Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2 : 2024
ISSN :**1906-9685**

ROAD ACCIDENT PREDICTION USING MACHINE LEARNING TECHNIQUES

C JYOTHI SREE [1], Dr GNV VIBHAV REDDY [2]

[1] PG SCHOLAR IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SREEDATTHA INSTITUTE OF ENGINEERING AND SCIENCE SHEERIGUDA, IBRAHIMPATNAM HYDERABAD, TELANGANA, INDIA.

jyothikjr15@gmail.com.

[2] ASSOCIATE PROFESSOR IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SREE DATTHA INSTITUTE OF ENGINEERING AND SCIENCE, SHERIGUDA IBRAHIMPATNAM HYDERABAD TELANGANA, INDIA.

ABSTRACT

Around 10.25 lakh people die and 5 crores people are injured in road accidents every year, making it one of the main causes of mortality globally. The frequency of road accidents is high. Everyone who lives in a large city like ours has probably heard of, seen, or been a part of at least one. Consequently, lives might be saved by a system that can forecast when road accidents or locations prone to accidents would occur. Although it is challenging, it is not impossible to foresee traffic accidents. Numerous variables, including drivers' health, vehicle characteristics, speed, traffic, road conditions, and weather, impact the likelihood of accidents; accidents do not occur randomly. Paper is a dynamic routine builder in deep learning for Python that aims to help users get a better night's sleep by providing them with information about tomorrow's traffic. We use a From more apparent variables like national holidays, the moon phase, and selective attention to less apparent ones like speed, traffic, accident numbers, road structure, and weather, there are many inputs to think about. The good news is that the public has access to a few of these accident data. Rich datasets of Road Traffic Accidents (RTAs) and their corresponding variables have been made accessible by several UK local and national governments. Our goal is to find patterns that may accurately anticipate when traffic accidents will happen by analysing these government information and additional data sources.

Keywords: Road Accidents, Municipal And National Government, Traffic.

INTRODUCTION

Approximately 1.25 million people die and 50 million are injured each year as a result of road traffic accidents (RTAs), making them a top cause of mortality worldwide. Efforts to reduce RTA have been made by transport authorities throughout the globe. But this is no easy feat; road traffic accidents (RTAs) have not diminished much despite the implementation of several safety laws. The difficulty in anticipating the location and timing of RTA is a contributing factor to this failure. Many variables are associated with RTAs, including but not limited to: weather, road structure, collision counts, traffic conditions, speed, and less evident influences like national holidays, the moon phase, and selective attention.

Several UK national and local administrations have made accessible extensive databases on RTAs and the variables that contribute to them.

We want to find trends that accurately forecast when and where RTA are likely to occur by studying these government databases and additional data sources. Our final objective is to develop a reliable RTA prediction model for the UK market with an intuitive online interface that includes:

The model itself, in a nutshell. Images that emphasise the significance of different elements in RTA prediction. The user may enter variables into an interactive dashboard and instantly retrieve the probability of RTA occurrence in different locations of the UK. The public at large in the United Kingdom stands to gain from this paper's findings because of the visualisation tool it will provide for conveying the likelihood of a traffic accident at a specific location. Also, the traffic authorities may use it to their advantage when coming up with plans to cut down on RTA. Our goal in writing this study was to provide a user-friendly, interactive traffic accident predictor that everyone could use. We settled on deploying a trained predictor on a website as the optimal means of accomplishing this objective. Here are some things that this predictor-website should be able to do: The optimum driving route connecting two points may be found by letting users provide an origin and a destination, provided that both are inside greater London. The user may choose the trip's start and end times, and within that time frame, the system will highlight spots along the route that are more likely to have accidents.

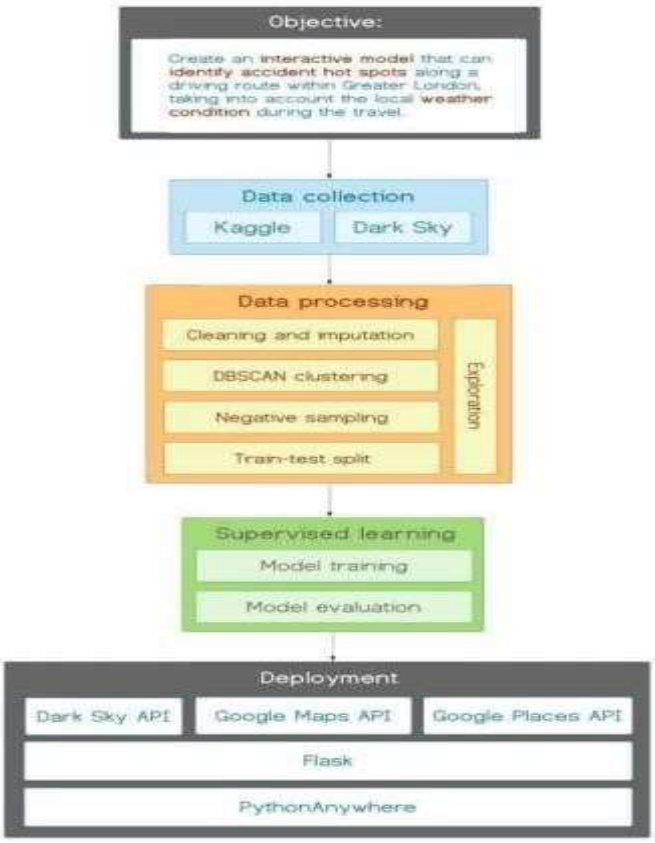

Figure 1: Steps in Process

LITERATURE REVIEW

Table 1 provides an overview of the state-of-the-art models as well as their limitations, which we shall attempt to improve upon. **Athanasios et al. [10]** proposed a model uses logistic regression to predict RTA. It considers RTA as rare events and subjects theresulting probability values to correction steps to account for the rarity of positive samples (accidents).

Milton et al. [13] proposed study examines random parameters approach – parameters can vary randomly across roadway segments to account for unobserved effects related to the environment, roadway characteristics, and driver behavior.

Alexandra et al. [11] proposed study examined the Empirical Bayesian model enhanced by Proportional Discordance Ratio(PDR) similarity technique.

Abdel-Aty et al Sun et al. [1] proposed study examined the Empirical Bayesian model enhanced by Proportional Discordance Ratio(PDR) similarity technique.

Wang et al. [] proposed study presents a two-staged model, Bayesian spatial model (for accident count data) and a mixed logit model (for severity level – slight, serious, fatal) to estimate accident frequency at different severity levels for London highways to identify accident h3otspots.

Prieto et al. [15] proposed study employed Rare Event Concentration Coefficient (RECC) to identify regions of high concentration of road accidents in London city and Mexico highways.

Table 1. Literature review of various models implemented to predict traffic accidents.

Authors	Limitations	Applications
Athanasios et al [10]	The model only takes into account trafficdata as its predictors and has limitedscope (highways in the city of Athens).	Logistics regression, together with the consideration of RTAas rare event, are among the algorithms we applied in our model development.
Milton et al [13]	The random parameter approach might introduce too much variance in the resulting model.	After experimenting, we found that the random parameter approach introduced too much variance in the resulting model.Thus, we decided to exclude it in our final model.
Alexandra et al. [6]	The accuracy of the model diminishes ifroad segments are not defined well enough by the state DOTs. For example,if the road segments are defined too short, there may be fewer accidents in each segment, thereby reducing the model's accuracy.	This study showed that the Empirical Bayesian (EB) method is preferred when conducting traffic safety analysis because itexcels in handling the regression to-the-mean bias.
Abdel-Aty et al Sun et al. [1,3]	The dataset used in the studies omitted important environmental variables such as weather and societal factors. The study also mentioned that the model suffer drop in accuracy after being deployed on other highways, hinting on a possibility of overfitting. Our project aims to use a broader range of predictor variables (such as weather data) to train a more- rounded model.	Our RTA prediction paper is a binary classification problem (e.g., Accident = 0 or 1). Therefore, Generalized Estimating Equation (GEE) [1] is not applicable, since it is more suitablefor Linear Regression. Support-Vector Machine (SVM) [3], while it is suitable for binary classification, takes extremely long time to train the model when the amount of data is huge(in our case, a few hundred thousand rows with more than 30features). We experimented with this model and found that ittakes too long (> 2 hours) to train. We decided to use other more efficient models instead.
Wang et al. [14]	The study has small dataset of 1k+ observations with limited features and only limited to highways.	RTA frequency and accident severity were modelled separately in this study. RTA data in the frequency model were aggregated at each road segment while individual accidents were used in the severity model. Both models examined the predictors of accident but this approach could not predict the probability of an accident occurrence given certain variables.

PROPOSED METHOD

In this model we have used five different methods such as Data Collection, Data Exploration, Feature Selection, Model Training for Prediction, Models to predict RTA.

Table 2

Data	Source	Size	Challenges
UK Accident recordsfrom year 2005 till 2014	Kaggle https://www.kaggle.com/daveianhickey/2000-16- traffic-flow-england-scotland-wales/version/8	1.6 million records 33 columns	Year 2008 data is missing Crucial information are coded in numerals e.g., Boroughs
Code to Text mapping for Accident records	UK Department for Transport https://data.gov.uk/dataset/cb7ae6f0-4be6-4935- 9277-47e5ce24a11f/road-safety-data	600 records	N.A.
Economic Policy Uncertainty Index (daily) from year 2012 to 2014	UK Daily Policy Data http://www.policyuncertainty.com/uk_monthly.htm	Approx. 1,000 records	N.A.

Data Collection: Data was collected from the sources shown in Table 2:

Data Exploration: A basic Exploratory Data Analysis (EDA) was performed on the datasets. The visualizations in Figure1 show that:

- Most accidents are of Severity 3 (Slight injury). Minority of accidents result in Severity 2 (Serious) and Severity 1 (Fatal)
- Most accidents happened on roads with relatively slow speed limit (30 miles/hour)
- There are lesser accidents on the first and last day of the week (Sunday and Saturday), which is also the weekends
- Surprisingly, most accidents happened on fine weather where the road conditions are dry

Feature selection: Figure 2 shows a correlation matrix among the numerical features of the dataset. For a set of highly correlated features, it is standard practice to exclude all but one feature to reduce impact of multicollinearity. However, certain models such as Random Forest, are relatively unaffected by multicollinearity. Therefore, the need to exclude correlated features will depend on the model used. The features and their importance scores for the Random Forest Classifier is shown in Figure 3.

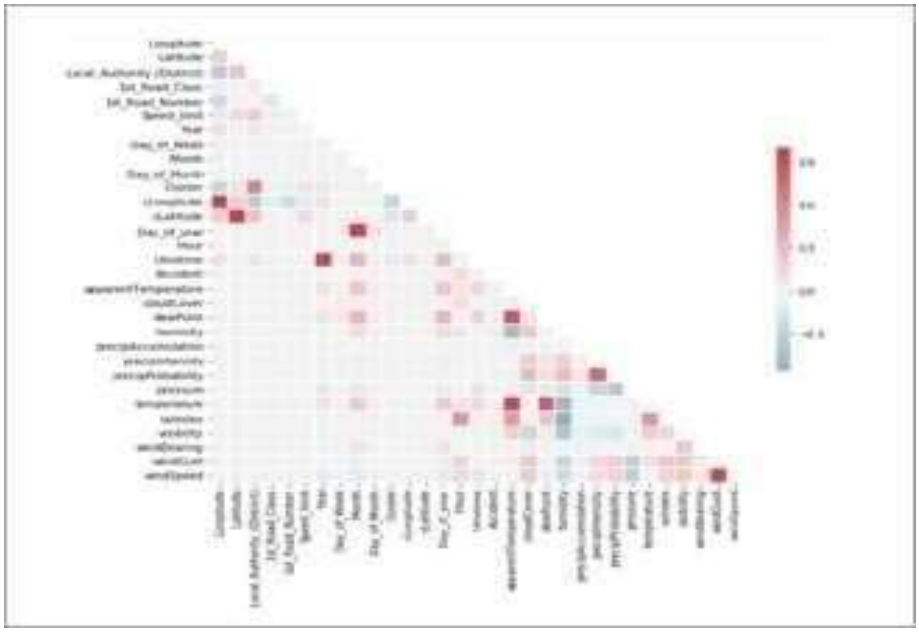

Figure 2: **Correlation matrix of numeric features in the UKRTA data set.**

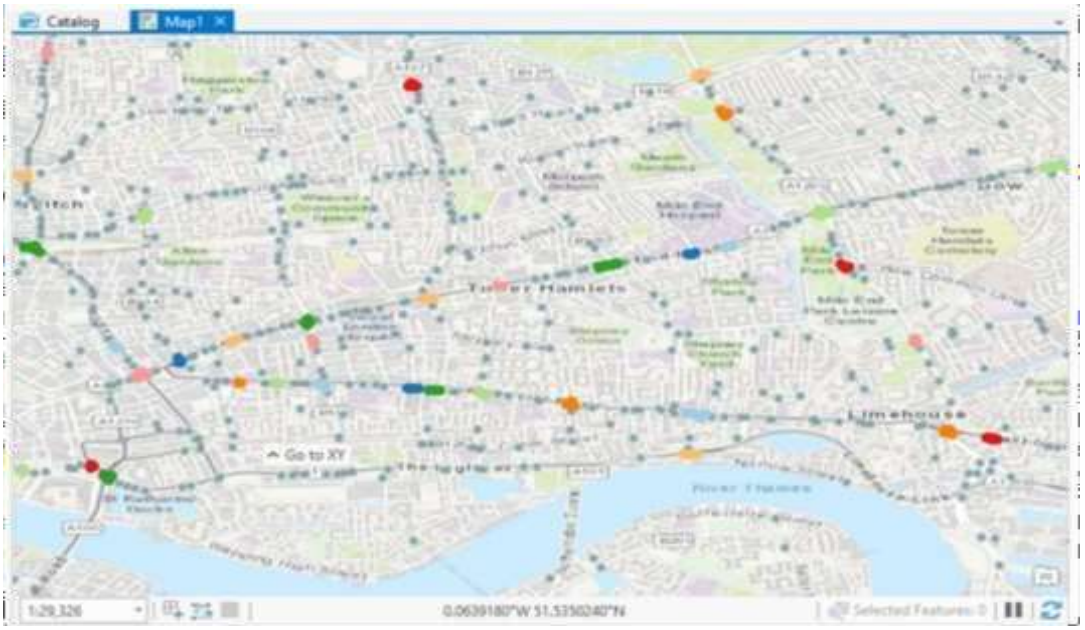

Figure 3: Clustered accident data points using ArcGIS. Colored points are hotspots with high density of accident o occurrence.

Model Training for Prediction: Clustering of Data Points Many of the accident data points are very close to one another. Some accidents occur frequently in a defined location, which we will label a "hotspot", i.e., signal. Other accidents occur in locations where accidents are likely to rarely occur and can be considered random events, i.e., noise. To define accidents as signal versus noise, we clustered all of the accident data points using ArcGIS with DBSCAN (Figure 3). This will improve the accuracy of traffic accident prediction.

Models to Predict RTAs: All model training was done in Python. Table 3 displays the models that were pursued and the measures that were used to evaluate their quality. Random Forest using just numerical/floating point predictors has shown to be the most effective model so far.

Table 3: Summary of algorithms used to build model and their performance based on accuracy.

Models	Predictors	Accuracy
Logistics Regression	25 Numerical / Floating point features	0.7611
Random Forest with only numerical predictors	25 Numerical / Floating point features	0.8347
Support Vector Machine (SVM)	21 Numerical features	Inconclusive since SVM ran for hours and has yet to complete.

IMPLEMENTATION

In this paper, the implementation is done in two steps:

1. Frontend Implementation
2. Backend Implementation

4.1 Backend Implementation: A web application has been built using the Flask framework. All the html pages, JavaScript libraries and CSS from front-end are integrated into the web application. Google Maps API and Google Places API are used for route planning and autocomplete function of places respectively. Weather forecasts is obtained by calling Dark sky API. In addition, a REST ful API module was built to handle users‟ requests of RTA predictions. Once a user enters the three inputs, i.e., date and time, traveling origin and destination, a POST request is sent to the backend framework. Google API is then called for route planning. Using the latitudes and longitudes on the route returned by Google, the backend calculates a radius of 50 meters from these points. We have a dataset of 9000+ past accident points that was a result from the DBSCAN clustering in previous section. Any past accident points in this dataset that do not fall within this 50m distance is filtered out. Next, for each unique cluster in the remaining accident points, Dark sky API is called. Each unique cluster will have the same weather forecast. This is a reasonable imputation as each cluster has a 25 meters radius and they should share the same weather. Instead of calling the weather API for many latitudes and longitudes, doing so allows our webpage to return the results to the users faster and reduces lag time. With the weather data, the final model is now loaded and predictions are made. For those duplicated latitudes and longitudes, i.e.,accidents had happened at the same spot multiple times, duplicates are removed. Frontend can now use the predictions to generate visualizations and highlight potential accident sites.

4.2 Frontend Implementation: The website will contain two main sections: “Exploration” and “Interaction”. The “Exploration” section includes a general background of the project as well as the most important takeaways of the EDA steps.The “Interaction” section contains an interactive map which will carry out RTA prediction. This visualization will allow users to input a specific particular date/time. Upon making this selection, the website will fetch weather information that correspond to the chosen date/time. These three inputs (date, time and weather) will be sent to our trained model, which in turn will predict probabilities on accident-prone spots. These spots will then be displayed on the map.

Figure 4: A screenshot of the website front page. The website will contain two main sections: “Exploration” and “Interaction”

After the user inputs their starting and ending points, the system will utilise Google Maps APIs to provide possible routes. Users may use the Google Place API to get recommended alternatives based on their starting and ending points. We will gather the GPS coordinates of several locations along the route and transmit them to our backend platform so that the algorithm can make predictions. On the route, you'll see danger symbols that represent the findings of the prediction model. These icons highlight the likelihood of accidents in the "hotspot" zones.

.

Figure 5: A screenshot of the Interaction page.

Figure 6: A screenshot of user-inputted origin and destination.

CONCLUSION

In this paper, we predicted probabilities of RTA for 32 boroughs of London for 48 hours in advance. We had successfully created an interactive web application that integrates Random Forest model (for prediction), Google API (for route suggestions). Doing a reality check on the output of the model, we found that it is reasonable in its prediction. For example, the model predicts that there will be a cluster of 15 accidents along the route from the Museum of London to Big Ben, both on a Friday 5pm and a Saturday 4am. Although the number of accidents predicted is the same on the two days, the probabilities are different. The cluster of predicted accidents yield 0.44±0.02 and 0.13±0.01 probabilities for Friday 5pm and Saturday 4am respectively. The difference in 26 predicted probabilities on different day and time coincides with our findings during data exploration stage. While we have largely met our project objectives, this paper has also exhibited a few limitations. The following tables show the limitations and how it can be improved in future studies:

Table 4

Limitations	Future Work
Prediction Accuracy (currently at 0.83) can be improved.	Introduce hyperparameter tuning for modelling. Experimentwith other models such as XGBoost and Neural network.
Latitude and Longitude of accident occurrence does not indicate direction of traffic. This may affect the probability of RTA prediction.	Include or find ways to extract this information for futurestudies.
Current model does not take into account traffic volume. If a particular road experience high number of accidents, it may beperceived as having a high accident probability. This may not be the case if the traffic volume is also high.	Incorporate traffic volume in future studies.
Current methodology in backend calculations, i.e., using the 50 meters radius for filtering prediction locations andeliminating multiple accident points in probability predictions were not extensively tested for yielding the best results.	Explore different combinations of parameters for optimization.Weighting method could be used on multiple accident points, e.g., assigning heavier weight to them. In this way, a single cluster will have different probabilities which is moreinformative.
Current model uses accident data from 2012-2014 which are more reflective of recent traffic laws, road conditions, speed limit change, population density, land usage etc.	Dataset could be enriched with more predictors such as population density, traffic volume, number of shops, number oftourist spots etc. More past data could be included in the model.

REFERENCES

[1] Mohamed Abdel-Aty, M. Fathy Abdalla (2004) "Linking Roadway Geometrics and Real-Time Traffic Characteristics to Model Daytime Freeway Crashes: Generalized Estimating Equations for Correlated Data", Transportation Research Record: Journal of the Transportation Research Board, Volume 1897, issue 1, pp. 106-115

[2] Azad Abdulhafedh (2017) "Road Crash Prediction Models: Different Statistical Modeling Approaches", Journal of Transportation Technologies", Volume 7, pp. 190-205

[3] Jian Sun, Jie Sun, and Peng Chen (2014) "Crash risk analysis for Shanghai Urban Expressways: Use of Support Vector Machine Models for Real-Time Prediction of Crash Risk on Urban Expressways", Transportation Research Record:

[4] Journal of the Transportation Research Board, Volume 2432, pp 91-98

[5] Fancello Gianfranco, Stefano Soddu & Paolo Fadda (2018) "An Accident Prediction Model for Urban Road Networks,"

[6] Journal of Transportation Safety & Security, Volume 10, issue 4, pp. 387-405

[7] Wen Cheng & Xudong Jia (2015) "Exploring an Alternative Method of Hazardous Location Identification: Using Accident Count and Accident Reduction Potential Jointly", Journal of Transportation Safety & Security, Volume 7, issue1, pp. 40-55

[8] Alexander S. Lee, Wei-Hua Lin, Gurdiljot Singh Gill & Wen Cheng (2018) "An enhanced empirical bayesian method for identifying road hotspots and predicting number of crashes, Journal of Transportation Safety & Security, pp.1-17,

[9] DOI: 10.1080/19439962.2018.1450314

[10] Maryam Dastoorpoor, Esmaeil Idani, Narges Khanjani, Gholamreza Goudarzi, Abbas Bahrampour (2016) "Relationship between air pollution, weather, traffic, and trafficrelated mortality", Trauma Mon, Volume 21, issue 4, pp. e37585. PMID:2818

[11] Guodong Liu, Siyu Chen, Ziqian Zeng, Hujie Cui, Yanfei Fang, Dongqing Gu, Zhiyong Yin, Zhengguo Wang (2018) "Risk factor for extremely serious road accidents: results from national road accident statistical annual report of China" PLoS One, 13(8):e0201587. PMID: 30067799.

[12] Lutz Sager (2016) "Estimating the effect of air pollution on road safety using atmospheric temperature" GRI Working Papers 251, Grantham Research Institute on Climate Change and the Environment.

[13] Athanasios Theofilatos, George Yannis, Pantelis Kopelias, Fanis Papadimitriou (2016) "Predicting Road accidents: a rare-events modeling approach, Transportation Research Procedia", Volume 14, pp. 3399- 3405

[14] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Francesca La Torre, Lorenzo Domenichini, Thomas Richter, Stephan Ruhl, Daniel Graham, Niovi Karathodorou (2016) "Road traffic accident prediction modelling: a literature review", Transportation, Volume 170, pp. 245-254

[15] Fred L. Mannering, Chandra R. Bhat (2014) "Analytic methods in accident research: Methodological frontier and futuredirections", Analytic Methods in Accident Research, Volume 1, pp. 1-22