Photo: DR / paperJam

Rapport de Data Mining

Accidents corporels de la circulation routière en 2016

Université Claude Bernard Lyon 1

Paul Peseux

27/01/2018

Introduction

Ce rapport résume le travail mené dans le cadre du cours de Data Mining faisant partie du M2 Data Science à l'Université Claude Bernard de Lyon. L'objectif de ce projet est de se confronter à un jeu de données réelles et de produire de la valeur en travaillant dessus. Le Data Mining, pouvant se traduire par "fouille de données", ce projet consiste essentiellement en une exploration d'un dataset choisi par mes soins. Ainsi à partir de données brutes, sous forme de fichiers .csv statiques, j'ai essayé d'en sortir de la valeur claire et de l'information concise. J'ai ainsi mis en place les techniques et outils qui nous avaient été introduit en cours de Data Mining, et j'ai pu confronter la théorie à la pratique.

J'ai volontairement choisi un dataset d'utilité publique, car je suis convaincus que le Data Mining et l'informatique en général peut être un formidable outil afin d'améliorer notre société. Ainsi j'ai choisi de m'intéresser aux accidents corporels de la circulation routière de 2016. En effet il m'a semblé judicieux de travailler sur un tel dataset, toute information que j'ai pu produire peut s'avérer cruciale afin d'aider et éclairer la mise en place d'une politique de sécurité routière.

L'objectif fondamental de ce travail est d'apporter des informations utiles à la réduction du nombre de tués et de blessés sur les routes françaises.

Ce rapport s'articule en quatre grande partie. Tout d'abord je présente en profondeur les données que j'ai récolté afin de bien situer notre travail. Ensuite je decris mon travail de mise en forme des données, qui est une étape cruciale en Data Mining. Ensuite j'explique les algorithmes que j'ai décidé de mettre en place afin de créer de la valeur sur les données. Finalement j'expose mes résultats et la valeur ajoutée aux données. Ce plan retranscrit l'enchaînement logique et chronologique du projet.

1.Présentation des données

1.1 Forme des données

Les données brutes que j'ai recueillies sont en libre service sur internet, sur un site public[1]. Elles s'articulent en quatre fichiers .csv qui contiennent les données et un .pdf qui contient les métadonnées essentielles à la compréhension et à la cohérence des fichiers précédemments cités.
Afin de comprendre la logique des données il est important de visualiser mentalement un accident de la circulation routière.
Un accident implique :

- une ou plusieurs personnes
- un ou plusieurs véhicules
- un lieu géographique
- des circonstances

Pour chacun de ces quatre points, il est possible de représenter l'information dans une ligne, qui sera respectivement stockée dans les fichiers suivants :

- usagers.csv
- véhicules.csv
- lieux.csv
- caractéristiques.csv

Ces données sont d'excellentes qualité car il y a très peu (moins de 1%) d'informations manquantes. De plus elles sont extrêmements complètes, et elles semblent décrire au mieux les accidents (dans la mesure de ce qui est faisable légalement, voir le prochain paragraphe).

[1] https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/

Il s'agit essentiellement de données labellisées, mais les données sont sous formes de nombres naturels. Ainsi, afin d'interpréter ces labels, la lecture attentive du .pdf qui regroupe les métadonnée est fondamentale. Par exemple, pour la colonne représentant la gravité d'un accident, les labels sont à interpréter dans ce sens :

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

Il n'y a donc aucune logique hiérarchique dans la labellisation des données. Ce fût un facteur important à prendre en compte dans la mise en place de certains algorithmes (en Machine Learning[2] notamment).

1.2 Taille des données

Il s'agit de données conséquentes, mais elles restent de tailles raisonnables :

- usagers.csv 5.3 MB
- véhicules.csv 3.6 MB
- lieux.csv 2.9 MB
- caractéristiques.csv 4.3 MB

Ainsi je n'ai connu aucun soucis de scalabilité, il fut aisé de faire tourner toutes les manipulations sur ma machine classique. Le fait que les données soient retranscrites en 1,2,3,... permet en autre de garder cette taille raisonnable.

1.3 Anonymat des données

Dans un cadre légal, il est nécessaire de respecter un certain anonymat dans ces données qui sont destinées à être mise en libre service sur internet. Ainsi je ne dispose pas d'informations privées sur les usagers (pas de noms ni de professions par exemple). De plus il est aussi nécessaire de ne pas avoir d'informations permettant de juger le tort des usagers. Ainsi nous ne disposons pas d'informations relatives à la vitesse des véhicules lors de l'accident, du taux d'alcoolémie des personnes impliquées. Ces informations qui semblent précieuses dans le cadre de notre projet, ne sont en fait pas nécessaires. En effet j'ai eu la volonté d'apporter quelque chose de nouveau. L'objectif n'est pas de sortir des informations déjà connues de tous. Il est en effet de notoriété publique que l'alcool et la vitesse sont les plus gros facteurs de risque au volant. J'ai souhaité aller plus loin.

[2] Ce projet réunit Data Mining et Machine Learning

2.Pré-processing

Cette étape fut fondamentale. J'ai dû bien définir notre objectif afin de formater les données de la façon la plus optimale. Je rappelle ici que l'objectif est d'apporter des informations utiles à la réduction du nombre de tués et de blessés sur les routes françaises. Ainsi j'ai décidé de regrouper ces données dans un tableau dans lequel chaque ligne correspond à un usager concerné par un accident de la route. En cela il se rapproche de usagers.csv, mais ce tableau contient toutes les informations que j'ai jugée utiles pour notre travail.

2.1 Variables utilisées

On propose ici un aperçu de la donnée finalement utilisée :

	catu	grav	sexe	secu	senc	obs	obsm	choc	manv	lum	agg	int	atm	col	mom	age	catv_gen	catvopp_gen
0	1.0	1.0	2.0	11.0	0.0	0.0	0.0	1.0	1.0	1.0	2.0	1.0	8.0	3.0	3.0	2.0	2.0	1.0
2	1.0	3.0	1.0	11.0	0.0	6.0	0.0	1.0	1.0	1.0	2.0	6.0	1.0	6.0	4.0	3.0	2.0	2.0
3	2.0	3.0	1.0	11.0	0.0	6.0	0.0	1.0	1.0	1.0	2.0	6.0	1.0	6.0	4.0	1.0	2.0	2.0
4	2.0	3.0	2.0	11.0	0.0	6.0	0.0	1.0	1.0	1.0	2.0	6.0	1.0	6.0	4.0	3.0	2.0	2.0
5	1.0	1.0	1.0	11.0	0.0	0.0	1.0	6.0	1.0	1.0	1.0	1.0	1.0	6.0	4.0	1.0	2.0	2.0

On propose en annexe la signification des 14 premières colonnes, car elles sont directement issues de la donnée brute.
Pour ce qui est des quatre dernières colonnes, elles sont le fruit de ma réflexion, et elles ont pour but de condenser de l'information que j'ai considéré trop éparpillée sinon.

Ainsi j'ai créé les colonnes :

i) *mom*

Cette colonne représente le moment de la journée auquel l'accident à eu lieu. Il regroupe ainsi les informations contenues dans *hr* et *mn* qui sont à mes yeux trop précises pour apporter de l'information globale. J'ai donc fait un choix de découpe de la journée:

0 : 00h-06h
1 : 06h-12h
2 : 12h-14h
3 : 14h-18h
4 : 18h-22h
5 : 22h-00h

Cette découpe est arbitraire et biaisée. Il s'agit cependant d'un choix que je pense tout à fait légitime. Il conditionne donc tous les résultats obtenus quant aux horaires des accidents.

ii) *age*

Il m'a semblé que l'âge est une variable trop précise là encore. J'ai donc fait le choix de découper la population en tranche d'âge comme c'est souvent le cas lors d'études statistiques :

0 : 00 - 15 ans
1 : 15 - 25 ans
2 : 25 - 45 ans
3 : 45 - 65 ans
4 : 65 et + ans

Là encore la découpe est arbitraire mais assumée.

iii) *catv_gen* et catvopp_gen

Ces deux catégories représentent respectivement la catégorie du véhicule associée à l'usager et la catégorie du véhicule opposé (0, respectivement 0, si pas de véhicule, respectivement si pas de véhicule opposé). Cette colonne est issue de la colonne *catv* qui est extrêmement exhaustive[3] :

01 - Bicyclette
02 - Cyclomoteur <50cm3
03 - Voiturette (Quadricycle à moteur carrossé) (anciennement "voiturette ou tricycle à moteur")
04 - Référence plus utilisée depuis 2006 (scooter immatriculé)
05 - Référence plus utilisée depuis 2006 (motocyclette)
06 - Référence plus utilisée depuis 2006 (side-car)
07 - VL seul
08 - Catégorie plus utilisée (VL + caravane)
09 - Catégorie plus utilisée (VL + remorque)
10 - VU seul 1,5T <= PTAC <= 3,5T avec ou sans remorque (anciennement VU seul 1,5T <= PTAC <= 3,5T)
11 - Référence plus utilisée depuis 2006 (VU (10) + caravane)
12 - Référence plus utilisée depuis 2006 (VU (10) + remorque)
13 - PL seul 3,5T <PTCA <= 7,5T
14 - PL seul > 7,5T
15 - PL > 3,5T + remorque
16 - Tracteur routier seul
17 - Tracteur routier + semi-remorque
18 - Référence plus utilisée depuis 2006 (transport en commun)
19 - Référence plus utilisée depuis 2006 (tramway)
20 - Engin spécial
21 - Tracteur agricole
30 - Scooter < 50 cm3
31 - Motocyclette > 50 cm3 et <= 125 cm3
32-Scooter >50cm3 et<=125cm3
33 - Motocyclette > 125 cm3
34 - Scooter > 125 cm3
35 - Quad léger <= 50 cm3 (Quadricycle à moteur non carrossé)
36 - Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé)
37 - Autobus
38 - Autocar
39 - Train
40 - Tramway
99 - Autre véhicule

Ainsi il m'a paru obligatoire de regrouper cela en trois grandes catégories de véhicules :

1 : les véhicules légers non fermés
2 : les véhicules classiques
3 : les véhicules lourds

L'appartenance d'une sous catégorie à nos nouvelles catégories me semble clair. Elle est disponible en annexe.

Pour le véhicule opposé, il a été nécessaire de faire matcher les véhicules en collision, cela a été rendu possible grâce au numéro d'accident, référencé dans les fichiers.

[3] Cela provient directement du fichier de métadonnées.

2.2 Variables abandonnées

Comme mentionné précédemment, les données sont extrêmement complètes. Cela contraint à laisser de côté de nombreuses variables, qui pourraient faire l'objet d'autres études. Par exemple j'ai fait le choix d'abandonner la notion géographique d'un accident (région, département , …). Cela est un choix volontaire qui découle du fait que j'ai souhaité être réaliste quant au travail que j'étais capable de fournir. La liste complète des variables abandonnées est disponible en annexe.

2.3 Détection d'outliers et de données aberrantes

Comme Dans tout jeu de données, il a été nécessaire de détecter des outliers et des données aberrantes. La quantité de variables étant imposante, j'ai décidé d'appliquer l'algorithme Isolation Forest, qui est directement disponible dans sklearn sous Python. Ainsi j'ai pu nettoyer les données d'éventuels outliers.

3.Algorithmes

3.1 Apriori

Cet algorithme est celui qui nous a été présenté en cours. Je l'ai utilisé afin de rechercher les données les plus fréquentes lors des accidents. Je l'ai mis en place sur des sous ensembles des données afin que celui ci apporte une information précise. En effet vu la diversité des données, faire tourner Apriori sur l'ensemble des données, n'aurait pas eu de sens.
Afin de faire tourner Apriori, il a été nécessaire de binariser les données[4].

La démarche était de trouver une population P parmis les données, rechercher les données fréquentes dans P, puis extraire une sous population P' de P et comparer les données fréquentes de P et celles de P'

De plus, après avoir mis en place la recherche des données les plus fréquentes, il était tout à fait logique de mettre en place la recherche des règles d'associations de ces données fréquentes.

[4] pour cela, nous avons utilisé la fonction pandas.get_dummies sous Python

3.2 Tree Classifieurs

Cette Méthode, qui est à la limite de Data Mining et Machine Learning, a permis de visualiser les données discriminantes pour déterminer le sort de l'usager.
Ainsi en appliquant le critère du Gini pour discriminer la population, l'algorithme mis en place offre des résultats visuels et un outil simple qui permet d'exhiber le caractère discriminant de certaines données.
J'ai privilégié des arbres de petite taille afin de ne pas tomber dans le particulier, mais d'obtenir des résultats significatifs.

4.Résultats

On présente ici les résultats obtenus qui semblent les plus pertinents.

4.1 Motifs fréquents

Population P : Véhicule léger non fermé et pluie battante
Population P' : 15-25 ans

Motif	Fréquence dans P	Fréquence dans P'
Homme et en agglomération	46%	55%
Blessé hospitalisé et voiture classique heurtée	47%	55%

Population P : Véhicule classique et pluie légère
Population P' : entre 22h et minuit

Motif	Fréquence dans P	Fréquence dans P'
Sans éclairage public contre voiture classique	23%	72%
Blessé hospitalisé ayant percuté une voiture classique hors agglomération et hors intersection	24%	37%

Population P : Collision frontale entre deux voitures classiques

Population P' : les plus de 65 ans

Motif	Fréquence dans P	Fréquence dans P'
Femme, de plein jour avec des conditions atmosphériques normales	20%	29%
Blessé léger sans percussion d'obstacle fixe	27%	20%

4.2 Règles d'association

Ici je n'ai pas réussi à sortir de l'information pertinente. En effet les seules associations que j'ai réussi à sortir n'apportent aucune plus-value.

4.3 Tree Classifieurs

Population P : les 15-25 ans ayant eu un accident dans des conditions météorologiques normales :

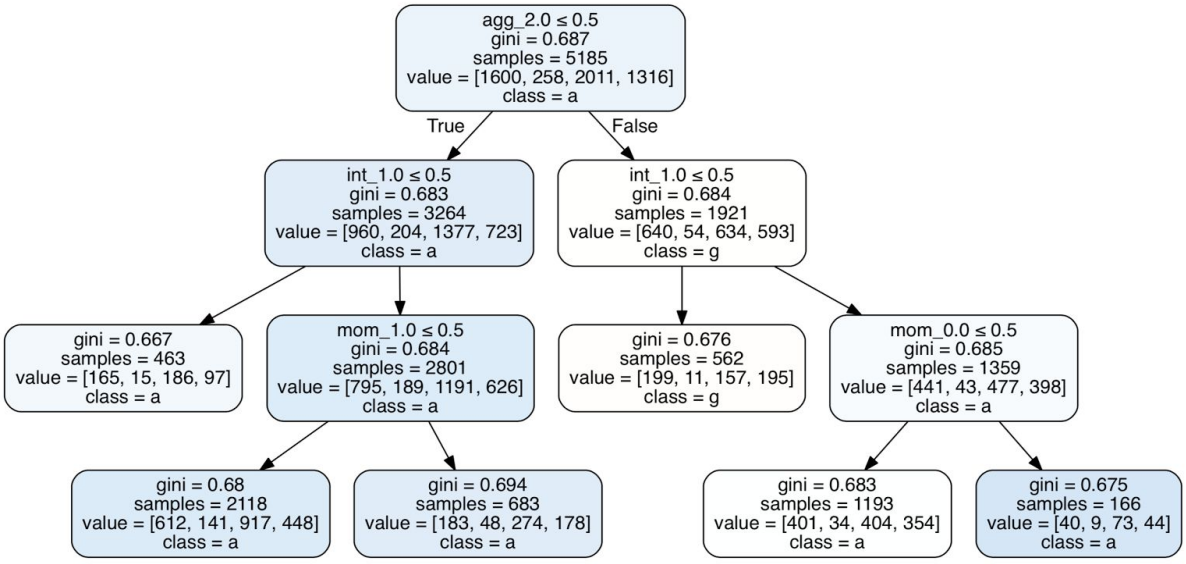

Figure 1 : output graphique de notre tree classifieur

Ce genre de résultats graphiques est très parlant et permet de catégoriser très facilement un accident en fonction des autres.

Population P : les véhicules légers rentrant en collision avec un véhicule léger

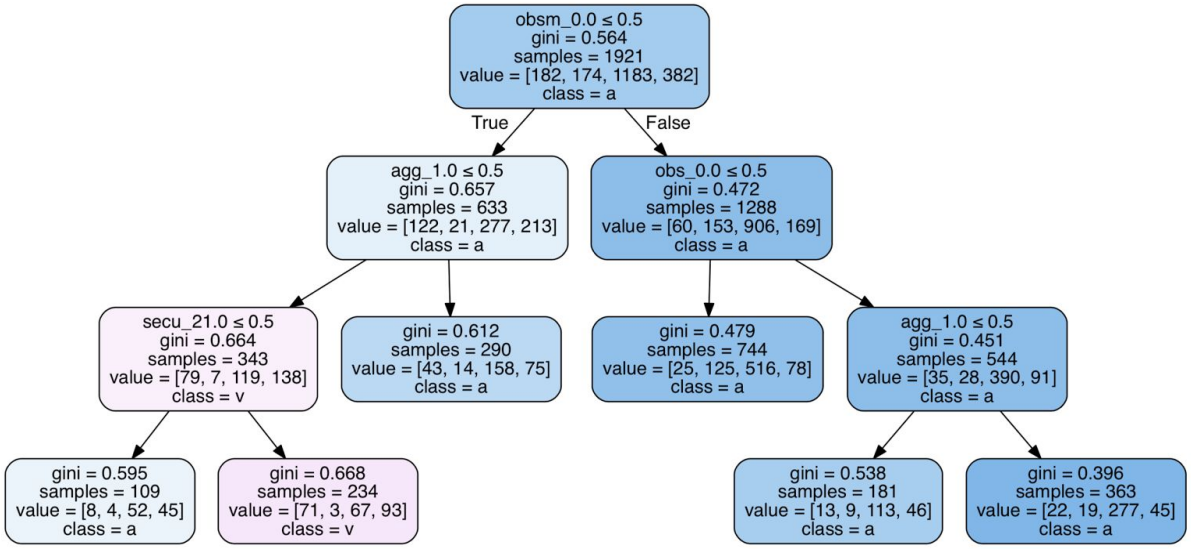

Figure 2 : output graphique de notre tree classifieur

Conclusion

Ce projet s'inscrit au coeur de notre formation de Data Scientist et m'a permis de fouiller des données réelles.

J'ai mis un temps certain à les découvrir et à les mettre en forme. Une fois cette mise en forme faite, j'ai appliqué dessus des techniques classiques de Data Mining afin d'en sortir de la valeur.

Les résultats exposés ici sont un résumé des différentes fouilles, ils ne résument pas toutes les informations qui peuvent être sorties d'un tel jeu de données. Il est en effet crucial de savoir sortir les chiffres les plus utiles, car il est en effet tentant de submerger le lecteur d'informations. J'ai donc choisi d'être concis.

Toute cette étude peut servir d'introduction au projet de Machine Learning qui s'inscrit en parallèle de celui-ci.

Les résultats sortis de cette étude s'inscrivent en complément des connaissances déjà établies sur la sécurité routière.

Annexes

Signification des colonnes :

catu : catégorie d'usager
grav : gravité de l'accident
sexe : sexe de l'usager
secu : existence et type d'équipement de sécurité
senc : sens de la circulation
obs : obstacle fixe heurté
obsm : obstacle en mouvement heurté
choc : point de choc initial
manv : manoeuvre
lum : luminosité
agg : agglomération
int : intersection
atm : atmosphère
col : collision

Appartenance du à la catégorie de véhicule :

véhicule classique = 7,10
véhicule air libre = 1,2,30,31,32,33,34,35,36
véhicule lourd = 13,14,15,16,17,37,38,39,40

Variables abandonnées :

- num_veh
- Num_Acc
- locp
- actp
- etatp
- catv
- hrmn
- mois
- jour
- catvopp
- occutc
- place
- an_nais