Proceedings of Machine Learning Research 1–5

Prédiction des Crimes à New York

BELDI Chaima CHAIMA.BELDI@SUPCOM.TN
HAMHOUM Wissal WISSAL.HAMHOUM@SUPCOM.TN
MNASSER Mariem MNASSER.MARIEM@SUPCOM.TN

1. Résumé

Les activités criminelles préoccupent depuis longtemps tous les pays. L'analyse des données sur la criminalité a été un élément-clé, mais un défi considérable pour découvrir les tendances de la criminalité et réduire la criminalité. Au cours de la dernière année, en plus de l'élaboration de techniques de collecte et d'exploration de données, de nombreuses études liées aux mégas données ont été menées pour analyser les données sur la criminalité. Pour combler ces lacunes de recherche, cet article propose d'analyser les activités criminelles et de l'expérimenter sur New York. Le cadre proposé combine des algorithmes d'apprentissage automatique pour construire un modèle qui peut prédire la probabilité d'un crime peut se produire un certain jour et à un endroit spécifique en fonction de certaines caractéristiques comme l'âge, sexe, date, lieu, etc...
Mots clès: Prédiction, Crime, Machine Learning, New York City

2. Introduction

Les crimes sont des problèmes sociaux courants, dont aucune société n'est exempte. Ils affectent fortement la réputation et la croissance économique de tout pays. D'autre part, la sécurité est un facteur de décision lorsqu'il s'agit de déménager dans un nouveau lieu, de se déplacer à certaines heures et de visiter certaines zones. New York est l'un des États les plus attrayants et les plus célèbres des États-Unis. Comme il offre de nombreuses possibilités d'expérimenter l'art, la danse, la musique et le théâtre, il est l'une des destinations les plus recherchées par de nombreuses personnes. Cependant, il ne fait pas exception à ce phénomène social. En 2020, le taux de criminalité dans la ville de New York a augmenté de 11,2 % par rapport à l'année précédente [1]. Les crimes tels que les vols, les agressions criminelles et les vols de voitures ont augmenté respectivement de 15,8%, 13,8% et 15%. Ces statistiques ne peuvent qu'accroîtr e le malaise tant pour les habitants que pour les visiteurs de cet état,ce qui rend la mise en œuvre d'une solution qui peut aider à décider où et quand sortir à New York nécessaire pour réduire le nombre des victimes de ces crimes.
Et puisque, les événements criminels sont connus pour révéler des patterns spatio-temporels, l'utilisation de l'analyse de localisation renforcée par les algorithmes d'apprentissage automatique, peut créer la solution adéquate pour prédire les crimes et par conséquent les risques d'occurrence de ces évéments.
Dans cet article, nous expliquons en détails le pouvoir prédictif des facteurs de prédiction de la criminalité dérivés d'informations provenant de l'historique des complaintes de New York City. On commencera par étudier l'existant en donnant un bref aperçu de sur les modèles de prédiction des crimes existants dans la communauté d'exploration de données. Ensuite, nous détaillons les différentes étapes de l'élaboration de notre sollution. En commençant par l'exploratione et le nettoyage de données en corrigeant les valeurs manquantes et les enregistrements inexactes. Puis nous choisissons les les valeurs à prédire, en expliquant l'algorithme de prédiction choisi et en décrivant l'application côté client où l'utilisateur peut entrer l'âge le jour, le mois, le sexe, la race et l'emplacement après le modèle calculera la prévision et l'imprimera dans l'interface côté client. Finalement, on clôture avec une conclusion.

3. Etat de l'art

Les activités criminelles sont si courantes dans le monde entier que beaucoup de chercheurs ont réalisé de nombreux travaux sur ce sujet. Le sujet est profond de sort qu'il y avait eu plusieurs approches, certains ont eu recours à l'analyse des relations entre les activités criminelles et les variables socio-économiques telles que le chômage, la pauvreté, l'exclusion sociale et la pauvreté, d'autres se sont plutôt focalisé sur l'analyse des données spatio-temporelles pour déterminer les hot spots des crimes.

Parmi ces travaux nous puvons citer [2]. qui ont exploré la base de données des crimes à Chicago qui contient des information telles que le type de crime, la date, l'heure, la distance etc. Leur classifieur est créé en utilisant le KNN et plusieurs autres algorithmes d'apprentissage machine, et il a permis d'attendre une bonne précision.
En contre partie [3] se sont focalisés sur la prédiction des tendances des crimes en se basant sur la zone géographique et non pas et non pas sur la période.

Nous pouvons aussi mentionner l'approche de [4] qui ont opter pour les methodes de classification embarqués, où ils ont utilisé le Naive Bayes et les réseaux de neurones artificiels pour identifier un modèle des crimes, ce qui a permis de prédir le type de crime qui peut avoir lieux dans une zone spécifique avec plus de précision.

D'autre part [5], ont tiré avantage des données spatio-temporelles pour créer un outil interactif à l'aide de Google Maps, permettant la visualisation des patterns de crimes. Ils ont utilisé plusieurs algorithmes comme le KNN et la Naïve Bayes etc...

Nous pouvons remarquer dans les travaux cités que la majorité des solutions sont créées pour être utilisées par la police, les résultats sont certes bons mais ils peuvent être améliorés si on implique la partie la plus concernée par l'affaire qui est la potentielle victime. La victime porte plusieurs features (sexe, âge...) qui peuvent augmenter les risques de certains crimes et diminuer les chances pour d'autres. Les informations portées par la potentiel victime et les données spatio-temporelles combinées ensemble, peuvent aboutir à des solutions plus robutes et surtout utilisable par les citoyens.

4. Solution proposée

4.1. Exploration des données

L'exploration de données, ou Data Exploration, est la première de l'analyse de données. Elle consiste à explorer un large ensemble de données pour y découvrir des tendances, caractéristiques et corrélations à examiner plus en profondeur par la suite. On utilise diverses techniques statistiques pour définir les caractéristiques de l'ensemble de données : taille, quantité, qualité, nature...
Cette première exploration a pour but d'offrir une première vue d'ensemble sur les points d'intérêt d'un dataset et d'extraire d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Notre dataset (Table 1) comprend 6 982 505, lignes et 41 colonnes, tous les crimes, délits et infractions valides signalés au département de police de New York (NYPD) de 1028 à la fin de l'année dernière (2017).

Colonne	Description
CMPLNT_NUM	ID persistant généré aléatoirement
CMPLNT_FR_DT	Date exacte d'occurrence de l'événement signalé
CMPLNT_FR_TM	Heure exacte de l'événement signalé
CMPLNT_TO_DT	Date de fin de l'événement signalé
CMPLNT_TO_TM	Fin de l'heure de l'événement pour l'événement signalé
ADDR_PCT_CD	L'enceinte dans laquelle l'incident s'est produit
RPT_DT	L'événement de date a été signalé à la police
KY_CD	Code de classification des infractions à trois chiffres
OFNS_DESC	Description de l'infraction correspondant au code clé
PD_CD	Code de classification interne à trois chiffres (plus granulaire que KY_CD)
PD_DESC	Description de la classification interne correspondant au code
CRM_ATPT_CPTD_CD	Indicateur du succès de la criminalité
LAW_CAT_CD	Niveau d'infraction : felony, misdemeanor, violation
BORO_NM	Le nom de l'arrondissement dans lequel l'incident s'est produit
LOC_OF_OCCUR_DESC	Lieu précis de l'événement dans ou autour des lieux
PREM_TYP_DESC	Description spécifique des lieux
JURIS_DESC	Description of the jurisdiction code
JURISDICTION_CODE	Juridiction responsable de l'incident.
PARKS_NM	Nom du parc ou des espaces verts de New York
HADEVELOPT	Nom du développement immobilier NYCHA de l'occurrence
HOUSING_PSA	Code de niveau de développement
X_COORD_CD	Coordonnée X, Plan de l'État de New York (FIPS 3104)
Y_COORD_CD	Coordonnée Y, Plan de l'État de New York (FIPS 3104)
SUSP_AGE_GROUP	Groupe d'âge du suspect
SUSP_RACE	Description de la race du suspect
SUSP_SEX	Description sexuelle du suspect
TRANSIT_DISTRICT	district de transit dans lequel l'infraction s'est produite.
Latitude	Latitude (EPSG 4326)
Longitude	Longitude (EPSG 4326)
Lat_Lon	(Latitude, Longitude)
PATROL_BORO	Le nom de l'arrondissement de patrouille dans lequel l'incident s'est produit
STATION_NAME	Nom de la station de transport en commun
VIC_AGE_GROUP	Groupe d'âge de la victime
VIC_RACE	Description de la race de la victime
VIC_SEX	Description du sexe de la victime

Figure 1: Description des colonnes du Datasset

4.2. Nettoyage des données

Le Data Cleaning ou nettoyage de données est une étape indispensable en Data Science et en Machine Learning. Elle consiste à résoudre les problèmes dans les ensembles de données, afin de pouvoir les exploiter par la suite. Le Data Cleaning englobe plusieurs processus ayant pour but d'améliorer la qualité des

données. Il existe de nombreux outils et des pratiques permettant d'éliminer les problèmes dans un dataset. Nous avons opté à des processus qui servent à corriger et à supprimer les enregistrements inexacts dans la base de données. En plus on a identifié et remplacé des données incomplets, inexacts, corrompus et manquant de pertinence.

4.3. Elaboration du modèle

L'objectif principal est de former le modèle le plus performant possible en utilisant les données pré-traitées. Pour cela, il est nécessaire d'avoir une image claire des données. Il est donc important de :

• Catégoriser le problème ; les données sont étiquetées ou non et la sortie est un nombre, une classe ou un ensemble de groupes d'entrée.

• Comprendre les contraintes ; la capacité de stockage des données et la prédiction et la vitesse d'apprentissage souhaitées.

Nous avons utilisé un modèle séquentiel qui se compose de 5 couches dense dont 2 couches consistent l'input et l'output. La couche d'entrée se compose d'un vecteur d'entrées qui constitue le nombre de features ou colonnes correspondant à 9 features qui sont Latitude,Longitude, age, année, mois, jour et l'heure de l'incident . La profondeur des couches sont 500, 500, 200, 64 et 5.La fonction d'activation utilisée est Relu. La couche de sortie se compose de 6 classes identifiant le type de crime basé sur KYCD. La fonction d'activation Softmax est utilisée pour classifier les crimes. On a choisi ADAM comme optimiser et 256 comme la taille du lot avec la fonction de perte categoricalcrossentropy suite à la nature de notre sortie affiché sur 6 classes qui sont Drugs crimes, Killing crimes, Violent crimes, Sexual crimes, Theft crimes, Other type of crimes. Le modèle est representé dans la figure 9.

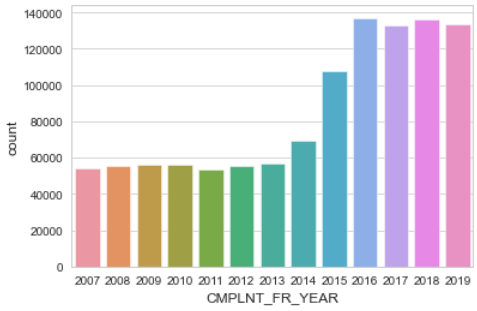

Figure 2: Nombre d'infractions par ans

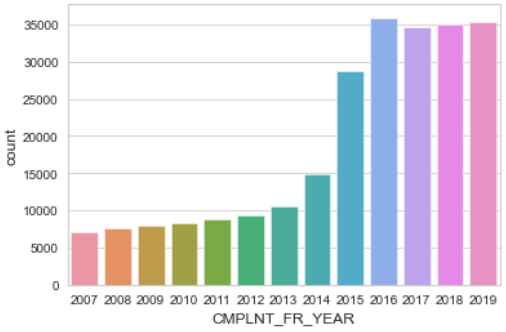

Figure 3: Nombre de Crimes par ans

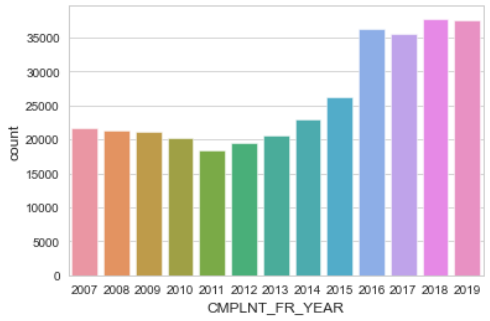

Figure 4: Nombre de Violations par ans

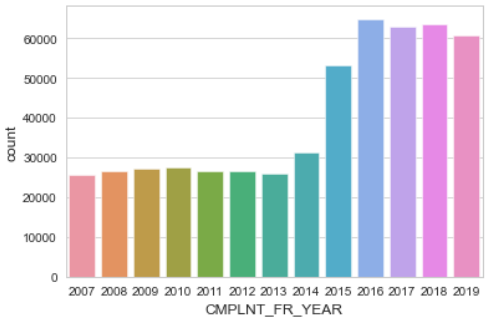

Figure 5: Nombre de Délits par ans

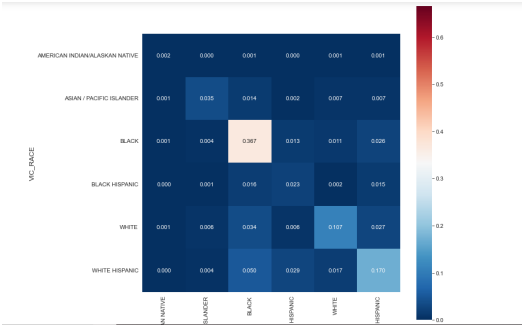

Figure 6: Corrélation entre la victime et suspect

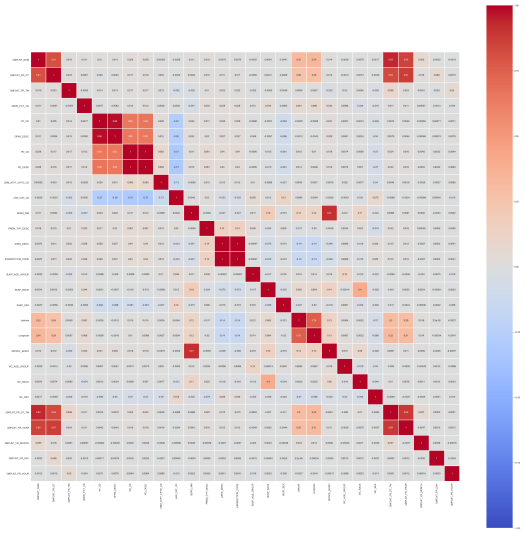

Figure 7: Corrélation des données

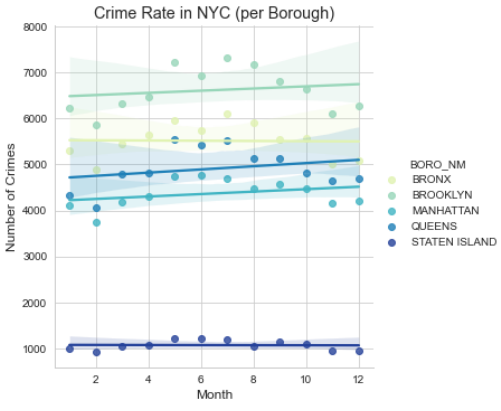

Figure 8: Taux de criminalité mensuel pour les arrondissements

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_10 (Dense)             (None, 500)               5000
_________________________________________________________________
dense_11 (Dense)             (None, 500)               250500
_________________________________________________________________
dense_12 (Dense)             (None, 200)               100200
_________________________________________________________________
dense_13 (Dense)             (None, 64)                12864
_________________________________________________________________
dense_14 (Dense)             (None, 6)                 390
=================================================================
Total params: 368,954
Trainable params: 368,954
Non-trainable params: 0
_________________________________________________________________
```

Figure 9: Modèle

4.4. Résultat et Visualisation

La précision de notre modéle est de valeur 54.98% lors de l'évaluation comme affiché dans la figure 10. La prédiction du type de crime est affiché sur l'application client dévéloppée par Flask accompagné avec openLayers. Pour la prediction, l'utilisateur doit se localiser, sélectionner son age, date et heure et son sexe, lors de la prediction la résultat s'affiche dans un popup comme montré dans la figure 11.

```
# Evaluate model on test data
scores = model.evaluate(X_valid, y_valid)
print("\n%s: %.14f%%" % (model.metrics_names[1], scores[1]*100))

3454/3454 [==============================] - ETA: 0s - loss: 1.0566 - accuracy: 0.54 - 5s 2ms/step - loss: 1.0565 - accuracy: 0.5489

accuracy: 54.898817403793334%
```

Figure 10: Précision du modèle

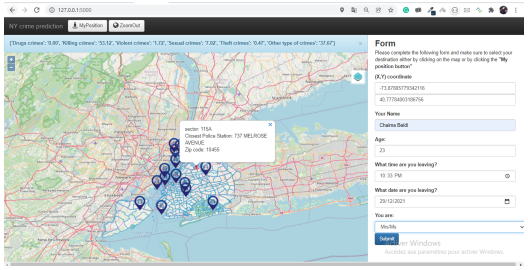

Figure 11: Résultat

4.5. Application coté client

Le modèle d'apprentissage machine déjà décrit précédemment , est servi aux utilisateurs grâce à une application web. Sa partie frontend est developpée en HTML, CSS et javascript alors que la partie back-end est en python Flask. L'application utilise aussi la bibliothèque openLayers, qui est une bibliothèque javascript opensource [], pour l'affichage de la carte et

le traitement des données géographiques. L'interface de l'application est divisé en 2 partie:

- la partie à gauche affiche la carte de New york avec les limites de chaque secteur de cet état ainsi que des marques représentant la localisation des stations de police.
- La partie à droite contient un formulaire à remplir par l'utilisateur. Les informations récoltées de ce formulaire servent par la suite comme input du modèle de prédiction.

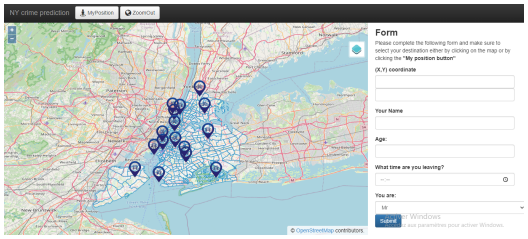

Figure 12: Interface principale

L'application offre la possibilité d'effectuer les prédictions en tenant compte de ses informations spatio-temporelles actuelles, en cliquant sur le bouton "MyPosition" qui le localise et prend ses coordonnées géographiques. L'utilisateur peut également consulter des prédictions sur une destination et un heur qu'il précise lui-même en cliquant sur la carte et remplissant le champ "What Time are You Leaving" dans le formulaire.

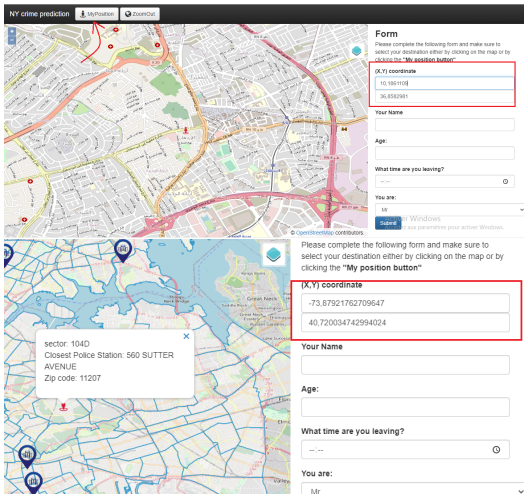

Figure 13: Remplir les champs de X et Y

5. Conclusion

La prédiction des crimes est l'un des sujets de recherche les plus importants de ces dernières années qui visent à protéger la vie des gens. Ces études analytiques pour les points chauds criminels sont fréquemment demandées par les organismes d'application de la loi. Dans notre approche, nous avons remédié ce problème en faisant une analyse des données et en élaborant un modèle pour détecter les crimes à New York. L'utilisateur peut utiliser l'application Web pour vérifier la probabilité qu'un crime puisse se produire ou non basé sur un algorithme d'apprentissage automatique. Le résultat peut être considérablement amélioré en prenant en compte d'autres entrées a notre modèle pour alimenter le modèle.

References

[1] FOX 5 NY Staff. *Violent crime continues to surge in NYC*. (Visited on 07/23/2021).

[2] Alkesh Bharati and Dr Sarvanaguru RA.K. "Crime Prediction and Analysis Using Machine Learning". In: (2018).

[3] Vijaya Pinjarkar Ankit Sangani and Chirag Sampat. "Crime Prediction and Analysis, 2nd International Conference on Advances in Science Technology." In: (2019).

[4] Ayisheshim Almaw and Kalyani Kadam. "Survey Paper on Crime Prediction using Ensemble Approach". In: (2017).

[5] Bhavna Saini Hitesh Kumar Reddy ToppiReddy and Ginika Mahajan. "Crime Prediction Monitoring Framework Based on Spatial Analysis". In: (2018).