Travail pratique

Contexte

Vous travaillez à titre de stagiaire en actuariat au sein d'une compagnie d'assurances vie œuvrant partout au Canada. Votre équipe (de 4 personnes ou moins) est en charge de la tarification des produits d'assurance invalidité.

Le mandat qui vous est confié pour ce travail consiste à utiliser la régression multiple que vous avez étudié au cours de votre formation afin de proposer le « meilleur » algorithme pour prédire la durée en jours des invalidités qui surviennent. En d'autres mots, une fois qu'on sait qu'il y a une réclamation en assurance invalidité, on veut prédire combien de temps cette dernière durera.

Pour ce faire, le département d'actuariat met à votre disposition les logiciel R, SAS et Excel (dans lesquels vous devez faire tous vos calculs), ainsi qu'un jeu de données nommé « MODELING_DATA.xlsx » (se référer à l'annexe I pour plus de détails sur les données).

Au final, vous devrez rédiger un rapport de qualité professionnelle pour expliquer votre démarche et interpréter vos résultats (se référer à l'annexe II pour les directives détaillées concernant le rapport).

Vous devrez remettre votre rapport électroniquement en plus d'une copie du fichier « EVALUATION_DATA.xlsx » dans lequel vos valeurs prédites pour la variable nommée « Duree_Invalidite » devront être ajoutés. Cette variable devra contenir la « meilleure » estimation, selon vos justifications, de la durée des invalidités des 1500 profiles fournis (se référer à l'annexe VI pour les détails sur cette partie).

Bien que vous soyez libre d'explorer de multiples avenues pour bâtir votre algorithme de prévision, votre patron s'attend à ce que les aspects suivants soient couverts par votre analyse :

1. Nettoyage et transformation des données (voir l'annexe III);
2. Analyse unidimensionnelle de la (log) durée des invalidités (voir l'annexe IV);
3. Sélection et validation du meilleur algorithme pour la durée d'une invalidité d'un assuré en fonction de ses caractéristiques (voir l'annexe V);
4. Prédiction de la durée des 1500 invalidités du jeu de données « EVALUATION_DATA.xlsx » à l'aide de l'algorithme sélectionné à l'étape précédente (voir l'annexe VI).

Annexe I : Description des données

Le jeu de données « MODELING_DATA.xlsx » est une extraction de 5000 exemples d'invalidités passés observés par votre employeur entre 2006 et 2016. Chaque ligne du jeu de données correspond donc à un sinistre passé en assurance invalidité.

À titre informatif, on mentionne que ces données ont été simulées pour les besoin du présent travail pratique. Les structures et tendances sont toutefois inspirées d'un projet réel de modélisation de durée des invalidités.

Le tableau suivant contient une brève description des 10 colonnes du jeu de données pour votre information:

Variable	Description
An_Debut_Invalidite	Année du début de l'invalidité
Mois_Debut_Invalidite	Numéro du mois où à commencé l'invalidité. Ex. : 1 = Janvier, 2 = Février, …, 12 = Décembre
Duree_Delai_Attente	Délai d'attente avant le commencement des prestations d'invalidité.
FSA	3 premiers caractères du code postal de la résidence habité par l'assuré au moment du début de l'invalidité.
Sexe	Sexe de l'assuré. Ex. : F = Femme, M = Homme.
Annee_Naissance	Année de naissance de l'assuré.
Code_Emploi	Code d'emploi de l'assuré. Ex. : 1 = Gestionnaire ou professionnel; 2 = Professionnel travaillant assis; 3 = Professionnel travaillant debout; 4 = Autre travailleur - formé pour son emploi; 5 = Autre travailleur – non formé pour son emploi.
Description_Invalidite	Champs texte contenant une description en anglais de l'invalidité
Salaire_Annuel	Salaire annuel de l'assuré au moment du début de l'invalidité.
Duree_Invalidite	[Variable réponse] : Durée en jours de l'invalidité.

Annexe II : Directives concernant le rapport

Le rapport écrit devrait résumer la procédure suivie pour l'analyse des données et se concentrer sur la justification du modèle et des résultats. Il n'est donc pas nécessaire — ni même souhaitable — de présenter de nouveau la théorie.

On doit donc viser la concision et la précision. Évidemment, le rapport devra être préparé à l'ordinateur avec une présentation générale de qualité professionnelle. Il devra de plus respecter strictement les exigences suivantes :

- Maximum de 15 pages;

- Utilisation d'une police de caractères de 12 points, interligne simple;

- Marges de 3 centimètres partout (haut, bas, gauche, droite);

- Les graphiques et tableaux doivent se trouver soit dans le texte, ou bien en annexe à la fin du rapport (assurez-vous de bien faire référence aux numéros de tableaux et graphiques si vous placez ces derniers à la fin de votre rapport);

- Le code informatique pertinent sera fourni dans un fichier séparé de votre rapport (utiliser une police non proportionnelle telle que `Courier` pour la présentation du code informatique);

- Utiliser un outil approprié pour les équations et autres notations mathématiques.

Annexe III : Nettoyage et transformation des données

Puisque les données représentent la base de tout modèle prédictif, et que plusieurs intervenants sont généralement impliqués dans le traitement des bases de données des assureurs, il est de mise de commencer votre analyse par l'identification et la correction des erreurs et/ou incohérences dans le jeu de données.

Dans le jeu de données, essayer de repérer le plus de problèmes que vous pouvez. Rapporter ces problèmes dans votre rapport, expliquer comment vous avez choisi de corriger la situation et appliquer les correctifs appropriées aux données.

Le volet transformation des données est aussi une partie très importante, car cette étape permettra d'augmenter la taille et la qualité des variables explicatives qui entreront dans vos modèles de régression multiples. Vous devriez passer beaucoup de temps sur cette section, et définir au moins une dizaine de nouvelles variables explicatives potentielles. La liste suivante contient quelques pistes (ne vous limitez pas à cette liste) pour la transformation de vos données :

1. Création d'une variable « Saison de l'invalidité » à partir de « Mois_Debut_Invalidite »;

2. Création d'une variable à dimension réduite (ex. : petit/moyen/grand) à partir de «Duree_Delai_Attente » ;

3. Création d'une variable « Province » à partir du premier caractère de « FSA »;

4. Création d'une variable urbain/rural à partir du deuxième caractère du « FSA ». **Astuce** : si le deuxième caractère est un « 0 », alors Postes Canada définit la région comme étant rurale;

5. Ajout de variables géo-démographiques (ex. : taux de chômage, nombre moyen d'enfants, densité de population, taux d'invalidité, température moyenne, … dans le FSA) à partir de « FSA ». **Astuce** : Chercher sur internet sur des sites comme Statistiques Canada, Environnement Canada, etc. Il n'y a pas de limite au nombre de nouvelles variables explicatives que vous pouvez trouver dans l'univers du Big Data. On vous demande d'ajouter **<u>un minimum de 5 variables</u>** à cette section;

6. Création de la variable « Âge de l'assuré au début de son invalidité » à partir de « Annee_Naissance »;

7. Dériver des indicateurs ou des variables ciblant des catégories d'âges particulières (ex. : ind(Age<=25), max(Age – 65,0), Age^2, …);

8. Création de multiples indicateurs de présences de mots clés (ex. : cancer, depression, ake, …) dans « Description_Invalidite ». **Astuce** : tentez de repérer les mots les plus fréquents et utiliser votre jugement ainsi que votre créativité;

9. Création de catégories de salaires (ex. : sous la moyenne / classe moyenne / élevé /…) à partir de « Salaire_Annuel »;

10. Transformation de la variable « Salaire_Annuel » (log, carré, cube, min, max, …).

Annexe IV : Analyse unidimensionnelle

L'analyse unidimensionnelle (ou descriptive) est la méthode traditionnellement utilisée par les compagnies d'assurances (particulièrement avant l'émergence de la régression dans ce domaine) pour construire leurs algorithmes de prévision.

À toutes fins pratique, cette méthode se résume à construire un tableau contenant la durée moyenne des invalidités observés (ou encore le log de de cette moyenne) pour chaque valeur possible de la variable explicative étudiée.

Voici une méthode possible pour effectuer ce travail pour la variable « An_Debut_Invalidite» avec la fonction `tapply` de R :

```
D<-tapply(dataset$ Duree_Invalidite,dataset$An_Debut_Invalidite,mean)
```

À partir de ces tableaux, faire ensuite des graphiques illustrant le comportement de la durée moyenne des invalidités (ou du log de la durée moyenne des invalidités selon votre choix et votre jugement) pour chaque variable explicative. Dans votre rapport interpréter ces graphiques et tableaux de sorte à déceler les tendances, et à proposer des transformations potentielles (si nécessaire) pour les variables explicatives (ex. : ln(x), x^2, e^x, etc.) pour l'analyse de l'annexe III.

Notons que ce travail est très répétitif en pratique, vous êtes donc encouragés à utiliser la programmation informatique afin de générer automatiquement tous vos tableaux et graphiques.

Annexe V : Meilleur modèle pour prédire la durée d'une invalidité

Cette analyse consiste à comparer plusieurs alternatives de régressions multiples servant ultimement à prédire la durée d'une invalidité en fonction des caractéristiques observables au début de cette invalidité. Vous pouvez examiner plusieurs avenues, mais votre patron s'attend à ce que vous compariez les 2 méthodes suivantes (entre elles et avec les résultats de l'analyse unidimensionnelle) :

1. Modèle de régression multiple **additif (aussi appelé Gaussien)**

- $Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

2. Modèle de régression multiple **multiplicatif (aussi appelé log-normale)** :

- $Z = \ln(Y) = \ln(\text{Duree\_Invalidite})$
- $Z = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

Vous devez analyser et commenter les aspects suivants dans votre rapport:

Les questions 1 à 5 doivent être faite en général, peu importe le modèle final choisi :

1. Comparer les 2 méthodes proposées précédemment. Notamment, expliquer les similitudes, les différences, ainsi que les avantages et désavantages des méthodes. Comparer aussi les pouvoirs prédictifs de chacun des modèles. Utiliser aussi des statistiques plus formelles étudiées en classe tels que le coefficient de détermination ajusté et/ou le F de Fisher. Votre patron aimerait aussi que vous fassiez des recherches sur le critère AIC et que vous tentiez d'utiliser ce critère pour comparer les 2 modèles ensembles.
2. Analyse des variables exogènes redondantes (ou dépendantes entre elles), si applicable.
3. Analyse et création si nécessaire de variables indicatrices pour les variables catégoriques.
4. Analyse et création si nécessaire de variables indicatrices pour les variables numériques.
5. Analyse des transformations requises **repérées lors de l'analyse unidimensionnelle** (ex. : ln(x), x^2, e^x, etc.) sur les variables numériques (si nécessaire).

Les questions 5 à 10 ne doivent être faites que sur votre modèle final, le modèle que vous aurai choisi pour faire vos prévisions sur le jeu de données « EVALUATION_DATA.xlsx » :

6. Application des procédures de sélection des modèles (forward, backward ou stepwise – choisir au moins une méthode) pour ne conserver que les variables utiles dans les modèles.
7. Calcul des intervalles de confiance pour les paramètres estimés et interprétation.
8. Construction de graphiques illustrant durée des invalidités moyennes prédites pour le modèle retenu. De plus, comparer les tendances observées avec celles de l'analyse unidimensionnelle.
9. Analyse de la variance (ANOVA), du coefficient de détermination (R^2) et du test global de la significativité du modèle obtenu.
10. Analyse des résidus du modèle final et construction du graphique des observations en fonctions des prévisions pour la variable dépendante. Observez-vous une droite à 45°?

Annexe VI : Prévision sur les données « EVALUATION_DATA.xlsx »

Le jeu de données « EVALUATION_DATA.xlsx » est, au même titre que le fichier de données « MODELING_DATA.xlsx », une extraction de 1500 sinistres en assurances invalidités pour laquelle la variable « Duree_Invalidite » a été volontairement effacée.

Noter que ce nouveau jeu de données est construit sous la même structure que celui dont vous disposez pour faire votre analyse de régression.

Vous devez donc utiliser l'algorithme de prévision que vous avez sélectionné afin de prédire la durée de l'invalidité de chacun des 1500 cas de ce jeu de donnée. Pour ce faire, vous devez remplir la colonne nommée « Duree_Invalidite » du fichier « EVALUATION_DATA.xlsx » avec le résultat de votre prévision.

Important : Vous devez annexer à votre rapport électronique une copie du fichier « EVALUATION_DATA.xlsx » pour lequel la variable « Duree_Invalidite» contiendra des valeurs prédites par votre modèle.

La qualité de votre prévision sera évaluée par la l'erreur quadratique (EQ) totale entre les vraies valeurs réellement observées pour la variable « Duree_Invalidite » et vos prévisions dans le fichier « EVALUATION_DATA.xlsx ». La formule suivante sera utilisée pour calculer le pointage de chaque équipe à la section ***Prévisions du jeu de données « EVALUATION_DATA.xlsx »*** de la grille de correction :

$$\text{Points}\left(\text{Votre Équipe}\right) = 15 \times \frac{EQ\left(\text{Pire Équipe}\right) - EQ\left(\text{Votre Équipe}\right)}{EQ\left(\text{Pire Équipe}\right) - EQ\left(\text{Meilleure Équipe}\right)}.$$

Annexe VII : Évaluation

Voici la grille de correction pour l'évaluation de ce travail. La note finale sera ramenée sur 8 dans le calcul de la note totale pour l'ensemble du cours.

	Points
Nettoyage et transformation des données	
Discussion des problèmes possibles	/2
Identification des erreurs dans les données et corrections (si applicables)	/2
Transformations des données :	
- Transformation et créations de variables explicatives additionnelles (non géo-démographiques)	/5
- Ajout d'au moins 5 variables géo-démographiques (exemple 5 de la liste de l'annexe III) à l'aide du FSA	/10
Analyse unidimensionnelle	
Exactitude des tableaux	/3
Exactitude des graphiques	/3
Interprétation des tendances	/10
Analyse de régression	
Comparaison des 2 méthodes entre elles (avantages et désavantages)	/5
Comparaison des 2 méthodes avec l'analyse unidimensionnelle (concordance ou non)	/5
Analyse des variables exogènes redondantes	/2.5
Création de variables indicatrices pour les variables non-numériques et commentaires	/2.5
Création de variables indicatrices pour les variables numériques et commentaires	/2.5
Transformations requises sur les variables numériques	/2.5
Comparaison des algorithmes de sélection des modèles	/5
Calcul des intervalles de confiance pour les paramètres estimés et interprétation	/5
Graphiques des coefficients de régression	/5
Analyse ANOVA, du R^2 et du test global de significativité du modèle	/5
Graphiques des observations en fonction des prévisions pour les variables réponses	/5
Analyse générale des résidus	/5
Prévisions du jeu de données « EVALUATION_DATA.xlsx »	
Allocation des points en fonction de l'erreur quadratique de votre équipe	/15
TOTAL	**/100**