Mining Toronto Fire Services Incident Data

CKME 136 Capstone Project

Prepared for Dr. Ceni Babaoglu
Ryerson University
by Geoffrey Clark
July 14th, 2018
Project repository: https://github.com/gffryclrk/toronto.fire.incident.data

Outline

- Data Preparation
  - Import Data
  - Data Description
- Initial Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
- Exploratory Analysis
- Dimensionality Reduction
- Experimental Design
- Modeling
- Evaluation
- Improving the Model
- Conclusions

Data Preparation

Import Data

The Toronto Fire Services Incident Data was obtained from the Toronto Open Data Catalogue in May, 2018. Observations represent a single incident involving Toronto's Fire Services including Medical, Fire and other types of events. Time data includes the years 2011 through 2016 and is provided as XML documents with the following nested tree structure:

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<FIRE_DATA>
      <INCIDENT>
            <RespondingUnits>
            </RespondingUnits>
      </INCIDENT>
</FIRE_DATA>
```

The first step in this project was to convert the dataset to a usable rectangular structure. XQuery was used for the purpose of extracting INCIDENT observations from the XML documents as .CSV data tables for each year. The breakdown of INCIDENT observations by year is available in **Figure 1**. This

chart shows a consistent trend of 110,000 - 120,00 for all years in the data set except for 2011, the first year of the series, which is higher with around 145,000.

Data Description

The INCIDENTS data set has the dimensions of 720,370 observations across 100 features. A detailed description of each individual feature is provided in Appendix-1. Generally speaking the majority of the variables are categorical and have been provided by the Office of the Fire Marshal's (OFFM) Standard Incident Report Codes List or the RMS System.

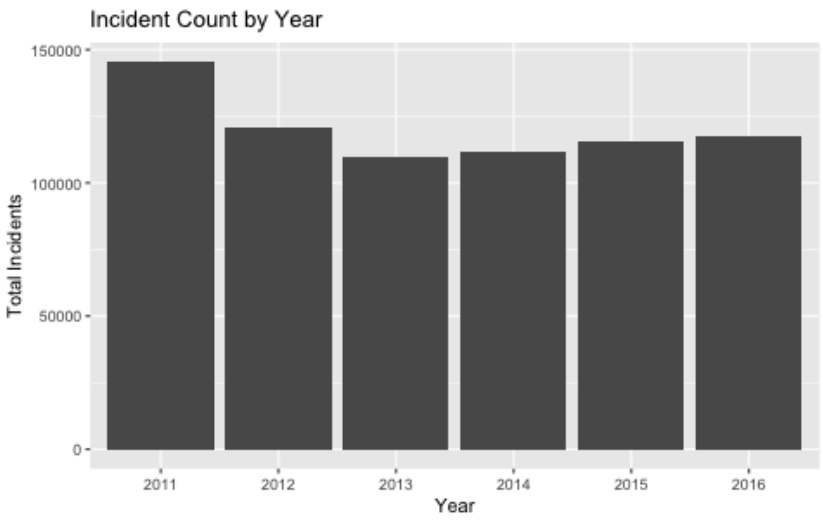

Figure 1

The features for all years are the same except for 2016 which includes an additional feature: EVENT_ALARM_LEVEL. This feature corresponds to the highest alarm level of the event and, since is only available in 2016, was discarded for the use in this project.

Initial Analysis

Univariate Analysis

As mentioned in the Data Description the data set consists primarily of categorical features across approximately 100 features. The first task involves encoding the features in meaningful, useful way.

Event Type & Property Type

Event types correspond to the type of incident being responded to. The 114 distinct event type levels can be assigned to 14 distinct groups (in descending order of incidents per group): Medical, Alarm, Vehicle, Fire, Carbon Monoxide, Rescue, Check Call, Utility, Natural Gas, Hazmat, Police Assist and Other.

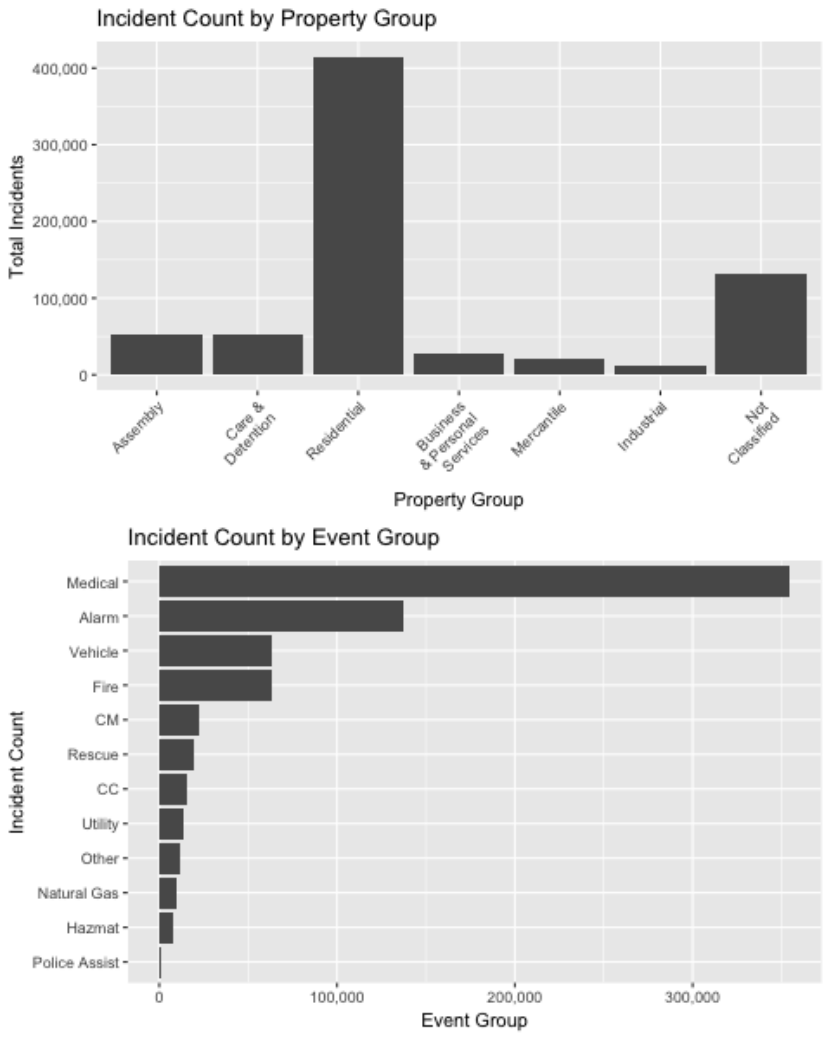

Figure 2

Property types are encoded in the data set according to the OFM Standard Incident Report Codes List. The PROPERTY attribute itself includes 350 unique categorical variables with 9863 NA (1.37%). The OFM Standard Incident Report Codes list also includes 7 distinct category groupings for these categories: Assembly, Care & Detention, Residential, Business & Personal Services, Mercantile, Industrial and Other (Structures/Properties not classified by the Ontario Building Code). **Figure 2** displays Incident Count breakdowns by Property & Event Group.

Date & Time

The data set includes several temporal features pertaining to the time of initial call, responding units' dispatch time, incident on scene arrival time and the corresponding dates. There is also date and time data pertaining to when control of the incident was obtained.

Interesting patterns can be observed from the dates and times. A breakdown of the INITIAL_CALL_HOUR is represented in **Figure 3**. It is observed that incident calls begin to be received around 8am and

increase steadily until around 7pm where they begin to taper off and remain relatively low during the night. This could be attributed to reduced activity both at home, in the workplace and on roads.

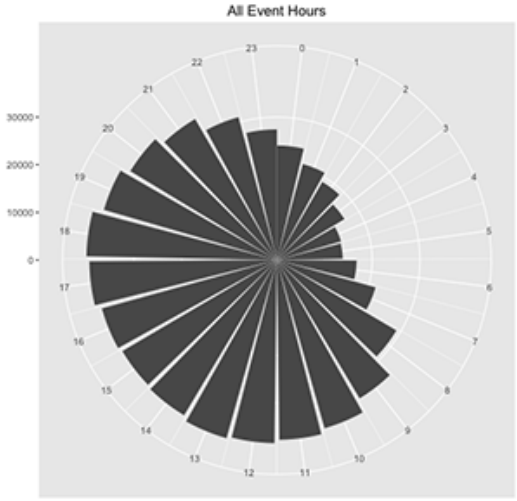

Figure 3

Figure 4 reveals a daily breakdown of all incident call hours. It is observed that weekdays experience similar hourly call distributions as the aggregate but that the weekends, Saturday & Sunday, deviate from this norm and experience higher relative call volumes. This is most prominent in the relative call volumes on Sunday between the hours of Midnight and 3am. Such a deviation is perhaps attributed to a difference in activity and behaviour on weekends among individuals and groups.

The data set includes INITIAL_CALL, DISPATCH, and ONSCENE times each encoded as HOUR, MIN and SEC as separate features. Combining these times with the corresponding date values enabled the creation of a TTA (Time-to-arrival) feature that is the difference between INITIAL_CALL and ONSCENE times.

Injuries and Fatalities

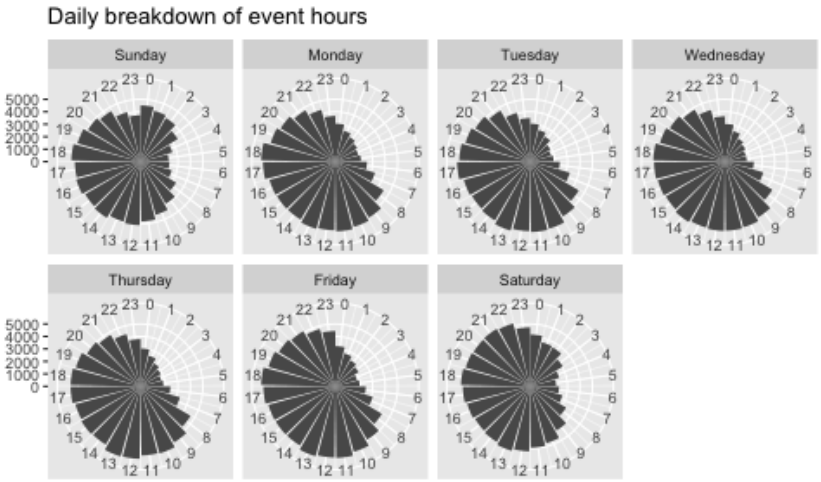

Figure 4

The data set includes counts of fatalities for both Fire Fighters and Civilians. However, the count of fire fighter fatalities is 0 for every observation and is thus dropped (low variance filter). There are 891 incidents that had one of the three remaining injury or fatality features: 213, 51 and 653 non-zero observations for fire fighter injuries, civilian fatalities and civilian injuries respectively. There is some overlap in the data: the combinations of events with two of the three features ranges from 5 to 18 but there are only two observed incidents that had non-zero entries for firefighter injury, civilian injury, and civilian fatality.

Removing Features

ARRIVE_DATE & FF_FATALITIES are mentioned previously...

Some features in the data set are either redundant or not able to be used. These features are removed. An example of a redundant feature would be one of DISPATCH_DATE and ARRIVE_DATE: since these two features are identical there is no lost information by removing one. In this analysis ARRIVE_DATE was arbitrarily chosen to be removed. FF_FATALITIES is an aptly named feature which records the

quantity of fire fighter fatalities which occurred during the incident. However, every observation is 0 and so this feature is dropped also.

A further questionable feature is FD_STATION. This is reportedly the "Fire Department Station", however, the data is inconsistent. To begin with, there are 375,478 unique entries whereas the Toronto Website (https://www.toronto.ca/community-people/public-safety-alerts/understanding-emergency-services/fire-station-locations/) only lists 85 different Fire Stations. Further, in all but 30 observations this feature is either the INCIDENT_NUMBER, the INCIDENT_NUMBER padded with trailing zeroes, or simply zeroes. Accurate fire department station data would be useful for an analysis but since this column doesn't appear to contain such data it is safely dropped.

Bivariate Analysis

Alarm to Fire Department, Event Group

The Alarm to Fire Department feature, ALARM_TO_FD, describes from where the incident was reported. EVENT_GROUP is a grouping of different types of event. The features are related in the sense that ALARM_TO_FD is the input source (911, Police, etc) whereas EVENT_GROUP describes what type of incident the call turned out to be. Perhaps a reasonable expectation would be that certain types of calls correspond to certain types of events. **Figure 5** shows the portions of Event Type group by Call source.

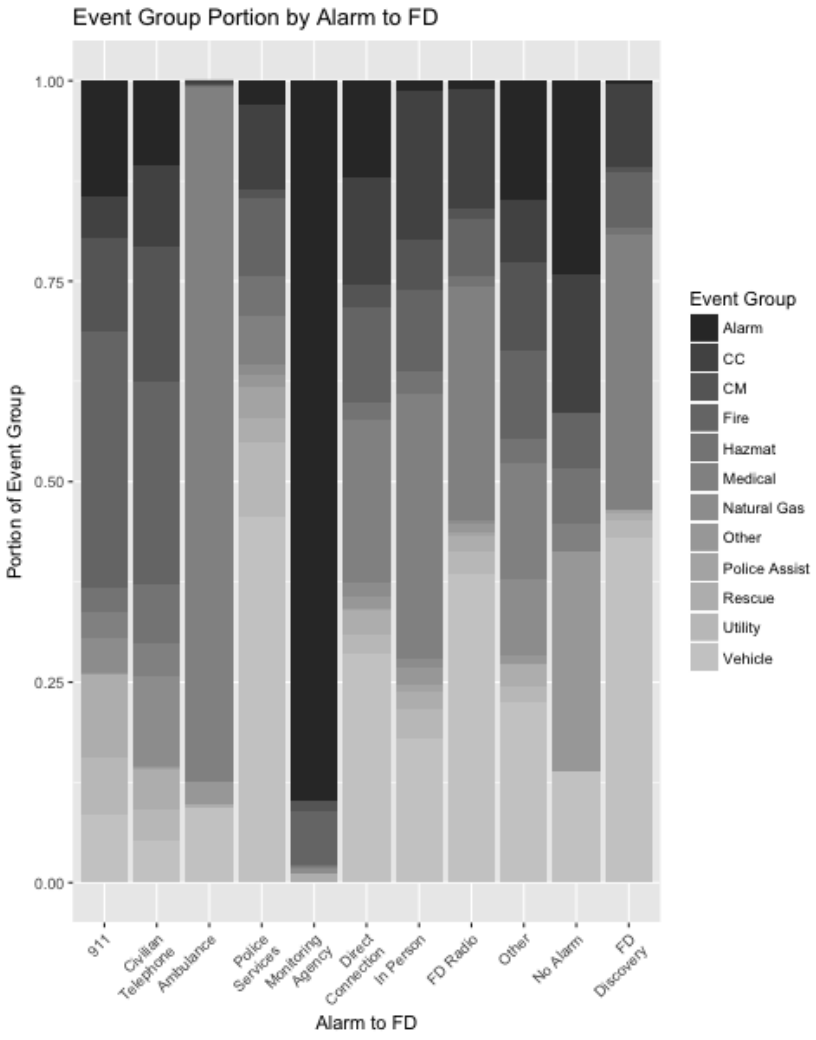

Figure 5

To test the hypothesis that EVENT_GROUP and ALARM_TO_FD are independent a Chi square test of independence was used. The test of hypothesis was as follows:

H_0: EVENT_GROUP and ALARM_TO_FD are independent

H_1: EVENT_GROUP and ALARM_TO_FD are dependent

The result of Pearson's Chi-Square test of independence concludes that there is statistically significant evidence to reject the null hypothesis that EVENT_GROUP and ALARM_TO_FD are independent. This conclusion supports the notion that the source of incident call impacts the probability of incident type.

Note: I tried a few different settings using R, both shown below. I experimented with merging columns so that there were fewer 0s (notably in columns 10 & 11, both of which are similar categories. None of my attempts made much difference: if I set 'simulate.p.value' = T then degrees of freedom would be NA (df = NA). Alternatively, if I didn't set this parameter in the chisq.test() method call it returned a df = 121 but gave me a warning that the Chi-squared approximation may be incorrect. I am confused about this and then was informed, by Dr. Ceni, that this might have been an improper test to use in this scenario.

```
> chisq.test(table(I$EVENT_GROUP, I$ALARM_TO_FD), simulate.p.value = T)

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(I$EVENT_GROUP, I$ALARM_TO_FD)
X-squared = 1091400, df = NA, p-value = 0.0004998


> chisq.test(table(I$EVENT_GROUP, I$ALARM_TO_FD))

        Pearson's Chi-squared test

data:  table(I$EVENT_GROUP, I$ALARM_TO_FD)
X-squared = 1091400, df = 121, p-value < 2.2e-16

Warning message:
In chisq.test(table(I$EVENT_GROUP, I$ALARM_TO_FD)) :
  Chi-squared approximation may be incorrect


> table(I$EVENT_GROUP, I$ALARM_TO_FD)

                  1     2      3     4      5     6     7     8     9    10    11  <NA>
  Alarm       21346  1748    128   605 113377   143    52    36   322     7     2     0
  CC           7854  1698   1754  2219    129   162   725   503   171     5    45     1
  CM          17199  2811    327   197   1545    33   246    42   238     0     3     0
  Fire        47269  4171    315  1993   8439   143   395   235   241     2    31     0
  Hazmat       4483  1227    690  1049    169    26   110    45    67     2     3     0
  Medical      4964   697 344383  1270    431   241  1301   984   316     1   151     3
  Natural Gas  6321  1849     60   256    810    22    39    16   206     0     0     0
  Other         237    48  11088   293     74    18    85    33    22     8     0     0
  Police Assist 221    24     15   803      4     1    39    15     2     0     2     0
  Rescue      15205   850   1136   643   1156    37    86    65    60     0     4     0
  Utility     10561   623     53  1924     63    29   143    96    42     0    10     0
  Vehicle     12590   880  37338  9419    126   341   703  1293   490     4   189     0
```

1	911	5	From Monitoring agency	9	Other alarm
2	Telephone From Civilian (other than 911)	6	Direct Connection	10	No alarm Received - No response
3	From Ambulance	7	Verbal Report to Station (in person)	11	No alarm rcv'd - incident discovered by FD personnel
4	From Police Services	8	Two-way radio (fire department)		

Table 1: ALARM_TO_FD Levels

Event Group, Time to Arrival

The box plot in Figure 6 depicts Time to Arrival against Event Groups. This relationship might suggest different response times depending on the type of incident. All Event Group categories, save Medical, are statistically significant predictors of Time-to-arrival as displayed in the below linear model summary.

```
Call:
lm(formula = TTA ~ EVENT_GROUP, data = I)

Residuals:
   Min     1Q Median     3Q    Max
  -973    -75    -18     46  82928

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               301.248      1.397 215.626  < 2e-16 ***
EVENT_GROUPCC             138.562      4.490  30.861  < 2e-16 ***
EVENT_GROUPCM              79.791      3.721  21.441  < 2e-16 ***
EVENT_GROUPFire            60.224      2.495  24.137  < 2e-16 ***
EVENT_GROUPHazmat          83.990      6.025  13.939  < 2e-16 ***
EVENT_GROUPMedical          2.218      1.652   1.343  0.17941
EVENT_GROUPNatural Gas     45.950      5.480   8.384  < 2e-16 ***
EVENT_GROUPOther           15.271      5.013   3.046  0.00232 **
EVENT_GROUPPolice Assist  201.005     15.954  12.599  < 2e-16 ***
EVENT_GROUPRescue          45.990      4.086  11.256  < 2e-16 ***
EVENT_GROUPUtility        673.128      4.684 143.698  < 2e-16 ***
EVENT_GROUPVehicle         54.480      2.533  21.506  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 518.2 on 704908 degrees of freedom
  (15450 observations deleted due to missingness)
Multiple R-squared:  0.03235,   Adjusted R-squared:  0.03233
F-statistic:  2142 on 11 and 704908 DF,  p-value: < 2.2e-16
```

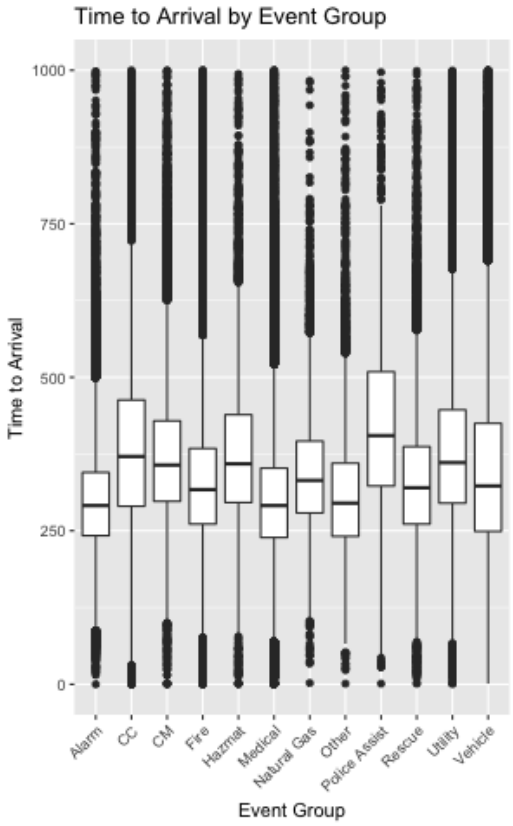

Figure 6

Multivariate Analysis

Fire Classification

In *Modeling the risk of structural fire incidents using a self-organizing map* Asgari et. al classify fire risk level into 5 distinct categories based on fatalities, injuries and damage. The levels are defined as follows:

- **Very Low**: 0 Fatalities, 0 Injuries, Damage < $5000
- **Low**: 0 Fatalities, 0 Injuries, $5000 <= Damage < $10,000
- **Moderate**: 0 Fatalities, 0 Injuries, $10,000 <= Damage < $50,000 *or* 0 Fatalities, 1 Injury, Damage < $50,000
- **High**: 0 Fatalities, Injuries <= 1, $50,000 <= Damage < $100,000 *or* Injuries > 1, Damage < $100,000
- **Very High**: Fatalities > 0 *or* Damage >= $100,000

Table 2 contains a yearly breakdown of incident classification by year and displays relatively consistent class levels for all years in the data set. However, there is a major class imbalance problem when classifying fires in this way. The 'Very Low' level contains 99.2% of the incidents. Such discrepancy between classes introduces potential problems with machine learning algorithms and classifiers (KrishnaVeni, Rani 2011).

Year	VL	L	M	H	VH
2011	144517	235	399	102	112
2012	119651	223	442	103	126
2013	108675	224	422	113	142
2014	110805	229	488	126	146
2015	114749	218	451	118	128
2016	116437	233	483	123	150
Total	714834	1362	2685	685	804

Table 2

Class imbalance problems such as this are a prevalent problem in machine learning and thus potential solutions have been the subject of various research (López et al.). Techniques to deal with this issue include undersampling, oversampling or hybrid methods which combine both. Synthetic Minority Oversampling TEchnique (SMOTE) and Cost-Sensitive Learning are two popular techniques to deal with class imbalance issues.

Class Imbalance: Binary Class & Features

It is possible to consider additional features and thresholds in the above classification in order to help reduce the class imbalance problem. The following are some potential suggestions:

1. Reduce class levels to binary, *Critical* or *Non-critical*: In the previous classification *Very Low* is considered *Non-critical* whereas *Low, Medium, High, and Very High* are considered critical.
2. Damage (EST_LOSS): There are a total of 10,480 incidents with Damage > 0. Perhaps all of these should be considered *critical* incidents.
3. Responding Units: The 99th percentile of RESPONDING_UNITS is 7 units. Perhaps consider any incident with seven or more responding units *critical.*
4. Rescues: There are 7907 incidents where RESCUES > 0.
5. Office of the Fire Marshall Investigations: there are 1144 observations in the data set where OFM_INVESTIGATIONS_CONTACTED is TRUE.
6. Estimated value at risk: This data set also contains a feature EST_VALUE_AT_RISK which is categorical and contains levels pertaining to dollar value estimates. There are 9565 observations with non-zero value estimates for this feature.

After implementing the above features the class imbalance problem is reduced to a 95.9 to 4.1% ratio. This is a significant improvement with absolute values depicted in **Table 3**.

Class Imbalance: Subsetting

	Critical	Non-critical
Count	29782	690588
%	4.1	95.9

Table 3

The class levels described by Asgari et al.—applied to this data set—were originally proposed for fire incidents. Although not exclusive, most features used to describe the fire classification levels pertain to fire related incidents. Fire incidents account for only 8.8% of the incidents but contain 74% of the *Low* through *Very High* incidents. Vehicle incidents account for most of the remaining *Low, Very Low, Medium, High and Very High incidents.* **Table 4** illustrates this relationship.

	VL	L	M	H	VH
Alarm	137691	24	42	5	4
CC	15243	11	10	0	2
CM	22639	0	2	0	0
Fire	59153	848	1888	598	747
Hazmat	7866	2	2	0	1
Medical	354719	0	22	1	0
Natural Gas	9574	2	3	0	0
Other	11895	1	5	3	2
Police Assist	1126	0	0	0	0
Rescue	19238	0	3	0	1
Utility	13541	0	3	0	0
Vehicle	62069	474	705	78	47

Table 4

By subsetting the data to include only fire, or perhaps fire and vehicle, incidents the class imbalance problem can be reduced significantly. Furthermore, by applying subsetting and the aforementioned binary class and augmented threshold techniques previously mentioned together the ratio of *Non-critical* to *critical* incidents can be further balanced to 79% and 21% respectively.

Association Rules

Association rules were run on all observations of the following features: EVENT_GROUP, ALARM_TO_FD, RESPONSE_GROUP, PROPERTY_GROUP, MONTH, DAY, HOUR, EVENT_TYPE_CD. No rules were found with the three temporal features (MONTH, DAY, HOUR), however, other interesting patterns were discovered.

#	Rules	Support	Confidence	Lift	Count
1	{EVENT_GROUP=Alarm,PROPERTY_GROUP=C} => {ALARM_TO_FD=5}	0.1019	0.8294	4.7289	73398
2	{EVENT_GROUP=Alarm} => {RESPONSE_GROUP=E}	0.1564	0.8178	3.9498	112659
3	{EVENT_GROUP=Alarm,ALARM_TO_FD=5} => {RESPONSE_GROUP=E}	0.1282	0.8145	3.9340	92345
4	{RESPONSE_GROUP=I,EVENT_TYPE_CD=Medical} => {EVENT_GROUP=Medical}	0.1920	1.0000	2.0307	138287
5	{ALARM_TO_FD=3,EVENT_TYPE_CD=Medical} => {EVENT_GROUP=Medical}	0.2123	1.0000	2.0307	152920

Table 5: Association Rules & Corresponding Support, Confidence, Lift and Total Count.

1. {Event group = Alarm, Property Group = Residential} => {Alarm to Fire Department = From Monitoring Agency}

This rule describes incidents where the Fire Services responded to Residential Alarms. Of these alarms, 82% (confidence) were received from a monitoring agency.

2. {Event group = Alarm} => {Response Group = False Fire Calls}

This simple statistic, also with high lift, shows that 81% of the alarms that the fire services respond to are false fire calls. This is an incredibly high percentage and represents approximately 15% of all incidents in this dataset.

3. {Event group = Alarm, Alarm to Fire Department = From Monitoring Agency} => {Response Group = False Fire Calls}

This rule elaborates on rule #2 by showing that a high percentage (~ 82%) of aforementioned false alarm calls are received from a monitoring agency.

4. {Response group = Medical/resuscitator call, Event type call as dispatched = Medical} => {Event group => Medical}

Letter	Response Group
A	Property Fires/Explosions
B	Overpressure rupture/explosion (no fire)
C	Pre fire conditions/no fire
D	Burning (controlled)
E	False fire calls
F	CO False calls
G	Public Hazard
H	Rescue
I	Medical/resuscitator call
J	Other response

Table 6: Response Group Levels

Letter	Property Group
A	Assembly
B	Care and Detention
C	Residential
D	Business & Personal Services
E	Mercantile
F	Industrial
O	Structures/Properties not classified by the Ontario Building Code

Table 7: Property Group Levels

This rule perhaps seems obvious but is worth noting: with a confidence of 100%, all incidents responded to as medical calls, which were dispatched as medical calls, ended up being classified as medical calls. This data set has a high proportion of medical calls. Note that only 138,000 of the some 350,000 medical incidents in this data set fit that pattern. This suggests that there is some variability among these features of medical calls.

5. {Alarm to Fire Department = From Ambulance, Event type call as dispatched = Medical} => {Event group => Medical}

As with rule #4 this is perhaps a sanity check (despite a lift value of 2) and may seem as expected: Every incident received from an Ambulance is dispatched as Medical with an event type categorized as Medical. This stipulates that ambulances make no other calls to the fire services besides medical calls.

Predictive Modeling: Classification

Training, Cross Validation & Test Sets

In order to run a classifier on the data set it was important to first subset the data into training, cross validation and test sets which respectively contain 60, 20 and 20 percent of the observations. Two different partitions were created: Training, Validation and Test sets for only the Fire Incidents (FIRE_GROUP = 'Fire') and also for the entire data set. Some minor adjustments were required to ensure that all factor levels were available in the training set so that predictions could be made with the model. The models were then compared on the basis of Accuracy, Precision, Recall and F_1-Score.

Fire Incidents: Binary Class with Logistic Regression (No Interaction)

Logistic Regression was the first classifier used. As the class feature, CRITICAL, was created using features known after the incident (INJURIES, FATALITIES, EST_LOSS, etc), it was then regressed onto features which are apparently available at call time in an effort to contribute to the research question, "What further information, if any, can be provided to first responders at the time of incident call?"

The logistic model was initially trained on the FIRE incidents without any interaction:

CRITICAL ~ EVENT_TYPE_CD + ALARM_TO_FD + RESPONSE_TYPE + AID_TO_FROM_OTHER_DEPTS + EST_KM + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INCIDENT_YEAR

After training the model on the training set, probabilities were assigned to observations in the Validation set. Observations with a probability of critical greater than 0.5 were classified as CRITICAL and all other observations were classified as NON-CRITICAL. The confusion matrix of this classification is presented in **Table 8** and contains favourable results.

		Actual Class	
		0	1
Predicted Class	0	9098	743
	1	176	1657

***Table 8**: Confusion Matrix of Logistic Regression classifier on FIRE incidents. 0 and 1 represent CRITICAL and NON-CRITICAL incidents, respectively.*

Statistic	Value
Accuracy	0.89
False-Negative Rate	0.31
Precision	0.90
Recall	0.69
F1-Score	0.78

Table 9

The choice of evaluation metrics, particularly F1-Score, were chosen due to the relative large proportion of NON-CRITICAL versus CRITICAL incidents in the data set: approximately 20% of fire incidents and 5% of all incidents. A high level of accuracy could thus be obtained by simply assigning NON-CRITICAL to all incidents and so the goal was to focus on Recall (True Positive Rate), Precision (Positive Predictive Rate), as well as the harmonic mean between the two (F_1-Score). The corresponding metrics are presented in **Table 9** and the F_1-Score of 0.78 represents a promising initial success rate for this classifier.

In this model none of the days were significant in predicting the class feature. However, the other temporal features MONTH and YEAR were both significant, especially the month of November which has a statistically significant positive coefficient.

Fire Incidents: Binary Class with Logistic Regression (Interaction)

The above model was run on the Fire event type incidents without attempting to have the model account for interaction between features. Two additional models were run that included additional interactive terms: the first model with EVENT_TYPE * ALARM_TO_FD and the second with both EVENT_TYPE * ALARM_TO_FD and EVENT_TYPE * RESPONSE_TYPE. The Analysis of Deviance Table (**Table 10**) for the three models shows a decreasing residual deviance which corresponds with a better fitting model. However, the time to train this model —even on a smaller subset of only fire incidents— was significant for what amounted to only marginal improvements in the evaluation metrics, as shown in **Table 11**.

Model 1	CRITICAL ~ EVENT_TYPE_CD + ALARM_TO_FD + RESPONSE_TYPE + AID_TO_FROM_OTHER_DEPTS + EST_KM + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INCIDENT_YEAR			
Model 2	CRITICAL ~ EVENT_TYPE + EVENT_TYPE_CD + ALARM_TO_FD + RESPONSE_TYPE + AID_TO_FROM_OTHER_DEPTS + EST_KM + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INCIDENT_YEAR + EVENT_TYPE:ALARM_TO_FD			
Model 3	CRITICAL ~ EVENT_TYPE + EVENT_TYPE_CD + ALARM_TO_FD + RESPONSE_TYPE + AID_TO_FROM_OTHER_DEPTS + EST_KM + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INCIDENT_YEAR + EVENT_TYPE:ALARM_TO_FD + EVENT_TYPE:RESPONSE_TYPE			
	Resid. Df	Resid. Dev	Df	Deviance
1	35050	15507		
2	34958	15418	92	88.67
3	34601	14851	357	567.55

Table 10: Logistic Model Analysis of Deviance Table

Statistic	Model 1	Model 2	Model 3
Accuracy	0.89	0.92	0.92
False-Negative Rate	0.31	0.31	0.29
Precision	0.90	0.90	0.87
Recall	0.69	0.69	0.70
F1-Score	0.78	0.78	0.78

Table 11: A Logistic Model comparison of evaluation metrics

All Incidents: Binary Class with Logistic Regression

After applying logistic regression to the Fire incidents subset a similar classifier was then run on the training set of all incidents: the only difference being that EVENT_TYPE was also used as a predictor.

The resulting confusion matrix and evaluation metrics are presented in **Tables 11 and 12**. Although at a glance the model might appear less applicable to all incidents the overall performance is reasonable considering the significantly increased size (approximately 12 times larger) and variability among features. There is a much more pronounced increase in the class imbalance for the whole data set where critical incidents represent only 4% of observations. Despite these challenges the classifier correctly identified 43% of the critical incidents.

		Actual Class	
		0	1
Predicted Class	0	126537	3109
	1	287	2391

Table 12: Confusion Matrix of Logistic Regression Classifier on all incidents.

Statistic	Value
Accuracy	0.97
False-Negative Rate	0.56
Precision	0.89
Recall	0.43
F1-Score	0.58

Table 13: Evaluation Metrics of Logistic Regression Classifier on all incidents

Naive Bayes

		Actual Class	
		0	1
Predicted Class	0	131273	1803
	1	6831	4163

Table 14: Confusion Matrix of Naive Bayes Classifier on all incidents.

Statistic	Value
Accuracy	0.94
False-Negative Rate	0.30
Precision	0.37
Recall	0.69
F1-Score	0.49

Table 15: Evaluation Metrics of Naive Bayes Classifier on all incidents

Appendix: Data Dictionary

Original Features

Feature	Type	NA	Description	Summary
INCIDENT_NUMBER	character	6	Unique ID	Primary Key
EVENT_TYPE	factor	80	Event Type as Dispatched	114 Levels
DISPATCH_DATE	date	16802	On-scene date & time	2011:2016
ARRIVE_DATE	date	16802	Identical to DISPATCH_DATE	2011:2016
EVENT_TYPE_CD	factor	32	Event Code Dispatched	130 Levels
MAIN_STREET	factor	420556	Main Street of Incident	8932 Levels
CROSS_STREET	factor	390349	Closest Cross Street	11026 Levels
FSA	factor	355138	Postal Code FSA	119 Levels
RESPONDING_UNITS	continuous	32	No. of units responding	Min 1 / 1 1 / Med 1 / Ave 2.3 / 3 4 / Max 453
FD_STATION	factor	0	Fire Department Station	375,478 levels
OFM_INVESTIGATIONS_CONTACTED	factor	0	???	Binary: (0, 1) = (719226, 1144)
AID_TO_FROM_OTHER_DEPTS	factor	0	???	2 Levels: (1, 4) = (12408, 707962)
INCIDENT_DATE	date	0	Initial Call Date	2011:2016 Inclusive
INITIAL_CALL_HOUR	continuous	0	Initial Call Hour	0:23
INITIAL_CALL_MIN	continuous	0	Initial Call Minute	0:59
INITIAL_CALL_SEC	continuous	0	Initial Call Second	0:59
DISPATCH_HOUR	continuous	177	Dispatch Hour	0:23
DISPATCH_MIN	continuous	177	Dispatch Minute	0:59
DISPATCH_SEC	continuous	177	Dispatch Second	0:59
ONSCENE_HOUR	continuous	15371	On-scene Hour	0:23
ONSCENE_MIN	continuous	15371	On-scene Minute	0:59
ONSCENE_SEC	continuous	15371	On-scene Second	0:59
INITIAL_UNIT_PERSONNEL	continuous	0	Number of People Responding on First Apparatus	Min 0 / 1 4 / Med 4 / Ave 3.8 / 3 4 / Max 75
TOTAL_NUM_PERSONNEL	continuous	0	Total number of personnel responding	Min 0 / 1 4 / Med 4 / Ave 8.0 / 3 13 / Max 1277
EST_KM	continuous	0	Estimated Distance from station to incident	Min -1 / 1 2 / Med 2 / Ave 2.5 / 3 3 / Max 99
ALARM_TO_FD	factor	1	Alarm Source	11 Levels
RESPONSE_TYPE	factor	0	Response Type	68 Levels
RESCUES	continuous	0	Total no. of persons rescued	Min 0 / 1 0 / Med 0 / Ave 0.0 / 3 0 / Max 43
FF_INJURIES	continuous	0	No. of fire fighters injured	Min 0 / 1 0 / Med 0 / Ave 0.0 / 3 0 / Max 4
FF_FATALITIES	continuous	0	No. of fire fighter fatalities	Min 0 / 1 0 / Med 0 / Ave 0.0 / 3 0 / Max 0
AGENT_APP_HOUR	continuous	708040	Application of agent or decision to defer hour, min, sec	0:23
AGENT_APP_MIN	continuous	708040		0:59
AGENT_APP_SEC	continuous	708040		0:59
CONTROL_DATE	continuous	478833	Date fire under control	2000:2016
CONTROL_HOUR	continuous	707821	Hour, min, sec fire under control	0:23
CONTROL_MIN	continuous	707821		0:59
CONTROL_SEC	continuous	707821		0:59
STATUS_ON_ARRIVAL	factor	708040	Status on Arrival	9 Levels
WATER	factor	708040	Hydrant proximity	6 Levels

FIRE_CONTROL	factor	708040	Extinguish Action Taken							5 Levels
PROPERTY	factor	9863	Residential, Industrial,...							350 Levels
AREA_OF_ORIGIN	factor	708040	Outside, Storage, Functional,...							72 Levels
IGNITION_SOURCE	factor	708040	Cooking, heating, electrical,...							84 Levels
FUEL_OF_IGNITION_SOURCE	factor	708040	Gasoline, diesel, propane,...							18 Levels
OBJECT_OR_MATERIAL_FIRST_IGNITED	factor	708040	Building, furniture, soft goods...							54 Levels
POSSIBLE_CAUSE	factor	708040	Intentional, unintentional,...							22 Levels
VEH_PURPOSE	factor	718631	Vehicles only Purpose							10 Levels
VEH_FUEL	factor	718631	Vehicles only Fuel							9 Levels
EST_LOSS	continuous	0	Estimated $ loss (dollars only)	Min	1	Med	Ave	3	Max	
				0	0	0	465	0	1.3e7	
INSURANCE_ESTIMATE	factor	707958	Binary Insurance Estimate: Yes, No, NA, Undetermined							4 Levels
EST_VALUE_AT_RISK	factor	707958	Value based estimate categories							11 Levels
CIVILIAN_FIRE_INJURY	continuous	0	Number of Civilian fire injuries & fatalities	Min	1	Med	Ave	3	Max	
				0	0	0	0.0	0	12	
CIVILIAN_FIRE_FATALITY	continuous	0		Min	1	Med	Ave	3	Max	
				0	0	0	0.0	0	3	
RESCUED_CHILDREN	continuous	0	Numbers of children, adults and seniors rescued	Min	1	Med	Ave	3	Max	
				0	0	0	0.0	0	6	
RESCUED_ADULTS	continuous	0		Min	1	Med	Ave	3	Max	
				0	0	0	0.0	0	40	
RESCUED_SENIORS	continuous	0		Min	1	Med	Ave	3	Max	
				0	0	0	0.0	0	10	
PHYSICAL_CONDITION_1	factor	708041	Rescued Persons' Physical Condition							10 Levels
PHYSICAL_CONDITION_2	factor	720323								6 Levels
PHYSICAL_CONDITION_3	factor	720363								3 Levels
CIV_FIRE_CONTROL	factor	708041	Civilian Action: Fire Control							5 Levels
CIV_EVACUATION	factor	708041	Civilian Evacuation: All, None, etc							5 Levels
CIV_EVACUATION_REASON_1	factor	708041	Civilian Action: Reasons for not evacuating							9 Levels
CIV_EVACUATION_REASON_2	factor	720265								8 Levels
OPP	factor	719431	Other agencies, municipalities and services contacted							1 Level
MOE	factor	719011								1 Level
TSSA	factor	719822								1 Level
ESA	factor	720080								1 Level
MOL	factor	720080								1 Level
EMS	factor	720325								1 Level
CANUTEC	factor	720138								1 Level
GAS	factor	715150								1 Level
HYDRO	factor	715966								1 Level
MUNICIPAL_BUILDING_OFFICE	factor	720142								1 Level
MUNICIPAL_HEALTH_OFFICE	factor	719008								1 Level
MUNICIPAL_POLICE	factor	718973								1 Level
OTHER	factor	717388								1 Level
INITIAL_DETECTION	factor	711584	How fire was detected							9 Levels
EXTENT_FIRE	factor	711585	Extent of Fire: confined, spread etc							12 Levels
EXTENT_SMOKE	factor	711585	Spread of smoke: Confined, spread...							10 Levels
EST_NUM_PERSONS_DISPLACED	continuous	0	Estimated number of persons displaced	Min	1	Med	Ave	3	Max	
				0	0	0	0.2	0	999	
POSSIBLE_BUSINESS_IMPACT	factor	711585	Possible business impact							7 Levels
COMPLEX	factor	711589	Type of complex (airport, apartment, etc)							23 Levels
OCC_STATUS	factor	711589	Occupancy Status: Permanent, seasonal...							8

OCC_TYPE	factor	711589	Occupancy Type: Owner, renter...	7 Levels
BLD_STATUS	factor	711589	Building Status: Normal, reno...	7 Levels
BLD_HEIGHT	factor	711589	Number of Stories, Other or Undetermined	70 Levels
LEVEL_OF_ORIGIN	factor	711589	Floor, Basement, Other, NA or Undetermined	57 Levels
AGE_OF_STRUCTURE	factor	711589	Historic, before 1946, after 1946...	7 Levels
SMOKE_ALARM_PRESENCE_AND_OPERATION_MAIN_FLOOR	factor	711589	No smoke alarm, present and operated....	5 Levels
SMOKE_ALARM_FAILURE_TO_OPERATE	factor	711589	No battery, dead battery...	11 Levels
SMOKE_ALARM_TYPE	factor	711589	Battery, hardwired, wireless...	6 Levels
SMOKE_ALARM_OTHER_FLOOR_PRESENCE	factor	711589	No smoke alarms, present and operated...	6 Levels
SMOKE_ALARM_ON_ALL_FLOORS	factor	711589	Yes, not all operational,...	6 Levels
SMOKE_ALARM_IMPACT_ON_EVAC	factor	711589	All persons self-evacuated,...	7 Levels
SMOKE_ALARM_IMPACT_ON_NUM_EVAC	continuous	0	Continuous & Categorical	0:99, 99+, Undetermined
FIRE_ALARM_SYSTEM_PRESENCE	factor	711589	Fire alarm system present, no fire alarm,...	4 Levels
FIRE_ALARM_SYSTEM_OPERATION	factor	711589	Operated, did not operate,...	4 Levels
FIRE_ALARM_SYSTEM_IMPACT	factor	711589	All persons evacuated, some persons....	7 Levels
SPRINKLER_SYSTEM_PRESENCE	factor	711589	Full sprinkler system, partial...	4 Levels
SPRINKLER_SYSTEM_ACTIVATION	factor	711589	Activated, did not activate, ...	7 Levels

Data Sources
All data obtained from either RMS or Standard Incident Report

Above table information is prior to cleaning & not always consistent with "acceptable" values (i.e negative distance, years outside of range, etc) or possible values as described on OFM Standard Incident Report Codes List.

Dropped Features

Feature	Reason
ARRIVE_DATE	Identical to DISPATCH_DATE
FF_FATALITIES	Every observation is 0
FD_STATION	No information: Either INCIDENT_NUMBER, 0, or INCIDENT_NUMBER padded with one or more 0s
EVENT_ALARM_LEVEL	This feature was only available for the incidents in the year 2016.

Added Features

Feature	Type	Description	
TTA	Continuous	Time to Arrival in. Difference between ONSCENE time and INITIAL_CALL time measured in seconds.	
INCIDENT_DAY	Factor	Day of week extracted from INCIDENT_DATE	7 Levels
INCIDENT_MONTH	Factor	MONTH extracted from INCIDENT_DATE	12 Levels
INCIDENT_YEAR	Factor	YEAR extracted from INCIDENT_DATE	6 Levels
EVENT_GROUP			
PROPERTY_GROUP			
RESPONSE_GROUP			
FIRE_CLASS	Factor	Severity of fire classification based on parameters put forth by Asgari et. al	5 Levels
CRITICAL	Factor	Binary fire class feature derived from various features whose values are available after incident response	2 Levels: {0, 1}

Glossary

RMS - ???
CAD - Computer Aided Dispatch
OFM - Office of the Fire Marshall
FSA - Forward Sorting Area

References

Mendenhall, W., Beaver, R., & Beaver, B. (2013) *Introduction to Probability and Statistics*. Boston, MA: Brooks/Cole

KrishnaVeni, C.V, & Rani, T.S. (2011) On the Classification of Imbalanced Datasets. *International Journal of Computer Science & Technology*. 145-148. Retrieved from http://ijcst.com/icaccbie11/sp1/krishnaveni.pdf

López, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2013) An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences, 250,* 113-141.doi:10.1016/j.ins.2013.07.007

Asgary, A., Naini, AS., & Levy, J. (2012). Modeling the risk of structural fire incidents using a self-organizing map. *Fire Safety Journal, 49,* 1-9.doi:10.1016/j.firesaf.2011.12.007