Ernst & Young - Project Report

Authors: Dhruv Raval, Aditya Pandya, Kayaan Tharani, and Aarnav Noble

Submission Date: February 23rd, 2024

Abstract

In response to Alberta's pressing wildfire issues, this project utilizes data analysis and machine learning with the objective of predicting wildfire severity and area burned. By reading the data from an extensive datasheet, the team's methodology involves establishing a clear vulnerability criteria and machine learning model implementation. The team aims to draw strong conclusions, potentially aiding with resource allocation and evacuation strategies. This report consists of our methodologies and key findings, as shown through our experiments and results, hopefully being of value to help combat this serious issue.

Introduction

Through learning about Alberta's escalating wildfire challenges, the team's initiative, led by Ernst & Young, is to integrate data science and machine learning techniques to combat the often unpredictable nature of wildfires. This project aims to provide an innovative strategy towards enhancing wildfire prediction and safeguarding nature, communities, and the economy. The team hopes to leverage the capabilities of machine learning to predict the severity and final burned area alongside displaying extensive data in a visually insightful way that improves clarity. The machine learning model aims to provide accurate predictions, be a valuable asset for effective resource allocation, and aid with evacuation efforts. Delving into this complex task requires a strong methodology that aims to find subtle correlations within overwhelming data to mitigate the wildfire's impact. This introduction provides a comprehensive overview, aiming to shed light on this dire situation and provide our aim for effective decision-making. In this report, our team aims to present our clear methodology, share important discoveries and insights, and address the dangerous challenges posed by wildfires in Alberta.

Methodology

Data Integration

Using the wildfire dataset provided to us, the team created a dataframe by reading the data from the excel file, electing to use Pandas library and numpy for data processing, as well as matplotlib for graphing. The team then created multiple dictionaries to help categorize and access specific parts of the data frame easily, filling the dictionary with key values such as total fires, with it being further categorized by its size class to differentiate the approximate area covered by each fire. A similar approach was taken with wildfire causes, categorizing them into general, activity-related, and true causes for each FSA region.

When the wildfire causes graphs were created, the team noticed that "activity_class" and "true_class" had many empty spaces for those categories, causing NaN values to take up a significant amount of space in each stacked bar graph. As such, some data preprocessing was done to remove the NaN values altogether, making the graphs visually clear to the viewer and more representative of the dataset's values.

Criteria for Most Vulnerable FSA

To derive the most vulnerable FSA region, the team started by looking through the dataset's categories to find ones that relate to vulnerability. After discussing, 'size_class', 'current_size', and the total number of wildfires were selected. 'Size_class' displayed the severity of each fire, with this severity being quantified in "current_size". The total number of wildfires displayed the likelihood/frequency of a fire, contributing to how vulnerable an FSA region would be. Due to the sheer amount of data, our team decided to create various stacked bar graphs and other visuals to help find the most vulnerable FSA, plotting regions versus frequency of fires, as well as regions versus average size burned. This criteria combines the likelihood and capability of damage capable by a wildfire.

When categorizing the data in a dictionary, the team noticed that Calgary had the highest number of wildfires with no other region coming close. On the other hand, places like High Level, Fort McMurray, and Slave Lake had much greater Class C, D, and E wildfires, making those places more vulnerable to wildfires. As such, a dictionary was made to take the weighted average of each FSA region, giving a more accurate representation of the region's vulnerability. This was done by dividing the total area burned by the total number of fires, giving greater priority to the damage caused by each fire and less priority to the total number of fires.

Approach to Finding Main Reasons for Wildfires:

To analyze the damage and severity caused by wildfires in vulnerable FSA regions, the team isolated 3 metrics: "general_cause_desc", "activity_class", and "true_cause". These three categories stored data on how each fire started, which when comparing the distributions for these categories, allowed the team to hone in on key reasons that caused severe and frequent wildfires. To maintain consistency by showcasing both frequency and severity, the team used both "size_class" and the total number of fires for each category

Deriving the Vulnerable Population Profile:

When assessing the impact of wildfires on Indigenous people, the team ended up going for two different approaches, one using just the information provided in the dataset and the other one using data from Statistics Canada. The team accumulated the number of Indigenous people and total number of people in a table (fig. 1) that used to show the average vulnerability in Indigenous vs non Indigenous places. Considering Grande Prairie consists of a city and a town and Rocky consists of a municipal district and a town, the team found the sum and made that the number used in our data.

Fig.1: Statistics Canada Table that shows the Indigenous and Total population in each FSA region.

Area	Total Population	Indigenous Population
Calgary	1,291,770	41,355
Edson	8,130	865
High Level	3,550	1,250
Grande	24,170	9,425

Prairie +		
Lac La Biche	7,565	1,915
Fort McMurray	330	310
Peace River	6,465	1,200
Rocky +	46,976	2,005
Slave Lake	6,660	1,780
Whitecourt	9,855	1,260

Prediction Models:

Considering the goal of predicting the severity and final size burned for wildfires, the team decided to use "current_size" as the target variable (y). "Current_size" accounts for the final burned area, and can be translated into severity classifications as well - as seen in the "size_class" label. The ranges provided for each class can be used to process the final size burned into severity, thus completing all the goals of the model. The team believed that a regression model would be simpler and more robust for future analysis and improvement, compared to a classification model for this dataset.

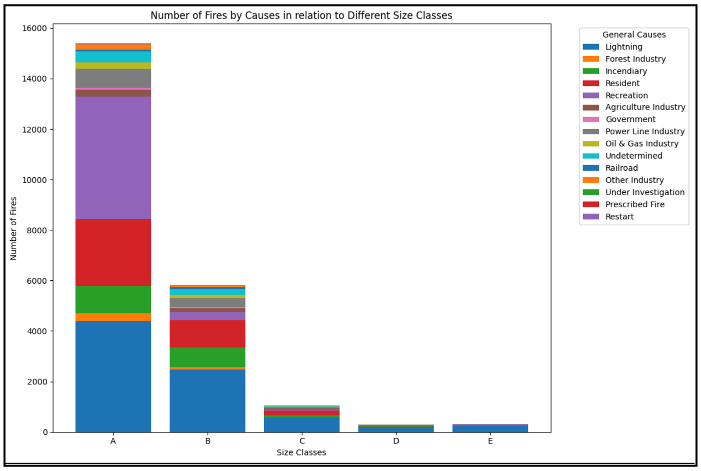
Final burned size being a numerical value meant the team was tackling a regression problem. As such, all the input variables had to be numerical and not contain any NaN values, as that would disrupt the fit of a regression model.

The team decided to use linear and ridge regression if the data were to be considered linearly related. Otherwise, the team would use kernel ridge regression, SVM regression and Random Forest Regression for non-linear relations. These models were selected for their adaptability to large datasets and relatively straightforward approach.

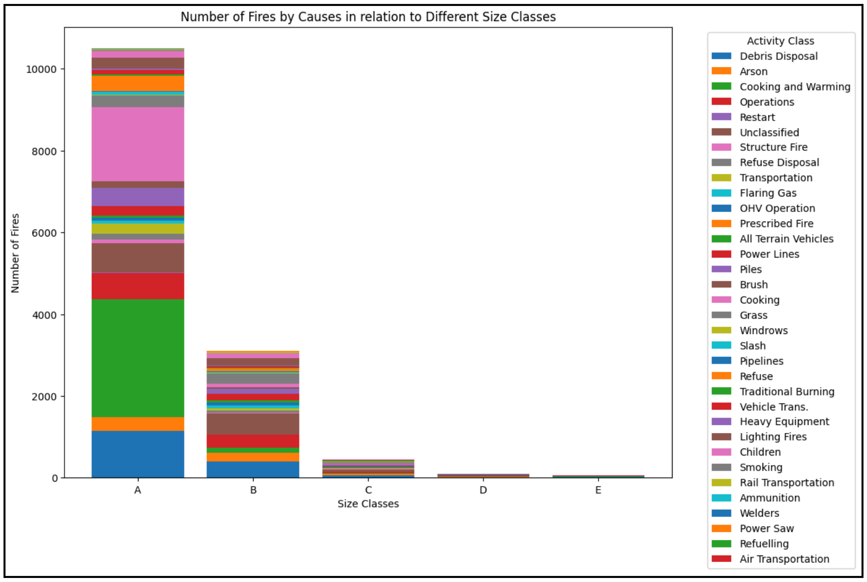
Experiments & Results:

Visual Representation of Wildfire Causes (Size Classes)

Recreation is one of the biggest causes of wildfires that caused a small burned area, closely accompanied by lightning and residents. This indicates that many small wildfires are linked to residents and recreational activities. For wildfires with a slightly larger final area burned (Class B and C), lightning, residents, and incendiaries are the primary causes. Residents contributing to such fires might imply that people must be educated about wildfire safety to avoid risks of moderat wildfires.

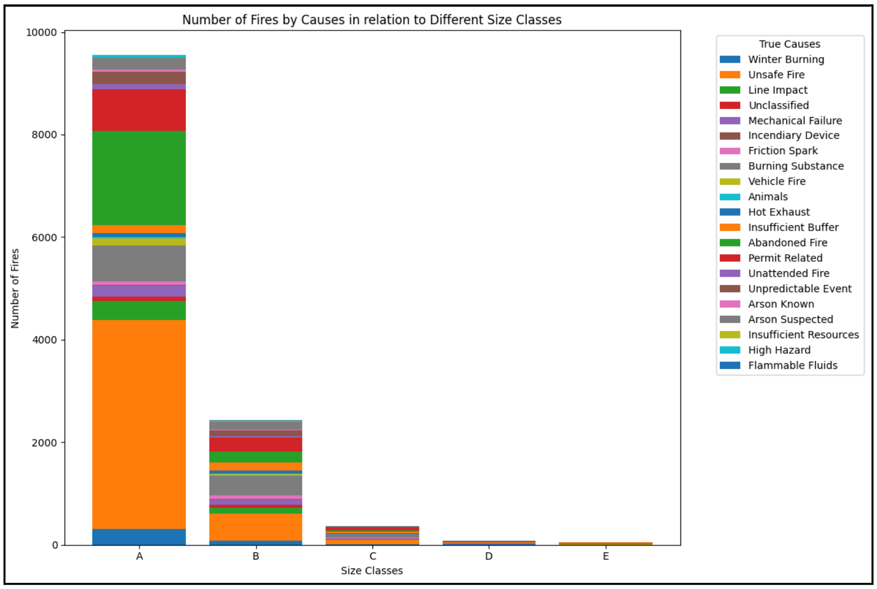


Lightning appears to be a significant factor as the burned area starts to increase, raising concern about the destructive impact of natural factors. Industries such as oil and gas, forest, agriculture, and power line contribute to the amount of relatively smaller fires (Class A and B). This could be due to quicker detection, containment, and suppression efforts for fires associated with human activities in industrial settings.



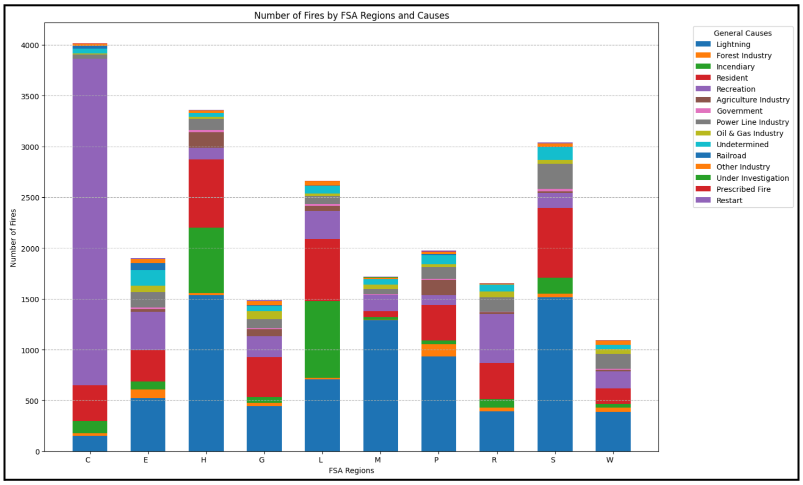
The majority of Class A fires were caused by cooking and warming, and debris, with Class B showing a similar trend but with less significance. The high number of fires caused by these activities highlight the importance of preventative measures, with resource allocation being better focused to prevent these activities and in-turn minimize damage and lost resources. The number of fires drastically decreased in Class C, D, and E, with various causes having similar amounts of fires in similar distribution.

The true causes for the majority of Class A fires were unsafe fires, with abandoned fires and permit related fires contributing a decent amount. As for Class B, the true causes were considerably more even, with unsafe fires, burning substances, abandoned fires and permit related fires being notable causes. With the remaining classes, the amount of wildfires drop considerably, with unsafe fires and burning substances having the highest amounts.

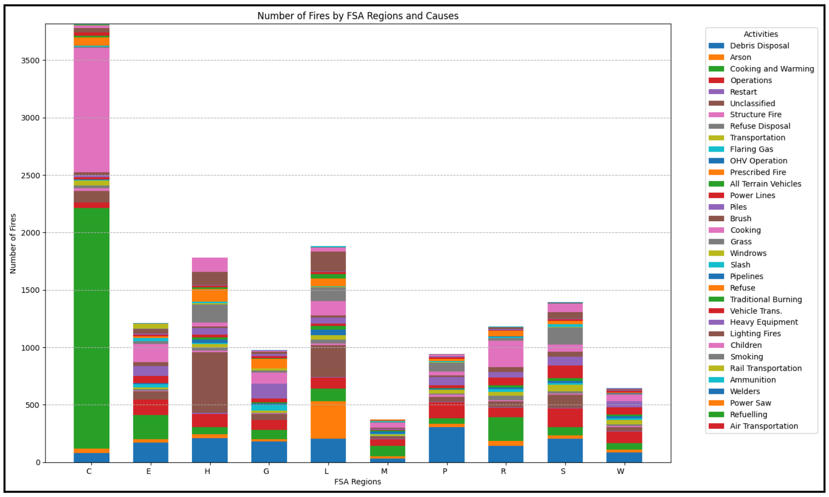


Through these graphs, a helpful takeaway would be to focus on preventative measures for both abandoned fires and burning substances, as this would help maintain these fires before they cause excessive damage. This is especially the case for burning substances as the drop for Class A to B is not as much in comparison to the other causes, with it being a prominent cause even in more dangerous classes.

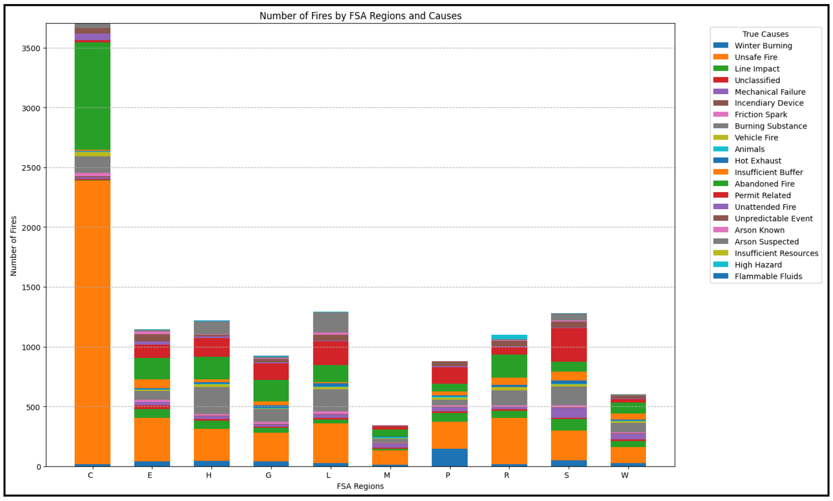
Visual Representation of Wildfire Causes (FSA Regions)



Calgary experiences a significant number of fires through recreation, with other places like High Level, Lac La Biche, and Slave Lake experiencing wildfires by residents. Due to the high number of human-caused wildfires in Alberta, it is important to educate the public and design targeted public awareness campaigns to ensure the numbers are not as high. From the previous graphs, the team can see that recreation and residents rarely caused fires in the C, D, and E Classes, meaning that the fires were extinguished before covering a large area which is certainly a positive. The impact of lightning throughout all FSA regions seems to be prominent, especially in High Level, Fort McMurray, and Slave Lake. Since this natural ignition source is impossible to reduce, a way to help reduce the impact would be to allocate more resources in these regions to help extinguish these fires before they spread rapidly and cause more damage.



Calgary experiences a significant number of fires through cooking and warming, with other places having a more dispersed list of activities that caused wildfires. This, once again, brings up the concerning actions of humans, reinforcing the importance of educating the public through awareness campaigns. Additionally, Debris disposal having a consistent impact in all places except Fort McMurray could be from poor debris disposal methods and must be investigated further.

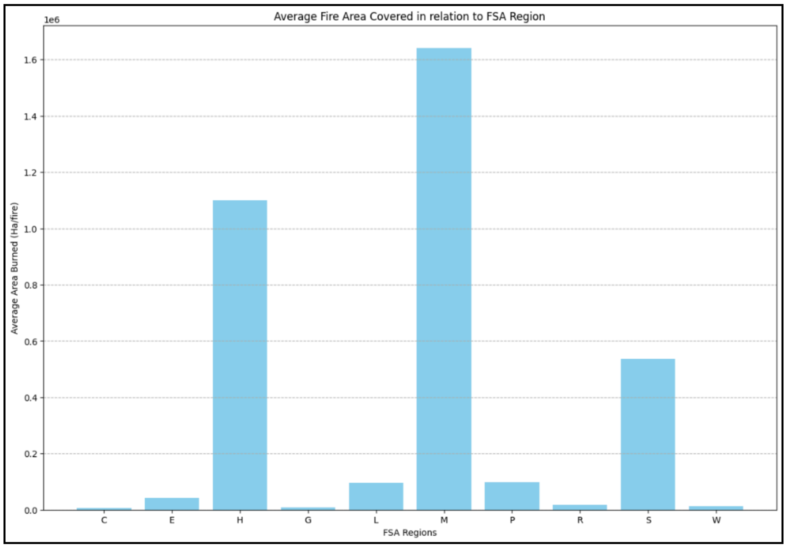


The impact of unsafe and abandoned fires are evident, being prominent causes in every region, especially in Calgary, causing approximately 88% of the wildfires. To address these dangerous issues, it is important, if not necessary, to raise public awareness by emphasizing the importance of following safe burning practices amongst the community and strengthening enforcement of burning regulations to ensure compliance with strict requirements.

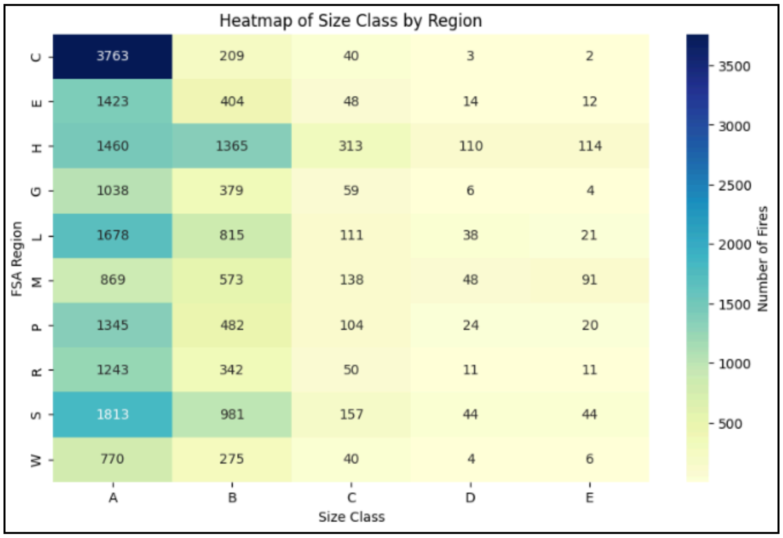
In addition, burning substances and permit related activities caused fires that were considerable in all regions, reestablishing the importance of adhering to permit conditions and making them stricter if need be.

Weighted Average Visual Representation

To show the average area covered by a wildfire in each FSA Region graphically, the team decided to find the average area covered by each wildfire in each FSA Region, accounting for the total number of wildfires. By dividing the total area burned by the total number of wildfires in each region, places with multiple small scale fires will not be as big as places with large scale fires. As shown by the graph here, the average area burned in Calgary is almost 0 whereas places like High Level, Fort McMurray, and Slave Lake are way higher.

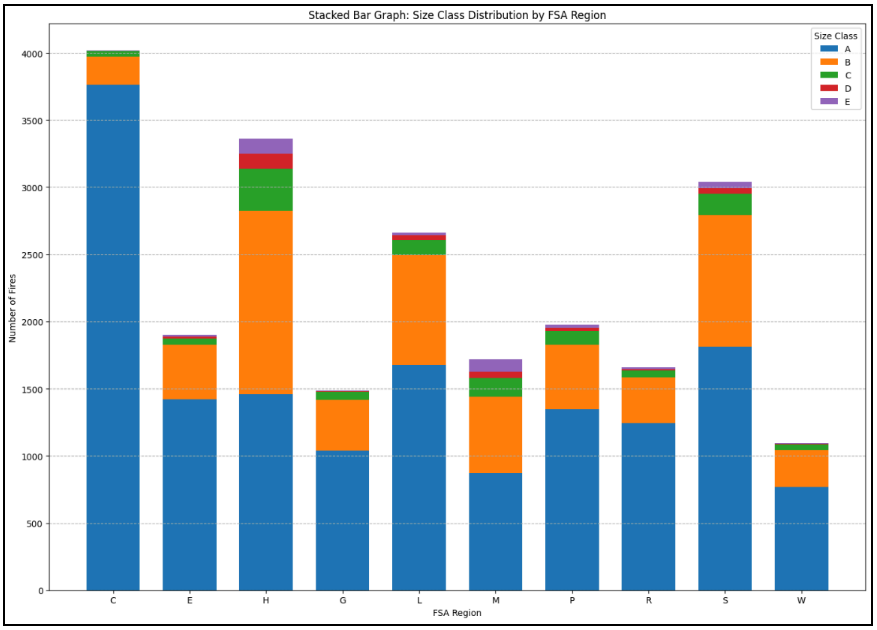


Size Class by Region - Visual Representation



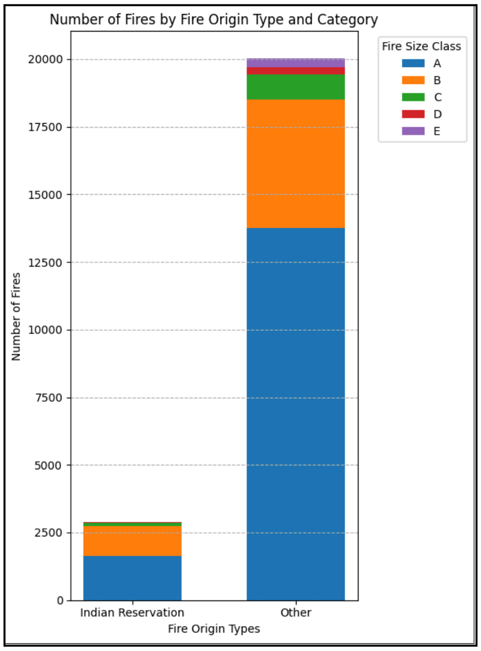
This heatmap system shows the number of fires in each FSA region alongside its size class to show both the frequency and size, giving its vulnerability. This provides more specific values, making it easier to find connections. From the heatmap, it is clear that Calgary has the most fires, with almost all of them being wildfires that covered only 0 - 0.1 ha. The data for High Level is interesting as a decent amount of the fire covered an area equal to and exceeding 4.0 ha, making it more vulnerable to damage. The data for the wildfires at Lac La Biche gives more valuable information, with it being the only region in which the Class E fires almost double the Class D wildfires in the region, resulting in it being vulnerable to damaging wildfires.

Size Class by Region - Numerical Representation

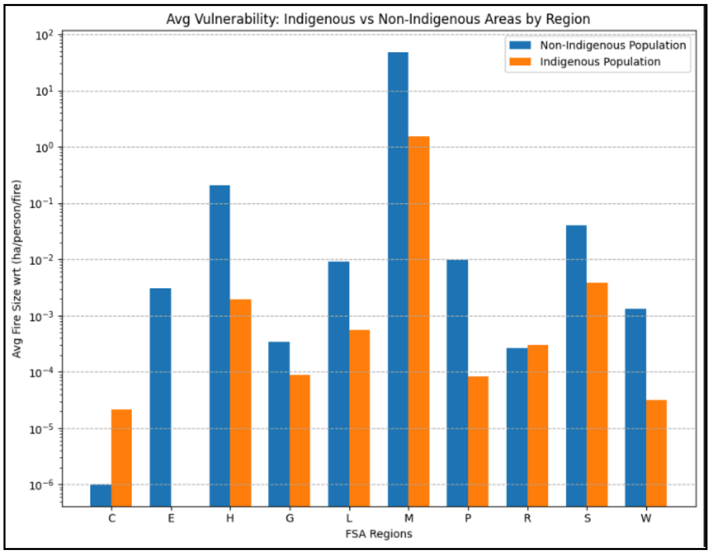


This gives a more visual and overarching representation of the heatmap. Almost all regions have most of their wildfires in the 0 - 0.1 ha range, showing that ways of mitigating fires before growth seems to be working, especially in Calgary where almost all wildfires fall in this range. Proportionate to the number of fires, High Level, Lac La Biche, Fort McMurray, and Slave Lake all have a large number of wildfires in the 0.1 - 4.0 ha range, showing some areas of vulnerability. This is shown through the relatively high number of fires exceeding the 4.0 ha range, indicating that those regions are more susceptible to devastating wildfires and resource allocation methods must be improved to mitigate damage.

Indigenous vs Non-Indigenous - Size Class Comparison



This comparison bar graph was made to see if there is any correlation present between both the number and size of fires in Indian Reservations when compared to the other fire origins. There is a big difference between the number of fires between both origins, which is to be expected considering the vast types of fire origins. One interesting factor is when comparing the ratio of the Class A and B sized fires between the two. For Class A, there is approximately a 8:1 difference whereas with Class B, there is a 4:1 difference. This could suggest that the fire mitigation method and the resources available in Indian Reservations might not be as effective as the other origins.



Indigenous vs Non-Indigenous - Average Fire Size

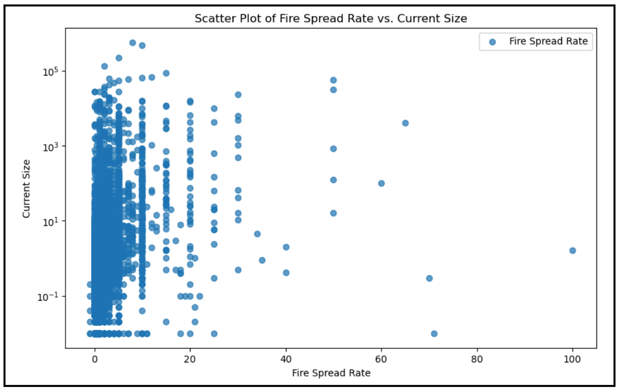
This visual breaks down wildfires on FSA regions and the impact on the Indigenous and Non-Indigenous population. The team decided to incorporate the Indigenous and total population values in each FSA region as mentioned in the Statistics Canada Website. The average size (in ha) per fire is calculated for each region, and then added to either the Indigenous or non Indigenous tally, based on whether the fire was started on an Indian Reservation or not. This data is then divided by the respective population, creating an average fire size "per person".

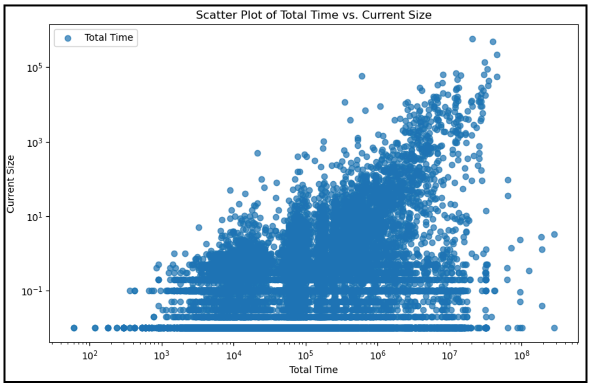
Selected Features

From the 50 columns present in the given dataset, it was evident that only a few are necessary in creating a model that predicts severity and final burned size. Through logical reasoning, the team selected "fire_spread_rate" to be an important feature, as it directly controls how fast a fire can spread. The team also included a new column, "total_time", which is the difference between "ex_fs_date" and 'fire_start_date". This feature represented the amount of time the fire was burning for, and when combined with the fire_spread_rate, could yield a prediction. Important to note, the team ignored the first cell in the data, since it contained incorrect information in the 'fire_start_date' category, indicating the fire started in 2010, while every single other entry included years between 2017 and 2021 respectively. This start date was deemed incorrect due to the "ex_fs_date" for that cell showcasing a time in 2021, meaning that this fire had lasted almost 11 years, which is impossible. This led to the conclusion that the given data, especially for total time, may have some inaccuracies, which can affect the model, especially wild outliers such as the one mentioned above.

The two input variables were then graphed against the target variable.

Due to the discrete values present in the fire spread rate, the distribution showed almost no correlation and seemed random. However, the team planned that along with the other input variable, together, they would form a decent correlation. Even though there is a lot of noise in the second graph, it is clear that there is some positive correlation between the variables.



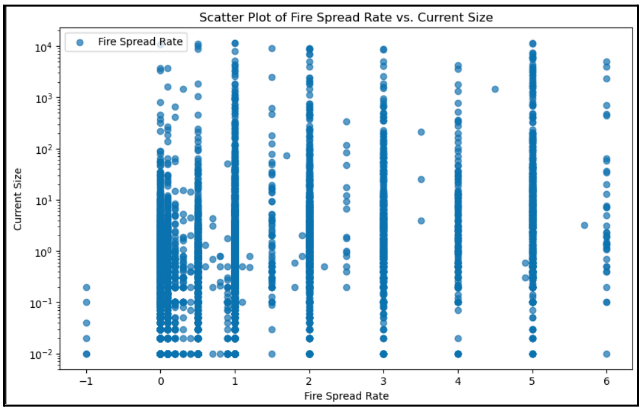


Data Processing

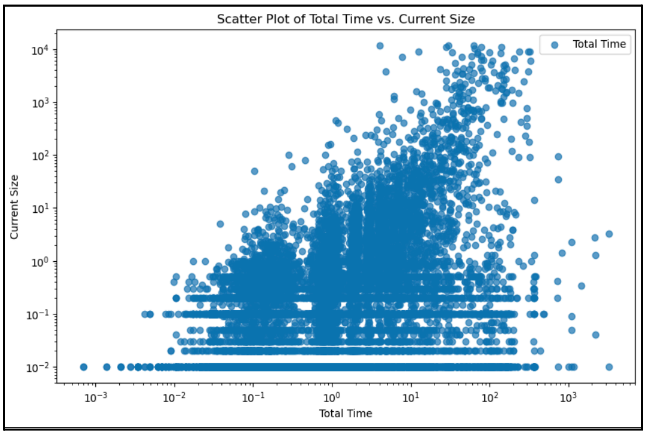
Through analyzing the graphs, the team realized the features need to be scaled to be inputted into machine learning algorithms better. Thus, the team scaled the target and input variables using StandardScaler, with a mean of around 0 for each distribution, ensuring that the correlations stayed the same. Additionally, the team split the data into 80% training and 20% testing so that the models would not be overfitted and in turn see its true accuracy.

Modeling:

Since the team did not know which model would be the best fit for the data, linear regression and ridge regression was initially tried, as they were the most straightforward for the data. Their coefficient of determination was low (around 10%), and it was clear they did not fit the data well. However, after looking at the mean squared error as well as the plotted graphs above, the outliers in our dataset were affecting the model. Thus, it was decided to remove outliers in 2 ways. Firstly, using IQR (InterQuartile Range) the team could statistically categorize the outliers in the dataset and remove them (an outlier is beyond the bounds indicated by (lower_bound = Q1 - 1.5 * IQR, upper_bound = Q3 + 1.5 * IQR).



Secondly, a 95% confidence interval was used on the processed data, to further remove outliers and make our processing more robust. After cleaning, the data was plotted again to see any differences.

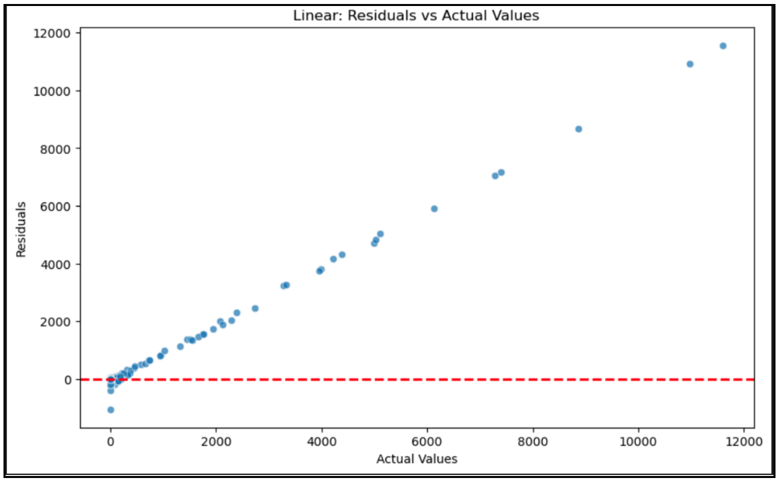


This input variable is now on full display as being discrete, and it is clear that there is little correlation between itself and the target variable. Although straight lines are an indication of discrete data and low correlation, when incorporated with other models, a correlation may occur. This input variable, without outliers, has a large spread and is generally showcasing a positive correlation.

Modeling cont.

The team conducted the same pre-processing on the new data and now wanted to see how the linear models would react. Therefore, the same type of models were trained on this data. While the coefficient of determination reduced for both linear and ridge regression to around 5%, the mean squared error nearly halved in value. This made the team realize that it was the outliers that were creating a fake linear correlation with the data, and now without them, these models are performing even worse. To take it a step further, plotted a residual plot from our predictions from the linear model.

The residual plot does not show residuals in random places, rather it shows a linear pattern, showcasing that linear model is not the right fit for this model.



After this realization, our team began training our data on non-linear models, such as SVR (Support Vector Machine Regression), Random Forest Regressor and Kernel Regressor.

Since these models were non-linear, the team could not use a coefficient of determination and rather minimized Mean Squared error (MSE). All these models yielded much lower (MSE) and fitted better to the data. However, the team realized that even with these models, the data was not fitting as expected. The team decided to try out a model that would incorporate a bunch of these non-linear models together, in hopes of creating a better fit, thus, decided to utilize an Extreme Gradient Boosting Regressor. This model yielded our lowest MSE.

Navigating hyper parameters:

While using machine learning models, specifically the non-linear models, which include hyperparameters such as gamma, alpha, the team saw a vast difference in the performance of a model when tweaking these parameters. Therefore, the team established a grid search algorithm, which enabled the team to isolate the hyper parameters that best suited our data for each model.

Connection to Objective:

Given the objective of predicting final burned size and severity of wildfires, the team implemented a non-linear model using an Extreme Gradient Boosting Regressor, which was a combination of other non-linear models the team tested. While the MSE for this model was high, it was still possible to categorize the severity of the data and predict final burned size. The target variable had an unpredictable distribution and considering all the confounding variables that existed outside of our model, it was difficult to establish an extremely accurate regression model.

Conclusion:

Through plenty of effort and persistence, the team believe that this project has the potential to make significant strides in combating Alberta's wildfire challenges through the team's use of data analysis and machine learning. Through using an Extreme Gradient Boosting Regressor, this model has the capability to categorize wildfire severity and predict final area burned, clarifying crucial aspects that are needed for effective resource allocation. Moving forward, the project would certainly benefit from an enhanced feature selection process, potentially employing sequential feature selectors for improved model accuracy. Additionally, optimizing data scaling by considering min-max scaling instead of standardizing it would certainly help. Finally, delving deeper into the dataset's categories for a more comprehensive feature identification would certainly help with future development. These developments would potentially produce a wildfire prediction model with enough accuracy and consistency to address the alarming landscape of Alberta's wildlife management.

References:

DataFrame. pandas. (n.d.). https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

API reference. matplotlib. (n.d.). https://matplotlib.org/stable/api/index.html

XGBoost. NVIDIA Data Science Glossary. (n.d.). https://www.nvidia.com/en-us/glossary/xgboost/#:~:text=XGBoost%2C%20which%20stands%20for%20Extreme,%2C%20classification%2C%20and%20ranking%20problems.

Brownlee, J. (2021, April 26). Extreme Gradient Boosting (XGBoost) Ensemble in Python. Machine Learning Mastery. https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/

Linear Regression Example. scikit-learn. (n.d.). https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

Machine Learning - Linear Regression. w3schools. (n.d.). https://www.w3schools.com/python/python_ml_linear_regression.asp

sklearn.linear_model.LinearRegression. scikit-learn. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

sklearn.linear_model.Ridge. scikit-learn. (n.d.-c). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Sethi, A. (2024, February 9). Support Vector Regression Tutorial for Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

Government of Canada, Statistics Canada. (2023, June 21). Indigenous Population Profile, 2021 Census of Population. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/ipp-ppa/index.cfm?Lang=E

sklearn.svm.SVR. scikit-learn. (n.d.-d). https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Support Vector Regression (SVR) using linear and non-linear kernels. scikit-learn. (n.d.-e). https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html

Support Vector Regression (SVR) using Linear and Non-Linear Kernels in Scikit Learn. GeeksforGeeks. (2023, January 30). https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/

sklearn.kernel_ridge.KernelRidge. scikit-learn. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html

sklearn.ensemble.RandomForestRegressor. scikit-learn. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html