignements anonymes, lorsque le mécanisme d'anonymisation est réversible ou lorsqu'un pseudonyme remplace un identifiant. Lorsque l'anonymat n'est pas garanti, les obligations de protection conférées par la loi aux renseignements personnels demeurent applicables.

Par conséquent, il est impératif que les fichiers qui seront transmis aux chercheurs ne puissent, par aucun moyen raisonnablement disponible, être soumis à des processus qui engendreraient la réidentfication.

Il est connu qu'il existe une possibilité théorique de réidentification de fichiers qui ont été anonymisés. Mais le seul fait que cette hypothèse ne peut être entièrement écartée ne transforme pas des données anonymisées en renseignements personnels. Raisonner de cette façon reviendrait à considérer que toute donnée est une donnée personnelle puisqu'il serait impossible d'exclure la possibilité qu'un jour, quelqu'un invente un procédé qui permettrait

de l'associer à une personne identifiable. Ce n'est pas ainsi que les lois actuelles définissent la notion de renseignement personnel.

Le critère retenu est plutôt celui de l'existence de possibilités raisonnables que les données soient soumises à des procédés qui engendreraient l'identification. Or, il n'y a pas dans la présente situation, d'indication que ce genre d'opérations pourrait être tentée.

À tout évènement, l'équipe aurait avantage à s'engager auprès des collèges à s'abstenir de tenter de réidentifier les données reçues. Alors cela couperait court à toute possibilité que l'on puisse prétendre que ces données sont réidentifiables.

Ainsi, dès lors que les renseignements sont anonymisés au moyen du procédé retenu dans le cadre du projet et que cette anonymisation est effectuée avant les données soient transmises à l'équipe de recherche, celle-ci ne reçoit pas de données personnelles au sens de la *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels.*
Je vous prie d'accepter, cher collègue, l'expression de ma considération la plus distinguée.

Pierre Trudel, avocat, professeur

---

[1] < http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/A_2_1/A2_1.html >

[2] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 59(1)-59(3).

[3] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 59(4).

[4] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 59(5). Dans ce cas, une autorisation de la Commission d'accès à l'information est exigée; Voir : *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 125.

[5] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 64.

[6] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, art. 64, al. 2. En vertu de l'alinéa 3, cette collecte « s'effectue dans le cadre d'une entente écrite transmise à la Commission. L'entente entre en vigueur 40 jours après sa réception par la Commission ».

[7] Voir : < http://www.fastsum.com/support/md5-checksum-utility-faq/md5-hash.php>

[8] ARTICLE 29 DATA PROTECTION WORKING PARTY, *Opinion 4/2007 on the concept of personal data*, june 2002,< http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf>

CHAPITRE 3

DEUX MODÈLES D'ATTRITION

Les méthodes de forage de données et d'apprentissage machine statistique peuvent être grossièrement subdivisées en modèles *prédictifs* et en modèles *explicatifs* (Shmueli, 2010). Ces deux paradigmes se manifestent aussi bien dans le domaine du forage de données en *éducation* (Liu et Koedinger, 2017).

Recherches antérieures sur l'attrition dans l'enseignement supérieur

Certains des premiers travaux théoriques sur la persévérance scolaire dans l'enseignement supérieur (Tinto, 1975) suggèrent que le décrochage scolaire dans l'enseignement supérieur est le fruit d'un processus social qui découle d'une mauvaise intégration des élèves au sein de leur établissement. Conséquemment, la probabilité qu'un élève demeure à l'école est non-seulement déterminée par le degré de correspondance entre la motivation de l'élève et son aptitude aux études, mais aussi par la culture sociale et scolaire de l'établissement (Cabrera, Nora et Castaneda, 1993). Puisque l'attrition dépend de la culture sociale et scolaire de l'établissement, ce « ***modèle d'intégration*** des élèves » va jusqu'à suggérer que l'attrition peut servir de « baromètre de la santé sociale et intellectuelle de la vie collégiale » (Tinto, 1987).

Un autre modèle théorique concurrent, qu'on appelle généralement le « ***modèle d'attrition*** des élèves », s'inspire de travaux sur le « roulement du personnel » dans l'industrie et les entreprises. Ce modèle propose que l'attrition, que ce soit en entreprise ou à l'école, est principalement affectée par la satisfaction des participants (étudiants ou employés). L'intention d'abandonner les études (ou leur emploi) dépend alors de la perceptions de la valeur de leurs propres travaux (Bean, 1980; Bean, 1983). Bien que le « modèle *d'intégration* des élèves » accorde plus d'importance aux variables *propres à l'établissement*, le « modèle *d'attrition* des élèves » attribue ce qui est observé au collège comme étant une *conséquence*, façonnée par des facteurs externes au collège, comme le soutien social et parental ou la présence d'autres opportunités. Dans les travaux de Tinto, les notes obtenues au collège sont considérées comme étant le résultat de l'intégration d'un élève au sein de l'établissement. Par contre, dans le modèle de Bean (1985), ces mêmes notes sont la *cause* du « syndrome de décrochage ».

Les modèles longitudinaux d'attrition des élèves sont importants si nous voulons tenir compte de la nature évolutive de la décision ultime d'un élève d'abandonner l'école (Tinto, 1975) et la façon dont le moment des interventions peut avoir un impact sur le résultat (Ishitani et Desjardins, 2002; Simons, 2011). Les recherches sur l'attrition dans les établissements d'enseignement supérieur (Singell et Waddell, 2010) se sont servies de modèles pour examiner la rétention dans une grande université publique, leur objectif étant d'identifier les élèves à risque de décrochage assez tôt pour être en mesure d'intervenir et de réduire le nombre de décrocheurs.

Bien qu'un modèle théorique propose que le décrochage est due principalement à des facteurs extérieurs tandis que l'autre propose plutôt que le décrochage est due principalement des facteurs propres à l'étudiant, une troisième approche est possible, soit celle du forage de données. Cette approche ne présuppose pas de modèle théorique à priori. En effet, le forage de donnée, part des données colligées et ne privilégie pas certains facteurs à priori qui pourrait causer le décrochage scolaire. L'approche de forage des données peut se voir comme agnostique quant aux facteurs internes ou externes aux étudiants. Cela dit, l'approche

privilégie les facteurs associés à des données mesurables et donc à un biais partiel envers les certains facteurs culturels qui sont plus difficiles à mesurer. Le forage de donnée préconise simplement qu'un échantillonnage important de données soient colligées et préconise la production de modèles statistiques prédictifs provenant directement des données. C'est précisément cette approche que nous préconisons dans les chapitres suivants.

RÉFÉRENCES

Bean, John P. 1980. "Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition." Research in Higher Education 12 (2). Springer : 155-87.

Bean, John P. 1983. "The Application of a Model of Turnover in Work Organizations to the Student Attrition Process." The Review of Higher Education 6 (2). The Johns Hopkins University Press : 129-48.

Bean, John P. 1985. "Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome." American Educational Research Journal 22 (1). Sage Publications : 35-64.

Cabrera, Alberto F., Amaury Nora et Maria B. Castaneda. 1993. "College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention." The Journal of Higher Education 64 (2). Taylor et Francis : 123-39.

Ishitani, Terry T. et Stephen L. Desjardins. 2002. "A Longitudinal Investigation of Dropout from College in the United States." Journal of College Student Retention: Research, Theory & Practice 4 (2). SAGE Publications Sage, CA : Los Angeles, CA : 173-201.

Liu, Ran et Kenneth R. Koedinger. 2017. "Going Beyond Better Data Prediction to Create Explanatory Models of Educational Data." In The Handbook of Learning Analytics, edited by Charles Lang, George Siemens, Alyssa Friend Wise, and Dragan Gaševic, 1re éd., 69–76. Alberta, Canada : Society for Learning Analytics Research (SoLAR).

Shmueli, Galit. 2010. "To Explain or to Predict?" Statistical Science 25 (3). Institute of Mathematical Statistics : 289-310.

Simons, Jill M. 2011. "A National Study of Student Early Alert Models at Four-Year Institutions of Higher Education." Thèse de doctorat, UNIVERSITÉ DE L'ÉTAT D'ARKANSAS.

Singell, Larry D. et Glen R. Waddell. 2010. “Modeling Retention at a Large Public University: Can at-Risk Students Be Identified Early Enough to Treat?” Research in Higher Education 51 (6). Springer : 546-72.

Tinto, Vincent. 1975. “Dropout from Higher Education: A Theoretical Synthesis of Recent Research.” Review of Educational Research 45 (1). SAGE Publications Sage, CA : Thousand Oaks, CA : 89-125.

Tinto, Vincent. 1987. Leaving College: Rethinking the Causes and Cures of Student Attrition. ERIC.

CHAPITRE 4:

ÉCHANTILLONNAGE ET DONNÉES DESCRIPTIVES

Pour mieux comprendre la population étudiée, certaines données démographiques sont nécessaires. Ces données démographiques nous permettent de quantifier les caractéristiques de la population étudiante des trois collèges participant à cette étude : le Collège John Abbott, le Collège Dawson et le Collège Vanier. Parmi les points communs évidents à ces trois cégeps, notons que ce sont tous des cégeps anglophones de la région de Montréal qui ensembles représentent la majorité de la population étudiante anglophone du Québec.

Notons que dans la première partie de ce rapport nous focalisons notre attention sur la cohorte d’élèves ayant commencé leurs études collégiales à l’automne 2010 sous l’égide de la « réforme » en éducation, ou plus formellement, le Programme de formation de l'école québécoise. Nous restreignons notre analyse à cet échantillon qui représente les étudiants actuels et nous permet de comparer des étudiants ayant suivi une formation similaire au secondaire et au collégial. Nous commençons par examiner les différences entre les sexes pour

les trois cégeps globalement, et ce, pour toutes les cohortes d'élèves qui auraient commencé le cégep à partir de l'automne 2010 jusqu'à la session d'hiver 2017. Sur notre échantillon d'élèves entrant au collégial et ayant un registre démographiques complet (N= 65 865) nous observons que 55.2% de l'échantillon est composé de filles alors que 44.8% est composé de garçons. Il est aussi possible d'illustrer la distribution de garçon et filles par langue maternelle, ce qui donne le portrait suivant :

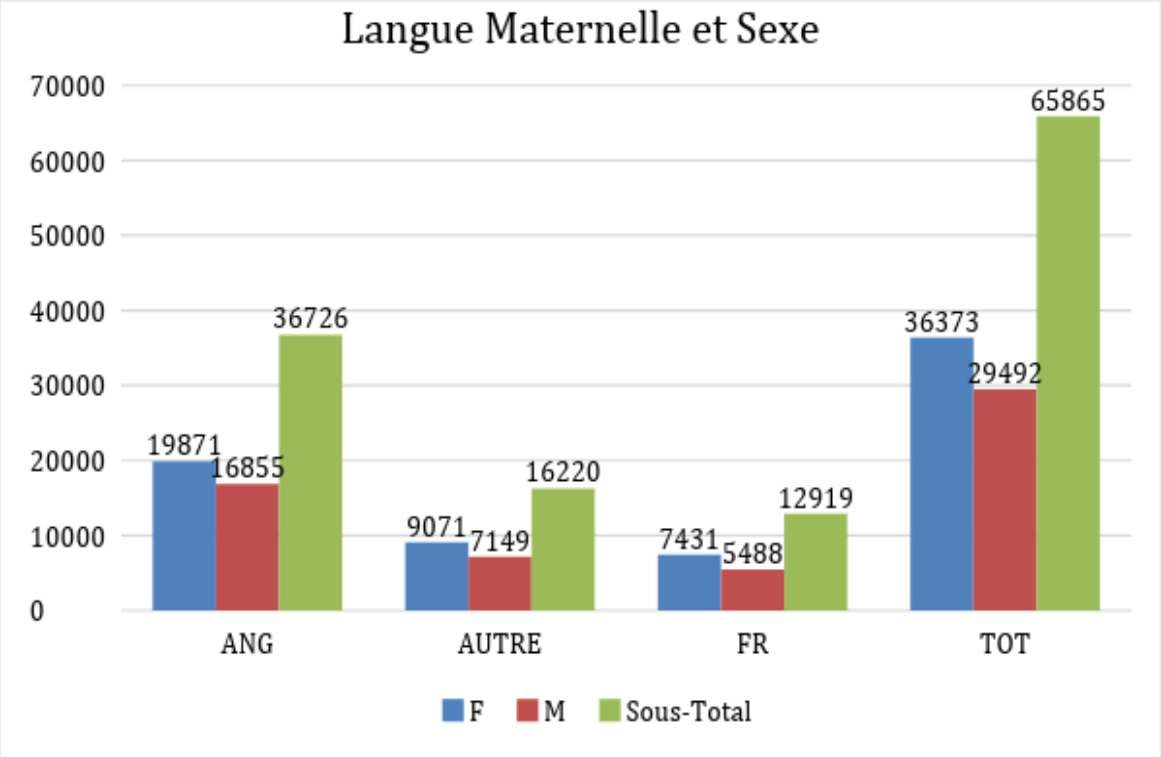


Figure 4.1 Distribution d'étudiants enregistrés par langue maternelle

Ce portrait rapide de notre échantillon révèle une population majoritairement anglophone (55.8%), bien que près d'un cinquième (19.6%) est francophone et près d'un quart est allophones (24.6%). De plus, on observe que dans chaque sous groupe linguistique, les filles sont plus nombreuses que les garçons.

Nous cherchons ensuite à rentrer dans le vif du sujet et obtenir une impression quantitative de la probabilité qu'un étudiant inscrit ne gradue pas. Pour ce faire nous examinons notre échantillon de plus de 65 000 élèves inscrit aux collégial depuis la réforme. Nous considérons comme ayant décroché tous les élèves inscrits à un moment et n'étant pas actuellement

enregistrés ou n'ayant pas gradués. En partant de notre échantillon initial complet (N=65 561), nous observons que près d'un tier est encore actif (N=18 966) et le reste peut être réparti entre ceux qui ont gradué et ceux qui ont décroché puisqu'ils ne sont pas actifs et n'ont pas gradué. Il y a évidemment des limites claires à cette définition. Tout d'abord, nous ne sommes pas en mesure de suivre les étudiants qui quittent un collège pour aller dans un autre ou quittent la province pour compléter leurs études. Il est aussi possible que certains étudiants soit catégorisés comme inactifs (décrochage) mais que leur date de graduation soit en dehors de l'interval de notre étude. Les étudiants que reviennent et gradueront dans le futur sont actuellement comptés comme des décrocheurs. La figure 4.2 ci-dessous donne une image de la proportion d'étudiants ayant été inscrit, mais qui sont soit inactifs (décroche) ou qui ont gradués, et ce par le sexe masculin ou féminin.

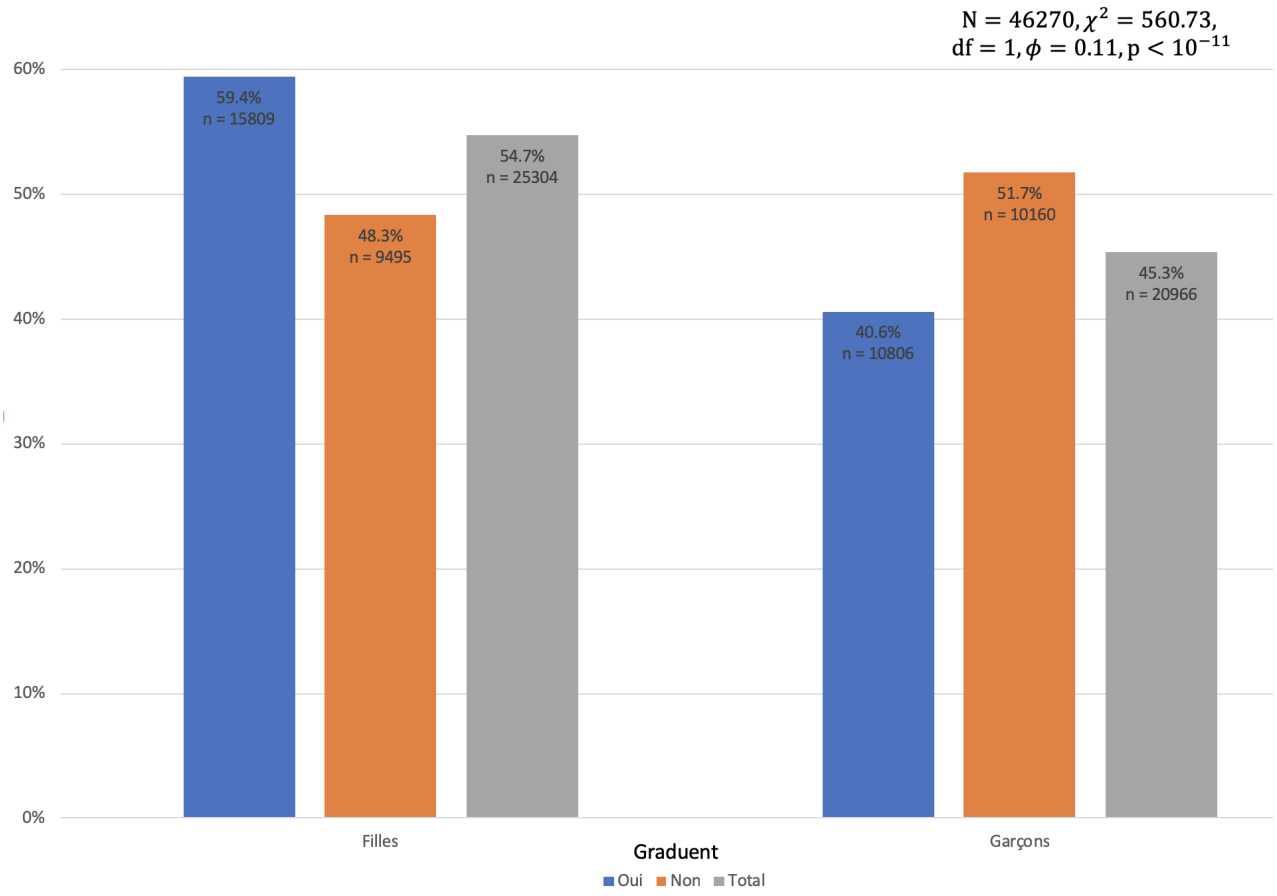


Figure 4.2 Taux de graduation par sexe

Notons que dans ce graphique, les barres d'une même couleur ont une somme de 100%. Tout d'abord, ayant exclu les étudiants actifs, on obtient un échantillon ayant approximativement 55% de filles et 45% de garçons. Cependant, la répartition de gradués comprend proportionnellement plus de filles ayant gradués (59.4% filles vs 40.6% garçons). En revanche, la répartition d'étudiants n'ayant pas gradués et n'étant pas inscrits (notre définition de décrochage) comprend proportionnellement plus de garçons (51.7% garçons vs 48.3% filles). Bien que non indiqué dans le graph, il est intéressant de noter que le pourcentage de filles qui graduent par rapport à l'échantillon total de filles est de 62.5% (15 809/25 304). En revanche, seulement 51.5% des garçons (10806/20966) dans notre échantillon complètent leurs études. Il est donc possible de calculer un Rapport de Cote ("Odds Ratio") pour déterminer la différence relative de probabilité de graduation entre les filles et les garçons. Pour ce faire, nous comparons la cote de graduation des filles (15809/9495) et la comparons à la cote de graduation des garçons (10806/10160). Nous obtenons un rapport de cote OR=1.565 (intervalle de confiance 95%, IC95%=1.51-1.62, $p<0.0001$) indiquant que d'être une fille confère 56.5% de plus de chances de graduer pour une fille que pour un garçon.

Il est aussi possible d'examiner l'association de la langue maternelle sur le taux de graduation. Dans la figure 4.3 ci-dessous, les barres d'une même couleur ont une somme de 100%. Tout d'abord, ayant exclu les étudiants actifs, on obtient un échantillon ayant approximativement 55.6% d'anglophones, 20.2% de francophone et 24.2% d'allophones.

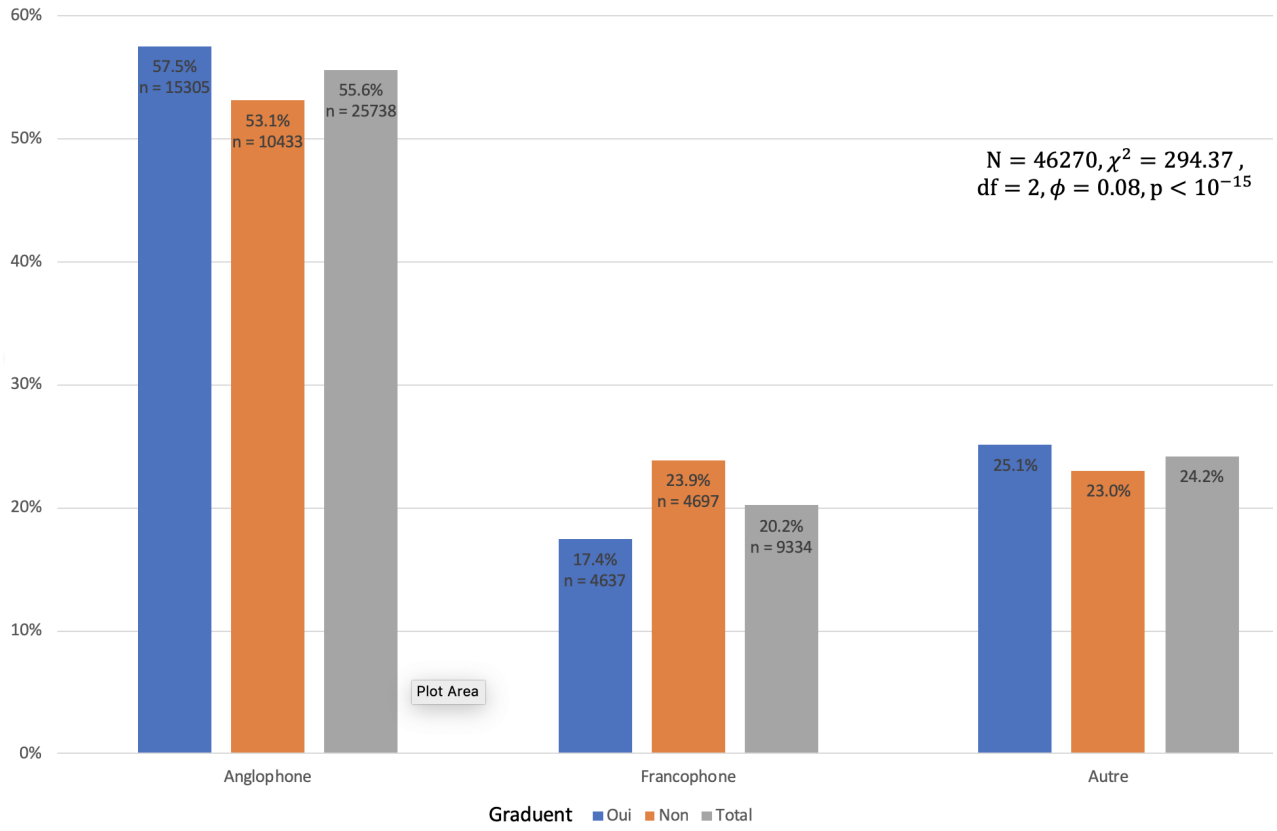


Figure 4.3 Taux de graduation par langue maternelle

La répartition de gradués comprend une proportion légèrement supérieure d'anglophones (57.5%) et quelque peu inférieur de francophones (17.4%) ainsi qu'un poids démographique à peu près équivalent à 1% près d'allophones. Parmi ceux qui ne sont pas actifs et qui n'ont pas gradués, on trouve les francophones sur-représentés par rapport à leur poids démographique tandis que les anglophones et allophones sont légèrement sous-représentés. Il est également intéressant de noter que le pourcentage d'anglophones qui graduent par rapport à l'échantillon total d'anglophones est de 59.5% (15 305/25 738). En revanche, seulement 49.7% des francophones (4637/9334) dans notre échantillon complètent leurs études. Il est donc possible de calculer un Rapport de Cote ("Odds Ratio") pour déterminer la différence relative de probabilité de graduation entre les anglophones et les francophones inscrits dans ces trois collèges anglophones. Pour ce faire, nous comparons la cote de graduation des anglophones (15305/10433) et la comparons à la cote de graduation des garçons (4637/4697). Nous obtenons un rapport de cote OR=1.486 (IC95%=1.42-1.56,

p<0.0001) indiquant que d'être anglophone confère 48.6% plus de chances de graduer pour les anglophones que pour les francophones inscrits dans ces trois collèges anglophones.

Une dernière analyse intéressante examine le lieu de naissance des étudiants et son association avec les taux de graduation ou de décrochage. Dans la figure 4.4 ci-dessous, environ un quart (23.9%) des étudiants inscrits au collégial ne sont pas nés au Québec. Ce graphique montre aussi que ces étudiants sont légèrement sous-représentés parmi les gradués et sur-représentés parmi les décrocheurs. Parmi les étudiants nés au Québec, 59.0% (20 765/35 223) graduent. En revanche, seulement 53.0% des étudiants nés hors Québec (5850/11047) complètent leurs études. Nous calculons encore le Rapport de Cote pour déterminer la différence relative de probabilité de graduation entre les étudiants nés au Québec et ceux nés hors Québec et obtenons un rapport de cote OR=1.276 (IC95%=1.222-1.332, p<0.0001) indiquant que d'être nés au Québec confère 27.6% plus de chances de graduer.

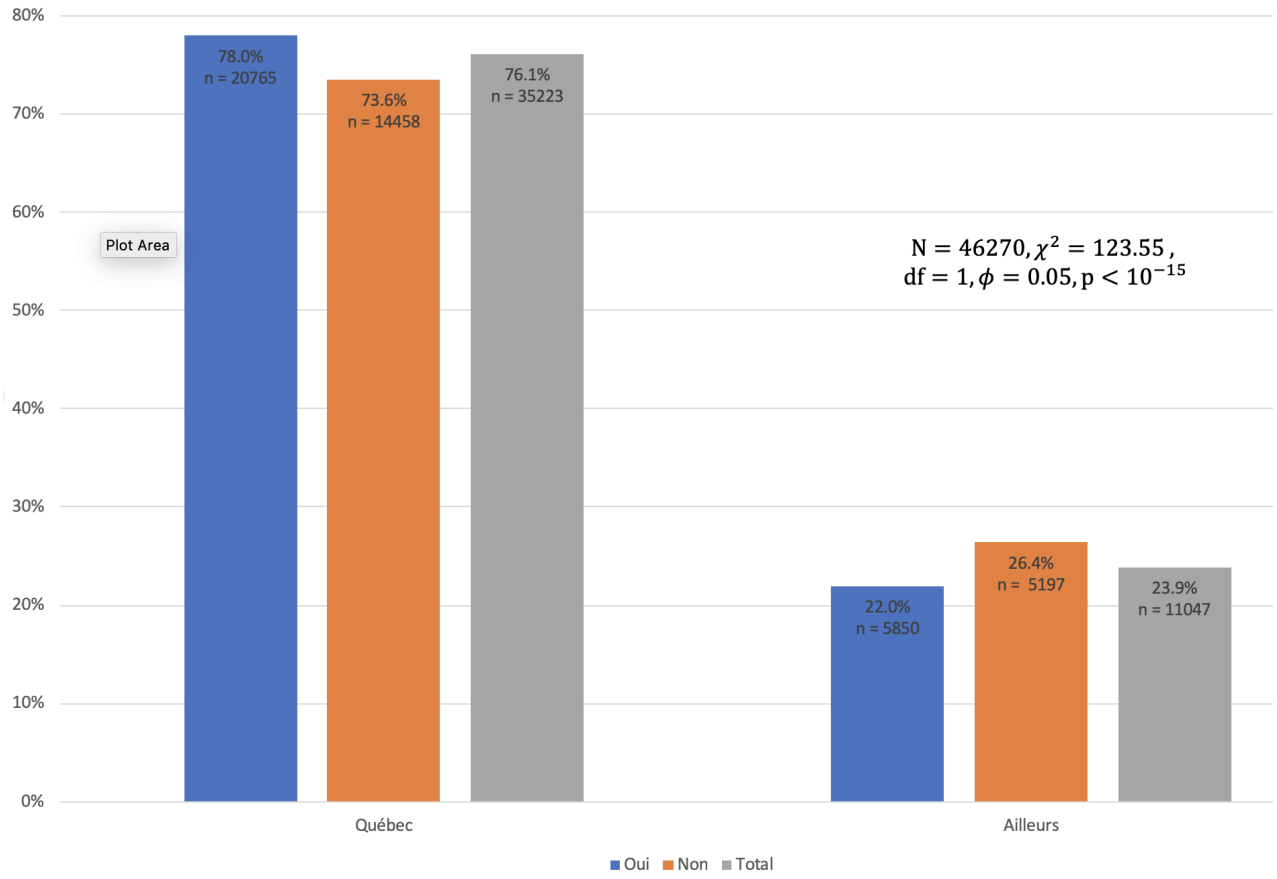


Figure 4.4 Taux de graduation par lieu de Naissance

Ayant examiné ces facteurs associés, nous posons une question simple. Peut-on examiner la probabilité qu'un étudiant inscrit lors d'une session académique, ne soit plus inscrit la session suivante?

Il ne s'agit pas de décrochage nécessairement mais plutôt d'une discontinuité dans le cursus académique d'un étudiant. Nous examinons ceci dans un sous échantillon d'étudiants inscrits entre l'automne 2010 et l'hiver 2017 et obtenons les figures 4.5 à 4.8 ci-dessous. Nous commençons par examiner la distribution d'étudiant inscrits et déterminons la fraction d'entre eux qui ne sont pas enregistré lors de la session académique suivante, sans pour autant avoir gradué. La figure 4.5 donne un aperçu de la fraction de fille et de garçons absents lors de la session académique suivante.

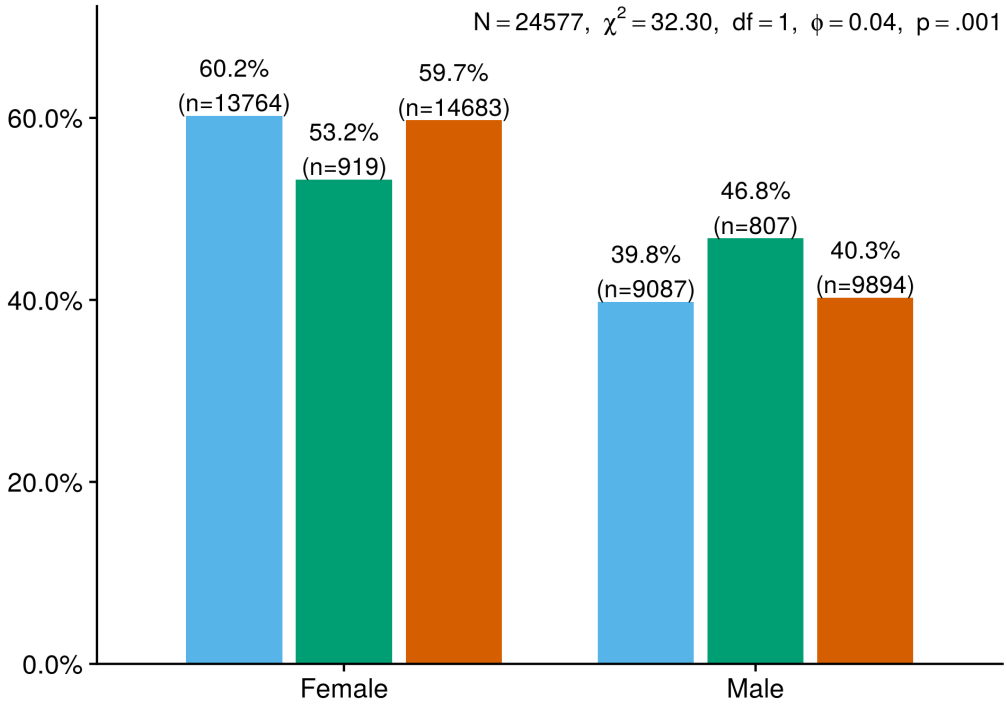


Figure 4.5 Absent la session suivante, par sexe

Dans ce graphique, nous comparons le taux d'étudiants enregistrés qui sont absent une session académique plus tard, sans avoir complété leurs études. Notons encore que dans ce graphique, les colonnes d'une même couleur ont une somme de 100%. Trois de ces comparaisons sont faites. Les barres oranges montrent le total de filles et de garçons analysés dans ce sous-échantillon (59.7% de filles vs 40.3% de garçons). Les barres vertes montrent la quantité d'étudiants, parmi tous ceux examinés, qui s'absente une session plus tard du programme dans lequel ils sont inscrits, sans pour autant l'avoir complété. Les barres bleues montrent la grande majorité d'étudiants qui sont enregistrés lors de deux sessions académiques consécutives. Ce qui est intéressant à propos de ce graphique, c'est que nous observons une perte attendue dans une session donnée de 7.0 % des étudiants. En effet, ce graphique montre une perte de 919 filles et de 807 garçons (pour une attrition totale de 1726 élèves), soit 7% (1726/24577) de perte d'une session à l'autre. Bien que le pourcentage de garçons qui décrochent est plus élevé (8.2% = 807/9894) que le pourcentage de filles qui décrochent (6.3% = 919/14683), en valeur absolue, nous observons que plus de filles décrochent à la première session (919) que de garçons (807). Lorsque nous examinons des données similaires, mais en fonction de la langue maternelle, nous notons ce qui suit :

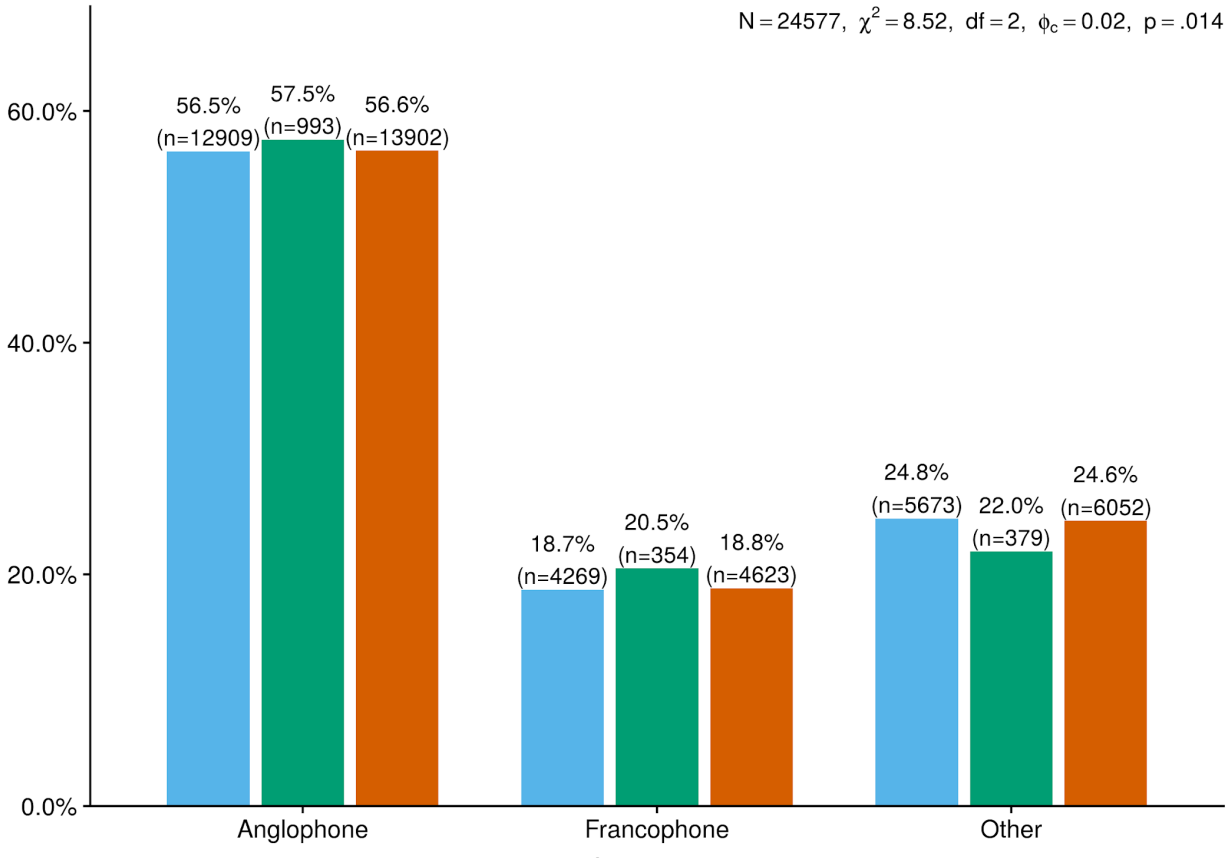


Figure 4.6 Absent la session suivante, par langue maternelle

Ce graphique montre clairement que, dans ces collèges anglophones, la langue maternelle de la majeure partie de la population étudiante est l'anglais. Le graphique précédent montre que la langue maternelle n'est pas à elle seule une variable qui permet de discriminer entre les élèves qui s'absentent la session académique suivante de ceux qui persévèrent. En effet, nous observons des différences bien plus petites en ce qui concerne la langue maternelle que ce que nous avons observé pour le sexe. Il existe certaines différences légères : les anglophones sont environ 1 % plus à risque de décrochage que leur poids démographique; les francophones sont environ 2 % plus à risque de décrochage que leur poids démographique le suggère et, enfin, les allophones sont un peu plus de 2 % moins à risque de décrochage que leur poids démographique ne le suggère. Finalement, nous comparons les taux de décrochage à la première session en fonction du lieu de naissance : Québec vs ailleurs.

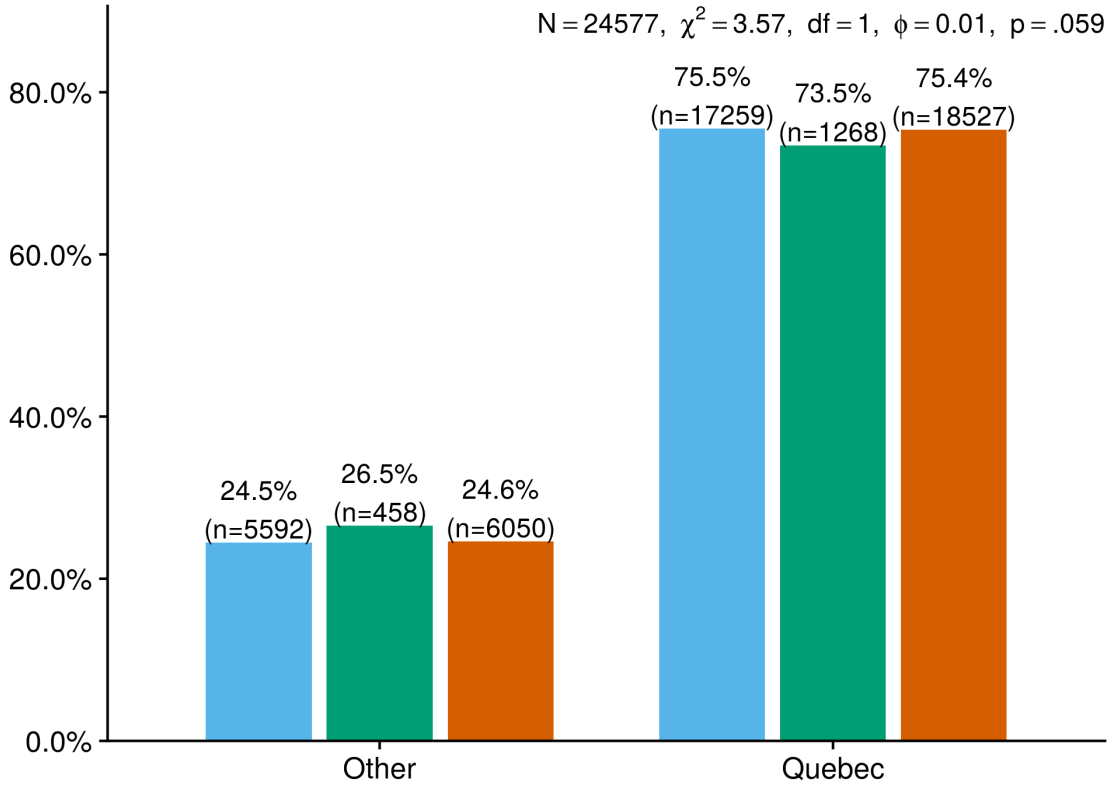


Figure 4.7 Absent la session suivante, par lieu de naissance

Ce graphique montre clairement, qu'environ trois-quarts des élèves sont nés au Québec. Les élèves qui sont nés au Québec sont environ 2 % moins à risque de quitter une session plus tard que leur poids démographique ne le suggère. En revanche, les élèves nés ailleurs sont environ 2 % plus à risque de quitter que ne le suggère leur poids démographique.

Ce portrait de la situation, comme illustré dans les graphiques précédents utilisent des éléments d'information que les cégeps ont à leur disposition dès le premier jour où les élèves commencent leurs études collégiales. Les résultats pertinents de cette analyse préliminaire nous permettent de confirmer ce que d'autres études ont démontré comme le fait que les garçons sont plus à risque de décrochage. Cependant, notre analyse permet aussi de quantifier le risque et permet de montrer de façon statistiquement significative que le filles ont 56% plus de chances de graduer que les garçons, les anglophones ont 49% plus de chances que les francophones de graduer et que les étudiants nés au Québec ont 28% plus de chances de graduer. Nos données suggèrent donc qu'il faut tenir compte de trois groupes statistiquement significativement ($p<0.001$) plus à risque de décrocher, soit les garçons, les francophones et

les étudiants nés hors Québec. Dans un chapitre subséquent de ce rapport, des méthodes de prédictions d'étudiants à risque de décrocher seront présentées.

Si l'on regroupe ces trois catégories de données (sexe, langue maternelle et lieu de naissance), il est possible d'obtenir le portrait typique du décrocheur : un garçon francophone qui n'est pas né au Québec. Bien que cette information soit utile à connaître le premier jour de classe, elle est stable et non malléable. En effet, ce profil type ne change pas au fil du temps. Nos efforts visent cependant aussi à identifier les variables malléables qui pourraient nous permettre de minimiser les taux de décrochage et maximiser les taux de graduation de tous les élèves inscrits. Avec ce projet de recherche, nous avons pu mettre le microscope le plus puissant à ce jour sur le phénomène des décrocheurs au collégial, dans le prochain chapitre, nous utiliserons parmi les données les plus granulaires qui existent afin d'outrepasser les phénomènes déjà connus dans le décrochage.

CHAPITRE 5

Rendement scolaire et décrochage

Plusieurs hypothèses ont été avancées, pour expliquer les raisons pour lesquelles certains étudiants décrochent, mais nous savons actuellement peu de choses qui privilégie concrètement une hypothèse par rapport à une autre. Les hypothèses classiques qui ont été avancées incluent des facteurs propres à l'étudiant comme une mauvaise préparation à l'école secondaire menant à de mauvais résultats scolaires au cégep ou une faible motivation. D'autres hypothèse privilégient des facteurs sociaux comme le manque de soutien à l'école ou de support dans le réseau social des étudiants qui pourraient les aider à rester à l'école. Il est important de noter qu'avec les données que nous avons colligés, nous pouvons nous pencher sur certaines de ces hypothèses qui varient dans le temps. Plus précisément, le manque de ressources financières et l'absence de mécanismes de soutien sont des causes qui demeurent à peu près constantes pendant toute la durée du séjour des élèves. D'un autre côté, les problèmes de rendement et de motivation peuvent fluctuer d'une session à l'autre. Bien qu'avec les sources de données à notre disposition, il ne nous est impossible de déterminer toutes les causes qui pourraient mener au décrochage scolaire, il est néanmoins possible pour nous d'étudier l'effet du rendement scolaire de ces élèves sur la poursuite ou l'interruption de leurs études. Cet ensemble de données permet de savoir si l'un de ces deux paramètres, soit le rendement scolaire, fluctue suffisamment dans le temps pour diriger les élèves de la voie de l'obtention du diplôme vers la voie du décrochage scolaire. Cette section aborde beaucoup plus en détail le rendement scolaire mesuré à l'aide de trois paramètres : la moyenne générale dans les sessions qui mènent à un décrochage, la note la plus basse dans chaque session précédant un décrochage ou la diplomation et, finalement, les résultats aux évaluations de la mi-session qui précède un décrochage.

Il est intéressant d'examiner ces paramètres puisqu'ils permettent de poser des questions qui pourraient modifier la représentation mentale que nous avons des élèves qui décrochent. Que

faire si nous constations qu'une importante proportion des élèves qui décrochent ont de bonnes notes, n'ont aucun échec dans leurs cours et n'ont aucune caractéristique apparente qui les distingue des élèves qui obtiennent un diplôme?

Cela voudrait dire que, soit l'école échoue à garder ces élèves motivés, ce qui signifie qu'ils ne voient plus le but de poursuivre leurs études, soit qu'ils n'ont tout simplement pas décroché, c'est-à-dire qu'ils ont seulement changé de système en délaissant leur institution ou le système québécois. Ils pourraient très bien être diplômés d'autres établissements. En raison de la difficulté à suivre les élèves en dehors d'une institution, il est impossible de déterminer si ces étudiants sont de véritables décrocheurs ou seulement des étudiants qui ont changé de système ou de lieu géographique.

Les résultats scolaires

La première question que nous posons est: quelle est la moyenne générale des élèves qui décrochent, en général et dans les sessions qui mènent à leur décrochage?

1. Les résultats scolaires des élèves qui décrochent sont-ils très différents de ceux de leurs camarades qui poursuivent leurs études?
2. Les élèves qui décrochent le font-ils en raison des mauvaises notes?
3. Quelle proportion des élèves est considérée, année après année, comme ayant décroché et étiquetée comme étant des problèmes à résoudre par le système, alors qu'en fait ce sont des élèves exemplaires en ce qui concerne leur rendement scolaire?

Grâce à notre ensemble de données, nous pouvons obtenir la réponse à ces questions.

Les moyennes générales

Commençons par examiner la moyenne générale des élèves qui quittent l'établissement collégial par rapport à celle des élèves qui graduent. Ces résultats scolaires représentent les moyennes générales pour l'ensemble des cours que ces étudiants ont pris dans un établissement collégial.

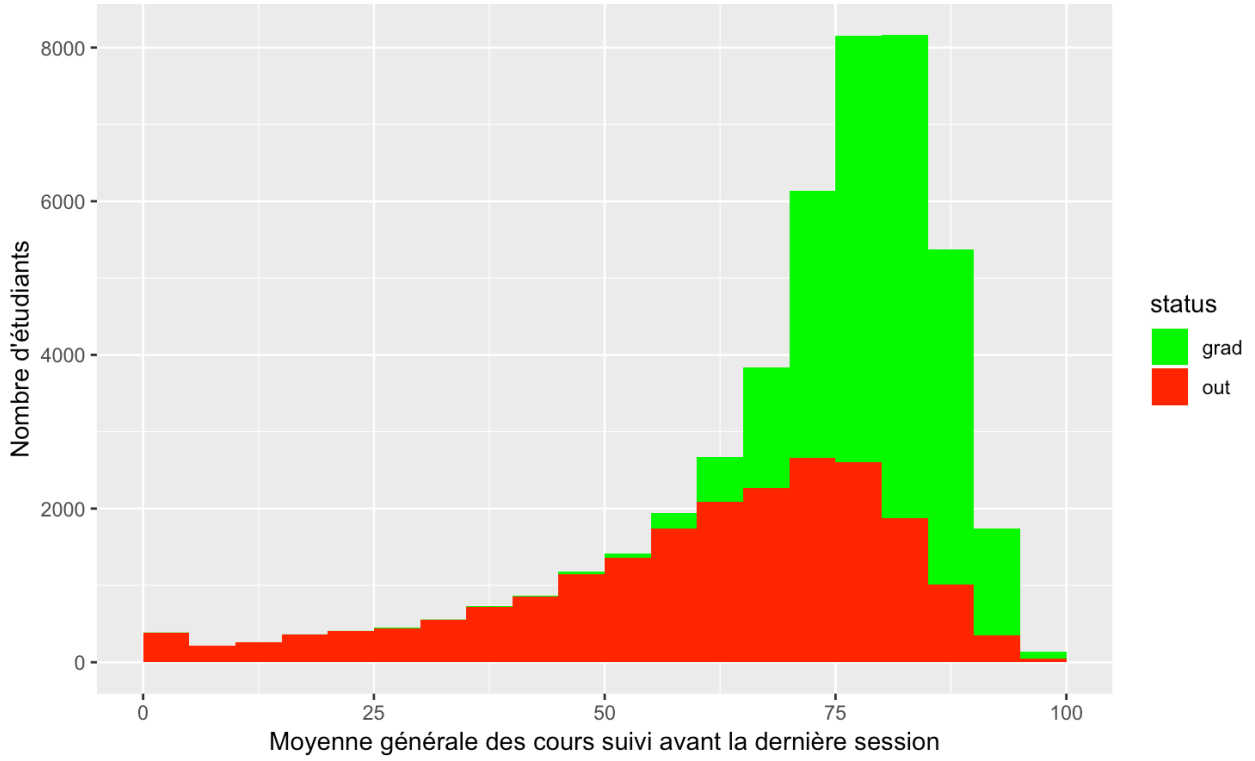


Figure 5.1 Moyenne générales chez étudiants qui quittent (rouge) versus qui graduent (vert)

La Figure 5.1 ci-dessous montre que la plupart des élèves (61%) qui quittent ont néanmoins des moyennes générales au-dessus de 60%. En effet, la moyenne générale 'moyenne' de cet échantillon d'étudiants qui quittent l'établissement sans avoir gradué (N=21 320) est de est de 60.9% tandis que la médiane des moyennes générales est de 65,4%. En revanche, la moyenne générale parmi les gradués (N= 23 653) est de 79.5% avec une médiane de 80.3%. De façon générale, nous observons que la moyenne générale des deux groupes est sensiblement différente, soit 80 % versus 61 % respectivement pour les diplômés et les décrocheurs. Cependant, il est important de noter que la distribution de notes en rouge est très dispersée. En effet, la variance des notes de ceux qui décrochent est donc bien plus large que la distribution des notes de ceux qui persévèrent. En effet, la plupart (61%) des décrocheurs potentiels obtiennent une moyenne générale au-dessus de 60% avec 27.6% des décrocheurs potentiels qui obtiennent une moyenne générale supérieure à 75 %.

Nous pouvons aussi prendre la moyenne générale de 60% comme seuil pour évaluer un Rapport de Cote ("Odds Ratio") nous permettant de déterminer la différence relative des chances de graduation entre ceux qui ont une moyenne générale au dessus de 60% et ceux qui ont une moyenne générale en dessous de 60%. Parmi les étudiants ayant une moyenne générale au dessus de 60%, 23 321 obtiennent un diplôme tandis que 12 985 quittent. En revanche, parmi ceux qui obtiennent une moyenne générale en deçà de 60%, seulement 332 graduent tandis que 8335 quittent l'établissement. Nous obtenons ainsi un rapport de cote OR=45.1 (IC95%=40.3-50.4, $p<0.0001$) indiquant que les étudiants ayant une moyenne générale en dessous de 60% ont 45 fois plus de chances de quitter que les étudiants ayant une moyenne générale au dessus de 60%.

Ces données suggèrent deux résultats qui pourraient à priori sembler contradictoires. Tout d'abord, les élèves ayant des moyennes générales en dessous de 60% sont très largement et statistiquement significativement plus à risque de décrocher. Cependant, l'inverse n'est pas vrai. Avoir une moyenne générale au dessus de 60% ou même de 75% n'assure pas qu'un étudiant graduera. En effet, une fraction importante des élèves qui décrochent pourrait être considérés comme ayant de 'bons' résultats scolaires. Comment se peut-il que près de la moitié des élèves qui décrochent a une moyenne au-dessus de 60% dont une fraction a même une moyenne générale au dessus de 75%? Une hypothèse pour expliquer cette situation est que les élèves qui finissent par décrocher commencent par avoir de bonnes sessions et voient leur rendement diminuer à mesure qu'ils approchent de leur dernière session. Il est aussi possible que les élèves obtiennent de bonnes notes dans certains cours mais en échouent d'autres, de sorte à ce que leur moyenne générale est tout de même au-dessus de 60% mais qu'un nombre important de cours ont été échoués. Nous vérifions ces deux hypothèses en analysant les moyennes générales des élèves session par session, examinant la note la plus basse obtenue par les élèves chaque session.

Les moyennes au cours de leur dernière session

Examinons donc le rendement des décrocheurs comparativement aux diplômés, session par session. Commençons par examiner la moyenne générale des élèves lors de la dernière session

au cours de laquelle ils sont inscrits dans un collège. Ce faisant, nous cherchons à comparer la dernière session d'élèves qui décrochent à celle d'élèves dans leur dernière session précédent l'obtention d'un diplôme.

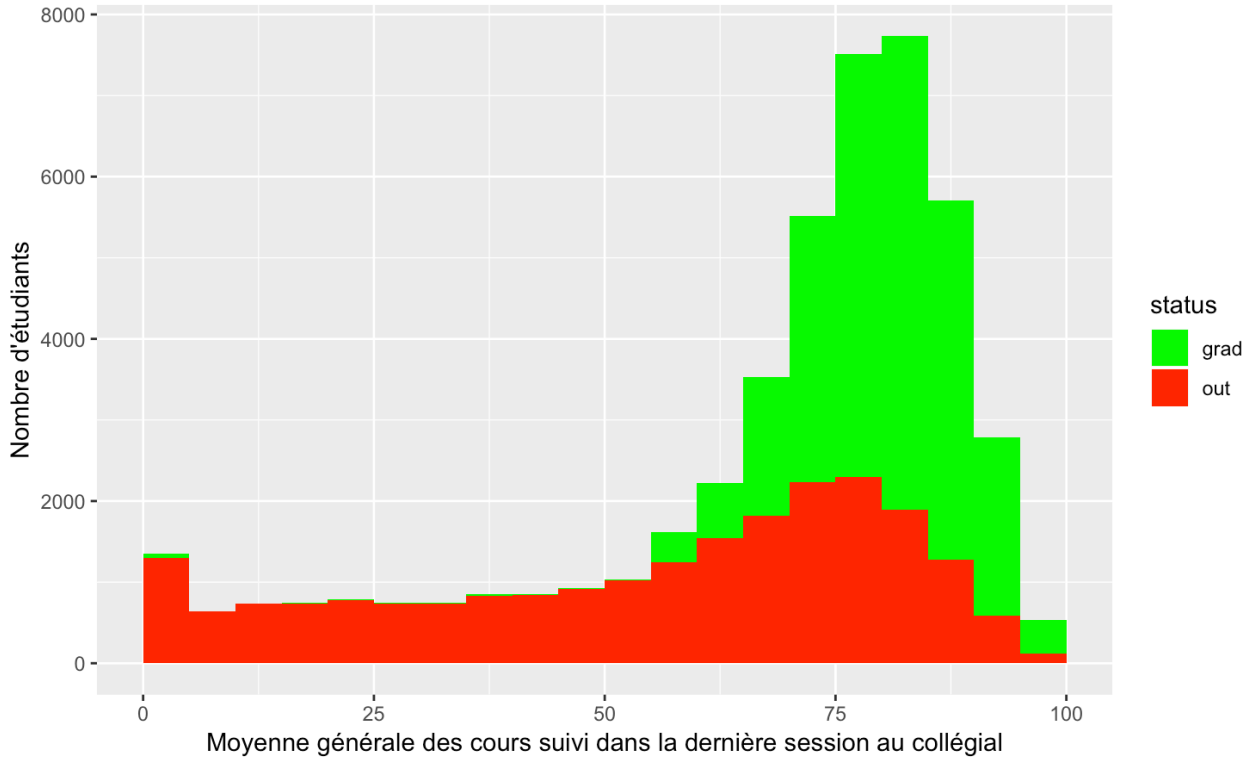


Figure 5.2 Moyenne générales lors de la dernière session pour étudiants qui quittent (rouge) et qui graduent (vert)

En comparant la moyenne générale des décrocheurs (N= 22 296) et des diplômés (N=24265), nous observons que les distributions sont significativement différentes (moyenne de 80.07±0.06 % vs 55.0±0.2 %). Ce qui est surprenant, c'est que plus de la moitié (54.2 %) des décrocheurs ont une moyenne générale supérieure à 60 % et que 28.7 % (N=6408) d'entre eux ont une moyenne générale supérieure à 75 %. Bien que plus de trois-quart (76.3%) des gradués ont des moyennes générales au dessus de 75%, on s'explique mal pourquoi plus du quart des étudiants qui quittent ont des moyennes au dessus de 75%. **En d'autres termes, la moitié des élèves qui quittent obtiennent une moyenne générale de passage dans leur dernière session.** Cela nous oblige certainement à reconsidérer la représentation classique

qui dépeint le décrocheur comme étant nécessairement un élève qui obtient de mauvais résultats scolaires.

Nous pouvons aussi prendre la moyenne de 60% comme seuil pour évaluer un Rapport de Cote ("Odds Ratio") nous permettant de déterminer la différence relative des chances de graduation entre ceux qui ont une moyenne générale au dessus de 60% et ceux qui ont une moyenne générale en dessous de 60%. Parmi les étudiants ayant une moyenne générale au dessus de 60%, 24 109 obtiennent un diplôme tandis que 12 083 quittent. En revanche, parmi ceux qui obtiennent une moyenne générale en deçà de 60%, un total de 156 graduent tandis que 10 213 quittent l'établissement. Nous obtenons un rapport de cote OR=131 (IC95%=111-153, $p<0.0001$) indiquant que les étudiants ayant une moyenne générale en dessous de 60% lors de leur dernière session au collégial ont 111 fois plus de chances de quitter que les étudiants ayant une moyenne générale au dessus de 60%.

Ces données sont à la fois claires et intrigantes. Tel qu'on pourrait s'y attendre, avoir une moyenne générale au dessous de 60% indique que l'étudiant a du mal à réussir et pourrait échouer une partie importante de ses cours. Donc, nous ne sommes pas extrêmement surpris de constater que les étudiants ayant une moyenne générale en dessous de 60% ont plus de 100 fois plus de chances de décrocher. Par contre, notre échantillon montre que plus de 12000 étudiants décrochent tout en ayant des résultats acceptables (>60%) y compris plus de 6400 d'entre eux ayant des moyennes générales au dessus de 75%. Que se passe-t-il?

Prenons la moyenne de 75% comme seuil pour évaluer un Rapport de Cote ("Odds Ratio") nous permettant de déterminer la différence relative des chances de graduation entre ceux qui ont une moyenne générale au dessus de 75% et ceux qui ont une moyenne générale en dessous de 75%. Parmi les étudiants ayant une moyenne générale au dessus de 75%, 18 508 obtiennent un diplôme tandis que 6408 quittent. En revanche, parmi ceux qui obtiennent une moyenne générale en deçà de 75%, un total de 5757 graduent tandis que 15 888 quittent l'établissement. Nous obtenons un rapport de cote OR=8.0 (IC95%=7.6-8.3, $p<0.0001$) indiquant que les étudiants ayant une moyenne générale en dessous de 75% lors de leur

dernière session au collégial ont 8 fois plus de chances de quitter que les étudiants ayant une moyenne générale au dessus de 75%.

Notre seconde hypothèse était qu'il est également possible que les élèves qui décrochent aient des moyenne au-dessus de 60% tout en ayant des cours qu'ils échouent. En effet, un étudiant peut avoir une moyenne au dessus de 60% en ayant des notes de passage dans tous leurs cours juste au dessus de 60% ou bien en ayant certaines notes élevées et certains échecs qui ramènerait la moyenne légèrement au dessus de 60%. Pour déterminer l'impact d'un échec sur le décrochage, nous examinons donc la note la plus basse obtenue par des étudiants, lors de leur dernière session, que ce soit avant de décrocher ou avant l'obtention d'un diplôme.

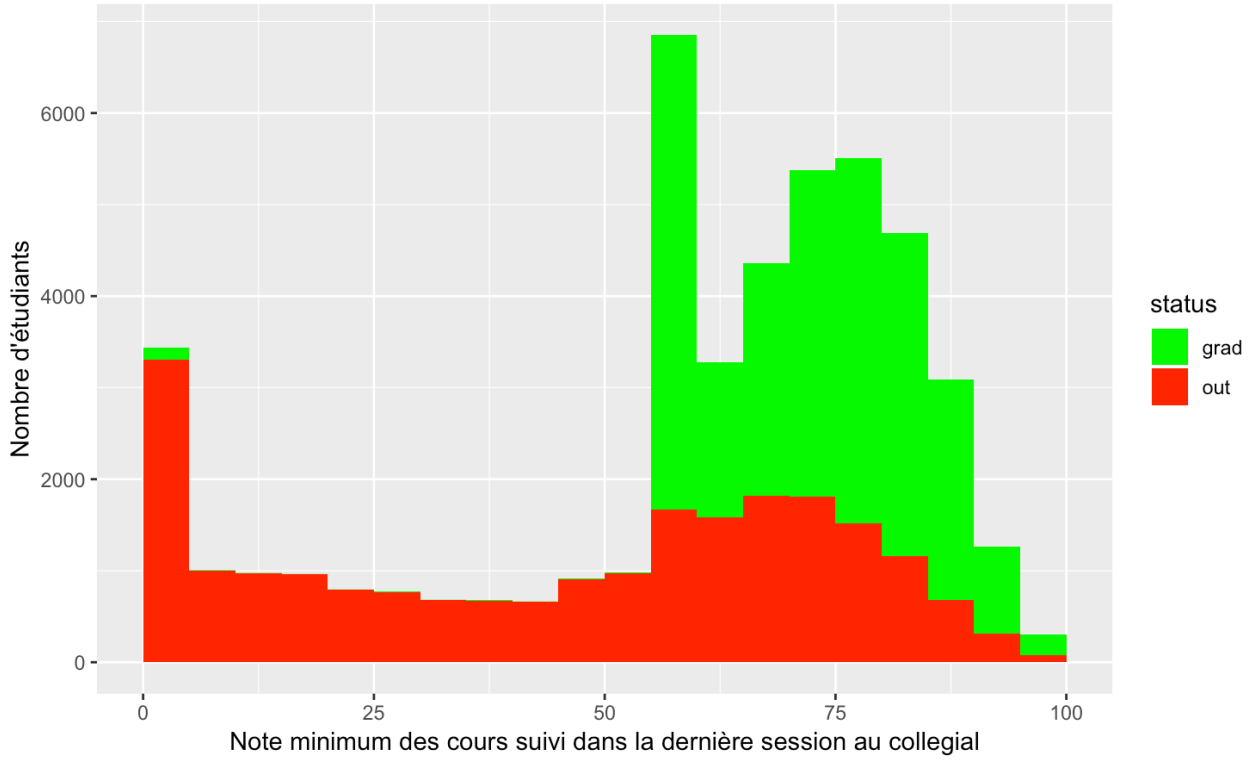


Figure 5.3 Note minimum lors de la dernière session pour étudiants qui quittent (rouge) et qui graduent (vert)

Il est important de souligner que, comme prévu, les notes les plus basses obtenues par les élèves diplômés sont supérieures à 60%, ce qui signifie qu'ils ont passé tous leurs cours lors de la dernière session. En revanche, un peu plus de la moitié (53.7%) des décrocheurs ont une

note minimale lors de leur dernière session qui est en deçà de la note minimale de passage de 60%. Donc, nous observons qu'un peu plus de la moitié des étudiants qui décrochent ont échoué au moins un cours dans la session académique précédant leur départ. Néanmoins, il existe tout de même près de la moitié (46.3 %) de ceux qui quittent pour lesquels la pire note à leur dernière session inscrite au collégial est supérieure à 60 %. Pour près de la moitié de ces étudiants qui quittent, aucun cours n'a été échoué lors de la dernière session. De plus, chez 18.6% des décrocheurs, la pire note obtenue au cours de leur dernière session est supérieure à 75 %! S'ils n'avaient pas décroché, ces étudiants auraient été considérés par toutes les autres méthodes comme étant d'excellents élèves.

Ces données peignent un tableau intéressant de la population de décrocheurs. Il est maintenant possible de dire que la représentation mentale qui associe le décrocheur à un élève obtenant nécessairement de mauvais résultats scolaires ne doit plus être le modèle de référence puisqu'une partie importante des décrocheurs NE répond PAS à ce critère.

Il est également vrai que d'échouer un cours lors de la dernière session est un facteur de risque pour le décrochage. En effet, est possible de prendre la présence ou absence d'un échec lors de la dernière session pour évaluer un Rapport de Cote nous permettant de déterminer la différence relative des chances de graduation entre ceux qui ont échoué un cours et ceux qui n'ont pas échoué un cours lors de leur dernière session au collégial. Parmi les étudiants n'ayant pas échoué de cours, 24 067 obtiennent un diplôme tandis que 10 333 quittent. En revanche, parmi ceux qui ont échoué au moins un cours lors de leur dernière session, seulement 198 d'entre eux graduent tandis que 11 963 quittent l'établissement. Nous obtenons un rapport de cote OR=141 (IC95%=122-162, $p<0.0001$) indiquant que l'échec d'un cours dans la session augmente les chances de quitter de plus de 140 fois. Nous utilisons aussi la note minimale de 75% pour reproduire ce rapport de cote et obtenons OR=4.3 (IC95%= 4.1-4.4) indiquant que les élèves ayant une note minimale de 75% ont 4 fois plus de chances de graduer, ou de façon équivalente que ceux ayant une note minimale au dessous de 75% ont 4 fois plus de chances de quitter l'établissement collégial.

Bien que l'échec d'un cours est donc un facteur de risque clair et significatif du décrochage, de n'avoir pas échouer un cours lors de la dernière session n'est pas une garantie de réussite puisque plus d'un cinquième de notre échantillon (22.2% ou 10 333 étudiants) n'a échoué aucun cours lors de la dernière session académique. Les décrocheurs sont-ils alors le fruit de plusieurs sessions de mauvais résultats scolaires? Ou bien sont-ils la conséquence d'un long processus de mauvais résultats scolaires sur plusieurs sessions différentes? Si cette dernière hypothèse s'avérait correcte, les données sur la distribution des notes au fil du temps devrait le corroborer. On examine la session précédant la diplomation ou le décrochage scolaire (la troisième session pour un diplômé typique d'un programme de deux ans) et la session survenue un an avant leur diplomation ou leur décrochage (la deuxième session pour un diplômé typique d'un programme de deux ans). Commençons par étudier cette question en examinant la moyenne générale des élèves pendant la dernière session au cours de laquelle soit ils décrochent, soit ils obtiennent leur diplôme.

Moyennes générales dans la session précédant la diplomation ou le décrochage

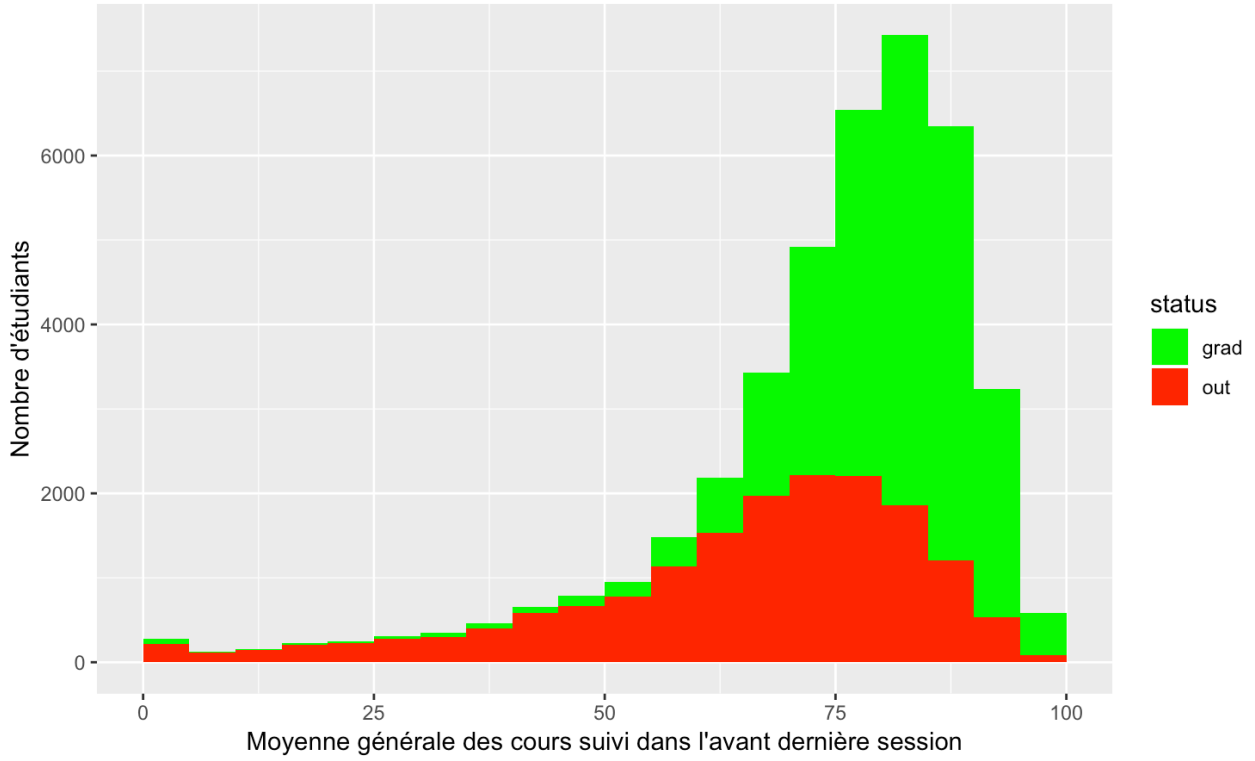


Figure 5.4 Moyenne générale lors de l'avant dernière session pour étudiants qui quittent (rouge) et qui graduent (vert)

En examinant ce graphique, nous pouvons voir que les deux distributions se rapprochent l'une de l'autre. Leurs moyennes générales sont maintenant de 80.23±0.07 % et de 65.28±0.15 % respectivement. En remontant dans le temps, il est facile de remarquer qu'il y a de moins en moins de différence apparentes dans les distribution de résultats scolaires entre les futurs diplômés et les futurs décrocheurs. Il est important de noter que 36.7 % des décrocheurs ont une moyenne générale supérieure à 75 % au cours de la session ayant précédé celle lors de laquelle ils ont décroché. De plus, 71.1 % des décrocheurs ont une moyenne générale supérieure à 60 % au cours de leur avant-dernière session. Les données suggèrent clairement qu'une partie des décrocheurs ne décroche pas en raison d'une tendance à la baisse de leur rendement scolaire, mais plutôt possiblement en raison de causes plus spontanément propres à une session en particulier.

Il est encore possible de prendre la moyenne générale de 60% comme seuil pour évaluer un Rapport de Cote ("Odds Ratio") nous permettant de déterminer la différence relative des chances de graduation entre ceux qui ont une moyenne générale au-dessus de 60% et ceux qui ont une moyenne générale en dessous de 60% lors de leur avant dernière session au collégial. Parmi les étudiants ayant une moyenne générale au-dessus de 60%, 23 164 obtiennent un diplôme tandis que 11 852 quittent. En revanche, parmi ceux qui obtiennent une moyenne générale en deçà de 60%, seulement 865 d'entre eux graduent tandis que 4809 quittent l'établissement. Nous obtenons un rapport de cote OR=10.9 (IC95%=10.8-11.7, $p<0.0001$) indiquant que les étudiants ayant une moyenne générale en dessous de 60% lors de leur avant dernière session au collégial ont 10,9 fois plus de chances de quitter que les étudiants ayant une moyenne générale au-dessus de 60%. On voit donc que d'avoir une moyenne générale en dessous de 60% dans l'avant dernière session est un prédicteur significatif de décrochage. Néanmoins, d'avoir au-dessus de 60% n'est pas une garantie de réussite puisque plus la plupart des étudiants (71%) qui quittent l'établissement ont une moyenne générale au-dessus de 60% est plus de la moitié d'entre eux (36.7%) ont une moyenne générale au-dessus de 75%. Il est aussi possible de calculer un rapport de cote pour les élèves ayant une moyenne générale au dessus de 75% et ceux qui ont une moyenne générale en dessous de 75% lors de leur avant dernière session au collégial. Nous obtenons un rapport de cote OR=5.8 (IC95% 5.6-6.1)

indiquant que les élèves ayant des moyennes générales au dessous de 75% ont 5.8 fois plus de chances de quitter l'établissement.

Il est possible que malgré une moyenne générale acceptable, certains futurs décrocheurs aient un ou deux cours avec de très mauvaises notes bien que la moyenne générale resterait parfaitement acceptable. Il est alors intéressant de comparer l'avant dernière session des élèves ci-dessus en ce qui concerne les notes les plus basses obtenues. Le tracé ci-dessous permet de faire cette comparaison.

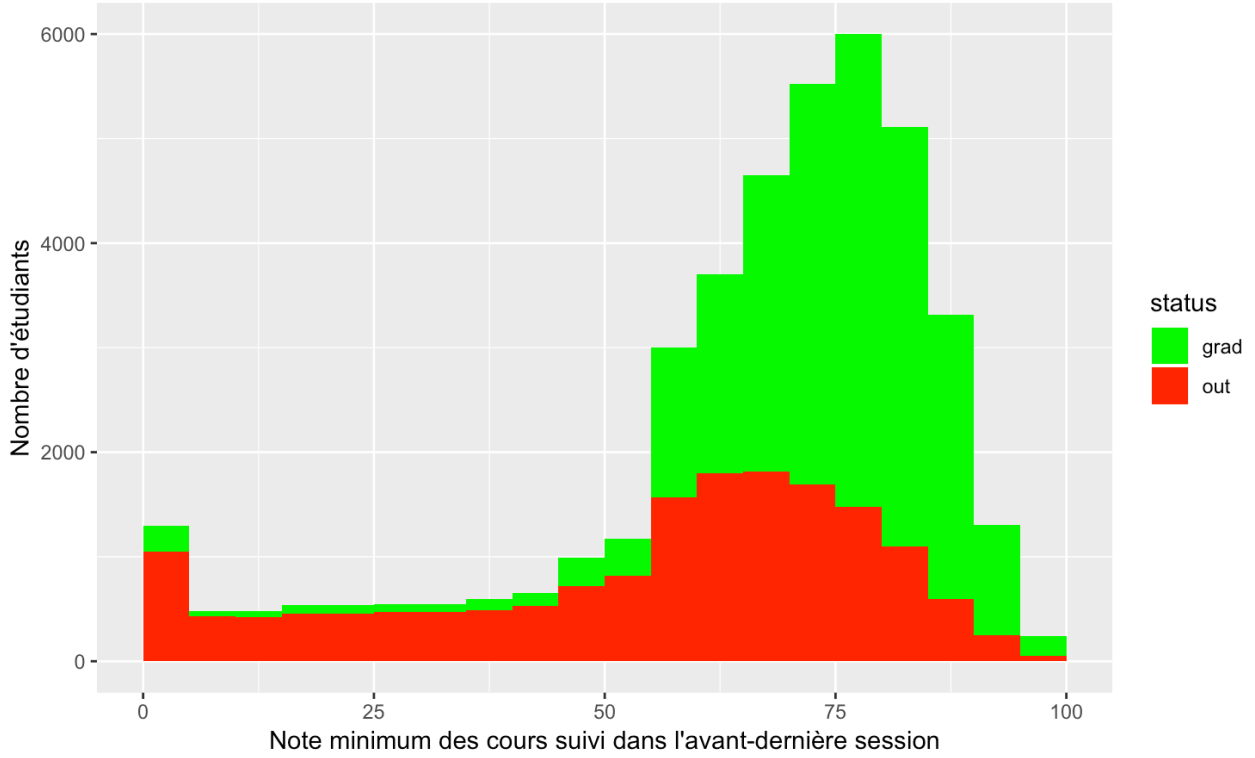


Figure 5.5 Note minimale lors de l'avant dernière session pour étudiants qui quittent (rouge) et qui graduent (vert)

Dans la figure 5.5 ci-dessus, nous examinons la note minimale lors de l'avant dernière session pour étudiants qui quittent (rouge) et qui graduent (vert). Ce graphique montre que, lorsque l'on remonte dans le temps dans le rendement scolaire des élèves, on constate une amélioration du rendement. Dans ce cas, nous voyons que la pire note obtenue au cours de l'avant dernière session des étudiants qui quittent est en moyenne de 55.1±0.2 % comparativement à 45.6±0.2% lors de leur dernière session. Ceci est néanmoins bien en dessous de la moyenne de la note la plus basse des futurs diplômés (73.71±0.09 %) pour

l'avant dernière session. Cependant, si l'on se concentre toujours sur les décrocheurs à haut rendement, nous constatons que, chez 60.7 % des décrocheurs, la pire note obtenue au cours de cette avant dernière session est supérieure à 60 %. De plus, chez 23.3% des décrocheurs, la pire note obtenue est supérieure à 75 %.

Il est possible de prendre la note la plus basse obtenue avec 60% comme seuil. Nous évaluons ainsi un Rapport de Cote nous permettant de déterminer la différence relative des chances de graduation entre ceux qui n'ont pas d'échec (note minimale au-dessus de 60%) de ceux qui ont au moins un échec (note minimale en dessous de 60%) lors de leur avant dernière session au collégial. Parmi les étudiants n'ayant aucun échec, 22 406 obtiennent un diplôme tandis que 10 113 quittent. En revanche, parmi ceux qui ont au moins un échec, seulement 1633 d'entre eux graduent tandis que 6548 quittent l'établissement. Nous obtenons un rapport de cote OR=8.9 (IC95%=8.4-9.4, $p<0.0001$) indiquant que les étudiants ayant au moins un échec lors de leur avant dernière session au collégial ont 8,9 fois plus de chances de quitter que les étudiants n'ayant aucun échec. On voit donc que d'avoir un échec dans l'avant dernière session est un prédicteur significatif de décrochage. Néanmoins, d'avoir une note minimale au-dessus de 60% n'est pas une garantie de réussite puisque qu'une partie importante de cet échantillon (10 113 étudiants) quittent l'établissement même avec une note minimale au-dessus de 60%. De plus, il est possible de reproduire le rapport de cote avec une note minimale étant soit en dessous, soit au-dessus de 75%. Nous obtenons un rapport de cote OR=3.3 (IC95%= 3.1-3.4) indiquant que les étudiants ayant une note minimale en dessous de 75% ont 3 fois plus de chances de quitter l'établissement collégial.

L'évolution des données en matière de rendement scolaire, entre la session au cours de laquelle ils décrochent ou obtiennent leur diplôme et la session précédant celle-ci, nous démontre que les distributions sont de plus en plus rapprochées les unes des autres et ne restent pas très éloignées. Ceci peut être observé à la fois dans les graphiques illustrant les moyennes générales des élèves et les notes les plus basses des élèves. Il est maintenant intéressant de vérifier si cette tendance se poursuit dans cette direction où les deux distributions se rapprochent de plus en plus à la fois en ce qui concerne les moyennes

générales et les notes les plus basses. Savoir si l'écart entre les deux groupes demeurera (comme c'est le cas dans le graphique de la session au cours de laquelle ils ont décroché ou obtenu leur diplôme) ou s'il commence à diminuer (comme le graphique précédent) nous en dira beaucoup sur la nature du processus menant au décrochage.

Les moyennes des deux sessions précédant la diplomation ou le décrochage

Finalement, penchons-nous sur les deux sessions précédant la diplomation ou le décrochage.

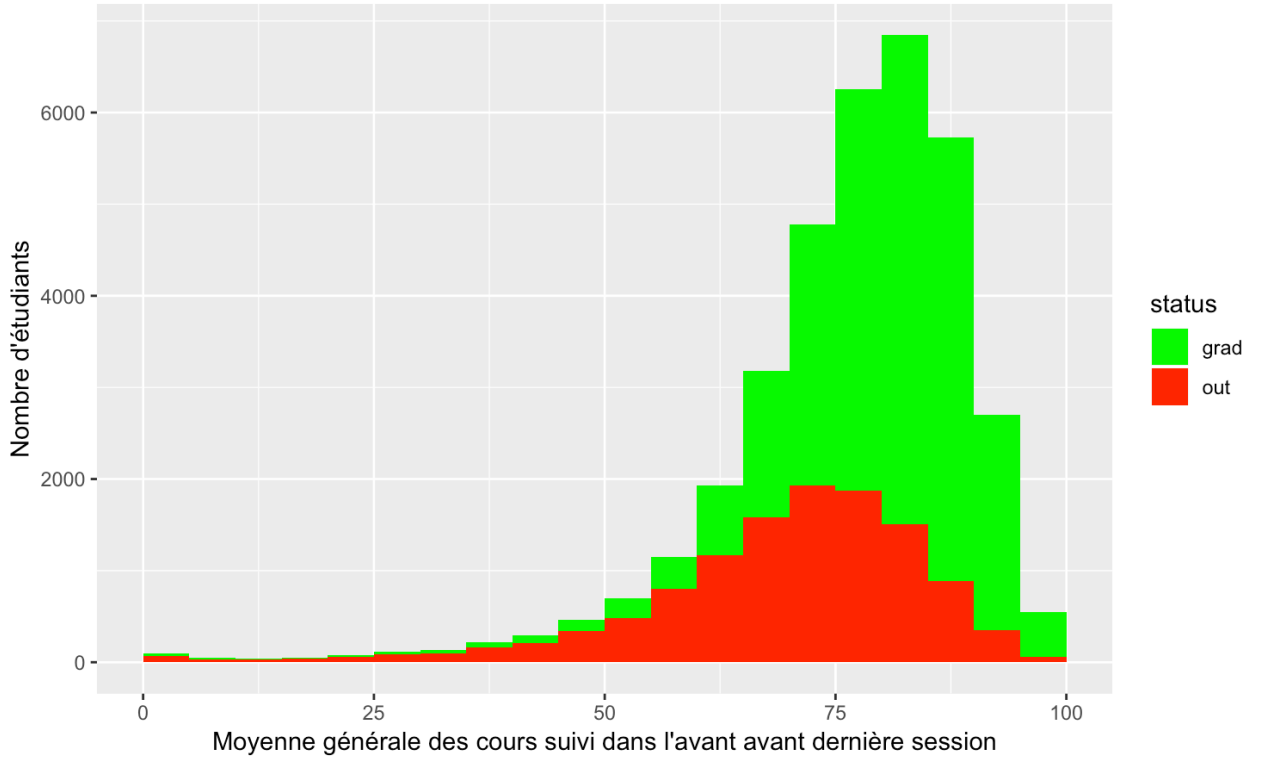


Figure 5.6 Moyenne générale dans l'avant-avant dernière session, étudiants qui quittent (rouge) et qui graduent (vert)

Il ressort clairement de ce graphique que l'écart de rendement entre les diplômés et les décrocheurs se réduit à des moyennes de 79.75±0.07 % et 69.6±0.1%, respectivement. De plus, comme auparavant, une bonne proportion des décrocheurs éventuels obtiennent de bonnes notes deux ou trois sessions avant de décrocher. En effet, 81.1% d'entre eux ont une moyenne supérieure à 60 % et plus de la moitié de ceux-là (41.2%) ont une moyenne supérieure à 75 %. Nous évaluons aussi le Rapport de Cote permettant de déterminer la

différence relative des chances de graduation entre ceux qui ont des moyennes générales au-dessus de 60% dans leur avant-avant dernière session et les comparons à ceux qui ont des moyennes générales en dessous de 60% lors de leur avant-avant dernière session au collégial. Parmi les étudiants ayant une moyenne générale au-dessus de 60%, 22 704 obtiennent un diplôme tandis que 9528 quittent. En revanche, parmi ceux qui ont une moyenne générale en dessous de 60%, seulement 903 d'entre eux graduent tandis que 2226 quittent l'établissement. Nous obtenons un rapport de cote OR=5.9 (IC95%=5.4-6.4, p<0.0001) indiquant que les étudiants ayant une moyenne générale en dessous de 60% lors de leur avant-avant dernière session au collégial ont 5,9 fois plus de chances de quitter que les étudiants ayant une moyenne générale au-dessus de 60%. Néanmoins, d'avoir une moyenne générale au-dessus de 60% n'est pas une garantie de réussite puisque qu'une partie importante de cet échantillon (9528 étudiants) quittent tout de même l'établissement. De plus, il est possible de reproduire le rapport de cote avec une moyenne générale étant soit en dessous, soit au-dessus de 75%. Nous obtenons un rapport de cote OR=4.3 (IC95%= 4.1-4.5) indiquant que les étudiants ayant une note minimale en dessous de 75% ont 4 fois plus de chances de quitter l'établissement collégial.

Examinons finalement les notes les plus basses de ces élèves pour voir si elles suivent la même tendance. La figure 5.7 ci-dessous montre que la tendance suivie par les notes les plus basses est la même que la tendance pour les moyennes générales. Par souci d'exhaustivité, notons que, chez 70.1 % des décrocheurs éventuels, la note la basse obtenue deux sessions avant leur dernière est supérieure à 60 % et, dont 24.4 % d'entre eux ayant la pire note de la session supérieure à 75 % (contre 52.3 % pour les diplômés potentiels). Évaluons le Rapport de Cote permettant de déterminer la différence relative des chances de graduation entre ceux qui n'ont pas d'échec (note minimale au-dessus de 60%) de ceux qui ont au moins un échec (note minimale en dessous de 60%) lors de leur avant-avant dernière session au collégial. Parmi les étudiants n'ayant aucun échec, 21 614 obtiennent un diplôme tandis que 8235 quittent. En revanche, parmi ceux qui ont au moins un échec, 1993 d'entre eux graduent tandis que 3519 quittent l'établissement. Nous obtenons un rapport de cote OR=4.6 (IC95%=4.4-4.9, p<0.0001) indiquant que les étudiants ayant au moins un échec lors de leur avant dernière

session au collégial ont plus de 4-fois plus de chances de quitter que les étudiants n'ayant aucun échec.

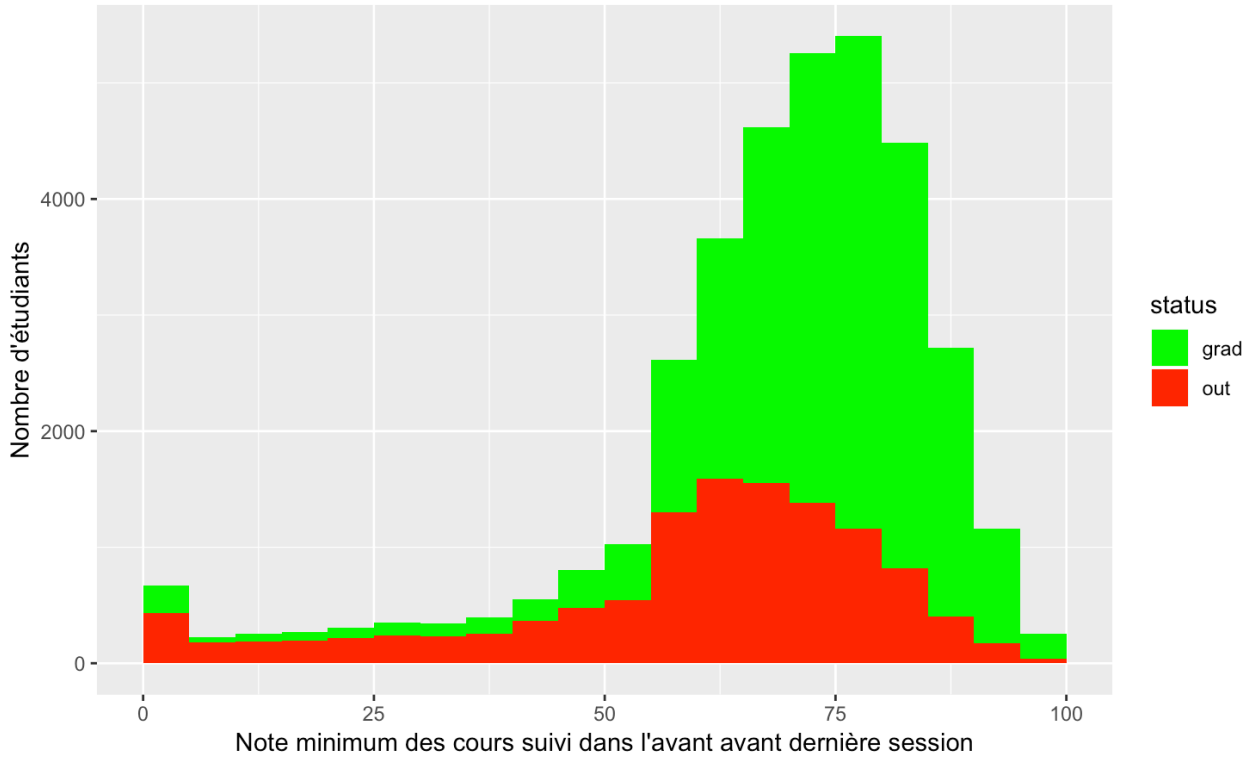


Figure 5.7 Note minimale, avant-avant dernière session pour étudiants qui quittent (rouge) et qui graduent (vert)

On voit donc que d'avoir un échec dans l'avant-avant dernière session est un prédicteur significatif de décrochage. Néanmoins, de n'avoir aucun échec n'est pas une garantie de réussite puisque qu'une partie importante de cet échantillon (8235 étudiants) quittent l'établissement même avec aucun échec. De plus, il est possible de reproduire le rapport de cote avec une note minimale étant soit en dessous, soit au-dessus de 75%. Nous obtenons un rapport de cote OR=3.4 (IC95%= 3.2-3.6) indiquant que les étudiants ayant une note minimale inférieure à 75% ont 3 fois plus de chances de quitter l'établissement collégial.

Résumé des résultats sur le rendement scolaire

Les résultats de cette section peuvent être résumés en deux temps. Premièrement, tel qu'on pourrait s'y attendre, le rendement scolaire affecte grandement et significativement les

chances de décrocher. Les figures 5.8 et 5.9 montrent l'impact sur le risque de décrochage que procure une moyenne générale en dessous de 60% et de 75% respectivement.

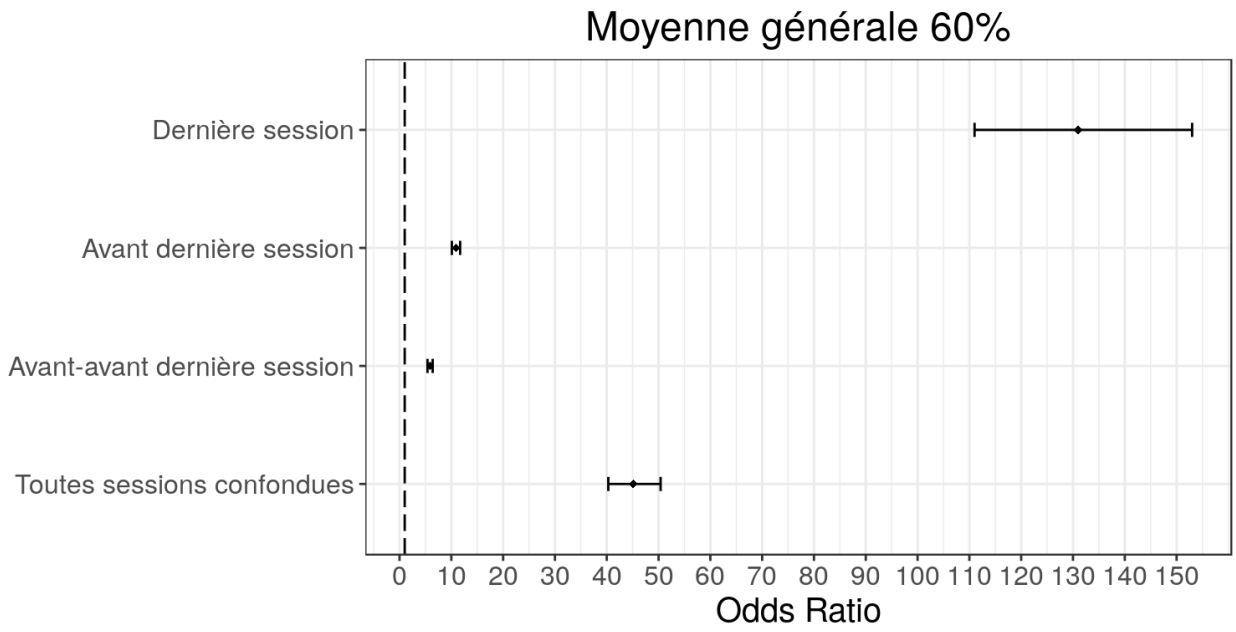


Figure 5.8 Augmentation des chances de décrocher d'entre 6 à 130 fois pour les étudiants ayant une moyenne générale au-dessous de 60%

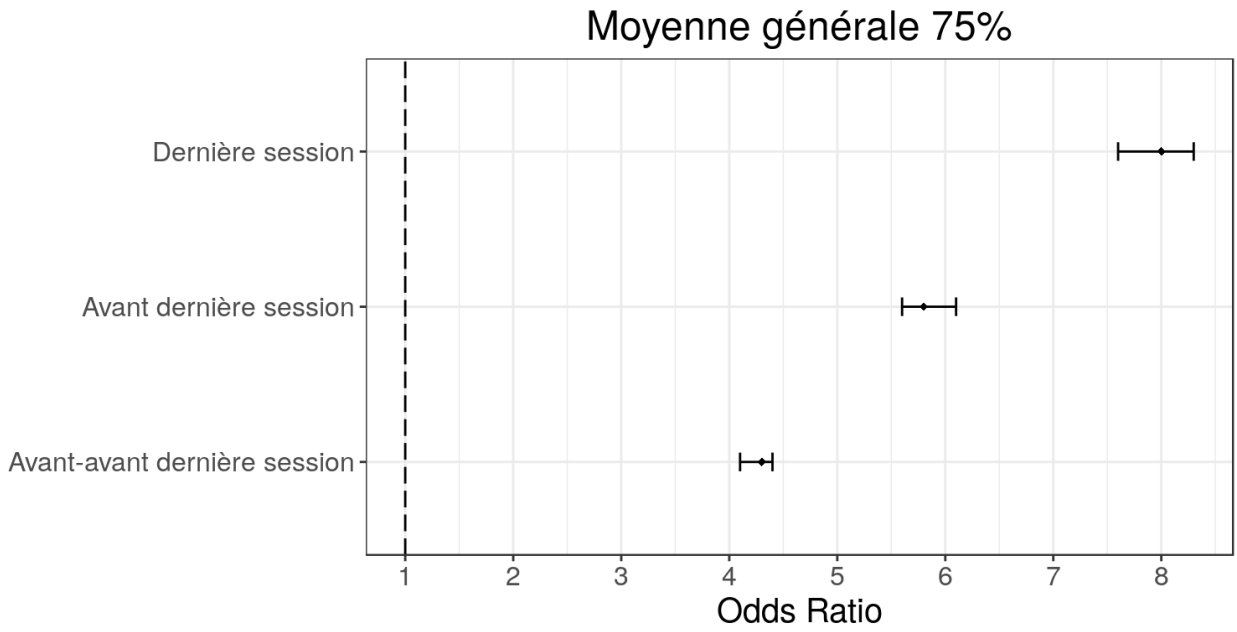


Figure 5.9 Augmentation des chances de décrocher d'entre 4 à 8 fois pour les étudiants ayant une moyenne générale au-dessous de 75%

La figure 5.10 montre l'impact important que procure l'échec d'un cours ou plus sur le risque de décrocher. Les chances de décrochage variant entre une augmentation de 4 fois à 140 fois les chances de décrocher par rapport aux étudiants n'ayant pas d'échec. À noter aussi que l'échec dans une dernière session est très différent que l'échec dans les sessions précédentes. Une des possibilités expliquant la différence entre une session régulière et la dernière session d'un étudiant est que la dernière session d'un décrocheur peut aussi bien être la quatrième que la première session du programme. il suffit qu'il n'y ait pas d'information colligées pour cet étudiant après une session donnée pour que la session soit caractérisée comme étant la dernière. Néanmoins, ces données suggèrent que l'échec en dernière session pourrait être un phénomène différent d'un échec dans une session normale.

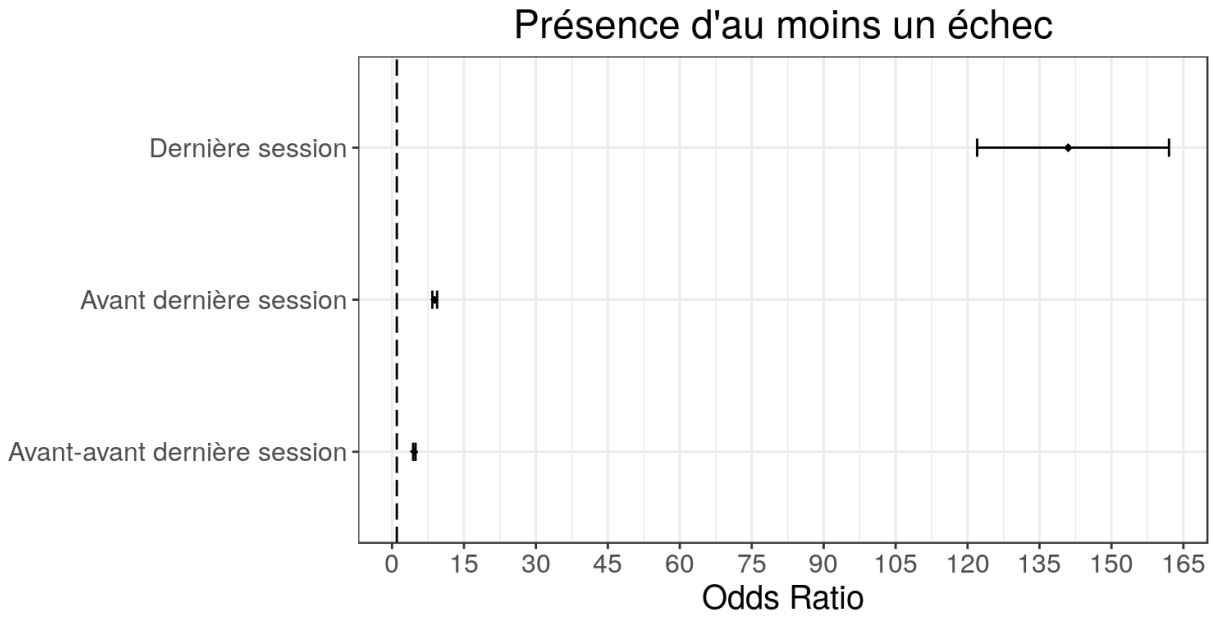


Figure 5.10 Augmentation des chances de décrocher d'entre 4 à 140 fois pour les étudiants ayant un échec ou plus

Finalement, la figure 5.11 part de la prémisse que les étudiants arrivants au cégep ont en moyennes des notes au secondaire avoisinant les 75% et évalue l'impact d'avoir au-dessus ou en dessous de 75% sur le risque de décrocher. On observe ci-dessous que les chances de décrocher augmentent de 3.3 à 4.3 fois pour les étudiants dont la plus basse note dans la session est inférieure à 75% comparés à ceux qui n'ont aucune note inférieur à 75%.

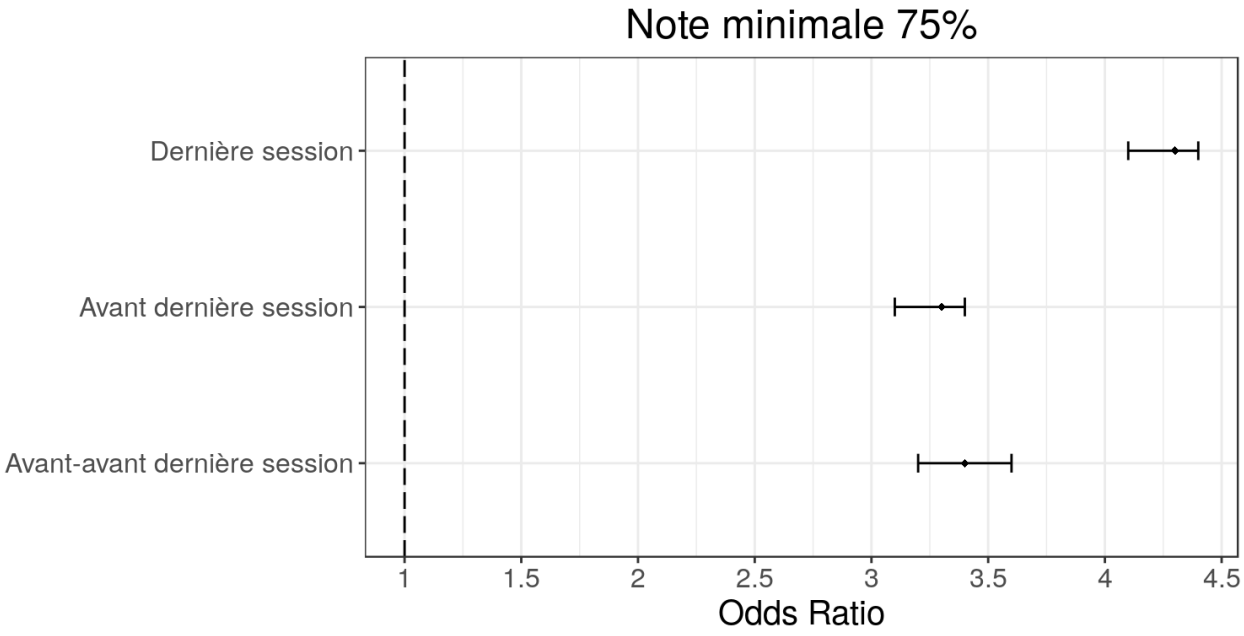


Figure 5.11 Augmentation des chances de graduer pour les étudiants qui n'ont aucune note en dessous de 75%

Le deuxième message est quelque peu contre-intuitif par rapport au premier et souligne la grande fraction d'étudiants qui n'ont pas les facteurs de risques comme la présence d'un échec ou des notes faibles et qui néanmoins décrochent. La figure 5.12 prend comme échantillon l'ensemble des étudiants ayant une moyennes générale au-dessus de 60% (et au-dessus de 75%) et montre la fraction d'entre eux qui décroche. Les données montrent que près d'un tiers des étudiants ayant des moyennes générales au-dessus de 60% et près d'un quart de ceux ayant des moyennes générales au-dessus de 75% finissent néanmoins par décrocher. La figure 5.13 examine les étudiants n'ayant aucun échec (ou note minimale 75%) et montre la fraction d'entre eux qui décroche. Le graphique montre que parmi les élèves n'ayant aucun échec, plus d'un quart décrochent (et 20% de ceux ayant 75%+).

Ces résultats mettent en tension les deux idées complémentaires de ce chapitre, soit que le rendement scolaire a un impact important sur le taux de décrochage, mais qu'en revanche le bon rendement scolaire n'est pas une garantie de persévérance puisque un fraction importante d'étudiants ont de bons rendements et décrochent néanmoins.

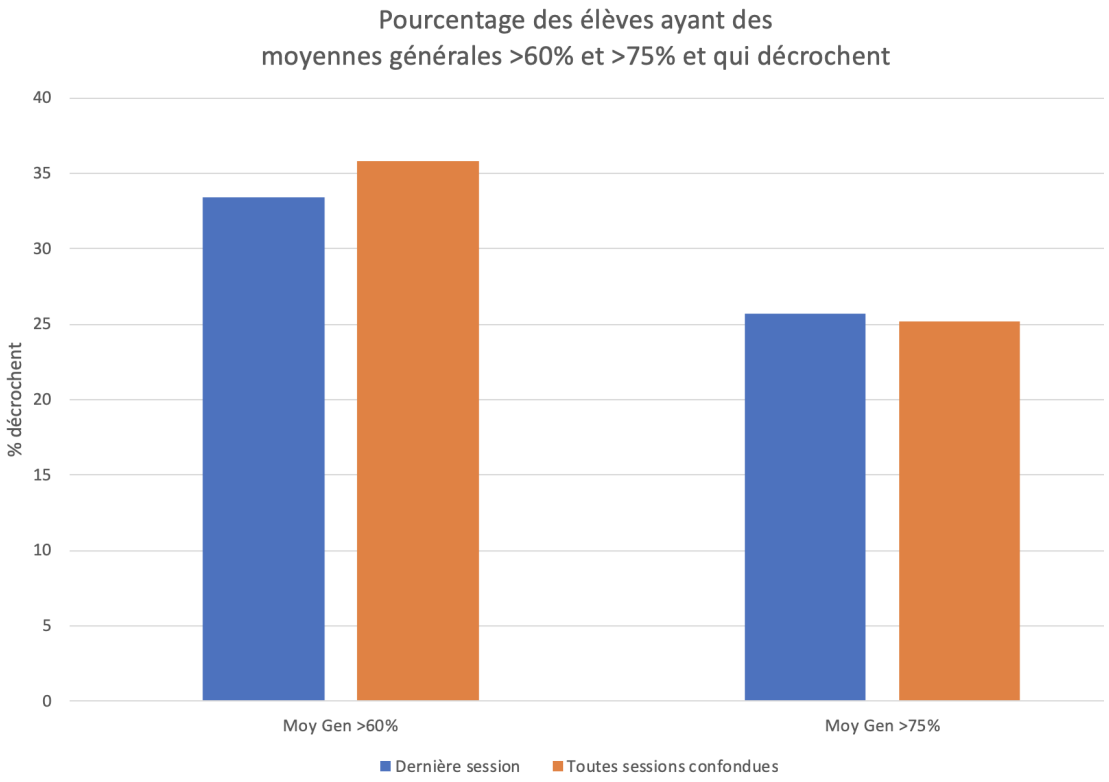


Figure 5.12 Fraction des étudiants ayant des moyennes générales >60% ou >75% et qui décrochent

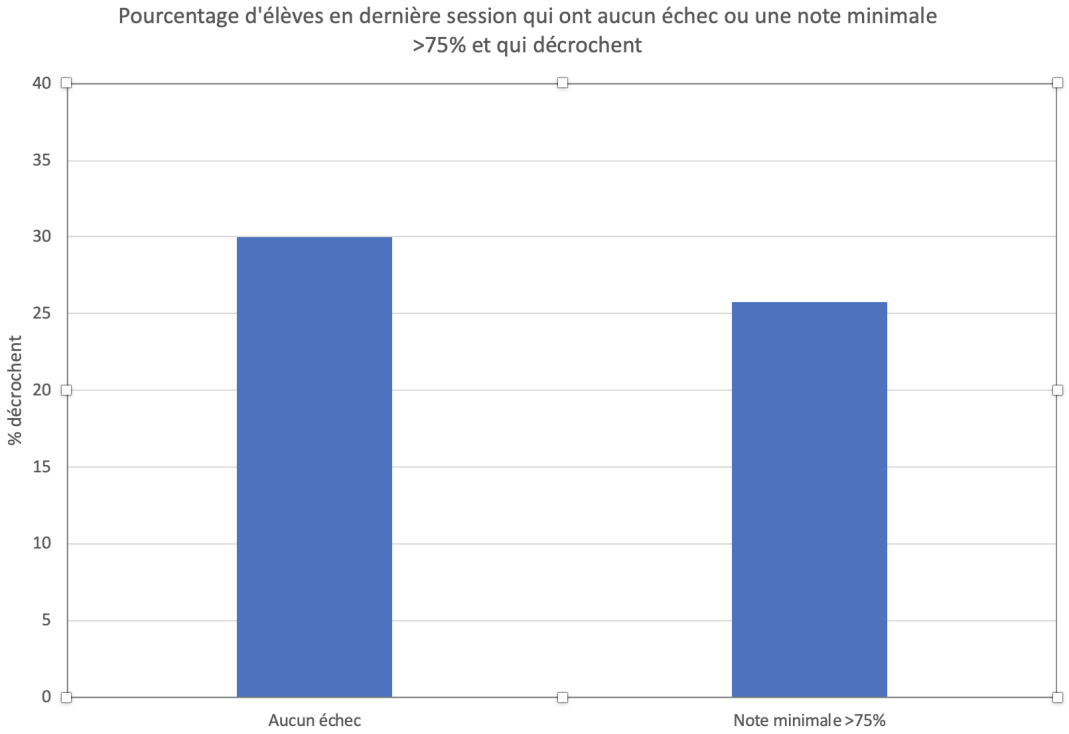


Figure 5.13 Fraction des étudiants ayant aucun échec ou notes minimales >75% et qui décrochent

Il faut donc réfléchir à la façon dont nous pouvons mieux identifier ces élèves puisque, à chaque session qui passe, leur rendement scolaire diverge de plus en plus, et ce, jusqu'à ce qu'ils en viennent à décrocher. La figure 5.14 montre comment l'écart entre les gradués et le décrocheurs s'accentue à mesure que les élèves se rapprochent de leur diplomation ou de leur décrochage.

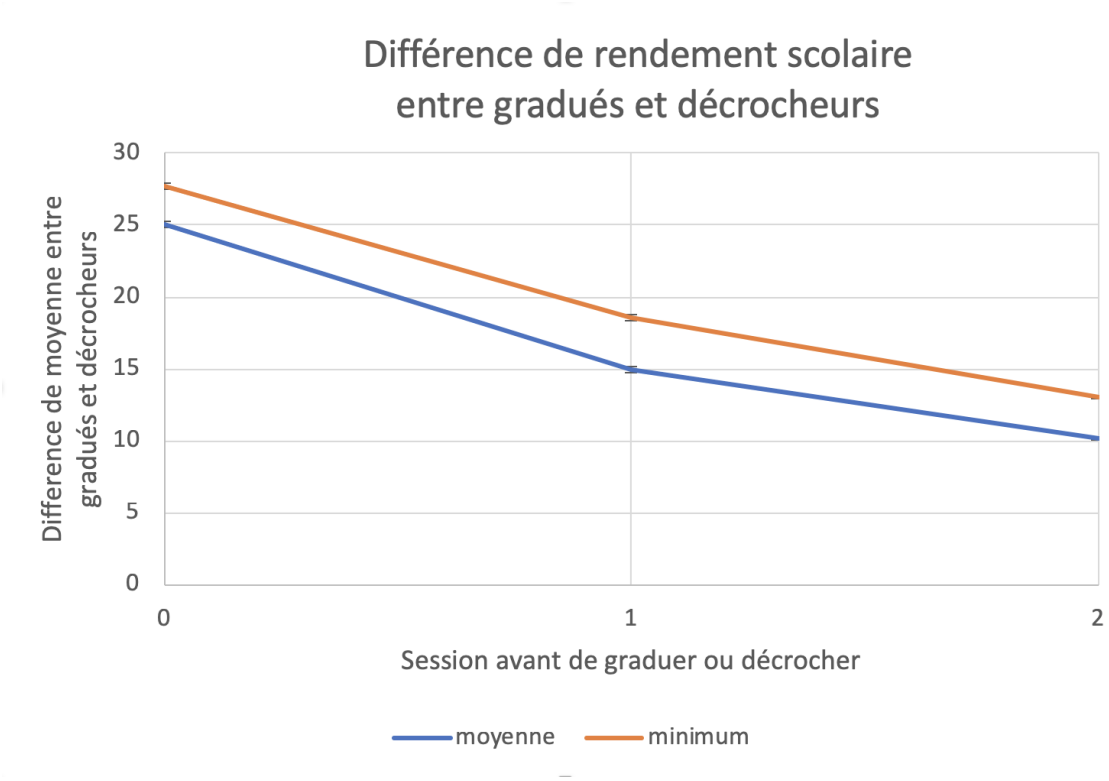


Figure 5.14 Divergence du rendement scolaire par session entre les gradués et les décrocheurs

La figure 5.14 montre la différence (que ce soit au niveau de la moyenne générale ou de la note la plus basse) entre les diplômés et les décrocheurs, à mesure que chaque session passe. Ce différence commence à 10 points et croît au fur et à mesure de sorte à ce qu'une session avant de décrocher la différence est de 15 points et la dernière session la différence croit à 25 points. Ces données mettent encore plus de pression sur le système pour qu'on identifie et aide ces élèves dès que possible. Dans la section suivante, nous allons examiner comment dont nous pourrions faire mieux.

Notons donc les trois points principaux importants à retenir sur cette section :

1. L'échec d'au moins un cours et le pauvre rendement scolaire en général sont des facteurs de risques significatifs pour le décrochage.
2. Une fraction importante des étudiants n'ayant aucun des ces facteurs de risque reliés au rendement scolaire décrochent néanmoins.
3. La différence entre le rendement scolaire des décrocheurs et des diplômés s'accentue lorsque la session de la diplomation ou du décrochage approche.

En prenant du recul par rapport à ces résultats, force est de constater qu'une autre hypothèse doit être proposée pour expliquer la fraction importante de décrocheurs qui ont des rendements acceptables ou même excellents. Pourquoi ces décrocheurs quittent-ils?

Il est peu probable que leur réseau ou leur situation financière personnelle aient sensiblement changé, car ce sont des paramètres relativement fixes. Il est possible que leur motivation ait chuté malgré le fait que leurs notes étaient bonnes. Une autre raison pourrait expliquer pourquoi ces étudiants ayant de bons résultats scolaires décrochent. Il est possible qu'ils ne décrochent pas, mais qu'ils s'en vont de ce cégep en particulier ou quittent le système québécois en entier. Il est donc possible que les élèves soient considérés comme des décrocheurs dans notre étude (et d'autres) alors qu'ils aient tout simplement décidé de poursuivre leurs études ailleurs. Il y a une brèche claire dans les données et dans la façon dont on fait le suivi des renseignements à propos des élèves. Si l'on supposait que tous les décrocheurs « très performants » ont poursuivi leurs études ailleurs et ne sont donc pas des décrocheurs, le problème du Québec en matière de persévérance scolaire disparaîtrait quantativement.

Par ailleurs, il est essentiel de mentionner que la majorité des méthodes actuelles de prévention du décrochage sont axées sur l'examen du rendement scolaire des élèves. Bien que nos données corroborent l'importance du rendement scolaire, nous constatons clairement, à partir de nos données, que même au cours de la session lors de laquelle ils décrochent, plus de

la moitié des décrocheurs (54.2 %) obtiennent la note moyenne de passage. En d'autres termes, les méthodes actuelles de prédiction du décrochage sont axées sur et s'intéressent à une unité de mesure inefficace pour la moitié de la population qu'ils essaient de détecter.

Si le décrochage s'avérait être un événement spontané, l'indicateur d'alerte le plus précoce que l'on peut obtenir des élèves pourrait être l'évaluation de mi-session administrée par l'enseignant. Nous allons maintenant analyser les résultats de ces évaluations de mi-session.

CHAPITRE 6

Rôle d'évaluations de mi-session (ÉMS) dans la prédiction du statut final de l'élève

Les évaluations de mi-session sont utilisées dans la sphère académique depuis de nombreuses décennies. Elles ont d'abord été utilisées comme des outils de rétroaction permettant aux élèves de communiquer avec l'enseignant et lui fournir des commentaires et suggestions quant au rendement rythme, au contenu ou à tout autre aspect du cours. Ces éléments de rétroaction étaient utiles aux enseignants puisqu'ils leur permettaient d'apporter des modifications à leurs cours en fonction des rétroactions réelles obtenues des élèves. Maintenant obligatoires dans plusieurs collèges et universités, certaines d'entre elles sont maintenant personnalisables par l'enseignant, par exemple si l'enseignant a des questions spécifiques à poser aux élèves pour savoir si certaines des tactiques d'enseignement utilisées sont aussi perçues comme étant utiles par les élèves.

Un autre usage des évaluations de mi-session (ÉMS) consistait à donner aux élèves une rétroaction à propos de leur rendement en classe même s'il n'y a généralement pas eu beaucoup d'évaluations de faites avant la mi-session. Ces ÉMS ont aussi été conçues comme un mécanisme de rétroaction de l'enseignant à l'élève. En effet, les EMS permettent à l'enseignant de faire savoir à l'élève si, en fonction de son rendement actuel, l'élève est susceptibles de réussir (« Réussite ») ou d'échouer (« Échec ») à l'examen ou se situaient quelque part au milieu (« À risque »). Il est important de mentionner que ces résultats ne sont pas notés et ne sont généralement pas remplis par les enseignants. En remplissant simplement les ÉMS en utilisant les notes actuelles des élèves, ils sont beaucoup plus englobants. Si l'enseignant remarque qu'un élève obtient la note de passage lors de l'ÉMS, mais a de la difficulté à saisir un concept clé, l'enseignant a encore la possibilité d'attribuer une note autre que « Réussite » à l'élève sachant ce qui s'en vient dans le contenu du cours. Mais en général,

les ÉMS ne sont pas perçus comme des instrument précis, en partie pour la simple raison qu'à notre connaissance, leur efficacité n'a jamais été documenté.

Certains des trois collèges pour lesquels nous avons des données utilisent les ÉMS comme moyen d'identifier les élèves à risque de décrocher. Chaque école a une approche légèrement différente, bien que toutes les trois ont de bonnes intentions en essayant d'agir en fonction des renseignements disponibles le plus tôt possible. Pour mieux comprendre comment un collège peut utiliser les ÉMS pour diminuer le nombre de décrocheurs, prenons un exemple. Un des collège de notre ensemble de données inscrit les élèves sur leur « Liste d'élèves à risque de décrocher » s'ils reçoivent deux mentions « Échec » ou plus lors d'une ÉMS. Tous les élèves qui répondent à ces critères recevront un courriel à l'interne en leur proposant quelques ressources par lesquelles ils peuvent obtenir de l'aide, ainsi que la disponibilité d'un conseiller aux élèves qui peut discuter plus en profondeur des difficultés de l'élève et le jumeler à un pair ou à toute autre ressource qu'il juge utile. Il y a plusieurs moyens par lesquels les données peuvent nous aider à améliorer ce processus.

La première manière dont un ÉMS peut être amélioré consiste à ne pas utiliser les résultats scolaires comme prédicteur principal du risque de décrochage. Comme nous l'avons vu à la section précédente, 47 % des décrocheurs ont une note de passage lors de la session pendant laquelle ils décrochent. Par conséquent, tous les mécanismes de dépistage fondés sur les notes manqueront probablement un grand pourcentage de décrocheurs.

La deuxième manière dont les ÉMS peuvent être amélioré consiste à utiliser toutes les données à notre disposition pour faire des prédictions sur la réussite ou l'échec d'un élève à un cours et s'il poursuivra ses études ou décrochera. Dans le cadre de ce projet de recherche, nous avons établi les fondements permettant d'obtenir ces données et d'utiliser des algorithmes d'apprentissage machine pour en arriver à effectuer de très bonnes prédictions qui tiennent compte de tous les renseignements disponibles au sujet d'un élève et d'effectuer des prédictions basées sur tous les renseignements, pas seulement une seule donnée. Imaginez une fonction qui permette, lorsque toutes les ÉMS sont saisies par les enseignants pour les

élèves, que le dossier complet de chaque élève soit comparé aux algorithmes de classification de l'apprentissage machine et classé selon l'une des deux catégories suivantes : « Réussite » ou « Échec » et aussi, possiblement, « Poursuivra ses études » ou « À risque de décrocher ». Cela permettrait vraiment de tirer parti de tous les renseignements auxquels les cégeps ont accès qui pourraient permettre d'améliorer les résultats pour les élèves.

Finalement, la troisième manière dont ce processus pourrait être amélioré consiste à commencer à mesurer l'efficacité des méthodes de rétention utilisées par les différents cégeps. Pour poursuivre avec l'exemple précédent, parmi tous les courriels qui sont envoyés aux élèves à risque, combien font l'objet d'une réponse? De ceux qui ont répondu, combien d'entre eux prennent un rendez-vous avec un conseiller pour qu'ils puissent se faire aider? De ce nombre, combien font le suivi, se rendent au rendez-vous et apportent certaines modifications à leurs méthodes d'étude? Il est évident pour le lecteur que ce processus a une très faible chance de succès compte tenu du fait que les élèves reconnaissent rarement que la solution à leurs problèmes est externe à eux-mêmes. Étant donné la quantité de littérature sur la façon de modifier le comportement de recherche d'aide des élèves, ce problème est suffisamment vaste pour justifier son propre projet de recherche. L'argument évoqué ici est simplement qu'un cégep ne doit pas seulement mettre en place un moyen pour tenter d'aider les élèves et ensuite considérer que le problème est réglé.

Avec tous les renseignements à notre disposition sur les données démographiques, les résultats scolaires et leur évolution, les ÉMS de la population étudiante, nous pouvons désormais tirer parti des techniques modernes et mettre au point des modèles permettant de mieux identifier les élèves les plus susceptibles de décrocher. Au meilleur de notre connaissance, jamais les données intra-sessions n'avait été analysées auparavant car celles-ci n'existaient pas de façon uniforme dans tous les cégeps. Avec cet ensemble de données, qui comprend non seulement les renseignements session par session, mais aussi à mi-chemin dans une même session, nous mettons au point le prédicteur de décrochage scolaire le plus précis que la province de Québec n'ait jamais eu à sa disposition. Le fait d'avoir un point d'intervention au milieu de la session nous aide non seulement à mieux prédire le futur de cette session pour les étudiants, mais nous permet d'avoir une granularité jamais atteinte pour le problème du décrochage scolaire au cégep.

Modéliser la réussite scolaire et l'attrition au cégep

Les travaux précédents les plus importants dans notre contexte de réussite et persévérance (Jorgensen, Fichten et Havel, 2009) consistait à déterminer les facteurs les plus importants d'attrition dans un cégep anglophone de Montréal. Après avoir examiné les relevés de notes de plus de 40 000 élèves, les auteurs ont constaté que, de façon générale, le plus important prédicteur d'attrition était les notes obtenues au secondaire. L'effet du sexe était plus prononcé chez les élèves plus faibles, les taux de décrochage scolaire étant jusqu'à 10 % plus élevés chez les garçons que chez les filles lorsque leur moyenne générale au secondaire était inférieure à 80 %. Bien que ces travaux comprenaient des résultats de sondage, l'approche méthodologique suivie (Desjardins, Ahlburg et McCall, 1999), où l'analyse principale portait sur de grands ensembles de données, permet d'éviter de coûteux sondages (qui obtiennent souvent de faibles taux de réponse) et a seulement utilisé des variables qui sont conservées dans des bases de données scolaires. L'objectif de notre projet est de poursuivre ces travaux, mais avec plusieurs extensions. Certaines de nos modifications incluent :

- examen de l'attrition dans le contexte plus large de multiples établissements du réseau collégial,
- application d'un ensemble plus varié de modèles statistiques et de techniques d'apprentissage machine,
- usage de données longitudinales, à savoir : quel point de vue supplémentaire peut être obtenu en modélisant l'attrition en fonction d'observations récurrentes du rendement scolaire des élèves à chaque session?
- examen de questions importante sur l'« attrition » tel que suggéré par (Tinto, 1975) : comment pouvons-nous faire la distinction entre les désistements volontaires et les décrochages associés à l'échec scolaire? Comment prendre en considération les absences temporaires et les transferts de programme?

Dans ce document, nous allons décrire : l'ensemble de données; — les méthodes par lesquelles nous identifions les élèves comme étant à risque; — les distributions des élèves à risque par : — les indicateurs démographiques; — les indicateurs relatifs au dossier d'admission session par session.

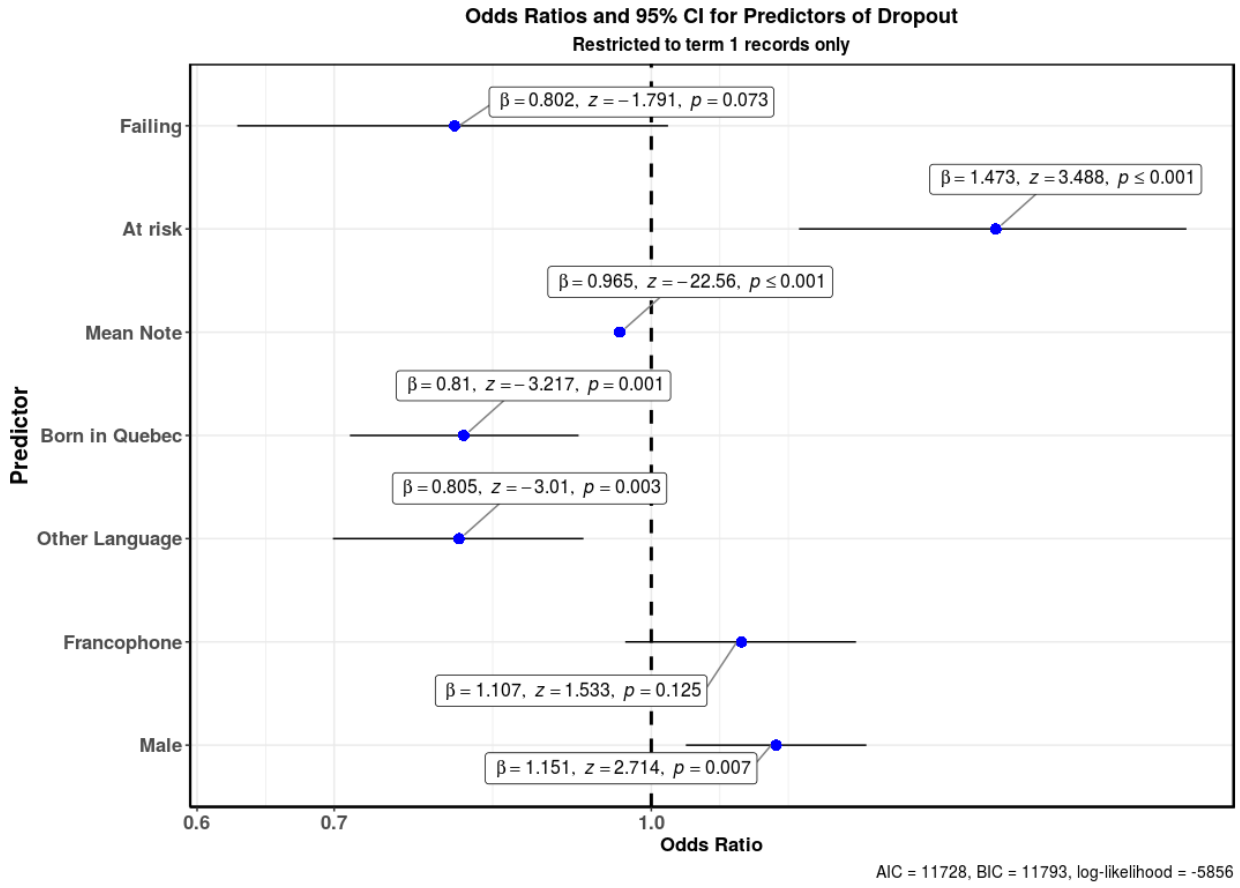
Il y a deux différences principales entre les régressions logistiques qui sont utilisées ici par rapport à celles déjà utilisées dans les études précédentes dans le système collégial : 1) la présence de résultats aux évaluations de mi-session et 2) l'approche longitudinale session par session pour prédire ce que l'élève fera au cours de la session suivante.

En d'autres termes, nous ne modélisons pas le résultat global de leur parcours au cégep, mais plutôt la prochaine étape qu'ils prendront dans leur parcours. Le haut niveau de granularité dans le temps (par session ou à l'intérieur d'une même session) qui est offert par les évaluations de mi-session nous permet d'examiner le problème des décrocheurs avec le plus puissant microscope utilisé pour examiner ce problème à ce jour.

Première session

Avec ce vaste ensemble de données, nous pourrions prédire les résultats en matière de décrochage à l'aide de la régression logistique. Les prédicteurs comprennent notamment le sexe, la langue, le lieu de naissance, le résultat à l'évaluation de mi-session (ÉMS) au cours d'une session donnée et la moyenne des résultats aux ÉMS au cours d'une session donnée.

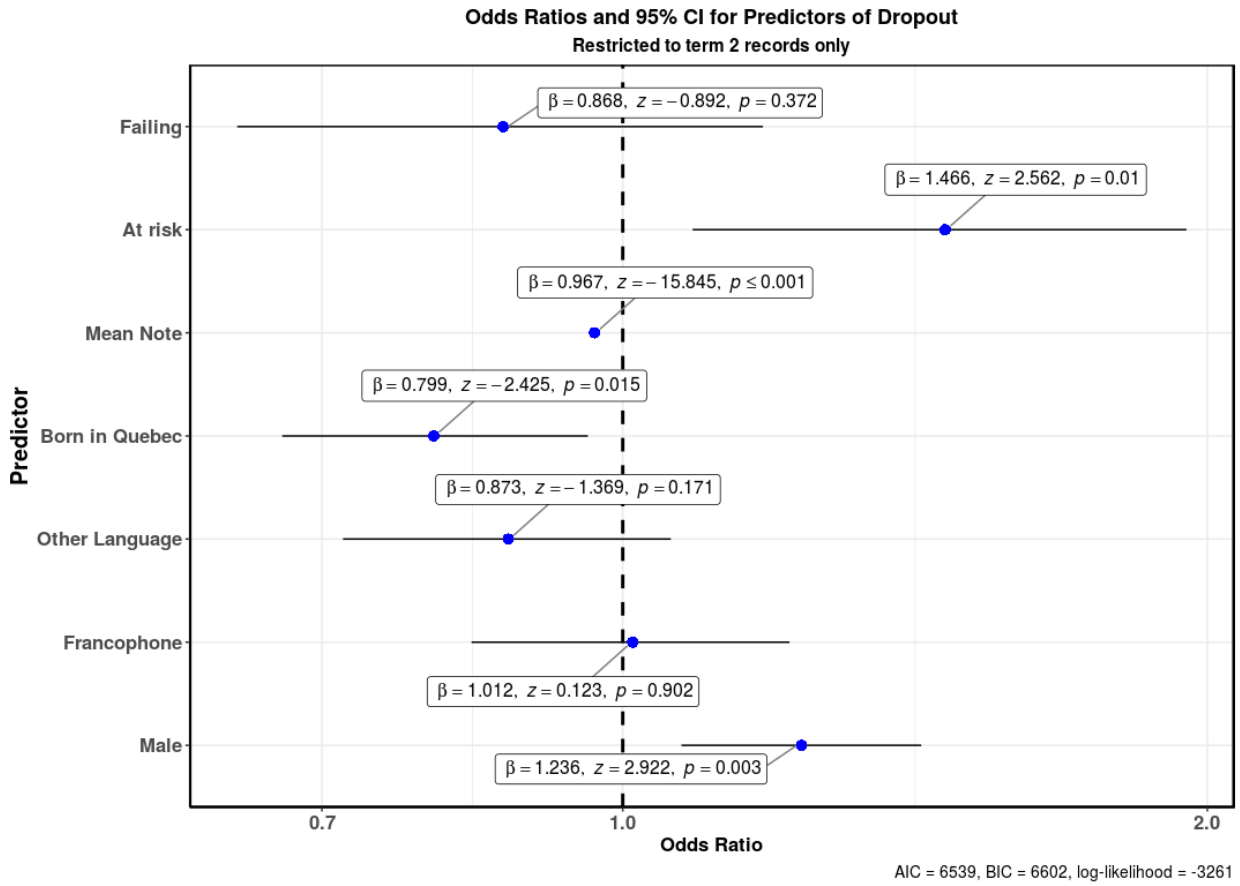
Les catégories de base sont « fille » pour le sexe, « anglophone » pour la langue, « née à l'extérieur du Québec » pour le lieu de naissance et « Réussite » pour le résultat à l'ÉMS. Le résultat consiste à déterminer si l'élève a décroché au cours de la session suivante. Les figures suivantes montrent les rapports de cotes et les intervalles de confiance de 95 % ainsi que les cotes Z et les valeurs de P qui ont été calculées à l'aide du test de Wald. Notons que dans les graphiques de forêt ci-dessous, une variable est statistiquement significative lorsque la barre d'erreur ne coupe pas l'unité.



La figure ci-dessus porte seulement sur les élèves de première session. Chaque élève passe une évaluation de mi-session par cours et session par session. Parmi les élèves de première session, les élèves chez qui la majorité des évaluations de mi-session affichent la mention « À risque » sont 47 % plus à risque de décrochage. Étonnamment, les élèves qui ont principalement échoué la majorité des ÉMS sont 20 % moins susceptibles de décrocher que les élèves qui ont réussi, bien que cette différence frôle le seuil significatif. C'est une observation très importante étant donné que la plupart des méthodes collégiales visant à identifier et aider les élèves à risque sont fondées sur les échecs aux ÉMS. Les élèves nés au Québec étaient et les allophones sont tous deux près de 20 % moins susceptibles de décrocher. Par ailleurs, les garçons avaient 15% plus de chances de décrocher que les filles, un résultat qui est bien connu et qui a été reproduit dans de nombreuses études faites sur les décrocheurs.

Ce qui est important de retenir de cette première analyse est que le profil des élèves qui sont à risque de décrocher au cours de la première session est principalement prédit par un paramètre qui, jusqu'à présent, n'a pas fait partie de ce type d'analyses, soit une ÉMS ayant la mention « À risque ». Ceci est un élément essentiel, car il nous fournit un signe très précoce et bien plus précis que le fait d'être un garçon ou de s'exprimer en français. De façon encore plus importante, ce paramètre change à chaque session et n'est pas fixe comme le sexe ou la langue maternelle. Examinons maintenant à la deuxième session pour voir si ces effets maintiennent leur puissance prédictive lors de la deuxième session.

Deuxième session

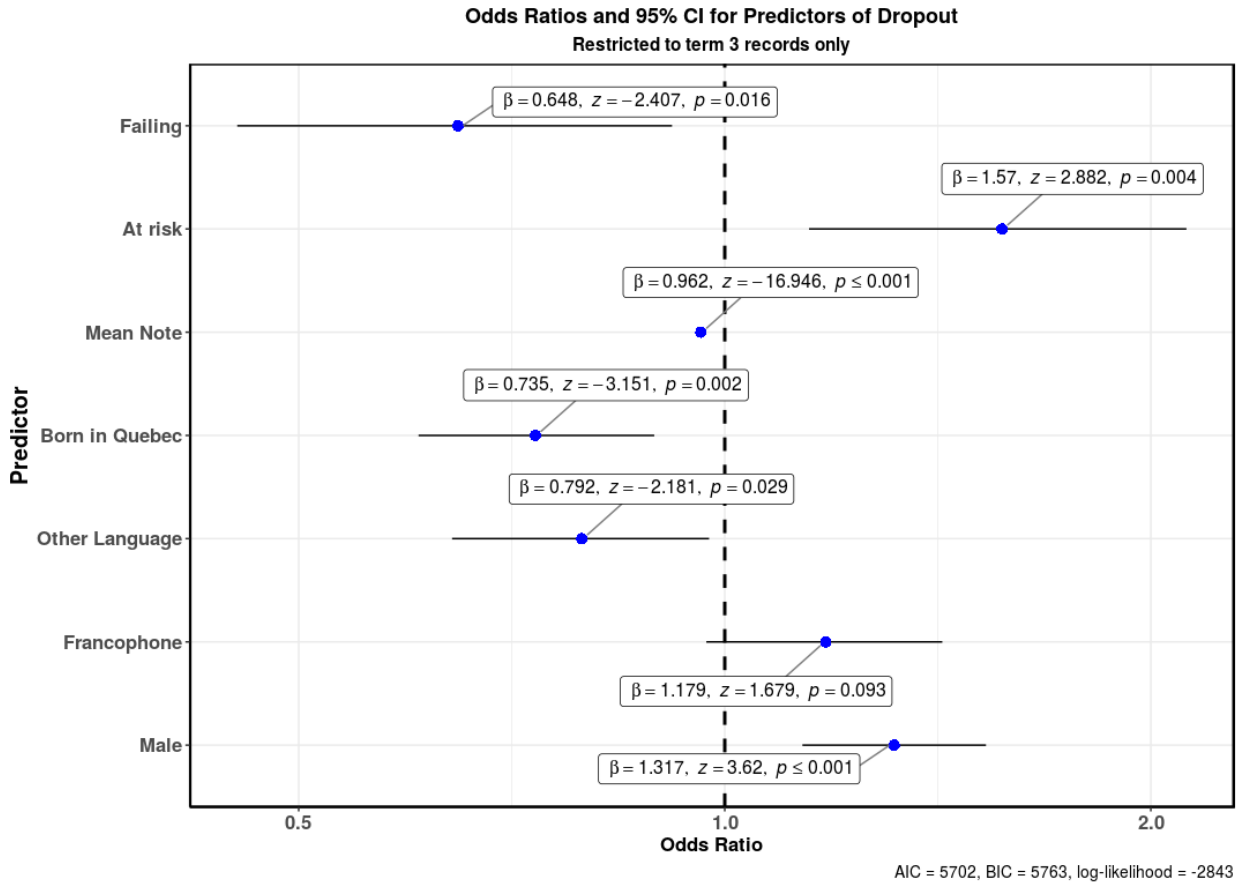


En regardant le graphique ci-dessus à la deuxième session spécifiquement, on peut déjà commencer à voir des différences dans la puissance des prédicteurs. Tout d'abord, seulment trois prédicteurs émergent de cette analyse comme étant statistiquement significative. Ces trois prédicteurs sont: la mention « À risque », être un garçon et être né au Québec. Tout d'abord, nous voyons que la force prédictive de la mention « À risque » aux ÉMS se maintient. En effet, les élèves ayant une mention « À risque » dans la ÉMS la plus récente sont 47 % plus susceptibles de décrocher que les élèves ayant réussi leurs ÉMS. C'est une autre preuve qui indique que les collèges ne doivent pas simplement s'intéresser aux ÉMS ayant obtenu la mention « Échec » puisque ces mentions d'échec sont encore moins susceptibles de prédire avec précision le décrochage des élèves. Le fait d'être un garçon demeure fortement associé au décrochage puisque les garçons sont 24 % plus susceptibles de décrocher que les filles. Ces deux prédicteurs continuent d'avoir une forte puissance prédictive même si la session pour laquelle nous tentons d'effectuer des prédictions a changé. De plus, comme pour la session précédente, être né au Québec procure 20 % de moins de chances de décrocher. Les autres variables telles que la langue maternelles ne demeurent pas des prédicteur faibles de décrochage à la deuxième session.

Étant donné que la rétention des élèves en troisième session est un facteur de mérite pour les cégeps, il devient très important de tenir compte des facteurs qui, au cours de la deuxième session, sont à l'origine du décrochage scolaire en troisième session. Nos données suggèrent que pour colmater la brèche qui est le décrochage des élèves, il faudrait mettre l'accent sur les garçons qui ont beaucoup d'ÉMS « À risque ».

Troisième session

Nous continuons cette analyse et regardons les mêmes variables à la troisième session pour voir si les mêmes modèles gardent leur validité.

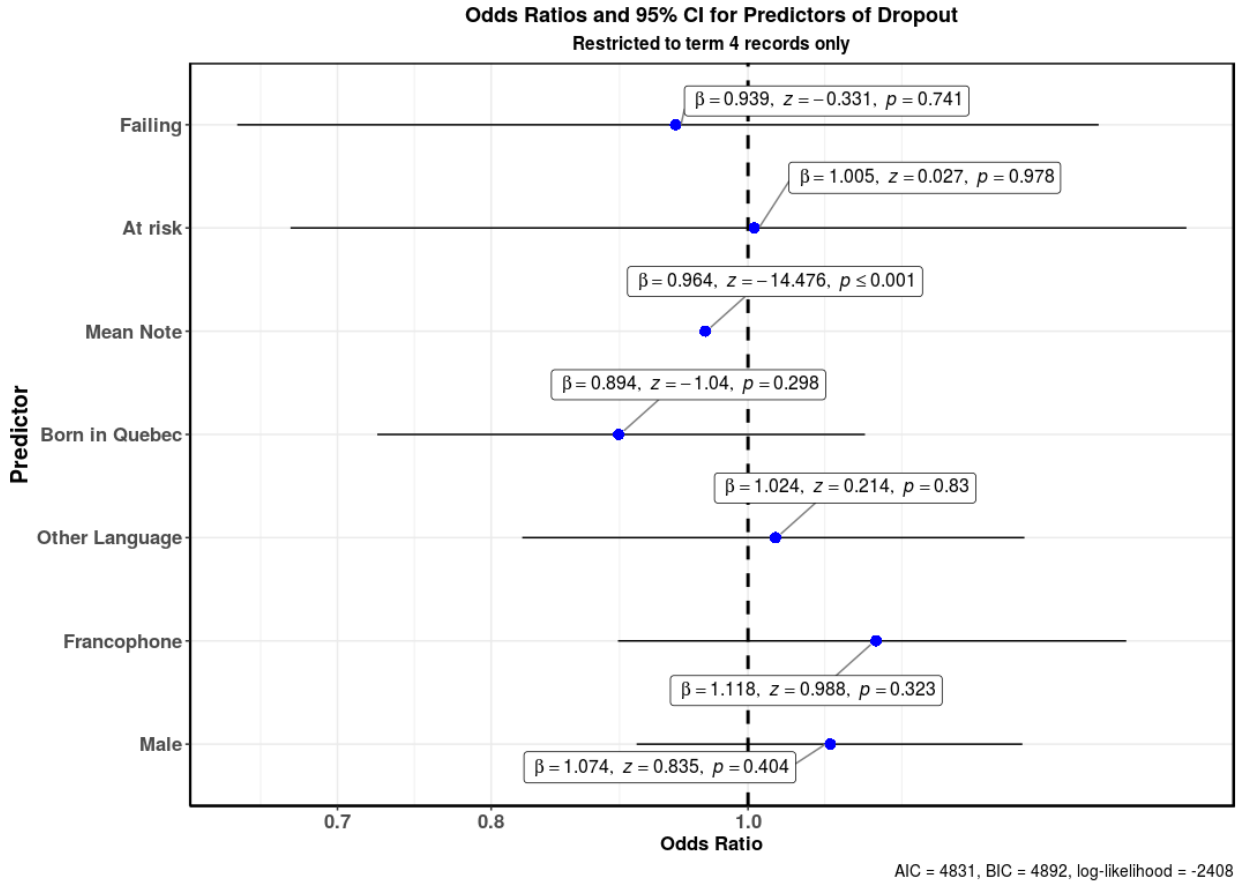


Le graphique ci-dessus montre qu'à part être francophone, toutes les variables dans ce modèle sont statistiquement significatives lors de la troisième session. Ce qui ne change pas par rapport à la session précédente, c'est que le fait d'obtenir une mention « À risque » aux ÉMS est le plus grand facteur prédicteur de décrochage avec 57 % plus de chances de décrocher à la 4e session qu'un élève qui réussit ses ÉMS. De plus, le fait d'être un garçon rajoute 32 % de plus de chances de décrocher à la 4e session que le fait d'être une fille. Les autres variables dont l'importance émergent à la troisième session incluent l'échec de la majorité des ÉMS. Notons encore que c'est la variable principale que les collèges utilisent pour prédire quels élèves doivent être aidés pour prévenir le décrochage. Cette variable est prédictive lors de la troisième session et confère 35 % de **moins** de chances de décrocher que l'élève ayant réussi la majorité de ses ÉMS. Il est possible que ce résultat est le fruit d'une intervention qui donne à l'étudiant un message le poussant à se reprendre pour ne pas échouer, mais nous n'avons pas de données concrètes pour supporter cette hypothèse. En revanche, les élèves ayant des mentions « À risque » aux ÉMS pourraient penser qu'ils n'échouent pas et négliger le message d'avertissement. Cependant, nous n'avons pas plus de données concrètes

pour supporter cette hypothèse. Finalement, être né au Québec se traduit par une diminution de 26 % de la probabilité de décrocher par rapport au fait d'être né à l'extérieur de la province tandis que d'être allophone se traduit par une diminution de 20 % de la probabilité de décrocher lors de la troisième session.

Quatrième session

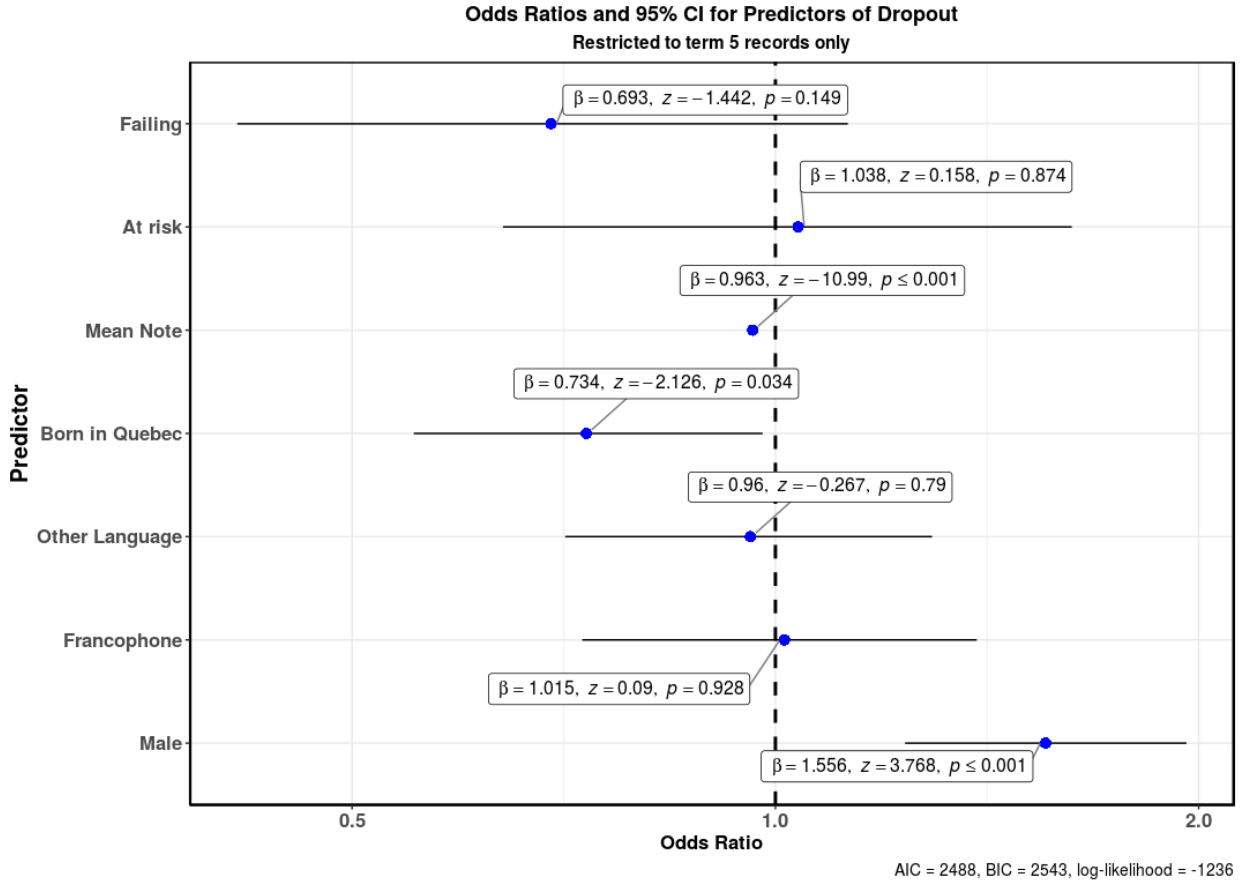
Nous allons maintenant, encore une fois, nous déplacer un peu plus loin dans le temps pour voir l'évolution des variables importantes et la puissance de ces variables comme facteurs prédictifs du décrochage scolaire.



Nous pouvons voir, dans ce graphique, qu'aucune variable ne prédit significativement le décrochage lors de la quatrième session puisque toutes les variables traversent l'unité.

Cinquième session

Voyons maintenant la dernière session pour laquelle nous allons poursuivre cette analyse puisqu'il y a de moins en moins d'élèves qui peuvent être modélisés à leur 6e session.



Une fois de plus, comme ce fut le cas pour l'analyse à la 4e session, aucune des variables analysées n'augmente le risque de décrocher sauf d'être un garçon (probabilité 56% plus élevée de décrocher). D'être né au Québec donne aussi une réduction modérée (27%) de la chance de décrocher lors du cinquième trimestre.

Points importants

Cette section nous permet d'être beaucoup mieux informés sur les types d'élèves qui décrochent et surtout sur les premiers signes qui peuvent être utilisés par les directions d'établissements pour identifier et aider les élèves à risque de décrocher.

Nos données illustrent clairement que, bien qu'il y ait des caractéristiques communes permettant d'identifier les élèves qui sont à risque de décrocher, de nombreux autres critères changent en

fonction de la session au cours de laquelle se trouve l'élève. Par conséquent, utiliser une série de critères fixes pour identifier les décrocheurs, qui constitue la façon principale de fonctionner de nombreux cégeps, nous amène à rater complètement la cible. **Les critères permettant d'identifier les élèves à risque de décrochage doivent être repensés et possiblement ajustés session par session.**

Dans l'éventualité où les cégeps ne pourraient pas utiliser de critères variables qui tiennent compte de la session d'étude, il convient de noter que, suite à cette analyse, le fait d'utiliser les ÉMS en « Échec » comme seul prédicteur définitif de décrochage est manifestement inadéquat. Seul la mention « À Risque » est un prédicteur stable de décrochage. En effet, la mention « Échec » soit ne prédit pas le décrochage ou bien agit en facteur protecteur. Une des raisons possibles pour expliquer ce résultat surprenant serait que les élèves qui obtiennent de nombreuses mentions « Échec » aux ÉMS réagissent à cette rétroaction et changent leurs façons de faire. Cela voudrait dire que l'ÉMS est une intervention qui marche. Pour que les collèges deviennent plus efficaces dans l'identification des étudiants susceptibles de décrocher, ils devraient se concentrer sur les élèves ayant obtenu des mentions « À risque » aux ÉMS plutôt que ceux ayant obtenu des mentions « Échecs ». Si les élèves qui sont en « Échec » aux ÉMS, ils sont moins susceptibles de décrocher que les élèves qui obtiennent la mention « Réussite » aux ÉMS. Est-ce que les élèves qui décrochent le font pour des raisons académiques? Si c'était le cas, on s'attendrait à ce que la mention « Échec » aux ÉMS soit le meilleur prédicteur du décrochage scolaire. Ce n'est pas le cas.

Possédant une méthode par laquelle il est possible de mieux prédire quels élèves décrocheront lors de la prochaine session, un test à faire consisterait à évaluer la façon dont cette nouvelle méthode se compare à la méthode traditionnelle utilisée actuellement par les cégeps.

Pour tout problème de prédiction, il est important de définir quel paramètre nous voulons optimiser. Comme cette étude vise à prédire les décrocheurs, le consensus est qu'on préférerait étiqueter par erreur un diplômé potentiel comme étant un décrocheur potentiel (et lui fournir des interventions pour prévenir le décrochage scolaire) plutôt que d'étiqueter par erreur un décrocheur comme étant un diplômé potentiel et ne lui fournir aucune intervention potentielle. Par ailleurs, il est essentiel de comprendre que le facteur de mérite qui sera utilisé par nos modèles est la sensibilité de la prédiction.

En d'autres termes, parmi tous les vrais décrocheurs, quel pourcentage réel d'entre eux ont été correctement prédits par nos modèles par rapport à la méthode actuelle utilisée par les cégeps.
Pour comparer l'efficacité des deux méthodes, voici les matrices de confusion comparant la méthode utilisée actuellement par les collèges (l'élève a deux mentions « Échec » ou plus aux ÉMS) à la séquence de prédicteurs issus de nos modèles de régression pour la première session (critères plus complexes).

Première session						
Nouvelle méthode	**Réel**			**Méthode actuelle**	**Réel**	
Prédit	Décroche	Poursuit ses études		**Prédit**	Décroche	Poursuit ses études
Décroche	1087	6516		Décroche	327	1477
Poursuit ses études	639	16 335		Poursuit ses études	1399	21 374

Il est intéressant de noter que notre modèle prédit trois fois mieux la probabilité de décrocher que les méthodes actuelles. Cependant, il y a un prix à payer, soit que nos modèles ne prédisent pas aussi bien qui poursuivra ses études. Il est évident qu'on ne s'inquiète moins d'identifier ceux qui continuent leurs études que d'identifier ceux qui les interrompent. Nos modèles montrent donc une grande augmentation de la sensibilité, même si c'est au dépend de la spécificité. En effet, notre méthode permet d'identifier correctement 63 % des élèves qui finiront par décrocher comparativement à 19 % avec la méthode actuellement utilisée par les cégeps. En utilisant cette nouvelle méthode, il est désormais possible de détecter trois fois plus de décrocheurs que l'ancien système.

Deuxième session						
Nouvelle méthode	**Réel**			**Méthode actuelle**	**Réel**	
Prédit	Décroche	Poursuit ses études		**Prédit**	Décroche	Poursuit ses études
Décroche	543	4832		Décroche	168	1354
Poursuit ses	295	14 024		Poursuit ses études	670	17 502

études						

Pour la deuxième session, les bonnes nouvelles se poursuivent, avec une sensibilité accrue de 65 % comparativement à 20 % par rapport à la méthode actuelle. Encore une fois, cette nouvelle méthode permet de tripler l'efficacité avec laquelle il est possible d'identifier les candidats probables au décrochage parmi le bassin d'élèves.

Troisième session						
Nouvelle méthode	**Réel**			**Méthode actuelle**	**Réel**	
Prédit	Décroche	Poursuit ses études		**Prédit**	Décroche	Poursuit ses études
Décroche	480	3731		Décroche	123	747
Poursuit ses études	300	10 434		Poursuit ses études	657	13 418

Pour la troisième session, la tendance se poursuit même si les pourcentages en soi descendent un peu. La sensibilité de la nouvelle méthode est de 62 % comparativement à 16 % avec la méthode actuelle, soit encore plus du triple d'efficacité.

Quatrième session						
Nouvelle méthode	**Réel**			**Méthode actuelle**	**Réel**	
Prédit	Décroche	Poursuit ses études		**Prédit**	Décroche	Poursuit ses études
Décroche	355	3623		Décroche	91	681
Poursuit ses études	261	10 319		Poursuit ses études	525	13 261

Pour la quatrième session, la sensibilité du nouveau modèle a atteint 58 % comparativement à 15 % avec la méthode actuelle, frôlant le 400% d'augmentation de sensibilité.

Cinquième session						
Nouvelle méthode	**Réel**			**Méthode actuelle**	**Réel**	
Prédit	Décroche	Poursuit ses études		**Prédit**	Décroche	Poursuit ses études
Décroche	179	1538		Décroche	45	388
Poursuit ses études	142	5585		Poursuit ses études	276	6735

Finalement, pour la cinquième session, la sensibilité du nouveau modèle a atteint 56 % comparativement à 14 % avec le modèle actuel, soit une sensibilité accrue de quatre fois ce qu'elle était avant notre étude.

Notons que les cégeps de partout dans la province peuvent utiliser les données qu'ils ont déjà à leur disposition (tous nos modèles sont basés sur les données que les cégeps ont en leur possession) pour créer des modèles qui permettent d'améliorer la détection des décrocheurs potentiels de plus de 300 %! **Dans une province où la lutte contre le problème de décrochage scolaire à tous les niveaux d'enseignement est considérée comme la plus grande priorité, il est surprenant qu'on ne s'appuie pas sur ces données pour mieux résoudre ce problème.**

CHAPITRE 7

APPRENTISSAGE MACHINE ET RÉSEAUX DE NEURONES

La réussite et la persévérance scolaires peuvent aussi faire l'objet d'une étude par le biais d'une autre stratégie d'exploration de données : l'apprentissage machine. Contrairement aux méthodes statistiques bien établies décrites dans les autres chapitres de ce rapport, l'apprentissage machine est une méthode non fondée sur des principes pour la découverte de modèles prédictifs. Plutôt que d'identifier les facteurs de causalité et, par la suite, de concevoir un modèle statistique qui tient compte de ces liens sous-jacents, l'apprentissage machine consiste à entraîner un système informatique afin qu'il puisse progressivement améliorer sa performance prédictive (c.-à-d., « apprendre ») sans être explicitement programmé.

Contexte

Les stratégies de mise en œuvre de l'apprentissage machine comportent souvent l'utilisation de réseaux de neurones artificiels. Ce sont des systèmes informatiques conçus pour reproduire approximativement la connectivité entre les neurones biologiques et leur activation non linéaire dans le cerveau des animaux. Les progrès récents dans le domaine des algorithmes et du matériel informatique ont suscité un regain d'intérêt envers les études sur les réseaux neuronaux et leurs applications dans l'apprentissage machine (van Gerven & Bohte, 2018).

Une brève introduction aux réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) sont composés d'unités intégrées appelées neurones artificiels (ou, par convention, juste neurones). Tout comme leur analogue biologique, ces unités génèrent une réponse non linéaire (biologiquement, un potentiel d'action) en fonction de la somme pondérée de ses entrées (synapses). Les poids des synapses et le seuil d'activation (biais) de la

réponse non linéaire sont des paramètres qui sont appris par l'entraînement (décrit ci-dessous). Par conséquent, la sortie y_i du i^e neurone est calculée comme étant

$$y_i = \sum_{k \in K_i} f(w_{ki} y_k + b_i)$$

où $f(x)$ est la fonction d'activation non linéaire du neurone, les w_{ki} représentent les poids des connexions (synapses) K_i agissant comme entrée du i^e neurone, et b_i est le seuil d'activation du neurone (biais).

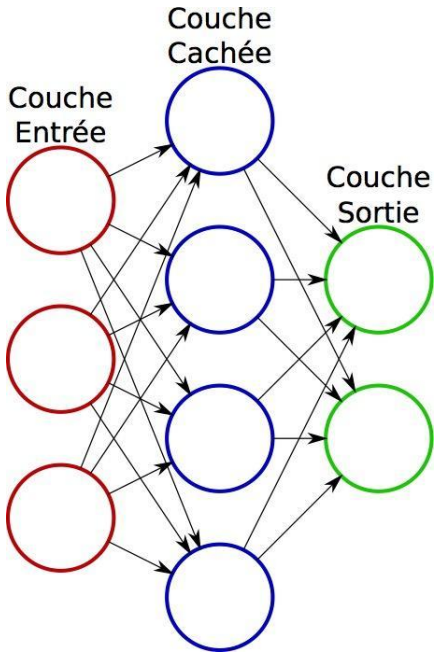


Figure 1 : Schéma d'un petit réseau de neurones à trois couches avec circulation d'informations en aval. Ce système prend un vecteur d'entrée de 3 unités de longueur et produit un vecteur de sortie de 2 unités de longueur (p. ex., la classification binaire). Les cercles représentent les neurones disposés en trois couches (entrée, cachée et sortie). Les flèches indiquent les connexions (synapses) entre les neurones. Chaque neurone comporte un paramètre de biais (seuil d'activation) et chaque synapse comporte un paramètre de poids. La sortie de ce réseau est une fonction fortement non linéaire du vecteur d'entrée paramétré par les 26 paramètres libres (20 poids synaptiques et 6 biais).

L'organisation en couches des neurones indépendants de façon à ce que les sorties d'une couche servent d'entrées pour la couche suivante (voir la Figure 1 pour un exemple simple d'une telle structure) se traduit par un système appelé réseau de neurones artificiels. Ces réseaux intègrent généralement plusieurs neurones et, par conséquent, un très grand nombre de connexions. Ils sont

donc en mesure de saisir les relations fortement non linéaires entre les entrées et les sorties. Il existe un très grand nombre de paramètres ajustables dans ces systèmes : chaque neurone comporte un paramètre de biais et chaque synapse comporte un paramètre de poids. Par exemple, le très simple RNA décrit par Figure 1 comporte 26 paramètres libres : 20 paramètres de poids (12 entre la couche d'entrée et la couche cachée + 8 entre la couche cachée et la couche de sortie) et 6 paramètres de biais (en lien avec les neurones de la couche cachée et la couche de sortie; aucun biais présumé pour la couche d'entrée puisqu'elle est censée servir simplement d'entrée au RNA). Ce sont ces paramètres qui sont ajustés au cours de l'entraînement de façon à ce que la sortie du réseau reflète plus étroitement les résultats correspondants de l'ensemble d'entraînement. En effet, ce sont ces paramètres qui encodent les fonctions apprises.

Choisir différentes topologies pour les connexions entre les neurones permet le traitement de différents types de données (p. ex., retrouver des objets identifiables dans les images, prédire le prochain mot d'une phrase, etc.).

Entraînement des RNA

L'entraînement supervisé d'un RNA pour qu'il effectue une tâche (p. ex., la catégorisation) consiste à fournir un ensemble d'entrées appariées et leurs sorties correspondantes attendues. Le RNA est généralement initialisé par l'attribution de valeurs aléatoires à chacun des paramètres de poids et de biais. Les entrées sont introduites dans le réseau et les sorties sont comparées aux sorties d'entraînement correspondantes, la différence (erreur) étant quantifiée par une fonction de perte. Par la suite, en mode entraînement, le RNA est un système qui cartographie les entrées multidimensionnelles pour obtenir une seule valeur réelle de perte de fonction. En traitant les données d'apprentissage comme des constantes connues, le système peut être considéré comme une cartographie de la configuration des poids et des biais (très haute dimensionnalité) et de la valeur de perte simple.

L'entraînement du RNA consiste ensuite à trouver des valeurs pour les poids et les biais qui réduisent au minimum la perte sur l'ensemble des exemples servant à l'entraînement; cela peut être considéré comme un problème d'optimisation globale de la fonction de perte par rapport au caractère

hautement dimensionnel des poids (souvent de l'ordre de 10^{-5} ou 10^{6}). Cette optimisation s'effectue habituellement à l'aide de méthodes semblables à la méthode de Newton : en utilisant le gradient de la fonction de perte relatif à l'espace des paramètres pour calculer les corrections optimales à appliquer aux valeurs des paramètres. Pour des raisons d'efficacité de calcul, ceci est généralement réalisé à l'aide d'un processus connu sous le nom de rétropropagation (Rumelhart, Hinton et Williams, 1986) dans lequel le gradient est calculé de manière inversée, couche par couche (c'est-à-dire, en commençant par la couche de sortie et en travaillant en sens inverse du réseau vers la couche d'entrée). De cette façon, les gradients calculés aux niveaux supérieurs (les couches les plus proches de la sortie) peuvent être réutilisés dans le calcul des gradients au niveau inférieur au moyen de la règle de la chaîne de calcul.

Une fois l'erreur minimisée sur l'ensemble d'entraînement, le RNA est considéré comme entraîné. À ce stade, il peut être utilisé pour faire des prédictions en fournissant de nouvelles entrées et en enregistrant leur sortie correspondante.

Difficultés associées aux RNA dans la pratique

La description ci-dessus n'est qu'une très courte introduction aux RNA et omettait donc de nombreux détails importants qui doivent être pris en considération lors de l'utilisation de ces systèmes dans la pratique. Parmi ceux-ci, notons :

- Quels types de topologies de réseau sont nécessaires pour refléter la structure sous-jacente aux données que nous souhaitons analyser?
- Comment peut-on s'assurer de la généralisabilité d'un RNA (c.-à-d., comment peut-on éviter le surapprentissage des paramètres réseau envers les données d'entraînement)?
- Existe-t-il des moyens de représenter les données pour mettre en évidence des relations significatives qui peuvent améliorer la convergence et la performance du RNA?
- Comment un très grand RNA peut-il être préentraîné de manière à ce que la convergence se produise dans un délai raisonnable?

Ces questions seront abordées dans le chapitre suivant, où nous concevons un RNA conçu pour signaler à l'utilisateur les élèves qui peuvent être à risque d'échouer une partie importante de leurs cours ou d'abandonner leurs études.

Références

van Gerven, M., & Bohte, S. (2018). Artificial neural networks as models of neural information processing: Frontiers Media.

SA Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533.

CHAPITRE 8

EXAMEN DE LA RÉUSSITE ET LA PERSÉVÉRANCE

PAR RÉSEAUX DE NEURONES

L'objectif de cette phase du projet était de concevoir un système d'apprentissage machine qui pouvait prévoir les problèmes associés à la réussite et la persévérance scolaires afin que les services de soutien (p. ex., le conseiller scolaire, le conseiller en orientation) puissent être avisés et peut-être intervenir pour aider les élèves ainsi identifiés. Dans ce contexte, les objectifs de cette phase étaient les suivants :

- Développer un RNA permettant d'identifier les élèves à risque d'échouer une partie substantielle des cours collégiaux auxquels ils sont inscrits au cours d'une session donnée.
- Développer un RNA permettant d'identifier les élèves à risque d'abandonner leurs études collégiales au cours d'une session donnée.

Les données à notre disposition sont les mêmes dossiers historiques obtenus des trois cégeps participant à l'étude (Collège Dawson, Collège Vanier et Collège John Abbott) qui ont été utilisés dans les autres sections de ce rapport. Ces données comprenaient les dossiers des cours (p. ex., les cours suivis, les notes obtenues) provenant de l'école secondaire des élèves et leur cheminement au cégep, les dossiers de la session (p. ex., le programme d'études, le statut à temps plein ou à temps partiel), ainsi que les dossiers des élèves (p. ex., leur sexe, leur langue maternelle, leur âge).

En tenant compte de ces données et des objectifs décrits ci-dessus, nous avons procédé à la conception d'un RNA permettant de prédire (a) la partie des cours qu'un étudiant devrait terminer avec succès et (b) la probabilité d'obtenir un DEC au cours de leur dernière session d'études.

1. Le RNA est-il un moyen viable de prédire quels sont les élèves à risque d'échouer une partie substantielle des cours collégiaux auxquels ils sont inscrits, au cours d'une session donnée,

en tenant compte de leurs antécédents scolaires, des cours auxquels ils sont inscrits actuellement et d'un minimum de renseignements démographiques?

   a. Comment la notion de « à risque d'échouer une partie substantielle de ses cours collégiaux » doit-elle être définie afin de mieux identifier les élèves à risque?

2. Le RNA est-il un moyen viable de prédire quels sont les élèves à risque d'abandonner leurs études collégiales, au cours d'une session donnée, en tenant compte de leurs antécédents scolaires, des cours auxquels ils sont inscrits actuellement et d'un minimum de renseignements démographiques?

   a. Comment la notion de « à risque d'abandonner ses études collégiales » doit-elle être définie afin de mieux identifier les élèves à risque?

Nous abordons ces questions par la conception d'un RNA multiniveaux qui peut intégrer les renseignements relatifs au cours, à la session et à l'élève en entraînant, validant et testant le modèle à l'aide des données historiques.

Initialement, nous avions prévu effectuer un « essai en direct » de la puissance prédictive du RNA, c'est-à-dire, de tenter d'identifier les élèves à risque qui sont inscrits présentement au cégep au cours d'une session donnée. Malheureusement, des retards imprévus dans l'obtention des données (l'obtention de l'approbation et la réalisation de l'extraction des données ont tous deux pris beaucoup plus longtemps que prévu), des difficultés rencontrées dans l'analyse et l'interprétation des données (des incohérences dans les dossiers qu'il fallait réparer manuellement), ainsi que des difficultés techniques imprévues (ensembles de données trop grands pour tenir dans la mémoire vive de l'ordinateur) nous ont empêchés de terminer la phase finale de ces travaux avant la fin de la session d'hiver 2018. Le modèle, cependant, a été mis au point et ce test pourrait possiblement être réalisé à l'automne 2018.

Les données

Compte tenu du vaste ensemble de données à notre disposition, il convenait de limiter notre analyse aux élèves qui avaient un dossier complet (c.-à-d., tous les cours avaient une note ou une remarque autre que « En cours ») et qui étaient inscrits à au moins une session d'études dans l'un des trois

cégeps participants. Une fois cette restriction en place, notre ensemble de données comportait 122 824 dossiers scolaires au total. Ceux-ci ont été répartis dans les groupes suivants :

- 61 412 dossiers scolaires (50 %) ont été utilisés pour l'entraînement du RNA;
- 30 706 dossiers (25 %) ont été retenus pour la validation du modèle (c.-à-d., sélectionner le meilleur modèle, déterminer le moment où le modèle était entraîné de façon optimale et déterminer les valeurs optimales des hyperparamètres);
- 30 706 dossiers (25 %) ont été utilisés pour tester le modèle et rédiger un rapport.

Structure des données

N'importe quel système qui cherche à tirer des conclusions à partir d'un vaste ensemble de données doit tenir compte de la structure des données. Comme indiqué ci-dessus, les données représentent les renseignements à trois niveaux différents : des cours, de la session et de l'élève.

Champs associés aux cours

Pour chaque élève, les cours suivis et les résultats correspondants ont été sauvegardés, session par session, tant en ce qui concerne leur dossier à l'école secondaire qu'au cégep. Le Tableau 1 et le Tableau 2 présentent la provenance de ces données.

Quantité	Champ CLARA	Type de données brutes	Type de données réencodées
Identifiant du cours	Étudiants ÉtudiantCoursSecondaire. IDCoursSecondaire.	Nombre entier	Vecteur 14-D[3]
Note	Étudiants ÉtudiantCoursSecondaire. Résultat	Flottante	Flottante
Remarque	Étudiants. ÉtudiantCoursSecondaire. CodeRemarque	Catégorielle	One-hot 11-D[4]
Moyenne de groupe	Étudiants. ÉtudiantCoursSecondaire. MoyenneGroupe	Flottante	Flottante
# étudiant	Étudiants. ÉtudeAntérieure. IDÉcoleSecondaireX (X =1.. 5)[5]	Texte	Vecteur 13-D[1]

[3] Voir la section Les cartes d'intégration pour une description de la manière dont ces vecteurs ont été générés.

[4] Des remarques ont été réencodées pour unifier les codes de remarques du secondaire et du collégial. Onze catégories de remarques de ce type ont été trouvées et représentées comme des encodages « one-hot » à 11 dimensions (c.-à-d., un vecteur dont toutes les valeurs sont nulles sauf une).

[5] Dans notre ensemble de données, ceux-ci ont été sauvegardés en utilisant le nom de leur école secondaire. En fin de compte, il semble que ces champs étaient associés à un identifiant comme IDÉcoleSecondaire5.

Tableau 1 : Données des cours (dossier de l'école secondaire)

Quantité	Champ CLARA	Type de données brutes	Type de données réencodées
Identifiant du cours	Inscriptions. Inscription. IDGroupe	Nombre entier	Vecteur 14-D[1]
Note	Inscriptions. Inscription. Note	Flottante	Flottante
Remarque	Inscriptions. Inscription. CodeRemarque	Catégorielle	Encodage one-hot 11-D[2]
Moyenne de groupe	Inscriptions. Inscription. CodeRemarque	Flottante	Flottante
Numéro étudiant	(Nom du cégep)	Texte	Vecteur 13-D[1]

Tableau 2 : Données des cours (dossier du cégep)

Pour chaque cours (c.-à-d., pour chaque cours suivi par chaque élève dans chaque session de leur dossier), un vecteur de cours 40-D a été créé à l'aide de ces données :

- 14-D pour l'identification du cours
- 1-D pour la note obtenue
- 11-D pour la remarque
- 1-D pour la moyenne de groupe
- 13-D pour l'identification de l'école

Les données des cours de chaque élève consistaient ensuite en une série de ces vecteurs descriptifs des cours en 40 dimensions. De l'espace était alloué pour permettre jusqu'à 30 dossiers de cours de ce type pour chaque session associée à un dossier scolaire.

Il est important de remarquer que ces vecteurs n'ont pas d'ordre naturel à l'intérieur d'une session donnée. [3] Dans notre ensemble de données, ceux-ci ont été sauvegardés en utilisant le nom de leur école secondaire. En fin de compte, il semble que ces champs étaient associés à un identifiant comme IDÉcoleSecondaire5.

Voir la section Les cartes d'intégration pour une description de la manière dont ces vecteurs ont été générés.

[2] Des remarques ont été réencodées pour unifier les codes de remarques du secondaire et du collégial. Onze catégories de remarques de ce type ont été trouvées et représentées comme des encodages « one-hot » à 11 dimensions (c.-à-d., un vecteur dont toutes les valeurs sont nulles sauf une).

Champs associés à la session

En plus des dossiers cours par cours décrits ci-dessus, chaque session du dossier scolaire de l'élève comportait aussi des variables de contexte pertinentes (Tableau 3 et Tableau 4).

Quantité	Champ CLARA	Type de données brutes	Type de données réencodées
Session	Étudiants. ÉtudiantCoursSecondaire. Année + Étudiants. ÉtudiantCoursSecondaire. Session	Nombre entier	Nombre entier [6,7]
Âge	(calculé à partir de la date de naissance de l'élève et de la session en cours)	Nombre entier	Nombre entier[8]

Tableau 3 : Données de la session (dossier de l'école secondaire)

Quantité	Champ CLARA	Type de données brutes	Type de données réencodées
Session	Étudiants. ÉtudiantSession. AnSession	Nombre entier	Nombre entier[5]
Âge	(calculé à partir de la date de naissance de l'élève et de la session en cours)	Nombre entier	Nombre entier[6]
SPE	Étudiants.ÉtudiantSession.SPE	Nombre entier	Nombre entier
Programme	Étudiants. ÉtudiantSession. IDProgramme	Nombre entier	Vecteur 13-D[1]
Statut	Étudiants. ÉtudiantSession. TypeFréquentation	Catégorielle	One-hot 5-D

Tableau 4 : Données de la session (dossier du cégep)

[6] Les dossiers scolaires comportent le mois où la note a été inscrite; ceux-ci ont été convertis sous une forme année-session (p. ex., 20 181 pour l'hiver 2018).
[7] Référencé à la première session d'inscription au cégep (le signe négatif indique avant l'inscription au cégep).
[8] Exprimé en nombre de sessions depuis la naissance.

Ces données de la session forment un ensemble vectoriel distinct à 21 dimensions pour chaque session du dossier scolaire d'un élève. Puisque les SPE et les renseignements sur le programme ne sont pas applicables à l'école secondaire, ces champs ont été saisis comme nuls pour les sessions ayant précédé le début du cégep.

Champs associés à l'élève

Enfin, certaines des données (Tableau 5) intégrées étaient associées à un élève donné plutôt que session par session ou cours par cours. Celles-ci comprennent la date et le lieu de naissance de l'élève, sa langue maternelle, son code postal, son sexe et ses études antérieures. Le vecteur descriptif de l'élève constitue alors l'entrée en 41 dimensions associée à chaque élève.

Quantité	Champ CLARA	Type de données brutes	Type de données réencodées
Date de naissance	Étudiants. Étudiant. DateNaissance	Date	Nombre entier[5]
Sexe	Étudiants. Étudiant. Sexe	Catégorielle	One-hot 3-D
Langue maternelle	Étudiants. Étudiant. LangueMaternelle	Catégorielle	One-hot 4-D
Lieu de naissance	Étudiants. Étudiant. CodeLieuNaissance	Texte	Vecteur 6-D[1]
Code postal	Coordonnées.Adresse.CodePostal	Texte	Vecteur 5-D[1]
Déficiences majeures	Étudiants. Étudiant. IndicateurDéficienceFonctionnelleMajeure	Nombre entier	Nombre entier
Cote R	Étudiants. Étudiant. CoteR	Flottante	Flottante
Études secondaires antérieures	Étudiants. ÉtudeAntérieure. ÉtatÉtudeSecondaire	Catégorielle	One-hot 5-D
Études secondaires antérieures aux adultes	Étudiants. ÉtudeAntérieure. ÉtatÉtudeSecondaireAdulte	Catégorielle	One-hot 5-D

Études collégiales antérieures	Étudiants. ÉtudeAntérieure. ÉtatÉtudeCollégial	Catégorielle	One-hot 5-D
Études universitaires antérieures	Étudiants. ÉtudeAntérieure. ÉtatÉtudeUniversitaire	Catégorielle	One-hot 5-D

Tableau 5 : Données de l'élève

Les cartes d'intégration

La représentation significative de variables catégorielles de grande dimension peut être améliorée par des intégrations préentraînement, une méthode souvent utilisée pour représenter les mots et expressions pour le traitement du langage naturel (Socher, 2014). Plutôt que de représenter une variable catégorielle comme un vecteur « one-hot » (c.-à-d., un vecteur binaire indiquant la catégorie sélectionnée comme étant 1 dans un vecteur autrement nul), les vecteurs intégrés représentent chaque valeur possible de la variable catégorielle comme un point dans un espace abstrait. Plutôt que des représentations mutuellement orthogonales et équidistantes du vecteur one-hot, les relations entre les variables catégorielles peuvent être trouvées et encodées en donnant une mesure significative de « proximité ».

Par exemple, 16 687 cours différents sont répertoriés dans l'ensemble de données. Les représenter comme des vecteurs one-hot se traduirait par des vecteurs très clairsemés à 16 687 dimensions. De plus, des cours très connexes (p. ex., Calcul différentiel et intégral I et II) ne sembleraient pas plus proches que toute autre paire (p. ex., Calcul différentiel et intégral I et un cours en sciences humaines). Toutes les relations significatives entre les cours devraient alors être déduites par le réseau de neurones.

En revanche, les intégrations peuvent être utilisées pour trouver ou définir une impression de proximité qui peut fournir des indices importants aux couches d'apprentissage machine. Dans le cas du traitement des langues naturelles, les intégrations de mots préentraînement sont généralement apprises en examinant les cooccurrences de mots à l'intérieur d'un vaste corpus textuel : le sens du mot est appris « par les liens qu'il conserve » (Socher, 2014).

Par analogie, nous avons déterminé qu'une intégration efficace des cours pourrait être apprise en examinant les séquences des cours suivis par chaque élève. De plus, d'autres informations comme les écoles où les cours ont été suivis et le programme d'étude pourraient aussi être intégrées de façon significative. En traitant les séquences de cours, les écoles et les programmes comme s'ils étaient des phrases dans un vaste corpus textuel, des routines existantes pour générer des intégrations mot-vecteur pourraient être utilisées pour créer des intégrations cours-vecteur, école-vecteur et programme-vecteur significatives dans un espace abstrait.

À cette fin, nous avons créé une séquence de numéros de cours, d'écoles et de programmes d'études pour chaque dossier scolaire de nos données et nous avons utilisé une routine d'intégration des mots appelée GloVe (Pennington, 2014) pour produire un ensemble conjoint de vecteurs d'intégration à 15 dimensions. Nous avons ensuite séparé les intégrations de cours, d'écoles et de programmes et effectué séparément une analyse en composantes principales (ACP) pour réduire la dimensionnalité. En utilisant un seuil de variance de 95 %, nous avons conservé 14 dimensions pour les vecteurs de cours, 13 dimensions pour les vecteurs d'écoles et 12 dimensions pour les vecteurs de programmes.

Les cartes d'intégration qui en ont découlé représentent une initialisation utile de la représentation de ces variables catégorielles, mais elles ne sont pas alignées vers l'objectif ultime de prédire la réussite et la persévérance scolaires. Chacune des cartes d'intégration avait donc la possibilité d'évoluer lors de l'entraînement du réseau neuronal (c.-à-d., traiter chacune des variables simplement comme un autre élément de poids à optimiser au cours de l'entraînement).

Les cartes d'intégration ont aussi été utilisées pour encoder le code postal et le pays de naissance de l'élève. Contrairement aux cours, aux écoles et aux programmes d'études, aucune séquence naturelle n'était présente dans les données desquelles on devait tirer une représentation utile. Pour fournir une représentation significative, des cartographies ad hoc ont été créées :

Codes postaux (trois premiers caractères) :

- Latitude et longitude centroïdes (normalisées) pour saisir l'emplacement géographique;

- Trois dimensions initialisées aléatoirement pour offrir un espace pour apprendre d'autres caractéristiques (p. ex., liées à la situation socio-économique).

Pays de naissance :

- Latitude et longitude (normalisées) de la capitale du pays pour saisir l'emplacement géographique;
- PIB nominal par habitant (normalisé);
- Indice de développement humain (normalisé);
- Trois dimensions initialisées aléatoirement pour offrir un espace pour apprendre d'autres caractéristiques.

Nous prévoyons que, si l'une ou l'autre de ces dimensions est importante pour prédire la réussite et la persévérance scolaires, le RNA sera en mesure d'établir des relations significatives entre les dimensions libres initialisées aléatoirement.

Conception de réseaux

Le traitement des données à différents niveaux (des cours, de la session et de l'élève) a nécessité une planification minutieuse de la topologie du réseau. Simplement déverser des données à titre d'entrées non structurées serait peu susceptible de révéler des caractéristiques utiles. Nous avons choisi de concevoir un réseau multiniveau dans lequel les données de différents niveaux étaient traitées puis fusionnées pour servir d'entrées à deux classificateurs distincts, l'un pour prédire la réussite scolaire et l'autre pour prédire la rétention des élèves.

Le réseau a donc été conçu en quatre sections différentes :

- Une section qui analysait les données des cours pour une même session, ce qui permettait d'obtenir un seul vecteur pour chaque session de chaque dossier scolaire qui était ensuite concaténé avec les données de la session correspondante, ce qui permettait d'obtenir un vecteur descriptif par session.
- Une section qui analysait la séquence des vecteurs descriptifs par session de chaque élève pour générer un vecteur relatif aux antécédents scolaires.

- Une section qui concaténait le vecteur relatif aux antécédents scolaires de l'élève avec les données de l'élève pour générer un vecteur descriptif de l'élève.
- Une section classification qui apprenait à prédire les mesures associées à la réussite et la persévérance scolaires.

Chacune de ces sections traitait des données de structure et de dimensionnalité différentes, comme décrit ci-dessous.

Analyse associée aux cours

Chaque dossier de cours de chaque élève comportait l'équivalent de 30 sessions (10 ans) de données, soit jusqu'à 30 vecteurs de cours ayant chacun 40 dimensions. En raison de l'utilisation prévue de ces données, la dernière session du dossier était traitée différemment.

Les données ont donc été organisées comme suit :

- Un tenseur de dimensionnalité (N_échantillons, 29, 30, 40) pour les données historiques (antécédents scolaires) de l'élève (pouvant équivaloir jusqu'à 29 sessions). Ce tenseur encode l'historique des cours suivis par l'élève et leurs résultats. Nous l'appelons le tenseur d'antécédents scolaires de l'élève.
- Un tenseur de dimensionnalité (N_échantillons, 1, 30, 40) décrivant la dernière session de l'élève, les cours suivis et les résultats obtenus, comme indiqué ci-dessus. Ce tenseur est utilisé comme cible dans le processus d'entraînement, c'est pourquoi nous l'appelons le tenseur cible de la dernière session.
- Un tenseur de dimensionnalité (N_échantillons, 1, 30, 40) donnant une description modifiée de la dernière session de l'élève. Ici, le résultat (note et moyenne de groupe) est supprimé et la catégorie de remarque est changée pour « en cours ». Nous appelons ce tenseur modifié de la dernière session l'entrée de la dernière session. Il sert à simuler les renseignements qui seraient disponibles au début de la session : les cours choisis, mais aucune information sur leurs résultats. Ce tenseur est utilisé comme entrée pour le réseau de classification, c'est pourquoi nous l'appelons le tenseur d'entrée de la dernière session.

L'objectif était de produire une représentation significative de la session d'un élève encodée comme un vecteur unique. Étant donné la nature non ordonnée des cours pris à l'intérieur d'une session donnée, il importait de concevoir cette section du réseau pour qu'elle ne tienne pas compte de l'ordre dans lequel les dossiers de cours apparaissaient (cela signifie que les dossiers de deux sessions, identiques à l'exception de l'ordre dans lequel les cours apparaissaient, devraient produire des vecteurs de sortie identiques).

Pour ce faire, une architecture de réseau neuronal *à convolution* (Bengio, 2009), suivie d'une sélection de valeurs, a été utilisée. Les réseaux à convolution, inspirés par la structure du système visuel des animaux, consistent à appliquer le même réseau sur différentes parties d'une image (Figure 2).

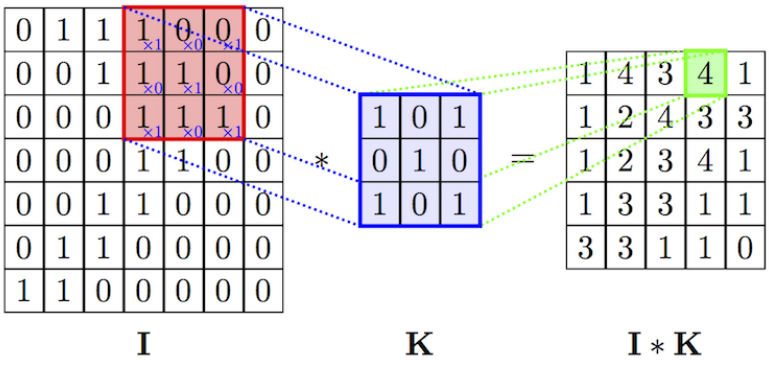


Figure 2 : Exemple d'un simple noyau de convolution 3 ✕3 appliqué à une image 7 ✕7. Chaque cellule de la couche de sortie est calculée en appliquant le même noyau de convolution à différents ensembles de cellules de la couche d'entrée.

Le fait d'appliquer le même noyau de convolution à chaque sous-section de la couche d'entrée mène à une invariance translationnelle inhérente de ces réseaux. Nous utilisons cette fonction pour générer un vecteur descriptif par session en appliquant un certain nombre de ces noyaux de convolution à chacune des 30 descriptions de cours dans une session donnée et en trouvant l'activation moyenne de chaque filtre sur ces 30 cours (un processus connu sous le nom de sélection des valeurs moyennes). Ceci a fourni le mécanisme ayant permis la conversion des matrices 30 ✕40 de données des cours en des informations d'encodage à vecteurs uniques sur l'ensemble de la session (voir Figure 3).

Des essais et erreurs et l'intuition nous ont conduits à choisir 256 filtres de ce genre, ce qui signifie que la sortie de cette étape du RNA est un vecteur de 256 unités de longueur. Les poids et les biais des filtres sont, comme les autres poids et biais du réseau, « appris » durant le processus d'entraînement pour fournir une représentation qui est utile à la tâche finale de classification. Cela signifie que les noyaux de convolution évoluent pour devenir des détecteurs de caractéristiques utiles pendant l'entraînement.

Chacun des trois tenseurs associés aux cours — le tenseur d'antécédents scolaires de l'élève, le tenseur de cible de la dernière session et le tenseur d'entrée de la dernière session — a été traité par ce même filtre convolué et cette même couche de sélection des valeurs, donnant respectivement lieu au vecteur d'antécédents scolaires de l'élève, au vecteur de cible de la dernière session et au vecteur d'entrée de la dernière session.

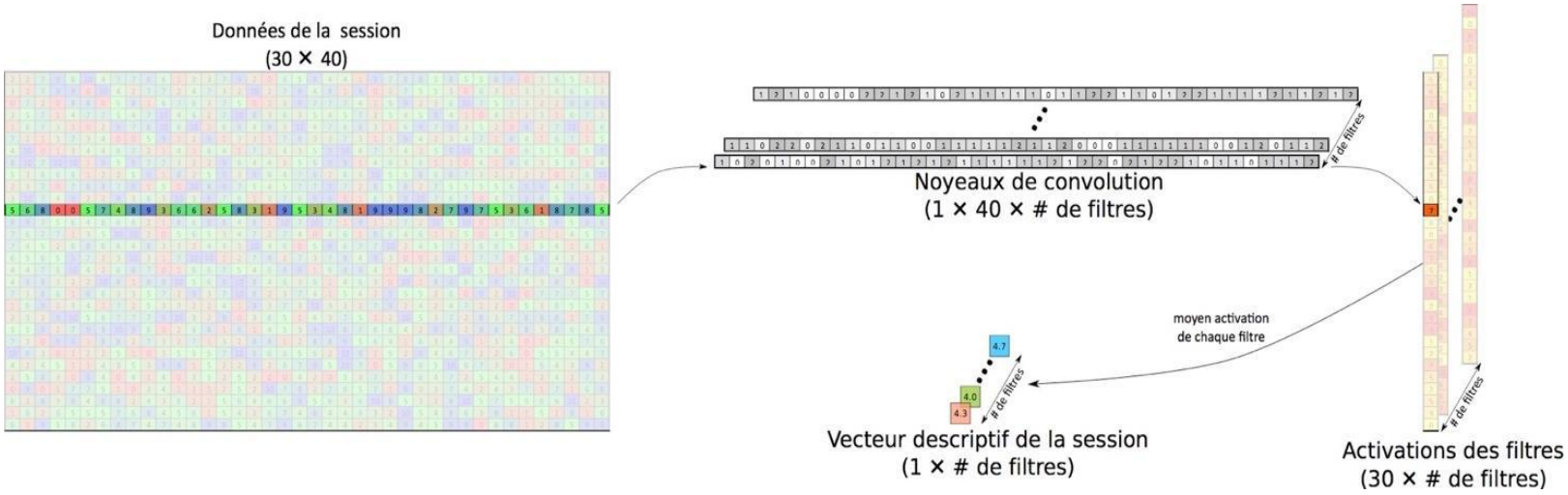


Figure 3 : Utilisation de noyaux de convolution et de la sélection des valeurs moyennes pour produire un vecteur descriptif de la session.

Analyse de la session

Les données des cours de l'élève pour chaque session étant représentées comme un vecteur unique à 256 dimensions, il est maintenant possible d'intégrer judicieusement les données contextuelles de la session. Nous y sommes arrivés par la concaténation de la sortie de la section à convolution (ci-dessus) avec les vecteurs de sessions à 21 dimensions décrits précédemment (qui décrivent l'âge de l'élève, ses SPE, son programme d'études et son statut). La session de l'élève contient donc une

séquence de 30 vecteurs (le nombre maximal de sessions possible), chacun de 277 unités de longueur.

Pour générer une description des antécédents scolaires d'un élève à partir de sa séquence de descriptions des sessions, on doit avoir recours à une autre topologie de réseaux neuronaux. Contrairement aux vecteurs des cours dont la nature ne dépend pas de la position qu'ils occupent dans une session donnée, l'ordre dans lequel ces sessions surviennent est sans doute très important. Les résultats scolaires récents sont probablement un indicateur plus solide de la réussite et de la persévérance scolaires que la performance remontant à plus loin en arrière.

Pour ce faire, une couche de réseau MCLT (mémoire à court et à long terme) a été ajoutée. La MCLT est un type d'architecture de réseau de neurones récurrents, c'est-à-dire, un type de réseau de neurones qui intègre des boucles qui fusionne une partie de ses sorties pour les retourner vers de nouvelles entrées. De cette façon, la séquence session par session des fonctions détectées peut être utilisée pour prédire les résultats à venir.

Pour simplifier, la sortie de la MCLT a été considérée comme égale à la longueur d'un vecteur descriptif de session : 277 unités de longueur. Cela a permis une interprétation simple de la sortie MCLT : faire en sorte qu'elle puisse prédire le prochain vecteur descriptif de session de 277 unités de longueur, même si cela a généré un très grand nombre de paramètres ($> 6 \times 10^5$).

À ce stade, les antécédents scolaires de l'élève (à l'exclusion de la dernière session) sont encodés dans un vecteur unique de 277 unités de longueur. Pour cela, nous concaténons les renseignements démographiques associés au vecteur descriptif de l'élève en 41 dimensions, ainsi que le vecteur d'entrée de la dernière session de 277 unités de longueur. Cela simule les connaissances qui seraient accessibles à l'administration du cégep au début d'une session : les résultats scolaires de l'élève lors des sessions précédentes, les renseignements sur les élèves eux-mêmes, ainsi qu'une liste de cours auxquels l'élève est inscrit. Au total, cela correspond à un vecteur unique de 595 unités de longueur qui sert d'entrée pour les sections des classificateurs qui font les prédictions réelles sur les résultats.

Couches de classification

À ce stade, la description du réseau est assez simple. Cette section du RNA comprend une première couche cachée partagée par les deux sorties du classificateur qui trouve une représentation parcimonieuse et utile des entrées en 595 dimensions et produit un vecteur intermédiaire de 72 unités de longueur. Cela a pour but de trouver une représentation des antécédents scolaires des élèves qui est utile pour prédire à la fois la réussite et la persévérance scolaires. La dimensionnalité de la sortie a été choisie de façon quelque peu arbitraire : comme le nombre de filtres convolués et la dimension de sortie de la MCLT, cela représente un hyperparamètre qui, en principe, pourrait être optimisé pour affiner la sortie.

Les sections du classificateur sont ensuite construites en parallèle avec la même topologie de réseau, mais déconnectées l'une de l'autre. Chaque classificateur contient une couche cachée de 12 unités de longueur suivie d'une couche de sortie de 6 unités de longueur. Les couches de sortie ont une fonction d'activation spéciale appelée fonction exponentielle normalisée (softmax).

Lorsqu'il est entraîné à l'aide d'une fonction de perte *catégorielle à entropie croisée*, la sortie de la fonction exponentielle normalisée est interprétée comme étant la probabilité relative qu'un exemple d'entrée soit associé à l'un des K résultats possibles. Plus précisément, la probabilité relative à la fonction exponentielle normalisée qu'un vecteur x soit associé à la classe de sortie j (d'une catégorie K potentielle) est

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}}$$

où sont les w_j sont les poids appris par la couche. Les deux sorties du réseau sont alors des vecteurs normalisés comportant chacun 6 dimensions qui peuvent être comparés significativement aux vecteurs one-shot à 6 dimensions qui catégorisent la réussite (% de cours réussis) et la persévérance (type de diplôme obtenu) scolaires de l'élève et qui servent de cibles d'entraînement. De plus, la sortie de ces couches en mode prédiction peut être interprétée comme la probabilité que l'entrée corresponde à chacun des six résultats possibles.

L'ensemble du réseau (y compris quelques couches supplémentaires décrites ci-dessous qui aident à éviter le surapprentissage du réseau) est présenté à la Figure 4. Ce réseau a été mis en œuvre dans Python 2.7.5, en utilisant la version 2.1.6 de la bibliothèque Keras et TensorFlow 1.8.0 en tant que programme secondaire.

Entraînement du RNA

L'entraînement final de ce réseau de classification consiste à fournir les données d'entrée, produire les sorties, comparer les sorties du réseau aux cibles d'entraînement et ajuster les poids. Les fonctions de perte pour les deux classificateurs étaient des fonctions de perte catégorielle à entropie croisée (adaptée aux tâches de catégorisation avec une sortie à fonction exponentielle normalisée) et ont simplement été calculées ensemble pour définir la fonction de perte totale du système. L'optimiseur choisi était l'optimiseur Nesterov Adam (Dozat, 2016) en utilisant des taux d'apprentissage initiaux allant d'un maximum de 0,002 pour le préentraînement à un minimum de 0,000 01 pour les derniers stades de polissage de l'ensemble du réseau.

Profondeur du réseau

Ce type de RNA multicouches est souvent considérée comme un réseau de neurones profonds puisqu'il implique plus d'une seule couche cachée. Ces types de réseaux sont devenus de plus en plus populaires, ces dernières années, en raison de l'augmentation de la puissance de calcul, et des algorithmes d'entraînement efficaces ont été mis au point.

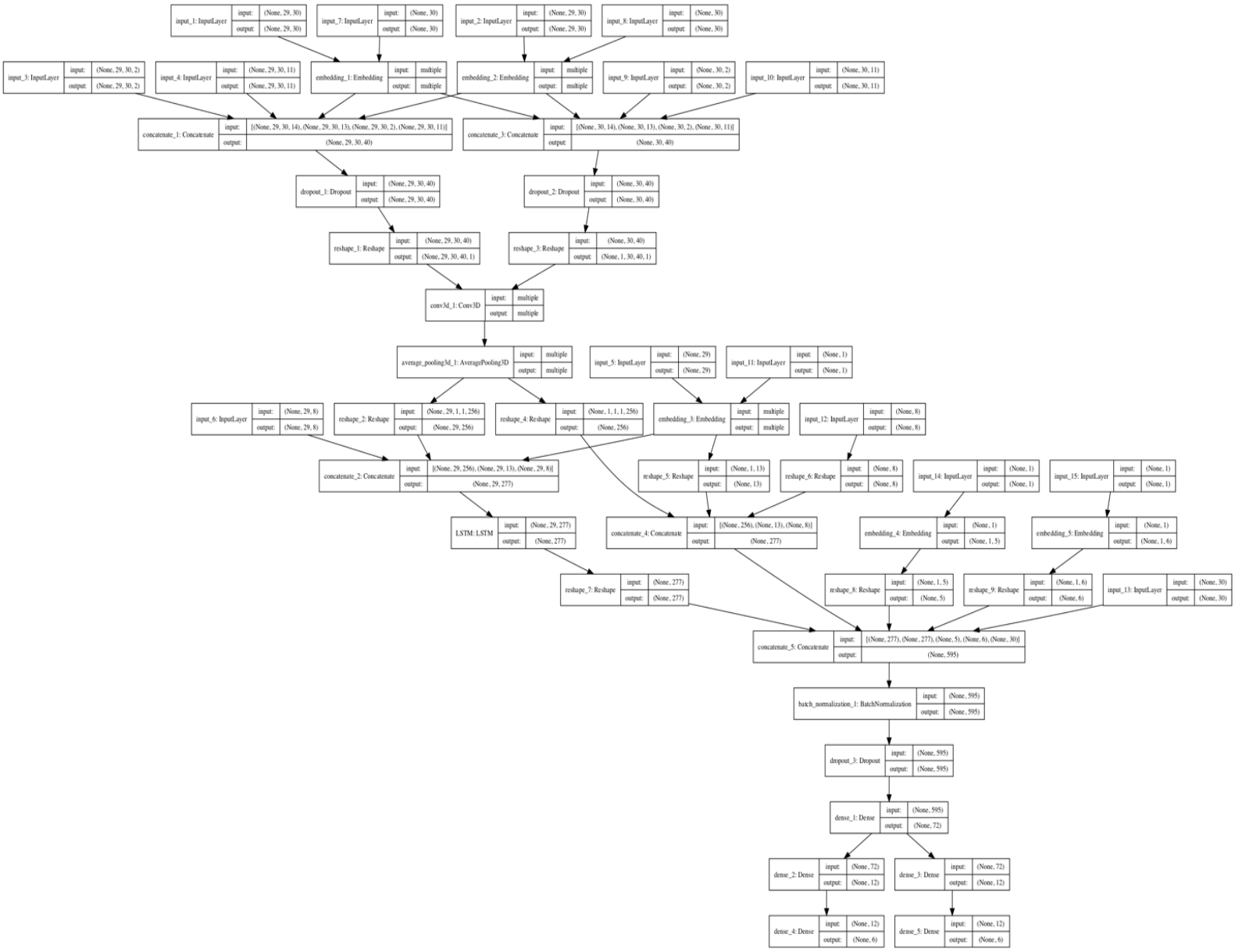


Figure 4 : L'ensemble du réseau de neurones (résumé du modèle de RNA généré par la bibliothèque d'apprentissage profond Keras).

Si l'on inclut les cartes de la couche d'intégration, le réseau en entier intègre près d'un million (1 000 000) de paramètres d'entraînement. La nature à grande échelle de ce réseau pose deux défis : l'entraînement d'un vaste réseau multicouche dans un délai raisonnable et la façon d'éviter le phénomène de surapprentissage. Les stratégies utilisées pour résoudre ces importantes problématiques sont décrites ci-dessous.

Favoriser la généralisabilité

L'une des difficultés avec de grands réseaux de neurones est leur capacité de surapprentissage. Cela signifie que le réseau a souvent la capacité suffisante pour apprendre des caractéristiques de l'échantillon d'entraînement qui ne sont pas représentatives de la distribution observée dans l'ensemble de la population (voir la Figure 5 pour un exemple). Une façon d'éviter le surapprentissage consiste à choisir soigneusement la taille des couches du réseau de façon à s'assurer qu'il n'y ait pas de capacité excédentaire pour apprendre de telles caractéristiques. Cependant, il n'y a aucun moyen de savoir a priori quel doit être la taille du réseau. Avec des réseaux de plus petite taille ou lorsqu'on réseau a beaucoup plus de puissance de calcul à sa disposition, une stratégie qui est souvent employée est l'optimisation des hyperparamètres. Il s'agit d'optimiser l'ensemble des paramètres autres que le poids et le biais (hyperparamètres) du modèle (p. ex., la taille de chaque couche, le nombre de régularisation et la perte d'informations à appliquer) en entraînant entièrement le modèle à plusieurs reprises et en sélectionnant l'ensemble qui fonctionne le mieux sur un autre ensemble de données. La taille de notre modèle, cependant, n'a pas permis que ce soit possible puisque cela aurait nécessité beaucoup plus de puissance de calcul que ce qui était à notre disposition.

Nous avons plutôt choisi d'utiliser d'autres stratégies courantes permettant de limiter le surapprentissage : une régularisation l_1 rigide pour obtenir un réseau clairsemé, la perte d'informations pour favoriser l'apprentissage des caractéristiques essentielles des données et l'arrêt précoce pour sélectionner les modèles les plus généralisables au cours de l'entraînement. Ces stratégies sont décrites ci-dessous.

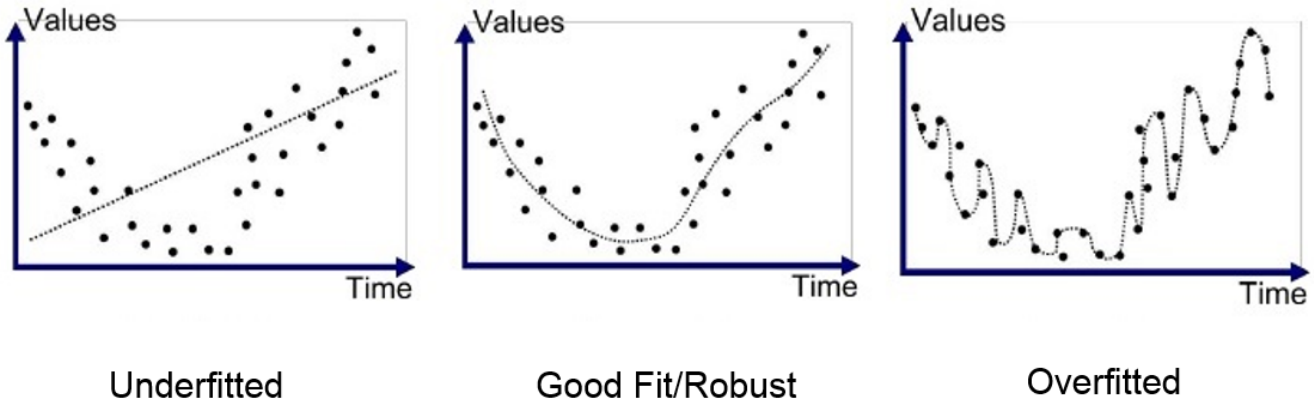


Figure 5 : Le sous-apprentissage et le surapprentissage des données. (Gandhi, 2018). Bien qu'il décrive « mieux » les données d'entraînement que l'apprentissage rigide, l'exemple de surapprentissage serait peu susceptible d'obtenir de bons résultats lorsqu'il est appliqué à un autre échantillon de population.

La régularisation

Une façon de prévenir le surapprentissage du RNA consiste à modifier la fonction coût pour pénaliser certains modèles de croissance au sein du réseau. Nous avons utilisé une régularisation l_1 assez agressive pour favoriser les interconnexions clairsemées entre les neurones, simulant ainsi l'effet d'un réseau plus petit, si possible. Cette forme de régularisation ajoute à la fonction coût C un terme proportionnel à la somme de la valeur absolue des poids du réseau :

$$C = C_0 + \frac{\lambda_1}{n}\sum_{w} |w|$$

où λ_1 représente les forces de régularisation, n est le nombre de poids dans le réseau, w représente les poids du réseau et C_0 est la fonction coût non régularisée. Cette régularisation a pour effet d'introduire un terme dans le gradient qui tente de réduire les poids à zéro à un taux constant. Cette forme de régularisation a alors pour effet de concentrer le poids du réseau dans un nombre relativement petit de connexions à haute importance alors que les autres poids sont menés à zéro.

Les valeurs typiques de λ_1 vont de 10^{-6} à 10^{-5}. Cependant, étant donné que nous n'avons pas les ressources nécessaires pour effectuer une optimisation d'un hyperparamètre, nous avons opté pour

la conception d'un vaste réseau (probablement trop vaste), mais avec un fort paramètre de régularisation de 10^{-4} appliqué à chaque couche. Cependant, les grands poids sont admis s'ils améliorent considérablement les performances du réseau.

D'autres formes de régularisation existent, mais n'ont pas été utilisées dans cette étude. La régularisation l_2 consiste à pénaliser la somme des carrés des poids :

$$C = C_0 + \frac{\lambda_2}{2n} \sum_w \quad w^2$$

Comme pour la régularisation l_1, cela tend à réduire la valeur des poids appris. Ce qui diffère, cependant, c'est que le terme qui en résulte dans le gradient n'abaisse plus les poids à un taux constant : il agit plutôt de façon à faire en sorte que les poids soient plus uniformément répartis en pénalisant plus fortement les poids plus grands que ceux plus petits.

La perte d'informations

La perte d'informations est une autre méthode que nous avons utilisée pour empêcher le surapprentissage. Dans cette stratégie, un certain pourcentage des synapses est ignoré au hasard (réglées à zéro) au cours de chaque itération d'entraînement. Cela a pour effet de réduire l'apprentissage interdépendant entre les neurones pour éviter les coadaptations complexes du réseau aux données d'entraînement (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Cela oblige le réseau à apprendre un ensemble plus rigoureux de représentations des données. En contrepartie, il nécessite un plus grand nombre d'itérations d'entraînement pour en établir la convergence.

Dans notre RNA, une perte d'informations de 30 % a été utilisée de façon constante entre les sections d'intégration et de convolution, dans la MCLT et dans les couches de classification à la fois pour l'entraînement et le préentraînement, à une exception près : une plus grande perte d'informations a été utilisée dans le préentraînement des couches de convolution afin de mettre beaucoup d'accent sur le développement de représentations efficaces des données plutôt que d'apprendre une simple transformation identitaire.

L'arrêt précoce

Finalement, l'arrêt précoce a été utilisé comme dernière mesure dans la prévention du surapprentissage. Après chaque époque d'entraînement (ou de préentraînement), la performance du réseau était comparée à des données (appelées les données de validation) qui n'avaient pas été utilisées dans l'entraînement. Cela nous a permis de surveiller si et quand le réseau avait atteint son point d'entraînement optimal. Un réseau qui aurait surappris commencerait à démontrer une piètre performance vis-à-vis l'ensemble de validation en raison de la sensibilité qu'il avait apprise dans l'analyse des caractéristiques de l'ensemble d'entraînement (Figure 6). L'arrêt précoce consiste à exercer un contrôle sur la perte de validation (erreur) lors de l'entraînement et d'arrêter l'entraînement lorsque la perte de validation commence à augmenter.

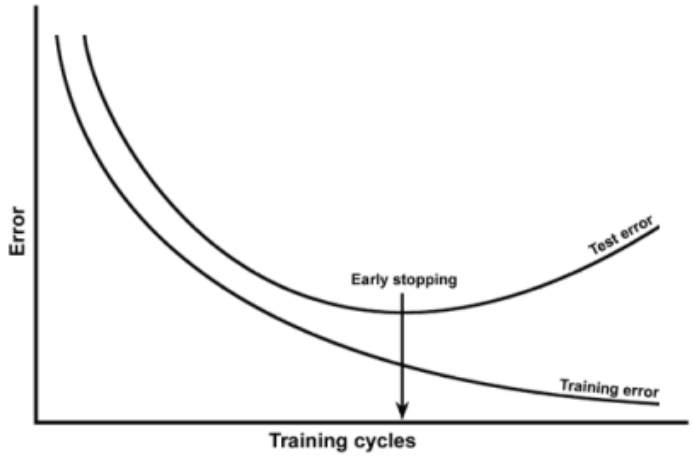


Figure 6 : Arrêt précoce de l'entraînement du réseau neuronal pour empêcher le surapprentissage.

Bien qu'utilisé au cours de l'entraînement de l'ensemble du réseau, l'arrêt précoce était principalement bénéfique au cours du préentraînement des couches du réseau (voir Entraînement des RNA profonds ci-dessous). Nous pensons que ce constat découlait en partie de l'efficacité de la régularisation et de l'arrêt précoce utilisés, ainsi que le fait que nous étions limités par le nombre d'époques que nous étions en mesure d'exécuter en raison de contraintes de temps et de la taille du réseau.

Entraînement des RNA profonds

Plusieurs problèmes peuvent survenir dans l'entraînement de ce RNA profond. Premièrement, il y a des difficultés liées à la multitude de paramètres : un réseau aussi vaste prend beaucoup de temps à entraîner à partir de zéro. Une séance d'entraînement complète prenait habituellement 50 époques ou plus (exécution complète des données d'entraînement). Ceci s'est avéré peu pratique puisque chaque époque d'entraînement nécessitait environ quatre jours à exécuter avec le matériel disponible.

Un autre problème qui est endémique aux réseaux de neurones profonds s'appelle le problème de disparition du gradient (Deng, Hinton, & Kingsbury, 2013; Szegedy et al., 2015). Ce phénomène a longtemps retardé la mise en œuvre pratique des réseaux de neurones profonds. En bref, le problème de disparition du gradient fait référence à une situation dans laquelle le gradient des neurones des niveaux inférieurs est très près de zéro parce que la valeur des activations des neurones des niveaux supérieurs est presque saturée. Cette situation entrave le processus d'apprentissage lié au gradient et allonge d'autant plus la durée de l'entraînement.

Pour atténuer ces problèmes, nous avons adopté une stratégie à deux volets. Premièrement, nous avons sélectionné des fonctions d'activation et des architectures de réseau qui sont moins touchées par le phénomène de saturation qui mène à la disparition des gradients. La couche de convolution et les couches denses de la section classification utilisaient toutes la fonction d'activation *linéaire rectifiée* plutôt que les fonctions tangentes hyperboliques ou sigmoïdes, plus classiques. Il a été démontré que les activations linéaires rectifiées, bien que non analytiques à leur seul point non linéaire (Deng et al., 2013), amélioraient la performance de l'entraînement du réseau dans les systèmes de RNA profonds. La dernière section impliquant des activations non linéaires est la couche MCLT. La MCLT est une forme de réseau de neurones récurrents dont l'architecture a été soigneusement conçue pour éviter le problème de disparition du gradient, et ce, en dépit de l'utilisation des fonctions d'activation sigmoïdes. L'utilisation de la MCLT pour les fonctions d'analyse séquentielle et d'activation linéaire rectifiée pour les autres couches est souvent suffisante pour éliminer complètement le problème de disparition du gradient.

L'autre stratégie que nous avons adoptée est l'utilisation du préentraînement vorace, une couche à la fois, de notre réseau. Cette stratégie, une forme de *transfert d'apprentissage* (Bengio, Lamblin, Popovici, & Larochelle, 2006), consiste à entraîner le réseau neuronal couche par couche pour des tâches plus simples qui illustrent certains aspects de la distribution des exemples qui sont utiles à la tâche finale. De cette façon, le préentraînement permet une initialisation du RNA qui implique un ensemble de représentations intermédiaires des données qui sont cohérentes à travers les couches du réseau. Une fois totalement préentraîné, le réseau est « poli » par l'exécution d'un entraînement complet beaucoup plus court pour optimiser les représentations intermédiaires pour l'ultime tâche de catégorisation.

Dans notre cas, il y avait quatre niveaux de préentraînement : le préentraînement des cartes d'intégration (décrit ci-dessus à la section Les cartes d'intégration), le préentraînement des filtres convolués, le préentraînement de la couche MCLT et le préentraînement des catégorisateurs de niveau supérieur. Ces étapes ont été réalisées dans cet ordre. À chaque étape, une représentation de l'ensemble de données a été construite en fonction de la sortie de la couche qui était préentraînée pour servir d'entrée à la couche suivante. Chacune de ces séances de préentraînement utilisait les stratégies de régularisation, de perte d'informations et d'arrêt précoce décrites à la section Favoriser la généralisabilité ci-dessus.

Préentraînement des cartes d'intégration

Les cartes d'intégration décrivant les cours suivis, les écoles et les programmes d'études ont été préentraînées, comme décrit ci-dessus, en se basant sur la cooccurrence, dans le corpus de séquences de cours suivis, par l'utilisation de routines créatrices de mots-vecteurs de la trousse GloVe (Pennington, 2014). Les intégrations du code postal et du pays de naissance ont été créées à la main en se basant sur l'emplacement géographique et, dans le cas du pays de naissance, les données économiques et relatives au développement humain. Ces deux dernières intégrations comportaient aussi trois dimensions initialisées au hasard pour une rétropropagation ultérieure. L'ensemble de données d'entraînement a été traité par ces cartes d'intégration. Ces données d'entraînement prétraitées ont servi d'entrée pour le préentraînement de la couche de convolution.

Préentraînement de la couche de convolution

Le préentraînement de la couche de convolution a été réalisé à l'aide d'un autoencodeur débruiteur (Vincent, Larochelle, Bengio, & Manzagol, 2008). Cela a nécessité l'utilisation de la couche de convolution pour encoder les données des cours de manière à ce qu'elles puissent être reconstituées, et ce, même en présence d'une entrée quelque peu endommagée. Ceci permet d'entraîner les 256 fonctions de convolution pour qu'elles soient à tout le moins en mesure d'élaborer une représentation des données ou même de procéder à un traitement significatif des données.

Afin d'encourager le système à apprendre une représentation significative plutôt que de simplement apprendre la transformation de l'identité, les entrées du système ont été corrompues par le mécanisme de perte d'informations (dans notre cas, à l'aide d'un ratio de perte d'informations de 70 %). La sortie de la couche de convolution a été transmise à une couche de décodage (à usage unique) ayant la même dimensionnalité que l'entrée. Le système a ensuite été entraîné pour essayer de reproduire les données (non corrompues) originales (voir la Figure 7 pour un exemple d'autoencodeurs de débruitage appliqué à un problème de reconnaissance d'image). Pour ce faire, il doit apprendre les principales caractéristiques des données plutôt que d'apprendre simplement une transformation d'identité.

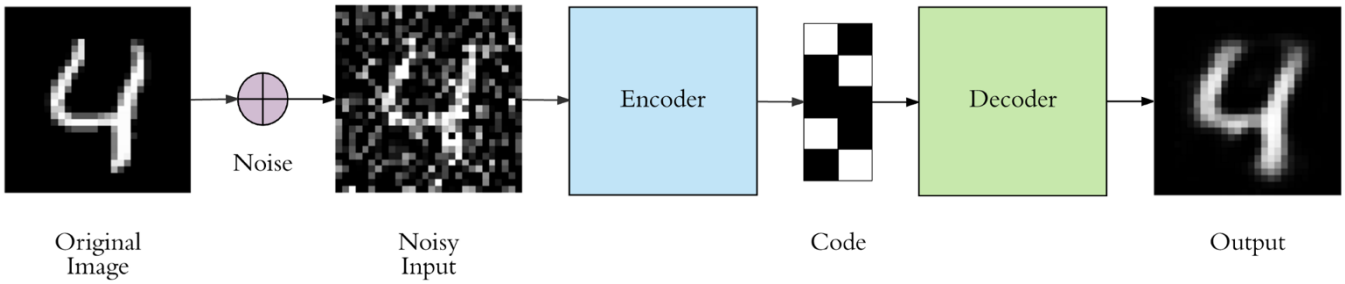


Figure 7 : Exemple d'autoencodeur débruiteur utilisé pour la reconnaissance de formes. (Dertat, 2017)

En plus de la perte d'informations de 70 % utilisée, nous avons aussi utilisé la régularisation l_1 rigide pour obtenir une représentation clairsemée des données d'entrée comme protection supplémentaire contre le surapprentissage.

Le système a été entraîné sur 200 époques des 5,5 millions de vecteurs de cours dans l'ensemble de données. En raison de l'empreinte mémoire relativement faible du noyau de convolution (seulement 10 496 paramètres d'entraînement), il a été possible d'utiliser des lots de très grande taille contenant 4 096 exemples d'entraînement pour chaque mise à jour d'entraînement. La fonction de perte utilisée était la différence entre la moyenne des carrés des vecteurs d'entrée et de sortie (non corrompus). À chaque époque, la performance de l'autoencodeur dans l'encodage des données provenant de la validation a été mesurée et le meilleur modèle a été conservé. Le modèle final retenu présentait une erreur moyenne de 0,1 par dimension de l'entrée.

Une fois entraînée, la couche de décodage a été éliminée et les filtres de préentraînement ont été utilisés pour encoder les données non corrompues des cours, y compris les entrées de la dernière session et les données cibles de la dernière session. Ces représentations fixes des données ont ensuite été utilisées dans le préentraînement de la couche MCLT.

Préentraînement de la couche MCLT

La couche MCLT est une forme de réseau de neurones récurrents destiné au traitement de séquences de données temporelles. Pour son préentraînement, nous avons choisi de nous concentrer sur sa capacité à reconstituer le vecteur réel observé de la dernière session (c.-à-d., les données cibles de la dernière session pour lesquelles les résultats étaient connus) compte tenu de la séquence de données de la session précédente. Les entrées du MCLT sont les séquences des vecteurs de la session de 277 unités de longueur correspondant aux sessions précédentes. La sortie du MCLT a aussi été choisie pour être un vecteur de 277 unités de longueur pour qu'il puisse être entraîné pour prédire les données cibles de la dernière session en tenant seulement compte des données des sessions précédentes. Encore une fois, la perte d'informations (30 %) et la régularisation l_1 ont été utilisées pour permettre d'obtenir une représentation significative. Le préentraînement a pris 100 époques en utilisant un lot de 2048 et une moyenne d'erreur de l'ordre de 0,01 par dimension. Une fois préentraînée, la MCLT a été utilisée pour créer des vecteurs d'antécédents scolaires encodés comme entrée du préentraînement des catégorisateurs.

Préentraînement des catégorisateurs

Les catégorisateurs ont été préentraînés en prenant comme entrée la sortie du préentraînement de la MCLT et en l'utilisant pour prédire les résultats en matière de réussite et de persévérance scolaire des élèves utilisés dans le mode d'entraînement complet. Cela signifie que cette étape était, en ce qui concerne ces couches supérieures, identique aux entraînements réels qui suivraient. La seule différence était que les erreurs n'étaient pas autorisées à se propager vers les niveaux inférieurs du réseau. Ceci a été exécuté pendant 50 époques de préentraînement.

Entraînement du modèle en entier

Une fois que les différents niveaux du modèle ont été préentraînés, l'ensemble du réseau connecté a été chargé avec les poids préentraînés et l'ensemble du réseau a été entraîné avec les données d'entraînement, soumises à un arrêt précoce. Comme mentionné, la taille du réseau signifiait que l'entraînement se produisait lentement, soit environ quatre jours par époque. Très tôt, la performance du catégorisateur de réussite scolaire a démontré un comportement préoccupant : sa précision catégorique est passée de ~ 50 % à 15 % avant de remonter à > 50 % après la deuxième époque d'entraînement. En même temps, la précision catégorielle du catégorisateur de persévérance s'est améliorée, passant de ~ 45 % à > 80 %. Deux autres entraînements ciblés ont été effectués avec des poids d'apprentissage réduits pour produire le modèle complet. Après ces longues séries d'entraînement, l'état final du système a été sauvegardé (Tableau 6) et le réseau mis à l'essai avec l'ensemble de données d'essai (non utilisé jusqu'à présent).

	Ensemble de données d'entraînement	Ensemble de données de validation	Ensemble de données d'essai
Perte globale	4,563 2	4,308 1	1 688
Perte de catégorisation de la réussite	3,816 6	3,629 6	1 058
Perte de catégorisation de la persévérance	0,371 3	0,331 1	0,290 1

Précision catégorielle de la réussite	58,10 %	62,57 %	65,49 %
Précision catégorielle de la persévérance	88,12 %	90,90 %	89,34 %

Tableau 6 : Résultats de l'entraînement de l'ensemble du réseau RNA.

En examinant le Tableau 6, nous constatons que la performance du réseau sur les données de validation et les données d'essai est meilleure que celle sur les données d'apprentissage. Ceci peut être expliqué en considérant que les essais de validation éliminent les pertes d'informations et les essais utilisant les données d'essai éliminent à la fois les pertes d'informations et la régularisation dans leurs calculs. La performance supérieure associée aux données de validation et d'essai, cependant, n'appuie pas l'affirmation selon laquelle le surapprentissage n'était pas un problème avec ce réseau. En effet, nous croyons qu'il est possible d'effectuer davantage d'entraînement avant que le surapprentissage devienne un problème significatif (il n'y avait aucun signe de diminution des améliorations avec plus d'entraînement). De plus, il devrait être possible d'ajuster les différents hyperparamètres pour atteindre des performances encore meilleures. Cependant, les limites de nos ressources informatiques ont retardé ces investigations pour le moment.

Comme indiqué ci-dessous, les précisions catégoriques ne sont pas des mesures directes indiquant si le RNA signale les élèves à risque de façon appropriée : cela signifie simplement que la cellule du vecteur de sortie qui a la plus grande valeur coïncide avec la sortie connue. Pour la mise en contexte, les catégories réussite et persévérance possèdent 6 dimensions; cela signifie que le choix d'une catégorie au hasard coïnciderait avec la bonne réponse 16,7 % du temps. Les catégorisateurs de réussite et de persévérance surpassent cette valeur de référence par des facteurs de 3,93 et 5,36, respectivement.

CHAPITRE 9

RÉSULTATS DE L'ANALYSE PAR RÉSEAUX DE NEURONES

La sortie de la réponse du réseau à l'ensemble de données d'essai (c'est-à-dire, les données qui n'ont jamais été utilisées pour l'entraînement ou la validation du modèle) a été comparée avec les résultats réels du pourcentage de cours réussis (réussite de l'élève) et des observations selon lesquelles un élève a terminé son DEC ou abandonné ses études (persévérance). Cette analyse a été effectuée dans le contexte du développement d'un système de signalement précoce pour identifier rapidement les élèves à risque de façon à ce que les systèmes de soutien du cégep puissent être mobilisés et des interventions envisagées.

Dans ce contexte, il est important de considérer non seulement la sensibilité du classificateur (c.-à-d., le pourcentage de cas positifs identifiés comme tels), mais aussi sa spécificité (c.-à-d., le pourcentage de cas rejetés). Il est très facile d'identifier tous les élèves à risque si l'on est disposé à accepter un grand nombre de faux positifs. Inversement, il est très facile d'éviter les faux positifs si l'on est disposé à accepter un grand nombre de faux négatifs. De toute évidence, il y a un compromis à faire entre la sensibilité et la spécificité d'un modèle de classificateur et c'est dans ce contexte que nous évaluons la performance du classificateur du RNA.

Les sorties des classificateurs de RNA sont constituées de vecteurs normalisés de 6 unités de longueur qui sont interprétés comme des probabilités d'appartenance à une classe. La conversion de ces vecteurs en classificateurs binaires (à risque ou non à risque) consistait à décider quelles classes

constituaient les classes « à risque » et à additionner les valeurs correspondantes pour déterminer la probabilité que l'élève soit à risque. Pour quantifier la performance des classificateurs, il fallait ensuite comparer ces coefficients de probabilité aux résultats correspondants.

Pour étudier le comportement de notre modèle dans ce contexte, nous avons examiné les fonctions d'efficacité de l'observateur (ROC) des classificateurs pour quantifier leur performance, ainsi que déterminer les seuils optimaux pour l'identification des élèves à risque. Ces courbes ROC représentent simplement la relation entre la sensibilité et la spécificité des classificateurs binaires évalués en fonction de tous les seuils possibles.

L'aire sous la courbe ROC (ASC-ROC) est une mesure de l'efficacité d'un classificateur binaire à distinguer les deux classes. L'ASC-ROC permet d'interpréter le rang moyen de l'ensemble des exemples positifs. Cela signifie que, dans un système ayant une ASC-ROC de 0,95, un exemple positif moyen obtiendrait un rang plus élevé (plus probablement positif) que 95 % des exemples négatifs. Les systèmes présentant des valeurs d'ASC-ROC de 0,8 à 0,9 sont considérés comme « d'excellents » discriminateurs alors que ceux ayant des valeurs supérieures à 0,9 sont considérés comme « exceptionnels » (Mandrekar, 2010).

La détermination du seuil optimal t à partir duquel signaler l'élève comme étant « à risque » a été réalisée en maximisant la statistique Youden $J(t)$ du système. Cette mesure s'exprime simplement de la façon suivante :

$$J(t) = Sensitivity(t) + Specificity(t) - 1$$

Le seuil optimal t est alors le point d'équilibre entre la sensibilité et la spécificité, et la valeur J correspondant à ce point est interprétée comme la probabilité de prendre une décision éclairée plutôt que de deviner au hasard (Powers, 2011). Considérons maintenant ces analyses appliquées aux classificateurs de la réussite et de la persévérance scolaires des élèves.

Classificateur de la réussite scolaire des élèves

Les 6 catégories du vecteur de la réussite scolaire correspondent au pourcentage des cours auxquels l'élève est inscrit et qu'il a réussis (Tableau 7). Étant donné que la probation scolaire donnée est habituellement activée à 50 %, nous avons choisi de définir comme étant « à risque » les élèves appartenant à l'une des trois premières catégories (c.-à-d., jusqu'à 60 % de cours réussis). Pour chaque élève, la probabilité d'être dans ce groupe était simplement la somme des résultats à ces trois premières catégories. Ceux-ci ont été calculés et comparés aux résultats observés.

Catégorie	% de cours réussis
0	Jusqu'à 20
1	Entre 20 et 40
2	Entre 40 et 60
3	Entre 60 et 80
4	Entre 80 et 100
5	100

Tableau 7 : Classes de réussite scolaire des élèves

Le ROC a été calculé par l'entremise du logiciel R 3.5.0 en utilisant la bibliothèque pROC v. 1.12.1 (Robin et al., 2011) et présenté à la Figure 8.

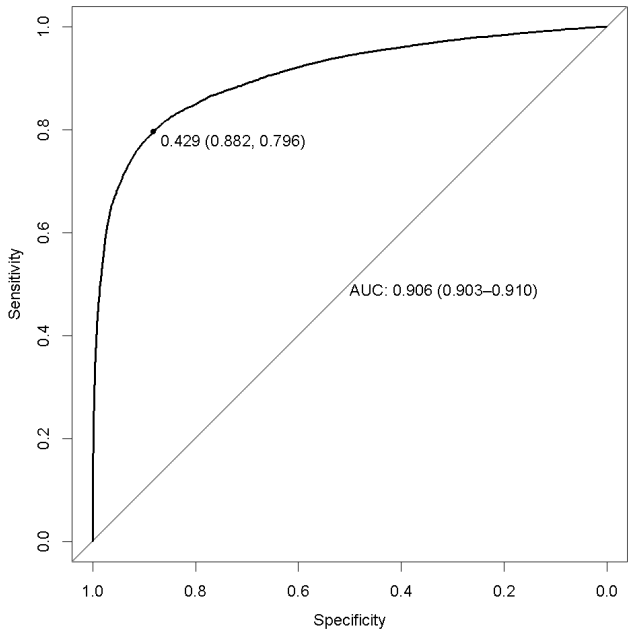


Figure 8 : Courbe ROC du classificateur de la réussite scolaire des élèves.

L'ASC-ROC a été estimée à 0,906 ± 0,004, ce qui indique une performance globale « exceptionnelle » comme classificateur.

Trouver le seuil de probabilité optimal pour identifier les élèves susceptibles d'échouer nécessitait de découvrir le point qui maximisait le *J* de Youden. Dans ce cas, le seuil optimal (mis en évidence à la Figure 8) était de 0,429 5. Ce point, correspondant à une spécificité de 0,882 1 ± 0,004 9 et une sensibilité de 0,796 2 ± 0,006 8, ce qui donne une valeur *J* de Youden de 0,678 3 ± 0,008 4. Si l'on choisit un seuil de 0,429 5 comme probabilité au-dessus de laquelle les élèves doivent être jugés à risque, alors cette catégorisation pleinement éclairée se produit avec une probabilité de 67,8 %.

Ces résultats indiquent que, malgré une précision catégorielle relativement faible (65,5 %) de la sortie brute du RNA, la performance du classificateur est adéquate. Étant donné la probabilité qu'un entraînement et un affinement supplémentaires du réseau permettraient d'améliorer davantage ses performances, nous concluons que ces systèmes sont viables pour un signalement précoce des élèves à risque d'échouer une partie significative des cours auxquels ils sont inscrits.

Classificateur de la persévérance scolaire des élèves

La catégorisation du diplôme prévu de l'élève ne comporte qu'une seule catégorie attribuée pour l'abandon des études. Par conséquent, aucun regroupement de catégories n'était nécessaire. Il fallait plutôt procéder à l'identification directe de la probabilité prédite d'abandon des études et la comparer aux résultats observés.

La courbe ROC a été calculée (Figure 9) et l'analyse a été effectuée comme ci-dessus. L'ASC-ROC s'est avérée être de 0,905 ± 0,004, ce qui indique, une fois de plus, une performance « exceptionnelle » du classificateur pour établir une distinction entre les deux cas.

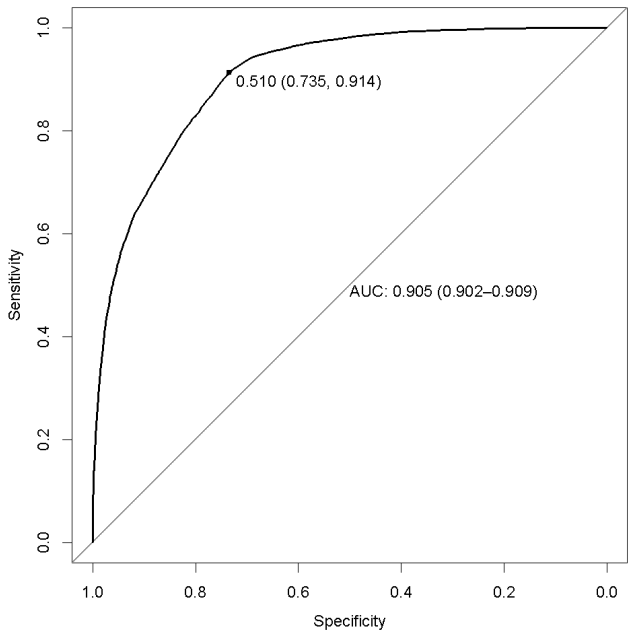


Figure 9 : Courbe ROC du classificateur de la persévérance scolaire des élèves.

Le seuil optimal pour l'identification des élèves à risque d'abandonner leurs études, selon le critère de Youden, s'est avéré être de 0,509 8. À ce stade, le classificateur présentait une sensibilité de 0,9137 ± 0,0047 et une spécificité de 0,7352 ± 0,00635, ce qui donne une valeur *J* de Youden de 0,6489 ± 0,0079. Le choix d'un seuil de probabilité de 0,5098 pour signaler les cas à risque a ensuite mené à une catégorisation pleinement éclairée avec une probabilité de 64,9 %.

Une fois encore, ces résultats indiquent que le classificateur du RNA arrive à identifier adéquatement les élèves à risque d'abandonner leurs études. Contrairement au classificateur de la réussite scolaire des élèves décrit ci-dessus, cependant, il semblait y avoir peu d'amélioration dans la précision ou la perte catégorielle de la dernière époque de l'entraînement du RNA. Ceci suggère qu'un entraînement supplémentaire serait peu susceptible d'améliorer ses performances de manière significative sans d'importants changements dans l'architecture du réseau.

Discussion

L'identification précoce des élèves à risque d'échouer une grande partie de leurs cours (réussite de l'élève) ou d'abandonner leurs études serait une étape importante permettant de mobiliser plus efficacement les systèmes de soutien du cégep. Nos résultats suggèrent que l'utilisation des systèmes de classification des réseaux de neurones profonds est une stratégie viable pour permettre cette identification.

Améliorations possibles

Dans la mise au point et l'évaluation de ce réseau de classification, un certain nombre d'orientations pour fins d'études ultérieures sont devenues évidentes. Nous les présentons ici pour suggérer aux parties intéressées des orientations de recherches.

Optimisation améliorée

Comme nous l'avons expliqué dans les sections précédentes, le réseau qui a été mis au point n'a pas été entièrement optimisé en ce qui concerne ses hyperparamètres. Des constantes comme la taille des couches, les constantes de régularisation, les ratios de perte d'informations, les taux d'apprentissage, etc., ont toutes des répercussions sur le temps nécessaire pour entraîner un réseau neuronal et sa performance finale. Il existe des mécanismes efficaces pour trouver de tels hyperparamètres optimaux (e.g., Bengio, 2000). Cependant, ceux-ci reposent sur la création de nombreux modèles entièrement entraînés pour chaque point d'espace des hyperparamètres. Compte tenu de la durée de chaque époque (~ 4 jours), cela n'était pas pratique. Des diminutions importantes dans la durée d'entraînement pourraient être obtenues en allouant plus de ressources informatiques au problème. Effectuer l'entraînement du réseau neuronal à l'aide d'unités de traitement graphique

(UTG) plutôt que d'unités centrales de traitement (UCT) entraînerait d'importantes diminutions dans la durée de l'entraînement, de même que la parallélisation des calculs sur plusieurs unités informatiques. Cependant, la taille de l'ensemble de données présente ses propres difficultés puisqu'elle serait probablement supérieure à la capacité de la plupart des UTG disponibles sur le marché.

Une autre stratégie d'optimisation consisterait à utiliser l'état actuel du système comme point de départ pour des classificateurs plus spécialisés. En ce moment, la majeure partie du RNA est partagée par les classificateurs associés à la réussite et la persévérance scolaires. Cela signifie que la majorité du réseau sous-jacent doit accommoder des représentations qui sont utiles simultanément pour prédire deux différents types de résultats. Il est concevable que des améliorations des performances puissent être obtenues si ces deux objectifs étaient dissociés. Cela permettrait aux couches inférieures (les sections MCLT, convolution et intégration) d'ajuster leurs représentations à une tâche de classification unique plutôt que la tâche commune exigée d'eux présentement. Ceci pourrait se faire assez facilement en entraînant davantage le modèle actuel à l'aide d'une modification de la fonction coût dans laquelle seul le classificateur d'intérêt contribue au taux d'erreur du réseau.

Développer l'ensemble du réseau

La mise en œuvre d'un tel système dans l'ensemble du réseau améliorerait vraisemblablement ses capacités de plusieurs façons :

- Cela augmenterait de façon importante les données disponibles pour l'entraînement;
- Cela permettrait de mieux refléter la diversité culturelle de la province : les trois cégeps inclus dans cette étude sont des établissements de langue minoritaire et, par conséquent, représentent un mélange culturel différent de l'ensemble du réseau.
- Intégrer davantage de données provenant des régions permettrait au modèle de mieux apprendre le contexte d'élèves qui ne proviennent pas du milieu urbain ou de la banlieue de Montréal.
- Un meilleur suivi des élèves à travers le système. En ce moment, l'abandon des études dans un établissement est identifié comme du décrochage scolaire. Cependant, certains de ces cas consistent probablement en un transfert vers un autre établissement. Même si nous avions

été en mesure de suivre les transferts entre nos trois collèges, nous n'avons aucun moyen de comptabiliser les transferts d'élèves vers d'autres établissements.

Cependant, il ne s'agit pas d'une recommandation anodine puisque le développement d'un tel système exigerait un effort important dans la collecte et le traitement des données provenant de chaque établissement. En effet, l'obtention des données sur lesquelles ce modèle a été construit a constitué un obstacle majeur pour l'ensemble de l'étude. Une deuxième tâche associée aux données consistait à les normaliser puisque de nombreuses incohérences ont été trouvées dans plusieurs des champs. L'école secondaire où l'élève a étudié et le pays de naissance, par exemple, se traduisaient tous les deux par de multiples clés différentes faisant référence à la même entité. Ces dernières ont dû être réparées à la main, une tâche très laborieuse.

De plus, l'élargissement de la portée des données nécessiterait une augmentation correspondante de la puissance de calcul, ce qui pourrait représenter une dépense importante.

Comprendre ce qui motive les prédictions

C'est possible, du moins en principe, d'explorer les poids appris du modèle pour examiner quels aspects sont à l'origine de ses décisions en matière de classification. En particulier, examiner comment les cartes d'intégration et les filtres convolués évoluent lors de l'entraînement pourrait permettre de faire la lumière sur les aspects du dossier scolaire de l'élève qui sont importants pour la tâche de classification. Cela pourrait fournir les fondements sur lesquels peaufiner les modèles statistiques davantage fondés sur des principes à propos de ce qui prédit le mieux la réussite et la persévérance scolaires.

Modèles de données plus utiles

Le modèle tel qu'il est conçu concentre toute son attention sur l'ensemble des dossiers scolaires. C'est possible, cependant, que ce ne soit pas l'organisation optimale des données. Au lieu de cela, les données pourraient être réparties en ensembles de « 1 session d'expérience collégiale », « 2 sessions d'expérience collégiale », etc., en tronquant la séquence des sessions, selon le cas. Chacun de ces ensembles de données tronquées pourrait alors être utilisé pour entraîner son propre réseau, en fournissant des classificateurs spécialisés pour la 1re session, la 2e session, etc. Cela pourrait s'avérer fructueux puisque la grande majorité des mêmes données serait réutilisable entre les

modèles. En effet, il pourrait être avantageux de commencer l'entraînement avec le modèle actuel complet comme point de départ.

Encore une fois, les ressources informatiques et le temps sont les facteurs limitatifs associés à une telle enquête. Au lieu d'entraîner un seul réseau, en prenant des jours ou des semaines d'entraînement, nous devrions répéter ce processus pour plusieurs réseaux de ce type, un pour chaque session d'intérêt.

Une avenue possible pour développer le réseau

Les Collège Bois-de-Boulogne et le Collège John Abbott viennent de recevoir du financement pour la création d'un centre collégial de transfert de technologie pour des applications en intelligence artificielle (CCTTia). Il pourrait s'agir d'une occasion intéressante de rassembler l'expertise et les ressources informatiques nécessaires pour entreprendre certains des développements décrits ci-dessus.

Éviter les dérives

Il y a un dernier point sur lequel nous croyons qu'il faut se pencher. Dans une conversation informelle avec des collègues et des administrateurs au sujet de ce projet, une suggestion a été faite à maintes reprises : que de tels systèmes pourraient être utilisés pour améliorer les décisions d'admission au cégep. Nous voulons noter avec véhémence notre objection à cette notion. Un modèle qui n'est pas fondé sur des principes clairs, comme ce classificateur de RNA, ne serait pas idéal pour les admissions et pourrait être éthiquement questionnable. Supposons que certains facteurs invisibles à nos yeux contribuent à la sensibilité et la spécificité du RNA et que certains de ces facteurs soient dérivés du statut économique et la race. Le modèle ayant compilé ces données augmenterait le biais contres les étudiants sous-représentés qui ont le plus besoin d'éducation. Bien qu'il soit possible d'utiliser ces systèmes pour sélectionner les élèves ayant une plus grande probabilité d'obtenir leur diplôme, sans bien comprendre la façon dont le classement est fait, les décisions découlant de ces classifications non fondées sur des principes ne devraient pas être utilisés pour sélectionner l'élève. Alors que l'activation du système de soutien du cégep pour aider un élève

incorrectement signalé comme étant « à risque » n'a pas d'impact négatif sur l'élève, leur refuser l'entrée dans le réseau collégial en a certainement un. En bref, les systèmes comme celui-ci ne devraient être utilisés que dans un contexte de soutien mutuel à faible risque.

Conclusions

En conclusion, nous avons été en mesure de répondre aux deux questions de recherche que nous avons étudiées.

1) Les RNA semblent être un moyen viable de prédire quels sont les élèves à risque d'échouer une partie substantielle des cours collégiaux auxquels ils sont inscrits au cours d'une session donnée compte tenu de leurs antécédents scolaires, des cours auxquels ils sont inscrits actuellement et d'un minimum de renseignements démographiques.
    a. La définition d'élèves à risque avec ce réseau neuronal correspond à une probabilité plus grande que *0.4295* d'échouer 40 % ou plus de leurs cours. Ce point correspond à une *sensibilité* de 0,796 2 ± 0,006 8 et une *spécificité* de 0,882 1 ± 0,004 9.

2) Les réseaux de neurones artificiels semblent être un moyen viable de prédire quels sont les élèves à risque d'abandonner leurs études collégiales au cours d'une session donnée en tenant compte de leurs antécédents scolaires, des cours auxquels ils sont inscrits actuellement et d'un minimum de renseignements démographiques.
    a. La définition de « à risque d'abandonner leurs études collégiales » avec ce réseau de neurones correspond à une probabilité de plus de 0,509 8 de ne pas revenir au cégep au cours de la session suivante. À ce stade, le système démontre une *sensibilité* de 0,913 7 ± 0,004 7 et une *spécificité* de 0,735 2 ± 0,006 35.

Enfin, il y a de nombreuses améliorations qui peuvent être apportées à ce système (décrites Améliorations possibles ci-dessus) et qui pourraient accroître considérablement les performances et l'applicabilité de tels réseaux de neurones artificiels.

Contribution informatique au réseau collégial : Lien GitHub

Étant donné qu'une des contributions majeures de cette recherche consiste en des modèles informatiques incluant du code et des modèles prédictifs de RNA, nous incluons ci-dessous un lien vers un dépôt GitHub informatique ouvert et disponible au réseau collégial.

https://github.com/sameerbhatnagar/studentsuccess_finalreport

Références

Bengio, Y. (2000). Gradient-Based Optimization of Hyperparameters. *Neural Computation, 12*(8), 1889-1900. doi:10.1162/089976600300015187

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1-127. doi:10.1561/2200000006

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). *Greedy layer-wise training of deep networks*. Paper presented at the Proceedings of the 19th International Conference on Neural Information Processing Systems, Canada.

Deng, L., Hinton, G., & Kingsbury, B. (2013, 26-31 May 2013). *New types of deep neural network learning for speech recognition and related applications: an overview.* Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Dertat, A. (2017, 3-Oct-2017). Applied Deep Learning - Part 3: Autoencoders. Retrieved from https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Dozat, T. (2016, May 2016). *Incorporating Nesterov Momentum into Adam.* Paper presented at the ICLR 2016, San Juan, Puerto Rico.

Gandhi, R. (2018). Improving the Performance of a Neural Network. Retrieved from https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol, 5*(9), 1315-1316. doi:10.1097/JTO.0b013e3181ec173d

Pennington, J. S., Richard; Manning, Christopher D. (2014). *GloVe: Global Vectors for Word Representation.* Paper presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

Powers, D. M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. doi:10.1186/1471-2105-12-77

Socher, R. (2014). *Recursive deep learning for natural language processing and computer vision.* (PhD), Stanford University, California, USA.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting* (Vol. 15).

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015, 7-12 June 2015). *Going deeper with convolutions.* Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

van Gerven, M., & Bohte, S. (2018). *Artificial neural networks as models of neural information processing*: Frontiers Media SA.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). *Extracting and composing robust features with denoising autoencoders*. Paper presented at the Proceedings of the 25th international conference on Machine learning, Helsinki, Finland.