e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

# ROAD ACCIDENT PREDICTOR USING MACHINE LEARNING

**Dipanshu Gupta*[1], Vagisha Goel*[2], Rithik Gupta*[3], Mohd Shariq*[4], Rajesh Singh*[5]**

*[1,2,3,4,5]Department Of Computer Science, Meerut Institute Of Engineering And Technology,
Meerut, Uttar Pradesh, India.

## ABSTRACT

Road Accident are a major cause of death worldwide leading to around 10.25 lakh deaths and 5 crores injuries ever year. Road accidentsare extremely common. If you live in a sprawling metropolis like we do, chances are that we have heard about, witnessed, or even involved in one. Therefore, a system that can predict the occurrence of traffic accidents or accident-prone areas can potentially save lives. Although difficult, traffic accident prediction is not impossible. Accidents do not arise in a purely stochastic manner; their occurrence is influenced by a multitude of factors such as drivers physical conditions, car types, driving speed, traffic condition, road structure and weather. Paper is a deep learning python dynamic Routine maker which is designed to give the user a better understanding with the next day Traffic and help user to fulfill his/her sleep. Our mode Consider the following inputs speed, traffic condition, crash counts, road structure and weather to less obvious factors such as national holidays, the moon cycle and selective attention. Fortunately, several of such accident records are publicly available. Various municipal and national government in the UK have made available rich datasets of Road Traffic Accidents (RTA) and their associated factors. By exploring this government datasets and external data sources, we aim to discover patterns that predict with high accuracy tells road accident to happens.

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

## I. INTRODUCTION

Road Traffic Accidents (RTA) are a major cause of death globally leading to around 1.25 million deaths and 50 million injuries every year. Transport authorities worldwide have been striving to implement strategies to minimize RTA. This however, is a difficult task – despite the adoption of various regulations and safety measures, RTA have not decreased significantly. This failure partially stems from the difficulty in predicting when and where RTA may occur. The occurrence of RTA is correlated with a multitude of factors – from speed, traffic condition, crash counts, road structure and weather to less obvious factors such as national holidays, the moon cycle and selective attention. Various municipal and national governments in the UK have made available rich datasets of RTA and their associated factors. By exploring these government datasets and external data sources, weaim to discover patterns that predict with high accuracy when and where RTA are likely to occur. Our end goal is to create an accurate RTA prediction model in the UK and user-friendly web interface that incorporates:

- A simple explanation of the model itself.
- Visuals that highlight the importance of various factors in predicting RTA.
- An interactive dashboard that allows user to input values and immediately obtain probabilities of RTA occurring in various parts of the UK The outcome of this paper will benefit the general UK public in providing a visualization tool that will communicate the probability of a traffic accident occurring in an area of interest. In addition, it will help the traffic authority in devising strategies to reduce RTA.

Before embarking on this paper, we set ourselves a clear objective: we wanted to create an interactive traffic accident predictor thatwould be easily accessible by anyone. We decided that the best way to achieve this goal was to deploy a trained predictor on a website. This predictor-website should be capable of doing thefollowing:

- Allow users to input an origin and a destination (both of which have to be in greater London) and find the best driving route thatconnects the two.
- Allow user to pick the date/time they plan to make2the trip and identify areas along the route that are particularly accident-pronewithin that time window.

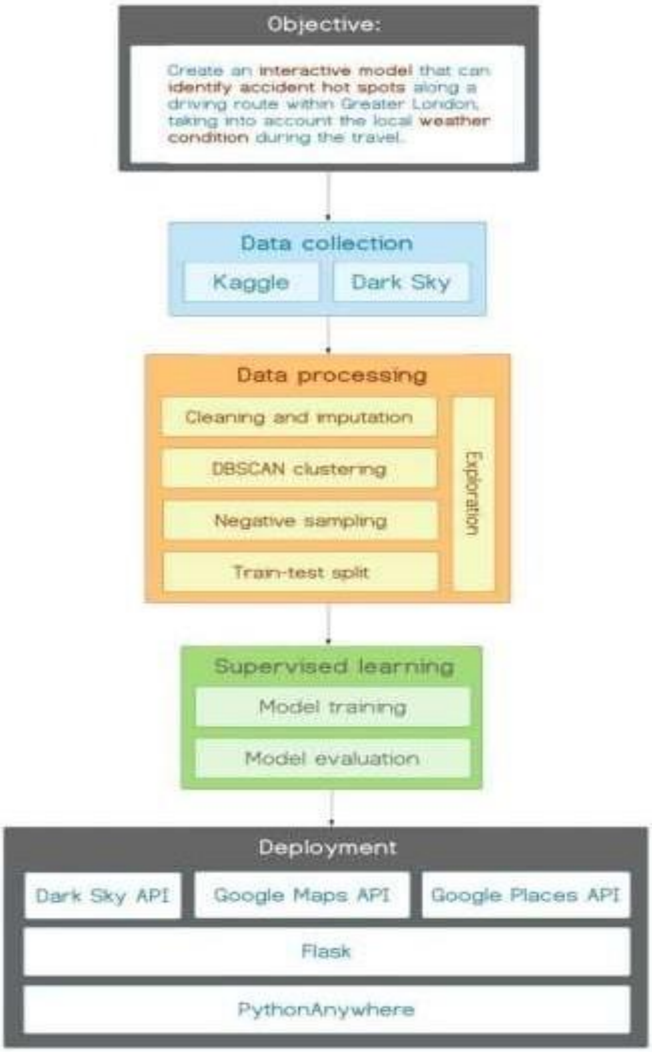


**Figure 1:** Steps in Process

## II. 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 hotspots.

**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. To do predictions, we need negative examples. |
| Prieto et al. [15] | The methodology in the study could not be used for predictions. Hexagonaltessellation was used to providevisualization of the data. Our project will use machine learning for accident prediction and interactive visualization to allow effective communication of insights. | RTA is a rare event but it is also highly concentrated in certainroad segments. Tessellation of space could help with visualization of dense data. Urban city and 3 motorways have very different accident distribution. We have not tested this method yet. |

## III. PROPOSED METHOD

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

Table 2

| Data | Source | Size | Challenges |
|---|---|---|---|
| UK Accident recordsfrom year 2005 till 2014 | Kaggle<br>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<br>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<br>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

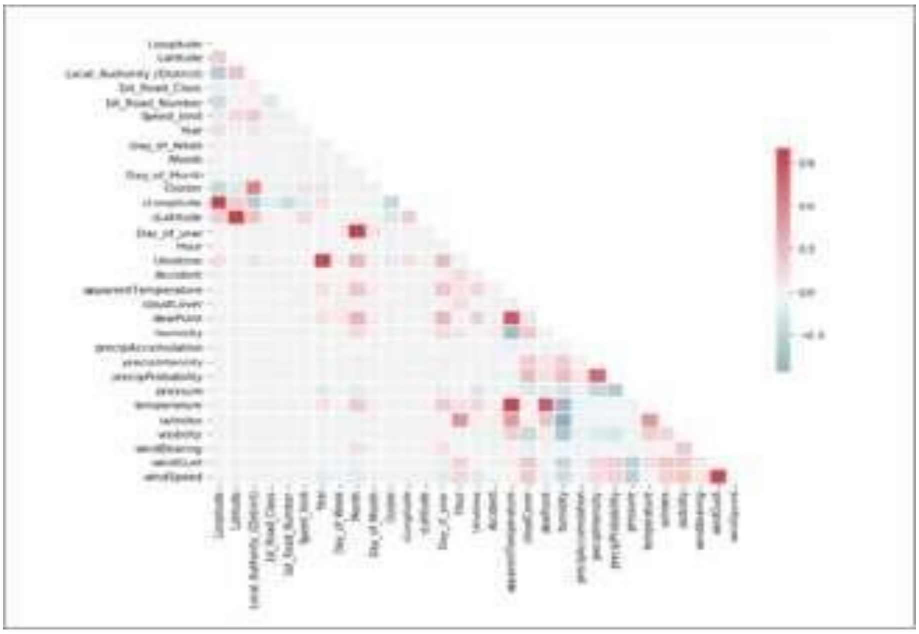


**Figure 2**: Correlation matrix of numeric features in the UKRTA data set.

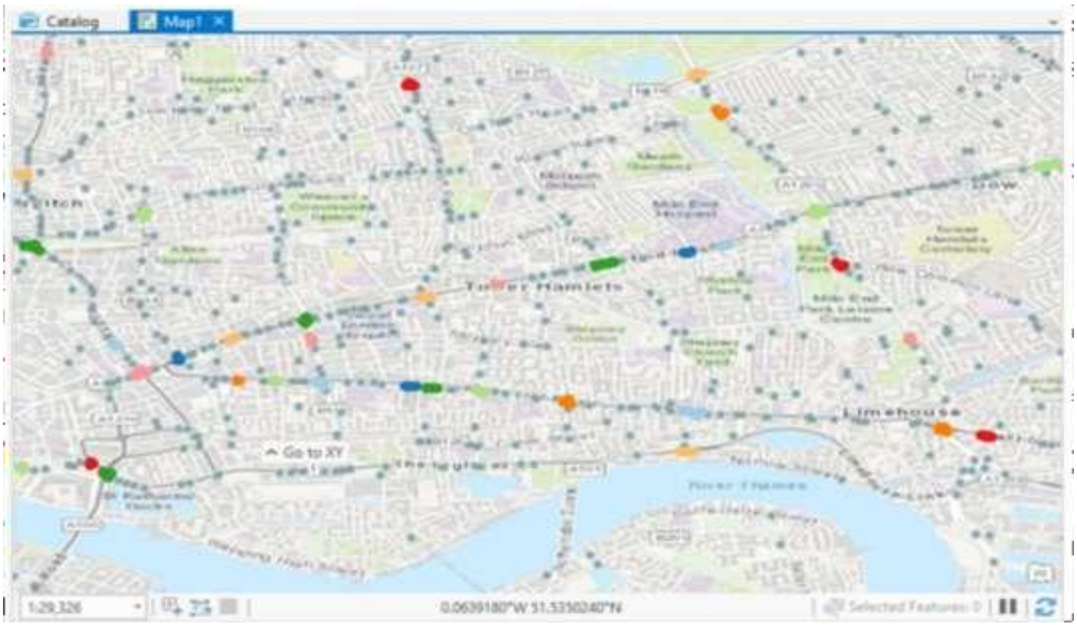


**Figure 3:** Clustered accident data points using ArcGIS. Colored points are hotspots with high density of accident o occurrence.

**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**: Python was used for all model training. Table 3 shows the models attempted and metrics used to compare the quality of the model. The best performing model so far is Random Forest with only numerical/floating point predictors.

**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. |

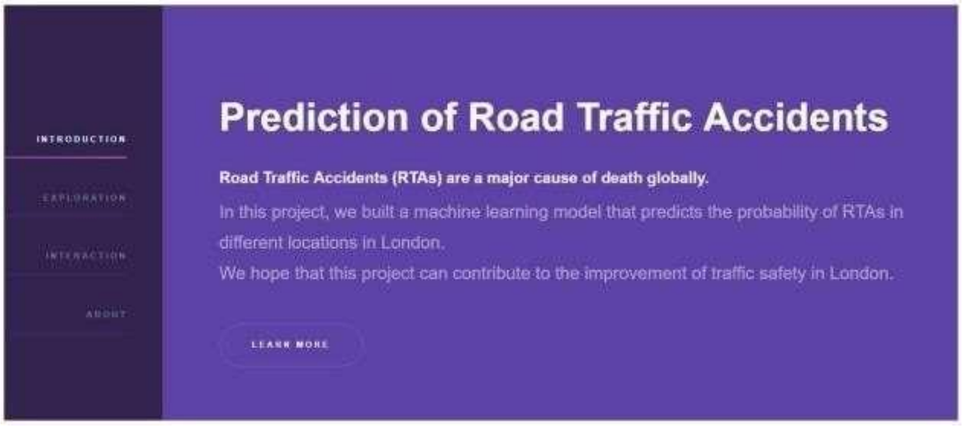
## IV. IMPLEMENTATION

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.



Prediction of Road Traffic Accidents

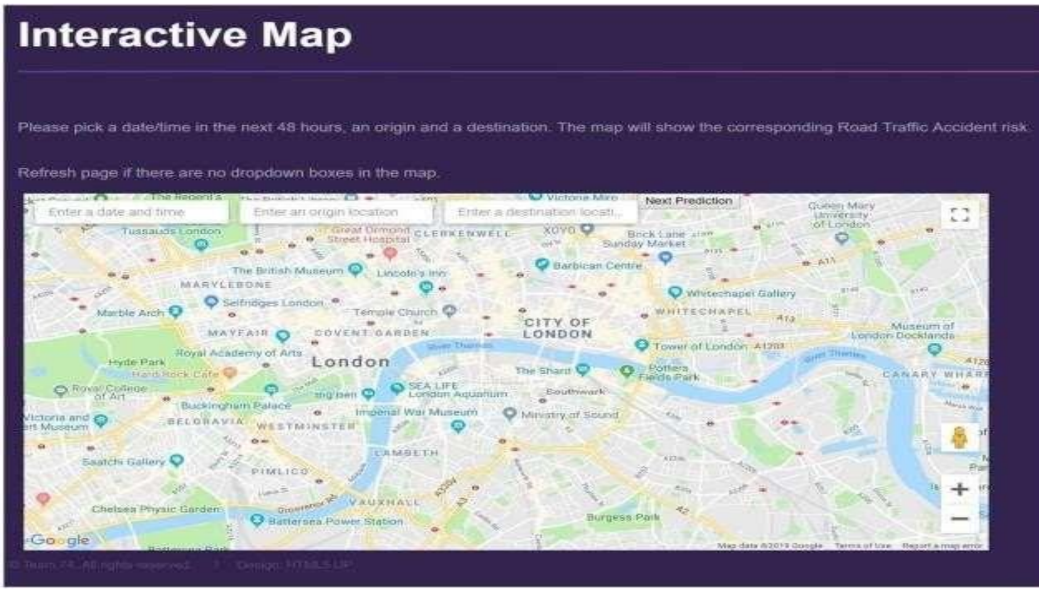
Road Traffic Accidents (RTAs) are a major cause of death globally.

In this project, we built a machine learning model that predicts the probability of RTAs in different locations in London.

We hope that this project can contribute to the improvement of traffic safety in London.

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

Google Maps APIs will be called to show proposed routes based on user-inputted origin and destination. Users are able to input the origin and destination with suggested options presented by Google Place API. We will collect the latitude-longitude coordinates of various places along the determined route and send these coordinates to our backend platform for model prediction. The returned results from the prediction model will be displayed as hazard icons on the route to show probabilities of accidents in the "hotspot"areas.



Interactive Map

Please pick a date/time in the next 48 hours, an origin and a destination. The map will show the corresponding Road Traffic Accident risk.

Refresh page if there are no dropdown boxes in the map.

© Team 74. All rights reserved. | Design: HTML5 UP

**Figure 5**: A screenshot of the Interaction page.

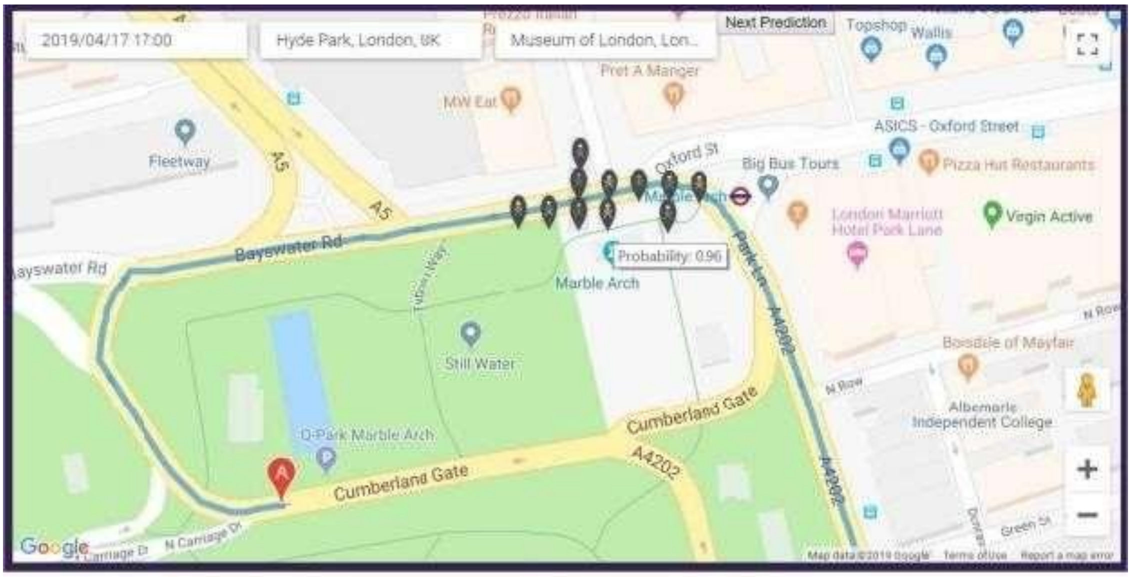


**Figure 6**: A screenshot of user-inputted origin and destination.

## V. 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. |

## VI. 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

[16] J. Milton, V. Shankar, F. Mannering (2008) "Highway accident severities and the mixed logit model: an exploratory empirical analysis", Accident Analysis & Prevention, Volume 40, pp. 260-266

[17] Chao Wang, Mohammed A.Quddus, Stephen G.Ison (2011) "Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model", Accident Analysis & Prevention, Volume 43,issue 6, pp. 1979- 1990, DOI: 10.1016/j.aap.2011.05.016

[18] Prieto Curiel R, Gonzalez Ramirez H, Bishop SR (2018) "A novel rare event approach to measure the randomness and concentration of road accidents" PLoS ONE 13(8): e0201890.