Data Science Capstone Project

High Fire Risk Areas in Montreal Prediction

A PROJECT REPORT

Submitted by:

YCBS 299 - Team 1
Ahmed Ibrahim,
Eunseo Lee,
Pradip Kumar,
Pranavkumar Pathak

McGill University
School of Continuing Studies
Montréal, Québec, Canada
March 14th, 2023

Executive Summary

The city of Montreal's initiative on fire prevention strategy has focused on identifying the high fire-risk areas to efficiently allocate resources. With extensive research on existing state-of-the-art prediction models, we have studied tree-based models such as decision tree, random forest, XGBoost models to predict the fire risk of Montreal areas monthly. Relevant datasets from 2015 to 2022 were retrieved from Ville de Montreal website and Government of Canada which are publicly open for use. A fire risk scoring system was implemented to categorize the areas to high, medium, and low fire-risk. We have observed most of the areas had low fire-risk especially the outskirts of Montreal. To address the imbalance of the data, we have applied various techniques to minimize potential misclassification. Out of the explored models, XGBoost had the highest prediction accuracy. We recommend consulting the prediction results into consideration when planning fire prevention strategies.

Problem Statement

"Predicting the high-risk fire areas in the city of Montréal."

Fire has been essential to humanity's history, growth, and survival since its inception. Fire is one of the most common types of disaster faced by humans.

There are about 67 million fires in the world every year, and 60000 to 70000 people die in them. 85% of fire incidents in Canada are associated with human behaviors such as cooking, smoking, and using electronic devices, and one in two residential fires are caused by human mistakes [1], [2]. In Canada, buildings must comply with the guidelines and requirements of the National Fire Code (NFC) which municipalities also add in conjunction with their potentially more stringent standards [3].

Montreal is the largest city in Quebec and the second largest in Canada. The city of Montreal is responsible for reviewing, updating, and enforcing fire-related codes and ordinances to reduce the risk of structure fires and ensure properties are properly constructed by their fire codes. There are nearly 400 000 residential, commercial, industrial, and institutional buildings in Montreal but unfortunately, with limited resources, only about 7500 of those buildings are inspected annually [4].

Business Problem

Assessing fire risk is a crucial process that involves predicting, analyzing, and evaluating potential hazards. By employing an effective fire risk prediction tool, city personnel and fire departments can optimize their resources and time to prevent fire-related disasters. Accurate forecasting of fire risk is crucial in reducing economic losses and preventing loss of life. To this end, we are developing a 3-level fire-risk prediction model that can classify 1 km^2 square grids monthly in Montreal. This model will enable fire departments to enhance their inspection plans and allocate their resources more efficiently, ultimately leading to a safer community.

Data Sources

Fire risk assessment remains a challenging task due to its intricate factors that can contribute to a fire, such as spatial connections, time-based dependencies, and external factors. For example, a region's fire risk can be affected by both internal factors, like previous fire risk records, and external factors, such as weather. Herein, various datasets have been leveraged comprised of fire incidents, crimes, property assessments, demographics, weather conditions and districts all pertaining to the greater city of Montreal and its affiliated cities. Besides the weather and population datasets, *Ville de Montréal* publishes the datasets.

The spatial file used for this project is the Montreal shapefile, and its coordinate reference system (CRS) is EPSG 4326 (datum WGS84), which is compatible with latitude and longitude coordinates. Five different spatial files are provided: "dbf," "geojson," "prj," "shp," and "shx." For this project, the shapefile "LIMADMIN.shp" is used. This shapefile contains a "geometry" column that includes a polygon or multipolygon for each borough or affiliated city, as well as other columns that provide the name, IDs, surface area (m^2), and perimeter (m) of each district [5].

The fire incidents dataset (donneesouvertes-interventions-sim.csv) is curated by the *Service de sécurité incendie de Montréal (SIM)*. This dataset provides the necessary data to construct the target and critical features for this project. The selected version of the dataset includes data from January 2015 to February 2023. Earlier incidents were recorded but lacked localization data, which is a crucial component for aggregation purposes. The dataset includes columns for a unique ID, timestamp, coordinates, and other details for each incident [6].

The crime incidents dataset (actes-criminels.csv) is curated by the *Service de police de la Ville de Montréal (SPVM)*. Spatial files for this repository can also be found, but they are not utilized for this project. For each criminal event, a timestamp, criminal category, and coordinates are attributed [7].

The property assessment dataset (uniteevaluationfonciere.csv) contains numerous attributes and dimensions related to properties, but it lacks timestamps. Moreover, each assessment provides street addresses but no geographical coordinates. Some of the columns for this dataset include the number of floors, building surface area, and unit evaluation category [8].

The demographic dataset (census_proportionate.csv) is curated by Statistics Canada. It provides the population per borough or affiliated city and a yearly timestamp. No geolocation data nor monthly timestamps are found. The columns include the borough or affiliated city code and name, population, and year [9].

The weather dataset (open_meteo_historical_data.csv) is published by *Open-Meteo*. This API provides extensive historical weather data for multiple cities around the world. Although it doesn't have weather conditions per borough, it was possible to request data for each day between January 2015 and February 2023. The columns include temperature (min., max., mean), shortwave radiation sum, total rain, total snowfall, wind speed, wind gust, and more [10].

Data Exploration and Cleaning

The first task for data cleaning consisted of handling missing values. Missing values were either omitted or treated by replacing them with values to mitigate potential skewness. Each CSV file was also scanned for its data type; all columns related to time or date were converted into the datetime64 data type. This was necessary to facilitate date manipulations using these columns. For instance, for the fire and crime incidents dataset, their respective date columns were used to generate day, month, and year columns. All datasets with timestamps were filtered to only include rows dated from January 2015 to January 2023. Columns that weren't needed for mapping or aggregating were dropped. This includes unused coordinates, columns with similar information (i.e., different names for identifying boroughs), and categories with excessive cardinality.

No further data cleaning was required for the incident dataset, crime dataset, demographics dataset, and weather dataset. However, some data transformations were required for the property assessment dataset. This dataset did not have a singular column labeling each assessment with a borough ID or an ID for affiliated cities, as found in the fire incident and crime datasets. As a result, mapping was done to combine boroughs and unidentified boroughs as municipalities or their associated cities in Montreal. This column was named CODEMAMROT, which matches the names in the other dataset to facilitate data integration. Additionally, nearly 4.5% of the dataset (19116 out of 424916) had invalid values (i.e., year 9999). We treated these entries during the aggregation process.

As part of data exploration, we showcase the key trends and distributions for the fire incidents, crimes, and property assessment dataset at a monthly, grid, or borough level. The designation of grid IDs for each qualifying dataset will be described in the feature engineering section of this report. The first trend to explore was the number of fire incidents per month (see figure 1).

Figure 1. Number of fire incidents per month for 2015-2022.

A degree of variance in incidents per month can be observed especially during the COVID-19 period and some level of seasonality can also be observed. The distribution of fire incidents per grid by count and mean has been explored (see figure 2 and 3).

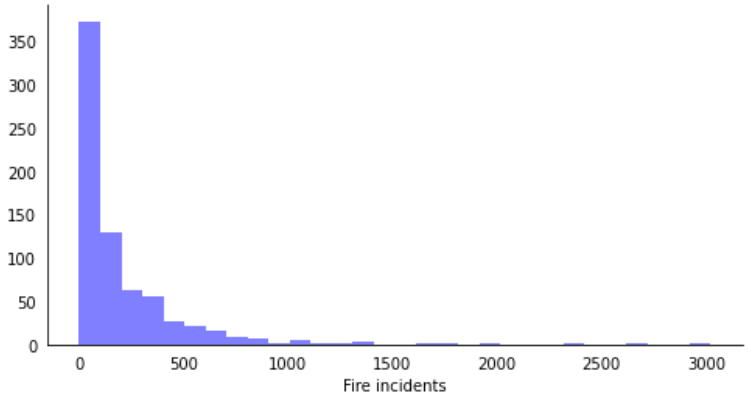

Figure 2. Histogram of fire incidents per grid. 714 grids on the Montreal shapefile and a bin size of 30.

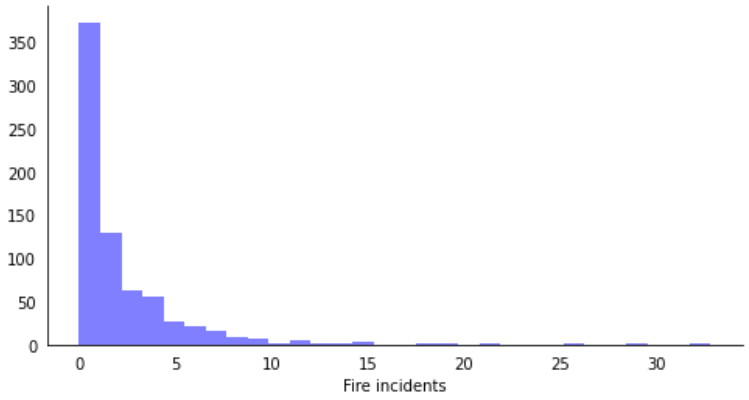

Figure 3. Histogram of mean fire incidents per grid and month. 714 grids on the Montreal shapefile and the binning size is set to 30.

The maximum number of fire incidents per grid is 3021, while the maximum number of fire incidents per month and grid is almost 33. However, more than half of the grids have less than 100 fire incidents per grid and less than 1 fire incident per grid and month. Thus, the dataset is highly skewed and imbalanced here. A heatmap showing the distribution of total fire incidents per grid can be found in Appendix C.

For the crime dataset, the distribution of crime incidents per grid was explored. The maximum number of crime incidents per grid is 4184, while more than half of the grids have less than 100 crime incidents over the full time. The distribution is also skewed (see histogram in Appendix C).

As for the property assessment dataset, the number of property assessments per borough was explored. The average count of assessments per borough is 14752, while the smallest count is 74 and the highest count is 42241. The distribution is less skewed compared to the other distributions discussed (see Appendix C).

Subsequently, correlations between each dataset and its attributes were analyzed (refer to correlation matrix in Appendix C). We removed attributes showing high correlation between each variable to reduce redundancy and those that had low correlation to the target variable.

We observed that temporal data, i.e., datasets with a time component, had a higher correlation with fire risk scores than static features, such as property assessments and fire stations. This finding aligns with research from Soongsil University, which emphasizes the impact of natural environmental factors like temperature, humidity, precipitation, wind speed, and direction on fire incidents [12]. Among the temporal data, we also noticed that weather information had the strongest relevance to the target variable.

Feature Engineering

The Montreal shapefile provided the geometry of the districts, but not the necessary grids for this project. Therefore, a Python pipeline was designed to generate a shapefile containing the geometry of the city with a square tessellation. Each grid was set to be 1km by 1km, except for the squares that overlapped with the city's boundary edges. For each grid, a unique ID, area, and centroid were assigned. To assign a district to a grid, the grid must partially or fully overlap with a district. Some grids overlapped with multiple districts, so the district with the most overlap was attributed (see map in Appendix C).

Using the grid shapefile, a table was generated with all grid IDs in one column, the associated district for each grid, and the months for the full period for every grid. Among the datasets, only the fire incidents and crimes had compatible coordinates with our grid solution. A spatial join was performed with each set of latitude and longitude from the fire and crime rows to attribute a grid ID to each of these events. The property assessment and population datasets were only aggregated by district, but the population dataset was also aggregated by year. The weather dataset was aggregated by the timestamps of the fire and crime datasets. For each crime or fire incident, weather conditions were matched to aggregate the weather conditions associated with these events. For the rows of grids and/or months lacking either of these events, the average weather condition for the month was imputed.

The fire incidents dataset was used to generate the number of incidents, fire stations, divisions involved, units deployed, and the average number of units deployed per grid and month. Additionally, the number of incidents per grid and month for each description group was aggregated. Similarly, for the crime features, the number of crimes and the number of crimes of each type were aggregated per grid and month. The property assessment features include the number of property assessments, average height above ground, average dwelling found in property assessments, average construction year (void years after 2023), average land area (m^2) and average building area (m^2) per borough or affiliated city for each UEF category. These rows are repeated for each month. For the population dataset, synthetic features generated being the total population and population density per borough or affiliated city and year.

The weather features require the incidents and crime datasets, with the weather dataset containing the weather conditions for each day of the study period. By joining these two datasets with the weather dataset via a day column, a table containing the weather conditions with an associated grid ID is generated. This table is then aggregated with the main integrated data tables containing all grid IDs and months for the study period. The

average weather conditions associated with all events are computed for each grid and month. For the grid and month pairs without such events, the average weather conditions for all days of a given month are used. Some of the average weather conditions include temperature, precipitation sum, rain sum, snowfall sum, wind speed, wind gusts, wind direction, and shortwave radiation sum.

For every computed feature, a 4-month backward shift equivalent is generated. All the above features are aggregated per month or associated with a month. Fourteen quarterly features were generated using the 4 months prior to a given month. These features included the quarterly total of units deployed, mean fire incidents, maximum number of fire incidents, total number of fire incidents, crime count, total incidents count, mean temperature, total rain, total snow, mean wind speed, and mean wind gust.

To compute the fire-risk levels, the fire-risk rank is computed using the monthly fire incident count, and the second fire-risk rank is computed using the quarterly fire incident count. All these ranks are aggregated in descending order. All 714 grids receive two ranks for each month, and a combined rank is generated using a duplicate of the first rank. In the case of equal ranks for rows in the same month, the second rank is used as a tiebreaker. If rows have equal values for both ranks, then they remain equal. Using this approach, the top 5% of grids are labeled high fire-risk, the next 15% as medium fire-risk, and the rest as low-fire risk. This approach constructs a fire scoring system based on monthly and quarterly fire incidents and attributes the most eventful grids as high fire-risk. Lastly, a forward shift of the fire-risk is stored in a new column, serving as the target as each row is used to predict the next month's fire-risk level.

The non-shifted features are dropped, including columns used solely to aggregate data, such as the year and district names, and columns closely related to the target, such as the fire incident quarterly mean, the fire incident count, the fire-risk monthly rank (1st rank), the fire-risk quarterly rank (2nd rank), the cumulative fire-risk rank, and fire-risk levels. The target is also label encoded.

To conduct feature filtration, three different methods are used: ANOVA F-test, chi-squared test, and mutual information. The dataset is split into a target variable and numerical features. The ANOVA F-test is used to select numerical features with a p-value less than 0.05. Similarly, the chi-squared test is used to select numerical features with a p-value less than 0.05. Finally, mutual information is calculated between each numerical feature and the target variable to select features with an MI score greater than 0.05. The selected features from all three methods are then combined to create a final list of selected features. Out of the 50 features injected into this process, 43 satisfied the three filtration checks. A correlation matrix is also generated by studying the correlations between filtered features (see Appendix C). Some of the fire incident and weather features seem to be highly correlated.

Tools and Techniques Used

A variety of tools and techniques were used in different steps of the project. Exploratory analysis was conducted using Python, Excel (pivot table), and Alteryx. Data cleaning, pre-

processing, model development, and visualizations were primarily accomplished using Python (Jupyter notebook and libraries) and Tableau.

Summary of Modelling Techniques Evaluated

Our objective can be defined as a classification problem as we are classifying areas in Montreal into three classes by predicting high (0), medium (1), and low (2) fire risks for the upcoming month. A decision tree model was constructed as a baseline model to compare with other models we have chosen – random forest and XGBoost (a variant of random forest). As a part of pre-processing, the aggregated feature dataset was ordered by year and month. The label (fire risk classes) was then encoded from 0 to 2.

Table 1. Fire-risk levels.

Classification	Level
High Fire-Risk	0
Medium Fire-Risk	1
Low Fire-Risk	2

Our full dataset ranged from January 1st, 2015, to December 31st, 2022. The first 92 months (up to July 31st, 2022) were used for training and validation. However, since a 4-month shift was used for the features, 87 months remained for training and validation. We maintained around a 16:1 ratio to ensure enough data was provided for training. Our dataset was scaled using robust scaler to standardize probable outliers.

Due to the binning strategy described in the previous feature engineering section, our dataset was imbalanced per class. Most of the data points fell under the Low Fire-Risk category, which could influence the likelihood of the model to predict most of the cases to the Low Fire-Risk class. To avoid skewed predictions, we applied up-sampling technique (SMOTE) in each cross-validation fold to add synthetic data to minority classes. For each model, 62,118 entries with 42 selected features were trained.

For each model, stratified cross-validation (5 folds) was used to validate the trained model. This method allowed us to handle potential data leakage from test knowledge being exposed in the training process. A grid search was conducted to search through a defined hyperparameter space for each model, and the optimal hyperparameters were found by optimizing the ROC-AUC (one vs rest) score for each model. Refer to the appendix for visualization on 5-fold stratified cross-validation for our dataset and find the optimal hyperparameter (see Appendix C).

Beginning with the decision tree model (baseline), the balanced validation accuracy yielded approximately 0.67, and the test accuracy yielded 0.64. The random forest model resulted in balanced validation and test accuracy of 0.75 and 0.72, respectively. The XGBoost model had slightly lower balanced accuracies of 0.74 (validation) and 0.71 (test).

Although accuracy can be a simple method to assess model performance, it is not a good choice for our multi-classification problem as we have highly imbalanced classes. For thoroughness, we have explored different evaluation measures to find the best model for our problem. With fire predictions, minimizing false negatives (ex. predicting it is not high fire-risk area, but it is) is more critical than minimizing false positives (ex. predicting it is high fire-risk area, but it is not). With this in mind, we have focused on recall, F1-score, and AUC among other measures (see Appendix C for ROC-AUC curve (One vs Rest) for each model).

Table 2. Performance evaluation comparison between models.

Model	Precision	Recall	F1-score	AUC
Decision Tree	0.87	0.76	0.80	0.74
Random Forest	0.90	0.90	0.90	0.90
XGBoost	0.90	0.91	0.90	0.91

Modelling Results

The comparison of metrics in table 3 shows that the random forest and gradient boosted tree models are the most performant among the three. However, both models are equally good at distinguishing between positive and negative classes, as they properly classify true positives among all predicted positives. By looking at the confusion matrix for all three models (see figure 4), both models predict low-risk and high-fire risk grids correctly, with the XGBoost model performing slightly better for high-risk grids and significantly better for low-risk grids. However, the XGBoost model misclassifies a higher percentage (52%) of medium-risk grids as low compared to the random forest model (43%). Overall, both models perform similarly but excel differently depending on the fire-risk level being focused on. The XGBoost model could be considered a superior choice for this business problem as identifying high-risk grids more efficiently is of higher interest, but the random forest model performs better at classifying the medium-risk grids and is more computationally efficient for training and inference.

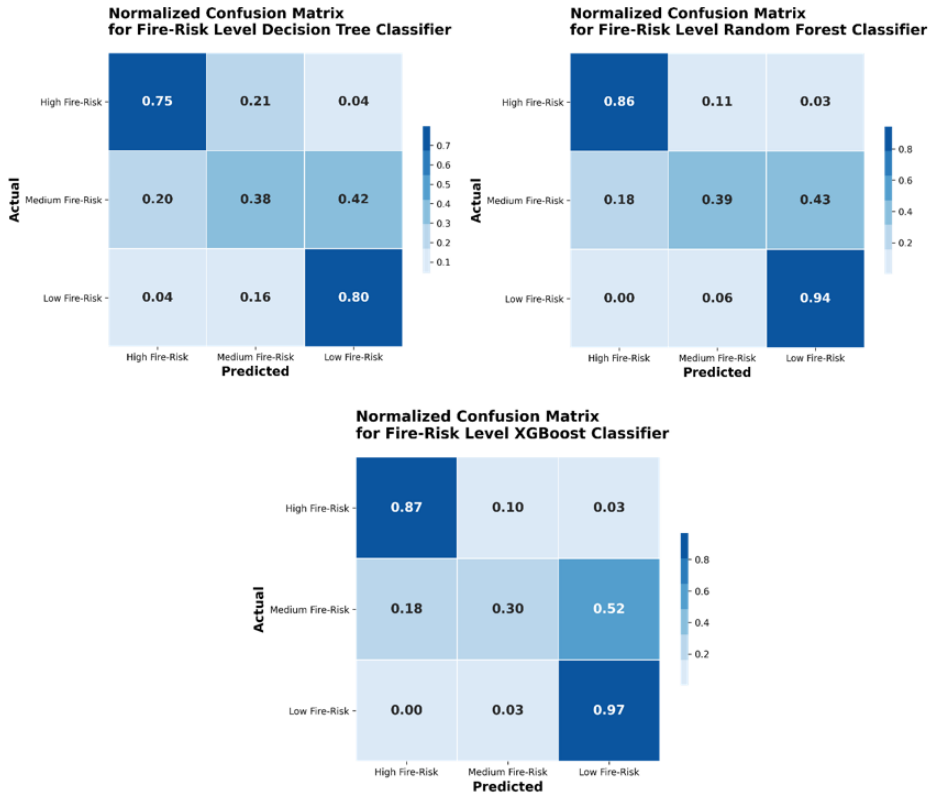

Figure 4. Confusion matrices for the decision tree, random forest and XGBoost classifiers.

The performance of the XGBoost and random forest models was evaluated using testing data from August 1, 2022, to December 31, 2022, and a map was generated to showcase the average false positive rate for each model. The XGBoost model had a low average false positive rate, misclassifying only 4 out of 714 grids at an 80% rate. In contrast, the random forest model had a higher average false positive rate, misclassifying 13 grids at an 80% rate.

Although the last month of the period lacked a calculated fire-risk level due to the 1-month forward shift of the target, the fire-risk level for the grids of that month (February 2023) was predicted for both models and can be seen in the maps provided in the Appendix.

To improve the model's performance, the top 10 features that impacted both the XGBoost and random forest classifiers by feature importance were identified. It was found that fire incident and weather features dominated the list in terms of feature importance. These findings can be used to enhance the current model or develop new ones for future fire risk prediction.

Insights and Challenges

During the project, we faced several challenges related to data integration and preprocessing. One of the main challenges was obtaining reliable data from various sources and integrating them together to create a comprehensive dataset. We had to ensure that the data was

accurate, up-to-date, and compatible with our other data sources. In some cases, we had to perform data cleaning and formatting to ensure consistency and quality.

Another challenge was deciding on the appropriate aggregation level geographically and temporally. We had to balance between the need for high spatial resolution to capture local variations in fire risk and the need for a sufficient sample size to train our models. We chose to aggregate the data by monthly and quarterly to predict for the subsequent month; however, we intend to explore quarterly predictions as a next step to incorporate the time required for the data to be published and provide realistic timeline for planning ahead. In addition, we explored various sizes of grids such as 500m by 500m and 750m by 750m to determine the optimal grid size for our models.

Having domain knowledge on fire systems, Montreal geography, and current strategies would have been useful when making decisions. We could have consulted with experts in the field to gain insights and validate our approach.

We also intended to explore a larger hyperparameter space, which could have improved the performance of our models. However, due to computational limitations, we had to limit the search space.

Furthermore, an emerging trend for solving predictive problems is to use both deep learning and decision-trees together via an ensemble learning approach, which could have yielded a more potent solution. This is something we could explore in the future to further enhance our models.

Conclusions

In conclusion, our project aimed to develop a model to identify high fire-risk areas in Montreal to assist the local government in allocating resources and implementing prevention strategies. We accomplished this by integrating various datasets related to fire incidents, weather conditions, population, and property assessments. We then engineered features from these datasets and trained three models: decision tree, random forest, and XGBoost.

After evaluating these models, we found that both random forest and XGBoost models performed well in accurately classifying high and low fire-risk areas, but XGBoost model was able to identify high-risk areas more efficiently. Furthermore, we were able to generate a map showcasing the average false positive rate for the XGBoost model and random forest model to visualize their performance.

Overall, we recommend the use of our XGBoost model to identify high fire-risk areas in Montreal, and to extend the model's scope to predict for longer periods, allowing for the prioritization of high-risk areas in 1-year and 5-year plans. We believe this information will assist the local government in allocating resources and implementing effective prevention strategies to reduce the potential risk of fires in Montreal.

References

1. J. Clare and H. Kelly, *Fire and at-risk populations in Canada Analysis of the Canadian National Fire Information Database*, Natl. Fire Prot. Assoc., no. December, 2017.
2. T. C. Press, *1 in 2 fires is due to human error: Quebec public safety minister, CTV News*, 2022.https://montreal.ctvnews.ca/1-in-2-fires-is-due-to-human-error-quebec-public-safetyminister-1.6102520 (accessed Jan. 22, 2023).
3. Quebec Safety Code, *Chapter VIII – Building, and National Fire Code of Canada 2010 (amended)*, Natl. Res. Counc. Canada, pp. 1–325, 2010.
4. Ville de Montréal, *Diversity of duties, Service de sécurité incendie de Montréal*, 2021. https://ville.montreal.qc.ca/sim/en/diversity-duties (accessed Jan. 24, 2023).
5. Ville de Montréal, *DÉSUET: Limite administrative de l'agglomération de Montréal (Arrondissements et Villes liées)*, 2020. https://donnees.montreal.ca/ville-de-montreal/polygones-arrondissements (accessed Jan. 24, 2023).
6. Ville de Montréal, *Actes criminels*, Service de police de la Ville de Montréal (SPVM), 2023. https://donnees.montreal.ca/ville-de-montreal/actes-criminels (accessed Jan. 24, 2023).
7. Ville de Montréal, *Interventions des pompiers de Montréal*, Service de sécurité incendie de Montréal (SIM), 2021. https://ville.montreal.qc.ca/sim/en/diversity-duties (accessed Jan. 24, 2023).
8. Ville de Montréal, *Unités d'évaluation foncière*, 2023. https://donnees.montreal.ca/ville-de-montreal/unites-evaluation-fonciere (accessed Jan. 24, 2023).
9. Statistic Canada, *Census of Population*, 2020. https://www12.statcan.gc.ca/census-recensement/index-eng.cfm (accessed Jan. 24, 2023).
10. Open-Meteo, *Historical Weather API*, 2023. https://open-meteo.com/en/docs/historical-weather-api (accessed Jan. 24, 2023).
11. Ville de Montréal, *Arrondissement de la Ville de Montréal - Liste*, 2016. https://donnees.montreal.ca/ville-de-montreal/arros-liste (accessed Jan. 24, 2023).

Appendix

Appendix A [Lists]

List of columns for Montreal shapefile (*LIMADMIN.shp*)

- MUNID: Identifier for the administrative division of municipalities in Quebec, (MAMROT).
- CODEID: Unique identifier.
- CODEMAMROT: Identifier for the administrative division - unique identifier with MAMROT district code as prefix.
- NOM: Name of the administrative division - as described by the Quebec Government Toponymy Commission.
- TYPE: Type or entity of the administrative division - e.g., Borough, Associated City.
- ABREV: Abbreviation for the definition of boroughs and associated cities.
- NUM: Internal alphanumeric identifier (geomatics).
- AIRE: Official non-calculated area in square meters.
- PERIM: Official non-calculated perimeter in meters.
- GEOM: Administrative division geometry formatted according to the Well-known text standard.

List of columns for fire incidents dataset (donn*eesouvertes-interventions-sim.csv*)

- INCIDENT_NBR: Unique ID for incident.
- CREATION_DATE_TIME: Timestamp of incident.
- DESCRIPTION_GROUPE: Grouping of intervention types into 6 categories: *Building Fires, Other Fires, Non-Fire, Fire Alarms, First Responders, False Alerts/Cancellations*.
- INCIDENT_TYPE_DESC: Detailed incident type.
- CASERNE: Number of the fire stations responsible for the area where the event occurred.
- NOM_VILLE: Name of the city where the incident occurred.
- NOM_ARROND: Name of the borough where the incident occurred.
- DIVISION: SIM division responsible for the area where the event occurred.
- LONGITUDE, LATITUDE: Geographic location of the event after obfuscation at an intersection according to the WGS84 geodetic reference.
- NOMBRE_UNITES: Number of vehicles deployed to respond to the event.

List of columns for criminal incidents dataset (*actes-criminels.csv*)

- CATEGORIE: Nature of the event. 6 categories include: break and enter, theft from motor vehicle, motor vehicle theft, mischief, robbery, and criminal offense causing death.
- DATE: Timestamp of criminal event.
- QUART: Time of day when the event was reported to the SPVM. Options include day (8:01 a.m. and 4:00 p.m.), evening (4:01 p.m. and midnight) and night (12:01 a.m. and 8:00 a.m.).

- PDQ: Number of the police station covering the area where the event occurred.
- X: Geospatial position according to the MTM8 projection (SRID 2950).
- Y: Geospatial position according to the MTM8 projection (SRID 2950).
- LATITUDE: Geographic location of the event after obfuscation at an intersection according to the WGS84 geodetic reference.
- LONGITUDE: Geographic location of the event after obfuscation at an intersection according to the WGS84 geodetic reference.

List of columns for property assessment dataset (*uniteevaluationfonciere.csv*)

- ID_UEV: Unique system identifier.
- CIVIQUE_DEBUT: Civic number (range - start).
- CIVIQUE_FIN: Civic number (range - end).
- NOM_RUE: Street name.
- SUITE_DEBUT: Unit number (apartment or local).
- ETAGE_HORS_SOL: Maximum number of floors:
- If the UEF includes a single building: Number of floors of the building.
- If the UEF includes multiple buildings: Number of floors of the building with the most floors (maximum).
- NOMBRE_LOGEMENT: Number of housing units.
- ANNEE_CONSTRUCTION: Year of construction.
- CODE_UTILISATION: CUBF coding.
- LETTRE_DEBUT: First letter of the apartment.
- LETTRE_FIN: Last letter of the apartment.
- LIBELLE_UTILISATION: CUBF description.
- CATEGORIE_UEF: Unit evaluation category (Regular or Condominium).
- MATRICULE83: Roll number (NAD83 MT8 geospatial system).
- SUPERFICIE_TERRAIN: Land area for property assessment purposes (square meters).
- SUPERFICIE_BATIMENT: Building floor area, i.e. gross floor area corresponding to the sum of the areas of each of the whole floors of the main building and, if applicable, those of the attic, integrated garage and integrated greenhouse (square meters).
- NO_ARROND_ILE_CUM: Borough identifier (MAMROT reference identifier).
- MUNICIPALITE: Internal municipality identifier.

List of columns for population dataset (*census_proportionate_v5.csv*)

- CODEMAMROT: Identifier for the administrative division - unique identifier with MAMROT district code as prefix.
- NOM: Name of the administrative division - as described by the Quebec Government Toponymy Commission.
- YEAR: Yearly timestamp.
- POPULATION: Total number of individuals.

List of columns for weather dataset (*open_meteo_historical_data.csv*)

- Time: UNIX timestamp.

- Temperature_2m_max (°C): Maximum temperature for the day.
- Temperature_2m_min (°C): Minimum temperature for the day.
- Temperature_2m_mean (°C): Mean temperature for the day.
- Shortwave_radiation_sum (MJ/m²): The sum of solar radiation on a given day in Megajoules.
- Precipitation_sum (mm): Sum of daily precipitation (including rain, showers, and snowfall).
- Rain_sum (mm): Sum of daily rain.
- Snowfall_sum (cm): Sum of daily snowfall.
- Windspeed_10m_max (km/h): Maximum wind speed on a day.
- Windgusts_10m_max (km/h): Maximum wind gusts on a day.
- Winddirection_10m_dominant (°): Dominant wind direction.
- Wt0_fao_evapotranspiration (mm): Daily sum of ET_0 Reference Evapotranspiration of a well watered grass field.

Appendix [Tables]

Table 3. List of hyperparameters obtained through grid search for each model.

Decision Tree	Random Forest	XGBoost
max_depth=100, min_samples_split=2, min_samples_leaf=1, criterion='gini', max_features='sqrt', splitter='random', random_state=42	n_estimators=900, max_depth=100, min_samples_split=5, min_samples_leaf=2, max_features='sqrt', class_weight='balanced', criterion='gini', bootstrap=True, oob_score=True, random_state=42, n_jobs=-1	num_class=3, learning_rate=0.1, max_depth=60, n_estimators=600, subsample=0.6, colsample_bytree=0.6, eval_metric='aucpr', objective='multi:softprob', tree_method='gpu_hist', gpu_id=0,

Appendix C [Figures]

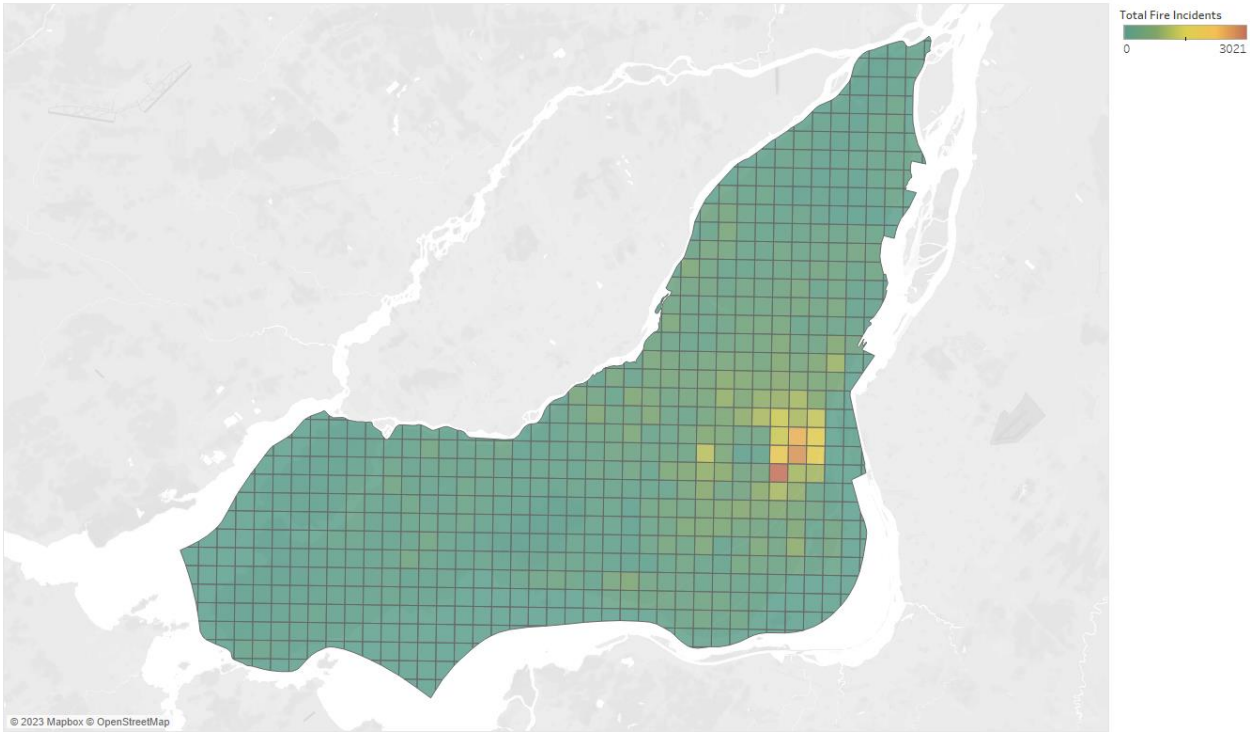

Figure 5. Fire incidents per grid on the Montreal shapefile.

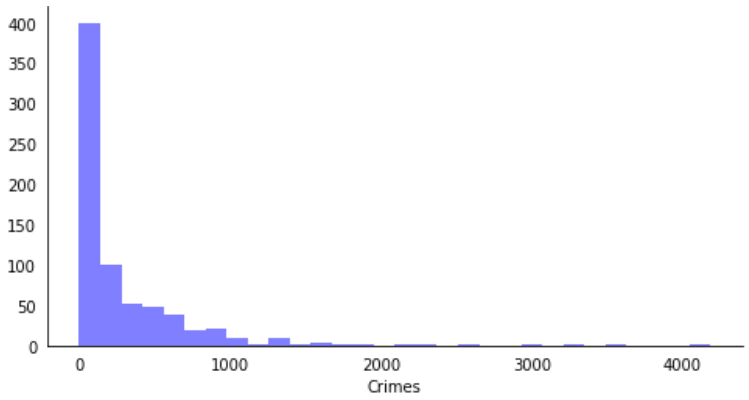

Figure 6. Histogram of crimes per grid. 714 grids on the Montreal shapefile and the binning size is set to 30.

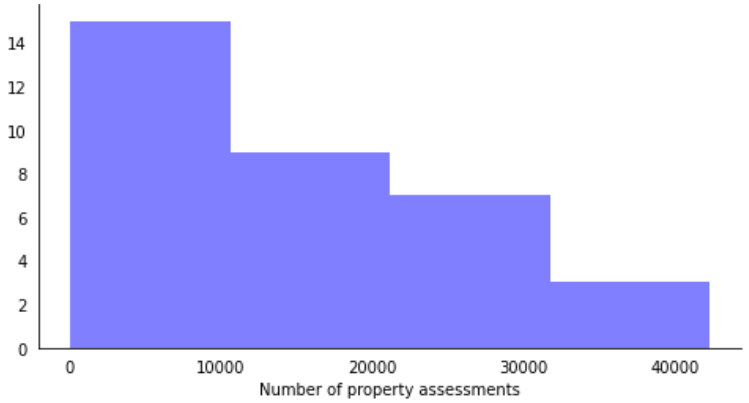

Figure 7. Histogram of property assessments per district. 714 grids on the Montreal shapefile and the binning size is set to 4.

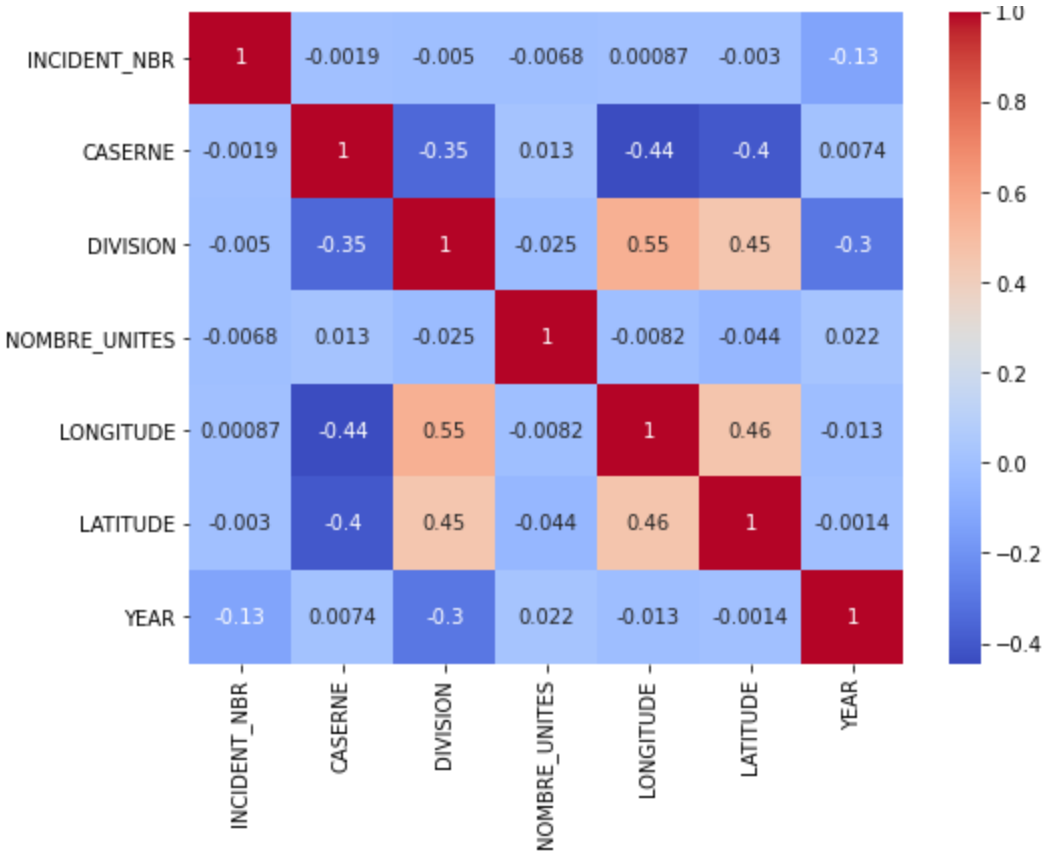

Figure 8. Correlation matrix for numerical features from fire incidents dataset.

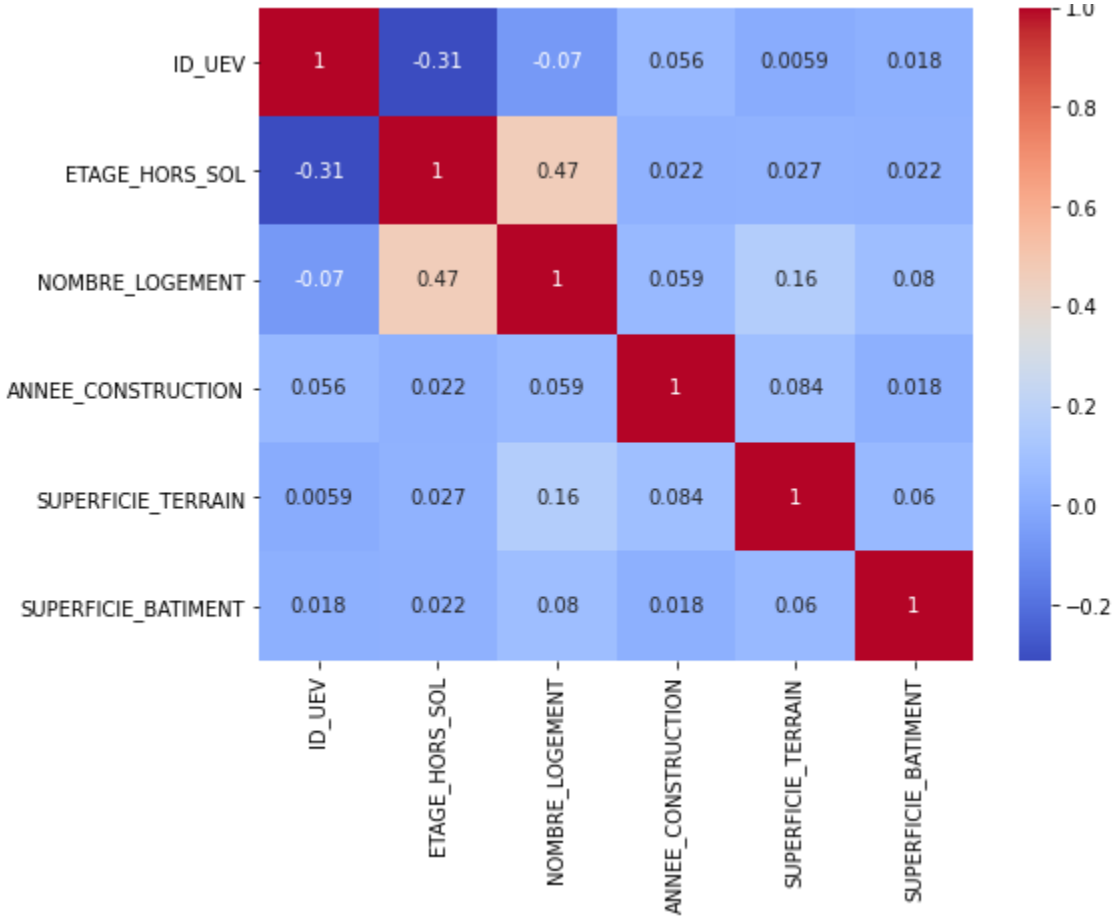

Figure 9. Correlation matrix for numerical features from property assessment.

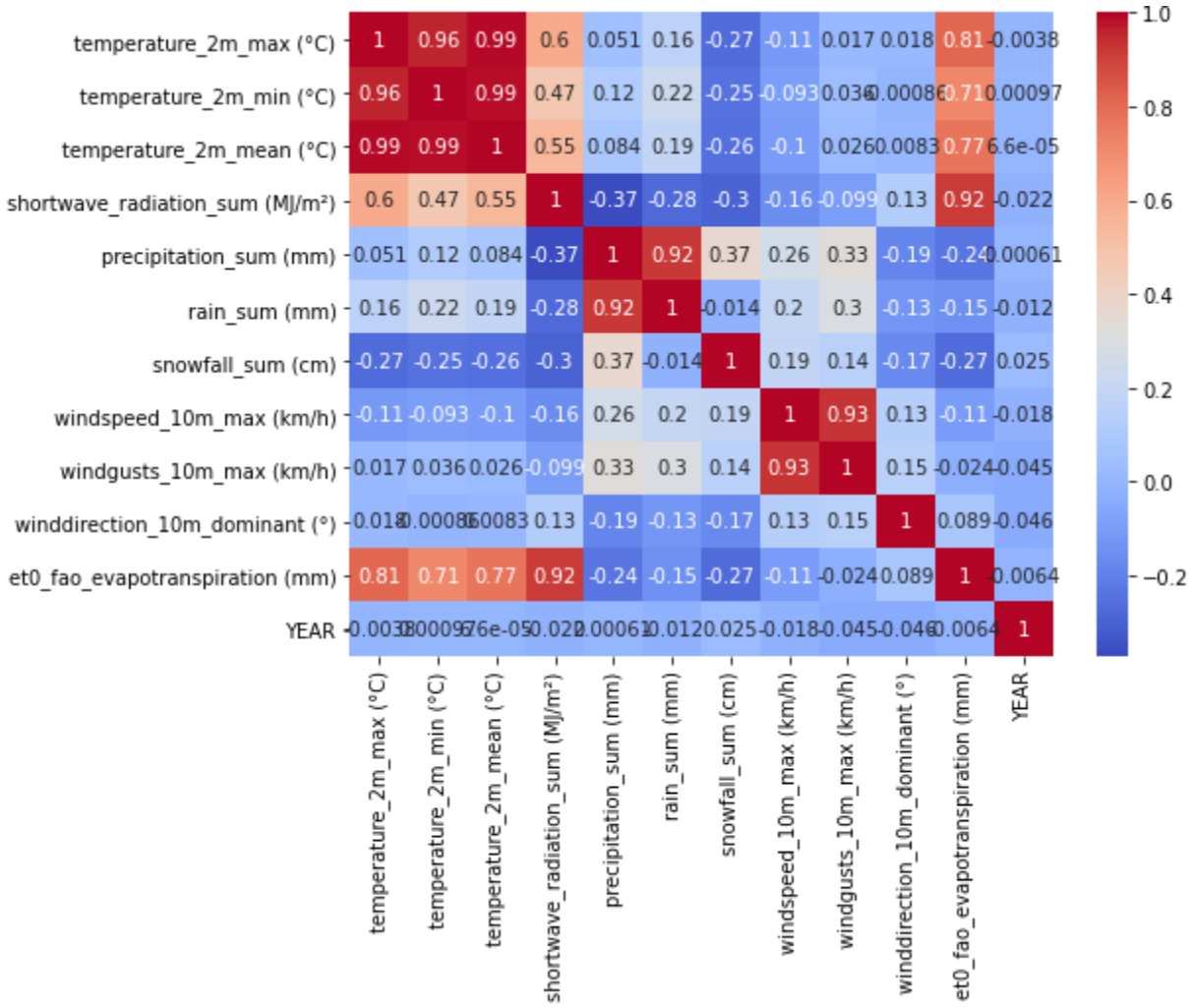

Figure 10. Correlation matrix for numerical features from the weather dataset.

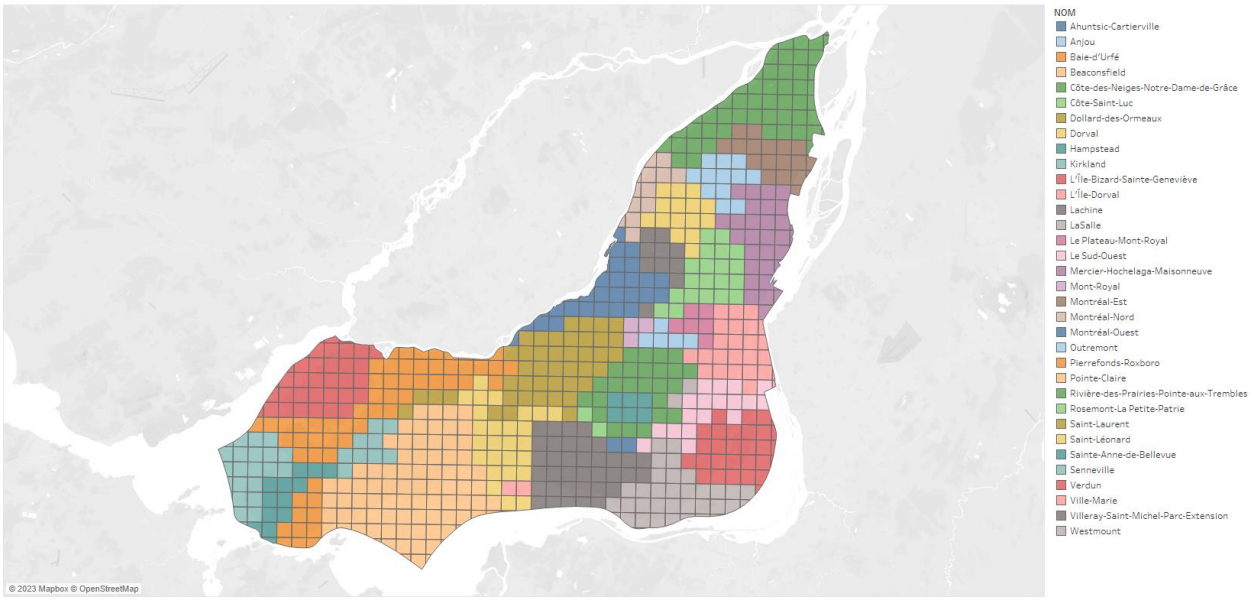

Figure 11. Montreal shapefile with square tessellation and assigned districts.

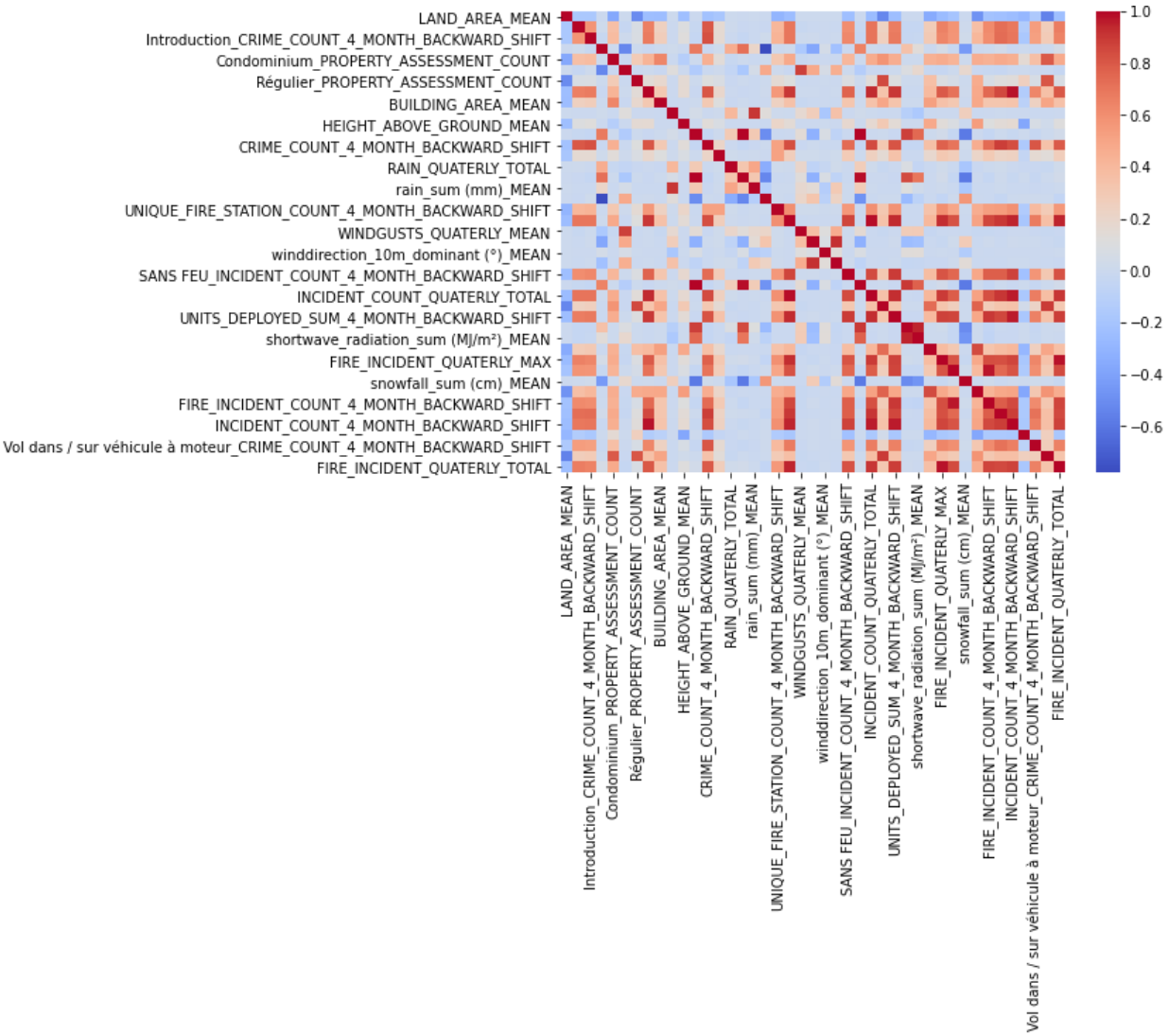

Figure 12. Correlation matrix for filtered features.

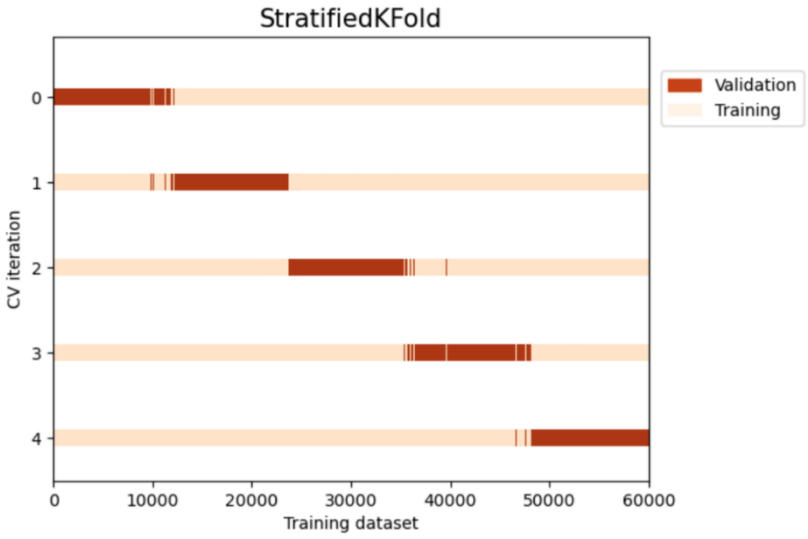

Figure 13. 5 iterations of stratified cross validation with training dataset.

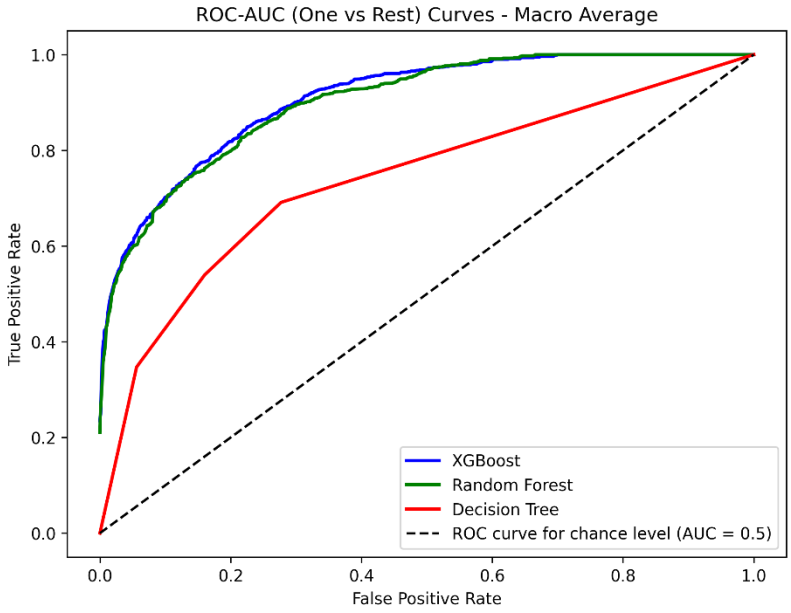

Figure 14. ROC-AUC curve (One vs Rest) for XGBoost, random forest and decision tree classifiers.

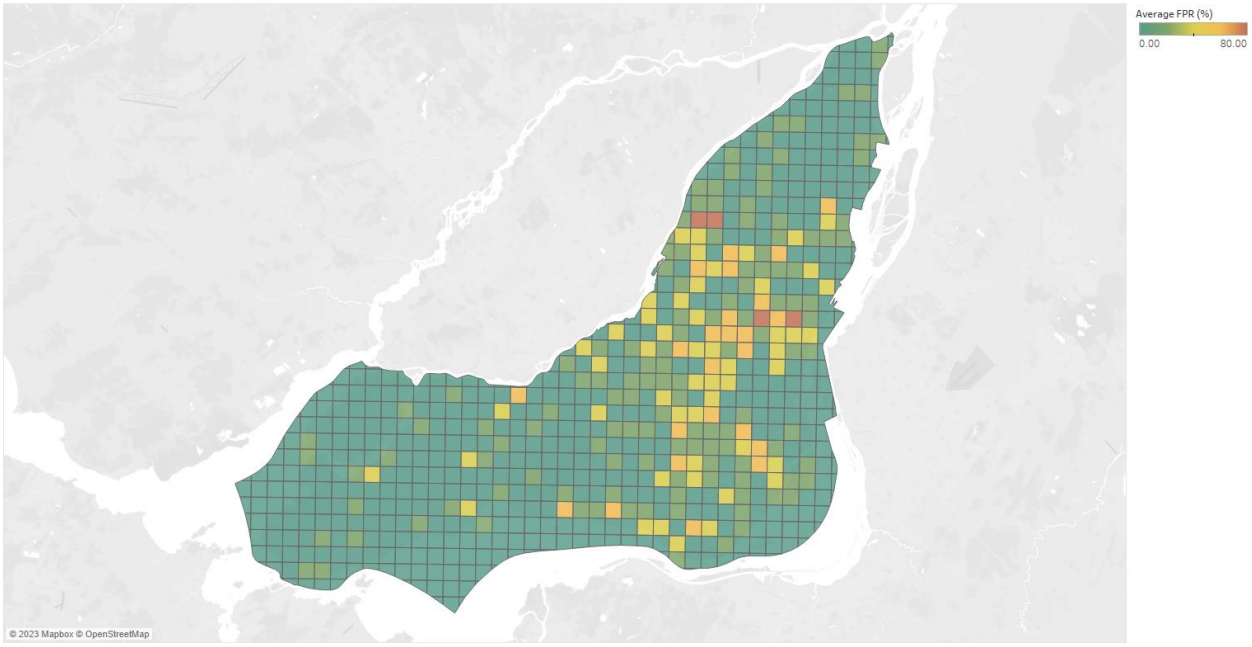

Figure 15. Average false positive rate for fire-risk prediction of Montreal using a XGBoost classifier.

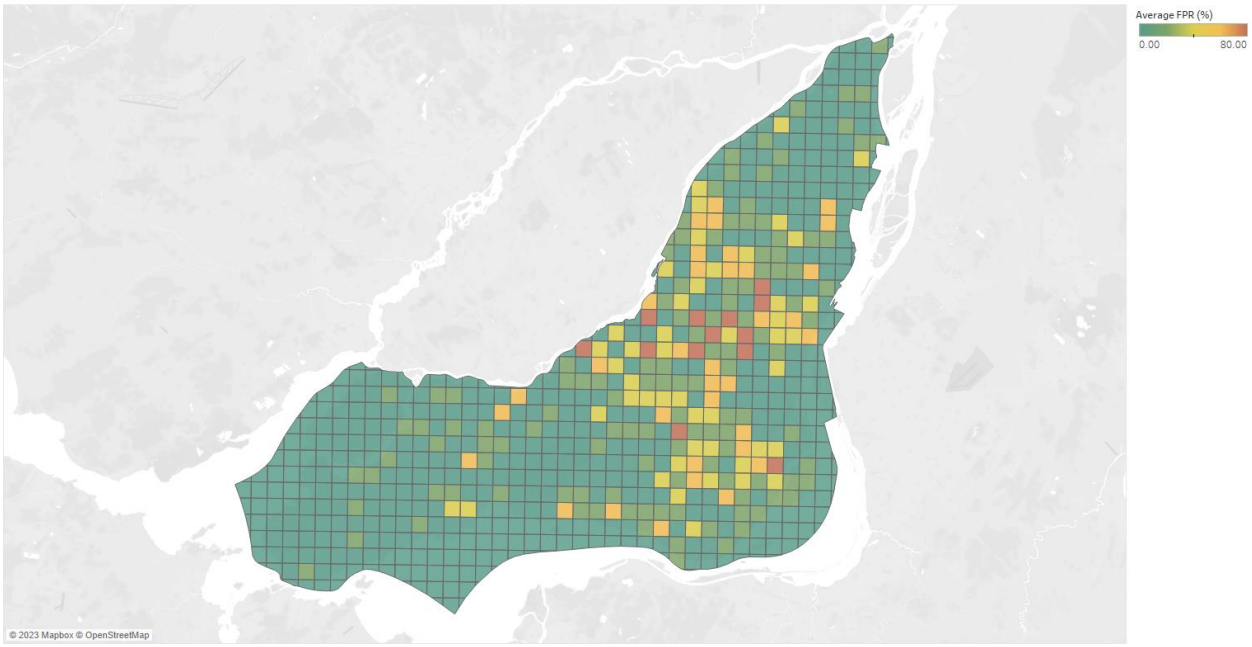

Figure 16. Average false positive rate for fire-risk prediction of Montreal using a random forest classifier.

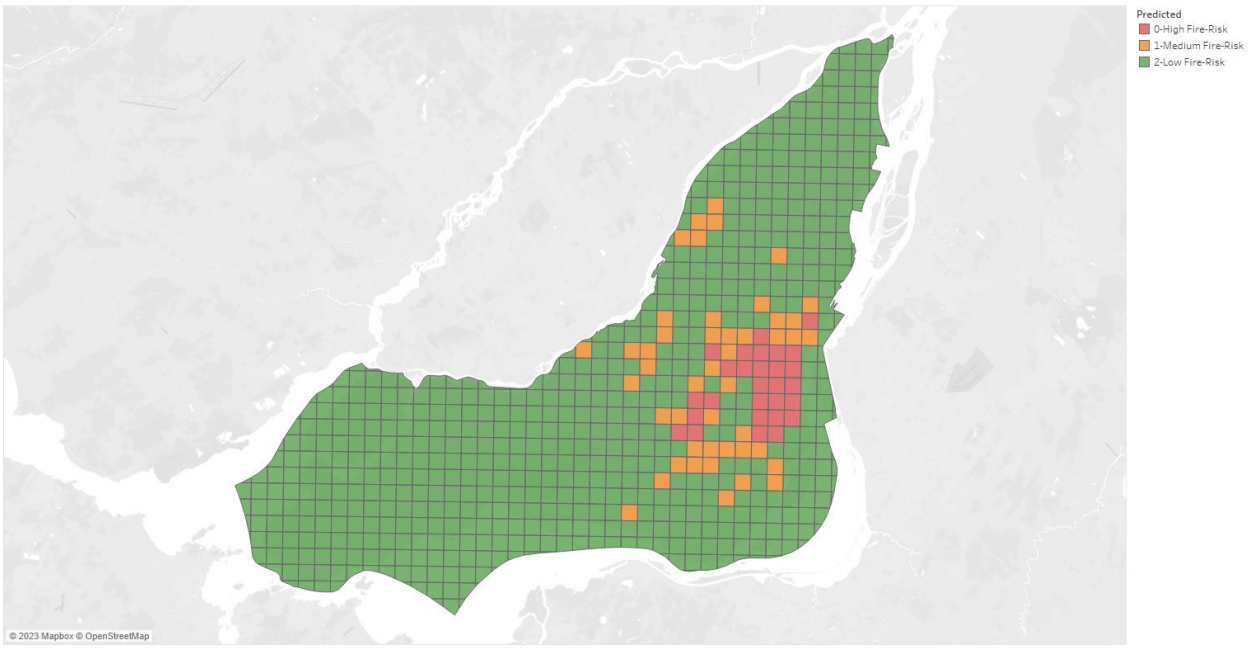

Figure 17. Predicted fire-risk levels for the city of Montreal for February 2023 using XGBoost classifier.

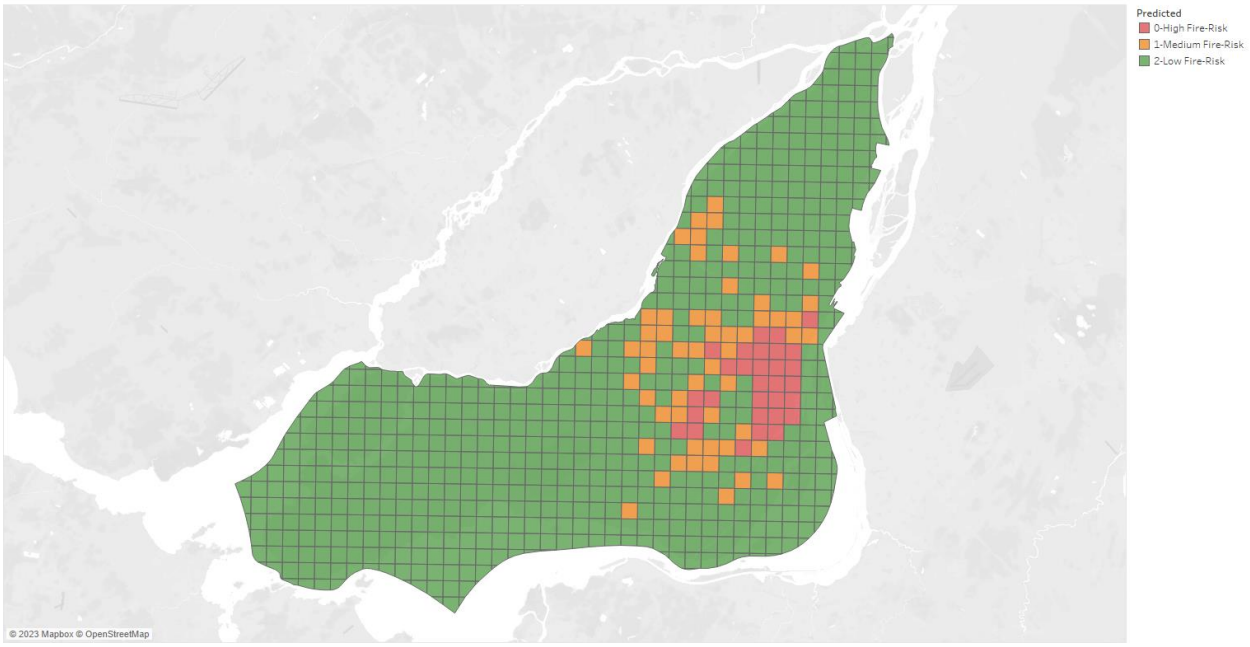

Figure 18. Predicted fire-risk levels for the city of Montreal for February 2023 using random forest classifier.

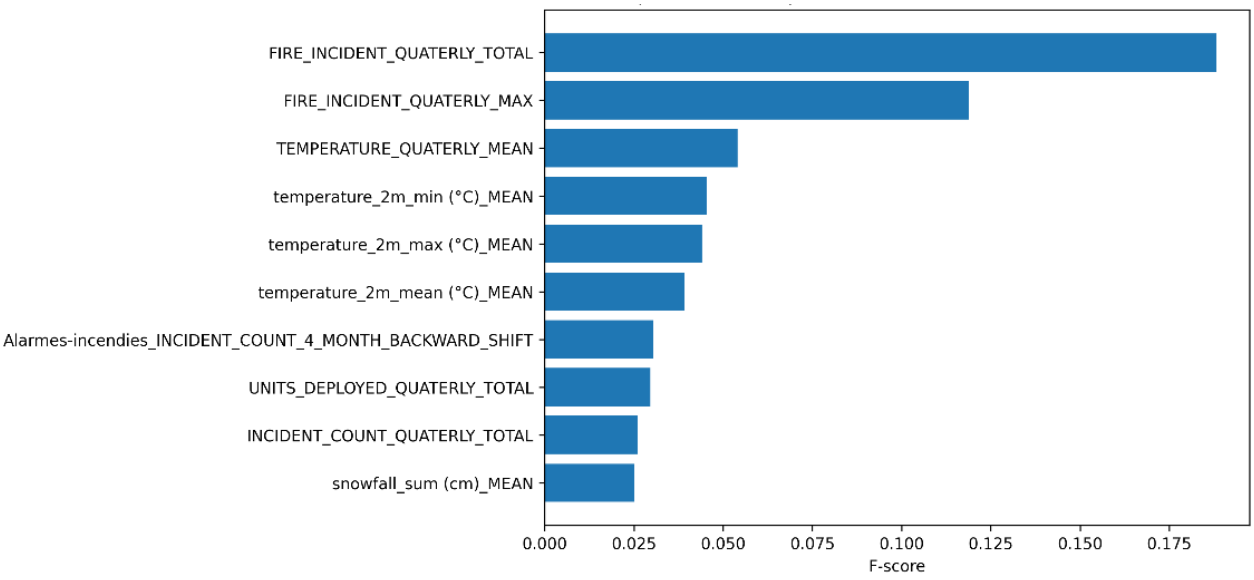

Figure 19. Top 10 features ranked by feature importance for the XGBoost classifier.

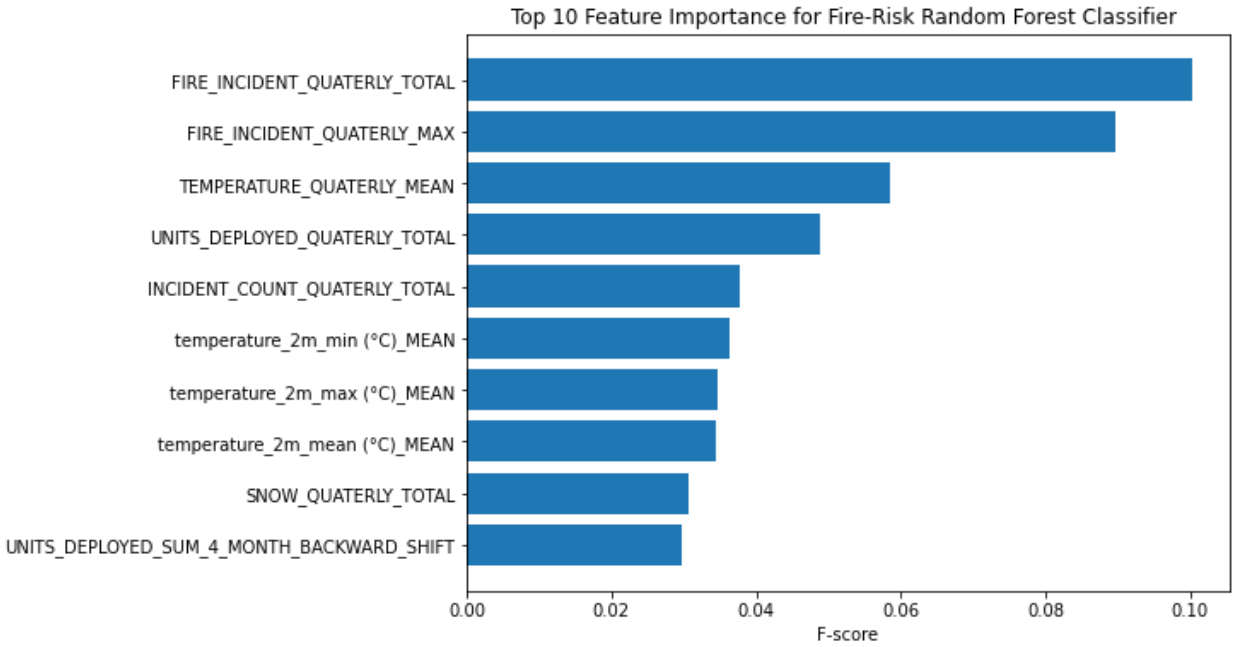

Figure 20. Top 10 features by feature importance for the random forest classifier.