"The goal is to turn data into information, and information into insight." – Carly Fiorina, former executive, president, and chair of Hewlett-Packard Co.

Prediction of high risk fire area

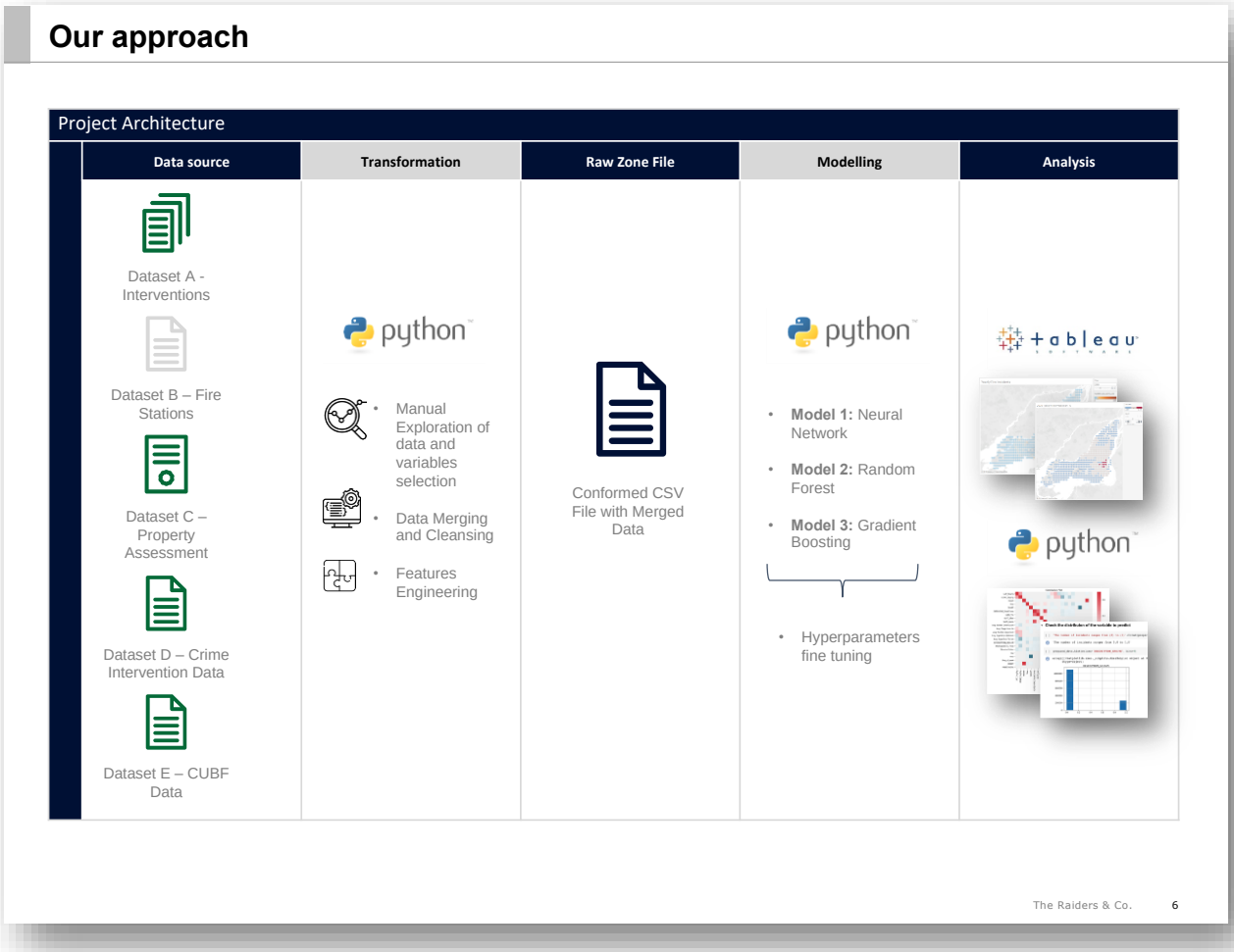
City of Montreal

Ikram Mecheri
Maria Papadopoulos
François St-Amant
Stanley Tran

Table of Contents

1. Introduction

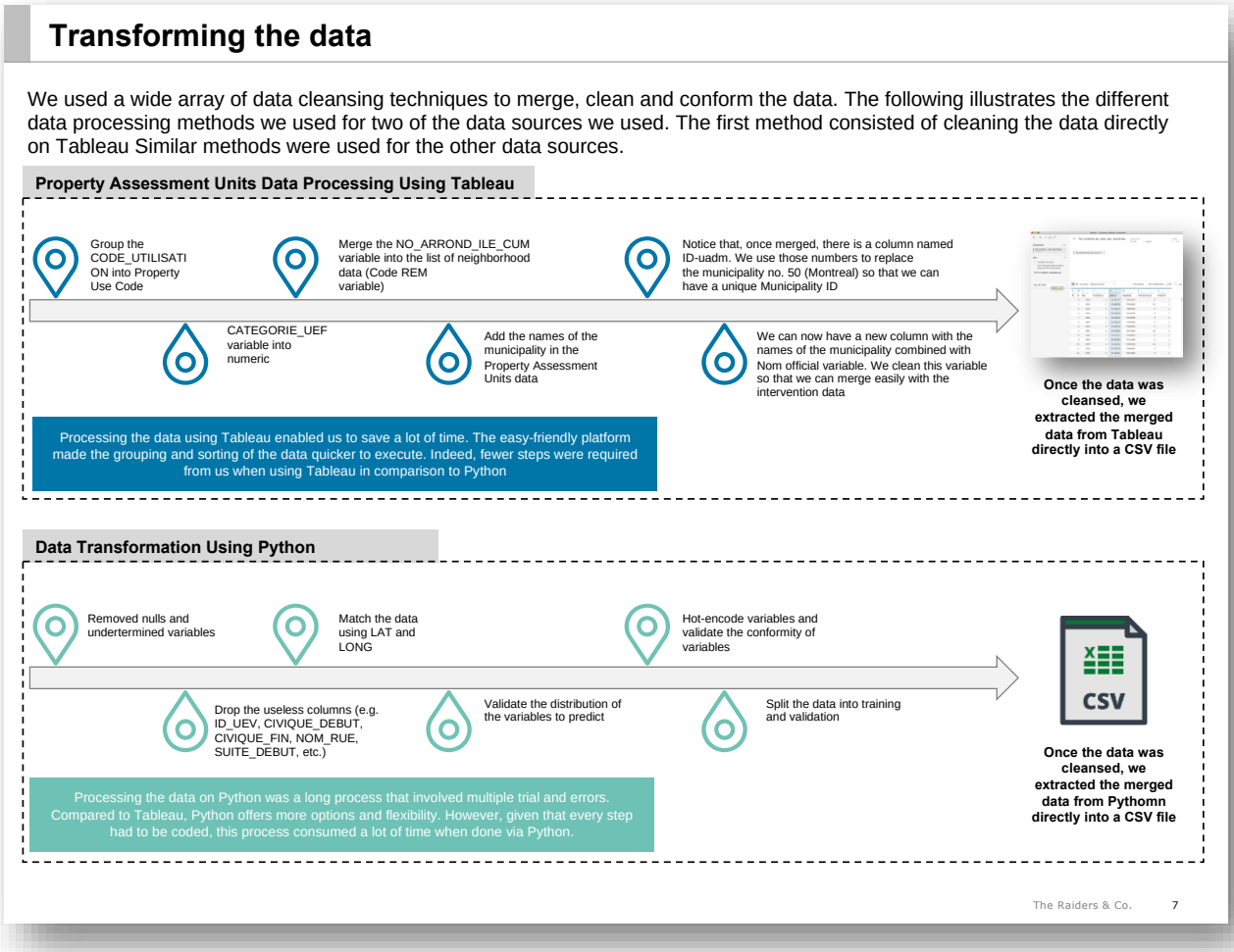
The city of Montreal, the second largest city in Canada, holds population estimated at 1.75 million as of 2016. The city offers to share their data across their website. In the course of this project, we are trying to analyze the firefighter's dataset in order the predict high risk fire areas in the city of Montreal. Since 2005, data has been collected by the Montreal Fire Department (including interventions, coordinates of the Fire Departments of the city, property assessment data and criminal acts) to produce reports required by the Ministry of Public Security, also reported to the Police department of the City of Montreal. We will be using those collected data to analyze, do fire predictions (using statistical and machine learning models), in which we can navigate and generate insights.



2. Data analysis and mining process

Before any type of analysis can be done, data must first be gathered and cleaned. For this project, the challenge was to clean each large dataset and merge them into one comprehensive dataset. It would be difficult to simply take datasets to generate insights and make predictions using machine learning models. For Project A, 4 datasets concerning statistics on the city of Montreal were given:

interventions, fire stations, property assessment and crime. Each dataset had a different structure and content making it difficult to merge without proper transformations.

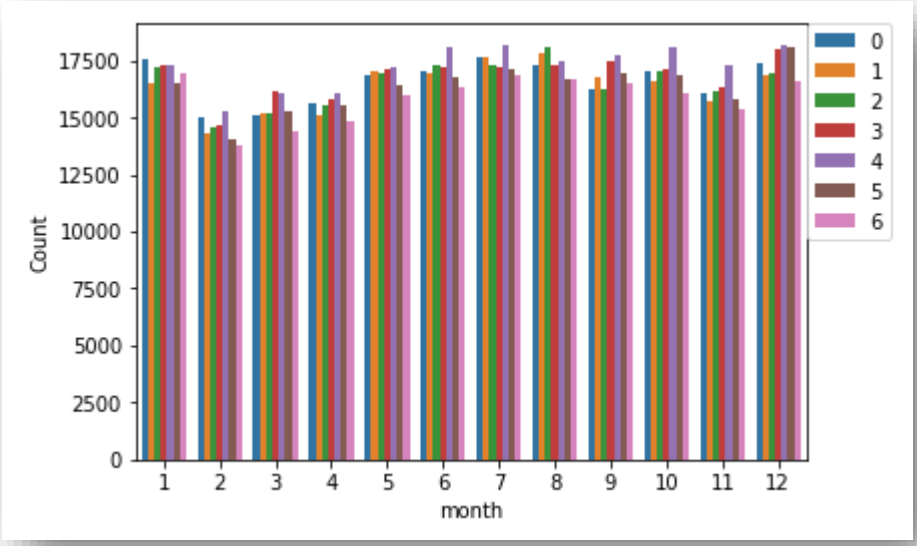


a. Interventions dataset

This dataset described the interventions taken by the city of Montreal for various emergencies. Since we want to predict fire, certain features were dropped since they were a result of an emergency and therefore did not cause an emergency. For example, the type of intervention cannot predict fire but location and date might. The date column was further separated into year, month, time of day and weekend categories. The feature “Description Group” was the main category of describing the type of intervention taken. This group was split into fire incidents (ex: “other fire”) and non-fire incidents (ex: “first response”). Using this method of separating the “description group” results to a dataset that contains ample amount of fire and non-fire data without having to generate non-fire data for the future predictive model. Although this method resulted to an unbalanced dataset, it was not decided to not do under sampling or oversampling.

Some general analysis was done to help with the story and approach to modeling. Some points include how the winter months tend to generally have lower amount of fires. Fridays appears to

generally have the highest amount of fires throughout the months. Influence of location was done at a later step when Tableau was used.



It's important to note that the interventions dataset had been anonymized before publication on the City of Montreal's website. This means that the latitude and longitudes of incidents in the dataset had been rounded to the nearest street intersection instead of the address of the building. For this reason, in contrast to the accuracy that was demonstrated in the Ville de Montréal's presentation, the insights of our project will not be able to achieve the high resolution of risk of fire per building. Our model will therefore be cast on a broader area instead. Since the borough can be yielded across all given datasets, it was chosen as the locational feature for our models.

b. Firefighter Stations dataset

After some analysis, this dataset was discarded from the analysis since the goal of this project was to predict the risk of fire. The firefighter station location would be more beneficial for projects with a different scope such as the analysis of the time between fire stations and fire incidents and so forth.

c. Property Assessment dataset

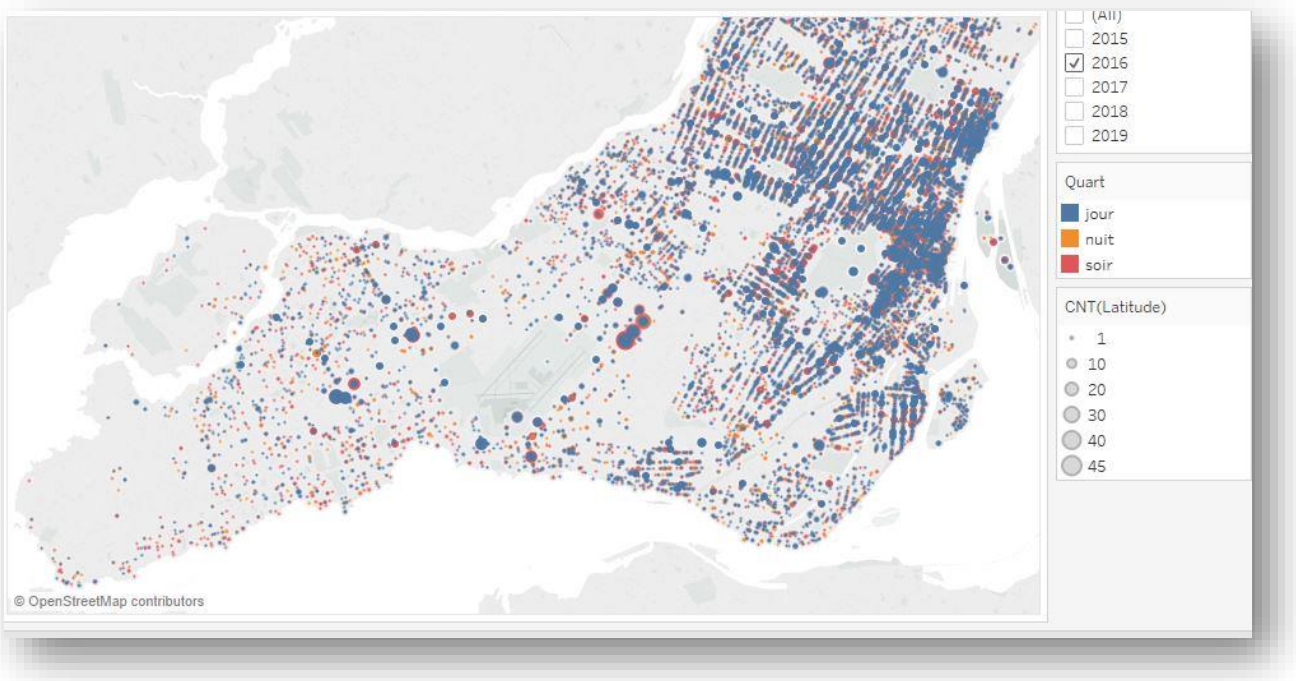
The first step taken was finding an approach to clean the property assessment dataset for it be easily merged into the intervention's dataset. The interventions dataset contains the locational features latitude and longitude as well as city and borough names. The property assessment dataset had multiple features referring to location. First, the property assessment data was merged with an external dataset that listed the borough names using the REM code as the common variable in both datasets. This resulted in the property assessment having a clean feature of borough names. Second, the feature concerning municipality in the property assessment data was given as a numerical value. Using the property assessment data website, this numerical feature was decrypted into a feature containing the name of the municipality. Further cleaning was needed for the borough and municipality names to enable a proper join between this dataset and the intervention data.

Additional features were created for future analysis by our training models. Examples of such features include average year of construction, average building area and so forth.

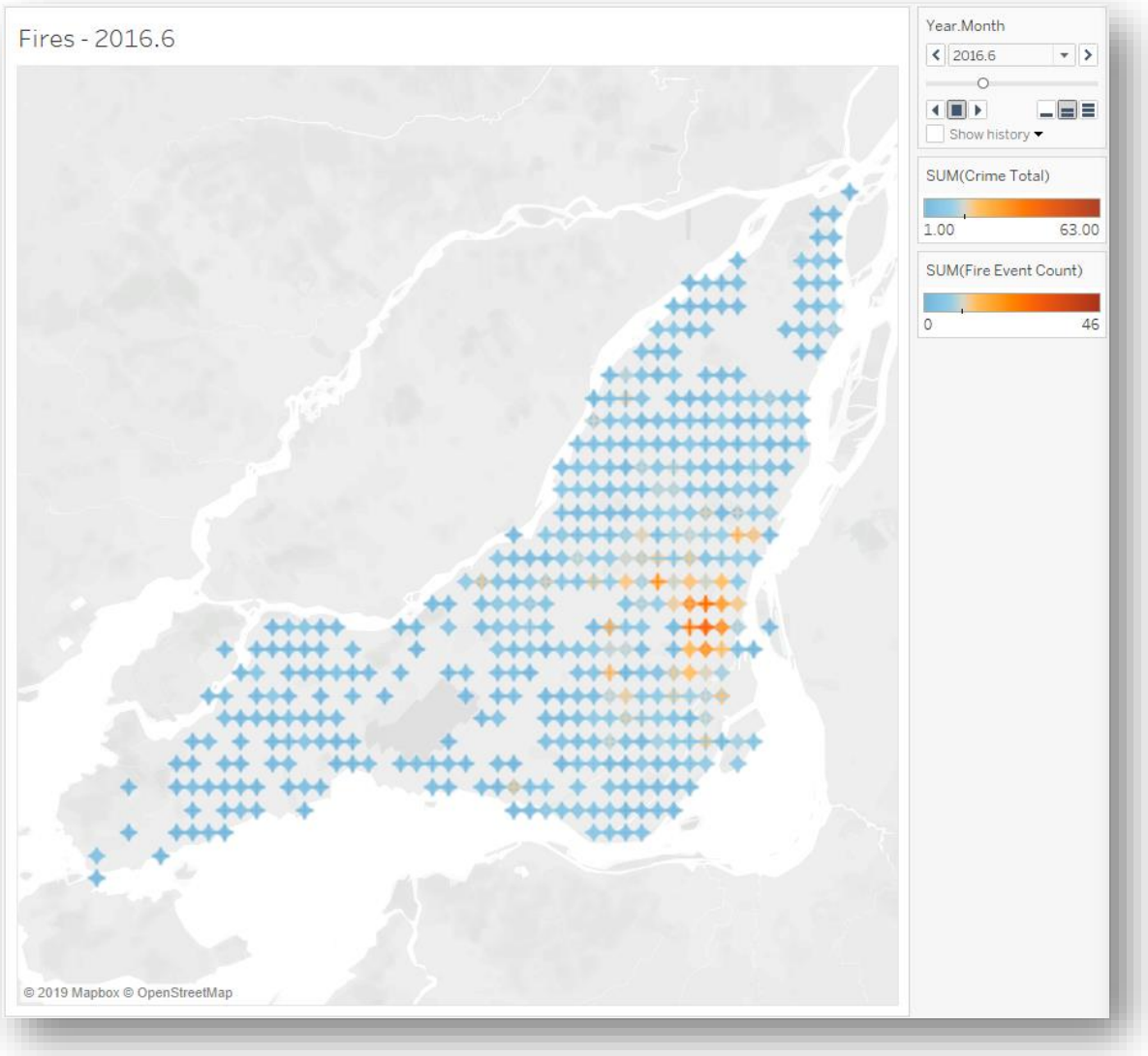
d. Crime dataset

The crime dataset only contained latitude and longitude as locational features. Reverse geocoders were attempted to convert this type of geographical data into borough and municipality datasets. Due to the amount of data, timeout errors and other challenges occurred. In order to merge with the intervention's dataset, latitude and longitude were binned into grid locations. This was achieved by rounding the given six significant figures into three. The date feature was separate into year, month and weekday features. New features were created as well. The average amount of crime per grid location, year, month, weekday and time of day was done. All these features are mergeable with the intervention's dataset. It is worth nothing, however, that this analysis would be considered static if used for modeling. In other words, 2019 predictions would assume a certain location has high/medium/low crime. Future work would be to create a model unique to crime to predict 2019 crime rates per grid location to then input into the fire risk model's dataset.

Tableau was used to see the general distribution of crime per area given the year, month, and time of day.



We then merged the Fire dataset with the Crime dataset to investigate any potential relationship between the two.



3. Features Engineering

a. Defining the problem

Before starting any feature engineering, we had to settle on the variable parameters in the question we were looking to answer. Though we want to predict the areas with a high risk of fire in the city of Montréal, the size of the area and the period for each prediction was left open for us to decide.

b. Area

We were first tempted to use the borough as the area but then discovered that if we chopped a few decimals from the GPS coordinates, we could easily group incidents by a smaller area. We settled for 2 decimals, which yielded areas of about 1 km^2.

c. Period

The exact date and time of each fire is available in the incident dataset. However, knowing that the firefighters would use the prediction to plan the building inspections for the upcoming weeks, it quickly became obvious that a monthly prediction would make sense.

d. Official problem definition

We are trying to predict, on a monthly basis, the risk of fire for each km² in the City of Montréal.

e. Building the features

For our model to make a prediction, we need features that we can easily generate for the future. That rules out the crime dataset as well as the demographics dataset since we would have to build models that make predictions on each of these and feed the results back into the fire risk prediction model we want to build. Although it would be an interesting exercise, we feel that we did not have the time to achieve this within the few weeks that last the class. We therefore chose to use the intervention dataset, since it contains what we want to predict, and the buildings dataset, since it varies less year over year.

f. Interventions

Our choice of area and period being made, we now have to group the intervention data by these criteria. This was achieved through some python code that looped through the raw intervention data and compiled several statistics for any given area + period combination. The figure below pretty much sums it up.

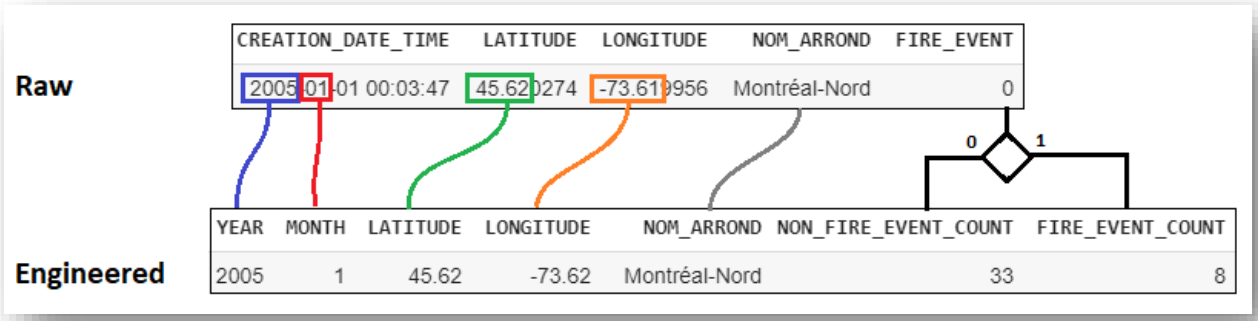
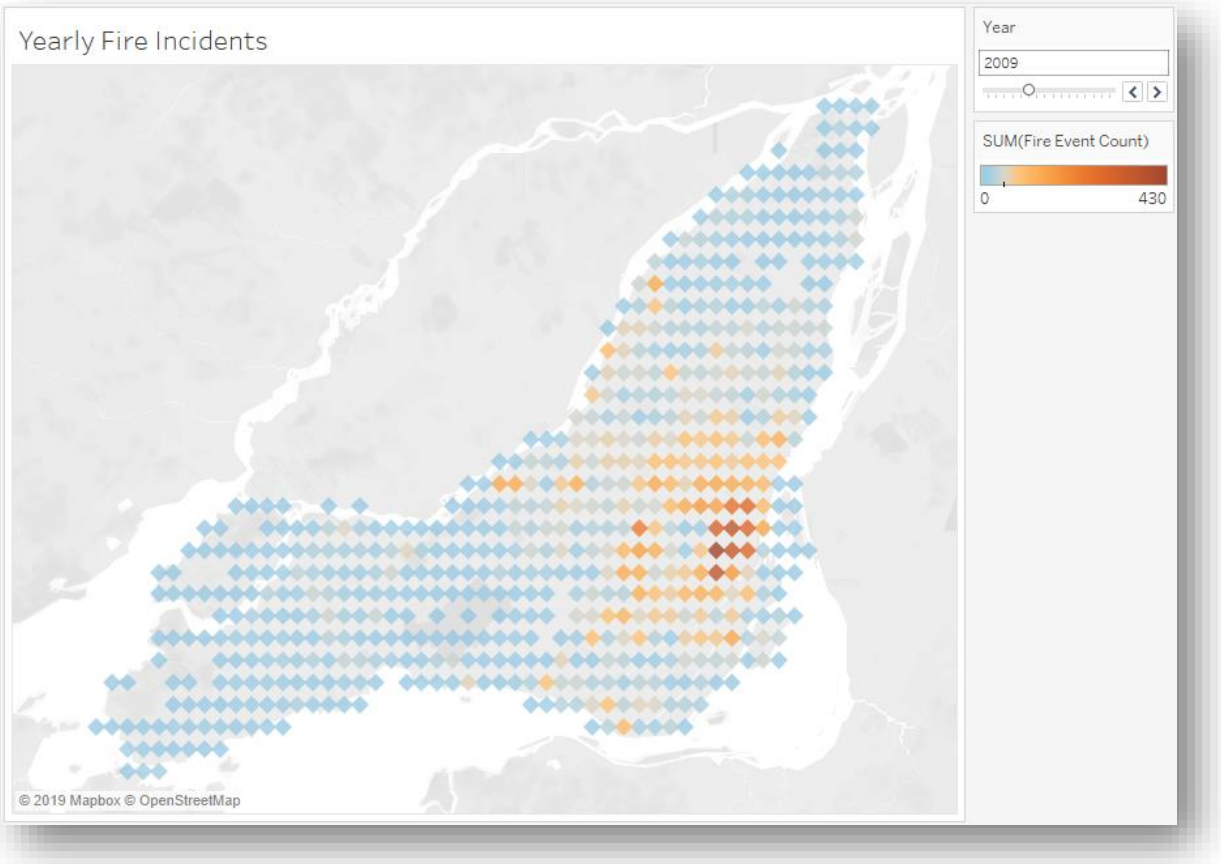


Figure 1- Feature generation from the raw interventions dataset

We kept the NOM_ARROND column as it will be useful later to merge the Buildings feature since they lack the GPS coordinates.

With the data binned as such, we can now visualize the historical data on the actual map of the city of Montréal, which can help us get some insight of what could drive the risk of fire up in a given area.



According to the observation, it seems like the risk of fire is higher around certain educational institutions: Cégep du Vieux-Montréal, UQAM and Université de Montréal. It would be interesting to see it the risk of fire goes down during the summer break in those areas.

g. Buildings

The buildings dataset contains the address of every building in the city of Montreal, but not their GPS coordinates. There were services to convert the address to coordinates, but we either had to pay for them or they timed out before we got the results. We therefore settled to grouping the buildings by their borough.

We think the number of buildings, the number of stories, the number of individual units, the average age, the land and building areas are all interesting features to calculate for each borough.

Furthermore, the buildings are also associated by their usage. In fact, there are 523 different usages in the available dataset! Since we selected 6 features to extract and did not want to end up with 3138 more features (6 features X 523 usages), we binned the usages into 16 different groups, including a special group for student housing in the light of our insights while engineering the features from the interventions data.

To match the interventions features as much as possible, we compiled the stats for each borough by year. That meant leaving out of the equation buildings that were built in years after the current year we were compiling the data for.

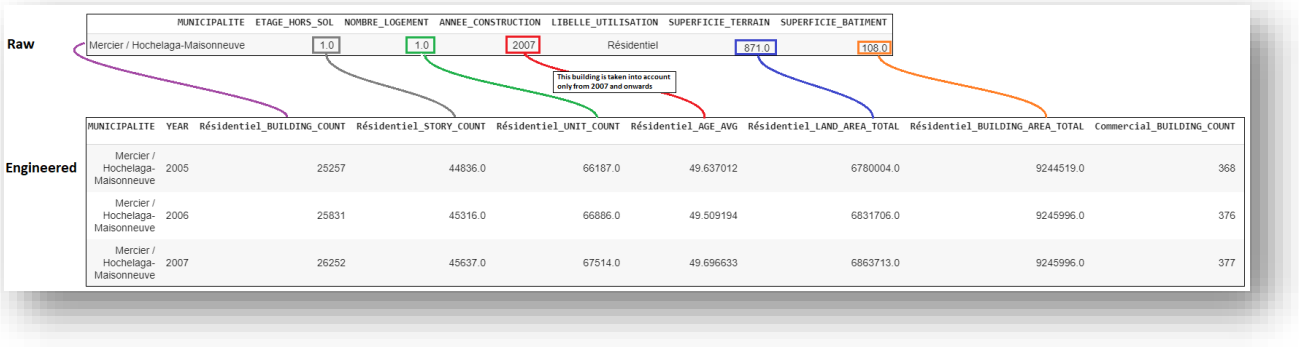


Figure 2- Feature engineering from the buildings dataset

Merging the features

We used the borough as the common key to merge the interventions and the buildings feature datasets. For each entry in the interventions features, we added the buildings feature from the matching borough and year.

4. Statistical/Machine Learning Modeling

a. Random Forest Classifier

Final data preparation

We binned the number of fire per month per borough into 3 risk categories:

- Low: 0-4 fire events
- Medium: 5 to 19 fire events
- High: 20 fire events or more

This generalization of the possible outcomes will allow our model to make more accurate predictions.

We separated the data with years 2005 to 2017 as the training data and kept year 2018 as the testing data. Once the data was split, we dropped the YEAR column as we did not want it to influence the prediction.

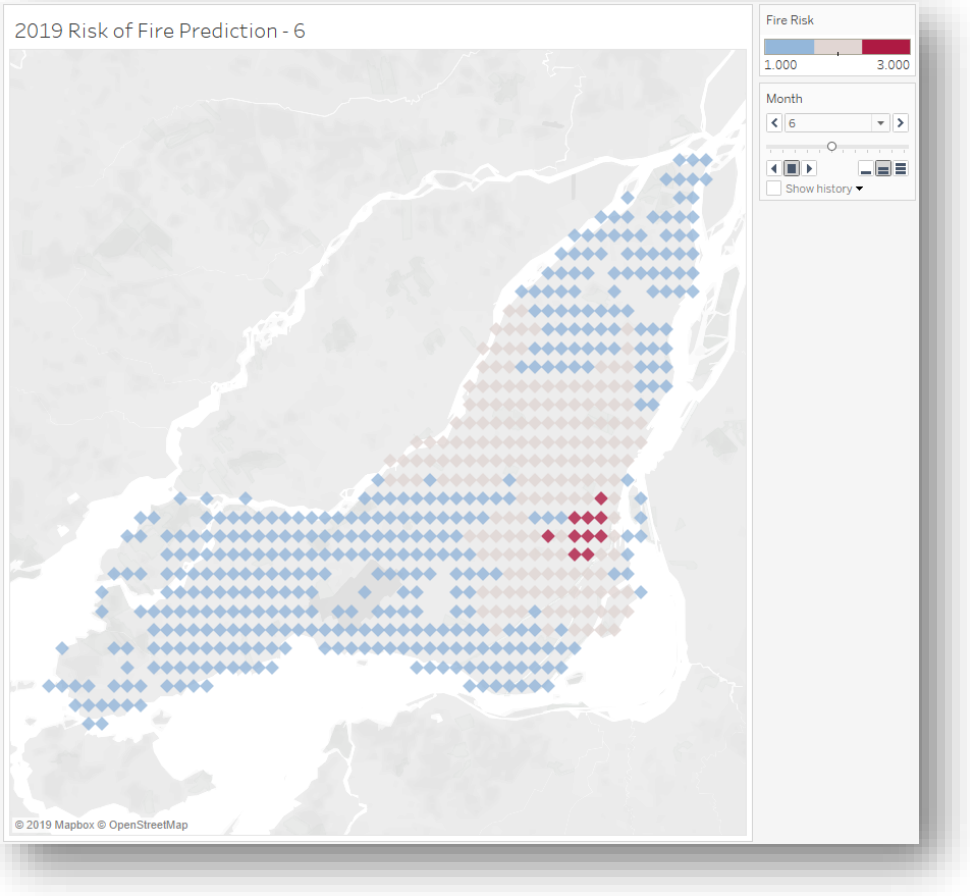
Predictions

On predicting the 2018 risk of fire, our random forest classifier model scored an 85.64% accuracy. This was achieved through fine-tuning the hyperparameters of the random forest model as well as using features that had an influence greater than 1.5%. The hyperparameters used are 200 estimators and a depth of 18. The features used and their contribution are as follows:

LONGITUDE	0.097930777
LATITUDE	0.087687498
Loisirs_BUILDING_COUNT	0.060631984
Construction_STORY_COUNT	0.057189645
Etudiants_BUILDING_AREA_TOTAL	0.054656795
Sante_BUILDING_COUNT	0.047336502
Commercial_UNIT_COUNT	0.043910328
Etudiants_LAND_AREA_TOTAL	0.040082593

Etudiants_BUILDING_COUNT	0.03373051
Etudiants_STORY_COUNT	0.030330696
MONTH	0.026565314
Lieu de culte_AGE_AVG	0.019353091
Industriel_UNIT_COUNT	0.019029218
Commercial_BUILDING_COUNT	0.018583265
Education_STORY_COUNT	0.017258286
Residentiel_BUILDING_AREA_TOTAL	0.016595754
Securite publique_AGE_AVG	0.016081102
Industriel_AGE_AVG	0.015520171

To predict the 2019 risk of fire, we simply took the 2018 data and added a year to every building age average before asking the model to make its predictions. The figure below shows the results.



5. Exploring more data to assess significance

After achieving a prediction using our feature engineering, we wanted to assess if the crime data could help us make a better prediction.

a. Neural Network with Adabound optimizer

Since we have a lot of data, we can use Neural Network to predict fire. The optimization algorithm is the main approach used in order to minimize errors. So far, the most common used are Adam and

SGD. In recent papers, we discovered that there is a model that is as fast as Adam and as good as SGD, which is called Adabound. From our understanding, Adabound is an algorithm that employs dynamic bound on learning rate using Adam and gradually transitions into SGD[11]. The model is built with the following settings:

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 1000)	22000
batch_normalization_1 (Batch	(None, 1000)	4000
dropout_1 (Dropout)	(None, 1000)	0
dense_2 (Dense)	(None, 500)	500500
batch_normalization_2 (Batch	(None, 500)	2000
dropout_2 (Dropout)	(None, 500)	0
dense_3 (Dense)	(None, 1)	501

Total params: 529,001
Trainable params: 526,001
Non-trainable params: 3,000

Instead of training the model with data from 2005 to 2017 then predict 2018, we use the whole dataset from 2005 to 2018 and split randomly to have a batch of training sets (70%) and testing sets (30%). This way, we can train the model with data that are more uniform, with different years well distributed. In addition to this, we also normalize all the numerical data so that all values are between 0 and 1 to improve data integrity. In other words, we ensure that all variables of the data are on the same scale.

Without the feature engineering process, we are able to achieve an accuracy of 86.4% testing sets. With feature engineering, we achieved an accuracy of 79.7% testing sets.

b. Gradient Boosting with Feature Engineering

We tested another model, gradient boosting using our featured engineered data. We were able to achieve a 81.59% accuracy. Gradient boosting uses regression trees to predict a future state based on the historical data. Gradient boosting transforms weak learners into strong learners. This is particularly useful when the data is sparse and comes from disparate sources. We believe that the low accuracy is due to the fact that Gradient Boosting is Sensitive to overfitting when the data is

[1] https://syncedreview.com/2019/03/07/iclr-2019-fast-as-adam-good-as-sgd-new-optimizer-has-both/

noisy and that we would need to increase the number of estimators to train our model. Unfortunately this option would need more computing resources.

6. Insights discovery and future improvement process initiatives

We used Tableau throughout this journey to process and visualize our data. In this dashboard, we combine the Intervention Data, along with the Crime Data to investigate the relationship between these two variables. As shown by the figures above, there seems to be a correlation between the areas with higher crime rates that also have a high risk of fire.

We can also observe that areas around educational institutions, such as Côte-des-Neiges or Berri-UQAM where a lot of students live (Cégep du Vieux-Montréal, UQAM) are also associated with a higher risk of fire.

In order to predict Fire risk using the Crime data, we would need to create a model that would also predict the risk of crime and then merge it with the Fire risk data to investigate this relationship.

One of the challenges is the time constraints with this project. With more time, here is a list of initiatives we can work on in the future:

- Merge crime data to do further analysis and understand the relationship between the crime and the fire intervention data
- Gather external data from Environment and Climate Change Canada. In addition to the crime data, we can also study the correlation of climate change with the fire occurrences.

7. References

[1] ICLR 2019 | 'Fast as Adam & Good as SGD'— New Optimizer Has Both

https://syncedreview.com/2019/03/07/iclr-2019-fast-as-adam-good-as-sgd-new-optimizer-has-both/