**Data Summarization**

We had an idea that we wanted to classify the outcomes of the fires as either good or bad. Depending on the number of casualties, the number of people displaced and the estimated damages in CAD$ so those were the attributes we evaluated.

We began by looking at the estimated damages in CAD dollars, we chose to use a boxplot, as can be seen in the following graph

Chart

Description automatically generated

However the outliers prevent the values in the boxplot from being distinguishable so, a version without the dots representing the outliers was created to get a better idea of scale, as can be seen here, the first quantile consists of estimated damages being equal to 0$:

Chart, box and whisker chart

Description automatically generated

Next, we looked at the casualties and found that the vast majority of the data fell at 0 casualties, skewing us towards good outcomes. As is shown in the following graph:

Chart, histogram

Description automatically generated

Similar findings were observed for the number of people displaced, where the majority of cases have 0 people displaced but here the outliers are much more drastic, as can be observed in the histogram and boxplot below:

Chart, histogram

Description automatically generatedA picture containing scatter chart

Description automatically generated

Upon finding that both of the above typically have low values, we wanted to see if they typically matched, that is to say that having 0 casualties usually meant having 0 people displaced, so we plotted the sum of the two measures to see if we would have more numbers above 0 or if the sum quantity of 0s would stay roughly consistent. The results can be found in the figures below:

Chart, box and whisker chart

Description automatically generatedA picture containing chart

Description automatically generated

Chart, histogram

Description automatically generated

**Feature Selection**

The features that we selected for our classification algorithms included:

|  |  |  |
| --- | --- | --- |
| **Dimension** | **Features** | **Notes** |
| Location dimension | Longitude  Latitude | This dimension contained other data such as the postal code, nearest intersection, and dissemination area id. We decided to use only the latitude and longitude because they were the most precise numerical values to use in the AI and the other location data would have been redundant anyway. |
| Date dimension | Month  Day  Day of week  Holiday  Time of day | This dimension contained other data as well such as the exact timestamp of the previous fires. Since those exact timestamps will never occur again, they are likely to overfit the AI to the training data. Instead, we only used the date features that are repeatable. |
| Weather dimension | Temperature  Relative humidity  Precipitation  Snow  Wind direction  Wind speed | We used almost all the features from this dimension because they could all potentially have a huge impact on how frequently fires can start and how damaging they may be. |
| Demographic dimension | Population  Median age  Total dwellings  Average household size  Median household income  Mother tongue official percentage  Mother tongue unofficial percentage | We used almost all the features from this dimension because, again, they could all potentially have an impact on the frequency and damage of fires. |
| Fire ward dimension | Stations in ward | This is the only feature in this dimension and it could have an affect on the response time to fires, thus limiting the damage. |

We left out all the measures from the fact table itself because most of them were measures that could only be determined during or after a fire. Since the purpose of our AI is to predict if some conditions could result in a ‘bad’ or ‘good’ fire, we can’t use any variables that occur during or after the fire to predict this. Instead, we categorized all the fires as bad or good based on the number of casualties, the number of people displaced, the damage in Canadian dollars, and the response time. A fire was considered bad if the total number of casualties plus the number of people displaced was 5 or more, the damage was $10,000 or more, or the response time was more than 20 minutes. A fire was considered ‘good’ or ‘acceptable’ otherwise.

In addition, we performed under-sampling of the bad outcomes because there was about a 3:1 ratio between the bad and good outcomes. We reduced the number of bad outcome samples to equal those of the good outcomes.

**Imputation**

Out of the features that we chose, there was only one that had any missing values. The holiday feature was mostly missing because only a small percentage of the fires in the data occurred on holidays. Since over 95% of the data was missing for that feature, we created a new value for it called ‘Not holiday’ and filled those missing values. The figure below shows the number of missing values in each column before imputation.

Text

Description automatically generated

**Handling categorical attributes**

When viewing the features, most of them were numerical, but holiday and time of day where categorical. Holiday contained the name of the holiday that was occurring and the time of day could have been night (midnight to 6am), morning (6am to noon), afternoon (noon to 6pm), or evening (6pm to midnight). We used one-hot encoding to split both of those into multiple columns with values of 0 or 1. This next figure is the type of data that is held by each column before the one-hot encoding.

Text, timeline

Description automatically generated with medium confidence

The following figure is the type of data that is help by each column after the one-hot encoding.

Text, timeline

Description automatically generated

**Normalization**

We normalized all the columns independently so they would be worth the same to the classification algorithms later. The following image is a sample of a few of the features after normalizing.

Text

Description automatically generated

**Classification algorithms**

After running the algorithms multiple times with different data subsets, we noticed some patterns emerging. The decision tree was always the fastest to train, followed by the random forest. The gradient boosting was always the slowest. By contrast, the accuracy was always the opposite. The gradient boosting was always the most accurate while the decision tree was the least accurate. The following three images show a classification report for each of the three different AI types.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated