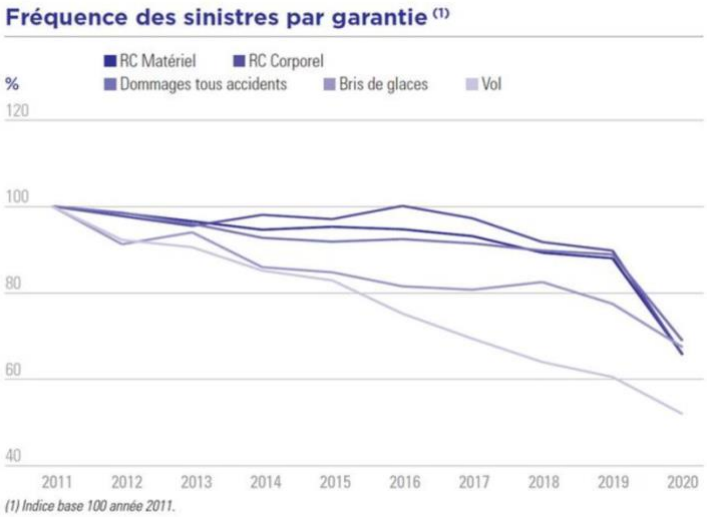
Rapport pour le projet d'Apprentissage Supervisé

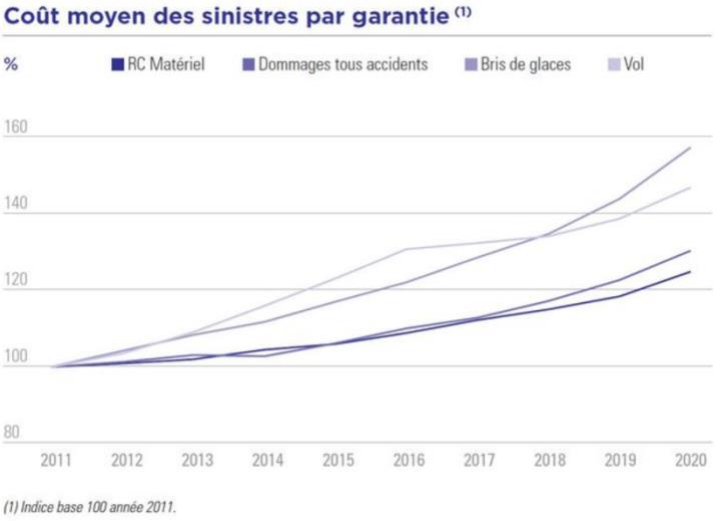
Machine learning appliqué à l'accidentologie en France

Sommaire

Résumé

Aujourd'hui, la survenue d'accidents corporels dans la circulation routière reste un événement très couteux pour les compagnies d'assurance.
Pour illustrer ce propos, jetons un œil sur l'évolution des charges et fréquences des sinistres moyennes en France en assurance auto ces dernières années : les coûts croissent chaque année.





Grâce à l'Open Data mise à disposition par le Ministère de l'intérieur Français plus précisément l'ONISR[1], les bases de données des accidents corporels de la circulation routière sont une mine d'or pour l'assureur.
Son enjeu de demain est double : mieux gérer et anticiper les accidents à caractère grave et onéreux pour réduire ses coûts, tirer profit de cette richesse d'informations pour se développer sur l'assurance télématique.

C'est dans ce contexte que cette étude s'est menée pour apporter un éclairage en matière de prédiction de l'accidentologie en France entre 2015 et 2019 par la mesure de l'impact de certaines caractéristiques liées aux accidents.

Pour ce faire, nous allons modéliser cette gravité au moyen de diverses approches.
Dans un premier temps et en première intention, nous implémenterons une régression GLM classique de la fréquence par model point, pour un premier aperçu.
Dans une deuxième étape, nous mettrons en œuvre différentes méthodes de machine learning afin de trouver le modèle prédisant le mieux les accidents graves.

Pour finir, nous aborderons l'interprétabilité du modèle retenu afin de mettre en lumière les différentes catégories de contextes de conduite relatifs à la gravité de l'accident.

En guise de conclusion nous ouvrirons notre réflexion vers les usages et applications qui peuvent en découler dans un contexte assurantiel.

[1] ONISR Observatoire National Interministériel de la Sécurité Routière

Remerciements

En premier lieu, je souhaite remercier l'ensemble des professeurs du Certificat en Data Science pour l'Actuariat de l'Institut du Risk Management mais également de l'Université Paris-Dauphine PSL pour la qualité de leurs enseignements et leurs pédagogies respectives.

Je remercie chaleureusement André GRONDIN, mon tuteur, pour la disponibilité et la qualité des échanges qu'il m'a accordées, ainsi que tous mes camarades de promotion de la formation Data science pour l'Actuariat et de l'Université Paris-Dauphine PSL pour leurs conseils, et les bons moments d'échanges et de partages vécus pendant ces 18 mois de formation.

Pour finir mais qui n'est pas des moindres, je tiens à exprimer ma profonde gratitude envers mon mari, ma famille et mes proches pour m'avoir soutenue et encouragée à accomplir cette formation exigeante et à réaliser ce mémoire et, par-dessus tout, mes enfants Andri-Yann, Andy et Ally-Soa, pour les nombreuses heures que je leur ai volées.

Introduction

Dans un contexte où les accidents corporels graves ne sont jamais choses rares, mon choix d'étude s'est porté sur une Open data à disposition pour étudier les influences de notre environnement de conduite habituel sur la potentielle gravité de l'accident.

L'objectif de ce mémoire est de parvenir à prédire au mieux les contextes de conduite qui favorisent les accidents graves. Celui-ci se propose notamment d'établir une analyse comparative de différents modèles prédictifs de la gravité des accidents corporels reposant sur des techniques statistiques classiques ou de machine learning.

Il sera décomposé en trois parties.

La première grande partie s'attache aux éléments constitutifs à la base de données d'étude ainsi qu'à l'analyse exploratoire détaillée de cette dernière.

Dans une deuxième partie, nous nous intéresserons à la mise en œuvre technique des modèles prédictifs en précisant toutes les étapes de la méthodologie et des choix retenus.

Une troisième partie sera consacrée à l'interprétabilité du modèle en identifiant le ou les segments de risque pour lesquels notre modèle aura généralisé le mieux ainsi qu'à l'analyse de ces segments.
Dans cette dernière partie, nous visualiserons la gravité des accidents sur une carte.

1. Présentation des données

1.1 Description Open Data et périmètre d'étude

1.1.1 Base de données principales : BAAC

Les données principales proviennent des bases des accidents corporels en France pendant 5 ans entre 2015 et 2019. Pour la construction de modèles dans ce mémoire, nous avons choisi cette période d'observations de 5 ans même si les données sont disponibles jusqu'en 2021. En effet, ce choix permet d'inclure un volume d'observations déjà suffisamment important pour la robustesse des modèles mais aussi et surtout pour être en phase avec les données supplémentaires de l'Insee choisies qui ne sont disponibles que jusqu'en 2019 uniquement.
L'information à disposition sur le site data.gouv.fr se décompose en quatre parties distinctes. (cf. Annexe sur le descriptif détaillé)
Ces bases sont structurées différemment : celles prénommées « Caractéristiques » et « Lieux » comportent une ligne pour un accident (impliquant 2 ou plusieurs véhicules), celles prénommées « Usagers » et « Véhicules » comportent une ligne par usager et/ou véhicule impliqué.

En annexe (cf. Annexe sur le recensement des variables et leurs traitements), nous retrouvons un récapitulatif de toutes les variables utilisées et/ou retraitées dans chaque rubrique de table : ces variables sont pour certaines conservées sous leur forme initiale mais pour la plupart ont été retraitées reformatées et/ou recodées.

C'est dans la table « Usagers » ci-dessus que se trouve la variable à expliquer « grav » qui, comme indiquée, a été retraitée et remplacée par une variable binaire 0 ou 1.

Variables explicatives et regroupement

Certaines variables que nous utilisons dans ce mémoire disposaient de beaucoup trop de modalités. C'est pourquoi, des regroupements ont été opérés.

Les regroupements et transformations réalisés sont :

- La Classe age est traduit en 12 modalités : un regroupement par tranches de 10 ans a été fait permettant de créer ces classes

- Gpe_dep : un découpage par la méthode des quantiles (ici déciles) a été opéré en s'appuyant sur le taux de fréquence des accidents graves par département
(Cf. annexe pour le regroupement des départements)

- Quant aux autres variables créées, elles ont été extraites de variables déjà existantes et/ou nouvellement créée.

1.1.2 Enrichissement avec des données INSEE supplémentaires

Cette deuxième source de données provient de l'Insee. Il est apparu judicieux de chercher à ajouter des données supplémentaires à l'étude sur les activités des habitants leurs catégories-socio-professionnelles les densités de population etc… donnant des informations intéressantes au regard des départements de survenance des accidents.

En effet, la variable "département" contient beaucoup d'informations cachées potentiellement très intéressantes comme l'état des routes, la catégorie socio-professionnelle des habitants, la densité de population etc. Avec ces nouvelles données, nous avons ainsi rajouté des variables à la maille « département » plus facilement interprétables qui sont les suivantes :

Activités des résidents

Nom de variable	Définition de la variable
COM	commune
P15_ACT1564	actifs 15-64 ans en 2015
P15_CHOM1564	chômeurs 15-64 ans en 2015
P15_INACT1564	inactifs 15-64 ans en 2015
P15_RETR1564	retraité ou pré-retraités 15-64 ans en 2015
C15_ACTOCC15P_PAS	Actifs occupés >15 ans n'utilisant pas de transport pour aller au travail en 2015
C15_ACTOCC15P_MAR	Actifs occupés >15 ans marchant à pied pour aller au travail en 2015
C15_ACTOCC15P_DROU	Actifs occupés >15 ans utilisant un deux-roues motorisé pour aller au travail en 2015
C15_ACTOCC15P_VOIT	Actifs occupés >15 ans utilisant la voiture pour aller travailler en 2015
C15_ACTOCC15P_TCOM	Actifs occupés >15 ans utilisant les transports en commun pour aller travailler en 2015

Revenus

Nom de variables	Définition de la variable
CODGEO	département
LIBGEO	nom du département
TP0615	taux de pauvreté en 2015
MED15	médiane du niveau de vie en 2015

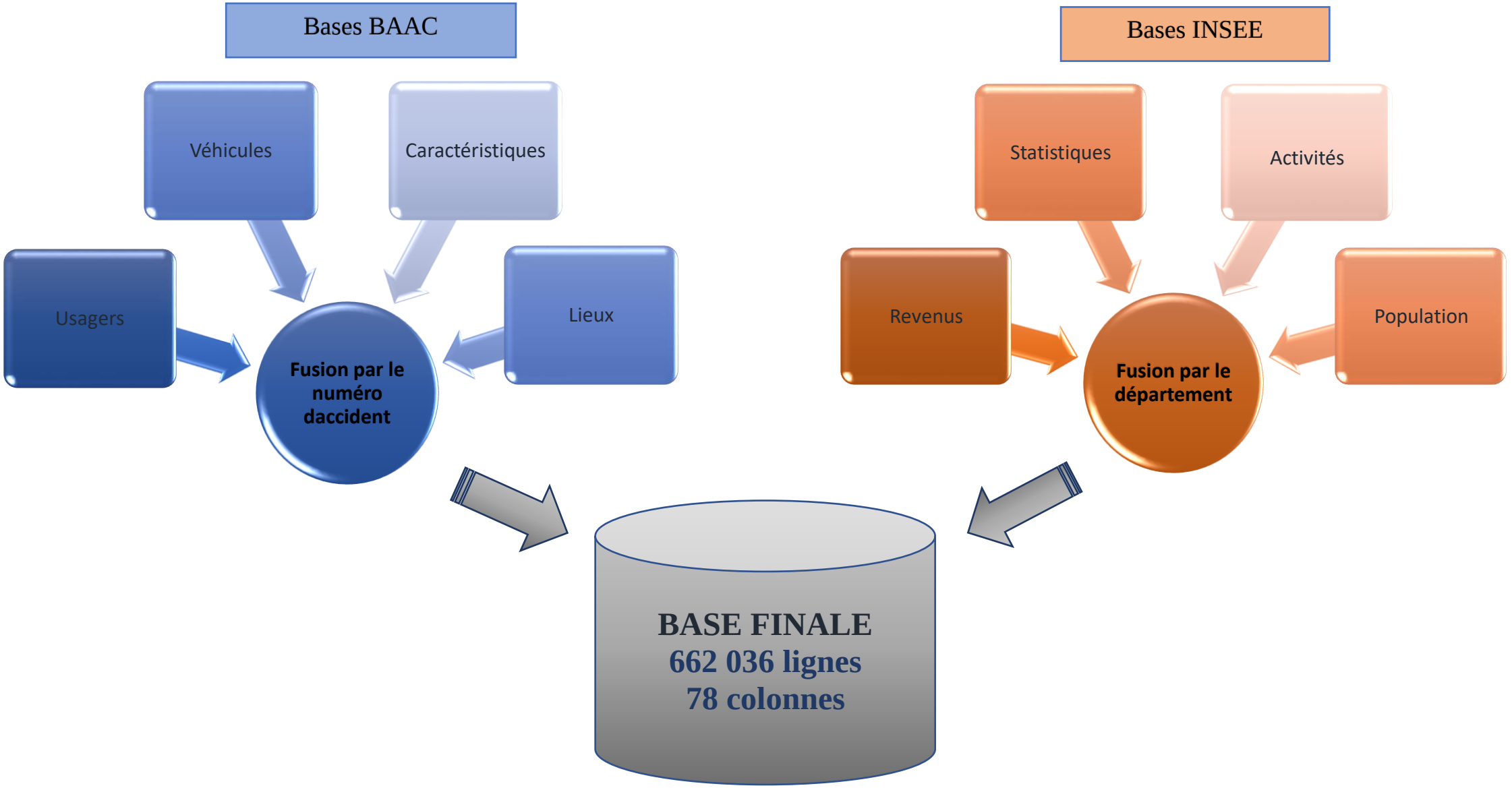
Statistiques

Nom de variable	Définition de la variable
DEP	département
SUPERF	superficie

Revenus

Nom de variables	Définition de la variable
COM	commune
P15_POP	population en 2015
P15_POPH	population masculine en 2015
P15_POPF	population féminine en 2015
C15_POP15_CS1	exploitants en 2015
C15_POP15_CS2	population >15 ans dans la CSP Artisans, Commerçants, Chefs d'entreprises en 2015
C15_POP15_CS3	population >15 ans dans la CSP Cadres, Professions intellectuelles supérieures en 2015
C15_POP15_CS4	population >15 ans dans la CSP Prof. intermédiaires en 2015
C15_POP15_CS5	population >15 ans dans la CSP Employés en 2015
C15_POP15_CS6	population >15 ans dans la CSP Ouvriers en 2015
C15_POP15_CS7	population >15 ans dans la CSP Retraités en 2015
C15_POP15_CS8	population >15 ans dans la CSP Autres en 2015
P15_POP_FR	population de nationalité française
P15_POP_ETR	population de nationalité étrangère
P15_POP_IMM	population immigrée

1.2 Architecture de la construction de la base d'étude



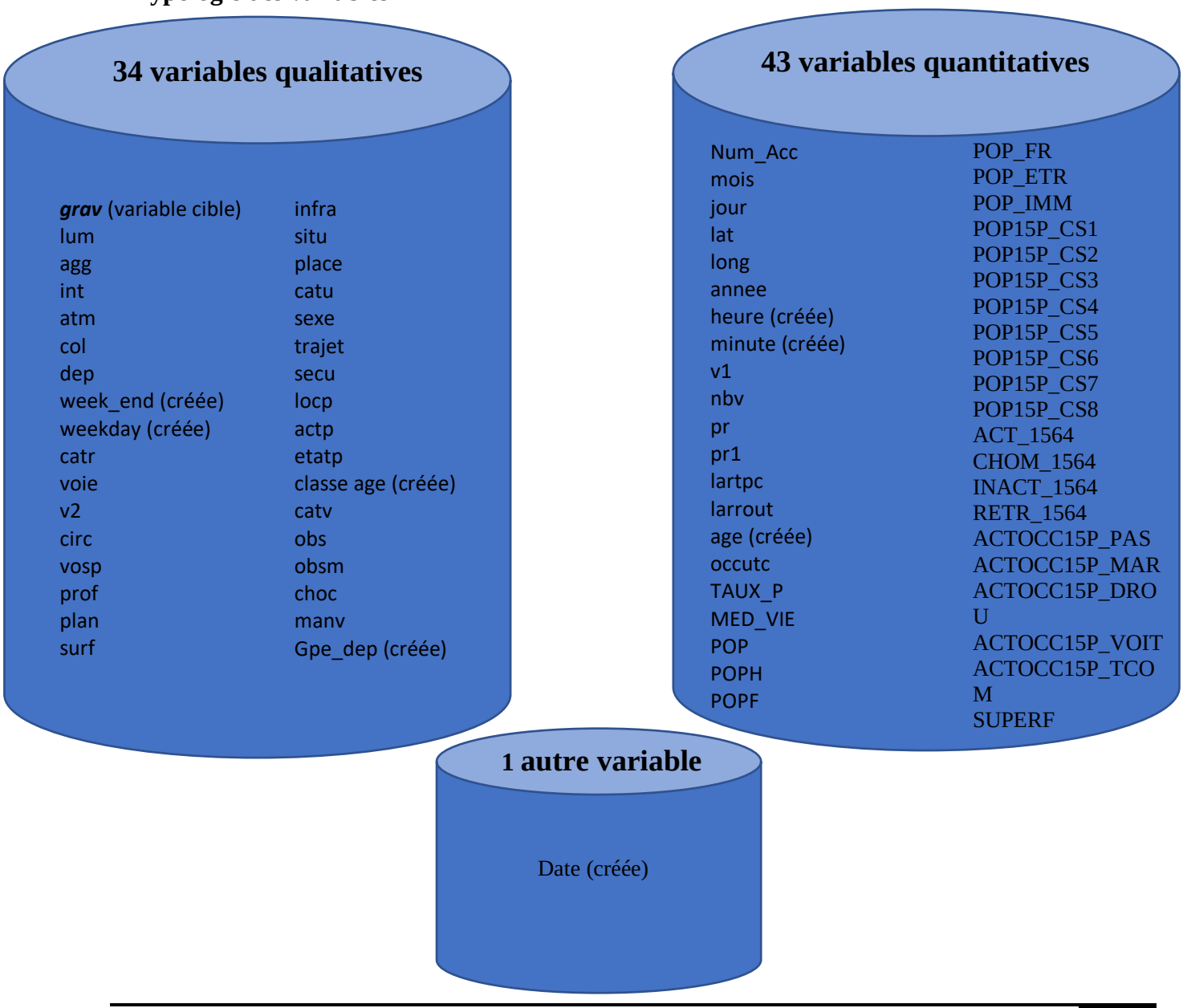
1.3 Analyse Exploratoire des Données

Avant toute modélisation, il est nécessaire d'évaluer la qualité des données et de réaliser des traitements adéquats sur les données. En effet, des données incomplètes ou erronées peuvent engendrer des biais dans les modèles et détériorer la qualité de la prédiction. Les analyses réalisées ici ont consisté à :

- Analyser la forme et le fond de notre dataset
- Retraiter les données manquantes
- Analyser la répartition du portefeuille par variable
- Identifier et tester les relations entre les variables

1.3.1 Analyse de la forme

Typologie des variables



1.3.1.1 Analyse des données manquantes

Sur le graphe présenté en annexe, les valeurs manquantes représentées sont celles indiquant les NaN. Or, comme énoncé dans le descriptif des données, il y a parmi les données manquantes celles qui sont également codifiées en "0" "-1" "Non renseigné" et celles réaffectées dans les variables OBS_NI, OBSM_NI, CHOC_NI, MANV_NI (lors de l'étape de formatage plus haut).
Le heatmap représenté ici ne prend en compte que les valeurs NaN restées inchangées dans notre dataset.
Plus loin, les valeurs manquantes et toutes les autres recodifiées en manquantes seront analysées dans sa globalité.
Le graphe indique la présence de valeurs manquantes pour plusieurs variables : lat, long, voie, v1, v2, pr, pr1, prof, plan, lartpc, larrout, circ, nbv, vosp, surf, infra, situ, classe age et de TAUX_P jusqu'à DENSITE.
Ce qui est tout de suite remarquable, c'est qu'il apparaît souvent comme des lignes au niveau des valeurs manquantes : ce sont en fait les mêmes observations qui sont concernées par plusieurs variables prouvant potentiellement un lien entre ces variables.

Après observations plus fines, certaines variables ont manifestement beaucoup de valeurs manquantes en NaN : v1(83%), v2(95%), pr (43%), pr1(43%), lartpc(37%), larrout(36%) avec au moins 36% de valeurs manquantes.
Un premier tri de nos variables a donc été fait ici en supprimant ces colonnes. De même, il a été décidé de supprimer également la variable voie qui contient une information texte difficilement exploitable.
Sur les variables quantitatives TAUX_P à DENSITE, il reste également environ 37% de données manquantes qui sont potentiellement liées entre elles car formant des lignes blanches.

1.3.2 Analyse du fond

1.3.2.1 Traitement des données manquantes marquées comme telles

Aussi, attardons-nous sur la plupart des variables qui comme évoqué au descriptif peuvent contenir des cellules vides des 0 des points et aussi des -1 mais sont au même titre que les valeurs manquantes marquées en NaN. Cela mérite une analyse approfondie.

Examinons les groupes de variables avec modalités à -1 "Non renseigné" ou vides ou "." ou 0 "sans objet" dont certaines ont été reformatées en remplissant aussi les NaN en -1 ou 0 précédemment prétraitées lors de la phase préliminaire.

Après avoir examiné ce groupe de variables et avoir rassemblé les modalités qui correspondent à aucune information, nous avons finalement beaucoup de données manquantes encore dont le pourcentage est très important (>80%). C'est pourquoi elles ont été supprimées.

La dataset finale après suppression de ces variables comporte alors 66 colonnes.

1.3.3 Analyse statistique descriptive

1.3.3.1 Analyse univariée et bivariée

Variable grav

Nos classes sont déséquilibrées. Pour rappel, la variable initiale était décrite comme suit :

```
1 -> indemne
2 -> tué
3 -> blessé hospitalisé
4 -> blessé léger
```

Cette variable grav a été transformée en ayant regroupé 1 "Indemne" et 4 "blessé léger" en 0 et 2 "tué" et 3 "blessé hospitalisé" en 1.

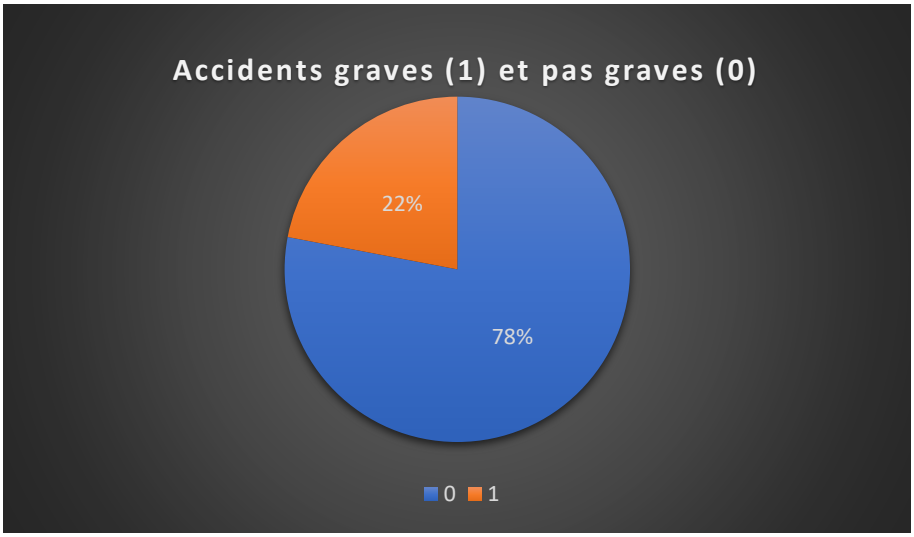


Figure 1 : Répartition des accidents graves et pas graves

Exploration basée sur le moment où a lieu l'accident

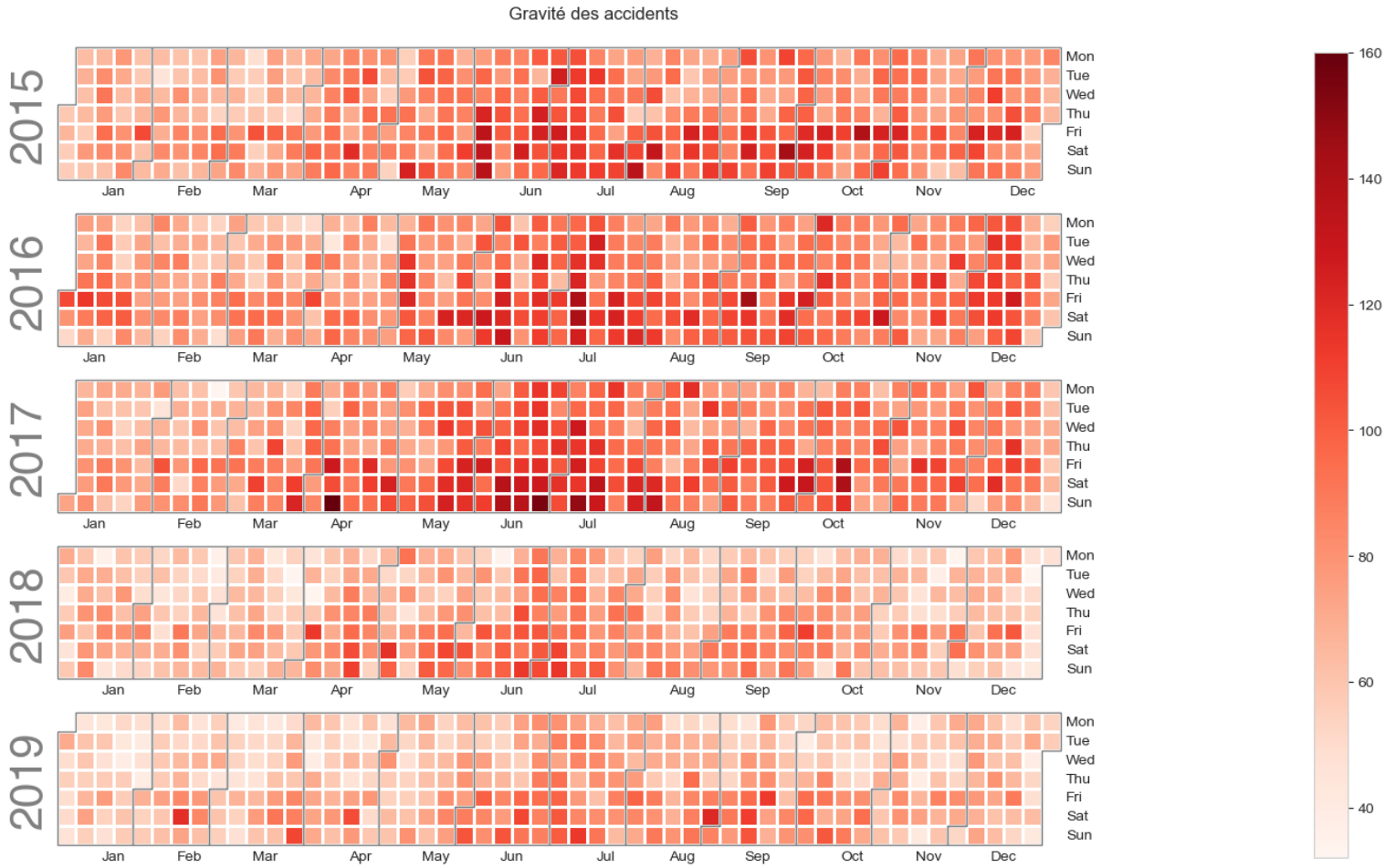


Figure 2 : Calendrier des accidents par gravité

Observations sur le calendrier des jours d'accidents

De manière globale, chaque année est caractérisée par ce que nous avons déjà constaté plus haut : les jours les plus à risques sont les fins de semaine (vendredis, samedis et dimanches), les périodes de grandes vacances scolaires ((juin-juillet) ou de veille de vacances et parfois de lendemain de jours fériés ou jour de fête.
Plus précisément, en comparant les années les unes par rapport aux autres, l'année 2017 est quand même nettement plus marquée par ces accidents graves surtout les mois de juin, juillet. Les jours particulièrement dangereux sont les fins de semaine : vendredi à dimanche.

En 2016, les jours avec beaucoup d'accidents graves sont :
- vendredis 9 et 30 septembre
- week-ends au mois de juin-juillet

En 2017, les jours où se sont produits le plus d'accidents graves sont :
- dimanche 9 avril week-end des vacances scolaires (>130 accidents)
- vendredi 13 octobre (>120 accidents)
- les week-ends au mois de juin-juillet
- le lendemain de l'Ascension le 26 mai

En 2018, les jours où se sont produits beaucoup d'accidents graves sont :
- vendredi 6 avril début de vacances scolaires
- vendredi 5 octobre
- week-ends de mai juin juillet

En 2019, ce sont les mêmes tendances.

- Les accidents ont-ils augmentés en fréquence par année ? (De 2015 à 2019)

2015
- Nombre d'accidents graves : 31 294
- Nombre d'accidents pas graves : 98 938

2016
- Nombre d'accidents graves : 31 994
- Nombre d'accidents pas graves : 101 139

2017
- Nombre d'accidents graves : 32 566
- Nombre d'accidents pas graves : 103 260

2018
- Nombre d'accidents graves : 25 549
- Nombre d'accidents pas graves : 104 362

2019
- Nombre d'accidents graves : 24 368
- Nombre d'accidents pas graves : 108 566

- Quels sont les mois où les fréquences d'accidents sont les plus élevées ?
- Quel jour dans le mois est le plus sécurisé pour conduire ?

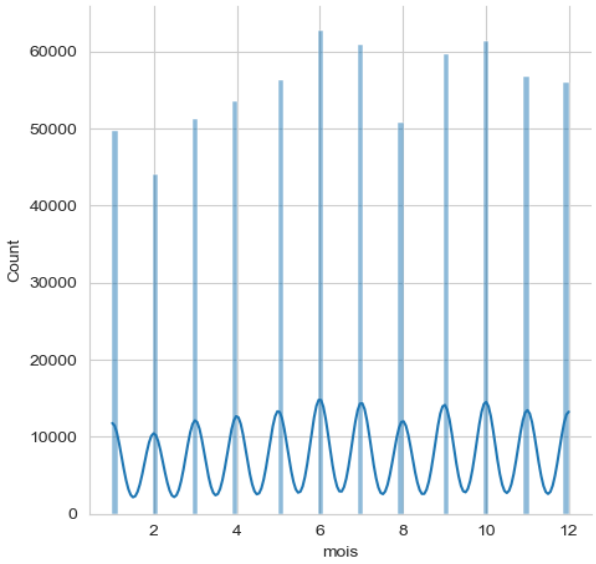


Figure 3 : Répartition des accidents par mois

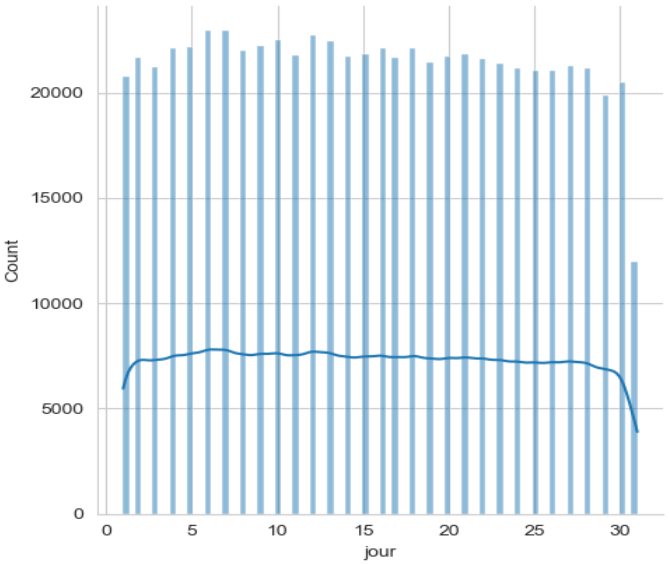


Figure 4 : Répartition des accidents par jour du mois

Sur les mois de l'année, il apparaît que ce sont les mois de juin, juillet, septembre et octobre qui enregistrent les pics d'accidents. A contrario, les mois de février et avril sont les mois avec le moins d'accidents en France sur ces années.
Concernant les jours, ce sont généralement les 6, 7, 12, 13 et 16 du mois où les accidents ont lieu.

Quant au moment de la journée, le créneau horaire enregistrant le plus d'accidents est 17h-18h c-à-d aux heures de pointe des usagers de la route (à la fin de la journée). La distribution est asymétrique.

Le graphe illustrant la distribution des minutes nous indique que les accidents ont souvent lieu en heure pleine (00 min) ou à la demi-heure (30 minutes). A 0 ou 30 minutes, nous avons des pics, puis une forte baisse, puis une hausse entre 10 et 15 minutes et à nouveau une baisse aux 20-25 minutes de chaque heure.

Exploration basée sur le type de route de survenance de l'accident

- Quel type de routes sont les plus risquées ?

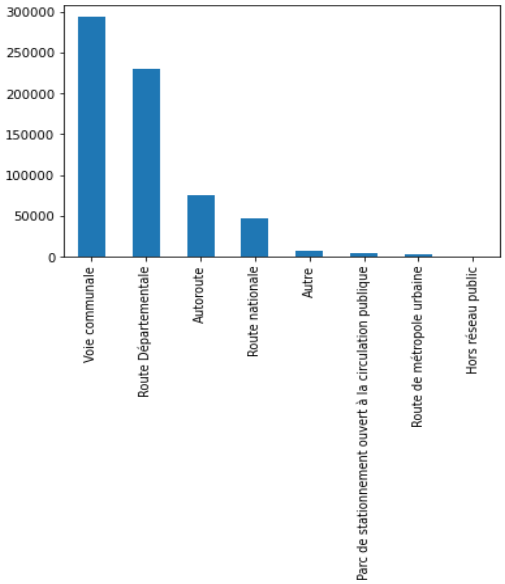


Figure 5 : Fréquence d'accidents par type de route

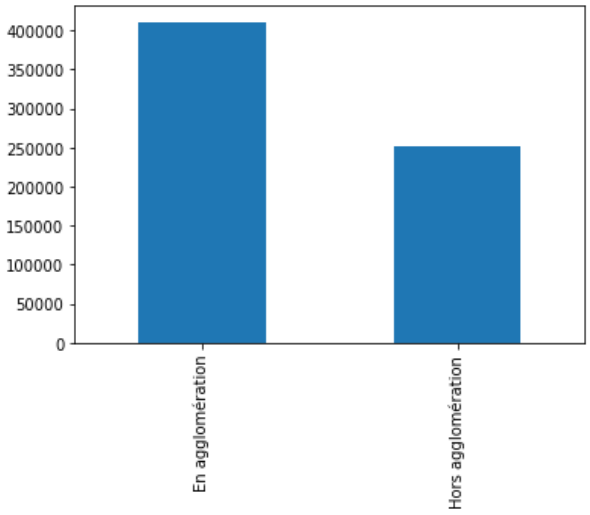


Figure 6 : Fréquence des accidents par type d'agglomération

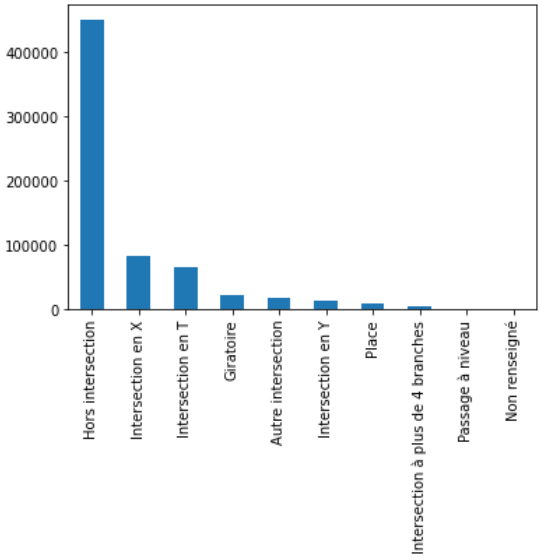


Figure 7 : Fréquence d'accidents par type d'intersection

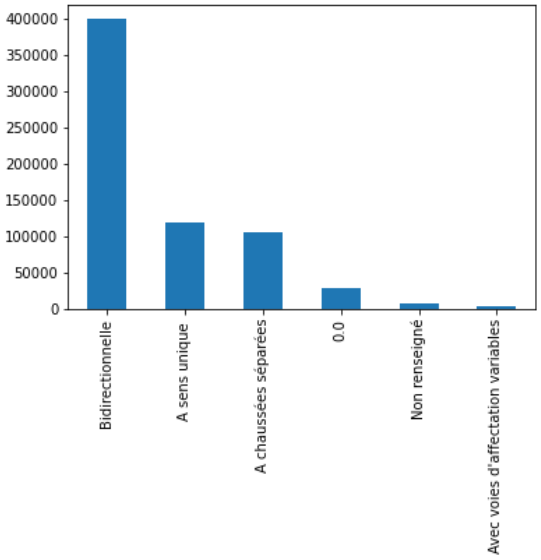


Figure 8 : Fréquence des accidents par type de régime de circulation

La plupart des accidents répertoriés ont eu lieu dans ces circonstances : en agglomération, sur des voies communales ou routes départementales, sur un régime de circulation bidirectionnelle, en profil plat, sur une partie rectiligne, en situation de surface de route normale (sans pluie), sur une chaussée, sur des trajets promenade-loisirs et essentiellement dans le département 75 (paris).

A noter également que les routes ayant 2 voies semblent être les plus accidentogènes.

Exploration basée sur les conditions des usagers victimes après l'accident

- Quelle est la tranche d'âge des victimes impliquées dans les accidents ?

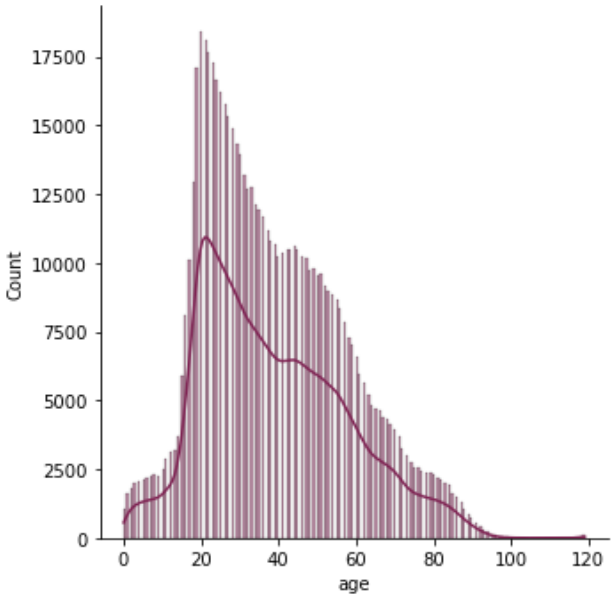


Figure 9 : Fréquence des âges des victimes

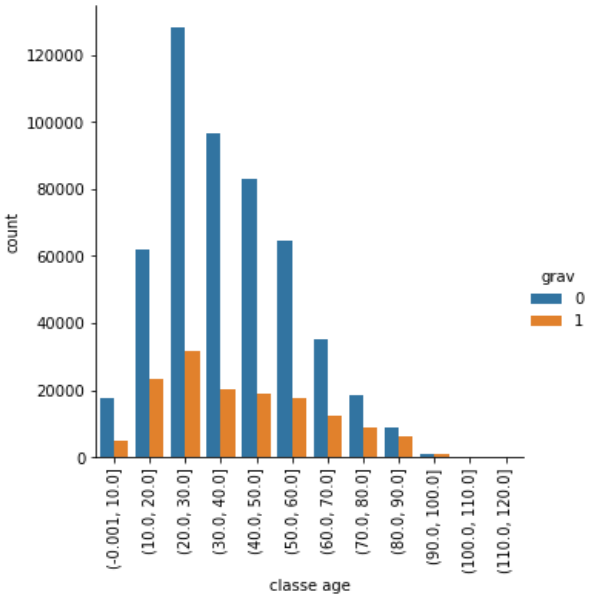
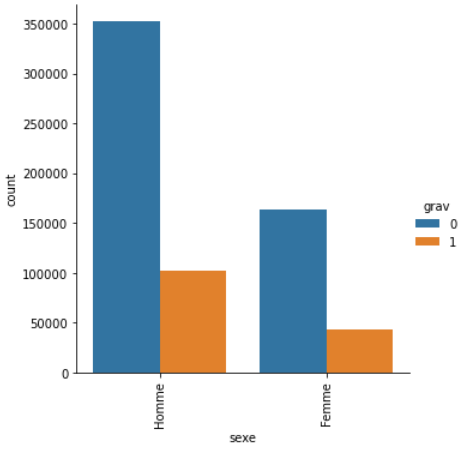


Figure 10 : Fréquence des classes d'âges par type de gravité

Après avoir remplacé les 3 lignes ayant pour année de naissance 0 (probablement des erreurs de saisies) le graphe illustrant l'âge présente une asymétrie à droite. Les individus impliqués dans le plus d'accidents sont les jeunes âgés de 20-30 ans environ. Ceux ayant le moins d'accident sont plus âgés (>60 ans).

- Quel est le genre des usagers les plus souvent victimes d'accidents ?



Les personnes impliquées dans les accidents les plus fréquents sont ceux situés sur la place 1 dans la grande majorité donc les conducteurs et de sexe masculin.

Concernant l'âge, la classe la plus touchée est celle des 20-30 ans dans tous les cas, accidents graves ou pas graves. Cependant, on peut remarquer que dès 45-50 ans, les courbes s'inversent et font apparaître plus d'accidents graves que de moins graves. A partir de cette tranche d'âge, il y a un gros facteur de risque d'apparition d'accidents corporels graves.

Figure 11 : Fréquence d'accidents par sexe

Exploration sur l'utilisation de l'équipement de sécurité

- Quels sont les équipements de sécurité utilisés ?

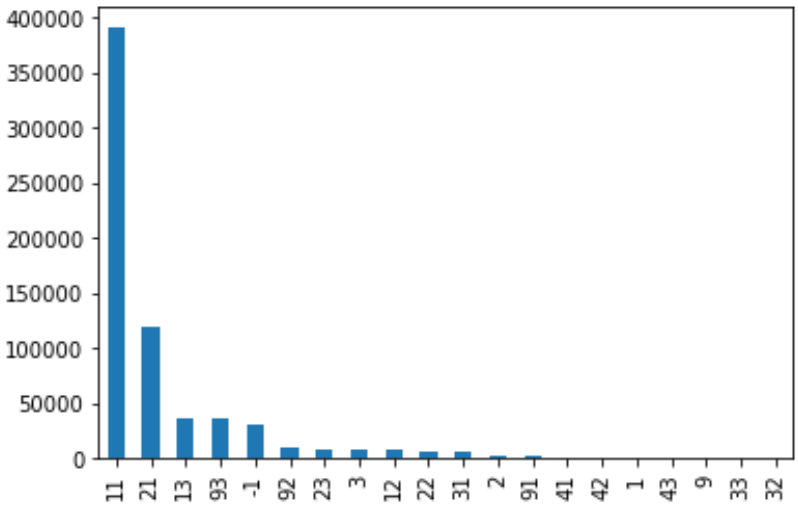


Figure 12 : Fréquence d'accidents par équipement de sécurité

- Est-ce-que le port de l'équipement de sécurité a eu un impact sur la gravité de l'accident ?

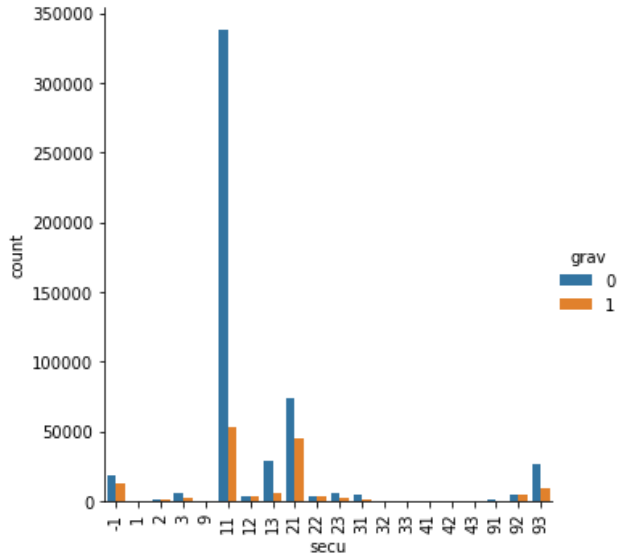


Figure 13 : Fréquence d'accidents par type d'équipement de sécurité et gravité de l'accident

Autres observations sur les variables qualitatives

- Les caractéristiques environnementales météorologiques tempore
lles :

Les accidents ont lieu plus fréquemment en plein jour en conditions météorologiques normales hors week-end et plus précisément les journées du vendredi.

Autres observations des variables quantitatives

- Caractéristiques des individus :

Autour des véhicules :

La plupart des variables de comptage sur les véhicules, les obstacles, les chocs et les manœuvres présentent une densité avec une forte concentration sur quelques valeurs (faibles) seulement : ceci signifie qu'il y a le plus souvent peu de véhicules impliqués lors d'un accident mettant en cause peu d'obstacles fixes ou mobiles en faisant peu de manœuvres et causant peu de chocs. C'est ce contexte d'accident qui est le plus reproduit.

- Caractéristiques démographiques et sociales :

L'immense majorité des accidents ont lieu dans des départements ayant un taux de pauvreté dans la fourchette [10% ;20%]. Quant au niveau de vie, ce sont les départements de revenus médians [20 000;23 000] qui sont les plus accidentogènes.
Il est à noter que là où l'on enregistre le plus d'accidents (pic) nous nous trouvons en lieu où la population avoisine les 22M, une population à faible représentativité en terme d'étrangers et d'immigrés.
Les graphes sur la catégorie socio-professionnelle apporte également de l'information : les

endroits peu peuplés en CS1 (agriculteurs) et CS3 (cadres) sont très accidentogènes. Aussi les courbes de représentations pour les CS4 CS5 et CS8 connaissent à peu près une tendance similaire avec un pic aux points aux environs de 200 000-300 000.

Rappelons les CSP ici :
- CS1 : Agriculteurs exploitants
- CS2 : Artisans Commerçants Chefs d'entreprise
- CS3 : Cadres et Professions intellectuelles supérieures
- CS4 : Professions intermédiaires
- CS5 : Employés
- CS6 : Ouvriers
- CS7 : Retraités
- CS8 : Autres sans activité professionnelle

Pour terminer quant à la densité l'accident est plus fréquent là où celle-ci est la plus faible.

Autres observations des croisements variable cible-variables quantitatives

- Caractéristiques démographiques et sociales :

La gravité de l'accident augmente avec le nombre d'occupants dans le transport en commun impliqué. De façon identique si plusieurs véhicules et/ou piétons sont impliqués la gravité s'accroît si le nombre de véhicules et/ou piétons augmente (2 roues, voitures...).
Chez les revenus médians faibles à modérés, la gravité des accidents s'avère plus importante. De plus, nous constatons que là où la population est le plus faible, il y a plus d'accidents graves c'est notamment les endroits à faible population étrangère et immigrée.

Plus tôt nous avons vu que les départements touchés par beaucoup d'accidents sont ceux où il y a peu de populations de CS1 ; mais il s'agissait de tous les accidents confondus. Nuançons ce point car ici force est de constater que ce sont plus souvent des accidents moins graves qui surviennent en ces lieux précités et les accidents plus graves se produisent aux endroits à plus forte concentration en population de CS1. Sur les autres courbes, nous remarquons que les accidents les plus graves ont lieu dans les zones où la population CS2 CS3 CS4 CS5 CS6 CS7 CS8 sont faibles.
Aussi, sur les courbes des actifs chômeurs inactifs et retraités cette tendance se dessine également : les accidents les plus graves arrivent aux endroits où il y a le moins de population. Sur les habitudes de déplacement, ce sont aussi aux endroits les moins fréquentés par les usagers qu'arrivent les accidents les plus graves (excepté pour les déplacements en voiture).
Nous allons vérifier plus loin les corrélations avec le V de Cramer.

Autres observations des relations variable cible-variables qualitatives

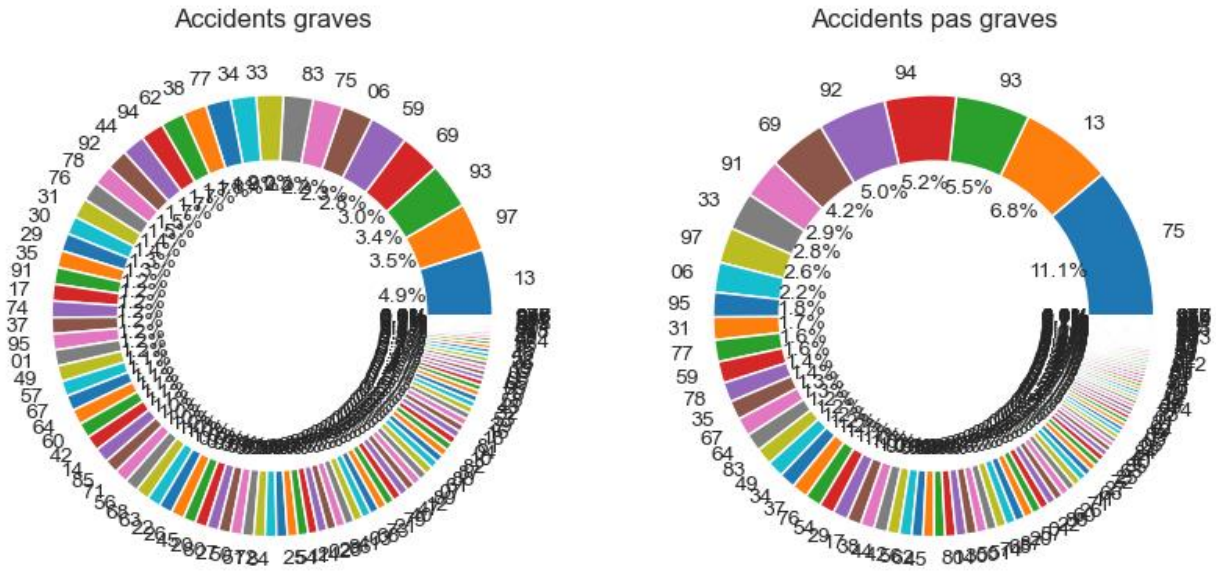


Figure 14 : Répartition des accidents graves et pas graves par département

- Les caractéristiques routières :

Les routes départementales sont les endroits les plus sujets aux accidents graves que moins graves. Aussi les chemins en courbe sont les plus favorables aux accidents graves.

- Les caractéristiques des victimes et véhicules impliqués :

Les hommes sont les victimes les plus représentées des accidents graves.

Mesures de liaisons entre les variables et la variable cible (tests statistiques)

- V de Cramer pour les variables qualitatives et grav

On peut appliquer les tests de chi-2 sous réserve de respecter les conditions de son application. En effet, pour pouvoir être valide, ce test requiert que chaque cellule des tableaux croisés constitués des modalités des variables testées contienne au minimum un effectif de 5. Cela signifie que pour les mesures entre nos variables et grav il faut qu'il y ait au minimum 5 accidents pour chaque modalité de la variable testée.
Les résultats des tests sont fournis en annexe (Cf. Annexe sur les mesures de liaisons entre variables qualitatives et grav).

Pour toutes les variables, les valeurs sont plus proches de 0 que de 1, il y a donc peu de liaisons avec la variable grav. Celles qui ont le coefficient le plus proche de 1 sont : **secu, Gpe_dep, col, obs et catr.**

- Test de Student pour les variables quantitatives avec grav

Dans ce test statistique, nous allons tester l'hypothèse nulle à savoir que les moyennes sur chaque groupe (grav=0 et grav=1) sont égales. (Cf. Annexe Mesures de liaisons entre variables quantitatives et grav).

Les variables mois et jour ont des p-value>5% et aboutissent à un non-rejet de l'hypothèse nulle.
En revanche, pour toutes les autres variables, les p-value <5% signifient que l'hypothèse nulle est rejetée et qu'il y a donc bien un lien entre ces variables et grav. Ces variables seront intéressantes pour la suite.

1.3.3.2 Analyse multivariée

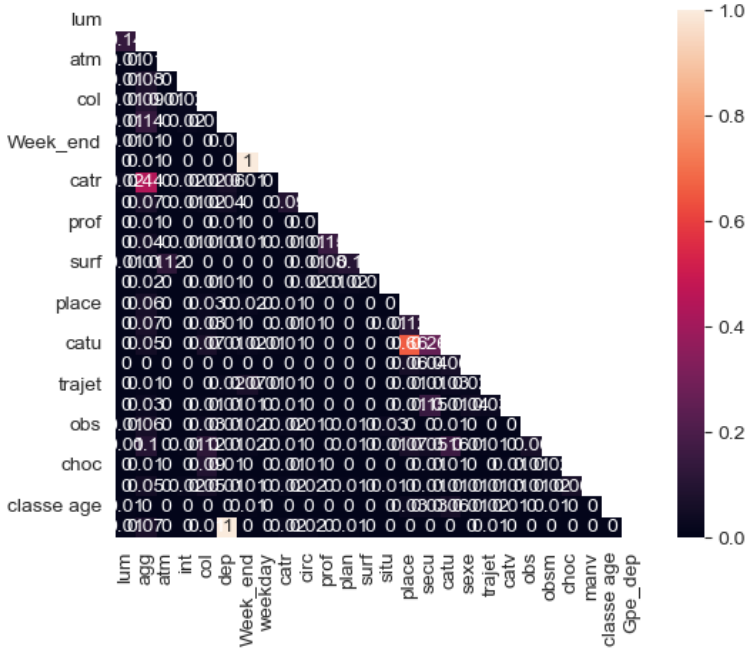
- Quantitative / quantitative

Bases Principales (cf. annexe : matrice de corrélation des variables quantitatives)

Bases INSEE (cf. annexe : matrice de corrélation des variables quantitatives)

Dans le groupe des variables de nos données principales, il y a certaines corrélations très fortes (>08) : occutc-catv_tc (08) obs_ni-obsm_veh (085) et catv_drou-catv_voit (-054) obsm_aucun-obsm_veh (-056) obsm_veh-obsm_pieton (-051).
En revanche, dans le groupe des variables recueillies de l'INSEE, nous observons de très fortes corrélations parmi les variables relatives à la population exceptée POP15P_CS1 (qui d'ailleurs, nous l'avons vu était bien corrélée à grav).
D'autre part, DENSITE et TAUX_P ont également peu de corrélations avec les autres variables.

- Qualitative / qualitative



Dans le groupe de ces variables, il y a très peu de corrélations : seules les variables catr-agg (042) et catu-place (039) apparaissent avec les plus fortes valeurs de corrélations mais elles seront néanmoins conservées.

Figure 15 : V de Cramer sur variables qualitatives

Conclusion de l'Analyse Exploratoire

Les variables corrélées à grav et intéressantes à conserver (non corrélées entre elles) sont dans le tableau ci-dessous :

Variables quantitatives	Variables qualitatives
Heure	Catu
Minute	Sexe
Annee	Trajet
Age	Secu
Nbv	Classe age
Lat	Catr
Long	Circ
TAUX_P	Prof
POP	Plan
POP15P_CS1	Surf
MED_VIE	Situ
SUPERF	Lum
	Agg
	Int
	Atm
	Col
	Week_end
	Weekday
	Gpe_dep
	Senc
	Catv
	Obs
	Obsm
	Choc
	Manv

Identification des outliers

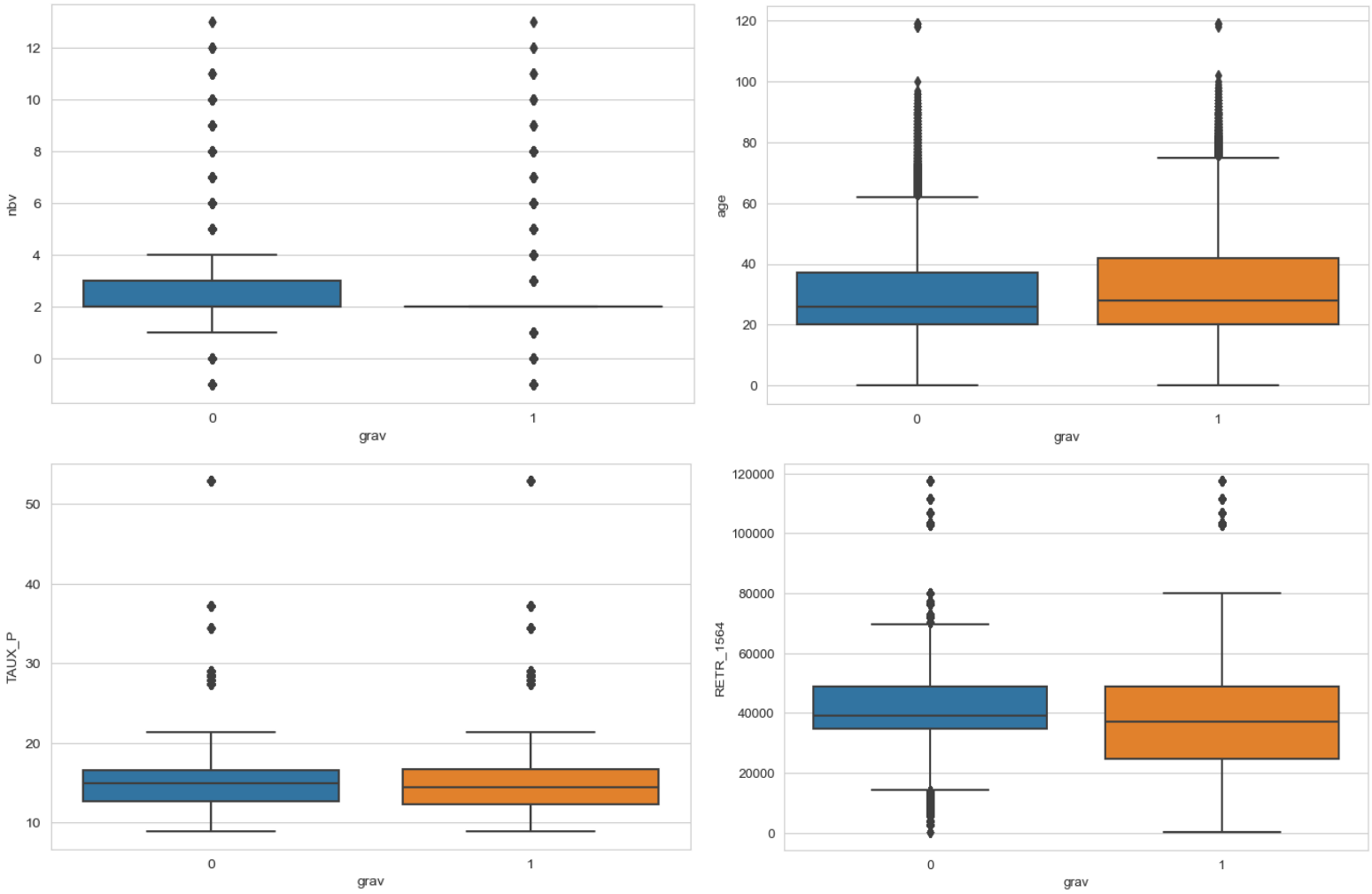


Figure 16 : Boxplot sur variables qualitatives nbv age TAUX_P RETR_1564

Observations sur les boxplot

Pour les variables quantitatives présentées ci-dessus nbv age TAUX_P et RETR_1564, les boxplots révèlent l'existence de valeurs aberrantes. En effet, comme nous l'avions constaté sur l'analyse univariée pour l'âge, par exemple, certains usagers ont presque 120 ans. Par exemple, sur les revenus médians (MED_VIE) les 2 boxplot montrent des différences : la moyenne est sensiblement plus faible dans le groupe grav=1 que dans l'autre ce qui signifie que les accidents ont une gravité plus importante chez les populations dont la moyenne des revenus médians est située en dessous de 21000 alors que celle chez les populations ayant un accident moins grave se situe aux environs de 22000. De plus, pour les accidents graves, il y a des valeurs aberrantes au-delà des moustaches.
Il semble y avoir un lien entre la gravité de l'accident et la population de la catégorie CS1. Dans les autres CSP, la tendance est inversée : c'est la population moyenne chez les accidents moins graves qui est plus élevée.
En ce qui concerne la densité de population, les boxplots de grav et pas grav sont très différentes avec des points aberrants sur le groupe grav=1 et des moyennes à peu près au même niveau.
Pour conforter nos constats, nous allons plus loin réaliser des tests d'hypothèses afin de vérifier les relations évoquées.

2. Modélisation

2.1 Preprocessing préliminaire : standardisation, encodage, ACP

Après les étapes de preprocessing de base (standardisation, encodage), une Analyse en Composantes Principales est effectuée pour pouvoir visualiser les observations sur un plan factoriel.

Les trois composantes principales réunissent 23% de la variance totale, et voici les contributions des variables sur chaque composante.

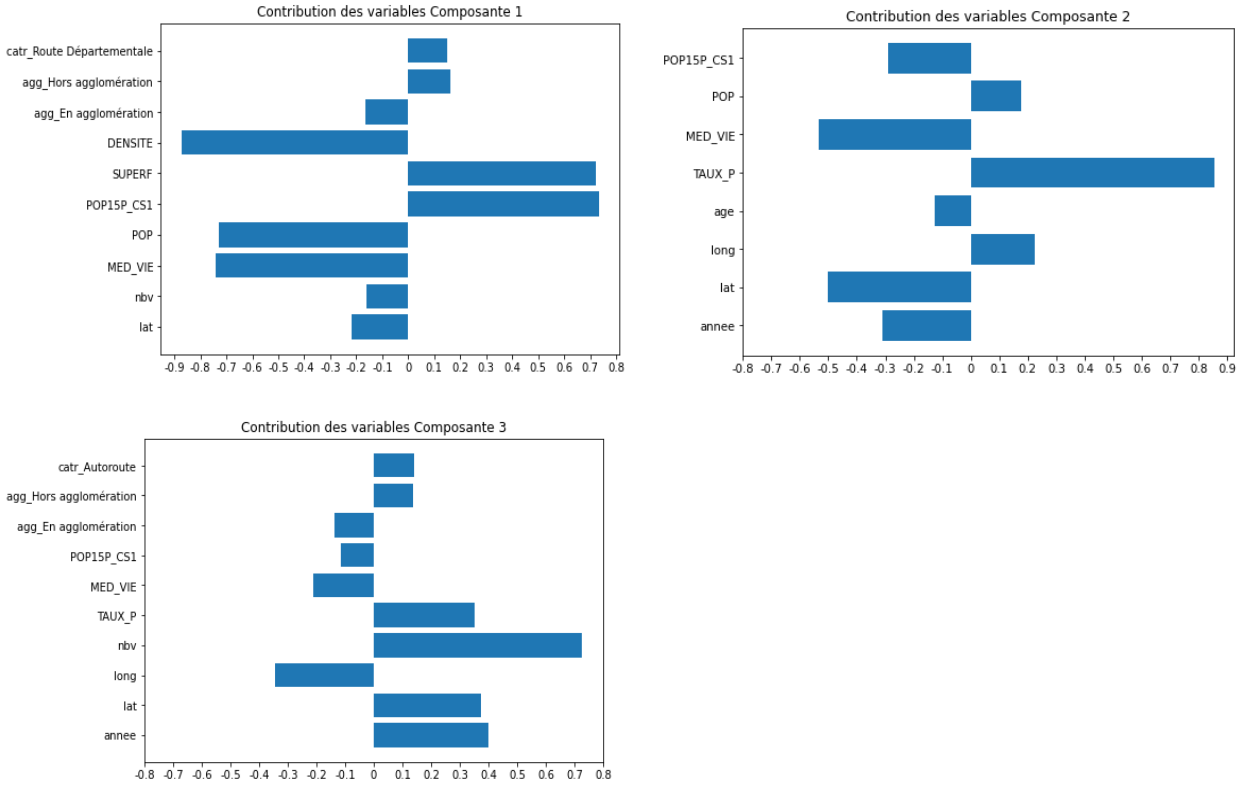


Figure 17 : Contribution des variables sur les composantes

La visualisation des contributions à la formation des axes montre que :

-SUPERF et POP15P_CS1 (et agg_Hors agglomération dans une moindre mesure) contribuent positivement à la construction de l'axe 1 ainsi que POP, MED_VIE et DENSITE mais de façon négative.

-TAUX_P contribue très fortement (positivement) à la formation de l'axe 2 et MED_VIE et lat contribuent également mais de façon négative.

- nbv est celle qui contribue le plus à l'axe 3 (avec annee, lat et TAUX_P mais dans une moindre mesure).

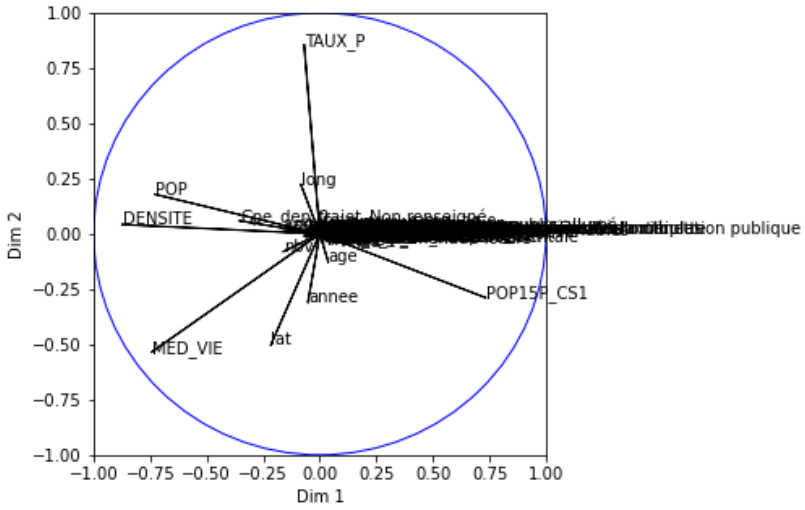


Figure 18 : Cercle des variables sur le plan factoriel

Avec le cercle des variables, nous nous confortons dans les constats précédemment évoqués :

- sur **l'axe 1**, les accidents ayant lieu dans les endroits à superficies élevées où loge la population de la catégorie CS1 (agriculteurs exploitants) s'opposent aux accidents aux endroits à forte densité de population. Avec la première analyse ci-dessus nous pouvons résumer sur cet axe les accidents ayant lieu hors agglomération (larges superficies avec des agriculteurs) par opposition à ceux en agglomération (population dense revenus élevés).

- **l'axe 2** oppose quant à lui les accidents ayant lieu autour des zones dont le taux de pauvreté est très élevé avec les accidents survenant aux zones où le revenu médian de vie est important : cet axe-ci retrace les accidents aux lieux de niveau de vie différents.

- **l'axe 3** se définit par les accidents ayant lieu aux zones où le nombre de voies est élevé.

Après cette fine analyse, nous visualisons les individus sur le nuage (premier plan factoriel retenu)

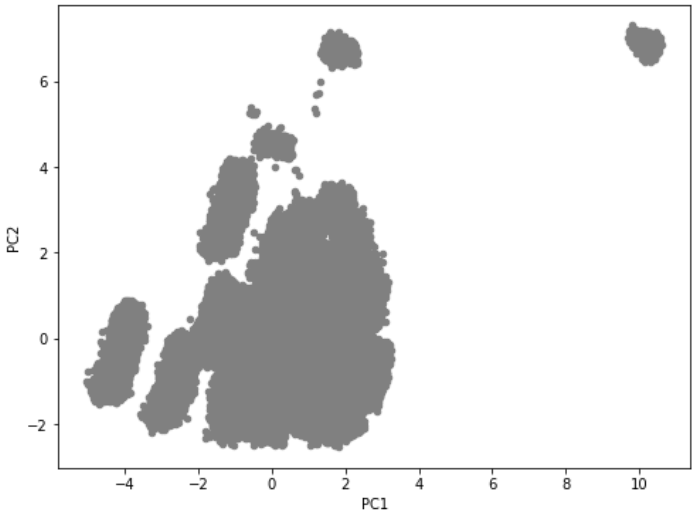


Figure 19 : Nuage des individus dans le plan factoriel (1,2)

Observations :

Le nuage dessine plusieurs clusters, environ 7 dont un gros cluster au milieu du plan. Il y a un premier cluster (le premier depuis la gauche), très proche de l'origine : ce sont donc des accidents survenus dans des endroits où il y a le moins d'agriculteurs-exploitants en zone de superficie peu élevée, et dont le taux de pauvreté est faible.
Le deuxième cluster juste à sa droite connaît à peu près les mêmes caractéristiques avec une représentativité d'agriculteurs-exploitants plus important que le précédent, mais dont le taux de pauvreté est plus faible que le premier cluster.
Sans détailler chaque cluster, nous observons qu'en haut à droite, un cluster se démarque complètement avec une situation des accidents à forte représentativité d'agriculteurs-exploitants, en zone de superficie le plus élevée et où le taux de pauvreté est très élevé.

Toutes ces composantes construites par combinaison linéaire seront conservées dans notre dataset pour construire les modèles.

2.2 Premier modèle de régression GLM par model point

Une première idée de modélisation serait de faire une approche classique GLM sur la fréquence de gravité par model point c'est-à-dire par agrégation de données. C'est une méthodologie courante, bien connue dans le domaine de la tarification en assurance.

En effet, nous allons d'abord construire des Model Points en regroupant les accidents selon des caractéristiques communes : dans l'idée, nous pourrions faire un vrai regroupement utilisant des techniques comme le clustering K-means pour identifier des cohortes d'accidents, mais, ici pour faire plus simple, nous allons regrouper simplement les variables selon les variables jugées les plus corrélées à la variable cible (vu dans l'analyse exploratoire).

A l'issue de cette étape, sera modélisée la fréquence de gravité des accidents de la nouvelle dataset par model point.

2.2.1 Premiers résultats

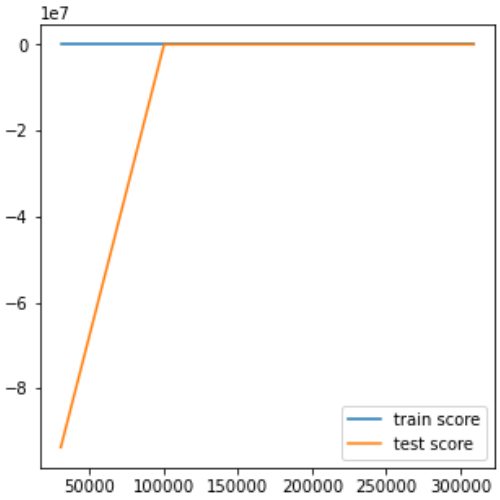


Figure 20 : Learning curve modèle de régression

Observations :

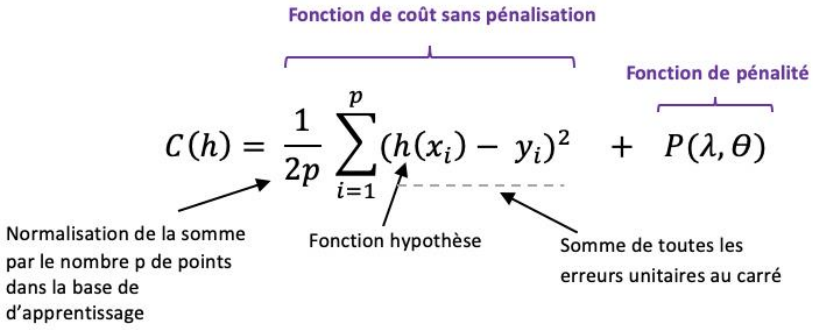
D'après le graphe, la courbe du train score et celle du test score se rejoignent et se superposent complètement, il n'y a pas d'overfitting. Nous pouvons en déduire que le modèle a parfaitement appris sur les données d'entraînement et généralise bien sur des nouvelles données.

2.3 Régression pénalisée ELASTICNET

Cadre théorique de la régularisation Elasticnet

Dans le contexte de la statistique et en particulier de l'apprentissage automatique, la régularisation est le nom donné à un processus visant principalement à réduire des problèmes liés au surapprentissage. Et ce au travers d'une réduction de la variance ou d'une sélection du nombre de paramètres employés.
Ce processus n'est pas neutre puisqu'il consiste à introduire une nouvelle information au problème dans l'objectif de le "simplifier" que ce soit en introduisant une pénalité d'autant plus grande que la complexité du modèle l'est ou en "imposant" une distribution a priori des paramètres du modèle.
Le principe général de la régularisation consiste à pénaliser les valeurs extrêmes des paramètres (ce qui conduit souvent à la variance du surapprentissage). Nous cherchons alors à minimiser à la fois le modèle par rapport à notre métrique de choix ainsi qu'une métrique sur la taille et le nombre de paramètres employés.
Pour ce faire, il s'agit de modifier un peu la fonction de coût du problème de régression linéaire en la complétant par un terme de pénalité.

La fonction de coût avec pénalisation s'écrit dans un cadre général de la façon suivante :



C'est la fonction P (λ θ) qui va gérer la pénalité selon un paramètre lambda λ que l'on fixe empiriquement de façon à obtenir les meilleurs résultats.

Dans la régression Elasticnet la fonction de coût devient :

$$C(h) = \frac{1}{2p}\sum_{i=1}^{p}(h(x_i) - y_i)^2 + \lambda \sum_{j=1}^{n}[\frac{1}{2}(1 - \alpha)\theta_j^2 + \alpha\,|\Theta_j|]$$

Où le paramètre α est un paramètre définissant l'équilibre entre ridge et lasso.

Et il est possible d'ajuster la pénalisation en fonction du cas d'application.

Mise en pratique de la régularisation Elasticnet sans hyperparamétrage

A l'issue de la régularisation, 117 variables sont retenues.

2.4 Modèles de machine learning sur modèle réduit par Elasticnet

Après avoir effectué ces différentes étapes préliminaires de traitement des variables et pensé à un premier modèle sur la fréquence de gravité par model point, nous pouvons à présent réaliser les modèles de prédiction des accidents graves à partir de l'ensemble des accidents corporels survenus entre 2015 et 2019 sur la dataset réduite par elasticnet.

Dans un premier temps, une méthode classique sera mise en œuvre, une régression logistique qui servira de référence. Ensuite, les méthodes de machine learning seront mises en avant.

Ce mémoire sera donc consacré à quatre méthodes fréquemment utilisées parmi les méthodes d'apprentissage adaptées à l'assurance :

- Régression logistique
- Arbre CART
- Random Forest
- Extreme Gradient Boosting

Tous sont d'abord entraînés avec des paramètres par défaut en première intention. (cf. Annexe sur les learning curve et matrices de confusion).
Sur nos 4 modèles de base, nous voyons que sur le premier souffre d'overfitting (régression logistique). Sur les modèles des arbres et le random forest, les courbes ont du mal à se rapprocher, montrant une difficulté au modèle de bien généraliser. Enfin, le 4ème modèle XGBoost semble être le plus performant, les courbes se rejoignent nettement mieux.

L'étape d'hyperparamétrage de chaque modèle sera ensuite effectuée.

Métriques de performance

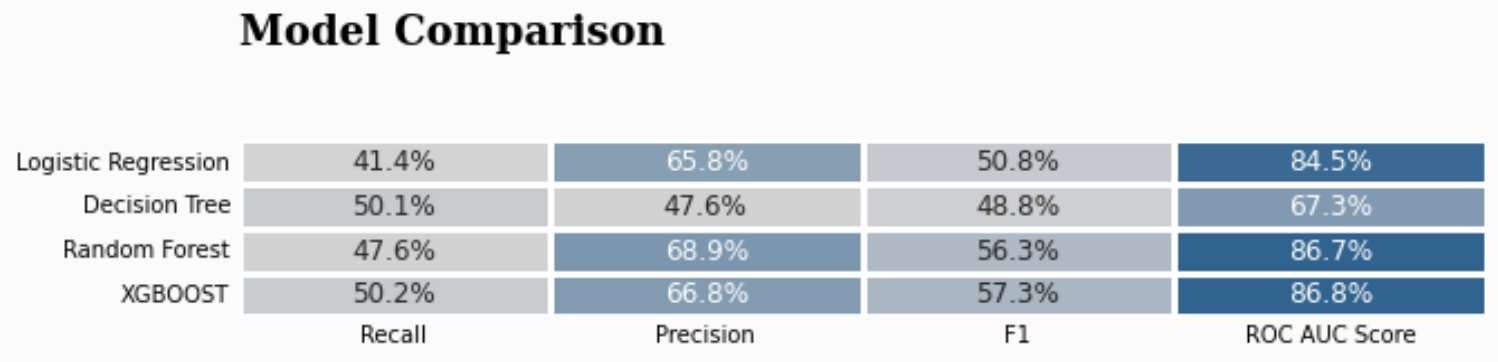
Les métriques de performance doivent être définies en accord avec notre objectif de modélisation.

Qu'est-ce-que nous cherchons à prédire ?

- ⇨ **La gravité de l'accident pour l'usager concerné en minimisant l'erreur globale**

Puisque notre variable à prédire est binaire, notre priorité est définie telle qu'on puisse prédire le mieux possible les accidents graves.
Dans ce sens, nous savons que parmi nos métriques, nous devons donc nous baser sur le **recall** (taux des vrais positifs parmi les réels positifs), et aussi sur la **precision** (taux des positifs parmi les prédictions positives), le f1-score (moyenne harmonique de ces indicateurs).
Dans la pratique, et comme il est davantage important de prédire les cas graves, un bon taux sera privilégié au taux de precision.

2.4.1 Premiers résultats



Sans hyperparamétrage, le modèle qui semble être le plus performant à ce stade est le XGBoost selon les métriques considérées.

2.4.2 Optimisation des hyperparamètres

Les étapes d'hyperparamétrage suivantes ont été réalisées avec une GridSearchCV.

- Régression logistique

Les principaux paramètres retenus sont ceux ci-dessous :
- penalty : l1
- C : 10
- fit_intercept : False

- Arbre CART

Les principaux paramètres retenus sont ceux ci-dessous :
- max_depth : 20
- min_samples_leaf : 10
- min_samples_split : 2

Les autres valeurs ont été laissées à leur valeur par défaut

- Random Forest

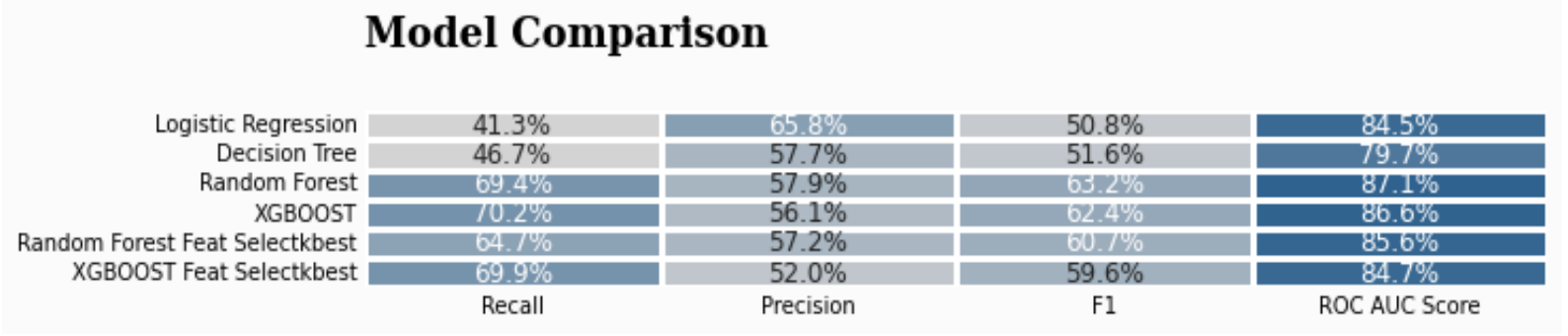
Les principaux paramètres retenus sont ceux ci- dessous :
- n_estimators : 600
- criterion : entropy
- min_samples_split : 12
- min_samples_leaf : 4
- max_features : auto
- class_weight : balanced_subsample

Les autres valeurs ont été laissées à leur valeur par défaut

- XGBoost

Les principaux paramètres retenus sont ceux ci- dessous :
-n_estimators : 100
- learning_rate : 0.3
- max_depth : 10
- gamma: 1
- subsample: 1
- colsubsample_bytree : 0.5
- scale_pos_weight : 2.5
Les autres valeurs ont été laissées à leur valeur par défaut

Les résultats sont les suivants :



Le XGBoost est le modèle qui donne le meilleur **recall** (métrique que nous cherchions à maximiser), même si le Random Forest fonctionne le mieux en terme de f1-score.

2.4.3 Feature Selection sur les meilleurs modèles Random Forest et XGBOOST

En amont de la modélisation, nous avons déjà réalisé une première sélection de variables avec Elasticnet, dans le but de réduire au mieux nos paramètres et obtenir un modèle plus parcimonieux.

Ici, nous allons considérer les techniques supplémentaires de sélection de variables qui seront mises en œuvre sur les modèles retenus Random Forest et XGBOOST.

Les trois approches pour la sélection des prédicteurs dans une démarche de machine learning sont : filter, embedded & wrapper.

L'approche Filter utilise des outils statistiques ou liés à a théorie de l'information pour isoler les variables les plus significatives parmi un ensemble, et cela indépendamment de tout algorithme d'apprentissage. C'est le cas par exemple de SelectKBest basé sur le test ANOVA ou le chi2.

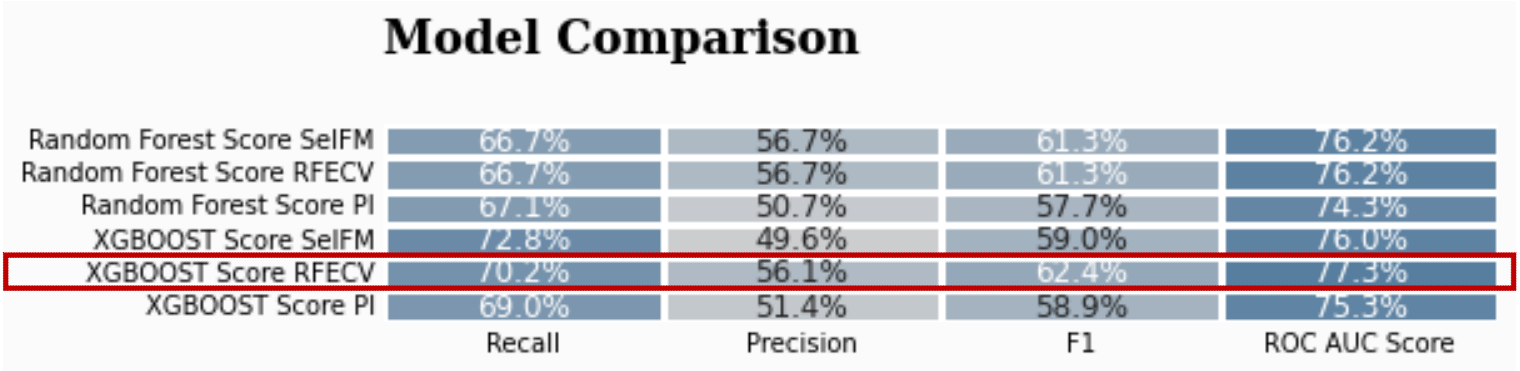
Pour l'approche Embedded, Il s'agit ici de méthodes de sélection de variable intégrées à la construction d'un modèle, à l'instar des régressions de régularisation LASSO, RIDGE et ELASTICNET.

Enfin, l'approche Wrapper est basée sur des sélections qui sont aléatoires initialement, et qui sont renforcées en les modifiant de manière répétée jusqu'à qu'une bonne solution soit trouvée. Cette approche nécessite d'évaluer les propositions une à une avec une cross-validation reproduisant au plus près le scénario d'utilisation des modèles choisis. Ces méthodes sont généralement très coûteuses en temps de calcul.

Dans ce mémoire, nous allons en tester une approche filter (selectKBest), deux approches embedded (SelectfromModel et RFECV Recursive Feature Elimination avec Cross-validation). RFECV est un algorithme d'optimisation gourmand qui vise à trouver le sous-ensemble de variables le plus performant. Il crée donc à plusieurs reprises des modèles et met de côté la variable la plus performante ou la moins performante à chaque itération. Il construit ensuite le modèle suivant avec les variables de gauche jusqu'à ce que toutes les variables soient épuisées. Il classe ensuite les variables en fonction de l'ordre de leur élimination.

Aussi, nous examinerons la méthode de Permutation Importance comme méthode de sélection de variables, en prenant les 15 premières.

2.4.4 Comparaison des modèles

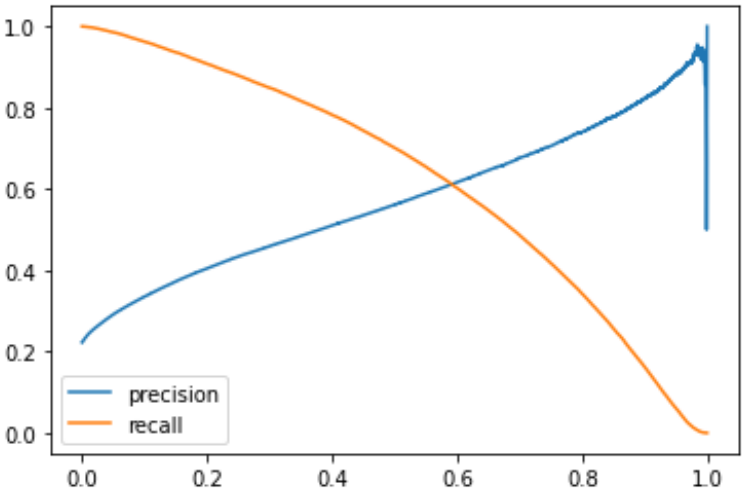


Même si le meilleur recall est celui du XGBoost avec la sélection SelectfromModel, le modèle que nous allons retenir est le **XGBoost avec la sélection de variables RFECV** qui présente un ensemble de scores meilleurs.

2.4.5 Ajustement du seuil et modèle final

Pour prendre en compte le compromis entre la Precision et le Recall, nous nous intéressons à une courbe retraçant les deux métriques et résume la performance globale du modèle.

Pour le modèle retenu XGBoost, le graphique ci-dessous nous montre où se situe les valeurs optimales pour les métriques.
Préférant privilégier le recall au profit de la precision, nous pouvons alors situer un threshold à 0.44, ce qui nous maximisera notre **recall à 0.73** pour une **precision à 0.56** et un **f1-score à 0.62.**
Ainsi, notre modèle final aura donc pour paramètre de seuil 0.44.



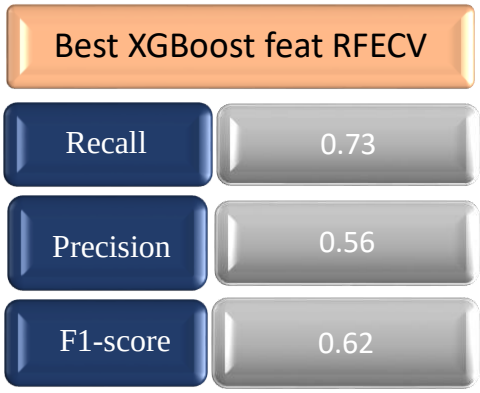


Figure 21 : Courbe precision recall Best XGBoost

3. Interprétabilité des résultats (SHAP)

L'interprétabilité d'un modèle peut être obtenue de deux façons :

• Par transparence - le modèle, comme les modèles classiques (GLM, arbres de décision simples), par essence et par construction est interprétable.

• Par interprétation post hoc - en exploitant des méthodes d'interprétation sans chercher à expliquer exactement comment le modèle fonctionne.

C'est cette dernière approche agnostique que nous mettrons en œuvre dans ce mémoire.

Les méthodes sont agnostiques au modèle, c'est-à-dire qu'elles peuvent être utilisées sur n'importe quel modèle. Elles sont toutes utilisables pour la régression, ainsi que la classification lorsque l'output du modèle est une probabilité ou un score lié à l'appartenance à une classe.

L'interprétabilité peut être envisagée selon deux angles :

- L'interprétation globale (explique le comportement du modèle de façon globale)
- L'interprétation locale (explique la prédiction d'une seule instance de données)

L'avantage du premier angle vient du côté plus synthétique tandis que le second permet d'être plus précis.

Pour qu'un modèle soit compréhensible pour un humain, il doit être particulièrement simple (de type arbre ou régression classique). Le dilemme de l'interprétation globale est donc d'arriver à construire un modèle suffisamment simple pour être compris, mais suffisamment complexe pour approximer le modèle initial avec un degré de précision acceptable.

Nous utiliserons les attributs summary_plot, force_plot, dependence_plot de SHAP.

SHAP Summary Plot

Le graphique combine feature importance et feature effects et permet d'avoir une interprétation globale ainsi que le sens de l'impact des variables. Chaque point du summary représente la Shapley Value pour une observation. Leur position sur l'axe des abscisses indique la Shapley value (la contribution) et leur couleur indique la valeur x_{ij} de la variable explicative (le joueur).

Les Shapley values proviennent initialement de la théorie des jeux et ont été conçues comme mesure de partage des gains entre joueurs dans un jeu coopératif Shapley [1953].

Les variables sont ordonnées par ordre d'importance dans le modèle. Ce résultat en termes d'ordre d'importance est conforme à celui de la permutation feature importance.

SHAP Dependence Plot

C'est la représentation des couples $\{(x_j^{(i)}, \phi_j^{(i)})\}_{i=1}^{n}$

Autrement dit, ce graphique présente la répartition des valeurs de Shapley d'une variable en fonction des valeurs de celle-ci. Il permet de visualiser la forme de la dépendance (ou de liaison) du modèle à la variable en question plus précisément que le graphique précédent, mais une variable à la fois.

SHAP Force Plot

Ce type de graphe donne l'explication d'impact des variables pour un exemple du dataset. En rouge, les variables qui ont un impact positif (contribuent à ce que la prédiction soit plus élevée que la valeur de base) et, en bleu, celles ayant un impact négatif (contribuent à ce que la prédiction soit plus basse que la valeur de base)

En pratique

Pour notre modèle XGBoost, nous avons calculé les SHAP values sur un sous-échantillon du jeu d'entraînement de 15000 observations stratifiées sur la prédiction du modèle, car le calcul est trop fastidieux si nous devions réaliser sur l'intégralité de l'échantillon.

3.1 Interprétation globale en moyenne sur un sous-échantillon représentatif

3.1.1 Feature importance et Summary plot

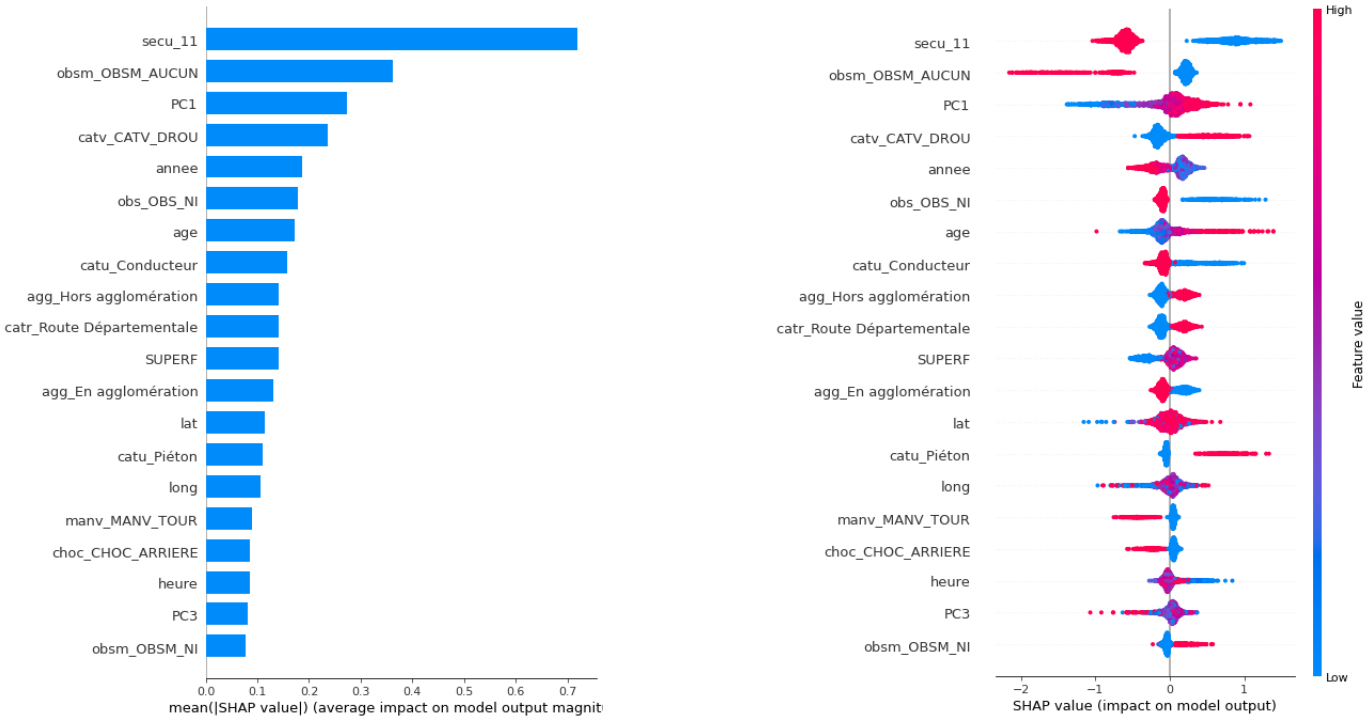


Figure 22 : Influence moyenne globale des variables pour un ensemble représentatif de prédictions de la gravité de l'accident par le modèle XGBoost

Observations (sur les 4 premières variables)

Sur le graphe feature_importance, nous avons la contribution marginale de chaque variable à la prédiction. Le top 3 des variables les plus contributrices sont secu_11, obsm_OBSM_AUCUN et PC1.
Pour rappel, secu_11 correspond à l'utilisation de l'équipement de sécurité, obsm_OBSM_AUCUN correspond au fait qu'il ne s'agit d'aucun obstacle mobile heurté et PC1, était la première composante principale de l'ACP, qui correspond à la superficie et à la présence d'une population d'agriculteurs exploitants (CS1).

Ainsi, pour la variable secu_11 par exemple, beaucoup de points rouges dans les Shapley Values s'entassent à une même valeur négative (-0.8 environ), et les points bleus s'étalent sur les valeurs positives. Cela montre que les valeurs faibles de secu_11 (c-à-d 0, ne pas porter l'équipement) participent à augmenter la gravité de l'accident mais que les valeurs élevées (c-à-d 1) ne participent pas à diminuer ce risque par rapport à la moyenne. Autrement dit, le fait de ne pas porter l'équipement de sécurité contribue fortement à l'aggravation de l'accident.

Pour la variable obsm_OBSM_AUCUN (aucun obstacle mobile heurté), beaucoup de points rouges (c-à-d 1, soit aucun obstacle) s'étalent vers les valeurs négatives alors que les bleus (valeurs faibles c-à-d 0) s'entassent à une valeur Shapley positive juste après 0. On en déduit que le petit nombre des accidents aux variables faibles (c-à-d 0) ne participe pas à augmenter le risque d'accident grave (car tout proche de 0), alors que le fait qu'il n'y ait aucun obstacle mobile heurté contribue à faire diminuer la gravité de l'accident, à rendre tout simplement celui-ci moins grave. Rappelons que dans la définition, l'obstacle mobile heurté peut-être un piéton, un animal sauvage, etc.

Pour la troisième variable la plus importante PC1 (combinaison linéaire des variables de l'ACP), nous remarquons que beaucoup de points rouges (valeurs élevées de superficies et de population de CS1) s'entassent à des valeurs Shapley proche de 0 mais s'étalent aussi vers les valeurs Shapley positives sur l'axe. Ceci montre qu'il y a un impact positif sur la gravité de l'accident avec des valeurs de PC1 élevées. De l'autre côté, les bleus (valeurs faibles) s'étalent, eux, vers les valeurs Shapley négatives et participent donc à diminuer la gravité de l'accident. Ceci confirme les analyses faites plus haut lors de l'analyse exploratoire, que plus on se trouve en zones de grandes étendues avec une population à majorité de CS1 (agriculteurs-exploitants), plus les accidents sont graves.

Pour la variable catv_CATV_DROU, nous voyons tout de suite que les valeurs positives rouges s'étalant sur les valeurs Shapley positives, contribuent donc fortement à rendre l'accident grave, et les valeurs bleus (indiquant l'absence de 2 roues motorisés) n'exerce aucune influence sur la gravité de l'accident (proche de 0).

3.1.2 Dépendance Partielle 1D

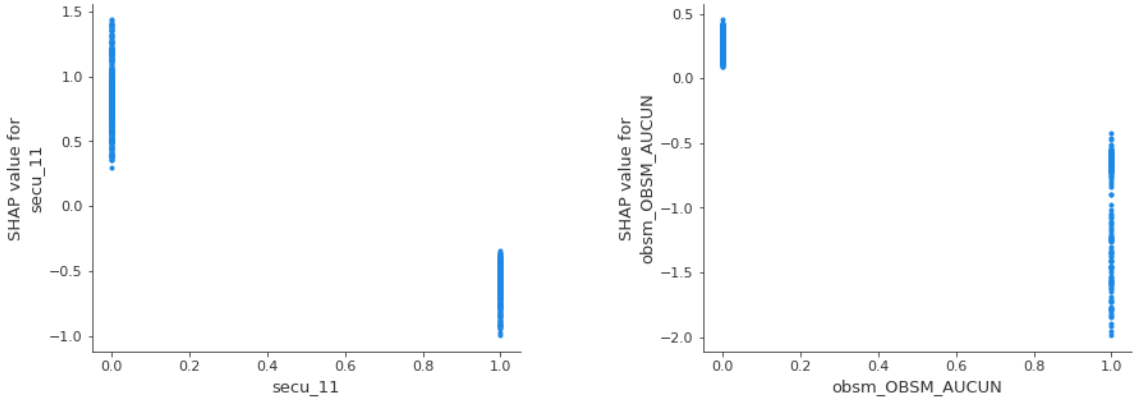


Figure 23 : Dépendance partielle plot 1D : secu_11, obsm_OBSM_AUCUN

Observations :
Sur les 2 graphes, nous pouvons visualiser la forme de la dépendance du modèle à la variable en question, ici secu_11 et obsm_OBSM_AUCUN.
Pour secu_11, les valeurs à 0 prises par secu_11, font croître la gravité de l'accident et les valeurs à 1 de secu_11 font diminuer le risque de gravité de celui-ci. Ceci confirme l'interprétation précédente.

Quant à obsm_OBSM_AUCUN, les valeurs à 0 de la variable font légèrement augmenter le risque d'accident grave, tandis que les valeurs à 1, elles, ont une influence négative forte sur la gravité de l'accident. Cela reste en phase avec l'interprétation précédente.

3.1.3 Dépendance Partielle 2D

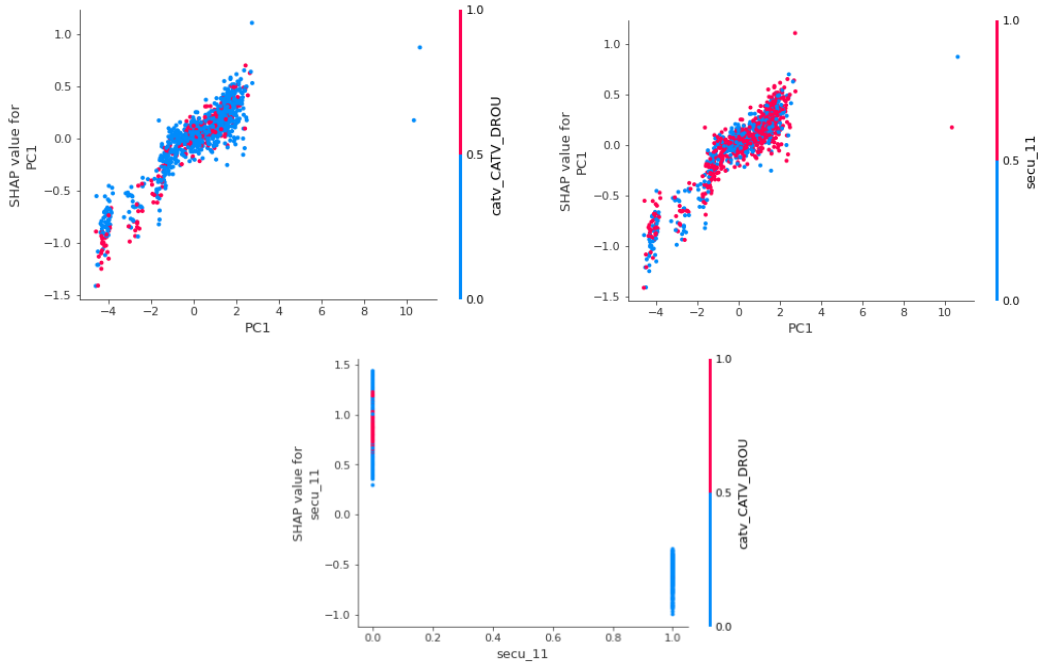


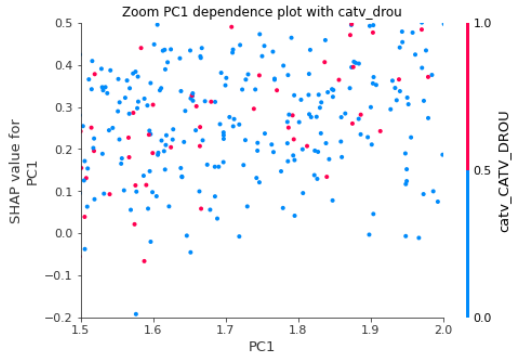
Figure 24 : Dépendance partielle plot 2D :PC1-catv_CATV_DROU, PC1-secu_11, secu_11-catv_CATV_DROU

Observations sur les interactions :

Étudions les interactions des premières variables :

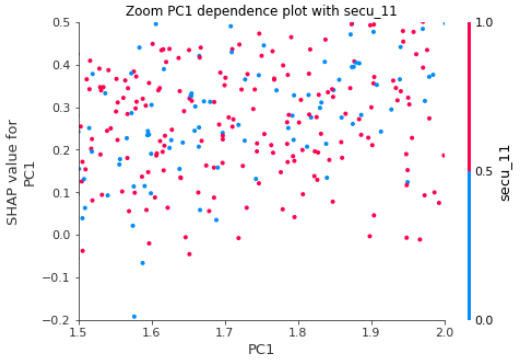
- Entre PC1 et catv_CATV_DROU :

Pour des valeurs faibles (négatives) de PC1, nous constatons que le risque de gravité de l'accident s'accroît chez les accidents en l'absence de 2 roues motorisés. Pour les valeurs plus élevées et positives de PC1 (à partir de -1), la gravité de l'accident croît aussi en l'absence de 2 roues motorisés (plus de points bleus que rouges). Si nous zoomons un peu plus sur des valeurs élevées entre 1.5 et 2 de PC1, l'influence de la variable catv_CATV_DROU fait que les points bleus sont situés plus bas en général que les rouges (plus on fait croître PC1 vers 2), ceci signifie que ces valeurs de PC1, associées à l'absence de 2 roues motorisé, poussent davantage la donnée vers la classe 0 qu'une donnée avec le même PC1 mais associée à la présence de 2 roues motorisé.



- Entre PC1 et secu_11 :

Pour les valeurs faibles négatives de PC1, le risque de gravité de l'accident s'accroît lorsque la variable secu_11 est à 1 (point rouge). Pour les valeurs plus élevées de PC1, si nous zoomons sur 1.5 à 2, les points rouges sont prédominants partout et les points bleus, eux, sont placés plus en haut au niveau de valeurs Shapley vers 0.3 qui les poussent davantage vers la classe 1.



- Entre secu_11 et catv_CATV_DROU :

Pour la valeur 0 de secu_11, nous voyons un mix entre une partie des points bleus d'abord puis de rouges et à nouveau de bleus puis rouges. Il est difficile de donner une unique interprétation.
Les valeurs de secu_11 à 0 associées à la présence de 2 roues motorisés vient accroître la gravité de l'accident pour certaines valeurs Shapley uniquement.
Les valeurs de secu_11 à 1 associées à l'absence de 2 roues motorisé viennent pousser à l'appartenance à la classe 0.
En d'autres termes, l'utilisation de l'équipement de sécurité associée à l'absence de 2 roues motorisé rendent l'accident moins grave. Le non-port de l'équipement de sécurité en présence ou en l'absence de 2 roues peut aggraver l'accident.

3.2 Interprétation Locale Détaillée sur un Sous-Échantillon

Dans un second temps nous cherchons à comprendre plus en détail sur quelles modalités précisément le modèle s'appuie pour ses prédictions. Pour cela, nous allons sélectionner des observations qui ont été mal classées par le modèle.

Représentation d'observations False Negative

Figure 25 : Influence des variables pour une observation précise (false negative)

Observations :

Dans le premier exemple, la valeur prédite est à -1.93 (à droite de la base value), classée en 0. Comme les valeurs SHAP de chaque variable sont proportionnelles aux tailles des flèches, nous constatons que c'est la variable secu_11 qui contribue le plus à atteindre la valeur prédite, avec un impact négatif. Aussi, dans le cas de cette donnée, les valeurs des variables col_Deux véhicules – par l'arrière (impact négatif), obsm_OBSM_AUCUN et agg_En agglomération (impact positif) sont aussi déterminantes pour prédire que la donnée appartient à la classe 0. Ici, il s'agit d'un accident survenu avec l'utilisation de l'équipement de sécurité, pendant une collision de deux véhicules par l'arrière, hors agglomération.

Dans le deuxième exemple, la valeur prédite est à -0.71 et a donc été classée en 0.
Les variables les plus contributrices sont secu_11 (impact positif) et catv_CATV_DROU (impact positif), et choc_CHOC_NI avec un impact négatif. Il s'agissait d'un accident en conditions de non-port de l'équipement de sécurité, sur un véhicule 2 roues motorisé, avec un choc non identifié.

Représentation d'observations False Positive

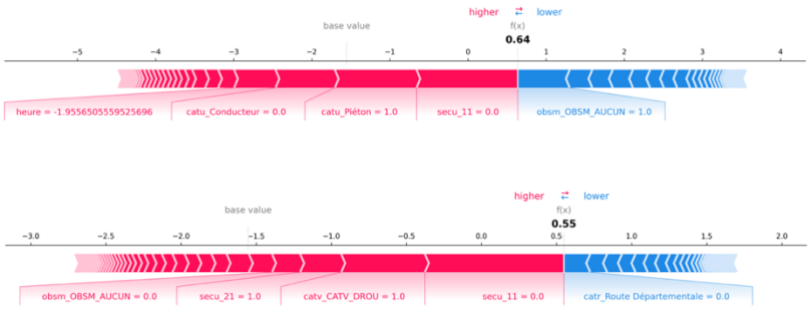


Figure 26 : Influence des variables pour une observation précise (false positive)

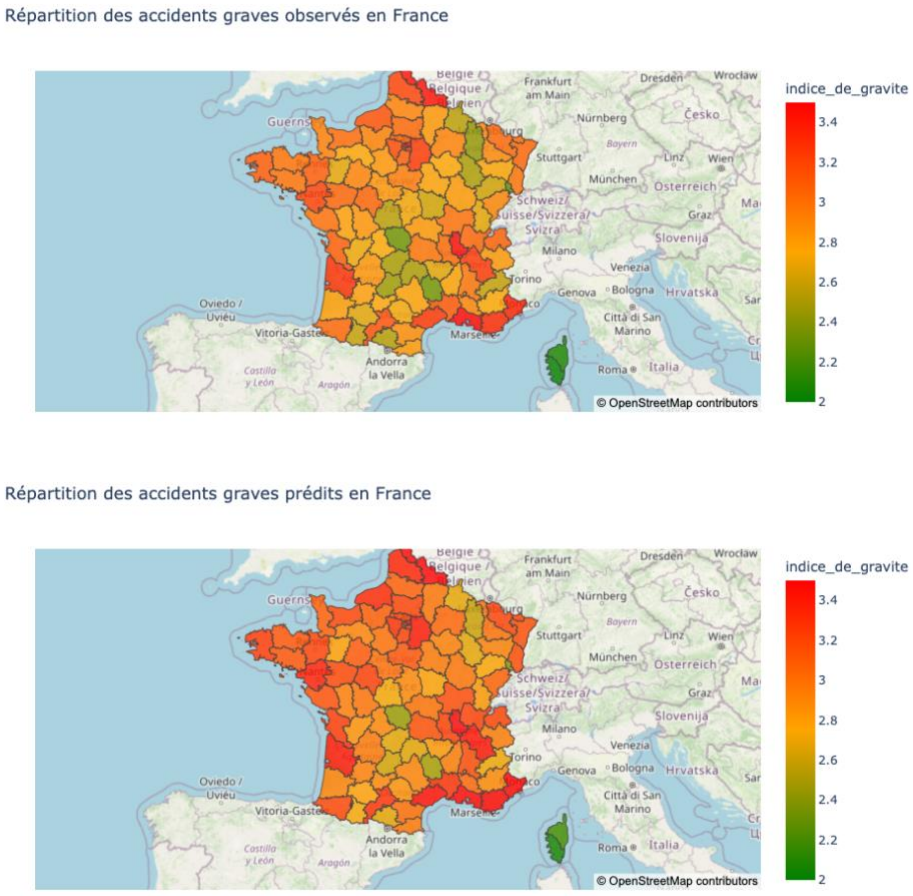
Observations :

Dans le premier exemple, la valeur prédite est à 0.64 (d'où la classe 1). Les variables et modalités qui ont le plus contribué sont secu_11=0 et catu_Piéton=1 avec un impact positif (en rouge), obsm_OBSM_AUCUN=1 (en bleu, impact négatif). Cet accident est survenu en conditions de non-port de l'équipement de sécurité, dont la victime est un piéton, sans obstacle mobile heurté.

Le deuxième exemple de False positive a obtenu une valeur prédite à 0.55 (classe 1).
Les variables qui ont contribué le plus en impact positif sont secu_11, catv_CATV_DROU et secu_21 et en impact négatif la variable catr_Route Départemenale.
Les circonstances de cet accident classé grave sont : non-port de l'équipement de sécurité principal mais port d'un équipement de sécurité supplémentaire, la catégorie du véhicule est un 2 roues motorisé, et pas sur une route départementale.

3.3 Visualisation des observations et des prédictions

Base Train



Observations :
Sur la base train, les accidents observés sur certains départements gardent la même tendance de couleur sur la carte des prédictions. C'est le cas particulier de la Corse, ou encore dans le 13, 69, 59, 90. Dans l'ensemble, cela dit, les nuances se retrouvent d'une carte à l'autre mais avec une accentuation sur la carte des prédictions. Le modèle a tendance à aggraver les départements.

Figure 27 : Carte des accidents graves réels et prédits par département (base train)

Base Test

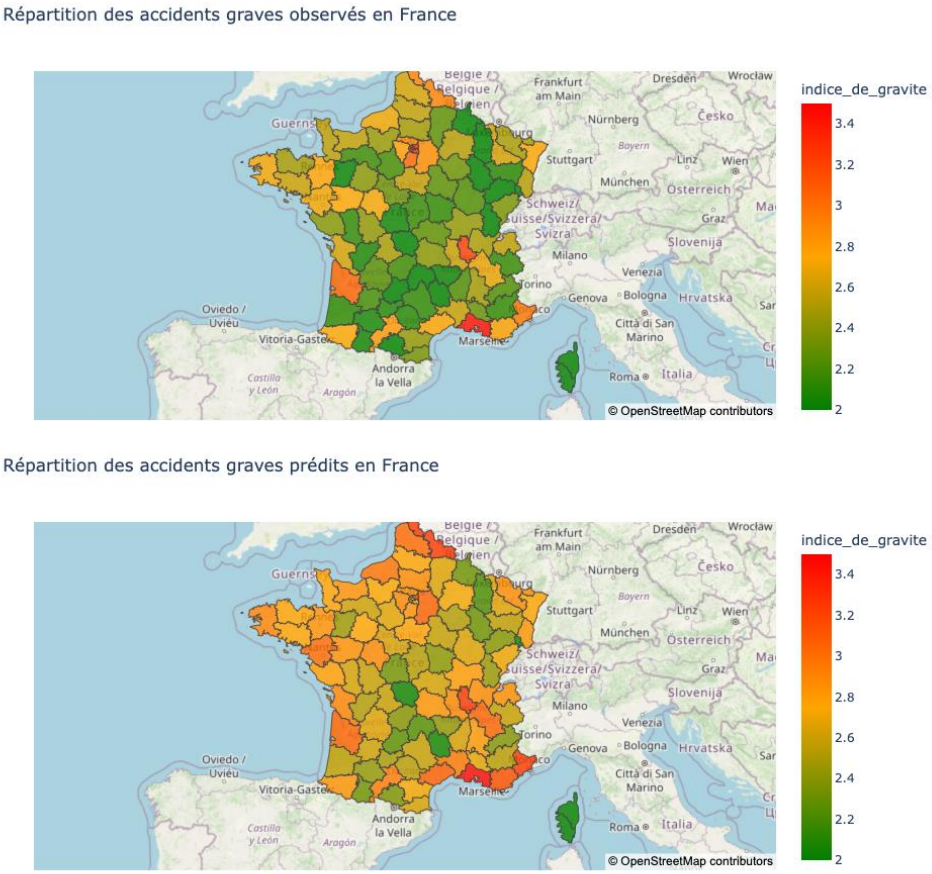


Figure 28 : Carte des accidents graves réels et prédits par département (base test)

s

Observations :

Sur la base test, les accidents observés sur certains départements gardent aussi la même couleur sur la carte des prédictions, comme le cas de la Corse toujours.
Nous remarquons aussi que sur les départements 13, 23, 69, 33, 67, 90, 64, nous avons les mêmes indices de gravité (a priori la même couleur) sur les 2 cartes.
L'écart entre les observations et les prédictions se manifeste par une accentuation des couleurs sur la carte des prédictions.

Notre modèle semble donc bien apprendre sur certains départements, et sur d'autres avec une accentuation de la gravité.

4. Conclusion

Dans ce mémoire, la problématique de prédiction de la gravité des accidents en France a conduit à construire plusieurs modèles de machine learning.

En amont, grâce aux données en Open Data mises à disposition par le gouvernement français et les données recueillies de l'Insee, nous avons été en capacité de construire une base relative aux accidents corporels de la route en France sur un historique de 5 ans entre 2015 et 2019. Pour le retraitement et le reformatage des données, un énorme travail a été mené, car les bases brutes par année n'étaient pas toujours très homogènes. Il a fallu faire un gros travail d'uniformisation par rubrique, étape cruciale pour la modélisation ensuite.

La base établie a permis ensuite de modéliser dans un premier temps un modèle en première intuition basé sur des model point, dans le but de prédire la fréquence de gravité des accidents présents dans une même cohorte.
Pour ce faire, après avoir agrégé les lignes selon des caractéristiques similaires (basé sur les premières variables montrant la meilleure corrélation avec la variable cible), un modèle de régression a été construit donnant les résultats vus au paragraphe concerné.

Nous avons explicité quel devait être l'objectif de notre prédiction selon trois mesures judicieusement choisies par le contexte : le recall, la precision et le f1-score (l'auc score a également été recueilli) en maximisant le recall car il est déterminant de pouvoir bien prédire les accidents graves. Sur cette base, nous avons construit le modèle paramétrique classique, une régression logistique, selon les règles de l'art et nous avons développé trois modèles de machine learning pour le challenger : un arbre de décision CART, un random forest et un XGBoost.
Ces méthodes de machine learning, surtout l'algorithme du XGBoost qui a été retenu, offrent de meilleures performances techniques que la régression logistique tout en alliant une bonne robustesse. L'inconvénient en revanche est la perte de lisibilité des arbres de décision.

Pour finir, ce mémoire s'achève en introduisant la méthode d'interprétation globale et locale SHAP sur notre modèle XGBoost.

A l'issue de ces différents modèles proposés et des interprétations qui en découlaient, on peut en déduire que la gravité d'un accident n'est pas la simple conséquence d'une conduite à risque, mais qu'elle est fortement liée au contexte dans lequel l'accident survient (route départementale, port de l'équipement de sécurité, présence de deux-roues motorisé, …).

Si nous devons donc nous resituer dans le contexte assurantiel, tous ces différents modèles construits à partir de données en Open Data nous font prendre conscience des multiples usages et applications qu'on peut en tirer notamment dans le domaine de la prévention et de l'assistance qui doivent venir compléter les efforts déjà mis en place.

Comme évoqué en introduction également, avec l'arrivée en force de l'assurance télématique, l'assureur de demain se doit de maîtriser l'utilisation de ces Open Data pour leur permettre de proposer des solutions innovantes afin de poursuivre leur rôle essentiel au sein de notre société et dans l'économie.

ANNEXES

Description des variables dans chaque table

Base Véhicules :

Num_Acc : numéro d'identifiant de l'accident
id_vehicule : identifiant unique du véhicule repris pour chacun des usagers occupant le véhicule (code numérique)
Num_Veh : identifiant unique du véhicule repris pour chacun des usagers occupant le véhicule (code alphanumérique)
senc : sens de circulation

-1 -> Non renseigné
0 -> Inconnu
1 -> PK ou PR ou numéro d'adresse postale croissant
2 -> PK ou PR ou numéro d'adresse postale décroissant
3 -> Absence de repère

catv : catégorie du véhicule

00 -> indéterminable
01 -> bicyclette
02 -> cyclomoteur <50cm3
03 -> voiturette
04 -> scooter immatriculé
05 -> motocyclette
06 -> side-car
07 -> VL seul
08 -> VL+caravane
09 -> VL+remorque
10 -> VU seul 15T <= PTAC <= 35T avec ou sans remorque
11 -> VU seul 15T <= PTAC <= 35T + caravane
12 -> VU seul 15T <= PTAC <= 35T + remorque
13 -> PL seul 35T < PTAC <= 75T
14 -> PL seul > 75T
15 -> PL > 35T + remorque
16 -> tracteur routier seul
17 -> tracteur routier + semi-remorque
18 -> transports en commun
19 -> tramway
20 -> engin spécial
21 -> tracteur agricole
30 -> scooter <50cm3
31 -> motocyclette > 50cm3 <= 125cm3
32 -> scooter > 50cm3 <= 125cm3
33 -> motocyclette > 125cm3
34 -> scooter > 125cm3

35 -> quad léger <=50cm3
36 -> quad lourd > 50cm3
37 -> autobus
38 -> autocar
39 -> train
40 -> tramway
41 -> 3RM <= 50cm3
42 -> 3RM > 50cm3 <=125cm3
43 -> 3RM > 125cm3
50 -> EDP à moteur
60 -> EDP sans moteur
80 -> VAE
99 -> Autre véhicule

obs : obstacle fixé heurté

-1-> Non renseigné
0 -> Sans objet
1 -> Véhicule en stationnement
2 -> Arbre
3 -> Glissière métallique
4 -> Glissière en béton
5 -> Autre glissière
6 -> Bâtiment mur pile de pont
7 -> Support de signalisation verticale ou poste d'appel d'urgence
8 -> Poteau
9 -> Mobilier urbain
10 -> Parapet
11 -> Ilot refuge borne haute
12 -> Bordure de trottoir
13 -> Fossé talus paroi rocheuse
14 -> Autre obstacle fixe sur chaussée
15 -> Autre obstacle fixe sur trottoir ou accotement
16 -> Sortie de chaussée sans obstacle
17 -> Buse - tête d'aqueduc

obsm : obstacle mobile heurté

-1-> Non renseigné
0 -> Sans objet
1 -> Piéton
2 -> Véhicule
4 -> Véhicule sur rails
5 -> Animal domestique
6 -> Animal sauvage
9 -> Autre

choc : point de choc initial

-1-> Non renseigné

0 -> Aucun
1 -> Avant
2 -> Avant droit
3 -> Avant gauche
4 -> Arrière
5 -> Arrière droit
6 -> Arrière gauche
7 -> Côté droit
8 -> Côté gauche
9 -> Chocs multiples (tonneaux)

manv : manœuvre principale avant l'accident

-1-> Non renseigné
0 -> Inconnue
1 -> Sans changement de direction
2 -> Même sens même file
3 -> Entre 2 files
4 -> Marche arrière
5 -> A contre sens
6 -> En franchissant le terre-plein central
7 -> Dans le couloir bus dans le même sens
8 -> Dans le couloir bus en sens Inverse
9 -> En s'insérant
10-> En faisant demi-tour sur la chaussée

Changeant de file

11-> A gauche
12-> A droite

Déporté

13-> A gauche
14-> A droite

Tournant

15-> A Gauche
16-> A droite

Dépassant

17-> A Gauche
18-> A droite

Divers

19-> Traversant la chaussée
20-> Manoeuvre de stationnement

21-> Manoeuvre d’évitement
22-> Ouverture de porte
23-> Arrêté (hors stationnement)
24-> En stationnement (avec occupant)
25-> Circulant sur trottoir
26-> Autres manœuvres

motor : type de motorisation du véhicule

-1-> Non renseigné
0 -> inconnue
1 -> Hydrocarbures
2 -> Hybride électrique
3 -> Électrique
4 -> Hydrogène
5 -> Humaine
6 -> Autre

occutc : nombre d’occupants dans le transport en commun

Base Lieux :

Num_Acc : numéro d’identifiant de l’accident
catr : catégorie de route

1 -> autoroute
2 -> route nationale
3 -> route départementale
4 -> voie communale
5 -> hors réseau public
6 -> parc de stationnement ouvert à la circulation publique
9 -> autre

voie : numéro de la route
V1 : indice numérique du numéro de route
V2 : lettre indice alphanumérique de la route

circ : régime de circulation

-1 -> non renseigné
1 -> a sens unique
2 -> bidirectionnelle
3 -> a chaussées séparées
4 -> avec voies d’affectation variable

nbv : nombre total de voies de circulation

vosp : signale l’existence d’une voie réservée indépendamment du fait que l’accident ait lieu ou non sur la voie

-1 -> non renseigné
0 -> sans objet
1 -> piste cyclable
2 -> bande cyclable
3 -> voie réservée

prof : profil en long décrit la déclivité de la route à l’endroit de l’accident

-1 -> non renseigné
1 -> plat
2 -> pente
3 -> sommet de côte
4 -> bas de côte

pr : numéro du PR "point de référence" de rattachement (numéro de la borne amont)
pr1 : distance en mètres du PR "point de référence" (par rapport à la borne amont)
plan : tracé en plan

-1 -> non renseigné
1 -> partie rectiligne
2 -> en courbe à gauche
3 -> en courbe à droite
4 -> en "S"

lartpc : largeur du terre-plein central (TPC) s’il existe
larrout : largeur de la chaussée affectée à la circulation des véhicules ne sont pas compris les bandes d’arrêt d’urgence les TPC et les places de stationnement

surf : état de la surface

-1 -> non renseigné
1 -> normale
2 -> mouillée
3 -> flaques
4 -> inondée
5 -> enneigée
6 -> boue
7 -> verglacée
8 -> corps gras - huile
9 -> autre

infra : aménagement - infrastructure

-1 -> non renseigné
0 -> Aucun
1 -> souterrain- tunnel
2 -> pont - autopont
3 -> bretelle d’échangeur ou de raccordement

4 -> voie ferrée
5 -> carrefour aménagé
6 -> zone piétonne
7 -> zone de péage
8 -> Chantier
9 -> Autres

situ : situation de l'accident

-1 -> non renseigné
0 -> Aucun
1 -> sur chaussée
2 -> sur bande d'arrêt d'urgence
3 -> sur accotement
4 -> sur trottoir
5 -> sur piste cyclable
6 -> sur autre voie spéciale
8 -> autres

env1 : point école proximité d'une école
vma : vitesse maximale autorisée sur le lieu et au moment de l'accident

Base Caractéristiques :

Num_Acc : Numéro d'identifiant de l'accident
jour : jour de l'accident
mois : mois de l'accident
an : année de l'accident
hrmn : heure et minutes de l'accident
lum : lumière conditions d'éclairage dans lesquelles l'accident s'est produit

1 -> plein jour
2 -> crépuscule ou aube
3 -> nuit sans éclairage public
4 -> nuit avec éclairage public non allumé
5 -> nuit avec éclairage public allumé

dep : département code Insee
com : commune code Insee
agg : localisation

1 -> hors agglomération
2 -> en agglomération

int : intersection

1 -> hors intersection
2 -> intersection en X
3 -> intersection en T
4 -> intersection en Y
5 -> intersection à plus de 4 branches
6 -> giratoire
7 -> place
8 -> passage à niveau
9 -> autre intersection

atm : conditions atmosphériques

-1 -> non renseigné
1 -> normale
2 -> pluie légère
3 -> pluie forte
4 -> neige - grêle
5 -> brouillard - fumée
6 -> vent fort - tempête
7 -> temps éblouissant
8 -> temps couvert
9 -> autre

col : type de collision

-1 -> non renseigné
1 -> deux véhicules - frontale
2 -> deux véhicules - par l'arrière
3 -> deux véhicules - par le côté
4 -> trois véhicules et plus - en chaîne
5 -> trois véhicules et plus - collisions multiples
6 -> autre collision
7 -> sans collision

adr : adresse postale
gps : codage gps

M -> Métropole
A -> Antilles (Martinique ou Guadeloupe)
G -> Guyane
R -> Réunion
Y -> Mayotte

lat : latitude
long : longitude

Base Usagers :

Num_Acc : numéro d'identifiant de l'accident
Num_Veh : identifiant du véhicule repris pour chacun des usagers occupant ce véhicule
place : permet de situer la place occupée dans le véhicule par l'usager au moment de l'accident
catu : catégorie d'usager

1 -> conducteur
2 -> passager
3 -> piéton

grav : gravité de l'accident

1 -> indemne
2 -> tué
3 -> blessé hospitalisé
4 -> blessé léger

sexe : sexe de l'usager

1 -> masculin
2 -> féminin

an_nais : année de naissance de l'usager

trajet : motif du déplacement au moment de l'accident

-1 -> non renseigné
0 -> non renseigné
1 -> domicile - travail
2 -> domicile - école
3 -> courses - achats
4 -> utilisation professionnelle
5 -> promenade - loisirs
9 -> autre

secu : sur 2 caractères (jusqu'en 2018) le premier concerne l'existence d'un équipement de sécurité

1 -> ceinture
2 -> casque
3 -> dispositif enfant
4 -> équipement réfléchissant
9 -> autre

le second concerne l'utilisation de l'équipement de sécurité

1 -> oui
2 -> non
3 -> non déterminable

secu1 : le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité

-1 -> non renseigné
0 -> aucun équipement
1 -> ceinture
2 -> casque
3 -> dispositif enfant
4 -> gilet réfléchissant
5 -> airbag (2RM/3RM)
6 -> gants (2RM/3RM)
7 -> gants + airbag (2RM/3RM)
8 -> non déterminable
9 -> autre

secu2 : le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité

-1 -> non renseigné
0 -> aucun équipement
1 -> ceinture
2 -> casque
3 -> dispositif enfant
4 -> gilet réfléchissant
5 -> airbag (2RM/3RM)
6 -> gants (2RM/3RM)
7 -> gants + airbag (2RM/3RM)
8 -> non déterminable
9 -> autre

secu3 : le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité

-1 -> non renseigné
0 -> aucun équipement
1 -> ceinture
2 -> casque
3 -> dispositif enfant
4 -> gilet réfléchissant
5 -> airbag (2RM/3RM)
6 -> gants (2RM/3RM)
7 -> gants + airbag (2RM/3RM)
8 -> non déterminable
9 -> autre

locp : localisation du piéton
Sur chaussée

1 -> A >50m du passage piéton
2 -> A <50m du passage piéton

Sur passage piéton

3 -> Sans signalisation lumineuse
4 -> Avec signalisation lumineuse

Divers

5 -> sur trottoir
6 -> sur accotement
7 -> sur refuge ou BAU
8 -> sur contre allée
9 -> inconnue

actp : action du piéton
se déplaçant

-1 -> non renseigné
0 -> non renseigné ou sans objet
1 -> sens véhicule heurtant
2 -> sens inverse du véhicule

divers

3 -> traversant
4 -> masqué
5 -> jouant-courant
6 -> avec animal
9 -> autre

etatp : permet de situer si le piéton accidenté était seul ou non

-1 -> non renseigné
1 -> seul
2 -> accompagné
3 -> en groupe

Recensement des variables et leurs traitements

Usagers

Nom de variable	Définition de la variable	Type de traitement de la variable
Num_Acc	numéro d'identifiant de l'accident	
id_vehicule	Identifiant unique du véhicule - code numérique	supprimée
num_veh	identificant du véhicule - code alphanumérique	supprimée
place	place occupée dans le véhicule	reformatée
catu	catégorie d'usager	recodage du libellé des modalités
grav	**gravité de l'accident (1 : indemne, 2 : tué, 3 : blessé hospitalisé, 4 : blessé léger)**	**retraitée en variable binaire : - 0 accident pas grave (1,4) - 1 accident grave (2,3)**
sexe	sexe de l'usager	recodage du libellé des modalités
an_nais	année de naissance de l'usager	reformatée
trajet	motif du déplacement au moment de l'accident	reformatée, recodage du libellé des modalités
secu	existence et utilisation d'un équipement de sécurité (sur 2 caractères)	
secu1	présence et l'utilisation de l'équipement de sécurité	retraitée sous le même format que secu (sur 2 caractères)
secu2	présence et l'utilisation de l'équipement de sécurité	supprimée
secu3	présence et l'utilisation de l'équipement de sécurité	supprimée
locp	localisation du piéton	
actp	action du piéton	
etatp	permet de situer si le piéton accidenté était seul ou non	
an	année de survenance	nouvelle variable créée
age	âge de l'usager	nouvelle variable créée
classe age	classe d'âge de l'usager	age (tranches d'âge de 10 ans entre 0 et 120 ans)

Caractéristiques

Nom de variable	Définition de la variable	Type de traitement de la variable
Num_Acc	numéro d'identifiant de l'accident	
an	année de l'accident	retraitée
mois	mois de l'accident	
jour	jour de l'accident	
hrmn	heure et minutes de l'accident	supprimée
lum	lumière, conditions d'éclairage	recodage de libellé des modalités
agg	localisation en agglomération ou hors agglomération	recodage de libellé des modalités
int	intersection	recodage de libellé des modalités
atm	conditions atmosphériques	retraitée et recodage de libellé des modalités
col	type de collision	retraitée et recodage de libellé des modalités
com	commune code Insee	supprimée
adr	adresse postale	supprimée
gps	codage gps	supprimée
lat	latitude	retraitée
long	longitude	retraitée
dep	département code Insee	retraitée
heure	heure de l'accident	nouvelle variable créée (sur la base de hrmn)
minute	minute de l'accident	nouvelle variable créée (sur la base de hrmn)
Date	date de l'accident	nouvelle variable créée (sur la base de jour, mois, année)
Week_end	week-end ou non	nouvelle variable créée (sur la base de Date)
weekday	jour de la semaine	nouvelle variable créée (sur la base de Date)
Gpe_dep	groupe de départements	nouvelle variable créée (sur la base de dep)

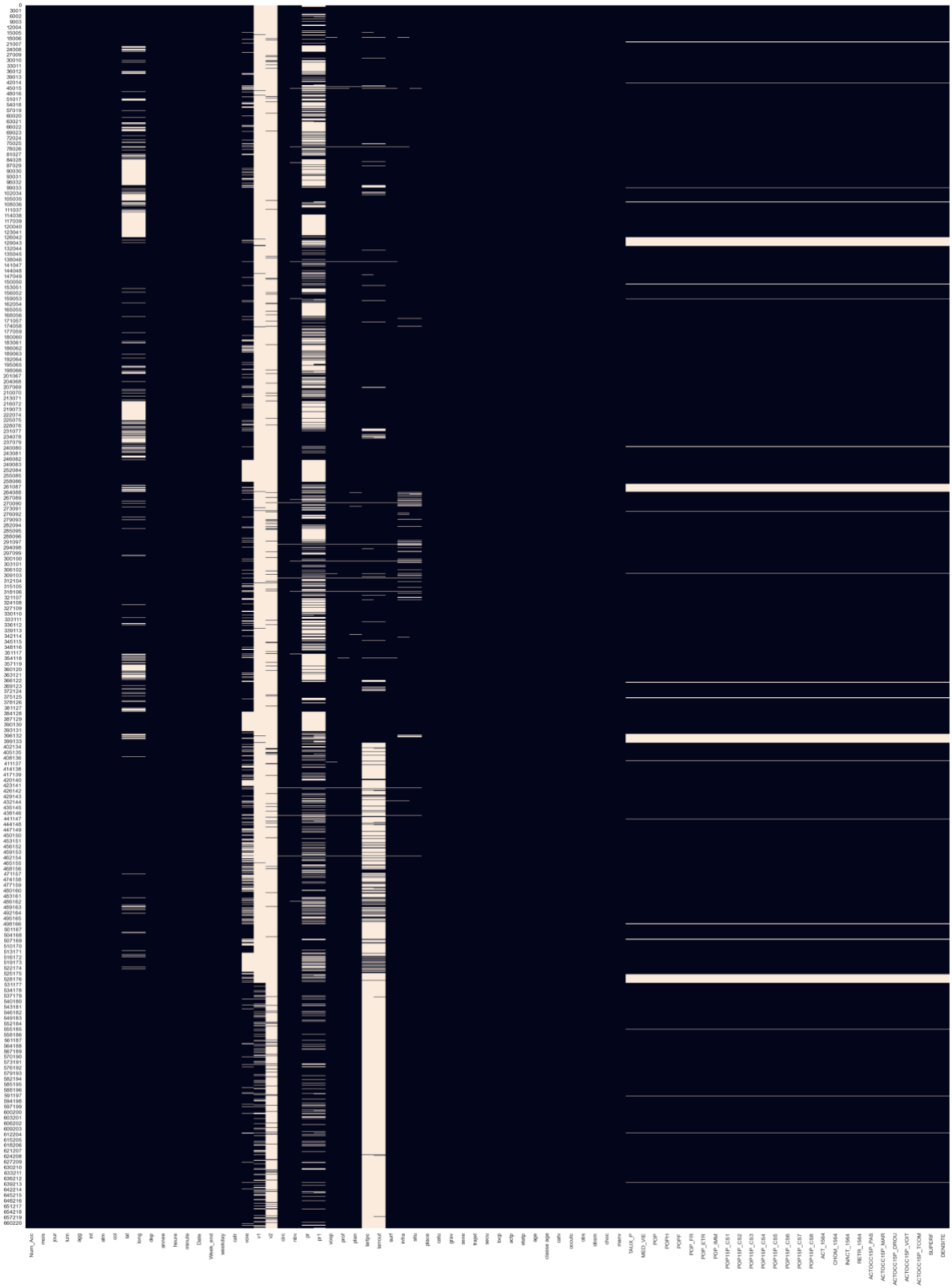
Lieux

Nom de variable	Définition de la variable	Type de traitement de la variable
Num_Acc	numéro d'identifiant de l'accident	
catr	catégorie de route	recodage de libellée des modalités
voie	numéro de la route	
v1	indice numérique du numéro de route	
v2	lettre indice alphanumérique de la route	
circ	régime de circulation	recodage de libellée des modalités
nbv	nombre total de voies de circulation	
pr	numéro du PR "point de référence"	reformatée
pr1	distance en mètres du PR "point de référence"	reformatée
vosp	existence d'une voie réservée	recodage de libellée des modalités
prof	profil en long décrit la déclivité de la route	recodage de libellée des modalités
plan	tracé en plan	recodage de libellée des modalités
lartpc	largeur du terre plein central	
larrout	largeur de la chaussée	
surf	état de la surface	recodage de libellée des modalités
infra	aménagement - infrastructure	recodage de libellée des modalités
situ	situation de l'accident	recodage de libellée des modalités
env1	point école	supprimée
vma	vitesse maximale	supprimée

Véhicules

Nom de variable	Définition de la variable	Type de traitement de la variable
Num_Acc	numéro d'identifiant de l'accident	
senc	sens de circulation	reformatée
num_veh	alphanumérique	supprimée
catv	catégorie du véhicule	retraitée
occutc	nombre d'occupants dans le transport en commun	reformatée
obs	obstacle fixé heurté	reformatée, retraitée et recodage de libellé des modalités
obsm	obstacle mobile heurté	reformatée et recodage de libellé des modalités
choc	point de choc initial	reformatée, retraitée et recodage de libellé des modalités
manv	manoeuvre principale avant l'accident	reformatée, retraitée et recodage de libellé des modalités
id_vehicule	identifiant unique du véhicule	supprimée
motor	type de motorisation du véhicule	supprimée

Heatmap des données manquantes

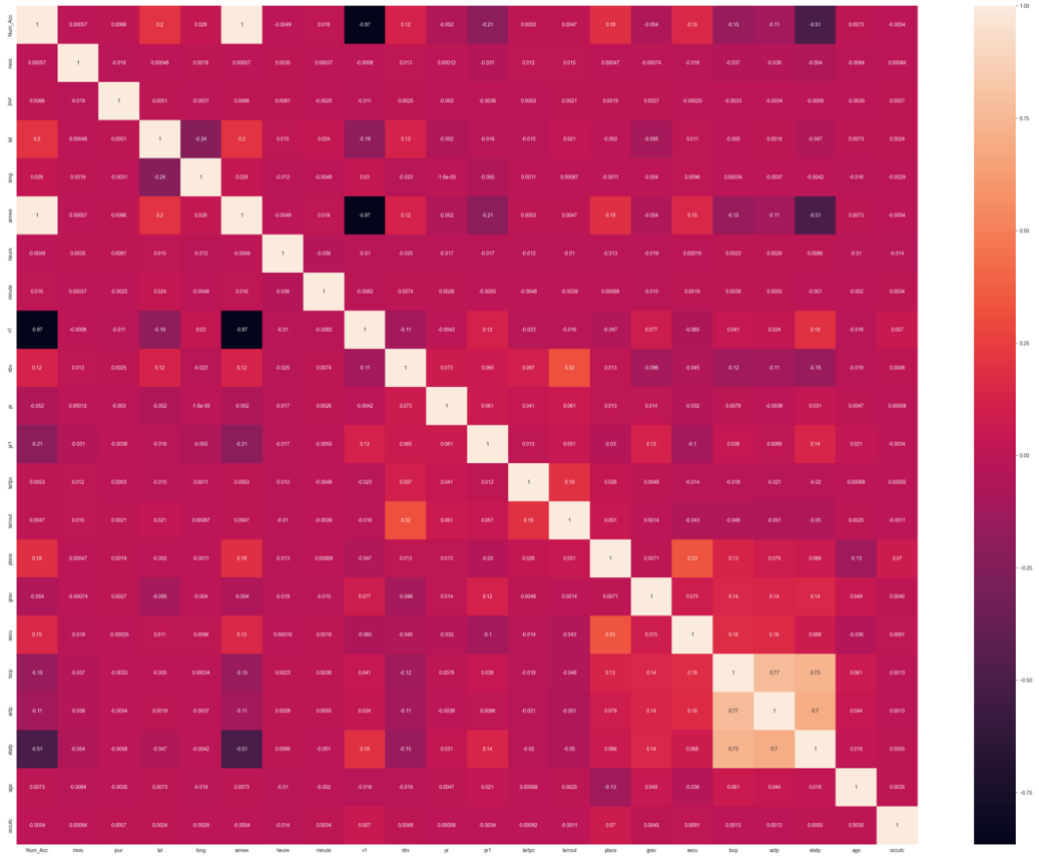


Mesures de liaisons entre variables qualitatives-grav (tableau de gauche) et entre variables quantitatives-grav (tableaux de droite)

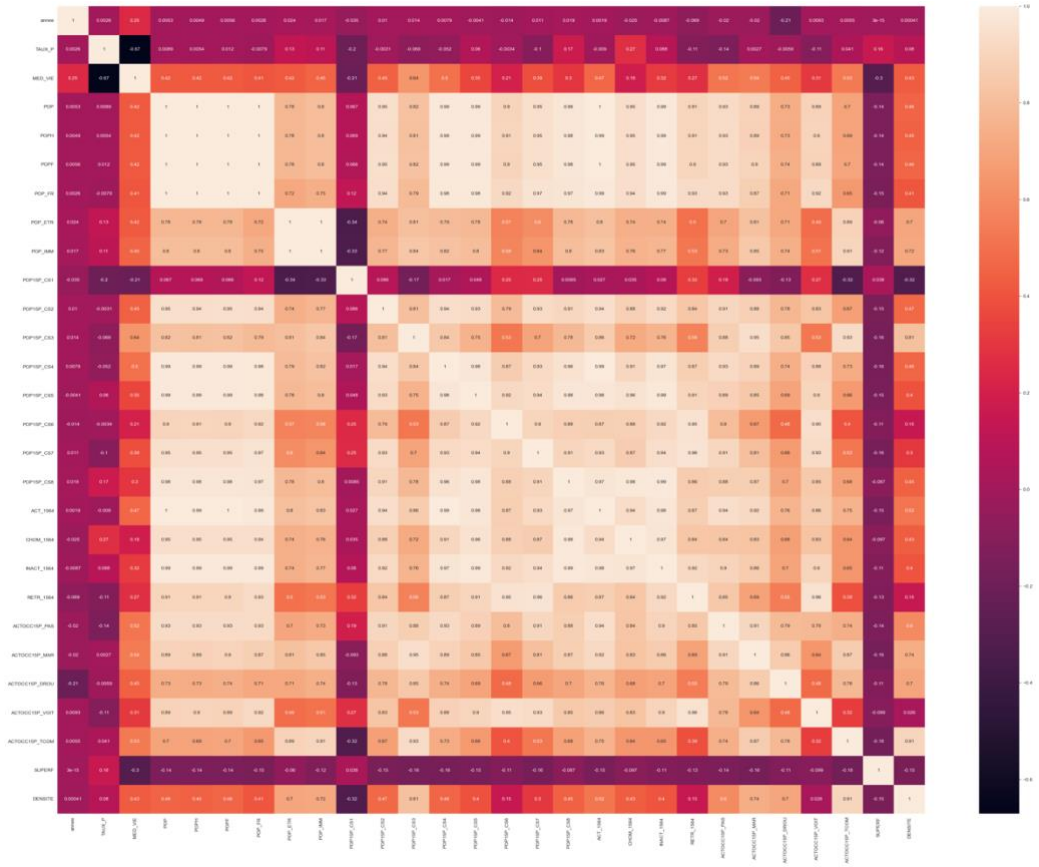
secu	0.08
Gpe_dep	0.06
col	0.05
obs	0.05
catr	0.04
catv	0.04
manv	0.03
obsm	0.03
classe age	0.02
choc	0.02
trajet	0.02
catu	0.02
place	0.02
agg	0.02
circ	0.02
situ	0.01
plan	0.01
prof	0.01
lum	0.01
sexe	0.00
surf	0.00
weekday	0.00
Week_end	0.00
atm	0.00

	statistic	pvalue
Num_Acc	-35.615838	3.160471e-277
mois	-0.264962	7.910390e-01
jour	1.634708	1.021113e-01
annee	-35.670732	4.498123e-278
heure	-11.800930	3.925853e-32
minute	-9.496402	2.188226e-21
nbv	-66.660963	0.000000e+00
age	30.678554	2.350988e-206
lat	-54.311537	0.000000e+00
long	-2.055162	3.986435e-02
occutc	1.996858	4.584152e-02
TAUX_P	-22.755203	1.623530e-114
MED_VIE	-108.492348	0.000000e+00
POP	-114.328385	0.000000e+00
POPH	-114.007131	0.000000e+00
POPF	-114.805371	0.000000e+00
POP_FR	-103.807006	0.000000e+00
POP_ETR	-138.146253	0.000000e+00
POP_IMM	-142.669409	0.000000e+00
POP15P_CS1	95.634815	0.000000e+00
POP15P_CS2	-114.853546	0.000000e+00
POP15P_CS3	-140.834903	0.000000e+00
POP15P_CS4	-121.735512	0.000000e+00
POP15P_CS5	-103.385042	0.000000e+00
POP15P_CS6	-30.464255	1.683919e-203
POP15P_CS7	-78.950321	0.000000e+00
POP15P_CS8	-108.443376	0.000000e+00
ACT_1564	-125.933310	0.000000e+00
CHOM_1564	-98.467010	0.000000e+00
INACT_1564	-98.520071	0.000000e+00
RETR_1564	-32.105600	9.578014e-226
ACTOCC15P_PAS	-116.472427	0.000000e+00
ACTOCC15P_MAR	-135.243735	0.000000e+00
ACTOCC15P_DROU	-116.352258	0.000000e+00
ACTOCC15P_VOIT	-11.112331	1.108031e-28
ACTOCC15P_TCOM	-144.708650	0.000000e+00
SUPERF	83.093207	0.000000e+00
DENSITE	-128.147669	0.000000e+00

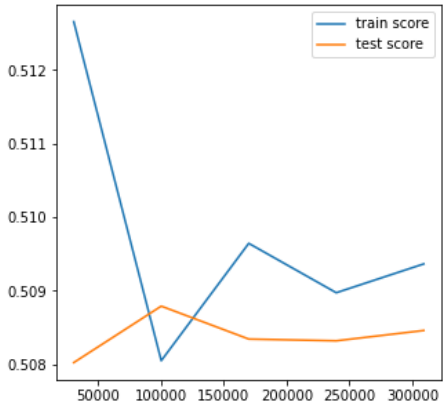
Matrice de corrélation de la base principale (variables quantitatives)

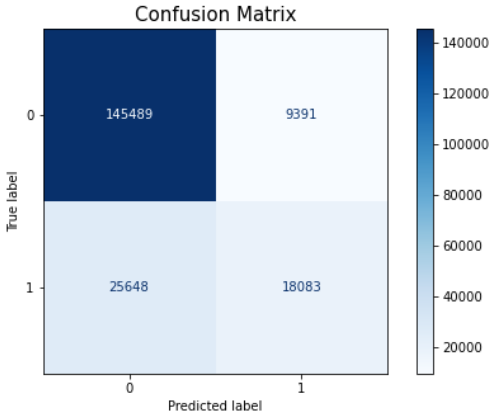


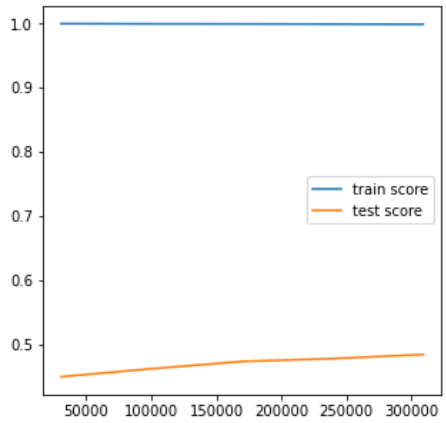
Matrice de corrélation de la base Insee supplémentaire (variables quantitatives)

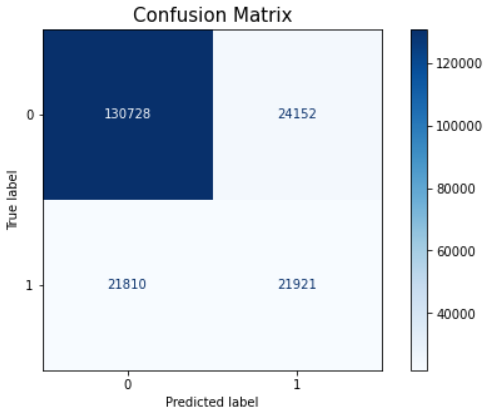


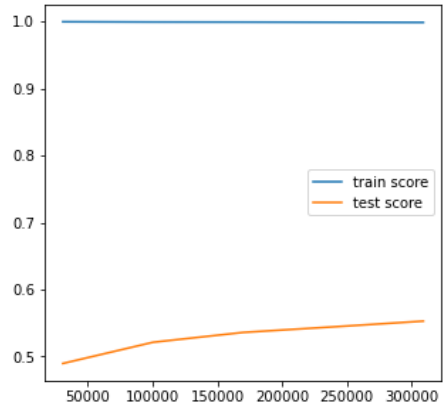
Learning curve et matrices de confusion des 4 modèles de base

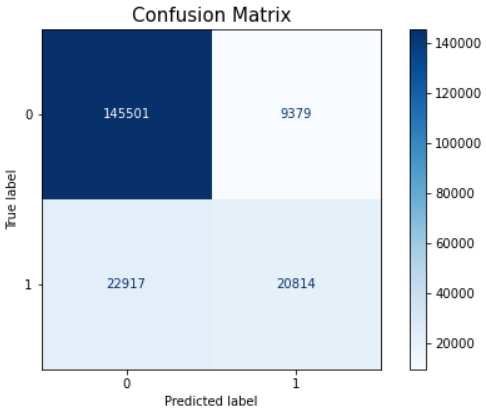


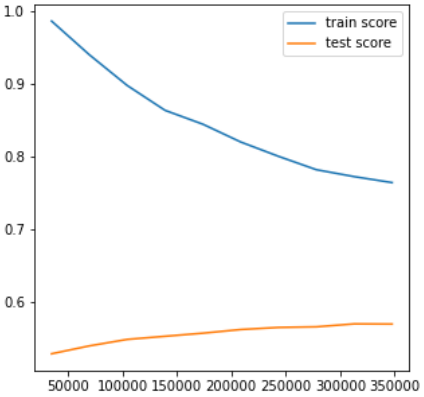


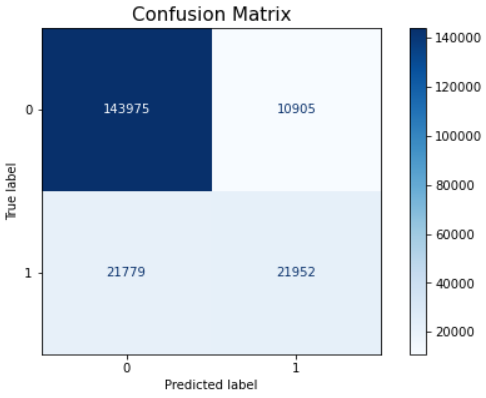












Regroupement des départements par Gpe_dep

	dep	Gpe_dep
0	01	6
1	02	7
2	03	6
3	04	7
4	05	3
5	06	2
6	07	6
7	08	8
8	09	8
9	10	3
10	11	4
11	12	8
12	13	1
13	14	4
14	15	8
15	16	8
16	17	3
17	18	6
18	19	2
19	20	3
20	21	7
21	22	4
22	23	9
23	24	9
24	25	5
25	26	5
26	27	5
27	28	7
28	29	3
29	2A	1
30	2B	1
31	30	5
32	31	1
33	32	7
34	33	0
35	34	4
36	35	1
37	36	3
38	37	1
39	38	5
40	39	9
41	40	9
42	41	6
43	42	3
44	43	9
45	44	5
46	45	2
47	46	9
48	47	8
49	48	8
50	49	1
51	50	2
52	51	2
53	52	7
54	53	9
55	54	0
56	55	7
57	56	2
58	57	5
59	58	4
60	59	6
61	60	6
62	61	8
63	62	6
64	63	7
65	64	1
66	65	4
67	66	3
68	67	1
69	68	4
70	69	0
71	70	9
72	71	6
73	72	5
74	73	7
75	74	5
76	75	0
77	76	3
78	77	2
79	78	2
80	79	8
81	80	2
82	81	9
83	82	9
84	83	5
85	84	7
86	85	8
87	86	2
88	87	1
89	88	8
90	89	6
91	90	3
92	91	0
93	92	0
94	93	0
95	94	0
96	95	0
97	97	3
98	971	4
99	972	0
100	973	2
101	974	1
102	975	4
103	976	0
104	977	6
105	978	7
106	986	5
107	987	9
108	988	4

Rappels sur les méthodes utilisées : Régression logistique, Arbre CART, Random Forest, XGBoost

Méthode de modélisation prédictive classique : régression logistique

La régression logistique est ici un modèle de régression binomiale. La variable endogène est de type binaire :

$$Y_i = \begin{cases} 1 \; si \; l'évènement \; s'est \; produit \\ 0 \; sinon \end{cases}$$

Dans notre cas, il s'agira de déterminer si l'accident est grave ou non.

Par ailleurs, la fonction de lien entre la variable à expliquer et les variables explicatives est la fonction de répartition d'une loi logistique, à savoir :

$$p_i = P(y_i = 1 \,|x_i) = \frac{1}{1 + \exp(-x_i\beta)}$$

Où :
- p_i est la probabilité que l'évènement se réalise pour un individu i
- y_i est la variable expliquée pour un individu i
- x_i correspond au vecteur des variables un individu i
- β correspond au vecteur des coefficients estimés

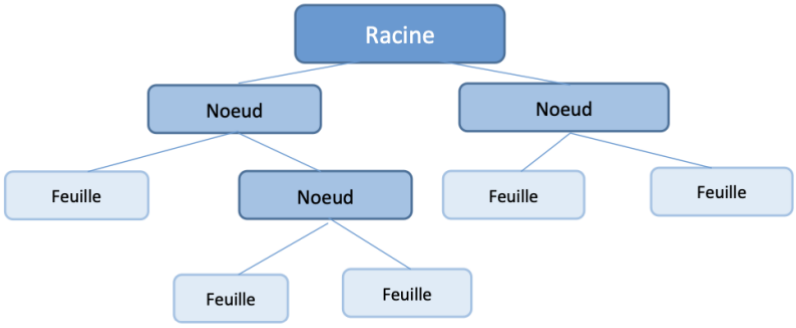
Méthodes d'ensemble de modélisation prédictive

Le principe des méthodes ensemblistes repose sur le fait qu'une agrégation de plusieurs modèles peut se révéler plus performante qu'un unique modèle complexe, ce qui se prénomme la méthode de « wisdom of the crowd ».

- Arbre CART

Les arbres de décision sont utilisés aussi bien pour expliquer une variable qualitative que quantitative. Dans le second cas, on emploiera le terme d'arbre de régression.
Ces arbres de décision vont permettre de répartir les variables explicatives en groupes homogènes selon la variable à prédire.
L'algorithme d'un arbre de décision, notamment ceux de type CART, repose sur la logique suivante :
- Au départ, nous avons une racine qui va être divisée nos individus en k classes (souvent k=2, par exemple pour CART) pour expliquer la variable de sortie. La première coupe va être obtenue en sélectionnant la variable explicative qui donne la meilleure explication de la variable à prédire (pour CART, il s'agira de l'indice de Gini). L'échantillon initial sera donc divisé en k sous- échantillons de populations. Ces sous-populations constitueront des nœuds de l'arbre. A chaque nœud on associe une mesure de proportion qui expliquera l'affiliation à cette classe
- On va ensuite réaliser une subdivision des nœuds obtenus afin de séparer les nœuds parents en nœuds enfants.

- On continue ainsi de suite jusqu'à ce que les regroupements d'individus en sous-classe soient relativement homogènes. On obtiendra alors des nœuds terminaux qui constitueront les « feuilles » de notre arbre.



- Random Forest

Les Random Forest (ou Forêts Aléatoires) reposent sur le fait d'utiliser plusieurs arbres de décision pour en faire des « forêts ». Cette méthode repose sur deux aspects, à savoir, le « Bagging » et le « Feature sampling ».

Le fondement du « Bagging », pour « Bootstrap Aggregating », introduit et popularisé par Leo Breiman (2001) consiste à générer différents échantillons aléatoires d'observations, en tirant avec remise dans l'échantillon d'origine, comme dans le cas d'un bootstrap. Chaque échantillon ainsi généré permet d'estimer un nouvel arbre de classification, créant ainsi une forêt d'arbres. La prévision va donc être amenée par l'agrégation de tous ces arbres comme la moyenne de la valeur prédite par chacun des arbres.
Le Feature sampling est similaire au Bagging en proposant non pas des observations aléatoires à chaque arbre, mais en proposant des variables explicatives tirées aléatoirement. Le Random Forest s'appuiera donc sur la combinaison du Bagging et du Feature Sampling. De cette façon, en plus de reposer sur le principe du « Bagging », lors de la construction d'un arbre de classification, à chaque branche, on procèdera également à une sélection d'un sous-ensemble de m covariables choisies aléatoirement. De cette façon, il y aura une plus forte variabilité des différents arbres puisque chaque branche d'un arbre ne s'appuiera pas sur le même ensemble de covariables. De cette façon, la forêt serait composée d'arbres moins corrélés les uns aux autres.
On pourrait penser dans ce cas que les Random Forest sont insensibles aux effets de corrélation entre les différentes variables explicatives du fait que les interactions entre les variables sont a priori prises en compte automatiquement. Cependant, comme l'expliquent B. Gregorutti, B.Michel et P.Saint-Pierre dans leur article « Corrélation et importance des variables dans les forêts aléatoires », des études numériques ont montré que les mesures d'importance des variables sont touchées par la corrélation entre les variables dans le cadre des forêts aléatoires. En particulier, Tolosi et Lengauer (2011) ont montré que les valeurs d'importance varient en fonction du nombre de variables corrélées et du niveau de corrélation.

- XGBoost

XGBoost est une abréviation pour « e**X**treme **G**radient **Boost**ing », indiquant donc qu'il s'agit d'une version « extrême » de Gradient Boosting.

Qu'est-ce-que le Gradient Boosting ?

Soient T_t un arbre obtenu à l'étape t et f_t le modèle constitué à l'étape t tel que f_i=$\sum^{t}_{j=1} T_t$. Supposons également que nos données d'apprentissage intègrent les entrées X et sorties Y tels qu'à l'étape t, le modèle renvoie $f_i(X)$ comme estimation de Y.

En ayant également en tête que la « *fonction objectif* » qu'on cherche à optimiser dans le cadre de ce type d'algorithme intègre deux composantes, à savoir une fonction de coût (ou de perte) L et un terme de régularisation Ω :

$$Obj(.) = L(.) + \Omega(.)$$

Dans ce cas, on prendra comme fonction de coût la fonction des moindres carrés définie de la façon suivante :

$$L = \sum_{i=1}^{m} l(Y_i, f(X_i))$$

Avec X_i les données de la ligne i de la matrice X et $j(a_1, a_2) = \frac{(a_2 - a_1)^2}{2}$

On considèrera dans ce cas le gradient de L par rapport à $f(X_i)$:

$$\frac{\partial L}{\partial f(X_i)} = \frac{\partial \sum_{c=1}^{m} l(Y_c, f(X_c))}{\partial f(X_i)} = f(X_i) - Y_i$$

Si l'on considère T_t l'arbre de décision créé à l'étape t, tel qu'à cette étape, la prédiction du modèle :

$$\hat{Y}^t = \hat{Y}^{t-1} + \eta T\ (X).$$

Ce paramètre η va correspondre au pas du gradient et peut varier pour diminuer l'effet de sur-ajustement dans le cas où le nombre d'arbres est élevé.
Le but ici va être à chaque étape de minimiser la fonction objectif $Obj(\theta)$ en ajoutant un nouvel arbre.
En considérant la fonction objectif $j(\theta) = L(\theta) + \Omega(\theta)$, en prenant l'erreur quadratique moyenne (MSE) dans son cas comme fonction de perte L (dans notre cas, la fonction L prend 0 si $y = f(x)$, sinon 1), en développant les différents facteurs et en utilisant un développement de Taylor à l'ordre 2, la fonction objectif à l'étape t s'écrira :

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i T_t(X_i) + \frac{1}{2} h_i {T_t}^2(X_i) \right] + \Omega(T_t)$$

Avec : $\begin{cases} g_i = \frac{\partial l(\hat{Y}_i^{(t-1)}, Y_i)}{\partial f(X_i)} \\ h_i = \frac{\partial^2 l(\hat{Y}_i^{(t-1)}, Y_i)}{\partial f(X_i)} \end{cases}$

De cette façon, on aura une baisse de la fonction objectif dans le cas de l'ajout d'un nœud à une feuille de l'arbre sous la forme suivante :

$$\frac{1}{2}\left[\frac{G^2{}_L}{H_L + \lambda} + \frac{G^2{}_R}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_L + \lambda}\right] - \gamma$$

Où :

- $\frac{G^2{}_L}{H_L+\lambda}$ $(resp.$ $\frac{G^2{}_R}{H_R+\lambda}$) le score sur la partie gauche (resp. droite) de la séparation
- $\frac{(G_L+G_R)^2}{H_L+H_L+\lambda}$ le score sur la feuille originale de l'arbre avant séparation
- γ le surplus de complexité du modèle induit par la séparation.