INF2044 - Statistiques et Analyse des données

Examen de Travail Pratique

Analyse préliminaire des données de remboursement d'empreint d'une banque

Le jeu de données fourni avec ce TP contient des données des prêts des clients d'une banque aux états unis qui ont été remboursés ou non, qui sont en difficulté de remboursement et mis en recouvrement sans rembourser leur prêt et ses intérêts et ceux qui n'ont remboursé qu'après leur mise en recouvrement. Il faut noter que les empreints peuvent être remboursés avant la date d'échéance prévu.

Ce travail pratique consiste à faire des analyses primaires sur ce jeu de données. On pourrait s'intéresser ensuite à la construction des modèles de prédiction de remboursement d'empreint pour un client non enregistré, ceci peut servir au banquier pour savoir si oui ou non il peut accorder du crédit à un client qui en sollicite ou bien apporter des ajustements pour le crédit octroyé à un client selon les résultat du modèle.

Le questionnaire qui suit va guider votre analyse :

0.1 Importation des données et informations sur le jeu de données

1. Importer dans une session R, le jeu de données contenu dans le fichier **Loan_payments_data.csv** sous forme de **data.frame.**

2. Dans ce jeu de données, nous pouvons voir les colonnes suivantes :
   a) **Loan_ID** : L'identifiant de prêt d'un client donné
   b) **Loan_status** : Le statut d'un prêt, il peut être soit "PAIDOFF" (Le prêt a été remboursé dans les délais), "COLLECTION" (le remboursement n'est pas remboursé et est en recouvrement), "COLLECTION_PAIDOFF" (Le prêt a été remboursé en recouvrement).
   c) **Principal** : Montant principal de base du prêt en dollars $.
   d) **terms** : Termes de remboursement, pouvant être semaine (7), bi-semaine (15) et mensuel (15)
   e) **effective_date** : La date de prise en compte du prêt
   f) **due_date** : La date d'échéance du prêt
   g) **paid_off_time** : La date et l'heure du remboursement du prêt. Si vide, alors le prêt n'a pas été remboursé
   h) **past_due_days** : Le nombre de jours après la date d'échéance prévu pour le remboursement
   i) **age** : L'âge du client ayant contacté le prêt
   j) **education** : Le niveau scolaire du client ayant contacté le prêt

k) **Gender** : Le sexe du client ayant contacté le prêt

Combien variables sont contenu dans ce jeu de données et combien d'observations ?

3. Donner le type de chaque variable en colonne du jeu de données. Afficher les 10 premières et les 10 dernières observations.

4. Certains colonnes possèdent des valeurs vides ou des valeurs **NA** (Non Available) qui signifient que les valeur n'est pas disponible. Donner pour chaque variable le pourcentage des observations nulles (NA ou vides).

0.2 Analyse descriptive avec les graphiques

La colonne **Loan_ID** est juste un identifiant et donc peut être ignorée.

1. **loan_status** : Représenter un diagramme en barre des effectifs, un diagramme en barre des effectifs cumulés et un diagramme en secteurs pour cette variable. Quelle conclusion faites-vous à partir de ce diagramme ?

2. **Principal** : Tracer une boite à moustaches pour cette variable pour chaque modalité de **loan_status** ainsi qu'un histogramme de cette variable. Quel est d'après-vous le montant le plus sélectionné d'après l'histogramme? Dans un tableau, donner le nombre d'observations par statut et par montant du prêt.

3. **terms** : Représenter un diagramme en barre des effectifs, ensuite un diagramme en barre des effectifs par valeur de **loan_status**. Quelle conclusion faites-vous ?

4. **effective_date** : Représenter un diagramme en barre des effectifs par date effective de remboursement et statut de remboursement. Quelle conclusion faites-vous ?

5. **age** : Représenter un histogramme sur les ages et des boites à moustaches des ages par statut du prêt. Quelle conclusion faites-vous ?

6. **education** : Représenter un diagramme en barre des effectifs de **education** et un diagramme en barre des effectifs de **education** en fonction du statut du prêt, ensuite interpréter.

7. **gender** : Représenter un diagramme en en barre des effectifs de **gender** et un diagramme en barre des effectifs de **gender** des en fonction du statut du prêt, ensuite interpréter.

8. Tracer à la main le cas échéant un diagramme en barres des effectifs, un diagramme en secteurs, un diagramme en barres des effectifs cumulés, un diagramme en bâtonnets et un histogramme (des densités d'effectif et de fréquence).

0.3 Analyse descriptive avec les mesures

Certaines questions ici nécessitent d'écrire une ou plusieurs fonctions.

1. Quelle est la moyenne des montants des prêts contactés dans cette banque? et la variance.

2. Quelle est le pourcentage de prêts au dessus de 900 $? Interpréter.

3. Quelle est le moyenne d'âge des clients ayant effectué un prêt ?

4. Quelle est la moyenne d'âge des clients ayant effectué des prêts en dessous de 600 $?

5. Quel est le nombre de prêts avec un terme d'une semaine ?

6. Quel est le niveau d'étude majoritaire des clients ayant effectué un prêt supérieur ou égal à 900 $?

7. Écrire une fonction qui accepte deux dates d_1 et d_2 sous forme de chaîne de caractères (exemple **"9/14/2016"** et donc le format **"%m/%d/%Y"**) et retourne le nombre de jours qui sépare les deux dates. Vous pouvez la modifier pour qu'elle prenne en compte l'heure comme "9/14/2016 19:31".

8. Quel est le pourcentage de prêts qui ont été remboursés dans les délais ?

9. Entre les hommes et les femmes, qui sont les plus loyaux dans le remboursement des prêts (y compris sans recouvrement) ?

10. Entre les hommes et les femmes, qui contactent des prêts au dessus de 900 $?

11. Quel est le niveau d'étude majoritaire des moins fidèles au remboursement (même avec recouvrement) ?

12. Quel forme de terme possède le plus grand nombre de remboursements ?

13. Quel est la moyenne d'âge des clients qui ne remboursent pas leurs prêts ?

14. Pour les plus futés, on souhaite savoir le client le plus honnête dans le remboursement (ou encore au mieux l'identifiant du prêt), pour cela, il faudra séparer les observations par montant de prêt contacté, ensuite calculer le nombre de jours écoulés pour rembourser le prêt en question (On ne tient que compte de ceux qui remboursent au plus à la date d'échéance prévu et pas avec le recouvrement). Avec ces données, il est possible de déduire qui est le plus fidèle dans le remboursement.