École d'actuariat
Université Laval

Travail pratique partie II - révisé
Apprentissage statistique en actuariat
ACT-3114

Professeure : Marie-Pier Côté
Hiver 2020

1 Consignes générales

- Le travail doit être effectué avec les mêmes équipes que le premier rapport.
- 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

— **22 avril 2020 à 8 h 15** : date limite de remise du deuxième rapport dans la boîte de dépôt sur le site de cours.
— 22 et 23 avril 2020 de 8 h 30 à 11 h 30 : présentations orales en visioconférence. Remise des évaluations par les pairs le jour de l'exposé de chaque équipe. Votre présence aux exposés est nécessaire puisqu'une partie de votre note à l'exposé sera sur la pertinence des commentaires que vous ferez sur les travaux des autres équipes.
— 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 Deuxième rapport

Dans le deuxième rapport, vous devez résoudre le problème qui vous intéresse à l'aide d'une analyse de données pertinente.

Utilisez les données corrigées et pré-traitées que vous avez obtenues dans la première partie du travail. Séparez ces données en un échantillon d'entrainement et un échantillon test (au moins 15 % des observations).[1] L'échantillon test **ne doit pas** être utilisé pour ajuster les modèles ni pour choisir les hyperparamètres. L'échantillon test n'est utilisé qu'à la toute fin pour comparer les performances des modèles.

En utilisant l'échantillon d'entrainement seulement, ajustez au minimum

1. Un bon modèle de base (régression linéaire ou GLM)
2. Un modèle linéaire (généralisé) avec une régularisation (Ridge, Lasso ou Elastic Net)
3. Un modèle des k plus proches voisins
4. Un arbre de décision
5. Un ensemble d'arbres de décision agrégées par *bagging*
6. Une forêt aléatoire
7. Un modèle de *gradient boosting*

Analysez les résultats chacun des modèles et comparez les performances sur l'échantillon test. Interprétez les effets des deux meilleurs modèles. Plus de détails sur le contenu du deuxième rapport sont donnés à la section suivante.

1. Si vous avez des données débalancées, il est approprié de faire un échantillonnage stratifié.

4 Contenu du rapport

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

Table des matières

Introduction : Rappelez brièvement le problème qui vous intéresse pour ce travail. Mentionnez le jeu de données que vous avez utilisé et la source.

Modèle de base : En utilisant seulement l'échantillon d'entrainement, proposez un modèle simple d'apprentissage supervisé pour votre problème. Utilisez une (ou des) technique(s) étudiées dans le cours ACT-2003 Modèles linéaires en actuariat, par exemple la régression linéaire multiple ou un modèle linéaire généralisé. La distribution choisie doit être appropriée pour votre problème.

Ajustement des modèles : En utilisant seulement l'échantillon d'entrainement, ajustez les modèles listés dans la section 3. Pour chacun des modèles, justifiez la distribution ou la fonction de perte utilisée, détaillez la procédure pour obtenir le modèle final. Présentez les hyperparamètres optimaux.

Comparaison des modèles : Comparez la performance prédictive des différents modèles obtenus sur les données de test à l'aide de métriques appropriées pour votre problème. Visualisez les résultats.

Interprétation des meilleurs modèles : Interprétez les deux meilleurs modèles obtenus. Expliquez les variables importantes pour la prédiction et leur effet sur la variable réponse. Détectez la présence d'interactions importantes entre certaines variables explicatives. Critiquez les modèles en présentant les avantages et les désavantages dans le contexte de votre problème.

Conclusion : Concluez votre rapport en expliquant quel est le meilleur modèle et si vous avez réussi à résoudre votre problème de façon satisfaisante avec les modèles étudiés. Faites une ouverture (par exemple, mentionnez d'autres possibilités de modèles ou des améliorations qui pourraient être apportées si plus de données ou d'autres variables explicatives étaient disponibles).

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. Pour savoir comment citer un paquetage R, utiliser la commande `citation(package = "nom-du-paquetage")`.

5 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 4 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 expliqué brièvement.
8. Un échantillon de test de 15 % est conservé pour la comparaison des modèles et n'est pas utilisé pour ajuster les modèles.
9. Le modèle de base choisi est pertinent.
10. Les modèles listés dans la section 3 sont tous ajustés et la procédure est expliquée.
11. La notation mathématique est bien définie.
12. Les hyperparamètres ont été choisis de façon appropriée.
13. Les titres des tableaux et graphiques sont clairs.
14. Les tableaux et graphiques sont tous mentionnés dans le texte.
15. Les légendes et les titres des axes des graphiques sont clairs.
16. La police dans les tableaux et graphique est lisible.
17. Les commentaires sont pertinents et concis.
18. Les modèles sont comparés à l'aide de métriques appropriées pour le problème.
19. Les deux meilleurs modèles sont interprétés.
20. Les outils d'interprétation des meilleurs modèles pertinents sont utilisés (importance des variable, graphiques de dépendance partielle, courbe ROC, levier de ratio de perte, etc.)
21. Il y a des commentaires sur les interactions entre les variables explicatives.
22. La conclusion est correcte.
23. La conclusion se termine par une ouverture.
24. La bibliographie contient toutes les sources utilisées pour le travail.
25. Les références sont citées dans le texte lorsqu'elles sont utilisées.
26. La bibliographie est présentée selon les normes APA.
27. Le rapport en format pdf est remis dans la boîte de dépôt.

Amusez-vous bien !