Année 2021-2022

Survival and longitudinal data analysis

Exercice 1

Le but de ce TP est d'utiliser les outils de l'analyse de survie pour développer un outil de maintenance prédictive pour une machine. "La maintenance prévisionnelle (aussi appelée « maintenance prédictive », ou encore « maintenance anticipée ») est une maintenance conditionnelle basée sur l'anticipation (...) qui permet de donner l'état de dégradation du bien avant sa détérioration complète." (voir https://fr.wikipedia.org/wiki/Maintenance_prévisionnelle).

L'intérêt de la maintenance prédictive est donc de pouvoir faire des réparations avant la panne, qui pourrait endommager l'équipement. Cela permet aussi d'éviter les accidents qui pourraient être dus à la panne, un autre but est d'éviter les contrôles trop rapprochés qui coutent chers à l'entreprise. Cela est rendu possible par le développement de l'internet of things (IoT), qui permet de collecter facilement et à moindre coût des données enregistrées par des capteurs placés dans ou à proximité de la machine.

Les données contiennent des mesures sur 1000 machines, l'évènement d'intérêt est la panne. On veut prédire avec précision l'évènement `panne dans les 2 semaines` afin d'alerter à l'avance l'équipe de maintenance. Ce sont des données d'entraînement où on suppose que les valeurs enregistrées par les capteurs ne varient pas avec le temps. En situation réelle, ces données dépendraient évidemment du temps.

Les données `predictive_maintenance.csv` qui contiennent les colonnes :

Nom	Type	Description
`lifetime`	numérique	Nombre de semaines de fonctionnement de la machine
`broken`	numérique	Indicatrice de censure : 1 si la machine est en panne, 0 sinon
`pressureInd`	numérique	Mesure de pression dans les tuyaux
`moistureInd`	numérique	Mesure de l'humidité ambiante
`temperatureInd`	numérique	Mesure de la température ambiante
`team`	catégorique	Nom de l'équipe utilisant la machine
`provider`	catégorique	Nom du constructeur

1. Importer les données, vérifier le type des variables, faire les changements nécessaires.
2. Vérifier si les données contiennent des `NA` ou des lignes dupliquées.
3. Faire un histogramme pour la variable `lifelines` en colorant suivant la valeur de `broken` et calculer le pourcentage de censure dans le jeu de données. Que remarquez vous ?
4. Faire des histogrammes pour les covariables continues et des diagrammes en bâton pour celles discrètes.

5. Grâce à la librairie `corrplot`, représentez graphiquement les corrélations entre covariables (attention à d'abord transformer les données qu'elles soient entièrement numériques, voir code joint).

6. Représentez graphiquement les fonctions de survie dans les sous-groupes définis par la variable `team` d'une part et `provider` d'autre part. Que remarquez vous ?

7. Créer une partition 80/20 des données en `train` et `test` via la library `caret`. Attention à bien stratifier sur la variable de censure (voir code joint).

8. Faire un premire modèle de Cox sur le `train`, on ne pourra pas inclure `provider`, expliquer pourquoi (essayer avec).

9. Grâce à la librarie `riskRegression`, représenter le score de Brier en fonction du temps (voir code joint). Puis coder une fonction qui calcule le score intégré sur le `test`, voir https://square.github.io/pysurvival/metrics/brier_score.html pour des définitions.

10. Répéter les deux dernières questions avec une forêt aléatoire (dans un premier, on prendra les paramètres par défaut) de la libraire `randomForestSRC`. Attention, il y a un bug dans la fonction `Score` de `riskRegression`, dans le code joint, une solution de correction est proposée.

11. Quel modèle préférez vous pour la prédiction ?

12. On considère une machine dont les features sont données par `"id" = 2001,"pressureInd" = 96.4, "moistureInd" = 107.2,"temperatureInd" = 101.8, "team" = "TeamA", "provider" = "Provider2"` et qui a déjà fonctionné 63 semaines, donner une estimation de la probabilité qu'elle fonctionne encore deux semaines avec votre meilleur modèle.

13. Essayer de trouver un modèle qui améliore les performances des précédents.