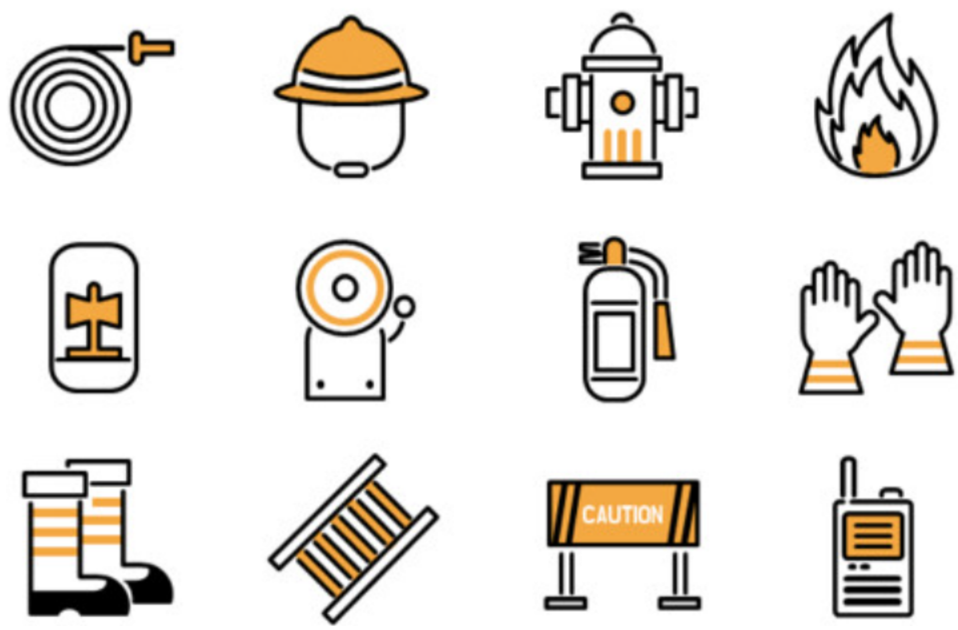
Executive Report



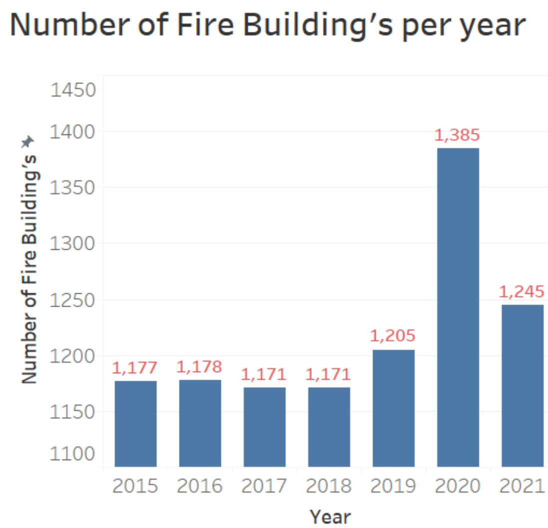
Team 4

MARIE-NOËL LEPAGE et al.

Problem Statement

We know the 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 km^2 and it contains ~334,000 residential and non-residential buildings and the population density is about 4105.85 (per km^2). 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]



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 devised by the administrative boundary and by 1 km square.
- **Time Frame:** The data are aggregate 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 building fire inspections are targeted and the city of Montreal find 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

[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 incident, property assessment, crime, census data. Also, we use the geo-localization of 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 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 administrative boundary.

Some data sets have static date. In this case for our model, we use this data for each year. One of possible amelioration in the future should be growth these data with a rate.

	2015	2016	2017	2018	2019	2020	2021
Property							
Crime							
Fire							
Census							

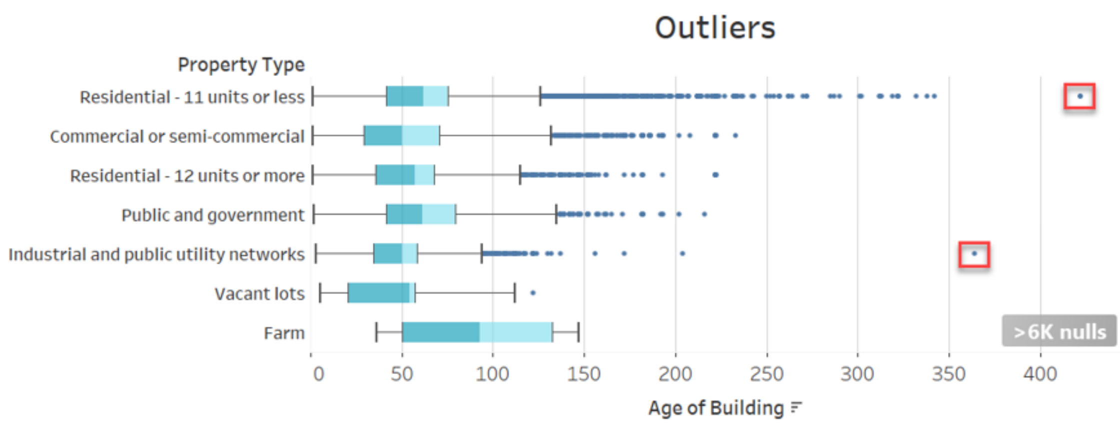
Static Data
Non-Static Data

[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 show in the table of 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 buildings 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 code that not related to the building (see the appendix 2 for more detail).
- **Crime**: we aggregate all category for count the number of crimes. For the
- **Fire**: we keep only some description group (fire building, fire alarm, false alarm and without fire). For fire alarm and without fire, we drop some type description in this category (see the appendix 3 for more detail).
- **Census**: We imputed median. Mean or 0 for the null data of Ile-Dorval.

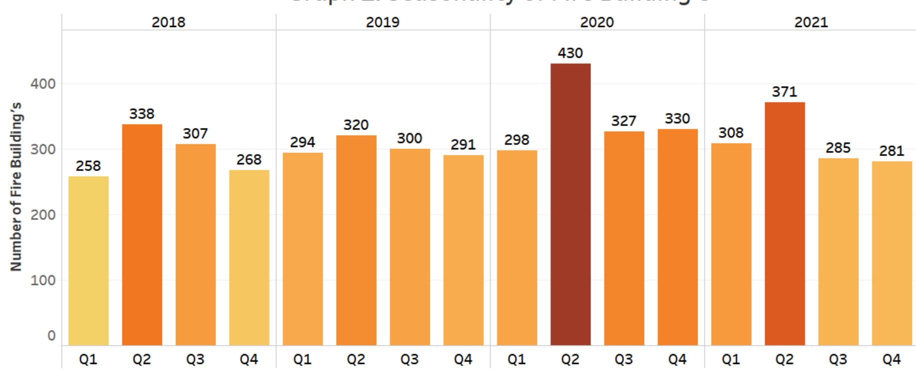


For the property data sets, we keep all data for the buildings but for some outliers and null value feature, we populate 0 in our data sets. Also, for the decision trees classification and ensemble methods are not impacted by the outliers in the data as the data is split using scores which are calculated using homogeneity of the resultant data points.

- Years of building: Two years of buildings have 1600 years and make no sense with some search on the web.[7] Also, more of 6k years are 9999 years.
- A lot of floors are null (around 14k) and the most frequently time is for the parking inside and storage.
- One building have 3100 units and with validation on google map make not sense. Also, a lot of buildings has 0 units (around 19k). The most frequently time is for parking inside, storage and the commercial.
- One building have an excessive area and make not sense if we check on google map. Also, a lot of buildings (around 39k) has 0 area.

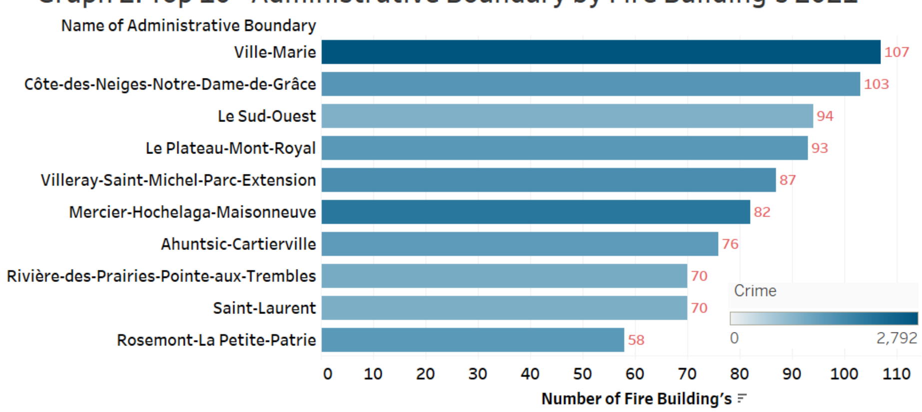
[7] https://montreal.ca/articles/la-maison-le-ber-le-moyne-le-plus-ancien-batiment-complet-de-montreal-12465#:~:text=La%20Maison%20Le%20Ber%2DLe%20Moyne%20et%20sa%20D%C3%A9pendance%20sont,Ber%20et%20Charles%20Le%20Moyne.

Graph 1: Seasonality of Fire Building's



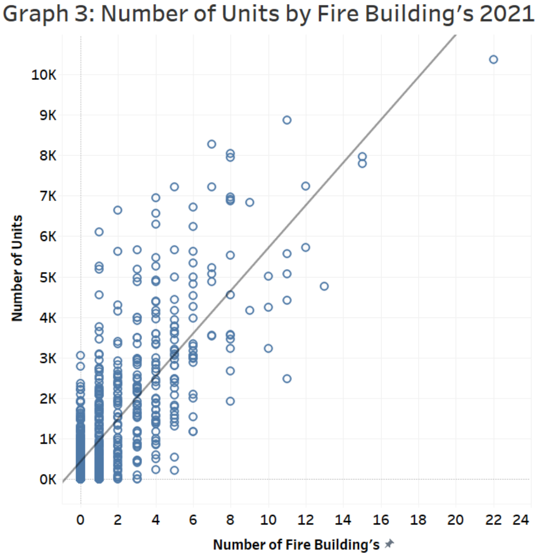
- There are significantly more fire building's during the Q2 in the last 2 years.
- We recommend conducting the rectangular data by season.

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

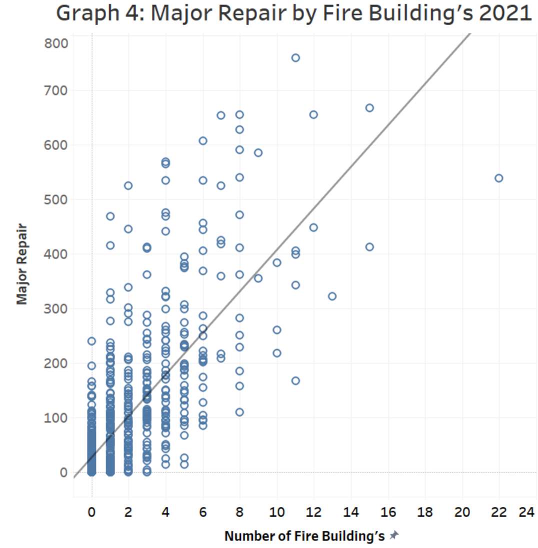


- Some administrative boundaries have more fire building's incident and seems correlated with the crime dataset.
- We recommend the rectangular data by a grid of 1km square of administrative boundary.

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



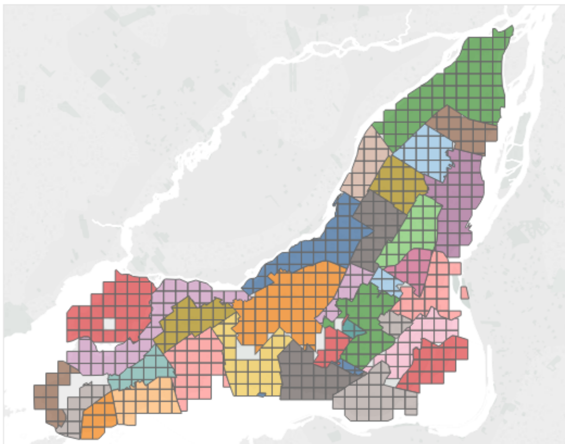
Graph 4: Major Repair by Fire Building's 2021



- 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

Grid: Administrative Boudary of Montreal by 1 km



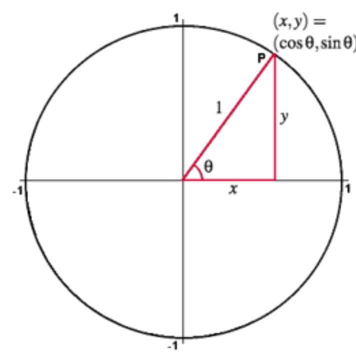
Area and Time Frame:
We joined the data sets for each feature and the label to obtain the most complete data set information for an aggregation level, i.e., by a square of 1 km (area) and for every fixed time frame of season (time).

Properties Types:
With the utilization code and codification CUBF, we created the properties types feature. We categorized the properties in less complexity of the literature but should be ameliorated to see if give better performance in the future.[8] (see the appendix 2 for more detail). 7 categories are created for the buildings and are transformed by two class of our representative percentage by grid (residential & commercial and the 7 categories) to see these features performed in the modelling machine learning.

Census data:
For Ile-Dorval, we use mean, median or imputed 0 for the null 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:
For 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 problem. 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 4 lag and check the performance of the model with different lag to see with one perform better.

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

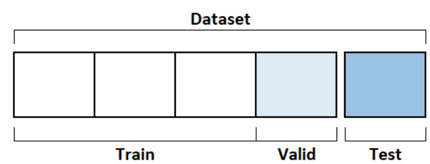
[8] https://cdn-contenu.quebec.ca/cdn-contenu/adm/min/securite-publique/publications-adm/publications-secteurs/securite-incendie/soutien-municipalites-incendie/guide_planification_activites_prevention/guide_planification_activites_annexe1.pdf?1623760272

Tools and Techniques Used

alteryx +ableau python Microsoft Teams

For our production environment, we use different tools:

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



Split our dataset in Train / Valid / Test:
We split in 3 our dataset.

- 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 chosen to take the ordinal for the Grid Name for eliminate columns to train our model. Also, these 2 techniques give similar result.

Rescale data methods:

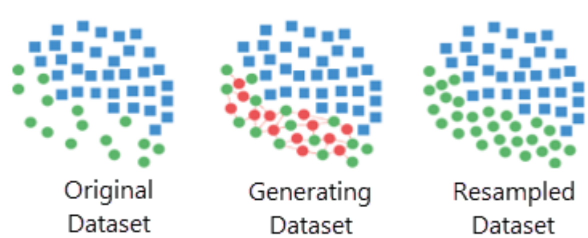
$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

We rescale our data for better performance of our model. A lot of technique can used: Normalization, Standardization, Log transformation, Power Transformer, etc. We try some of them but finally go ahead with the normalization. Not required feature scaling to be performed as they are not sensitive to the variance in the data.

Creation of our classification label:
For our label, we use the number of fire buildings by grid for create a classification in 3 levels. We can remark our data are unbalanced and need take in 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 pose a challenge for our predictive modelling. This results in models that have poor predictive performance, specially for the minority class. We try different technique to solve this problem: Smote, balanced weight, random over / under. We can remark better performance with the smote technique and choose this method in our final model.



Summary of Modelling Techniques Evaluated

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

For classification model, we use some metrics to avoid subjective validation:
- **Confusion Matrix**: It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and the AUC-PR curves.

  True Positive (TP): You predicted positive and it's true.
  True Negative (TN): You predicted negative and it's true.
  False Positive (FP): You predicted positive and it's false.
  False Negative (FN): You predicted negative and it's false.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- **Recall** (TP / (TP + FN)): From all the positive classes, how many we predicted correctly.
- **Precision** (TP / (TP + FP)): From all the classes we have predicted as positive, how many are actually positive.
- **Accuracy:** From all the classes (positive and negative), how many of them have predicted correctly.
- **F-measure** (2 * Recall * Precision / (Recall + Precision)): Helps to measure Recall and Precision at the same time. It uses Harmonic mean in place of arithmetic mean by punishing the extreme values more.
- **AUC-PR:** Recommended for highly skewed domains where ROC curves may provide an excessively optimistic view of the performance (imbalanced dataset).
- **Cohen's Kappa:** Can handle very well both multi-class and imbalanced call problems.

We tried some classification models (see in next section - modelling results). The model selected should be flexible and fit into the data and explain the insight of the data statistical estimation from.

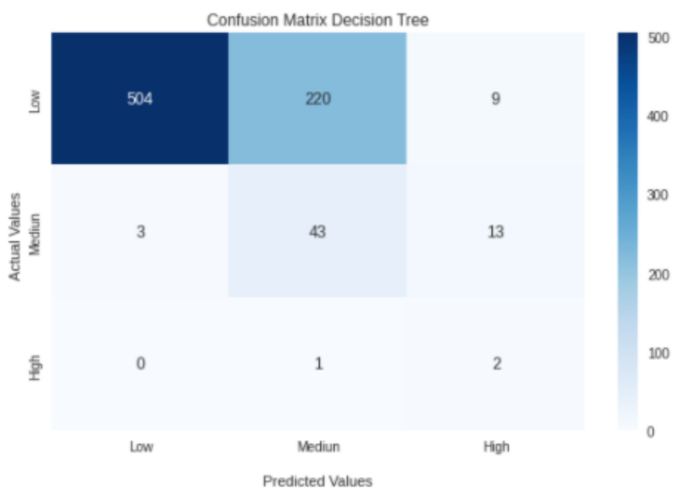
The bias and variance are directly proportional to each other and we try to do a good balance. The bias is an error that has been introduced in our model due to the oversimplification and can underfitting our prediction. The variance is an error that has been introduced in our model due to the selection of a complexity resulting in high sensitivity and overfitting.

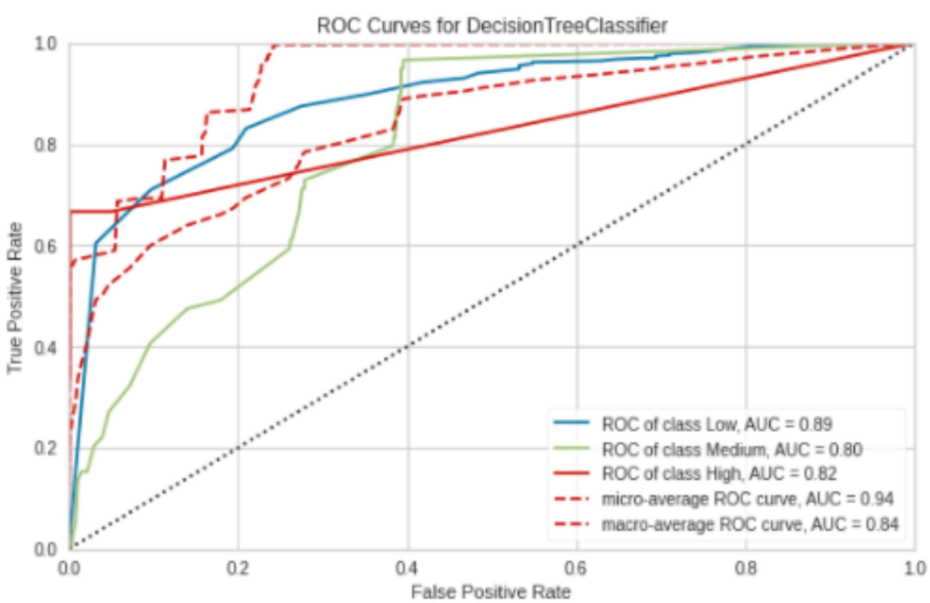
At the end, the objective is to estimate the model with some metrics to find the best should be generalized the form.

Modelling Results

Model Building	F1 Score	Precision	Recall	Kappa
Baseline				
Naïve Bayes				
K – Nearest Neighbors				
Decision Tree				
XGBoost				
Random Forest				
Linear Discriminant Analysis				
Easy Ensemble Classifier				

We chose the random forest for the performance and the interpretability of the model. With this model, we fine-tuning to improve the forecasting. Finally, we obtained these results:





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 the fire to a building and predict the risk by area.

- Some data sets have static date. For the building, should be interesting to update the list we the new construction each year. For the census data, we need to calculate our data by grid of 1km. Also, should be interesting to growth these data with an appropriate rate for each static year.

- When we join our data for the rectangular set, some data falling on boundary line between the grid of 1km square (Fire: 113, Buildings: 11, Crime: 3). With future we can optimize the accurate spatial data with vector quantization for represented by the closest centroid. Course 7 – valid that please. I think team 5 use that

- Additional data may be used for creating and interpreting the result of a predictive model. As example, the 311 datasets[9] contains requests for information and requests for services submitted to the City of Montreal via the 311-call system, by email and at "Bureaux Accès Montréal" (BAM). It should be noted that service requests are subdivided into three categories: requests, complaints, and comments.

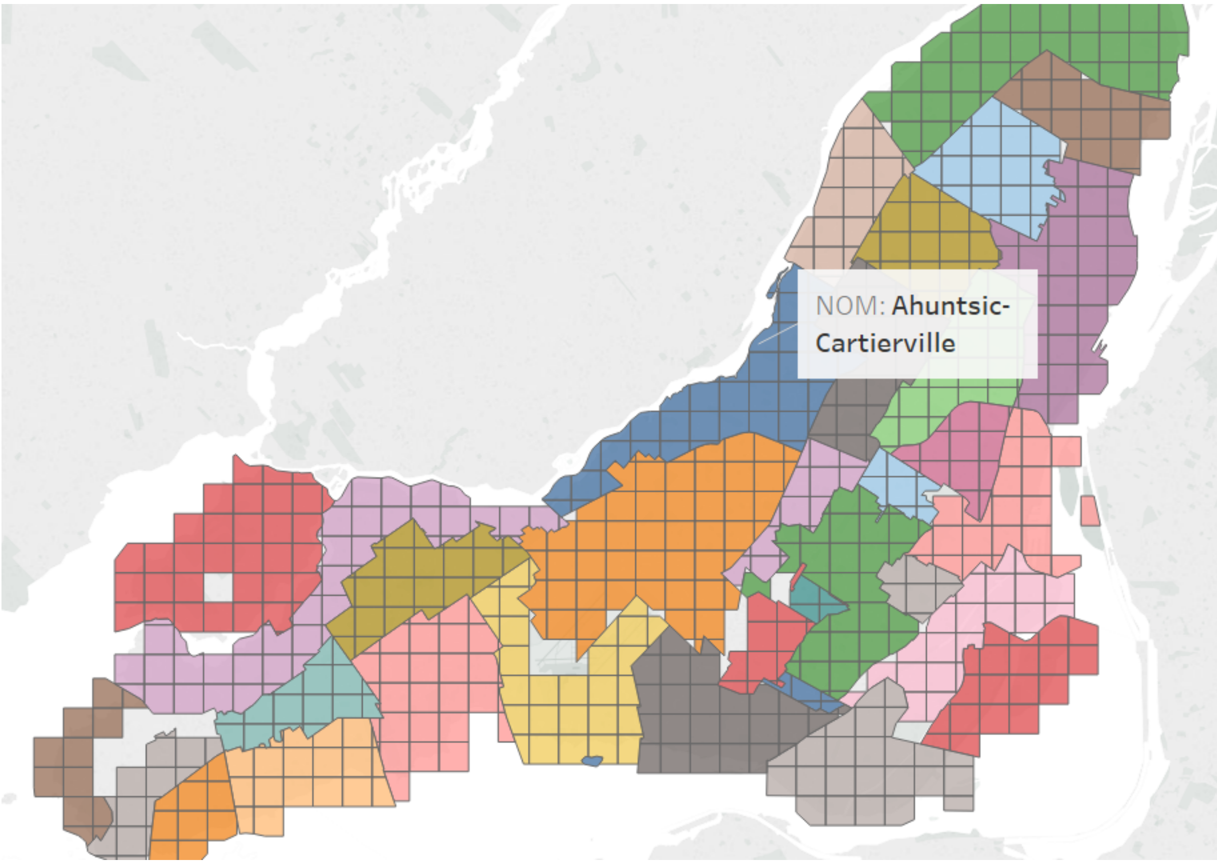
- With more time, we can go in more detail for more precision and performance for our model. As example, for search some material of construction using in time to select more efficiency the year of buildings related to the risk.

- 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, balanced weight, random over / under).

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

[9] https://donnees.montreal.ca/ville-de-montreal/requete-311

Conclusions



Show the map of Montreal with predicted risk class. We can annotate the number of residential and commercial as example. Should be updated with our predicted model.

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 type to be inspected in this zone.

We expected the utilization of past data to predict the next season help the inspections targeted in the city of Montreal. We hope a best way to allocate resources more efficiently, as a result save live 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) [10]	• 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 [11]	• 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 [12]	• 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) [13]	• 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 [14]	• 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%).

[10] http://eddiecopeland.me/wp-content/uploads/2015/11/Big-Data-in-the-Big-Apple-Report.pdf
[11] https://faculty.cc.gatech.edu/~dchau/papers/16-kdd-firebird.pdf
[12] https://scholar.harvard.edu/jonjay/blog/how-we-predicted-building-fires-baton-rouge-la-working-version
[13] http://michaelmadaio.com/Metro21_FireRisk_FinalReport.pdf
http://michaelmadaio.com/NeurIPS_2018_FireRisk.pdf
[14] 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 number of units not null	
		Number Units <= 11	Residential – 11 units or less
		Number Units >=12	Residential – 12 units or more
		When number of units is null but number of floors 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 marshalling yard)
4112	Railway switch and marshalling 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

For Fire Alarm, we drop the type 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 description should be related to a fire building.

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 traffic light
Suspicious smell - gas 14.
Electrical problems
Dangerous structure
Surplus oil
Transshipment mat. explosions