Towards Big Data Analytics and Mining for UK Traffic Accident Analysis, Visualization & Prediction

Mingchen Feng[1], Jiangbin Zheng[1], Jinchang Ren[2] and Yanqin Liu[3]
[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China
[2]Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, G1 1XW, U.K
[3]Department of Sports, Xi'an Fanyi University, Xi'an, China
(86)18189143290, (44)141 5482384
mingchen@mail.nwpu.edu.cn; jinchang.ren@strath.ac.uk

ABSTRACT
Road traffic accident (RTA) is a big issue to our society due to it is among the main causes of traffic congestion, human death, health problems, environmental pollution, and economic losses. Facing these fatal and unexpected traffic accidents, understanding what happened and discover factors that relate to them and then make alarms in advance play critical roles for possibly effective traffic management and reduction of accidents. This paper presents our work to establish a novel big data analytics platform for UK traffic accident analysis using machine learning and deep learning techniques. Our system consists of three parts in which we first cluster accident incidents in an interactive Google map to highlight some hotspots and then narratively visualize accident attributes to uncover potentially related factors, finally we explored several state-of-the-art machine learning, deep learning and time series forecasting models to predict the number of road accidents in the future. The experimental results show that our big data processing platform can not only effectively handle large amount of data but also give new insights into what happened and reasonably prediction of what will happen in the future to assist decision making, which will undoubtedly show its great value as a generic platform for other big data analytics fields.

CCS Concepts
•Information systems→Data Mining • Mathematics of computing→Exploratory data analysis

Keywords
Traffic Accident Analysis; Big Data Analytics; Deep Learning; Time series Forecasting.

1. INTRODUCTION
Road traffic accident (RTA) is a major but neglected public health challenge that requires concerted efforts for effective and sustainable prevention. According to a survey, it is estimated that there are 1.2 million deaths and 50 million injuries in road crashes worldwide each year [1]. In UK, it is reported that from 2017 to 2018, there were 1770 road deaths and 26,610 people killed or seriously injured in traffic accident, resulting in totally 165,100 casualties [2]. So new technologies and ways need to devised to uncover accident-relate factors and to identify time and space that are risky, thus supporting traffic accident data analysis in decision-making processes.

In this paper, we proposed a novel big data analytics platform for UK traffic accident analysis using data mining and deep learning techniques. We first displayed the data on an interactive map, which will effectively highlight accident hotspot in accordance with time and space. Then we visualized some attributes in the data to find key factors that related to traffic accident. Finally, we utilized state-of-the-art time series and deep learning algorithms to forecast accident in the future.

2. RELATED WORK
Various studies have addressed the different aspects of RTA, while most of which focus on using machine learning [3] and data mining [4] techniques to analyze traffic accidents. Kumar et al. [5] used cluster methods to identify high-frequently accident areas and then applied association rules to uncover factors that have an effect on road accidents at that locations. Shanthi et al. [6] carried out gender based classification, which RndTree and C4.5 using AdaBoost Meta classifier were utilized to obtain road accident patterns. Chen et al. [7] proposed a procedure which evaluates clusters of traffic accident and organized them according to their significance, thus identified high-risk locations (hotspot). Kumar et al. [8] proposed a data mining framework to analyze road accident, k-modes clustering and association rules are used to discover potential patterns and trends. In Xi et al. [9], association rules are used to categorize accidental factors and analyzed the degree of an accident or the level of influence. Park et al. [10] utilized Hadoop framework to handle large data set and built a predictive model to settle data imbalance problem. Shahrokh et al. [11] assessed trends of traffic accidents and then explored time series models to forecast them in the next years. He et al. [12] applied BP neural network to analyze the relationship between accident evaluating index and forecast the number of accident in the future. Chee et al. [13] improved hybrid artificial neural network to visualize multidimensional data without increasing the number of neurons. Bakir et al.[14] provided a robust forecasting model to predict phone prices in European markets using Long Short-Term Memory (LSTM) neural network and Support Vector Regression (SVR).

Studies above for exploring road traffic related factors and patterns are essential and necessary for certain methods and application scenario, but each traffic accident dataset itself has their specific meaning and unique inner-relationship, which are hard to extract. Moreover huge volume of data highlight the

ICMLC 2020, February 15–17, 2020, Shenzhen, China

ACM ISBN 978-1-4503-7642-6/20/02…$15.00
DOI: https://doi.org/10.1145/3383972.3384034

requirements for high speed and accuracy which stresses the limitations of existing models and algorithms.

3. DATA ANALYSIS, VISUALIZATION & PREDICTION

3.1 Data

Our dataset comes from Department for Transport, UK [15], which amassed traffic data from 2005 to 2017, recording over 2 million accidents in the process.

Each data record has the following 12 features:

1) Index - Index of traffic accident;

2) Longitude - Longitude of the location of an accident scene;

3) Latitude - Latitude of the location of an accident scene;

4) Severity - The severity of the accident;

5) Number of Vehicles – The amount of vehicles involved in the accident;

6) Casualties - The number of person injured in the accident;

7) Date – Date of the accident;

8) Time - Timestamp of the accident;

9) Date of Week - Day of the week that accident occurred;

10) Road type - The type of the road where accident occurred;

11) Speed limit - The speed limit of the road where accident happened;

12) Weather - Weather condition at the time of the accident;

3.2 Narrative Visualization

Considering the geographic nature of traffic accident, an interactive map based on google map was used to display accident incidents, where RTAs are clustered according to their geographic coordinate information. As shown in Figure 1. the round label with numbers are accident hot-spots and the associated number of traffic accidents. Our platform allows users to select time period to show incidents, it can also help us manage multiple markers at different zooming levels, corresponding to various spatial scales or resolutions. We can found that England has more traffic problems than Scotland, especially in London, Manchester, Birmingham, and Leeds, which are top 4 cities with the most populations in UK. Besides, we also found that most of fatal accidents happened locally within cities instead on highways as the red dots in the map clustered in the inner city. It could be the reason that the traffic is more congested locally than on highways.

Figure. 2 summarized the yearly traffic accident, The graph shows accidents count on daily basis. We can quickly identify interesting trends in this time-series data that from 2005 to 2017, the amount of accident incidents seems decrease yet following certain patterns each year. It seems that accidents is fluctuating greatly throughout the year where it is high at the beginning of the year, and then decline for several months then increase till mid of the year, after that it will decrease for one month and then arising until reach its peak at the end of the year. February has achieved the least accidents throughout the years. On the other hand, November accounts for the most accidents, followed by a decrease going into the holidays and New Years! Summer months reported average accidents with slight ups and downs.

The hourly trends unravel some interesting accident facts as shown in Figure 3., the graph shows the number of crashes by day of the week and by hour of the day. As seen, there are more incidents on weekdays than on weekends, this may due to that on work days people are going out for work resulting more drivers on the road, which increase the risk for traffic accidents. There are also several peaks emerge, but the strongest is associated with the workweek at 8 and between hours 15 and 17, which are rush hours in a day when it has the most traffic moving such as people leaving for/from work. The rate of accident is observed to increase from Monday to Friday and decline over the weekends. The number of reported accident seems to be the maximum on Fridays. Interestingly, It also shows that on weekends there are more accidents between 0 a.m. and 5 a.m., when people tends to go out for parties and back home.

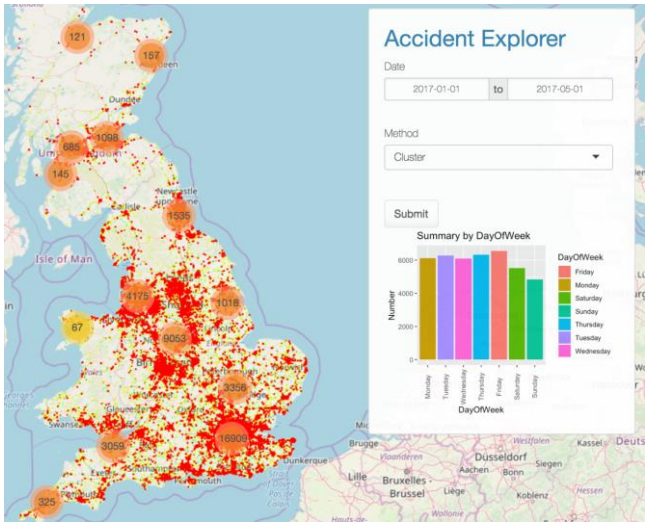

Figure 1. UK RTAs Hotspots displayed on interactive map

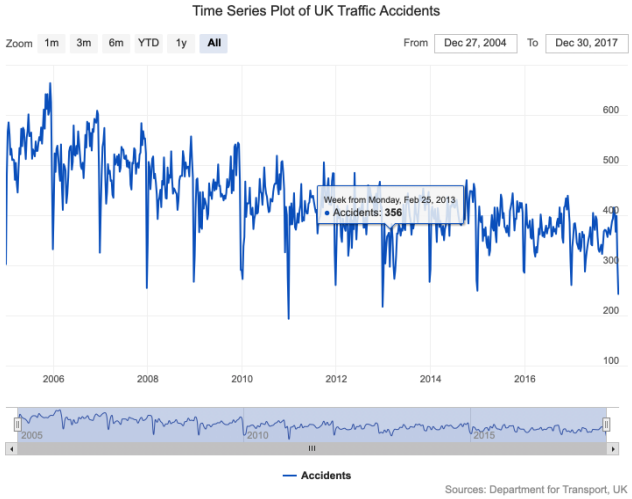

Figure 2. Time series plot of RTAs from 2005 to 2017

The dataset recorded 6 road types where accident happens. Figure 4. demonstrated the percentage of each road type and the severity for accident using sunburst chart, as seen single carriageway is 5 times more dangerous than dual carriageway as 74.6% of the accident occurred on it. Some of the accidents on single

carriageway could be cause of stop sign, changing lanes or turning into parking lot etc. And fortunately, most of the accidents are slight.

In Figure 5, we explored the effect of speed limit and weather on traffic accidents. The result of this analysis surprises us because it was the exact opposite of what we had expected. From our intuition, we would have expected that extreme weather conditions, such as snowing days with high winds or raining days with high winds, would have led to the most accidents to occur. But on the heatmap, it is shown that a normal/fine day and a rainy day without any high winds are the two kinds of weather to have led to the most number of traffic accidents. A possible reason for this could be that rather than causing more accidents, the poor weather reduces them, by discouraging driving, leading to fewer drivers, less congested roads and lower speeds. The heatmap also showed that when speed limit of a road is 30 miles/h it will be very dangerous to drive outside, however this is reasonable since most accidents are categorized as slight, so it is more likely to happen in roads with low speed limit. In addition when speed limit is 60 miles/h, it is also prone to cause accidents.

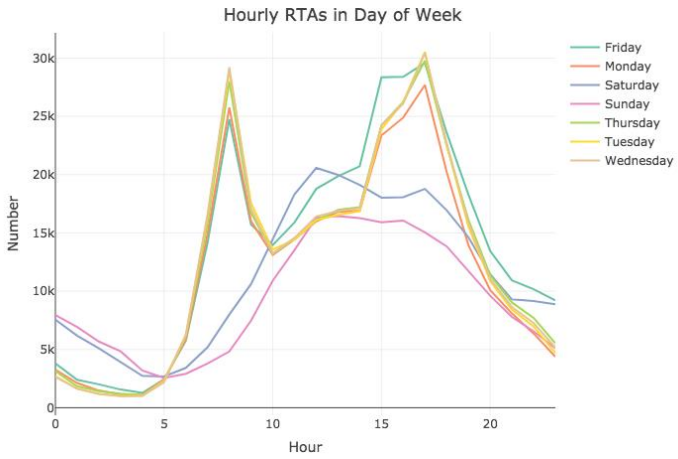

Figure 3. Hourly RTAs trends in weekdays and weekends.

Figure 4. Sunburst chart for accident-prone road type and the severity

3.3 Accident Trends Forecasting

In order to tackle the problem of accident trends forecasting we explored several state-of-the-art deep learning algorithms and time series models. We first counted the daily number of traffic accident and then transferred them to time series format. For performance evaluation of each model, the Root Mean Square Error (RMSE) and spearman correlation are used in terms of different parameters and different sizes of training samples. Finally we divided the data into two parts: 2005-2016 as training set and 2017 as testing set. Using keras stateful LSTM model and Facebook Prophet model, as they have achieved great success in crime prediction [16]. The experimental results showed that the optimal years of training sample is 3 years in terms of the minimum RMSE and the highest correlation as shown in Table 1 & Table 2, and Prophet model performs better than LSTM on this task. While the best echos for LSTM is 500, and the number of optimal layers of neural network is 50.

After gaining the optimal parameters and training size, we predict daily traffic accident in 2018 as shown in Figure 6. and Figure 7.

Table 1. Comparison of Prophet model in terms of RMSE and spearman correlation under different sizes of training samples

Years for training	RMSE-Prophet	Correlation-Prophet
10	48.21	0.735
9	48.70	0.741
8	46.18	0.717
5	45.70	0.711
4	46.18	0.728
3	**42.18**	**0.798**
2	91.93	0.702
1	100.56	0.628

Table 2. Comparison of LSTM model in terms of RMSE and spearman correlation under different sizes of training samples

Years for training	RMSE-LSTM	Correlation-Prophet
10	46.96	0.655
9	45.34	0.678
8	43.89	0.697
5	43.39	0.738
4	43.65	**0.785**
3	**43.37**	0.781
2	90.15	0.702
1	109.75	0.456

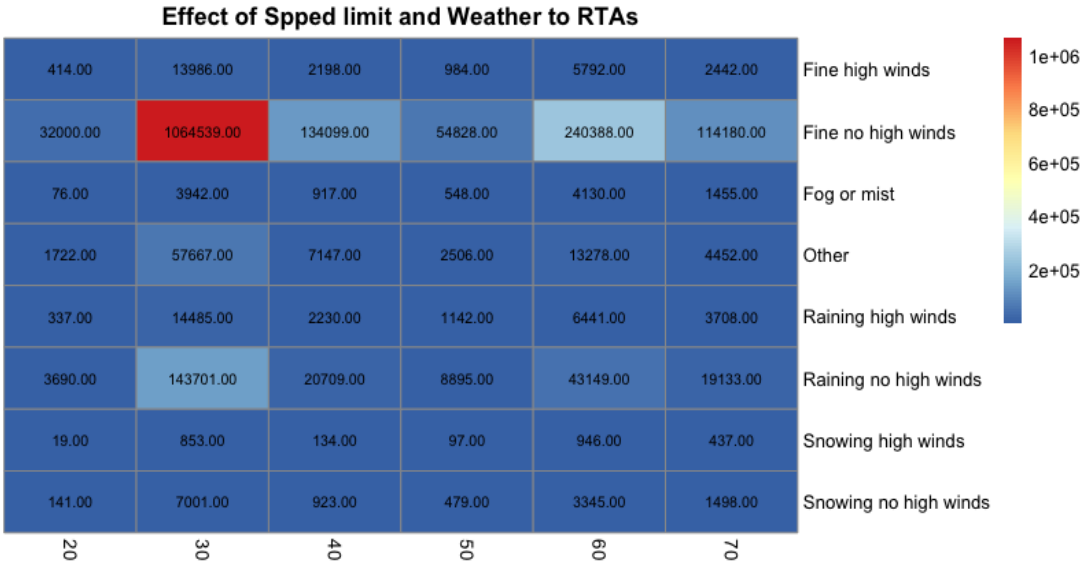

Figure 5. Heatmap for effect of speed limit and weather on accident

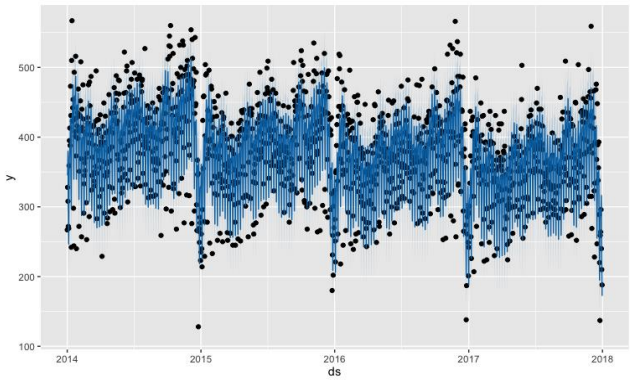

Figure 6. Prophet model for accident forecasting in 2017

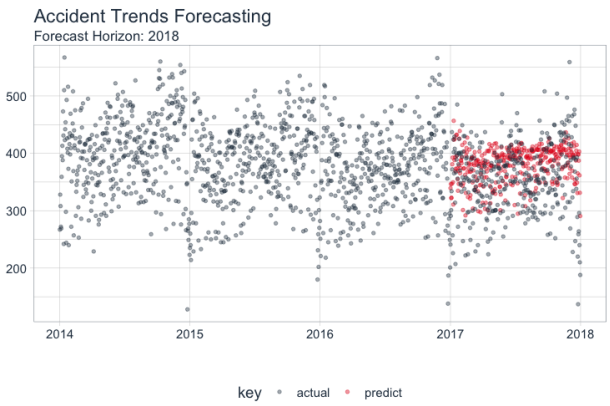

Figure 7. LSTM model for accident forecasting in 2017

4. CONCLUSION & FUTURE WORK

In this paper, we proposed a novel big data analytics platform for analyzing UK traffic accident data. By displaying the data on interactive map and using visualization techniques, some meaningful trends and patterns were discovered. Besides, we also explored novel time series and deep learning algorithms to forecast trends in the future. We found the optimal time period for training sample is 3 years so that to obtain the best forecasting performance. In the future, as other deep learning algorithms has achieved great success, we plan to apply more deep learning methods and add more feature selection and classification task on our data, and perform graph mining and association rules to uncover more interesting and meaningful patterns.

5. ACKNOWLEDGMENTS

This work has been supported by HGJ, HJSW and Research & Development plan of Shaanxi Province (Program No. 2017ZDXM-GY-094， 2015KTZDGY04-01) and by The 13th Five-Year Plan of Education Science in Shaanxi Province (Program No. SGH18H457) and Key Project in Xi'an Fanyi University (Program No. J18A07).

6. REFERENCES

[1] Pawłowski W, Lasota D, Goniewicz M, et al. The effect of ethyl alcohol upon pedestrian trauma sustained in traffic crashes. International journal of environmental research and public health, (April 2019), 16(8): 1471. DOI = https://doi.org/10.3390/ijerph16081471.

[2] Reported road casualties in Great Britain, provisional estimates: year ending June 2018: https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-provisional-estimates-year-ending-june-2018.

[3] Nandurge P. A. Nandurge and Dharwadkar N. V., Analyzing road accident data using machine learning paradigms, In 2017 Int. Conf. on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, 2017, 604-610. DOI= https://doi.org/10.1109/I-SMAC.2017.8058251.

[4] Li L., Shrestha S. and Hu G., Analysis of road traffic fatal accidents using data mining techniques,In 15th Int. Conf. on Software Engineering Research, Management and Applications (SERA), London, 363-370. DOI= https://doi.org/10.1109/SERA.2017.7965753.

[5] Kumar S, Toshniwal D. A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24, 1, (2016), 62-72. DOI= https://doi.org/10.1007/s40534-016-0095-5.

[6] Shanthi S. and Geetha Ramani R., Gender specific classification of road accident patterns through data mining techniques, In IEEE Int. Conf. On Advances In Engineering, Science And Management (ICAESM -2012), Nagapattinam, Tamil Nadu, 2012, 359-365.

[7] Chen X, Huang L, Dai D, et al. Hotspots of road traffic crashes in a redeveloping area of Shanghai. In International journal of injury control and safety promotion, 25, 3, 2018, 293-302. DOI= https://doi.org/10.1080/17457300.2018.1431938.

[8] Kumar S , Toshniwal D . A data mining framework to analyze road accident data. Journal of Big Data, 2, 1, 2015, 1-18. DOI= https://doi.org/10.1186/s40537-015-0035-y.

[9] Xi J., Zhao Z., et al., A Traffic Accident Causation Analysis Method Based on AHP-Apriori, In Procedia Engineering, 137, 2016, 680-687. DOI= https://doi.org/10.1016/j.proeng.2016.01.305.

[10] Park S , Kim S , Ha Y . Highway traffic accident prediction using VDS big data analysis. In Journal of Supercomputing, 72, 7,2016, 2832-2832. DOI= https://doi.org/10.1007/s11227-016-1655-5.

[11] Shahrokh Y C , Fatemeh R T , et al., A Time Series Model for Assessing the Trend and Forecasting the Road Traffic Accident Mortality. In Archives of Trauma Research, 5, 3, 2016, .DOI= https://doi.org/10.5812/atr.36570

[12] He M. and Guo X., The Application of BP Neural Network Principal Component Analysis in the Forecasting the Road Traffic Accident. In Proceedings of the 2009 Second International Conference on Intelligent Computation Technology and Automation, 1, (ICICTA '09), IEEE Computer Society, Washington, DC, USA, 107-111. DOI= https://dx.doi.org/10.1109/ICICTA.2009.35.

[13] Teh C S, Yii M L, Chen C J. Dimensional Reduction and Data Visualization Using Hybrid Artificial Neural Networks. International Journal of Machine Learning and Computing, 2015, 5(5): 420. DOI= https://dx.doi.org/10.7763/IJMLC.2015.V5.545.

[14] Bakir H, Chniti G, Zaher H. E-Commerce price forecasting using LSTM neural networks. Int. J. Mach. Learn. Comput, 2018, 8: 169-174. DOI= https://dx.doi.org/10.18178/ijmlc.2018.8.2.682.

[15] UK Road Safety Dataset: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

[16] Feng, M., Zheng, J., et al.: Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access 7, 1, 2019, 106111 – 106123. DOI= https://doi.org/10.1109/ACCESS.2019.2930410.