UK Traffic Accidents:
Spatiotemporal visual analysis

Zohra Bouchamaoui

Abstract— This paper aims to analyse visually road traffic accidents in the United Kingdom between 2005 and 2014. Temporal and spatial data is analysed to identify seasonality patterns while focusing on the number and the severity of accidents over time and geographically. From there, supervised machine learning methods were used to help preventing future accidents by predicting the severity of the accidents using surrounding factors such as weather, road type, light, speed limit. Logistic Regression and Naïve Bayes algorithms are used and their performance is measured using metrics such as F1-score, precision, recall and AUC. Logistic Regression performs better than Naïve Bayes with an accuracy of 0.53% compared to 0.50% .

1 Problem Statement

According to the Department for Transport(2019), road accidents account for 1,752 deaths and 25,945 serious injuries in the United Kingdom. In order to gain an understanding of the problem at hand and to develop prevention mechanisms to road accidents, we will analyse the UK traffic accident data provided by the Department for Transport on a spatiotemporal dimension. The visualisations in this paper will allow humans to interpret their outputs and recommend a course of action. Additionally, if the hotspots where accidents happen could be predicted, this could potentially help decrease the number of casualties each year. For this purpose, we will be answering the following questions in this paper:

- When do accidents usually happen and what is their frequency pattern?
- What are the main factors that characterises the cause of an accident? And what is their relationship with accident severity?
- How accurately can we predict the severity of accidents?

This dataset contains temporal(date, time) and spatial information(latitude, longitude) making it suitable for our analysis. Furthermore, the data contains sizeable variety of factors which will prove useful when predicting the severity of an accident. In the following section, we investigate papers with similar research questions or dataset and their visualisation tools.

2 State of the Art

Fang et al.(2018) research examines spatiotemporal visualisation of expressway traffic accident information. The aim of this paper is to improve the collection of traffic accident information. This paper uses GIS road map data to improve spatiotemporal visualisation analysis. For temporal analysis, calendar graphs were used to show the frequency of daily traffic accidents over the months. Visual mapping helps transform data by depict the attributes to the encoding of visual items. In our paper, our temporal analysis will be visualised using heatmaps and focus at first on the frequency of car accidents on an hourly basis, throughout the days of the week and then look at the frequency on weekdays throughout the months. To create scalable figures for effective visual analysis, we aggregate the number of accidents. For this, Tableau will be used as the visual interface which will allow the user to select attributes or metrics to specify the outlook of the analysis.

Bhawker (2018) identified factors that can influence traffic accidents in the UK for the years 2016 and 2017. The main

aspects their paper focuses on are the location of the accidents, speed limits, and the age and gender of the drivers. Another aspect of their studies is sentiment analysis of people when accidents occur. Bar charts and stacked bar charts were used for feature analysis. For spatial analysis they used Tableau to create maps displaying the speed limits using different colours.

We will be using Tableau for our visualisations and the focus of our study will be on the severity of the accidents rather than on the factors. We will display the accident severities as points, with different colours, on maps using the longitude and latitude information for each year of the dataset. This is because there is no appropriate way to represent both temporal and spatial aspects together(Adrienko et al., 2017). Given the information provided in our dataset we will not be doing a sentiment analysis but rather a prediction using machine learning methods, similarly to Almamlook et al.(2019).

In their paper Almamlook et al. (2019) focused on predicting traffic accident severity. The aim of their investigation is to help prevent injuries, death and property damage. To visualise their results, bar charts were used as well as tables displaying the metrics of the machine learning algorithms. Random Forest, Naïve Bayes, Logistic Regression, and AdaBoost were used for their predictions. The outcome of the paper calls attention to Random Forest as a promising tool for traffic accident prediction with an accuracy of 75.5% compared to 74.5%, 73.1%, and 74.5% accuracy for Logistic Regression, Naïve Bayes and AdaBoost, respectively.

In our paper, we will be using Logistic Regression and Gaussian Naive Bayes to predict the severity of car accidents. Due to the large size of the dataset and memory costs, we will be using a small proportion of the dataset for our forecasts. Additionally, we will visualise the results of the recall, precision, F1-score metrics, as well as the area under the curve (AUC) using bar charts.

3 Properties of the Data

3.1 Dataset

This dataset (Kaggle, 2017) contains detailed road safety data in Great Britain between 2005 and 2014, which was reported by the police using the STATS19 accident forms. It is published by the British Department of Transport under an Open Government Licence (Data.gov.uk, 2021).
The data contains three merged tables which include information about the accidents and their severity (fatal, serious, slight), the casualities (driver or rider, passenger, pedestrian) and the type of vehicles (pedal cycle, motorcycle, car, bus or coach, tram).

The data has 4,287,593 rows and 67 features, and all the variables are coded numerically instead of being categorical. We are provided with both the location and the time of the accidents, with the latitude, the longitude as well as the date, hour and day of the week information. This will allow us to perform a temporal analysis of the data over multiple angles, i.e., weekdays and hours, months and weekdays for each year.

The data also contains features such as weather, road type, speed limit, and the road surface condition. We will look at the correlation between these factors and the severity of an accident.

In *Figure 1*, we can see that there is a high imbalance between the severity of the accidents. The majority of the accident are considered as '*Slight'* meaning that the injuries were minimal or none.

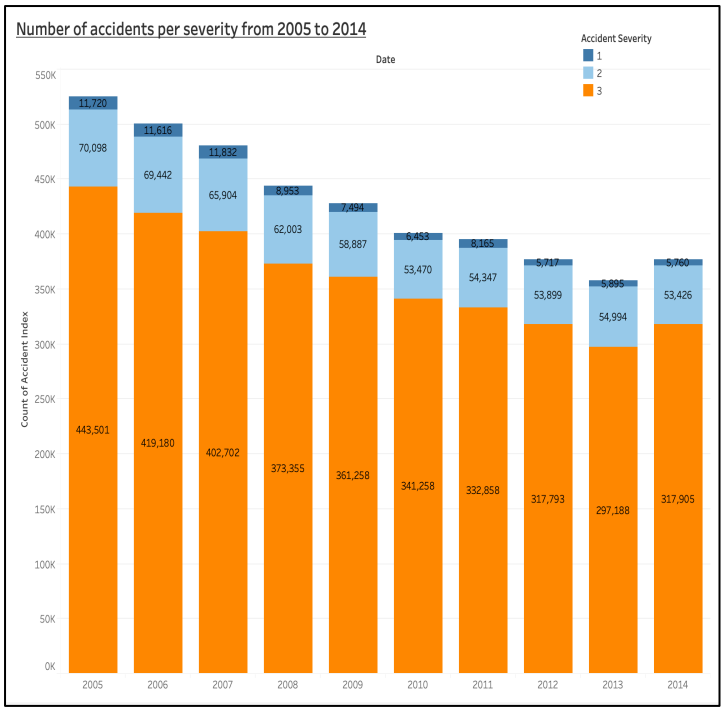

Fig.1. Number of accidents over the years, grouped by severity: 1 = Fatal, 2 = Serious and 3 = Slight.

3.2 Missing values

Before looking at the missing values, we started by dropping columns that are redundant: *'Location_Easting_OSGR'* and '*Location_Northing_OSGR'* as we will not be using QGIS for this analysis but *Tableau* and *Python*. We also dropped the '*LSOA_of_Accident_Location'* column as it is redundant with the '*Longitude'* and '*Latitude'* columns.

There are 138 missing values in the '*Longitude'* and '*Latitude'* columns as well as 151 missing values in the '*Time'* column. Due to the large size of our dataset, we decided to drop the rows with the empty values as there will be no noticeable impact on our analysis.

3.3 Outliers

After plotting the number of accidents and casualities between the years 2005 and 2014, we can see that the number of car accidents surpasses significantly the number of victims. *Figure 2* shows an irregular value on the 6th of May 2013 where the number of casualities increased sharply and exceeded the highest number of accidents recorded. We then examined the accident data on that date to identify if this peak is an outlier. As can be seen on the map in *Figure 2*, the accidents are dispersed widely around the UK. Therefore, this peak is not an outlier as it does not represent one region and one accident, and we will not remove it from our data during the analysis.

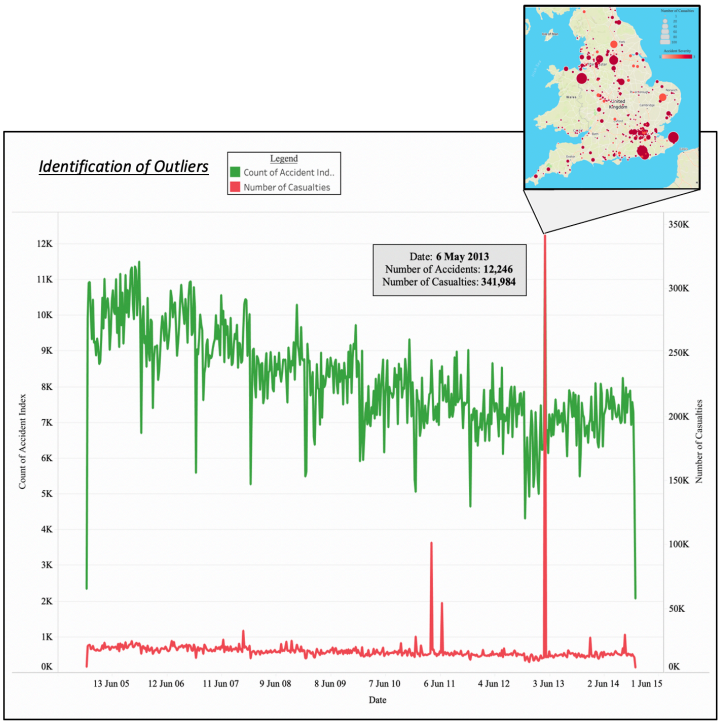

Fig.2. Identification of Outliers: Temporal visualisation of the number of accidents (Green) and casualties (Red)

4 Analysis

4.1 Approach

For the purpose of this paper, Tableau and Python will be used to plot visualisations and predictions. Tableau will allow us to enable human input as well as facilitate the analysis of the visualisations' outputs. Initially, the user will derive information about the data properties using computational and inform about the data attributes, errors and uncertainties. After creating the visualisations, the user will interact with the output and derive new knowledge which we can use to improve the visualisations (*Figure 3*). This is done by correcting the data and managing uncertainties in the initial models, then using the information collected about the pattern types and methods to improve the models and visual analytics. Furthermore, human interaction is required to help evaluate the model and represent the evaluation results.(Collins et al.,2018)

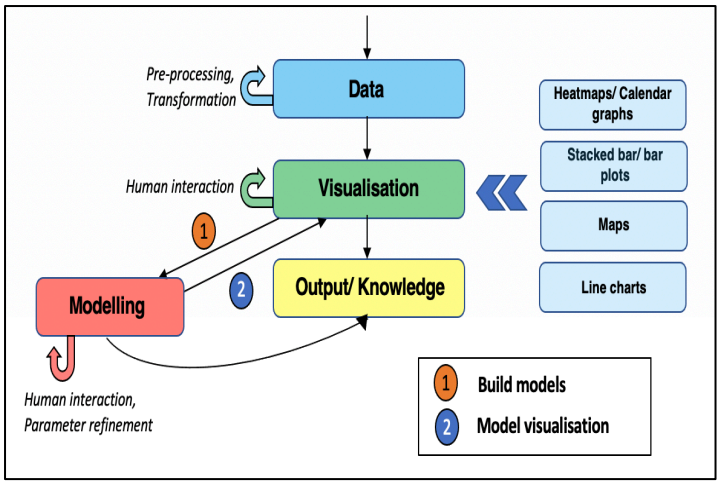

Fig.3. Diagram representing the visual analytics process inspired by Keim et al.(2008) visual analytics process diagram

4.1.1 Temporal Analysis

1) First, we transformed the temporal features, and we derived the months, years, days, and hours of the accidents using datetime and timestamp. Also, the values used to display the frequency of accidents are aggregated.
2) Before looking at the frequency of accidents, we created a line graph to show the number of accidents and casualties (*Figure 2*). This allows us to identify if any unusual statistics lie within our dataset.
3) Plot the frequency of the number of accidents using two plots: the first one using hours and weekdays and the second one looking at the frequency in the weekdays, every month between 2005 and 2014.

4.1.2 Spatial Analysis

1) As mentioned previously, there are 138 values missing for the *Longitude* and *Latitude* columns which were dropped.
2) Using the *Longitude* and *Latitude* data, we plot the accident occurrences for each year as well as their severity on a map. The number of casualities will also be represented on the maps as circles with different sizes.

4.1.3 Correlation

We looked at the correlation between features of the data. This allows us to determine whether there is any relationship between these features and the severity of accidents.

4.1.4 Prediction

Lastly, we looked at predicting the severity of the accidents using two machine learning models: Gaussian Naïve Bayes and Logistic Regression.

1) We looked at the correlation matrix. Then we use PCA to avoid multicollinearity.
2) Due to large size of the dataset, we decided to use a small portion of the data(2000 variables).
3) Implement SMOTE to oversample the minority classes in the imbalanced dataset.
4) We split the data into train and test sets, 75% and 25% respectively.
5) We then created a model with initial parameters for both methods.
6) After visualising the results, we attempted to improve the models by hyper tuning the parameters. For this we firstly tried and manually changed the parameters. Then we use Grid Search and found that this tool provides us with the best accuracies.
7) Finally, we computed the accuracy, recall, f1-score and precision of the models, and ROC curve.

4.2. Process

4.2.1 Temporal Analysis

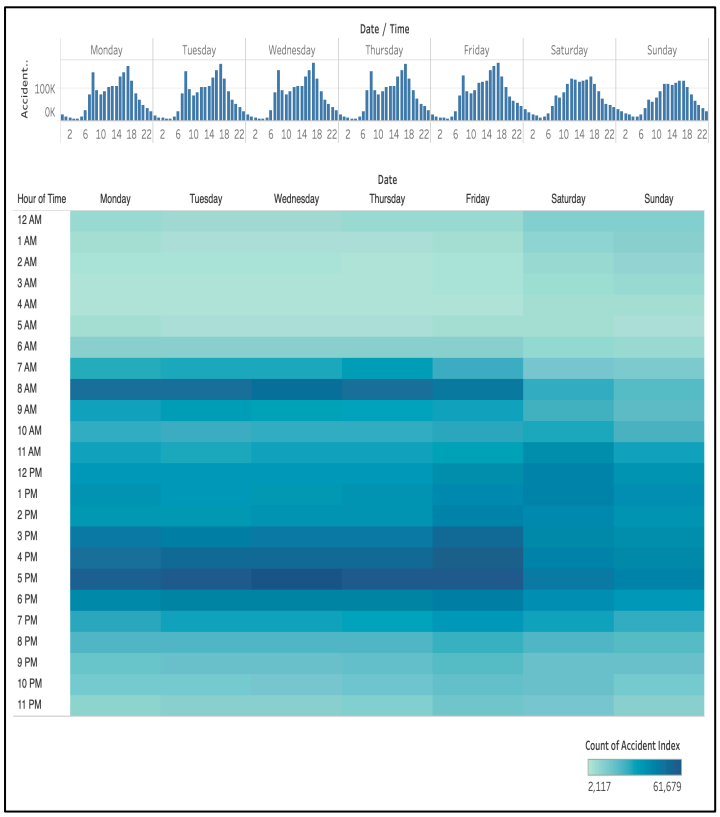

Fig.4. Temporal visualisation: Frequency of the number of accidents during the hours of weekdays

Figure 4 shows that the highest number of car accidents is focused around 7-8 AM and between 3 and 5 PM. Moreover, it appears that the frequency of accidents is high between Monday and Friday. This is reasonable as it is consistent with workdays and rush-hours. As mentioned by Cabrera-Arnaud et al. (2020), this might be due to factors such as higher levels of stress, drivers being exhausted before and after a day at work, reduced visibility in the evening, a very congested traffic flow or a combination of these. On Fridays we can see that the frequency of accidents is higher than on the other days, between 3 and 5 PM. This can be explained by the fact that people might leave work early before the weekend.

The frequency of car accidents during those hours in the weekends is reduced as people are not likely to travel to work. However, the number of accidents, between 11 AM and 5 PM are still slightly high as people might be going out shopping, seeing their friends and family and so on during those hours.

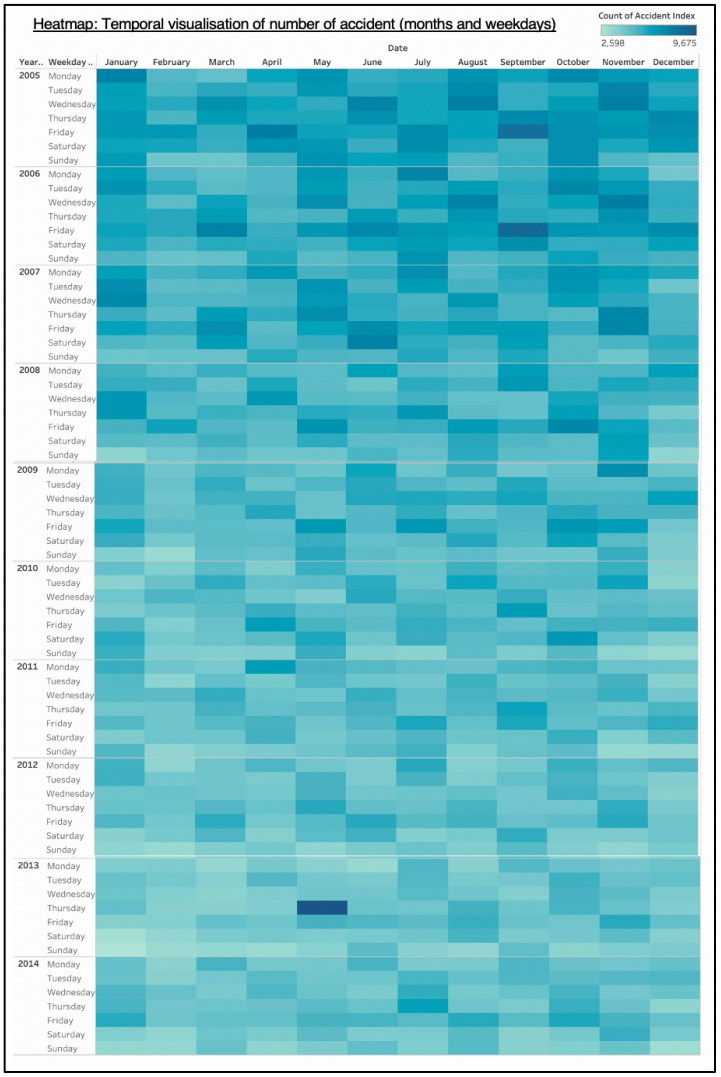

Fig.5. Temporal visualisation: Frequency of the number of accidents during weekdays throughout the months for each year

In *Figure 5*, we can see that throughout the years the number of accidents has been decreasing. Nevertheless, for the years 2005 and 2006, January has shown a constant frequency throughout the days of the week. February appears to be the month with the lowest frequencies overall. Comparably to the unusual value in *Figure 2*, we can see in *Figure 5* that in May 2013 the frequency of accidents was at its highest.

When looking at the other frequencies, there does not seem to be any particular pattern. However, there seems to be a resonance between *Figure 4* and *5* as weekends have a smaller frequency of accidents than the one of the weekdays.

Henceforward, we will perform a spatial analysis to visualise the frequency of accidents over time, as well as their severity, using dotted maps.

4.2.2 Spatial Analysis

Fig.6. Spatial representation of the severity of accidents through the years: (A) 2005, (B) 2006, (C) 2007, (D) 2008, (E) 2009, (F) 2010, (G) 2011, (H) 2012, (I) 2013, (J) 2014

From *Figure 6* we can see the spatial representation of accident severity and casualties in the UK from 2005 to 2014. The most severe accidents appear to be focused on some of the most visited cities in the UK London, Birmingham, Nottingham, Manchester and Leeds, with London, Birmingham and Manchester being the largest hotspots. This is not surprising as London is ranked 6th in the world's most congested city (London Councils, 2019).

The spatiotemporal analysis helped determine whether similar hazardous locations are subject to temporal fluctuations in road traffic accidents in the UK. As seen in *Figure 4* and *5*, the frequency of accidents varies over time (between weekdays and weekends, between different hours of the days and throughout the years). It is evident that road accident frequency fluctuates heavily throughout the hours of the days, with the highest frequencies being focused during rush-hours. On the other hand, there does not seem to be any patterns amongst seasons throughout the years. As shown in *Figure* 6, accidents tend to be more clustered in big cities and high areas of congestion. Furthermore, the accidents appear to be relatively uniformly distributed throughout the years.

The results revealed that the frequency of road accidents fluctuates through both time and space. The level of variations seems to depend on various factors, for example accidents are more likely to happen during rush-hours and are more frequent in highly congested traffic areas. (Harirforoush et al., 2019)

Looking past spatial and temporal analysis in the instance of the number of accidents, it is important to be able to identify their severity to support and improve road safety. If it is possible to predict the severity of accidents accurately, the management of traffic accidents could be crucially improved. This could help mitigate the impacts of the accidents and improve the effectiveness of the transportation system. Henceforth, the next section will aim to predict the severity of accidents using two machine learning methods, Logistic Regression and Naïve Bayes.

4.2.3 Correlation Analysis

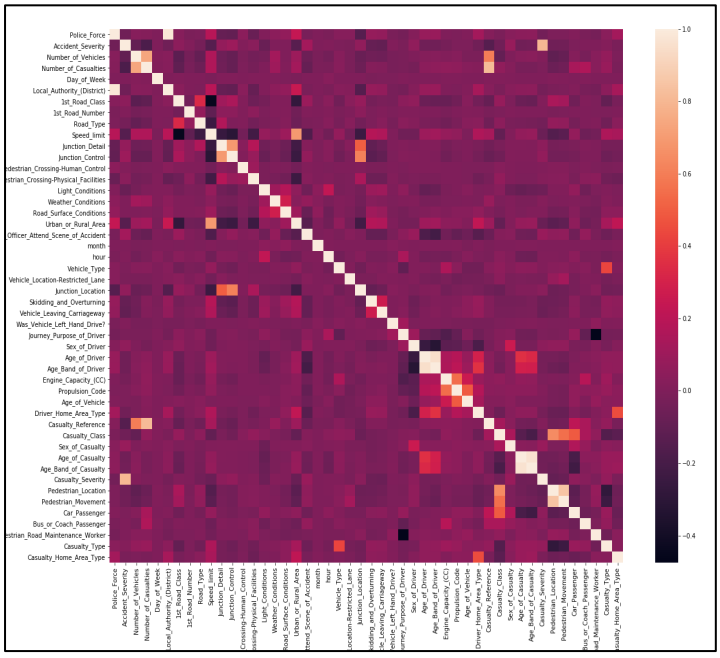

Fig.7. Pearson Correlation Matrix

Figure 7 displays the correlation between the factors of the dataset. As the main focus of this paper is on accident severity, we will look at the correlation of other factors with this variable. Most features seem to have a low or even negative correlation to the target variable, although, the severity of casualties is highly correlation to the accident severity. Surprisingly, the road surface, weather, light, speed limit, road type conditions have a very low correlation with the severity of the accidents. Hence, in order to avoid multicollinearity issues when making our model predictions, we will be using Principle Component Analysis as a dimension reduction algorithm.

On the other hand, the number of casualties and their type: pedestrians, cyclists, car occupants, motorcycle riders or passengers, minibus occupants, mobility scooter riders, goods vehicles occupants (over 3.5 tonnes or over 7.5 tonnes), have a significant correlation with the number of vehicles and their type. This could be explained by the fact that bigger vehicles may induce a greater amount of damage in accidents. For instance, a cyclist's accident with a pedestrian might cause a light to no injuries while if a goods vehicle over 7.5 tonnes collides with a car containing several occupants, this could result in a higher number of injuries or even fatalities.

4.2.4 Accident Prediction

To be able to identify and predict the severity of an accident, supervised machine learning methods were used. Since the dataset is highly unbalanced (as can be seen in *Figure* 1) we used the Synthetic Minority Oversampling Technique to increase the size of the minority classes by creating synthetic observations based on the original data.

By examining the severity of an accident, we will be able to identify whether or not an accident is likely to happen. This will allow the department of transportation to put forward a prevention plan in order to reduce the number of accidents and fatalities.

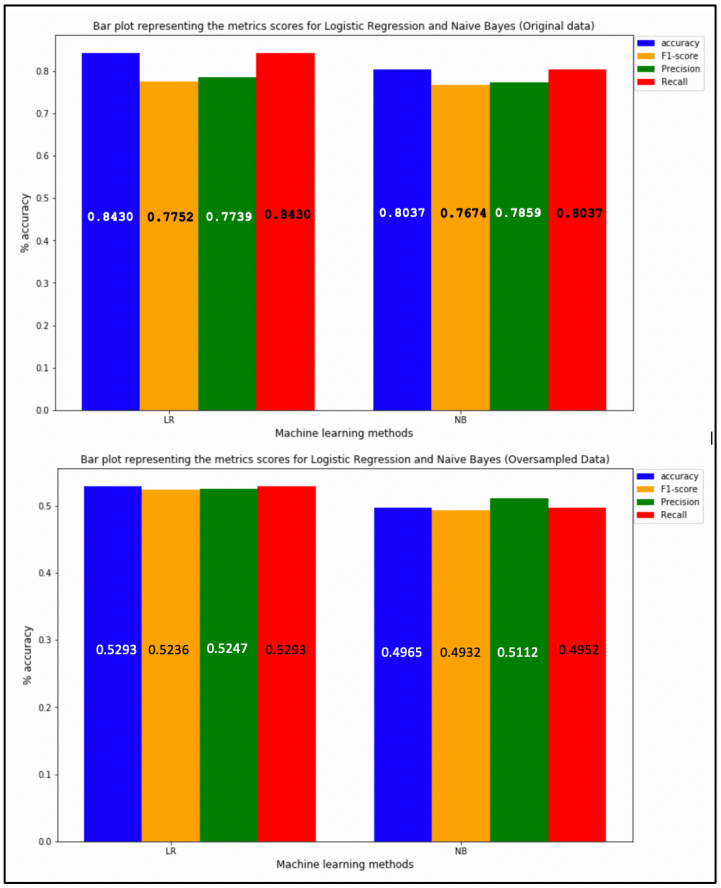

Fig.8. Prediction Metrics for Logistic Regression and Gaussian Naïve Bayes algorithms. (1) Using Imbalanced data (2) Using Oversampled data (SMOTE)

Figure 8 shows the metric values for the Logistic Regression and Naïve Bayes methods for both the originally imbalanced data and the newly oversampled data. In the case of the imbalanced data, Logistic Regression appears to perform

slightly better than Naïve Bayes with an F1-score of 0.7752 compared to 0.7674 and higher Precision and Recall values of 0.7739 and 0.8430 against 0.7859 and 0.8037. As both F1-scores are high, we can say that our models performed well given the imbalance of the data.

When looking at the oversampled modelled data results, the metrics are significantly lower. This is due to the fact that when oversampling the smaller classes, the likelihood of predicting from the previously largest class is decreased, leading to a fall in accuracy (from 0.84 to 0.53 for Logistic Regression and from 0.80 to 0.50 for Naïve Bayes).

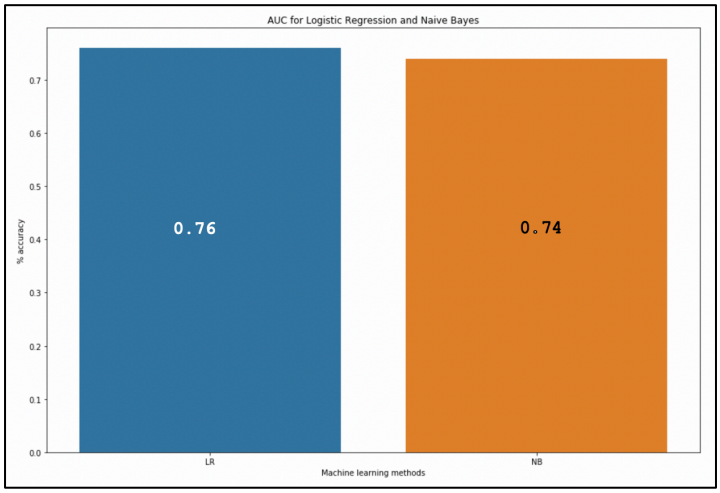

Fig.9. AUC for Logistic Regression and Gaussian Naïve Bayes algorithms

Figure 9 displays the AUC metric value for both models. Logistic Regression appears to have a higher AUC than the one of Naïve Bayes which means that the proportion of positive data points which were predicted correctly is higher than the number of false positives. The greater the area, the better the performance, therefore according to the area under the curve the Logistic Regression model outperforms Naïve Bayes. This confirms the results provided by the previously mentioned metrics which suggested that the Logistic Regression algorithm has a better performance than Gaussian Naïve Bayes.

4.3. Results

From the analysis of the temporal data, we can conclude that the physical and emotional state of the drivers can be a strong factor in the cause of accidents. This is shown in Figure 4 as the severity of the accidents increases during rush hours. Overall, the number of road accidents in the UK has been decreasing over the years (*Figure 10)*.

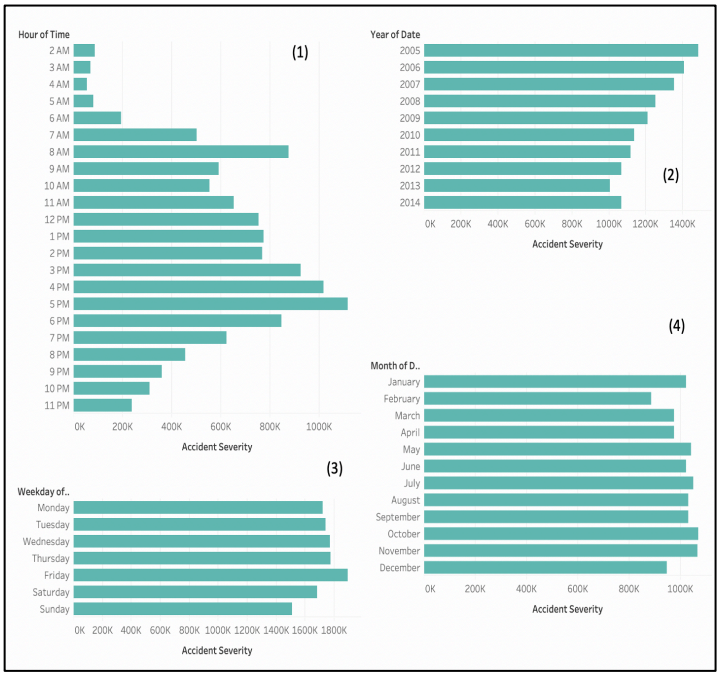

Fig.10. Temporal identification summary (1)Hour (2)Year (3)Weekday (4)Month

For the spatial analysis, we are unable to summarise the results of the investigation as the data is spread out all over the UK. The easiest visualisation is *Figure 6* from which we can distinguish hotspots, i.e., areas of high congestion, where severe accidents happen. These usually appear in big or touristic cities in the UK.

Concerning the prediction, as expected, Logistic Regression performed better than Naïve Bayes with accuracies of 0.8430 and 0.8037, respectively using an imbalanced data and 0.53 and 0.50 using an oversampled data. Moreover, according to *Figure 9*, Logistic Regression has a greater area under the curve with a value of 0.76 while Naïve Bayes has a value of 0.74. As a result, Logistic Regression should be preferred to Naïve Bayes for traffic accident severity prediction.

5 Critical Reflection

Heatmaps are great way to show the frequency of accidents on a temporal dimension. In terms of temporal analysis, this paper highlights that the severity of the accident increases within

morning and evening rush-hours. This can be due to the inattention of individuals while driving when wanting to arrive home or leave work quickly. With regards to the speed limit, the study has not shown a correlation with the accident severity. However, a deeper analysis could be done by investigating the number of accidents where the speed limit was exceeded and compare it to each severity class separately. Despite the lack of correlation, traffic regulations should be refreshed and reinforced to the individuals involved in road accidents. Additionally, increasing the number of police officers around high congestion areas such as the centre of London, Manchester and Birmingham could help fortify the sense of road safety in such areas. This can allow a future drop in the number of accidents.

When visualising the spatial information, a density-based clustering be used to display the concentration of several events considering a filter (Adrienko et al., 2017). For instance, instead of mapping the number of accidents by severity using points, a density-map could potentially accentuate hotspots better.

In terms of the methodology, it would be interesting to implement different approaches and learning machine methods such as Random Forest and comparing it to a regression model such as Poisson's Regression. Moreover, the models could focus on different features of the data such as indicators of cyclists' and pedestrian's accidents and not only investigate the severity of accidents. Furthermore, another appealing aspect which could be investigated is the aggregation of different types of casualties using clustering algorithms like K-means, Density-Based Spatial Clustering of Applications with Noise or Expectation-Maximization Clustering using Gaussian Mixture Models. Another recommendation for future work could be to analyse this data from the perspective of urban and non-urban areas while considering different types of roadways.

For the purpose of this paper, Naïve Bayes and Logistic Regression were used due to the ease of their implementation, their efficiency to train models as well as their rapidity while keeping a good accuracy (Tsangaratos and Ilia, 2016). Due to the large size of the dataset, Random Forest could not be implemented as the trees are highly memory-costly (Kiran and Serra, 2017).

The Synthetic Minority Oversampling Technique was used as a balancing procedure in this study to solve the issue of imbalanced dataset. Alternative methods could have been used such a under sampling the larger class and using K-Fold cross validation.

Table of word counts

Problem statement	200
State of the art	500
Properties of the data	480
Analysis: Approach	500
Analysis: Process	1304
Analysis: Results	200
Critical reflection	424

REFERENCES

[1] Adrienko G., Andrienko N., Chen W., Maciejewski R., Zhao Y. (2017). *'Visual Analytics of Mobility and Transportation: State of the Art and Further Research Directions'.* [Online]. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7891950> [Accessed on 29th December 2020]

[2] Almamlook, R., Kwayu K., Alkasasbeh M., Frefer A. A. (2019). *'Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity'.* [Online]. Available at: <https://www.researchgate.net/publication/333229225_Comparison_of_Machine_Learning_Algorithms_for_Predicting_Traffic_Accident_Severity> [Accessed on 5th January 2021]

[3] Bhawkar, A. (2018). *'Severe Traffic Accidents in United Kingdom'.* [Online]. Available at: <https://www.researchgate.net/publication/330676135_Severe_Traffic_Accidents_in_United_Kingdom> [Accessed on 2nd January 2021]

[4] Cabrera-Arnau C., Curiel R. P., Bishop S. R. (2020). *'Uncovering the behaviour of road accidents in urban areas'*. [Online]. Available at: <https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.191739> [Accessed on 29th December 2020]

[5] Chong, M. M., Abraham, A., Paprzycki, M. (2005). *'Traffic Accident Analysis Using Machine Learning Paradigms.'*. [Online]. Available at: <https://www.researchgate.net/publication/220166391_Traffic_Accident_Analysis_Using_Machine_Learning_Paradigms> [Accessed on 5th January 2021]

[6] Collins C., Andrienko N., Schreck T., Yang J., Choo J., Engelke U., Jena A., and Dwyer T. (2018). *'Guidance in the human-machine analytics process'*. [Online]. Available at: <22D19A11C1D0CA62615C6AB00772D3E3A1712C6BF86CA7581E160BE8AFC700BF7B50D95872FEB6924BFA89493FBCE0F8> [Accessed on 5th January 2021]

[7] Data.gov.uk (2021). '*Road Safety* Data'. [Online]. Available at: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> [Accessed on 29th December 2020]

[8] Department for Transport (2019). '*Road Casualties Great Britain, Main Results 2019'*. Department for Transport Statistics Bulletin. [Online]. Available at :< http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/downloadable/dft_transstats_038554.pdf> [Accessed on 2nd January 2021]

[9] Fang A., Peng X., Zhou J., Tang L. (2018). *'Research on the Map-matching nd Spatial-temporal Visualization of Expressway Traffic Accident Information'.* [Online]. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8492572> [Accessed on 5th January 2021]

[10] Feng M., Zheng J., Ren J., Liu Y. (2020). *'Towards Big Data Analytics and Mining for UK traffic Accident Analysis, Visualization& Prediction'.* [Online]. Available at: <https://dl.acm.org/doi/10.1145/3383972.3384034> [Accessed on 30th December 2020]

[11] Hébert, A., Guédon T., Glatard T., Jaumard B. (2019). *'High-Resolution Road Vehicle Collision Prediction for the city Montreal'.* [Online]. Available at: <https://arxiv.org/pdf/1905.08770.pdf> [Accessed on 5th January 2021]

[12] Harirforoush H., Bellalite L., Benie G. B. (2019). *'Spatial and Temporal Analysis of Seasonal Traffic Accidents'*. [Online]. Available at: < https://www.researchgate.net/publication/333601530_Spatial_and_Temporal_Analysis_of_Seasonal_Traffic_Accidents > [Accessed on 5th January 2021]

[13] Kaggle (2017). *'Uk Car Accidents 2005-2015'*. [Online]. Available at: <https://www.kaggle.com/silicon99/dft-accident-data> [Accessed on 26th December 2020]

[14] Keim D. A., Andrienko G., Fekete J. D., Görg C., Kohlhammer J., Melançan G. (2008). *'Visual Analytics: Definition, Process, and Challenges'*. [Online]. Available at: <https://www.researchgate.net/publication/29637192_Visual_Analytics_Definition_Process_and_Challenges> [Accessed on 29th December 2020]

[15] Kiran B. R., Serra J. (2017). *'Cost-Complexity Pruning of Random Forests'.* [Online]. Available at: <https://www.researchgate.net/publication/315116126_Cost-Complexity_Pruning_of_Random_Forests> [Accessed on 1st January 2021]

[16] London Councils (2019). *'London ranked 6th most congested city in the world'.* [Online]. Available at: <https://www.londoncouncils.gov.uk/press-release/18-february-2019/london-ranked-6th-most-congested-city-world> [Accessed on 2nd of January 2021]

[17] Tsangaratos P. and Ilia I. K. (2016). *'Comparison of logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size'.* [Online]. Available at: <https://www.researchgate.net/publication/303880287_Comparison_of_a_logistic_regression_and_Naive_Bayes_classifier_in_landslide_susceptibility_assessments_The_influence_of_models_complexity_and_training_dataset_size> [Accessed on 9th January 2021]