**Data mining**

Data Mining and Analysis

A picture containing logo

Description automatically generated

Dataset London Fire Bridget

London South Bank University

Contents

[1. Problem Identification 3](#_Toc101303960)

[Introduction: 3](#_Toc101303961)

[Dataset 3](#_Toc101303962)

[Data type and description 3](#_Toc101303963)

[Problem Statement 5](#_Toc101303964)

[Data mining tasks 5](#_Toc101303965)

[2. Data preparation 5](#_Toc101303966)

[Tools 5](#_Toc101303967)

[Data transform for SAS 5](#_Toc101303968)

[Selected Variable analysis 12](#_Toc101303969)

[3. Model Construction and interpretation 17](#_Toc101303970)

[Decision tree 17](#_Toc101303971)

[K-nearest neighbor 22](#_Toc101303972)

[Clustering 25](#_Toc101303973)

[Bar chart 31](#_Toc101303974)

[4. Model interpretation and evaluation 38](#_Toc101303975)

[Interpret the descriptive models created. 38](#_Toc101303976)

[Compare Performance 39](#_Toc101303977)

[5. Conclusion & and Future Work 42](#_Toc101303978)

[6. References 43](#_Toc101303979)

# Problem Identification

# Introduction:

The London fire brigade is basically the rescue, and the fire service for the London, the capital of the United Kingdom; moreover, it was formed in 1865 by the metropolitan act. Furthermore, it was founded in 1833, and it has 103 pump stations, including river station. (Anon., 2022)

# Dataset

* The dataset is about the London fire brigade from 2017 to 2021; moreover,
* The data set is consisting of 492999 tuples, and 38 columns

Graphical user interface

Description automatically generated

# Data type and description

Text

Description automatically generated

Text

Description automatically generated with medium confidence

* The data type of integer is 5, float is 12, and the rest of the features are object.
* The object data type may be the referenced, or non-primitive data type like array, string, classes, or etc.
* The data is real time and taken from the United Kingdom government website based upon the London fire brigade
* 499299 records with 38 columns
* Chart, histogram

  Description automatically generated

Chart

Description automatically generated

This is the visualization of some features of whole dataset to find the outliers and its possible values with respect to its frequency

# Problem Statement

The problem statement of my interest after observing the dataset, and its features is to find when, and where most of incident group (Fire, and Special Service) happened, so this can help London fire brigade to take positive or proper measures minimize the false alarm cases.

# Data mining tasks

For the problem of my interest that is shown above we must correlate the incident category with different features to find the trend, and patterns; however, the incident group is consist of three possible values

* Fire alarm
* Special service
* Fire

we will apply data datamining rules such as

* Problem definition
* Data gathering and preparation
* Model building and gathering
* Knowledge deployment (Anon., 2022)

# Data preparation

# Tools

For the dataset of the London fire brigade, I will use two different frameworks

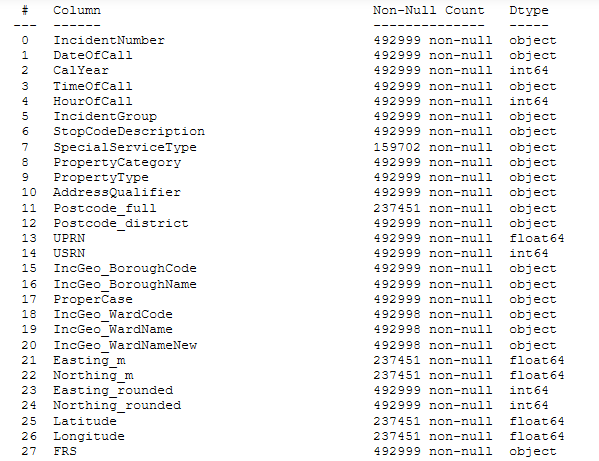
* Python (Jupiter notebook)
* SAS

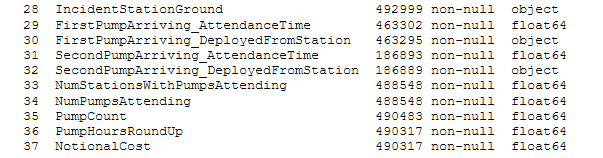
# Data transform for SAS

Firstly, I load the dataset into python, and apply different processing steps to transform the data for SAS.

Table

Description automatically generated





* The total numbers of features with name in the dataset
* The number of values in the dataset (492999) by each column
* Missing values in it out of 492999

Graphical user interface, text, application

Description automatically generated

* Total record 492999

Text

Description automatically generated

Target class label incident group for predictive models

* False alarm: the call of incident was fake there were no fire on that time
* Special service: the call was genuine, and the case was critical, so need of special service
* Fire: the call was real, and there was a fire there, but not critical case.

Then I checked the possible value of the property category where the incident was happened.

Table

Description automatically generated with medium confidence

Graphical user interface, text, table

Description automatically generated

This is the values, and occurrence of longitude, and latitude features in the dataset; however, there are so much missing values in it, as the total number of tuples is 492999, but this feature have 205678, and 204121 respectively.

Graphical user interface, text, application, email

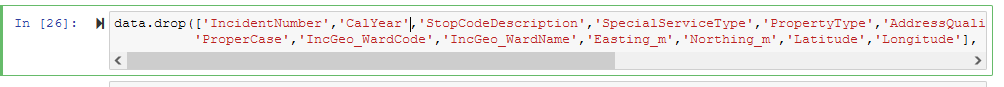
Description automatically generated

The fire service ground (FRS) is London, this dataset of London fire brigade only covers the information of the London.

Moreover, the nation cost of the whole incidents from 2017 to 2021 on London fire brigade is 490317 pounds.

**Feature drop and its justification as below:**

|  |  |
| --- | --- |
| **Drop Feature Name** | **Justification** |
| Incident number | incident number just shows the unique or specific number in the London fire brigade data |
| Cal year | it shows only year of call, and it is understood that the information of the dataset is from 2017 to 2021. |
| Special Service Type | It is further division of incident group |
| Property Type | It is further division of the feature property category |
| Address Qualifier | Property category is divided into property type, and the property type further divided into the address qualifier, as it has also many possible values to handle |
| Postal code full | There is dozen postal code in the London, but it is hard and complex to handle the data that have many possible values |
| UPRN | UPRN stands for unique property reference number, as it is primary or unique in the dataset |
| USRN | Unique street reference number is similar problem like UPRN |
| Borough code | This feature is very important in the dataset, but the problem with this feature is that it shows the same information as borough name, and proper case as well, and possible values is a code, but borough name shows the name of borough |
| Proper Case | Proper case and borough name have the similar values in each tuple. |
| Ward code | Same problem as borough code |
| Ward name | this feature shows the information of ward names with old names, and has been not called now, so it’s better to take ward new name feature instead of ward name. |
| Easting | which shows the same information of easting, but a few missing values; however, this feature have a lot of missing values |
| Northing | Same problem of easting |
| Longitude | Due to lack of missing values in it. It is better to drop it and take the same data from easting or northing rounded. |
| Latitude | Same problem of longitude |
| FRS | First service ground has only one possible value which is LONDON, throughout the whole dataset. |



So, how I dropped the columns from the London fire brigade dataset and justify the reason of dropping the column.

Graphical user interface

Description automatically generated

After dropping the columns, the new dataset looks like that above.

Graphical user interface, text, application

Description automatically generated

Moreover, the dataset is now about London fire brigade about the **Croydon** area.

* **19178** **tuples**
* **18 columns**

I have created another file CSV file name (lfbdm) of transformed data then load into the SAS.

* **Most of the selected features don’t have missing values**
* **There are some missing values in the attributes, but it does not mean that data is noisy; however, some of the missing values means there is no possible values for that.**
* **For instance, the second pump count tuple is null it means that there was no need of second pump on that time.**

Table

Description automatically generated

Then I have loaded the save python data into SAS studio.

**Table

Description automatically generated**

**This is the selected variable, and its level**

**Graphical user interface, application, Word

Description automatically generated**

# Selected Variable analysis

* **Date of Call**

This graph shows the information from 1-1-2017 to 31-12-2021; moreover, there is no sudden change in this graph throughout the graph, as there is consistency from 2017 to 2021, as most of the cases is reported of fire throughout the four year is similar approximately, and consider to be outliers, and most of the cases of 1/1/2017, and others are outliers.

* **Time of call**

This graph shows the information of exact time when the fire was held, as it has been observed that most of the fire events held in between the 11:59 to 21:35 throughout the day, and the graph is dramedy increase from 9:35, and increased with 3 to 5 percent till 21:35, then sudden decreased with the percentage of 20 percent. Thus, most of the cased in Croydon held in between 11:59 morning to 21:35 night.

* **Hours of call**

This graph shows the information regarding the hours when the most of fire event was held, as it has been clearly observed that from 11 to 21 has mostly fire cases, and after 13 pm there was no abrupt change in the graph and still to 19 pm then after gradually decreased.

* **Incident Group**

This feature is most important among all the features; moreover, this feature further elaborate the London fire brigade dataset, as the fire event was based upon the Fake call or False alarm, real fire, or special service. Thus, the most of cases reported in this data set is based upon the false alarm in Corydon.

* **Property Category**

**Chart, bar chart

Description automatically generated**

This graph shows the information which property type is mostly victim of the fire, as it has been clearly mentioned that the dwelling property type have most of the cases. Thus, there is need of positive measure in the dwelling type to overcome this issue in London. However, other types also got victim of fire, but a few cases were register compared with dwelling.

* **Postal Code District**

This graph shows that most of the area affect from the fire in the Croydon was CRO, compared with others, moreover the CR2, CR7, and SE25 had been same; however, the Croydon CR3, CR4, and CR6 has minimum cases, and its outliers in it.

* **Easting and northing rounded**

Chart

Description automatically generated

So, this is the geographical representation of the Croydon, London where most of the fire events happened including the possible values of the incident category, as this graph is combined with two features northing rounded as Y, and east Rounded as X.

* **Ward new name**

Graphical user interface, application

Description automatically generated

This bar chart shows the information of ward names that got victim of fire, as most of the ward in the Croydon was Fairfield, then the graph is gradually decreased with consistency ratio. Thus, after the sudden drop the ward with no case reported was Shirley.

* **Incident station ground**

**Chart, bar chart

Description automatically generated**

This bar chart shows the information about the fire station location on the borough name of Croydon, so there are 9 stations in the Croydon, and when most of the incident reported in the Croydon station ground; however, the Buckenham, Mitcham, wellington, and west now consider to be outliers

* **First pump arriving time**

This bar chart shows the information of the time in seconds the fire brigade responds, and the vehicle departed from the station in between the 100 to 500 sec, and other time in second are outliers.

* **First pump deploys from station**

A picture containing graphical user interface

Description automatically generated

This bar charts shows the information regarding the first pump arrival when the fire incident was happened, so it has been observed that most of the vehicle departed from the Croydon pump station, then woodside, and then Norbury, and others are outlier. Thus, most of the incident covered by the Croydon pump station.

* **Second Pump Arrival time**

This chart shows the information regarding the time taken by the second pump when the fire was not controlled, so it would take only a few second to depart the vehicle form the second pump.

* **Second pump deploy from station**

**A picture containing graphical user interface

Description automatically generated**

In this chart, it has been also observed that most of the departed vehicle from the Croydon pump station again when the fire event was not mange, and when the need of help or another vehicle it was mostly departed from the Croydon, Norbury, Addington, wood site, Buckenham, and purely station, and rest are outliers.

* **Number of stations with pump attending**

This graph shows the number of station that pump attended; however, most of the incident events was covered or handle by one station either first deploy or second. Note if the fire event was handled with only one pump like Croydon for the first and second time as well, as the location was same, so the pump count would be 1 or 2 and rest are outliers 3,4,5,6, (.).

* **Pump Count**

This graph shows the total number of pump that each incident event covered, thus most of the events was covered by the 1 or 2 pumps mostly; however, there are value of 175 which means some of the fire events could attended by the 175 pumps that is impossible, and it is outliers. Note if the first pump, and second pump was different then pump count is 2, otherwise if the pump was same n first and second time then pump count was 1.

* **Pump hours round up**

The Pump Hour Round Up is an estimate of total time taken to deal with the incident and will count all pumps in attendance including non-emergency pumps sent as “reliefs”, and most of value in between 1, and rest of all out are outliers

* **National cost**

This bar charts shows the clear information that most of fire events estimated cost on each event is approximately from 326 pounds to 28265 pounds and others are outliers.

# Model Construction and interpretation

**Predictive model**

* Decision tree
* K nearest neighbor

# Decision tree

Statistical analysis

A picture containing graphical user interface

Description automatically generated

This is the statistical analysis or correlation of each variable with respect to incident group class label. So, after observing the correlation and complexity of data so I take variable as below.

Table

Description automatically generated

These are the input variable that I am taking for the deployment of the machine leaning model Decision Tree.



The class label is incident group.

Diagram

Description automatically generated

Import London fire brigade data of Croydon -> filter to handle the outliers-> data partition->Decision Tree.

Table

Description automatically generated

This is setting of decision tree node

Diagram

Description automatically generated

This is the decision tree with root node of incident group and property type, as if the property type is outdoor, so the probability of fire is 59%.

Diagram

Description automatically generated

**IF** (Property category: dwelling->Number of pumps Attending :1 ->property category: dwelling -> incident station ground: purely ->ward new name: purely) **then** False Alarm is 50%

Table

Description automatically generated

The rules of each node in detail as above.

Table

Description automatically generated

New setting of decision tree node

Diagram

Description automatically generated

Text

Description automatically generated

Node rule using different setting of decision tree.

# K-nearest neighbor

Table

Description automatically generated

These are the input variable that I am taking for the deployment of the machine leaning model Decision Tree.



The class label is incident group.

Diagram

Description automatically generated

Croydon data then filter to handle outliers then data partition as default and k nearest neighbor

Table

Description automatically generated

The value of K is **6.**

Table

Description automatically generated

* The wrong classification in train set is 170, 3368 in validation, and 3332 in test, as the prediction and actual result in test, and validate is very poor.

Table

Description automatically generated

* The correct classification from prediction to actual is 3 in both cases.

Graphical user interface, chart, line chart

Description automatically generated

The line chart shows the accuracy of the k nearest neighbor at start it goes down sharply, and keep consistent, then goes up.

Table

Description automatically generated

The value k is **4**

Table

Description automatically generated

* The wrong classification in train is now 165, and 3392, 3349 in validation, and test respectively

Table

Description automatically generated

The accuracy of training test is: 4602/7614=0.60\*100=60

Chart, line chart

Description automatically generated

The two parallel lines meet at 0.42 and till at 0.57.

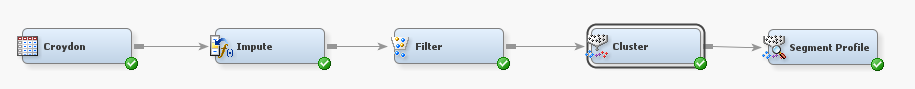
# Clustering

Table

Description automatically generated

* Input variables





Table

Description automatically generated

Chart, pie chart

Description automatically generated

* This is the graphical visualization of 5 cluster.
* Cluster 1 frequency 4154
* Cluster 2 frequency 2960
* Cluster 3 frequency 5746
* Cluster 4 frequency 3834
* Cluster 5 frequency 2337

Chart, bar chart, treemap chart

Description automatically generated

* Hours of call: All the cluster have the approximately same value no significant change; however, each cluster have same pattern of numeric value of 0 to 24 with same frequency with minor change.
* Number of pumps attending: each cluster have the number 1 value with higher frequency instead of 3 it has high frequency value of 2
* Ward new name: cluster have most of the value is Fairfield, and ward don; however, the rest of the cluster have mixed value.
* Incident group: cluster one has false alarm, and special service in high frequency, cluster two have fire in high frequency; however, the cluster 4, and cluster 5 have special case, and false alarm respectively.
* Incident station ground: cluster one is full of Croydon; however, the other four clusters have mixed values,
* Postcode district: CR0 is mainly in each cluster.
* Property category: the significant values in each cluster are dwelling, nonresidential, and road vehicle

Table

Description automatically generated

The most important feature in the clustering is property category, and then ward newname, postal code district, and incident group.

Shape

Description automatically generated

Segment 3 with maximum count of 5756, and percentage of 30.19 among all segments.

A picture containing application

Description automatically generated

* Segment 1: ward new name, incident group station, postal code district is in high volume.
* Segment 2: incident group, property category is in high volume
* Segment 3: numbers of pump attending is in high volume
* Segment 4: incident group in high volume
* Segment 5: incident group, property category, numbers of pump attending, postal code district are in high volume.

Graphical user interface, text, application

Description automatically generated

The number of clusters are 3.

Chart, pie chart

Description automatically generated

* Segment 1: 6581
* Segment 2: 6991
* Segment 3: 5459

Chart, bar chart

Description automatically generated

* Hours of call: same formatted value in each cluster
* Numbers of pump attending cluster 1 and 3 have same value 1, and 2 cluster have 2 values in most
* Ward new name: cluster 1 have Fairfield, and wad don in most, and rest of cluster have mixed formatted value
* Incident group: fire is lesser in each cluster
* Incident group station: Croydon is most in cluster 1 and 2; however, the Norbury is most in cluster 3 only
* Postal code district: CR0 is in most in cluster 1 and 2
* Property category: dwelling is most in each cluster.

A picture containing text, screenshot, receipt

Description automatically generated

* The importance of the inputs with respect to the clustering with the value of 3.

Shape, circle

Description automatically generated

* Cluster 2 have the maximum number of counts 6991 with the percentage of 26.73%

Table

Description automatically generated

Table

Description automatically generated

* The segment or cluster variable worth along with the rank.

A picture containing application

Description automatically generated

* Cluster one has most of the values are ward new name, postcode district, incident station ground, and numbers of pump attending
* Cluster 2 have numbers of pump attending in most
* Cluster 3 post code district, ward new name, incident station ground, and number of pump attending in most.

# Bar chart

Every problem in the datamining needs preprocessing; however, the preprocessing needs to consume the 70 percent of the attention and 30 percent for the modelling.

Diagram

Description automatically generated

Import the data then check the relationship between the incident group and the hours of call to determine in which time most of the fire incident happened.

So, I just change the data type of the hours of call to nominal.

**Input: hours of call**

**Target: incident group**

Chart, bar chart

Description automatically generated

It has been observed that most of the incident was false alarm, fire cases a few; moreover, the frequency of special cases are more compared with fire.

**Fire:** 1hour,12hour, 0hour, 12hour, 22hour

**Special Cases:** 1hour,9hour, 13hour, 14hour, 10hour, 23hour, 0hour, 18hour

A screenshot of a computer

Description automatically generated with medium confidence

There are no missing values in it.

Chart, bar chart

Description automatically generated

**Fire: 2017,18,19,20**

**Special Case: 2017,18,19,20,21**

It has been observed that false alarm of fire is in mode in above graph, and mostly false alarm has gradually increased with 10 percent from 2017 to 2021, and sharply decline in 2021.

Chart, bar chart

Description automatically generated

**Fire: CR0, CR7, SE25**

**Special Case: CR0, CR2, CR7, SE25**

**A screenshot of a computer

Description automatically generated with low confidence**

**No missing value.**

**Chart, bar chart

Description automatically generated**

**Fire: dwelling, outdoor, outdoor structure**

**Special Case: dwelling, road vehicle.**

**Interesting feature no false alarm complaints in road vehicle, outdoor structure, and mostly fire alarm complaints in dwelling, and nonresidential Ares.**

**Table

Description automatically generated**

**No missing values.**

**A picture containing timeline

Description automatically generated**

Fire: Fairfield, broad green, wad don

Special Case: Fairfield, new Addington north, purely and wood cote, Selhurst, south Croydon, south Norwood, Thornton heath, wad don, west Thornton, woodside, wad don

A screenshot of a computer

Description automatically generated with medium confidence

No missing values.

Chart, bar chart

Description automatically generated

Fire: 1pump, 2pump, 3pump

Special case: 1pump, 2pump

Table

Description automatically generated

140 missing values in number of pumps attending, so this missing value shows that that number of pump attend was 0 the fire incident was handled, and no need of pump there.

Application

Description automatically generated with medium confidence

Fire: 4000 pounds

Special Case: 0 pounds, 4000 pounds

Table

Description automatically generated

97 missing values, so it means 97 events had not cost.

# Model interpretation and evaluation

# Interpret the descriptive models created.

**The interpretation of descriptive model is above.**

K means clustering: The k means is used for both supervised as well as unsupervised learning clustering problem. Import the data then apply the impute that is used to handle missing value then apply filter to handle or manage the outliers, and then apply cluster node first I will take the value of K is 5, then 3.

Bar charts import the data then connect the node of multiplot and the correlation of input features and target feature.

# Compare Performance

Table, Excel

Description automatically generated



Diagram

Description automatically generated

The model comparison node is used to find the comparison with respect to accuracy, and other evaluation measure to find the best algorithms.

Table

Description automatically generated

Table

Description automatically generated

According to the desire setting I did the comparison of the two different predictive model.

Table

Description automatically generated.A screenshot of a computer

Description automatically generated with low confidence

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

**Error Rate**

It has been observed that the error rate of KNN is greater than Decision Tree, thus the Decision tree proved good in term of error rate

KNN =1-0.60= 0.40

Decision Tree = 1-0.85=0.15

Table

Description automatically generated

**Accuracy:**

KNN =4602/7614=0.60\*100=60%

Decision Tree = 6459/7614=0.85\*100=85%

**Chart, line chart

Description automatically generated**

**Graphical user interface, chart, line chart

Description automatically generated**

**Table

Description automatically generated**

**Overfitting:**

The overfitting lies in the KNN, as the graph shows the ups and down trend to find the target value; however, the decision tree shows the linear line to find the class label.

# Conclusion & and Future Work

The problem was to find out the when, and where the most of fire, and special cases happened in the Croydon, London; however, as per the requirement I applied two different types of models

* Predictive
* Descriptive

In predictive the decision tree was so much helpful, meaningful, and useful to find out the business problem in case what are the parameters or association rules to find out that it would be special case or fire incident in future on test data.

Secondly, in descriptive model the bar charts were very useful, and beneficial to find out the correlation with respect to incident group, as in which postcode district, property type, hours, and wards most of the fire, and special cases incident happened.

However, the focus of this report is to find out in which postal code district, ward areas, and further its property type the most of fire, and special cases happened.

In future I would like to work with London fire brigade to minimize the cases related to fire, and special cases in these wards, so the cost of each incident would be decreased, and also beneficial or helpful for the government budget.

# References

Anon., 2022. *Data mining.* [Online]   
Available at: https://en.wikipedia.org/wiki/Data\_mining  
[Accessed 19 04 2022].

Anon., 2022. *London Fire Brigate.* [Online]   
Available at: https://www.london-fire.gov.uk/  
[Accessed 19 04 2022].