100% ACTUAIRES & 100% DATA SCIENCE

INSTITUT DES ACTUAIRES

Apport des données télématiques dans la connaissance du risque

Kent Aquereburu

Marc Juillard

Yolan Honoré Rougé

16 Novembre 2018

Hôtel Marriott Rive Gauche

Paris 14ème

Telematic: est une méthode d'administration d'un véhicule en utilisant le GPS et des diagnostics embarqués pour enregistrer les mouvements sur une carte informatisée. Grâce à cela, de nombreux services peuvent être proposés

- Dans le cadre de son programme smart car, Société Générale Insurance a lancé un programme de véhicules connectés en Italie qui lui à permis d'acquérir une base de données conséquente (>200 Giga). Sur la base de cet historique, les équipes du Data Lab ont mis en place un score télématique permettant de faire le lien entre comportement de conduite et sinistralité.

- Les différents travaux réalisés lors de cette étude sont repris ci-après :

  - ❑ Prétraitement et enrichissement des données

  - ❑ Application d'un algorithme de transparence permettant d'expliquer les résultats de l'algorithme pour chaque trajet, mais permettant également d'extraire la part « durée de conduite » afin d'obtenir un modèle purement comportemental

  - ❑ Utilisation du modèle pour construire le score de conduite à la maille journée, analyse du sa stabilité à la maille mensuelle et intégration dans un modèle actuariel classique.

- Ces différents ces éléments sont repris ci-après.

- **Il convient de garder en mémoire qu'au-delà de la mise en place d'un modèle transparent, de nombreux autres chantiers sont à mettre en place dans le cadre de la télématique : plateforme de traitement en temps réel, mode de collecte, portabilité des algorithmes et structuration d'une offre.**

Données brutes de l'étude …

7,4 millions de journées	**54 000** véhicules	**2566** sinistres	**636 millions** de km	**185 millions** d'évènements	**Entraînement du modèle** (*Données 2017*)
3,7 millions de journées	**55 000** véhicules	**1221** sinistres	**294 millions** de km	**95 millions** d'évènements	**Validation du modèle** (*Données 2018*)

Un événement c'est…

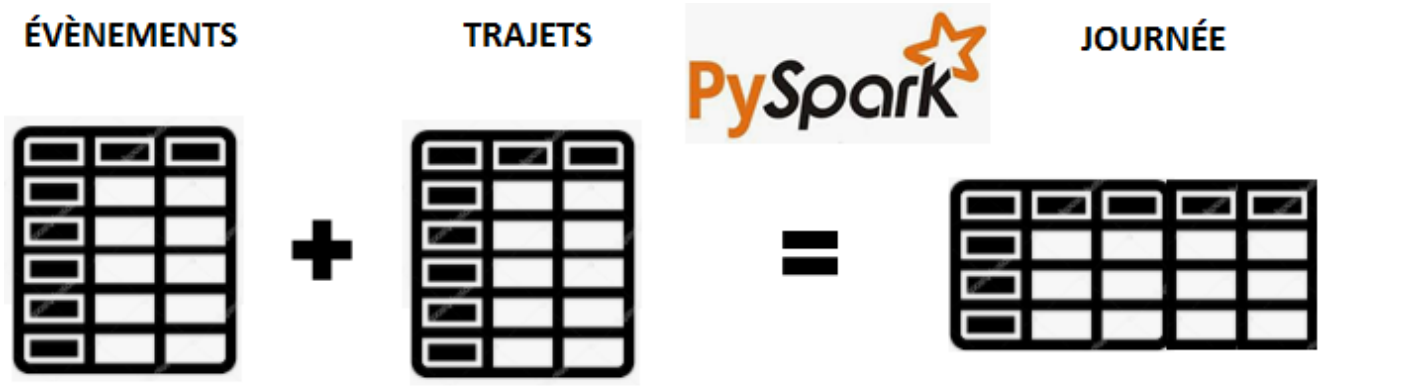
Virage / Freinage / Accélération
=
« Une variation de vitesse »
X
« Une durée »
X
« Un angle »

Enrichissement effectué

- Contextualisation
- *Proxy* de fluidité du traffic
- Test de la météo

Prétraitements des données : gestion de la volumétrie

- En 2018, les flux de données représentent 10 Go de données hebdomadaires sous forme de points GPS.
- Afin de pouvoir manipuler et modéliser cette volumétrie de données, il faut se ramener à une maille de calcul plus agrégée. Pour effectuer l'agrégation, Spark est utilisé.
- La maille de calcul retenue est la journée (à la maille « trajet » il y a trop peu d'événements et la notion de « trajet » est mal définie : la maille « mensuelle » est trop agrégée pour obtenir des résultats comportementaux fins).



Prétraitements des données : enrichissement

Données contextuelles

- Le type de route (autoroute, nationale, ville) est rajoutée à la maille point GPS
- Cette valeur est calculée par un prestataire externe, et contrôlée avec un algorithme de *map matching*
- Cette information a un fort impact sur le niveau de risque

Données météorologiques

- Des informations météorologiques sont rattachées à la maille point GPS (pluie, vent, neige, brouillard)
- La qualité des données est assurée par recoupement de plusieurs sources. L'accès à un historique est payant.
- Les résultats n'ont pas mis en évidence d'impact marqué. Cela est du à la dilution d'information (pas de sinistres observés sous la neige dans la base par exemple)
- Leur utilisation n'est pas retenue dans le cadre du score et fera l'objet d'une étude plus détaillée ultérieure.

Données de trafic

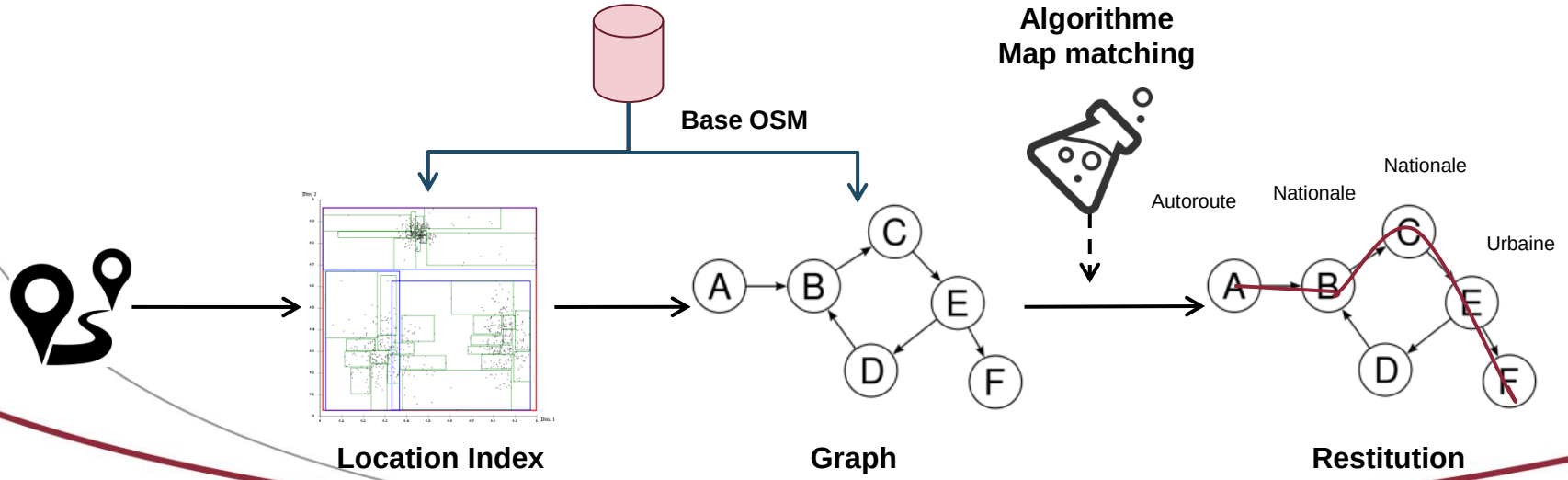
- Les données de trafic n'ont pas été historisées au moment de la collecte des points GPS et aucun prestataire ne fournit d'historique
- Il est supposé que l'heure de la journée et le jour de la semaine captent partiellement la notion de trafic. Un *proxy* est également utilisé : le trafic est considéré comme dense lorsque l'individu roule à moins de 30% de la vitesse limite.
- Ceci est corroboré par les résultats qui mettent en évidence une augmentation du risque pour les individus roulant très en dessous de la vitesse limite, notamment aux horaires de bureau (7-10h et 17-20h)

Qualité des données et loyauté de l'algorithme

Dans une optique de restitution à l'assuré ou d'usage tarifaire, il est essentiel de s'assurer que l'algorithme est « loyal », i.e. que **les traitements appliqués sont conformes à la réalité annoncée**.

A ce titre, l'application d'une contextualisation lors de la construction d'un score impactant l'offre d'assurance nécessite de s'assurer que cette dernière est exacte. Dans le cadre de l'étude, une contextualisation basée sur la route (effectuée par le fournisseur du boîtier) présente un impact fort sur le résultat. Il est donc nécessaire de pouvoir mesurer la pertinence de cet enrichissement. Cette vérification est effectuée avec un algorithme de *map matching* appliqué sur la base *Open street map*.

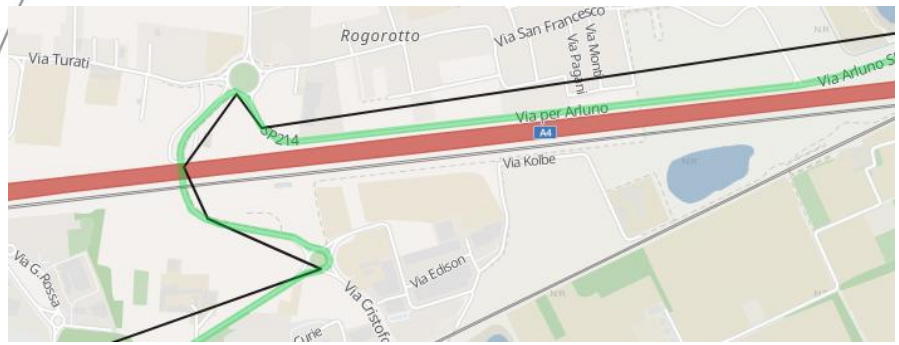
A noter : *L'usage de données météorologiques n'a pas été retenu pour cette version du score, mais cela nécessiterait également un contrôle avancé : il n'est pas acceptable pour l'assuré de voir son score se dégrader car l'algorithme considère qu'il a plu pendant son trajet alors que ce n'est pas le cas)*



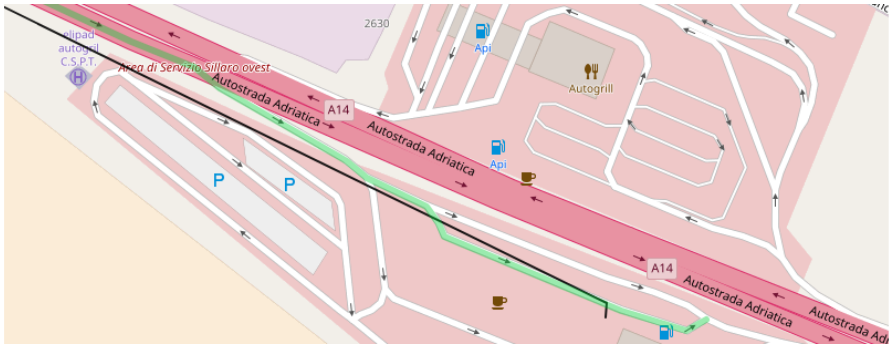
Prétraitements des données : gestion de la qualité des données

Application :

- ❑ **Le traçage en noir lie les points GPS remontés par le boitier**
- ❑ Le traçage vert représente le trajet reconstruit par l'API du Data Lab



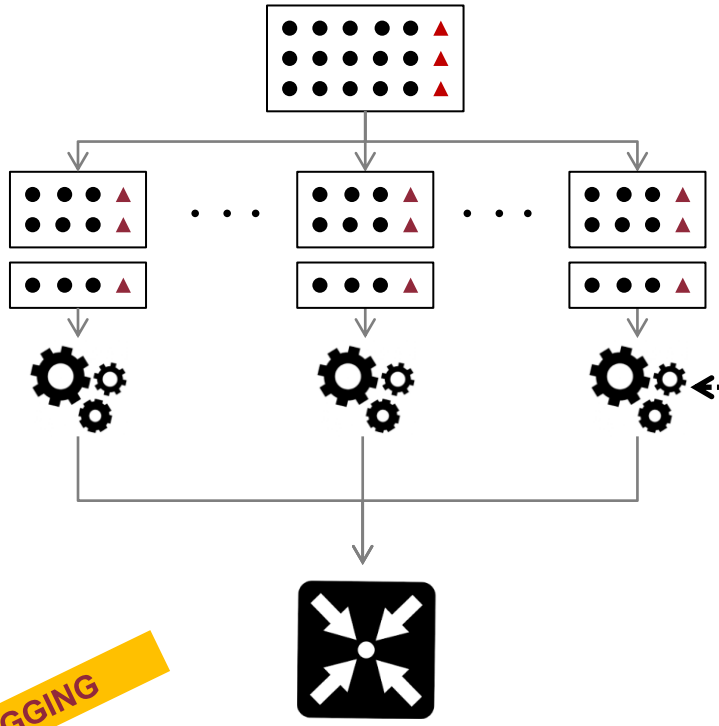
Le prestataire considère à tort que le trajet est effectué sur l'autoroute quand le trajet longe l'autoroute



Permet de détecter les arrêts au stations service (et donc de raccorder des sous-trajets à un trajet)

Permet d'affiner la connaissance du trajet (ici les échangeurs autoroutiers avec virage dangereux).

Prétraitements des données : gestion de la base déséquilibrée



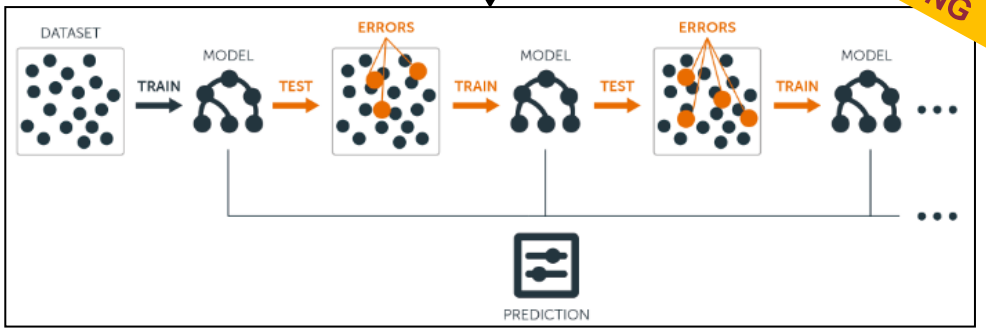
Données initiales :
- 2 566 sinistres
- 7 400 000 journées

SAMPLING

100 couples de bases (apprentissages / te
- (1 600 / 700) sinistres
- (160 000 / 70 000) journées

Un modèle par base

BOOSTING



BAGGING

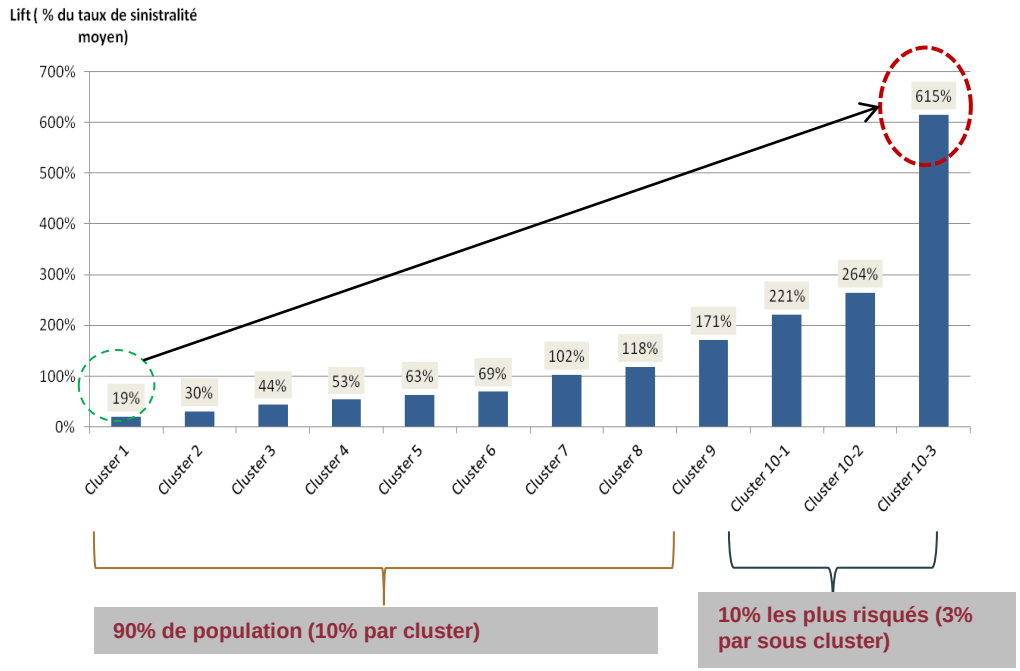
« Méta-modèle » qui agrège les prédictions

Modèle prédictif complet : résultats sur la base de test

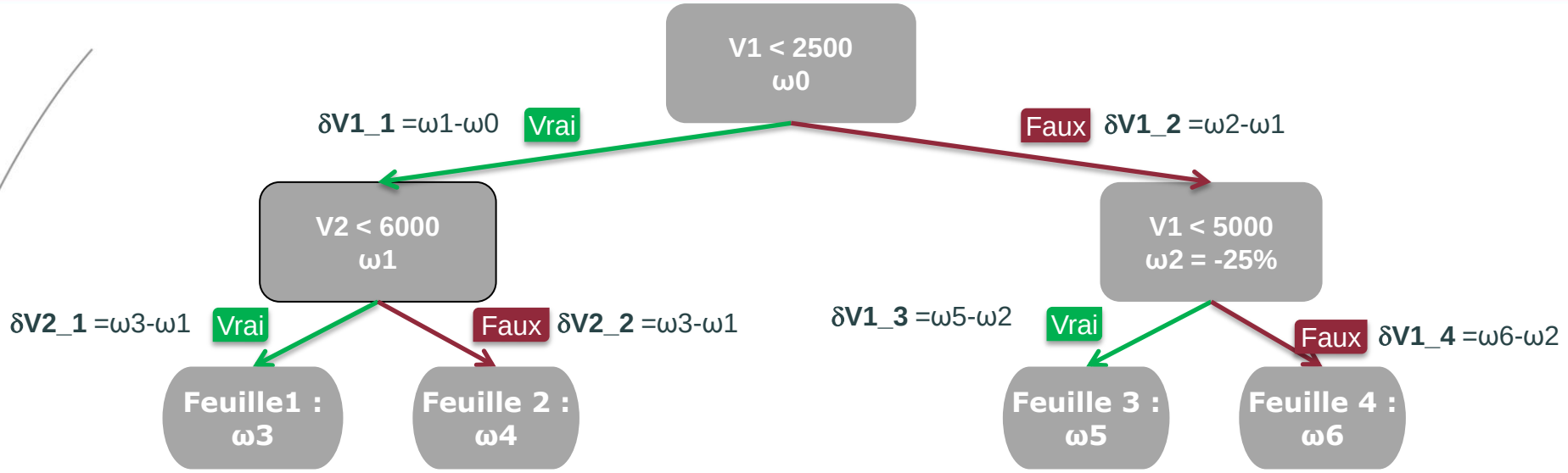
En ne se basant que sur des variables liées aux trajets effectués (le modèle n'intégrant pas de données propres au véhicule ni au conducteur), le modèle conduit à un S/P de 106% sur la base de test.

- On observe une évolution exponentielle du taux de sinistralité par cluster de risque (le cluster 10 étant le plus risqué).
- Le cluster identifié comme étant le moins risqué présente une sinistralité **5 fois moins importante que la sinistralité moyenne.**
- Le cluster identifié comme le plus risqué par le modèle présente une sinistralité **6 fois plus importante que la sinistralité moyenne.**

La capacité de classement du modèle valide sa pertinence pour la construction d'un score lié à la sinistralité



TRANSPARENCE DE L'ALGORITHME : DESCRIPTIF METHODLOGIQUE



On définit **l'impact d'une variable pour l'individu comme la somme des variations pour chaque nœud traversé.**

Individu	Feuille	Impact_V1	Impact_V2
(V1 = 1000, V2 = 5000)	1	**δV1_1**	**δV2_1**
(V1 = 3750, V2 = 8000)	3	**δV1_2**+**δV1_3**	0

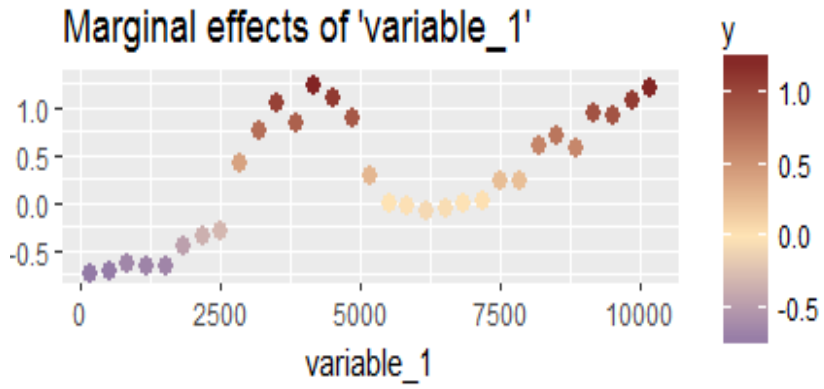
- ω est la valeur de la log-côte si le nœud était une feuille terminale.
- On attribue la variation de côte à la variable sur laquelle on fait le test logique dans chaque nœud.
- On réitère cette opération pour tous les arbres.

TRANSPARENCE DE L'ALGORITHME : DU GLOBAL AU LOCAL

Individu	Impact_V1	Impact_V2
(V1 = 2700, V2 = 8500)	**δV1_1**	**δV2_1**
(V1 = 3500, V2 = 100)	**δV1_2**+**δV1_3**	0

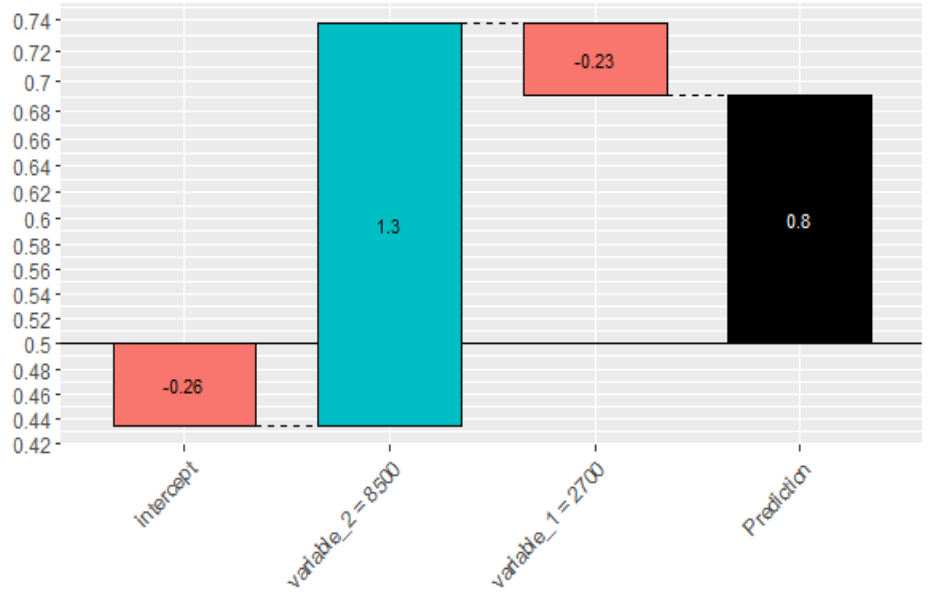
Interprétation « GLOBALE »

Il est possible de tracer l'impact moyen en fonction des valeurs prises par chaque variable. Cela permet de visualiser les décisions prises par le modèle.

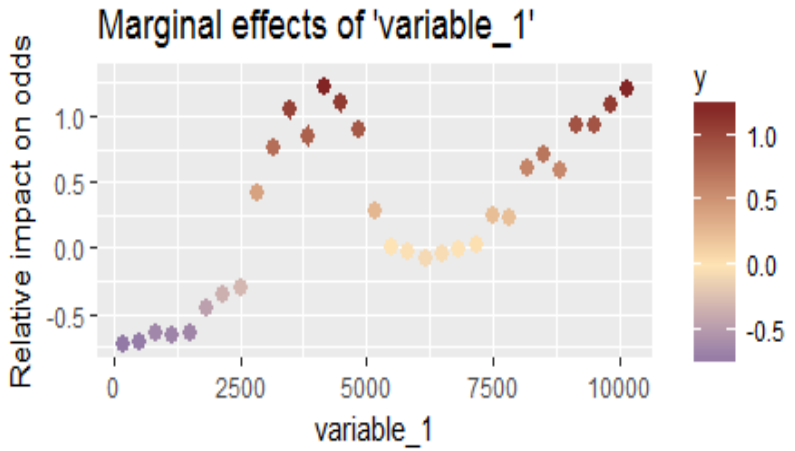


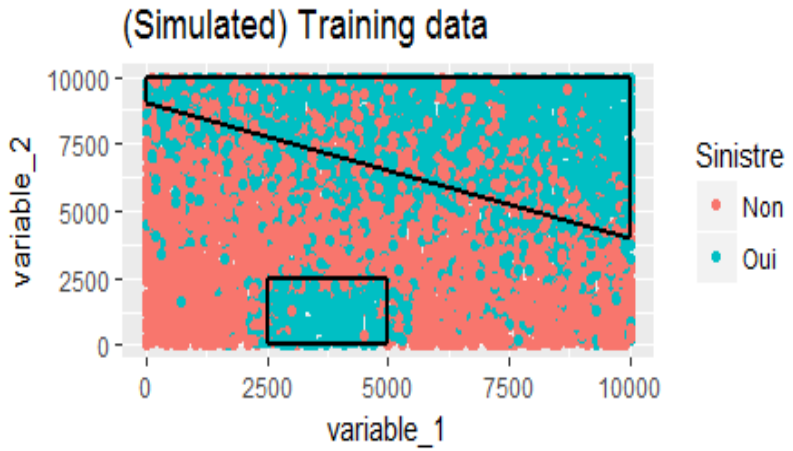
Explicabilité « LOCALE »

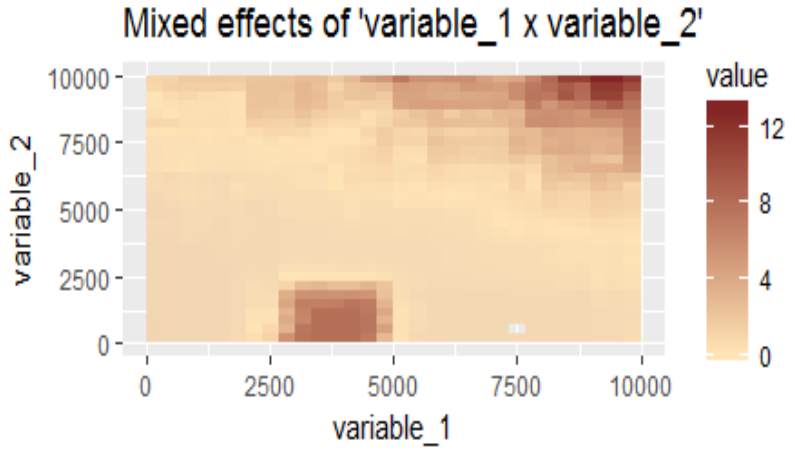
Il est possible de décomposer de façon exacte la contribution de chaque variable à la prédiction d'un individu pour expliquer la prédiction du modèle

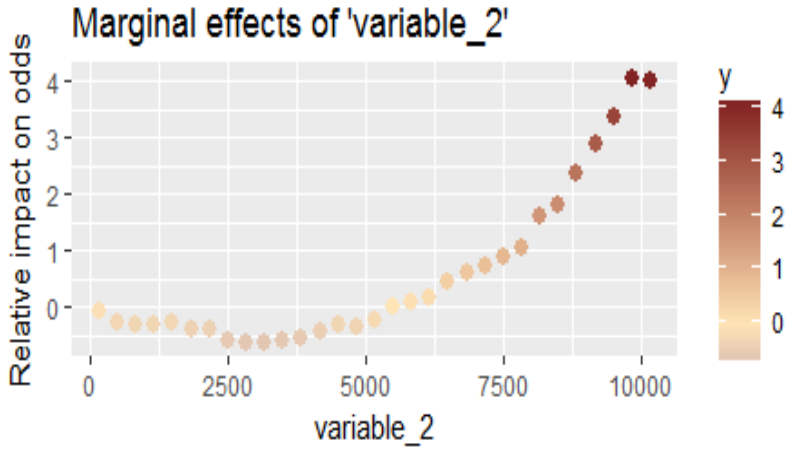


TRANSPARENCE DE L'ALGORITHME : EXEMPLE FICTIF DE TRANSPARENCE GLOBALE

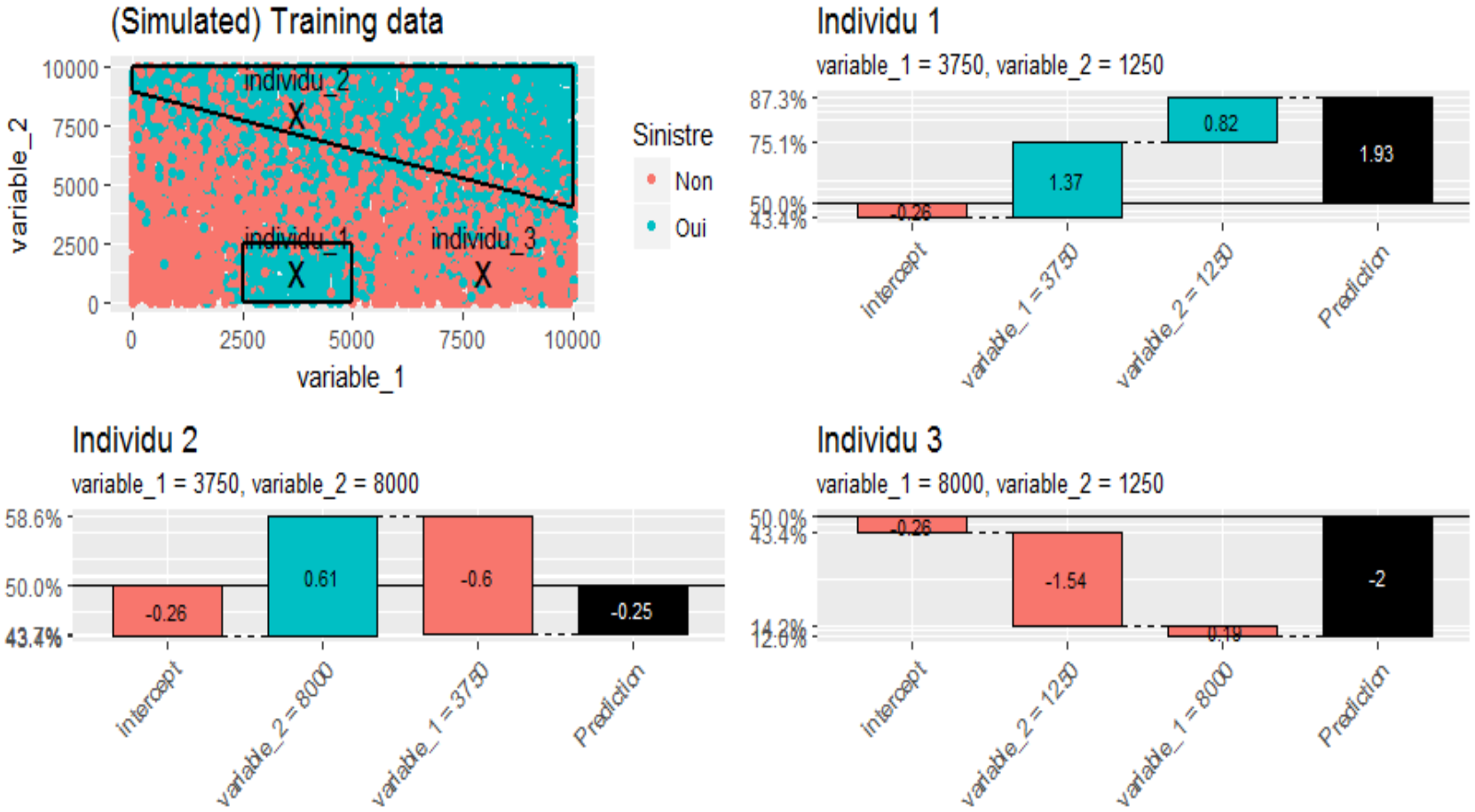








TRANSPARENCE DE L'ALGORITHME : EXEMPLE FICTIF DE TRANSPARENCE LOCALE



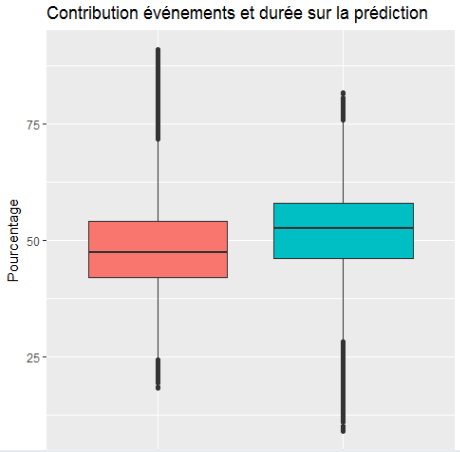
Transparence aspect pratique : global versus local

La mise en place d'un score purement comportemental nécessite de pouvoir extraire la part de la prédiction liée à **la sinistralité comportementale (*)**. Ceci peut conduire à changer fortement le score d'un individu.

Individu	Score xgb	Durée du trajet	Nombre événements	Score comportemental
N°1	10	30 min	21	20
N°2	10	5h 30 min	2	90

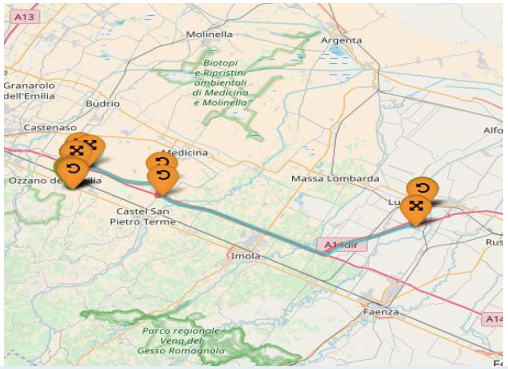
Cette procédure est ensuite appliquée sur 100% de la base :

- Les variables comportementales expliquent dans la majorité des cas **plus de 40% des prédictions faites par le modèle**.
- **Seule la sinistralité comportementale prédite par le modèle est conservée dans la suite de l'étude (**la durée n'est pas retenue dans la construction du score)



(*) Sinistralité « comportementale » : *sinistralité prédite par le modèle sans prise en compte de la durée. Cette part de la sinistralité est approximée à dire du modèle Xgboost (cf annexes)*

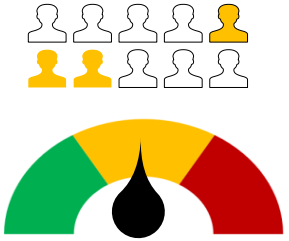
Construction du SCORE : 3 principaux styles de conduite détectés par le modèle

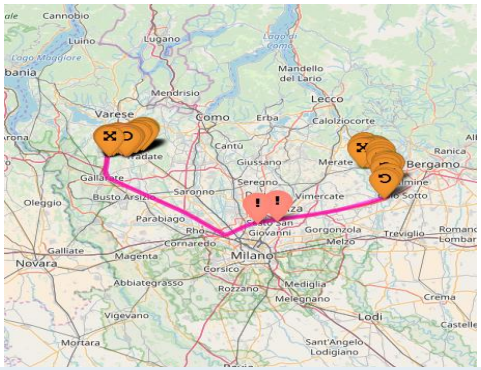


Distance totale : 90 Km
Nombre d'événements : 16

Quasiment aucun événement comportemental

Lift sinistralité « comportementale » :[65% ;81%]

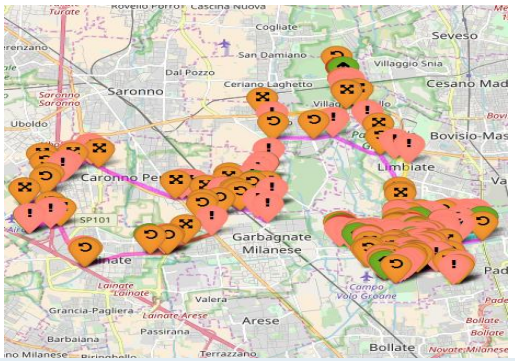




Distance totale : 195 Km
Nombre d'événements : 90

Taux d'événement dans la moyenne avec une légère surreprésentation des freinages et mouvements latéral

Lift sinistralité « comportementale » : [87% ;106%]



Distance totale : 98 Km
Nombre d'événements : 284

Taux d'événement **~2 fois supérieur à la moyenne avec une surreprésentation de tous les événements**.

Lift sinistralité « comportementale » : >119%

Construction du SCORE *: intelligibilité du modèle*

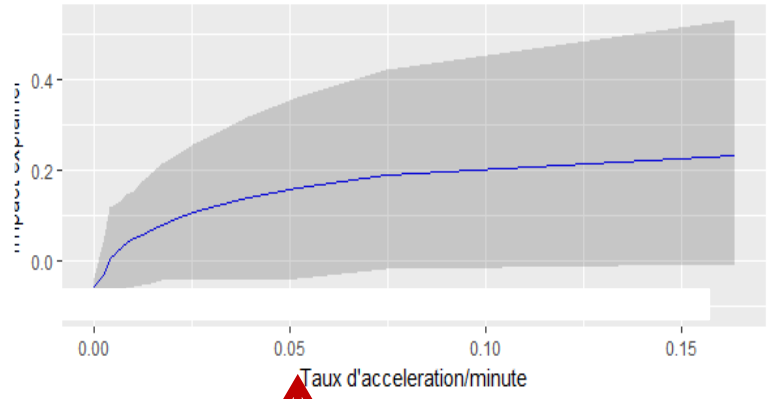
Une fois le modèle prédictif mis en place l'étape suivante consiste à transformer les prédictions en sous-scores communicables à un client.

Si en 1ère approche, le score peut se construire en se basant sur la fonction de répartition empirique du modèle, cette technique peut conduire à un score peu compréhensible par le client. En effet :

- Le modèle prédictif est fortement contextualisé (45 variables)
- La simplicité de communication du score, va conduire à retenir un nombre restreint de sous-score sans prise en compte réelle du contexte (en terme d'affichage client).

Ces 2 points peuvent donner naissance à des incompréhensions :

- **2 styles de conduite proches** (au sens de la métrique de communication du score) …
- **peuvent avoir 2 scores différents** (la différence étant due à la prise en compte du contexte global de la journée par le modèle prédictif).

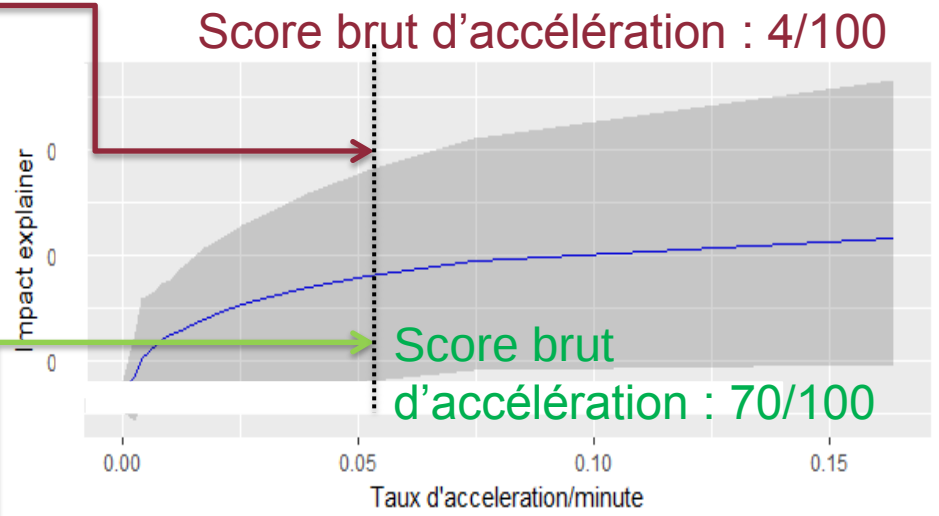


Impact moyen sur la probabilité de sinistre prédite par le modèle et intervalles de confiance 95%

Construction du SCORE *: intelligibilité du modèle*

Une fois le modèle prédictif mis en place l'étape suivante consiste à transformer les impacts bruts du modèle en sous-scores communicables à un client.

Taux d'accélération par minute en fonction du contexte (%)	Individu avec un « bon score brut »	Individu avec un « mauvais score brut »
Taux global	**~5%**	
Ville - Journée	0%	1,4%
Ville - Horaires de bureau	0%	3%
Ville - Nuit	0%	0
Autoroute - Journée	1 %	0,6%
Autoroute - Horaires de bureau	0%	0%
Autoroute - Nuit	0%	0%
Autres - Journée	4%	0%
Autres - Horaires de bureau	0%	0%
Autres - Nuit	0	0%

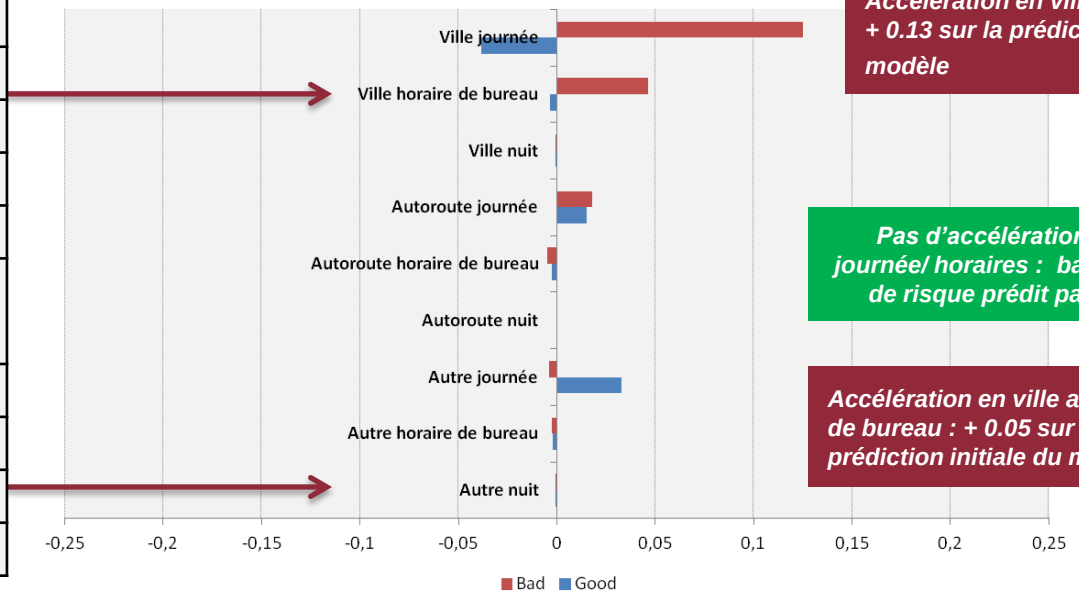


Pour un taux d'accélération global identique, une forte **variance** entre les impacts sur la prédiction du modèle est observée.
Une étape de « lissage » des scores est appliquée

Construction du SCORE *: intelligibilité du modèle*

Une fois le modèle prédictif mis en place l'étape suivante consiste à transformer les prédictions en sous-scores communicables à un client.

Taux d'accélération par minute en fonction du contexte (%)	Individu avec un « bon score brut »	Individu avec un « mauvais score brut »
Taux global	**~5%**	
Ville - Journée	0%	1,4%
Ville - Horaires de bureau	0%	3%
Ville - Nuit	0%	0
Autoroute - Journée	1 %	0,6%
Autoroute - Horaires de bureau	0%	0%
Autoroute - Nuit	0%	0%
Autres - Journée	4%	0%
Autres - Horaires de bureau	0%	0%
Autres - Nuit	0	0%



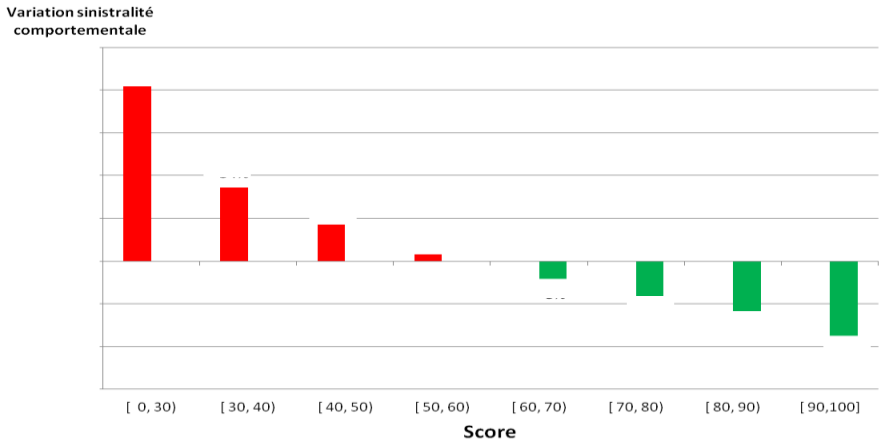
Accélération en ville en journée : + 0.13 sur la prédiction initiale du modèle

Pas d'accélération en ville en journée/ horaires : baisse du niveau de risque prédit par le modèke

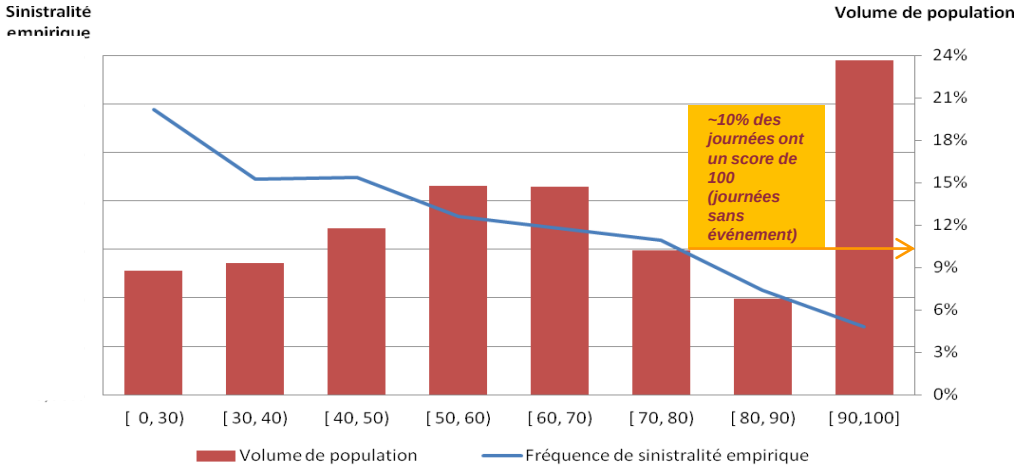
Accélération en ville aux horaires de bureau : + 0.05 sur la prédiction initiale du modèle

Score à la maille journée : lien entre score et sinistralité totale

1 – Score journalier et sinistralité comportementale



2 –Score journalier et sinistralité totale

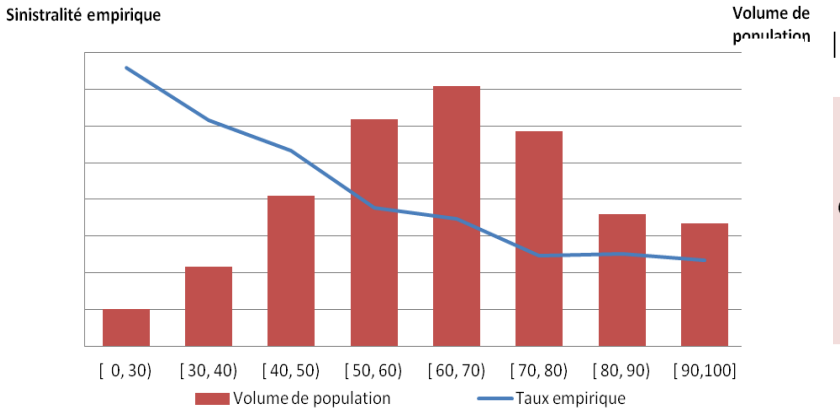


Bien que n'intégrant aucune variable liée à la durée, au véhicule ou au conducteur, le score présente une corrélation négative avec la **sinistralité empirique globale**. **Les variables comportementales expliquent une part suffisante de la sinistralité globale pour effectuer un « classement des risques »**

Score à la maille mensuelle : lien entre score et sinistralité totale

Le score mensuel est déterminé en agrégeant les scores journaliers du mois par la distance journalière :

1 – Score mensuel et sinistralité totale



2 – Score mensuel et stabilité temporelle

Cluster 2017	Répartition en 2018 [0, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100]	
[90 ,100]	0,0%	0,0%	0,0%	0,2%	0,7%	1,7%	11,3%	86,1%	100%
[80, 90)	0,0%	0,1%	0,6%	2,1%	5,7%	26,8%	53,9%	10,8%	100%
[70, 80)	0,1%	0,5%	2,1%	7,2%	28,2%	46,5%	14,2%	1,3%	100%
[60, 70)	0,2%	1,4%	6,8%	26,6%	44,5%	17,3%	3,1%	0,3%	100%
[50, 60)	0,6%	5,5%	24,3%	44,2%	19,5%	4,9%	1,0%	0,1%	100%
[40, 50)	4,1%	23,0%	43,5%	20,5%	6,3%	1,9%	0,6%	0,1%	100%
[30, 40)	22,8%	48,3%	19,5%	6,3%	2,1%	0,9%	0,2%	0,1%	100%
[0, 30)	72,7%	20,2%	4,5%	1,9%	0,2%	0,2%	0,0%	0,2%	100%

➢La corrélation négative entre le score et la sinistralité est maintenue à la maille mensuelle.

➢L'agrégation mensuelle permet d'obtenir une distribution des scores plus recentrée qu'à la maille journée. En particulier, la proportion de bons scores ([90,100]) est de ~ 10% à la maille mensuelle contre ~24% à la maille journée.

➢Le score calculé en période (t-1) peut donc être utilisé pour prédire le comportement de conduite à t.

➢Le score comportemental pourrait ainsi servir de bonus malus dans une perspective d'offre tarifaire

QUESTIONS / REPONSES