# Data Science & Machine Learning Capstone Project

## Six Months Fire Predictor Modeling

March 2019

TEAM
CARLOS FERREIRA
DIMITRIS KATSAOUNIS
HESHENG CHEN
LEIDY BARRERA
ANDRES URREGO

# Table of contents

# Objective

Predict if there will be a house fire incident in next 6 months for any location where the Fire Department intervenes. We base our prediction in the 6 months history of interventions at the specific location combined with data from 2016 Census Canada,  house data properties from Ville de Montreal, and Crime data from the city of Montreal.

# Motivation

Help the Fire Department of Montreal to prepare better for fire incidents by identifying the approximate location of a house fire intervention in the next 6 months  with good accuracy. A good prediction can help the FD better schedule its forces and arrange possible road works outside the risky window. And at the same time prepare the necessary infrastructure.

# Methodology

To address this problem with fire intervention prioritization, we developed a predictive model to determine  location-level fire risk (i.e. likelihood of a given location having a fire incident in a given 6-month window). To develop this model, we used Fire Department historical data of all types of incidents from 2009 to 2018 from Ville de Montreal, data from Census Canada with approximately 300 demographic features of Montreal, and Crime data.

We joined these data sets at incident intervention level (longitude, latitude). The values of the features used for the specific intervention are derived from the weighted average of the inverse distance of the values of the Census features of the 4 closest boroughs.  Also in a similar way we include house data for the intervention location from the Unités d'évaluation foncière.  Then we looked at the history of this intervention location for 6 months and add all intervention types separately. Next we look at the future of this location and check if there is a fire in the next 6 months. This way we enrich the interventions data with house and social features about each location, and history of interventions and fire occurrences within the next six months. So we train our models with the future fire incident as label and all the rest data as features.

The models we tried are: Decision Tree, Random Forest and XG boost. We evaluate our model on Precision, Recall, F1-score Kappa score and AUC metrics that are appropriate for this kind of problems and we come to the conclusion the the Random Forest model has the best combination of scores with relatively high accuracy and good recall score and the best F1_score.

# Initial problem proposed & final problem solved

**Initial problem**
Our first approach was to try to predict the daily number of interventions from each fire station in the city of Montreal.

Although some patterns were found in the data like the number of interventions during the night hours, this was not enough to help us come up with a clear prediction, even when we integrated **weather data** into our analysis. Unfortunately this initial approach proved to be too random and no prediction was producing good results.

**Final problem solved**
Finally we built a system that is able to predict the fires in the next 6 months with about 62% accuracy. Meaning that if the system predicts that there will be a fire, there is a 62% chance that it will be correct and there will be a fire.

The model can not tell the Fire Department exactly when the fire will occur but that there will be a fire in the next six months. The Fire Department can use this model for better planning of resources, making sure that it will have the right equipments to battle fires at the right place and making sure that there will be no road works in those locations during the 6-month window of fire prediction.This is a serious prevention tool. For each intervention they make they could run the model before they dispatch resources to the location of the incident, and the model will tell them if a fire is predicted in the next six (6) months at the intervention location they are going.

This prediction has about 62% of accuracy and is definitely great advancement. We expect that the model will continue to be updated with the intervention history, location from Census data, and housing data.

# Assumptions

1. To assign feature values to a location using the Census Canada data we assumed that in moving

for one borough to the next the feature value change smoothly. So two house next to each other but on belonging to different boroughs will share very similar feature values. In plain english this means that if in one house is located at one side of the street and belongs to Outremont and the next house at the other side of the same street belongs to Cote de Neiges their assigned feature values will be similar and not the average values of their respective boroughs. This creates a smoothing of the features.

2. The model used 6 months of historical data and looked 6 months into the future for an occurrence of a fire incident in the specific intervention location.
3. The six months is a logical period since it gives the FD the time to plan accordingly.

**Related Work**

We looked at other analyses that were made worldwide to help other fire departments in the prevention of building fire risk. The majority of the analyses pointed towards the prediction of risk of fires during a given timeframe.

So for example in the project to generate a building fire risk score for every address in Baton Rouge, LA the goal was to predict the fires in the next six months using data of the houses, income levels and crime data for the city of Baton Rouge.
Reference: "How we predicted building fires in Baton Rouge, LA -- working version"

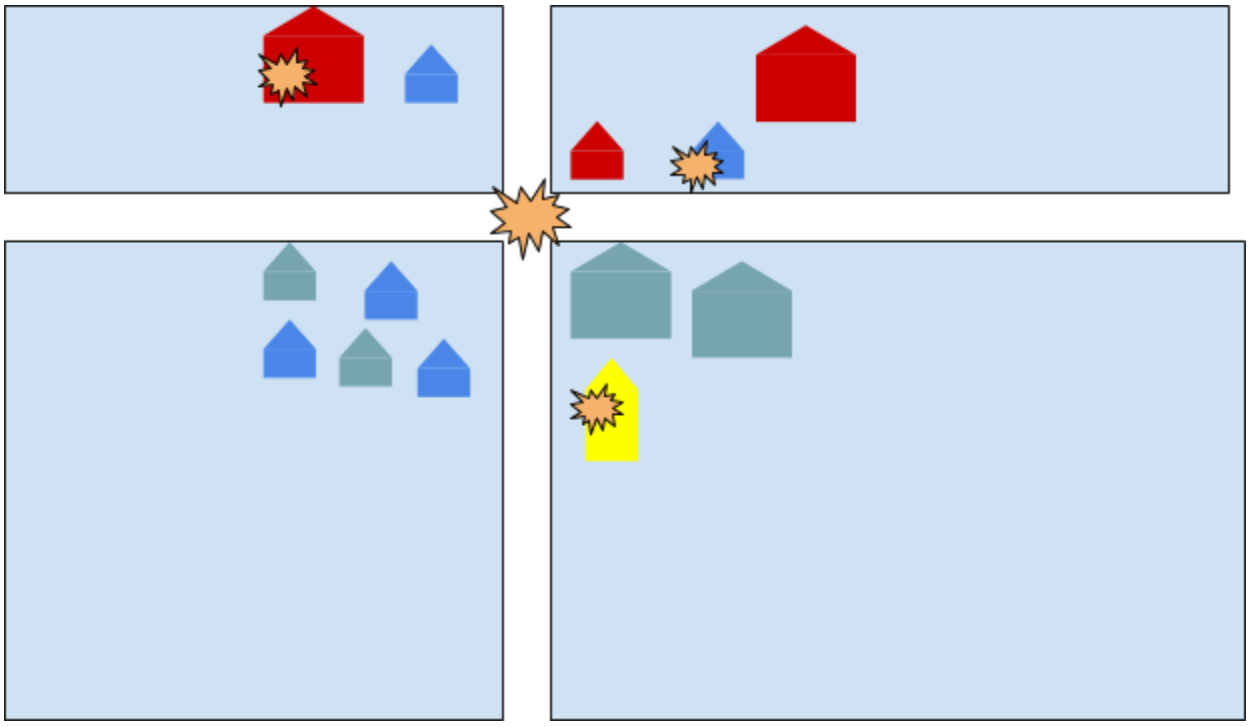
Another study we looked at was the predictive modeling of building fire risk for the city of Pittsburg, which tried to build a predictive model to prioritize property fire inspections at property-level that calculated the likelihood of a given property having a fire incident in a given 6-month time period.
Reference: "Predictive Modeling of Building Fire Risk, Metro21: Smart Cities Initiative"

After looking at those two studies and having realized that the initial approach was fruitless we decided to switch our approach and try to follow similar steps thus estimating the fire risk of each location in our dataset.

Unfortunately the data provided by the city of Montreal is obfuscated for privacy reasons and this creates a few interesting challenges that are explained below among the other difficulties we faced

- The interventions data is more accurate starting on year 2009 since in 2008 the Fire Department was assigned the task of First Responders which broadened the job of firefighters
    - To mitigate this we decided to use data after 2009-01-01 in our analysis
- The interventions are recorded in the closest intersection instead of the actual location they occured.
    - This was a more difficult problem to deal with since the real incident is not revealed in the data but rather a projection of the incident in a different subset of addresses. We can better understand this constraint through the figure 1:

Figure 1



Let's assume that a fire is recorded on the small blue house at the top right block. This incident will be recorded in our dataset on the marked street intersection. However, the same marking of street intersection will be assigned to a fire that occured in the large red house at the top of left block and in the yellow house at the lower right block.

- The same type of obfuscation is true for all types of Fire department interventions.
- This made it hard to replicate the processes used in the studies, and hard to enrich our dataset with accurate data regarding the fires like the property year built, the lot area, the dwelling size, the income of the people in the property, who was involved or even the family status at the place where the incident occured.
- To address this, we decided to proceed assuming that the street intersection where the incident is projected is were we will be predicting fire interventions in the next 6 month window for the purposes of our project.
- This assumption is very important in the continuation of our project as it allowed us to join data from different sources by projecting them to the intersection of the location. Having made this assumption enabled us to make data augmentation by adding external data to our data modeling.

# External Data Used

For our analysis we used the following data sources

- The interventions data from the fire department of the city of Montreal
- The data about Crime in Montreal
- Data from the 2016 Census for the Agglomeration de Montreal
  Annuaire statistique de l'agglomération de Montréal - 2016
- And data regarding the location and sizes year built, lot area etc. of the houses from the the city of Montreal.
  Unités d'évaluation foncière

# Data Analysis Tools & Techniques

**Data analysis tools and techniques used to solve the problem**

Data Visualization

- Tableau
- Python for Data Visualization

Data Preparation

- Alteryx
- Excel for Data Preparation

Machine Learning Tools

- Python
- Alteryx

The models we used for our prediction were **Decision Tree Analysis**, **Random Forest**, and we also tried the **XGBoost** model that seemed to have worked well in the Pittsburg case.

# Relevant Analyses

**Work relevant analyses, models, findings, and results**

We started by acquiring datasets from the sources available that contain data about the interventions of the Fire Department, and we identified the intervention categories as follows:

- Autres_incendies
- Incendie_de_batiments
- Premier_Repondant
- Sans_incendie
- False_Alertes_Annulations
- Alarmes_incendies

We also know the obfuscated location of each incident, therefore we could also join the following information to the dataset

- Crime data from the city of Montreal.

Now for each location in the interventions table we need more features. We acquire the data required by joining data from the other two datasets described above

- The unites de evaluation fonciere
- The 2016 Census of the agglomeration de Montreal.

To join the data we used the location field in both cases:

For the “unites de evaluation fonciere” we have

- location of each house
- year built
- area of the lot
- area of the house
- number of floors

In order to join the data with the interventions table we used the following logic:

1. For Each house in the “unites de evaluation fonciere” table there is a closest location in the interventions table. So we pointed each house to that intersection
2. Next for each intersection in our Interventions table we assigned each of the four (4) features
    a. Year built
    b. area of the lot
    c. area of the house
    d. number of floors

to be the average value of the houses features in the ““unites de evaluation fonciere” dataset that pointed to that particular intersection location.

For the “2016 Census of the agglomeration de Montreal.” it is a little harder to get useful information to fill our interventions data. The reason is that although we have a lot of features, we only have them for 33 boroughs and not for each of the thousand of intervention locations.

To work around that we decided to make the following assumption:
We assumed that the values of the features are smoothly distributed across the boroughs. This means that in the previous figure (figure 1), the houses around the intersection would all be similar in characteristics regardless if the street defines a change in borough. This is a very logical assumption if we look around us in Montreal, and helped us assign values to features we want for the intersections according to their distance from the center of the borough, using the following logic:

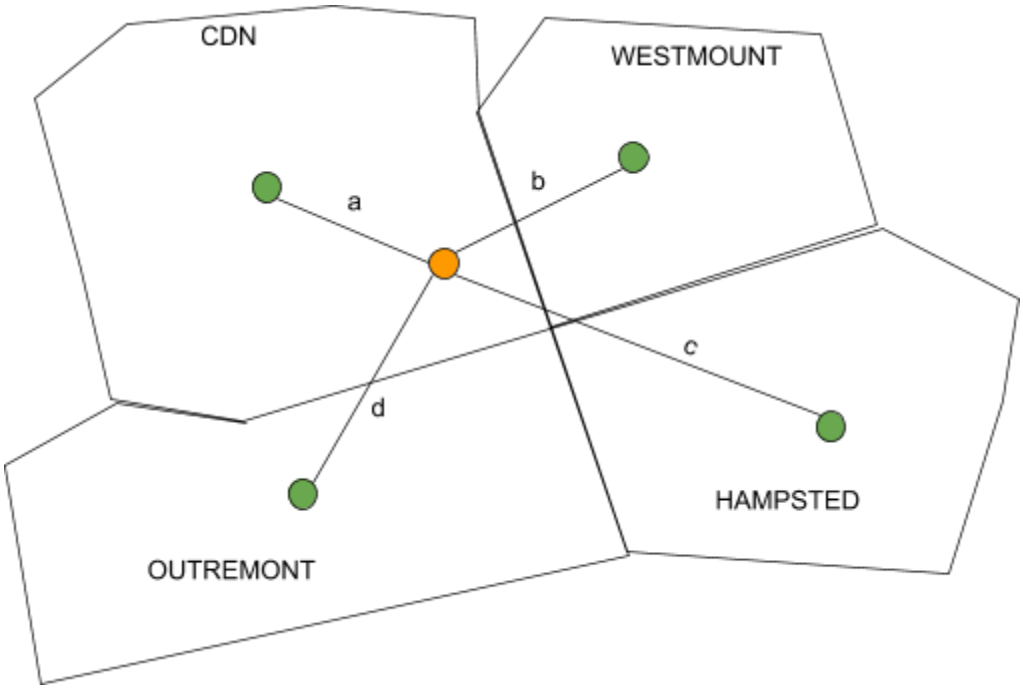
**The problem identified**
We have demographic and economical data from Statistics Canada for the 33 boroughs of Montreal but we do not have specific data for each incident location (Longitude and Latitude). One idea would be to assign each incident location the value of the borough where it belongs. However, for all incidents in the same borough we would have the same data. Therefore, all of our incidents would be grouped to 33 similar groups.

**The solution adopted**
Instead we decided to try the following approach:

1. We calculate the average Longitude and Latitude for each borough in Montreal, by grouping and averaging all the incidents since 2009-01-01 that have occured in that borough. For the purpose of this project let's call this value the Center of Borough (COB is presented as a green dot in figure 2 below)
2. Next for each incident (orange dot), we calculate the distance from the 4 closest centers of Borough (COB) shown in green in figure 2
3. Finally we assign a calculated value to all the features in the dataset for that specific location (orange dot) based on the weighted average value of each of the 4 values of the same feature in each borough

   Here is a graphical representation of the solution adopted (borough names, proximity, real areas used in the picture used for explanatory purposes only)

Figure 2




a = Distance of incident to CDN COB (green dot) = 1.3km
b = Distance of incident to Westmount COB (green dot) = 1.6km
c = Distance of incident to Hampstead COB (green dot) = 2.9km
d = Distance of incident to Outremont COB (green dot) = 2.2km

Also here are the average price of homes (numbers are for explanatory purposes only) for each area in the city as shown in the 2016 Census data for the agglomeration de Montreal:

- CDN = $350K
- Westmount = $850K
- Hampstead = $750K
- Outremont = $550K

Now we assign the weight for calculating the value of the house as it is affected by the four (4) closest boroughs, and we calculate the weighted average with this formula:

$$\text{Value of house} = (w_1 V_1 + w_2 V_2 + w_3 V_3 + w_4 V_4) \,/\, (w_1 + w_2 + w_3 + w_4)$$

The closer a house is to the COB the closer its price should be to that of the COB. To simulate that we assign as weight is the inverse of the distance to the COB. So if distance is 0 (the intersection coincides with the COB) then the wight is infinite and the feature value becomes the COB value

In our case
$w_1 = 1/1.3$, $w_2 = 1/1.6$, $w_3 = 1/2.9$, $w_4 = 1/2.2$ and $V_1 = 350K$, $V_2 = 850K$, $V_3 = 750K$, $V_4 = 550K$

So:

$$\frac{(\frac{1}{1.3}350K+\frac{1}{1.6}850K+\frac{1}{2.9}750K+\frac{1}{2.2}550K)}{(\frac{1}{1.3}+\frac{1}{1.6}+\frac{1}{2.9}+\frac{1}{2.2})} = 596,781$$

This calculated value is closer to the reality than assigning the value of $350K for that particular house since this is close to Westmount, Outremont and Hampstead, boroughs which have higher home values.

From the designed dataset we kept the following features:

**Location related features**

- Couples_No_Children
- CouplesWithChildren
- SizeOF House
- Detached_House
- Appartment_FiveFloors
- OtherType
- Semi_detached
- TownHouse
- Duplex
- Appartment_less_5Florrs
- OtherDtached
- MObileHome
- 1_4Rooms
- 5_rooms
- 6_rooms
- 7_rooms
- 8_RoomsOrmore
- Average_Rooms
- Simple_Maintenance
- Major_repairs
- Average_House_Value
- AverageHouseholdIncome2015

Now that we have filled our Intervention table with prepared data we can move to the final stage of our process which is to include history for each intervention. By history we mean that we will add features regarding the existence of the various types of incidents in the specific locations.

This is easy to do as it can be derived from the interventions table just by adding the type of incidents that were recorded for each particular location in the last six months.

Our idea was to see whether the history of a location plays a role in predicting fire risk given also other data about that location. Therefore, the new features are:

**History related features**

- Sum_Autres_incendies
- Sum_Incendie_de_batiments
- Sum_Premier_Repondant
- Sum_Sans_incendie
- Sum_Alarmes_incendies
- Sum_False_Alertes_Annulations
- Sum_Crime

The final version of the dataset looks like this:

| **Features not included in the model** | | **Features** | | **Label** |
|---|---|---|---|---|
| Location - not used as a predictor in the model, only to create the table | Date - not used as a predictor, only to create the table | **History related features** looking back 6 months (7 features)<br>Sum_Autres_incendies, Sum_Incendie_de_batiments, Sum_Premier_Repondant, Sum_Sans_incendie, Sum_Alarmes_incendies, Sum_False_Alertes_Annulations, Sum_Crime. | **Location related features** (25 features) | Was there a fire in the NEXT 6 months?<br>(look into the future to respond) |

**ML model prediction**

The models are fed with the **location related features** (house value, how many houses around have 1 room, 2 rooms, how many houses have couples and how many have couples with children, etc.) and the history of interventions **(history related features)** at each particular location during the last 6 months, and our **model will predict whether there will be a fire in the next six (6) months at the location of that particular intervention**.

# Results & Findings

It is proper to note that in this case because our dataset is skewed (94% - No, 6% - Yes), as we have very few positive responses in the total number of observations, the proper measurement metrics for the results are the **Precision**, **Recall**, **f1_score**, and **Kappa** metrics.

Here is a summary and comparison of the three models we implemented:

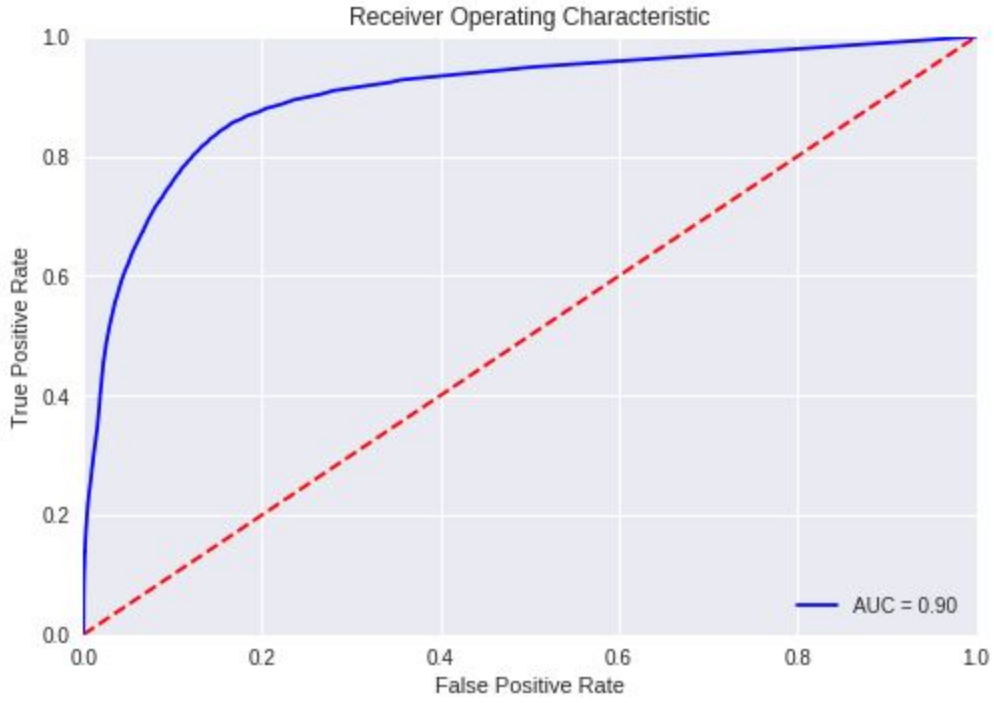
| Model | Recall | Precision | F1_Score | Kappa | AUC |
|---|---|---|---|---|---|
| Decision Tree | 45.5% | 43.9% | 44.7% | 40.3% | 0.72 |
| **Random Forest** | **44.6%** | **61.7%** | **51.8%** | **48.6%** | **0.90** |
| XGBoost | 12.9% | 81.4% | 22.3% | 20.7% | 0.90 |

From the results we see that the **Random Forest** model was the best in terms of performance.

**Confusion Table**

| | | Predicted | |
|---|---|---|---|
| | | No Fire | Fire |
| Actual | No Fire | 181,173 | 4,045 |
| | Fire | 8,105 | 6,524 |

Figure 3



**Results description**

- The 62% Precision means that when our model predicted a fire, 62% of the time it was correct
- The 45% Recall means that if 100 fires happen in the next six months at the intervention location, our model would have predicted correctly (captured) 45% of those fire incidents
- The balance between the two metrics (Precision & Recall) is very important and is captured by the two other metrics the **F1_score** and the **Kappa**

We want a model that not only is correct when it predicts but also does not miss fire incidents for the sake of accurate prediction. This is why the **XGBoost** model was not chosen.

These are very good results given the quality of the data we had and the assumptions we made.

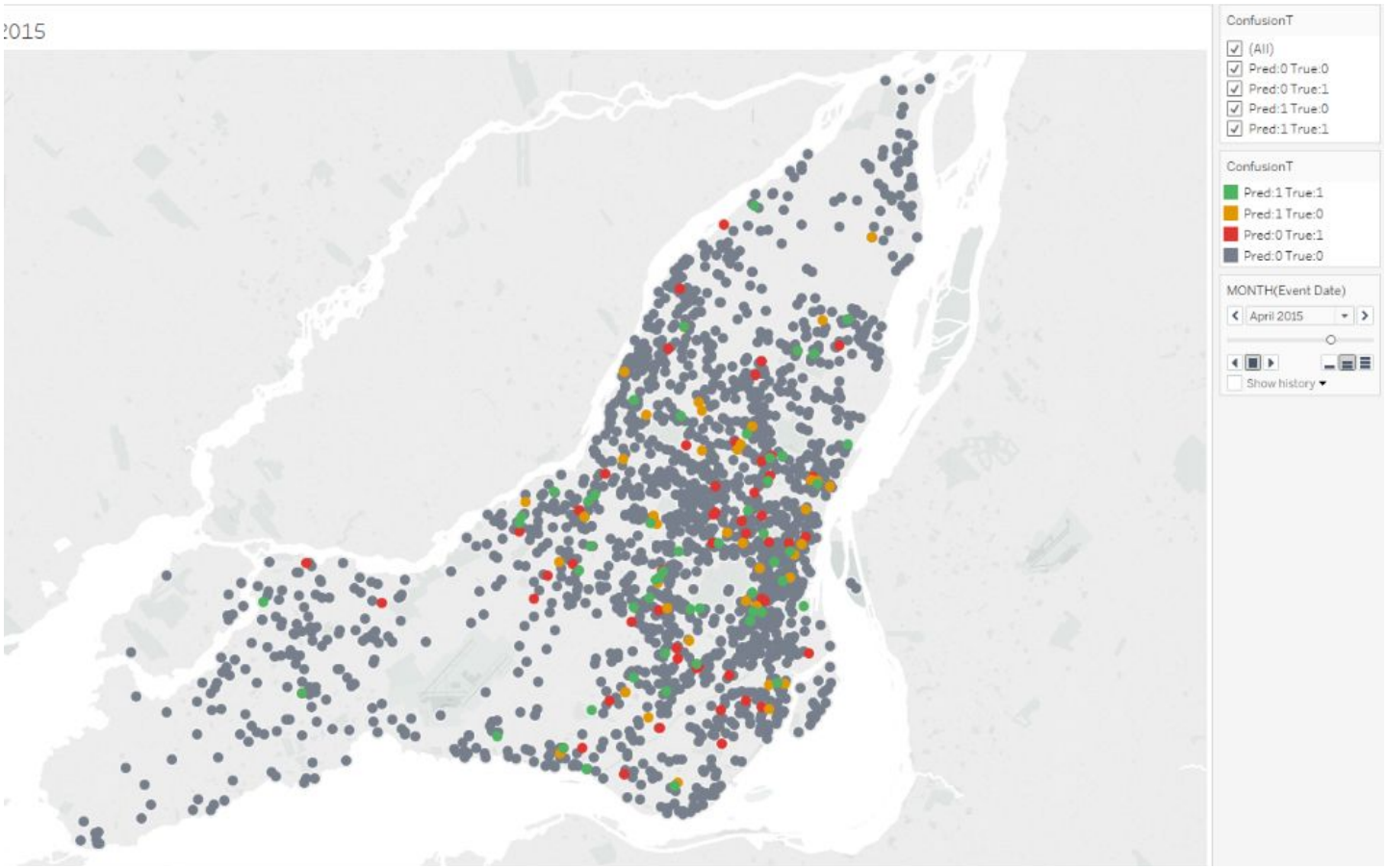
When we look at the importance of variable in our regression we see that **history related features** are playing a major role. Also as precision increases the importance of location specific data increases, and this can be seen from the results of the XG Boost model in appendix 1 (doc: CapstoneProject-FinalCode.pdf).

**Feature Importance**

| | Feature | Importance |
|---|---|---|
| 0 | Sum_Autres_incendies | 5.627817 |
| 1 | Sum_Incendie_de_batiments | 4.416537 |
| 2 | Sum_Premier_Repondant | 23.562459 |
| 3 | Sum_Sans_incendie | 14.230483 |
| 4 | Sum_Crime | 0.223628 |
| 5 | Sum_Alarmes_incendies | 12.848393 |
| 6 | Sum_False_Alertes_Annulations | 0.466472 |
| 7 | Couples_No_Children | 1.352183 |
| 8 | CouplesWithChildren | 1.376694 |
| 9 | SizeOF House | 1.478065 |
| 10 | Detached_House | 1.527576 |
| 11 | Appartment_FiveFloors | 1.497550 |
| 12 | OtherType | 1.347506 |
| 13 | Semi_detached | 1.426204 |
| 14 | TownHouse | 1.550613 |
| 15 | Duplex | 1.474393 |
| 16 | Appartment_less_5Florrs | 1.332883 |
| 17 | OtherDtached | 1.456652 |
| 18 | MObileHome | 1.499862 |
| 19 | 1_4Rooms | 1.328204 |
| 20 | 5_rooms | 1.365861 |
| 21 | 6_rooms | 1.373940 |
| 22 | 7_rooms | 1.387381 |
| 23 | 8_RoomsOrmore | 1.461079 |
| 24 | Average_Rooms | 1.424583 |
| 25 | Simple_Maintenance | 1.289666 |

```
26                  Major_repairs    1.351905
27           Average_House_Value     1.445947
28     AverageHouseholdIncome2015    1.485375
29                      Avg_flors    1.853714
30                  Avg_YearBuilt    1.669168
31                   Avg_LandArea    1.852248
32                   Avg_HomeArea    2.014959
```

Below is the map for the prediction of our model for the month of April 2015. Black (dark grey) dots and green dots are correct predictions. Red and orange are wrong ones as we can read in the Legend. Not bad!



# Conclusion

We were provided with data from the city of Montreal and we were able to join (using certain assumptions) with other external data to augment the data and improve the quality of the data we had. We used additional data from the Census Canada and from the city of Montreal.

We concluded that the history of fire department interventions for a specific location along with information regarding the dwelling properties like lot area, floors, etc. are very important for the prediction of fires in the short run.

Moreover it turned out that the history of Fire department interventions played a significant role in predicting future fire incidents in those locations. This makes sense if we think about it since if a location requires the fire department to intervene more times this might have to do with the number of people/families are living in that particular location and the house types and condition in that location.

Finally we believe that this work, along with the other deliverables produced, if leveraged with more and real data, would prove to be useful for the Fire Department as a tool to improve public safety, resources planning and allocation, and detect and combat fires in the city of Montreal.

# References


[01] How we predicted building fires in Baton Rouge, LA
January 26, 2017
How we predicted building fires in Baton Rouge, LA -- working version

[02] Predictive Modeling of Building Fire Risk
Designing and evaluating predictive models of fire risk to prioritize property fire inspections
Predictive Modeling of Building Fire Risk, Metro21: Smart Cities Initiative

[03] Interventions des pompiers de Montréal
The interventions data from the fire department of the city of Montreal

[04] Actes criminels
The data about Crime in Montreal

[05] Data from the 2016 Census for the Agglomeration de Montréal
Annuaire statistique de l'agglomération de Montréal - 2016

[06] Data regarding houses characteristics from the the city of Montreal
Unités d'évaluation foncière