Apprentissage machine pour la prédiction de l'attrition: une étude comparative

Louis Geiler [*] [,**], Séverine Affeldt[*] Mohamed Nadif [*],

[*] Université de Paris, CNRS, Centre Borelli UMR9010, 75006, France.
[**] Brigad, 34 Rue du Sentier, 75002, Paris

Résumé. La prédiction du taux d'attrition est une préoccupation économique majeure pour de nombreuses entreprises. Différentes approches d'apprentissage ont été proposées, toutefois le choix à priori du modèle le plus adapté reste une tâche non triviale car extrêmement dépendante des caractéristiques intrinsèques des données d'attrition. Notre étude compare huit méthodes d'apprentissage supervisé combinées à sept approches d'échantillonnage sur treize jeux de données publiques relatifs au désabonnement. Nos évaluations, rapportées en termes d'aire sous la courbe (AUC), explorent l'influence du rééquilibrage et des propriétés des données sur les performances des méthodes d'apprentissage. Nous nous appuyons sur le test de Nemenyi et l'Analyse des Correspondances comme moyens de visualisation de l'association entre modèles, rééquilibrages et données. Notre étude comparative identifie les meilleures méthodes dans un contexte d'attrition et propose une chaîne de traitements générique performante basée sur une approche ensemble.

1 Introduction

La mise en place d'une bonne gestion de la relation client est devenue un sujet crucial pour de nombreuses entreprises qui concentrent notamment leur attention sur la *rétention* des clients. En effet, il apparaît aujourd'hui clairement que les coûts d'acquisition d'un nouveau client peuvent être beaucoup plus élevés que les coûts de rétention d'un client existant (Yang et Peterson, 2004). Jusqu'à récemment, les services du marketing et de l'industrie financière préféraient l'exploitation des méthodes de modélisation statistique pour l'analyse et la prédiction du taux de *désabonnement*, telles que l'analyse de survie, la modélisation par équations structurelles ou l'analyse de la variance. L'étude que nous proposons n'explore pas ces approches traditionnelles et se concentre sur les techniques d'apprentissage machine qui sont de plus en plus utilisées dans le contexte du départ ou du désabonnement des clients.

Notre objectif principal est de comparer plusieurs variantes d'une chaîne de traitements pour l'analyse des désabonnements. Cette chaîne comporte *(i)* une étape de *rééquilibrage* des classes, *(ii)* une phase d'apprentissage supervisé et *(iii)* une procédure d'évaluation robuste (Fig. 1). Une analyse exhaustive de toutes les variantes des algorithmes de cette chaîne n'étant pas envisageable dans le cadre de cet article, nous nous concentrons sur les algorithmes d'apprentissage dans leur version originale. Par ailleurs, le départ d'un client étant un événement

rare, les jeux de données d'attrition présentent un déséquilibre de classes relativement important entre la classe minoritaire, composée des individus *désabonnés*, et la classe majoritaire. Ce déséquilibre induit une dégradation de la performance des classifieurs standards (García et al., 2012) qui peut être aggravée par un chevauchement des classes ou un morcellement de la classe minoritaire en sous-ensembles correspondant à des profils clients différents. Le problème de déséquilibre se retrouve dans de nombreux contextes pour lesquels le mauvais classement des instances minoritaires a un coût important (e.g. fraude à la carte de crédit, défaillance d'équipements de télécommunication ou prédiction de la survie des patients).

Ainsi, nous comparons dans ce travail les performances d'algorithmes d'apprentissage supervisé combinés à des approches de rééquilibrage des classes dans le contexte de la prédiction de l'attrition. Nos résultats nous permettent de formuler des recommandations pratiques et de proposer une approche *ensemble* générique et originale, qui est performante sur un large éventail de jeux de données d'attrition.

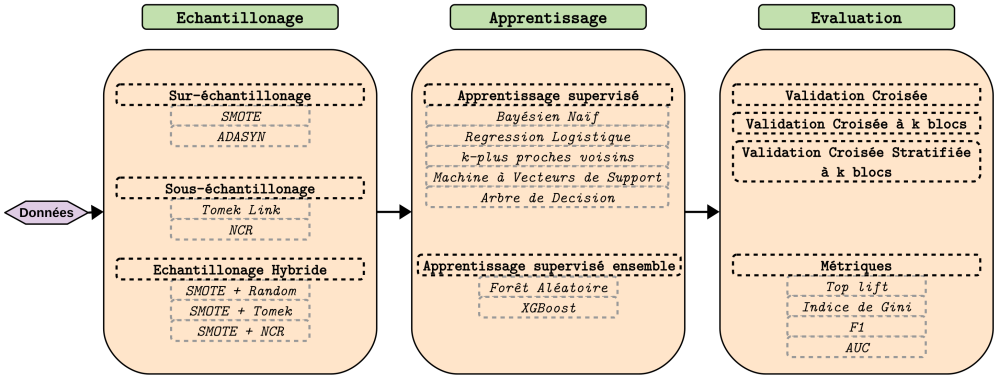


FIG. 1: Chaîne de traitements pour l'analyse et la prédiction d'attrition.

Nous présentons d'abord en Section 2, sept techniques d'équilibrage très répandues. La Section 3 donne un aperçu des huit techniques d'apprentissage supervisé considérées. Nous discutons également de la procédure d'évaluation et des métriques en Section 3.2 avant de fournir les résultats expérimentaux exhaustifs des combinaisons de notre chaîne de traitements en Section 4. Nos expériences sont réalisées *exclusivement* sur des données *publiques* - allant des ressources humaines aux télécommunications, en passant par l'Internet, les abonnements et le *streaming* musical - et reposent sur des librairies Python librement accessibles.

2 Méthodes de rééquilibrage

Les données d'attrition nécessitent l'utilisation de méthodes de rééquilibrage permettant de modifier la distribution des classes. Ces méthodes consistent à introduire des instances dans la classe minoritaire (*sur-échantillonnage*), à retirer des instances de la classe majoritaire (*sous-échantillonnage*) ou à combiner ces deux stratégies (*hybride*). Diverses méthodes de rééquilibrage ont été proposées (Chawla et al., 2002) et plusieurs études tendent à montrer que le sous-échantillonnage est plus performant que le sur-échantillonnage (Drummond et al., 2003).

2.1 Le sur-échantillonage

Cette technique consiste généralement à répliquer des instances de la classe minoritaire ou à en synthétiser de nouvelles. Le *sur-échantillonnage aléatoire* est une approche simple qui sélectionne aléatoirement les instances à répliquer. Cette approche peut fortement dégrader la qualité de la frontière de décision, en répétant par exemple des valeurs aberrantes. D'autres approches plus avancées ont été proposées, telles que *Synthetic Minority Oversampling Technique* (SMOTE, Chawla et al. (2002)) et *Adaptative Synthetic Sampling* (ADASYN, He, H., Bai, Y., Garcia, E., & Li (2008)). SMOTE sur-échantillonne la classe minoritaire en générant des instances synthétiques le long des segments créés par une approche $k-\texttt{NN}$. Les nouvelles instances SMOTE sont donc des observations plausibles qui permettent d'éviter le sur-apprentissage. Mais ces instances synthétiques peuvent être ambiguës en cas de chevauchement fort entre les classes. Des extensions ont été proposées pour surmonter ce problème, dont la très répandue ADASYN, qui génère de manière *adaptative* des instances minoritaires en fonction de leur distribution. Ainsi, beaucoup d'instances sont générées dans les régions de l'espace des caractéristiques où la densité d'observation est faible, et inversement.

2.2 Le sous-échantillonage

Le sous-échantillonnage supprime des instances de la classe majoritaire ou sélectionne un sous-ensemble. Le *sous-échantillonnage aléatoire*, une approche simple qui supprime aléatoirement des instances, peut conduire à la suppression d'instances importantes. Des stratégies plus avancées ont été proposées, telles que *Neighborhood Cleaning Rule* (NCR, Laurikkala (2001)) et Tomek Links (Tomek, 1976). NCR combine deux *règles* qui éliminent de la classe majoritaire les instances redondantes et ambiguës. La première règle, *Condensed Nearest Neighbor* (CNN, Hart (1968)), sélectionne un sous-ensemble d'instances de la classe majoritaire ne pouvant être classées correctement. Ces instances sont considérées comme pertinentes pour l'apprentissage. La deuxième règle, *Edited Nearest Neighbors* (ENN, Wilson (1972)), supprime les instances ambiguës via une approche $k-\texttt{NN}$. Une instance de classe majoritaire mal classée par ses voisins est supprimée du jeu de données. Une instance de classe minoritaire mal classée par ses voisins de classe majoritaire implique la suppressions de ces mêmes voisins. Tomek Links s'appuie sur CNN et identifie les paires d'instances *cross-class*. Ce sont les paires composées d'une instance de la classe majoritaire et d'une instance de la classe minoritaire identifiées comme plus proches voisins. Les instances majoritaires qui appartiennent aux Tomek Links sont bruitées et doivent être supprimées.

2.3 Les stratégies Hybrides

Diverses combinaisons de méthodes de sur- et sous-échantillonnage ont été proposées afin d'équilibrer les données en améliorant la séparation des classes. Une méthode hybride simple consiste à combiner SMOTE avec un *sous-échantillonnage aléatoire*. Chawla et al. (2002) ont montré que cette combinaison donne de meilleurs résultats que le simple sous-échantillonnage. Une combinaison plus complexe, proposée par Batista et al. (2003), combine SMOTE avec Tomek Links.

3 Apprentissage, évaluation et métrique

3.1 Les techniques d'apprentissage supervisé

Nous donnons dans cette section un aperçu des techniques d'apprentissage supervisé que nous avons intégrées dans notre chaîne de traitements, en raison notamment de leur utilisation croissante dans le contexte de l'attrition. Nous considérons uniquement les approches qui ne comportent pas de pondérations pour compenser le déséquilibre des classes et choisissons plutôt d'atténuer ce déséquilibre avec des approches de rééquilibrage. Ainsi, pour cette étude, nous comparons les performances de sept techniques d'apprentissage supervisé dont deux approches ensemble. Les modèles considérés sont le classifieur bayésien naïf (`Gnb`, John et Langley (2013); Hand et Yu (2001)), la Régression Logistique (`LR`), le k-Nearest Neighbors ($k-$`NN`), la Machine à Vecteur de Support (`SVM`, avec et sans kernel, Vapnik (1998)) et les arbres de décision (`DT`, Breiman et al. (2017); Hastie et al. (2009)).

Nous considérons aussi deux méthodes ensembles, *Random Forest* (`RF`, Hastie et al. (2009)) et *eXtreme Gradient Boosting* (`XGBoost`, Chen et Guestrin (2016)), qui sont généralement performantes sur des données d'attrition (Chen et al., 2004; Zhao et al., 2018). `RF` s'appuie sur le *bagging* (Breiman, 1996) et construit $\mathcal{C}$ arbres de décisions *profonds* à partir de $\mathcal{C}$ ensembles d'entraînement obtenus par *bootstrap*. Leurs prédictions sont combinées par un vote majoritaire. Un défaut connu du bagging est la tendance pour les classifieurs à être corrélés car ils partagent le même ensemble de propriétés. `RF` décorrèle les arbres en les contraignant à apprendre sur un sous-ensemble des propriétés choisies de façon aléatoire. `XGBoost` intègre la méthode du *boosting* qui, comme le bagging, combine les résultats de plusieurs classifieurs. Cependant, dans la stratégie de boosting, chaque modèle tente de minimiser les erreurs du modèle précédent. Les variantes bien connues du boosting sont *Adaboost*, *gradient boosting* et *stochastic gradient boosting*. Au lieu d'ajuster les poids comme *Adaboost*, la variante *gradient boosting* optimise une fonction de coût, tandis que la stratégie *stochastic gradient boosting* ajoute des observations et un échantillonnage de variables à chaque itération. `XGBoost` est l'implémentation la plus largement utilisée pour le boosting.

3.2 Les procédures d'évaluation

Un procédure simple d'évaluation est le *holdout set*, où un sous-ensemble de données non utilisé pour l'apprentissage est utilisé pour évaluer les prédictions du modèle entraîné. Un inconvénient est qu'une partie des données est *perdue* pour l'entraînement. La *validation croisée* permet de résoudre ce problème en définissant un ensemble d'entraînement et un ensemble de validation, puis en intervertissant ces ensembles avant de combiner les deux scores d'évaluation. Cette idée peut être étendue à plusieurs sous-ensembles ou *folds*. Les données sont divisées en K *folds* de taille équivalente et le modèle est entraîné sur $K-1$ *folds*. L'erreur de prédiction du modèle est ensuite calculée sur le K^{th} sous-ensemble. Cette stratégie est répétée K fois avant de combiner les K estimations. Il s'agit de la K*-fold cross-validation* (avec typiquement $K = 5$ ou 10). Cette validation n'est pas appropriée pour les données d'attrition en raison du fort déséquilibre. Il faut dans ce cas produire des *folds stratifiés*, garantissant que chaque *fold* respectera la distribution initiale des classes. La *stratified* K*-fold cross-validation* ($K = 5$) est la stratégie retenue pour nos expériences.

3.3 Les métriques

Le *top decile-lift* et le *coefficient Gini* sont les mesures d'évaluation préférentiellement utilisées par les services marketing pour évaluer les modèles prédictifs. Le *lift* considère les instances dans l'ordre de leur probabilité d'être dans la classe minoritaire. Si on se concentre sur les 10% des clients les plus risqués, le décile supérieur du lift donne le ratio entre la proportion de clients désabonnés dans le segment risqué, $\pi_{10\%}$, et la proportion totale de désabonnements dans l'ensemble de validation, π, $lift_{10\%} = \hat{\pi_{10\%}}/\hat{\pi}$. Cette mesure évalue si les clients prédits comme risqués le sont réellement. Le coefficient de Gini prend en compte les clients risqués et moins risqués. En apprentissage automatiques, le score $\mathbf{F_1}$ et l'aire sous la courbe (AUC) sont deux métriques recommandées dans le contexte de l'attrition. $\mathbf{F_1}$ est la moyenne harmonique de la *Precision* et du *Recall*. La *Precision* estime la capacité du modèle à obtenir des *vrais* positifs parmi ses prédictions positives. Cette mesure est complémentaire du *Recall* qui évalue la capacité du modèle à récupérer les *vrais* positifs. L'AUC nécessite d'exprimer la performance du modèle par une courbe *Receiver Operating Characteristic* (ROC) qui donne le taux de vrais positifs en fonction du taux de faux positifs pour une série de seuils de décision. Elle fournit donc une mesure de performance agrégée pour tous les seuils de classement possibles. L'AUC, adaptée aux jeux de données déséquilibrés, est la mesure retenue pour nos expériences.

4 Expériences sur les données publiques

Nos 13 jeux de données publiques ont des pourcentages d'attrition allant de 0.07% à 0.50% (Table 1, *%churn*), et accessibles en ligne (Table 1, *Accès*). Nous avons conservé les paramètres par défaut - fournis dans les packages Python `scikit-learn` (0.23.2) et `xgboost` (1.0.2) - pour les 8 techniques d'apprentissage considérées ; Régression logistique (`LR`), Machine à vecteurs de support linéaire (`SVM`) et avec fonction de base radiale (`SVM-rbf`), Classifieur bayésien naïf (`Gnb`), Forêt aléatoire (`RF`), Arbre de décision (`DT`), eXtreme Gradient Boosting (`XGBoost`) et K-plus proche voisins (k-`NN`). Ces approches sont évaluées sans et avec rééquilibrage (sur-/sous-échantillonage et stratégies hybrides ; voir Fig. 1, *Enchantillonage*).

TAB. 1: Jeux de données publiques pour l'attrition et accès en ligne.

Nom du jeu de données	**Accès**	Instances	Attributs	Attributs *numériques* [1]	%attrition	$\frac{attrition}{nonattrition}$
KDD-Cup 2009 small	K2009	$50,000$	230	$1,039$	0.07	0.08
WDSDM Cup 2018	KKbox	$970,960$	49	56	0.09	0.10
MLC churn	UCI	$5,000$	20	21	0.14	0.16
IBM Employee Attrition	HR	$1,470$	34	86	0.16	0.19
Telco-Europa	TelE	$190,776$	19	26	0.16	0.19
Newspaper	News	$15,855$	18	307	0.19	0.23
Bank	Bank	$10,000$	12	16	0.20	0.25
Mobile	Mobile	$66,469$	65	65	0.21	0.27
IBM Telco Churn	TelC	$7,043$	20	34	0.27	0.37
Cell2Cell	C2C	$71,047$	71	75	0.29	0.41
Membership Woes	Member	$10,362$	14	26	0.30	0.43
South-asian	SATO	$2,000$	13	29	0.50	1
DSN-telecom	DSN	$1,401$	15	32	0.50	1

(1) Avant d'ajuster un modèle, chaque catégorie des variables catégorielles est convertie en une variable binaire.

4.1 Comparaison des algorithmes de classification

Nous avons évalué la prédiction de l'attrition pour toutes les combinaisons de la chaîne de traitements de la Figure 1. L'évaluation suit une validation croisée stratifiée à 5 blocs. Les résultats sont exprimés en AUC sans échantillonnage, avec différentes approches de sur-/sous-échantillonnage et d'échantillonnages hybrides. La Table 2 est un extrait des résultats obtenus avec les 8 approches de rééquilibrage. Le rang moyen et l'AUC médian ($\widetilde{AUC}$) de chaque algorithme sont indiqués dans les deux dernières colonnes. Nos expériences indiquent que l'$\widetilde{AUC}$ est globalement peu influencé par le mode de rééquilibrage. Pour `RF`, le rééquilibrage dégrade l'$\widetilde{AUC}$ par rapport aux résultats obtenus sans échantillonnage (de 0.8095 à 0.7862). En moyenne, les performances de `XGBoost` sont légèrement améliorées avec Tomek Links ($+0.0014$) ou SMOTE combinée à NCR ($+0.0048$).

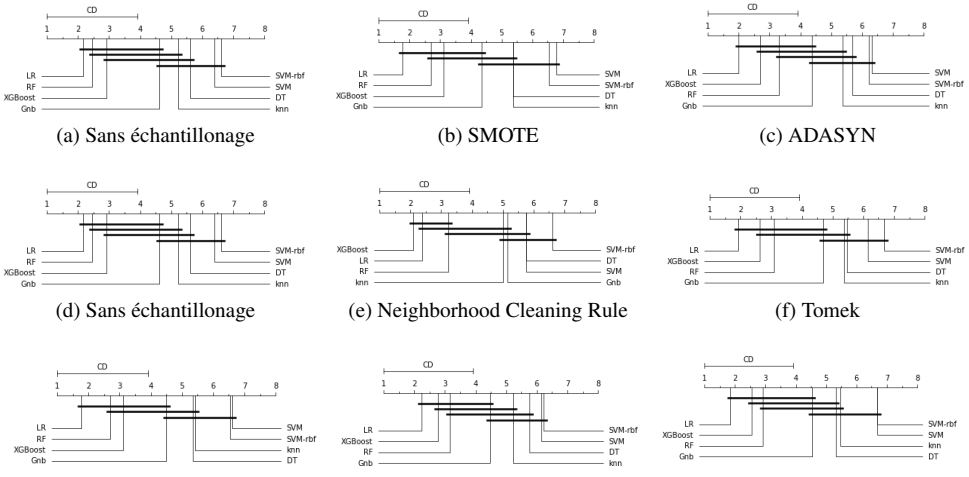
TAB. 2: Un extrait de nos résultats d'évaluation en terme d'AUC.

Données	Bank	C2C	DSN	HR	K2009	KKBox	Member	Mobile	SATO	TelC	TelE	UCI	News	$\overline{Rang}$	$\widetilde{AUC}$
	Approche *No Sampling* (*sans-échantillonnage*)														
k-NN	0.7768	0.4387	0.6576	0.6575	0.5004	0.5835	0.5827	0.7567	0.6900	0.7822	0.8226	0.7731	0.7484	5.23	0.6900
Gnb	0.7166	0.5181	0.6671	0.7442	0.5002	0.6468	0.5914	0.7201	0.7272	0.8245	0.7505	0.8477	0.5655	4.62	0.7166
LR	0.8322	**0.5222**	0.7319	**0.8596**	**0.5135**	0.6763	**0.6146**	**0.9030**	0.7594	**0.8458**	0.7584	0.8244	0.8369	**2.15**	**0.7594**
SVM	0.6645	0.4578	0.6868	0.8091	0.5052	0.5022	0.4874	0.4605	0.7116	0.6498	0.5335	0.5963	0.5958	6.38	0.5958
SVM-rbf	0.7248	0.4656	0.6293	0.4984	0.4989	0.4983	0.5088	0.5463	0.7153	0.6548	0.6098	0.7528	0.6227	6.62	0.6098
DT	0.6908	0.4440	0.7350	0.6053	0.4993	0.5302	0.5462	0.6660	0.6365	0.6555	0.8514	0.8447	0.6754	5.62	0.6555
RF	**0.8506**	0.3518	**0.8590**	0.7867	0.5114	0.6442	0.6130	0.8095	**0.7882**	0.8210	0.9380	**0.9182**	**0.8615**	**2.46**	**0.8095**
XGBoost	0.8216	0.3862	0.8516	0.7993	0.5112	**0.6800**	0.5987	0.7816	0.7396	0.7983	**0.9411**	0.9174	0.8323	**2.92**	**0.7983**
Max-Min	0.1861	0.1704	0.2297	0.3612	0.0146	0.1817	0.1272	0.4425	0.1517	0.1960	**0.4076**	0.3219	0.2960		
	Neighborhood Cleaning Rule *(sous-échantillonnage)*														
k-NN	0.7994	0.4069	0.6634	0.6761	0.5061	0.6099	0.5915	0.7274	0.7028	0.8028	0.8295	0.8052	0.7804	5.00	0.7028
Gnb	0.7460	0.4890	0.6328	0.7350	0.5004	0.6483	0.5886	0.7255	0.7348	0.8205	0.7468	0.8512	0.5672	5.15	0.7255
LR	**0.8313**	0.4985	0.7311	**0.8580**	**0.5146**	**0.6762**	**0.6209**	**0.8867**	**0.7645**	**0.8438**	0.7615	0.8234	0.8371	**2.38**	**0.7645**
SVM	0.6647	**0.5659**	0.7186	0.8332	0.5017	0.5353	0.4915	0.4912	0.7741	0.8007	0.4438	0.6309	0.6727	5.77	0.6309
SVM-rbf	0.7938	0.4533	0.6308	0.4984	0.5033	0.4797	0.5512	0.6077	0.7089	0.7920	0.6260	0.6288	0.6745	6.62	0.6260
DT	0.7327	0.4146	0.7214	0.6194	0.5027	0.5488	0.5693	0.6710	0.6615	0.7136	0.8583	0.8500	0.7306	5.77	0.6710
RF	0.8361	0.3527	0.8173	0.7430	0.5105	0.6397	0.6129	0.7862	0.7631	0.8201	0.9394	0.9145	0.8298	**3.23**	**0.7862**
XGBoost	0.8369	0.3668	**0.8672**	0.7918	0.5149	0.6824	0.6104	0.7745	0.7685	0.8216	**0.9417**	**0.9200**	**0.8399**	**2.08**	**0.7918**
Max-Min	0.1722	0.2132	0.2364	0.3596	0.0145	0.2027	0.1294	**0.3955**	0.1126	0.1302	0.4979	0.2912	0.2727		
	Approche ADASYN *(sur-échantillonnage)*														
k-NN	0.7647	0.4408	0.6576	0.6612	0.5007	0.5899	0.5791	0.6203	0.6900	0.7515	0.8248	0.7791	0.7377	5.38	0.6612
Gnb	0.7865	0.5031	0.6671	0.7241	0.4987	0.6421	0.5958	0.6814	0.7272	0.8311	0.7551	0.8293	0.5661	4.38	0.6814
LR	**0.8315**	0.5171	0.7319	**0.8476**	**0.5137**	**0.6777**	**0.6266**	**0.8848**	**0.7594**	**0.8444**	0.7634	0.8276	0.8309	**2.00**	**0.7634**
SVM	0.6403	**0.5271**	0.6869	0.6768	0.5032	0.5491	0.5015	0.1398	0.7116	0.4093	0.4678	0.5512	0.5467	6.31	0.5467
SVM-rbf	0.7123	0.4734	0.6297	0.5026	0.5053	0.5239	0.5304	0.4864	0.7153	0.6822	0.5559	0.7601	0.6419	6.23	0.5559
DT	0.6865	0.4401	0.7336	0.5814	0.4985	0.5268	0.5479	0.6644	0.6375	0.6546	0.8382	0.8483	0.6876	5.69	0.6546
RF	0.8197	0.3971	0.8038	0.7597	0.4945	0.6107	0.6092	0.7970	0.7494	0.8003	0.9364	0.9112	0.8107	**3.31**	**0.7970**
XGBoost	0.8225	0.3905	**0.8516**	0.7978	0.5013	0.6468	0.5973	0.7937	0.7396	0.7968	**0.9418**	**0.9156**	**0.8328**	**2.69**	**0.7968**
Max-Min	0.1912	0.1366	0.2219	0.3450	0.0192	0.1538	0.1251	**0.7450**	0.1219	0.4351	0.4740	0.3644	0.2861		

L'approche qui bénéficie le plus des stratégies de rééquilibrage est `LR`, avec une augmentation maximale pour l'$\widetilde{AUC}$ de 0.0051 en utilisant NCR. Finalement, les trois meilleures approches - quelque soit le jeu de données et la stratégie de rééquilibrage - sont `LR`, `XGBoost` et `RF`, avec un rang moyen de 2.01, 2.74 et 2.94 respectivement.

Pour certains couples 'jeu de données/technique', on peut observer une amélioration plus importante. Par exemple, la combinaison de SVM & NCR augmente l'AUC de 0.1081 sur *C2C*. L'AUC de XGBoost croît également lors de l'utilisation de l'échantillonnage hybride SMOTE & Tomek Links (de 0.8516 à 0.8694) sur *DSN*. Nous constatons une augmentation de l'AUC de 0.0124 en utilisant SMOTE & NCR sur *Member* avec LR. Par conséquent, bien qu'une amélioration globale de toutes les approches d'apprentissage ne puisse être observée, il existe des améliorations *locales*, en fonction des ensembles de données.

Nous proposons de visualiser les similarités et classement des techniques d'apprentissage via les diagrammes de Différence Critique (DC, Demšar (2006)) basés sur des comparaisons statistiques par paire calculées à partir de tous nos résultats AUC (Fig. 2). Pour ces comparaisons, nous considérons le test de Nemenyi ($\alpha = 0.05$). Des lignes horizontales relient les approches pour lesquelles nous ne pouvons pas exclure l'hypothèse que les rangs moyens sont égaux (Figure 2). Les diagrammes DC traduisent bien le fait que les stratégies d'échantillonnage ont peu d'effet sur le classement des approches d'apprentissage. Nous pouvons aussi facilement voir que RF est l'une des deux meilleures approches si on utilise SMOTE seul, ou en combinaison avec le sous-échantillonnage aléatoire, ou encore sans aucun rééquilibrage.



(a) Sans échantillonage (b) SMOTE (c) ADASYN

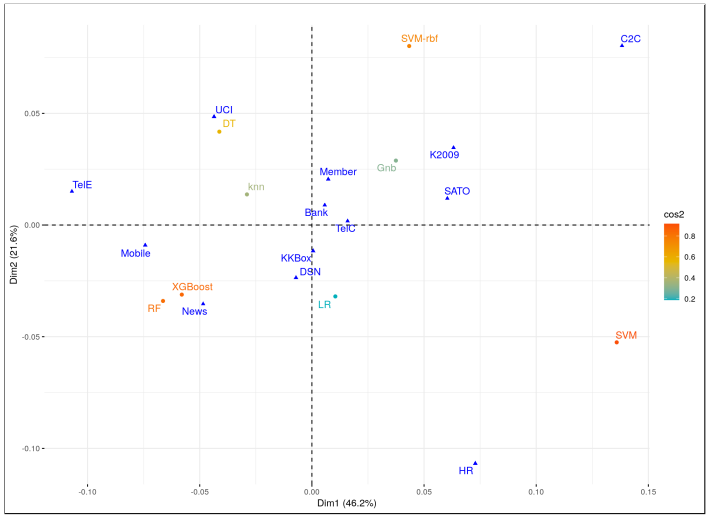
(d) Sans échantillonage (e) Neighborhood Cleaning Rule (f) Tomek

(g) SMOTE + Sous-échantillonage aléatoire (h) SMOTE + NCR (i) SMOTE + Tomek

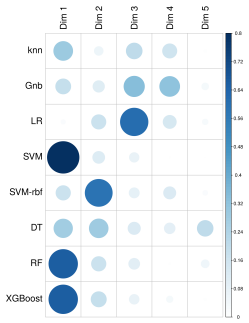
FIG. 2: Similarité et classement des approches via les diagrammes de Différence Critique.

4.2 Analyse des Correspondances entre modèles et données

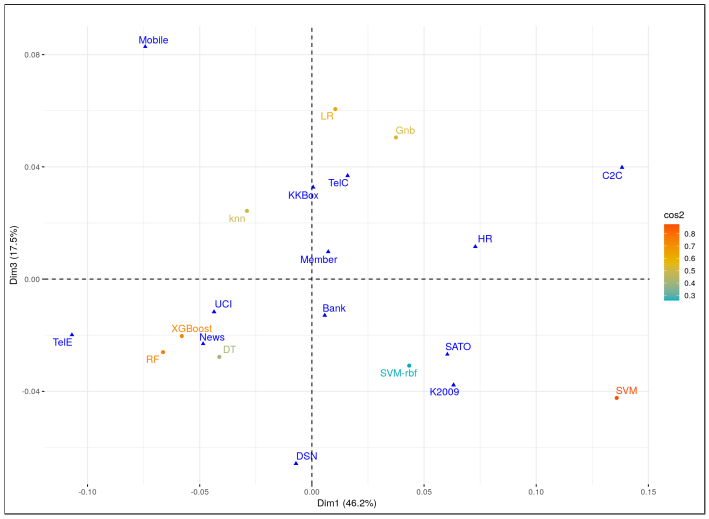
Nous proposons, pour visualiser les relations entre techniques d'apprentissage et jeux de données sur la base des AUC, d'utiliser l'Analyse des Correspondances (CA). La Figure 3 donne un aperçu des résultats de cette analyse et met en avant des comportements ou caractéristiques similaires. Ainsi, la Figure 3(a) indique un comportement similaire entre RF et XGBoost. Elle met aussi en avant la différence avec SVM et SVM-rbf. Le jeu de données *News* est associé à RF et XGBoost. Alors que la meilleure technique pour *Mobile* est LR, ce



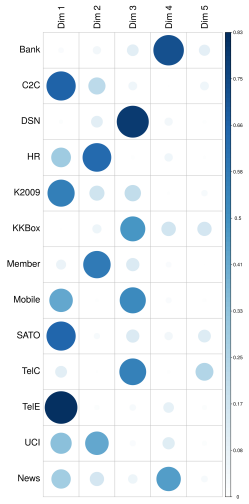
(a) *CA* biplot, dimensions 1 et 2, pas de rééquilibrage



(b) Qualité de Représentation



(c) *CA* biplot, dimensions 1 et 3, pas de rééquilibrage



(d) Qualité de Représentation

FIG. 3: Visualisation des associations entre approches d'apprentissage (sans rééquilibrage) et données d'attrition via l'Analyse des Correspondances et la Qualité des représentations.

jeu de données est dans le voisinage de `RF` et `XGBoost`. Une explication est que `LR` n'est pas correctement représentée par les deux premières dimensions de la CA (Fig. 3(b)). Ce sont donc les deuxièmes meilleures approches, `RF` et `XGBoost`, que nous trouvons dans le voisinage de *Mobile*. De même, *HR* est représenté comme associé à sa deuxième meilleure approche après `LR`, qui est `SVM`. Enfin, *C2C* est associé à sa troisième meilleure approche après `LR` et `Gnb`, qui est `SVM-rbf`. La Figure 3(c) remplace la deuxième dimension par la troisième, offrant une

bonne représentation de LR et Gnb. Les données *UCI*, *News* et *TelE* sont regroupées autour de RF et XGBoost, en accord avec nos résultats. On peut noter que *SATO* et *K2009* sont à *mi-chemin* entre {RF, XGBoost} et {LR, Gnb}. En effet, ces technique donnent les premiers et deuxièmes meilleurs scores AUC pour ces données. De même, *C2C* est *à la croisée* de {LR, Gnb} et SVM, ainsi que *KKBox* qui se situe entre {LR, Gnb} et {RF, XGBoost}. La Figure 3(c) souligne bien les jeux de données préférentiellement liés à LR et Gnb (*C2C*, *HR*, *Mobile*, *TelC*) et les jeux de données préférentiellement liés à RF et XGBoost (*DSN*, *SATO*, *TelE*, *UCI*, *News*). Cette visualisation issue de CA nous conforte dans l'idée qu'une approche ensemble regroupant LR, XGBoost et RF pourrait être performante sur une large partie des jeux de données d'attrition.

4.3 Étude comparative des méthodes ensembles

Les conclusions des sections précédentes motivent l'association de LR, XGBoost et RF dans une approche ensemble pour la prédiction de l'attrition. Plus spécifiquement, nous calculons pour chaque observation la moyenne des probabilités prédites par deux ou trois techniques prises parmi ces trois approches.

La Figure 4 montre, pour chaque stratégie de rééquilibragee, et pour tous les jeux de données, l'AUC pour LR, XGBoost et RF (gris claire), leurs ensembles par paire (orange claire), et la combinaison des trois méthodes (orange foncé). Il apparaît que l'approche ensemble LR|XGBoost|RF est la plus performante, suivie des approches ensemble par paire, LR|XGBoost et LR|RF.

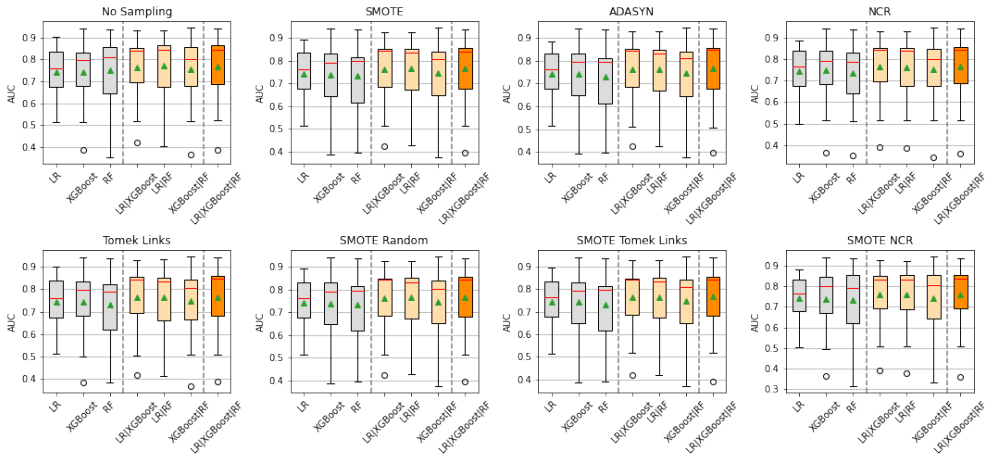


FIG. 4: AUC pour les approches ensemble basées sur les trois meilleurs techniques.

La Table 3 présente l'$\widetilde{AUC}$ sur tous les jeux de données pour les approches ensemble et non-ensemble. Alors que les résultats de XGBoost et RF surpassent ceux de LR ($\widetilde{AUC}$ de 0.7956 et 0.7953 vs. 0.7622), la combinaison des approches XGBoost et RF n'augmente pas significativement l'$\widetilde{AUC}$ (0.8061). En revanche, l'ajout de LR dans l'approche d'ensemble

(LR|XGBoost et LR|RF) augmente significativement l'$\widetilde{AUC}$ (0.8413 et 0.8365 resp.). Globalement, la meilleure approche ensemble est LR|XGBoost|RF, combinant les trois techniques, avec la stratégie de sur-échantillonnage ADASYN ($\widetilde{AUC} = 0.8483$).

TAB. 3: $\widetilde{AUC}$ sur tous les jeux de données pour les approches *ensemble* et *non ensemble*.

Echantillonage	no sampling	SMOTE	ADASYN	NCR	Tomek Links	SMOTE & R.U.	SMOTE & Tomek	SMOTE & NCR	$\widetilde{\widetilde{AUC}}$
LR	0.7594	0.7626	0.7634	0.7645	0.7589	0.7626	0.7628	0.7633	0.7622
XGBoost	0.7983	0.7905	0.7968	0.7918	0.7997	0.7905	0.7939	0.8031	0.7956
RF	0.8095	0.8007	0.797	0.7862	0.788	0.7947	0.7951	0.7911	0.7953
LR|XGBoost	0.8422	**0.8422**	0.8422	**0.8416**	0.8433	0.8422	0.8419	0.8346	0.8413
LR|RF	0.8440	0.8369	0.8339	0.8402	0.8358	0.8334	0.8349	0.8325	0.8365
XGBoost|RF	0.8028	0.8078	0.8115	0.8015	0.8055	0.8047	0.8094	0.8055	0.8061
LR|XGBoost|RF	**0.8443**	0.8401	**0.8483**	0.8413	**0.8468**	**0.8453**	**0.8434**	**0.8374**	**0.8434**

La Table 4 fournit la chaîne de traitements qui produit la meilleure AUC pour chaque jeu de données (colonnes '*Meilleure Chaîne de traitements*', '*Meilleur AUC*'). L'approche ensemble que nous recommandons - LR|XGBoost|RF & ADASYN - fournit un score AUC très proche de celui de la meilleure approche. La seule exception est pour *C2C*, pour lequel l'ensemble LR|Gnb sans rééquilibrage est un meilleur choix ($AUC = 0.5247$). Ainsi, en pratique, nous recommandons l'utilisation de l'approche ensemble LR|XGBoost|RF avec ADASYN pour l'exploration de nouveaux jeux de données d'attrition.

TAB. 4: Approche ensemble vs. meilleure approche individuelle par jeu de données.

	LR|XGBoost|RF & ADASYN	Meilleur AUC	Meilleure Chaîne de traitements
Bank	0.8492	0.8506	no sampling & RF
C2C	0.3962	0.5659	NCR & SVM
DSN	0.8486	0.8694	SMOTE-T.Links & XGBoost
HR	0.8483	0.8596	no sampling & LR
K2009	0.5070	0.5153	SMOTE-NCR & LR
KKBox	0.6778	0.6805	Tomek Links & XGBoost
Member	0.6209	0.6270	SMOTE-NCR & LR
Mobile	0.8788	0.9030	no sampling & LR
SATO	0.7703	0.7882	no sampling & RF
TelC	0.8302	0.8458	no sampling & LR
TelE	0.9408	0.9421	SMOTE & XGBoost
UCI	0.9214	0.9200	NCR & XGBoost
News	0.8574	0.8615	no sampling & RF
$\widetilde{AUC}$	0.8483	0.8506	

5 Conclusion

Cette étude comparative a pour but d'examiner, d'évaluer et de comparer plusieurs approches d'apprentissage machine populaires dans le contexte de la prédiction d'attrition. Elle propose également des analyses et des visualisations éclairantes, et fournit finalement une recommandation générale sur le choix d'une chaîne de traitements pour la prédiction de désabonnement basée sur une approche *ensemble*.

Ainsi, dans ce travail, nous avons présenté les jeux de données d'attrition disponibles publiquement. Ensuite, nous avons introduit les approches d'échantillonnage des données, qui se déploient en trois catégories, à savoir le sur-échantillonnage, le sous-échantillonnage et des stratégies hybrides. Nous avons également discuté les stratégies et les mesures de validation. Enfin, les approches d'apprentissage sont évaluées sur treize ensembles de données soumis à différents traitements de rééquilibrage. Nous avons résumé nos résultats en termes de score AUC. Finalement, nous avons proposé des visualisations synthétiques qui mettent en avant les associations entre classifieurs/méthodes d'échantillonnage et jeux de données. Plus important encore, nous avons présenté une chaîne de traitements générale pour l'analyse des désabonnements basée sur une approche *ensemble* simple qui peut être utilisée avec succès dans la pratique. Cette étude technique constitue donc une bonne référence pour les utilisateurs intéressés par les choix d'approches d'apprentissage machine dans le contexte de la prédiction d'attrition.

Références

Batista, G. E., A. L. Bazzan, M. C. Monard, et al. (2003). Balancing training data for automated annotation of keywords : a case study. In *WOB*, pp. 10–18.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (2017). *Classification and regression trees*. Routledge.

Chawla, N. V., K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer (2002). Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Chen, C., A. Liaw, L. Breiman, et al. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley 110*(1-12), 24.

Chen, T. et C. Guestrin (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research 7*, 1–30.

Drummond, C., R. C. Holte, et al. (2003). C4. 5, class imbalance, and cost sensitivity : why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, Volume 11, pp. 1–8. Citeseer.

García, V., J. S. Sánchez, et R. A. Mollineda (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems 25*(1), 13–21.

Hand, D. J. et K. Yu (2001). Idiot's bayes—not so stupid after all ? *International statistical review 69*(3), 385–398.

Hart, P. (1968). The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory 14*(3), 515–516.

Hastie, T., R. Tibshirani, et J. Friedman (2009). The elements of statistical learnin. *Cited on*, 33.

He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN : Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008. *IJCNN 2008.(IEEE World Congress on Computational Intelligence) (pp. 1322– 1328)* (3), 1322– 1328.

John, G. H. et P. Langley (2013). Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv :1302.4964*.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe*, pp. 63–66. Springer.

Tomek, I. (1976). Tomek Link : Two Modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics SMC-6*, 769–772.

Vapnik, V. (1998). Statistical learning theory wiley-interscience. *New York*.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* (3), 408–421.

Yang, Z. et R. T. Peterson (2004). Customer perceived value, satisfaction, and loyalty : The role of switching costs. *Psychology & Marketing 21*(10), 799–822.

Zhao, Z., H. Peng, C. Lan, Y. Zheng, L. Fang, et J. Li (2018). Imbalance learning for the prediction of n 6-methylation sites in mrnas. *BMC genomics 19*(1), 1–10.

Summary

Attrition rate prediction is a major economic concern for many companies. Different learning approaches have been proposed, however, the a priori choice of the most suitable model remains a non-trivial task as it is highly dependent on the intrinsic characteristics of the churn data. Our study compares eight supervised learning methods combined with seven sampling approaches on thirteen public churn data sets. Our evaluations, reported in terms of area under the curve (AUC), explore the influence of rebalancing and data properties on the performance of learning methods. We rely on the Nemenyi test and Correspondence Analysis as a means of visualization of the associations between models, rebalancing and data. Our comparative study identifies the best methods in an attrition context and proposes a powerful generic pipeline based on an ensemble approach.