Université de Montréal

Proposition du projet final: Crime Predictor

Groupe
Simon Janssen
Reda Kzaz
Kevin Lessard

Département de mathématiques et de statistique
Faculté des arts et des sciences

Travail présenté à Guy Wolf
Dans le cadre du cours STT3795
Fondements théoriques en science des données

February 26, 2022

Optimisation des activités de patrouilles policière en fonction des points chauds et la criminalité

Description et objectifs du projet

La SPVM en collaboration avec la ville de Montréal a rendu disponible en 2016 un jeu de données portant sur les crimes sur l'île de Montréal. Les crimes qui y sont reportés sont les introductions par effraction, les vols de véhicules, les vols dans les véhicules, les vols qualifiés, les méfaits, les accidents de circulation et les infractions entraînant la mort.

Une première analyse des données pourrait être d'analyser les taux de criminalité et les tendances sur une zone ou une période spécifique. En effet, il serait intéressant de trouver des motifs dans les données en fonction du moment de la journée, de l'endroit, du type de crime, du jour de la semaine, etc..

Ainsi, il serait possible de faire plusieurs types de prédictions sur ces données. Nous pouvons prédire dans le temps, selon le moment, les endroits où il y aurait des crimes ou selon le moment, la distribution des types de crimes. Nous pouvons aussi essayer de prédire dans l'espace, pour chaque endroit, les moments avec les plus hautes probabilité de crime et quel type de crime se passerait. Nous pourrons aussi tenter d'appliquer ces modèles afin d'optimiser les activités des patrouilles policières.

Taches du projet ainsi que les contributions prévues de chaque membre de l'équipe

Dans un premier temps on pourra définir la résolution du projet en plusieurs Taches:

- **Tache 0:** Extraire les données.
- **Tache 1:** Nettoyer et Transformer les données, ce qui va permettre d'avoir une analyse plus exacte de ces dernières ainsi qu'une utilisation plus pointue des données dans notre/nos modèles de prédiction.
- **Tache 2:** Analyser les données et en retirer des interprétations.
- **Tache 3:** Définir notre modèle de prédiction.
- **Tache 4:** Entraîner notre modèle de prédiction.
- **Tache 5:** Tester notre modèle de prédiction.
- **Tache 6:** Produire les résultats de nos prédictions en incluant des visualisations qui permettront concrètement d'afficher les résultats de notre travail.

Les membres contribueront également en effectuant différentes tâches sur la même étape du projet. Nous ferons des réunions fréquentes pour s'assurer que nous sommes tous au même niveau. Ainsi, chacun de nous aura une bonne compréhension de chaque partie du projet et pourra se concentrer sur les aspects techniques de manière autonome. Nous considérons diviser les tâches de recherches et de développements adéquatement entre les membres. Une partition initiale des taches serait comme suit:

- **Simon:** Entraînera et testera le modèle
- **Reda:** Produira les résultats et des visualisations
- **Kevin:** Préparera les données pour le projet

La tâche de concevoir et définir le modèle et l'écriture du rapport sera fait par tous les membres ensemble.
P.S: Sachez que les taches ainsi que les répartitions pourraient être sujette à modification.

Gestion de portée

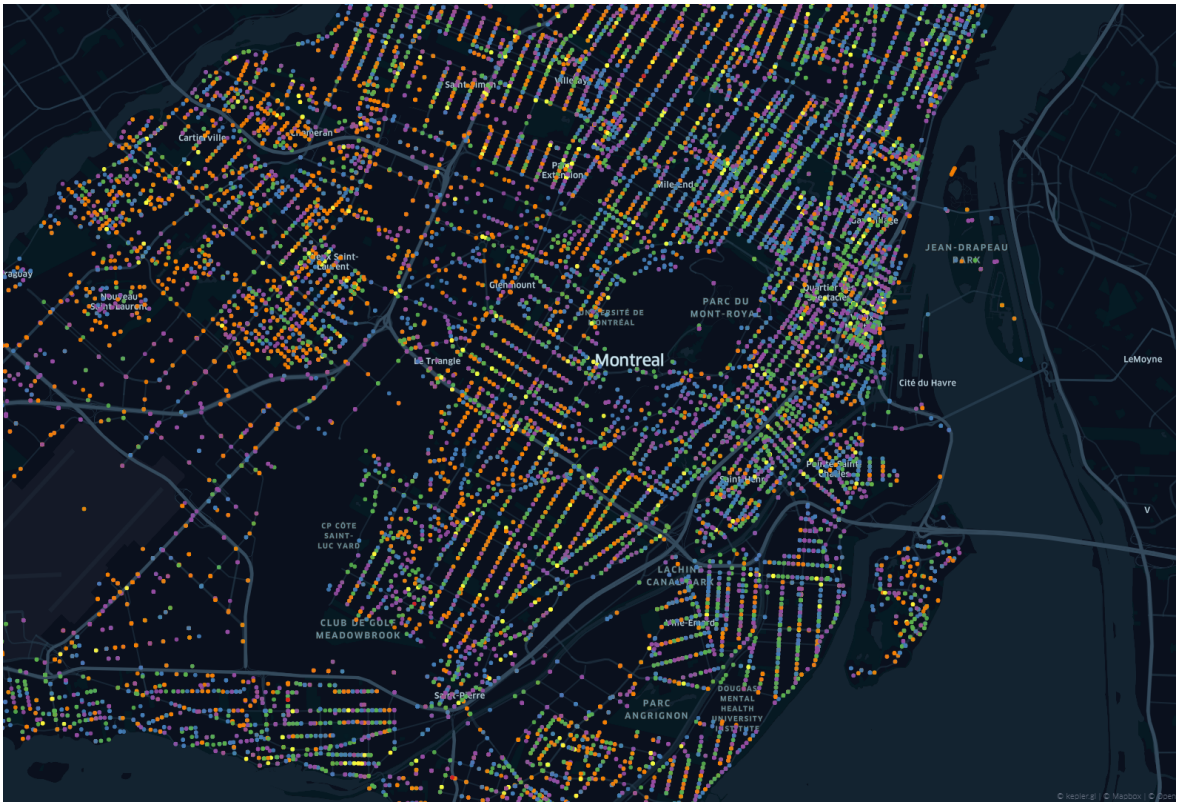
Nous avons écris que nous allons entraîner et tester un modèle dans la liste de tâches. Il se peut cependant que nous en faisons quelques-uns de plus si la portée du projet nous le permet.

Données utilisées et sources des données

Les données que nous allons utiliser pour le projet sont les données ouvertes de la ville de Montréal qui sont mises à jour quotidiennement. Les données sont disponibles ici.

Les données sont obtenues sous forme tabulaire. Chaque entrée a 8 attributs, 3 de type qualitatif (1 ordinal et 2 nominaux) et 5 de type quantitatif (attributs d'intervalles). Chaque entrée dans les données est attribuée une catégorie qui décrit le type de crime, une date du moment du crime, un quart de jour où c'est déroulé le crime (jour, soir ou nuit), le poste de police du quartier où c'est déroulé le crime, et les coordonnées géographiques de l'intersection le plus proche du crime. Les données sont donc de nature géographique et temporelle. Nous pourrions imaginer un espace trois dimensionnel dont deux axes correspondent aux coordonnées et le troisième ace correspond au temps. Dans cet espace sont les points de donnée.

On peut avoir une idée de l'ampleur des données avec cette visualisation sur Kepler.gl.



Technologies:

- On va principalement utiliser Python pour remplir quasiment toutes les taches mentionnées ci-dessus. Plus précisément, python va nous permette d'accéder a des librairies comme Pytorch pour la partie modèles de prédiction.
- Pour la partie Traitement de données, notre choix est tombe sur pyspark(Apache Spark sur python) afin de gérer nos données de façon optimale(100x plus rapide que Pandas)
- Github: afin de structurer nos échange de code.
- VS Code: pour IDE.
- Kepler.gl: pour les visualisations

Annexe

Liens utiles et possibles références modifiables par tous ici.