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

Nathaniel Lasry[1], Michael Dugdale[1], Jonathan Guillemette[1] et Sameer Bhatnagar[2]

Contexte

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? 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*. S'appuyant sur les recommandations de cette *Commission Parent,* le gouvernement provincial 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 en 1967. Le nombre de cégeps continue à croître de décennie en décennie, 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). Selon un rapport du MELS de 2014, le taux de diplomation au collégial a augmenté depuis la création des cégeps : de 22 % en 1976, 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, notre objectif est d'examiner ce qui peut être fait pour la moitié qui n'obtient pas de diplôme collégial.

Persévérance et réussite

Cet article présente les résultats d'une recherche PAREA de trois ans qui comporte des données colligées sur plus de 122 000 étudiants du collégial inscrits dans les cégeps Dawson, John Abbott et Vanier. L'objectif principal fut d'utiliser des méthodes d'analyses classiques et d'intelligence artificielle pour modéliser les facteurs qui prédisent le mieux la réussite et la persévérance scolaires au collégial. Bien que la persévérance et la réussite scolaires (PRS) sont d'intérêt public et priorisé par le gouvernement, jusqu'à présent, les études sur la PRS s'intéressent principalement à aux écoles primaires et secondaires. Pourtant, le passage au collégial est difficile pour plusieurs étudiants. Ils se retrouvent exposés à un nouveau système d'éducation simultanément plus flexible et plus contraignant. Plusieurs ont du mal à s'adapter au rythme collégial. 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 affectés au problème de la PRS et plusieurs facteurs influençant la PRS ont été identifiés. Ce que nous présentons 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

1- CEGEP John Abbott College
2- CEGEP Dawson College

décrochage avant qu'il se produise. Dans un deuxième 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 produisons des classificateurs puissants qui permettent d'identifier les étudiants les plus à risque.

Riches en données, pauvres en information

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 d'étude de l'élève, etc. Les établissements conservent aussi des données démographiques comme le sexe, l'âge et même le code postal; un indicateur du statut socioéconomique (Demissie, Hanley, Menzies, Joseph et Ernst, 2000; Deonandan, Ostbye, Tummon, Robertson et Campbell, 2000). 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. 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. 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.

Prédicteurs de persévérance et de réussite

L'objectif premier de cette étude est de maximiser la persévérance et la réussite scolaires. Nous utilisons le forage de données (connu sous le nom de « data-mining » en anglais), une méthode puissante et relativement nouvelle pour découvrir des structures émergentes dans de grands ensembles de données. Il s'agit d'un domaine interdisciplinaire à l'intersection de « l'apprentissage machine » et des statistiques appliquées. 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. 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.

En ce qui concerne la PRS, les études précédentes nous indiquent 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. Bien que certains liens sont prévisibles, comme les liens entre la présence en classe et l'achèvement des devoirs sur la réussite d'un cours, 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.

Les questions qui ont orienté la création de nos modèle informatiques sont de type : quelles variables permettent de prédire les étudiants les plus susceptibles d'échouer ou d'abandonner un cours, de changer de programme ou encore de décrocher complètement? Une des limitations de notre approche est qu'il nous est actuellement difficile à définir le décrochage avec certitude. Nous opérationnalisons le décrochage comme l'inscription d'un étudiant lors d'une session et son absence d'inscription à la session suivante. Cependant, cet étudiant 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 trois 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. Par conséquent, nos approches portent donc sur les prédicteurs de réussite, car celui-ci est facilement mesurable via les indicateurs de graduation.

Échantillonnage et données descriptives

Nous examinons des données historiques anonymisées à l'aide de clés unidirectionnelles de cryptage (SH-256) nous empêchant d'identifier un étudiant. Les données démographiques nous permettent de quantifier les caractéristiques de la population étudiante des trois collèges participant (John Abbott, Dawson et Vanier). Notons que ce sont tous des cégeps anglophones à Montréal qui ensembles représentent la majorité de la population étudiante anglophone du Québec. Notons aussi que nous focalisons notre attention d'abord 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. Cette analyse représente les étudiants les plus similaires aux 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.

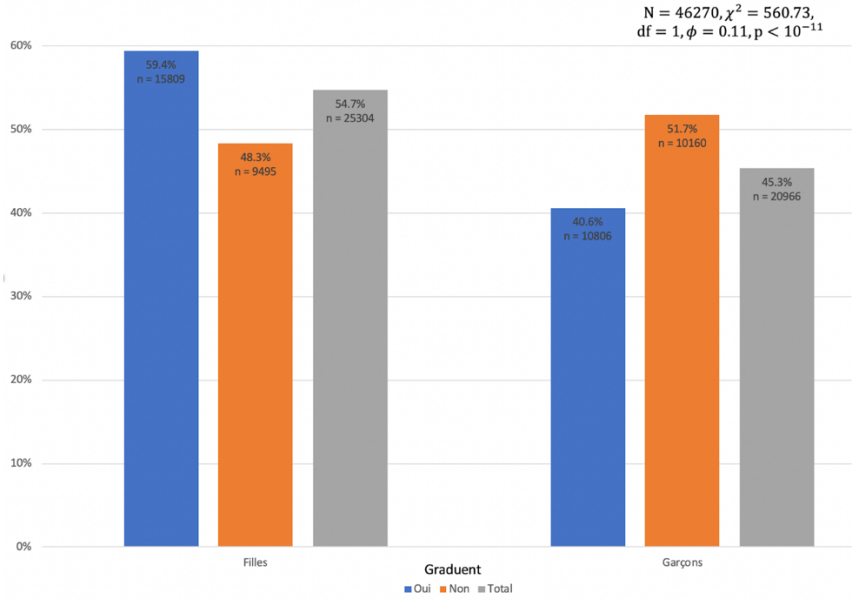

Figure 1. Taux de réussite et d'échec par sexe

La figure 1 montre que les filles graduent en proportions statistiquement plus grande ($p<0.001$) que les garçons avec 59.4% des gradués qui sont filles contre 40.6% qui sont des garçons. Bien que non indiqué dans le graph, notons 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. 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.

Nous examinons ensuite l'impact des notes sur le taux de graduation. Échouer s'avère être un indicateur clair de décrochage. En effet, nous obtenons 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%. Si nous prenons une moyenne de 75% comme seuil pour évaluer un Rapport de Cote ("Odds Ratio") 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%. Cependant, le portrait n'est pas aussi clair pour ceux ayant une moyenne générale au-dessus de 60% ou même de 75% puisque cela n'assure pas qu'un étudiant graduera. Nous analysons la distribution des notes de ceux qui graduent versus ceux qui décrochent. La figure 2 montre qu'à priori, il n'existe quasiment pas de différences dans les notes de ces deux groupes.

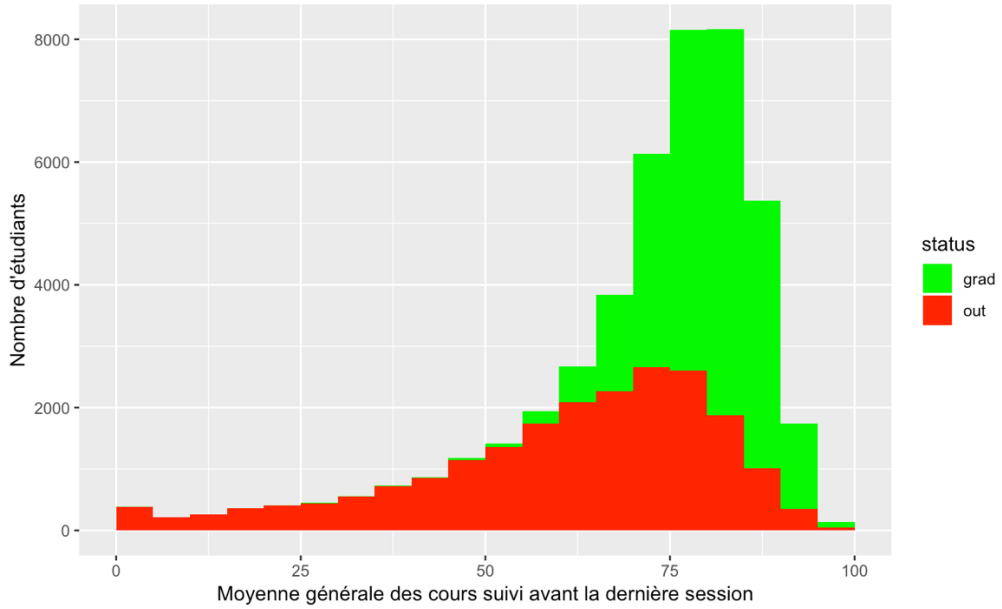

Figure 2. Distribution des moyennes générales pour gradués versus décrocheurs

Un des résultats marquant de la première phase de notre étude est que la distribution de notes, ou plus spécifiquement de la moyenne générale, est très similaire à celle des gradués (figure 2). Une possibilité est que le taux de graduation serait influencé par l'absence de cours échoués.

Bien que les échecs à répétition prédisent le décrochage, nos donées montrent que plus de 30% des étudiants qui décrochent n'ont en fait aucun échec à leur dernière session et que la grande majorité d'entre eux (27.6%) ont des moyennes générales de plus de 75%. Avoir une moyenne générale au-dessus de 60% ou même de 75% n'assure pas qu'un étudiant graduera. Nous retenons donc deux points importants. Premièrement, l'échec d'au moins un cours et le pauvre rendement scolaire en général sont des facteurs de risques significatifs pour le décrochage. Deuxièmement, une fraction importante des étudiants n'ayant aucun de ces facteurs de risque reliés au rendement scolaire décrochent néanmoins.

Prédire le décrochage

Nous examinons si des informations colligées d'étudiants en cours de session seraient utiles dans la prédiction du décrochage. Les évaluations de mi-session (EMS) 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 autres aspect du cours. Maintenant obligatoires dans plusieurs collèges et universités, certaines d'entre elles sont maintenant personnalisables par l'enseignant. Un autre usage des 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. Les EMS permettent aussi à l'enseignant de faire savoir à l'élève si, en fonction de son rendement actuel, l'élève est susceptible de réussir (« Réussite ») ou d'échouer (« Échec ») à l'examen ou se situaient quelque part au milieu (« À risque »). Mais en général, les ÉMS ne sont pas perçues comme des instruments précis, en partie parce qu'à notre connaissance, leur efficacité n'a jamais été documenté. Certains des trois collèges pour lesquels nous avons obtenu des données utilisent les ÉMS comme moyen d'identifier les élèves à risque de décrocher. Ayant déjà plusieurs 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'ont jamais été analysées auparavant, en partie parce que celles-ci n'existaient pas de façon uniforme dans tous les cégeps. Avec notre 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 le Québec n'ait eu à sa disposition.

Nous utilisons des régressions logistiques pour modéliser la prochaine étape qu'un étudiant prendra dans son parcours (réinscription à la prochaine session ou non). 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. Examinant la première session au collégial, nous cherchons à prédire les résultats en matière de décrochage à l'aide d'une régression logistique en utilisant des prédicteurs comprennant : 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 côtes et les intervalles de confiance de 95 % ainsi que les cotes Z et les valeurs de probabilité (P) qui ont été calculées (à l'aide du test de Wald). Notons que dans les graphiques de forêt ci-dessous (figure 3), une variable est significative quand la barre d'erreur ne coupe pas l'unité.

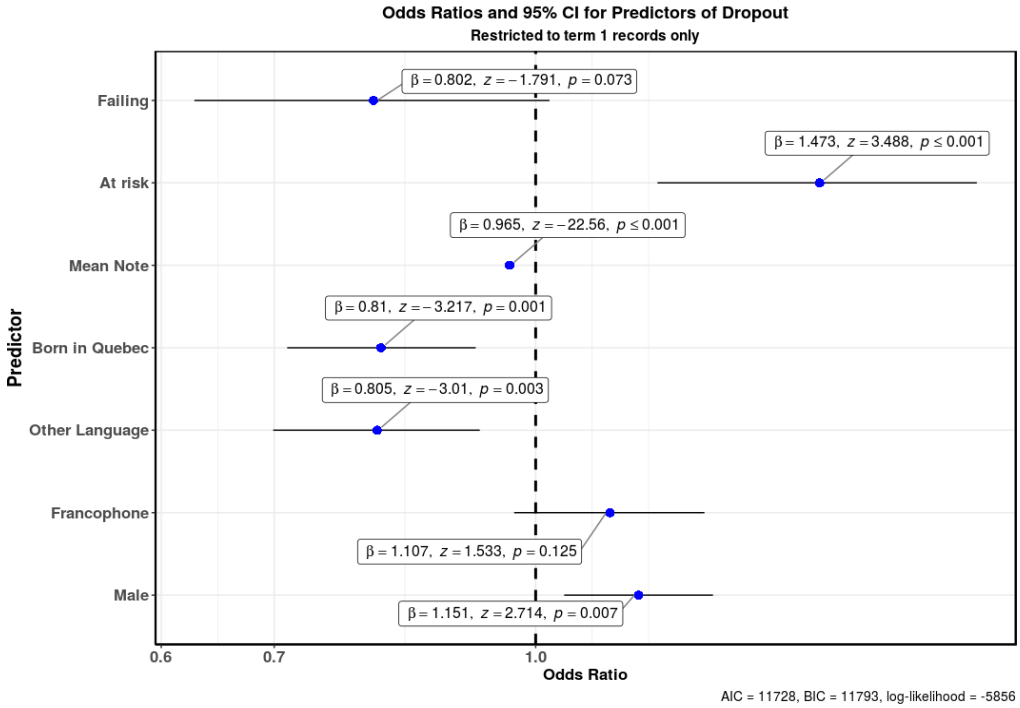

Figure 3. Régression logistique de paramètres à la première session

La figure 3 ci-dessus examine seulement les élèves de première session. Parmi ces élèves, ceux ayant une majorité d'évaluations portant 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 et les allophones sont tous deux près de 20 % moins susceptibles de décrocher. Par ailleurs, les garçons dès leur première session ont 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 utile que le fait d'être un garçon. De plus, ces paramètres changent à chaque session et ne sont pas fixes comme le sexe ou la langue maternelle. Examinons maintenant à la deuxième et la troisième session pour voir si ces effets maintiennent leur puissance prédictive. La figure 4 ci-dessous montre les résultats de la régression logistique pour les EMS de la seconde session tandis que la figure 5 montre les résultats de la régression pour la troisième session.

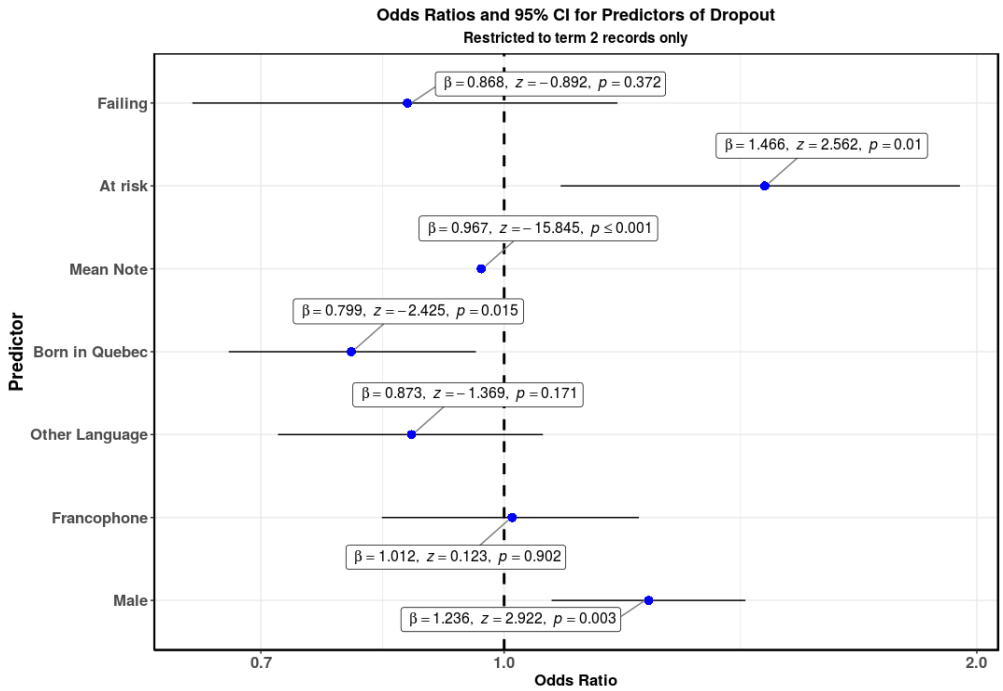

Figure 4. Régression logistique de paramètres à la seconde session

Les figures 4 et 5 montrent les résultats de régressions logistiques faites sur les données de seconde et troisième session. On observe dans les deux cas que les variables les plus stables pour la prédiction du décrochage sont une mention d'être « à risque d'échouer », d'être un garçon et d'être né à l'extérieur du Québec. Nos résultats suggèrent donc que les critères permettant d'identifier les élèves à risque de décrochage doivent être repensés et possiblement ajustés session d'une session à l'autre.

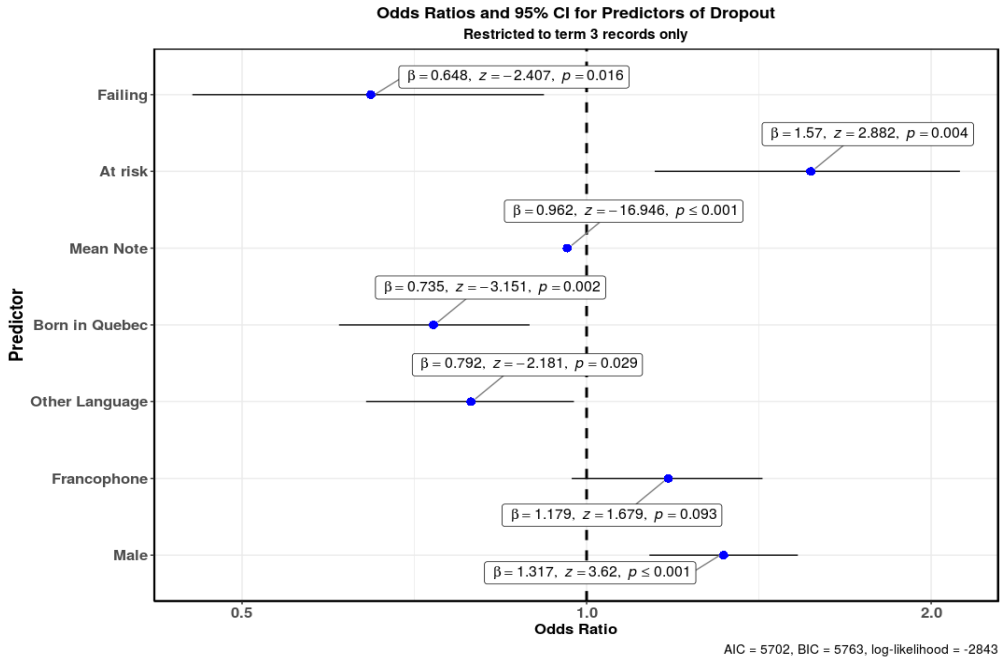

Figure 5. Régression logistique de paramètres à la troisième session

Ayant une méthode par laquelle il est possible de mieux prédire quels élèves décrocheront lors de la prochaine session, il est possible d'évaluer comment cette nouvelle méthode se compare à la méthode traditionnelle utilisée actuellement par les cégeps. Dans tout problème de prédiction, il est important de définir quel paramètre doit être optimisé. Comme nous visons la prédiction du décrochage, on préfère 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 il est possible de produire une matrice 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). Nos résultats de la matrice de confusion indiquent que notre modèle actuel prédit trois fois mieux la probabilité de décrocher que les méthodes actuelles. Il y a néanmoins un prix à payer, soit que nos modèles ne prédisent pas aussi bien qui poursuivra ses études. Cependant, on 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.

Réussite, persévérance et Intelligence Artificielle

La réussite et la persévérance scolaires peuvent aussi faire l'objet d'une étude par le biais d'une stratégie d'exploration de données par l'apprentissage machine. Contrairement aux méthodes statistiques précédentes, l'apprentissage machine est une méthode non fondée sur des principes pour la découverte de modèles prédictifs. Plutôt que d'identifier les facteurs causaux et, par la suite, concevoir un modèle statistique qui tient compte de ces liens sous-jacents, l'apprentissage machine consiste à entraîner un système informatique qui améliore progressivement sa performance prédictive (c.-à-d., « apprend ») sans être explicitement programmé. Les stratégies d'apprentissage machine comportent souvent l'utilisation de réseaux de neurones artificiels (RNA). Les RNA sont des systèmes informatiques conçus pour reproduire approximativement la connectivité entre neurones biologiques et leur activation dans les cerveaux biologiques.

L'objectif de cette phase du projet était de concevoir un système d'apprentissage machine qui pouvait prévoir 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. Les données à notre disposition sont les dossiers historiques complets obtenus des trois cégeps participant à l'étude (Collège Dawson, Collège Vanier et Collège John Abbott). 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. Compte tenu du vaste ensemble de données à notre disposition, nous limitons 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 %) retenus pour la validation du modèle
- 30 706 dossiers (25 %) utilisés pour tester le modèle.

Nous avons ensuite créé deux classificateurs, un pour la réussite et l'autre pour la persévérance. La figure 6 présente les courbe ROC des classificateurs de réussite (à droite) et de persévérance (à gauche) respectivement. Les aires sous les courbes de ces courbes ROC ont été estimée à plus de 0,90 dans les deux cas, indiquant une performance globale « exceptionnelle ».

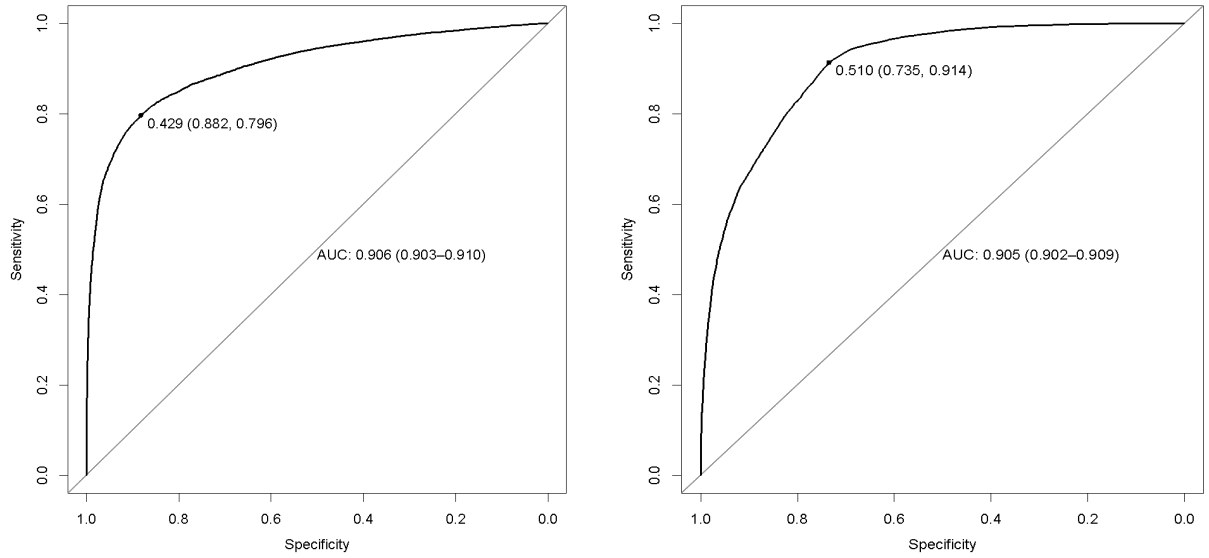

Figure 6. Courbe ROC pour classificateur de réussite (droite) et persévérance (gauche)

Trouver le seuil de probabilité optimal pour identifier les élèves susceptibles d'échouer (classificateur de réussite) nécessitait de découvrir le point maximisant un paramètre statistique (J de Youden). Nous obtenons un seuil optimal qui correspond à une spécificité de 0,882 ± 0,005 et une sensibilité de 0,796 ± 0,007. Ainsi cette catégorisation éclairée se produit avec une probabilité de 67,8 %. Ces résultats indiquent que, malgré une précision catégorielle relativement faible (65,5 %), la performance du classificateur est adéquate. Ce classificateur de réussite est donc viable pour un signalement précoce des élèves à risque d'échouer une partie significative des cours auxquels ils sont inscrits.

En ce qui concerne le classificateur de persévérance la catégorisation du diplôme prévu de l'élève ne comporte qu'une seule catégorie attribuée pour l'abandon des études. Nous identifions la probabilité prédite d'abandon des études et la comparons aux résultats observés. Nous obtenons

un seuil optimal pour l'identification des élèves à risque d'abandonner leurs études (selon le critère de Youden) de 0,510. À ce stade, le classificateur présentait une sensibilité de 0,9137 ± 0,0047 et une spécificité de 0,7352 ± 0,00635. Le signalement de cas à risque mène à une catégorisation éclairée avec une probabilité de 64,9 %. Ici encore, ces résultats indiquent que le classificateur de persévérance arrive à identifier adéquatement les élèves à risque d'abandonner leurs études.

Une réflexion finale pour éviter les dérives

Bien que nous sommes fort satisfaits des résultats de notre étude, tant sur le plan des analyses statistiques communes et des approches d'intelligence artificielle, nous avons certaines réservations quant aux applications de nos résultats. Il y a donc un dernier point sur lequel nous croyons qu'il faut se pencher. Dans plusieurs conversations informelles 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 ou de la race de l'étudiant. Le modèle ayant compilé ces données augmenterait le biais contre 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.

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.