Équipe 8

Charles Comeau
(111 185 421)

Nicholas Langevin
(111 184 631)

Andréanne Larouche
(111 190 518)

Apprentissage statistique en actuariat
ACT-3114

Analyse des données de renouvellement d'assurance

présenté à
Marie-Pier Côté

École d'actuariat
Université Laval
27 février 2020

Table des matières

Introduction

Les données qui seront analysées dans ce rapport proviennent du jeu de données “eudirectlapse” du paquetage “CASdatasets” de R. Dans le but de modéliser le statut de renouvellement d’une police d’assurance, il sera d’abord nécessaire de visualiser et de pré-traiter les données observées des 23 060 polices d’assurance. Il est à noter que la durée d’observation est de un an et que l’année visée et la compagnie demeurent inconnue. Ainsi, “lapse”, variable de type catégorielle qui indique si l’assuré résignera lors de son renouvellement d’assurance ou non, est la variable réponse qui sera intéressante.

Analyse exploratoire des données

Autres sections

Analyse en composantes principales

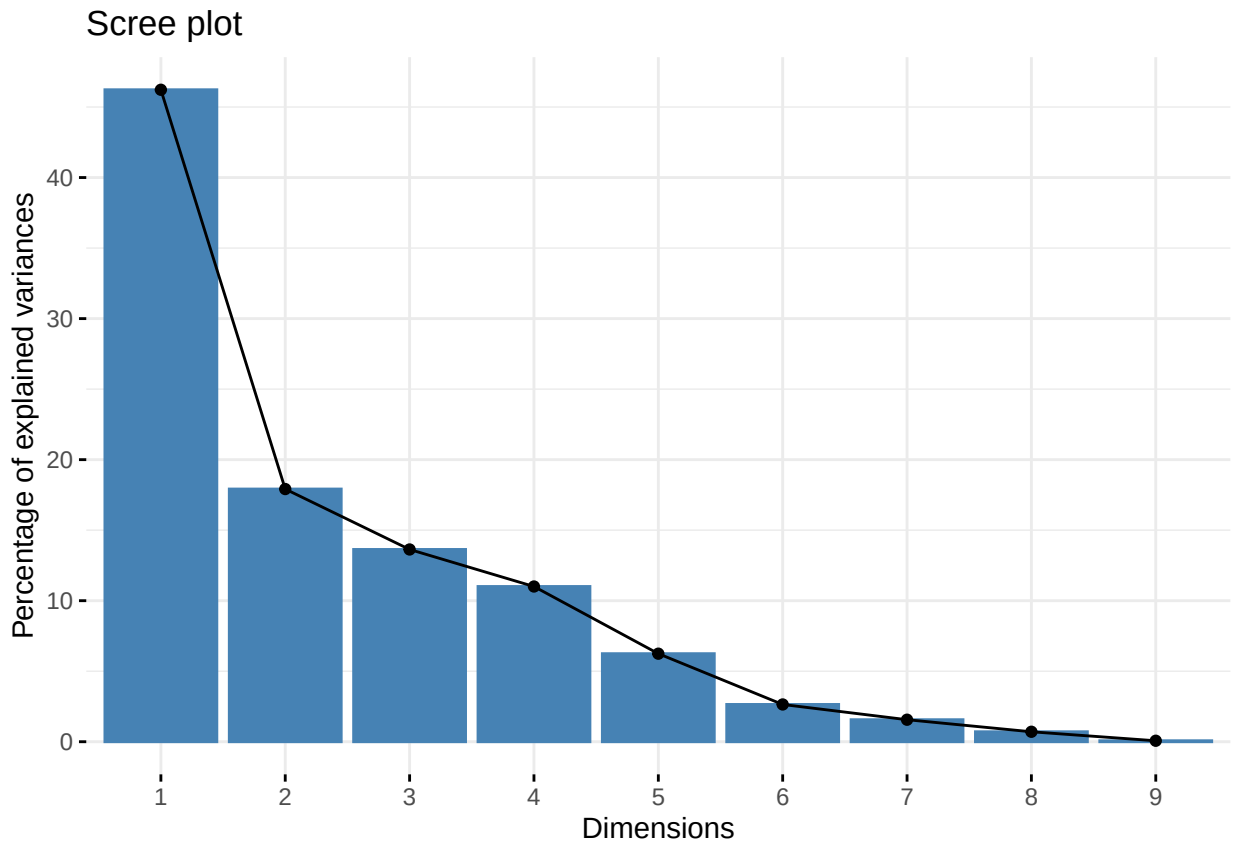
Étant donné que notre jeu de données contient `R nrow(Donnees_tempo)` observations, il peut être utile de visualiser les données à l'aide de l'analyse en composantes principales, appelé ACP. En effet, ce type d'analyse permet de mieux visualiser un jeu de données lorsque celui-ci est de grande dimension. Il sera ainsi possible de voir quelles variables explicatives sont plus intéressantes par leur impact sur la variance des composantes principales. Il est à noter qu'en général, on garde assez de composantes pour représenter entre 80 et 90 % de la variance totale.

Pour que cette méthode de visualisation puisse être utilisée, il sera nécessaire de prendre seulement les variables explicatives numériques de ce jeu de donnée. Les variables catégorielles ne seront pas analysées dans cette section car même en les transformant en variables numériques, elles ne seront pas représentative de leur signification.

On doit ensuite choisir le nombre de composantes principales. Cette étape peut être complétée en spécifiant le pourcentage de variance expliquée pour obtenir la cible précisée précedemment ou utiliser la méthode du coude sur le diagramme d'éboulis.

À l'aide du graphique ACP des variables on peut observer que pour la première composante principlae, une valeur élevé indique un contrat ayant une prime élevé, que ce soit la prime du marché, la prime pure, la prime finale ou la prime chargée lors du dernier renouvellement. Par contre, un assuré âgé et une prime élevé aura une valeur moindre qu'un assuré en bas âge. Un score élevé représente donc un assuré en bas âge ayant une prime élevé tandis qu'un score faible représente une personne plus âgée avec une faible prime d'assurance.

La deuxième composante principale représente, quant à elle, l'âge du véhicule assuré. Un score élevé est associé à des véhicules de moindres valeurs mais risquant d'avantage un bris de veillesse. Plus les polices d'assurance sont récentes et plus le score en sera augmenté. Ainsi, les polices d'assurances récente ayant des véhicules de l'année représenteront les scores les plus faibles pour cette composante.



```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.160051371            46.22279302                           46.22279
## comp 2 1.612312623            17.91458470                           64.13738
## comp 3 1.226885579            13.63206199                           77.76944
## comp 4 0.990548182            11.00609091                           88.77553
## comp 5 0.561893743             6.24326381                           95.01879
## comp 6 0.237840256             2.64266952                           97.66146
## comp 7 0.140583466             1.56203851                           99.22350
## comp 8 0.063602117             0.70669019                           99.93019
## comp 9 0.006282662             0.06980736                          100.00000
```

Selon le diagramme d'éboulis, il sera nécessaire de conserver 6 composantes principales et on observe, à l'aide des valeurs propres de la matrice de corélation, que 6 composantes principales permettent d'expliquer 97% de la variance totale.

À l'aide de la fonction PCA, nous pouvons voir la gravité des contributions pour chacune des variables sur chacune des composantes principales retenu. Ainsi, on remarque que la variable *prem_final* est celle ayant la plus grande contribution sur la première composantes principale, tout comme les variable *prem_last*, *prem_market* et *prem_pure*. Quand à la deuxième et à la sixième composante principale, c'est les variables *vehicl_age* et *vehicl_agepurchase* qui fournissent la plus grande contribution. Les variables explicatives *polholder_age* et *policy_age* ont les contributions les plus importantes pour les troisième et cinquième composantes principales. Pour ce qui est de la variable *policy_nbcontract*, elle est la variable apportant la plus grande contribution sur la quatrième composante principale.

```
##                        Dim.1        Dim.2      Dim.3        Dim.4       Dim.5
## polholder_age       4.2103455 3.497867e+00 37.6244932 1.550453e+00 51.04078055
## policy_age          3.3812324 1.431283e+01 29.6532650 4.729139e-01 43.40212041
## policy_nbcontract   0.1503547 6.839103e-04  2.5250221 9.711547e+01  0.08900955
## prem_final         22.0021626 3.247937e+00  1.4208179 3.647355e-05  0.04852303
## prem_last          20.9216918 2.662232e+00  0.8949391 3.793099e-07  0.15168005
## prem_market        21.0542024 7.166569e-01  0.1236893 1.248747e-04  2.64828035
## prem_pure          21.6552294 3.345531e+00  1.6673400 4.127283e-04  0.01605639
## vehicl_age          0.1524604 3.680304e+01 23.8993158 8.392614e-01  2.50874370
## vehicl_agepurchase  6.4723209 3.541323e+01  2.1911176 2.132797e-02  0.09480597
##                         Dim.6
## polholder_age       1.82242261
## policy_age          4.57649720
## policy_nbcontract   0.11661131
## prem_final          0.12108900
## prem_last           0.06253184
## prem_market         2.57053258
## prem_pure           0.21721009
## vehicl_age         35.29386259
## vehicl_agepurchase 55.21924278
```

En illustrant les contributions des variables pour chacune des composantes principales, il est plus facile de visualiser l'intensité de leur contribution.

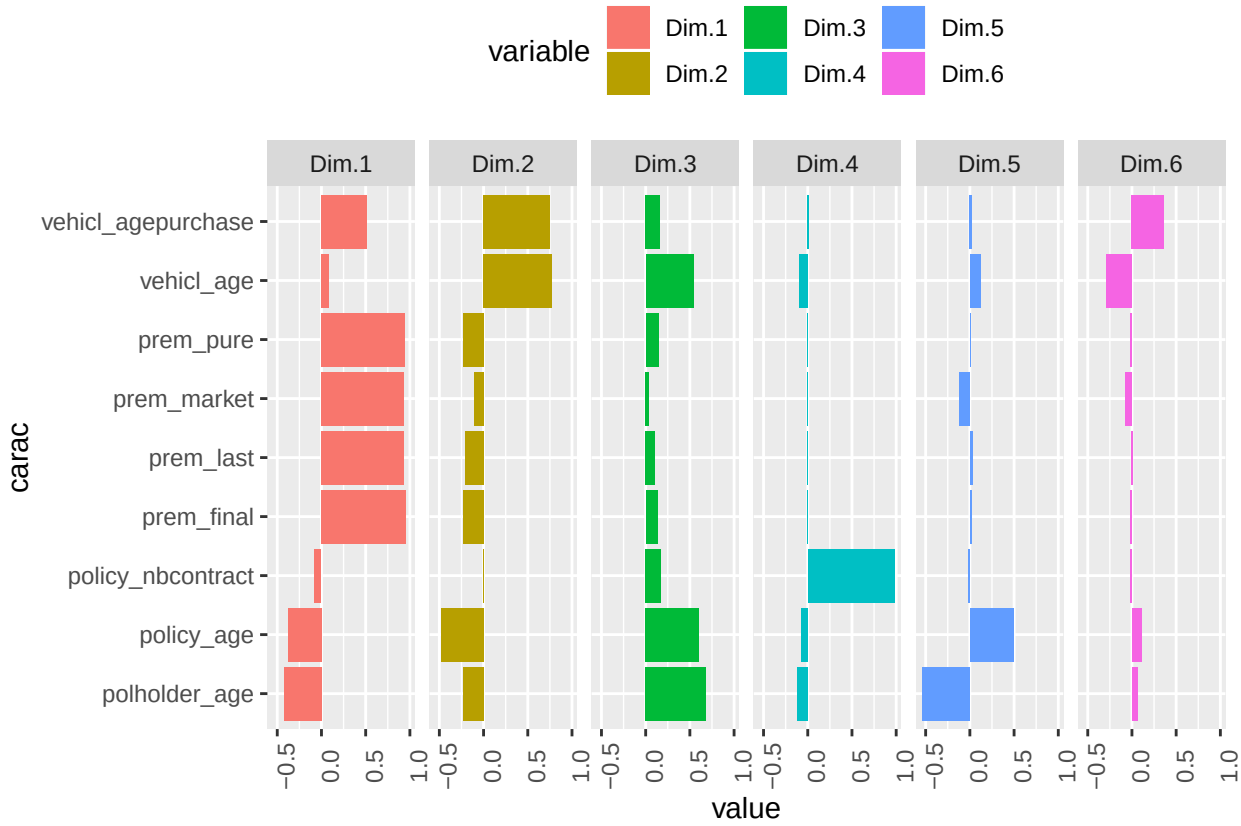
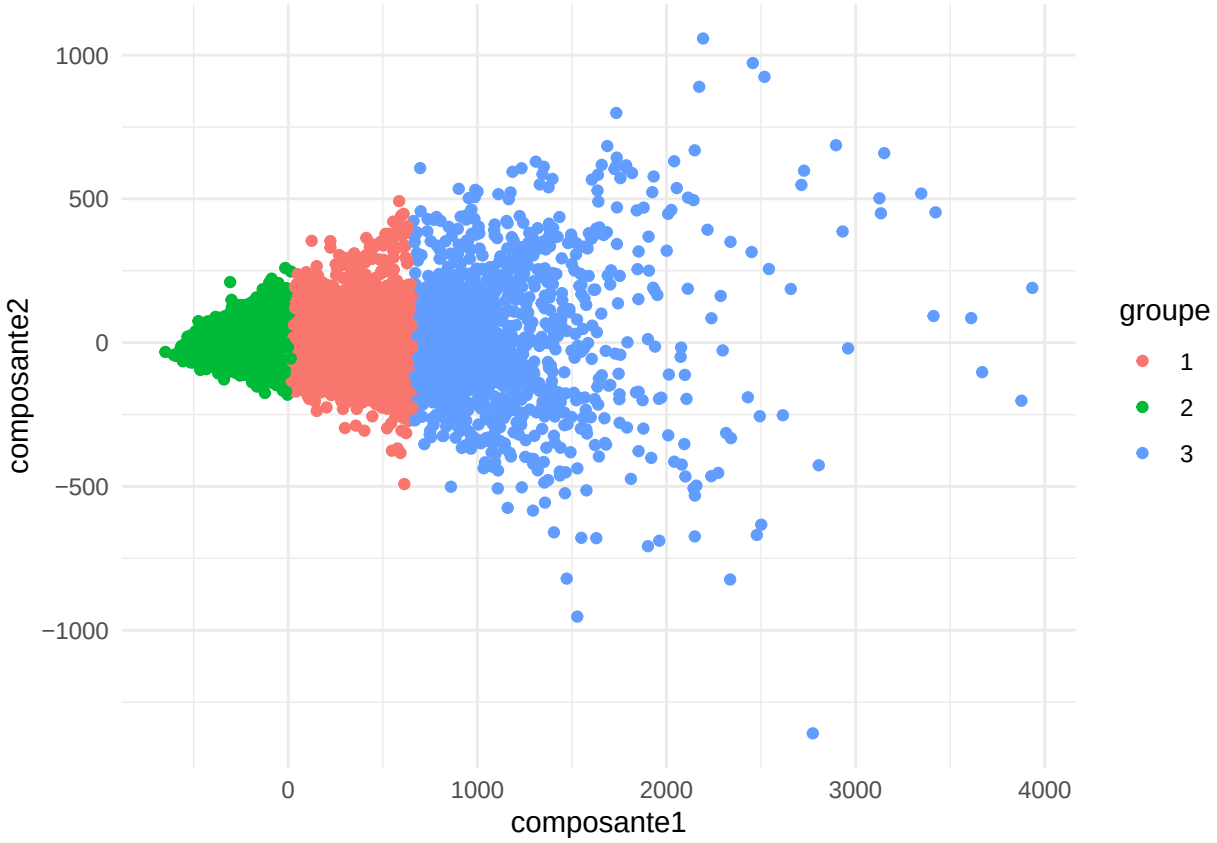


Diagramme de points (modèle) : À COMPLÉTER ET VÉRIFIER S'IL NE FAUT PAS INCLURE LES VARIABLES CATEGORIELLE ORDINALE TRANSFORMÉES EN VARIABLE NUMÉRQUES

Partitionnement en k moyennes

Le partitionnement en k moyennes est utiliser pour classifier les observations en k groupes distincts. La valeur de k est une valeur qu'on transmet pour indiquer le nombre de partitions désirées. Chaque observation sera ensuite assigné à un seul groupe. L'algorithme utilisée pour ce type de partitionnement a pour objectif de minimiser la variance intra-groupe.



Conclusion

Annexe

Notre jeu de données représente le statut de renouvellement pour 23 060 polices d'assurance basées sur un an d'observation. Les données receuillies proviennent d'une compagnie d'assurance inconnue dont l'année d'observation est également inconnue.