Rapport statistique sur les accidents corporels de la circulation

Florine GIRAUD et Muruo WANG

14 janvier 2019

INTRODUCTION

Nous avons choisi d'étudier la base de données des accidents corporels de la circulations routière en 2017 disponible ici: https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/

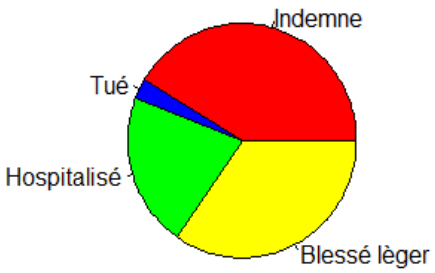
Un accident corporel de la circulation routière est un accident qui implique au moins une victime (c'est-à-dire une personne non indemne) et au moins un véhicule.

Les usagers impliqués dans l'accident sont classés dans différentes catégories :

- Les personnes **indemnes**
- Les personnes **tuées** : personnes qui décèdent du fait de l'accident, sur le coup ou dans les trente jours qui suivent l'accident.
- Les blessés dits « **hospitalisés** » : victimes hospitalisées plus de 24 heures
- **Les blessés légers** : victimes ayant fait l'objet de soins médicaux mais n'ayant pas été admises comme patients à l'hôpital plus de 24 heures

Pour notre étude, nous utilisons 3 bases de données de 2017 :

- **CARACTERISTIQUES** (date de l'accident, luminosité, commune, conditions météorologiques…)
- **LIEUX** (type et forme de route, état de la surface, infrastructure, présence d'une école…)
- **USAGER** (catégorie, sexe, année de naissance, gravité, place dans le véhicule, dispositif de sécurité)



La variable que nous allons étudier est la variable **gravité** qui prends les valeurs suivantes :

1. Indemne
2. Tué
3. Blessé léger
4. Blessé hospitalisé

Pour notre étude, nous avons aussi créée une variable **Vivant** qui est dichotomique (*Vivant = 0 signifie être mort*)

ENJEU

L'étude des facteurs qui influent sur la gravité des accidents de la route peut aider à améliorer la prévention en mettant l'accent sur les facteurs les plus importants (port de la ceinture de sécurité, prudence en cas de non-visibilité …). De plus, on pourrait prévoir les moments et les lieux les plus opportuns pour faire cette prévention si l'on arrive à établir un lien entre la gravité et le lieu et moment de l'accident.

IMPORTATION DU JEU DE DONNEES

Concernant l'importations de données, nous n'avons pas eu de difficulté particulière, nous avons fusionné les 3 tables qui nous paraissaient les plus intéressantes (usager, lieux et caractéristiques) en utilisant la colonne donnant le numéro de l'accident Num_Acc qui est présente dans les 3 tables.

1) IDENTIFICATION DES FACTEURS D'INFLUENCE

1.1) INFLUENCE DE L'AGE

Dans un premier temps, on veut regarder si l'âge de la personne accidenté influe sur son état après l'accident.

D'après le graphique ci-contre, les personnes tuées sont en moyenne plus vieilles que celles qui survivent. On peut donc supposer que l'âge est un facteur qui influe sur la gravité de l'accident.

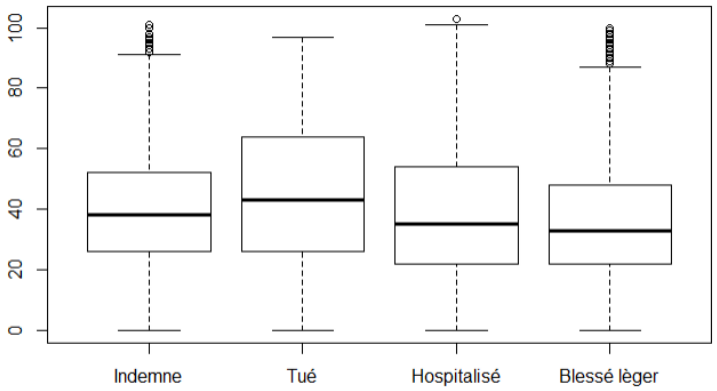


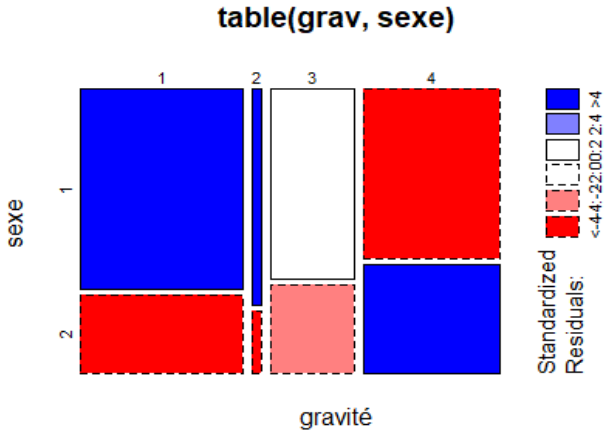
Figure 1-Boxplot de l'âge en fonction de la gravité

Nous faisons le test du chi2 pour vérifier cette hypothèse:

```
##  Pearson's Chi-squared test
##
## data:  accident$grav and accident$age
## X-squared = 5542.8, df = 306, p-value < 2.2e-16
```

La p-valeur obtenue est très faible donc on peut en conclure que les variables ne sont pas indépendantes.

1.2) INFLUENCE DU SEXE



Nous regardons si la répartition des genres des personnes accidentées pour voir si le sexe (1-Masculin/2-Féminin) peut influencer sur la gravité d'un accident. Cela pourrait servir à identifier la cible majoritaire de personnes à sensibiliser sur les accidents de la route.

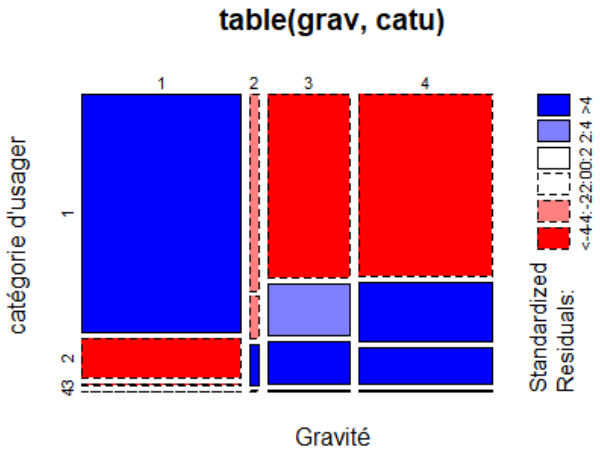
D'après le tableau croisé ci-contre, dans tous les cas de gravité, la proportion d'homme accidenté est supérieure à celle des femmes.

Nous faisons le test du chi2 pour vérifier que les variables **grav** et **sexe** ne sont pas indépendantes:

```
##  Pearson's Chi-squared test
##
## data:  accident$grav and accident$sexe
## X-squared = 1557.4, df = 3, p-value < 2.2e-16
```

La p-valeur obtenue est très faible donc on peut en conclure que les variables ne sont pas indépendantes. Le sexe est un facteur qui influe sur la gravité de l'accident.

1.3) INFLUENCE DE LA CATEGORIE D'USAGER



La table étudiée recense plusieurs catégories d'usagers (1-Conducteur/2-Passager/3-Piéton/4-Piéton en roller ou en trottinette), on veut regarder l'influence de cette caractéristique sur la gravité d'un accident.

D'après le graphique ci-contre, on voit les piétons sont les usagers les plus vulnérable. En effet, la proportion de piéton tués ou gravement blessés est beaucoup plus importante que celle des piétons indemnes qui est minime.

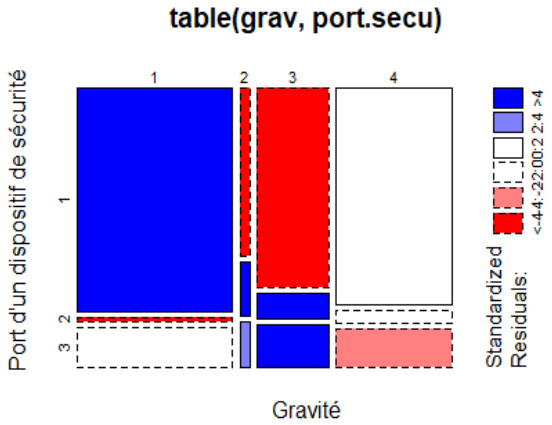
Nous faisons le test du chi2 pour vérifier que les variables **grav** et **catu** ne sont pas indépendantes :

```
##
##  Pearson's Chi-squared test
##
## data:  accident$grav and accident$catu
## X-squared = 10198, df = 9, p-value < 2.2e-16
```

La p-valeur obtenue est très faible donc on peut en conclure que les variables ne sont pas indépendantes.

1.4) INFLUENCE DE L'USAGE D'UN DISPOSITIF DE SECURITE

A partir des variables disponible dans les tables, on crée un variable **port.secu** qui indique si l'usager portait un dispositif de sécurité (ceinture, casque…) lors de l'accident:1-Oui/2-Non/3-On ne sait pas



D'après le tableau croisé ci-contre, on voit que la proportion des usagers qui ne portait pas de dispositif de sécurité (ceinture, casque …) et qui est tué est beaucoup plus important que celle qui s'en sortent indemne qui est minime.

On suppose donc que l'usage d'un dispositive de sécurité est un facteur d'influence de la gravité d'un accident.

Nous faisons le test du chi2 pour vérifier que les variables **grav** et **port_secu** ne sont pas indépendantes :

```
##
##  Pearson's Chi-squared test
##
## data:  accident$grav and accident$port.secu
## X-squared = 4328.4, df = 6, p-value < 2.2e-16
```

La p-valeur obtenue est très faible donc on peut en conclure que les variables ne sont pas indépendantes.

2) REGRESSION LOGISTIQUE

Dans un premier temps, nous avons choisi un modèle de régression pour exprimer la variable Vivant en fonction des autres variables disponibles et ensuite faire de la prédiction.

2.1) REGRESSION PAR L'AGE

D'abord, on choisit de faire une régression logistique de la variable catégorielle **Vivant** par la variable quantitative **age** (ANNEXE1 : REGRESSION DE VIVANT PAR AGE). L'objectif est de regarder comment l'âge influe sur la mortalité dans les accidents de la route.

Les graphes ci-dessous montre les probabilités a posteriori obtenues avec le modèle de régression logistique. Globalement, on voit que la probabilité de sortir vivant (non tué) d'un accident de la route est élevée mais cette probabilité est plus faible pour les personnes plus âgées.

Dans un second temps, on a enlevé les lignes contenant des valeurs d'âge supérieures à 90 ans car pour ces personnes l'effectif n'est pas assez conséquent pour pouvoir dresser un modèle fiable. Le modèle obtenu est plus proche de la réalité avec cette modification.

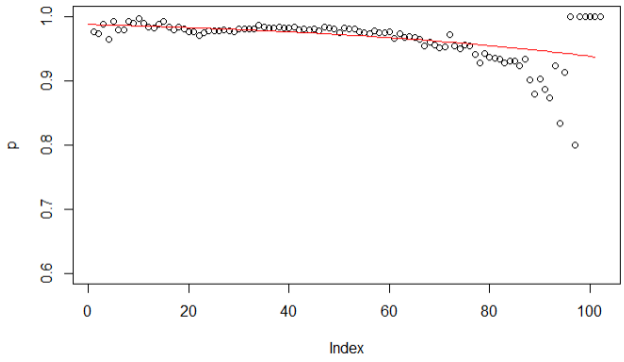


Figure 3-Probabilité de survie selon l'âge

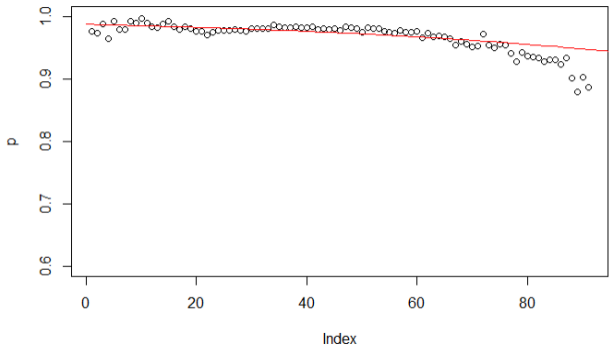
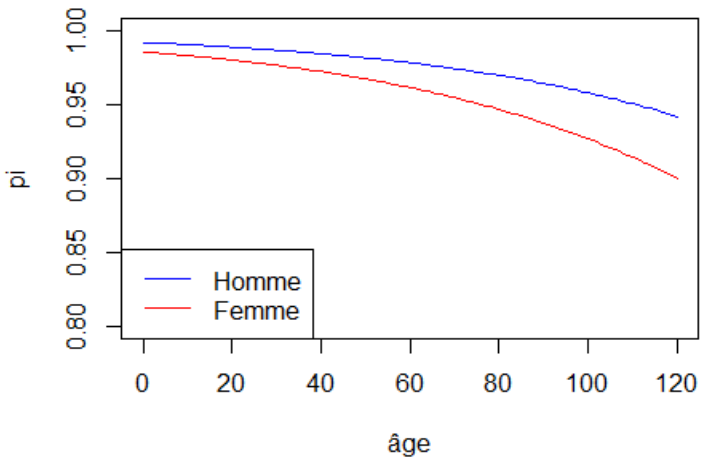


Figure 2- Probabilité de survie selon l' âge (âge <90 ans)

2.2) REGRESSION PAR L'AGE ET LE SEXE



Ensuite, on cherche à voir comment le sexe de l'usager influe sur la gravité de l'accident.

Le graphe ci-contre présente la probabilité de survie à un âge donnée pour les hommes et pour les femmes.

Ainsi, on voit qu'à âge égal, un homme a plus de chance de survivre à un accident de voiture qu'une femme.

On aurait aussi pu faire un test de Wald pour tester l'effet de la variable **sexe** sur notre modèle.

2.3) MODELE COMPLET AVEC LES VARIABLES DE LA TABLE USAGER

2.3.1) Création du modèle

On crée maintenant un modèle de régression logistique avec toutes les variables utilisables de la table usager : **sexe**, **age**, place dans le véhicule (**place**), catégorie d'usager (**catu**) et type de trajet effectué (**trajet**) et port d'un dispositif de sécurité (**port.secu**) (voirANNEXE2 : REGRESSION DE VIVANT PAR LES VARIABLES DE LA TABLE USAGER).

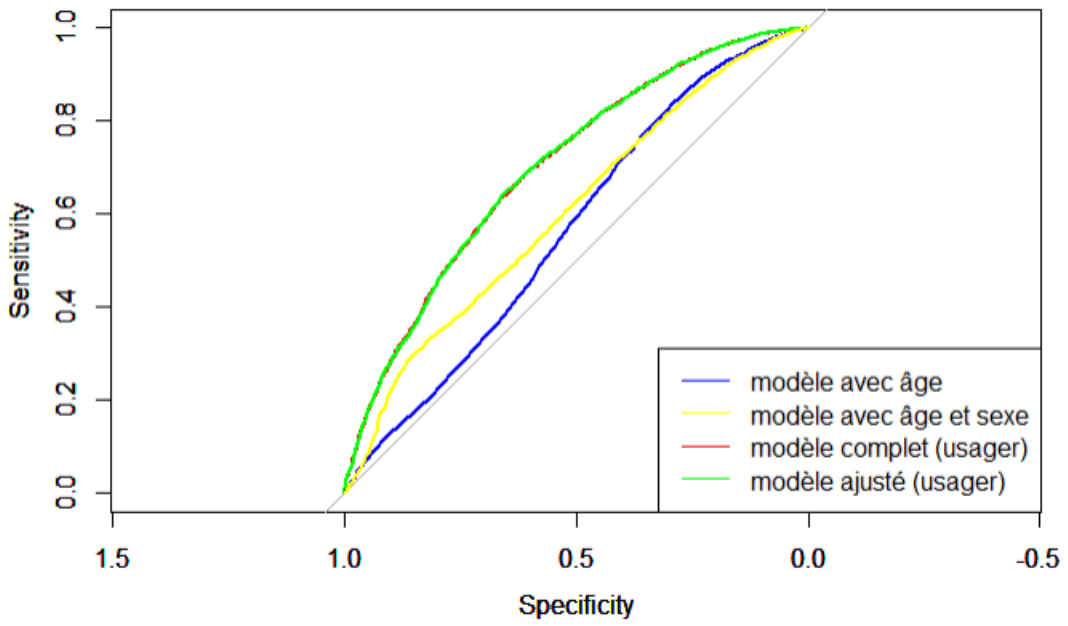
2.3.2) Ajustement du modèle

On veut ensuite regarder quels sont les facteurs prédominants dans la mortalité routière car cela pourrait aider à mieux cibler la prévention qui est faite pour réduire le nombre de mort sur la route.

On utilise un algorithme de type backward pour garder les variables utiles au modèle qui s'avèrent ici être : **sexe, age, trajet, port.secu** (voir ANNEXE 3 : BACKWARD MODELE USAGER).

2.3.3) Qualité du modèle

La courbe ROC ci-dessous nous permet de juger de la qualité des différents modèles dressés:



On voit que lorsque l'on rajoute des variables, le modèle est meilleur. En revanche, les modèles complets et ajusté sont aussi efficaces. C'est logique car on a enlevé les variables qui n'était pas utiles donc on n'a pas retiré de l'information entre les 2 modèles.

Le modèle n'est pas très performant, on va essayer de l'améliorer en ajoutant des données concernant les lieux des accidents et les caractéristiques (humidité du sol, largeur de la route, date, heure de la journée…)

2.4) MODELE COMPLET AVEC LES VARIABLES DE LA TABLE USAGER, CARACTERISTIQUE, LIEUX

2.4.1) Création du modèle

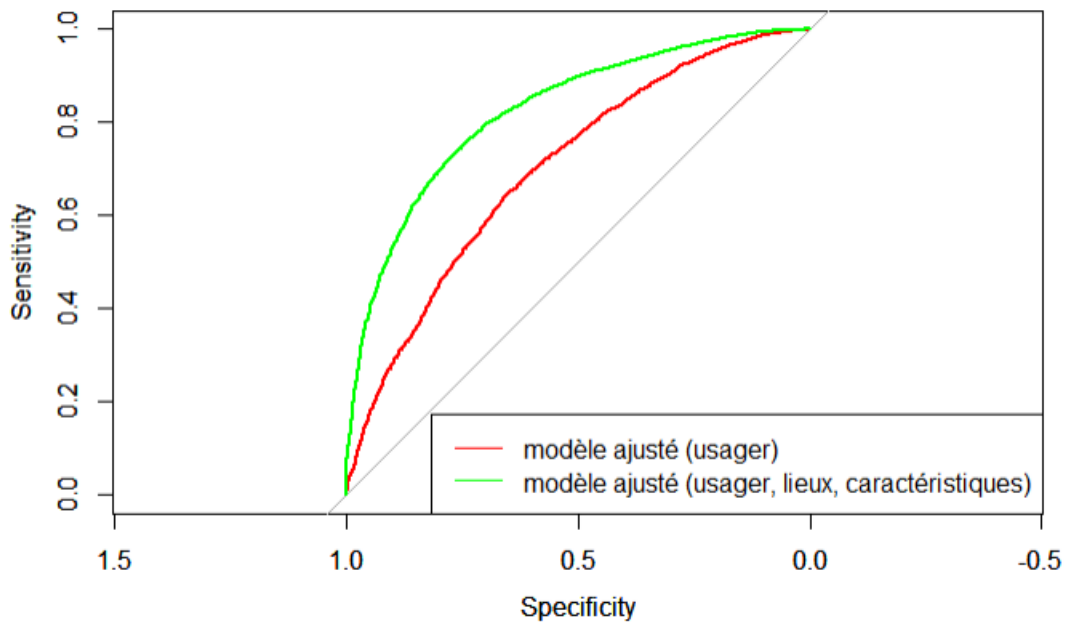
On crée maintenant un modèle de régression logistique avec toutes les variables utilisables des tables usager, caractéristique et lieux (voir ANNEXE 4 : REGRESSION DE VIVANT PAR LES VARIABLES DE LA TABLE USAGER, CARACTERISTIQUES ET LIEUX).

2.4.2) Ajustement du modèle

Toujours avec l'objectif d'identifier les facteurs prédominants dans la mortalité routière, on veut retirer les variables qui ne sont pas utiles au modèle. On utilise un algorithme de type backward pour garder les variables utiles au modèle (voir ANNEXE 5 : BACKWARD MODELE USAGER, CARACTERISTIQUES ET LIEUX) qui s'avèrent être ici : département(**dep**), **sexe**, **age**, **trajet**, **place**, **port.secu** , **catu**, largeur de la route (**larrout**), agglomération (**agg**), type d'intersection (**int**), catégorie de route (**catr**), état de la surface de la route (**surf**), forme de la route (**plan**), infrastructure (**infra**). L'algorithme a retiré les variables **mois** et conditions météorologiques (**atm**).

2.4.3) Qualité du modèle

La courbe ROC ci-dessous nous permet de juger de la qualité des différents modèles dressés:

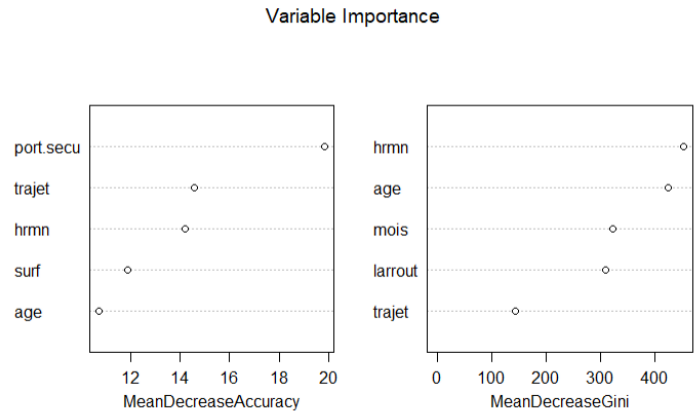


On voit ici que le modèle ajusté (après backward) est bien meilleur lorsqu'il prend en compte les variables des 3 tables que lorsqu'il prenait en compte seulement les informations usager. On peut en conclure que notamment les caractéristique temporelles et spatiales ont une importante dans la gravité des accidents de la route.

3) PREDICTION AVEC RANDOM FOREST

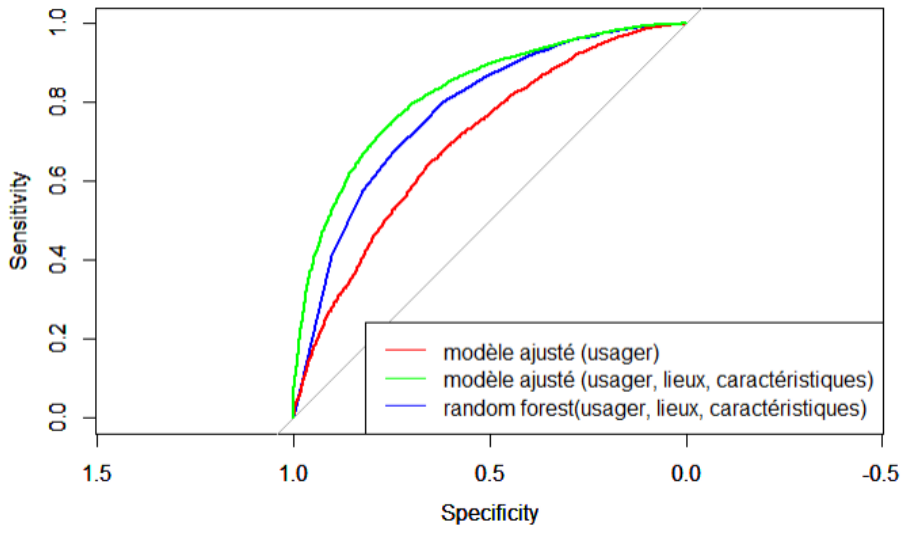
On a voulu utiliser un nouveau modèle (les forêts d'arbre décisionnels) pour essayer d'améliorer la qualité de notre prédiction (voir annexe 4).

Les 5 variables les plus importante sont données selon 2 méthodes (précision du modèle et valeur Gini) par les graphes ci-dessous :



On retrouve les variables **age** et **trajet** qui figure dans les 5 premières en termes d'importance selon les 3 méthodes utilisées (AIC, accuracy, Gini value). On peut donc penser qu'il est important de cibler la population pour faire de la prévention en insistant sur le fait qu'il faut être prudent dans tous ces trajets (même si l'on fait chaque jour le même chemin que l'on connaît très bien par exemple)

En termes de prédiction, ce modèle ne s'avère par meilleur que la régression logistique d'après la courbe ROC ci-dessus lorsque l'on utilise les mêmes variables.



CONCLUSION

Les modèles créés peuvent permettre de mieux cibler les campagnes de prévention concernant les accidents de la route. En effet, on a identifié les facteurs prédominants qui entraîne les morts sur la route. Les modèles permettent donc de déterminer le thème le plus important lors des campagnes de prévention (port de la ceinture…). On peut aussi prévoir la cible concernée par ces campagnes (âge, sexe, piéton…) et les périodes et les lieux dans lesquelles ces campagnes seraient les plus pertinentes (là où la mortalité est la plus élevée).

Notre modèle pourrait être amélioré. Il pourrait être plus performant si les tables contenaient des informations tels que le taux d'alcoolémie du conducteur, la vitesse lors de l'accident, l'éventuel non-respect du code de la route. On pourrait aussi utiliser les données des années précédentes pour être plus précis sur le modèle établi.

ANNEXE1 : REGRESSION DE VIVANT PAR AGE

```
modele_age1 <- glm(accident$Vivant ~ age, family=binomial(link='logit'))
summary(modele_age1)

##
## Call:
## glm(formula = accident$Vivant ~ age, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9596   0.1918   0.2099   0.2392   0.3614
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.366992   0.048777   89.53   <2e-16 ***
## age         -0.016548   0.001028  -16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25459  on 111260  degrees of freedom
## Residual deviance: 25209  on 111259  degrees of freedom
## AIC: 25213
##
## Number of Fisher Scoring iterations: 6

confint(modele_age1)

## Waiting for profiling to be done...

##                   2.5 %      97.5 %
## (Intercept)  4.27186249  4.46308228
## age         -0.01855935 -0.01452966
```

ANNEXE2 : REGRESSION DE VIVANT PAR LES VARIABLES DE LA TABLE USAGER

```
modele_complet1 <-glm(Vivant ~ sexe + age + trajet + port.secu + catu +place,family=binom
ial(link='logit'))
summary(modele_complet1)

##
## Call:
## glm(formula = Vivant ~ sexe + age + trajet + port.secu + catu +
##     place, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3357   0.1535   0.1899   0.2338   0.8936
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.723823   0.071634  65.944  < 2e-16 ***
## sexe2        0.646604   0.050617  12.774  < 2e-16 ***
## age         -0.017693   0.001029 -17.191  < 2e-16 ***
## trajet1     -0.047861   0.076190  -0.628  0.52988
## trajet2     -0.243604   0.173658  -1.403  0.16068
## trajet3     -0.390063   0.127321  -3.064  0.00219 **
## trajet4      0.713615   0.104986   6.797 1.07e-11 ***
## trajet5     -0.596012   0.054026 -11.032  < 2e-16 ***
## trajet9      0.302467   0.099360   3.044  0.00233 **
## port.secu2  -1.706049   0.055811 -30.568  < 2e-16 ***
## port.secu3  -0.349492   0.055405  -6.308 2.83e-10 ***
## catu2       -0.048724   0.072959  -0.668  0.50424
## place       -0.011304   0.021482  -0.526  0.59875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25459  on 111260  degrees of freedom
## Residual deviance: 23868  on 111248  degrees of freedom
## AIC: 23894
##
## Number of Fisher Scoring iterations: 7
```

ANNEXE 3 : BACKWARD MODELE USAGER

```
backwards = (step(modele_complet1, print=TRUE))
```

```
## Start:  AIC=23893.96
## Vivant ~ sexe + age + trajet + port.secu + catu + place
##
##             Df Deviance   AIC
## - place      1    23868 23892
## - catu       1    23868 23892
## <none>            23868 23894
## - sexe       1    24049 24073
## - age        1    24152 24176
## - trajet     6    24258 24272
## - port.secu  2    24578 24600
##
## Step:  AIC=23892.23
## Vivant ~ sexe + age + trajet + port.secu + catu
##
##             Df Deviance   AIC
## - catu       1    23870 23892
## <none>            23868 23892
## - sexe       1    24049 24071
## - age        1    24153 24175
## - trajet     6    24259 24271
## - port.secu  2    24582 24602
##
## Step:  AIC=23892.06
## Vivant ~ sexe + age + trajet + port.secu
##
##             Df Deviance   AIC
## <none>            23870 23892
## - sexe       1    24053 24073
## - age        1    24158 24178
## - trajet     6    24272 24282
## - port.secu  2    24588 24606
```

ANNEXE 4 : REGRESSION DE VIVANT PAR LES VARIABLES DE LA TABLE USAGER, CARACTERISTIQUES ET LIEUX

```
modele_complet2 <-glm(Vivant ~  dep+sexe + age + trajet+place + port.secu +catu + larrout
+hrmn + mois +lum+atm+agg+int + catr+surf+plan+infra ,family=binomial(link='logit'))
summary(modele_complet2)

##
## Call:
## glm(formula = Vivant ~ dep + sexe + age + trajet + place + port.secu +
##     catu + larrout + hrmn + mois + lum + atm + agg + int + catr +
##     surf + plan + infra, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6906   0.0933   0.1380   0.2222   1.3918
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.425e+01  6.169e+02   0.023 0.981570
## dep          3.101e-04  7.087e-05   4.375 1.21e-05 ***
## sexe2        5.687e-01  5.162e-02  11.017  < 2e-16 ***
## age         -1.834e-02  1.106e-03 -16.579  < 2e-16 ***
## trajet1      2.420e-01  7.860e-02   3.080 0.002073 **
## trajet2      1.824e-01  1.852e-01   0.985 0.324796
## trajet3     -1.517e-01  1.314e-01  -1.155 0.248289
## trajet4      9.162e-01  1.073e-01   8.543  < 2e-16 ***
## trajet5     -1.331e-01  5.645e-02  -2.359 0.018334 *
## trajet9      3.910e-01  1.018e-01   3.841 0.000123 ***
## place        5.458e-02  2.158e-02   2.529 0.011439 *
## port.secu2  -1.832e+00  6.048e-02 -30.290  < 2e-16 ***
## port.secu3  -2.780e-01  5.755e-02  -4.830 1.36e-06 ***
## catu2        1.583e-01  7.382e-02   2.144 0.032054 *
## larrout     -1.265e-03  2.637e-04  -4.797 1.61e-06 ***
## hrmn         2.391e-04  3.408e-05   7.016 2.29e-12 ***
## mois2        8.156e-02  1.134e-01   0.719 0.472121
## mois3       -2.742e-02  1.067e-01  -0.257 0.797167
## mois4        4.758e-02  1.082e-01   0.440 0.659996
## mois5       -1.006e-01  1.057e-01  -0.952 0.341003
## mois6       -1.936e-02  1.042e-01  -0.186 0.852550
## mois7       -6.147e-02  1.030e-01  -0.597 0.550788
## mois8       -9.689e-02  1.058e-01  -0.916 0.359690
## mois9       -2.890e-02  1.056e-01  -0.274 0.784372
## mois10      -6.614e-02  1.037e-01  -0.638 0.523490
## mois11       1.221e-01  1.076e-01   1.135 0.256255
## mois12       1.231e-01  1.060e-01   1.162 0.245221
## lum2        -3.505e-01  7.750e-02  -4.522 6.11e-06 ***
## lum3        -5.830e-01  5.545e-02 -10.514  < 2e-16 ***
## lum4        -8.941e-01  1.864e-01  -4.796 1.62e-06 ***
## lum5        -2.985e-01  7.962e-02  -3.750 0.000177 ***
## atm2        -1.298e-02  1.005e-01  -0.129 0.897288
## atm3         1.123e-01  1.596e-01   0.704 0.481608
## atm4         3.372e-01  3.152e-01   1.070 0.284701
```

```
## atm5         8.369e-03  1.613e-01   0.052 0.958627
## atm6        -5.651e-01  2.567e-01  -2.202 0.027694 *
## atm7        -1.354e-01  1.557e-01  -0.869 0.384692
## atm8        -1.533e-01  1.122e-01  -1.366 0.171955
## atm9        -1.520e-01  2.169e-01  -0.701 0.483405
## agg2         1.293e+00  6.068e-02  21.303  < 2e-16 ***
## int1        -8.823e+00  6.169e+02  -0.014 0.988589
## int2        -8.277e+00  6.169e+02  -0.013 0.989295
## int3        -8.118e+00  6.169e+02  -0.013 0.989501
## int4        -8.294e+00  6.169e+02  -0.013 0.989273
## int5        -8.142e+00  6.169e+02  -0.013 0.989469
## int6        -8.146e+00  6.169e+02  -0.013 0.989465
## int7        -7.474e+00  6.169e+02  -0.012 0.990334
## int8        -9.985e+00  6.169e+02  -0.016 0.987086
## int9        -7.904e+00  6.169e+02  -0.013 0.989777
## catr2       -8.116e-01  9.630e-02  -8.428  < 2e-16 ***
## catr3       -1.355e+00  7.787e-02 -17.398  < 2e-16 ***
## catr4       -6.265e-01  9.823e-02  -6.378 1.80e-10 ***
## catr5       -8.717e-01  7.284e-01  -1.197 0.231414
## catr6       -1.331e-01  5.159e-01  -0.258 0.796369
## catr9       -1.384e+00  1.936e-01  -7.147 8.85e-13 ***
## surf1       -7.570e-01  2.559e-01  -2.959 0.003090 **
## surf2       -7.234e-01  2.678e-01  -2.701 0.006916 **
## surf3       -1.308e+00  3.874e-01  -3.376 0.000736 ***
## surf4       -1.897e+00  6.145e-01  -3.087 0.002019 **
## surf5       -1.029e+00  4.616e-01  -2.228 0.025859 *
## surf6       -9.722e-01  6.612e-01  -1.470 0.141460
## surf7       -7.273e-01  3.301e-01  -2.203 0.027589 *
## surf8       -5.265e-01  5.756e-01  -0.915 0.360383
## surf9       -1.196e+00  3.165e-01  -3.780 0.000157 ***
## plan1       -1.959e-01  9.235e-02  -2.121 0.033934 *
## plan2       -4.441e-01  1.045e-01  -4.249 2.14e-05 ***
## plan3       -5.915e-01  1.028e-01  -5.752 8.84e-09 ***
## plan4       -2.068e-01  1.635e-01  -1.265 0.206008
## infra1       3.649e-01  2.750e-01   1.327 0.184547
## infra2      -3.527e-01  1.220e-01  -2.890 0.003851 **
## infra3       3.048e-01  1.802e-01   1.692 0.090653 .
## infra4      -1.540e+00  3.047e-01  -5.055 4.30e-07 ***
## infra5      -8.966e-02  1.296e-01  -0.692 0.488910
## infra6      -3.399e-01  3.282e-01  -1.035 0.300502
## infra7       1.034e+01  9.374e+01   0.110 0.912205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25459  on 111260  degrees of freedom
## Residual deviance: 21231  on 111186  degrees of freedom
## AIC: 21381
##
## Number of Fisher Scoring iterations: 13
```

ANNEXE 5 : BACKWARD MODELE USAGER, CARACTERISTIQUES ET LIEUX

```
backwards = (step(modele_complet2, print=TRUE))
```

```
## Start:  AIC=21380.64
## Vivant ~ dep + sexe + age + trajet + place + port.secu + catu +
##     larrout + hrmn + mois + lum + atm + agg + int + catr + surf +
##     plan + infra
##
##             Df Deviance   AIC
## - mois      11    21243 21371
## - atm        8    21241 21375
## <none>            21231 21381
## - catu       1    21235 21383
## - place      1    21237 21385
## - surf       9    21254 21386
## - dep        1    21250 21398
## - larrout    1    21251 21399
## - infra      7    21268 21404
## - hrmn       1    21280 21428
## - plan       4    21292 21434
## - int        9    21328 21460
## - lum        4    21356 21498
## - sexe       1    21363 21511
## - trajet     6    21405 21543
## - age        1    21499 21647
## - catr       6    21708 21846
## - agg        1    21740 21888
## - port.secu  2    21956 22102
##
## Step:  AIC=21370.85
## Vivant ~ dep + sexe + age + trajet + place + port.secu + catu +
##     larrout + hrmn + lum + atm + agg + int + catr + surf + plan +
##     infra
##
##             Df Deviance   AIC
## - atm        8    21252 21364
## <none>            21243 21371
## - catu       1    21247 21373
## - place      1    21250 21376
## - surf       9    21268 21378
## - dep        1    21262 21388
## - larrout    1    21263 21389
## - infra      7    21279 21393
## - hrmn       1    21296 21422
## - plan       4    21306 21426
## - int        9    21340 21450
## - lum        4    21362 21482
## - sexe       1    21376 21502
## - trajet     6    21421 21537
## - age        1    21508 21634
## - catr       6    21722 21838
## - agg        1    21757 21883
## - port.secu  2    21972 22096
```

```
##
## Step:  AIC=21364.12
## Vivant ~ dep + sexe + age + trajet + place + port.secu + catu +
##     larrout + hrmn + lum + agg + int + catr + surf + plan + infra
##
##             Df Deviance   AIC
## <none>           21252 21364
## - catu       1    21257 21367
## - place      1    21259 21369
## - surf       9    21277 21371
## - dep        1    21271 21381
## - larrout    1    21273 21383
## - infra      7    21288 21386
## - hrmn       1    21305 21415
## - plan       4    21315 21419
## - int        9    21349 21443
## - lum        4    21372 21476
## - sexe       1    21385 21495
## - trajet     6    21431 21531
## - age        1    21518 21628
## - catr       6    21737 21837
## - agg        1    21769 21879
## - port.secu  2    21984 22092
```

ANNEXE 6 : RANDOM FOREST

```
# Load library
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

sample.ind <- sample(2, nrow(accident),    replace = T,prob = c(0.6,0.4))
accident.dev <- accident[sample.ind==1,]
accident.val <- accident[sample.ind==2,]
accident.dev <-accident.dev[1:length(accident.val$Vivant),]


varNames <- names(accident.dev)
# Exclude ID or Response variable

accident.rf <- randomForest(Vivant ~  age+sexe  + trajet + port.secu +catu+place+ larrout
+hrmn + mois +lum+atm+agg+int + catr+surf+plan+infra ,accident.dev, ntree=100,  importanc
e=T)


#
# Variable Importance Plot
varImpPlot(accident.rf,sort = T,main="Variable Importance",n.var=5)
```

```
accident.val$predicted.response <- predict(accident.rf ,accident.val,type=c("prob"))

g5 <- roc(accident.val$Vivant,accident.val$predicted.response[,2])
plot(g5, col="blue")
par(new=TRUE)
plot(g,col="red")
par(new=TRUE)
plot(g2,col="green")
par(new=TRUE)
legend(legend = c("modèle ajusté (usager)","modèle ajusté (usager, lieux, caractéristique
s)","random forest(usager, lieux, caractéristiques)"),
       lty =rep(1,2,3) ,col=c("red","green","blue"),x = "bottomright")
```