Quebec Wildfire Project

December 1, 2023

The increasing wildfires in Quebec demand a clear understanding for effective strategies. This data science project explores wildfire data through analytics and machine learning to unveil crucial patterns for decision-making. By leveraging data, our objective is to contribute valuable knowledge to enhance comprehension of wildfire dynamics using weather data in Quebec. This report outlines our methodologies, findings, and implications derived from analyzing wildfire and weather data in the region.

1 Data Preparation

1.1 Gathering Data

At the project's start, we collected weather data for Quebec, handling over 14 CSV files via [1]. We consolidated the data into a single dataset named "weather data." Additionally, we incorporated wildfire data for Canada from Kaggle [2], enabling a comprehensive examination of wildfire occurrences.

1.2 Restricting Data

Our focus narrowed to wildfires in Quebec between 2011 and 2021 caused by lightning ('L'). Duplicates in two locations were identified and addressed. Geographical boundaries were refined, removing weather stations falling outside specified coordinates.

1.3 Preparing Data

Building upon our commitment to a comprehensive analysis, we introduced the "Location" column in the wildfire dataset using already known latitude and longitude of the fire. This addition serves a dual purpose—enabling us to categorize each fire based on its geographical coordinates and associating each fire with the nearest weather station. Our next endeavor involved the matching of each fire with its closest weather station. To achieve this, we leveraged the spatial.KDTree() functionality [3].

When the task of matching each fire to its respective weather station accomplished, the next decision point centered around establishing a temporal window for predictive modeling. Drawing inspiration from a relevant article [4], we opted for a 15-day period preceding a fire event. To facilitate predictive modeling, we replicated each row 15 times, aligning with the chosen temporal window of 15 days preceding a fire.

Adjusting the dates using the days15() function ensured a consecutive sequence within this timeframe. The merged dataframe, now featuring 15 rows for each fire with adjusted dates, was joined with the weather data corresponding to the weather stations' locations and dates.

2 Data Refinement

After merging, duplicates were eliminated, resulting in a structured format where rows at indices 0, 15, 30, and so forth represent the original fire data, while rows from indices 1 to 14, 16 to 29, etc., encompass non-fire instances. To distinguish non-fire instances, we assigned a size of 0 to their corresponding rows.

Further refinement involved addressing NaN values, particularly in columns indicative of snow and rain variables. Given that NaN values in "SNOW ON GROUND", "TOTAL RAIN", "TOTAL PRECIPITATION", and "TOTAL SNOW" signify the absence of snow or rain, we filled these with 0.

Specifically, for columns listed in the "col to be filled" list, we employed aggregation methods tailored to each variable. Mean temperature, min and max temperature were filled using mean() aggregation. For max and min relative humidity, we opted for pchip interpolation. This decision aligns with best practices outlined in relevant literature, [5], [6], acknowledging the common interpolation practices employed by weather stations to estimate missing values in these critical features.

The adoption of pchip interpolation ensures a balanced and informed approach to maintaining the integrity of our dataset, crucial for robust analyses and predictions. With the successful reduction of missing values, the refined wildfire dataset was now poised for use, but a pivotal element remained—the addition of a 'class' column.

Given that our dataset only comprised instances where fires occurred, we recognized the necessity for a counterpart non-fire dataset. To address this, we crafted a non-fire dataset, ensuring a balanced representation of both classes. With the integration of the 'class' column, distinguishing between fire and non-fire instances became explicit, marking the conclusion of the dataset preparation phase and paving the way for subsequent analyses and predictive modeling.

In the creation of the non-fire dataset, our focus shifted to locations and dates where no fires occurred. To construct this dataset, we collected weather data for the 15 days preceding the date that is 30 days before the fire. Mirroring the processes applied to the fire dataset, we undertook steps to handle missing values, impute appropriate values for snow and rain variables, and aggregate the data using tailored methods.

The subsequent task centered on aggregating the values of the 15 days preceding a fire into the corresponding fire row and applying the same aggregation methods for the non-fire dataset. The choice of aggregation method was a critical consideration, varying based on the nature of the variable. For temperature-related features, mean aggregation provided a representative value, while total rain and snow necessitated a summation approach.

3 Dataset Integration and Shuffling

Upon reducing missing values, both fire and non-fire datasets were aggregated, concatenated, and shuffled to ensure unbiased representation in subsequent analyses.

With these preparations, our dataset is poised for predictive modeling, offering nuanced insights into factors influencing wildfire occurrences in Quebec.

4 Data Modelling

An exploration of recent literature on the topic of wildfire occurrences suggests that it is not necessarily evident a priori which classification model would be best to use for our data set (c.f. [7, 8]). We consider all of the basic classification models:

- decision tree,
- random forest,
- gradient boosting,
- logistic regression,
- k-nearest neighbours,
- support vector machine,
- naïve bayes.

We perform a simple test to see if the model seems viable (using a single-train test), and proceed with a cross-validation and hyperparameter tuning if the preliminary test suggests that the model has potential.

4.1 Model Results

All of the model succeeded the preliminary test, so we followed with cross-validation hyperparameter tuning for each of them. The results of the tuned models is display in Figure 1. We can see that all of the tuned models performed very well on the test data.

	Model	Accuracy of tuned model
0	Decision Tree	0.990566
1	Random Forest	0.962264
2	Gradient Boosting	0.990566
3	Logistic Regression	0.962264
4	K-Nearest Neighbours	0.990566
5	Support Vector Machine	0.981132
6	Naïve Bayes	0.924528

Figure 1: Table of accuracy on test set with best hyperparameters.

4.2 Data Limitations

The main limiting factor in the strength of the models is that data itself. In principle the goal of our models is to make a prediction using two very non-linear meta-features: recent atmospheric conditions, and the occurrence of a lightning storm. The former being a somewhat of a measure of how combustible an area is (e.g. an area that has been rained on for many days in a row would be less likely to ignite or spread a fire than an area with no rain), and the latter being a measure of when combustion will occur (i.e. we expect rain to accompany the lightning).

A complication with trying to build the model around these meta-features is that neither one is a feature that is measured. The relevant data that we could obtain was weather data, by which atmospheric conditions would be built out of several days of data (e.g. considering a weeks worth of weather conditions), and the presence of a thunderstorms utilizes the weather data of a single day (i.e. does the data suggest a storm). In practice, we made a decision to consider focus mainly on recent atmospheric conditions, by aggregating our weather day over a span of days. A more robust data set would perhaps consider collections of individual days rather than a single aggregate. This should make the importance on "recent" more clear (i.e. how many days back are useful), and give single-day data with which to determine thunderstorms.

4.3 Feature Importance

We investigate how various models (the tree models and logistic regression) weigh the different features, especially as compared to the correlations among the features themselves (see Figure 2). This gives both an insight into the models themselves, and allows us to make sense of how the models value them as compared to what we would expect.

There are several conclusions that we can draw from this information, regarding both the features and the models.

All of the above models rank total precipitation as most important, and notably having a positive correlation with class 1 (fire). This aligns with the correlations between class and the other features (as seen in the heat map), but is a interesting result given how the data was created. From the setting of the problem, this aligns with what we would expect, as we expect lightning (the cause of the fire) to be accompanied by some form of precipitation, however we built our data set to be aggregated data (i.e. recent atmospheric data). Built this way, we might expect that the correlation would be negative, as precipitation would suggest the conditions leading up to the fire were wet. This might suggest that precipitation on a given day is far more correlated to a fire than recent precipitation is to a non-fire. This kind of idea could be addressed more clearly in the "more robust data set" mentioned above.

All three tree models weigh the features in a similar order, with total precipitation, relative humidity and minimum temperature being the top three. This also showcases some of the failings of the decision

```
Feature Importance from the Decision Tree Model
                Feature  Importance
4  TOTAL_PRECIPITATION    0.913215
0     MAX_REL_HUMIDITY    0.057185
6      MIN_TEMPERATURE    0.029600
1     MEAN_TEMPERATURE    0.000000
2     MIN_REL_HUMIDITY    0.000000
3       SPEED_MAX_GUST    0.000000
5      MAX_TEMPERATURE    0.000000


 Feature Importance from the Random Forest Model
                Feature  Importance
4  TOTAL_PRECIPITATION    0.388423
0     MAX_REL_HUMIDITY    0.333690
6      MIN_TEMPERATURE    0.089818
5      MAX_TEMPERATURE    0.078998
2     MIN_REL_HUMIDITY    0.058029
1     MEAN_TEMPERATURE    0.047533
3       SPEED_MAX_GUST    0.003509


 Feature Importance from the Gradient Boosting Model
                Feature  Importance
4  TOTAL_PRECIPITATION    0.904993
0     MAX_REL_HUMIDITY    0.058354
6      MIN_TEMPERATURE    0.030315
3       SPEED_MAX_GUST    0.002575
5      MAX_TEMPERATURE    0.001992
2     MIN_REL_HUMIDITY    0.001733
1     MEAN_TEMPERATURE    0.000038


 Feature Importance from the Logistic Regression Model
 Feature Importance (Coefficients):
TOTAL_PRECIPITATION: 4.45053187413822
MEAN_TEMPERATURE: 0.7803340474848195
MIN_TEMPERATURE: 0.3773134618473127
MAX_TEMPERATURE: 0.3334226291456579
MIN_REL_HUMIDITY: 0.05676628630121049
MAX_REL_HUMIDITY: -0.3728939629111549
SPEED_MAX_GUST: -0.9695765769096939
```

Figure 2: Feature importance for decision tree, random forest, gradient boosting, and logistic regression.

tree model. While it does perform very well and would be faster than the other tree models given a larger data set, the tuned decision tree model only considers three of the features.

References

[1] https://climatedata.ca/download/station-download

[2] https://www.kaggle.com/datasets/ulasozdemir/wildfires-in-canada-19502021/data 1

[3] https://docs.scipy.org/doc/scipy/reference/generate/scipy.spatial.KDTree.html

[4] R.Bu, Y.Chang, H.Chen, Z.Zhu, Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China (2013), Landscape Ecology 28(10), DOI:10.1007/s10980 − 013 − 9935 − 4. 1

[5] P.Kourtzand, B.Todd, Predicting the Daily Occurence of Lightning-Caused Forest Fires, Petawawa National Forestry Insitute, Information Report PI-X-112. 1

[6] B.S.Negara, R.Kurniawan, M.Z.A.Nazri, S.N.H.S.Abdullah., R.W.Saputra, A.Ismanto, Riau Forest Fire Prediction using Supervised Machine Learning, Journal of Physics: Conference Series, Volume 1566, 4th International Conference on Computing and Applied Informatics 2019 (ICCAI 2019) 26-27 November 2019, Medan, Indonesia. 1

[7] Lan, Yongcui Wang, Jinliang Hu, Wenying Kurbanov, Eldar Cole, Janine Sha, Jinming Jiao, Yuanmei. (2022). Spatial pattern prediction of forest wildfire susceptibility in Central Yunnan Province, China based on multivariate data. Natural Hazards. 116. 1-22. 10.1007/s11069-022-05689-x. 2

[8] Xiao Y, Zhang X, Ji P (2015) Modeling Forest Fire Occurrences Using Count-Data Mixed Models in Qiannan Autonomous Prefecture of Guizhou Province in China. PLOS ONE 10(3): e0120621. https://doi.org/10.1371/journal.pone.0120621 2

3