Utilisation des méthodes d'apprentissage automatique pour prédire le décrochage au secondaire.

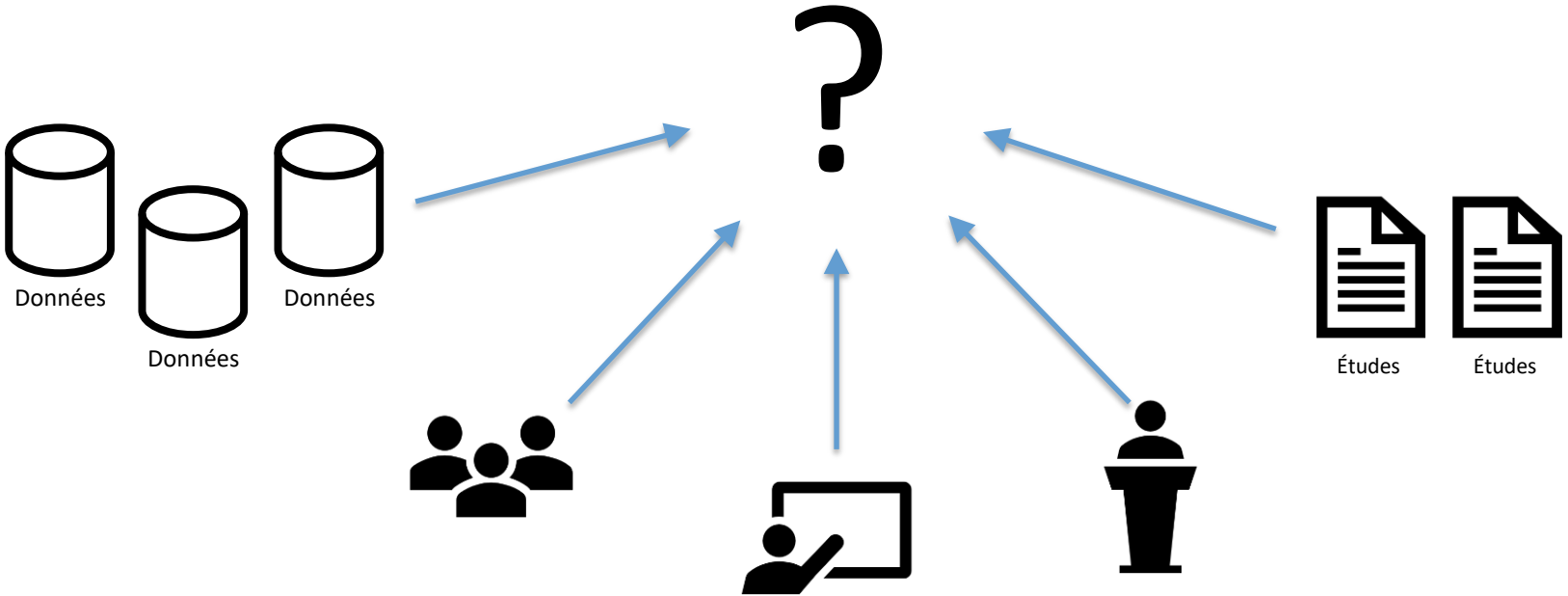
Moyneur Erick MA[1], Daniel Bellemare[2]

[1] StatLog Econometrics Inc., Québec, QC, Canada.
[2] Commission scolaire au Cœur-des-Vallées, Gatineau, QC.

STATLOG

Commission scolaire au Cœur-des-Vallées

Problématique



Problématique

- Le décrochage scolaire est un problème pour toutes les organisations scolaires, d'ici et d'ailleurs.
- Plusieurs méthodes ont été développées et plusieurs études réalisées au cours des 10 dernières années pour identifier les décrocheurs et mieux comprendre le décrochage scolaire.
- Le défi est « d'utiliser » ce savoir pour améliorer le taux de décrochage scolaire.
    - Démocratisation des données et des résultats.
    - Démocratisation des modèles et des méthodes.

Objectif

L'objectif de ce projet est en deux étapes:

1. Établir un modèle fonctionnel, réutilisable et accessible pour aider les intervenants à mieux identifier les élèves les plus à risque de ne pas obtenir un diplôme d'études secondaire (DES) ou un diplôme d'étude professionnel (DEP) ou une qualification.
2. Évaluer *l'amplitude* des variables explicatives et leurs interactions pour mieux comprendre les facteurs de risque de décrochage *propres à notre commission scolaire* et ainsi identifier les mesures à mettre en place pour atténuer le risque.

Approche

- **Économétrique**: construire des modèles probabilistes permettant de décrire des phénomènes économiques ou statistiques.
  - Régression logistique.
- **Science des données** (intelligence artificielle): utiliser des algorithmes qui vont apprendre de leurs erreurs dans le but d'obtenir la meilleure prédiction possible.
  - Apprentissage automatique (Machine Learning):
    - Régression logistique;
    - Arbres de classification et forêts d'arbres décisionnels (Random Forest);
    - Support Vector Machine;
    - Etc.

Données

- La Commission scolaire au Cœur-des-Vallées (CSCV) dessert 26 municipalités dans la région de l'Outaouais.
  - 17 écoles primaires.
  - 5 écoles secondaires.
  - 2 centres de formation professionnelle.
  - 2 centres de formation générale adulte.
- Environ 6000 élèves et 600 adultes.

Données

- Les données proviennent du système LUMIX (2005 à 2018).

- Un élève a 7 ans pour se diplômer ou obtenir une qualification, sinon quoi il sera non-diplômé (décrocheur).

- La diplomation peut être déterminée que pour les élèves qui commencent leur secondaire à la CSCV, même s'ils quittent pour une autre commission scolaire.

- Suivi de la réussite scolaire des élèves qui entrent en 1ère secondaire à la Commission scolaire au Cœur-des-Vallées.

Modèles

Cinq modèles de profondeur 1 et 2 sont estimés:

- Modèle 1ère secondaire
- **Modèle 2ème secondaire**
- Modèle 3ème secondaire
- Modèle 4ème secondaire
- Modèle 5ème secondaire

Taille des Échantillons

Modèle de 2e secondaire utilisant leurs données du 1er secondaire

Année de l'entrée au secondaire	Nombre d'élèves retenus pour les analyses	Nombre d'élèves diplômés	Taux de diplomation	Nombre d'élèves décrocheurs	Taux de décrochage
2007	537	379	71%	158	29%
2008	483	370	77%	113	23%
2009	446	331	74%	115	26%
2010	414	317	77%	97	23%
2011	445	317	71%	128	29%
2012	425	294	69%	131	31%
Total	**2 750**	**2 008**	**73%**	**742**	**27%**

Note: élèves ayant des données complètes dans le système.

Variables de modélisation

- Variables personnelles élèves et familiales
- Variables difficultés d'apprentissage
- Variables assiduité
- Variables comportementales
- Variables académiques
- Variables des mémos internes (messages)

Au total, plus de 450 variables ont été sélectionnées ou construites.

Résultats du modèle économétrique

Variable	Ratio d'incidence	P-value	Variable	Ratio d'incidence	P-value
Père inscrit comme répondant	0.742	0.0058	Nombre de retards	1.011	0.1884
EHDAA + plan d'intervention	1.371	0.0108	Programme international	0.338	0.0028
EHDAA seulement	1.319	0.0558	Nombre de mémos: suivis	1.289	0.0306
Suspension: 1-8 périodes	1.377	0.0428	Nombre de mémos: violence	1.049	0.7687
Suspension: 9+ périodes	1.635	0.0081	Nombre de mémos: drogue	1.844	0.3104
Local de retrait (TES): 1-9 périodes	1.307	0.1403	Min. des notes annuelles	0.977	0.0004
Local de retrait (TES): 10+ périodes	0.581	0.0169	Min. des notes en mathématiques	0.990	0.0416
Sécher les cours: 1-4 périodes	1.256	0.1710	Min. des notes en français	0.984	0.0061
Sécher les cours: 5+ périodes	1.758	0.0528	Min. des notes en géo-histoire	0.994	0.2651
Absences non-motivées	1.014	<.0001			

- Régression logistique du décrochage à partir de la 2ème secondaire en utilisant les données de la 1ère secondaire **(N=2750).**
- Les autres variables explicatives incluent les trois premiers caractères des code postaux.
- EHDAA = élèves handicapés ou en difficulté d'adaptation ou d'apprentissage; TES=technicien(ne)s en éducation spécialisée.

Résultats du Machine Learning

	Performance (AUC de la ROC)				
Méthode	À la fin de la **6ème année**	À la fin de la **1ère secondaire**	À la fin de la **2ème secondaire**	À la fin de la **3ème secondaire**	À la fin de la **4ème secondaire**
Modèle logistique	0.75	0.77	**0.81**	0.81	0.83
Arbre de classification	0.73	0.76	**0.77**	0.76	0.84
Forêt d'arbres décisionnels	0.77	0.77	**0.81**	0.81	**0.86**
Support Vector Maching	0.71	0.66	**0.72**	0.76	0.76
Gradient Boosting	0.77	0.75	**0.79**	0.81	0.85

Note: La méthode de rééchantillonnage itérative a été utilisée pour entrainer les modèles.

Résultats du Machine Learning

Tolérance probabiltité	Décrocheur prédit décrocheur	Décrocheur prédit diplômé	Diplômé prédit décrocheur	Diplômé prédit diplômé	Sensibilité	Précision
0	131	0	294	0	100.00%	30.82%
0.1	124	7	170	124	94.66%	58.35%
0.2	100	31	105	189	76.34%	68.00%
0.3	74	57	67	227	56.49%	70.82%
0.4	58	73	49	245	44.27%	71.29%
0.5	35	96	32	262	26.72%	69.88%
0.6	29	102	21	273	22.14%	71.06%
0.7	17	114	14	280	12.98%	69.88%
0.8	9	122	8	286	6.87%	69.41%
0.9	3	128	1	293	2.29%	69.65%
1	0	131	0	294	0.00%	69.18%

Note: Prédiction du décrochage à la fin de la 1ère secondaire en utilisant le modèle logistique et les données de 2012 (N=425).

Conclusions

- Le modèle économétrique nous apprend que la situation familiale, les difficultés d'apprentissage, le succès scolaire, l'assiduité en classe et les problèmes de comportement sont d'importants facteurs de risque de décrochage.

- Ces résultats sont similaires à ceux de la littérature.

- Parmi les autres facteurs identifiés dans la littérature, on compte le sexe*, la santé mentale, la situation économique, l'encadrement scolaire, le sentiment d'appartenance et l'engagement scolaire.

Note: Le sexe n'est pas inclus dans la régression économétrique, car il est fortement corrélé avec d'autres variables.

Références: Fortin, L., Royer, É., Potvin, P., Marcotte, D. et Yergeau, É., 2004. La prédiction du risque de décrochage scolaire au secondaire: facteurs personnels, familiaux et scolaires. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *36*(3), p.219; Gendron, M., Mélançon, J., Hébert, MH., et Frenette, E. Rapport de recherche sur la persévérance scolaire en Chaudière-Appalaches. 2012. Section 3: 17-38. Institut de la statistique du Québec (ISQ). 2004. Décrochage scolaire chez les élèves du secondaire du Québec, santé physique et mentale et adaptation sociale : une analyse des principaux facteurs associés. *Zoom santé*. Sep 2014(46).

Conclusions

- Grâce au *machine learning*, il est possible de différencier un décrocheur d'un diplômé dans 86% des cas.

- Il est possible d'ordonner les décrocheurs identifiés selon leur risque de décrochage.

- Il y a un équilibre à trouver entre l'identification du maximum de décrocheurs et l'identification de faux décrocheurs.

- Durant les 7 ans qui séparent la prédiction de sa réalisation, les élèves continuent de recevoir des interventions. Ainsi, les mesures de performance « standards » ne sont pas directement applicables aux modèles.

Et maintenant?

Prochaines étapes:

- Utilisation du traitement automatique du langage naturel (Natural Language Processing) pour tirer profit des commentaires des enseignants et des intervenants.
- Personnalisation des modèles selon l'année scolaire (différents ensembles de variables).
- Mise en place d'une procédure d'évaluation et de suivi de nos prédictions.
- Mise en valeur des données et des modèles.

Dîner-conference de l'Association des économistes québécois
UQAM, 15 avril 2019

Daniel Bellemare bellemare.daniel@cscv.qc.ca
Erick Moyneur emoyneur@statlogeconometrics.com

Ce qui est mesurable peut être amélioré
http://statlogeconometrics.com/

STATLOG

Commission scolaire au Cœur-des-Vallées