École d'actuariat
Université Laval

Travail pratique partie I
Apprentissage statistique en actuariat
ACT-3114

Professeure : Marie-Pier Côté
Hiver 2020

1 Consignes générales

- Le travail doit être effectué en équipe de quatre personnes.
- Le rapport en format pdf doit être remis dans la boite de dépôt du site de cours. Aucun rapport imprimé ne sera accepté.
- Le rapport, incluant les formules mathématiques, doit être rédigé avec un logiciel de traitement de texte (LaTeX, Rmarkdown, Word, etc.) Un gabarit de rapport produit avec LaTeX est disponible sur le site du cours pour les intéressés.
- Le rapport doit être rédigé de façon structurée. Les graphiques et tableaux doivent tous être expliqués dans le corps du texte.
- 10 % des points sont accordés pour la qualité du français, à raison de 1/2 point par faute d'orthographe ou de grammaire. Un délai de grâce de 2 fautes sera accordé.
- 10 % des points sont accordés pour la qualité de la présentation, et pour la présence d'une page titre appropriée, d'une table des matières et d'une pagination correcte. Les tableaux et graphiques doivent tous avoir un titre clair. Les textes des graphiques doivent être assez gros pour être lisibles, et les titres ou légendes doivent être clairs.
- 10 % des points de cette évaluation provient d'une évaluation par les pairs de la contribution au travail d'équipe.

2 Dates importantes

— 23 janvier 2020 à 8 h 30 : date limite de formation des équipes sur le site de cours.
— 5 février 2020 à 8 h 30 : date limite pour partager la description de vos données sur le forum afin de faire valider par la professeure. L'approbation peut prendre jusqu'à 48 h.
— **27 février 2020 à 23 h 59** : date limite de remise du premier rapport dans la boîte de dépôt sur le site de cours.
— 28 février 2020 à 23 h 59 : date limite pour compléter l'évaluation de la contribution au travail d'équipe concernant le premier rapport sur le site de cours.
— 15 avril 2020 à 8 h 30 : date limite de remise du deuxième rapport.
— 15, 16, 22 et 23 avril 2020 : présentations orales pendant les séances de cours et de laboratoire. Présence obligatoire. Remise des évaluations par les pairs le jour de l'exposé de chaque équipe.
— 24 avril 2020 à 8 h 30 : date limite pour compléter l'évaluation de la contribution au travail d'équipe concernant le deuxième rapport et la présentation sur le site de cours.

3 Choix du projet

Sélectionnez un jeu de données contenant une variable réponse (continue, binaire, catégorielle ou discrète), une mesure d'exposition (si approprié) et au moins 5 variables explicatives. Ce jeu de données ne doit pas avoir été utilisé dans le cadre du cours.

Votre jeu de données devra vous servir pour un modèle dans un contexte actuariel. Par exemple, vous pouvez considérer un problème de tarification ou de réserves en assurance générale. Vous pouvez aussi choisir de modéliser une variable en assurance vie ou santé (espérance de vie, incidence de maladies critiques ou dépression, frais médicaux).

Voici quelques possibilités de sources pour votre jeu de données :

— Le paquetage `CASdatasets` contient un grand nombre de jeux de données actuariels.
— Les données de triangles de développement disponibles sur le site de la CAS, pour l'étude d'un problème de réserves agrégées.
— Des données ouvertes d'un concours Kaggle passé, par exemple les données d'espérance de vie de l'Organisation mondiale de la santé ou l'un des jeux de données d'assurance.
— Les données disponibles sur Gapminder.

Vous devez faire approuver votre choix de jeu de données par la professeure. Sur le forum `Choix des sujets de travaux pratiques` du site de cours, faire une brève description du jeu de données sélectionné par votre équipe avant le 6 février 2020. Inclure dans votre description :

— le nom du jeu de données
— la source
— une brève description des données (environ deux phrases)
— la variable réponse et son type
— la mesure d'exposition (s'il n'y en a pas, le mentionner)
— cinq variables explicatives et leur type,
— la taille du jeu de données (nombre d'observations et de variables).

Vous ne pourrez pas choisir le même problème qu'une autre équipe : premier arrivé, premier servi ! (La validation peut prendre jusqu'à 48 h suivant le message sur le forum.)

Le jeu de données choisi vous servira pour toute la session, puisqu'il fera l'objet du premier rapport, du deuxième rapport et de l'exposé oral.

4 Premier rapport

Dans le premier rapport, vous devez expliquer le problème qui vous intéresse, faire l'analyse exploratoire des données à l'aide de `ggplot2`, faire toutes les étapes de pré-traitement des variables (détection et correction des erreurs, traitement des données manquantes, création de nouvelles variables explicatives), et utiliser les techniques de réduction de la dimension et de classification non-supervisée afin de mieux comprendre le contenu de votre jeu de données. Plus de détails sur le contenu du premier rapport sont donnés à la section suivante.

5 Contenu du rapport

Page titre : N'oubliez pas d'écrire votre numéro d'équipe sur la page titre. Donnez un titre spécifique à votre travail.

Table des matières

Introduction : Expliquez le problème qui vous intéresse pour ce travail. La description du problème doit clairement détailler la variable réponse et la mesure d'exposition (s'il y a lieu). Mentionnez le jeu de données que vous avez utilisé et la source. Si le jeu de données a été étudié dans un livre ou un article actuariel, décrivez brièvement l'analyse des auteurs (inclure la source).

Une analyse exploratoire des données : Dans cette section, présentez un résumé clair et concis des données. Cela peut être fait à l'aide de tableaux de fréquence, d'histogrammes ou d'autres types de diagrammes. Les moyennes, écarts-types, médianes, minimums et maximums sont également intéressants pour des variables continues. Expliquez brièvement les données. Faites les graphiques en nuage de points ou les diagrammes en boîtes à moustaches de la variable endogène en fonction de chacune des variables exogènes à considérer. Utilisez les outils graphiques du paquetage `ggplot2`.

Si vous identifiez des erreurs dans les données, elles doivent d'abord être corrigées. Mentionnez ces erreurs et la façon dont elles ont été traitées dans cette section.

D'autres sections : Ajoutez ici les sections appropriées selon les particularités de votre jeu de données. Vous pouvez aborder la réduction de la dimensionalité, la création de nouvelles variables explicatives, la traitement des valeurs manquantes, la classification hiérarchique ou l'algorithme des k-moyennes. Vous devez obligatoirement inclure au moins une section parmi celles-là.

Conclusion : Rappelez le problème à l'étude et les points à retenir de votre analyse exploratoire et de votre pré-traitement de données. Concluez en mentionnant quels types de modèle de base seraient appropriés pour résoudre le problème : une régression linéaire simple, multiple, pondérée, régularisée (Ridge ou Lasso), une régression Poisson, Binomiale négative, Poisson gonflée à zéro, logistique, etc.

Bibliographie : Utilisez les normes de présentation bibliographique APA décrites ici ou le style bibliographique `apalike` avec BibTeX. Vous devriez avoir au minimum une référence à la source de votre jeu de données. Si vous utilisez d'autres sources pour votre travail, listez les aussi. Consultez le site de la bibliothèque pour de plus amples détails sur la citation des sources.

Annexe : Inclure ici la description du jeu de données que vous avez soumise sur le forum pour approbation.

6 Liste de vérification

1. Le problème a été validé par la professeure sur le forum du site de cours.
2. Le rapport contient toutes les parties décrites dans la section 5 de cet énoncé.
3. Il n'y a pas de fautes d'orthographe ou de grammaire.
4. La page titre présente le titre du travail, la date, les noms et le numéro d'équipe.
5. La table des matières est correcte.
6. La pagination est correcte.
7. Le problème étudié est bien expliqué et abordé de façon intéressante.
8. Le jeu de données est bien décrit.
9. La variable réponse, la mesure d'exposition (s'il y a lieu) et les variables explicatives sont données.
10. L'analyse exploratoire est pertinente.
11. Les titres des tableaux et graphiques sont clairs.
12. Les tableaux et graphiques sont tous mentionnés dans le texte.
13. Les légendes et les titres des axes des graphiques sont clairs.
14. La police dans les tableaux et graphique est lisible.
15. Les commentaires sont pertinents et concis.
16. La notation mathématique est bien définie.
17. Les erreurs dans les données sont détectées et corrigées, s'il y a lieu.
18. De nouvelles variables explicatives sont créées, si possible.
19. Les valeurs manquantes sont traitées, s'il y a lieu.
20. Une analyse en composante principale permet de réduire la dimension, si approprié.
21. Une classification non-supervisée (classification hiérarchique ou algorithme des k-moyennes) permet de mieux comprendre les données, si approprié.
22. La conclusion est adéquate.
23. La bibliographie contient toutes les sources utilisées pour le travail.
24. Les références sont citées dans le texte lorsqu'elles sont utilisées.
25. La bibliographie est présentée selon les normes APA.
26. L'annexe contient la description et toutes les composantes listées dans la section 3.
27. Le rapport en format pdf est remis dans la boîte de dépôt.

Amusez-vous bien !