IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

A journey to data scientist - UE coeur BC - TAF Data Sciences

Intitulé du projet :Analyse de l'Impact de l'Implantation des Écoles sur la Criminalité

“Analyse Prédictive : Impact des innovations éducatives sur les taux de criminalité à Chicago”

Groupe 12

Soumiya RAZZOUK
Chenjie QIAN
Yann LEGENDRE
Hicham CHEKIRI

I - Choix du modèle :

Variables :

Nous rappelons que les variables suivantes ont été utilisées comme entrées pour nos algorithmes de prédiction. Ces variables sont résultantes d'un codage (one-hot), transformation d'une variable existante, ou création à partir des autres features.

After_School_Hours : entier (minutes)
School_Hours : entier (minutes)
Dress_Code : booléen
Student_Count_Total : entier
Is_High_School : booléen
Is_Middle_School : booléen
Is_Elementary_School : booléen
Is_Pre_School : booléen
Count_High_School_Near : entier
Count_Middle_School_Near : entier
Count_Elementary_School_Near : entier
Count_Pre_School_Near : entier
Crime_Count : entier

Construction des bases de données :

Pour notre étude, nous nous intéressons à voir l'impact d'une école sur l'évolution de taux de criminalité dans son entourage délimité par un cercle de périmètre = 1 km dans un an, deux ans, trois ans, …. , sept ans.

Pour cela, nous avons construit des combinaisons de base de données comme suit :
df1 : écoles 2016 - crime 2017
df2 : écoles 2016 - crime 2018
df3 : écoles 2016 - crime 2019
df4 : écoles 2016 - crime 2020
df5 : écoles 2016 - crime 2021
df6 : écoles 2016 - crime 2022
df7 : écoles 2016 - crime 2023

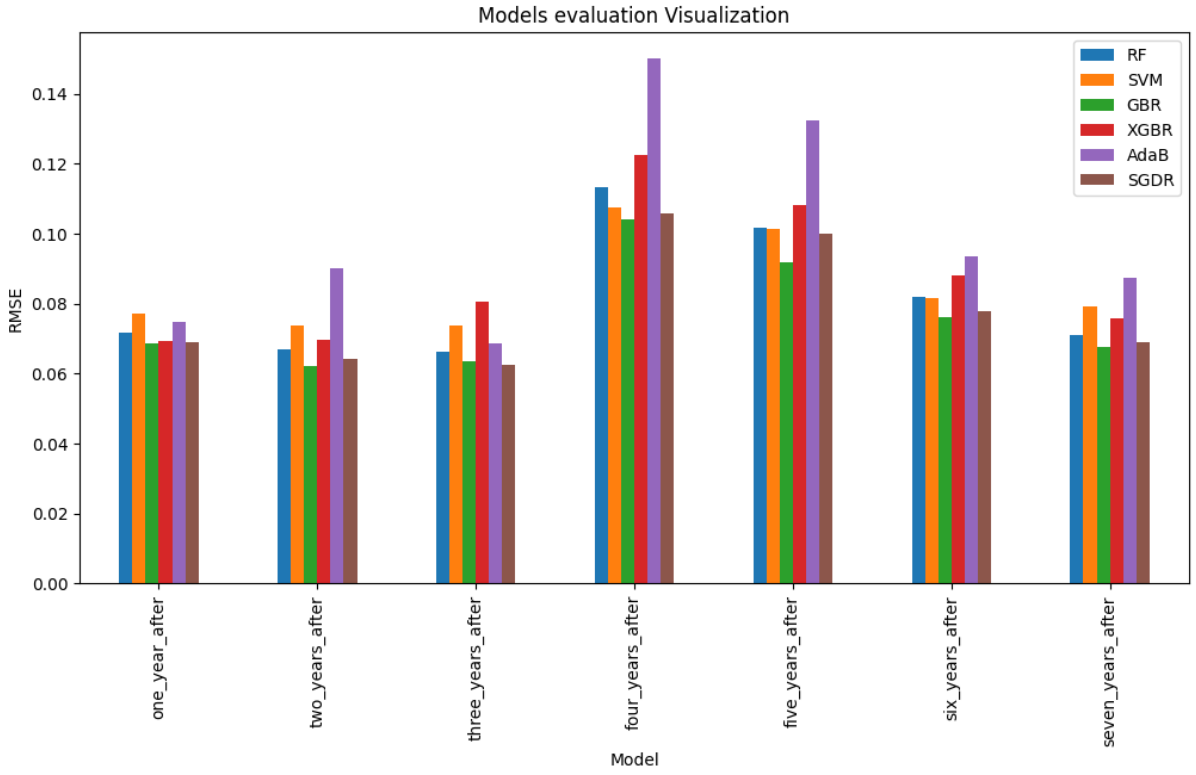
Les algorithmes utilisés :

Pour faire ces prédictions, et puisque le but est de faire un apprentissage supervisé sur la variable cible **nombre de criminalité** (*crime_count*), nous avons sélectionné des algorithmes de machine learning qui vont nous permettre à la fois de faire les prédictions et de révéler les liens, entre variables, non explorés par le calcul des corrélations.

- **Random Forest Regressor**
- **SVM Regressor**
- **Gradient Boosting**
- **XGBoost**
- **AdaBoost**
- **SGDRegressor**

Afin de maximiser la performance des algorithmes, nous avons testé les 6 modèles sur nos 7 dataset, et nous avons retenu les algorithmes avec le minimum MAE et RMSE.
Notre modèle alors, est une combinaison des algorithmes qui ont montré la meilleure performance (pour prédire un an après, … , 7 ans après).

Le tableau ci-dessous résume les RMSE et MAE des algorithmes retenus pour chaque df que nous avons nommé (**Model_1, … , Model_7**).



Et voici les algorithmes que nous avons sélectionné :

df1 (2016 - 2017) : Gradient Boosting
df2 (2016 - 2018) : Gradient Boosting
df3 (2016 - 2019) : SGDRegressor
df4 (2016 - 2020) : Gradient Boosting
df5 (2016 - 2021) : Gradient Boosting
df6 (2016 - 2022) : Gradient Boosting
df7 (2016 - 2023) : SGDRegressor

Modèle	MAE	RMSE
Model_1	0.067998	0.261042
Model_2	0.062460	0.249750
Model_3	0.062500	0.250000
Model_4	0.102262	0.319701
Model_5	0.099939	0.303650
Model_6	0.077846	0.276445
Model_7	0.068611	0.261938

Validation croisée des algorithmes :

Les 7 algorithmes ont été entraînés sur des database de presque 655 lignes chacune, il était intéressant de tester la validation croisée pour vérifier que les modèles ne sont pas en train de surajuster les données d'entraînement. Et voici les résultats :

Modèle	MAE	RMSE
Model_1	0.0610	0.0892
Model_2	0.0593	0.0878
Model_3	0.0646	0.0896
Model_4	0.1089	0.1471
Model_5	0.1001	0.1353
Model_6	0.0795	0.1117
Model_7	0.0756	0.0994

```
for i, df in enumerate(normalized_dfs):
    # Separate features (X) and the target variable (y)
    X = all_data[df].drop([target_variable, 'Crime_level'], axis=1)
    y = all_data[df][target_variable]

    model = best_models[i]

    # K-fold cross-validation
    k_fold = KFold(n_splits=15, shuffle=True, random_state=42)

    # Evaluate the model, use RMSE
    rmse_scorer = make_scorer(lambda y_true, y_pred: np.sqrt(mean_squared_error(y_true, y_pred)), greater_is_better=False)
    cross_val_results_rmse = cross_val_score(model, X, y, cv=k_fold, scoring=rmse_scorer)

    # Evaluate the model, use MAE
    mae_scorer = make_scorer(mean_absolute_error, greater_is_better=False)
    cross_val_results_mae = cross_val_score(model, X, y, cv=k_fold, scoring=mae_scorer)

    # Output the results
    print(f'Cross validation df{i+1}')
    print(f'Model used: {model}')
    print(f'Average RMSE: {-1*cross_val_results_rmse.mean():.4f}')
    print(f'Average MAE: {-1*cross_val_results_mae.mean():.4f}')
    print('='*50)
    print()
```

Feature importance :

Une analyse de l'importance des variables fourni par nos modèles et en tenant compte des faibles corrélations entre les variables, montre que :

- Le taux de criminalité (***crime_count***) est plus associé à la présence d'une école dans une zone (délimitée par un cercle de périmètre 1 km), et des écoles à côté qu'aux caractéristiques intrinsèques à l'écoles (Dress code, school hours, after school hours).

```
Feature Importances:
                        Feature  Importance
7             Student_Count_Total    0.336039
8          Count_High_School_Near    0.136003
5                    School_Hours    0.103437
10  Count_Elementary_School_Near    0.098650
9        Count_Middle_School_Near    0.092297
11          Count_Pre_School_Near    0.075218
4              After_School_Hours    0.073326
6                      Dress_Code    0.026233
3                  Is_Pre_School    0.023502
1               Is_Middle_School    0.017082
2            Is_Elementary_School    0.009705
0                  Is_High_School    0.008509
```

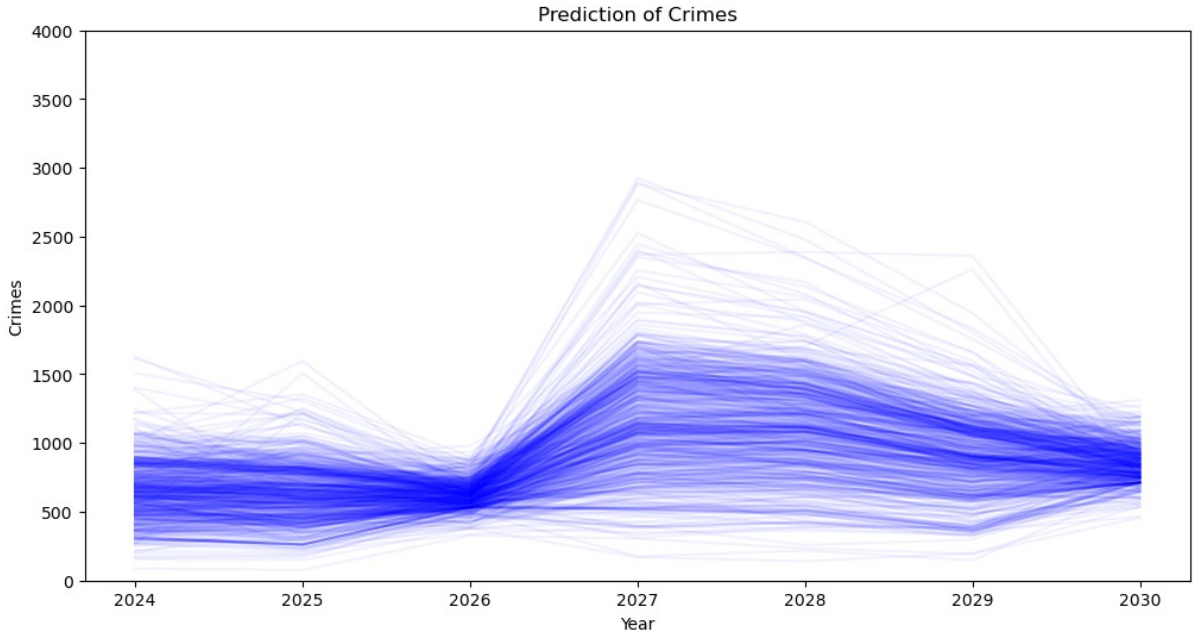
II - Prédictions et résultats du modèle :

Il est important de rappeler que notre modèle ainsi construit permet de faire :

- **1-** une simple prédiction sur le taux de criminalité sur 7 ans à venir en prenant en compte les écoles qui existent déjà. Exactement, on peut prédire des crimes en (2024, …. , 2030) en prenant en entrée les données d'école en 2023.

Le tableau montre un exemple du résultat obtenu. Chaque ligne correspond à une école.

2024	2025	2026	2027	2028	2029	2030
88	77	328	305	234	195	456
314	374	595	567	541	540	822
796	792	689	1441	1291	1310	985
548	743	708	1245	983	875	989
1508	1306	823	2895	2352	1757	1100
...	...	...	...	...	...	...
369	468	505	691	688	736	731
856	777	788	1511	1405	1159	1086
361	309	662	557	503	386	918
561	515	611	979	1010	860	822
742	924	701	1697	1667	1221	957

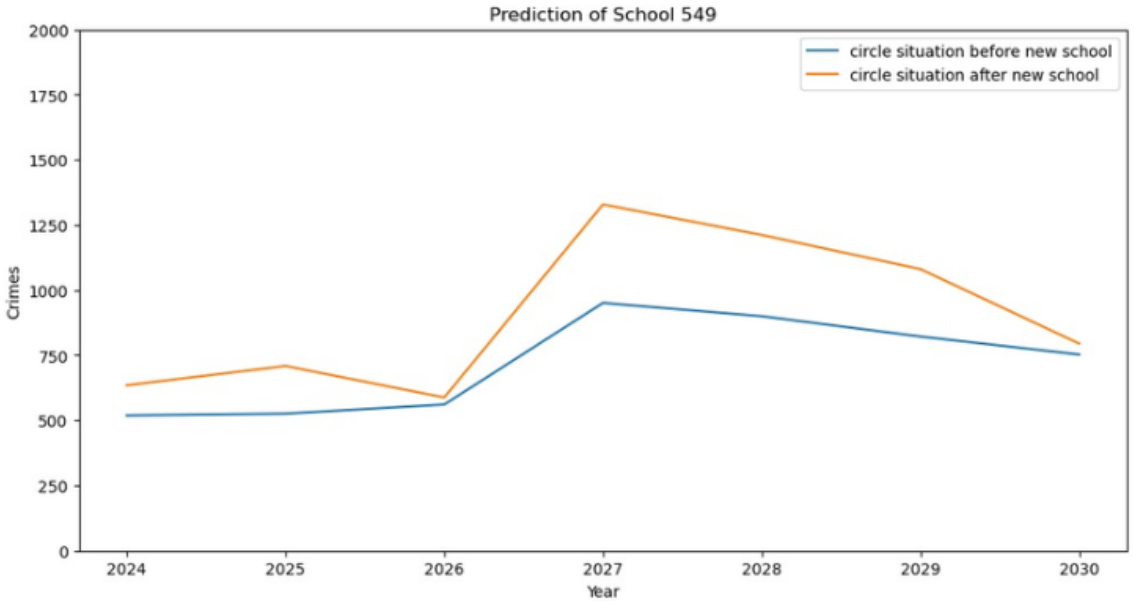


- **2-** simuler l'implantation d'une nouvelle école dans une zone précise (***longitude, latitude***) et ses caractéristiques, mesurer l'impact dans 7 ans à venir de cette implantation sur le cercle de l'école et les écoles à côté (en se concentrant sur les intersections des cercles). Plus précisément, on s'intéresse à voir l'état du cercle de 1 km avant et après l'introduction de cette école. Pour montrer comment notre application fonctionne, nous allons montrer un exemple.

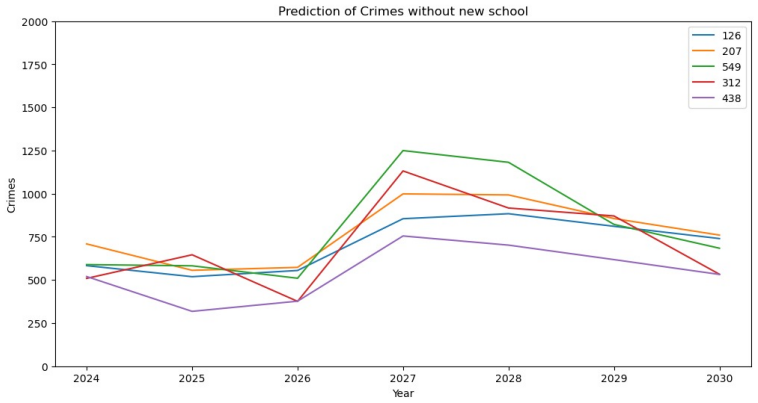
Considérons l'école suivante qu'on lui attribue l'ID 549 appelée ***new_school***

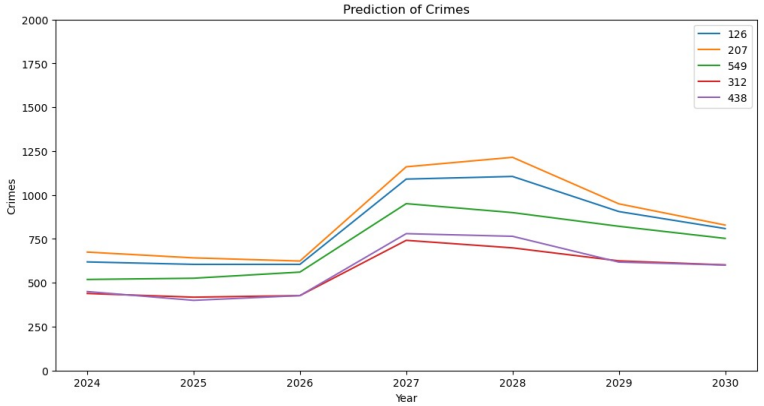
School_Latitude	School_Longitude	Is_High_School	Is_Middle_School	Is_Elementary_School	Is_Pre_School	After_School_Hours	School_Hours	Dress_Code	Student_Count_Total	Count_High_School_Near	Count_Middle_School_Near	Count_Elementary_School_Near	Count_Pre_School_Near
41.858374	-87.668724	0	0	1	0	20	470	1	550	1	5	5	3

En se basant sur sa position, le modèle nous permet de faire une comparaison de la situation des crimes avec et sans cette école dans cette position. Et en plus, voir les écoles à côté qui seront impactés et modéliser cet impact. Et voici le résultat :

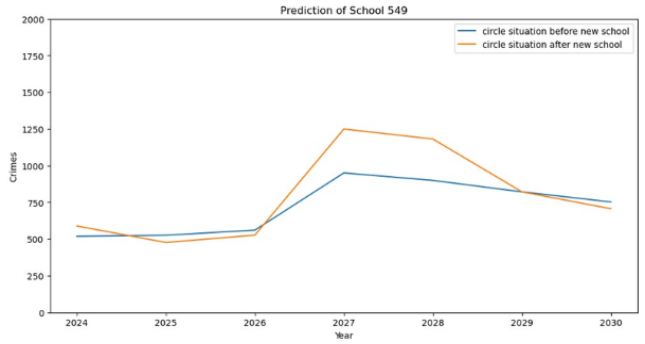


Et voici l’impact qu’aura cette école sur les écoles qui l'entourent :

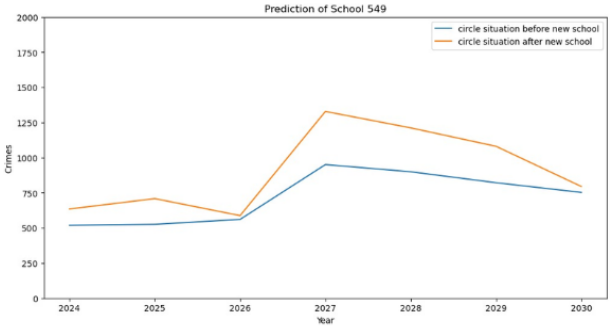




Pour une deuxième simulation, nous avons gardé la même école en augmentant la valeur de la variable after school hours (devenue égale à 60 minutes) pour vérifier s'il l'impact de cette feature est mesurable. Et voici, le résultat obtenu :



after school hours : 60 min



after school hours : 20 min

Pour plus de visibilité, voici un tableau qui permet de comparer la variation du taux de crime entourant cette école (***new_school***) avec un programme après cours de 20 et 60 minutes.

After School Hours	2024	2025	2026	2027	2028	2029	2030
60 min	13,49%	-9,31%	-3,03 %	31,44 %	31,33%	0%	-3,03 %
20 min	22,35%	34,79 %	4,81% %	39,75 %	34,67%	31,51%	5,58%

On remarque qu'une augmentation de ***after school hours*** de 20 minutes à 60 minutes permet une diminution de l'impact mesuré sur son entourage.

III - Conclusion :

Les résultats obtenus à travers notre modèle fournissent une base solide pour les décideurs locaux de Chicago afin de prendre des décisions en ce qui concerne la construction d'écoles dans des zones spécifiques. Les corrélations entre les écoles et le taux de criminalité sont relativement faibles, ce qui souligne l'importance de considérer divers facteurs lors de la planification de nouvelles infrastructures éducatives (postes de polices, état du quartier …) et de prendre en considération les indicateurs démographiques, la situation financière des étudiants recrutés par chaque école pour plus d'alignement avec les articles sur lesquels nous avons basés nos hypothèses.

Notre modèle nous a permis de valider l'hypothèse selon laquelle les programmes parascolaires peuvent influencer le taux de criminalité, même si leur impact est assez faible et n'est pas le principal facteur d'influence. En fournissant des prédictions sur le taux de crime, il est intéressant de développer notre modèle de façon à ce qu'il nous rapporte une idée sur la répartition géographique de ces crimes autour de l'école.

L'application de la méthodologie CRISP-DM s'est révélée particulièrement bénéfique tout au long du processus, nous guidant de manière structurée depuis la phase initiale de définition de la problématique jusqu'à l'analyse approfondie des résultats obtenus. Cette approche méthodique a grandement facilité la gestion et l'optimisation du flux de travail, renforçant ainsi notre capacité à prendre des décisions éclairées tout au long du projet. En somme, l'expérience de mettre en œuvre la méthode CRISP-DM a enrichi notre compréhension des processus analytiques et a contribué de manière significative à la réussite globale de cette initiative.