Final Report for Capstone Project- Predicting Neighborhood Fire Risk in Montreal

Sam Wanis, Hammed Akande, Pak Lo, Ekue Afanou

October 2025

Contents

1 Problem Statement **1**

2 Data Sources **2**
2.1 Neighborhood characteristics (Statistics Canada Census) . . . 2
2.2 Fire incidents (2005–2025) . . . 2
2.3 Montreal Building Assessment dataset (municipal property assessment) . . . 2
2.4 Integrated modeling dataset . . . 3

3 Data Exploration and Cleaning **3**
3.1 Fire . . . 3
3.2 Census . . . 4
3.3 Buildings . . . 4

4 Feature Engineering **4**
4.1 Features or explanatory variables . . . 4
4.2 Target Variable . . . 5

5 Tools and Techniques Used **5**

6 Summary of Modeling Techniques Evaluated **6**

7 Modeling Results **6**
7.1 Performance Summary . . . 6
7.2 What Drives Fire Risk Predictions . . . 6
7.3 October 2025 Predictions and Resource Allocation Strategy . . . 6
7.4 Operational Impact and Recommendations . . . 7

8 Insights and Challenges **8**

9 Conclusions **9**
9.1 Key Findings . . . 9
9.2 Limitations . . . 10
9.3 Future Directions . . . 10

A Appendix **i**
A.1 Data Dictionary / Variable Definitions i
A.2 Modeling Details i
A.3 October 2025 Predictions and Resource Allocation Strategy i
A.4 Rationale for Predictive Fire-Risk Modeling iv

1 Problem Statement

Structural fires represent a significant public health and public safety risk in large urban areas. Globally, fire-related incidents are estimated to cause approximately 401,000 deaths per year, a figure exacerbated by rapid urbanization, high population density, and aging building infrastructure (Zhang, 2023). In Canada, fire departments responded to more than 39,000 fire incidents in 2021; 42% of these were structural fires, the majority of which occurred in residential settings (Statistics Canada, 2023). In the United States, fire departments responded to an estimated 1.39 million fires in 2023, resulting in 3,670 civilian deaths and approximately $23 billion in direct property damage. These figures indicate that fire remains both a life-safety issue and a significant economic burden at the municipal level.

Montreal is subject to similar pressures. High-density residential areas, mixed-use commercial corridors, and the presence of industrial zones produce a spatially uneven fire risk between neighborhoods. Despite this heterogeneity, current fire prevention practices in Montreal are primarily reactive. Fire services respond to emergencies as they occur and conduct routine inspections, but there is no systematic mechanism to anticipate where an increased fire risk is likely to arise next. The Montreal Fire Safety Service currently records detailed operational data for every intervention, including date, location, and call type. However, while these historical records are collected and consulted, they are not yet being leveraged in an optimal, forward-looking way to produce systematic risk forecasts. This limitation leads to three specific operational constraints:

1. **Resource limitations**. Inspection capacity is finite. Inspectors cannot visit all high-concern buildings at a sufficient frequency, and frontline personnel must often prioritize urgent response over proactive prevention.

2. **Uneven spatial risk**. Neighborhoods differ markedly in building age, structural density, land use, and socioeconomic characteristics. Fire risk is therefore not uniform across the city; some areas may be systematically more vulnerable than others.

3. **Lack of predictive targeting**. Without a predictive framework, preventive measures such as safety education, smoke alarm distribution, or code enforcement are not necessarily delivered in the areas where they would have the greatest prospective impact.

The objective of this project is to address that gap. We propose a data-driven methodology to generate monthly, neighborhood-level forecasts of structural fire risk. More specifically, for each neighborhood in Montreal, we estimate the probability of experiencing structural fire incidents in the upcoming month. Each neighborhood-month is then assigned to a discrete risk category. A monthly forecast horizon is intentional: it avoids the unrealistic task of predicting the exact timing and count of individual fires, while aligning with operational planning cycles for inspection routing and community outreach. This predictive framing enables three direct applications for the fire service:

- **Inspection planning**. Direct limited inspection resources toward neighborhoods with elevated predicted risk.

- **Targeted outreach**. Focus prevention and public education (e.g. smoke alarm checks, door-to-door safety visits) in areas where population exposure and structural vulnerability are high.

- **Operational posture**. Support proactive allocation or pre-deployment of resources in neighborhoods expected to experience higher incident volume.

In summary, this work addresses the need to improve existing fire prevention practices in Montreal by introducing an evidence-based, predictive risk framework for structural fires. The contribution of this project is a supervised, neighborhood-level, monthly model that estimates and classifies near-term fire risk, in order to support prevention, inspection strategy, and the equitable allocation of municipal safety resources.

2 Data Sources

To produce neighborhood-level monthly fire risk forecasts, we construct a supervised learning dataset that integrates operational fire records, neighborhood characteristics from census data, and building assessment data. The principal data sources are summarized below.

2.1 Neighborhood characteristics (Statistics Canada Census)

Neighborhood geography is defined using small-area boundaries derived from Statistics Canada census tracts. These official statistical units are used to delimit consistent neighborhood areas across the city of Montreal. The census tract primarily serves as the spatial reference frame for the rest of the analysis. In practice, this means that:

- We use the census-based neighborhood areas as the base unit of analysis;
- We assign (spatially join) all other data sources — historical fire incidents, building assessment records, and monthly aggregations — to those same areas;
- We then construct a panel where each row corresponds to a (neighborhood, month) pair using this common spatial definition.

This approach ensures that fire activity data and building assessment data are expressed on the same spatial footprint. In other words, the census geography provides the stable neighborhood boundaries into which all other datasets are integrated, rather than serving as a direct source of explanatory socioeconomic features in the predictive model.

2.2 Fire incidents (2005–2025)

The primary outcome variable is fire activity. We use nearly two decades of emergency call records handled by Montreal firefighters. These operational records from the city's open data infrastructure identify when and where each fire-related intervention occurred. All incidents are:

- Spatially assigned to an official Montreal neighborhood (based on census tract), and
- Related to a monthly time resolution.

This helps to calculate for each neighborhood $\times$ month, the observed fire incident count. These neighborhood-month counts serve both as:

- The target to be predicted in future months, and
- Lagged historical predictors (autoregressive terms) that characterize recent local fire activity.

2.3 Montreal Building Assessment dataset (municipal property assessment)

This dataset describes the physical and functional characteristics of buildings and lots, including:

- Building type and use (e.g. single-family residential, condominium / multi-unit residential),
- Number of units / dwelling density,
- Construction year or effective age of the structure.

These indicators are relevant because aging electrical systems, informal subdivisions, and high residential density are repeatedly associated with elevated structural fire risk in North American cities. By aggregating parcel-level assessment data to the neighborhood level, we obtain structural vulnerability profiles for each neighborhood (for example, concentration of older housing stock, prevalence of multi-unit buildings).

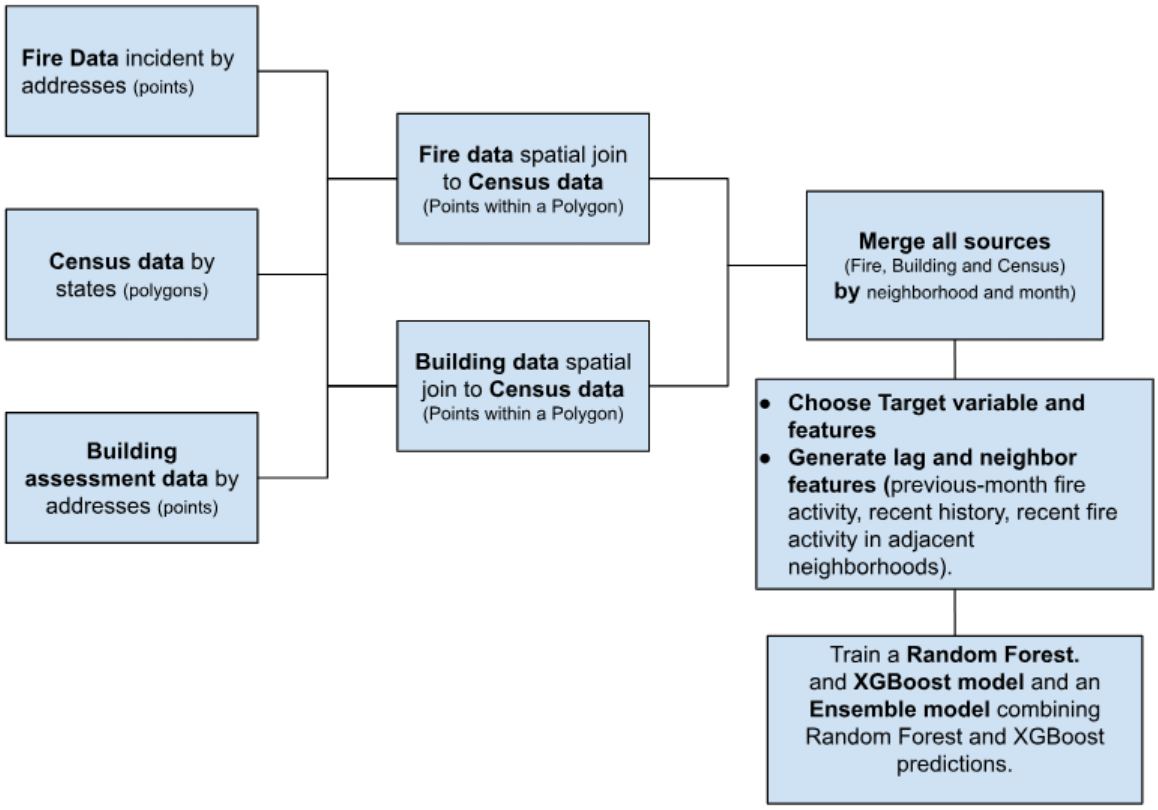


Figure 1: Pipeline of data integration and modeling

2.4 Integrated modeling dataset

As illustrated in Figure 1, all sources are harmonized to:

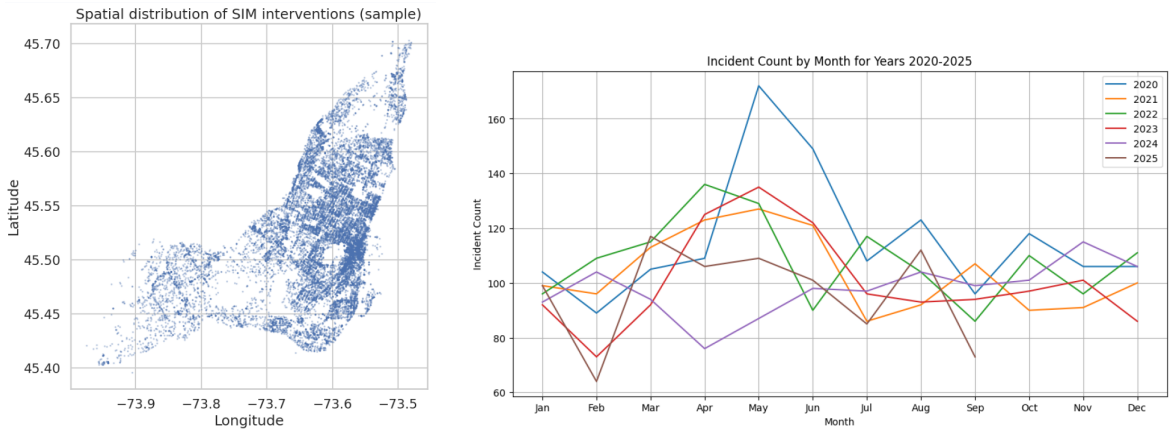
- A common spatial unit (the official census area neighborhood), and
- A common temporal unit (the calendar month).

This integrated data forms the input to supervised classification models that estimate, for each neighborhood, the probability of elevated fire activity in the following month. The model output is expressed as a binary risk label (high risk = 1, otherwise = 0). This output is intended to inform proactive inspection scheduling, targeted prevention campaigns, and equitable deployment of municipal fire safety resources.

3 Data Exploration and Cleaning

3.1 Fire

The original fire datasets covered four periods (2005–2014, 2015–2022, 2020–2023, and 2024–2025). Since 2015–2022 and 2020–2023 overlap, duplicates were removed during concatenation. The "CREATION DATE TIME" column had varying formats (e.g., YYYY-MM-DD HH:MM:SS vs. YYYY/MM/DD), so dates were standardized to YYYY/MM/DD, keeping only the date. A new "YearMonth" column (e.g., 2018-02 from 2018/02/27) was created to aggregate interventions monthly. Records without coordinates were excluded. Longitude and latitude were used to generate point geometries to link interventions to census tracts. Finally, the coordinate reference system was converted to NAD83 to ensure accurate representation of North American locations.



(a) Spatial Distribution of Interventions

(b) Time Plot of Interventions by Month

Figure 2: Spatial distribution of SIM Interventions and Time Plot

Figure 2a shows that interventions are densely clustered in Ville-Marie, suggesting that adjacent census tracts' fire risks should be considered when estimating a tract's risk. Figure 2b indicates monthly variation in intervention counts, highlighting the need to examine how month influences fire risk.

3.2 Census

We downloaded the census or demographics data across all of Canada, and the census tract data across all of Canada. Using Alteryx, we extracted those census tracts located in the Montreal Metropolitan Area. We then spatially joined these two datasets together so we have the census tracts within the Montreal Metropolitan Area, which includes not only the Island of Montreal, but also Laval and the South Shore. We therefore overlapped the joined dataset with the firefighters' stations cover map to filter those census tracts on the Island. The original census dataset uses the World Geodetic System 1984 (WGS 84) datum as the CRS. We have converted it to NAD83.

3.3 Buildings

We did cleaning for the building dataset. For instance, we found that some of the buildings have 9999 as the construction year. They are removed from the dataset before further processing. The original CRS is WGS84, which is converted to NAD83 for spatially joining datasets in the later stage. The spatial distribution of buildings is visualized in Figure 3.

4 Feature Engineering

4.1 Features or explanatory variables

Before modeling, interventions within each census tract and month are aggregated to obtain the total interventions per tract per month. Building data are also aggregated to calculate the number of buildings, average floors, and average age per tract. All datasets are then merged using the unique census tract ID (CTUID), after which additional features are created based on the aggregated data.

- Lagged variables: We included interventions from twelve months ago to capture the annual cycle and from one month ago to account for delayed fire mitigation, as fire risk may remain high shortly after incidents.

Figure 3: Spatial Distribution of Buildings

- Seasonality was accounted for by converting the month variable into sine and cosine functions, making months like January and December, though eleven apart, temporally close. Specifically, we consider

$$\sin\left(\frac{2\pi t}{12}\right), \cos\left(\frac{2\pi t}{12}\right),$$

  where t takes value on $1, 2, \ldots, 12$, corresponding to each month of the year.

- We added the previous month's fire counts from neighboring census tracts, as a tract's fire risk increases when nearby tracts have high fire risk.

4.2 Target Variable

We frame this as a binary classification: high fire risk (Class 1: > 2 incidents) vs. low fire risk (Class 0: ≤ 2 incidents), resulting in a balanced dataset with 44.04% Class 1 and 55.96% Class 0 after spatial and temporal aggregation. A threshold of 2 was selected because it corresponds to the median of the data, providing a balanced division between lower and higher values.

5 Tools and Techniques Used

The tools and techniques employed to build the neighborhood-level monthly fire risk prediction system for Montreal are :

- **Alteryx** handled heavy ETL and aggregation.
- **Python** (**Colab** for exploration, **VS Code** for production) handled spatial joins, feature engineering.
- Rolling Origin Evaluation, Spatial Lag Computation, Ensemble Learning, and Model Training.
- **Visualization** provided interpretability and validation at every step.

6 Summary of Modeling Techniques Evaluated

We developed and evaluated three machine learning models (Random Forest, XGBoost, and an Ensemble approach) to predict fire risk across Montreal's 541 census tracts using historical data from September 2005 to September 2025. The Ensemble model combines predictions from both Random Forest and XGBoost using majority voting—requiring both models to agree before flagging a census tract as fire risk. This consensus approach minimizes false alarms while maintaining strong detection capability.

Our models utilize 11 predictive features across four categories: building characteristics (density, height, age), temporal patterns (fire history from 1 and 12 months prior), spatial patterns (neighboring area fire incidents), and seasonality (monthly and annual trends). Analysis reveals that past fire history is the strongest predictor, accounting for 60% of predictive power, followed by building characteristics (21%) and spatial patterns (14%) (See Figure 6 for the graphic presentation and Section 7.2 for breakdown of these numbers at the feature level).

Performance was carefully evaluated using Rolling Origin Evaluation across 37 consecutive months (September 2022 to September 2025), where models were retrained monthly on all historical data and tested on the upcoming month—simulating real-world deployment conditions.

7 Modeling Results

7.1 Performance Summary

Table 1 presents model performance across the 37-month evaluation period. The Ensemble model achieved the best overall performance with F-1 score of 0.69, successfully identifying 7 in 10 census tracts that will experience fire incidents while maintaining balanced precision to minimize false alarms.

Table 1: Model Performance Summary (Rolling Origin Evaluation: Sep 2022 - Sep 2025)

Model	Balanced Acc	Precision	Recall	F1 Score
Random Forest	0.71	0.69	0.69	0.69
XGBoost	0.70	0.73	0.59	0.65
Ensemble	**0.71**	**0.69**	**0.69**	**0.69**

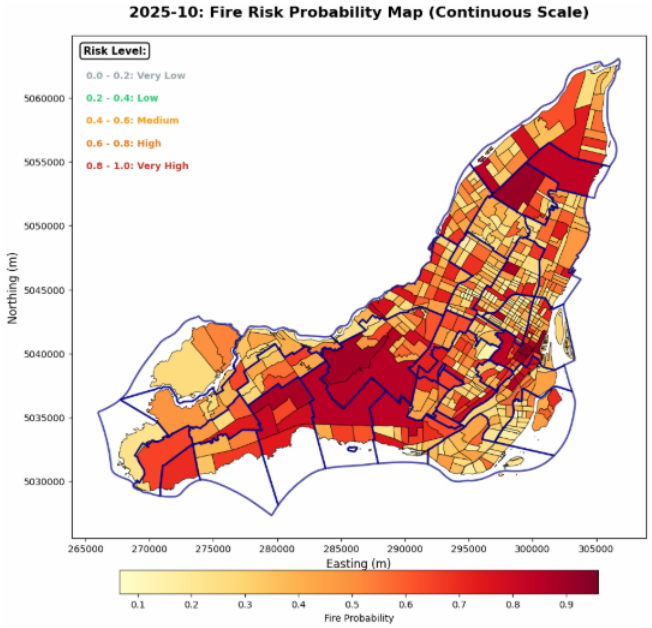
The model demonstrates strong temporal stability, maintaining consistent F1 scores (0.55-0.77) across the 37-month evaluation period. This reliability ensures consistent prediction quality across different seasons and years, enabling effective operational planning.

7.2 What Drives Fire Risk Predictions

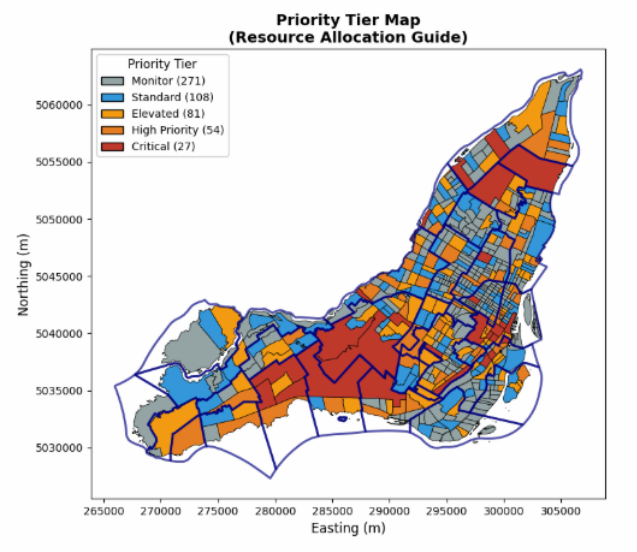
Figure 6 shows the relative importance of factors driving fire risk predictions. The top five predictors are: previous month's fire incidents (35%), fire incidents from 12 months prior (25%), building density (11%), neighboring area fire activity (10%), and building height (5%). This distribution confirms that recent fire history is the dominant indicator of future risk. Census tracts with fire incidents in the past 1-12 months are significantly more likely to experience future fires, supporting prevention strategies that prioritize these high-recurrence areas. The presence of spatial patterns (neighboring area fires contributing 10%) indicates that fire risk clusters geographically, informing neighborhood-level intervention zone design.

7.3 October 2025 Predictions and Resource Allocation Strategy

The ensemble model generated fire risk predictions for all 541 Montreal census tracts for October 2025. Figure 4a visualizes the geographic distribution of predicted fire risk, with 168 census tracts (31%) classified as fire



(a) Predicted fire risk across Montreal (October 2025). Darker red indicates higher risk, highlighting spatial clusters for targeted intervention.



(b) Five-tier fire risk map for October 2025, prioritizing resource deployment; the top 5% (critical tier) marks areas needing immediate intervention.

Figure 4

risk and 373 tracts (69%) as no immediate risk.

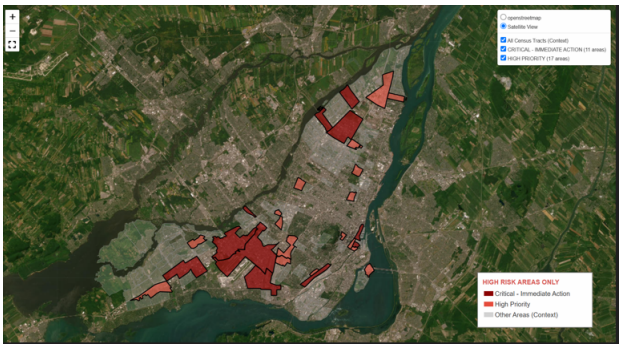
To support operational resource allocation, we stratified all census tracts into five priority tiers based on predicted risk probability (Figure 4b). Figure 5 shows the High-Risk Fire Zones Across Montreal (Figure 5a) and the Top 10 Most Critical Fire-Risk Areas (Figure 5b).

- **Critical** (27 tracts, 5%): Immediate building inspections, enhanced fire prevention education, and proactive safety interventions
- **High Priority** (54 tracts, 10%): Enhanced monitoring programs and community safety initiatives
- **Elevated** (81 tracts, 15%): Proactive community education and routine inspections
- **Standard** (108 tracts, 20%): Routine monitoring protocols
- **Monitor** (271 tracts, 50%): Standard background surveillance

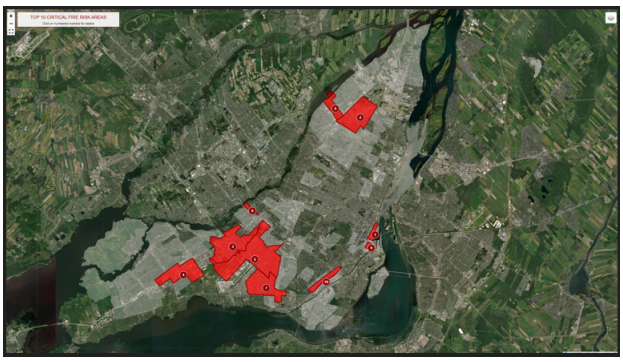
7.4 Operational Impact and Recommendations

Detection Capability: The model successfully identifies 69% of at-risk census tracts, enabling proactive fire prevention rather than purely reactive emergency response. This detection rate translates to approximately 7 in 10 areas that will experience fire incidents being flagged in advance for preventive intervention.

Resource Efficiency: By concentrating fire prevention resources on the top 15% risk tier (Critical + High Priority = 81 census tracts), the fire department can achieve focused, high-impact intervention while avoiding inefficient dispersion of resources across all 541 census tracts. This represents an 85% reduction in geographic coverage area while maintaining high detection rates.



(a) High-Risk Fire Zones Across Montreal.



(b) Top 10 Most Critical Fire-Risk Areas.

Figure 5

Prediction Reliability: The ensemble approach requires both independent models to agree on fire risk classification, providing high-confidence predictions that reduce false alarms. The consistent performance over time ensures reliable operational planning without frequent model recalibration.

Actionable Strategy: Feature importance analysis validates a two-way prevention approach: (1) prioritize census tracts with recent fire history (past 1-12 months) for immediate intervention, and (2) design neighborhood-level prevention zones based on geographic clustering patterns. This data-driven strategy optimizes resource allocation while addressing the root drivers of fire risk recurrence.

8 Insights and Challenges

This section outlines how the team identified, analyzed, and overcame key technical and domain challenges while developing the ML model to predict high fire-risk areas in Montreal.

- **Challenge**: The team lacked prior experience with fire incident datasets, making it difficult to select appropriate machine learning models and techniques initially.
  - **Insights**: We reviewed existing global research on similar fire risk prediction problems. This helped identify effective ML models, key techniques, and common pitfalls to avoid. As a result, we established a solid foundation for the project, guiding our approach from the outset.
- **Challenge**: The team lacked prior experience in spatial-temporal modeling and handling complex geospatial data formats.
  - **Insights**: We conducted a focused learning effort to master CRS and utilized GeoPandas for effective spatial feature engineering. This enabled a smooth transition from tabular to spatial-temporal analysis, ensuring accurate predictive modeling.
- **Challenge**: Optimizing model performance was difficult because the initial code used nested loops that processed data element by element, leading to slow execution.
  - **Insights**: The team improved efficiency by replacing loops with vectorized operations using optimized library functions. This change led to a reduced computation time, streamlined processing, and enhanced overall model performance.
- **Challenge**: Data Leakage in Model Training: From our initial training results, there was an unexpectedly high model performance suggesting data leakage or an imbalanced target variables.

  - **Insights**: This early spike in model metrics due to data leakage highlighted how essential continuous validation, and proper cross-validation splits are, in maintaining trust in performance results. The team reviewed the training process, identified and fixed the leakage source, and implemented balanced sampling to ensure reliable evaluation results.

- **Challenge:** Determining how to effectively apply ensemble learning to improve prediction accuracy was initially unclear. The team needed to understand the trade-offs between model complexity and performance gains.
  - **Insights:** We combined Random Forest and XGBoost models and compared their outputs through rigorous testing. This process showed that considering ensemble learning offered a slight improvement in accuracy.

Lessons Learned:

- **Data-Driven Fire Prevention Strategies:** The model enables the identification of high-risk zones, allowing city officials to proactively plan inspections and allocate resources more efficiently.
- The interdisciplinary nature of the project—combining data science, geospatial analysis, and municipal data—helped the team gain practical experience in integrating diverse data sources and applying analytical methods to a real-world urban safety problem.

Recommendations for Planning Inspections and Maintaining the Models

1. **Dynamic Inspection Prioritization:** Use model outputs to schedule inspections more frequently in high-risk areas, updating the priority zones monthly/quarterly based on new data and model retraining results.
2. **Regular Model Maintenance Cycle:** Establish a maintenance schedule that includes periodic retraining, feature updates, and performance monitoring (e.g., tracking AUC, recall, and precision drift) to ensure continued accuracy over time.

9 Conclusions

We developed a machine learning framework for predicting fire risk in Montreal census tracts, achieving an F1 score of 0.68 through an ensemble of Random Forest and XGBoost models. The analysis reveals important findings about urban fire risk and demonstrates the potential for data-driven fire prevention resource allocation.

9.1 Key Findings

Fire Risk is Highly Recurrent. Past fire history dominates prediction (60% of model power), with census tracts experiencing incidents in the previous 1-12 months facing significantly elevated future risk. Fire events follow predictable recurrence patterns, validating targeted prevention strategies focused on recent incident locations.

Geographic Clustering. Fire risk exhibits strong spatial patterns, with neighboring areas influencing each other's risk levels (10% predictive contribution). This clustering supports neighborhood-level prevention zone design rather than isolated interventions.

Seasonal Variation. Model performance varies across seasons (F1: 0.55-0.77), with higher accuracy during summer months, suggesting fire risk patterns are more predictable in certain seasons.

Ensemble Confidence. Requiring agreement between independent models provides operational confidence, enabling resource-intensive interventions for high-consensus predictions and lighter approaches for lower-confidence areas.

9.2 Limitations

The 31% false negative rate means the model misses approximately one-third of at-risk areas, requiring fire departments to maintain baseline prevention programs across all areas. The current feature set excludes socioeconomic factors, weather conditions, and building code compliance data that may influence risk. The model does not account for prevention intervention impacts, potentially overestimating risk in successfully mitigated areas. Strong associations between past fires and future risk do not imply direct causation; prevention strategies should address underlying root causes. Operational deployment requires balancing data-driven predictions with field expertise, community relationships, and equity considerations.

9.3 Future Directions

We recommend: (1) incorporating socioeconomic and weather data to capture additional risk dimensions, (2) implementing feedback mechanisms to track intervention effectiveness, (3) developing interpretation tools for fire prevention teams, and (4) establishing pilot programs to validate predictions before full-scale deployment.

References

Crowley, C., Miller, A., Richardson, R., and Malcom, J. (2023). Increasing damages from wildfires warrant investment in wildland fire management. U.S. Department of the Interior, Office of Policy Analysis. Retrieved from https://www.doi.gov/sites/doi.gov/files/ppa-report-wildland-fire-econ-review-2023-05-25.pdf.

Hall, S. (2024). Fire loss in the united states. National Fire Protection Association Research. Retrieved from https://www.nfpa.org/education-and-research/research/nfpa-research/fire-statistical-reports/fire-loss-in-the-united-states.

Roth, J. and Tarleton, J. (2014). Profiles in public service: The analytics of fire. Urban Omnibus. Retrieved from https://urbanomnibus.net/2014/06/the-analytics-of-fire/.

Statistics Canada (2023). Table 35-10-0192-01: International merchandise trade, by province and territory. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510019201. Accessed on November 1, 2025.

Zhang, C. (2023). Review of structural fire hazards, challenges, and prevention strategies. *Fire*, 6(4).

A Appendix

A.1 Data Dictionary / Variable Definitions

Table 2 includes a feature dictionary describing all variables used in the predictive model.

Name	Description	Source
num_buildings	Number of buildings in a census tract	Buildings
avg_floors	Average floors of buildings in a census tract	Buildings
avg_building_age	Average building age in a census tract	Buildings
fires_lag1m	Fire counts one month ago in a census tract	Fire incidents
fires_lag12m	Fire counts twelve months ago in a census tract	Fire incidents
month	month in which a fire incident occurred	Fire incidents
year	year in which a fire incident occurred	Fire incidents
month_sin	month converted to sine function	Fire incidents
month_cos	month converted to cosine function	Fire incidents
neighbors_fires_lag1m	Fire counts one month ago in the adjacent census tracts	Fire incidents
neighbor_count	Number of adjacent census tracts	Census

Table 2: Features used in our model

A.2 Modeling Details

We show the importance of each feature in Figure 6. In addition to the performance metrics report in Section 7.1, the confusion matrices of those three models are reported in Figure 7.

A.3 October 2025 Predictions and Resource Allocation Strategy

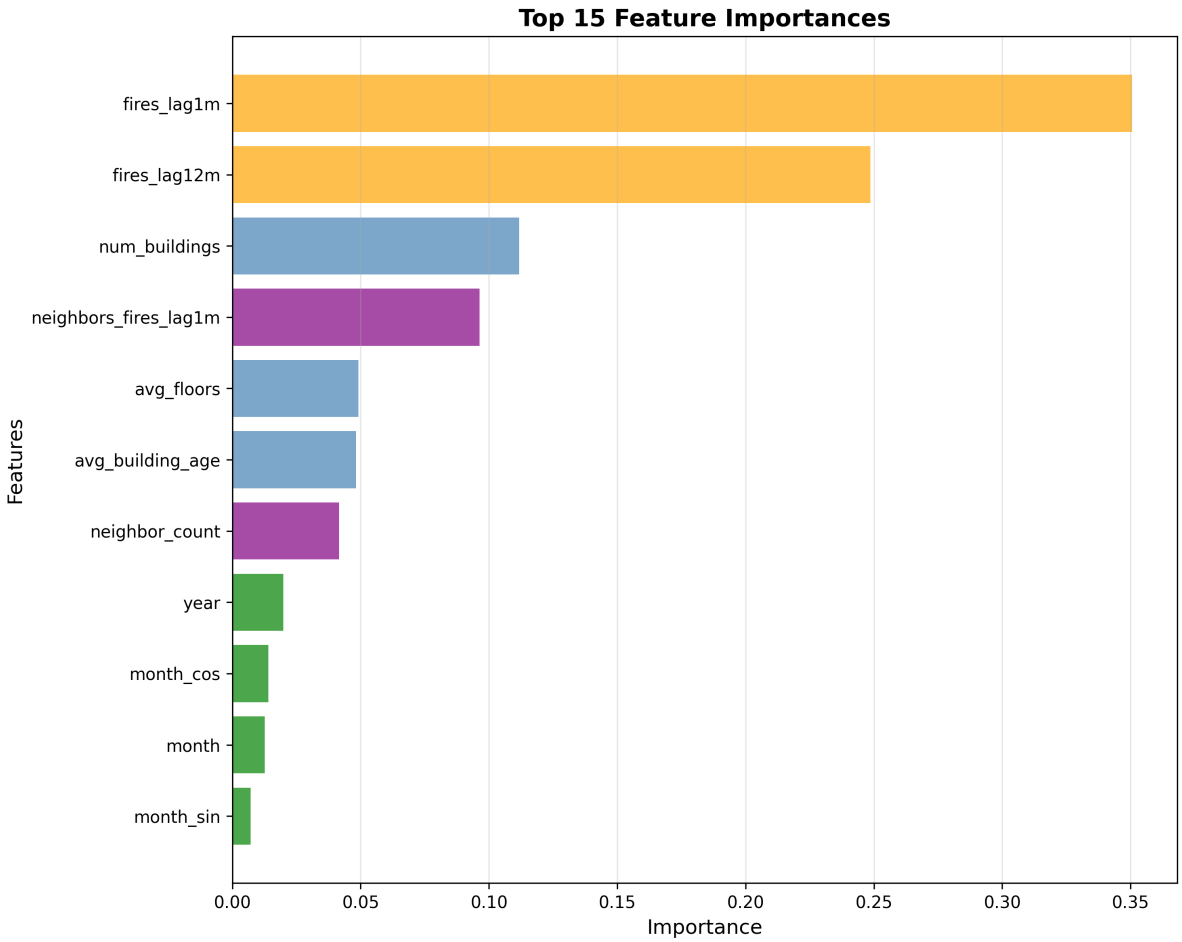
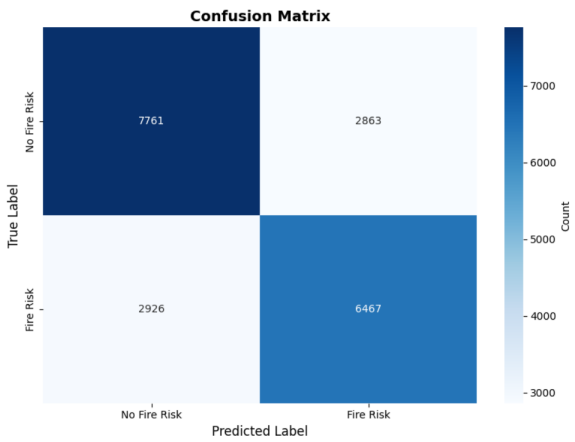
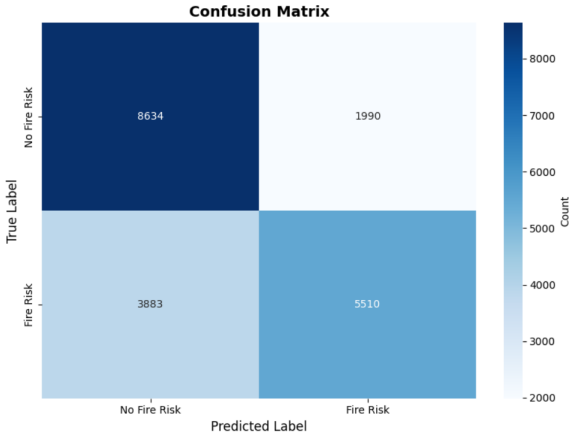


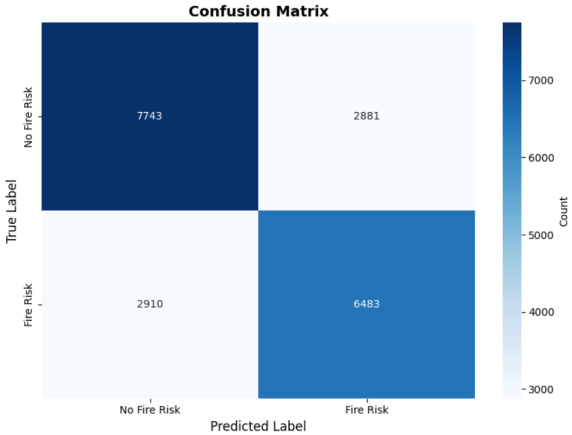
Figure 6: Predictive factors ranked by importance. Past fire history (temporal patterns) accounts for 60% of predictive power, validating focus on areas with recent fire incidents for prevention efforts.



(a) Random Forest Classifier



(b) XGBoost



(c) Ensemble Models

Figure 7: Confusion matrices

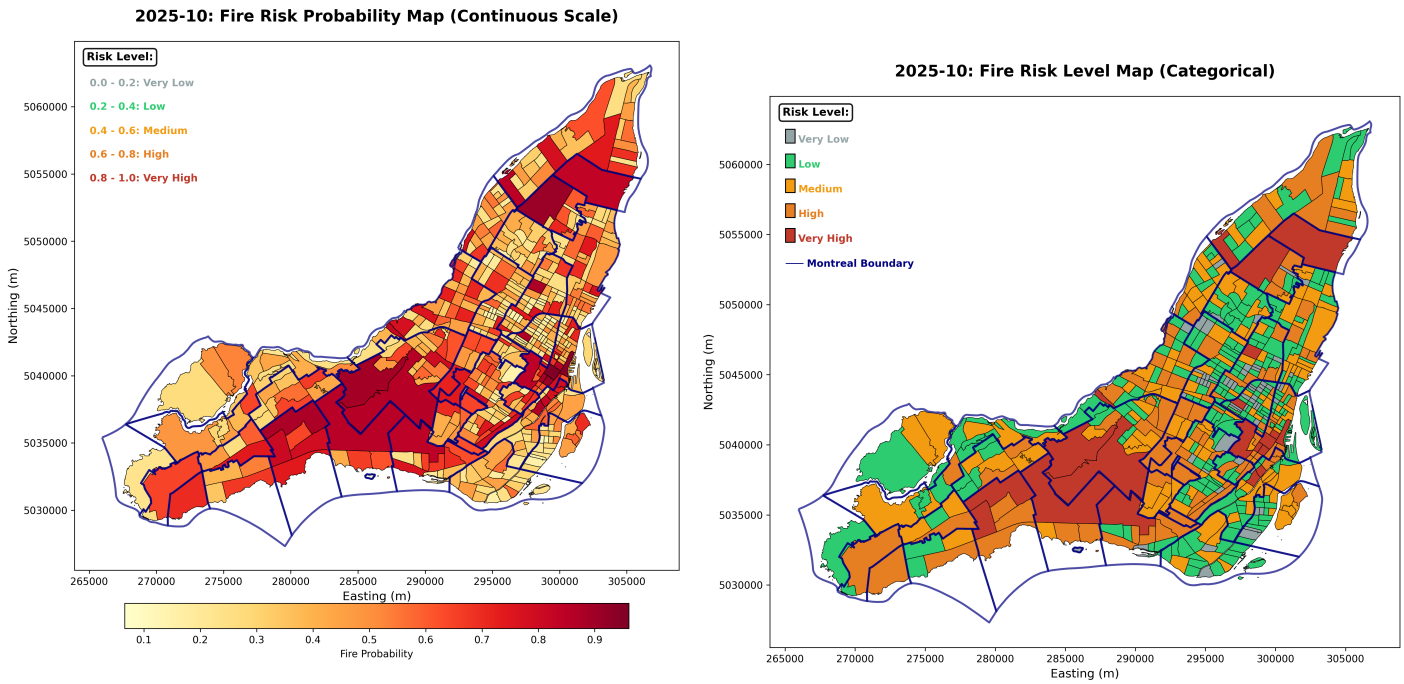


Figure 8: Predicted fire risk across Montreal (October 2025). Darker red indicates higher risk, highlighting spatial clusters for targeted intervention.

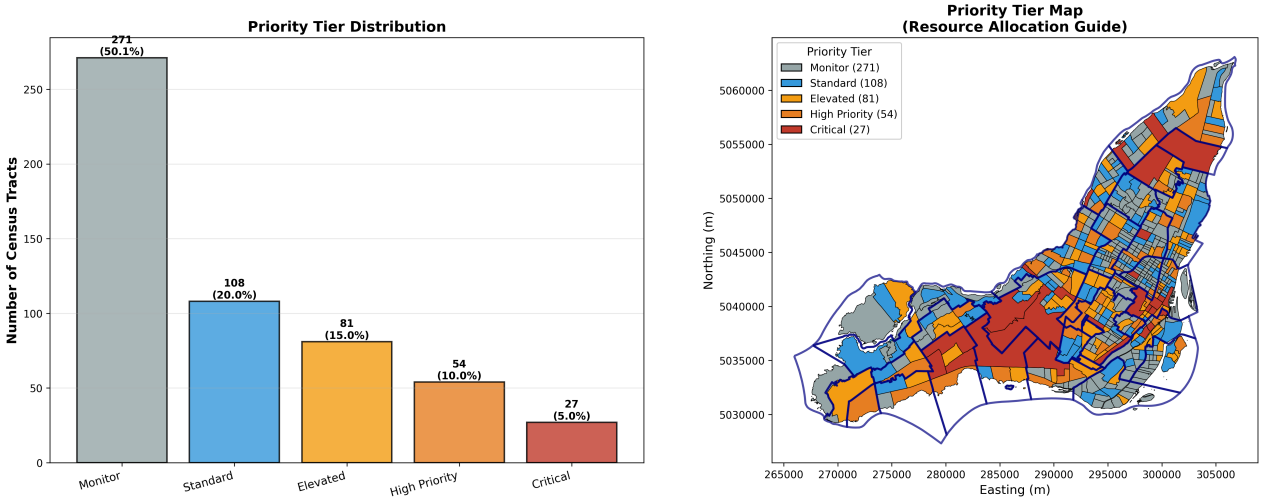


Figure 9: Five-tier risk stratification for October 2025, enabling focused resource deployment. Critical tier (5% of areas) represents highest-priority targets for immediate fire prevention intervention.

A.4 Rationale for Predictive Fire-Risk Modeling

The policy and operational motivation for the predictive framework are:

1. **Public safety**. Anticipating spatial concentrations of risk enables earlier, location-specific prevention (alarm distribution, code enforcement, resident education), which can reduce fatalities and injuries.

2. **Resource efficiency**. Fire services operate under inspection and staffing constraints. Jurisdictions such as the New York City Fire Department (FDNY) have adopted analytics-driven inspection targeting,

inspecting only 10% of 330,000 buildings annually while focusing on those with highest predicted risk (Crowley et al. (2023); Roth and Tarleton (2014)).

3. **Equity**. Fire impacts are not evenly distributed. Evidence from Shreveport indicates that fire-related harm is disproportionately concentrated in socioeconomically disadvantaged neighborhoods (Hall, 2024). Incorporating demographic indicators helps identify vulnerable populations and support equitable prevention.

4. **Modernization**. Cities such as New York and Pittsburgh have integrated predictive analytics into fire prevention and resilience planning. Montreal adopting a similar approach moves from reactive response ("where fires already happened") to proactive risk management ("where fires are most likely next").