Coursework 3 - Predicting Emergency Crime and Safety Incidences in London

Siripichaiprom, Kongpob 30583489
Clarke, Dario 30304377
Yang, Shuhan 30468701
Imre, Berathan 30493161
Jungsanguansit, Pittawat 30513928

***Abstract*—This task is aimed to predict the possibility of emergency crime and safety incidence in London. We used three data sets including crime, fire and accident. It also implement the prediction into a application so that the emergency response teams are able to use. Some suggestions about possible improvements are provided at the end of report.**

I. Introduction

Imagine, a police force that could respond and de-escalate incidences before anyone is hurt, fire brigades that ensure minimal property damage and less lives lost to traffic accidents. We believe that this goal is attainable, if emergency response teams were able to predict incidences and thus act to prevent them or minimize their impact. This project aims at providing a tool to make such predictions.

II. Objectives

A. Crime Prediction

Predict the probability of crime occurring at a given time and location in London.

B. Fire Incident Prediction

Fire is considered as serious problem especially in London. With lots of people living in the same area, even the minor alarmed from automatic fire alarm would be costly. Protecting it from happening beforehand would decrease the destruction. The result shows probability of fire incident occurs in London according to set of inputs, time and location for examples.

C. Accident Prediction

Accident can occur any time at any place. In general, when the accident occurs, the ambulance is called. Specifically, during rush hours, the accident is likely to happen the most. As most people will be on the road rushing for works, there will be lots of traffics throughout the city and as London, capital city, it would be the busiest time filled with cars. Since every single minute counts for life's saving and there are lots of traffics around the city, predicting the accident is required in this case. The result shows the probability of fire incident occurs in London from input such as time and location.

D. Application

Develop an interface through which emergency response teams can access these predictions. To generate an interface to aid those people we had to create a website with three different pages. Each page would communicate with their own model. The reason we chose that type of approach is that to provide a refined interface for users as in emergency situations they dont have to waste time through navigating through the web page. The page that we developed consists of; crime prediction, fire prediction, car accident prediction. To communicate with the back-end that we have written, we needed a web page to display those values. The model that we created is written on Python and the front-end is written with HTML and JavaScript. We generate HTTP GET requests to our flask server and retrieve the results using JavaScript. After receiving the values, web page displays it on the map accordingly.

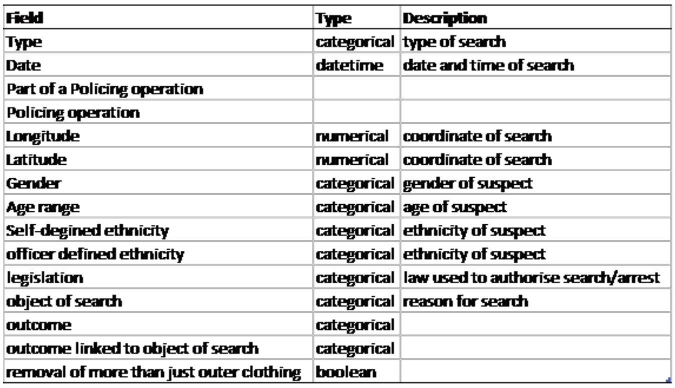
Webpage consists of several fields to fill and a map. The map had to be interactive and users should be able to navigate around the map. For these reasons, Google Maps was used to generate the map part of the website. There is a pin on the map and by moving it, users could select different parts of London and navigate with zooming tool as well. The address and coordinate data are fetched from Google. The address is displayed for user to see where they selected. There are other parameters that are required to be filled. Different models require different types of inputs. For example, accident dataset requires user to select the severity of the accident. However, all models need coordinates, time to predict the outcome. This type of approach is used to distinguish between critical crashes and slight damage accidents. If the severity is above a certain threshold which is 80% for our web page, a circle of radius 200 meters is coloured as red which indicates that, the region selected has a high chance of crash with the given parameters. However, if the value is lower than that threshold the circle is coloured as green. In the next phases of the development, website can be organized to enhance user experience such as using Bootstrap, CSS and other tools. Unfortunately, due to time constraints on our schedule we had to create a functional structure.

III. Data Collection

A. Crime Data

The data used for the crime predictions was obtained from three sources:

1) London Stop and Search Data: This is the main dataset used for the predictions and is a dataset provided by the British Home Office. It consists of stop and search reports from 2014 to mid 2017 and has the following field shown in figure 1 below:



Field	Type	Description
Type	categorical	type of search
Date	datetime	date and time of search
Part of a Policing operation		
Policing operation		
Longitude	numerical	coordinate of search
Latitude	numerical	coordinate of search
Gender	categorical	gender of suspect
Age range	categorical	age of suspect
Self-degined ethnicity	categorical	ethnicity of suspect
officer defined ethnicity	categorical	ethnicity of suspect
legislation	categorical	law used to authorise search/arrest
object of search	categorical	reason for search
outcome	categorical	
outcome linked to object of search	categorical	
removal of more than just outer clothing	boolean	

Fig. 1. Stop and search fields

2) LSOA Atlas: This dataset was provided by the greater London authority and consists of a list of all of the Lower Super Output Areas (LSOAs) in London along with demographic information for several years. However, we used it as a list of all the LSOA codes in London so we could join the crime dataset, to the post codes data.

B. London LSOA API

As the third data source we used an API provided by Chris Bell on his website doogal.co.uk. The API makes use of OS, Royal Mail and National Statistics data to provide statistics on all of the post codes within the specified LSOA code, including the coordinates, population and number of households within each post code. We iterated the LSOA Atlas to query the API and construct a postcode dataset.

C. Fire Data

The fire data were collected from Kaggle.com [1]. It contains 32,247 data points of London fire brigade service calls in 2017. Each data point has 35 features which can be found in appendix A-A.

D. Accident Data

The accident data were collected from Kaggle.com [2]. It contains accident data in London from year 2005 to 2014 splitted into three parts by year. The data features can be found in appendix A-B.

IV. Data Processing

A. Crime Data

The objectives of data processing for the crime data was to clean, join the datasets and to create a set of labels to use for training a machine learning algorithm.

1) Cleaning the Stop and Search Dataset: Cleaning of the stop and search dataset involved removing missing values. Figure 2 below shows the distribution of missing values among the columns: To clean the stop and search dataset, we first removed the completely empty columns and removed the age, gender and race columns. We also removed the rows without coordinates, however this had the unfortunate side effect of eliminating all of the thefts from the dataset. Finally, the datetime string field was tokenised

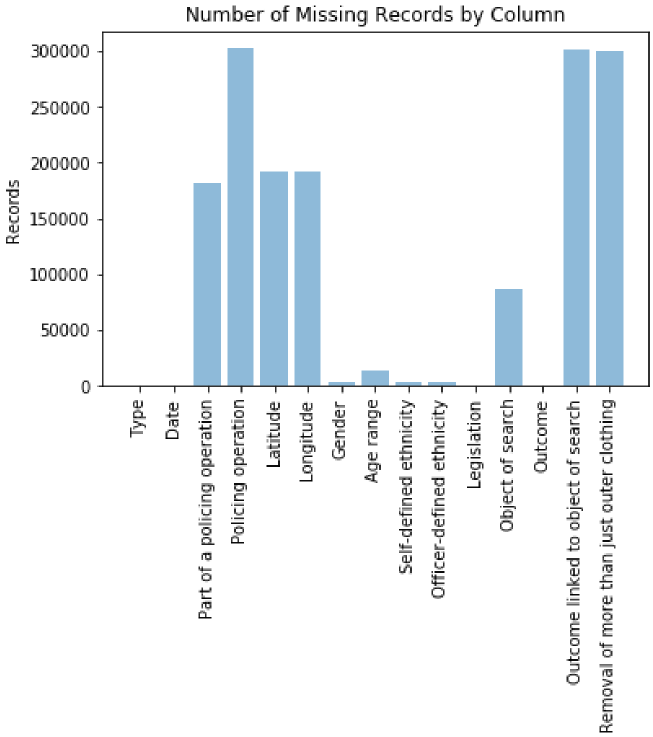


Fig. 2. Distribution of missing records for stop and search dataset

2) Cleaning the Post Code dataset: Cleaning the post code dataset involved removing the post codes that were no longer in use as they were replaced by newer post codes already in the dataset.

3) Joining the datasets: The population and the number of households was added to the stop and search dataset. These were taken from the post code dataset, rather than the LSOA dataset, because the post code dataset gave these features for more precise areas. To add them to the data set, we found the post codes within 0.001 latitude and longitude of the location of the stop and search and took the mean of the population and number of households for them. This was done for each data point in the stop and search dataset. The rationale behind this was to add information regarding the number of people living within an area of a crime as we wanted to investigate whether this impacts the volume or types of crime.

4) Creating Labels: The type of crime committed was inferred from the object of search, the legislation, outcome related to search. Code was written, based on combinations of these fields to generate the type of crime committed, if any at all. The type of crime committed was then used as the label. After the pre-processing of the datasets, the final set of features were as follows: Longitude, Latitude, Hour, Minute, Weekday, Month, Day of Month, Average Population, Average Households and Label. The longitude and latitude were selected as we wanted to investigate how crime differs from area to area. We also wanted to investigate what times throughout the day are crimes likely to occur. We also suspected that crime would be more prevalent during the weekend, and during holidays and so the weekday, month and day of month were selected. Finally, we believed that crime would be more prevalent in more densely populated areas and thus we included the average population and average households to confirm.

B. Fire Data

Feature selection and data cleaning processes are required with this data set.

1) Data Cleaning: Firstly, We removed feature that has no data recorded. Some features are filtered because of inconsistency to our prediction. The remaining features left are the date of fire, the coordinate, and the description of fire. Since the coordinates of data point contained in original data set is indicated by using polar coordinates, they are converted as the latitude and longitude. Then we sort the fire into different classifications in terms of their property category. From 1 to 8, it represents the different fire happened places, including Boat, Dwelling, Non Residential, Other Residential, Outdoor, Outdoor Structure, Rail Vehicle, and Road Vehicle. The conversion rules also displayed in appendix A-A. Meanwhile, we created a new feature called Severity. According to the stop code description and special service type of fire call, the severity level is defined as 0.3, 0.6 and 0.9 which means the probability of incident occurs are 30%, 60%, 90%, respectively. These numbers are calculating from the incident based on actual fire event which are Chimney Fire, Secondary Fire, and Primary Fire. Primary Fire is 90 percent chance, Secondary Fire is 60 chance, and Chimney is the least dangerous for 30 percent chance. This is used as the correspondingly labels for the fire incident likelihood. We renamed data set at the end of cleaning. All the useful features with new names could be found in appendix A-C.

C. Accident Data

1) Data Merge: Since there are three data sets for accident data with the same categories but different time period, we merge them together so that they can be cleaned more conveniently.

2) Data Cleaning: First, we removed the features with no record because it only needs the date, weather, coordinates, and location of accident to predict. In addition, we filtered the unwanted features. Table II in appendix A shows the conversion of accident data features. A new column called accident severity percent which shows the probability percent of each severity occurring. The probability of accident to occur is based on casualties and severity of the event calculated as percent. They will be the labels for each accident data point. Renamed each feature after cleaning it. All the useful features with new names could be found in appendix A-D.

V. Algorithm Choices and Results

A. Crime Data

At first, we tried to use a tree regression to predict the coordinates of a crime given time and date parameters. After tree was performing quite badly, we implemented Support Vector Machine to increase the accuracy of the model. Unfortunately, the accuracy difference was not significant and made us think that we needed a different approach. We realized that predicting the coordinate was nearly impossible to calculate with our existing dataset. So, we converted our problem into a classification problem How likely for a given region to have a crime? this was our investigation question that we wanted to solve. With this approach we can generalize the dataset as London has several places which crimes occur frequently. Finally, random forest algorithm was chosen because it can handle multilabel classification problems such as this one. When compared to decision trees, it is less likely to overfit data and it is currently regarded as one of the most powerful algorithms available [3]. Furthermore, in a similar project on crime prediction [4], Alves et al. found great success using this algorithm for making crime predictions.

B. Drug incidents of the dataset

The dataset that we use has different types of crime labels such as firearm possession, vandalism, drug possession etc. some regions in London have similar types of crime for. More than 400 incidents of drug related crimes occurred near the borough which is followed by northern part of London such as Camden and Tower Hamlets. These places are known to have drug abuse by police and media[5]. Following figure shows the drug incidents of the data set.

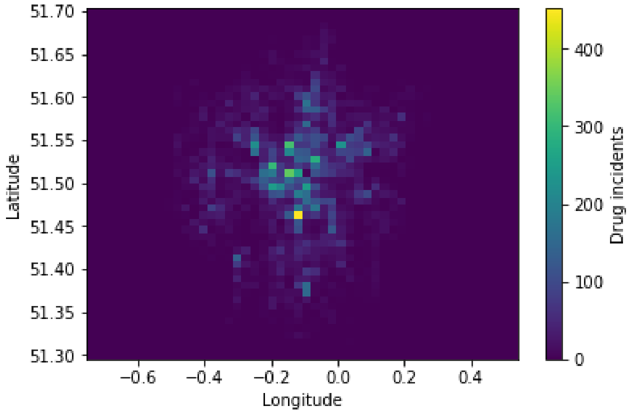


Fig. 3. Drug incidents of the dataset

C. Vandalism incidents of the dataset

There is another common problem in London which is vandalism. Vandalism, unlike drug related crimes dont have a specific region. Central London has high levels of vandalism rates. However, it is most likely due to high amount of police near populated areas. Our dataset has the incidents where the suspect is caught. On the other hand, in suburban parts of London would also have vandalism crimes but till police arrives, the suspects might escape. For this reason, central part of London seems like it has more crime, but outskirts of London may not report vandalism on public property incidents at all which makes it quite hard to catch the suspects. The heat map of vandalism can be seen below, the plot looks different than the previous plot. The incidents are spread around the city.

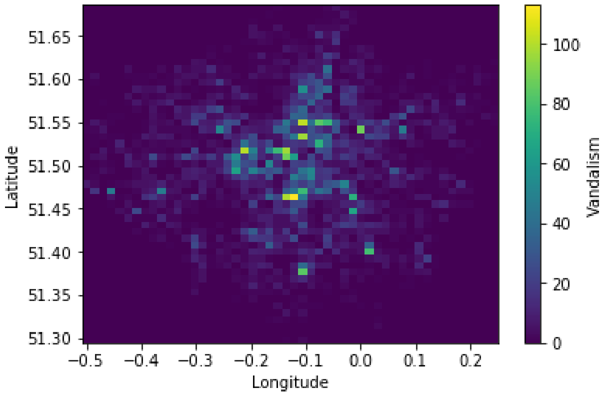


Fig. 4. Vandalism incidents of the dataset

D. Firearm possession incidents of the dataset

Even though there are strict regulations of firearm possession in UK, it is still a problem in London. As we expected the gun possession incidents were quite rare and they were distant from city centre. Since the dataset that we used had few gun possession incidents, our models were biased against other types of crimes. To solve that we used sampling to prevent the bias against firearm possession, thus our model at first would always predict the same crime due to unbalanced data points.

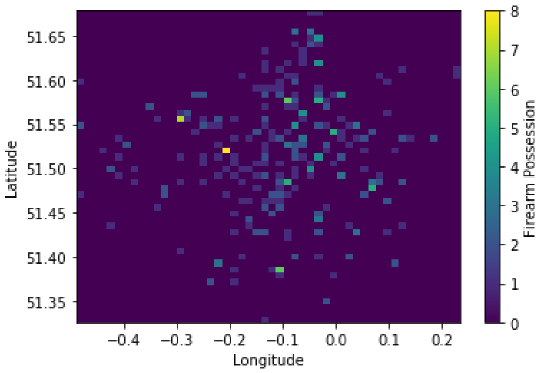


Fig. 5. Firearm possession incidents of the dataset

E. Importance of different features

The most important features were longitude and latitude values with 14% importance each. The next important features are the average population and household around the crime scene. The average households and average population affect the result 10% each. This is expected as more people living near the crime scene more people are likely to call the police and police has higher chance of catching the suspects in populated areas. On the other hand, in regions where population density is low has lower chance of witness calling the police. Weekday and Month seems to dont affect the possibility of crime happening. Day of Month feature was generated to see the trend of crime incidents in a month, at the end of the month between 25th till 31th crime rates rise significantly. This might be caused by pay check dates. Here is a bar chart shows the importance of different features:

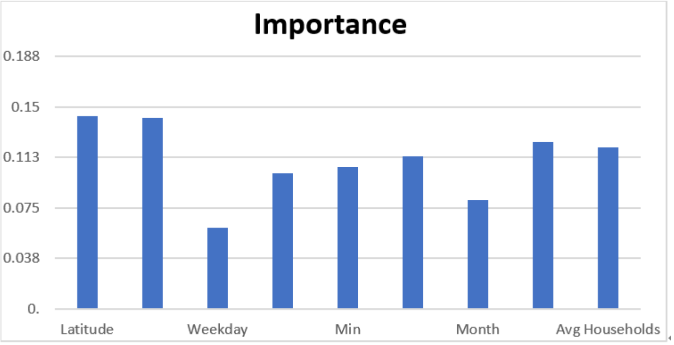


Fig. 6. Importance of different features

F. Fire Data

Firstly, the Linear Regression model, simplest machine learning algorithm, has been applied to the data set, and the Root Mean Square Error (RMSE) is used to measure the model accuracy. The training data set should large 70% of the total size, and the remaining is for the validation data set. The result RMSE 0.135, interpret as 13.5% of error in prediction. Next, the second algorithm is Lasso Regression with different alpha parameters range from -5 to 5. However the error in predicting results is equivalent as the linear regression model. Next, the decision tree model has been applied to the data set, and the predicting accuracy is acceptable. Therefore, Random Forest has been chosen as the optimal solution for the fire data set.

TABLE I
OPTIMAL MODEL PARAMETERS FOR RANDOM FOREST MODEL

Hyper Parameter	Value
max_depth	30
min_leaf_nodes	20
min_split	2
impurity	Gini

The result RMSE is 0.06, interpret as 6% of error in prediction.

G. Accident Data

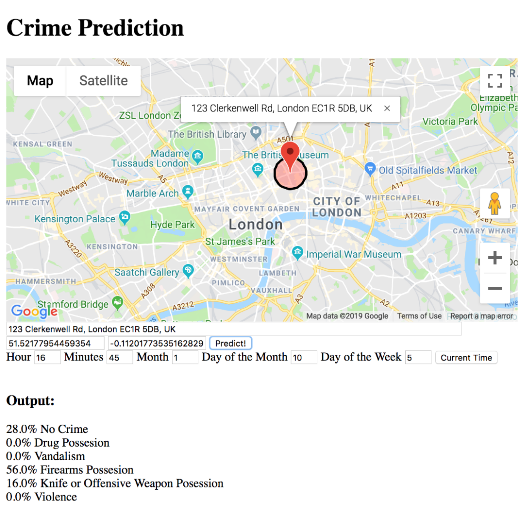
The convention of selecting suitable model for Accident data set is similar to the Fire data set, Linear Regression, Decision Tree and Random Forest are applied to the data set, and RMSE results are 0.004 and 0.007 for Linear Regression and

Random Forest, respectively. The Linear Regression model tend to fit the data well than other models, since it has the lowest RMSE. The optimal training set size is 66% of the total data set size, and the remaining is for the validation set size.

VI. Application

A.

Our prediction algorithms are fast, and they can create a responsive user interface. Even though we havent deployed the server on a real workstation, we were able to emulate the experience on localhost server of our machines. To make the form faster to fill, we added a button that fills the required fields with current time and date. It basically takes the time of the computer to fill the fields and if user desires to change some parameters again, they dont have to refill the form since their time is very precious. Following figure shows the prediction of application:



Crime Prediction

123 Clerkenwell Rd, London EC1R 5DB, UK

51.52177954459354 -0.11201773535162829

Hour 16 Minutes 45 Month 1 Day of the Month 10 Day of the Week 5

Output:

28.0% No Crime
0.0% Drug Possesion
0.0% Vandalism
56.0% Firearms Possesion
16.0% Knife or Offensive Weapon Posession
0.0% Violence

Fig. 7. Screenshot of Application

The full version can be found from https://github.com/jjayub/The-Final-Distribution.git

VII. conclusion

Bad thing is possible to occur at any time such as crime, accident and fire incident. The prediction of those incident beforehand is the solution to reduce the damage and loss of human lives. This project aims to achieve that objective with the intellectual of machine learning model. Random Forest, Linear Regression, Decision Tree, Lasso Regression are the algorithm which is chosen to accomplish this mission. The result shows that Fire Incident prediction and Crime prediction are favoured by Random Forest algorithm which held the most accuracy out of those algorithm choices. For the accident prediction, the linear regression is the optimal model used to predict the probability. Unfortunately, the result from fire incident prediction and accident prediction are heavily biased by the datasets which mostly contain records that has the severity of some value. This case the prediction might be biased to some prediction rather than the other value. In future works, improving the model is needed as well as gathering more data that is not bias or do some operation as the preprocessing to ensure that dataset is equally balance.

References

[1] J. Boysen (2017) *London Fire Brigade Calls* [Online]. Available: https://www.kaggle.com/jboysen/london-fire/version/1

[2] Visualise and analyse traffic demographics (2017) *EDA of 1.6 Mil Traffic Accidents in London* [Online]. Available: https://www.kaggle.com/yesterdog/eda-of-1-6-mil-traffic-accidents-in-london/data?fbclid=IwAR0tUmfPa4U90gIU5WqIei_3TuGO8ZMjRpDefCZeTtNvGT95caswYdeSJWI

[3] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, 2017

[4] L. Alves, H. Ribiero and F. Rodrigues, "Crime prediction through urban metrics and statistical learning," *Physica A*, vol. 505, pp. 435-443, 2018.

[5] F. Simpson(Dec 9, 2016) *Three London boroughs among UK's worst places for drug abuse* [Online]. Available: https://www.standard.co.uk/news/london/three-london-boroughs-among-uks-worst-places-for-drug-abuse-a3416136.html

Appendix A

A. Features of original fire data

TABLE II
Conversion of fire data features

Feature names	Feature description
Address qualifier	Location of fire call
Borough code	Borough of fire
borough name	Name of Borough
cal year	2017
date of call	Date of Call
easting m	Polar coordinate of easting
easting rounded	Polar coordinate of easting
first pump arriving attendance time	first pump arrived time
first pump arriving deployed from station	first pump deployed from station
frs	Fire and Rescue Service
hour of call	hour of call
incident group	fire incident
incident number	incident unique id
incident station ground	station incident location
northing m	Polar coordinate of northing
northing rounded	Polar coordinate of northing
Postcode district	Postcode district
Postcode full	Full postage
Proper case	Geolocation
Property type	Location of fire happened
second pump arriving attendance time	second pump arrived time
second pump arriving deployed from station	second pump deployed from station
special service type	description of special service
stop code description	classification of stop code
time of call	time of fire call
timestamp of call	standard timestamp of call
ward code	ward code of patient
ward name	ward name of patient
ward name new	new ward name
num pumps attending	number of pumps attend in fire

B. Features of original accident data

TABLE III
CONVERSION OF FIRE DATA FEATURES

Feature names	Feature description
Accident Index	Unique ID
Location Easting OSGR	Local British coordinates x-value.
Location Northing OSGR	Local British coordinates y-value
Longitude	Longitude
Latitude	Latitude
Police_Force	Number of police involved in the incident
Accident Severity	1 = Fatal, 2 = Serious, 3 = Slight
Number of Vehicles	Number of Vehicles
Number of Casualties	Number of Casualties
Date	In dd/mm/yyyy format
Day of Week	Numeric: 1 for Sunday, 2 for Monday, and so on
Time	Time the accident was reported, in UTC+0
Local Authority (District)	The number represent district in London
Local Authority (Highway)	The code represent highway in London
1st Road Class	This field is only used for junctions
1st Road Number	This field is only used for junctions.
Road Type	Some options are Roundabout, One Way, Dual Carriageway, Single Carriageway, Slip Road, Unknown.
Speed limit	limited speed
Junction Detail	Some options are Crossroads, Roundabouts, Private Roads, Not a Junction.
Junction Control	A person, a type of sign, automated, etc
2nd Road Class	This field is only used for junctions.
2nd Road Number	This field is only used for junctions.
Light Conditions	Day, night, street lights or not.
Weather Conditions	Wind, rain, snow, fog.
Road Surface Conditions	Wet, snow, ice, flood.
Urban or Rural Area	The number represent urban or rural area in the incident's location
Did Police Officer Attend Scene of Accident	Is police attend the scene of accident
LSOA of Accident Location	LSOA code of accident location
Year	Year of accident

C. Features of cleaned fire data

1) Property category
2) Severity
3) Day
4) Hour
5) Latitude
6) Longitude

D. Features of cleaned accident data

1) Longitude Latitude
2) Day of Week
3) Weather Conditions
4) Month
5) Day
6) Hour
7) Accident Severity Percent

E.

TABLE IV
COVERSION OF FIRE DATA FEATURES

Property Category	Numerical representation
boat	1
Dwelling	2
Non Residential	3
Other Residential	4
Outdoor	5
Outdoor Structure	6
Rail Vehicle	7
Road Vehicle	8

F.

TABLE V
COVERSION OF ACCIDENT DATA FEATURES

Weather Conditions	Numerical representation
Fine with high winds	1
Fine without high winds	2
Fog or mist	3
Raining with high winds	4
Raining without high winds	5
Snowing with high winds	6
Snowing without high winds	7
Other	8
Unknown	8