LOUIS AMSTUTZ

Rapport de projet

# MACHINE LEARNING

Professeur : JEFF ABRAHAMSON

Projet final

## Crimes à Chicago de 2001 à aujourd'hui

16/03/2017

Depuis longtemps, la ville de Chicago propose un portail de données ouvertes. Ils publient régulièrement leurs dernières données disponibles.

Parmi ces données, nous pouvons trouver les « crimes » depuis 2001 à aujourd'hui.

En anglais, les « crimes » correspondent aux délits (ex : les vols).

Le dataset comporte plus de 6 000 000 de lignes ! parfait pour du machine Learning. Néanmoins long à traiter. Pour alléger les traitements et soulager la machine, lors du nettoyage des données j'ai enlevé toutes le colonnes que je ne voulais pas directement traité, elle pourrait être utile pour améliorer la prédiction mais trop lourde dans mon cas.

Mon but est de prédire les crimes, plus précisément l'heure où a eu lieu un crime donné ou sa localisation.

# Nettoyage des données

Dans mon nettoyage de données, j'ai commencé par supprimer beaucoup de colonnes que je ne voulais pas utilisé, (trop encombrante et trop sale). J'ai gardé le type de crime, la date et la localisation (adresse ou coordonnées).

« Adresse ou coordonnées » car des coordonnées sont manquantes ( +82000 manquantes) mais toutes les adresses sont présentes.

Problème 1 : les adresses sont anonymes (il manque une partie du numéro de bâtiment) !

```
In [14]: dfTrain[pd.isnull(dfTrain['Latitude'])]['Block']
```

```
Out[14]: 73                    046XX W 49TH ST
         365                   009XX E 40TH ST
         454             045XX N CLARENDON AVE
         786               112XX S HERMOSA AVE
         814               042XX N ASHLAND AVE
         842               069XX S STEWART AVE
         867                   003XX E 29TH ST
         879                 082XX S COLES AVE
         1062                  003XX E 29TH ST
         1078                  015XX E 55TH ST
         1136                  023XX W OHIO ST
         1498              0000X S LA SALLE ST
         2234                 072XX S BELL AVE
         2368                  033XX W 47TH ST
         2394              050XX S WESTERN AVE
         2402                 050XX N BROADWAY
         2406       023XX N LINCOLN PARK WEST
         2411                  001XX N STATE ST
         2418              022XX N LINCOLN AVE
         2919              025XX N KIMBALL AVE
```

Problème 2 : comment retrouver les coordonnées à partir d'adresses ?

Pour résoudre ces problèmes, j'ai remplacé les 'X' ou 'XX' dans les adresses par 4 ou 49. (la moyenne des possibilités) nous serons suffisamment précis …

Puis j'ai utilisé l'API Google Géocoder pour récupérer les coordonnées géographiques à partir d'adresses. Le problème de l'api c'est qu'elle est payante si on veut faire plus de requêtes que 2500 par jours. J'ai donc fait cette opération 3 fois (3 jours) puis pour la suite j'ai supprimer les lignes manquantes.

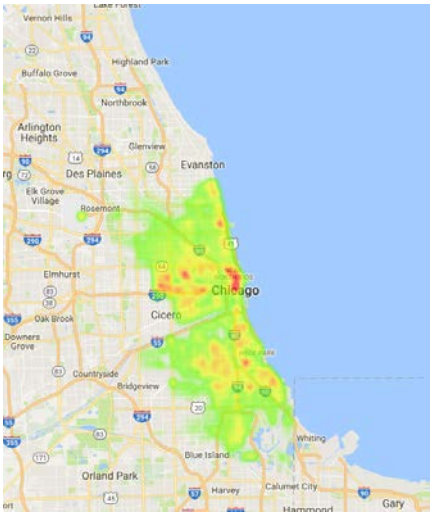
La dernière étape de nettoyage consiste à reformater la date pour pouvoir l'utiliser facilement et simplement. J'ai alors séparer le numéro du jour, du mois, l'année, et l'heure arrondie. Cela me permettra de faire très facilement des tris ou régression pour chaqu'une de ses parties.

Puis j'ai découper mon dataset en deux (30%/70%) pour l'entrainement et les tests.

# Partie 2 :

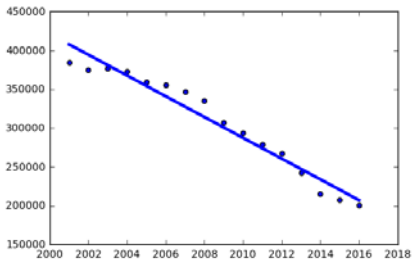
Dans cette partie, j'ai tout d'abord voulu savoir si il y a des quartiers plus dangereux que d'autres à Chicago et si oui, si il y a eu des changements au cours du temps.

J'ai donc mappé grâce a une librairie gmplot (qui utilise google maps) 1000000 de crimes pris au hasard dans le dataset.




En mappant par année on remarquera que les zones rouges ne changent quasiment pas.

Il est alors naturel de regarder si les nombres de crimes évoluent au fil des années.




En 15 ans, il y a deux fois moins de crimes ! la diminution est régulière, on réalise alors une régression linéaire.

Prédiction pour 2018 : 180295 crimes !

Nous voulons maintenant pouvoir prédire le nombre de crimes à chaque heure des journées, mais l'histogramme révèle une variation « sinusoidale » du nombre de crimes. J'ai alors isolé pour chaque mois de chaque année le nombre de crimes à chaque heure, et réaliser une régression sur ces données (une régression par heure).

Pour le futur de cette partie, il suffira de faire chaque prédiction pour une date donnée et nous pourrons récupérer l'évolution par heure pour cette date.

Cette partie est réitérable pour le nombre de crimes par mois 😊

## Aller plus loin

Je n'ai pas réalisé le machine learning pour les positions géographiques mais ma piste était d'utiliser un algorithme de clustering sur les positions en prenant en compte le type de crime (un algorithme qui me parait adapté : dbscan)

## Auto-évaluation :

Ce court de machine Learning m'a permis de progresser considérablement dans ce domaine que je ne connaissais que de nom et par ses applications. Je suis donc monté en compétences sur le langage de programmation python, l'utilisation de librairie comme scikitLearn et les algorithmes de machine Learning.

Néanmoins, j'ai encore beaucoup de mal à les mettre en application (surtout pour choisir quel algorithme utilisé et à quel endroit).

Il me reste à m'exercer et mettre en application chaque algorithme de façon concrète pour pouvoir « décoller ».

Durant le projet final, j'ai réalisé beaucoup de recherche mais je n'ai pas réussi à faire tout ce que je voulais (beaucoup d'idées mais difficultés à mettre en place)

Je pense mériter 14.85/20