Canadian Wildfire Analysis

Dhruvin Hasmukh Makwana
110052843
makwana7@uwindsor.ca

Darshil Hirenkumar Shah
110059176
shah5p@uwindsor.ca

Mansi Hitendrakumar Brahmbhatt
110068455
brahmb51@uwindsor.ca

Mahima Vadlamudi
110069331
vadlamum@uwindsor.ca

***Abstract*— Wildfire occurrence, frequency, and behavior have altered dramatically across time and location, owing primarily to the complex factors of climate change and variation. The annual amount of forest area burned by wildland fires in Canada's northwestern boreal regions increased steadily throughout the second half of the 20th century. "Climate change during the 21st century is expected to result in more frequent fires in many boreal forests, with severe environmental and economic consequences" [1]. Understanding the future activity of wildfires is a prevalent research topic in academia. To that end, we present an exploratory analysis of historical wildfire activity seen in Canada from 1930 to 2020 using Google's BigQuery platform and Data Studio. In addition, we apply various Machine Learning approaches to estimate the approximate amount of damage caused by fire based on key indices.**

Keywords—Wildfires, BigQuery, Data Studio, SVM, KNN, XGBoost.

I. Introduction

In the 21st century, wildfires have become one of the greatest contributors to the sudden surge in greenhouse gases globally. These wildfires can wipe out entire forests and cities within days, destroying everything in their path and polluting the air with smoke thick enough to be seen from space. According to Natural Resources Canada, the number of fires this year has gone beyond the historical average, based on the past 25 years of reporting [2]. There is a growing consensus that as wildland fire activity rises, suppression efforts required will become increasingly difficult. Thus, making it critical to implement a proper system to aid in improved wildfire management. One of the most important concerns for firefighters today is preventing aggressive and catastrophic wildfires before they spread across large areas. Using technology such as big data and visualizations to analyze wildfire occurrences coupled with Machine Learning models to detect and prevent fires could save wild habitats and native wildlife and lives and billions of dollars in clean-up and reconstruction costs.

II. Related Work

Many weather-based fire behaviour prediction techniques have been proposed in literature such as the Canadian Forest Fire Weather Index [3], which uses different weather patterns such as relative humidity, precipitation, wind, etc., to give numerical numbers ranging from 0 to 30 providing an estimation of risk of wildfire. Research from Jain et al. (2020) states various traditional ML approaches to predict various wildfires metrics. However, problem domains associated with predictive analytics have seen much less work with ML methods.

III. Proposed Model

A. *WildFire Dataset*

To understand perform the analysis of wildfire activity and provide actionable insights based on various key factors we have used the dataset National Fire Database Points Data [5] provided by the Canadian Wildland Fire Information System. This data contains 2M+ records of fires locations aggregated from various fire management agencies of various provinces and parks. To handle such a large dataset we have created a data pipeline that feeds the data extracted from the MongoDB database and feed it in Google's data warehouse BigQuery. For this purpose, a specific data pipeline is built which is illustrated below.

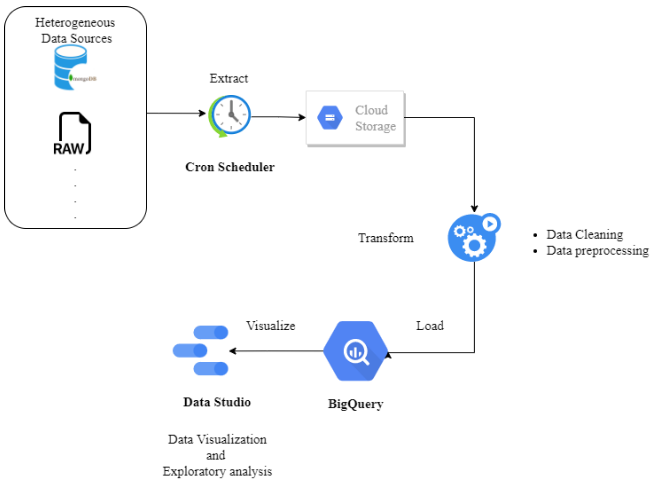


Figure 1 Data Pipeline

As illustrated in Fig 1 the data is extracted from the MongoDb dataset through an ETL bat script. This script is regularly run using a cron scheduler to first extract data from MongoDB documents. Next extracted data is uploaded to the cloud storage bucket for redundancy and backup. Data stored in Google cloud storage is then fed into the Google BigQuery where interim processing is done to pre-process and clean the incoming data. At last, cleaned data is merged with the main data table in BigQuery.

B. *Exploratory and Descriptive Analysis*

To better understand the relations between various components of the dataset Google's Data studio platform is utilized. Data studio connects with Big Query's data table and runs complex queries on very large datasets very quickly and efficiently. We have created a wildfire analysis dashboard [6] which depicts the various relations and findings from the dataset as illustrated below.

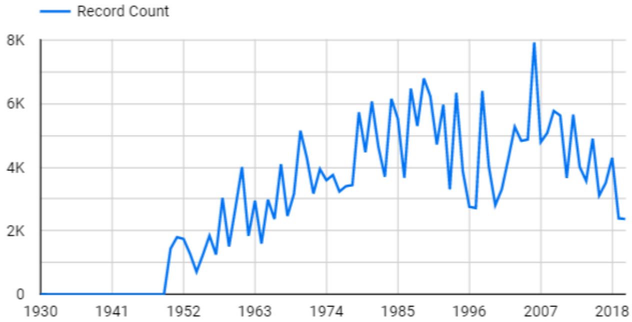


Figure 2 Wildfire trend over years

To first understand what is the trend of wildfires over years in Canada we plotted a graph of the number of wildfires in all of the provinces throughout Canada. Based on that we can see that there is an overall increasing trend in wildfire activity. However, in recent years, it can be seen that due to significant wildfire prevention techniques and resource allocation occurrences of wildfires have shown decreasing trend.

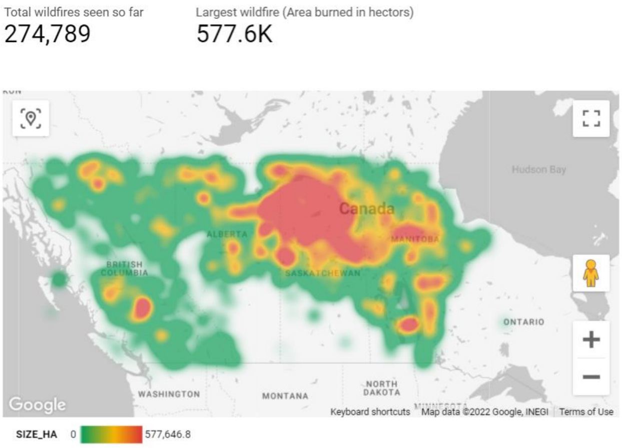


Figure 3 Wildfire activity heatmap

To find the key locations having high wildfire activity we have created a heat map of all occurrences of wildfire throughout the year. From the results shown in figure 2, it can be seen that the majority of wildfire activity is clustered in the northern Canadian regions. Following graph further analyses the provinces that have been impacted significantly among all others based on the number of wildfires reported which shows that British Columbia and Alberta have the highest wildfire activity compared to all other regions specifically.

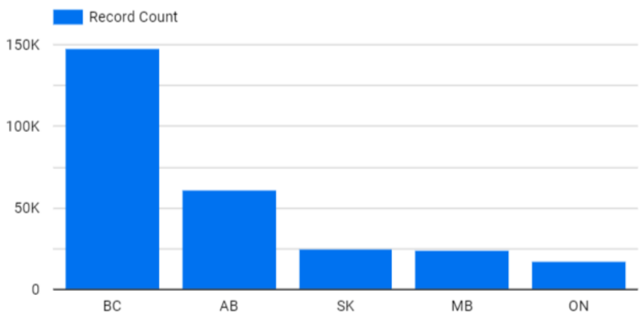


Figure 4 Frequency distribution of wildfires by provinces

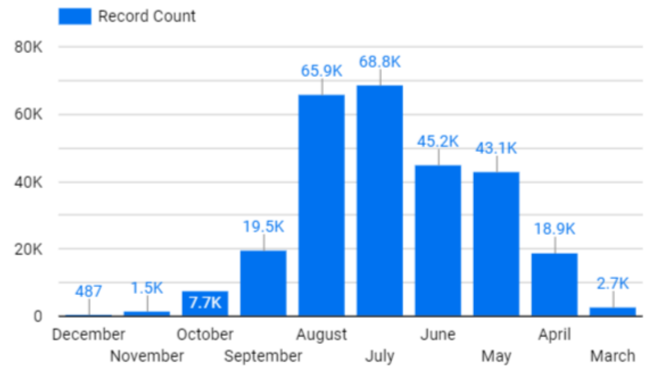


Figure 5 Frequency distribution of wildfire over calendar months

Moreover, the further analysis of data based on figure 4 shows that the months from August to may have the highest activity of wilder fires. This can be the result of high temperature and low humid weather which results in higher than normal occurrences of wildfires compared to other months.

To provide insights on how to reduce the occurrences of wildfires it is imperative to understand what are the biggest causes of the wildfires. From the dataset, we have extracted the information which describes major contributors to the wildfire in Canada.

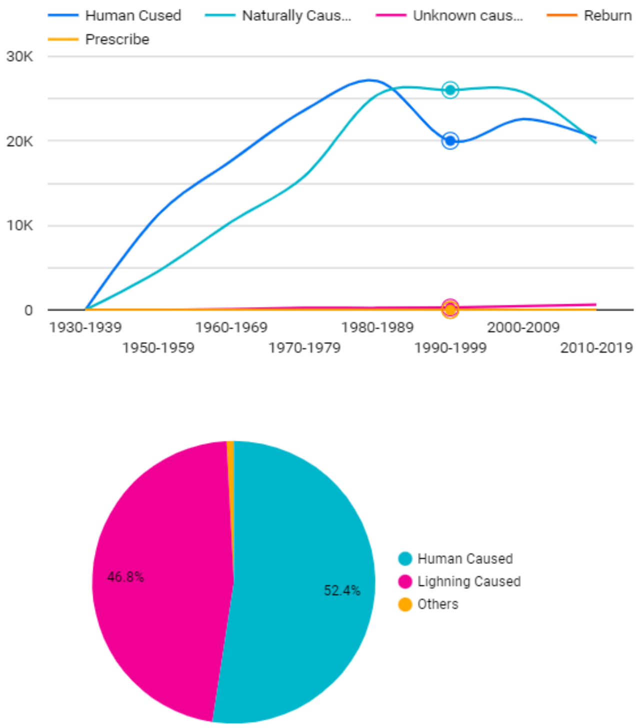


Figure 6 Causes of fires over decades

As per the graphs shown above it is evident that humans have played a major role in increased wildfire activity as out of all wildfires 52.4% of total fires can be contributed to human-related activities. Another major contributor is lightning causing 46.8% of total wildfires. This result is also supported by the graph in figure 4 which states that the majority of wildfire activity happens from august to June which is the rainy season in Canada.

C. *Predictive Analysis*

Here we have applied various machine learning algorithms to perform predictive analysis which aims to predict estimated damage caused by fire based on various key metrics. Based on exploratory and descriptive analysis performed in the previous steps we have chosen the dataset of the Alberta Smoke Plume Observation Study to perform predictive analysis. This dataset contains various key metrics related to fire and weather that can allow us to estimate the damage caused by fires. Following are the subset of features that we have used in our analysis.

Table 1 List of selected features

Feature name	Feature Description
Fire latitude	Latitude of the reported fire
Fire longitude	Longitude of the reported fire
Fire elevation	Elevation of the fire
Reported fire size	Size of fire collected from Alberta fire report; CWFIS report or MODIS Hotspot data
Temperature	Local noon temperature at fire location (as interpolated from CWFIS weather data)
Relative humidity	Local relative humidity at fire location (as interpolated from CWFIS weather data)
Wind speed	Local wind speed at fire location (as interpolated from CWFIS weather data)
Wind direction	The local wind direction at the fire location (as interpolated from CWFIS weather data)
24-hour precipitation	Local noon preceding 24-hour precipitation at fire location (as interpolated from CWFIS weather data)

To make sure the features we selected to relate to the Reported fire size we have used the Pandas corr() function to draw a correlation plot.

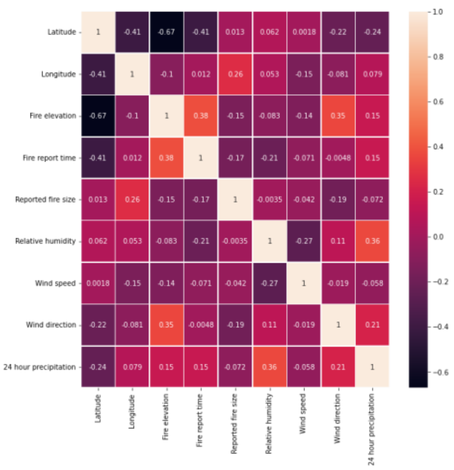


Figure 7 Correlation plot of selected features

IV. Results

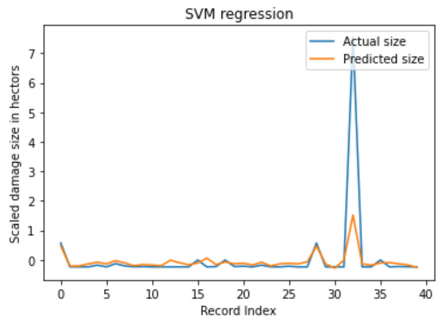
As we are planning to predict the size of damage caused by fires we have used multiple regression methods to predict the size of damage caused by wildfires which are SVM regression [7], Random Forest regression [8], Gradient Boosting regression [9] and K-Means regression [10]. To measure the performance of the model we use the mean absolute error metric of each model on the test dataset. The following table describes the mean absolute errors of each model on the dataset.

Table 2 Mean absolute error of selected models

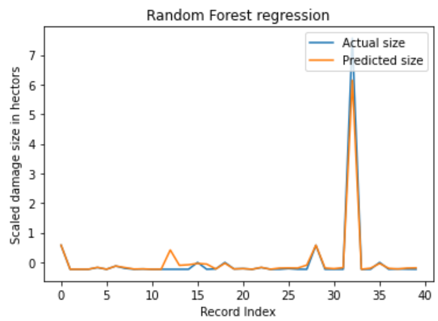
Regression Model	Mean Absolute Error
SVM	0.1552
Random forest	0.1374
Gradient Boosting	0.0110
K-Means	0.1998

Following are the result of various machine learning algorithms of the test dataset after training the models on training data.

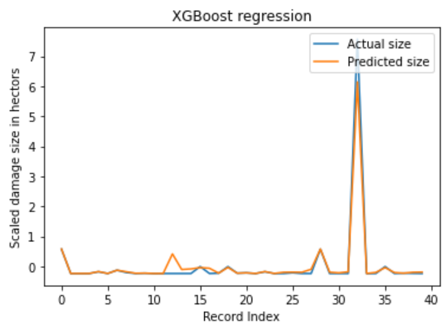
- SVM Regression



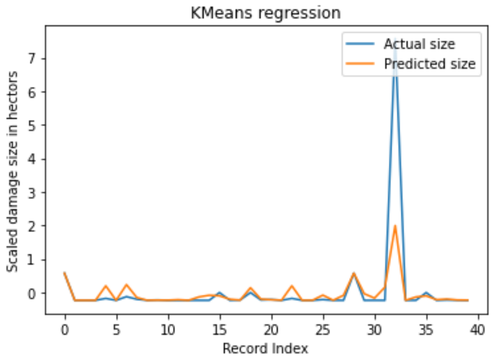
- Random forest regression



- XGBoost regression



- KMeans Regression



V. LIMITATIONS AND CHALLENGES

One of the biggest challenges we encountered was related to data cleaning and preprocessing tasks as the dataset contained various non-numeric data in numeric fields. In addition, the dataset we used only contained the data about the fire locations, date and size. It did not have any data related to weather conditions in that particular location. To solve this problem rather than focusing on the country-wide dataset we used the provincial dataset related to Alberta wildfires from 2010 to 2015.

VI. CONCLUSION AND FUTURE WORK

From our exploratory, descriptive and predictive analysis we can see that using machine learning algorithms can prove to be very useful in resource allocation and wildfire management efforts which is a very prevalent natural phenomenon in Canada. We described various machine learning algorithms and their performance in predicting the size of wildfires based on various weather patterns. From the results obtained it is evident that all models show good performance for generalized inputs however among all other models used only Gradient boosting shows the best results even with outliers.

Next, we will attempt to gather more weather data in wildfire locations by automating the data collection process from the government of Canada's climate data repository and perform further analysis on all wildfire activity seen in every province.

VII. REFERENCES

[1] Canada, N. (2020, July 07). Government of Canada. Retrieved March 13, 2022, from https://www.nrcan.gc.ca/our-natural-resources/forests/wildland-fires-insects-disturbances/climate-change-fire/13155

[2] World Health Organization. (n.d.). Wildfires. World Health Organization. Retrieved March 13, 2022, from https://www.who.int/health-topics/wildfires#tab=tab_1

[3] Anon. 1978. Canadian Forest Fire Weather Index Tables. Third Edition. Environment Canada, Forestry Service. Forestry Technical Rep. 25, Ottawa

[4] Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019

[5] Canadian Wildland Fire Information System. (n.d.). Canadian Wildland Fire Information System. https://cwfis.cfs.nrcan.gc.ca/datamart/metadata/nfdbpnt.

[6] *Wildfire analysis*. (n.d.-b). Google Data Studio. https://datastudio.google.com/reporting/132d4137-d997-4f58-8541-0ea7f6f4f176

[7] Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh (1992)

[8] MACQUEEN, J. (1967): Some methods for classification and analysis of multivariate observations. In: L.M. LeCam, J. Neyman (eds.): Proc. 5th Berkeley Symp. Math. Statist. Probab. 1965/66. Univ. of California Press, Berkeley, vol. I, 281–297.

[9] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

[10] Breiman, L. (2001). In Machine Learning (Vol. 45, Issue 1, pp. 5–32). Springer Science and Business Media LLC. https://doi.org/10.1023/a:1010933404324