Data Science & Machine Learning Capstone Project

House Risk Prediction

March 2019

TEAM

CARLOS FERREIRA
DIMITRIS KATSAOUNIS
HESHENG CHEN
LEIDY BARRERA
ANDRES URREGO

Contents

Introduction

Predicting likelihood of fire incidents for a specific location for a future time will give Montreal Fire Department (FD) time to prepare for the incidents. At the same time, we want to prevent the fire incident if possible. If we know fire risk level for each house for each region, FD can help to visit the higher risk house to have a precautious check.

In this project, we create a model to predict the house risk for any house in Montreal.

Model Limitation

In this model, we only handle residential house. Commercial properties are not in the scope, but the same method can be applied to it.

Idea behind the model prediction

Enriching fire incidents data with more features will lead use to fire incident prediction. If we can enrich the house data with fire incident information, we may get some features that are correlated to fire. But we can not simply combine the house information with fire incidents, we need to enrich the house data with some fire measurement or fire risk level. Once each house is labelled with fire risk level, we will be able to use ML to create a model for predicting the fire risk level and find out the feature importance for fire.

Assumption

- House Information:

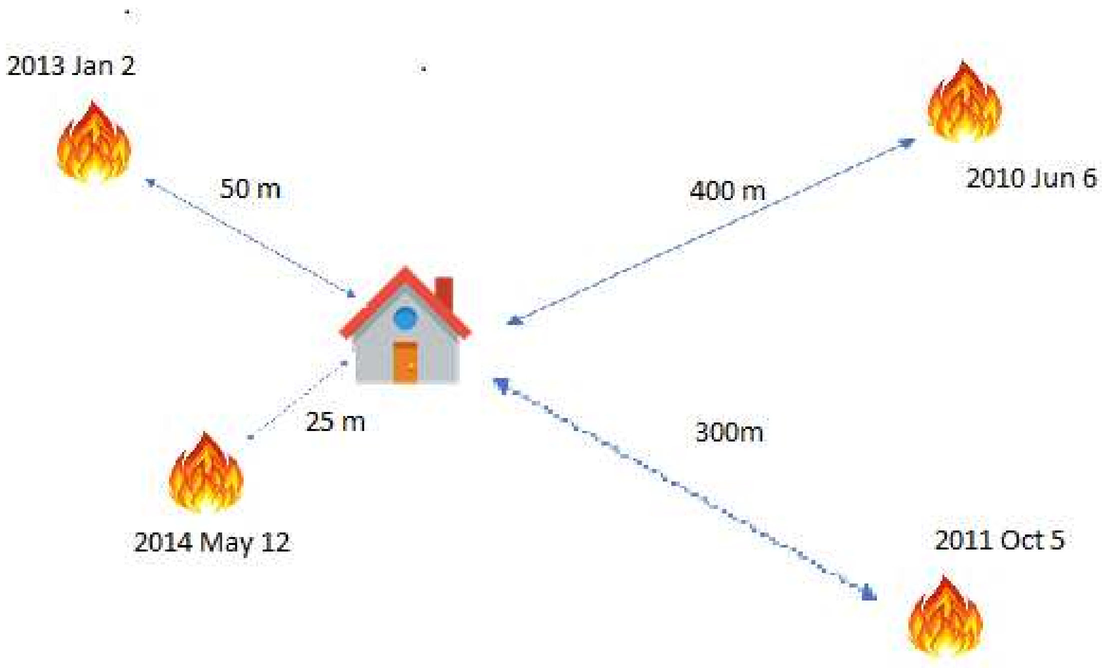
  Right now, we don't have all the Montreal house list, even we have, we may not have all the detail information or features. Instead of all the list, we can use a subset of the list, with the condition that this subset is well-distributed to each region in Montreal, also distributed to different house type, value and other information.

For this purpose, we are using Centris listings as Montreal house's subset, we believe the houses to be sell are well-distributed and satisfy above requirement. Also, the house information from Centris list contains longitude and latitude.

- Risk level

The provided fire incident data doesn't contain actual incident location, it is obfuscated to cross-street location. We can not locate which house had the actual incident. We need to have a reasonable risk level calculation.

House Risk Level Calculation



Risk level = 0.5 (for 25 m) + 0.25 (for 50 m) + 0.1 (for 300 m) + 0.05 (for 400 m) = 0.9

Basically, for any fire incident, we use the incident distance from the house to assign a value, and we add all the fire incident values together, the sum will be the house's risk level. The mapping from distance to value are arbitrarily assigned. This value scale will affect the actual risk level value, but the relative risk level should be kept. More detail discussion will be followed.

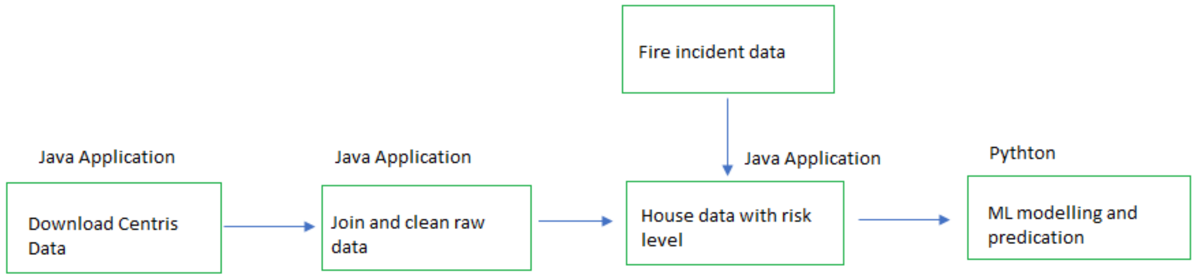
External Data

Besides FD incident data, we use the house information downloaded from Centris. The following is such sample information:

Title	Address	Rooms	Price	Latitude	Longititude	Type	Build year	Land size	Driveway	Basement
House for sale in Mercier/Hoc	4232 Avenue Mercier Mer	9 Rooms 2+1 Bedroor	349000	45.60188	-73.533093	Bungalow S	1953	2651 sqft	Driveway (2)	Basement 6 feet or +
Condo for sale in Mercier/Hoc	4270 Rue Adam Mercier/H	8 Rooms 0+2 Bedroor	420000	45.550884	-73.535767	Divided	1928	1237 sqft		Basement 6 feet or +
House for sale in Ville-Marie (I	41A Rue King Ville-Marie	8 Rooms 3+0 Bedroor	995000	45.497413	-73.552884	Two or mor	2003		Garage (1)	
Condo for sale in Le Sud-Oues	431 Rue Saint-Martin apt.	4 Rooms 1+0 Bedroor	301625	45.488294	-73.567984	Divided	To be built Ne	635 sqft		
Condo for sale in Ville-Marie (	2091 Rue Beaudry apt. 24(	6 Rooms 2+0 Bedroor	649000	45.522374	-73.563403	Divided	1923	1660 sqft	Driveway (1)	
Condo for sale in Lachine (Mo	4520 boulevard Saint-Jose	8 Rooms 2+0 Bedroor	375000	45.43679967	-73.7064554	Divided	1999	1006 sqft	Driveway (1) Ga	Located on a river
House for sale in Saint-Lauren	1500 Avenue Sainte-Croix	7 Rooms 3+0 Bedroor	559900	45.52037603	-73.6864371	Two or mor	1953	4275 sqft	Driveway (1) Garage (1)	
House for sale in LaSalle (Mon	8987 boulevard LaSalle La	8 Rooms 2+1 Bedroor	659000	45.416324	-73.633016	Two or mor	1976	3821 sqft	Driveway (2) Garage (1)	

Totally, we downloaded around 8,000 information that is currently listed for Montreal island.

Data Pipeline



1. Download Centris Data: create a java application to download all the summary list published on Centris for Montreal. For each house in the summary list, down the detail information.

2. Join and clean raw data: Extract the house information from the raw data. The extracted information contains: Title, Address, Rooms, Price, Latitude, Longitude, Type, Build year, Land Size, Driveway and Basement.

   From the Title information, extract the build type (House, Condo, Apartment, Townhouse) and City (or borough).

   Some houses are removed because of too many missing information.

3. Fire incident data: two incident data files provided in the class are joined together. Filtered out all the non-fire incident data rows.

4. House data with risk level:

   The following algorithm will be used for calculating the house risk level:

```
for each house in the house list:
      risk level ← 0
      for each incident in the incident data:
            distance ← calculate(house_lat, house_long, incident_lat, incident_long)
            risk level delta ← assign a value given distance
            risk level ← risk level + risk level delta
      house risk level ← risk level
```

   The following is the mapping from distance to assigned value.

Distance from fire (fire intersection)	Value assigned
50m	0.5
100m	0.2
200m	0.1
400m	0.05
800m	0.02
1000m	0.01

   Other distance thresholds and values are also tried and tested in this project.

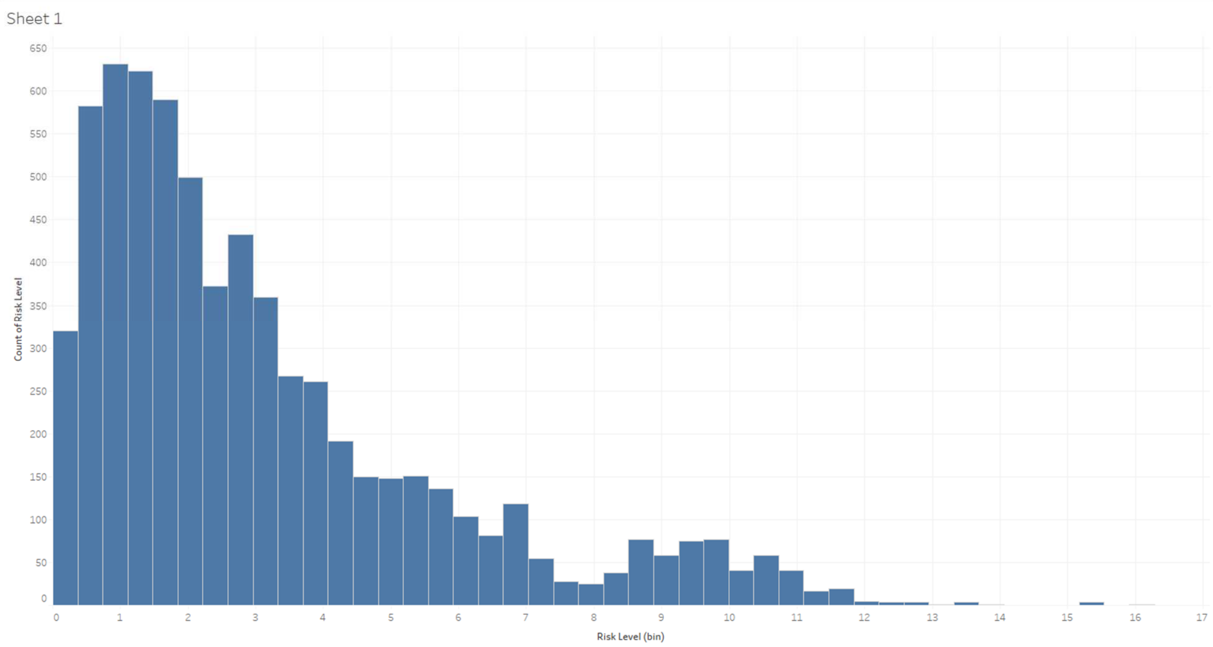
5. ML modelling and prediction:

   The house data with risk level are preprocessed: retrieved the number rooms, convert City, Building Type, Sub Type to numeric numbers so that it can be used by ML model.

   The house data are split to training and testing data sets. Decision tree is selected as the ML model. The features include Rooms, Price, Year, Lot Size, Driveway, Garage, htype (house type), subtype, cityIndex. The label is the calculated risk level.

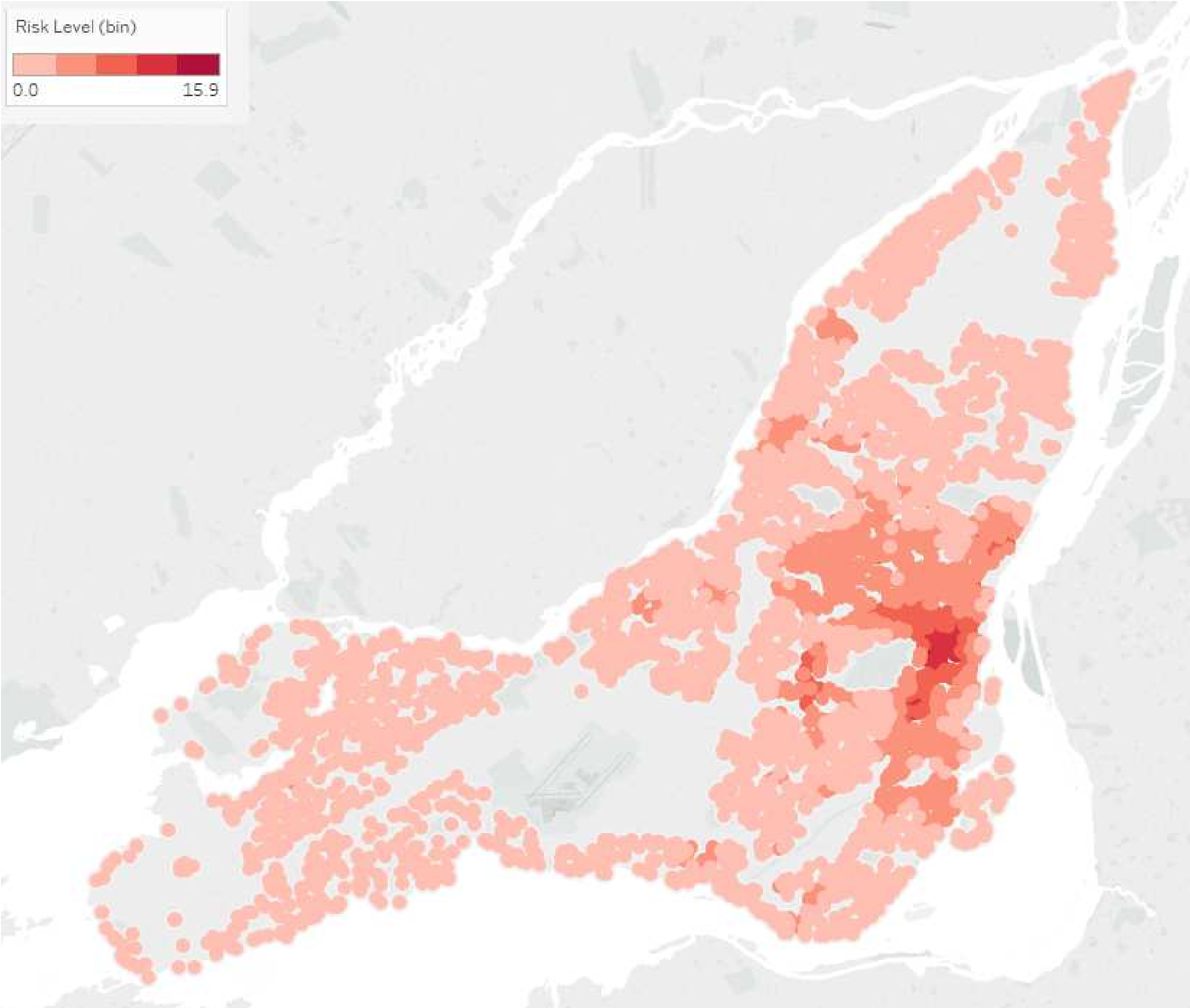
House Risk Level Visualization

The following is the histogram of House Risk vs. Number of houses in our data sets.



Interestingly, we can see there are two distributions around two different house risk level, one is around 1, the another one is around 9.5.

Also, the house risk levels are plotted on the map as shown in the following:



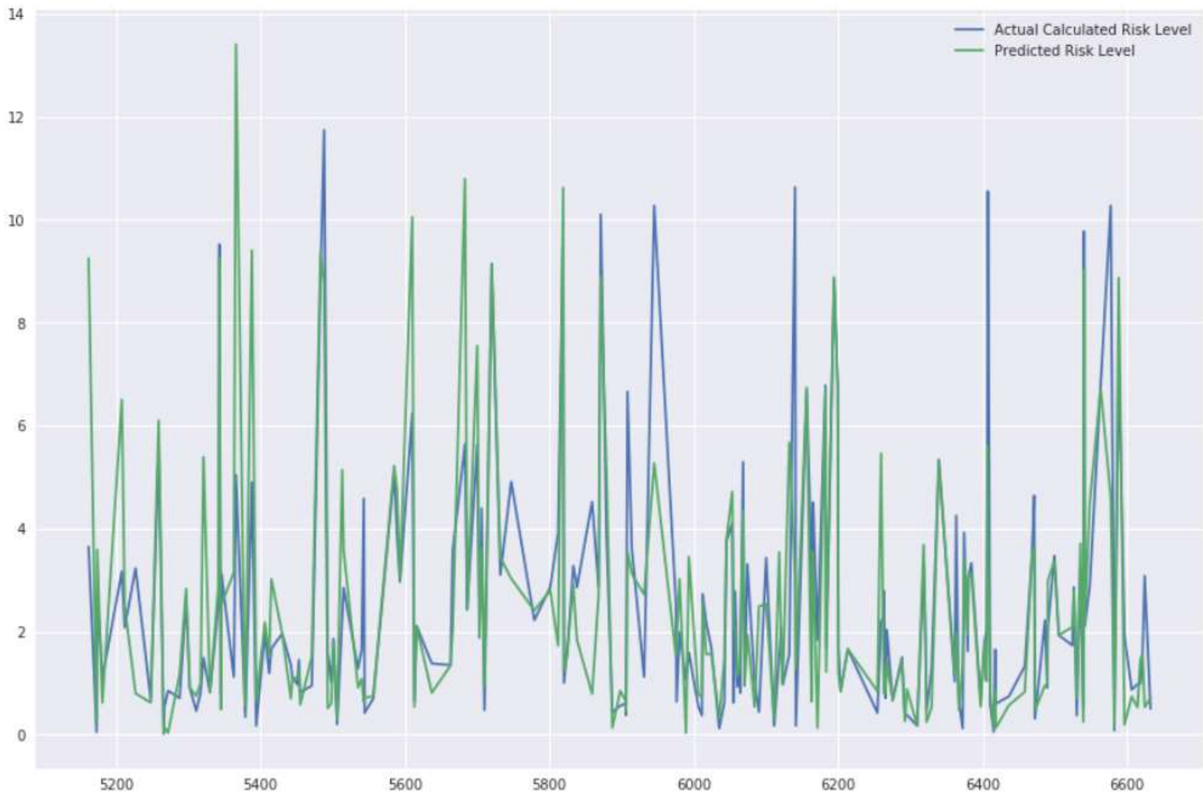
The houses around downtown areas have higher level risks.

Modelling Results

After training the model, the following is the comparison between the predicted risk level and actual calculated risk level:

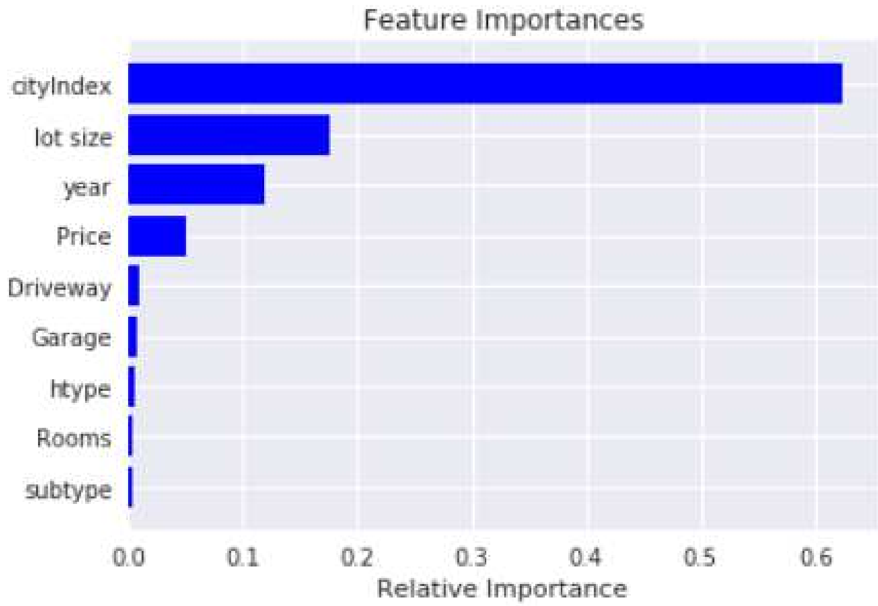
	Rooms	Price	year	lot size	Driveway	Garage	htype	cityIndex	subtype	Actual Calculated Risk Level	Predicted Risk Level
5388	1	499900	1983	1079.0	0	1	2	29	0	4.90	9.40
4820	1	319000	1994	935.0	0	1	2	28	0	0.99	1.08
82	1	1790000	1952	10200.0	4	1	4	15	1	1.59	1.59
1248	1	535800	1999	1262.0	0	1	7	23	3	2.80	1.87
726	1	388000	2005	575.0	0	1	2	29	0	8.36	5.48
3294	1	2150000	1993	2300.0	0	2	2	29	0	3.47	0.49
1112	1	649000	1999	2199.0	0	2	4	23	1	2.86	1.84
6530	1	528000	1991	6782.0	4	2	4	20	1	0.37	0.70
3966	2	1198000	2010	17432.0	2	2	4	10	1	0.20	0.04
670	2	1695000	1962	5857.0	2	2	4	8	1	1.44	1.44
5009	2	285900	1971	3610.0	3	0	4	17	1	5.51	1.55
1791	1	385000	1993	5750.0	1	1	4	22	1	0.86	1.08
5041	1	244900	2012	717.0	0	1	2	12	0	2.33	2.49
1323	1	2385000	1867	10867.0	5	2	4	5	1	1.82	1.59
2105	1	320000	1965	5054.0	2	1	4	22	1	1.27	0.89
567	1	359000	1870	806.0	0	0	2	29	0	6.66	6.66
2743	2	888000	2004	181.0	0	2	2	29	0	9.98	11.60
1280	2	619000	1942	5700.0	0	0	4	24	1	2.64	3.51
4678	2	340000	1988	4410.0	1	1	4	22	1	1.47	0.47
1719	1	358000	2013	678.0	0	1	2	29	0	6.09	11.25

The following is the graph comparison:



In general, the model predicted risk level does fit the calculated risk level. It can even predict quite some higher risk levels, which is useful for FD prevention purpose. The actual risk level differences between two are not that import. At the same time, it also produces some false alerts too.

We also calculated the importance of the features for the model. We can see that the four most important variables are the borough, lot size, year built, and the house price.

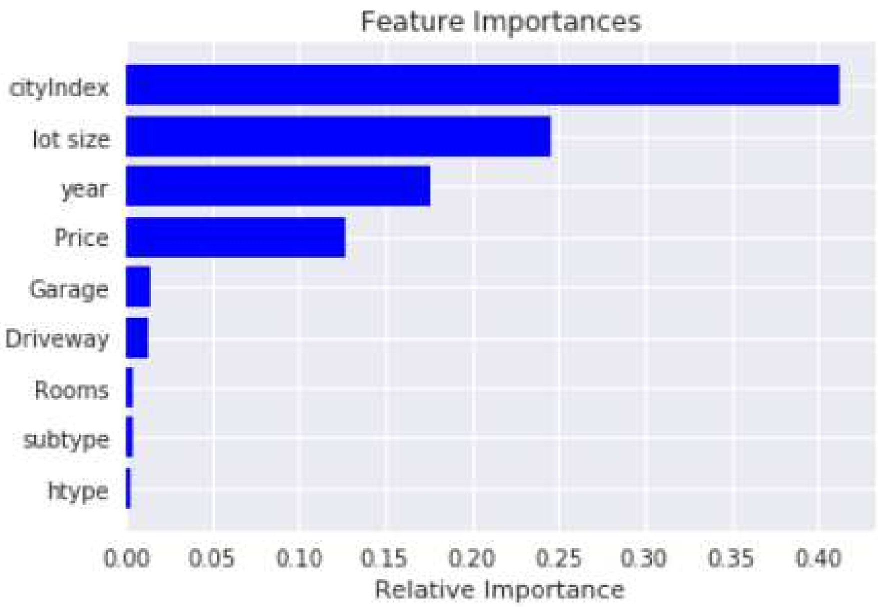


Another risk level calculation

The following distance to value mapping is also used to calculate another data risk level for each house.

Distance from fire (fire intersection)	Value assigned
50m	0.5
150m	0.4
200m	0.3
250m	0.2
300m	0.1
> 100m	0

Using this dataset, Decision Tree model generate the following feature importance:



The feature importance of two different risk level calculation are almost the same except the relative importance value changed.

Final Results

Based on the calculated risk level (0 to 14) we created a risk score to assess the risk of fire for each house in Montreal as below

Risk	% of all the houses	Risk Level
Low	75	0
Medium	20	3.5
High	5	9.2

Since we only care about the high-risk prediction, we output the confusion matrix from the model against test data for high risk:

	Predict Not High Risk	Predict High Risk
Actual Not High Risk	781	23
Actual High Risk	25	29

The model can predict a little more than half of the high-risk houses.

Conclusion

We believe that this model can predict, with a good level of precision, the risk level of fire for all houses in the city of Montreal if the model is fed with more data. The model will predict the fire risk level based on the houses' characteristics only, not considering any external additional data such as the weather data for example.

Future Works

- Divide borough into further smaller region for more accurate prediction.
- Add more feature data into the house information such as number of people living in the house, the ages of the residents…
- Further tuning the risk calculation using the actual incident location instead of using distance.