FINAL REPORT

Analysis and Prediction of Traffic Accidents in London

Team 74 : Siew Lee Koh, Katannya Kapeli, Roger Teo Kee Eng, Meraldo Antonio, Rajkumar Lalwani, Yeok Kwan Chong

TABLE OF CONTENTS

I. INTRODUCTION

Problem definition

Road Traffic Accidents (RTAs) are a major cause of death globally leading to around 1.25 million deaths and 50 million injuries every year [2]. Transport authorities worldwide have been striving to implement strategies to minimize RTAs. This, however, is a difficult task – despite the adoption of various regulations and safety measures, RTAs have not decreased significantly. This failure partially stems from the difficulty in predicting when and where RTAs may occur.

Objectives

The occurrence of RTAs is correlated with a multitude of factors – from speed, traffic condition, crash counts [5][6], road structure [6] and weather to less obvious factors such as national holidays, the moon cycle [19] and selective attention [20]. Various municipal and national governments in the UK have made available rich datasets of RTAs and their associated factors. By exploring these government datasets and external data sources, we aim to discover patterns that predict with high accuracy when and where RTAs 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:

- An simple explanation of the model itself
- Visuals that highlight the importance of various factors in predicting RTAs
- An interactive dashboard that allows user to input values and immediately obtain probabilities of RTAs occurring in various parts of the UK

The outcome of this project 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 RTAs.

II. LITERATURE SURVEY

Table 1 provides an overview of the state of the art models as well as their limitations, which we shall attempt to improve upon. Some of our reviewed studies specifically look at the effect of incorporating certain features in RTA prediction accuracy (Table 2).

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

Authors	Ref.	Brief explanation of model(s)	Limitations	Application to our study
Athanasios et al	10	The model uses logistic regression to predict RTA. It considers RTAs as rare events and subjects the resulting probability values to correction steps to account for the rarity of positive samples (accidents)	The model only takes into account traffic data as its predictors and has limited scope (highways in the city of Athens).	Logistics regression, together with the consideration of RTAs as rare event, are among the algorithms we applied in our model development.
Milton et al	13	This 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	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.
Coruh et al	22	The study analyses RTA frequency in 81 cities over three years. It uses random-parameters negative binomial panel count data models.	The paper did not provide detailed explanation as to why the random-parameter approach was able to attain the decent result.	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.

Abdel-Aty et al Sun et al	1,3	These studies examined how well Generalized Estimating Equation (GEE) and Support-Vector Machine (SVM) models real time traffic flow(loop-detector) and roadway geometric features.	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 project is a binary classification problem (e.g. Accident = 0 or 1). Therefore, Generalized Estimating Equation (GEE)[1] is not applicable, since it is more suitable for 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 30 features). We experimented with this model and found that it takes too long (> 2 hours) to train. We decided to use other more efficient models instead.
Wang et al	14	The 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.	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 study employed Rare Event Concentration Coefficient ("RECC") to identify regions of high concentration of road accidents in London city and Mexico highways.	The methodology in the study could not be used for predictions. Hexagonal tessellation was used to provide visualization 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 certain road segments. Tessellation of space could help with visualization of dense data. Urban city and motorways have very different accident distribution. We have not tested this method yet.
Yuan et al	16	The study used binary classification (Support Vector Machine, Decision Tree, Random Forest, and Deep Neural Network) to predict if an accident will happen for each road segment in Iowa, United States using informative sampling techniques and a heterogeneous dataset.	Weather data had a lot of missing values and were imputed by interpolation. Negative examples were generated using authors' own reasoning and the methodology that could be biased.	We used more training data as suggested as it led to better model performance according to the study. We also employed the negative example (no accident) generation method for predicting RTA because it addressed the class imbalance problem.
Fancello et al	4	The study used clustering to divide accident data into homogeneous clusters and subsequently applies Poisson or Negative Binomial ("NB") modelling to each cluster.	This study stresses that models developed were only basic models that will be good starting points for further studies. Improvements can be made by incorporating new explanatory variables that are related to the influence of human characteristics and behaviour on accident cause.	Based on learnings from this study, we have adopted a two-phase approach to our modelling. The first phase involved using density-based clustering (DBSCAN). This is to derive all the accident hotspots in the target area. Accidents away from hotspot clusters tend to be a result of randomness and hence will not be advisable to be used for modelling.

Alexandra et al	6	The study examined the Empirical Bayesian model enhanced by Proportional Discordance Ratio (PDR) similarity technique.	The accuracy of the model diminishes if road 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 it excels in handling the regression-to-the-mean bias.
Markus Deublein et al	21	The study compared Empirical Bayes model with the more recently created model Bayesian Probabilistic Network for traffic accidents.	The model looks at a limited number of datasets and only look at traffic-related factors.	The study suggests that Bayesian Probabilistic Networks performs better compared to Empirical Bayes method.
Lv et al	23	This study compared the use of K-nearest neighbours (KNN) and C-means clustering (CM) method in real time RTA prediction.	As with many other papers on real time RTA prediction, the data is mainly from inductive loop-detectors situated at various points of a particular road segment.	This study suggests that KNN is a simple, fast and effective model as compared to CM. Our project could include KNN as one of the models for comparison.

Table 2. Literature review of features used in RTA prediction models.

Reference authors	Ref.	Feature	Study conclusions	Application to our study
Maryam et al	7	Air quality	This studies did not find a significant correlation between RTAs and air pollution (measured by NO, CO, NO2, NOx PM10, SO2, and O3 rates) in Iran.	This study suggests that air quality may not be a strong feature in predicting RTAs for our model. We used this information in considering to use air quality data in our study.

Guodong et al	8	Air quality	This studies did not find a significant correlation between RTAs and air pollution (measured by NO, CO, NO2, NOx PM10, SO2, and O3 rates) in China.	This study suggests that air quality may not be a strong feature in predicting RTAs for our model. We used this information in considering to use air quality data in our study.
Sager et al	9	Air quality	This study found a positive correlation between RTAs and poor air quality in London.	This study suggests that air quality does impact the occurrence RTAs in London. Air quality measurements are recorded by the London Authority, but data was insufficient, i.e. missing, for 2012-2014 for the London boroughs. Data could not be imputed as more than 10% of data was missing. We use 10% as a general threshold for whether imputation is appropriate.
Vandoros et al	17	Commencement of austerity measures	This study examined the acute effect financial hardship and RTAs in Greece. A positive correlation was found between the start of austerity measures and frequency of RTAs. The authors surmise that stress caused by financial stress contributes to a rise in traffic accidents.	This study suggests that we should include data that measures the economic environment into our RTA prediction model. We have not obtained data yet to reflect the economic environment as side from the Economic policy uncertainty index (see Vandoros et al [5]).
Vandoros et al	18	Economic policy uncertainty Index	The economic policy uncertainty index is a derived daily from an analysis of UK newspapers. This index is an indicator of societal views of Britain's economic uncertainty. Authors found a positive correlation between this index and motor vehicle collisions in Great Britain.	This study suggests that including the economic policy uncertainty index as a feature in our RTA prediction model will improve our model. We decided not to incorporate it into our model due to lack of time in data mapping.

Wen et al	5	Accident count Index Accident Reduction Potential Index	This study aims to propose a method to rank hazardous traffic locations by using two criterion: accident count index & accident reduction potential index.	This study suggests that these criterion are widely used in identifying hazardous traffic locations and may be useful to include these in RTA prediction. We have not included this type of data into our model.
Carolina I.A. Pape-Köhlera et al	19	Moon cycle	There is an indirect relation i.e. moon cycle influences trauma and that in turn affects the road safety	It is valuable to consider other factors while creating the models as these factors may contribute to the error rate. However, we have not included this type of data into our model due to unavailability such data.
Kahneman, Daniel et al.	20	Selective attention	Selective attention causes people to ignore the dangers on the road and that leads to accidents	We have not identified a suitable dataset that measures driver attention, and therefore is not a feature in our model.

III. PROPOSED METHOD

Current approaches for RTA prediction

To date, numerous RTA prediction models have been developed, each with its own strengths and limitations. As RTAs are count data, the relatively simple Poisson regression formed the basis of initial efforts in the field. Over time, limitations of the simple Poisson regression became apparent and alternatives such as negative binomial regression, zero-inflated Poisson and neural network were developed. We have explored these different models in our project.

Limitations of current approaches to RTA prediction

Most of the current models suffer from the following constraints:

- *Inaccessibility*: None of these models presents its findings in an user-friendly manner that allows interaction and exploration. These models are usually published in academic journals and are explained in an abstruse language that is not accessible to the broader public. As a result, they have limited impacts towards changing government policies and behavior of road users.
- *Poor time resolution*: Existing models identify accident hotspots within relatively long time spans – weeks, months or even a year. These long time windows do not lend themselves to real time interactive prediction that we envisioned.

- *Limited features:* Most existing models have limited scope and/or incorporate only few features. Some of them only incorporate traffic conditions and geometric data [1][3][4], while some others concentrate on particular road segments [3][14]. In addition, as noted in previous studies, many existing models yield unsatisfactory results when applied to a different location[4][5][21][7].

Intuition: Why Our Approach Is Better

We have applied the following innovations to address the limitation of current state of the art:

1. ***Inclusion of more environmental features*** (more than 20 features from weather and geospatial data) for prediction of RTA. This is an improvement over existing models [3][4][23] whose predictors are limited to a handful of speed variance data collected by loop detectors placed at a particular road segment.

2. Our prediction also cover a wider range of road segments with a total of 473 hotspots spread across ***32 London boroughs***. Existing models for real-time RTA prediction [3][14] typically only cover a particular segment of a single highway.

3. Our prediction model is able to carry out prediction of RTA probabilities as far as ***48 hours in advance*** (with the help of a 48-hour weather forecast API). Existing models for real-time RTA prediction are limited to carrying out prediction for the next 30 mins to 1 hour.

4. An improvement in accuracy as compared to existing models to predict RTAs as depicted in the following table:

Reference	Model	Accuracy
Our project	*Random Forest*	***0.83***
Sun et al [3]	Support Vector Machine (SVM)	0.80
Lv et al [23]	K-nearest neighbours	0.80

5. The creation of a user-friendly and easily accessible platform for the London public and authorities to utilize our RTA prediction model to inform their driving behaviors.

IV. OUR INNOVATIVE APPROACH

A. Data Preparation

Data Collection

Data was collected from the sources shown in Table 3.

Table 3. Sources of data used in this project.

Data	Source	Size	Challenges
UK Accident records from 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.
Historical weather data (hourly) from year 2012 to 2014	Darksky API https://darksky.net/dev	26 thousand records 22 columns	Some days did not have weather data collected at all. Some days have missing data for random features

Pollution data (daily) from year 2012 to 2014	Openair (R package) http://davidcarslaw.github.io/openair/	Approx. 1,000 records	More than 10% data missing for 2012-2014 years; unable to impute data therefore cannot include as a feature in our model.
Economic Policy Uncertainty Index (daily) from year 2012 to 2014	UK Daily Policy Data http://www.policyuncertainty.com/uk_monthly.html	Approx. 1,000 records	N.A.

Data Exploration

A basic Exploratory Data Analysis (EDA) was performed on the datasets.

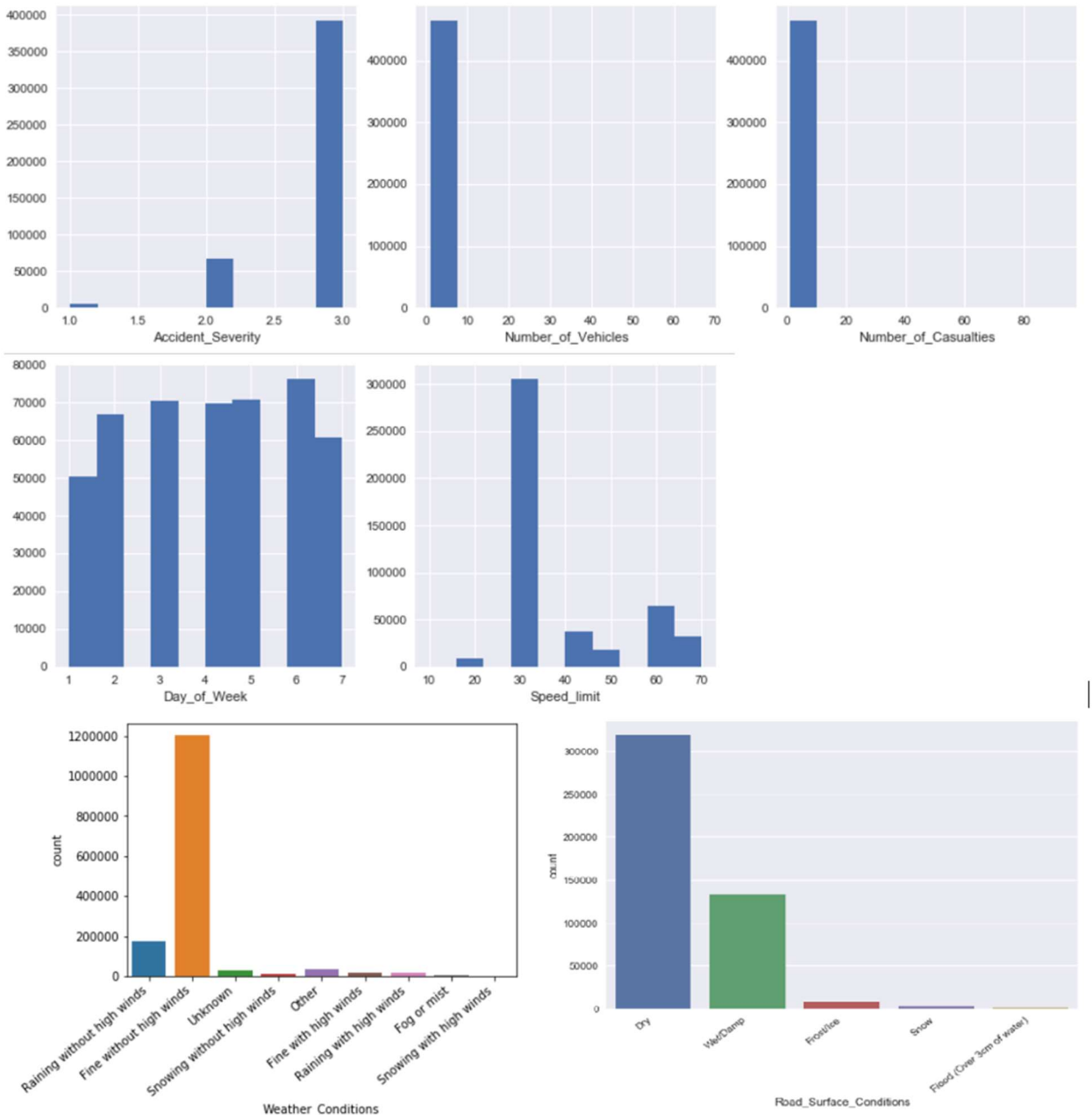


Figure 1: Visualization of some important features in the UK RTA dataset.

The visualizations in Figure 1 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.

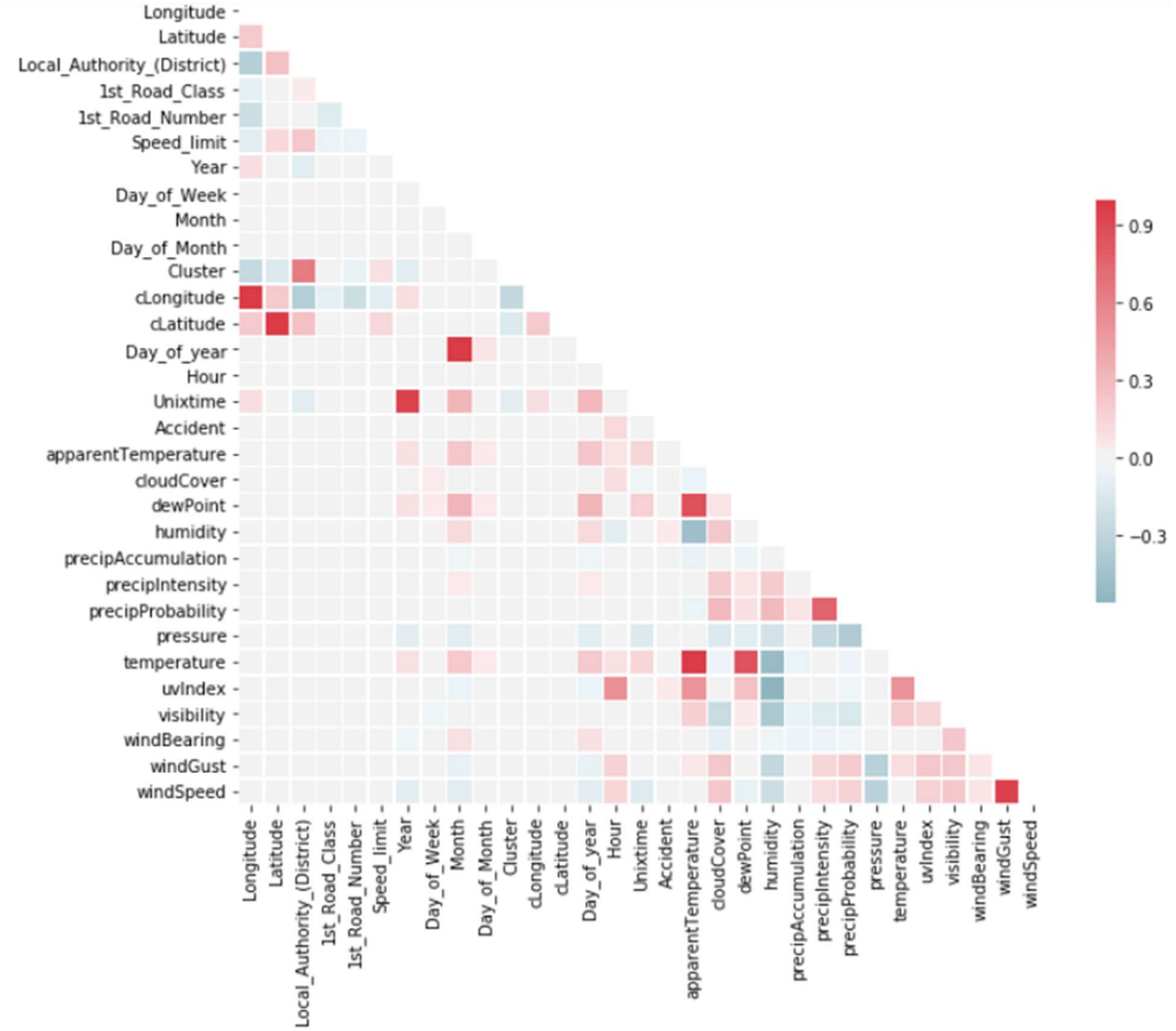


Figure 2. Correlation matrix of numeric features in the UK RTA data set.

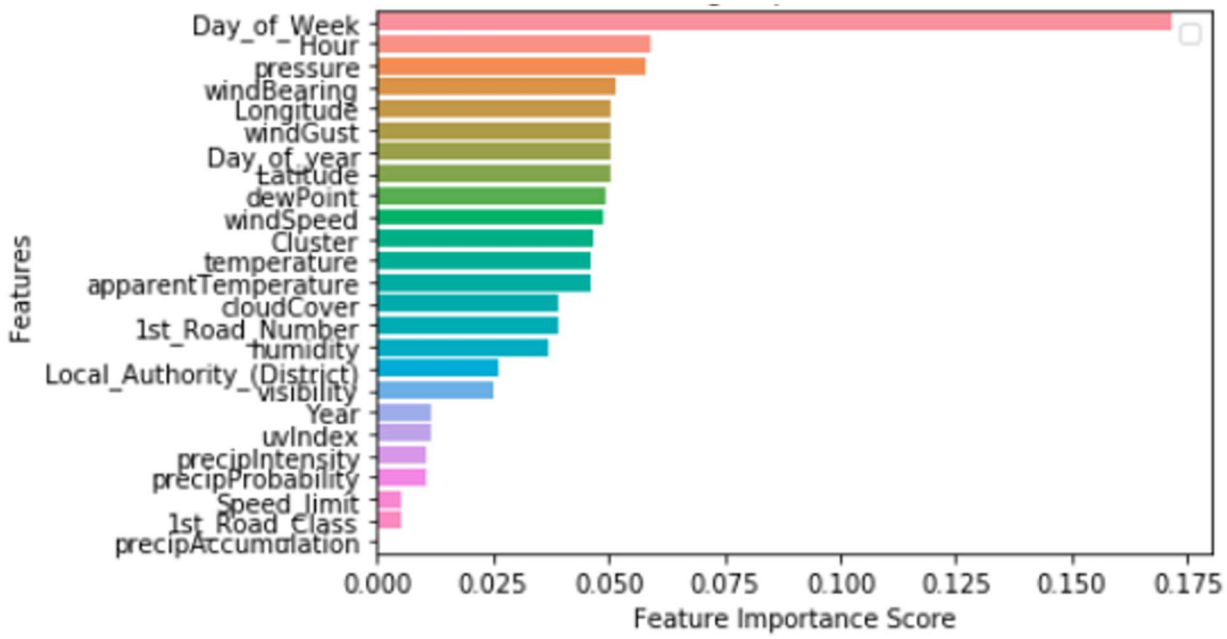


Figure 3. Importance score of features used in the Random Forest Classifier.

Data Imputation

As mentioned in the previous section, there are many missing data scattered across various data columns. Table 4 shows the imputation method based on different circumstances.

Table 4. Methods for imputation of missing data for data sets used in this project.

Extend of missing data	**Imputation Method**	**Examples of data imputed this way**
Affect small percentage of primary data or Does not severely impact modelling outcome if omitted	Omit entire rows, entire columns or entire data set	UK pollution data
Affect large percentage of primary data	Replace null values using monthly mean	Weather data - ‘Cloud cover’ and ‘Pressure’

Negative Samples Generation

Our primary accidents dataset contains features that were captured only *when an accident happens*, i.e. all labels = 1. Therefore, we needed to generate negative samples. Using a similar method described in Yuan et al [16], we created three negative samples for every accident record within the clusters. We randomly varied the day of the year or hour of a real accident data point, then checked if this synthesized negative sample matched a real data point. If not, we included this randomly-generated record as a negative sample.

B. Model Training for Prediction

Clustering of Data Points

Many of the accident data points are very close to one another as shown in Figure 4.

Figure 4. Initial visualization of all accidents data points in London using Tableau. Almost the entire road network are cluttered with accidents.

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 5). This will improve the accuracy of traffic accident prediction.

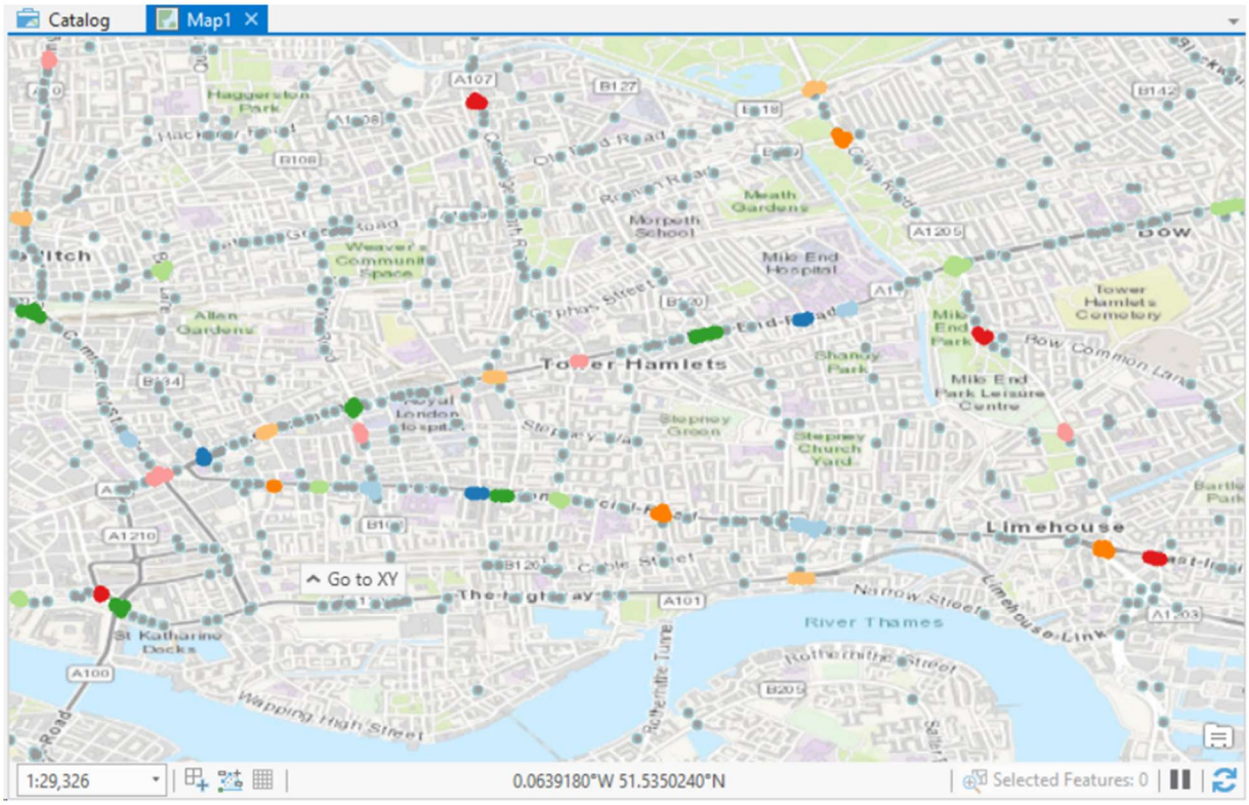


Figure 5. Clustered accident data points using ArcGIS. Colored points are 'hotspots' with high density of accident occurrence. Grey points are 'noises' (non-high density). Data points indicated as 'noises' are excluded from model training.

Using ArcGIS software, we clustered the data points into 473 clusters or RTA "hotspots" that will be used in our model training and visualization dashboard. The criteria for density-based clustering (DBSCAN) was set at a minimum of 14 accidents per cluster with a search distance of 25 meters. 'Noise' data points, which are excluded by DBSCAN, will not be used for modelling.

Models to Predict RTAs

Python was used for all model training. Table 5 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 5. Summary of algorithms used to build model and their performance based on accuracy and ROC-AUC values.

Models	Predictors	Evaluation Metrics	
		Accuracy	AUC
Logistics Regression	25 Numerical / Floating point features	0.7611	0.6753
Random Forest with only numerical predictors	25 Numerical / Floating point features	0.8347	0.8705
Random Forest with one-hot-encoding of categorical features	21 Numerical features with highly-correlated counterpart removed 2 Categorical features converted to 12 numeric features with one-hot-encoding	0.7865	0.499
Support Vector Machine (SVM)	21 Numerical features	Inconclusive since SVM ran for hours and has yet to complete.	Inconclusive

V. IMPLEMENTATION

A. Backend Implementation

Web development

A web application has been built using the Flask framework. Flask is based on Python as the Server Side Language and will be used to handle all the server side processes. 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. The first 40,000 calls for Routes API and the first 70,000

characters for the Places API in a month are free. Weather forecasts is obtained by calling Darksky API and it has free 1000 calls per day which is enough for our development use. In addition, a RESTful 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 metres 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 don't fall within this 50m distance are filtered out.

Next, for each unique cluster in the remaining accident points, Darksky API is called. Each unique cluster will have the same weather forecast. This is a reasonable imputation as each cluster has a 25 metres 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. We have not explored other methods of assigning higher probability to these accident-prone locations.

Frontend can now use the predictions to generate visualizations and highlight potential accident sites.

B. Frontend implementation

Creating a route planning tool with RTA hotspots: An interactive visual

Our project is presented as a website at http://kteo7.pythonanywhere.com/. This website contains two main sections: "***Exploration***" and "***Interaction***".

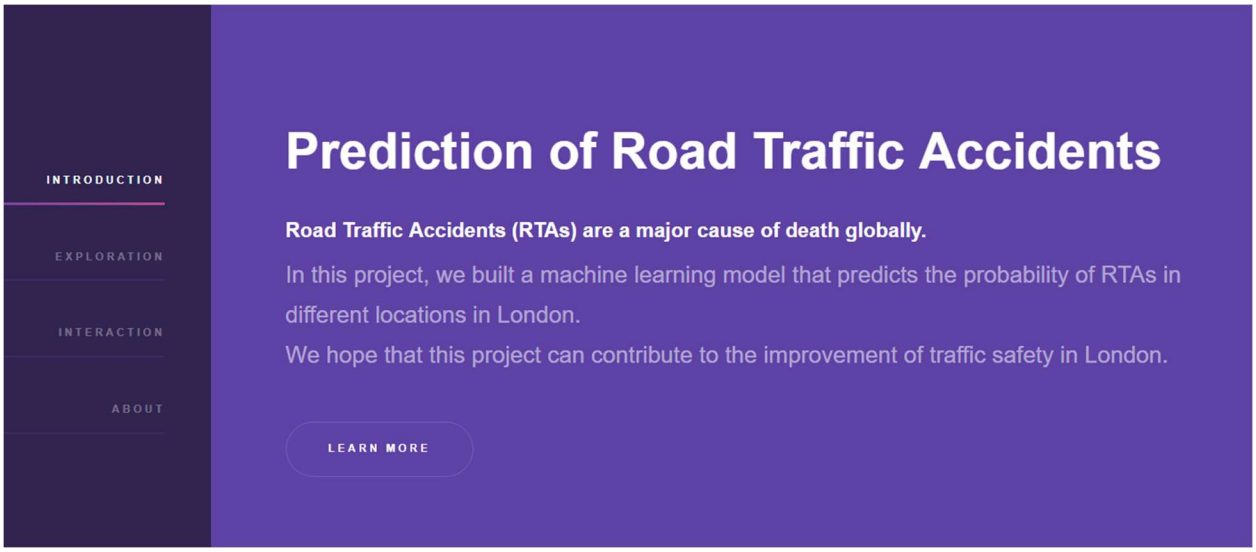


Figure 6. A screenshot of the website front page. The website will contain two main sections: "Exploration" and "Interaction".

The "***Exploration***" section (described in more detail in next section) includes a general background of the project as well as the most important takeaways of the EDA steps. It will present several figures that elucidate the importance of the features similar to those shown in above. The section also contains a brief explanation of model with the best performance.

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.

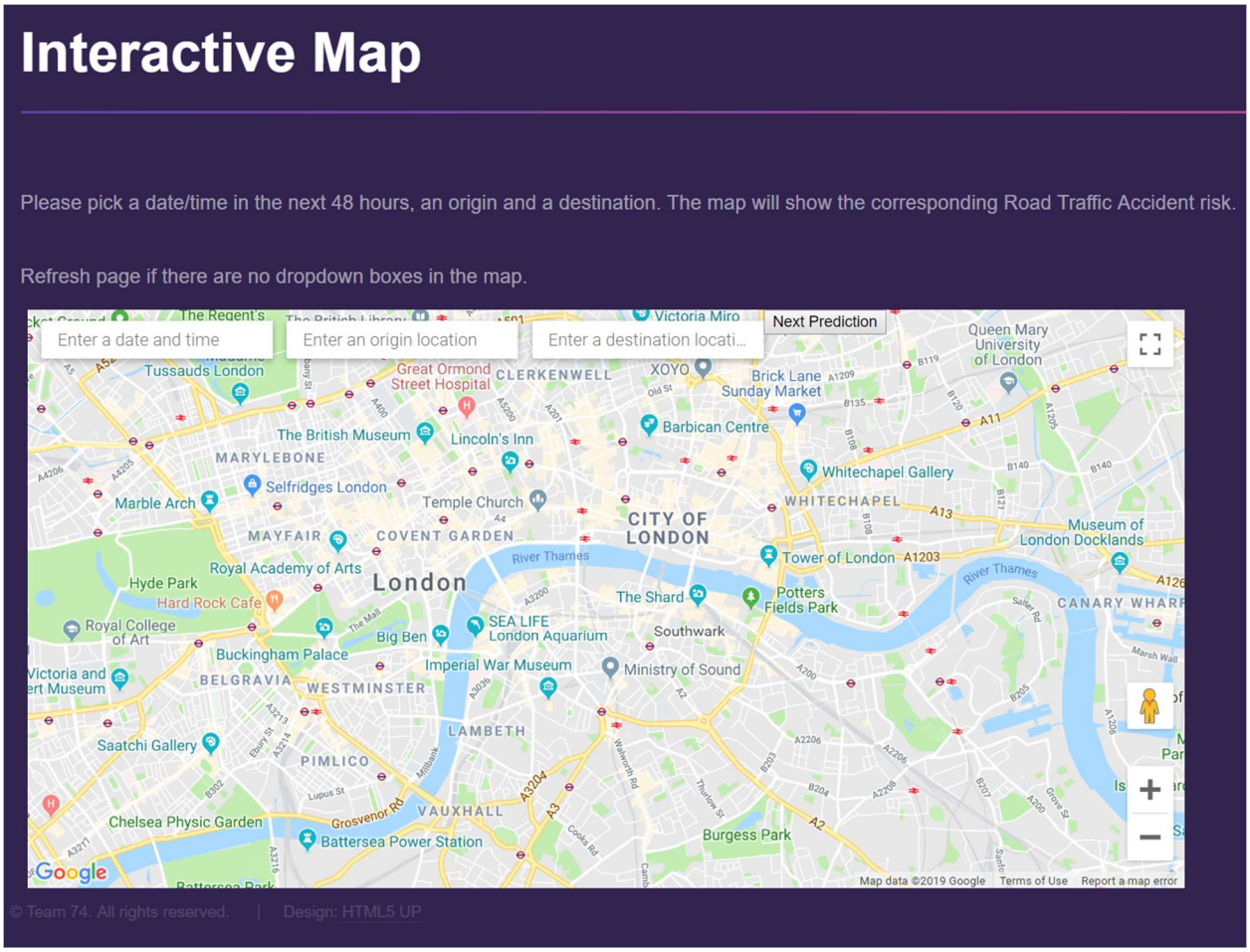


Figure 7. A screenshot of the ***Interaction*** page.

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. Figure 8 shows a screenshot of the map after prediction.

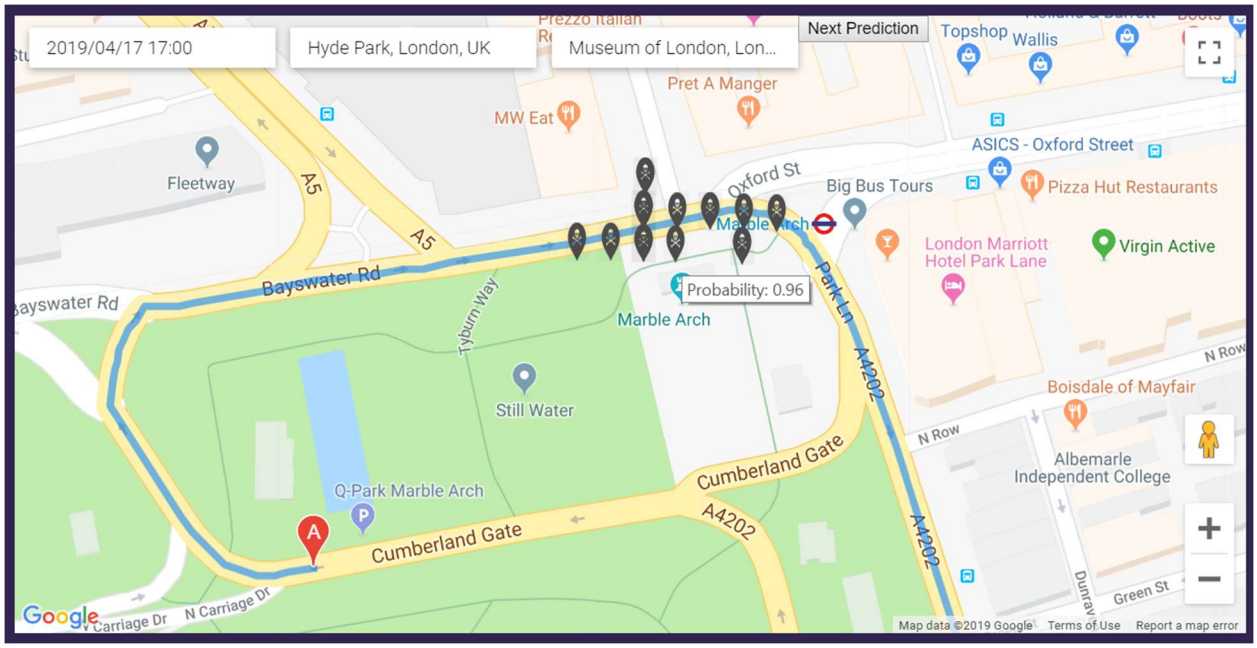


Figure 8. A screenshot of user-inputted origin and destination. Several latitude and longitude coordinates along the "best" route will be inputted to our model to predict accident points along the best route. These points are indicated with a hazard icon to inform the user. Hovering over the icons will show the predicted probability of an accident occurring.

Creating a visual exploration of UK Traffic data

Apart from the above visualization showcasing the predicted probabilities of RTA, the web application also presents some interactive visuals on explorative aspects of the RTA data. These informative visuals allow the user to “explore” general trends in RTAs in London, for example, which boroughs have the most and least RTAs (Figure 9) and which are the worst days of the week to drive by month (Figure 10) or by hour (Figure 11).

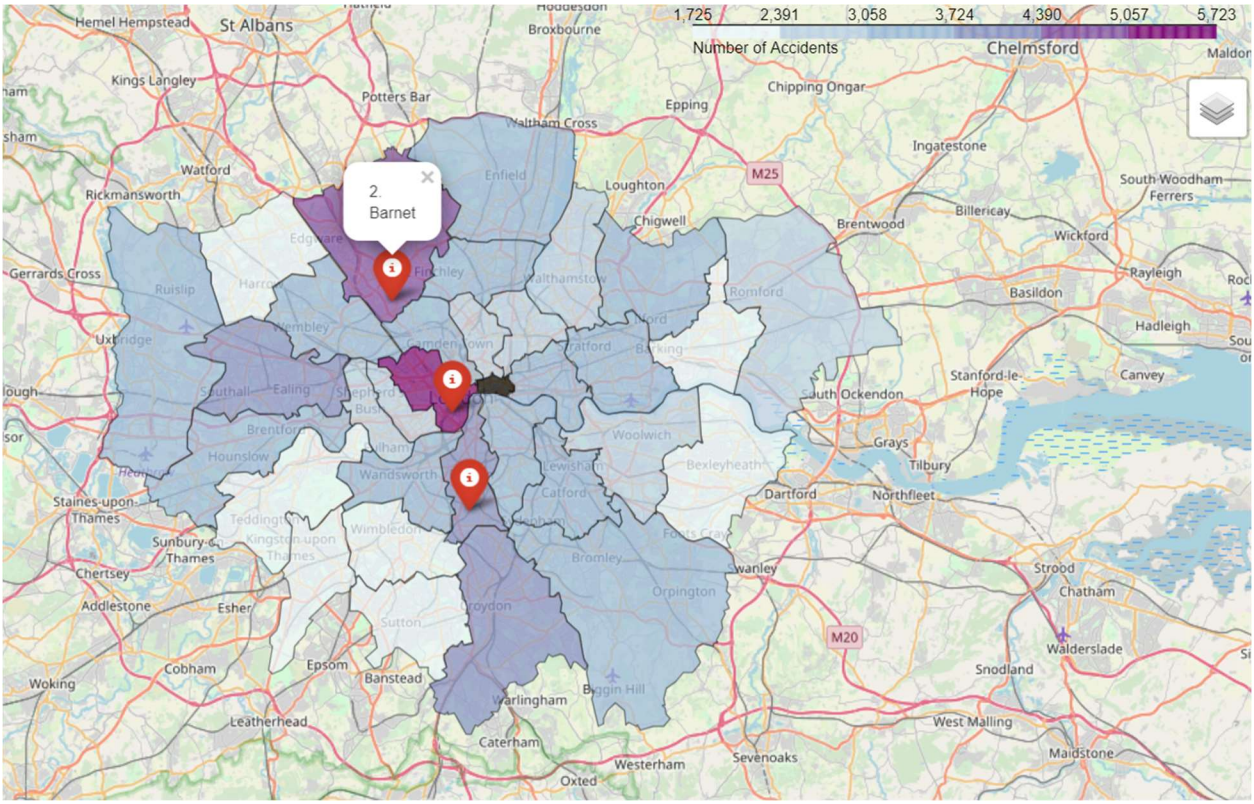


Figure 9. Choropleth (using 'Folium' package in Python) showing total RTAs in London boroughs from 2009-2014. The top 3 boroughs with the most number of RTAs are denoted with tooltips.

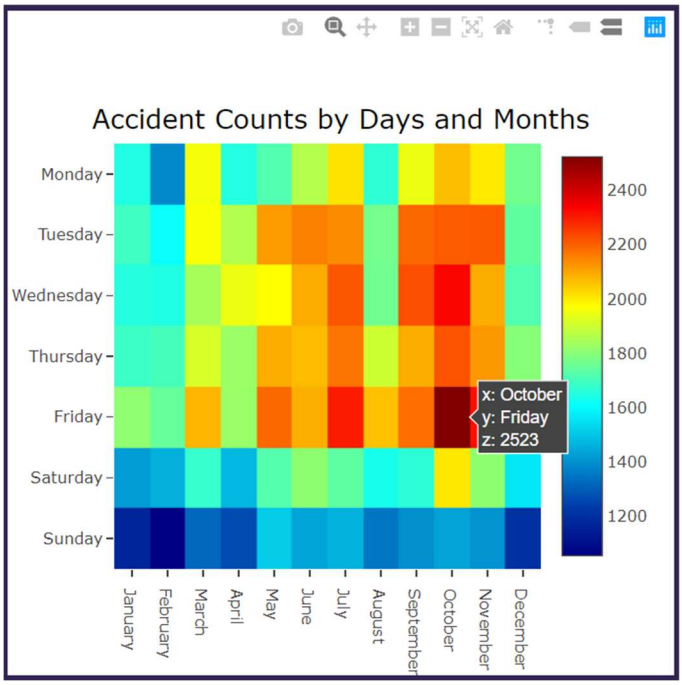


Figure 10. Heatmap showing the number RTAs in all London boroughs from 2009-2014 by day for each month. The visual suggests that Fridays generally have the most RTAs, with the fall months especially October having higher RTA counts.

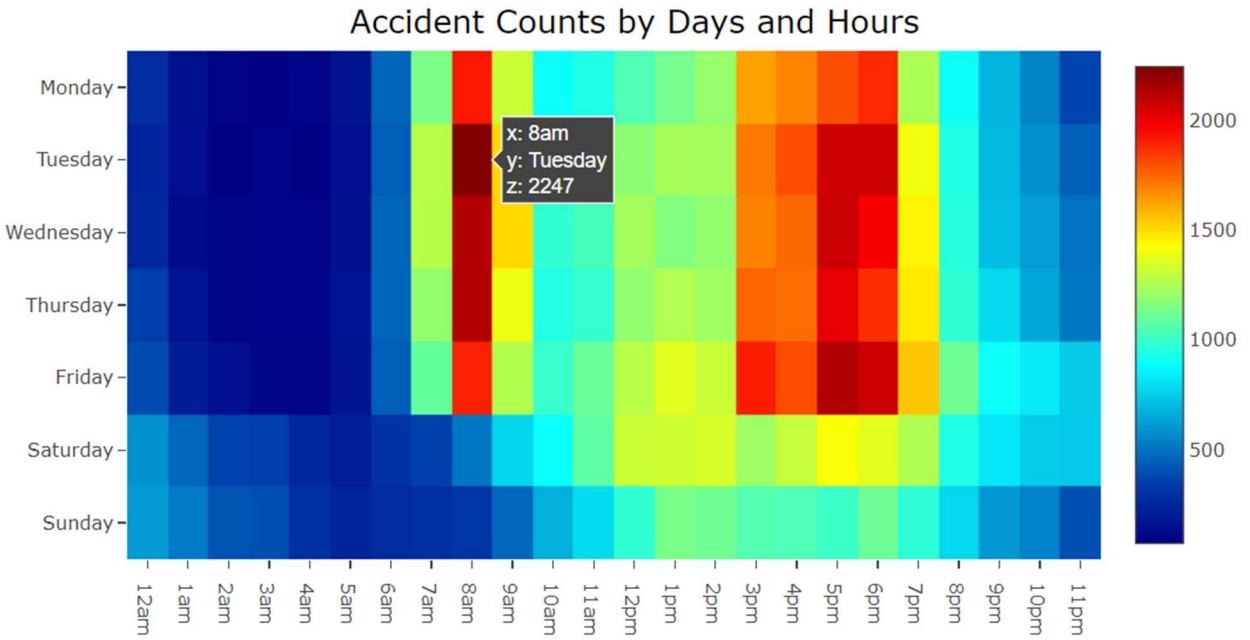


Figure 11. Heatmap showing the number RTAs in all London boroughs from 2009-2014 by day for each hour of the day. As expected, "rush hour" commuter times (8-9 am and 4-6 pm) have the highest RTA incidences.

VI. EXPERIMENTS AND EVALUATION

A. Experimental Approaches

Our experimental approach involves first doing data exploration and visualization on the UK Accidents Dataset to have an idea of the number of predictor variables we can use for modeling and also to know which predictor variables could have a larger effect on accident prediction. Upon visualization of the accidents' locations on a map, it is observed that accidents are mostly clustered around road intersections.

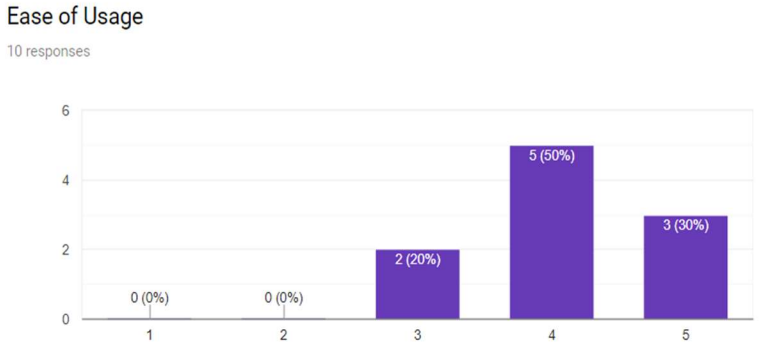
Next, we did ***density-based clustering*** (DBSCAN) to acquire all the accident hotspots. The result of the clustering exercise revealed 'hotspots' that clustered around road intersections and other accident-prone areas. Modelling our prediction for accidents in hotspots will greatly improve our model's prediction accuracy as accidents within hotspots are less likely a result of randomness.

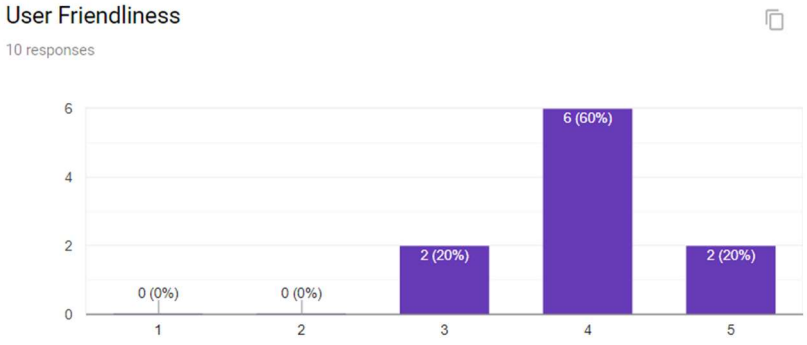
Lastly, in our approach, we used a more recent data set within the UK accidents data set and apply classification modelling using ***Logistics Regression, Random Forest and SVM***, choosing the model that yields the highest accuracy and AUC. Each data set is split into 70% training and 30% testing. Results are satisfactory (Table 5) but we believed doing hyperparameter tuning could yield us better results.

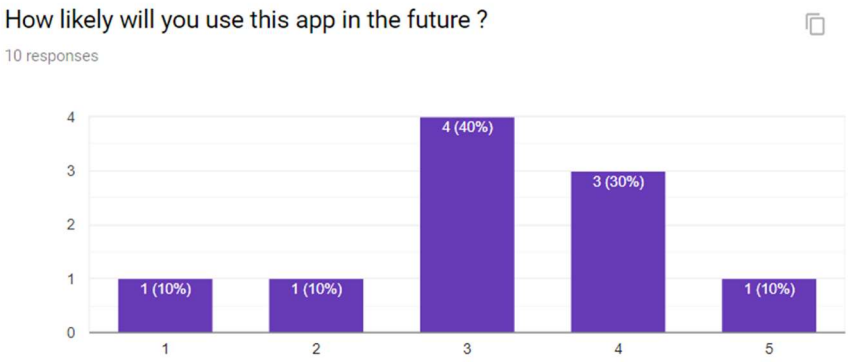
B. Evaluation

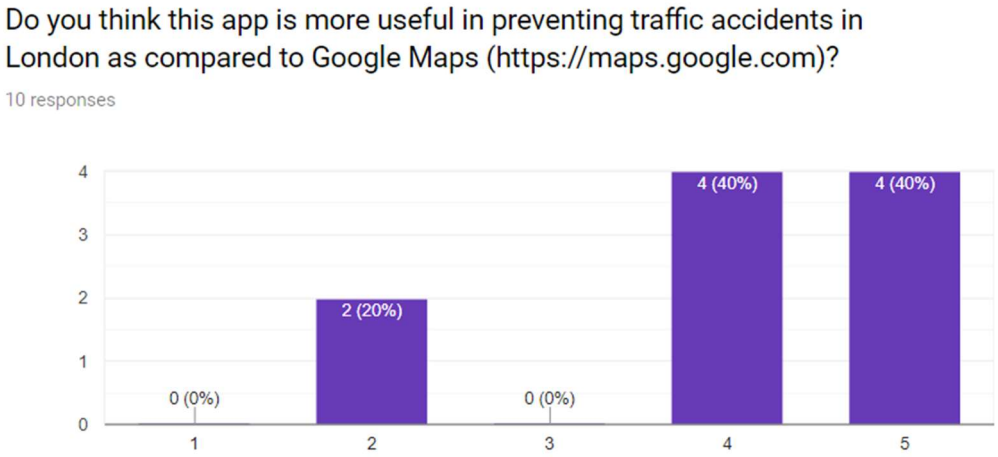
A total of 10 participants evaluated our RTA prediction app via a survey. Respondents were asked to compare our app with Google Maps with real time traffic information. Most respondents found our application:

- Easy to use and user friendly
- more useful in preventing accidents (as compared to Google Maps)









VII. DISCUSSION AND CONCLUSIONS

Current Limitations and Future Work

In this project, 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) and Darksky API (for weather forecast).

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

predicted probabilities on different day and time coincides with our findings during data exploration stage.

While we have largely met our project objectives, this project has also exhibited a few limitations. The following tables show the limitations and how it can be improved in future studies:

Limitations	**Future Work**
Prediction Accuracy (currently at 0.83) can be improved.	Introduce hyperparameter tuning for modelling. Experiment with 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 future studies
Current model does not take into account traffic volume. If a particular road experience high number of accidents, it may be perceived 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 metres radius for filtering prediction locations and eliminating 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 more informative.
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 of tourist spots etc. More past data could be included in the model.

Distribution of Team effort

The following table describes which team members are responsible for the various aspects of this project.

Category	Task	Tools	Members
Concept Development	Scope Finalization	Google Scholar Google Drive NCBI	All group members
	Business Understanding		
	Literature Review		
	Planning		
Model Development	Data Acquisition	Tableau Python Git/GitHub	Y.K. Chong
	Data Cleaning		
	Exploratory Data Analysis		K.E.R Teo
	Model training and optimization		
Frontend Development and Visualization	Mock-up visualization	HTML Javascript Python –Plotly, Folium Google Map API Darksky.net API Git/GitHub	M. Antonio
	Visualization of existing data and model output		
	UI design and development		R. Lalwani
	API integration		
	UI testing and validation		
Backend Development	API and web development	Python-Flask Git/GitHub	S.L Koh
	Testing and validation		K. Kapeli

Member contributions to the project were as follows:

9% Rajkumar Lalwani
18.2% Siew Lee Koh
18.2% Katannya Kapeli
18.2% Roger Teo Kee Eng
18.2% Meraldo Antonio
18.2% Yeok Kwan Chong

VIII. 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: Journal of the Transportation Research Board, Volume 2432, pp 91-98

[4] Fancello Gianfranco, Stefano Soddu & Paolo Fadda (2018) "*An Accident Prediction Model for Urban Road Networks,"* Journal of Transportation Safety & Security, Volume 10, issue 4, pp. 387-405

[5] 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, issue 1, pp. 40-55

[6] 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, DOI: 10.1080/19439962.2018.1450314

[7] Maryam Dastoorpoor, Esmaeil Idani, Narges Khanjani, Gholamreza Goudarzi, Abbas Bahrampour (2016) *"Relationship between air pollution, weather, traffic, and traffic-related mortality*", Trauma Mon, Volume 21, issue 4, pp. e37585. PMID: 28180125

[8] 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.

[9] 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. (link)

[10] 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 (link)

[11] 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 (link)

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

[13] 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 (link)

[14] 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

[15] 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. DOI: 10.1371/journal.pone.0201890

[16] Zhuoning Yuan, Xun Zhou, Tianbao Yang, James Tamerius, Ricardo Mantilla (2017) *"Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study"* PLoS ONE 13(8): e0201890. DOI: 10.475/123_4

[17] Vandoros S, Kavetsos G, Dolan P (2014). "*Greasy Roads: The Impact of Bad Financial News on Road Traffic Accidents*" Risk Analysis, Volume 34, issue 3, pp. 556-566.

[18] Vandoros S, Avendano M. Kawachi I (2018) "*The short-term impact of economic uncertainty on motor vehicle collisions"* Preventive Medicine Volume 111, pp. 87-93.

[19] Carolina I.A. Pape-Köhlera, Christian Simanskib, Ulrike Nienaberc, Rolf Lefering (2014) "*External Factors and the Incidence of Severe Trauma: Time, Date, Season and Moon"* Volume 45, pp. 93-99 (link)

[20] Kahneman, Daniel Ben-Ishai, Rachel Lotan, Michael Feishman, Edwin A. (1973) "*Relation of A Test of Attention to Road Accidents",* Journal of Applied Psychology, Volume 58, issue 1, pp. 113-115

[21] Markus Deublein, Matthias Schubert & Bryan T. Adey (2014) "*Prediction of road accidents: comparison of two Bayesian methods*", Structure and Infrastructure Engineering, Volume 10, pp. 1394-1416

[22] Emine Coruh, Abdulbaki Bilgic, Ahmet Tortum (2005) "*Accident analysis with aggregated data: The random parameters negative binomial panel count data model*", Analytic Methods in Accident Research, Volume 7 pp 37-49 (link)

[23] Yisheng Lv, Shuming Tang, Hongxia Zhao (2009) "*Real-Time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method*", *2009 International Conference on Measuring Technology and Mechatronics Automation*, DOI: 10.1109/ICMTMA.2009.657