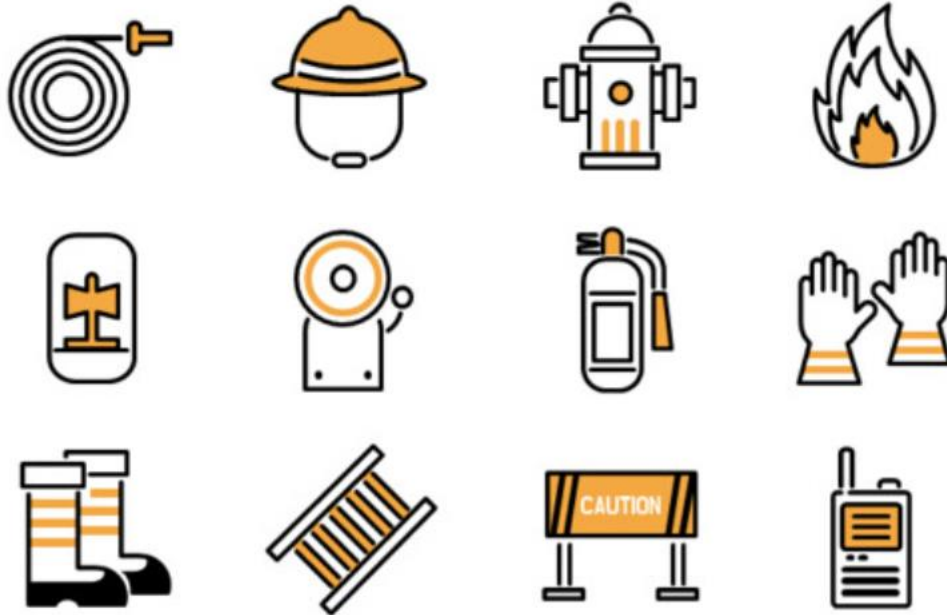# Executive Report

Team 4

NAZANIN HASHEMIPOOR
SOFIENNE SIRHI
NASIM AFZALI CHALI
MARIE-NOËL LEPAGE

Problem Statement

We know that Fire hazard is a real phenomenon and most of the time it happens by humans. Fires can cause costly property damages and significant economic losses. They are also a major source of severe injury and loss of human life in our urban and rural communities.

The total area of Montreal is ~499.30 km² and it contains ~334,000 residential and non-residential buildings and the population density is about 4105.85 (per km²). Last year, there were more than 1200 fires. Despite this, the city has only 67 fire stations and 2,694 full-time employees and can only inspect ~2% of the total buildings. [1]

Number of Fire Building's per year
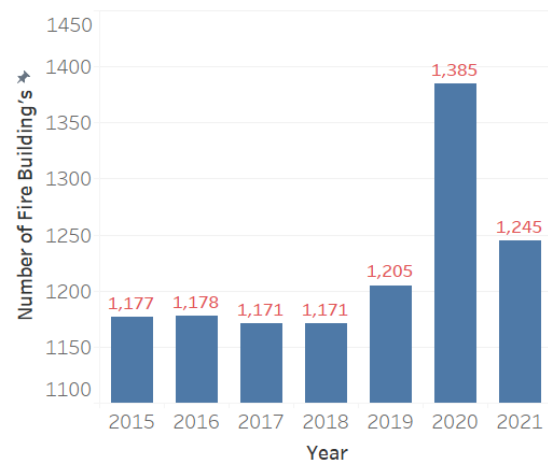
Municipal fire departments face the challenge of how to best politicize their inspection efforts. Fire departments rely on annual inspections and internal educational tools to reduce fire incidents. In the old day, it was a human to plan for the building inspection list and it was happening randomly without any previous data. If we look at most of the cities in the world, we will find out they rely on data and use them to control this hazard in the city (appendix 1 – State of the art review).

Our objective is to predict the risk of fire incidents for each area by considering the high-risk building with Machine Learning. In more detail, the granularity of the problem is:

- **Area:** Montreal is divided by the administrative boundary and by 1 km square.
- **Time Frame:** The data are aggregated for each season.
- **Fire Risk:** The number of incidents of building in 3 levels: Low, Medium, and High.

We hope the results of this project led to a change in how area buildings fire inspections are targeted, and the city of Montreal finds the best way to allocate resources more efficiently, and as a result, save lives of citizens and prevent loss of property.

Fire Risk Prediction Workflow

Scientific Literatures ⟹ Data Collection ⟹ Data Preprocess ⟹ Exploratory Result ⟹ Predictive Model

---

[1] https://ville.montreal.qc.ca/sim/rapport-des-activites

Data Sources

Based on the work completed in the scientific literature, we proposed to include fire incidents, property assessment, crime, census data. Also, we use the geo-localization of the administrative boundary of Montreal for the area grid.

Data	Description	Fields Used
Property [2]	Vector geospatial data of the division of properties in the Montreal agglomeration containing general information on property assessment units.	Floors, units, year of the building, area of the building, utilization code, codification CUBF, geo-localization.
Crime [3]	List of criminal acts recorded by the Service de police de la Ville de Montréal (SPVM).	Date, latitude, longitude, categories of crime.
Fire [4]	Data set listing the interventions carried out by the Montreal Fire Department (SIM).	Date, Description group, Type of description, longitude, latitude.
Census [5]	The socio-demographic profiles present data from Statistics Canada's census of population.	Density, Income, Building value, Minor & Major Repair, Name of the administrative boundary.
Area [6]	Polygons delimiting the boroughs of the City of Montreal, boroughs, and related cities constituting the agglomeration of Montreal.	Polygons, Name of the administrative boundary.

Some data sets have static data. In this case for our model, we use this data for each year.

	2015	2016	2017	2018	2019	2020	2021
Property	Static	Static	Non-Static	Static	Static	Static	Static
Crime	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static
Fire	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static	Non-Static
Census	Static	Non-Static	Static	Static	Static	Static	Static

Static Data
Non-Static Data

[2] https://donnees.montreal.ca/ville-de-montreal/unites-evaluation-fonciere
[3] https://donnees.montreal.ca/ville-de-montreal/actes-criminels
[4] https://donnees.montreal.ca/ville-de-montreal/interventions-service-securite-incendie-montreal
[5] http://ville.montreal.qc.ca/portal/page?_pageid=6897%2C68087646&_dad=portal&_schema=PORTAL
[6] https://donnees.montreal.ca/ville-de-montreal/polygones-arrondissements

Data Exploration and Cleaning

For all data sets we removal the non-relevant column and used only the fields shown in the table on the last page. Also, we take this level of granularity for:

- **Property**: We aggregate the residential data with the same address (average of building's age, maximum of floors, sum of the units, sum of the building's area, and spatial object). For the floors, we can remark a bias in this feature because is not possible to aggregate this variable to see exactly the number of floors in this case. Also, we drop some utilization codes that are not related to the building (see appendix 2 for more detail).
- **Crime**: We aggregate all categories to count the number of crimes.
- **Fire**: We keep only some description groups (fire building, fire alarm, false alarm, and without fire). For fire alarm and without fire, we drop some type of description in this category (see appendix 3 for more detail).
- **Census**: We imputed median. Mean or 0 for the null data of Ile-Dorval. The aggregation of this data as per the administrative boundary of Montreal needs to be adjusted.
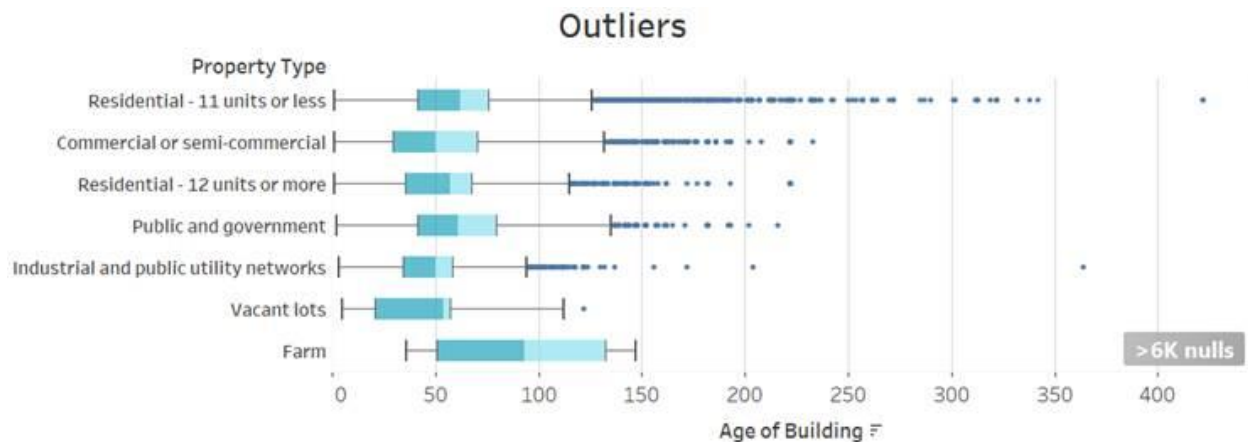
Decision trees classification and ensemble methods are not impacted by the outliers in the data as the data is split using scores which are calculated using homogeneity of the resultant data points. But for other technique models, we need to explore other tactics for the outlier. For this reason, we keep all outliers as they are but exclude these from our calculation for the sum of the area and the average for the year of buildings in our rectangular dataset.

- Years of building: More than 6k years are 9999 years.
- A lot of floors are null (around 14k), and the most frequent time is for the parking inside and storage.
- One building has 3100 units and validation on Google Maps does make no sense. Also, a lot of buildings have 0 units (around 19k). The most frequent time is for parking inside, storage, and the commercial.
- One building has an excessive area and makes no sense if we check on Google Map. Also, a lot of buildings (around 39k) have 0 areas.

Graph 1: Seasonality of Fire Building's

| 2018 | 2019 | 2020 | 2021 |
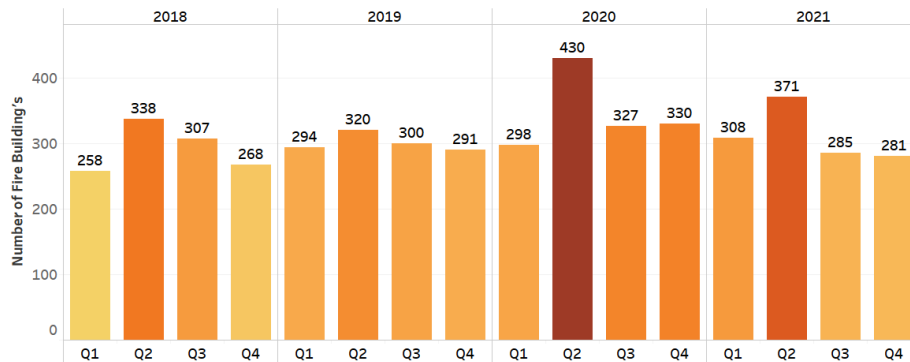
- There are significantly more fire building's during the Q2 in the last 2 years.

- We recommend conducting the rectangular data by season.

(2018) Q1: 258, Q2: 338, Q3: 307, Q4: 268
(2019) Q1: 294, Q2: 320, Q3: 300, Q4: 291
(2020) Q1: 298, Q2: 430, Q3: 327, Q4: 330
(2021) Q1: 308, Q2: 371, Q3: 285, Q4: 281

Graph 2: Top 10 - Administrative Boundary by Fire Building's 2021

Name of Administrative Boundary

| Ville-Marie | 107 |
Côte-des-Neiges-Notre-Dame-de-Grâce	103
Le Sud-Ouest	94
Le Plateau-Mont-Royal	93
Villeray-Saint-Michel-Parc-Extension	87
Mercier-Hochelaga-Maisonneuve	82
Ahuntsic-Cartierville	76
Rivière-des-Prairies-Pointe-aux-Trembles	70
Saint-Laurent	70
Rosemont-La Petite-Patrie	58

Crime: 0 — 2,792

Number of Fire Building's
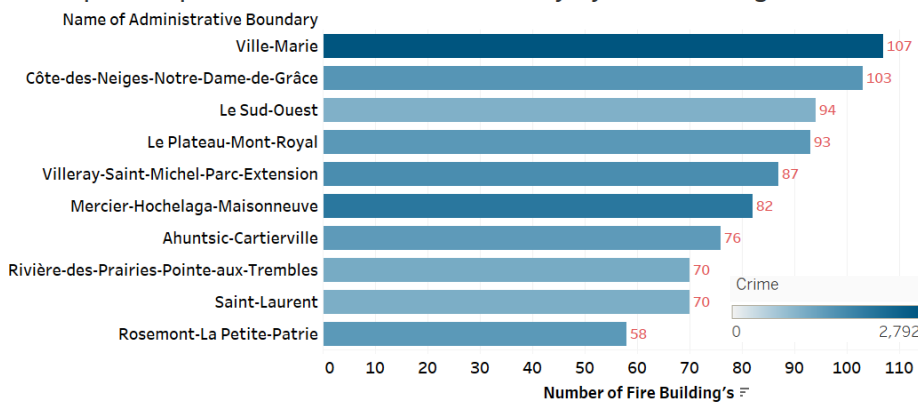
- Some administrative boundaries have more fire building's incident and seems correlated with the crime dataset.

- We recommend the rectangular data by a grid of 1km square of administrative boundary.

Graph 3: Number of Units by Fire Building's 2021

Number of Units — Number of Fire Building's
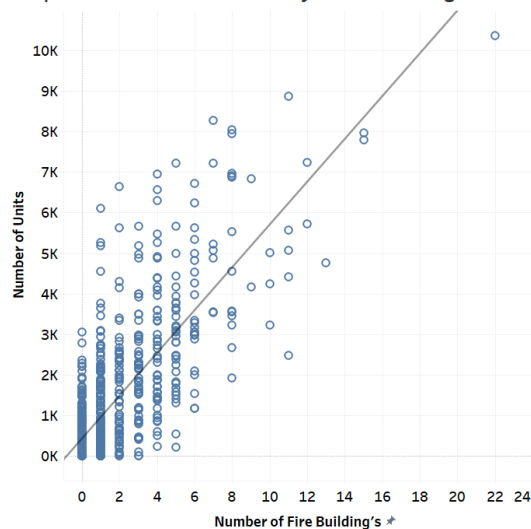
Graph 4: Major Repair by Fire Building's 2021

Major Repair — Number of Fire Building's
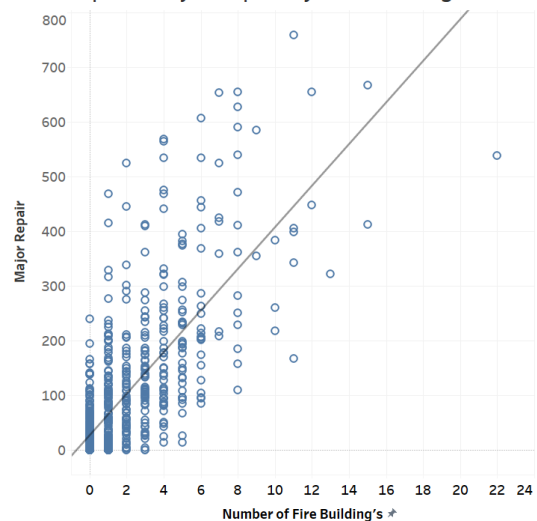
- The number of fire building's aggregates by grid of 1km square have a linear relationship with the number of units and the Major Repair.

Feature Engineering

Area and Time Frame:

We joined the data sets for each feature and label (count per box for each category) to obtain the most complete data set information for an aggregation level, i.e., by a square of 1 km (area) and for every fixed time frame of the season (time series resampling).

Property Types:

With the utilization code and codification of CUBF, we created the property types. We categorized the properties in less complexity of the literature but should be ameliorated to see if give better performance in the future.[7] (see appendix 2 for more detail). 7 categories are created for the buildings and are transformed by two classes of our representative percentage by grid (residential & commercial and the 7 categories) to see these features performed in the modeling machine learning.

Census data:

For Ile-Dorval, we use mean, median, or imputed 0 for the tkmnull value. Also, we adjusted some features (Density, Minor & Major Repair) for split by the grid of 1km square.

$$Feature = \frac{Number\ of\ units\ for\ the\ 1km\ square}{Number\ of\ units\ for\ the\ administrative\ boundary} * Feature \text{ (Administration boundary)}$$

Trigo Projection Trick:

To conserve the distances of the cyclic season, we use the Trigo Projection Trick.

$$sin = \ sin \frac{2\pi * \text{Modulo(Season Number, 4)}}{4}$$
$$cos = \ cos \frac{2\pi\ *\ Modulo(Season\ Number, 4)}{4}$$
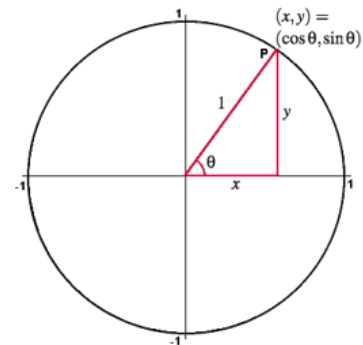
Lag Features:

The classical way that time series forecasting problems are transformed into supervised learning problems. We want to predict the value at the next season (t+1) given the value at the previous time (t-1). For this reason, we use the sliding window method with 3 lag and check the performance of the model with a different lag to see which one perform better.

Season	Lag(t-1)	Lag(t-2)	Lag(t-3)	Lag(t-4)
1	4	NaN	NaN	NaN
2	2	0	NaN	NaN
3	0	1	0	NaN
4	1	1	2	0

---

[7] https://cdn-contenu.quebec.ca/cdn-contenu/adm/min/securite-publique/publications-adm/publications-secteurs/securite-incendie/soutien-municipalites-incendie/guide_planification_activites_prevention/guide_planification_activites_annexe1.pdf?1623760272

Tools and Techniques Used

For our production environment, we use different tools:
- **Alteryx:** aggregate our rectangular data sets by the area (1km square grid) and a time frame (season).
- **Tableau:** for created visualization.
- **Python:** We used some libraries (NumPy, Pandas, Scikit-Learn, Seaborn, Yellowbrik) in our work to create and visualize the result of our model.
- **Teams:** To share our work in one place. With more time and in the future, Google Cloud Platform should be a more relevant tool.

Split our dataset in Train / Valid / Test:
We split our data into 3 datasets.
- The training dataset used to fit the model (18 seasons)
- The validation dataset to provide an unbiased evaluation of our model (6 seasons)
- The test dataset to provide an unbiased evaluation of the final model (1 season)
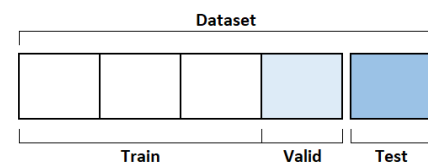
Ordinal or One-Hot Encoding for categorical Data (Grid Name):
We test these 2 techniques with our training set and chose to take the ordinal for the Grid Name to eliminate columns to train our model. Also, these 2 techniques give a similar result.

Creation of our classification label:
For our label, we use the number of fire buildings by the grid to create a classification in 3 levels. We can remark our data are unbalanced and need to be taken into consideration for the training step.

Risk	# of Fire	# of Risk Class
Low	= 0	14,448
Medium	= 1	2,952
High	>1	1,680

Imbalanced data classification problem:
The distribution of classification risk is skewed and poses a challenge for our predictive modeling. This results in models that have poor predictive performance, especially for the minority class. We try different techniques to solve this problem: Smote, the balanced weight, random over/under. We can remark better performance with the smote technique and choose this method in our final model.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$
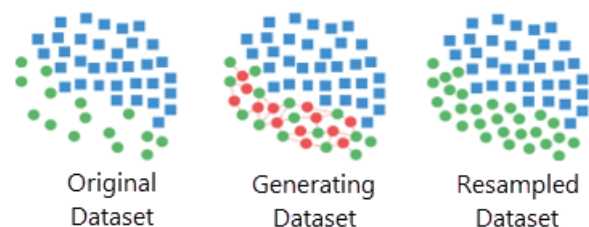
Note for smote technique to work best, you should previously normalize the data.

Summary of Modelling Techniques Evaluated

Baseline Model

A decent baseline model is created with these few requirements:
- A simple model for less likely to overfit.
- Interpretable to get a better understanding of our data and show a direction for feature engineering.
- The based tree model is non-parametric and does not require the data to be normally distributed.

For the classification model, we use some metrics to avoid objective validation. For someone who is essentially defined in the multiclass task, we use the macro for all classes are equally important (equal weight to each class).

Confusion Matrix

We use the below evaluation Matrix for model comparison.
- Kappa score
- AUC score
- Recall
- Precision

We evaluated multiple model types along the metrics described above in the table. Because of the substantial class imbalance in our data accuracy is not a useful measure of model performance. Therefore, the kappa score, which is, intuitively, the percent agreement accounting for the class imbalance, is a more useful metric. We also use the "recall" to evaluate the performance of our classifier.

Time series Cross-Validation

We used the grid and random search to learn the hyper-parameter of an estimator. For that, we use cross-validation for time series. The idea is to divide the training set into two folds at each iteration on the condition that the validation set is always ahead of the training set.

We tried some classification models (see in next section - modeling results). The model selected should be flexible and fit into the data and explain the insight of the data statistical estimation form. The bias and variance are directly proportional to each other, and we try to do a good balance. The bias is an error that has been introduced in our model due to oversimplification and can underfit our prediction. The variance is an error that has been introduced in our model due to the selection of a complexity resulting in high sensitivity and overfitting. In the end, the objective is to estimate the model with some metrics to find the best should be generalized the form.

Modeling Results

|	Model Building	Macro f1	Macro Precision	Macro Recall	Macro AUC	Kappa
Initial Models	Standard Decision Tree	0.46	0.47	0.46	0.61	0.22
	Naïve Bayes	0.46	0.56	0.48	0.79	0.22
	K – Nearest Neighbors	0.49	0.54	0.48	0.78	0.27
	Decision Tree (Balanced weight)	0.47	0.47	0.47	0.62	0.24
	XGBoost	0.49	0.54	0.48	0.85	0.30
	Standard Random Forest	0.50	0.54	0.49	0.80	0.28
	Linear Discriminant Analysis	0.48	0.54	0.49	0.80	0.26
	SVM	0.35	0.42	0.37	0.65	0.08
	Balanced Random Forest	**0.50**	**0.49**	**0.58**	**0.79**	**0.30**
	Random Forest-SMOTE Tomek Links	**0.51**	**0.51**	**0.52**	**0.81**	**0.35**
Final Models	**Balanced Random Forest**	**0.54**	**0.53**	**0.61**	**0.81**	**0.35**
After Tuning	**Random Forest-SMOTE Tomek Links**	**0.55**	**0.53**	**0.60**	**0.82**	**0.35**

We chose the balanced random forest model, and standard random forest using SMOTE Tomek Links resampling method as our final models. From Random search and Grid search tuning and based on the evaluation measures we decided to use Balanced Random Forest as our final model due to the larger Recall and less False Negative (FN) of high-risk class, 136 for Balanced Random Forest versus 145 for Random Forest-SMOTE Tomek Links.

The "confusion matrix" shows the relationship of the actual to predicted event Ideally, a predictive model will maximize the number of true positives (addresses that the model predicted would have a fire, and had a fire in the test set) and true negatives, while minimizing the false positives and false negatives. However, in some prediction cases, like fire prediction, for instance, false negatives (where the model predicts there will not be a fire, and they are was) are worse than false positives (model predicted fire, but there was not a fire in the next season of the test set). Therefore, we want to evaluate our model by how many true positives and true negatives it correctly classifies and how few false negatives it returns.
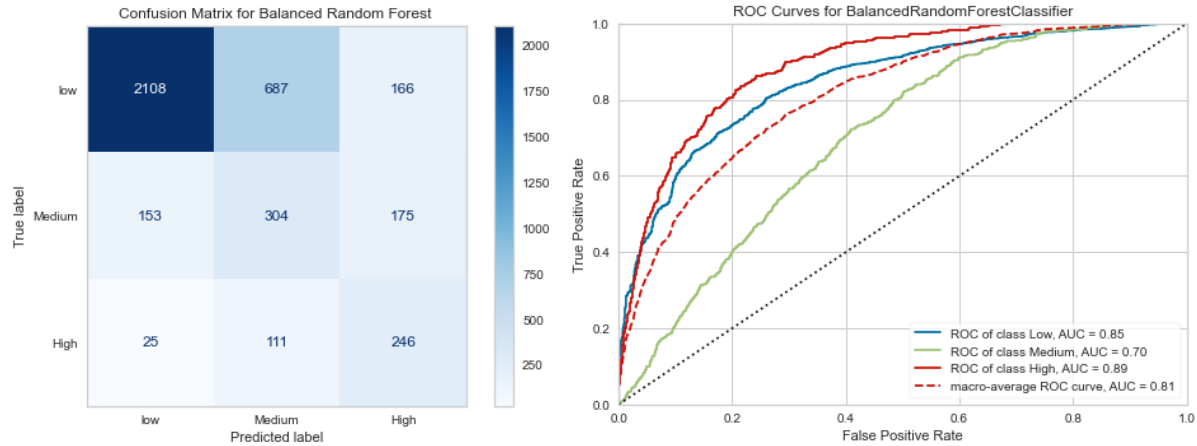
Feature importance based on our best model is as follow:
Influential features include the number of units, major and minor, building area, without fire, density, crime, number of floors, and number of buildings for each grid.
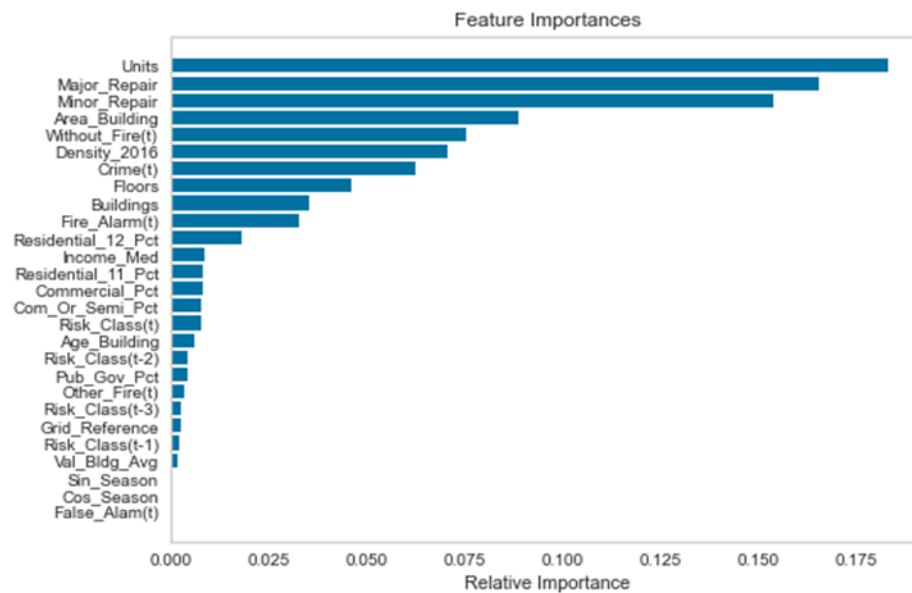

Insights and Challenges

- The fire data has been obfuscated and modified to ensure privacy and the protection of personal information. The geographical position of the event was located at one of the intersections of the street segment where the intervention took place. For this reason, it's not possible to rely on the fire to a building and predict the risk by area.

- Some data sets have a static date. For the building, it is important to update the list with the new construction each year. For the census data, we need to calculate our data by a grid of 1 square km. Also, it is necessary to grow these data at an appropriate rate for each static year.

- When we join our data for the rectangular set, some data falls on the boundary line between the grid of 1km square (Fire: 113, Buildings: 11, Crime: 3). In the future, we can optimize the accurate spatial data with a smoothing using gaussian for represented by the closest centroid.

- We see the risk class is an unequal distribution in the training dataset. A severe imbalance is more challenging to model and may require specialized techniques (smoke, the balanced weight, random over/under).

- With a dynamic environment, there is always a risk of the model becoming obsolete. Therefore, attention should be paid to any changes. For example, in 2020 the pandemic situation with the covid-19 maybe move some area risk with the telework and should be investigated.

Conclusions
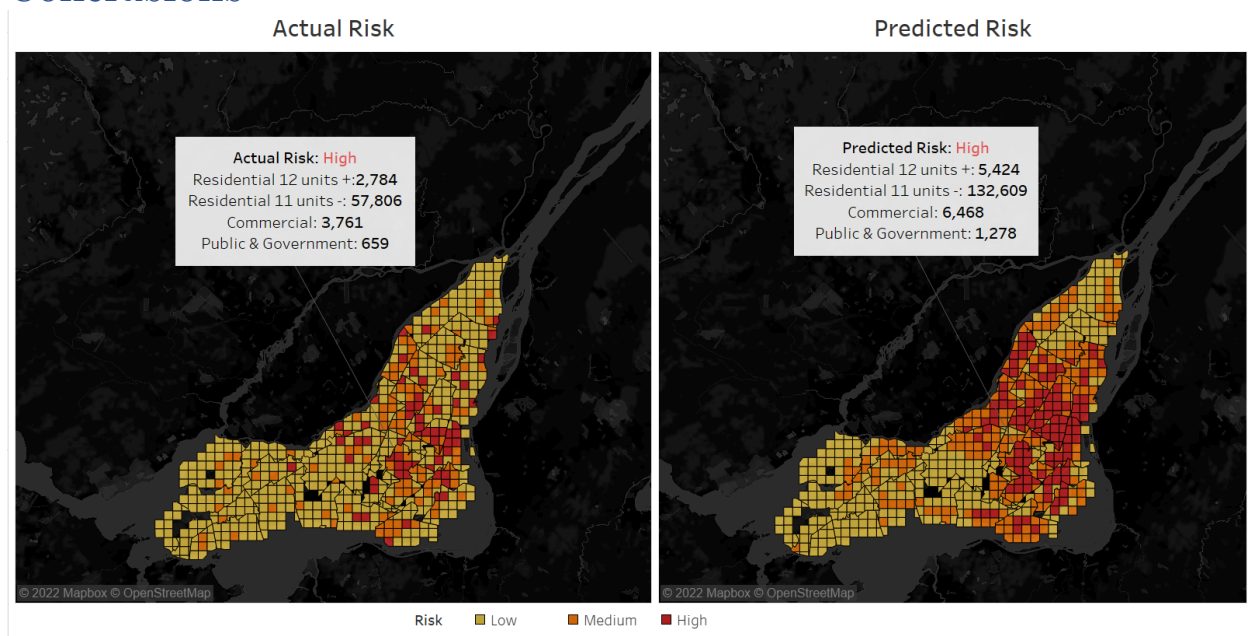
With our model, we can predict the area fire risk for the next season. For each area, we can see the number of different property types to be inspected in this zone.

We expected the utilization of past data to predict the next season to help the inspections targeted by area in the city of Montreal. We hope the best way to allocate resources more efficiently, as a result, saves the lives of citizens and prevent loss of property.

Appendix 1
State of the art review

Prediction Projects	Description
New York-MODA and Risk-Based Inspection System (RBIs) [8]	• Before applying MODA's data-driven analysis, the first 25% of FDNY inspections typically resulted in 21% of the most severe violations being discovered. • Using MODA's prediction model, the first 25% of inspections now result in more than 70% being discovered.
Atlanta Fire Department / Georgia Tech and Firebird [9]	• Firebird computes fire risk scores for over 5,000 commercial buildings in the city, with true positive rates of up to 71% in predicting fires. • Used models such as Logistic Regression, Gradient Boosting, Support Vector Machine (SVM), and Random Forest, SVM, and Random Forest performed the best.
Baton Rouge, LA- Predict building fires [10]	• Their objective was to generate a building fire risk score for every address in Baton Rouge, LA. • Tested models were logistic regression, Random Forest, and gradient-boosted trees. While the best gradient boosted trees model performed slightly better than the best Random Forest model by AUC, Random Forest performed better in their 2016 test. The Random Forest's AUC was 0.81, compared to boosted trees' 0.79 in 2016 testing.
Pittsburgh-CMU and Pittsburgh Bureau of Fire (PBF) [11]	• Developed a predictive model to determine property-level fire risk in the 6 months. • In the 6 months after the commercial risk model was deployed, 29% of the high-risk properties had some type of structure fire incident, compared to 5.2% of the medium-risk properties and 0.7% of the low-risk properties.
A Building Fire Risk Prediction Validation Project – Vancouver and New Westminster [12]	• Validation of the results of other smart-cities solutions to build a fire risk prediction assessment for two municipalities in Canada to prioritize fire inspections. • The models showed good prediction with approximately 70% of fires identified for the 2017 period (with the false positive rate of almost 25%).

---

[8] http://eddiecopeland.me/wp-content/uploads/2015/11/Big-Data-in-the-Big-Apple-Report.pdf

[9] https://faculty.cc.gatech.edu/~dchau/papers/16-kdd-firebird.pdf

[10] https://scholar.harvard.edu/jonjay/blog/how-we-predicted-building-fires-baton-rouge-la-working-version

[11] http://michaelmadaio.com/Metro21_FireRisk_FinalReport.pdf
http://michaelmadaio.com/NeurIPS_2018_FireRisk.pdf

[12] https://fireunderwriters.ca/media/bb737a67-f53f-4625-9cf8-d91e32c9fb7f/gtJiSg/FUS/Resources/Articles/FUS_Building_Fire_Risk_Validation_Project.pdf

Appendix 2
Cleaning property and creation of Property Type feature

Residentials (11 units or less / 12 units or more)

Utilization Code	Description	Property Type	
1***	Residentials *Some manipulation for aggregate data with the same address (all units in the same buildings)*	When the number of units is not null	
		Number Units <= 11	Residential – 11 units or less
		Number Units >=12	Residential – 12 units or more
		When the number of units is null but the number of floors is not null	
		Number of floors <= 3	Residential – 11 units or less
		Number of floors >= 4	Residential – 12 units or more
1921	Indoor parking	Residential – 11 units or less	
1923	Storage	Residential – 11 units or less	
Exclusion			
1701	Mobile home park (land only)		
1922	Outdoor parking		

Industrial and utility networks

Utilization Code	Description
2*** & 3***	Industrial and public utility networks
47**	Information industry and cultural industry
48**	Public Service (infrastructure)
8549	Other mining and quarrying of non-metallic minerals (except oil)
Exclusion	
4880	Snow deposit

Commercial or semi-commercial

Utilization Code	Description
49**	Other transportation, communication, and utilities (infrastructure)
5*	Commercial
60**	Office buildings
61**	Finance, insurance, and real estate service
62** to 66**	Personal, business, repair, professional, construction service
8221	Veterinarians and hospital service for farm animals
8399	Other services related to forestry
Exclusion	
6513	Hospital ward
6531	Reception center or curative establishment
6532	Local community service center (C.L.S.C.)
6533	Social service center (C.S.S. et C.R.S.S.S.)
6534	Self-help and community resource center (including housing, furniture, and food resources)
6539	Other social service centers or social worker offices

Public and government

Utilization Code	Description
41**	Railway and metro
42**	Motor vehicle transport (infrastructure)
43**	Air transport (infrastructure)
44**	Maritime transport (infrastructure)
46**	Parking lot and garage for vehicles
6513	Hospital ward
6531	Reception center or curative establishment
6532	Local community service center (C.L.S.C.)
6533	Social service center (C.S.S. et C.R.S.S.S.)
6534	Self-help and community resource center (including housing, furniture, and food resources)
6539	Other social service centers or social worker offices
67** to 69**	Government, educational and miscellaneous service
7***	Cultural, recreational, and leisure
Exclusion	
4111	Railway (except tourist train, switch, and marshaling yard)
4112	Railway switch and marshaling yard
4121	Subway track
4215	Bus shelter
4299	Other transport by motor vehicle
45**	Public highway
462*	Automobile parking lot and highway bed
4632	Outdoor parking
7223	Racetrack
7224	Toboggan run, bobsleigh, and ski jumps
7411	Golf course (without chalet and other sports facilities)
7421	Fun ground
7422	Playground
7423	Sports field
7431	Beach
7492	Wild camping and picnicking
7611	Park for general recreation
7620	Recreational and ornamental park

Farm

Utilization Code	Description
8***	Production and extraction of natural wealth
Exclusion	
8549	Other mining and quarrying of non-metallic minerals (except petroleum)
8221	Veterinarians and hospital service for farm animals
8399	Other services related to forestry

Vacant Lots

Utilization Code	Description
94**	Unoccupied floor space
95**	Building under construction

Appendix 3
Type of description selected for Alarm & Without Fire

Type Description Of Fire Alarm
Alarm / intrusion detection
Private or local alarm
Call of detection company

Type Description Of Without Fire
10-22 for airport call
Acc. no kills fire building
Acc. no victim fire tunnel
Acc. fire victim - building
Bomb Threat
Overheated foods
Call of detection company - GAZ
Horn - gas F7/GAS
Propane tank leak
External leak: hydrocar. liquid div.
Nat gas leak 10-07 F7/2GAZ
Natural gas leak 10-09
Natural gas leak 10-12
Natural gas leak 10-22
Int. leak: hydrocar. liquid div.
Fumigation F7/GAS
Interv. terminal or building
Hazardous materials / 10-07
Hazardous materials / 10-09
Hazardous materials / 10-22
Metro building /10-22 without a traffic light
Suspicious smell - gas 14.
Electrical problems
Dangerous structure
Surplus oil
Transshipment mat. explosions

For Fire Alarm, we drop the type of description alarm verification since according to our understanding, these are alarms that are triggered to validate systems.

For Without Fire, we keep only the type of description that should be related to a fire building.