LIT Model Project report

Group 6

Student 1 Peijie Lyu

Student 2 Vipul V Suresh

857-268-1445 (Tel of Student 1)

815-329-4856 (Tel of Student 2)

lyu.pe@husky.neu.edu

suresh.v@husky.neu.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Peijie Lyu

Signature of Student 2: Vipul V Suresh

Submission Date: October 10, 2019

The Lit Model

1. Problem setting

Public safety is a major domain which needs constant attention and quick resolutions during various situation arising at locations such as hospitals, universities, shopping malls and residential areas. Among the various emergencies occurring on a daily basis, fire prevention is most critical situations and readiness and urgencies by the fire department is of upmost importance. Considering Boston as an area of study, we see a lot of curiosity in the public minds when there is a fire truck driving around the block making loud siren and attending the fire alarm. But there is more to this than what meets the eye. The readiness and urgency from the department’s perspective is a completely different analytical scenario. Be it a true alarm or a false alarm there is movement of resources under multiple circumstances and there is room for improvement based on the patterns repeating on a daily basis.

2. Problem definition

Property loss is one of the end effect of a fire incident which can be controlled by taking preventive actions or it can be minimized by responding to the incident in a faster and efficient way by using predictive analysis based on the available data with variables such as incident type, incident time and date, estimated property losses, area zip code and neighborhood. Applying these variables to the data mining concept we can strive towards reduction in response time, property loss and other resource based on prioritization.

3. Data source

Boston fire department website https://data.boston.gov/dataset/fire-incident-reporting

4. Data description

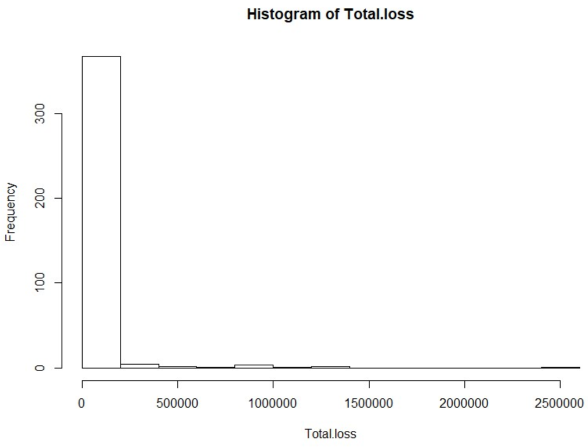
- Month based storage of fire incident data
- 24 attributes with 3000+ records on an average every month
- Dataset is processed to remove attributes with incomplete and redundant information
- Since the ratio of false alarm to true alarm was very low, records with true alarms leading to losses in content and property over a period of 6 months were gathered
- Link to processed data https://drive.google.com/open?id=1FVkQ0qagJRaMMAqVV8tXbzEo8zTLAUzF

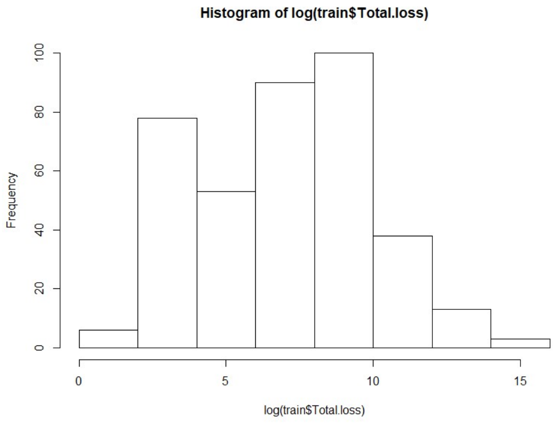
	Attribute Name	**Description**	**Attribute Type**
1	Incident date	Dispatch date.	Categorical
2	Incident type	Numeric code that identifies the incident description.	Categorical
3	District	District of Boston in which the incident has occurred	Categorical
4	Zip	Zip Code where incident occurred; special note that this	Categorical

		field is not always populated depending on the place of occurrence.	
5	Property use	Neighborhood where incident occurred (not always populated depending on the place of occurrence).	Categorical
6	Total loss	Estimated dollar amount of damage incurred to structure, personal property and belongings lost, as best determined by the Department at or around the time of the incident.	Numerical
7	**Fire**	**This is a manually generated binary classification response variable based on the total loss > $300 being considered a high risk incident**	**Categorical**

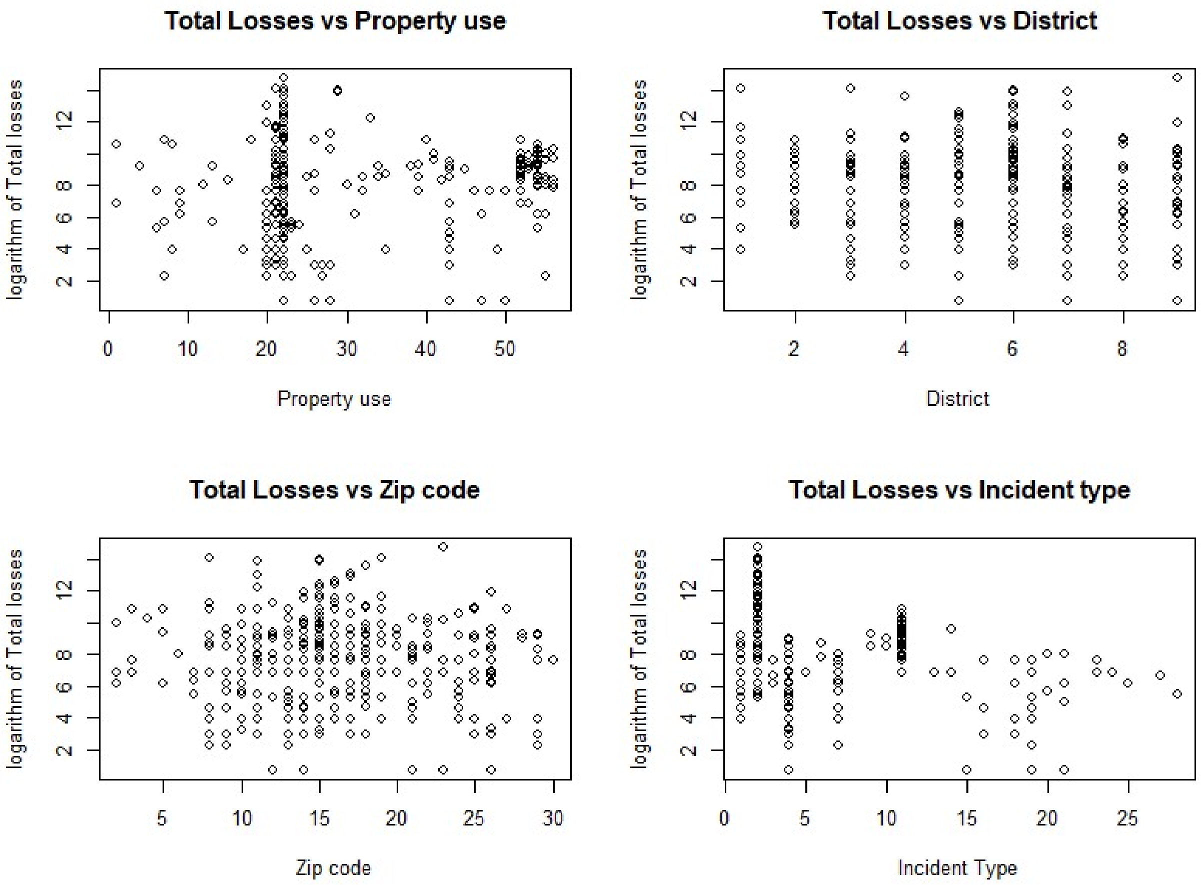
5. Data Exploration

- Histogram of total loss which is assumed as a function of the predictors such as incident type, property use, Zip and District as plot to analyze the distribution of the same
- Plot of the Total loss is transformed to its logarithmic value



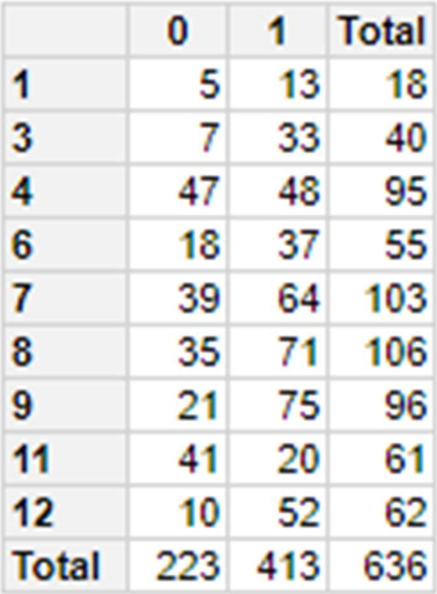


- The below plots helps to understand the dependencies of the predictors with the response variable



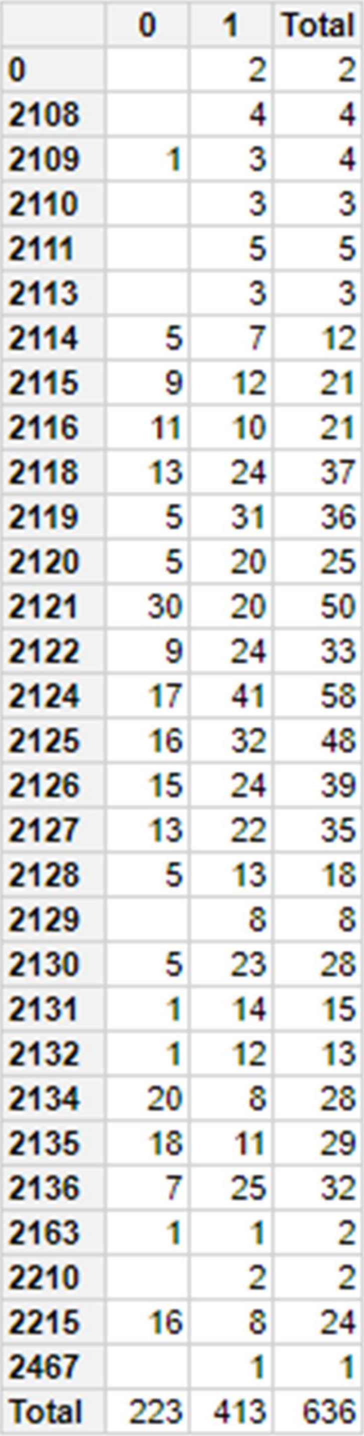
- A pivot chart was created to find the number of high and low risk incidents distributed in various Zip codes in the city

- Zip codes are also compacted and grouped to form Districts which helps in identifying which district is more prone to fire in the city of Boston. The second pivot chart explains the same



	0	1	Total
1	5	13	18
3	7	33	40
4	47	48	95
6	18	37	55
7	39	64	103
8	35	71	106
9	21	75	96
11	41	20	61
12	10	52	62
Total	223	413	636

District wise



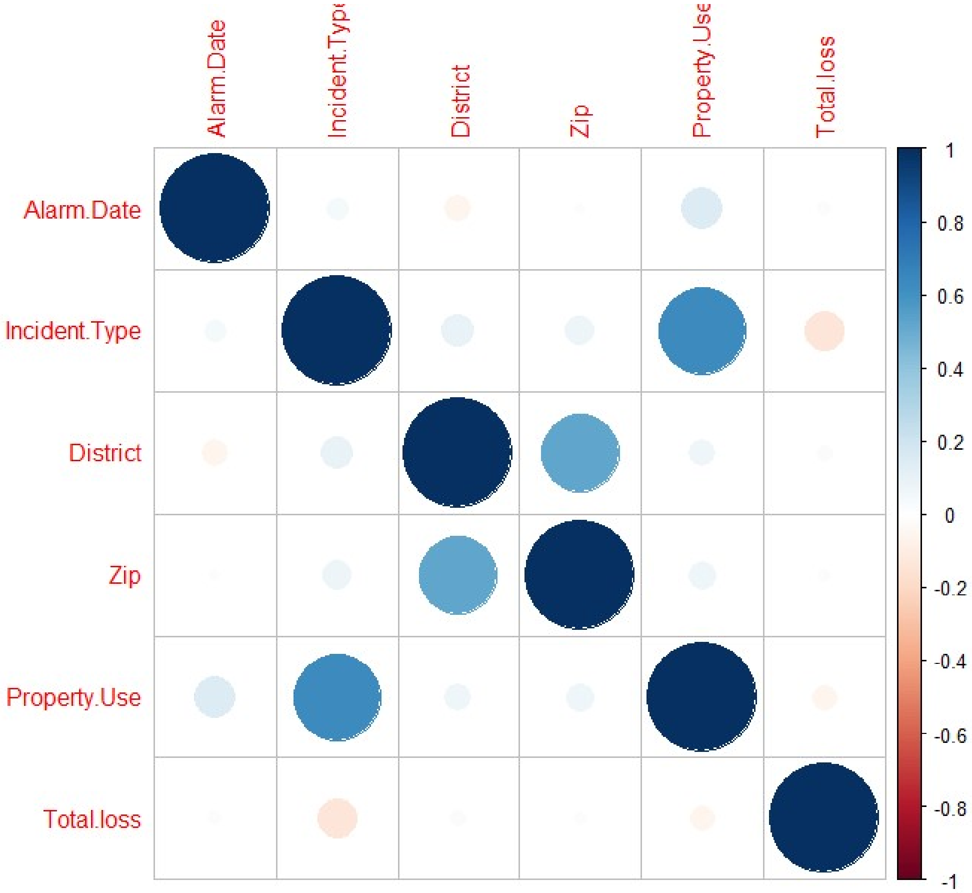
	0	1	Total
0		2	2
2108		4	4
2109	1	3	4
2110		3	3
2111		5	5
2113		3	3
2114	5	7	12
2115	9	12	21
2116	11	10	21
2118	13	24	37
2119	5	31	36
2120	5	20	25
2121	30	20	50
2122	9	24	33
2124	17	41	58
2125	16	32	48
2126	15	24	39
2127	13	22	35
2128	5	13	18
2129		8	8
2130	5	23	28
2131	1	14	15
2132	1	12	13
2134	20	8	28
2135	18	11	29
2136	7	25	32
2163	1	1	2
2210		2	2
2215	16	8	24
2467		1	1
Total	223	413	636

Zip code wise

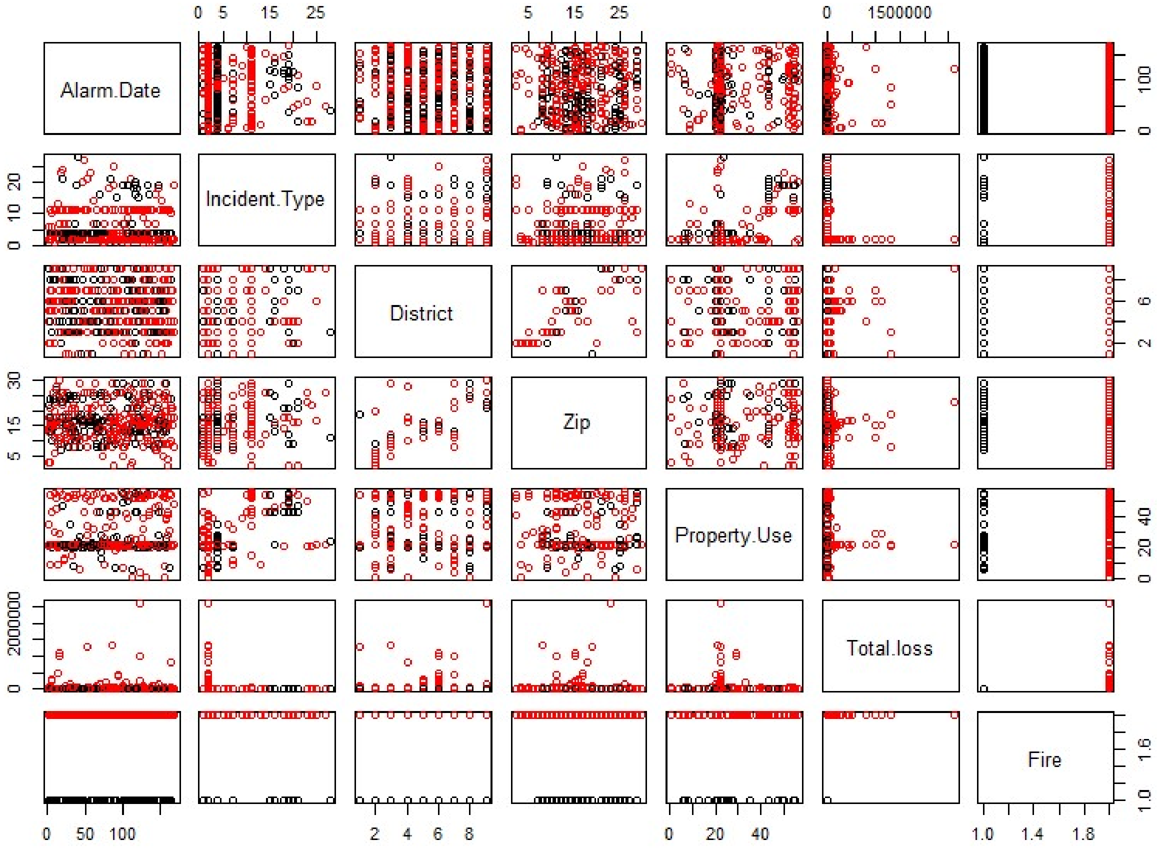
6. Data Mining tasks

- Incidents with total losses < $300 were classified as low risk incidents (class “0”) and incidents with total losses > $300 were classified as high risk incidents (class “1”)
- Predictors with incorrect classes for the ease of fitting the model were converted to factors and numeric
- The available data had 636 records and this was random sampled in to training and validating sample in a ratio of 60:40 percent
- Data reduction was carried out to eliminate the effect of spurious predictor information
- No missing values were encountered in the data

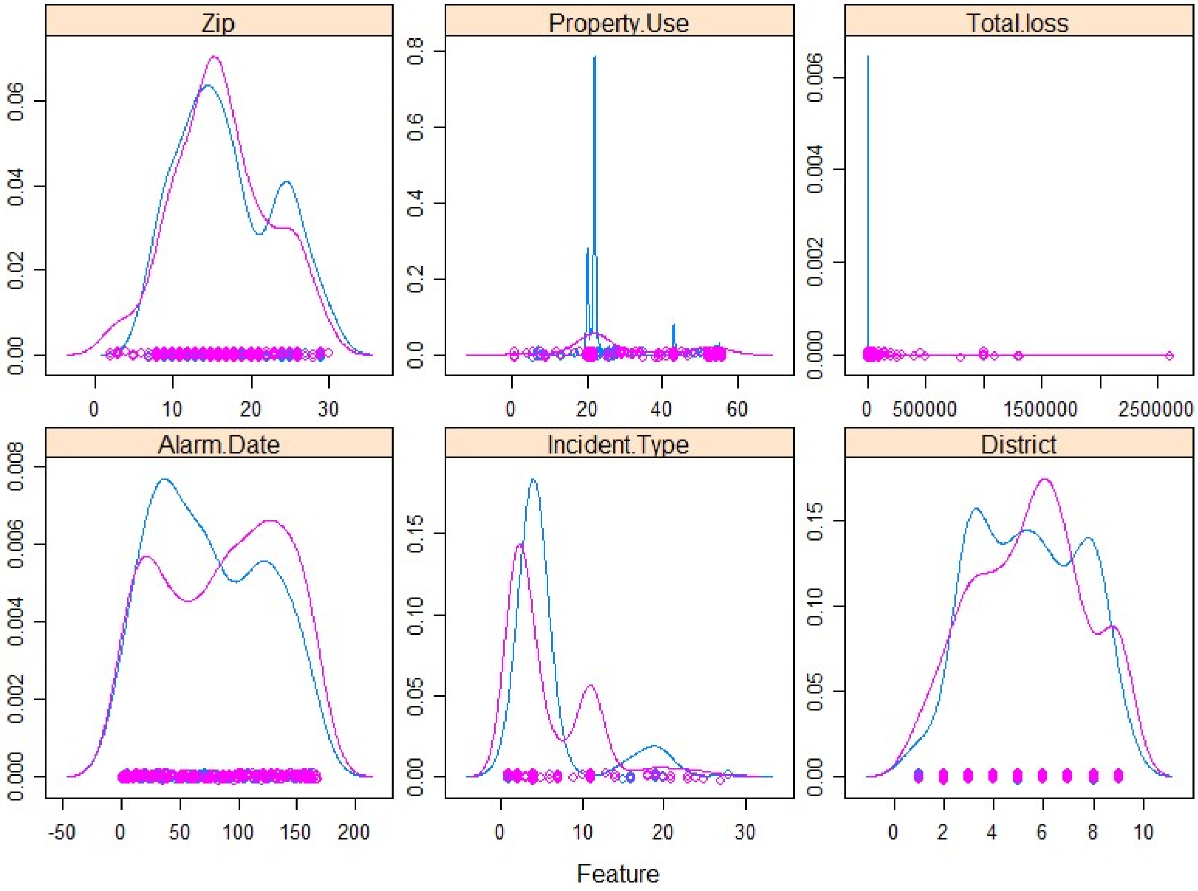
- A dot-representation was used where blue represents positive correlation and red negative. The larger the dot the larger the correlation



- Scatterplot matrix is plot and, Fire, our binary response, is the color indicator which is in red



- Density distribution of each variable broken down by Fire risk value. The density plot by Fire risk can help see the separation of High risk(class "1") and Low risk(class "0") It can also help to understand the overlap in Fire risk values for a variable



7. Data Mining Models/Methods

- After multiple iterations of different classification models, Logistic regression model was employed to address the classification of the new incidents into high and low risk

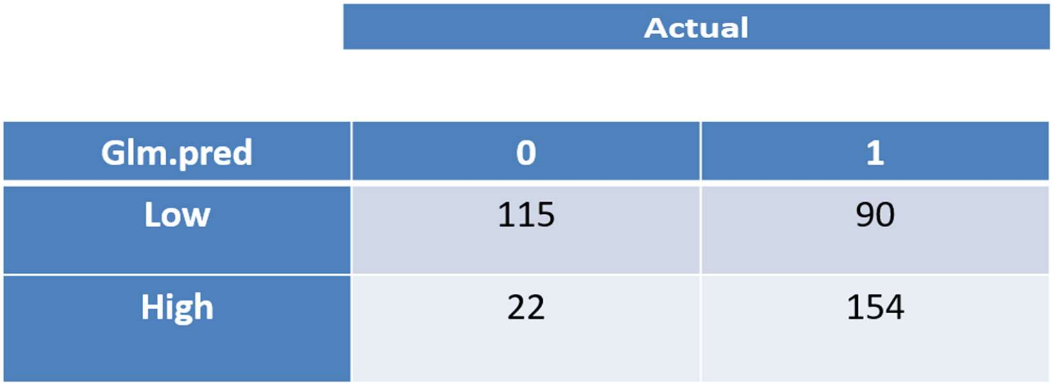
8. Model implementation:

- *logmodel<-glm(Fire~ Property.Use + Incident.Type + District +Zip, data = train, family = binomial)*

- The first 5 probability cut-offs for the above model were calculated glm.probs[1:5]

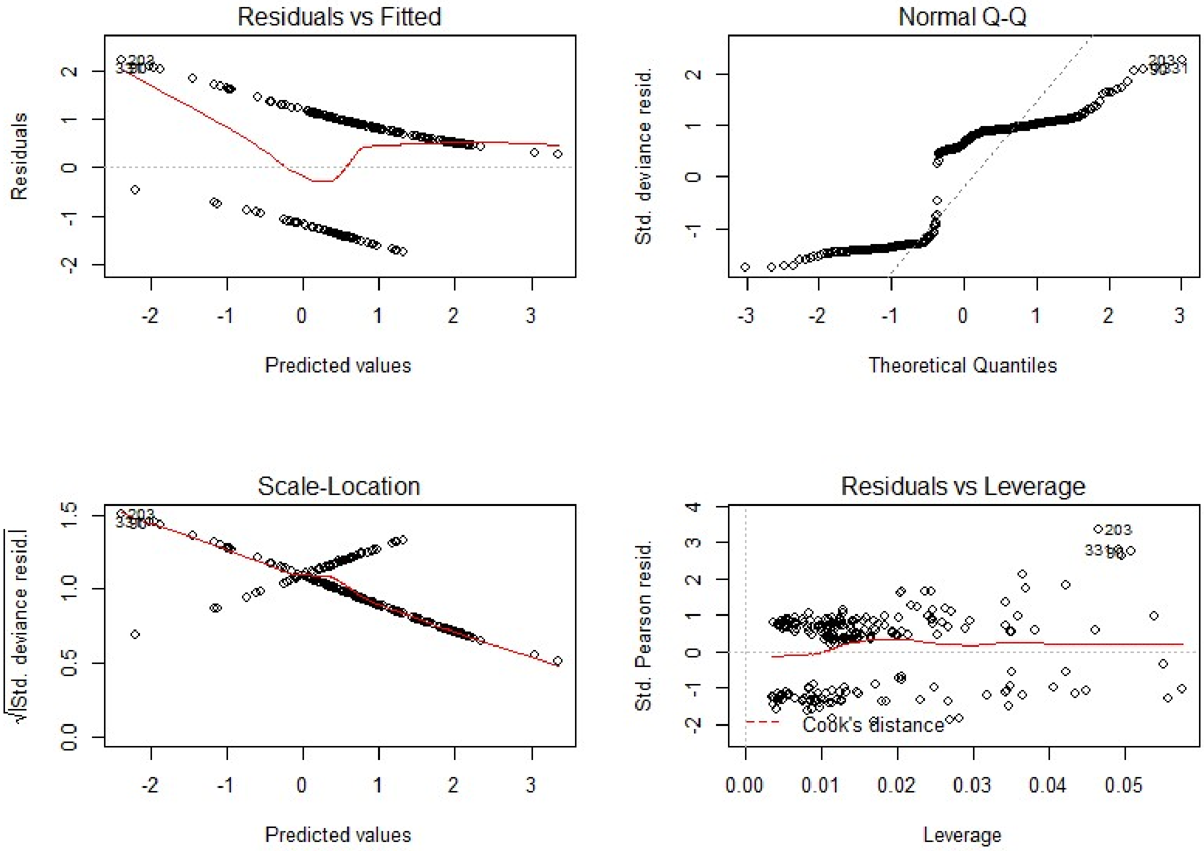
0.717	0.763	0.661	0.596	0.868

- Confusion matrix based on a cutoff of 0.67 was calculated which had the least misclassification as shown in the matrix below



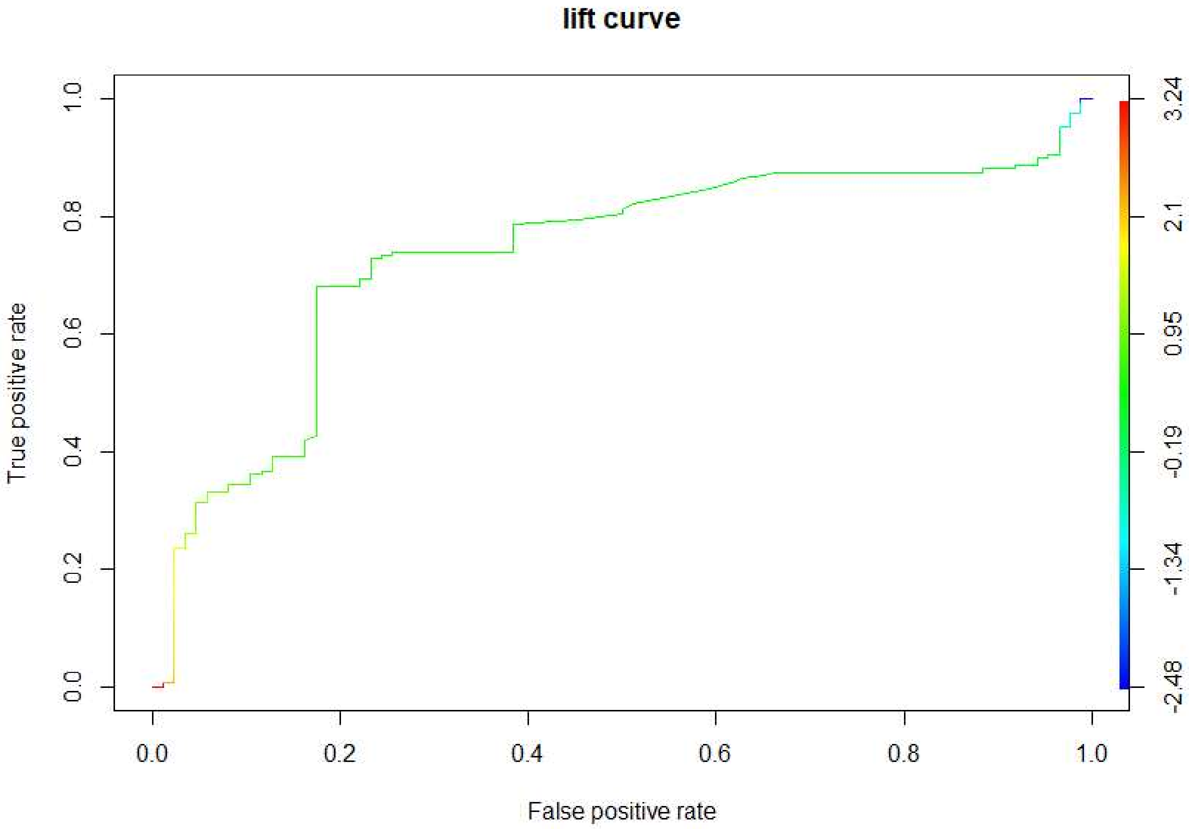
	Actual	
Glm.pred	0	1
Low	115	90
High	22	154

- **Summary plot of the model**



9. Performance Evaluation

- Area under lift curve was calculated and the results are shown as below



10. Project Results

- The Logistic regression model employed to classify the new dataset has an accuracy of 0.775 as shown below

```
>auc <- performance(pred,measure = "auc")
>auc <- auc@y.values[[1]]
>auc [1] 0.775
```

- With an accuracy of 0.775 the model performs okay. But this performance can be increased while changing the cutoff of classifying the records into low and high risk incidents which is currently at $300. This estimate is also a subjective estimate by the inspecting officers. Increasing the cutoff to >$300 will result in equal classification of the two classes which could possibly yield lesser misclassification errors