CLANET Guilhem
LELEU Anne
VIVIEN Cédric
BUT 3 FA EMS

SAE DATAMINING :

Mise en oeuvre d’un processus Datamining

S5
Encadré par M Sellami

IUT | Université de Lille

Sommaire

Introduction

En tant que Data Analyst au sein d'une institution bancaire, nous avons pour objectif d'aider à l'amélioration des campagnes de marketing directes. Pour cela, nous devons procéder à des techniques de fouille de données afin de cibler les clients susceptibles de souscrire à un compte à terme (CAT), une des offres de l'institution bancaire que nous cherchons à promouvoir.
L'objectif de cette étude est donc de développer un modèle prédictif qui permet de repérer les clients susceptibles de souscrire à un CAT pour les cibler dans nos campagnes marketing.

Cette étude est divisée en deux parties :

- Analyse et exploration des données
- Comparaison des modèles de classification

Les objectifs pédagogiques de cette SAE sont :

- Amener l'étudiant à percevoir l'importance de la préparation des données pour la qualité des résultats : nettoyage, transformations de variables, qualité des données, mise en place de données d'apprentissage et de données test
- Lui faire mesurer les différences entre plusieurs méthodes de data mining dans une démarche de sélection du modèle le plus adapté.
- Le faire réfléchir à la problématique du sur-apprentissage, à la validation du modèle et de sa robustesse
- L'amener à mesurer la difficulté d'une restitution adaptée à un commanditaire

1. Analyse et exploration des données

A. Présentation et modifications des données

Le jeu de données banque.csv compte 45211 lignes ou individus et 17 colonnes ou variables.
Voici un résumé des données **sans modifications :**

Variables	**Type**	**Nombre de modalités**
âge	texte	77
profession	texte	13
etat_civil	texte	4
éducation	texte	5
défaut	texte	2
solde	numérique	7168
prêt_logement	texte	2
prêt_personnel	texte	2
contact	texte	3
mois	texte	14
durée	numérique	1573
campagne	numérique	x
jours_précédents	numérique	x
précédent	numérique	x
résultat_précédent	texte	4
souscription	texte	2

Nous allons déclarer comme variable cible la souscription puisqu'elle répond à l'objectif de l'étude : repérer les clients susceptibles de souscrire à un dépôt à terme.
Les autres variables seront déclarées comme explicatives.

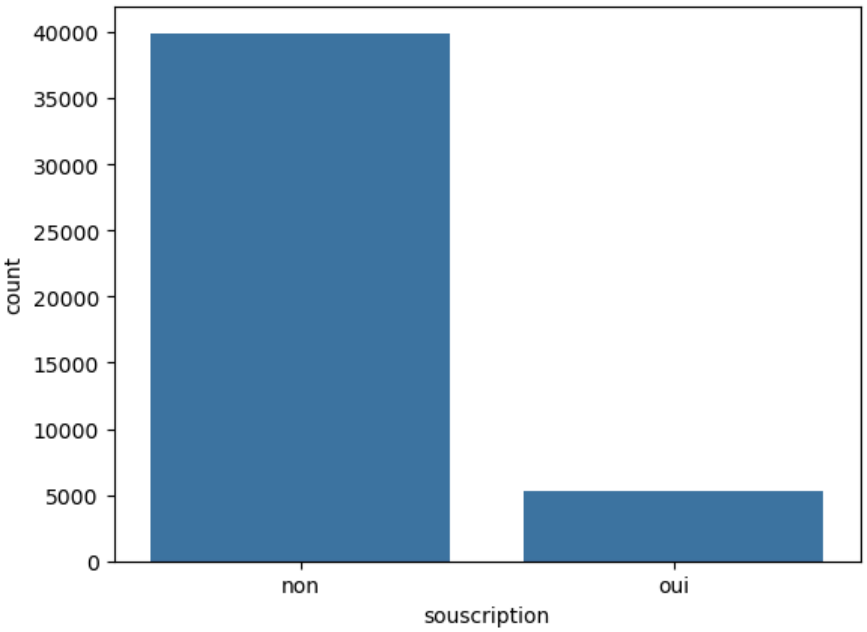
En termes de remplacement, nous avons opéré des remplacements “à la main” pour les modalités mal orthographiées ou renseignées avec des erreurs avec la fonction replace.
Exemple : “secondère” remplacé par secondaire.

Pour l’âge, une variable numérique qui comptait beaucoup d’incohérences, nous avons développé un modèle de régression qui a été déterminé par le test de 3 modèles différents : linéaire, Random Forest, et XGBoost.
La métrique d’évaluation utilisée est le R^2 qui indique l’adéquation des données au modèle.
Le R^2 le plus élevé était avec le modèle XGBoost qui s’élève à 0,46, nous avons utilisé XGBoost pour prédire les données manquantes concernant l’âge.

```
R² (r2_score) Modèle Linéaire :  0.20816326763252002
R² (r2_score) Random Forest :  0.4211808952797119
R² (r2_score) XGBoost :  0.45642896168668334
```

Pour l’état civil, le procédé est le même puisqu’on avait quelques “inconnu”.
Cette fois-ci, la variable étant qualitative, les modèles testés sont KNN et RandomForest.
Le modèle retenu est celui du RandomForest parce que l’accuracy s’élève à 1 (surement, car il y a peu de lignes à prédire).

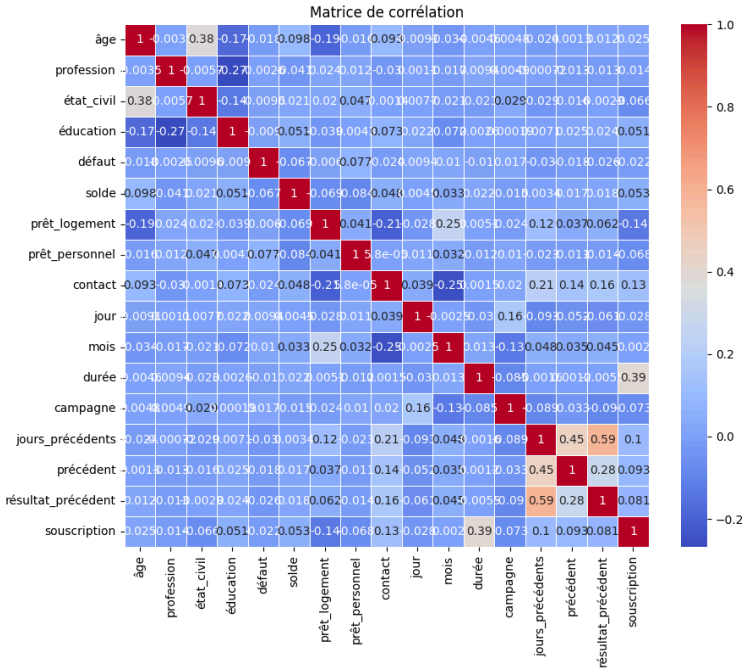
B. Exploration visuelle des données



L'histogramme de la variable cible, la souscription nous permet de voir que la plupart des clients ne souscrivent pas au CAT, la proportion des clients qui souscrivent ne s'élève qu'à environ 11%.

Avec la matrice de corrélation de toutes les variables, on observe que les variables les plus corrélées sont :

- le nombre de jours écoulés depuis le dernier contact avec le client lors de la précédente campagne avec le résultat de la campagne précédente (0,59)
- la souscription avec la durée de la dernière communication (0,39)



2. Comparaison de modèles de classification

A. Interprétation des métriques d'évaluation

Dans notre contexte qui est de cibler les clients prêts à souscrire au CAT, les 5 métriques d'évaluation utilisée sont les suivantes :

Accuracy : % de prédictions correctes (attention aux classes déséquilibrées)
Recall : Proportion de souscriptions réelles parmi les prédictions favorables et défavorables.

But : Minimiser les faux négatifs, ne pas passer à côté de clients potentiellement intéressés par le CAT.

Précision : Proportion de souscriptions réelles parmi toutes les prédictions positives.

But : Minimiser les faux positifs, éviter de cibler des clients pas intéressés par le CAT, perte d'argent dans les campagnes marketing.

F1 Score : Moyenne harmonique entre la précision et le rappel.

But : Prédire les souscriptions réelles en faisant un compromis entre le recall et la précision.

AUC : Mesure la capacité du modèle à séparer les classes

But : Déterminer la capacité de discrimination entre les souscriptions et les non-souscriptions

Voici un récapitulatif des métriques d'évaluation pour les 4 modèles : KNN, Random Forest, Arbre de décision et XGBoost.

	KNN	**Arbre de décision**	**XGBoost**	**Random Forest**
Accuracy	88,61%	87,58%	90,61%	90,16%
F1 Score	61,48%	70,84%	75,72%	72,46 %
Recall	58,77%	70,91%	73,51%	69,14 %
Précision	74,14%	70,77%	76,68%	78,32 %
AUC	83,43%	70,91%	92,96%	92,20%

KNN a le F1 score le plus faible, le modèle a un recall pour la classe 1 (les clients ayant souscrit) à environ 19,4%, ce recall faible montre que le modèle KNN a du mal à détecter les souscriptions. On peut l'enlever de notre sélection pour le meilleur modèle car ces scores montrent un déséquilibre dans la gestion des faux positifs et faux négatifs.

XGBoost a la meilleure accuracy.

Les autres scores étant sensiblement proches, on va s'intéresser à d'autres métriques d'évaluation comme l'AUC et la matrice de confusion.

Pour l'AUC, XGBoost et Random Forest ont des scores très bons et proches, autour de 92%, cela signifie qu'ils sont très efficaces pour détecter les souscriptions au CAT.
Le modèle qui a le plus de mal à détecter les souscriptions est l'Arbre de décision avec un score correct de 70,91%.

Le but étant de cibler au mieux les souscriptions, en vue d'une campagne marketing réussie, c'est-à-dire qui minimise les coûts tout en maximisant les profits, nous allons interpréter les FP et les FN de chaque modèle, donnés par la matrice de confusion :

Matrice de confusion	FP	FN
KNN	151	879
Arbre de décision	566	557
XGBoost	314	535
Random Forest	251	639

Les FP (Faux Positifs) correspondent aux clients qui sont catégorisés comme ayant souscrit alors qu'en vérité non.
Les FN (Faux Négatifs) correspondent aux clients qui sont catégorisés comme n'ayant pas souscrit alors qu'en vérité oui.

Bien que le modèle KNN prédit le moins de FP (les clients qui sont catégorisés comme ayant souscrit alors qu'en vérité non), il catégorise le plus de FN (clients qui sont catégorisés comme n'ayant pas souscrit alors qu'en vérité oui). Cela représente un manque en clients à cibler dans les campagnes marketing, des opportunités de profits ratées.

Le modèle Random Forest donne des performances moyennes sans écarts majeurs entre les FP et FN des autres modèles.

La force du modèle Random Forest réside dans sa capacité à minimiser les FP, ce qui revient à minimiser les pertes marketing. Si les moyens du service marketing sont limités, c'est ce que l'entreprise tend à faire : minimiser les faux positifs.

Le modèle XGBoost semble être un bon compromis, il minimise les clients manqués tout en minimisant les clients "faussement" intéressés.

B. Optimisation des hyperparamètres

Avec optimisation :	**KNN**	**Arbre de décision**	**XGBoost**	**Random Forest**
Accuracy	88,63%	89,44%	90,78%	90,25

Arbre de décision (89,44 %, +1,86 %) : le modèle qui a l'amélioration la plus notable, cela montre que l'arbre peut encore être affiné pour mieux généraliser
XGBoost (90,78 %, +0,17 %) : légère amélioration, le modèle était déjà bien robuste
Random Forest (90,25 %, +0,09 %) : amélioration la plus faible, modèle déjà robuste

C. ACP

Afin de pouvoir encore plus optimiser le programme, notamment en temps de calcul, il était intéressant de pouvoir réaliser une ACP. En effet, plus le volume de données est important, au plus le programme va mettre de temps pour calculer les modèles. C'est pour cela que nous avons aussi fait recours à une ACP pour pouvoir comparer les résultats des modèles avec et sans ACP.

Liste du nombre de classes pour lesquelles nous avons entraîné les modèles : [1,3,5,7,9,11,13,15,16]

Voici les résultats que nous obtenons avec la métrique accuracy_score

```
KNN score :
[0.87913, 0.88555, 0.88654, 0.88665, 0.88665, 0.88665, 0.88676, 0.88676, 0.88676]
RandomForest score :
[0.81743, 0.88566, 0.89251, 0.89528, 0.89605, 0.89716, 0.9007, 0.90059, 0.9007]
ArbreDecision score :
[0.87471, 0.88134, 0.87946, 0.88676, 0.88488, 0.88632, 0.88798, 0.88643, 0.88544]
XGBoost score :
[0.87935, 0.89296, 0.89329, 0.89672, 0.89959, 0.89893, 0.9039, 0.90147, 0.90158]
```

On remarque que les tendances sont les mêmes avec ou sans ACP, le modèle XG Boost reste le meilleur.

Conclusion

Pour répondre à l'objectif principal qui était de repérer les clients susceptibles de souscrire à un prêt à court terme, en vue d'une campagne marketing ciblée, nous avons :

- Nettoyé les données : remplacement, gestion des valeurs manquantes…
- Analysé et explorer les données nettoyées
- Entraîné, évalué et amélioré 4 modèles différents

En conclusion, le modèle XGBoost semble le plus complet, il fournit de bons résultats sur l'ensemble des données, même réduites *(cf. C.ACP)*.
Le modèle XGBoost paraît être un bon compromis pour minimiser le nombre de clients manqués tout en minimisant également le nombre de clients ciblés qui n'étaient pas susceptibles de souscrire au CAT. En d'autres termes, c'est le modèle le plus équilibré entre la gestion des coûts marketing et la performance (nombre CAT obtenus).
En parallèle, on peut noter que le modèle Random Forest peut être une autre option, moins performant globalement que XGBoost mais qui reste pertinent si l'entreprise a des ressources marketings limitées.