Bonjour, nous avons apporté quelques ajustements que nous avons jugées pertinents sur notre jeu de données afin qu’il puisse par la suite être utilisé pour l’ajustement des modèles.

D’abord nous avons dû faire un traitement des valeurs manquantes. On remarque qu’il y a trois variables pour lesquelles des valeurs sont manquantes : l’utilisation du véhicule (catégorielle à deux niveaux), le type de stationnement (catégorielle), et la différence de statut qui pourrait avoir entre le propriétaire de la police et le conducteur principal (catégorielle).

On constate que pour la variable de la différence de statut il y a un patron de non réponse monotone puisque l’utilisation du véhicule et le type de stationnement sont manquantes également. Pour caruse et le type de garage elles sont manquantes tous les deux à certains moments et parfois seulement une seule des deux est manquante.

Les trois variables sont de type catégoriel, nous avons effectué des tests de khi-carré au niveau pour voir si les données sont MCAR. Les tests ont révélé que les données n’étaient pas MCAR au niveau significatif de 0.1%. Il est donc très probable que les données soient MAR ou NMAR. Puisque nous n’avons pas d’information tant qu’à la collecte des données il nous est impossible de détecter des patrons indiquant que les données sont NMAR. Quoi qu’il en soit il y a un total de presque 20% des données qui sont touchées nous savons don décidé de procédé à une imputation multiple (imputation stochastique répété 5 fois, régression logistique pour caruse (deux niveaux) et régression polynomiale pour les deux autres variables (plusieurs niveaux)).

Une autre manipulation qui a été nécessaire c’est la réduction des niveaux pour la variable catégorielle de la puissance du moteur du véhicule. On constate qu’il y avait initialement un niveau regroupant les véhicules d’une puissance de 125 à 300 puis qu’il y a également des niveaux où la puissance du véhicule se situe entre cet intervalle. De plus, on constate que les niveaux se situant dans l’intervalle sont très peu représenter dans la base de données puisqu’ils ne contenaient peu d’observation. Nous avons donc trouvé justifiable d’ajouté ces observations dans le niveau général 125 à 300. La variable à donc subie une réduction de niveau, voici la distribution de la puissance du moteur avant et après la réduction de niveau.

Passons maintenant à la description des modèles. Pour le modèle de base, nous utiliserons un modèle linéaire généralisé avec une distribution Bernoulli. Le lien utilisé pour la modélisation de la probabilité de résignation est le lien logistique. Donc, nous avons que le logarithme de la prob divisé par 1 – la prob est égale à notre prédicteur linéaire. Une façon d’estimer les coefficients du glm est de maximiser la vraisemblance d’une Bernoulli avec cette équation. La fonction glm que nous avons utilisé en R pour calculer l’estimation de nos coefficients utilise plutôt le score de Fisher mais les méthodes donnent les mêmes résultats.

Ensuite pour rendre le modèle plus simple nous avons décidé d’effectuer une sélection de variable en effectuant des tests du rapport de vraisemblance avec niveau de confiance 99%. Au total 7 variables ont été rejetées rendant le modèle moins complexe. Cela veut dire que ces 7 variables n’ont pas d’incidence sur la probabilité de résignation prédite par le modèle de base.

L’avantage d’un glm dans notre situation de modélisation c’est que l’estimation des coefficients est fait sur un grand nombre d’observation réduisant ainsi la variance des estimateurs. Un autre avantage non négligeable est que le glm est facile à interpréter, on voit explicitement l’impact des variables sur la probilité de résignation (effet quantitatif).

Le deuxième modèle est un modèle linéaire avec régulation. La distribution choisie reste la distribution Bernoulli par contre une pénalité Lasso s’ajoute. Donc nous conservons le lien logistique pour la modélisation de la probabilité de résignation comme nous avions pour le glm de base. Les estimateurs sont obtenus en maximisant l’équation suivante qui correspond à la vraisemblance d’une Bernoulli pour laquelle on applique une pénalité fonction de l’hyperparamètre lambda et de la valeur absolue de coefficients. Avant d’estimer les coefficients il faut trouver le coefficient de pénalité optimal. La valeur optimale du lambda à été choisi par validation croisée à 10 plis selon le métrique AUC