DS Term Project: Phase Four

Source for phase 4 is found in /src/machine_learning/.

Introduction

Due to the nature of our dataset there is no appropriate supervised learning task to complete so under the guidance of our TA we opted to complete part C instead of part B. We perform analysis of our data to identify outliers in our data. Beginning with the suggested one-class SVM model, we then move on to other clustering algorithms to try and better identify outliers in our data. The vast majority of the data we inspected for outliers was by province simply due to the massive size of the data. Inspecting by province allowed us to better understand the data and identify outliers more effectively.

Data Preprocessing

The nature of our data is such that some features have intermittent values at best. For example both 'Humidity' and 'Windspeed' have many missing values as not all weather stations across our dataset collected those consistently or with the same sensitivity. As such simply having a value associated with these features can make a datapoint an outlier. We opted to perform clustering across three major tables ['Burns', 'Provinces', 'Weather'] to identify outliers in our data.

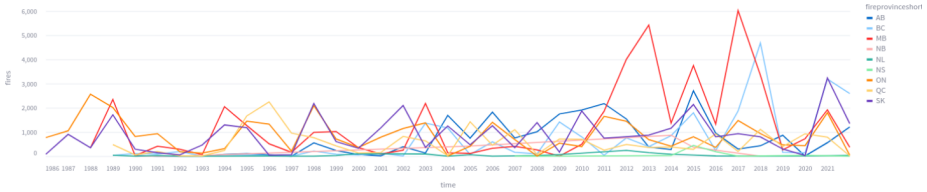
By and large the data was already in a format that was ready for analysis after being processed for database ingestion. We did not need to perform any additional data preprocessing for this phase of the project. We did however limit the information we analyzed to the most relevant features for our analysis. We did not choose to investigate attributes such as the cost of fires, as those would not have been particularly relevant to find outliers within as it is a feature entirely dependent on 'AreaBurned' and the given land cost for the province which would not add noise into the dataset compared to the other features.

NOTE: We opted to perform analysis by and large by province as wildfire characteristics vary greatly by province. For example, Alberta has a much larger area and more wildfires than Nova Scotia. As such we opted to perform analysis by province to better understand the data and identify outliers more effectively, reducing the noise and false positives in our analysis. This is also part of the reason that wildfire management is done provinially in Canada, as the characteristics of wildfires vary greatly by province.

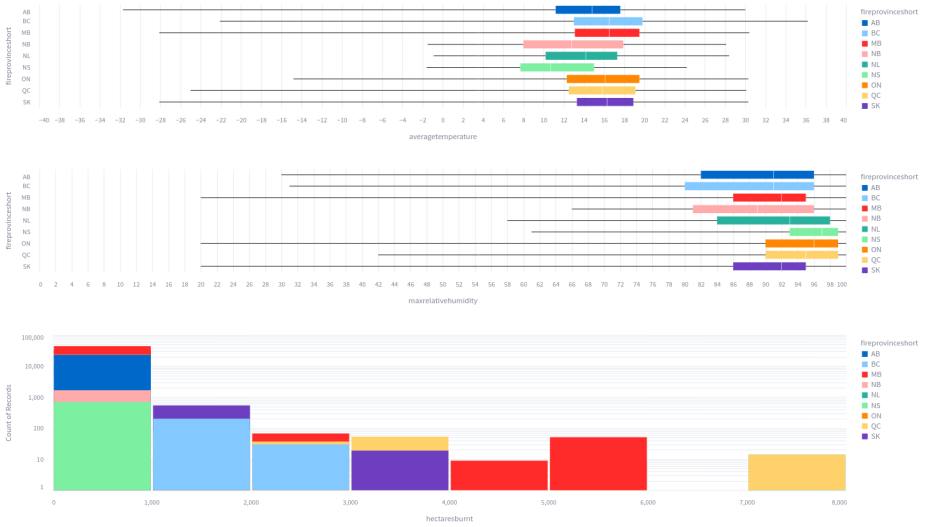
EDA

We ran some simple Exploratory Data Analysis to better understand the data we were working with. We looked at the distribution of the data and the relationships between the features.

When it comes to distribution of data we could utilize part of the infrastructure of the Dashboard created in the phase 3 and look at boxplots of different aspects of the data. We already knew from the graph below that the report rate of different years for fires may be the cause of batch-reporting which contribute sto outlier fires of much larger scales than the majority:

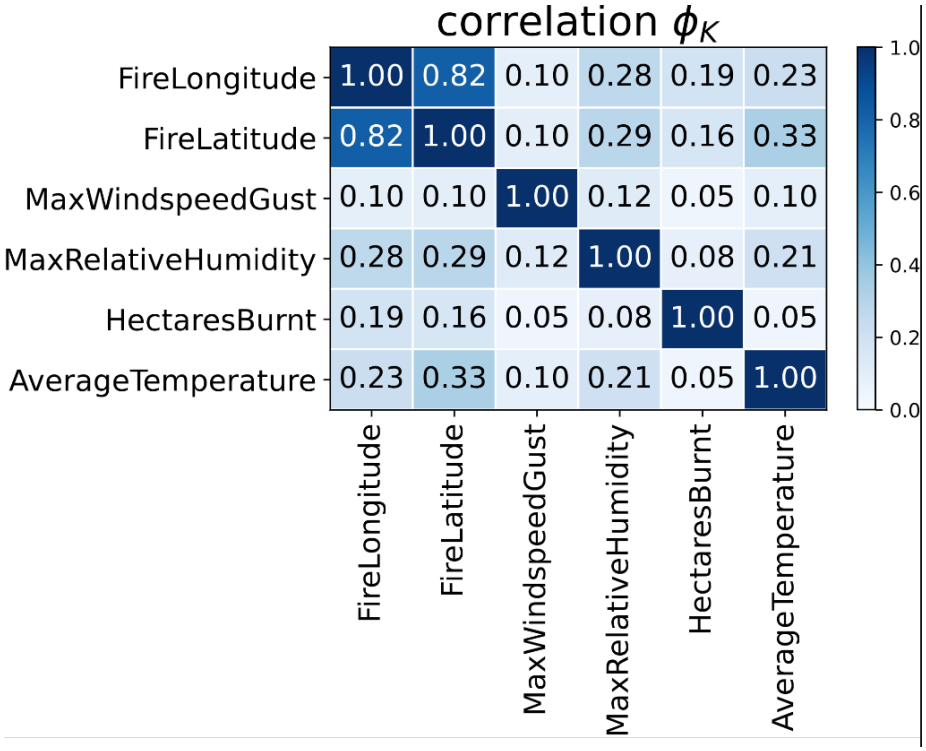


Looking at two more apt dimensions to investigate are average temperature of fires, humidity and hectares burnt for fires.



Note specifically the log-scaling of the last graph. This showed the suspected large distribution of data which makes outlier detection something which should be very fruitful for this data collection. Provincial differences are also wide which will prove important for the future analysis.

In terms of correlations and relations between dimensions we found that, sadly, the data was very simple and did not have many complex relationships between the features. Spefically we investigated PhiK correlation heatmaps to identify any relationships between the features. This method uses Pearson correlation to identify how closely variables are related. There were no strong relationships between the features in our data as seen in the picture below. Further analysis using regression based methods would likely be more useful but are out of scope.



PhiK correlation report can be found here: PhiK Correlation Report

One-Class SVM

The one-class SVM's can be used in one of two ways, outlier detection and novelty detection. In our case we are using it for outlier detection where the estimator attempts to fit to regions where the training data is the most concentrated. The model then predicts whether a sample is an outlier or not. The SVM does this by projecting points into higher dimensional space. The one-class SVM will then separate the data into two classes, inliers and outliers using a hyper-sphere that contains the inliers. The model will then predict whether a sample is an outlier or not.

In our case initial results from the one-class SVM model were not very promising. When working with finding positional outliers in our data, the one-class SVM model was not able to identify outliers effectively. Shown below the issue is that the estimator simply takes the centroid of the data and creates a hyper-sphere around it. When mapping back into the feature space this results in a large number of points seperated incorrectly by a simple distance from the mean rather than true outliers. We hypothesize that a One-Class SVM would be more effective in identifying outliers in a dataset with more features, more complex relationships between the features, and more noise. In this case the data is

simply too simple for the one-class SVM to be effective, especially when looking at positional outliers.

Poor Results of One-Class SVM Example

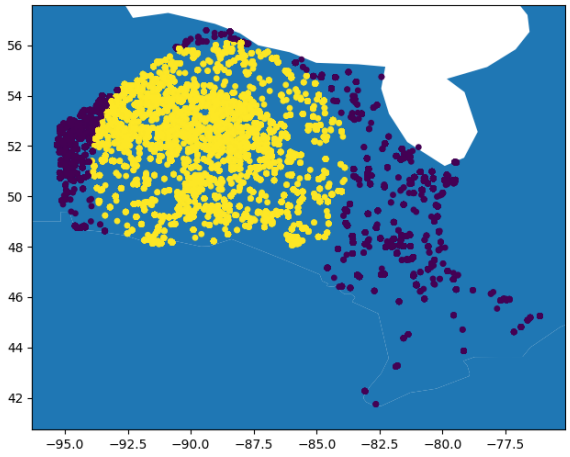


Figure 1: bad_one_class_svm_ontario

DBSCAN

DBSCAN is a clustering algorithm that groups together points that are closely packed together. The algorithm works by grouping points that are closely packed together and marking points that are in low-density regions as outliers. The algorithm works by defining two parameters, epsilon and min_samples. Epsilon is the maximum distance (lat/long degrees) between two samples for one to be considered as in the neighborhood of the other. Min_samples is the number of samples in a neighborhood for a point to be considered as a core point. By varying these two parameters we can identify outliers in our data most effectively.

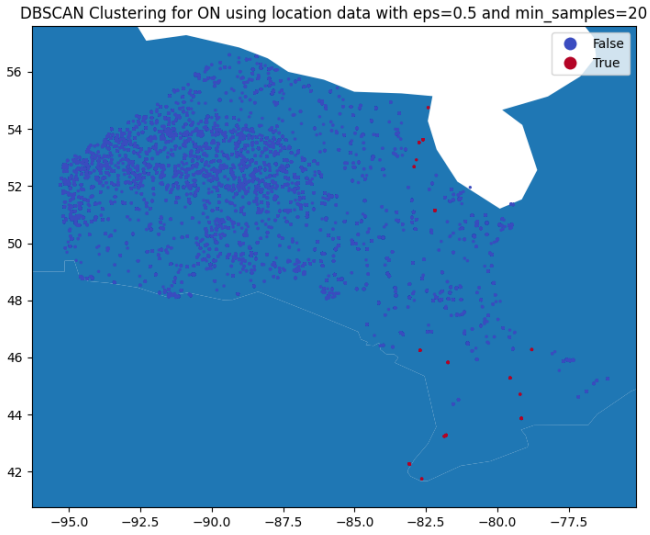
We looked at burn data across several dimensions including [‘HectaresBurnt’, ‘AverageTemperature’, ‘FireLatitude’, ‘FireLongitude’, ‘ProvinceID’]. Clustering with the same parameters will perform with different degrees of effectiveness depending on the dimensionality of the data. We note several interesting results from our clustering analysis. The results of this outlier analysis were much better, finding true outliers within the data.

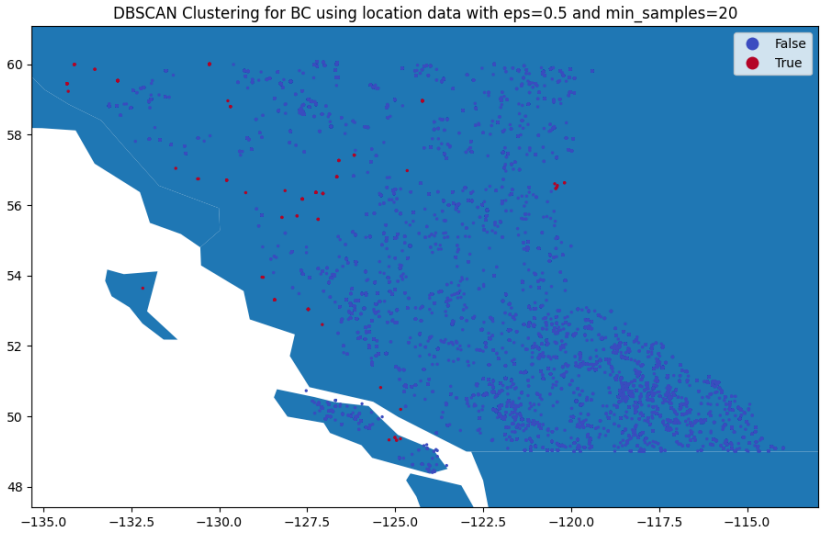
Investigating the actual data we found that the outliers identified by the DBSCAN algorithm were in fact true outliers. The DBSCAN algorithm was able to identify outliers in the data that were not simply due to the data being far from the mean.

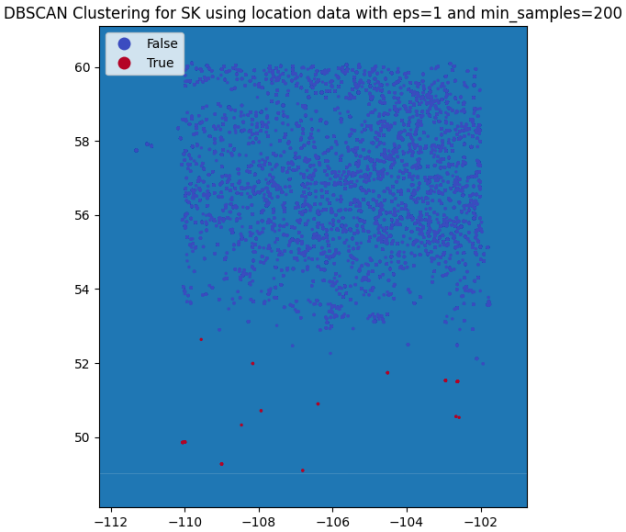
This included clusters of fires that occurred near Toronto for example. Given the nature of wildfires it is anomalous for a fire to occur in a major city like Toronto. The DBSCAN algorithm was able to identify these outliers effectively.

When including temperature in the clustering analysis we found that the DBSCAN algorithm was able to identify outliers across provinces that occured in likely locations but were not typical in other ways, such as a fire when the temperature is near freezing. Continuing to add variables when including the burn area size DBSCAN identified more outliers. Specifically, in Nova Scotia (as well as other provinces) we were able to identify a number of fires that fell into one of two categories, out of season fires, and very small in season fires. Both of these categories are outliers. The former intuitively makes sense as fires are not expected to occur in the winter, and if they are, they are expected to naturally be very small. The latter also makes sense as during hot dry seasons fires are expected to be able to grow to a certain minimum size very quickly before being addressed. Nova Scotia was picked as an example for the latter category as it provides a good example of how a smaller province can address fires much more quickly and effectively than a larger province like Alberta where it may be several days before a fire can be addressed.

Location Based Outliers Examples

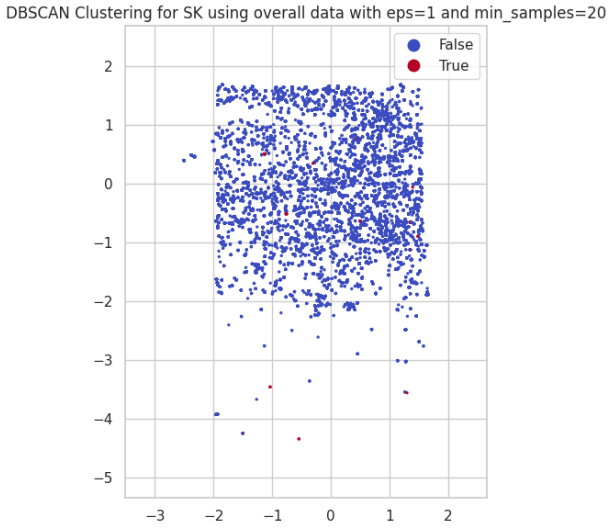


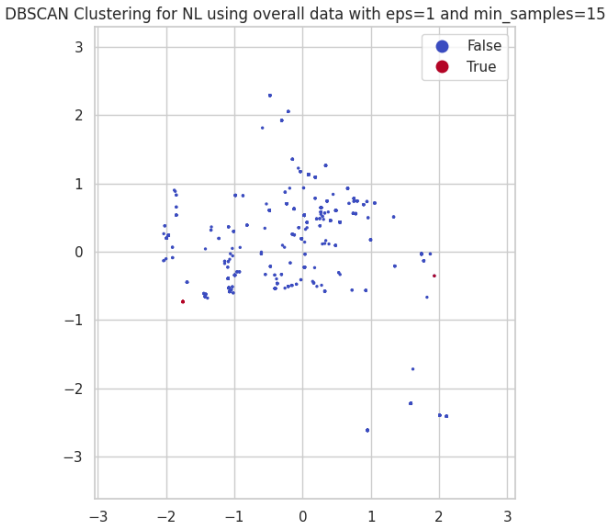




Temperature and Location Based Outliers Examples

In this case the input to the DBSCAN algorithm was normalized. The overlapping anomalies in the data were identified as anomalous due to their location. The visually identifiable outliers on the other hand were identified due to temperature and burn area, the two opposing categories mentioned above.




DBSCAN Clustering for NL using overall data with eps=1 and min_samples=15
False
True
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3