Université Paris Descartes

Module : Apprentissage Supervisé

PART II : Données Réelles

Réalisé par :

Ait Bachir Samy

I. Introduction

Dans cette seconde partie du projet, il s'agit de traiter de réelles données dans le but de pouvoir prédire selon les comportements de client d'une banque, si ces derniers sont susceptibles de posséder un carte « Visa Premier ».

Premièrement il s'agit de faire une première étude des données afin de mieux les comprendre, puis, dans un second temps, d'utiliser des méthodes connues d'apprentissage supervisé pour tenter de trouver des patterns indiquant la possession de carte « Visa Premier ». Finalement, il sera fait une comparaison des différentes méthodes utilisées, et ce grâce à des courbes « ROC »

II. Exploration des données

Comme mentionné en introduction, la base de données étudié ici est constitué de données réelles. Ces données-là décrivent des clients d'une banque et leurs comportements. Dans ce qui suite, nous allons voir plus en détails de quoi est constitué cette base de données.

II.1 Description de la base de données

La base de données, telle-que donnée pour le projet, est constituée de 1073 clients décrits par 47 variables et divisé en deux classes. Ces variables sont de types différents :

Variables quantitatives :

1. **age :** L'age (En années)
2. **anciente :** L'ancienneté (En mois)
3. **mtrejet :** Le nombre d'impayés en cours (tous à 0 dans cette base de données)
4. **nbopguic :** Les montants totaux des rejets en francs (La plupart sont nuls mais 13 d'entre eux sont négatifs)
5. **moycred3 :** Le nombre d'opérations par guichet dans le mois
6. **aveparmo :** La moyenne des mouvements créditeurs sur les trois derniers mois (En milliers de francs)
7. **endette :** Le total des avoirs monétaire (En francs)
8. **engagemt :** Le taux d'endettement (En pourcentage)
9. **engagemc :** Le total des engagements à court terme (En francs)
10. **engagemm :** Le total des engagements à moyen terme (En francs)
11. **nbcptvue :** Le nombre de comptes à vue
12. **moysold3 :** Moyenne des soldes sur 3 mois
13. **moycredi :** Moyenne des mouvements créditeurs (En milliers de francs)

14. **agemvt :** Le nombre de jours depuis le dernier mouvement
15. **nbop :** Le nombre d'opérations au mois dernier
16. **mtfactur :** Le montant des factures de l'an dernier
17. **engageml :** Total des engagements à long terme
18. **nbvie :** Nombre de produits contrat vie
19. **mtvie :** Montant de ces produits (En francs)
20. **nbeparmo** : Nombre de produits d'épargne monétaire
21. **mteparmo :** Montant de ces produits (En francs)
22. **nbeparlo :** Nombre de produits logement
23. **mteparlo :** Montant de ces produits (En francs)
24. **nblivret :** Nombre de comptes sur livret
25. **mtlivret :** Montant sur les livrets
26. **nbeparlt :** Nombre de produit épargne long termes
27. **mteparlt :** Montant de ces produits
28. **nbeparte :** Nombre produit épargne à terme
29. **mteparte :** Montant de ces produits
30. **nbbon :** Nombre de produits bons et certificats
31. **mtbon :** Montant de ces produits
32. **nbpaiecb :** Nombre de payement par carte le mois précédent
33. **nbcb :** Nombre total de cartes
34. **nbcbptar :** Nombre de cartes point argent
35. **avtscpte :** Total des avoirs sur tous les comptes
36. **aveparfi :** Total des avoirs épargne
37. **nbjdebit :** Nombre de jours de débit

Variables qualitatives :

1. **departem :** Le département ou est domicilié le client.
2. **ptvente :** Le point de vente de la banque rattaché au client.
3. **sitfamil :** La situation familiale du client.
4. **csp :** La catégorie socio-profetionnelle du client.

5. **codeqlt :** Le code qualité du client donné par la banque.
6. **sexer :** Le sexe du client en binaire.

Variables non-utilisées :

Certaines variables ne sont pas utilisées lors du traitement de cette base de données, et ces variables sont :

1. **matricul :** Représente l'identifiant du client mais celui-ci n'a de toute évidence aucune incidence sur la possession de carte visa premier.
2. **sexe :** Variable représentant le sexe du client sous la forme de chaîne de caractère. Cette variable n'est pas utilisée, car elle a un équivalant binaire nommé sexer.
3. **nbimpaye :** Cette variable qui représente le nombre d'impayés en cours n'est pas prise en considération, car elle vaut 0 et ce quel que soit le client.
4. **cartevp :** La possession ou non de carte premier visa, tout comme la variable sexe elle est remplacée par l'utilisation de la variable cartevpr.

Classe :

La classe est représentée par la variable **cartevpr** de la base de données qui est une variable binaire représentant la possession ou non d'une carte visa premier pour un client donné.

II.2 Pré-traitements sur la base de données

Avant de pouvoir appliquer un quelconque algorithme de classification sur la base de données des clients, plusieurs pré-traitements s'imposent. On se concentre sur les variables quantitatives, car seul celles-ci seront utilisées au cours de ce projet.

Valeurs manquantes :

Deux des variables quantitatives contiennent des valeurs manquantes. Il s'agit de :

1. **agemvt :** Ici il s'agit du nombre de jours passés depuis le dernier mouvement sur les comptes du client. Seul 6 valeurs sont manquantes, On remarque donc facilement que pour toutes les instances ou la valeur de agemvt est manquante, le client ne dispose pas de carte visa premier. Par ailleurs tout client n'ayant pas fait de mouvement depuis plus 48 jours ne possède pas de carte visa premier. Pour cette raison, on décide d'affecter la valeur maximale de la variable à toutes les instances ou celles-ci sont manquantes, sachant que le maximum vaut 944 jours.
2. **Nbpaiecb :** Pour le nombre de payement par carte effectués le mois dernier, lors de la présence de valeurs manquante, il existe certain cas ou le client possède une carte visa premier, et d'autres non, de plus le nombre de valeurs manquante est élevé. S'agissant d'une donnée quantitatives, les valeurs manquantes sont remplacées par la moyenne des nombres de payements par carte hors valeurs manquantes.

Élimination des colinéarités :

Plusieurs des variables quantitatives présentes dans la base de données peuvent être calculées à partir d'autres variables à l'aide d'un produit par un scalaire de la seconde variable. Ces premières représentent donc une redondance d'information pouvant fausser les résultats de classifications des algorithmes utilisés.

Afin d'éviter ce problème, les variables sont étudiées selon leurs facteurs d'inflation de la variance (VIF). Les variables causant les plus grandes hausse de VIF sont retirées. Plus précisément, on procède à la recherche de paires de variable ayant une forte corrélation (corrélation linéaire supérieur à 0,9), puis la variable appartenant à la paire ayant le plus haut score VIF et retiré des données utilisées. Puis, on les variables avec les plus grands VIF tour à tour en recalculant les VIF de toutes les variables après chaque élimination. Ces éliminations s'arrêtent lorsque toutes les variables ont un score de VIF inférieur à 10, sachant que 10 est une valeur dite large, car dans la littérature les valeurs de VIF sont souvent inférieures.

Les variables éliminées par ce procédé sont : mtbon, nbbon, mteparmo, aveparfi, moycred3, engageml, aveparmo, nbeparmo et avtscpte.

Normalisation :

Dans le but d'utiliser les algorithmes de classification connus en accordant une importance égale à toutes les variables, on applique sur les données une normalisation entre 0 et 1 pour chacune des valeurs des variables, et ce en appliquant la formule :

$$\textbf{Données\_normalisé}(d|X) = (d - \min(X)) / (\max(X) - \min(X))$$

Ou d est la donnée à normaliser et X est la variable dont d est une valeur.

III. Classification des données et résultats

Avant d'exécuter un quelconque algorithme sur les données, on sépare ces dernières en deux parties : Une première partie contenant trois quarts des données utilisée pour l'apprentissage des algorithmes, et une seconde partie contenant le quart restant et qui servira à valider les apprentissages.

Pour la classification des données plusieurs algorithmes ont été testés :

- K plus proches voisins (KNN).
- Analyse discriminante linéaire (LDA).
- Classifieur Bayésien Naif.
- Support Vecteur Machine (SVM).

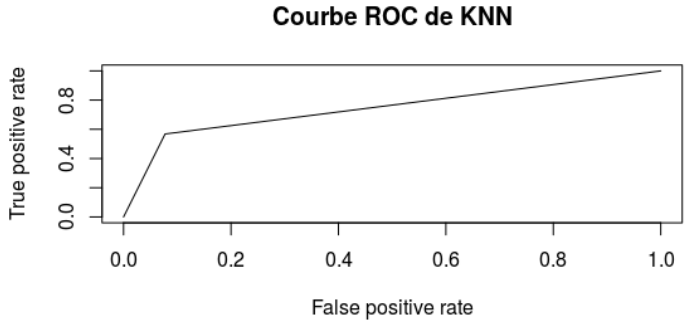
Dans ce qui suit sont détaillés les fonctions utilisées, les paramètres et leurs valeurs ainsi que les résultats de ces derniers.

III.1 K plus proches voisins

L'algorithme des K plus proches voisins est implémenté dans le package class sous R dans la fonction knn, cette fonction prend en paramètre les données d'entraînement et les classes correspondantes à chacune des entités représentées, les données de validation et le nombre de voisins à considérer K.

N'ayant pas de méthode de détermination à priori de K, la détermination se fait de manière empirique, en faisant varier le paramètre K de 1 à 50.

Les résultats obtenus sont moyennement satisfaisants, car la précision maximale, obtenue en prenant en compte 10 voisins, est de 80 %. et la sensibilité du modèle est montrée sur la courbe ROC suivante :

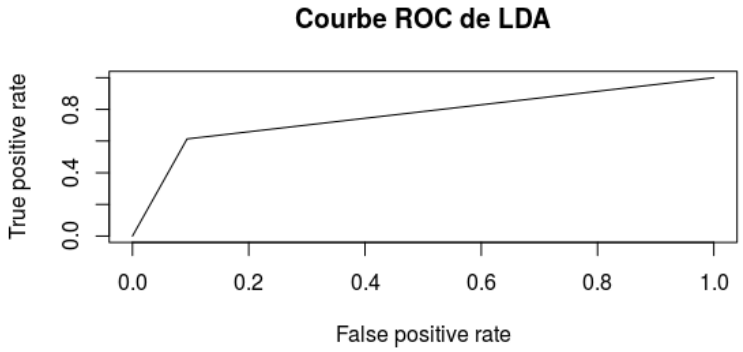


III.2 Analyse discriminante linéaire

La LDA est implémentée dans la fonction lda du package MASS de R. La fonction est très simple à utiliser elle prend en entrée les valeurs d'entraînement et les classes correspondantes et retourne un modèle discriminant. Aucun paramètre n'est à fixer.

Les résultats de la LDA sont de très peu meilleurs à ceux obtenus par KNN, avec une précision de 81 %. La figure suivante correspond à ce modèle :

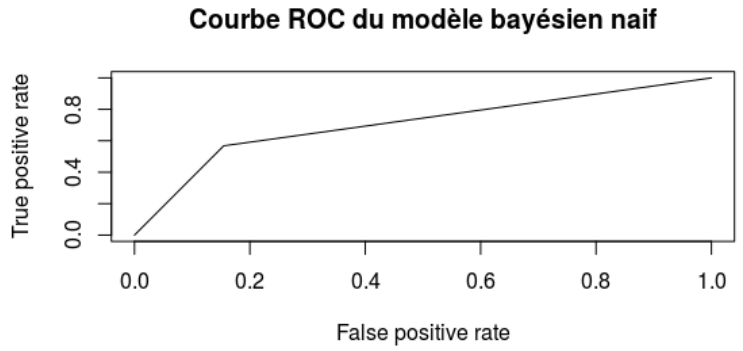
III.3



Classifieur Bayésien Naif.

Implémenté dans la même librairie que la LDA, le classifieur bayésien est implémenté dans la fonction naiveBayes, qui fonctionne de manière similaire à la fonction lda.

Ce type de classifieur ne donne pas de très bons résultats, en effet sa précision ne dépasse pas les 76 % de précisions. Et sa sensibilité est montée sur la courbe ROC représentée sur la figure suivante

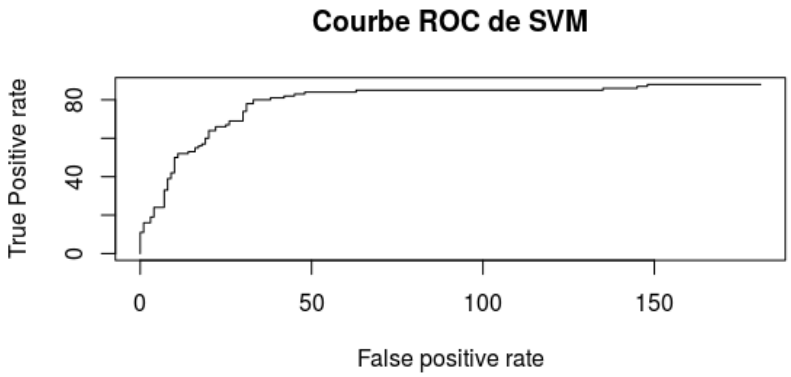


III.4 Support Vecteur Machine

SVM est une méthode de classification implémentée dans la library e1071 de R dans la fonction svm. La méthode en elle-même possède plusieurs paramètres à fixer selon le type des données en entrée, ces paramètres sont le kernel et le coût. Nous avons là des données quantitatives, le kernel radial serait à priori le meilleur choix. Pour aider à paramétrer le modèle, la librairie e1071 propose une méthode très utile nommée tune qui teste tour à tour tous les paramètres en entrée proposés.

La meilleure précision (88%) a été obtenue grâce a un kernel radial, en concordance avec l'hypothèse avancée, et un cout de 1. On remarque également que contrairement aux autres méthodes, aux travers de la courbe ROC obtenue (figure suivante) que le modèle SVM progresse plus aisément que les 3 autres.

IV.



Conclusion

Le classifieurs qui donne les plus mauvais résultats est le classifieur bayésien naif avec une précision de 75 %, sachant qu'un classifieur répondant toujours négativement aurait une précision de 66 %, ce score est très bas.

Les meilleurs résultats sur ces données de clients de banque sont obtenus en appliquant l'algorithme SVM dessus. La précision de ce dernier est de 88 %, ce qui est une nette amélioration par rapport au classifieur négatif.

Cette précision de 88 % n'est certes pas très haute, mais elle permettrait à la banque de viser plus efficacement les clients susceptibles d'acquérir une carte visa premier, en réalisant des compagnes les ciblant plus particulièrement.