Rapport d'analyse du projet

Introduction et méthodologie

Notre objectif principal est focalisé sur **la recherche des variables ayant un impact significatif sur la performance scolaire des étudiants**.
Lors du traitement des données, la première étape a été de changer le type de toutes les variables catégorielles (actuellement de type "*int*") en type "*factor*" afin que toutes nos fonctions traitent correctement les données de ces variables.
Une fois ceci fait, et ayant mis de côté les variables et les observations objectivement pauvres en informations, nous avons **scindé notre jeu de données en deux parties**. Une partie pour entraîner nos différents modèles (~80% du jeu de données), et une autre pour tester l'efficacité de ces derniers (~20% du jeu de données).
Pour pouvoir comparer l'efficacité de nos différents modèles, nous allons construire des **matrices de confusion** grâce aux données de test et calculer la précision de nos modèles en se basant sur ces dernières. Pour construire nos classes pour les prédictions, nous utiliserons un **seuil MAP** (0.5).

Modèles simples

Premièrement, la première idée qui nous vient naturellement serait de mettre de côté les variables non significatives pour notre régression linéaire. Cette idée est très intéressante, et nous avons ainsi vu dans le cours 3 **méthodes de sélections de variables** que l'on peut qualifier de "gloutonnes" : "**Forward selection**", "**Backward elimination**" et "**Stepwise regression**".

Modèle	Forward selection	Backward elimination	Stepwise regression
Matrice de confusion	Prediction True class 0 1 0 440 90 1 60 442	Prediction True class 0 1 0 437 93 1 60 442	Prediction True class 0 1 0 441 89 1 57 445
Précision	0.8546512	0.8517442	0.8585271
Temps écoulé *(sur ma machine)*	35.14 secondes	265.34 secondes	28.89 secondes
Variables principales (dans l'ordre)	-Curricular.units.2nd.sem..approved. -Curricular.units.1st.sem..approved. -Tuition.fees.up.to.date1	-Curricular.units.2nd.sem..approved. -Curricular.units.1st.sem..approved. -Tuition.fees.up.to.date1	-Curricular.units.2nd.sem..approved. -Curricular.units.1st.sem..approved. -Tuition.fees.up.to.date1

Ces méthodes sont dites incrémentales car elles ajoutent ou retirent une variable à la fois et réévaluent le modèle à chaque étape. Et donc, comme p > 30, elles sont très gourmandes en temps de calcul (surtout la "Backward selection"), mais elles fournissent de très bonnes précisions dans les matrices de confusion de nos données test.
Il reste alors pertinent de se mettre à la recherche de modèles moins gourmands en temps de calculs.

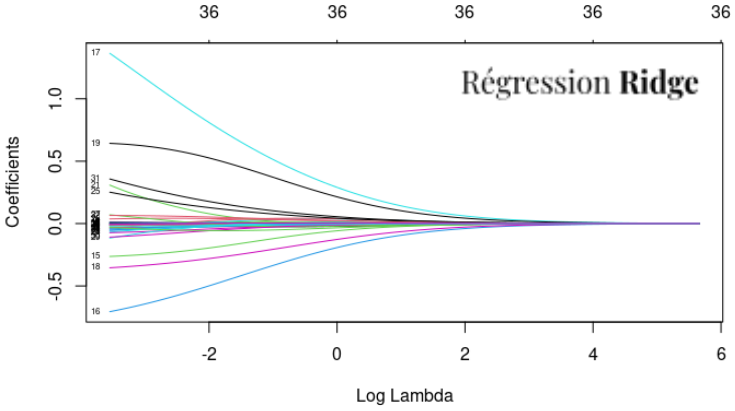
Modèles intermédiaires

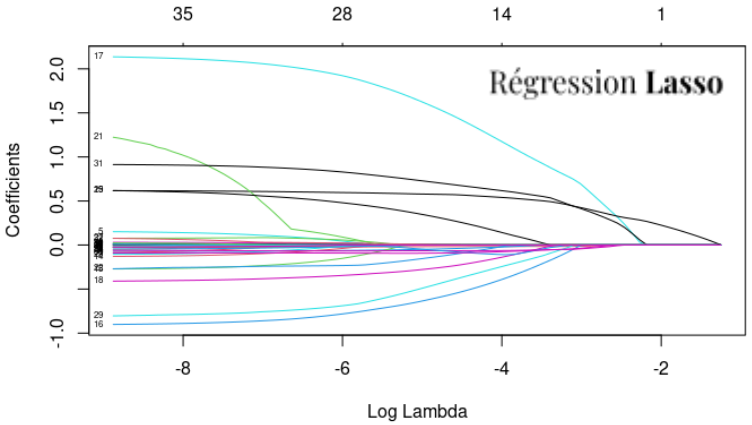
La régression Ridge (L2) et **la régression Lasso** (L1) sont des méthodes de régression pénalisée qui visent à résoudre ces problèmes d'une manière différente. En quelque sorte, elles construisent un modèle en **optimisant tous les coefficients simultanément** avec la contrainte de la pénalité k (ou λ selon les préférences).

Dans ces deux modèles de régression pénalisée, afin d'optimiser l'hyper-paramètre λ, nous avons effectué une validation croisée. La validation croisée nous a permis d'obtenir deux valeurs de λ :

- λ minimum (la valeur de λ qui donne le minimum d'erreur de validation croisée)

❖ λ 1se (la plus grande valeur de λ telle que l'erreur de validation croisée est dans une déviation standard du minimum)





Ces deux valeurs de λ nous permettent de développer deux modèles différents pour chaque régression. Voici ci-contre les différents résultats obtenus :

Modèle	Ridge λ min	Ridge λ 1se	Lasso λ min	Lasso λ 1se
Matrice de confusion	Prediction True class 0 1 0 404 114 1 46 468	Prediction True class 0 1 0 403 115 1 46 468	Prediction True class 0 1 0 418 100 1 54 460	Prediction True class 0 1 0 415 103 1 46 468
Précision	0.8449612	0.8439922	0.85077519	0.8556202
Temps écoulé	0.20 secondes		0.17 secondes	
Variables principales (dans l'ordre)	-Tuition.fees.up.to.date -International -Debtor	-Tuition.fees.up.to.date -Debtor -Scholarship.holder	-Tuition.fees.up.to.date -International -Curricular.units.2nd.sem ..approved.	-Tuition.fees.up.to.date -Curricular.units.2nd.sem ..approved -Scholarship.holder

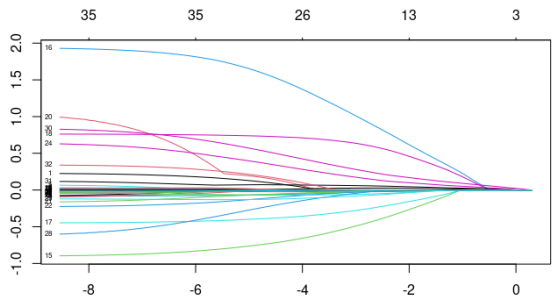
Nos quatre modèles montrent des performances similaires, légèrement inférieures aux modèles précédents, mais **beaucoup plus rapides**, avec des scores de prédiction entre 0.84 et 0.86.

Modèle avancé Elastic net

Le modèle Elastic Net est une technique de régression qui combine les pénalités de régularisation des modèles Lasso et Ridge. Dans Elastic Net, la régularisation est contrôlée par deux paramètres : λ (comme avant) et α, qui équilibre la proportion des pénalités L1 et L2. Afin d'optimiser le paramètre α, il est nécessaire d'effectuer une validation croisée. Après avoir effectué cela, nous avons obtenu α = 0.22, et le résultat suivant :

```
            Prediction
True class   0   1
         0 437  83
         1  61 451
```

Avec une précision de 0.8604651 (meilleure que celles obtenues via Ridge et Lasso)
Les variables principales sont les mêmes que celles de Ridge 1se.



Conclusion

Nos différents modèles nous ont permis d'avoir un premier aperçu des **différentes raisons qui peuvent entraîner une mauvaise performance scolaire des étudiants**. Les variables sélectionnées, telles que le paiement des frais de scolarité et le nombre d'unités de cours validées au second semestre, sont logiquement pertinentes et mises en avant par nos différents modèles, mais sont assez **évidentes**, et donc moins intéressantes à analyser en détail. Approfondir notre étude en ne tenant plus compte de ces variables, afin de trouver d'autres facteurs environnementaux, serait une piste intéressante à explorer pour développer de potentiels **nouveaux modèles**.
En revanche, nos modèles actuels révèlent tout de même des variables plus intrigantes, comme la dette envers l'école ou le statut d'étudiant étranger, qui apportent de nouvelles perspectives à notre étude.