Toronto Fire Services Incident Data: Revisited
By Geoffrey Clark, for Dr. Ceni Babaoglu
October 2nd, 2018

Methodology: Compare model trained with ROSE data vs. model trained with randomly sampled data. Both datasets created by sampling 60% of entire dataset. Classifier used was logistic regression with below formula and threshold of 0.5.

```
model_logistic_bal <- glm(CRITICAL ~ ., data=I_s_ROSE, family=binomial(link="logit"))
model_logistic_imb <- glm(CRITICAL ~., data = I_s[full_idx$train,], family=binomial(link="logit"))
```

<table>
<tr><td></td><td>Original Data (%)</td><td>ROSE Data (%)</td></tr>
<tr><td>Accuracy</td><td>97.4</td><td>91.8</td></tr>
<tr><td>False-Negative Rate</td><td>56.9</td><td>19.1</td></tr>
<tr><td>Precision</td><td>89.4</td><td>31.3</td></tr>
<tr><td>Recall</td><td>43.0</td><td>80.8</td></tr>
<tr><td>F1-Score</td><td>58.1</td><td>45.2</td></tr>
<tr><td>Confusion Matrix</td><td><table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>126544</td><td>3131</td></tr><tr><td>1</td><td>280</td><td>2369</td></tr></table></td><td><table><tr><td></td><td>0</td><td>1</td></tr><tr><td>0</td><td>117092</td><td>1051</td></tr><tr><td>1</td><td>9730</td><td>4448</td></tr></table></td></tr>
</table>

October 9th

Support Vector Machines

linear kernal, cost = 1, scale=T

```
model_svm_imb <- svm(formula = CRITICAL ~ ALARM_TO_FD + RESPONSE_TYPE + EST_KM + INITIAL_UNIT_PERSONNEL
+ INCIDENT_DAY + INCIDENT_MONTH + INITIAL_CALL_HOUR + INITIAL_CALL_MIN, data = I_s[full_idx$train, ],
kernel = "linear", cost = 1)

model_svm_bal <- svm(formula = CRITICAL ~ ALARM_TO_FD + RESPONSE_TYPE + EST_KM + INITIAL_UNIT_PERSONNEL
+ INCIDENT_DAY + INCIDENT_MONTH + INITIAL_CALL_HOUR + INITIAL_CALL_MIN, data = I_s_ROSE, kernel =
"linear", cost = 1)
```

<table>
<tr><td></td><td>Original Data (%)</td><td>ROSE Data (%)</td></tr>
<tr><td>Accuracy</td><td>97.2</td><td>85.6</td></tr>
<tr><td>False-Negative Rate</td><td>61.6</td><td>26.0</td></tr>
<tr><td>Precision</td><td>88.9</td><td>18.7</td></tr>
<tr><td>Recall</td><td>38.3</td><td>73.9</td></tr>
<tr><td>F1-Score</td><td>53.5</td><td>29.9</td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td>126552</td><td>3391</td></tr>
<tr><td>1</td><td>261</td><td>2108</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td>109239</td><td>1431</td></tr>
<tr><td>1</td><td>17574</td><td>4068</td></tr>
</table>
</td></tr>
</table>

October 10th

Support Vector Machines

radial kernal, cost = 1, gamma = 1, scale=T

```
model_svm_imb_radial11 <- svm(formula = CRITICAL ~ ALARM_TO_FD + RESPONSE_TYPE + EST_KM +
INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INITIAL_CALL_HOUR + INITIAL_CALL_MIN, data =
I_s[full_idx$train, ], kernel = "radial", cost = 1, gamma = 1)
```

<table>
<tr><td></td><td>Original Data (%)</td><td>ROSE Data (%)</td></tr>
<tr><td>Accuracy</td><td>97.0</td><td></td></tr>
<tr><td>False-Negative Rate</td><td>66.7</td><td></td></tr>
<tr><td>Precision</td><td>87.7</td><td></td></tr>
<tr><td>Recall</td><td>33.2</td><td></td></tr>
<tr><td>F1-Score</td><td>48.2</td><td></td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td>126558</td><td>3669</td></tr>
<tr><td>1</td><td>255</td><td>1830</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td></td><td></td></tr>
<tr><td>1</td><td></td><td></td></tr>
</table>
</td></tr>
</table>

October 11th

Support Vector Machines

radial kernal, cost = 1, gamma = 1, scale=T

```
model_svm_imb_radial11b <- svm(formula = CRITICAL ~ EVENT_TYPE_CD + ALARM_TO_FD + RESPONSE_TYPE +
EST_KM + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH + INITIAL_CALL_HOUR + INITIAL_CALL_MIN,
data = I_s[full_idx$train, ], kernel = "radial", cost = 1, gamma = 1)
```

<table>
<tr><td></td><td>Original Data (%)</td><td>ROSE Data (%)</td></tr>
<tr><td>Accuracy</td><td>96.8</td><td></td></tr>
<tr><td>False-Negative Rate</td><td>71.1</td><td></td></tr>
<tr><td>Precision</td><td>88.6</td><td></td></tr>
<tr><td>Recall</td><td>28.8</td><td></td></tr>
<tr><td>F1-Score</td><td>43.5</td><td></td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td>126609</td><td>3913</td></tr>
<tr><td>1</td><td>204</td><td>1586</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td></td><td></td></tr>
<tr><td>1</td><td></td><td></td></tr>
</table>
</td></tr>
</table>

```
<RespondingUnits>
  <INCIDENT_NUMBER>F11000010</INCIDENT_NUMBER>
  <CAD_UNIT_ID>P342</CAD_UNIT_ID>
  <DISPATCH_TIME>2011-01-01 00:04:13</DISPATCH_TIME>
  <ENROUTE_TIME>2011-01-01 00:05:45</ENROUTE_TIME>
  <ARRIVE_TIME>2011-01-01 00:10:02</ARRIVE_TIME>
  <CLEAR_TIME>2011-01-01 00:31:19</CLEAR_TIME>
</RespondingUnits>
<RespondingUnits>
  <INCIDENT_NUMBER>F11000011</INCIDENT_NUMBER>
  <CAD_UNIT_ID>A131</CAD_UNIT_ID>
  <DISPATCH_TIME>2011-01-01 00:04:46</DISPATCH_TIME>
  <ENROUTE_TIME>2011-01-01 00:07:30</ENROUTE_TIME>
  <ARRIVE_TIME>2011-01-01 00:09:02</ARRIVE_TIME>
  <CLEAR_TIME>2011-01-01 00:15:14</CLEAR_TIME>
</RespondingUnits>

> str(I_s)
'data.frame':   720370 obs. of  11 variables:
 $ EVENT_TYPE_CD         : Factor w/ 131 levels "CBRN1","CBRN2",..: 18 18 18 18 18 18 18 18 18 18 ...
 $ ALARM_TO_FD           : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 5 5 1 1 5 5 ...
 $ RESPONSE_TYPE         : Factor w/ 69 levels "1","2","3","11",..: 14 14 31 14 10 14 15 15 15 15 ...
 $ EST_KM                : num  3 3 3 3 4 5 2 1 3 1 ...
 $ INITIAL_UNIT_PERSONNEL: int  3 4 3 4 4 3 4 4 4 4 ...
 $ INCIDENT_DAY          : Factor w/ 7 levels "Sunday","Monday",..: 1 3 1 3 6 5 2 3 4 4 ...
 $ INCIDENT_MONTH        : Factor w/ 11 levels "January","February",..: 10 6 2 4 10 4 NA 5 9 6 ...
 $ INITIAL_CALL_HOUR     : int  20 13 15 15 13 9 15 14 17 14 ...
 $ INITIAL_CALL_MIN      : int  33 34 56 45 36 56 0 59 4 8 ...
 $ CRITICAL              : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 1 1 2 ...
 $ EVENT_GROUP           : Factor w/ 12 levels "Alarm","CC","CM",..: 1 1 1 1 1 1 1 1 1 1 ...

> str(I_s_ROSE)
'data.frame':   396801 obs. of  10 variables:
 $ EVENT_TYPE_CD         : Factor w/ 131 levels "CBRN1","CBRN2",..: 18 76 77 112 14 8 74 70 8 74 ...
 $ ALARM_TO_FD           : Factor w/ 12 levels "1","2","3","4",..: 1 3 3 1 5 1 3 3 1 3 ...
 $ RESPONSE_TYPE         : Factor w/ 69 levels "1","2","3","11",..: 14 48 48 40 10 21 48 48 34 48 ...
 $ EST_KM                : num  1.38 2.13 3.85 4.59 4.44 ...
 $ INITIAL_UNIT_PERSONNEL: num  3.95 4.21 4.05 4.13 3.81 ...
 $ INCIDENT_DAY          : Factor w/ 7 levels "Sunday","Monday",..: 4 3 1 3 1 6 6 2 2 5 ...
 $ INCIDENT_MONTH        : Factor w/ 11 levels "January","February",..: 9 2 2 11 11 9 6 11 3 9 ...
 $ INITIAL_CALL_HOUR     : num  5.99 5.59 15.96 23.77 14.48 ...
 $ INITIAL_CALL_MIN      : num  32.6 29.5 45.3 49.4 56.5 ...
 $ CRITICAL              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

October 16th

Support Vector Machines

linear kernal, cost = 1, gamma = 1, scale=T

```
svm_imb_linear_c1 <- svm(formula = CRITICAL ~ ALARM_TO_FD + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + IN-
CIDENT_MONTH + INITIAL_CALL_HOUR + INITIAL_CALL_MIN, data = I_s[full_idx$train, ], kernel = "linear",
cost = 1)
```

<table>
<tr><td></td><td>Original Data (%)</td><td>ROSE Data (%)</td></tr>
<tr><td>Accuracy</td><td>95.8</td><td></td></tr>
<tr><td>False-Negative Rate</td><td>1</td><td></td></tr>
<tr><td>Precision</td><td>NaN</td><td></td></tr>
<tr><td>Recall</td><td>0</td><td></td></tr>
<tr><td>F1-Score</td><td>0</td><td></td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td>126813</td><td>5499</td></tr>
<tr><td>1</td><td>0</td><td>0</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td colspan="2">truth</td></tr>
<tr><td>predict</td><td>0</td><td>1</td></tr>
<tr><td>0</td><td></td><td></td></tr>
<tr><td>1</td><td></td><td></td></tr>
</table>
</td></tr>
</table>

October 16th

Logistic Regression

```
> tfsi_formula <- "CRITICAL ~ ALARM_TO_FD + INITIAL_UNIT_PERSONNEL + INCIDENT_DAY + INCIDENT_MONTH +
INITIAL_CALL_HOUR + INITIAL_CALL_MIN"
> model_logistic_imb <- glm(formula = tfsi_formula, data = I_s[full_idx$train,],
family=binomial(link="logit"))

> ypred <- predict(model_logistic_imb, I_s[full_idx$cv,], type="response")
> fitted.results <- ifelse(ypred > 0.11,1,0) # Threshold 0.11
> table(predict=fitted.results, truth=I_s[full_idx$cv, 'CRITICAL'])
```

Threshold: 0.11

	Original Data (%)	ROSE Data (%)
Accuracy	80.8	4.2
False-Negative Rate	35.6	0.0
Precision	13.1	4.1
Recall	64.4	99.9
F1-Score	21.8	7.9
Confusion Matrix		

	truth	
predict	0	1
0	103382	1958
1	23442	3542

	truth	
predict	0	1
0	147	1
1	126677	5499

Threshold: 0.5

	Original Data (%)	ROSE Data (%)
Accuracy	95.8	62.7
False-Negative Rate	1	13.8
Precision	NaN	8.9
Recall	0	86.1
F1-Score	0	16.1
Confusion Matrix		

	truth	
predict	0	1
0	126824	5500
1	0	0

	truth	
predict	0	1
0	78340	760
1	48484	4740