The 10h IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2019
5-7 June, 2019, Leeds, United Kingdom

Data Analytics: Factors of Traffic Accidents in the UK

Steven Haynes, Prudencia Charles Estin, Sanela Lazarevski, Mekala Soosay, Ah-Lian Kor
School of Computing, Creative Technologies and Engineering,
Leeds Beckett University,
Leeds LS6 3QS, UK
S.Haynes6869@student.leedsbeckett.ac.uk ; P.EstinCharles8886@student.leedsbeckett.ac.uk;
{S.Lazarevski, M.Soosay, A.Kor}@leedsbeckett.ac.uk

***Abstract*— The traffic and accident datasets for this research are sourced by Data.gov.uk. The data analytics in this paper comprises three levels namely: descriptive statistical analysis; inferential statistical analysis; machine learning. The aim of the data analytics is to explore the factors that could have impact on the number of accidents and their associated fatalities. Some of the factors investigated on are: time of the day, day of the week, month of the year, speed limits, etc… Machine learning approaches have also been employed to predict the types of accident severity.**

Keywords— data analytics, machine learning, traffic accidents

I. INTRODUCTION

In this paper, we examine the relationship between traffic congestion, and traffic accidents. In addition, we investigate the level of traffic accident seriousness with traffic congestion as a possible underlying cause. In their article, [1] emphasise that different cities and towns will have their own constraints in dealing with traffic congestion. This is subject to various factors: geographical, historical, and socio-economic mechanisms, road networks of how cities were all developed in the first place. Furthermore, authors of [2] suggest that there could be many reasons for traffic congestion such as limited road capacity, time of the day, number of accidents, increase of vehicles at certain time of the day and certain parts of the road, road works that narrow the lanes, and affects traffic flow, as well as bad weather that could result in partial or full traffic congestion.

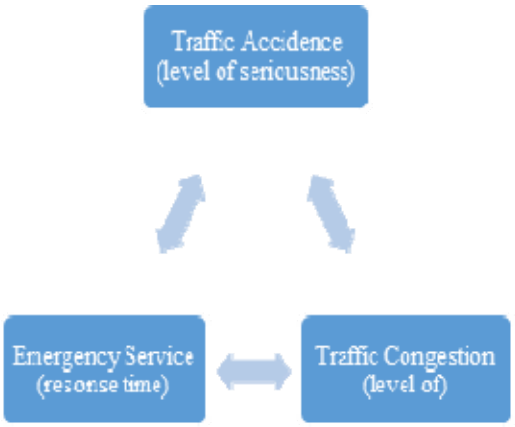

Figure 1: An overview of possible interactions amongst traffic accidents, emergency service and traffic congestion

Our research uses open date data published by UK central government [3, 4, 5] to establish the following research objectives:

- To investigate whether the time of the day has effect on the number of casualties per accident;
- To explore opportunities for reduction in the response time for the emergency services;
- To use machine learning to classify the severity of accidents at the first instance of being reported;
- To test the following hypothesis:

  Null Hypothesis (H_0):
  The mean number of casualties per accident is the same throughout the day.

  Alternative Hypothesis (H_1):
  The mean number of casualties per accident is not the same throughout the day.

We conduct a comparative analysis using datasets relating to traffic and emergency services. To support the analysis, tools such as Tableau, SAS and Python are used to prepare and analyse the data. Analysis is conducted on different types of accidents, casualties, and impact of speed limit on occurrences of accidents. A conducted literature review reveals the importance of data analytics on traffic and casualty. Several key factors of accidents (e.g. traffic congestion) are also reviewed. Statistical analysis methods used are descriptive statistics, ANOVA, inferential statistics and machine learning. An overview of the data analytics conducted in this research is depicted in Fig. 1.

This paper is structured as follows: Section II: Literature Review; Section III: Methodology; Section IV: Results and Discussion; Section V: Conclusion and Future Work.

978-1-7281-1733-1/19/$31.00 ©2019 IEEE

II. LITERATURE REVIEW

Undeniably, growth of urban areas is a cause of traffic congestion increase [6], few have discussed the increase in human population as one of the factors [7]. The population increase is expected to continue at an average annual growth of 0.51%. At this rate, the population of the United Kingdom is predicted to be 77.5m by 2050 [8]. However, there is no explicit link between traffic congestion and the type of accidents and possible road casualties. In 2016, there are 1,792 reported road deaths, an increase of 4% compared to 2015. The amount of traffic on Britain's roads also increases by 2.2% [9]. The point raised is how this would affect ambulance response time if there are minor changes to the infrastructure.

Literature shows that ambulance response times are critical to the following: hospitalization, rehabilitation, and survival following an accident, stroke or heart attack [10, 11, 12]. The author of [10] uses distance from the hospital as an instrument to investigate the impact of response time on mortality and hospital utilization. Although first emergency response services are typically expeditious, one factor beyond their control that a□ects response times is tra□c congestion. Other authors, [13] support this argument by pointing out that emergency services rely heavily on well-functioning road infrastructure. Moreover, [14] argues that the increase in traffic over past decades contributes to impact on ambulance response times across England. This results in missed target times when responding to life-threatening calls that require immediate attention. This increase in response times is mirrored when assessing the response times of the Fire and Rescue authority (FRA) [15]. As seen in Fig. 2**,** the average FRA response time has continually increased since 1994. With these statistics in mind, it could be said that the problem of fatalities on Britain's roads will not be improved in a business as usual context.

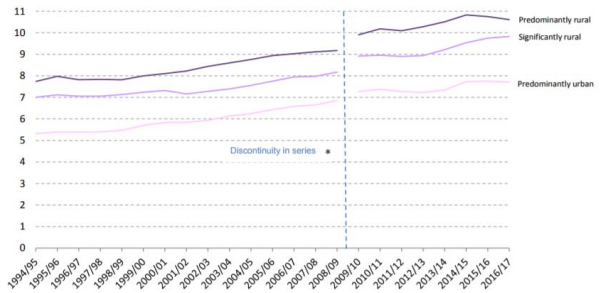

Figure 2. Average response times (minutes) by FRA type, England; 1994/95 to 2016/17 (source: Home Office, 2018).

III. METHODOLOGY

In this section, we discuss the data lifecycle with the following phases: data gathering, data exploration and cleansing, objectives identification, analysis and evaluation of results (see Fig. 3).

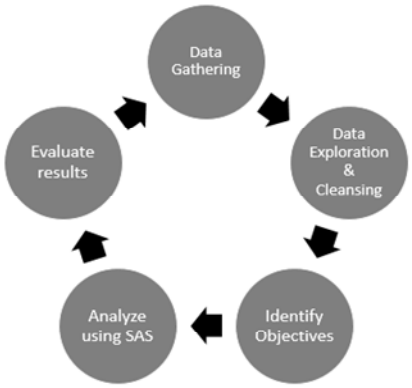

Figure 3: Data lifecycle

A. *Data Analysis Techniques*

The data obtained is already clean. All missing values have been imputed. Table 1 below, details the software packages used and their corresponding application.

TABLE I. SOFTWARE PACKAGES USED FOR DATA ANALYTICS

Software Package/Coding Library	Used For:
SAS Studio	Statistical testing – ANOVA, distribution analysis, descriptive statistics
Tableau	Interactive descriptive dashboard
Microsoft Excel	Data manipulation
Python / Jupyter Notebook	Machine learning & Data manipulation. Libraries used: **SciKit learn, pandas, numpy**
Matlab	Scientific graph plotting

B. *Data Gathering*

The datasets are obtained from the Department of Transport [5]. They provide detailed road safety data on all accidents which occur in the UK during 2016 [3, 4]. To reiterate, the number of records in the dataset is 136621. They comprise the following data: traffic and accidents; vehicles and drivers; and emergency services response time.

C. *Data Exploration and Cleansing*

Data that is not required for the analysis is filtered while missing data is imputed using means

D. *Identification of Objectives*

After the exploration of data, specific objectives are identified (see Section I).

E. *Data Analysis Using SAS and Evaluation of Results*

The chosen dataset for this research is analyzed using several statistical analysis techniques: descriptive statistics,

inferential statistics, and machine learning. Subsequently, analysis results are interpreted.

IV. RESULTS AND DISCUSSION

To reiterate, there are three levels of data analysis: descriptive statistics, inferential statistics and machine learning.

A. Descriptive Statistical Analysis

i. *Tableau interactive dashboard*

We create an interactive Tableau dashboard for data visualization. This helps us explore the data and also provides actionable insights. Fig. 4 shows an example of graphs produced It shows that in the year 2016, approximately 1.24% of the total traffic accidents in the UK is fatal. Additionally, single carriageway has the highest number of fatal accident occurrences. Fridays seem to have the highest number of road accident incidents while weekends, the lowest. Places with speed limits of 60 mph and 30 mph seem to be the highest contributors to fatal accidents.

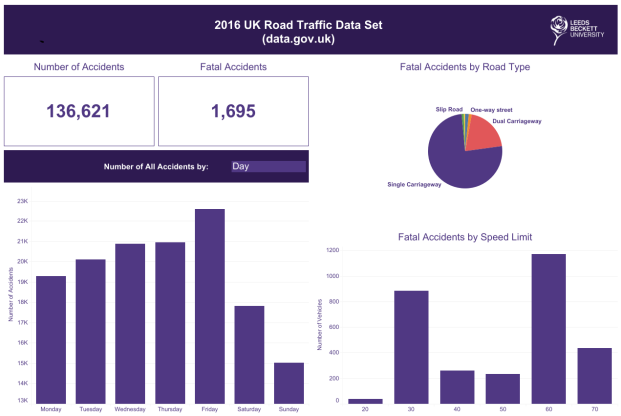

Figure 4. Dashboard – Accidents (Fatal), Road Type and speed limit (mph) by Day of the week.

ii. *SAS descriptive statistics*

TABLE II. DESCRIPTIVE STATISTICS FOR RELEVANT VARIABLES IN THE TRAFFIC ACCIDENT DATA SET (2006)

Variable	Mean	Std Dev	Minimum	Maximum	N
Accident_Severity	2.8161703	0.4181516	1.0000000	3.0000000	136621
Number_of_Vehicles	1.8481785	0.7101174	1.0000000	16.0000000	136621
Number_of_Casualties	1.3276436	0.7892962	1.0000000	58.0000000	136621
Day_of_Week	4.1158899	1.9172917	1.0000000	7.0000000	136621
1st_Road_Class	4.1436382	1.4562151	1.0000000	6.0000000	136621
1st_Road_Number	919.9971454	1753.63	0	9918.00	136621
Road_Type	5.1837346	1.6580716	-1.0000000	9.0000000	136621
Light_Conditions	1.9835091	1.6769377	-1.0000000	7.0000000	136621
Weather_Conditions	1.5530995	1.6905406	-1.0000000	9.0000000	136621
Road_Surface_Conditions	1.2916975	0.5882472	-1.0000000	5.0000000	136621
Did_Police_Officer_Attend_Scene_	1.2539653	0.4646575	1.0000000	3.0000000	136621
Speed_limit	37.9436830	14.0416695	20.0000000	70.0000000	136584

The total number of records in the dataset is 136621 and the descriptive statistics for relevant features are shown in Table 2.

Number of casualties: The mean for this variable is 1.328 while the range is considerably large (i.e. 57) and the standard deviation is 0.789.

Day of week: The mean for the day of the week is 4.116, range is 6 while the standard deviation is 1.917. **Speed limits:** Speed limits in the dataset are discrete with the following values: 20, 30, 40, 50, 60 or 70 mph. The speed limit range is 50 mph, standard deviation is 14.042, and mean is 37.944. **Number of vehicles:** The range value is 15, standard deviation is 0.710, and mean is 1.848.

iii. *Accidents and casualties by month*

The following graph (in Fig. 5) plots the number of accidents along with the number of fatalities by month. As expected, it shows a correlation between the number of accidents and the number of fatalities. It also shows that the *sum of casualties* seems to be constant (i.e. skewness value of 0.10) throughout the months while the *count of accidents* has a moderate negative skew (i.e. skewness value of -0.8). These findings suggest that the count of accidents slightly increases **a**t the end of the year.

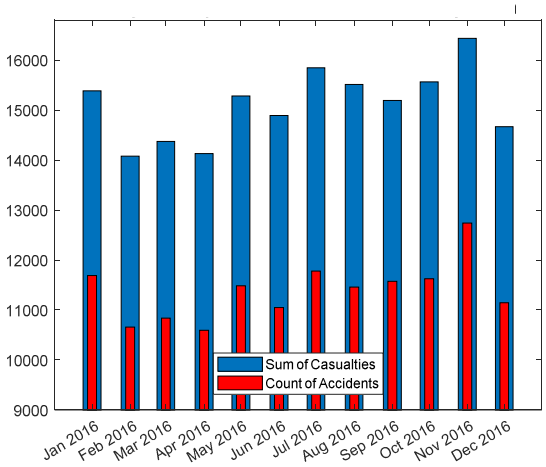

Figure 5. Overlay Bar graph – Accidents & Casualties by Month

iv. *Accidents and casualties by day of week*

The following graph (Fig. 6) plots the number of accidents along with the number of fatalities by day of the week. The *sum of casualties* has a skewness of -0.45 while the *count of accidents*, -0.94. Fig. 6 reveals that once again, Fridays peak

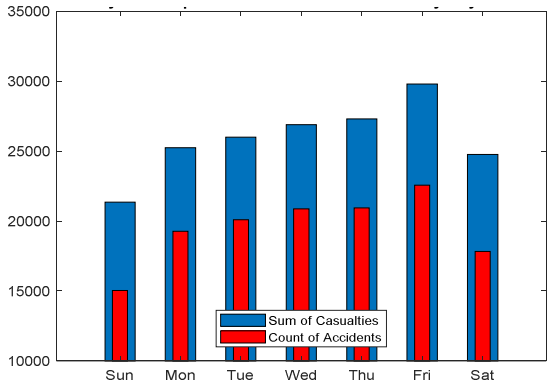

Figure 6. Overlay Bar graph – Accidents & Casualties by Day of Week

in the total number of accident occurrences and number of casualties. The kurtosis value for the *sum of casualties* is 1.41 while the *count of accidents* is 0.97. These positive kurtosis values are less than 3.00 (for normal distribution) and this implies that the distribution is shorter, and tails are thinner than the normal distribution.

v. Speed limits, fatalities and accidents

Mean casualties per accidents by speed limit

As seen in Fig. 7, the casualties mean seems to increase proportionately with the speed limit. As the speed limit increases, there is a slight increase in the number of casualties per accident.

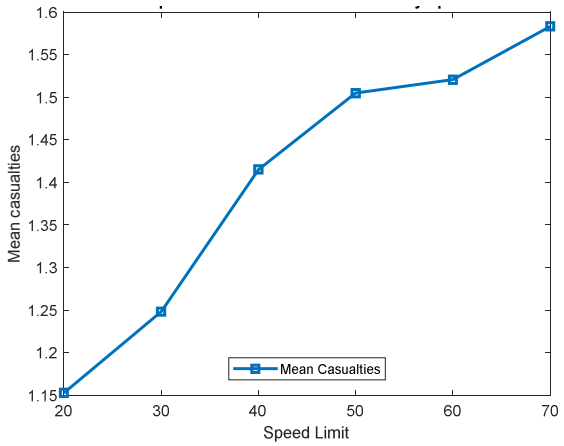

Figure 7. Line Graph– Mean number of casualties versus speed limit

Number of Accidents and Casualties by speed limit

Fig. 8 reveals that the majority of accidents occur on the 30 mph speed limit roads, resulting in the highest number of casualties compared to other speed limits. However, the average number of casualties (shown in Fig. 7) for the 30mph speed limit seems to be lower than speed limits from 40mph-70 mph. Additionally, Fig. 4 depicts the total number of vehicles involved in fatal accidents and it seems to indicate that roads with 60 mph speed limit have the highest number of fatal accidents followed by roads with 30mph speed limit.

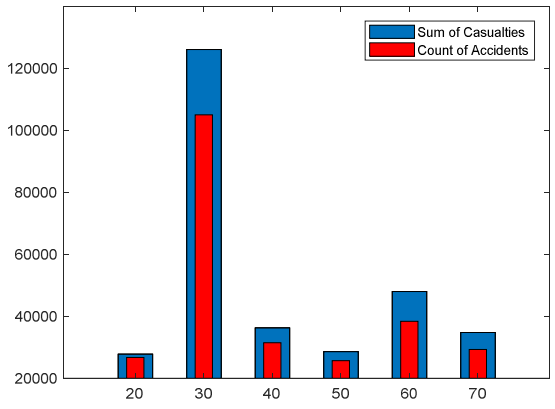

Figure 8. Sum of casualties and count of accidents by speed limit

Number of fatal accidents by speed limit

The number of accidents that result in at least one fatality is depicted in Fig. 9. Roads with 20mph speed limits seem to have the lowest number of accidents (with at least 1 fatality) while roads with 30mph and 60 mph speed limits seem to have the highest incidents.

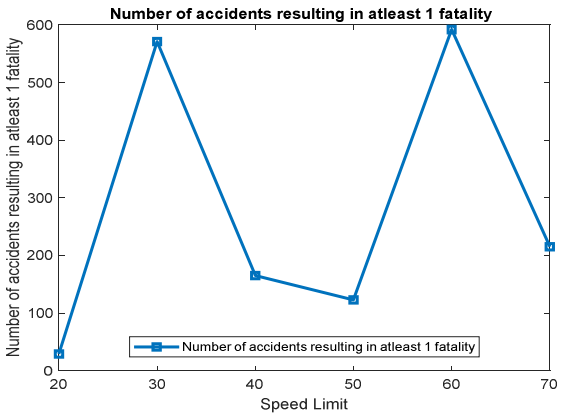

Figure 9. Line Graph– Number of accidents resulting in at least 1 fatality by speed limit

Percentage of fatal accidents by speed limit

Fig. 10 presents the percentage of accidents resulting in at least one fatality for speed limits ranging from 20mph to 70 mph. Roads with 60mph seems to peak while 30mph scores second lowest in this. However, when compared to Fig. 9, the 30mph and 60 mph speed limits seem to peak in the number of incidents. This implies that a low percentage of road accident occurrences results in fatality for the 30mph limit while the 60mph speed limit, conversely, has the highest percentage.

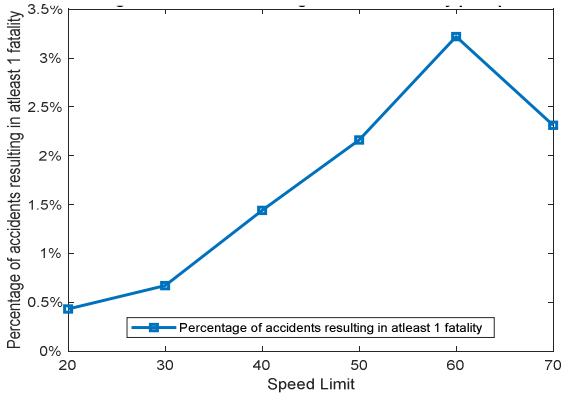

Figure 10. Line Graph– Percentage of accidents resulting in at least 1 fatality per speed limit

B. Inferential Statistical Analysis

Using the traffic dataset, we investigate whether the time of accident affects the number of casualties per accident. The test commences with a normality test for the dataset in order to decide on a subsequent parametric or non-parametric test.

i. Distribution Analysis - Histogram and Q-Q Plot

The null hypothesis Ho is that the dataset has a normal distribution while H1 is the alternative hypothesis (i.e. the dataset is not a normal distribution at a level confidence, α = 0.05). In Fig. 11, the p-value shown is 0.2069 and it is greater than α = 0.05. Thus, the null hypothesis, Ho is accepted, and this implies the dataset seems to have an approximate normal distribution (note: this is supported by the Q-Q plot).

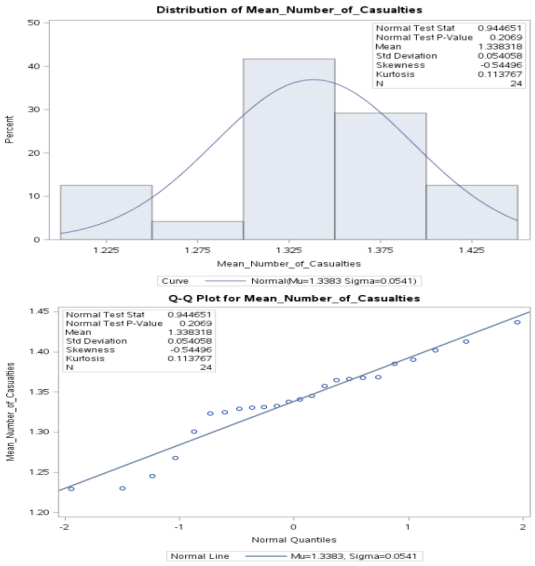

Figure 11. Histogram and Q-Q plot for mean number of casualties per

ii. One-way ANOVA test

After the normality test of distribution, we conducted a one-way ANOVA test to determine whether there is any significant difference of the number of casualties and the time of the day. Preparatory procedures entailed are: group the number of casualties by the hour of the day (1 to 24).

Class Level Information		
Class	Levels	Values
Time + 1	24	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Number of Observations Read	136621
Number of Observations Used	136619

Dependent Variable: Number_of_Casualties

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	289.43071	12.58394	20.26	<.0001
Error	136595	84822.69819	0.62098		
Corrected Total	136618	85112.12890			

Figure 12. ANOVA Results for number of casualties and time of the day

The dependent variable for the one-way ANOVA analysis is the Number of casualties and the categorical independent variable is Time (hour of the day) which takes on values from 1 to 24. The null hypothesis, Ho, for this test that there is no significant difference amongst the number of casualties for hour 1 to 24. As for the alternative hypothesis, H1, there is a significant difference amongst the *number of casualties* for hour 1 to 24. The p-value obtained for this analysis is < 0.001. This means that at a level of confidence, α = 0.05, H_o is rejected. This implies that there is a significant difference amongst the *number of casualties* for hour 1 to 24.

C. Machine Learning – Accident Severity Classification

Machine learning is the application of scientific algorithms that interpret data. These algorithms are fundamentally giving computers the ability to learn from the data that is input. There are different types of machine learning algorithms, each with their own specific purpose (see Fig. 13).

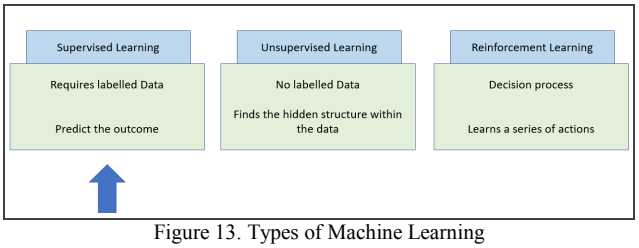

Figure 13. Types of Machine Learning

This paper presents models from the supervised learning category. The goal of supervised learning is to predict or group past observations using classification. Supervised learning follows an iterative process as seen in Fig. 14 with input boxes relative to the report.

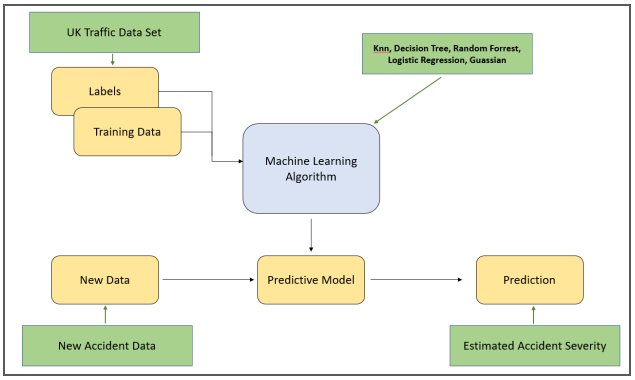

Figure 14. Supervised learning (adapted from Raschka, S et al (2017))

i. Accident Severity Classifier (Part 1)

Feature selection

In line with the goal of the model, relevant features are pre-determined. Table 3 documents the features and the characteristics.

Data Manipulation & Normalisation:

To maximise a model's performance and efficiency, it is appropriate to normalise the data. This ensures that the model achieves maximum accuracy on its predictions. Machine learning models are often computationally expensive, therefore the normalisation process rescales features in the range of [0,1]. This results in lesser computational requirements, resulting in faster run times. A well-known procedure that transforms a set of observations is the Principal Component

Analysis (PCA). Both the longitude and latitude features have been normalised for the model. The time feature is also transformed into a continuous variable, by calculating the minute of the day in which the accident occurred.

TABLE III. FEATURES SELECTION

FEATURES	VARIABLE TYPE	NOTES
Accident Severity	Categorical	1 = Fatal 2 = Serious 3 = Slight
Longitude	Continuous	
Latitude	Continuous	
1st road class	Categorical	Numerical identifier for road classes, i.e motors, A roads etc.
Day of week	Categorical	1 = Monday, 2 = Tuesday …
Light Conditions	Categorical	Non-linear numerical identifier for light conditions
Weather conditions	Categorical	Non-linear numerical identifier for weather conditions
Urban or rural?	Categorical	1 = urban, 2 = rural
Speed limit	Categorical	1 = 20, 2 = 30 …
Number of vehicles	Discrete	Number of vehicles involved in the accident
Road surface conditions	Categorical	Numerical identifier for the road surface conditions
Time	Continous	Minute of the day in which the accident happened

Model Selection and outcome

The non-linearly structured data and the prevalence of categorical variables mean that certain models are better suited to the problem than others. The total number of observations: for model building is 136,615. The train/test components are: 95,645 (70%) / 40,970 (30%). Table 4 depicts the performance of 4 different machine learning techniques with Random Forest showing the best performance (i.e. 82.54% accuracy score). Thus, Random Forest seems to be an effective classifier of accident severity.

TABLE IV. MASHINE LEARNING MODELS FOR ACCIDENT SEVERITY

MACHINE LEARNING MODEL	ACCURACY SCORE
Random Forest	82.54%
Gaussian	81.38%
KNN	80.39%
Decision Tree	71.08%

Fig. 15 presents the Random Forest Confusion Matrix. Results reveal that the percentages for the correct prediction of true positive values for: Accident severity type 1 is 0.42%; Accident severity type 2 is 8.23%; and Accident severity type 3 is 94.9%.

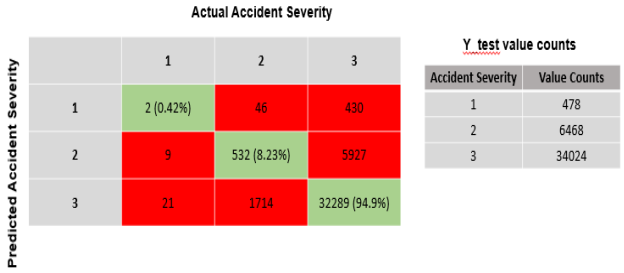

Figure 15. Random Forest Confusion Matrix

Hyper-tuning models could often result in accuracy improvements. The process of hyper-tuning includes the adjustment of a model's parameters. An example of hyper-tuning would be setting the maximum depth in a Decision Tree model which could result in an increase of the model's accuracy by a few percent. Considering the extremely low true positive value (for Accident Severity type 1) produced by the Random Forest model (shown in Fig. 15), the hyper-tuning process is abandoned. On the other hand, in order to improve the performance of the accident severity classifier, we search for additional data that could be incorporated during the initial reporting of an accident. As well as the traffic accident dataset, the Gov.UK website provides a dataset holding information on the vehicles and drivers involved in the reported accident [4]. This dataset is joined to the original dataset via the Accident Index. Random Forest version 2 is run on the aggregated dataset and the outcome is shown in the ensuing section.

TABLE V. ADDITIONAL FEATURES

Feature	Categorical, discrete or continuous?	Notes
Vehicle Type	Categorical	Non-Linear numerical representation of the type of vehicle
Age of the driver	Continuous	
Age of the vehicle	Continuous	

ii. Enhanced Accident Severity Classifier (Part 2) – Random Forest V2.0

To reiterate, the inclusion of additional features (i.e. information on vehicles and drivers – see Table 5) in the dataset seems to increase the accuracy of the Random Forest model (see Fig. 16).

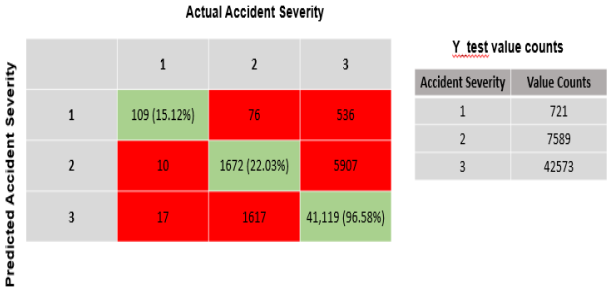

Figure 16. Enhanced Random Forest Confusion Matrix

The model's accuracy seems to have increased from 82.54% to 85.08%. The percentages for the correction prediction of true positive values are: Accident severity type 1 (increases from 0.42% to 15.12%); Accident severity type 2 (increases from 8.23% to 22.03%); and Accident severity type 3 (increases 94.9% to 96.58%).

V. CONCLUSION

Results from the inferential statistics suggest that road users are more susceptible to being involved in an accident at certain times of the day. Further analysis into this could highlight

specific times of the day when road users are more likely to be involved in road accidents.

The application of several machine learning models reveal that the accuracy score of the Random Forest model is approximately 85%, although the true positive rate for accidents of a fatal severity type 1 is less than 20%. This could be further improved by selecting only effective features. This could be done by applying associative rules to help uncover correlations between the dependent variable (Accident severity) and other features. Additionally, another possibility is to increase the number of records for Accident severity type 1.

VI. REFERENCES

[1] Wang, J., Dong, W., He., K., Gong, H., and Wang, P. (2015). Encapsulating Urban Traffic Rhythms into Road Networks. Scientific Reports 4, nNo. 1 (May 2015), Article: 4141. https://doi.org/10.1038/srep04141.

[2] Al-Kadi, O.mar, Osama Al-Kadi, O., Rizik Al-Sayyed, R., and Ja'far Alqatawna, J. (2014). Road Scene Analysis for Determination of Road Traffic Density, Frontiers of Computer Science 8, nNo. 4 (August 1, 2014), : pp. 619–28. https://doi.org/10.1007/s11704-014-3156-0.

[3] Data.gov.uk. (2017a). Road Safety Data – Accidents 2016. [online] Available at: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (Accessed 01 March 2018)

[4] Data.gov.uk. (2017b). Road Safety Data – Vehicles 2016. [online] Available at: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (Accessed 08 May 2018)

[5] Department of Transport. (2017). Road Safety Data. [online] Available at: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (Accessed 01 April, 2018).

[6] Kozlak, A., and Wach, D. (2018). Causes Of Traffic Congestion In Urban Areas. Case of Poland, SHS Web of Conferences 57, 01019, InfoGlob 2018, https://doi.org/10.1051/shsconf/20185701019

[7] Department for Transport. (2016b). Road Use Statistics Great Britian 2016, [online] Avaiable at https://www.gov.uk/government/statistics/road-traffic-estimates-in-great-britain-2016 (Accessed 25 April, 2019). (2016). Gov-UK.

[8] National Infrastructure Commission. (2016). "The impact of population change and demography on future infrastructure demand, [online] Available at https://www.gov.uk/government/publications/the-impact-of-population-change-and-demography-on-future-infrastructure-demand (Accessed 25 April, 2019). ". (2016). National Infrastructure Commission. Gov-UK: London, UK

[9] Department for Transport. (2016a). Reported road casualties in Great Britain: 2016 annual report, [online] Available at https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2016 (Accessed 25 April, 2019).

[10] Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes?, Health eEconomics, 2013, 22 (7), pp.790–806, doi: 10.1002/hec.2851.

[11] Emberson, J., Kennedy, R. L., Lyden, P., Blackwell, L., Albers, G., Bluhmki, E., Brott, T., et al. (2014). Effect of Treatment Delay, Age, and Stroke Severity on the Effects of Intravenous Thrombolysis with Alteplase for Acute Ischaemic Stroke: A Meta-Analysis of Individual Patient Data from Randomised Trials. The Lancet 384, nNo. 9958 (November 2014), : 1pp.1929–35. https://doi.org/10.1016/S0140-6736(14)60584-5.

[12] Jena, A.nupam B., N. Clay Mann, N. C., Leia N. Wedlund, L. N., and Andrew Olenski, A.. (2017). "Delays in Emergency Care and Mortality during Major U.S. Marathons." New England Journal of Medicine 376, nNo. 15 (13 April 13,, 2017), pp.: 1441–50. https://doi.org/10.1056/NEJMsa1614073.

[13] Beland, L., and Brent, D. (2018). Traffic Congestion, Transportation Policies, and the Performance of First Responders, 2018. https://doi.org/10.13140/rg.2.2.10201.42088.

[14] Morse, A. (2017). NHS Ambulance Services. National Audit Office: London, UK, [online] Available at https://www.nao.org.uk/wp-content/uploads/2017/01/NHS-Ambulance-Services.pdf (Accessed 25 April, 2019).

[15] Home Office. (2018). "Response times to fires attended by fire and rescue services: England, April 2016 to March 2017", [online] Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/675939/response-times-fires-england-1617-hosb0318.pdf (Accessed 25 April, 2019).