Toward Smart Cities: Data-driven Road Accident Prevention

Group 22 – A. Aleong, F. Cruz, Y. Ko, S. Yu, H. Hu, H. Xing

DriveAlive

Executive Summary

Road traffic accidents are the leading cause of death, particularly for the economically active age group around the world. Previous efforts have been made to mitigate road accidents in the form of stricter policies, enforcement, and improved infrastructure. However, implementing such strategies requires large capital investment and time. To aid with reducing traffic accidents, a data-driven classification model, DriveAlive, was created to quantify potentially hazardous areas within the city based on historical traffic accident data, current road conditions, and the weather. This model was created from a bottom-up approach by selecting relevant and important features based on datasets available from cities and historical traffic records. The City of Toronto was selected for the first application of this model, where data was retrieved from the Vision Zero Project and the Toronto Police Service [1-2]. Multiple classification models were tested including Random Forests and Deep Neural Networks. Among those used, the random forest model achieved the highest accuracy of 90.6% when predicting traffic accidents. By identifying features associated with high risk areas, DriveAlive can be extended to predict the safest optimal path between two points. With integration as an additional feature in mapping services, a recent market study in the City of Toronto indicated potential revenues ranging from CAD 100,000 upwards solely due to the large population commuting within, to and from the city. Millions more can be saved by factoring in reduced property damages, and healthcare costs. Furthermore, by recognising features associated with high risk areas, this application can also be used by local municipalities and governments for planning resource applications and provide insight for prospective infrastructure projects to mitigate the occurrence of accidents in the future. All in all, DriveAlive provides a simple, inexpensive, and robust alternative capable of addressing a growing safety concern in a world moving towards smart cities and higher rates of urbanization.

1.0 Introduction and Motivation

More than 1.2 million people die each year due to road traffic accidents [3]. Millions more sustain injuries and live on with long-term adverse health consequences. Based on a recent study conducted by the World Health Organization (WHO), road traffic accidents are the leading cause of death among those aged 15-29 years old [3]. This rise in fatalities is detrimental, particularly in developing countries, as they affect the economically active age group. Furthermore, this escalating death toll is exacerbated by the increase in urbanization and motorization which accompany economic growth that is common in these developing countries [3]. Local policies, enforcement, and infrastructure in these areas have not kept pace with rising vehicle use and traffic congestion. In fact, most crashes are both predictable and preventable. Countries that have implemented policies to reduce traffic deaths yielded promising results [3]. The aim of this project is to empower drivers to take safer routes. Implementing new policies and infrastructure takes time. Hence, a bottom-up approach has greater potential for earlier rewards and aligns with the smart city motif where the government and people work together toward a common goal.

In recognition of this public safety and development problem, there is growing momentum to increase international effort to achieve safer roads and modes of transportation. International cooperative guidelines such as the United Nations Sustainable Development Goals were created to increase global attention to this public safety issue [4]. Other global initiatives such as the Vision Zero project focuses on achieving a system with no fatalities or serious injuries involving road traffic, particularly in large economic cities in North America and Europe [1]. Of particular interest to this project and to this report, is its study and implementation in Canada's largest city, Toronto.

With around 3 million people moving around the city daily, Toronto's streets play an important role in the transportation of goods, services, and people [1]. Balancing the needs and safety of all road users is an exceptionally difficult task. Based on the Toronto Police Service Public Safety Data, the collision fatality count is the seventh largest among all cities in North America [5]. Many measures have been implemented to reduce road

traffic accidents. This includes multiple awareness campaigns centered on aggressive and drunk driving, numerous policies including reducing speed limits, stricter rules on passenger safety, distracted driving and many more [6]. Nevertheless, these existing solutions have not significantly improved the landscape of road safety in the city. Over the past three years, there has been an average increase of 22.5% per year in fatalities with minimal improvement in recent years [6]. Hence, this report aims to implement a data-driven approach using road infrastructure and environmental factors to predict the likelihood of an accident based on a set of features common in historical traffic accidents. The application of models and tools provided in this proposal will offer large potential to mitigate future damage, save lives, and more importantly, to make our roads safe for all.

2.0 Proposed Solution: Data-Driven Road Accident Prevention

This report details a data-driven approach to reducing road accidents by allowing users to make more informed decisions when navigating through the busy streets of their daily lives. In order to accomplish such a feat, a model that can reliably predict traffic accidents in real time is needed.

Historical traffic accident data was used to determine potentially dangerous locations through select features of the road and the surrounding environment. By identifying features associated with high risk areas, the output of this model can be extended to predict the safest optimal path between two points. Furthermore, these features can be used to inform local municipalities and governments about impending risks in the city for planning resource allocation and provide insight for infrastructure projects to mitigate the occurrence of traffic accidents in the future.

Taking the City of Toronto as a case study, the problem has been formulated into a binary classification problem. The model was designed to predict whether a traffic accident will occur based on the spatial and temporal features of a given road segment in the city.

The development of such model can be very challenging due to the following issues:

1) **Class imbalance**: Traffic accidents are rare occurrences and the negative samples greatly outnumber the positive ones.
2) **Spatial variation**: Different parts of the city are likely to have varying feature importance due to disparities in population density and/or land usage.
3) **Feature availability**: Human factors, such as distracted driving and inexperience, are some of the main causes of traffic accidents [7]. However, human behaviour cannot be realistically quantified and used in this application. This inherently limits the accuracy of the model as certain accidents will be caused by unpredictable human error.

The following sections will present the data pre-processing steps, model development strategy and model evaluation while addressing the challenges outlined above.

2.1 Datasets Description

All data sources were obtained through the Toronto Vision Zero challenge and the City of Toronto open data source.

- **Motor Vehicle Collision Events Dataset**: Basic information (e.g. time and date of accident, number involved, weather conditions, etc.) on all reported motor vehicle collisions for the City of Toronto from 2008 to 2017 [8].
- **Personnel Involved in Collision Dataset**: Additional information on the collision events, including actual speed of the vehicle, posted speed limit of the road, and more [8].
- **Traffic Volume Information Dataset**: Traffic volume count within 15-minute intervals at selected intersections and road segments from 1995 to 2017 [8].
- **Road Geometry Datasets**: Toronto Centerline, Posted Toronto Speed Limits, and the Toronto Centerline_Lanes datasets are used to obtain additional road geometry data [8].

The City of Toronto had over 486,000 reported traffic accidents in the last nine years as shown in Figure 1. However, only 10% of those will be used for this application due to missing data and computing power limitations. This resulted in a collision sample set of over 45,000 entries.

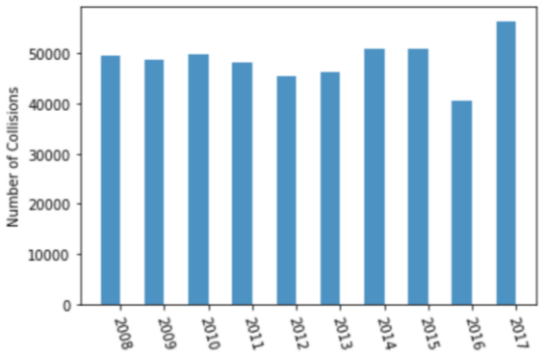

Figure 1: Traffic Collisions with k-clustering of 4

Figure 2: Traffic Collisions from 2008 to 2017

2.2 Feature Engineering and Negative Sampling

The following ten features were used as input for the model. The features were selected based on similar studies in literature and the availability of data [9-11]. The following features were identified as potential predictors of traffic accidents and they are visualized in Figure 3.

- **Temporal Factors**: month, weekday or weekend and hour
- **Weather Factors**: visibility and road surface condition
- **Spatial Factors**: average traffic volume per hour, posted speed limit, number of lanes, lane width and road class type

2.2.1 Spatial Variation: K-clustering

Though spatial variation can be captured to some degree by features such as population density and road geometry, traffic accident patterns still vary greatly for different locations within the city (i.e. downtown core versus residential areas). To account for this variation, k-mean clustering was used to divide the city into different regions. This cluster tag was then added as an extra feature to the model. An example with a k-means cluster number of 4 is shown in Figure 2.

2.2.2 Negative Sampling

Since all reported accidents are positive samples, negative samples were needed to build a binary classification model. With the assumption that all accidents were reported in the Motor Vehicle Collision Events Dataset, every single road segment and hour combination that did not result in an accident was a potential negative sample. This amounts to roughly 24.5 billion samples over the last nine years.

To avoid severe class imbalance, random sampling was used to extract a subset of the total negative samples. This subset was obtained by randomly selecting combinations of road segment, date and hour that did not result in an accident until a ratio of 3:1 for negative to positive samples is reached [11]. Road segment and time were chosen for this process as changing them can lead to the most diverse set of negative samples. The final data set contained a total of 201,000 examples.

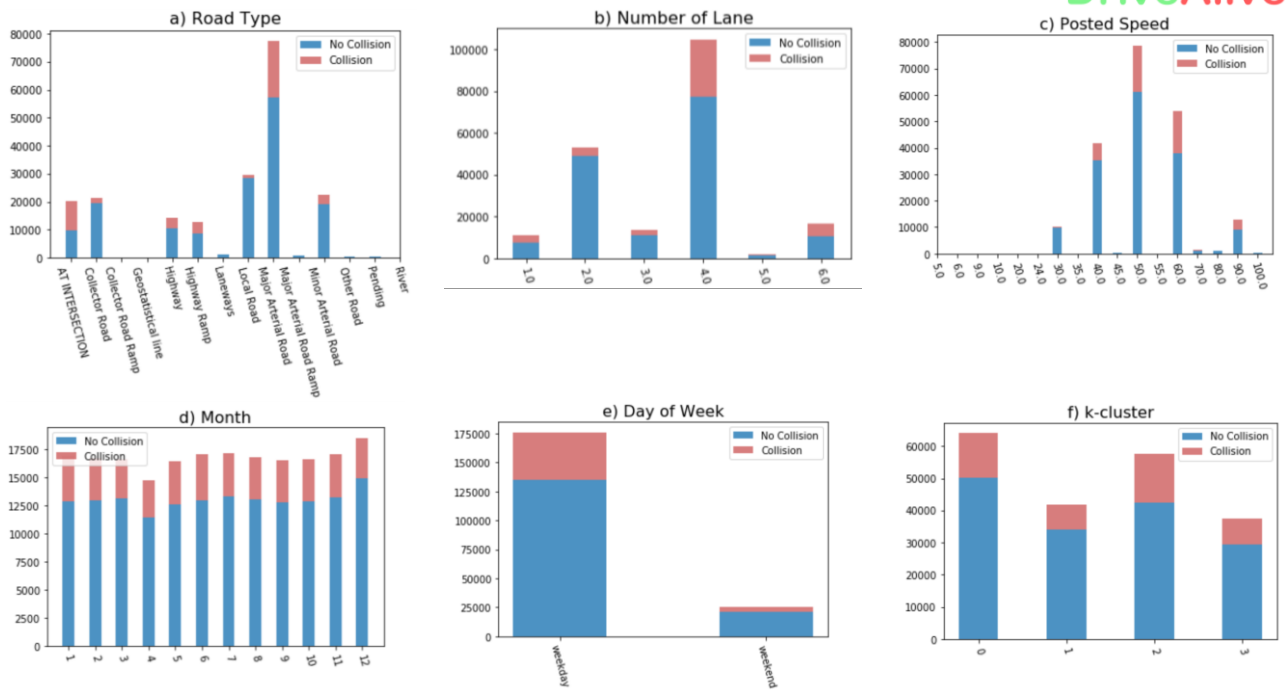

Figure 3: Visualization of traffic collisions vs collision related features

2.3 Classification Models and Results

Two supervised learning algorithms were compared for this problem, Random Forest (RF) and Deep Neural Networks (DNN). RF was chosen for its ease of implementation and resistance to overfitting for classification problems. On the other hand, DNN was chosen for its ability to learn and model non-linear and complex relationships. The results were cross-validated and the results are presented in Table 1.

	Clusters	Accuracy	Precision	Recall
RF	4	90.1	84.7	67.7
	20	90.4	85.5	68.2
	40	*90.6*	*86.3*	*68.9*
DNN	4	88.1	89.7	86.3
	20	89.1	89.9	85.2
	40	89.1	89.8	81.5

Table 1:Comparison of RF and DNN model

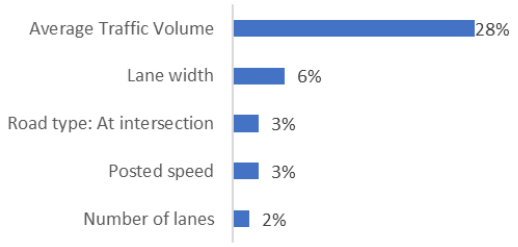

Figure 4: RF Variable Importance Rating

Based on results of Table 1, the two models are comparable. However, the DNN model is more computational expensive, and as such, the RF model with a k-cluster parameter of 40 was chosen. From Figure 4, traffic volume count and lane width are the most prevalent factors used to predict traffic accidents in the City of Toronto.

2.4 Planning Safer Routes

Using the above model, real time traffic accident prediction can be used to recommend routes with similar travel times. For example, each route can be discretized into smaller road segments where the model can predict the occurrence of an accident given the time, date, and weather conditions. The route with the minimum number of potential accidents

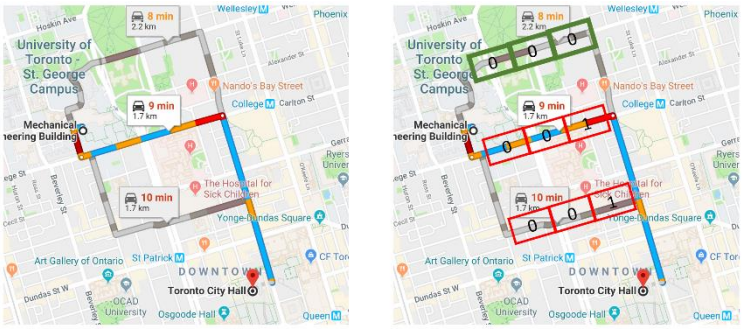

Figure 5. Safe Route Planning and Discretization Integrated on Google Maps (green indicates safe route option)

predicted will be recommended as the safest route as shown in green on Figure 5.

3.0 Business Plan and Market Strategy

DriveAlive can be added as an additional feature in mapping services to allow consumers to find routes to their destinations with the least likelihood of accidents. Mapping services were a USD 3 billion (CAD 4 billion) dollar industry in 2017, with almost 2 billion users worldwide, and is forecasted to grow to about USD 10 billion (CAD 13.2 billion) in 2020[12-14]. Being able to provide users with an additional useful service will therefore be the factor which affects users' choices of mapping service to use.

Toronto has a population of 2.7 million [5], with an additional 1.6 million people commuting by car, and 136,000 people carpooling into Toronto from the Greater Toronto Area [15]. Toronto also receives an average of 3.6 million tourists monthly [16]. Assuming an average carpool size of 2.5, and that 80% of drivers, 60% of the general population in Toronto, and 90% of tourists to Toronto use a mapping service to navigate within Toronto, approximately 3.1 million people use mapping services daily in Toronto. With 0.155% of total users of mapping services, Toronto is responsible for generating approximately USD 4.5 million (CAD 6 million) of revenue. As an additional feature, we expect DriveAlive to be responsible for approximately 2.5% of this revenue (CAD 100,000) annually.

However, the generation of monetary returns is not the only aspect of DriveAlive. By predicting where accidents are likely to occur, it is more likely for governments to improve infrastructure to reduce accidents, and for motorists to drive more carefully to avoid being involved in accidents. This would reduce both the social costs, such as from lost work days due to injuries, as well as healthcare costs on the individual and government. With the estimated average social costs per collision in Ontario to be about CAD 77,000 [17], and estimated healthcare costs per individual involved in a traffic accident to be approximately CAD 5,500 [18], this represents significant savings for the economy, government and individuals. Based on the current trend of traffic accidents in Toronto, there is still expected to be 50,000 to 65,000 accidents annually in the next 3 years [1]. Assuming that DriveAlive is able to prevent 25% of the accidents, it translates to about CAD 1 billion to CAD 1.2 billion in social costs and CAD 69 million to CAD 89 million in healthcare costs.

Considering that implementing DriveAlive in Toronto is relatively small scale, using an online cloud computing service such as Google Cloud Services will be sufficient [19]. It is estimated that 16 virtual CPUs with 208GB of memory will be capable for prediction of accidents on a continuous basis, while the model will be updated monthly with new traffic collision data using 96 CPUs and 624GB of memory for 10 hours. This adds up to a cost of USD 322 (CAD 427) monthly, or USD 3,900 (CAD 5,200) annually [19]. The total return on investment for implementing DriveAlive is approximately 210,000%, with a total benefit of CAD 1 billion at a cost of CAD 5,200 annually.

4.0 Conclusion

Despite numerous efforts, road traffic accidents remain a safety issue around the world. While traditional solutions employ policies and infrastructure improvements, these strategies require large amounts of resources and time. DriveAlive provides an alternative to reduce traffic accidents by identifying and quantifying high risk areas based on real-time road conditions and historical data using machine learning techniques. With integration with mapping services, DriveAlive can assist people with navigating through the busy streets by providing the safest optimal route around the city. DriveAlive can also assist local governments with infrastructure planning and resource allocation by identifying high risk segments that can be improved to mitigate the occurrence of future traffic accidents. Recent market studies suggest that an application to an economic hub such as Toronto has potential to generate CAD 100,000 in revenue with the use of existing mapping services and millions more in cost savings from healthcare, property damages, and from business due to reduced lost days from injuries. DriveAlive provides a simple, inexpensive, and robust solution capable of addressing a growing safety concern in a world moving towards smart cities and higher rates of urbanization.

Appendix - References

[1] Vision Zero Canada, "Vision Zero Canada", 2018. [Online]. Available: https://visionzero.ca/about/. Accessed: 22 November 2018.

[2] Toronto Police Service, "Traffic Collision Fatalities Dashboard", 2018. [Online]. Available: http://data.torontopolice.on.ca/pages/fatalities. Accessed: 22 November 2018.

[3] World Health Organization, "Global status report on road safety 2015", 2015. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. Accessed: 22 November 2018.

[4] United Nations, "Sustainable Development Goals", 2018. [Online]. Available: https://sustainabledevelopment.un.org/?menu=1300. Accessed: 22 November 2018.

[5] City of Toronto. Toronto at a Glance. Technical report, 2017. Accessed: 22 November 2018.

[6] City of Toronto, "2017-2021 Toronto's Road Safety Plan Vision Zero", 2017. [Online]. Available: https://www.toronto.ca/wp-content/uploads/2017/11/990f-2017-Vision-Zero-Road-Safety-Plan_June1.pdf. Accessed: 22 November 2018.

[7] J. Rolison, S. Regev, S. Moutari, and A.Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," *Accident Analysis & Prevention*, vol 115, pp 11-24, June 2018. Available: https://www.sciencedirect.com/science/article/pii/S0001457518300873. Accessed: 22 November 2018.

[8] City of Toronto, "VZ_Challenge". [Online]. Available: https://github.com/CityofToronto/vz_challenge. Accessed: 22 November 2018.

[9] T. Lu, Y.Lixin, Z. Dunyap and Z. Pan, "The traffic accident hotspot prediction: Based on the logistic regression method,"*The 3rd International Conference on Transporation Information and Safety*, June 2015. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7232194&tag=1. Accessed: 22 November 2018.

[10] C. Chen, "Analysis and Forecast of Traffic Accident Big Data," *ITM Web of Conferences*, vol 12, 2017. Available: https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf_ita2017_04029.pdf. Accessed: 22 November 2018.

[11] Z. Yuan, X. Zhou, T. Yang, K. Tamerius, and R. Mantilla, "Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study," *Proceedings of 6th International Workshop on Urban Computing*, August 2017. Available: http://urbcomp.ist.psu.edu/2017/papers/Predicting.pdf. Accessed: 22 November 2018.

[12] Avery Hartmans, "Here's how Google Maps could grow to be a $5 billion business by 2020". [Online]. Available: https://www.businessinsider.com/google-maps-could-be-a-5-billion-business-by-2020, 2017. Accessed: 22 November 2018.

[13] Felix Richter, "Google Maps is the Most-Used Smartphone App in the World". [Online]. Available: https://www.statista.com/chart/1345/top-10-smartphone-apps-in-q2-2013/, 2013. Accessed: 22 November 2018.

[14] Ben Popper, “Google announces over 2 billion monthly active devices on Android”. [Online]. Available: https://www.theverge.com/2017/5/17/15654454/android-reaches-2-billion-monthly-active-users, 2017. Accessed: 22 November 2018.

[15] Statistics Canada, “Commuters using sustainable transportation in census metropolitan areas. Technical report”, 2017. Accessed: 22 November 2018.

[16] Nick Westoll, “Toronto saw record number of visitors in 2017: tourism officials. [Online]. Available: https://globalnews.ca/news/3983802/toronto-tourism-2017-visitors/. Accessed: 22 November 2018.

[17] Keith Voddem, Douglas Smith, Frank Eaton, and Dan Mayhew, “Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario”. Technical report, Transport Canada, 2007. Accessed: 22 November 2018.

[18] Yu Qing Bai, Goncalo Santos, and Walter P. Wodchis, “Cost of Public Health Services for Ontario Residents Injured as a Result of a Motor Vehicle Accident”. Technical report, Health System Performance Research Network, 2016. Accessed: 22 November 2018.

[19] Google Cloud Services, “Google Cloud Pricing”. [Online]. Available: https://cloud.google.com/pricing/. Accessed: 22 November 2018.