Wildfire Classification and Size Prediction

Machine Learning Engineer Nanodegree Capstone Project

Brandon Avery, P. Eng

Feb 3, 2020

Table of Contents

Definition

Overview

Wildfires are increasingly intense natural events that often result in significant damage to many communities and effective wildfire management is becoming more challenging due to extreme weather conditions and limited resources. Classifying and predicting wildfire behavior would optimize wildfire management policy and minimize the losses attributed to these increasingly intense ecological events.

The Government of Alberta (GOA) maintains historical data sets of all wildfire incidents from 1961 to 2018, with 38 fields such as GPS coordinates, total area burned, fire start date, and initial suppression resources. The most recent data set contains data on over 18,000 wildfires between 2006 and 2018. I chose this data set as I previously worked as wildland firefighter in Northern Alberta and want to study the behavior of wildfires and the effectiveness of wildfire management strategy. These data sets are available on the provincial website.

Wildfires are defined as fires that have a point of origin within the provincial Forest Protected Area (FPA). They are detected by various groups such as fire tower lookout observers, aircraft patrols, and the general public. About 80% of fires are discovered by the lookout observers and the public, while the remaining 20% is typically discovered by air patrols. Early detection of fires is critical to the successful management of the forest resources throughout the province of Alberta.

Once a fire is detected, the closest fire center is notified, and initial action resources are mobilized. The GOA minimizes the response time through strategic placement of firefighting crews and air tankers based on the expected risk on a given day. This response time is tracked and used as a metric of the effectiveness of the initial action and is included in the provided data sets.

Wildfires exits in one of four states: *Out-of-Control (OC)*, *Being-Held (BH)*, *Under-Control (UC)*, and *Extinguished (EX)*. All detected wildfires begin in the OC state and can stay there even after suppression resources are deployed. The Incident Commander (IC), who manages the suppression efforts of the fire, is responsible to update the size and status of a wildfire. He or She does this typically by walking the perimeter of the fire setting GPS waypoints along the perimeter. To move from UC to BH, a wildfire must be momentarily contained and no longer growing. A fire in the UC state means that there are no observable flames and the fire is completely contained, despite hot spots existing in the burned area. An extinguished fire is completely void of hot spots throughout the entire burned area, on the surface or beneath the ground. The time and size when a fire advances to the next stage is documented in the provided data set.

Problem Statement

The problem to be solved is to optimize a wildfire management strategy by clustering wildfires in terms of the initial detected state into fire groups and predicting the final size of the burned area for each fire group based on the suppression response times. The fire groups can be used to identify how many and what combination of suppression resources work well on similar fires, and the estimated final burned area can be used to prioritize wildfires in situations where resources are limited.

Metrics

The data set contains a classification of wildfires based solely on the final size of the burned area, which will be a used as a metric to compare the fire groups to. The descriptive statistics of the wildfires in both classification systems will be compared to each other. The number of fire groups will be determined by maximizing the average Silhouette Coefficient for various numbers of fire groups. After the fire groups are created, the data set will be data used to create and validate the regression model. The desired accuracy of the predicted final burned area is at least 90%.

Analysis

Data Exploration and Visualization

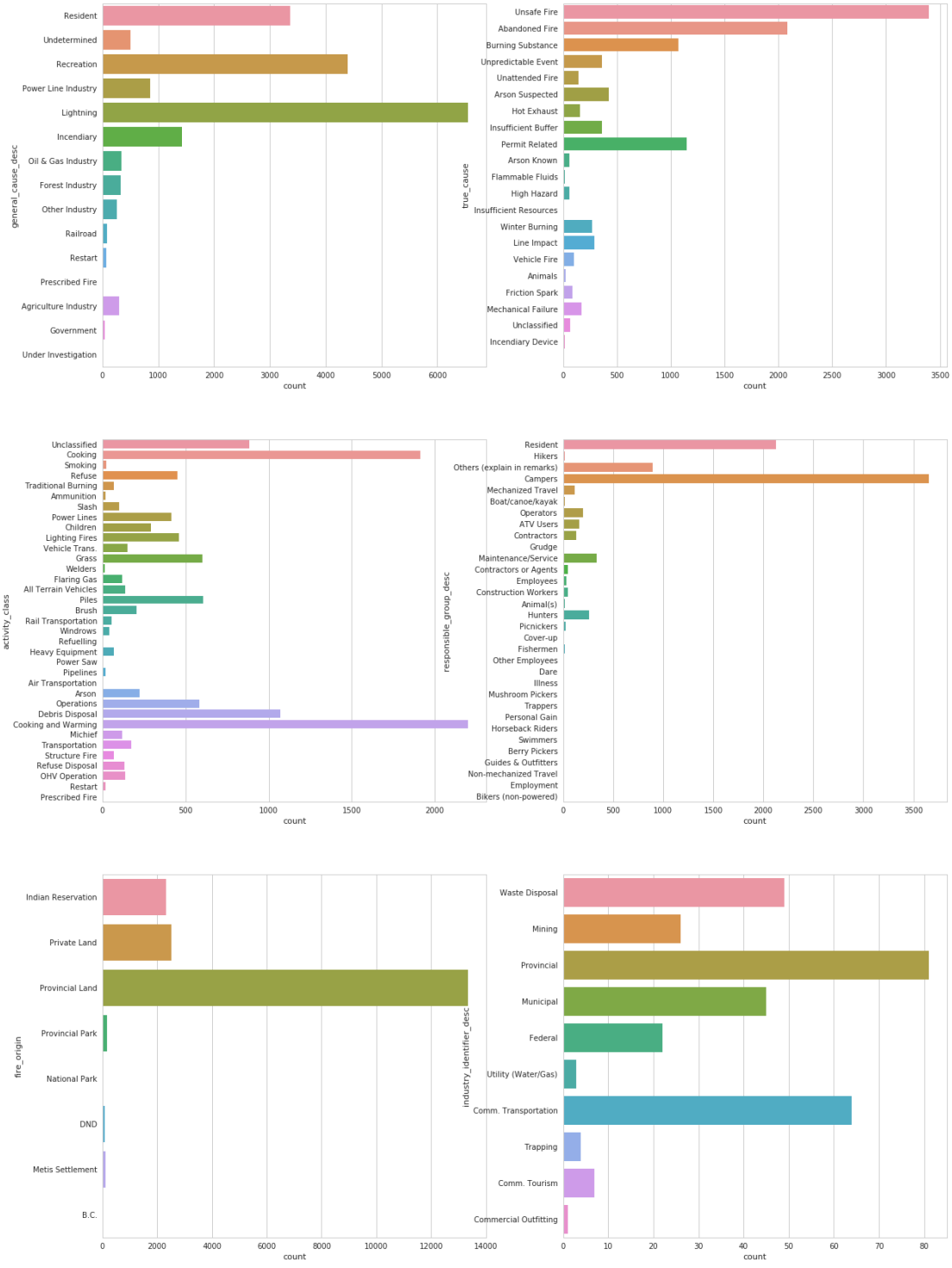
This data set has 18,565 rows and 39 columns, some of which are not relevant to this analysis. As a first step, I will drop some columns out of the dataframe that will not be used and use the fire_year and fire_number together as an index. The fire_year is the year in which the fire was first detected and the fire_number is specific to which fire district is responsible for managing the fire. The fire number contains a three-letter substring to identify the district and is reset to 1 every fire season. These two values combined create a unique identifier for each wildfire.

Between 2006 and 2017, there was 159 named fires. Wildfires are only given names if they become provincially important. These are typically a group of fires in close proximity of each there, called a fire complex, such as the Flat Top Complex that burned through Slave Lake, Alberta in 2011 or a fire that has grown to an immense size such as the Horse River Fire that burned through Fort McMurray, Alberta in 2016. The largest fire in the province in this time period, the second largest wildfire in Alberta history, also occurred in 2011 and was called the Richardson Fire. It was assessed the same day as the Flat Top Fire Complex and burn through approximately 577,647 hectares of forest north of the oilsands near Fort McMurray. These fires are shown below, and all started in May, which is early in the fire season in Alberta, on level ground on clear days and burned for months until they were extinguished. Despite the quick response time, as shown as the difference between the start_for_fire_date and fire_fighting_start_date, they grew rapidly and out of control. Some of the fires had known causes such as arson or insufficient buffers around power lines, however the cause of the larger Horse River Fire is still unknown.

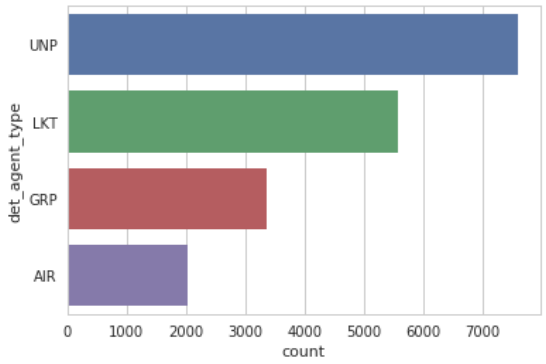
Index	Fire Name	Assessment Size	Current Size	Assessment Date	Start for Fire Date	Fire Fighting Start Date
2011-MWF007	Richardson Fire	4.0	577, 647	2011-05-14 14:25	2011-05-14 14:17	2011-05-14 17:45
2011-SWF056	Flat Top Complex	0.2	16, 011	2011-05-14 13:40	2011-05-14 13:49	2011-05-14 14:20
2011-SWF065	Flat Top Complex	5.0	3, 974	2011-05-14 18:01	2011-05-14 17:51	N/A
2011-SWF082	Flat Top Complex	30.0	425	2011-05-16 16:36	2011-05-16 16:31	2011-05-16 18:55
2016-MWF009	Horse River Fire	2.0	485, 124	2016-05-01 16:08	2016-05-01 16:03	2016-05-01 16:33

Not all fires appear to have a documented start or discovered date, but all have an assessment and reported date. This could be simply missing data during collection but for predicting wildfire sizes, using the reported or assessment date would be a more reliable data point. The primary cause of wildfire in the dataset was lightning strikes, as shown below, accounting for 6,549 or 35% of the reported wildfires. The top human induced causes are recreation activities, residents living in the wildland-urban interface, and intentional starts such as arson.

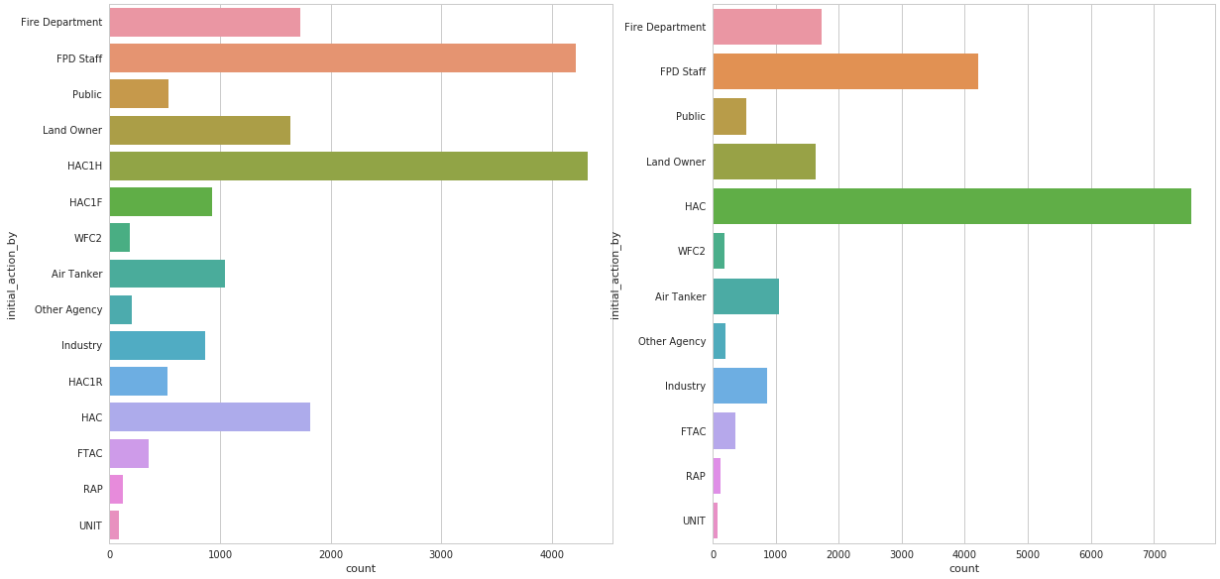
The most common cause of *human caused* wildfires given in this data set was unsafe and abandoned fires, typically from cooking and warming activities by campers and residents. These activities are associated with unsafe or abandoned campfires, wood stoves, smokehouses, sweat lodges and/or saunas and account for **5,624 (55%)** of the reported true causes. Combined with the 1072 cases of burning substances such as cigarette butts or smoldering materials from all-terrain vehicles, **6,696 (65%)** of all wildfires between 2006 and 2017 with reported true causes were from carelessness. The most dominant activity causing wildfires is cooking with 5,113 (45%) reported causes, followed by debris disposal with 1070 (9.4%) of reported causes. The distribution of true causes and activities are shown below. The fire origin and industry identifier indicate the administrator or owner of the land on which a given wildfire is detected and further details for the "other industry" category in the general cause description. Most fires were detected on provincial or crown land, followed by private lands and Indian Reserves.



Wildfires detection is sorted into four broad categories: *Unplanned (UNP)*, *Lookout (LKT)*, *Ground Patrol (GRP)*, *Air Patrol (AIR)*. The majority of detections are unplanned, primarily from the toll-free hotline 310-FIRE, with 7,601 reported cases. The number of detections by lookout towers at 5,585 cases is greater than the detections from air and ground patrols combined, equal to 5,377. The det_agent column contains detailed information regarding the specific agent that detected the wildfire, listing has all 127 fire towers, other government agencies, and types of aircraft. Apart from the 310-FIRE hotline, forest rangers (FRST), wildfire department personnel (LFS), the general public (PUB), and patrolmen (PATR) account for a large portion of detections.

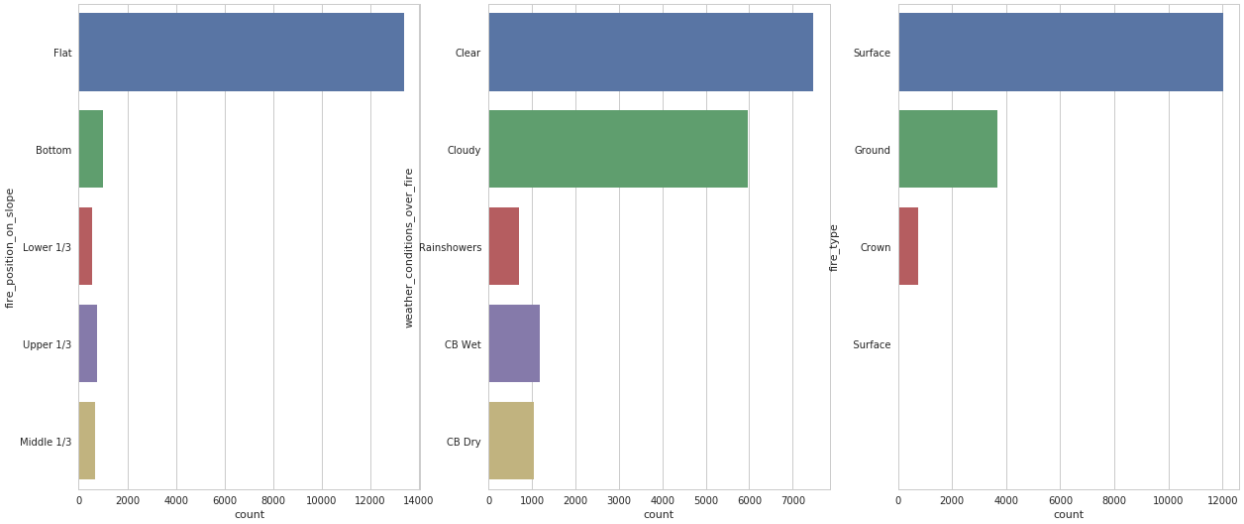


The initial_action_by column details the wildfire management resources that provided the first physical suppressive action. The top two categories are FPD staff, representing provincial employees working in wildfire management such as rangers or patrolmen, and Helitack firefighting crews (HAC) which are the primary initial action resource, as shown in the figure below. There appears to be duplicate categories for HAC, so that will be compiled together into one category.



Below are visualizations of the topographical and weather data provided. The most common initial conditions of wildfires appear to be surface fires on level ground on clear days. All three parameters have unbalanced distributions favouring the most common conditions. Given the geography of the province of Alberta, where the province is dominated by plains and forests except for the foothills and Rocky Mountains, flat slope is to be expected.

The weather conditions over the wildfires were primarily sunny or cloudy, with a small portion reported as rain showers and Cumulonimbus (CB) clouds, which are often seen on hot summer afternoons throughout Alberta. CB wet clouds are also known as thunderheads and are associated with lightning storms.



The fire type describes the predominant fire behavior that was observed when the fire was first assessed. There are three types: Ground, Surface, and Crown. A ground fire burns primarily in the ground fuel layer, such as grass or dead leaves. A surface fire burns in the surface fuels and involves less than 10% of the trees torching. Campfires, brush piles and windrows that have not burned are examples of these. Crown fire advance through the canopy layer and involves more than 10% of the tree crowns. The largest and most intense fires are typically crown fires.

The various types of fuels in which wildfires are assessed are shown below. The data consists of codes that correspond to the types of surface fuels and forest types native to Canada. The detailed descriptions are shown in the table below. The two largest categories are boreal spruce and matted grass. The boreal forest is the largest geographical region in the province and covers most of the north, so it makes sense that most of the reported wildfires are detected with this fuel type. The fuel codes C7, M3, and M4 only had 9, 2, and 1 samples in the dataset, respectively. Since there are so few of these codes, it would simplify the analysis to just remove these 11 records and reduce the number of fuel codes.

Code	Description
C1	Spruce-Lichen Woodland
C2	Boreal Spruce
C3	Mature Jack or Lodgepole Pine
C4	Immature Jack or Lodgepole Pine
C7	Ponderosa Pine and Douglas Fir
D1	Leafless Aspen
M1	Boreal Mixedwood-Leafless
M2	Boreal Mixedwood-Green
M3	Dead Balsam Fir Mixedwood–Leafless
M4	Dead Balsam Fir Mixedwood–Green
O1a	Matted Grass
O1b	Standing Grass
S1	Jack or Lodgepole Pine Slash
S2	White Spruce–Balsam Slash

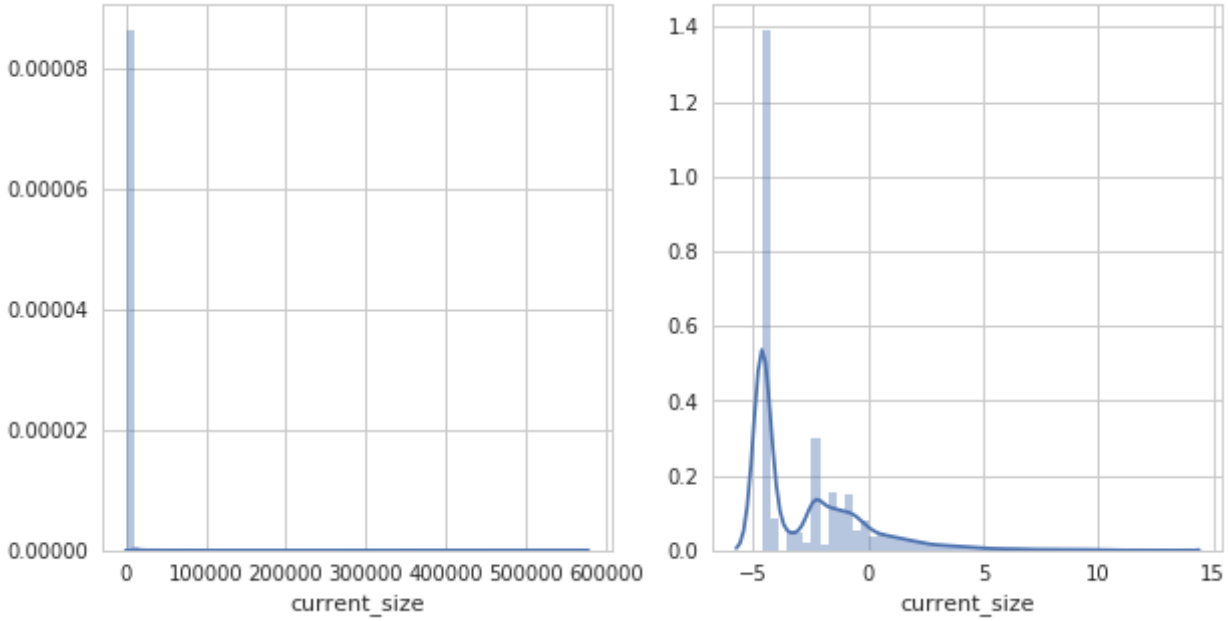
Numerical Data Exploration

Now to look at the numerical data provided which consists of the burned area of each wildfire, measured in hectares, at each stage of the wildfire life cycle. Below is the summary statistics of the wildfire sizes showing small mean sizes

with large standard deviations. This is expected as the province commits to fast response times to suppress wildfires before they grow too large. The large standard deviations are a result of the extreme outliers representing the largest wildfires.

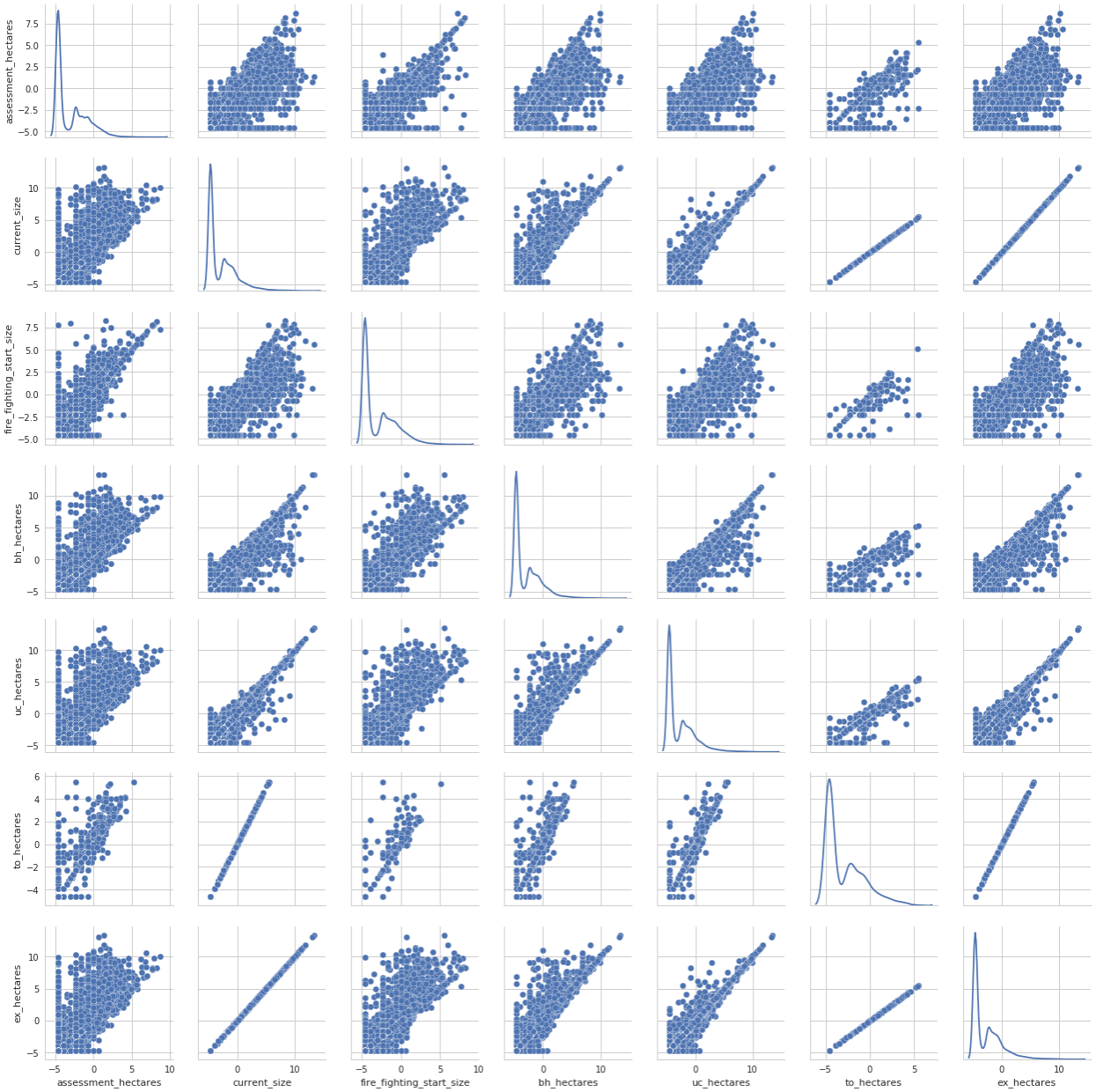
Value	Assessment Size (Ha)	Current size (Ha)	Fire Fighting Start Size (Ha)	BH Size (Ha)	UC Size (Ha)	TO Size (Ha)	EX Size (Ha)
Count	18,565	18,565	18,565	18,565	18,565	12,033	18,565
Mean	2.42	142.69	4.30	115.67	154.41	1.69	142.69
STD	66.61	5803.25	79.81	6299.47	6979.77	11.48	5803.25
Min	0.01	0.01	0.01	0.01	0.01	0.01	0.01
25%	0.01	0.01	0.01	0.01	0.01	0.01	0.01
50%	0.01	0.02	0.01	0.01	0.01	0.02	0.02
75%	0.02	0.03	0.28	0.28	0.30	0.20	0.30
Max	6,019.00	577,646.80	3,815.00	602,417.00	707,648.00	244.20	577,646.80

The first variable to look at is the current size, which is the final area burned by the fire, after it has been extinguished, further assessed using aerial photography, and interpolated to remove any unburned islands from the aera assessment. This is the value that I wish to predict and understanding the distribution of measurements is important to understand how the regression will be calculated. The vast majority of values are very small, skewing the distribution so much that it appears as a single column of values. The skewness and kurtosis of the original data was calculated to 85.7 and 7931.3 which show that distribution is extremely skewed towards lower sizes and is likely to contain large outliers, respectively. A log transformation is typically used to assess skewed data and to make it appear more normally distributed. A comparison between the original and transformed data is shown below. The transformed data had a skewness and kurtosis values of 4.5 and 9.0, respectively.



A pairwise comparison of the wildfire sizes, after applying a log transform, is shown below. This is useful to determine the distribution of each size variable as well as the correlation between them. The univariate distribution of each variable is similar in skewness and kurtosis. The size data show roughly linear relationships with each other, especially in the later stages of the wildfire life cycle. This is to be expected as the suppression efforts near the end of the lifecycle are usually enough to prevent further growth and is focused on finding hot spots and fully extinguishing the burned area. The bias on one side of the perfectly linear line between the values show the relationships between the variables (i.e., the uh_hectares is typical greater than the bh_hectares because the firefighters are still working to control the spread of the wildfire). There is exact linearity between the current_size, to_hectares, and ex_hectares. This is expected because these are all measurements of the wildfire after it has been

extinguished and assessed accurately for the actual burned area so this means that two of the three variables can be removed from the analysis without losing information in the models.



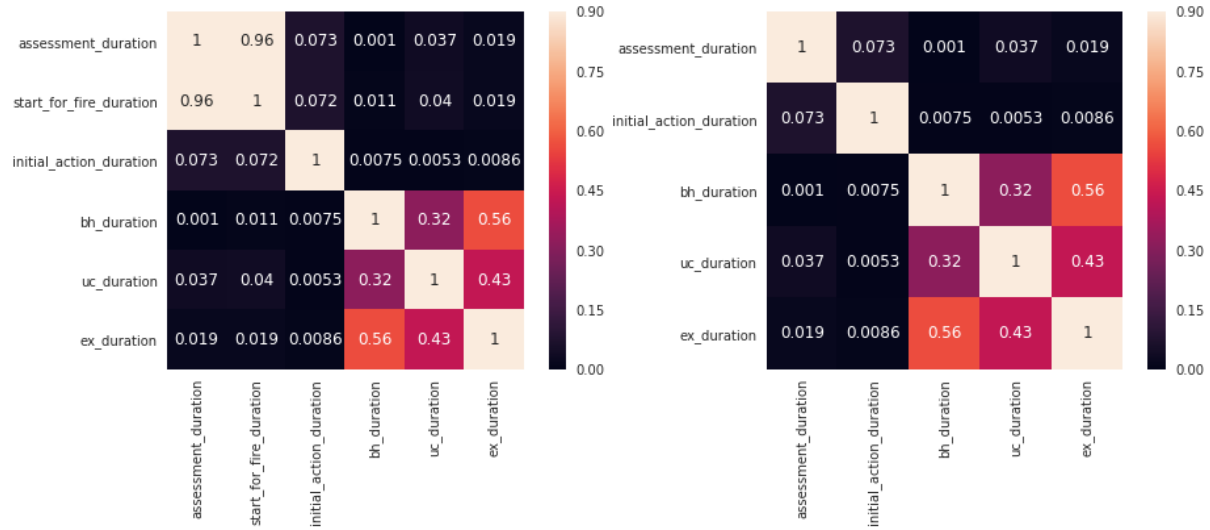
Missing Data

Not all variables have values for every reported wildfire. This could be a result of incomplete assessments, human error in the data entry process, or changes to the data tracking or priorities between 2006 and 2017. The percent of null values was calculated in the table below. The variables with the highest percent of missing values as well as redundant variables will be dropped with minimal information loss and to simplify the analysis. The fire_fighting_start date and fuel_type are critical variables in the analysis, so they cannot be removed but have 4,969 and 5,759 missing values, respectively. I will have to drop the rows with the missing values and not use the entire data set for the analysis. There remain a few missing values for the fuel_type, weather_conditions_over_fire, and fire_position_on_slope variables that I instead replaced the missing values with the most common values occurring in the data set. I chose this instead of just removing the rows to keep as much of the original data set as possible. After removing all missing values, 9,725 or 52% wildfire records remain of the original data set.

Category	Percent Missing
industry_identifier_desc	98.4%
fire_name	97.5%
permit_detail_desc	95.7%
to_hectares	89.0%
to_fs_date	89.0%
other_fuel_type	80.6%
responsible_group_desc	55.9%
true_cause	44.4%
activity_class	38.7%
fuel_type	31.0%
fire_fighting_start_size	26.8%
fire_fighting_start_date	26.8%
discovered_date	13.0%
weather_conditions_over_fire	11.9%
fire_position_on_slope	11.9%
fire_type	11.5%
fire_start_date	2.7%

Datetime Data

First, I want to check that there are no extreme dates within the data set. There are 5 dates in the fire_fighting_start_date column that seem to be typos where the year was inputted incorrectly. These are corrected to the most likely date from observing the other columns. The features to be used for the regression is the duration of time between fire stages. The first time delta is from when the fire is reported to when the fire is assessed, called the Assessment Duration. The Start for Fire Duration and Initial Action Duration are the differences between the start_for_fire_date and fire_fighting_start_date with the reported_date, respectively. The BH Duration is the difference between the bh_fs_date and the assessment_date. The UC and EX Durations are the time deltas between the bh_fs_date, uc_fs_date, and the ex_fs_date, chronologically. A correlation matrix is created of these time deltas to see if one or more of these variables can be removed without losing information. As shown in the heatmap below, there is high correlation between the Assessment and Start for Fire Durations. This is most likely due to the large amount of fires detected by personnel that initiate suppression actions, such as a HAC crew out on a patrol or a forest officer overseeing a permitted fire. I plan to take out the start_for_fire data columns as I believe the assessment time duration is a more meaningful metric to use in the regression. There is some correlation in the later time deltas, suggesting that the duration of wildfire suppression efforts is more evenly distributed amongst the wildfire stages. In other words, the time intervals between the BH, UC, and EX states are roughly equal.



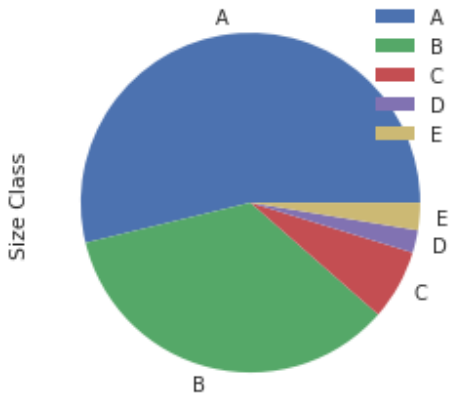
Methodology

There will be three types of algorithms to be used to classify and predict the final size of wildfires in Alberta: PCA, k-means clustering, and linear regression. The PCA algorithm will be used to perform dimension reduction on the processed data and extract a small number of features to explain the data. K-means clustering will be used to classify the wildfires into the fire groups, like the size class variable given in the original data set. The difference between these fire groups and the size_class variable is that they will consider all the available data to find relationships instead of relying solely on the final size of the burned area. Finally, a regression model on the fire groups using the duration between stages in the wildfire life cycle will be created to predict the final burned area of new wildfires. Linear regression will be explored first, as it is the simplest model, but more advanced models may be used to predict the size of new wildfires.

Benchmark

The data set contains a classification of wildfires based solely on the final size of the burned area, which will be a used as a metric to compare the new fire groups to. The distribution and descriptive statistics of the wildfires in both classification systems will be compared to discuss if the new classification method is relevant. The number of fire groups will be determined by maximizing the average Silhouette Coefficient for various number of groups. After the wildfires are classified, the data set will be split into training and test data used to create and validate the regression model. Each fire group will have its own regression model to estimate the final burned area of the type of wildfire as a function of the duration between wildfire management phases.

A pie chart of the size classes is shown below. The size classes are divided into five categories: A (less than 0.1 ha), B (between 0.1 ha and 4.0 ha), C (between 4.0 ha and 40.0 ha), D (between 40.0 ha and 200 ha), and E (greater than 200 ha). Approximately 90% of the reported wildfires were less than 4.0 ha with a combined burned area of 2,918 ha. 98% of the total burned area between 2006 and 2017 is due to only 253 wildfires.



Data Pre-Processing

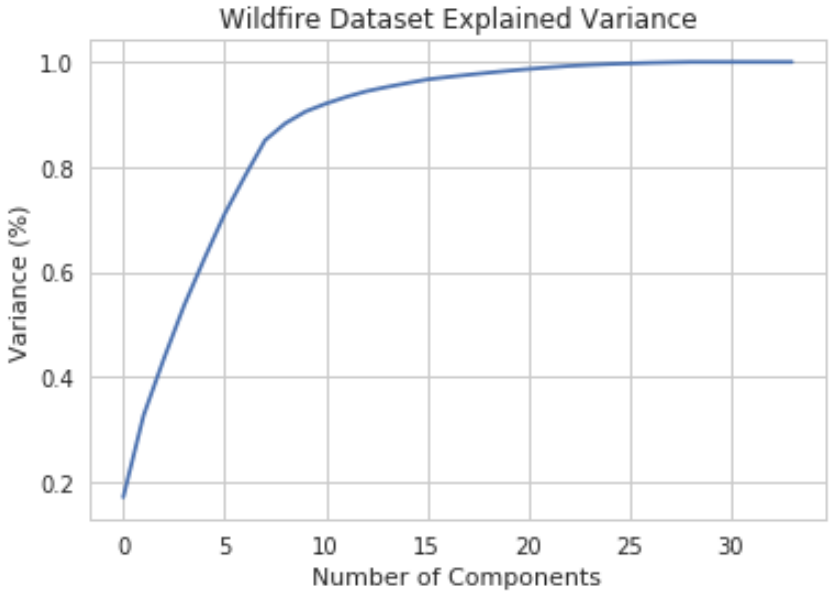
The first pre-processing step is to encode the categorical data in order to create models for classification and regression. The nominal data in the det_agent_type, initial_action_by, fire_position_on_slope, weather_conditions_over_fire, and fuel_type will be representing as integer values using the one-hot encoding method. The fire_type category can be represented using ordinal encoding to show the relative intensity between surface and crown fires and the size_class column will also be dropped since it is no longer needed. Since we will be classifying the wildfires on the initial observed state, the fire_fighting_start_size, bh_hectares, and uc_hectares will also be removed and the only remaining size measurement will be the assessment_hectares. The current_size and duration between wildfire states will be saved for training the regression algorithm later.

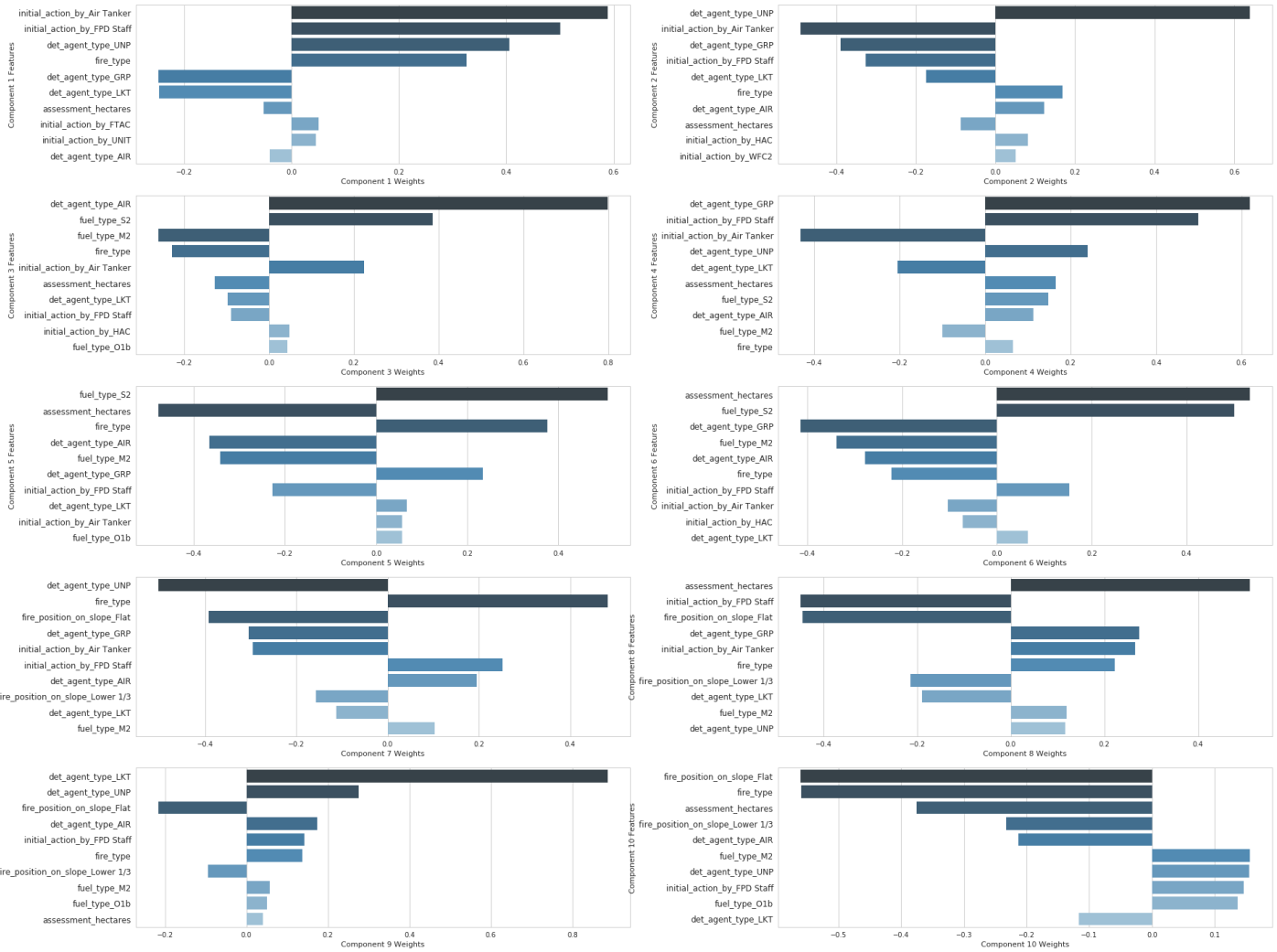
The next step is to scale and standardize the observed state data in order to prevent difficulties in the classification algorithms. Due to the presence of extreme outliers in this data set, which cannot be removed, the standard scalers will not work well so the power transformer from sklearn will be used to make the input data more Gaussian. The transformer will have to use the Yeo-Johnson transform due to the numerous zero values in the data.

Principle Component Analysis (PCA)

To complete the PCA I followed the procedure as shown in the example project of population segmentation provided here. The time duration data will be excluded from this analysis as these are the features I want to use in the regression analysis later. I used the PCA model from sklearn to compute the principle components and plotted the explained variance as a function of the number of components used (shown below), as detailed in this post. By using 10 components, just under one third of the original feature space, about 90% of the variance is captured.

The PCA is re-fit to the data using 10 components. To visualize the composition of the principle components, I modified the make-up visualization from the example project of population segmentation provided in this tutorial to plot the composition of all 10 principle components. Only the top 10 weights of the original data are displayed in each component. Through inspection of the composition plots, it appears that the type of initial action and detection resources, assessment size, and fire types are integral parameters to explain the variance in the data set. Lookouts, air and ground patrols appear to be dominant variables in explaining the variation in the data set.


Wildfire Dataset Explained Variance
Variance (%)
Number of Components
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
30

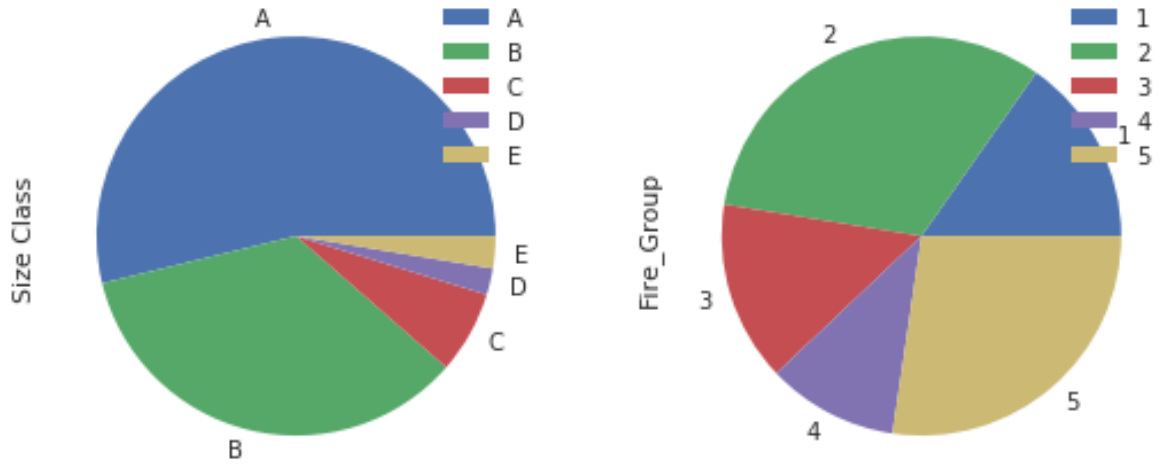

initial_action_by_Air Tanker
initial_action_by_FPD Staff
det_agent_type_UNP
fire_type
det_agent_type_GRP
det_agent_type_LKT
assessment_hectares
initial_action_by_FTAC
initial_action_by_UNIT
det_agent_type_AIR
Component 1 Features
Component 1 Weights
det_agent_type_UNP
initial_action_by_Air Tanker
det_agent_type_GRP
initial_action_by_FPD Staff
det_agent_type_LKT
fire_type
det_agent_type_AIR
assessment_hectares
initial_action_by_HAC
initial_action_by_WFC2
Component 2 Features
Component 2 Weights
det_agent_type_AIR
fuel_type_S2
fuel_type_M2
fire_type
initial_action_by_Air Tanker
assessment_hectares
det_agent_type_LKT
initial_action_by_FPD Staff
initial_action_by_HAC
fuel_type_O1b
Component 3 Features
Component 3 Weights
det_agent_type_GRP
initial_action_by_FPD Staff
initial_action_by_Air Tanker
det_agent_type_UNP
det_agent_type_LKT
assessment_hectares
fuel_type_S2
det_agent_type_AIR
fuel_type_M2
fire_type
Component 4 Features
Component 4 Weights
fuel_type_S2
assessment_hectares
fire_type
det_agent_type_AIR
fuel_type_M2
det_agent_type_GRP
initial_action_by_FPD Staff
det_agent_type_LKT
initial_action_by_Air Tanker
fuel_type_O1b
Component 5 Features
Component 5 Weights
assessment_hectares
fuel_type_S2
det_agent_type_GRP
fuel_type_M2
det_agent_type_AIR
fire_type
initial_action_by_FPD Staff
initial_action_by_Air Tanker
initial_action_by_HAC
det_agent_type_LKT
Component 6 Features
Component 6 Weights
det_agent_type_UNP
fire_type
fire_position_on_slope_Flat
det_agent_type_GRP
initial_action_by_Air Tanker
initial_action_by_FPD Staff
det_agent_type_AIR
fire_position_on_slope_Lower 1/3
det_agent_type_LKT
fuel_type_M2
Component 7 Features
Component 7 Weights
assessment_hectares
initial_action_by_FPD Staff
fire_position_on_slope_Flat
det_agent_type_GRP
initial_action_by_Air Tanker
fire_type
fire_position_on_slope_Lower 1/3
det_agent_type_LKT
fuel_type_M2
det_agent_type_UNP
Component 8 Features
Component 8 Weights
det_agent_type_LKT
det_agent_type_UNP
fire_position_on_slope_Flat
det_agent_type_AIR
initial_action_by_FPD Staff
fire_type
fire_position_on_slope_Lower 1/3
fuel_type_M2
fuel_type_O1b
assessment_hectares
Component 9 Features
Component 9 Weights
fire_position_on_slope_Flat
fire_type
assessment_hectares
fire_position_on_slope_Lower 1/3
det_agent_type_AIR
fuel_type_M2
det_agent_type_UNP
initial_action_by_FPD Staff
fuel_type_O1b
det_agent_type_LKT
Component 10 Features
Component 10 Weights

Implementation

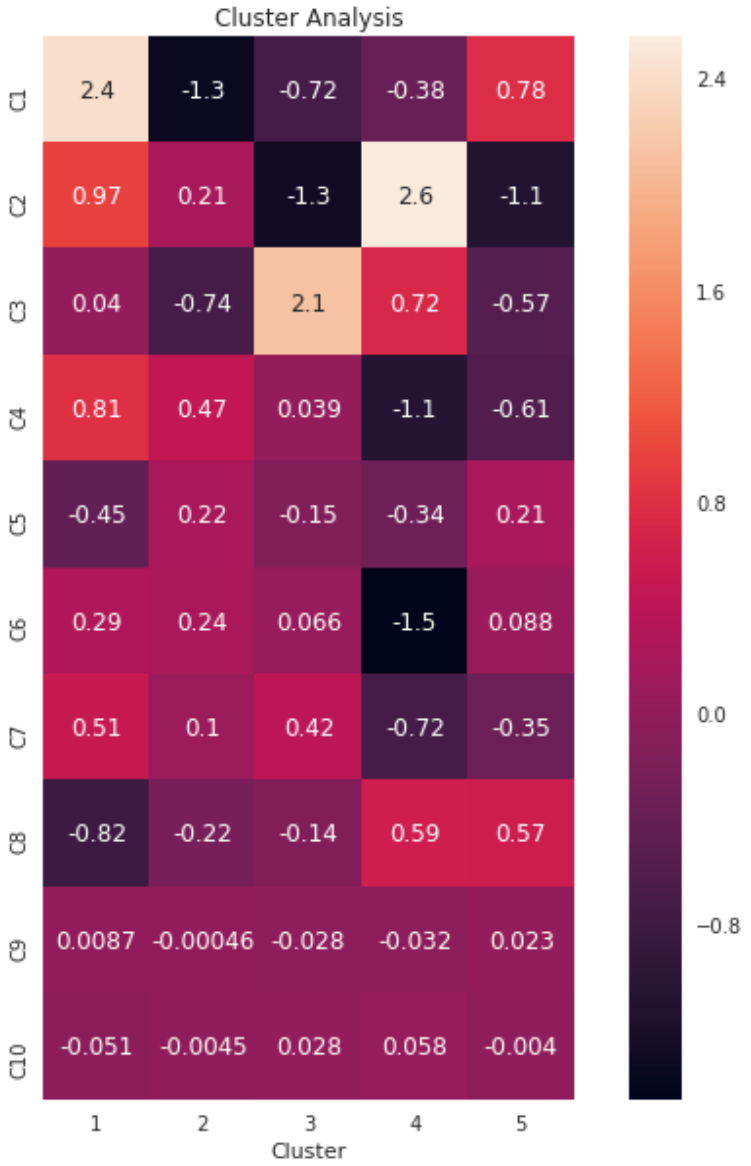
Fire Group Classification

To segment the wildfires using their PCA components in distinct fire groups based on the observed initial state, k-means clustering was used to find relationships on the principle components. To start, I used the same number of size categories given in the original dataset. This provided a direct comparison between the two classification systems. After, the quality of the clustering results was optimized using the average Silhouette Coefficient to determine an optimal number of groupings.

The labels from the k-means clustering algorithm were applied to the pre-processed data used to summarize the size classes. The fire groups are compared to the original size classes in the pie plots and summary table below. It appears to be more evenly distributed, which is to be expected using the k-means clustering algorithm, but the biggest difference is the distribution of the total burned area. The fire group with the largest burned area is group 2 but only accounts for 32% of the total burned area. Roughly half of the wildfires are classified in fire groups 2 and 5, which also account for 53% of the burned area.



To analyze the fire groups more closely, I will again borrow a visualization from this tutorial to study the distribution of principle components within the clusters by creating a heatmap of the cluster centers. Only the first eight components were really used to define the clusters. Cluster 1 primarily used component 1 which was heavily weighted towards the initial suppression and assessment detection resource type. Cluster 2 comprised mainly from components 1 and 3 which were greatly influenced by fuel types. Cluster 3 was mostly weighted on components 2 and 3. Cluster 4 had the highest weight across the clusters on component 2 which is also weighted towards the initial suppression and assessment detection resource type. Cluster 5 is more evenly weighted across most of the components.

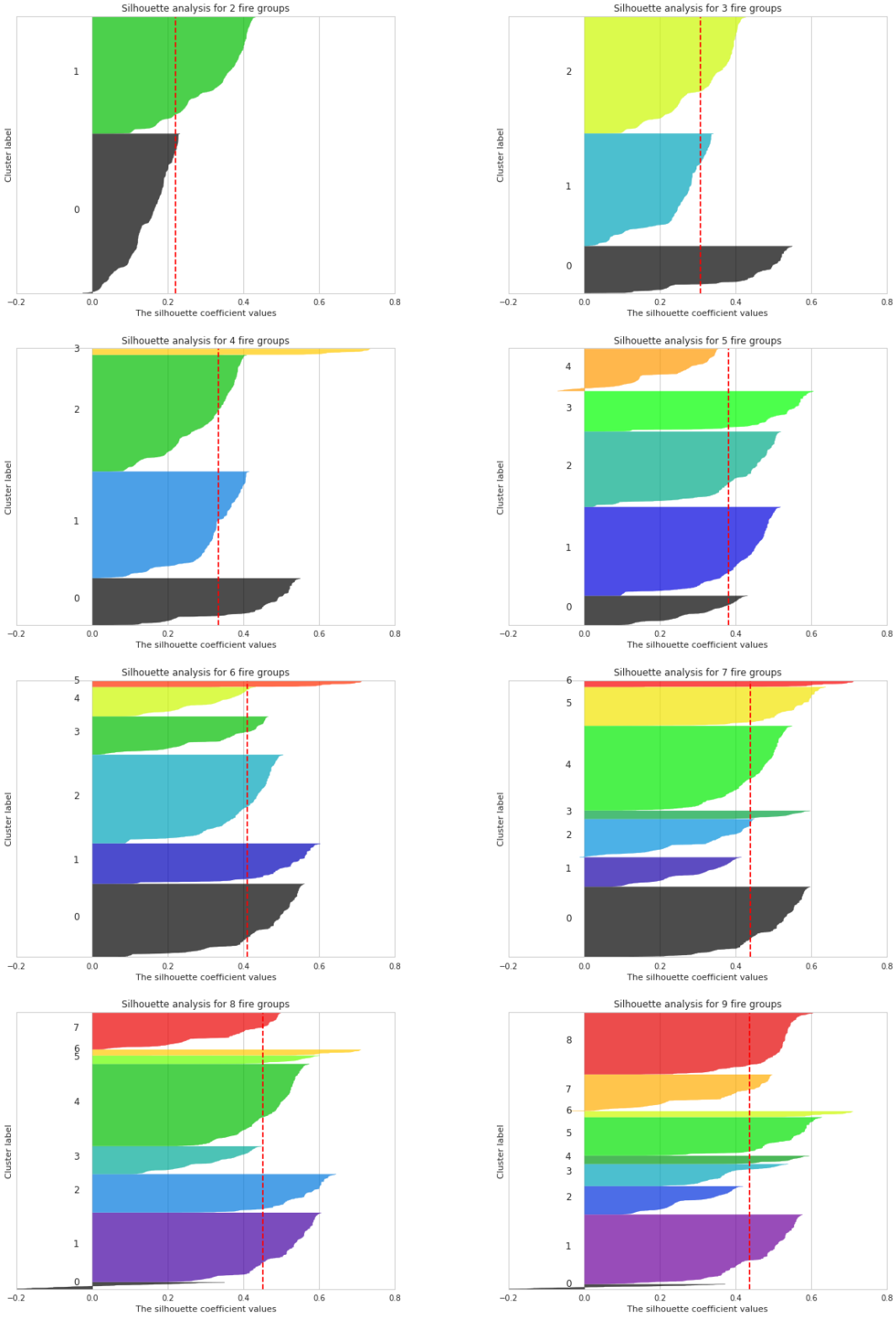


Silhouette Analysis

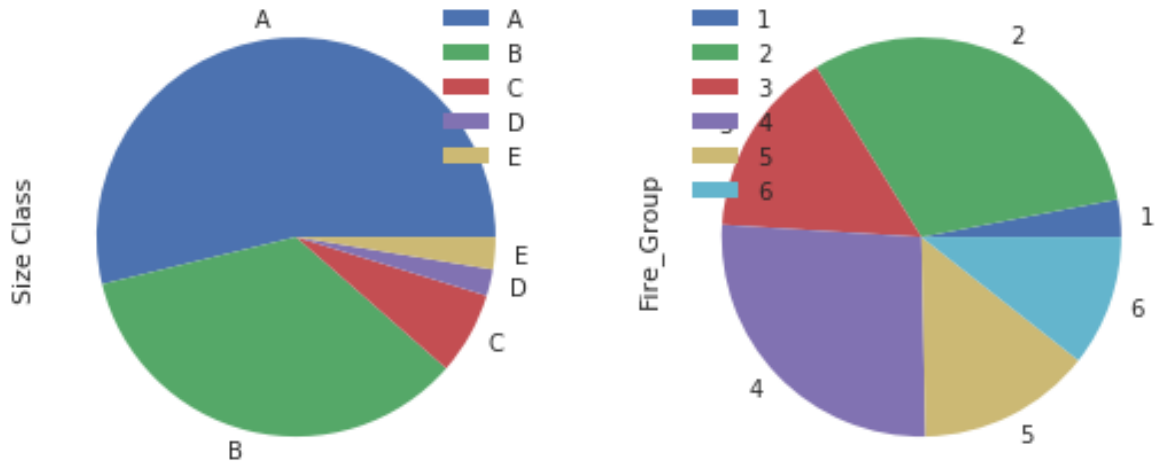
To determine the optimal number of fire groups, I will use silhouette analysis to visualize the separation between the groups. The silhouette coefficient is measure of how close samples in each group are to the group center and to points in the neighboring group. The coefficients range from -1, indicating group misclassification to 1 where clear separation exists between group and a value of 0 means that the sample is close to the boundary between groups. The average silhouette coefficient across all samples is used to assess the quality of clustering with the given number of fire groups.

Below is the result of the silhouette analysis. The number of proposed fire groups ranged from 2 to 9 and the average silhouette coefficient peaked at a value of 0.451 using 8 groups. The low values of average silhouette coefficient across the entire range of groups demonstrate that the data is not densely clustered and/or there is significant overlap between groups. For 5, 8, or 9 fire groups, there is some overlap as there are some negative values of the silhouette coefficient. I think the optimal number of fire groups to use for this analysis is 6. The average silhouette score for 6 fire groups was 0.409 which is only 4% less than the maximum average silhouette score calculated in this range. The

data was the most evenly distributed amongst 6 fire groups and had no negative silhouette scores which are indications for misclassifications. The data was reclassified into six fire groups and compared to the original size class system. The six fire groups more evenly distribute the burned area than the size class system, with no single group accounting for more than half of the reported number of wildfires.



Fire Groups	Average Silhouette Coefficient
2	0.2194
3	0.3068
4	0.3330
5	0.3821
6	0.4098
7	0.4396
8	0.4511
9	0.4369



Wildfire Size Prediction

The other objective of the analysis was to use the observed initial state of the fire and the estimated response times to predict the final size of the wildfire once it is extinguished in order to optimize the use of suppression resources. A regression model will be created for each fire group to predict the extinguished size of the wildfires. The fire group classification along with the time duration data between wildfire states will be used as input into the regression models. The data was sorted by the fire group, standardized using a power transform, and then split it into train/test sets. The data was split into 80/20% sets for the train and test sets, using k-fold cross validation to avoid overfitting the training data.

The first models to be considered are the linear ordinary least squares model, a random forest/decision tree, and a support vector regression machine (SVR). I chose these three models as they are some of the most common regression models. To choose which model to use, I fit a regressor using the default or recommended parameters to each fire groups' training data and calculate the R^2 score and used the average R^2 score across all 6 fire groups as a first pass check on which regression model to pursue further.

From the results below, the linear model performs poorly on the data, with an average R^2 of 44.6%. The random forest and SVR performed comparatively well, around a R^2 value of 77.0% and 73.6% respectively, with the random forest performing more consistently well across the fire groups than the SVR. These two models will be compared again after some hyperparameter tuning.

Fire Group	Ordinary Least Squares R^2	Random Forest R^2	Support Vector Machine R^2
1	0.4486	0.7485	0.6266
2	0.4950	0.7886	0.8031
3	0.4156	0.7570	0.7145
4	0.3449	0.7275	0.7028
5	0.3632	0.7622	0.7346
6	0.6097	0.8368	0.8368
Average	0.4462	0.7701	0.7364

Refinement

The optimization was completed using random parameter search to optimize the regressors as per the process detailed in the documentation. The hyperparameters of both models were varied over a wide range and cross-validated 3 times over the testing data for each fire group. Again, I am using the average R^2 score as measure of performance to compare both models. The results of the tuning resulted in approximately 10% increase in R^2 value in both models, with the random forest performing slightly better. For this reason, I will use the random forest model to predict the final size of wildfires.

Fire Group	Random Forest R^2	Support Vector Machine R^2
1	0.8558	0.9175
2	0.9157	0.8354
3	0.8832	0.7954
4	0.8243	0.7969
5	0.9026	0.8631
6	0.9348	0.8944
Average	**0.8861**	0.8504

The root mean square error is also calculated to quantify the accuracy of the regression model. The RMSE was calculated to be 0.552 or 55.2% of the standard variation of the transformed input data, which is relatively high. To make sense of the actual wildfire size predictions, the inverse transform was used to restore the data to the true scale. The calculated RMSE between all the predicted sizes and the current sizes, regardless of fire group, was calculated to be 7,338 Ha which is high. The mean, median, and maximum values between the predicted and current sizes are also quite different. This may indicate that the regression is not a good predictor of wildfire sizes.

Fire Group	Random Forest RMSE
1	0.7720
2	0.4692
3	0.6456
4	0.4296
5	0.4000
6	0.5954
Average	0.5519

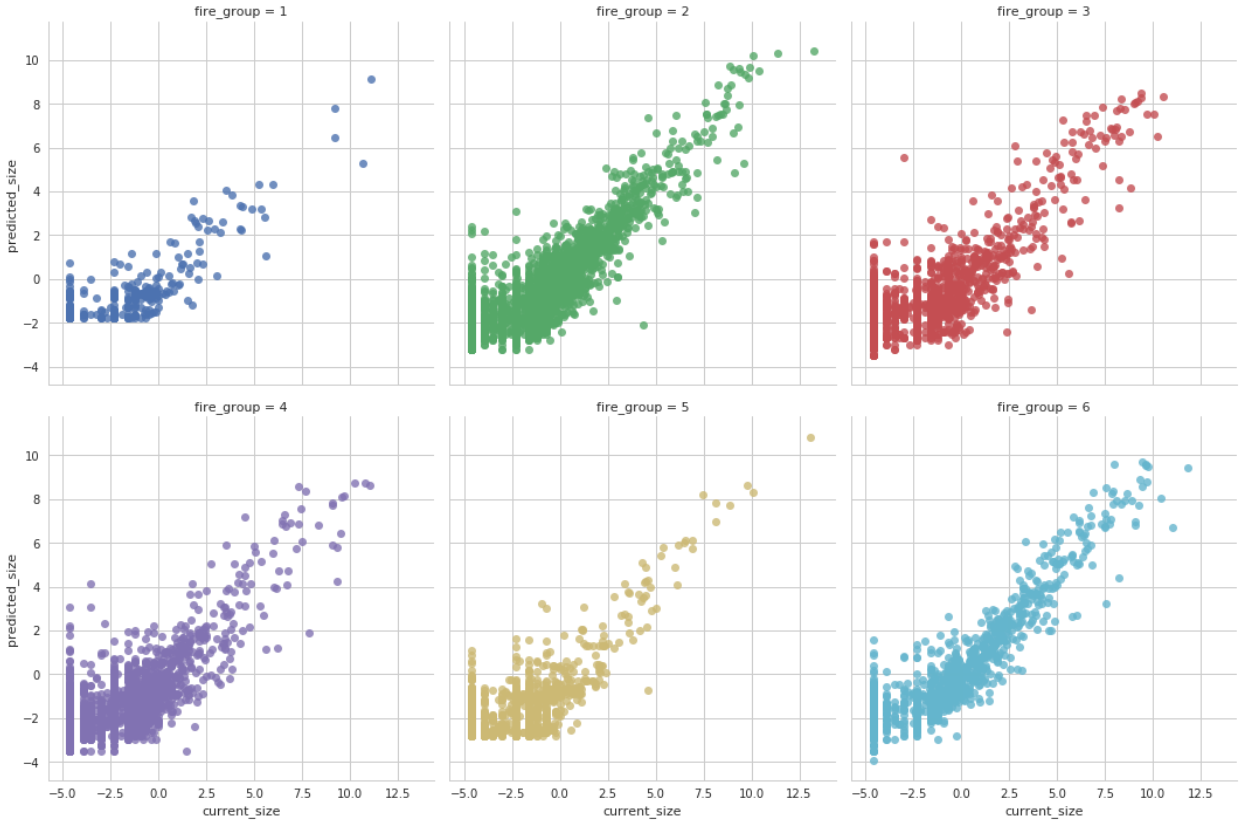
Parameter	Predicted Wildfire Size	Current Wildfire Size
Mean	68.99	268.19
Median	0.22	0.10
STD	970.87	8,014.14

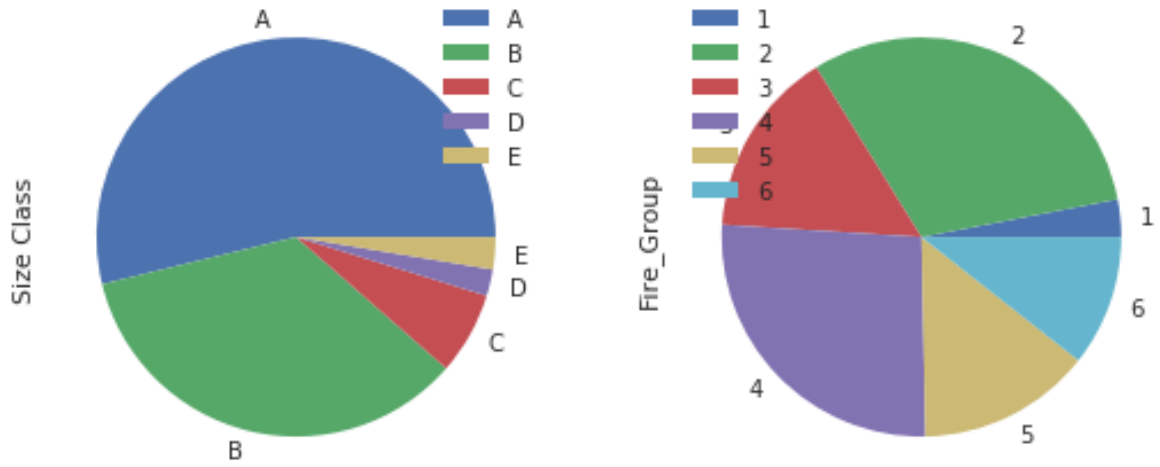
Conclusion

The wildfire data collected by the Alberta Government between 2006 and 2017 was analyzed to create distinct fire groups and used to create regression models to predict the final burned area of wildfires. The fire groups consisted of wildfires with similar initial conditions and detection/suppression resources. The regression models for each fire group used the estimated time between wildfire stages to predict the final size of the wildfire. The intent behind these predictors is that new fires would be put into the appropriate fire groups and suppression resources can be scheduled based on how long the wildfire exists in the wildfire lifecycle stage. For example, an air patrol could discover a wildfire that is defined as a fire group 1 fire. Based on their current resources, an estimated time to get control of the fire and change the state from UC to BH could be used to determine if additional resources are required.

Below are various plots showing the comparison of the predicted wildfire sizes to the actual final sizes of the wildfires for each fire group and the distribution of the data within the fire groups. The scatter plots display the log transform of the data to show more detail of the distribution due to the presence of outliers in the data. The relationship between the predicted and actual wildfire sizes is approximately linear but contains a lot of noise, especially in the smaller sizes.

The pie plot shows the optimized distribution of fire groups in the data, using the assessed conditions of the wildfires. The wildfire data was more evenly distributed amongst the fire groups compared to the size class. This implies that there are many similarities between the wildfire assessed conditions and the distinguishing features used to define the fire groups such as the initial suppression/detection resources, assessment size, and fire type define the boundaries between fire groups. The low silhouette scores for each group also indicated that there is little separation between the fire groups.





Reflection and Improvements

The original intent of the analysis was to use the assessed conditions of wildfires to identify distinguishing features used to create specific fire groups. However, the data provided was quite skewed and concentrated in few categories, like the fuel type, weather conditions, fire slope, fire type, and because of this the principle components that make up the fire groups were defined by few distinguishing features so there was no clear separation between the fire groups. More details in the fire assessments could provide more features to be used to further define the fire groups. Also, separating the physical properties of the wildfires from the utilized suppression resources to create more explainable classifications could have improved the analysis. In this case, the fire group, detection agent type, and initial action resource would be used in another layer of classification or as additional inputs in the regression analysis.

The regression analysis accuracy was calculated to be 89%, approximately equal to the target accuracy of 90%. Despite this, the RMSE was high and there was still a lot of spread in the predicted wildfire sizes. More optimization of the hyper-parameters used in the regression could improve the results, but I think reducing the number of inputs would have a larger effect. Instead of using the time between each wildfire lifecycle stage, maybe using the time between wildfire detection and when it is under control would give more accurate predictions. Another option is to remove the outliers in the size data corresponding to the largest wildfires and only predict wildfire sizes of moderate sizes instead of predicting all wildfire sizes. The regression would probably more closely predict the wildfire sizes and could be used to identify if a given fire would become an outlier with the estimated response times.