BARRY Alpha
Master 2

H3 HITEMA

Rapport Machine Learning

Accident de voitures aux Etats-Unis

H3 Hitema
ISSY

1. Introduction

Nombreux sont les accidents routiers qui arrivent chaque jour dans le monde. Les Etats-Unis n'échappent pas à la règle avec environ 38000 morts et plus de 4 millions de blessé par ces mêmes accidents. Aujourd'hui, l'objectif est de baisser le nombre d'accident de la route. Pour atteindre cet objectif, nous allons nous poser 2 questions :

- Quelles sont les régions les plus touchées par les accidents de voitures
- En cas d'accidents, quelles est la gravité de celui-ci.

Répondre à ces questions nous permettra de pouvoir prendre les mesures nécessaires dans les zones adéquats selon la gravité des accidents.

Pour répondre à ces questions, nous allons chercher à faire 2 choses :

1. Cibler les zones les plus touchées par les accidents de voitures
2. Prédire la gravité d'un accident

Prédire la gravité d'un accident nous permettra de savoir si en fonction de certain critère environnant, les accidents qui auront lieux dans une zone précise seront graves ou non.

2. Adaptation du Dataset

Dans un premier temps, on a dû adapter notre Dataset car il contenait beaucoup trop de données (quelques millions de lignes). Le problème était que, faute de la quantité de données, le temps d'exécution était augmenté et donc, les performances en été impacté. Il a donc été décidé de créer une version mini de note Dataset avec une diminution des données à 100000 permettant tout de mêmes d'avoir des résultats satisfaisant malgré la quantité de données réduite.

Dans un second temps, afin de pouvoirs appliquer nos différents algorithmes, une phase de cleaning a dû être réalisée sur nos données. Il y a également eu une réflexion autour du faite de garder certaines colonnes ou non (voir Liste_des_colonnes.docx).

Dans un dernier temps, un traitement des champs vide soit NaN ont été effectué. C'est-à-dire qu'il a été décidé d'enlever toutes les lignes contenant des valeurs NaN afin d'avoir une cohérence entre chaque ligne.

3. Les zones d'accidents

Nous allons tout d'abord essayer de voir quelles sont les états les plus touchées par les accidents de voitures :

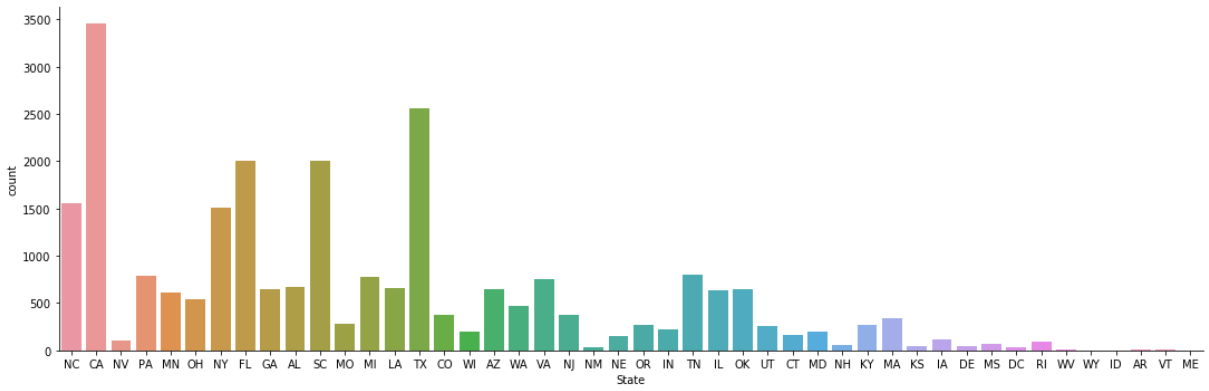

Figure 1 : Histogramme du nombre d'accidents par Etat

Nous pouvons voir dans le diagramme ci-dessus que certains états ont beaucoup plus d'accidents que d'autres. En l'occurrence :

- La Californie (CA) avec 3458 accidents
- Le Texas (TX) avec 2557 accidents
- La Floride (FL) avec 2010 accidents
- La Caroline du Sud (SC) avec 2006 accidents
- La Caroline du Nord (NC) avec 1555 accidents
- New York (NY) avec 1507 accidents

ont beaucoup plus d'accidents que la plupart des autres pays. En revanche, on va voir que les états avec le plus d'accidents n'ont pas forcement les villes plus accidents :

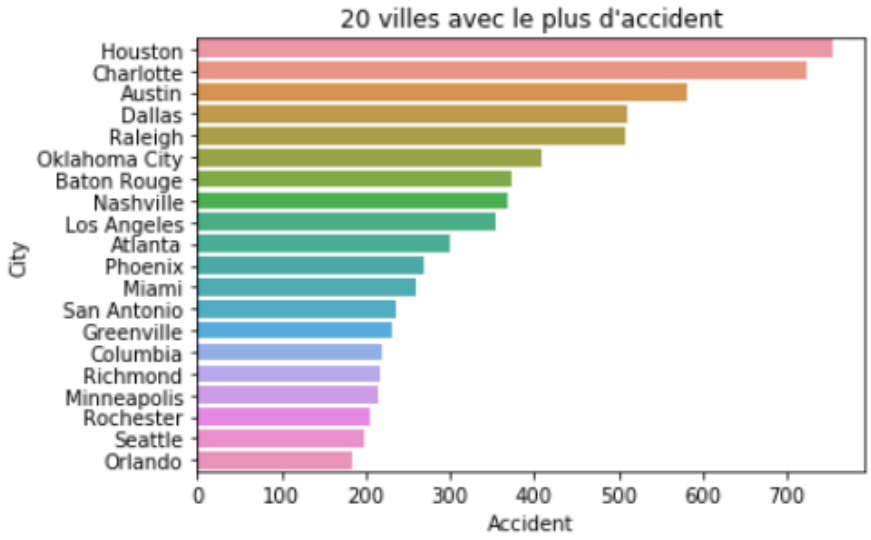

Figure 1 : Histogramme du nombre d'accidents par Ville

Comme nous pouvons voir, les états avec le plus d'accidents ne possède pas forcement les villes avec le plus d'accidents. Pour en citer quelques-uns :

- Oklahoma City qui se situe dans l'Oklahoma (OK)
- Baton Rouge en Louisiane (LA)
- Nashville au Tennessee (TN)
- Ou encore Atlanta en Géorgie (GA)

qui sont des villes situé dans des états avec beaucoup moins d'accidents que d'autres vu dans la **Figure1**.

C'est un point auquel il faudra prêter attention lorsque l'on voudra cibler les zones les plus accidents.

4. Prédiction des zones à forte gravité

Maintenant que nous avons cibler les zones d'accidents, nous voulons prédire quelles seront les zones touchées par les accidents graves. Pour cela, nous avons dû étudier plusieurs algorithmes afin de déterminer quelles étaient le plus intéressant à utiliser afin d'obtenir la meilleurs accuracy. 3 Algorithme ont été retenu :

- **k-nearest neighbors**
- **RandomForestClassifier**
- **GradientBoostingClassifier**

Voici les résultats obtenus avec chacun de ces trois algorithmes :

Algorithme	Précision (train-test)
k-nearest neighbors	70%-67%
RandomForestClassifier	74%-72%
GradientBoostingClassifier	92%-76%

Figure 3 : Tableau de la précision des prédictions de la gravité (Severity)

Le GradientBoostingClassifier est l'algorithme le plus intéressant car c'est celui qui nous a permis d'avoir le résultat le plus accurate. On a eu 76% lors du test mais lors de l'entrainement, ont été à 92% pour le même algorithme. Il y a donc eu de l'overfitting. On peut supposer que si on avait plus de données de tests, la précision aurait été bien supérieur.

Cependant, c'est un résultat qui faudra tout de même prendre avec des pincettes car comme nous pouvons le voir dans le graphique ci-dessous :

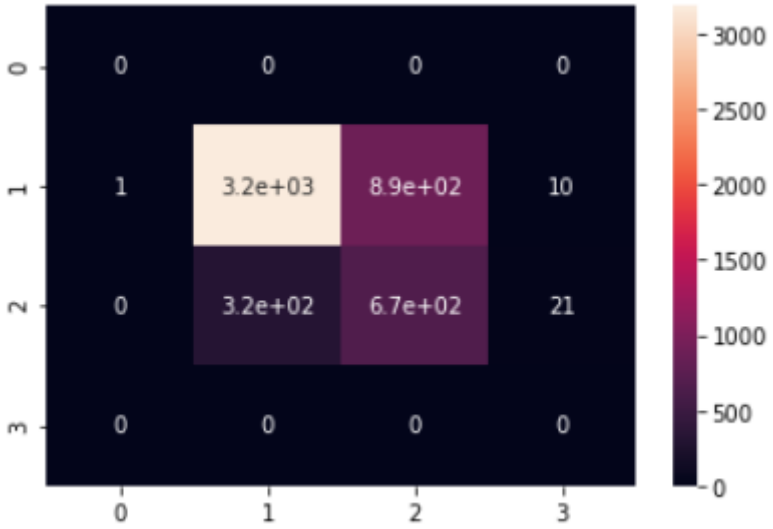

Figure 4 : Matrice de confusion

le graphique nous montrent que certaines gravité prédites/estimées sont en réalités soit plus graves, soit moins graves.

5. Conclusion

En conclusion, nous avons désormais une idée des régions les plus touchées par les accidents. Nous avons également pu prédire la gravité d'un accident à 76% d'accuracy. Il faudra cependant prendre ces résultats avec des pincettes car il y a toujours une probabilité de chance de se retrouver avec des faux positifs. Il faudra finalement tester l'algorithme choisie avec plus de données afin de voir si le résultat obtenu sera beaucoup intéressant.

6. Glossaire

Mot	Definition
k-nearest neighbors	k-nearest neighbors est un algorithme servant à estimer la sortie associée à une nouvelle entrée x, la méthode des k plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir.
RandomForestClassifier	RandomForestClassifier est une méthode d'apprentissage d'ensemble pour la classification, la régression et d'autres tâches qui fonctionnent en construisant une multitude d' arbres de décision au moment de l'apprentissage et en sortant la classe qui est le mode des classes (classification) ou la moyenne / moyenne de la prédiction (régression) des arbres individuels.
GradientBoostingClassifier	Gradient Boosting est un algorithme qui va assembler, unir en un tous des modèles élaborés séquentiellement sur un échantillon d'apprentissage dont les poids des individus sont corrigés au fur et à mesure. En bref, il va construire plusieurs modèles et à chaque nouveau modèle, il va, en cas d'erreur donner un poids supplémentaire à ces erreurs et se corriger.

7. Annexe

https://www.kaggle.com/sobhanmoosavi/us-accidents

https://www.nsc.org/road-safety/safety-topics/fatality-estimates

H3 Hitema
ISSY