## Q.

Act as an experienced machine learning and data analyst with strong analytical skills in studying, collecting, and understanding data, as well as identifying the potential sources of datasets, to successfully implement a high-performance machine learning model. Provide an accurate, clear, and practical response to help me in my below request:

Request: I am working with my team on a machine learning project to analyze residential fire incident datasets in Montreal, Quebec, Canada over the past few years. Based on the data shared with our team, I aim to identify the most dangerous and fire-prone residential geographical locations, with the ultimate goal of implementing proactive preventative measures for these locations. The dependent variable is planned to be either (high risk, medium risk, medium-low risk, low risk, or very low risk). Our team is currently studying and analyzing other datasets that could be combined with our existing dataset, including datasets of fire incidents from previous years, which should be included, and other important datasets that should be included.

Please provide a list of up to five other datasets that could be used in addition to the existing (joined) datasets. Please be sure to clearly indicate in what way how each of the suggested datasets may help us in to our project by providing more relevant data to achieve our goal of high-performance model. Provide your answer in a table with columns 1) for the dataset, 2) explaining the reasons for including it, 3) what important/key features are expected in this dataset that may be used in the project, 4) and the importance and/or significance of including this dataset in our project to improve the performance of a high-performance machine learning model (accuracy, precision, etc.). Please ask me questions, one by one, for a maximum of three questions to get a better and more accurate answer.

## A.

To tailor my recommendations to your project and propose the most relevant datasets, I need a bit more context about the data you already have.  
Here’s my first question:

**Q1.** Besides the basic fire incident records (time, date, location, cause, etc.), what additional information is already available in your current dataset?  
For example: building type/age, occupancy data, weather conditions at the time, fire department response time, or socioeconomic data.

## Q.

dataset is Interventions by Montreal firefighters (SIM interventions) available in <https://app.geo.ca/en-ca/map-browser/record/2fc8a2b9-1556-410e-a118-c46e97e9f19e>

## A.

Great — thanks for sharing that. Knowing you already have access to the **SIM interventions / fire-department dispatch / intervention** dataset is helpful. That gives you a strong basis of incident-level spatial & temporal data (dates, location, unit dispatched, incident types) in Montreal. ([Geo App](https://app.geo.ca/en-ca/map-browser/record/2fc8a2b9-1556-410e-a118-c46e97e9f19e))

To help you build a more predictive, high-performance risk model (classifying areas into risk tiers), here are up to **five** complementary datasets that, when joined or integrated (spatially, temporally, or both), can enrich feature space and help your model better discriminate high-risk from low-risk zones. After the table, I’ll comment on modeling and integration challenges. Also, I’ll ask one more question to refine further.

| **Candidate Dataset** | **Why Include It / Usefulness** | **Key Features / Variables Expected** | **How It Improves the ML Model (Significance)** |
| --- | --- | --- | --- |
| **Building inventory / Cadastre / Property attributes** | Fire risk depends strongly on building type, materials, age, size, occupancy, layout, number of dwelling units, height, etc. Without those, your model is blind to structural vulnerability. | For each building or parcel: year built, number of floors, building footprint area, building use (residential / mixed / commercial), number of units, construction materials, renovation date, building age, occupancy type (single family, multi-family), value or assessed value. | Adds crucial “vulnerability” features. Helps discriminate where fires are more likely or more severe. Should substantially boost precision / recall on “high risk” classes by giving the model more explanatory power beyond just past incident counts. |
| **Demographic & socioeconomic data at fine spatial granularity** | Socioeconomic conditions correlate with fire risk (e.g. lower income, higher density, older housing, maintenance lag). Incorporating these helps capture latent risk factors not visible in fire logs. | Census tract / dissemination area level: median income, proportion low income, education levels, population density, age of population, percent renters vs. owners, household size, vacancy rates. | Helps the model generalize beyond just historically burnt areas. Improves both sensitivity (catch areas that haven’t yet burned but are similar in socioeconomic risk) and specificity (avoid overpredicting risk in wealthy, low-density areas). |
| **Historical weather / climate / environmental data** | Weather and climate influence ignition, fire spread, and even secondary risks (e.g. power lines under high wind). Including weather variables (especially temporally aligned) helps adjust for “chance / conditions.” | Temperature, humidity, wind speed/direction, precipitation (recent days, cumulative), drought indices (soil dryness, relative humidity history), seasonal indicators. | Allows the model to account for temporal variation in fire likelihood. Improves time-aware predictions (e.g. certain months or dry spells). Reduces false positives/negatives driven by exceptional weather conditions. |
| **Fire safety infrastructure / response accessibility data** | The effectiveness of fire suppression depends on how close and well-equipped fire stations are, drive times, hydrant network, etc.—this impacts whether a fire escalates or is prevented early. | Locations of fire stations, dispatch zones, travel times / response times to each building / block, road network connectivity, fire hydrant network (hydrant density, distance to nearest hydrant), water pressure / pipe network if available. | Captures “mitigating capacity.” Areas far from fire stations, with sparse hydrants or longer response times, are more vulnerable. Including this helps the model penalize risk in such underserved zones, improving recall of at-risk spots. |
| **Land / vegetation / fuel load / urban forest / proximity-to-wildland data** | Even in urban areas, proximity to vegetation, brush, or green space can increase fire risk (ignition, spread). Also, vacant lots or abandoned buildings with overgrowth may act as fuel. | For each spatial unit: land cover (trees, shrubs, impervious surfaces), percentage green cover, distance to large parks or vegetation patches, fuel load index, recent vegetation dryness, land-use zoning (vacant vs occupied). | Helps the model capture ignition and spread risk beyond structural fires. This is especially useful for borderline or transition zones. It may help reduce false negatives in zones that historically had fewer fires but are at risk due to adjacent vegetation or fuel exposure. |

**Additional Remarks & Modeling Considerations**

* **Spatial and temporal alignment**: You’ll need to align all these datasets to a consistent spatial grid (e.g. small-area polygons, census dissemination areas, or even a hex grid) and temporal window (e.g. by year or seasonal windows).
* **Feature engineering**: For example, compute rolling averages (e.g. incident counts in past 1, 3, 5 years), indicator variables (e.g. whether building is older than 50 years), distance metrics (distance to nearest fire station, shortest path travel time), spatial lag features (neighboring area incident counts), interactions (e.g. building age × socioeconomic index).
* **Addressing imbalance**: “High risk” zones are likely a small fraction. You may need sampling strategies, class weighting, or anomaly detection techniques.
* **Model validation / spatial cross-validation**: Because of spatial autocorrelation, use spatial holdout (i.e., leave out spatial neighborhoods) or block cross-validation, not random splits.
* **Interpretability and causality caution**: Even if the model finds strong associations, policy or interventions should still be informed by domain expertise (fire safety engineers, urban planners).

Before I propose sources or access points for these datasets (for Montreal / Quebec) and refine further suggestions, here’s a second question:

**Q2.** What is the spatial granularity you intend to classify (e.g. building-level, street-block, census tract / dissemination area, grid cell), and how many (approximate) spatial units do you expect?

Once I know that, I can tailor which data sources are feasible and suggest access points.