Road Traffic Accident Prediction

CSE6242 Data Analytic and Visualization, Georgia Institute of Technology

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

Motivation

What is the problem?

Transport authorities worldwide implement strategies to minimize road traffic accidents (RTA). Despite their best efforts, RTAs have not significantly decreased significantly due to the difficulty in predicting when and where RTAs will occur.

Why is it important?

Road Traffic Accidents are a major cause of death globally, leading to a staggering **1.25 million deaths** and **50 million injuries** every year.

Objectives

1) Build a classification model to accurately *predict the incident of road traffic accidents* in London
2) Create an *interactive and user-friendly* application for the public to use our model to inform their driving decisions.

Limitations of current practice

- Some models predict where but not when accidents will occur (*Wang et al.)
- Some models are built using limited datasets that yield poor accuracy (*Sun et al)
- Studies focus which features are correlated with traffic accidents, not accident prediction (*Sager et al.)
- Inaccessible to the public: the majority of studies publish their findings in subscription journals in technical jargon.

*one of many examples analyzed in this study

Dataset

Road traffic collisions dataset

- A comprehensive dataset of RTAs in the UK from 2000-2016 (discontinuous)
- Data collected and hosted by UK Dept of Transport
- Compiled data available on Kaggle: 1.6 million accident records x 32 features
- Features include:
  - location: latitude/longitude coordinates
  - date and time
  - weather conditions

UK Historical Weather records

- Hourly readings from 2012-2014
- Darksky.net API
- Data collected: 26 thousand data points by 22 weather features

Air pollution Data

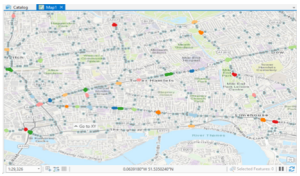
- Approximately one thousand rows of daily reading of +5 air pollutants

Department for Transport | kaggle | openair by David Carslaw

Building a predictive model

Clustering accident points to reduce noise

Accident "hotspots" were determined using ArcGIS DBSCAN. Hotspots are considered traffic accident signal, while accidents away from these clusters are more likely random and are considered noise (Francello *et* al.). A total of 473 'hotspots' clusters were generated. In the figure below, colored dots are signal hotspots while grey dots are noise.



Feature selection

We examined the correlation between features and excluded some features to reduce the effect of multicollinearity (figure below). For models such as random forest, multicollinearity is not an issue. We also excluded air pollution data due to too many data points missing.

Generating negative data points

Our accident dataset only contains positive data points (i.e. label = 0). Similar to Yuan et al., for every data point within a cluster we randomly changed the day, year, and/or hour. Three negative data points were created per positive data point.

Building and evaluating a predictive model

We test several algorithms to build our model and found that the random forest model using only numeric features resulted in the highest accuracy (0.83) and AUC (0.87). Based on these metrics, we used this model for prediction in our web application.

	No. features	Accuracy	AUC
Random forest (numeric)	25	0.8347	0.8705
Random forest	23	0.7865	0.499
Logistic regression	25	0.7611	0.6753
Support Vector Machine	21	*ND	*ND

* Not determined (ND). SVM model was stopped after running for two hours.

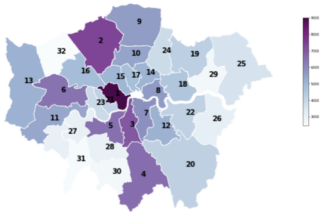
Implementation of an interactive web application

The Interactive *Road Traffic Accident Prediction* App

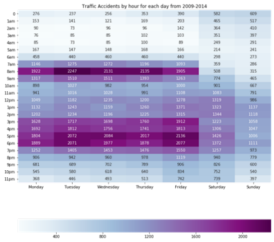
We will present our project as a website featuring "*Interactive*" and "*Exploration*" sections:

The *Exploration* section will take users through a story-line analysis of the traffic data. This visual presentation will ask questions such as:

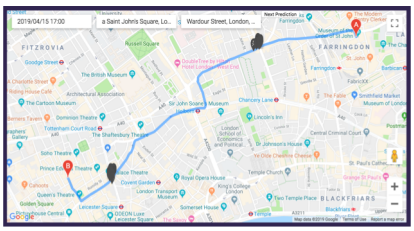
Which boroughs in London have the most and least number of traffic accidents?



Which times, days, and months do traffic accidents most frequently occur?



The *Interactive* section contains an interactive map where users input their start and end location within London and travel date and time (in the next 48 hours). A screenshot is shown below. Based on the above inputs, the backend engine will take user data, run the data point against our model, and return accident prone sites with the accident probability (displayed as crossbone markers)

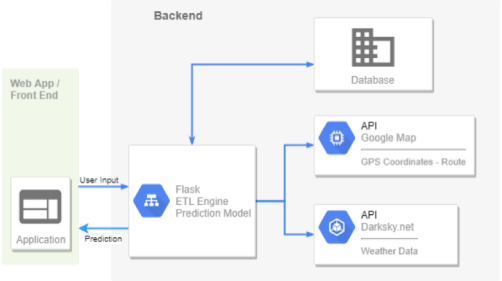


Integrating the front and back end

Front end: The front end will present a clean and simple map to collect user inputs

Back end: The web application was build using Flask, a Python-based server-side language. For the Interactive maps , the backend engine will do the following based on user input:

- Call Google Map API to return a proposed route with a set of GPS coordinates along the route
- Call Darksky.net API for weather forecast for the chosen date, time and GPS coordinates
- Feed all predictors from both API into the prediction model and return a probability of traffic accidents occurring along the route.
- Return the probabilities for each GPS coordinate along the route in question to the front end.



Evaluation

A total of ten participants evaluated our RTA prediction using a survey. Participants are asked to compare our app with Google Maps with real time traffic information. The were asked to rate and comment on the ease of use, user-friendliness, likelihood of using the information to inform their driving behavior, and likelihood of using an (improved) app like ours in the future. In summary, the results from the survey suggest that...

References:

Fancello et al. (2018) An accident prediction model for urban road networks, *Journal of Transportation Safety & Security*, 10: 387-405.
Sager (2016) Estimating the effect of air pollution on road safety using atmospheric temperature. GRI Working Papers 251.
Sun et al. (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, 2432: 91-98.
Wang et al. (2011) Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model, Accident Analysis & Prevention, 43:1979-1990.
Yuan et al. (2017) Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. PLoS ONE 13(8): e0201890.