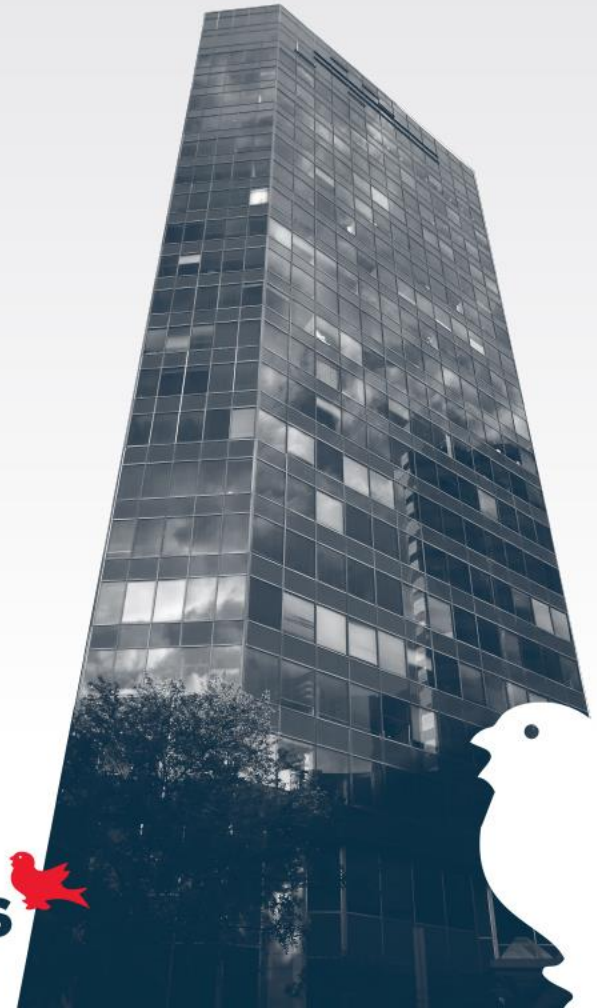
Data Science Capstone Project

Summer 2021

Team #02:

Zhongdi Wang
Vy Kiet Le



McGill | School of Continuing Studies | École d'éducation permanente

mcgill.ca/scs

Table of Contents

Problem Statement

Context

The city of Montreal is tasked with interventions on many levels to help protect lives and properties. One of the events that cause the most damage is undoubtedly when an apartment of a building catches fire and spreads it to other neighboring units.

Along with the city of Montreal, the firefighters of the SIM and the managers in charge of the fire prevention planning want to predict fire risk and distribute resources to areas in the city that most need it.

Problem

One of the biggest challenges that the city has faced in the fire prevention is to create the right model that would predict high fire risk areas and allocate the necessary resources to those areas for fire inspections and for taking appropriate fire intervention measures.

Solution

The goal of our team project is to predict fire risk level for each area delimited by the administrative limits in the city of Montreal for the next months.

Our solution approach

Data source

Dataset A: Interventions

Dataset B: Fire stations

Dataset C: Crime Incidents

Dataset D: Property Assessment

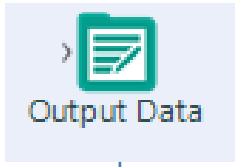
Dataset E: Administrative Limits of Montreal Area

Data preparation

alteryx

- Explore and analyze the data for outliers
- Add new field to indicate which intervention is considered as Fire incident based on description
- Aggregate data based on areas in Montreal city and on a monthly basis
- Merge datasets

Output merged file



Modelling

python

- Features selection
- Evaluate model performances
  - Model 1 : k-nearest neighbors
  - Model 2 : Support Vector Machine
  - Model 3 : Random Forest
  - Model 4 : XGBoost
- Predict fire risk level in each area

Visualization

+ableau

- Visualize fire risk level prediction for a given area of Montreal for the next months

Data exploration and cleaning

	Features				Response
Dataset	**Interventions 2015-2021**	**Crime Incidents 2015-2021**	**Property Assessment**	**Administrative limits of Montreal Area**	**Fire incidents 2015-2021**
Fields used	Monthly count of interventions in the previous month Month of incident	Monthly count of crimes in the previous month	Number of floors above ground level Number of apartments in the property Average Property Age Total Building Area Total Land Area	Code marmot for each administrative area in Montreal	Monthly count of interventions that resulted in Fire
Records	741,157	196,469	497,101	34	
% records retained	100% (dropped 3 records where area is not provided)	83% (dropped records where location is not provided)	66% (dropped records where building area is not provided)	100%	

Datasets used and features retained for modelling

Crime Incidents 2015-2021:

We have included crime incidents in our model because we believe that a fire can be a result of any type of crime incidence.

Property Assessment:

We have also included details on the properties in the city of Montreal because we believe that fire occurrence is relatively higher in administrative areas with huge property land area and higher number of accommodations.

Fire incidents 2015-2021:

Among the types of interventions classified by the city, we refer to a fire incident when the intervention is described either as "*Incendie*" or "*Autre feu*".

Feature engineering

Feature engineering:

- **Average Property Age**:

For each administrative area in Montreal city, we have calculated the average life of all the buildings and properties within that area.

- **Month of incident**

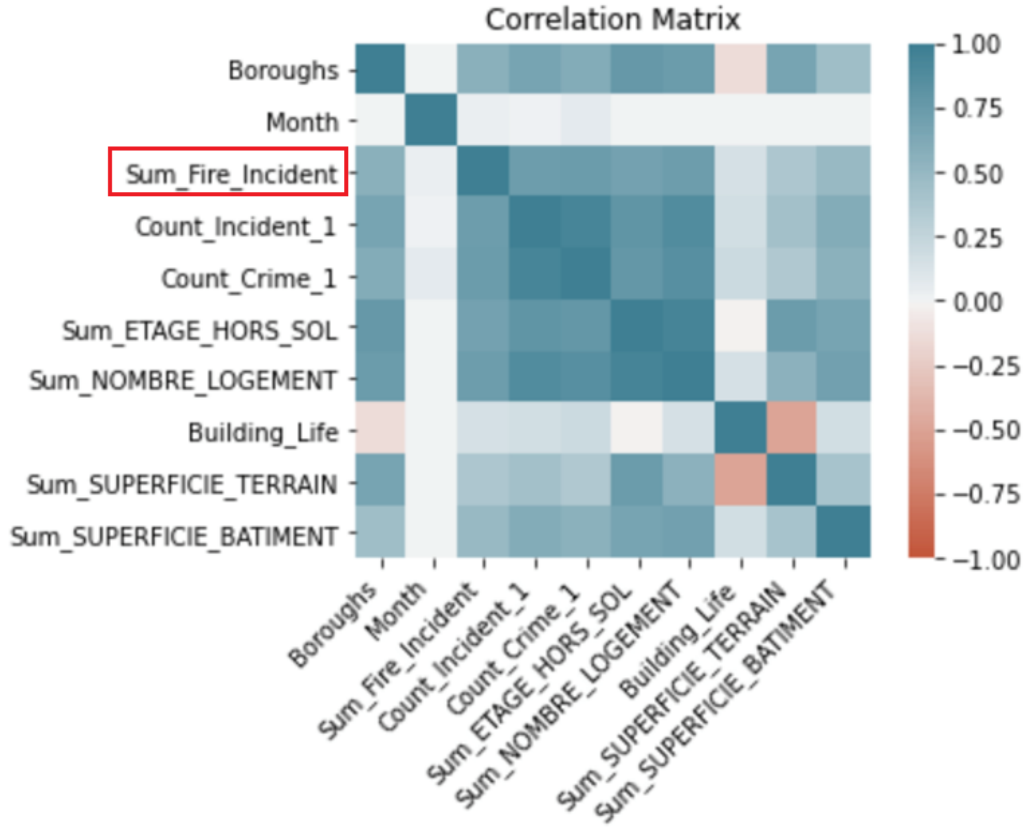
We add a new field for the month of the incidents (intervention, crime) to aggregate the data based on borough and month/year.

- **Lag features (interventions, crimes):**

We have introduced two variables, which are the counts of interventions and of crimes in the previous month. Under the belief that places with a history of interventions and crimes, fire incidents can have a higher probability of occurring in the future.

Correlation Matrix

High correlation between the fire incidents count and the variables such as number of interventions in the previous month, number of crimes in the previous month, number of building floors above ground as well as number of accommodations.



Modelling techniques

- **Categorization of the fire risk**: based on number of fire incidents per area per month, such that we get approximate same number of fire incidents across the levels

Count of Fire incidents	Fire risk level
[0, 2]	1 (low)
[3, 15]	2 (medium)
>15	3 (high)

- **Simple encoding**: convert areas/boroughs into numerical value before feeding the feature in the models

- **Normalization**: bring all numerical fields to use a common scale

Summary of the models and their results

Model	Decision Trees (DT) (Baseline)	Random Forest (RF)	K-Nearest Neighbors (KNN)	Support Vector Machine (SVM)	XGBoost (XGB)
Accuracy	• 74.4% with an R-score of 0.528 for a model based on 3 features	• 78.4% with an R-score of 0.588 for a model based on 5 features	• 78.4% with an R-score of 0.588 for a model based on all 9 features	• 70.1% with an R-score of 0.488 for a model based on 4 features	• 80.5% with an R-score of 0.627 for a model based on 7 features
Advantages	• Generally fast training time and simple tool.	• Generally fast training time and simple tool. • Decorrelate trees to reduce variance of predictions.	• No training period before making predictions • Easy to implement	• Works well if there is clear margin of separation between classes	• XGBoost performs very well on small dataset with subgroups and structured datasets with not too many features
Disadvantages	• Can overfit data and slow down predictions if there is a large number of trees in the algorithm	• Can overfit data and slow down predictions if there is a large number of trees in the algorithm	• Does not work well with large datasets (slow performance) • Does not work well high dimensions (features)	• Not suitable for large data sets • Does not perform well if target classes (fire risk levels 1-3) overlap	• Does not perform so well on sparse and unstructured data

Model used for prediction: XGBoost

We initially used all the 9 features that we believed are relevant in predicting fire risk level. The accuracy achieved was 79.3% with an R-score of 0.604.

After parameter tuning and using *sklearn.feature_selection.SelectFromModel* in Python, we reduce from 9 features to 7 based on the feature importance to the model. The accuracy achieved was then 80.5% with an R-score of 0.627.

The 7 features retained by the model are:

- Month
- Number of interventions
- Building Age
- Number of crimes
- Total number of accommodations
- Boroughs
- Total number of building floors

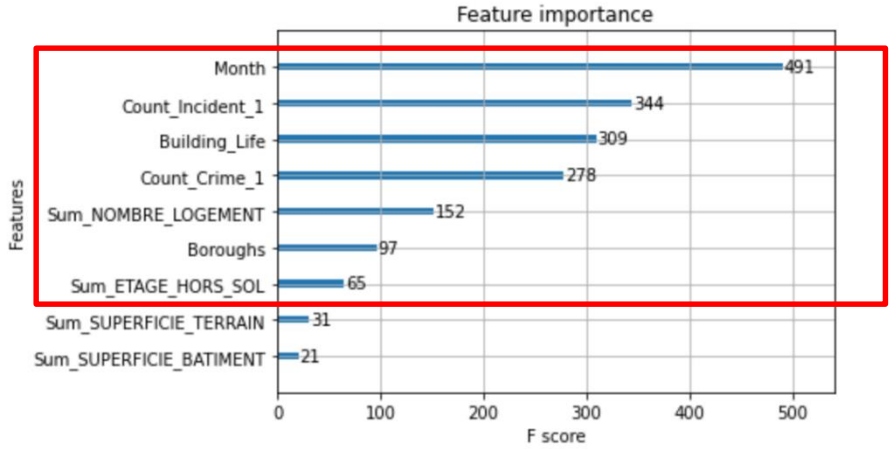
Evaluation metrics

XGB (9 features)

```
[[215  52   2]
 [ 53 275  31]
 [  0  29 149]]

Accuracy:  0.792803970223325

R Square Score: 0.6038704435498762

Classification Report:
              precision    recall  f1-score   support

           1       0.80      0.80      0.80       269
           2       0.77      0.77      0.77       359
           3       0.82      0.84      0.83       178

    accuracy                           0.79       806
   macro avg       0.80      0.80      0.80       806
weighted avg       0.79      0.79      0.79       806
```

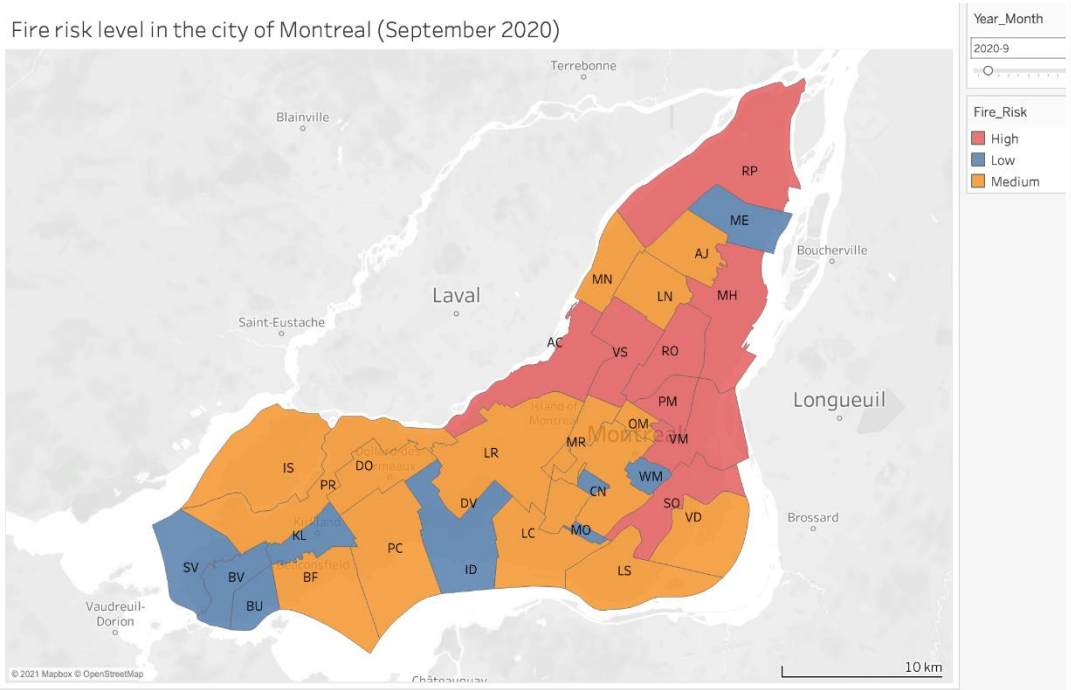
XGB (7 features)

```
[[215  52   2]
 [ 47 285  27]
 [  0  29 149]]

Accuracy:  0.8052109181141439

R Square Score: 0.6267681057724267

Classification Report:
              precision    recall  f1-score   support

           1       0.82      0.80      0.81       269
           2       0.78      0.79      0.79       359
           3       0.84      0.84      0.84       178

    accuracy                           0.81       806
   macro avg       0.81      0.81      0.81       806
weighted avg       0.81      0.81      0.81       806
```

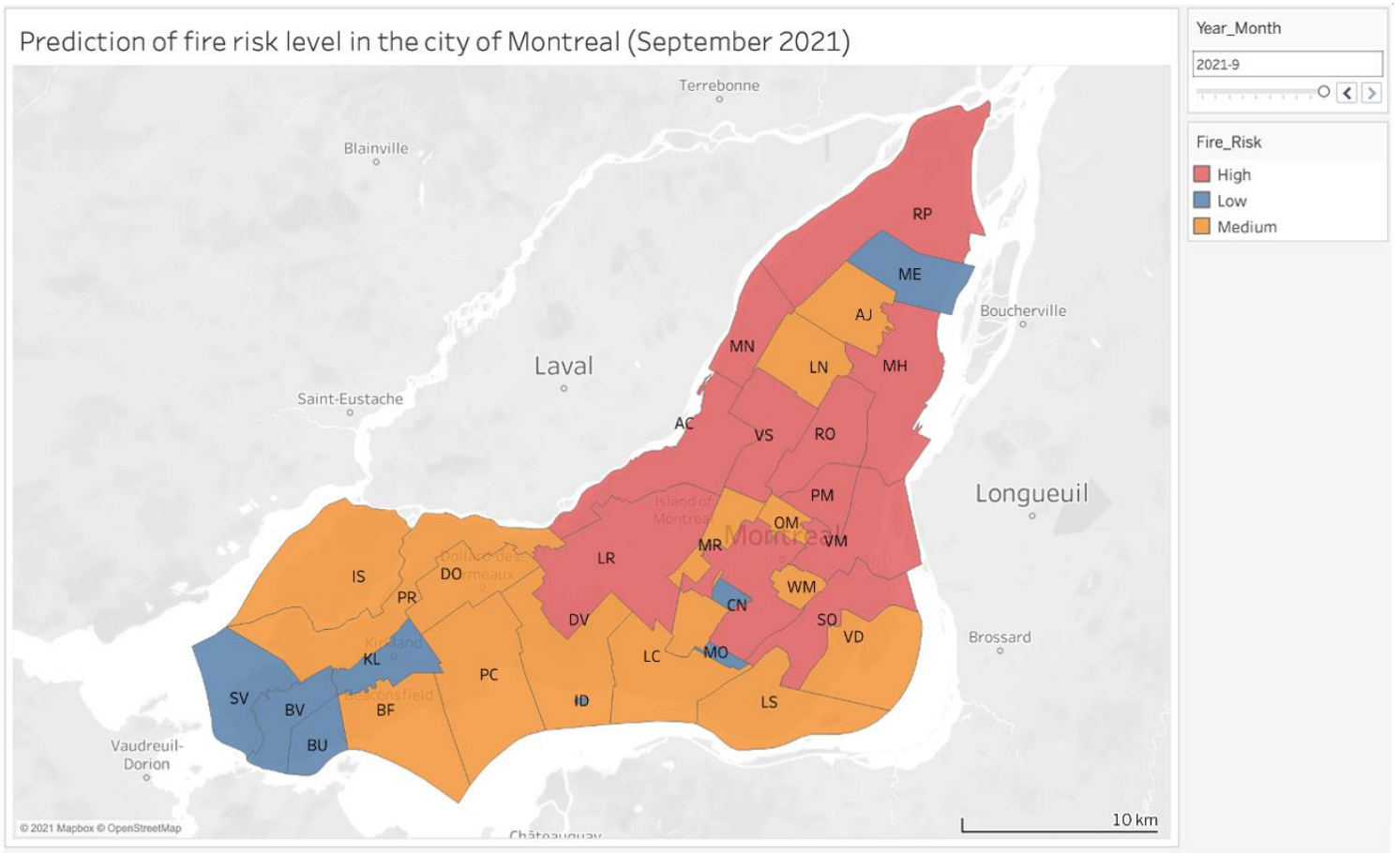


Visualization of prediction results

Historical Data



Predicted Data



As depicted in the above results:

- areas LR, CN and MN have changed from Medium risk to High risk this month of September one year later
- areas DV and WM have changed from Low risk to Medium risk this month of September one year later

Improvement initiatives and future work

- Explore more features from new datasets which may be more relevant to the fire risk, such as population of borough, education level, etc.
- Spend more time on feature engineering which helps select better feature to improve model performance.
- In the fire risk categorization, consider the size of areas/boroughs to compare the risk as the same number of fire incidents may have a different impact on boroughs with different sizes. For example, consider the dissemination blocks of 1 km^2.
- The severity of the incident may also be taken into consideration in the analysis. An area with low fire risk assigned by the model could have a high severity calculated in terms of significant loss of life and property damages.
- Consider other potential classification models to increase the chance of having a better prediction result.
- Consider the number of fire stations in each borough to help stakeholders better manage their resources.

Thank you!

McGill | School of Continuing Studies | École d'éducation permanente

mcgill.ca/scs