Comment réduire le nombre d’accidents de la route dans une ville et les frais des assurances liés à ces accidents ?

Les dommages économiques et sociétaux des accidents de la route.

Le prix le plus élevé que nous payons pour les accidents de la route est la perte de vies humaines, mais la société supporte également le poids des nombreux coûts liés à ces accidents.

Selon une étude publiée en mai 2014 par l’Administration nationale de la sécurité routière (NHTSA), les accidents de la route aux États-Unis en 2010 ont causé la mort de 32 999 personnes, blessé plus de 3,9 millions d'individus et endommagé plus de 24 millions de véhicules.

Ces pertes de productivité et humaines représentent un coût total de plus de 871 milliards de dollars avec près de 277 milliards de coûts économiques et 594 milliards de coûts humains et sociétaux (dommages causés par les pertes de vie, la douleur ou la diminution de la qualité de vie suite aux blessures).

Ces coûts pharaoniques sont pour la grande majorité pris en charge par les assurances privées, qui paient plus de 50% des coûts totaux aux Etats-Unis. Les assurances ont donc intérêt à réduire au maximum le nombre d’accidents dans une ville pour réduire ces frais et ces dépenses.

Comment réduire le nombre d’accidents ?

Une évolution récente sur le marché de l'assurance est celle des "contrats de télémétrie", qui incitent les conducteurs à adopter un comportement de conduite conforme au code de la route. Les voitures des entrepreneurs sont équipées d'un petit matériel qui surveille le comportement de conduite. Si, par exemple, la limite de vitesse n'est pas dépassée, le contrat de télémétrie est 10 % moins cher pour le conducteur car le risque d'accident pour la compagnie d'assurance est réduit. Ce contrat permet à la fois de réduire les frais de l’assurance en réduisant le nombre d’accidents de la route, tout en améliorant l’expérience du conducteur.

De la même manière, nous proposons un contrat de télémétrie basé sur le risque statistique d'accident à l'endroit où se trouve la voiture et sur les conditions environnementales. Le concept principal consiste à déterminer, à partir des rapports de police sur les accidents, les zones de la ville où le risque d'accidents est le plus élevé. Le conducteur

sera alors averti avant de pénétrer dans une telle zone à risque par un petit dispositif monté dans sa voiture. Ce dispositif a pour but d'augmenter la vigilance et le niveau d'attention du conducteur dans ces zones de danger pour réduire la probabilité d'avoir un accident. Réduire le nombre d'accidents permet à la fois d'éviter des coûts pour les compagnies d'assurance et permet à l'assureur d'être plus compétitif en proposant des contrats plus variés. Les assureurs proposent le dispositif matériel dans un contrat spécial qui comporte de meilleures conditions qu'un contrat habituel et le conducteur a moins d'accidents ainsi qu'un contrat moins cher.

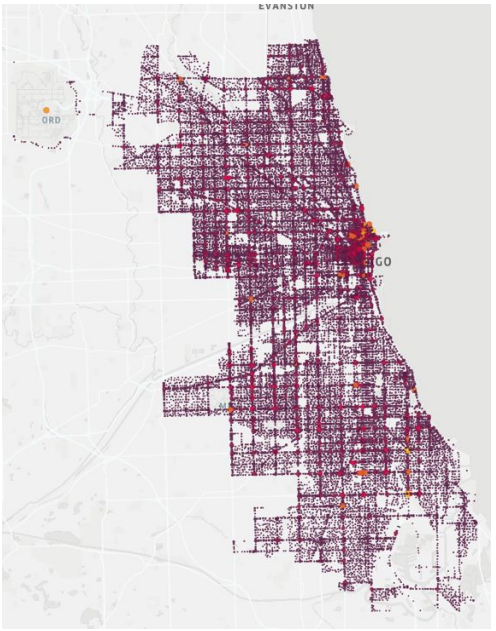
Facteurs clés de succès

Lors de l'étude des facteurs pouvant influencer les accidents de la route, nous avons identifié deux catégories d'influences : le comportement humain (niveau d'attention, fatigue, intoxication, humeur) et les facteurs environnementaux (état de la route, temps, lieu, conditions météorologiques). Si les données sur les facteurs environnementaux sont facilement accessibles et faciles à intégrer, celles sur le comportement humain au volant sont imprévisibles et ont des implications en matière de confidentialité des données. Ces facteurs liés au comportement humain peuvent d'ailleurs biaiser les résultats obtenus. Par conséquent, pour que le modèle soit performant, le risque d'accidents établi doit être lié aux conditions externes.

De plus, pour que le modèle soit efficace, il faut réduire autant que possible les faux positifs, même au prix de la précision. En effet, un système fiable qui n'avertit que dans 80% des cas une zone dangereuse a plus d'impact qu'un système qui détecte toutes les zones dangereuses et aussi d'innombrables zones non dangereuses. Dans le second cas, l'utilisateur s'enervera rapidement, et ne respectera plus les avertissements de notre module, qui pourra alors avoir un effet accidentogène. Il faut impérativement que les zones avertis par notre module soient véritablement dangereuses.

Quelles données avons-nous utilisé ?

Les données que nous utilisons sont des rapports d'accidents de la police de Chicago. L'ensemble des données contient près de 450 000 accidents étalés de 2015 à 2020 et caractérisés par une vingtaine de variables. Parmi ces variables, nous retrouvons la localisation, la date, l'heure, la météo, la visibilité, l'état de la route, la disposition de la route, l'état des panneaux de signalisation, la limitation de vitesse, les dégâts et le nombre de blessés.



Ces données sont fournies par la ville de Chicago elle-même, sont en libre accès et sont constamment mise à jour. Couplées aux données du trafic routier de Chicago, ces données nous permettent de déterminer la proportion d'accidents par zone de la ville en fonction des différents facteurs extérieurs et donc d'établir des zones à risque dans la ville.

Traitement des données

Si les données fournies par la ville de Chicago sont de très bonnes qualités, deux problèmes persistent.

Tout d'abord, on peut constater que les accidents en conditions idéales, c'est-à-dire se déroulant le jour, sous temps sec et route sèche, sont largement majoritaires dans notre base de données. Ainsi, les accidents causés par des facteurs extérieurs, sous temps de pluie ou la nuit, sont sous-représentés. Si la répartition des caractéristiques des accidents n'est pas équilibrée, nos algorithmes de Machine Learning ne pourront tirer aucune conclusion particulière : ils considèreront que les conditions en temps idéales étant majoritaires, elles sont responsables des accidents et on ne pourra pas étudier l'influence réelle des facteurs extérieurs.

Pour résoudre ce problème, nous avons essayé d'équilibrer la répartition des données en créant des données virtuelles avec la technique SMOTE-NC. Cette technique de sur-échantillonnage crée de nouvelles variables synthétiques de la catégorie minoritaire en modifiant légèrement des valeurs existantes. Nous avons alors pu créer des accidents virtuels se déroulant sur route mouillée, sous pluie, neige, brouillard, ou de nuit. Avec cet algorithme, nous sommes passés d'un ratio d'un accident de nuit pour 5 de jour à un ratio de 1 pour 2. Nous avons obtenu des ratios similaires pour les paramètres pluie/temps clair ou route sèche/route mouillée.

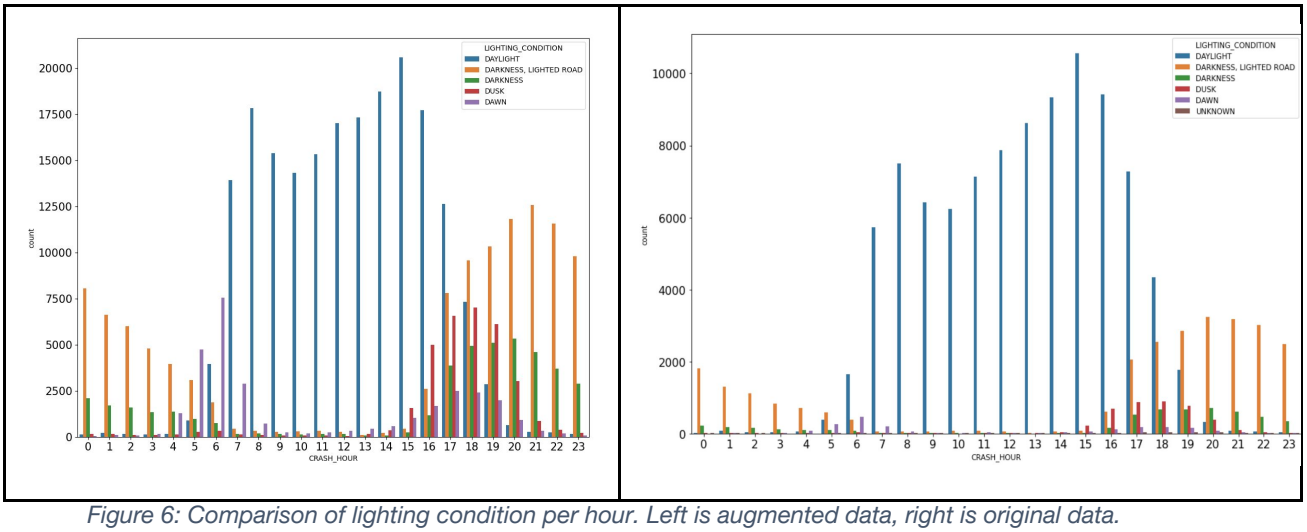


Figure 6: Comparison of lighting condition per hour. Left is augmented data, right is original data.

Le deuxième problème a été que nous ne disposions pas de données "non-accidents" nécessaire pour tout algorithme de classification. Nous avons alors dû utiliser un dataset donnant le trafic routier journalier dans les différentes rues de Chicago. Cela nous a permis de créer des non-accidents en gardant les répartitions des différentes variables de chaques paramètres extérieurs mais en les mélangeant de manière aléatoire selon le trafic des zones dans lesquels ces non-accidents sont situés.

Définition des zones d'étude

Comme dit précédemment, le but du projet est de définir selon les conditions météo, l'état de la route, la visibilité et le lieu, les zones de la ville de Chicago où le risque d'accident est le plus élevé. A chaque fois qu'un conducteur entrera dans une de ces zones à danger, il sera averti par un module dans sa voiture pour le pousser à être d'autant plus vigilant.

Il est alors indispensable de fractionner la ville de Chicago en zones. Pour cela, nous avons fait le choix de définir la ville de Chicago par un immense rectangle couvrant toute la ville. Ce rectangle est ensuite divisé en 67 * 67 cellules ayant une superficie de 0.25 km2. (voir figure ci-dessous). La taille de ces cellules a été choisie de manière à ce que les zones de danger soient les plus précises possible tout en ayant un nombre de données suffisant par zone pour établir un score de risque d'accidents de chacune des zones.

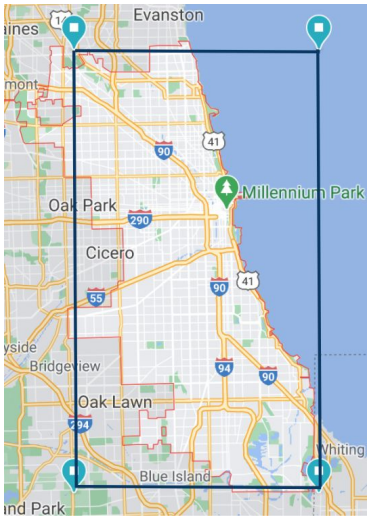


Figure 2: Left - Bounding box of the grid over Chicago. Note, a more precise and rotated bounding box caused technical difficulties and is not essential in building a PoC model version.

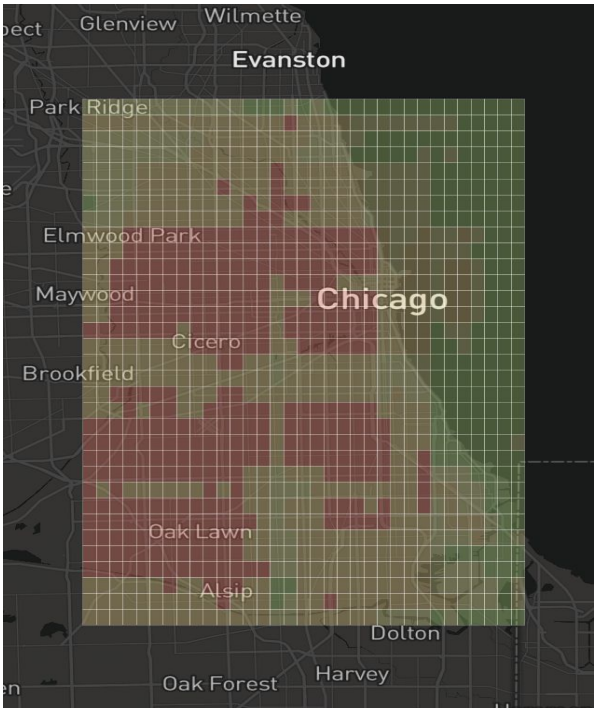
Figure 3: Right - Zooming into the grid. Right of the red line is the city area of Chicago.

Maintenant que la ville de Chicago est divisée en zone et que les données ont été nettoyées et traitées, nous pouvons commencer les algorithmes de Machine Learning et tenter de déterminer les zones les plus de la ville de Chicago où le risque d'accident est le plus élevé.

Prédiction des zones de risque d'accidents

Pour terminer, nous avons utilisé différents algorithmes de Machine Learning pour tenter de prédire le nombre d'accidents pouvant avoir lieu dans chacune des zones, en fonction notamment des conditions météo, de la visibilité, de l'heure, du jour et de l'état de la route. En connaissant ce nombre d'accidents par zone, nous pouvons déterminer quelles sont les zones les plus dangereuses et établir un score de danger par zone. On applique aux données accidents et "non accidents" créées virtuellement un réseau de neurones, un algorithme de prédiction Random Forest et un SVM (Support Vector Machine). On obtient les résultats suivants :

Algorithme	Temps d'entraînement	Précision
Réseau de neurone	20min	85,2%
SVM (linéaire)	23min	60,6%
Random Forest	6min	94,6%



Sur ces trois algorithmes, le modèle le plus performant et le plus précis est le Random Forest.
En faisant une visualisation de cette prédiction un mardi, sous temps clair à 5h de l'après midi sur toute la grille, on obtient la représentation ci-contre.

Cette représentation permet de mettre en évidence certaines zones de danger dans la ville (en rouge). Il reste cependant encore beaucoup de zones de danger maximum. Pour réduire le nombre de zones de danger max, on peut affiner les critères de définition d'une telle zone pour ne garder que les plus dangereuses. Évidemment ces zones de danger ne seront pas les mêmes si on lance notre modèle à une heure, un jour, un mois différent et avec d'autres conditions météo et de visibilité.

Conclusion et critiques

Nous avons donc des algorithmes nous permettant de prédire la dangerosité des zones qu'un conducteur va traverser en prenant en compte des paramètres extérieurs. Nous pouvons donc proposer notre algorithme à une assurance en tant que dispositif avertisseur de zones de dangers permettant de faire de la prévention en matière de sécurité routière. Il reste toutefois quelques interrogations à la fin de ce projet. Tout d'abord il faut s'assurer que l'assurance ne puisse pas détourner notre dispositif pour attribuer des bonus et des malus à ses clients en fonction de la dangerosité moyenne des zones qu'il traverse. Ensuite, nous ne savons pas si avertir un conducteur d'un danger non défini permet de réduire ses chances d'avoir un accident.

Authors :
Venkatesh Subramani : https://www.linkedin.com/in/venkatesh-subramani/
Julian Kopp : https://www.linkedin.com/in/julian-kopp-b786a211b/
Louis Lefebvre : https://www.linkedin.com/in/louis-lefebvre-971948194/
Mathieu Pierronne : https://www.linkedin.com/in/mathieu-pierronne-278949194/