KHALDI Jalal
Komlan Godwin AMEGAH

Analyse d'un dataset : classification binaire

Le dataset se nomme "bank-additional-full.csv".

On charge le dataset au format csv avec le séparateur " ;" et séparateur à décimales "."

Panel Dataset

On voit que le dataset n'a pas de valeurs manquantes sur toutes ces colonnes pas la peine de faire donc le preprocessing (bouton "Faire Preprocessing"). On vérifie si la catégorie de la chaque colonne (quantitative continue, quantitative discrète, qualitative nominale, qualitative ordinale) est bien correcte. Par exemple on voit que "age" est associée à une catégorie : quantitative discrète, c'est bien ce qu'on veut. Toutes les catégories de variables sont cohérentes pas besoin de les changer manuellement. Preprocessing a été fait et catégorisation des variables a été fait. On peut passer à l'analyse exploratoire pour visualiser la donnée. On sélectionne comme prédicteurs pour l'instant toutes les variables sauf la variable "y".

Panel Exploration

La variable à prédire est "y" à 2 modalités ("yes" ou "no"). On regarde à l'aide d'un diagramme à barres, à vue d'œil quelle est la proportion de classe majoritaire/ minoritaire. Les classes sont déséquilibrés. Cela se confirme avec le tableau effectifs/fréquences (89% pour "no" et 11% pour "yes").

On regarde les liaisons les plus fortes entre les prédicteurs et "y". Pour chaque prédicteur x qualitative, on va sélectionner un diagramme à barres pour voir empiriquement la liaison y/x.

Résumé :

'job' : influence, les distributions ne sont pas identiques

'marital' : pas d'influence, distributions identiques (presque)

'education' : une petite influence, la variance de la proportion de personnes positives (y == yes) conditionnée par l'éducation est faible (*)

'default' : influence, les personnes ayant un crédit en défaut n'adhèrent pas à un dépôt à terme bancaire

'housing' : pas d'influence, distributions identiques

'loan' : pas d'influence, distributions identiques

'contact' : influence, les distributions ne sont pas identiques

'month' : influence, les distributions ne sont pas identiques

'day_of_week' : pas d'influence, distributions identiques

'poutcome' : influence, les distributions ne sont pas identiques

Conclusion :

'job', 'default', 'contact','month','poutcome' ont une influence

Peut-être aussi un peu 'education'.

De la même façon, on regarde les liaisons pour les variables x quantitatives à l'aide de boxplot et des moyennes conditionnelles à l'aide de la visualisation "Statistiques".

Résumé :

'age' : pas d'influence, moyennes conditionnelles quasi identiques.

'duration' : influence, distributions différentes et moyennes conditionnelles différentes (196.74 pour y="no" et 471.45 pour y="yes").
'campaign' : influence, moyenne conditionnelle différente et médiane différente selon la modalité de "y"

'pdays' : influence, pdays=999 signifie que le client n'a pas été contacté avant (au cours de la précédente campagne), on voit pour y="no", pdays=984 en moyenne alors que pdays=792 en moyenne pour y="yes":donc les personnes qui refuse de déposer un compte à terme ont plus tendance en moyenne à ne pas être contacter au cours de la précédente campagne.

'previous' : influence, les moyennes conditionnelles sont très différentes (0.13 pour y="no" et 0.49 pour y="yes"). Une certaine logique se dégage "previous" est le nombre de contact effectué avant la campagne pour un client. Plus cette valeur est grande, plus on a de chance que le client dépose un compte à terme.

'emp.var.rate': influence clair, distributions donc moyennes conditionnelles (sachant y) très différentes

'cons.price.idx' : pas d'influence moyenne conditionnelles quasiment identique, ce prédicteur a une distribution quasi symétrique (médiane ~ moyenne) et a peu de lien avec y.

'cons.conf.idx' : idem pour "cons.conf.idx" que pour "cons.price.idx' : pas de lien

'euribor3m' : influence, distributions différentes, cette variable est discriminante pour y

'nr.employed': influence distributions différente, logique plus il y a d'employés dans la banque, plus de personnes pour faire du marketing et attirer de nouveaux clients pour souscrire à des comptes à termes

Conclusion : **'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'euribor3m', 'nr.employed'** qu'on va sélectionner

L'interface permet de sélectionner deux variables à comparer et voir directement les liaisons empiriques de celles-ci. Cela permet une approche individualisée de chaque prédicteur qui nous pousse à bien étudier chaque liaison pour chaque prédicteur.

Panel Apprentissage de Modèles

On va tout d'abord sélectionner que les variables quantitatives/ qualitatives qu'on a trouvé lors de l'analyse exploratoire : **'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'euribor3m', 'nr.employed'** et **'job', 'default', 'contact','month','poutcome'**

Ensuite on dummifie les variables qualitatives et on sélectionne toutes les variables : variables quantitatives sélectionnées + variables qualitatives dummifiées en supprimant pour chacune de ces variables une colonne dummifiée (une modalité précise) pour éviter la corrélation entre les prédicteurs.

Suppression de : job_admin, default_unknown, contact_cellular, month_apr, poutcome_nonexistent

Ensuite on va faire un oversampling/undersampling de y (dans le panel Apprentissage de Modèles) pour équilibrer les classes (50%-50%).

Régression Logistique :

Une fois ceci fait, dans l'onglet Apprentissage de Modèles, on sélectionne régression logistique avec un dataset d'entrainement de 79%, un seuil d'acceptation de 50% pour calculer les différentes métriques sauf l'AUC. La modalité positive de "y" est la classe minoritaire.

AUC : 0.933

Accuracy: 0.87

ErrorRate: 0.13

Precision: 0.85

Recall: 0.894

F1-score: 0.87

La métrique qu'on va utiliser est l'AUC.

On voit qu'on a une très bon AUC, on regardant la courbe roc, on voit qu'on a un seuil s* optimal telle que ce seuil et un trade-off specificity/ sensitivity qui maximise à la fois la specificity (specificity ~ 0.85) et maximise la sensitivity (sensitivity ~0.9). Ce sont de très bonnes valeurs. Les taux de faux positifs et faux négatifs vont être bas. Le seuil de 0.5 est intéressant car on équilibrant les classes on obtient un très bon f1-score, recall, precision. Un bon modèle qui prend en compte le nombre de faux positifs et de faux négatifs.

Dans un deuxième temps, on entraine un deuxième modèle (régression logistique) en essayant de supprimer quelques prédicteurs pour éviter l'overfitting, pour cela on va dans le panel Exploration, en sélectionnant comme type d'exploration "Multivariée", on peut obtenir la matrice de corrélations de notre DataFrame.

On voit que "emp.var.rate" est corrélé avec "euribor3m" (~0.97) et "nr.employed" (~0.91). On supprime donc "emp.var.rate" et "nr.employed" à l'aide du panel dataset simplement en désélectionnant ces deux variables.

AUC: 0.934

Accuracy: 0.87

ErrorRate: 0.13

Precision: 0.86

Recall: 0.89

F1-score : 0.87

On constante donc une amélioration de l'AUC avec une complexité diminuée ce qui est très bien.

Arbre de Décision CART :

On part de la sélection de prédicteurs qu'on a faite avant avec les mêmes paramétrages :

AUC: 0.911

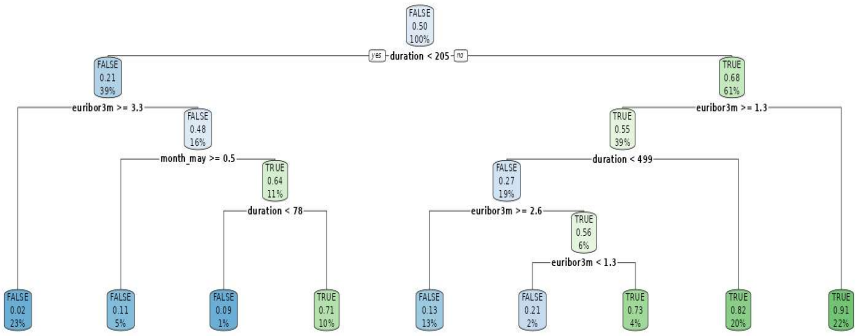
Accuracy: 0.87

ErrorRate: 0.13

Precision: 0.83

Recall: 0.94

F1-score: 0.88



L'arbre n'est pas trop complexe (on a fait un élagage : 3 prédicteurs seulement), c'est bien. Ce modèle est meilleur que la régression logistique car il réduit drastiquement la complexité du modèle avec que trois prédicteurs. Au niveau des métrique, l'AUC est un peu plus bas (~0.911) mais reste très bon. Au niveau d'un seuil de classification à 0.5 le recall est meilleur, une precision un peu moins bonne mais un f1-score plus élevé, le trade-off recall-precision est meilleur pour ce modèle.

Arbre de Décision CHAID :

On a décidé de sélectionner les variables qualitatives intactes c'est à dire en les dummifiant pas pour CHAID. De plus dans l'onglet "Nombres de bacs (var. quant.) correspond au nombre de bacs de discrétisation pour les variables quantitatives (toutes). On a choisi d'en mettre 3 bacs pour chaque variable quantitative.

AUC: 0.813

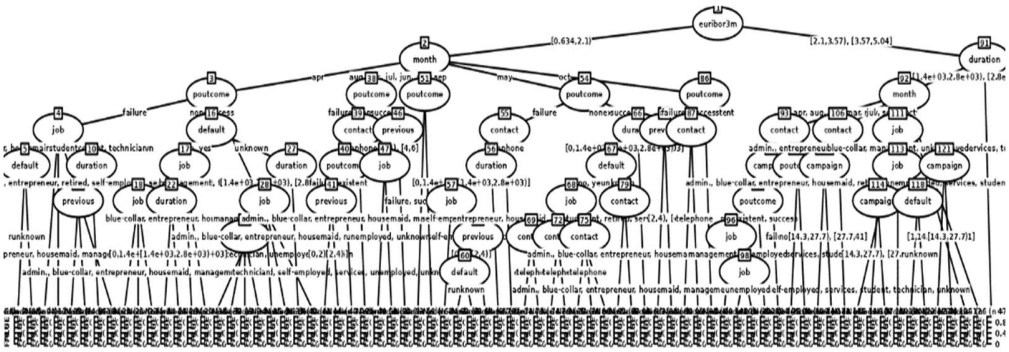
Accuracy: 0.76

ErrorRate: 0.24

Precision: 0.82

Recall: 0.66

F1-score: 0.73



L'arbre est très complexe et les métriques sont moins bonnes (AUC en l'occurrence) que les deux autres modèles. Ce modèle n'est pas adapté à cette configuration et à notre dataset

Conclusion :

Le modèle de CART (arbre de décision) semble le plus adapté. Un modèle simple avec des métriques (AUC, recall, precision, f1-score) qui ont des valeurs élevées qui réduisent à la fois le taux de faux positifs resp négatifs. Ensuite la régression logistique qui est un peu plus complexe mais a de bonnes performances. Puis finalement l'arbre de décision CHAID qui est très complexe, ce qui peut causer des problèmes lors de la généralisation à la population globale.

On a réalisé cette petite analyse de ce dataset, l'interface est fonctionnelle et permet d'analyser des datasets au format "csv" ou "xlsx" avec des modèles de classification.

Néanmoins, il reste encore des choses à améliorer notamment, des intégrations possibles de tests statistiques pour les variables pour vérifier la liaison des variables. Des ajouts de modèles de classification (SVM, LDA...) et régression (régression linéaire…). Des ajouts sur la possibilité de modifier les hyperparamètres des modèles. En conclusion cette interface constitue une bonne première approche pour évaluer quel modèle est le plus adapté à quel dataset.