Predicting Impact and Severity of Wildfires in Alberta using AI

Tuan Hiep Do*
Kevin Shi†
Tuong Phan‡
Yilei Wang§

University of Waterloo, Waterloo, Canada

February 24, 2024

*Email: th2do@uwaterloo.ca
†Email: c62shi@uwaterloo.ca
‡Email: tvphan@uwaterloo.ca
§Email: y362wang@uwaterloo.ca

1 Abstract **3**

2 Introduction **3**

3 Methodology **3**
3.1 Vulnerable FSA Regions . . . 3
3.2 Vulnerable Population . . . 3
3.3 Data Cleaning . . . 4
3.4 Feature Selection/Data Engineer . . . 4
3.5 Training, Validating, and Testing Sets . . . 4
3.6 Prediction Model . . . 4

4 Experiments and Results **4**
4.1 Vulnerable FSA Regions . . . 4
4.2 Causes of Wildfires and Impact on Environment/Residents . . . 4
4.3 Vulnerability of Population and Impacts on Indigenous Population . . . 5
4.4 Final Size Burned Prediction . . . 7
4.4.1 Model . . . 7
4.4.2 Interpretation . . . 7
4.4.3 Benchmarking/Comparison . . . 7

5 Conclusion **9**

6 References **10**

1 Abstract

In this report, we analyze the data provided by the Alberta Fire Department to find out the main causes of wildfires, assess the vulnerable population/regions, and predict the final size of burned area. The key findings are that lightning is the main cause of wildfires and the vulnerable population consists of demographics from 0 to 14 years old, 65 years old and over, and Indigenous people. Furthermore, 1b shows that the final size of area burned has extreme outliers which affect the predictive model in 4.4.1.

2 Introduction

In central Alberta, the threat of wildfires presents a major obstacle. These fires, with their unpredictable nature and varying degrees of severity, not only endanger the environment but also put communities and the economy at risk. In particular, the following important questions are considered

1. Which top FSA regions are more vulnerable with wildfires?
2. What are the main reasons causing wildfires near each vulnerable FSA region?
3. How to identify the impacts of wildfires on indigenous population and vulnerable population?
4. How to predict the severity and the final size of area burned by wildfires?

The issues addressed here are important because decision-makers in different industries such as forestry, agriculture etc. can have insights on what strategies to employ in order to preserve indigenous areas and protect life and property. As such, we have conducted a thorough data analysis to identify vulnerable FSA regions and the impacts of wildfires on indigenous population. Additionally, we also created a machine learning model to predict the severity and the final size of area burned by wildfires using the features available in the dataset.

In section 3, we describe the methodology used in data cleaning, feature engineer, and model development. In section 4, the answers to each of the questions above are provided along with visualizations and summary statistics.

3 Methodology

3.1 Vulnerable FSA Regions

The criteria that we use in defining most vulnerable FSA regions is the frequency of wildfires occurring in the region combined with the impacts of wildfires, e.g. final area burned. The areas with relatively high frequency of wildfire occurring and large area burned are considered vulnerable FSA regions. Distance from water source is another potential criteria to define vulnerable regions; however, this variate has a disproportionate amount of missing values. Thus, we exclude it from our criteria of vulnerable regions.

3.2 Vulnerable Population

The vulnerable population profile consists of the demographics from 0 to 14 years old, 65 years old and over, and Indigenous people. We derive such criteria based on Census Profile in 2021 from Statistics Canada as well as the characteristics of the demographic itself. For example, they either exhibit dependency on others for assistance in critical situations, have limited mobility, or does not have high concentration in population.

[3] shows that the demographic consisting of 0 to 14 years old and 65 years old and over contributes 33.2% to the overall population in Calgary which is quite significant. Additionally, Indigenous population only constitutes up to 3% of the overall population. We also observe a similar trend in other FSA regions.

3.3 Data Cleaning

Fortunately, most of the important features have little to no missing values. However, there are 2673 rows in which one of fire spread rate, weather conditions, temperature, relative humidity, wind direction, and wind speed is missing. These variables are important in our analysis and modeling; therefore, we have decided to exclude these rows from the dataset. The reason is that there are 22914 rows in total; therefore, dropping 2673 rows will not affect the amount of data significantly. We also noticed that the missing values in these variables are distributed evenly among each FSA region. As such, this will not create any significant imbalance in the dataset for each FSA region.

3.4 Feature Selection/Data Engineer

After conducting exploratory data analysis, the following variables are included in our final training and testing datasets: fire location latitude, fire location longitude, general cause description, fire spread rate, assessment hectares, weather conditions over fire, temperature, relative humidity, wind direction, and wind speed.

We included latitudes, longitudes, and FSA regions because as shown in 1b, the distribution of area burned varies quite significantly depending on regions. On the other hand, weather conditions, spread rate, causes of fire, and initial burned area intrinsically affect the ending result of wildfires as detailed in [1].

Finally, the typical approach in normalizing numeric values is used in our analysis, namely

$$\mathbf{x} \mapsto \frac{\mathbf{x} - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}.$$

The categorical variables are encoded using one hot encoding (indicator variables).

3.5 Training, Validating, and Testing Sets

After performing data cleaning, feature selection, and data engineering, we split the dataset into training set (60%), validating set (20%), and testing set (20%) using shuffling and random sampling.

3.6 Prediction Model

The model that we pick in this case is Extreme Gradient Boosting Regressor (XGBoost Regressor) [2]. We find that this model has a nice balance between interpretability and predictive power. Furthermore, the model provides feature importance ranking for the independent variables as well as great performance on structured data.

4 Experiments and Results

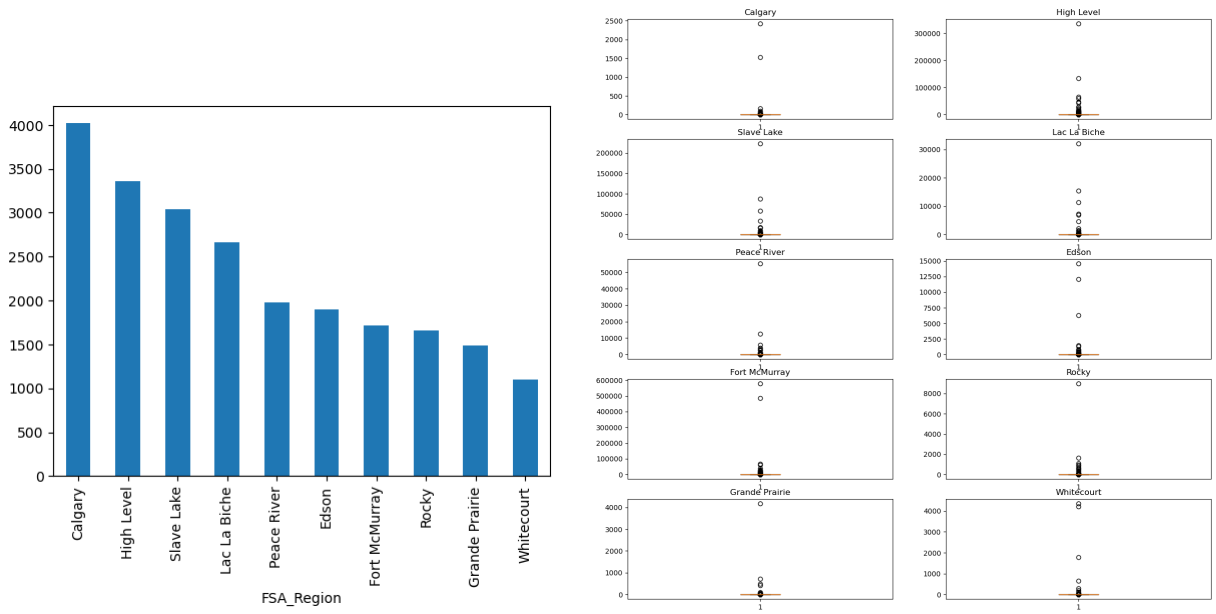
4.1 Vulnerable FSA Regions

From 1a, one sees that wildfires occur most frequently in Calgary, High Level, and Slave Lake. However, wildfires cause largest burn areas in High Level, Slave Lake, and Fort McMurray. Therefore, the most vulnerable regions based on this analysis are Calgary, High Level, Slave Lake, and Fort McMurray.

4.2 Causes of Wildfires and Impact on Environment/Residents

As illustrated in 2, lightning and recreation are the most common reasons that cause wildfires near each vulnerable FSA regions. From 4.1, wildfires causing largest burn areas occur in High Level, Slave Lake, and Fort McMurray. The most frequent cause for wildfires in these areas is lightning.

This makes sense because these vulnerable regions experience hot, dry summers with the potential for intense



(a) Number of wildfires in each region

(b) Distribution of Size Burned

Figure 1: Wildfires Frequency and Area Burned

thunderstorms. Lightning from these storms can ignite fires in dry vegetation. Also, lightning storms can result in multiple strikes over a wide area, increasing the likelihood of multiple ignition points and the rapid spread of fire. Furthermore, in forested areas, there is often an abundance of fuel in the form of trees, shrubs, and dry grasses. When lightning ignites this fuel, it can sustain large, intense wildfires.

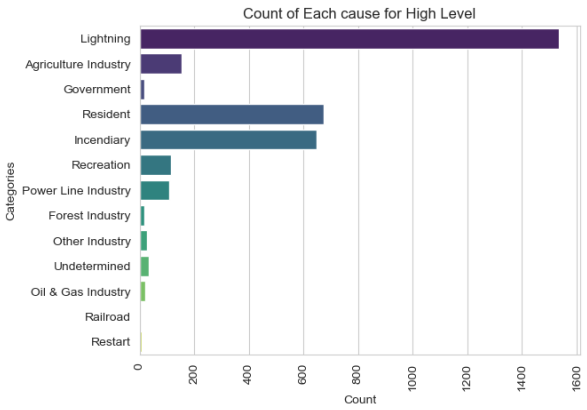
Wildfires wield profound effects on both the environment and nearby residents. They precipitate the loss of vegetation and disrupt wildlife habitats, thereby upsetting local ecosystems. These blazes can trigger soil erosion and degradation, degrade air quality, and have negative effects on water quality and aquatic life. Residents in proximity are exposed to health hazards stemming from wildfire smoke. Those residing in vulnerable areas often confront emotionally and financially taxing evacuations. Property damage or destruction due to wildfires looms as a significant worry, impacting homes and infrastructure alike. Moreover, these infernos carry substantial economic ramifications for affected communities, encompassing property loss, heightened insurance premiums, and the expenses tied to firefighting endeavors.

The immediate repercussions on both the environment and residents in close proximity to vulnerable burnt zones involve property destruction, wildlife habitat loss, and health hazards from smoke inhalation. Meanwhile, the enduring consequences have the potential to reshape ecosystems, economic landscapes, and community welfare for extended periods. Consequently, mitigation endeavors such as effective fire management techniques, community readiness initiatives, and adaptive strategies are imperative for reducing these effects and bolstering resilience against future wildfire occurrences.

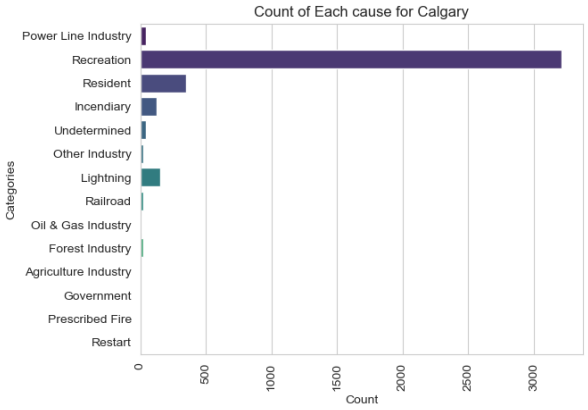
4.3 Vulnerability of Population and Impacts on Indigenous Population

As stated in 3.2, the vulnerable population consists of demographics from 0 to 14 years old, 65 years old and over, and Indigenous population. We found the following information about the four identified vulnerable regions in [3], [4], [5], [6] which are given in 3 and 4.

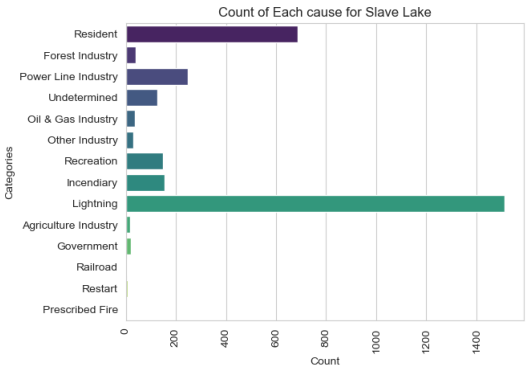
One sees that the proportion of 0 to 14 years and 65 years and over is significant in each FSA region. Moreover, the proportion of Indigenous population in the vulnerable FSA regions with large burned area (Slave Lake, High Level, Fort McMurray) is fairly high, ranging approximately from 10% to 30%. This



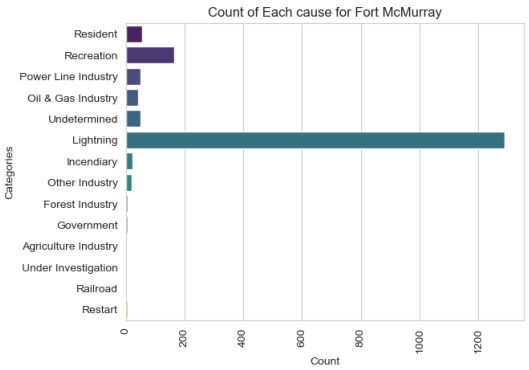
(a) High Level



(b) Calgary



(c) Slave Lake



(d) Fort McMurray

Figure 2: Causes Frequency

signifies that wildfires occuring in these areas will have the largest impacts on Indigenous Population in the following aspects: Loss of Traditional Lands and Resources, Health Effects, Increased Vulnerability to Climate Change, Environmental Degradation.

Age Distribution	Calgary	Slave Lake	High Level	Fort McMurray
0 to 14 years	18.0 %	23.5%	23.7%	23.0%
15 to 64 years	68.4%	67.2%	62.7%	72.8%
65 years and over	13.6%	9.3%	13.4%	4.2%
85 years and over	1.6%	1.1%	2.7%	0.2%

Figure 3: Age Distribution

Population	Calgary	Slave Lake	High Level	Fort McMurray
Indigenous	41,350	1,720	1,055	6,755
Total	1,306,784	6,542	3,461	68,002

Figure 4: Population Distribution

4.4 Final Size Burned Prediction

4.4.1 Model

We fit an Extreme Gradient Boosting Regressor on the training dataset, validate against the validating dataset, and test on the testing dataset. The result of training and validating can be seen from 6. The metric being used is mean absolute error to account for outliers. We have tuned the hyperparameters to get the optimal number of estimators, maximum depth, learning rate, and minimum child weight before making predictions.

Readers should note that the predicted values (area burned) have been scaled as described in 3.4 to make the algorithm converge faster. The validating errors are only slightly higher than the training errors (difference is factor of 0.001). Moreover, the scatterplot of predicted values and actual values 5 for test set clusters near 0 on the the identity line $y = x$ with outliers surrounding. The mean absolute error between predictions of testing data and actual values is approximately 0.00037.

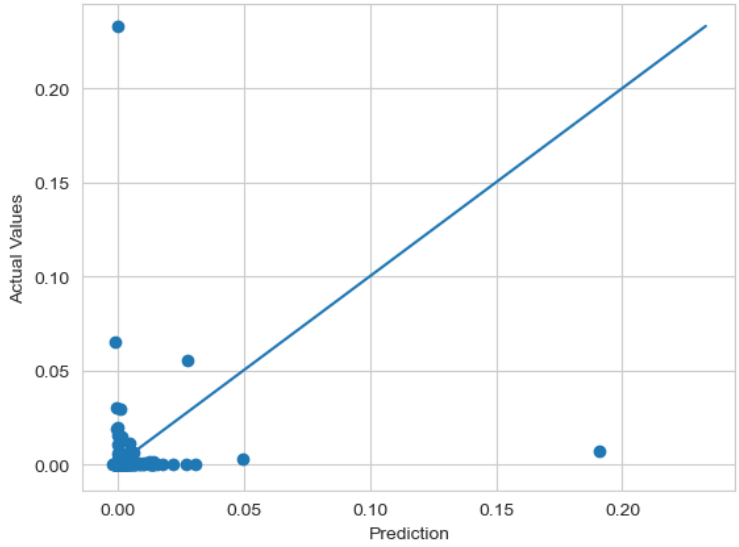


Figure 5: Test Set vs Predictions

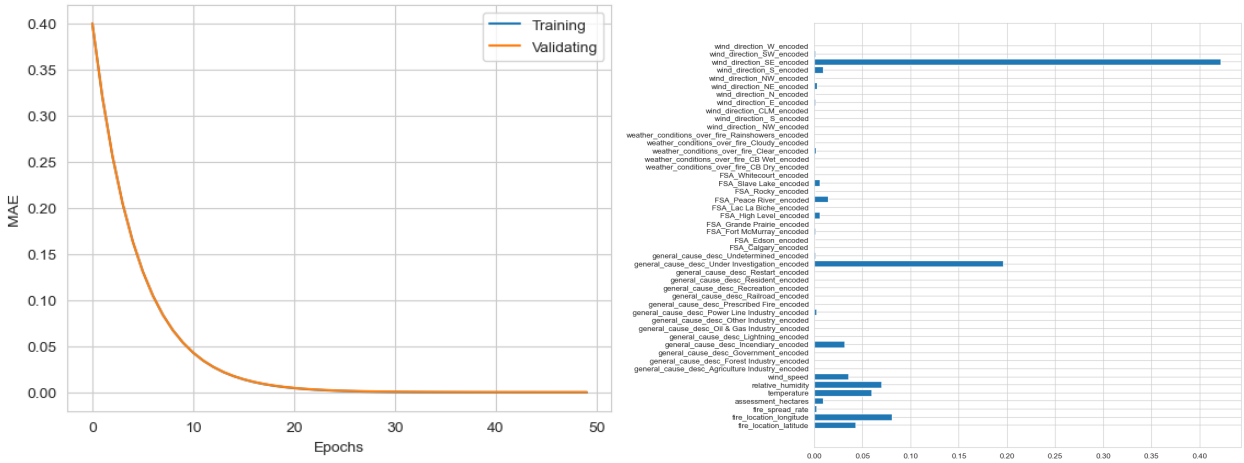
4.4.2 Interpretation

The clustering makes sense because as shown in 1b, the size of area burned has outliers in the most vulnerable regions (Slave Lake, High Level, Fort McMurray). However, most of the points are close to the identity line $y = x$.

6b shows the ranking of feature importance provided by the algorithm. We see that the variables wind direction from southeast to northeast are quite important. In [7], the government of Alberta confirms that wind from southeast to northeast can cause extreme fire behaviour. This aligns with the output of the model. The causes of the fires along with the weather conditions also play a role in predicting final size of burned area as expected.

4.4.3 Benchmarking/Comparison

In addition to the analysis and modeling above using Extreme Gradient Boosting, our team has also tried approaching the prediction task using Deep Learning. The architecture that we use consists of a dense layer of 16 neurons with ReLu activation, a dense layer of 8 neurons with ReLu activation, and a final dense layer with 1 output neuron and linear activation. The Adam optimizer is used in this case with loss being



(a) Training and Validating Losses

(b) Feature Importance

Figure 6: Losses and Feature Importance

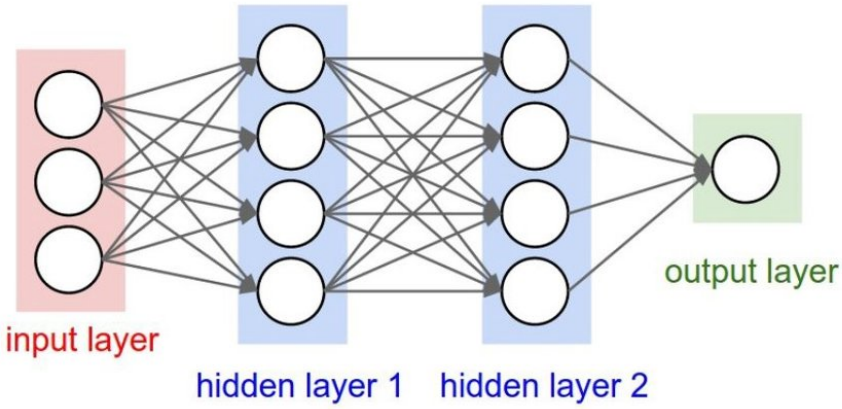


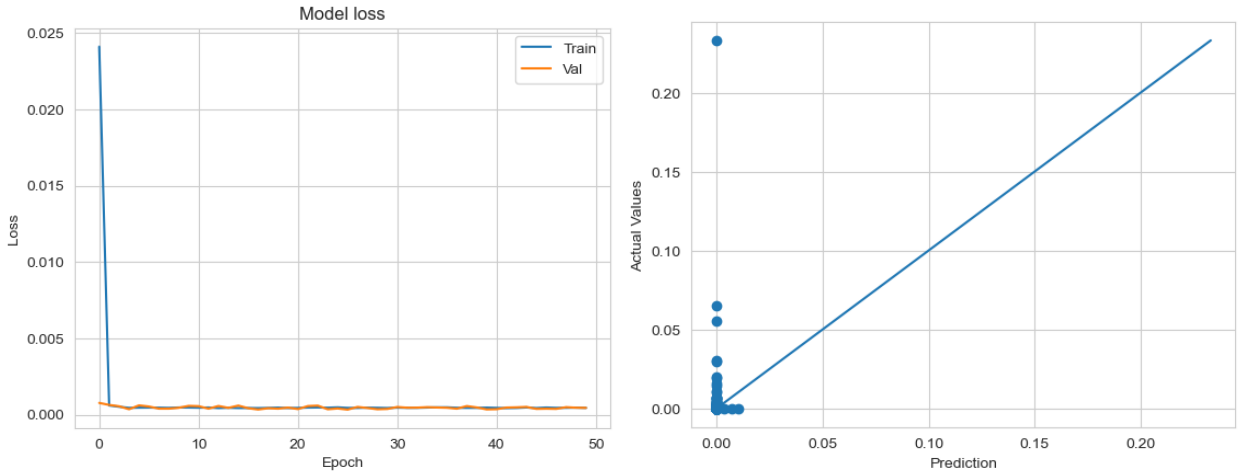
Figure 7: Neural Network Architecture

measured by mean absolute error to account for outliers. A visualization of such architecture is given in 7.

The training loss and validating loss are given in 8a and a scatterplot of predictions vs actual values on test sets is given in 8b. The mean absolute value error between the predicted values and the actual values on the test set is approximately 0.00028 which is about 22.85% lower than what we got using Extreme Gradient Boosting. However, interpretability is lost using this approach since we cannot rank which features contribute the most to the predictive task. As such, one cannot make a detailed interpretation as in 4.4.2.

Our team thinks this is a fair compromise since the mean absolute value error between predicted values and actual values on test set is still relatively low for the approach using Extreme Gradient Boosting while having nice interpretable results. This can be important to decision-makers in determining what actions to take to prevent disastrous wildfires.

Nevertheless, if the only important task is predicting the final size of area burned, then one can use a state-of-the-art approach such as neural network to make predictions. The clustering of points around the identity line $y = x$ close to 0 in 8b signifies that the predictions are fairly accurate.



(a) Training and Validating Losses

(b) Test Set vs Predictions

Figure 8: Losses and Feature Importance

5 Conclusion

In conclusion, we have addressed the main causes of wildfires, vulnerability of population/regions, and made predictions for the final size of burned area in FSA regions.

The report began by identifying vulnerable regions in central Alberta based on the frequency of wildfires and the size of the area burned. These regions included Calgary, High Level, Slave Lake, and Fort McMurray. For assessing vulnerable populations, our team considered demographics such as ages 0-14 and 65 and over, as well as Indigenous populations, due to their susceptibility to wildfire impacts. The report detailed steps in data cleaning, engineering, and model development while providing interpretations and comparisons of different models for business insights.

The issues addressed in the report are important because they allow decision-makers to answer the following important questions

1. Which top FSA regions are more vulnerable with wildfires?
2. What are the main reasons causing wildfires near each vulnerable FSA region?
3. How to identify the impacts of wildfires on indigenous population and vulnerable population?
4. How to predict the severity and the final size of area burned by wildfires?

Potential improvements for the methodologies used in this report could be finding better criteria to define vulnerable FSA regions/populations, coming up with a better algorithm that balances interpretability and predictive power, or performing more comprehensive feature engineer etc.

6 References

References

[1] Change, E. and C. (n.d.). Wildfire science. https://www.gov.nt.ca/ecc/en/services/wildfire-operations/wildfire-science#:~:text=The%20primary%20factors%20that%20influence,in%20the%20area%20is%20like)

[2] Chen, T., & Guestrin, C. (2016, June 10). *XGBoost: A scalable tree boosting system.* arXiv.org. https://arxiv.org/abs/1603.02754

[3] Government of Canada, S. C. (2023, February 1). *Census profile, 2021 census of Population profile Table.* Profile table, Census Profile, 2021 Census of Population - Calgary, City (CY) [Census subdivision], Alberta. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&GENDERlist=1&STATISTIClist=1&HEADERlist=0&DGUIDlist=2021A00054806016&SearchText=calgary

[4] Government of Canada, S. C. (2023b, February 1). Census profile, 2021 census of Population profile Table. Profile table, Census Profile, 2021 Census of Population - Slave Lake [Population centre], Alberta. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&SearchText=Slave+Lake&DGUIDlist=2021S05100764&GENDERlist=1%2C2%2C3&STATISTIClist=1&HEADERlist=0

[5] Government of Canada, S. C. (2023b, February 1). *Census profile, 2021 census of Population profile Table.* Profile table, Census Profile, 2021 Census of Population - High Level [Population centre], Alberta. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&SearchText=High+Level&DGUIDlist=2021S05100371&GENDERlist=1%2C2%2C3&STATISTIClist=1&HEADERlist=0

[6] Government of Canada, S. C. (2023b, February 1). *Census profile, 2021 census of Population profile Table.* Profile table, Census Profile, 2021 Census of Population - Fort McMurray [Population centre], Alberta. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&SearchText=Fort+McMurray&DGUIDlist=2021S05100292&GENDERlist=1%2C2%2C3&STATISTIClist=1&HEADERlist=0

[7] Ostendorf, V. (n.d.). *High level forest area wildfire Update - May 17, 2023 at 9:00 a.m.* High Level Forest Area Wildfire Update - May 17, 2023 at 9:00 a.m. https://srd.web.alberta.ca/high-level-area-update/may-13-2023-0