


BIENVENUE SUR NOTRE APPLICATION SHINY !

Le but de cette application est de comparer les performances de plusieurs méthodes de classification (Lasso, Adaptive Lasso, Elastic-Net, Random Forest, Boosting).

Lorsque vous lancez notre application, cette première fenêtre apparaît. Elle explique le but de l'application et permet de télécharger **ce document** qui vous en explique les fonctionnalités de l'application en cliquant sur : 



Notre application est composée de 4 onglets principaux :

- **PRÉSENTATION :**

Explications du travail effectué, de son utilité, et de la méthodologie utilisée.

Menu

Présentation

Application

Statistiques descriptives

Modélisation

Comparaison

À propos

Master ESA

Présentation

Le but de ce travail est de se focaliser sur un problème de classification. Plus précisément sur la probabilité de défaut. Dans ce contexte, les algorithmes ou méthodes de classification supervisées abordées (Lasso, Elastic-Net, Adaptive Lasso, Forêts aléatoires, Boosting) doivent être mis en compétition dans le but de sélectionner celui qui présente le plus fort pouvoir de généralisation.

Pour cela, nous allons appliquer ces algorithmes et méthodes à notre jeu de données (Kaggle : Give me some credit). Nous allons chercher à prédire le défaut de nos individus. Le défaut correspond au fait qu'un individu passe au moins 90 jours en défaut sur la période étudiée. Notre jeu de données dispose de 150 000 observations et 11 variables. Dans un premier temps, nous étudions quelques statistiques descriptives ainsi que la répartition de la variable cible et celle de toutes les variables explicatives afin d'avoir une première idée de nos données. Dans un deuxième temps, nous modélisons les cinq méthodes précédemment énoncées sur notre jeu de données et prenons en compte la matrice de confusion, l'aire sous la courbe ROC (AUC) et la précision de chaque modèle sur l'échantillon test. La dernière partie de notre travail consiste à comparer chaque méthode. Afin de ne retenir que la meilleure, nous comparons l'AUC et la précision des modèles mais également celles de l'échantillon d'apprentissage à celles de l'échantillon test afin de détecter un éventuel sur-apprentissage.

- **APPLICATION :**

Cet onglet est composé de 3 sous onglets :

- **Statistiques descriptives :**

Aperçu des statistiques descriptives (affichage de la table de base, de la distribution de la variable cible ainsi que des variables explicatives).

Explications sur les variables en cliquant sur :

Description des variables

Menu

Présentation

Application

Statistiques descriptives

Modélisation

Comparaison

À propos

Master ESA

Statistiques descriptives

Observer les données brutes est une étape primordiale avant de passer à la modélisation des données. Pour cela, affichons le début de notre jeu de données.

ser_delinquency	RU_unsecuredlines	age	nb_3059days	debt_ratio	income_month	nb_creditloan	nb_90days	nb_realEloanlines	nb_6089days
1	0.77	45	2	0.80	9120.00	13	0	6	0
0	0.96	40	0	0.12	2600.00	4	0	0	0
0	0.66	38	1	0.09	3042.00	2	1	0	0
0	0.23	30	0	0.04	3300.00	5	0	0	0
0	0.91	49	1	0.02	63588.00	7	0	1	0
0	0.21	74	0	0.38	3500.00	3	0	1	0

Afin de mieux comprendre le jeu de données, vous pouvez télécharger un PDF expliquant les différentes variables.

Description des variables

Nos données comportaient quelques valeurs manquantes sur deux variables quantitatives. Celles ci ont donc été imputées par la médiane. Cette décision se base sur la proportion faible de valeurs manquantes mais également par le fait que la médiane n'est pas affectée par les outliers contrairement à la moyenne.

○ **Modélisation** (composé des 5 méthodes sous formes d'onglets) :

Explication de la méthode utilisée, choix du paramètre de pénalisation, affichage des coefficients, de la matrice de confusion, ainsi que de l'AUC et de la précision du modèle.

Choix des paramètres de pénalisation grâce au curseur : 

Menu

Présentation

Application

Statistiques descriptives

Modélisation

Lasso

Adaptive Lasso

Elastic-Net

Random Forest

Boosting

Comparaison

À propos

Master ESA

Lasso

Cette méthode est adaptée lorsque le nombre de prédicteurs est plus élevé, voir beaucoup plus élevé que le nombre d'observations. Nous l'utilisons lorsqu'il y a une multicollinéarité, c'est-à-dire lorsque la matrice des observations n'est pas inversible, ou lorsque nous risquons d'avoir un estimateur MCO non unique et un grand risque de sur-ajustement. Nous pouvons aussi l'utiliser lorsque le nombre de prédicteurs est inférieur au nombre d'observations mais est tout de même élevé. Cette méthode sert ainsi à faire une sélection de variables et à éviter le risque de sur-ajustement.

L'hypothèse de base du Lasso est que le vecteur de paramètres est creux ou éclairci, c'est-à-dire que certains paramètres sont égaux à 0, ce qui est une hypothèse raisonnable dans le cadre du nombre élevé de prédicteurs. Pour cela, nous cherchons à trouver le vecteur β qui va minimiser $PRSS(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^p |\beta_j|$ avec $\lambda > 0$.

La sélection de variables permet donc, en plus de diminuer le risque de sur-ajustement et d'améliorer la prévision, de faciliter également l'interprétation du modèle.

Il est important de noter que la sélection effectuée par le Lasso a une efficacité supérieure à celles qu'offrent les méthodes de sélection traditionnelles, telles que le Forward le Backward, que ce soit pour l'invariance suite à la perturbation de l'échantillon d'apprentissage, ou pour l'exercice de prévision. De plus, la condition d'irreprésentabilité doit être satisfaite.

NB: Nous faisons varier λ entre 0 et 0.05. Il peut être supérieur à 0.05 mais avec $\lambda = 0.05$ nous avons déjà la quasi-totalité des paramètres qui sont nuls.

Choix du paramètre

Lambda

0 0.00001 0.010002 0.015003 0.020004 0.025005 0.030006 0.035007 0.040008 0.045009 0.05

Vous avez sélectionné un lambda de : 0.02505.

NB: Il est possible que la matrice de confusion, l'AUC et la précision ne soient pas modifiées par la valeur sélectionnée, car la modification sera trop faible pour être visible au vu des arrondis

Coefficients

Variables	Coefficients
Intercept	0.56
RU_unsecuredlines	0.00
age	-0.03
nb_3059days	0.02
debt_ratio	0.00
income_month	0.00
nb_creditloan	0.00
nb_90days	0.00
nb_realEloanslines	0.00
nb_6089days	0.00
nb_dependents	0.00

Matrice de confusion corrigée

	Prédiction 0	Prédiction 1
Réponse 0	33183	1611
Réponse 1	1429	875

0.7623

Précision

0.6553

AUC

- **Comparaison :**

Comparaison de la performance de chaque modèle.

Menu

Présentation

Application

Statistiques descriptives

Modélisation

Comparaison

À propos

Master ESA

Comparaison des méthodes

Nous pouvons à présent comparer chaque méthode afin de choisir laquelle convient le mieux à notre jeu de données. Pour comparer au mieux les modèles, nous avons déterminé les paramètres optimaux de chacune des méthodes par validation croisée.

Nous comparons les AUC ainsi que les précisions de chaque modèle sur les échantillons tests. Cependant, nous prôtons tout de même attention à celles des échantillons d'apprentissage afin de détecter un possible sur-apprentissage.

Méthodes	AUC.test	AUC.train	ACCURACY.test	ACCURACY.train
Lasso	0.665833823641863	0.665406529090658	0.76823548439269	0.767243868910786
Adapative Lasso	0.665037934082355	0.663936213668934	0.768127661868564	0.766550764148001
Elastic-Net	0.666182395171328	0.666376352660292	0.767885061189282	0.767486455577761
Random Forest	0.889782676054393	0.968521046022424	0.919672219526659	0.979322374576917
Boosting	0.867605339778501	0.868443986714389	0.901153701008141	0.900932225905946

Le meilleur modèle en tous points est le Random Forest car c'est la méthode avec le meilleur AUC ainsi que la meilleure précision. Cependant, si nous comparons les deux critères entre les échantillons d'apprentissage et test nous remarquons un écart important. Cet écart n'est pas présent pour les autres méthodes. Il y a donc du sur-apprentissage avec le Random Forest. Le Boosting pourrait être une bonne alternative car l'AUC et la précision sont proches de celles obtenues avec Random Forest mais d'après les résultats nous ne concluons pas à la présence de sur-apprentissage.

- **À PROPOS :** informations sur l'équipe.



Accès au LinkedIn en cliquant sur :

Menu

Présentation

Application

Statistiques descriptives

Modélisation

Comparaison

À propos

Master ESA

Notre Equipe

Cette application a été réalisée dans le cadre d'un projet de la formation Master économétrie et statistiques appliquées par deux étudiantes de deuxième année.

Ce projet est lié au cours Big Data Analytics (arbres de décisions, méthodes d'aggrégations et méthodes de pénalisations) enseigné par M. Tokpavi.

Pour toute question concernant cette application, n'hésitez pas à nous contacter.

Noeline LEPAIS

Lou DACCORD

En haut à droite, l'icône **Master ESA** permet l'accès direct au site du Master ESA.