Vincent BERNARD

Rapport : Apprentissage automatique

1. Introduction

Ce projet consiste à produire un modèle de machine learning ou de deep learning qui permette de prédire si un accident de la route est grave ou pas. Pour qu'un accident soit considéré comme grave il faut qu'au moins une personne soit hospitalisée ou tuée. Pour prédire cela, nous avions à disposition des informations sur les lieux, les véhicules, les usages et quelques autres informations. Ce projet est réalisé sous la forme de compétition Kaggle dans laquelle tous les étudiants de la matière participaient à la compétition. Le but était de maximiser l'AUC, en effet les données sont déséquilibrées car il y a plus d'accidents pas grave que d'accidents graves.

Tous les modèles rendus sont présents dans le notebook « modeles_projet_kaggle_Vincent_BERNARD », cela évite de devoir relancer le prétraitement des données 3 fois. Pour lancer le notebook il faut que tous les fichiers présents dans le rendu se situent à côté des dossiers TRAIN et TEST. Les autres fichiers permettent le traitement des différents datasets. Le run du notebook, le prétraitement des données et les modèles peuvent prendre plusieurs minutes.

2. Feature Engineering

J'ai décidé de traiter les données rubrique par rubrique (lieux, véhicules, ...). J'ai ensuite redéfinit la notion d'accident grave comme indiqué précédemment. Pour traiter les données, j'ai dû réaliser des regroupements de catégories. Pour faire ces regroupements, j'ai analysé la part d'accidents graves ou non pour chaque catégorie puis j'ai regroupé j'ai regroupé par part d'accidents graves similaires. J'ai remplacé les valeurs aberrantes ou manquantes par les valeurs les plus représentées. Si par exemple la part d'accidents graves pour une moto et un scooter est la même alors je vais les regrouper dans la même catégorie. J'ai créé des notions d'ordre en fonction de la part d'accidents graves de plus en plus importante (1 : en général pas très grave ,2 ,3 ... : la plupart des accidents sont graves).

Comme ceci :

```
# Regroupez les données par la colonne 'catv' et comptez le nombre d'accidents graves et non graves
result = df_nv.groupby('situ')['grave'].agg(graves='sum', non_graves=lambda x: len(x) - x.sum(), total='count').reset_index()

# on calcule le pourcentage d'accidents graves sur le total des accidents pour toutes les catégories
result['Pourcentage d\'accidents graves'] = (result['graves'] / result['total']) * 100

# pour afficher la part d'accident graves ou pas
result = result.rename(columns={'graves': 'Nombre d\'accidents graves', 'non_graves': 'Nombre d\'accidents non graves'})
```

2.1 Gestion des usagers

Pour les usagers, il y a plus de personnes impliquées dans un accident que d'accident. Ce qui est normal car par exemple si 2 voitures se rentrent dedans il va y avoir une ligne pour chaque occupant de chaque voiture et ces personnes font partie d'un seul et même accident. Or, il ne faut garder qu'une seule ligne par accident. J'ai donc décidé de compter pour chaque colonne le nombre d'apparition d'une variable, en faisant les regroupements indiqués précédemment. Par exemple, s'il y a 1 piéton dans un accident, il y a de grandes chances que cet accident soit grave. Mais il y a encore plus de chances qu'il soit grave s' il y a 2 piétons, puis 3,4,... et pareil pour les véhicules sans protection comme motos, quads, scooters, ... Les différentes variables que j'ai créé :

Nb_pers : Le nombre de personnes impliqués dans l'accident car j'ai remarqué que plus il y avait de personnes dans un accident plus il y avait de chances que celui-ci soit grave.

Nb_place_avant : Les personnes à l'avant ont plus de risque que les personnes à l'arrière donc plus il y a de personnes à l'avant plus il y a de chances que celui-ci soit grave. J'ai fait pareil pour le nombre de personnes au milieu et à l'arrière (**Nb_place_milieu et Nb_place_arrière)**

Nb_hom_cond et nb_fem_cond : Je me suis aperçu que la part d'accidents graves était plus importante lorsque des hommes étaient au volant plutôt que les femmes qui sont visiblement plus prudentes.

Moy_age : L'âge moyen par accident, j'ai pu observer que les jeunes conducteurs et ainsi que les personnes de plus de 60 ans étaient plus à même d'avoir un accident grave que les autres. J'ai donc décidé de regarder l'âge des personnes de l'accident en faisant l'année de l'accident « – » l'année de naissance des personnes. Puis j'ai fait une moyenne d'âge des personnes impliquées dans l'accident.

nb_dom_trav, nb_dom_eco, nb_achat, nb_prof, nb_loisir : j'ai compté le nombre de personnes qui réalisaient ce trajet lors de l'accident. Dans l'ordre, les personne domicile-travail ont moins de chances d'avoir un accident grave alors que les personnes qui se déplacent pour leur loisir sont ceux qui ont le plus de chances d'avoir un accident grave.

nb_conducteurs, nb_passagers, nb_piets : encore une fois, ces variables sont dans l'ordre. La part d'accidents graves pour les conducteurs est minime car il y a énormément d'accidents mais heureusement la plupart ne sont pas graves. Par contre, dès qu'un piéton est impliqué dans un accident, la part d'accidents graves est beaucoup plus importante. Et, plus il y a de piétons dans l'accident, plus il y a de chances que celui-ci soit grave.

Nb_pie_prot, Nb_sans_prot : pareil plus il y a de piétons plus il y a de chances que l'accident soit grave. Cependant, il y a moins de chances que l'accident soit grave si les piétons sont « protégés » (trottoir, refuges, …) que si les piétons ne sont pas protégés (loin d'un passage piéton, accotement, ….)

2.2 Gestion des lieux

Une nouvelle fois, j'ai fait des regroupements de catégories par part d'accident grave ou non pour ces catégories. Tous les regroupement de variables ont été effectués par ordre de dangerosité, j'indique d'ailleurs la part d'accidents graves pour chaque regroupements (ex : route départementale (+ de 40%), cela signifie que plus de 40% des accident sur une route départementale sont graves :

Surf : Regroupement en 1 : Normale, Mouillée et huile, dans 41% des cas, c'est un accident grave contrairement au regroupement des autres catégories dont plus de 57% des cas ce sera un accident grave (Neige, verglas, inondation, ..)

Situ : 1 : piste cyclable/voie spéciale car peu de personnes sont autorisées à rouler dessus (~23%), 2 : chaussée/arrêt d'urgence(~38%), 3 : trottoir(~54%), 4 : accotement (~62%)

Prof : 1 : plat(38%), 2 : pente (48%) et 3 : sommet et bas(~54%) (ordre de « dangerosité » basée sur la part d'accident grave ou non dans ces cas)

Vosp : 1 : piste cyclable (29%), 2 : bande cyclable/voie réservée (~32%) et 3 le reste (~42%)

Nbv : je me suis aperçu que le nombre de voies pouvaient être intéressant. En effet, la majorité des accidents graves interviennent lorsqu'il y a 2 voies (1 voie dans un sens et 1 dans l'autre).

Circ : 1 : sens unique (24%), 2 : chaussée séparés/ voie d'affectation (~34%), 3 : bidirectionnel (48%)

Infra : 1 : tunnels (~22%), 2 : raccordement ou chantier (~33% des cas c'est grave), 3 : voie ferrée, piétons, … (~41%), 4 : les ponts (~44%) et 5 : péage (infrastructures solides + possibilité de piétons) (57% des accidents dans ce cas sont graves)

Plan : 1 :rectiligne (37%), 2 : les courbes (56%)

Catr : 1 : voie communale (~28%), 2 : autoroute et route de métropole (~30%), 3 : route nationale, hors réseau public, … (~45%), 4 : route départementale (+ de 60%)

Vma : J'ai arrondi les différentes valeurs à 30,50,70,90,110 et 130km/h. Les valeurs aberrantes ou manquantes, je les ai remplacées par la vitesse moyenne en agglomération ou hors agglomération selon l'endroit où s'est déroulé l'accident. En observant les données, je me suis aperçu que la vitesse à laquelle la part d'accidents graves est la plus importante est 90 (61%) et 130 km/h (51%) alors qu'à 30km/h (~20%).

2.3 Gestion des véhicules

J'ai procédé de la même manière que pour les usagers, j'ai compté le nombre d'apparition de certains phénomènes par accident. Des « phénomènes» qui plus sont nombreux peuvent être dangereux, et qui parfois tout seul peuvent déjà l'être :

Sum_Occutc : J'ai remplacé les Na par la moyenne de personne par véhicule (arrondie à l'entier). On compte le nombre de personnes pour chaque accident.

Motor_1, motor_2, motor_3 : Je compte le nombre d'apparition de chacun de ces types de moteurs pour chaque accident. 1 : J'ai observé que la part d'accidents graves était la plus faible pour les moteurs électrique et hybride. (17%) 2 : hydrogène et humain (23%) et en 3 : hydrocarbures (principalement les vieilles voitures qui sont responsables d'accidents graves car elles n'ont pas toutes les aides d'aujourd'hui) (~26%).

nb_catv_1, nb_catv_2,..., nb_catv_7 : regroupement des véhicules par part d'accident grave ou pas avec. 1 : EDP, 3RM > à 50cm3 (~15%), 2 : VAE, Scooter (50cm3), ... (~23%), 3 :vélo, VU, VL (~28%), 4 : Engin spécial, ... autocar (~31%), 5 : PL, moto, ... semi-remorque (~37%), 6 : train, tracteur agricole (47%), 7 : très dangereux : les quads lourd sans protection (+ 55%)

obs : 1 : sans obstacle ou véhicule stationné(~27%), 2 : mobilier urbain, ilot (~34%), 3 : trottoir, glissière,... (~41%), 4 : poteau, parapet,... (~50%), 5 : arbre, fossé, aqueduc (63% de chance d'avoir un accident grave dans ces conditions (selon nos données)). On compte le nombre d'apparition de chaque catégorie par accident. Je considère que si sur son passage une personne heurte plusieurs obstacles fixes, les chances d'accident grave augmentent.

obsm : 1 : véhicules/véhicules sur rail (26%),**1 :** piéton (29%), 3 : animaux et autres (41%). On compte le nombre d'apparition de chaque catégorie d'obstacle mobile par accident. Je considère que si sur son passage une personne heurte plusieurs obstacles mobiles, les chances d'accident grave augmentent.

choc : 1 : arrière (les chocs à l'arrière sont réputés pour ne pas être très dangereux) (18%), 2 : à l'arrière sur les côtés (28%), 3 : chocs à l'avant (~31%), 4 : les tonneaux (~42%) . Je compte le nombre d'apparition de chacun de ces cas par accident car si lors d'un accident une personne tape avec son pare choc l'arrière d'une autre voiture, l'accident sera moins grave qu'un choc frontal entre 2 voitures.

Manv : 1 : ouverture de porte (11%), 2 : couloir de bus(bon sens), changement de file à droite, entre 2 files (17%), 3 : dépassement à droite, arrêt, trottoir (21%), 4 : tournant à droite, changement de file à gauche, ... (23%), 5 : demi-tour, tournant gauche (28%), 6 : dépassement à gauche, manœuvre d'évitement (31%) , 7 : traverser la chaussée, contresens (35%), 8 : se déporter à droite (39%), 9 : déporter à gauche, franchissement terre-plein central (~43%)

2.4 Gestion des caractéristiques

J'ai procédé de la même manière que précédemment :

An : utile pour le calcul de l'âge des personnes de l'accident puis avant le covid la part d'accident grave était plus importante. Puis le passage de 90 à 80km/h.

Mois : 1 : mois avec moins d'accident (39%), 2 : moins dont 41% des accidents sont graves, 3 : mois de juin et juillet (~46% des accidents pour ces 2 mois sont graves)

Dep : Je me suis aperçu qu'il y avait des départements plus propice à avoir un accident grave. Par exemple, à Paris il y a énormément d'accidents mais dans plus de 90% des cas il ne sont pas graves. Contrairement à certains départements en campagne ou en outre-mer.

Hrmn : la nuit entre minuit et 6h c'est le plus dangereux contrairement à 8,9,12,13h qui sont les heures les plus sûres.

Lum : 1 : nuit bien éclairé/ plein jour (29%), 2 : l'aube (33%), 3 : nuit sans éclairage(46%)

Int : 1 : plus de 4 branches, place (14%), 2 : autres intersections (25%), 3 : passage à niveau ou accident hors intersection (~34%)

Atm : 1 : normal et petite pluie (35%), 2 : forte pluie, temps couvert (41%), 3 : neige(50%), 4 : brouillard, vent fort et éblouissement (~60%)

Col : 1 : collision arrière ou 3 véhicules en chaîne (20%), 2 : 2 véhicules sur le côté (25%), 3 : collisions multiples et autres types de collision (37%), 4 : collision frontale (40%)

Agg : il y a beaucoup plus d'accident grave hors agglomération (59%) qu'en agglomération (31%)

3. Choix des modèles

J'ai testé différents modèles comme : Random Forest, l'ExtraTreeClassifier, la Regression logistique, les Arbres de décision, le GradientBoostingClassifier, le XGBClassifier qui est une implémentation optimisée du Gradient Boosting avec une grande capacité à traiter les grandes données), le LightGBM (Light Gradient Boosting Machine) qui a été développé par Microsoft qui utilisé une méthode d'optimisation de l'histogramme pour accélérer le processus d'apprentissage. AdaBoostClassifier (Adaptive Boosting ClassifierJ'ai ensuite essayé des réseaux de neurones comme un réseau fully connected. J'ai également essayé un réseau de neurones fully connected avec de la régularisation ainsi qu' un réseau de neurones récurrents. Puis des modèles d'ensemble que nous verrons par la suite.

Pour améliorer mes performances, j'ai dû trouver les meilleurs paramètres de mes modèles. Pour cela, j'ai dû utiliser **GridSeachCV** ainsi que **RandomizedSearchCV,** en leur donnant différentes possibilités de combinaisons, ils me ressortent la « meilleure » . Les modèles retenus :

Le **CatBoostClassifier**, développé par Yandex) qui a des performances élevées avec une gestion automatique des caractéristiques catégorielles. Le CatBoost intègre des techniques de régularisation et de diminution du taux d'apprentissage adaptatif pour prévenir le surajustement et améliorer la généralisation du modèle.

Un votting classifier (soft) : composé du LGBMClassifier, XGBClassifier, CatBoostClassifier et l'implémentation classique du grandient boosting classifier (GradientBoostingClassifier) (avec des paramètres plus ou moins optimisés pour chacun des modèles). Le modèle fait un calcul de probabilité pour chaque classe, les probabilitées sont agrégées en moyennes et la classe prédite est celle avec la probabilité moyenne la plus élevée.

Le **staking** utilisé est un modèle de stacking composé du LGBMClassifier, XGBClassifier, CatBoostClassifier et du GradientBoostingClassifier avec un méta-modèle de régression logistique. Il combine les prédictions de plusieurs modèles de base pour obtenir une meilleure prédiction finale. Il utilise les prédictions comme caractéristiques d'entrée pour un nouveau modèle qui apprend à fusionner ces prédictions pour produire une prédiction finale.

4. Conclusion

Lors de ce projet j'ai pu pleinement découvrir la Data Science. Il y avait plusieurs parties dans ce projet. Trouver les bonnes variables, trouver la meilleure manière de les exploiter. Ensuite trouver un modèle adapté au problème posé, ensuite l'ajuster avec les bons paramètres afin d'en tirer les meilleurs résultats.