LIZA BABAOGLU, BASc. Candidate[1]
E-mail: liza.babaoglu@mail.utoronto.ca
CENI BABAOGLU, Ph.D.[2]
E-mail: cenibabaoglu@ryerson.ca
[1] University of Toronto
44 St. George Street, Toronto, ON M5S 2E4, Canada
[2] The G. Raymond Chang School, Ryerson University
350 Victoria Street, Toronto, ON M5B 2K3, Canada

Safety and Security in Traffic
Original Scientific Paper
Submitted: 10 Jan. 2021
Accepted: 13 May 2021
DOI: 10.7307/ptt.v33i5.3782

PREDICTION OF FATALITIES IN VEHICLE COLLISIONS IN CANADA

ABSTRACT

Traffic collisions affect millions around the world and are the leading cause of death for children and young adults. Thus, Canada's road safety plan is to reduce collision injuries and fatalities with a vision of making the safest roads in the world. We aim to predict fatalities of collisions on Canadian roads, and to discover causation of fatalities through exploratory data analysis and machine learning techniques. We analyse the vehicle collisions from Canada's National Collision Database (1999–2017.) Through data mining methodologies, we investigate association rules and key contributing factors that lead to fatalities. Then, we propose two supervised learning classification models, Lasso Regression and XGBoost, to predict fatalities. Our analysis shows the deadliness of head-on collisions, especially in non-intersection areas with lacking traffic control systems. We also reveal that most collision fatalities occur in non-extreme weather and road conditions. Our prediction models show that the best classifier of fatalities is XGBoost with 83% accuracy. Its most important features are "collision configuration" and "used safety devices" elements, outnumbering attributes such as vehicle year, collision time, age, or sex of the individual. Our exploratory and predictive analysis reveal the importance of road design and traffic safety education.

KEYWORDS

fatality; collision; prediction; classification; data mining; road safety.

1. INTRODUCTION

Each year, traffic collisions kill approximately 1.35 million people around the world and are the leading cause of death for children and young adults [1]. In 2017, Canada's number of motor vehicle fatalities and injuries reached 1,841 and 154,886, respectively [2]. Meanwhile, the need for transportation steadily increases with the addition of more than half a million registered vehicles in Canada each year [3]. Therefore, collision analysis, prevention, and prediction deserve even greater importance.

The mission to prevent traffic injury requires correct road infrastructure and precise traffic controls, as well as educated users and safe vehicles. At the national level, Canada is implementing "Canada's Road Safety Strategy 2025" retaining the long term plan of making Canada's roads the safest in the world [4]. Also, more local road safety programs are being adopted through "Vision Zero", a road safety plan that aims to reduce traffic-related fatalities and serious injuries in cities, such as Edmonton [5] and Toronto [6].

We utilise Canada's National Collision Database (NCDB) to predict fatalities and to investigate patterns in collisions on Canadian roads. Our goals are (1) to discover associated rules and causations of fatal collisions (a collision with at least one fatality) through analysing collision data-elements; (2) to discover associated rules and causations of individuals fatalities through analysing all data-elements, including collision, vehicle, and personal data-elements; (3) to predict the individuals' fatal or non-fatal outcome by using all data-elements; and (4) to discuss the key contributing factors to the individuals' fatalities. Our findings show the importance of infrastructural design considerations, crash prevention and injury control systems, and public education.

This paper is organised as follows: in Section 2, the related background research is discussed; in Section 3, the data is described; in Section 4, the process and methodology is explained; in Section 5, our initial analysis is presented; in Section 6, results of our exploratory data analysis and predictions is presented; and lastly, in Section 7 and 8, the threats to validity and the discussion is showcased.

2. BACKGROUND

Analyses of collision databases are an ongoing process leading to numerous research worldwide. Research is undertaken in various countries with different datasets, each with unique research questions to address. For example, in the United States, the crash frequencies of Washington State are modelled given the collision and location type, the severity of the crash, and the number of vehicles involved [7]. Whereas in California, a focused study of the vehicle-by-vehicle crash data is conducted to extensively discuss the rear-end crashes [8]. In Japan [9], the vehicle-to-pedestrian-accidents data is used to predict the seriously injured body regions of pedestrians by considering various factors including the accident year, vehicle type, travel speed, and pedestrian gender and age. Meanwhile, the research in [10] used the data on road accidents with heavy-goods vehicles and buses for 27 European Union countries over 10 years and analysed safety parameters, such as area type, the season of the year, the weekday, casualty age and gender. Nevertheless, Colombia [11], Taiwan [12], and Serbia [13] separately examined spatial features such as road geometry and precipitation, as well as temporal attributes such as hour and day of the week, whereas India [14] analysed individuals' characteristics such as driving patterns and drunk driving to describe the traffic accidents and casualties.

In Canada, important research is conducted on severe collisions examining its causes and impacts. The research in [15] analysed all traffic collision events in which at least one person was killed or seriously injured in Toronto. The spatio-temporal and behavioural patterns are examined using the Killed or Seriously Injured dataset covering the years from 2007 to 2017. The results of this exploratory data analysis show the prevalence of collisions in intersections, in the spring and summer, as well as in the presence of aggressive and inattentive driving. On the other hand, the study on the Canadian National Population Health Survey for the years between 1994 and 2002, reported a higher percentage of subsequent injuries for binge drinkers, respondents with poor health, respondents with distress, and respondents using two or more medications [16]. The research in [17] provided a special focus on the back-over collisions with child pedestrians, utilising the Canadian Hospitals Injury Reporting and Prevention Program's injury dataset for the years from 1994 to 2003.

In the literature, there are some papers on this database, NCDB. One study [18], used two attributes, vehicle year and collision severity, and considered a subset of NCDB covering the years from 2001 to 2003. By combining these attributes with some external datasets, an association was found between older vehicles and mechanical failure as well as a higher rate of alcohol and drug use, unbelted occupants, and unlicensed drivers. Additionally, the research in [19] performed an analysis of the NCDB for the years between 1999 and 2012 to identify possible dangerous traffic scenarios that could result in injuries and fatalities. It reported a higher fatality rate in collisions involving streetcars driven by older drivers. Lastly, the research in [20] used a sub-dataset specialising in the survivability factors for the cyclists hit by motor vehicles, discussing the impacts of age, sex, helmet usage, and collision configuration. In our study, we consider all the available years (1999–2017) and utilise all the variables (collision, vehicle, and personal data-elements) presented in the NCDB. This allows us to extensively analyse and predict collisions in Canada.

3. DATA DESCRIPTION AND DATA PRE-PROCESSING

In this paper, we conduct analyses and predictions on all the police-reported vehicle collisions on public roads in Canada, from 1999 to 2017. Both the open-source database and the data dictionary are provided by Transport Canada at the Government of Canada's National Collision Database (NCDB) [21].

The NCDB dataset has 20 columns, excluding the ID columns, and 6,772,563 rows of observations, each representing a person involved in a collision. These reported observations have resulted from 2,570,235 collisions. The columns address collision, vehicle, and personal data-elements. Collision-related elements have temporal attributes including year, month, day of the week, and collision hour; spatial attributes including collision configuration, roadway configuration, number of vehicles involved in the collision, weather condition, road surface, and road alignment; as well as collision severity and traffic control attributes.

Vehicle-related elements contain vehicle type and model year, whereas personal-related elements contain the person's sex and age, their position in the collision, the road user class, safety devices used, and the individuals' injury severity. Our dependent variable is "individuals' injury severity". This has three classes: no injury, injury, and fatality. Fatality represents immediate death or death within 30 days of the crash, except in Quebec before 2007 (eight days) [2]. For the analysis of this paper, we merge no injury and injury classes into one class representing non-fatalities. We rename the dependent variable as "individual fatality."

In our initial analysis of the 19 years of data, from 1999 to 2017, we found that 98.41% of the collisions have resulted in no fatality, meanwhile, 1.59% have resulted in at least one fatality. Over the years, we observe a general decreasing trend in the number of collisions from 1999 to 2017, reaching its record minimum in 2017. Similarly, we observe a general decreasing trend in fatal collisions from 1999 to 2017, reaching its lowest and second-lowest values in 2015 and 2014, respectively. Then, we shifted our analyses from collisions to victims of these collisions, where *Figure 1* shows the decline in the number of victims over the years.

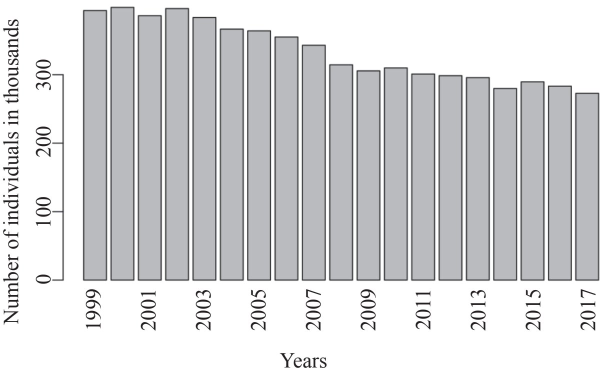

Figure 1 – The number of individuals involved in collisions over the years

For instance, over the 19 years, 43.30% of the people involved in the collisions had no injuries, whereas 55.97% were injured and 0.73% had died before having a chance to get any medical assistance. It is important to note that, the number of non-injuries and injuries has hit the lowest in 2017, with 118,199 and 152,772 individuals, respectively. Most importantly, 1,856 individuals were killed in a collision in 2017, showing a 37.72% decrease from 1999, as shown in *Figure 2*.

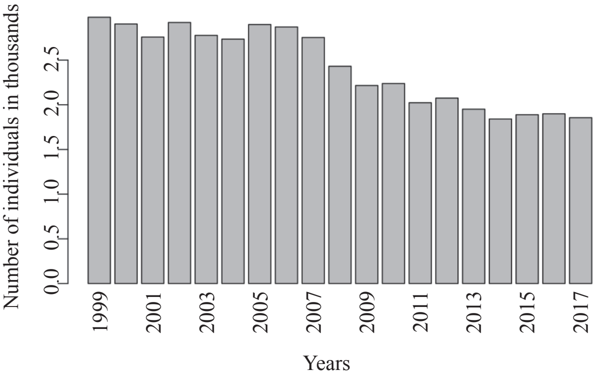

Figure 2 – The number of individuals killed in collisions over the years

We noted a gradual increase in the number of collisions from Monday to Friday, reaching its peak on Fridays. We found the highest frequency of collisions to occur between 3 pm and 6 pm. Interestingly, we observed the highest frequency in August, a summer month, and the least in February. This correctly matches our findings that most collisions occur in clear and sunny skies, dry and normal road surfaces, and straight and levelled roads, outnumbering collisions in harsher weather and road conditions. We, also, found that more than half of the collisions took place between two vehicles. In addition, we can conclude that most collisions have resulted in light-duty vehicles, such as passenger cars, vans, and pick-up trucks. Conversely, 77 vehicles, the highest number of vehicles, were reported in a collision on Friday, February 2006 and on Friday, January 2013. Of the 2,978,768 collisions, 1,001,610 collisions involved two cars traveling in the same direction, 1,006,752 collisions involved two cars traveling in the different directions, 930,766 collisions were with only a single vehicle in motion, and 39,640 collisions resulted by hitting a parked vehicle. However, we note head-on collisions as the deadliest of all configurations, because almost 9% of head-on collisions result in at least one fatality. Most of the collisions are reported to have taken place in non-intersections such as mid-block or at an intersection of at least two public roadways. It is important to note that only 0.07% of the collisions occurred while passing or climbing lanes, or in a freeway system. Most collisions occurred in areas of no traffic control, contributing to almost 60% of all collisions. Nevertheless, second and third most frequent collisions were reported where traffic signals were fully operating and a stop sign was in place, respectively. Similarly, we note that only 13% of the individuals involved in a collision were either not wearing a safety device or a child restraint, or there was no safety equipment, illustrating a bus setting. However, about 44% of such unsafe conditions resulted in individuals' instant death. Our analysis regarding the demographic of individuals showcases that 56% of individuals involved in a collision were male. Also, a significant number of fatalities have included individuals aged between 18 and 20 years, regardless of their lower death ratio from 60+ year-olds. On the other hand, although pedestrians are involved in less than 4% of all collisions, their fatalities make up more than 14% of all fatalities, signifying their vulnerability.

We conducted missing value imputations by mining specific story-based relationships between certain collision attributes. Firstly, within complete cases, we found that 96% of the rainy conditions had wet surfaces, allowing us to fill the unknown road surface values of all rainy conditions with wet surfaces. Then, similarly, 90% of the dry surfaces had clear and sunny skies, therefore we filled the unknown weather conditions of all dry surfaces with clear and sunny skies. Thirdly, 87% of the clear and sunny skies had dry road surfaces, therefore we filled the unknown road surface values of all clear and sunny conditions with dry surfaces. Additionally, pinpointing certain inconsistencies within personal and vehicle data-elements allowed us to correctly address more missing values. For example, by using the known positions of individuals in a vehicle, we imputed the unknown “road class user” attributes. We also imputed the missing values under the “number of vehicles” attribute with the maximum vehicle ID involved in that collision. We omitted 6.51% of the “individual fatality” attribute, which either had missing severity values or represented the hypothetical passengers in empty, parked cars. Lastly, we utilised Multivariate Imputation via Chained Equations (MICE) [22] to impute incomplete collision data-elements, where story-based imputations were not possible. As our method’s parameter, we selected Bayesian polytomous regression to impute our unordered categorical variables. Our pre-processed final dataset has 2,570,233 collisions, and 6,338,138 individuals involved in collisions. *Table 1* features the three most prevalent levels of each collision data element from a total of 2,570,233 collisions. *Table 2* and *Table 3* feature the three most prevalent levels of personal and vehicle data elements, respectively, from 6,338,138 observations.

Table 1 – Collision data elements

Variable	Most prevalent examples (quantity)
Collision year	2002 (156,415); 2000 (155,838); 2003 (152,980)
Collision month	August (233,382); July (231,301); December (231,171)
Collision day	Friday (433,487); Thursday (391,724); Wednesday (373,744)
Collision hour	16:00–18:00 (588,252); 13:00–15:00 (529,392); 10:00-12:00 (402,576)
Number of vehicles	2 (1,477,201); 1 (835,861); 3 (205,947)
Collision configuration	Rear-end collision (686,544); Right-angle collision (367,237); Other single vehicle collision (346,047)
Roadway configuration	At an intersection of at least two public roadways (1,235,694); Non-intersection (1,152,890); Intersection with parking lot entrance/exit, private driveway or laneway (131,160)
Weather	Clear and sunny (1,800,934); Overcast, cloudy but no precipitation (279,937); Raining (259,251)
Road surface	Dry, normal (1,737,494); Wet (467,530); Icy (169,805)
Road alignment	Straight and level (1,932,005); Straight with gradient (264,472); Curved and level (211,852)
Traffic control	No control present (1,514,287); Traffic signals fully operational (672,086); Stop sign (284,879)
Collision severity	Non-fatal (2,529,047); Fatal (41,186)

Table 2 – Personal data elements

Variable	Most prevalent examples (quantity)
Person's sex	Male (3455176); Female (2,760,739)
Person's age	18 (185,210); 19 (181,410); 20 (172,896)
Person's position during collision	Driver (4,228,869); Front row, right outboard, including motorcycle passenger in sidecar (1,000,314); Second row, right outboard (295,000)
Safety device used	Safety device used or child restraint used (4,803,467); No safety device equipped, such as buses (518,534); No safety device used or no child restraint used (221,414)
Road class user	Motor vehicle driver (3,996,636); Motor vehicle passenger (1,790,185); Pedestrian (244,830)
Individual fatality	Non-fatal (6,292,122); Fatal (46,016)

Table 3 – Vehicle data elements

Variable	Most prevalent examples (quantity)
Vehicle type	Light duty vehicle (5,221,555); Pedestrian (250,800); Other trucks and vans (184,556)
Vehicle model year	2000 (318,730); 2002 (302,609); 2003 (296,915)

4. METHODOLOGY

In our initial analysis, we conducted univariate, bivariate, and multivariate statistical methods. In our exploratory analysis, we used a classical data mining methodology, the a priori algorithm, to find frequent subsets and association rules in the dataset. We conducted this algorithm firstly, to find subsets in collision data-elements that lead to a fatal collision, and secondly, to find subsets in all data-elements that lead to a person's death. We list our findings with the highest lift values. Lift represents the strength of the association rules since it describes the likelihood of the outcome given a combination of dependent variables, while accounting for the popularity of the variables in the dataset.

Before creating our prediction models, we split our dataset into a 70% training set and a 30% test set. Our goal is to perform binary classification within the "individual fatality" attribute: fatality class vs. non-injury/injury class. We proposed the following three methods to treat these imbalanced classes in our training set. With undersampling, we selected a proportion of the observations from the majority class to create a balanced dataset. With Random Over-Sampling Examples (ROSE) [23], we populated the minority class by creating synthetic examples. As a third method, we used the combination of both under- and over-sampling methodologies [23].

We trained our training set by using two supervised learning classification algorithms: Lasso Regression (Least Absolute Shrinkage and Selection Operator) [24] and XGBoost (eXtreme Gradient Boosting) [25]. Lasso Regression is a special type of linear regression. It adds a penalty equivalent to the absolute magnitude of regression coefficients and tries to minimise them by performing variable selection, and regularisation to prevent overfitting by discouraging building complex, flexible models. We used Lasso Regression on a five-times cross-validated training set. Lasso Regression is given by the equation,

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{1}$$

where λ is the tuning parameter denoting the amount of shrinkage.

As λ increases, more variable coefficients approach to zero and get eliminated [24]. On the other hand, XGBoost is an ensemble learning technique that implements the gradient boosting on the decision tree learning algorithm. XGBoost aims to minimise the loss function,

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t}\Omega(f_k) \tag{2}$$

where

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\boldsymbol{w}_j^2 \tag{3}$$

Here, γ represents the complexity of each leaf, T represents the number of leaves, λ is a penalty scaling parameter, and $\boldsymbol{w}$ represents the vector of scores on leaves [25]. This boosting algorithm minimises error in sequential models through improving from the shortcomings of previous iterations. It conducts regularisation, and parallel computation to compute faster, allowing us to iterate our boosting algorithm 1200 times. Our model is based on a learning rate (eta) of 0.5 and a maximum depth of three. We used the objective of "binary:logistic" for our binary classification. We chose these supervised learning models, due to their promising performance with our large and sparse dataset. We evaluated our models on the test set with accuracy, sensitivity, and specificity measures. Accuracy is the ratio of the correctly predicted outcomes over the total number of samples in the test set. Sensitivity is defined as the true positive rate. In our dataset, it represents the ratio of the correctly identified fatalities to all fatality samples. Specificity is defined as the true negative rate, where it represents the ratio of correctly identified non-fatalities to all non-fatality samples.

5. RESULTS

5.1 Exploratory data analysis

The results for our exploratory analysis are shown in *Tables 4 and 5* with a minimum support value of 0.01 and a minimum confidence value of 0.05. The tables display higher lift values showcasing the importance and strength of the association rules on their respective outcomes: collision severity and individual fatality. *Table 4* shows the most significant combinations of collision data-elements that result in fatal collisions. The top six lift values range from 8.59 to 7.03 as shown below, whereas the minimum lift value observed in the association rules for collision severity is 3.15. Some collisions that result in at least one fatality have a mutual trait which is seen in every single rule, that is, head-on collisions. Head-on collisions are defined as two vehicles traveling in different directions. Specifically, head-on collisions at non-intersected road configurations with dry road surfaces are likely to result in fatal collisions (*Rule1*). Similarly, it is also probable for fatalities to occur during head-on collisions at non-intersected road configurations

Table 4 – Rules for collision severity

{Left-hand side} => {CollisionSeverity = At least one fatality}	Lift	Count
Rule1 {CollisionConfiguration = Head-on, RoadConfiguration = Non-intersection, RoadSurface = Dry/Normal} => {C_SEV = 1}	8.59	3,853
Rule2 {CollisionConfiguration = Head-on, RoadSurface = Dry/Normal, TrafficControl = No control present} => {C_SEV = 1}	7.77	4,205
Rule3 {CollisionConfiguration = Head-on, RoadConfiguration = Non-intersection, Weather = Clear/Sunny} => {C_SEV = 1}	7.67	3,830
Rule4 {CollisionConfiguration = Head-on, RoadConfiguration = Non-intersection, TrafficControl = No control present} => {C_SEV = 1}	7.12	6,615
Rule5 {VehiclesInvolved = Two, CollisionConfiguration = Head-on, RoadConfiguration = Non-intersection} => {C_SEV = 1}	7.04	5,630
Rule6 {CollisionConfiguration = Head-on, RoadConfiguration = Non-intersection} => {C_SEV=1}	7.03	6,699

Table 5 – Rules for individual fatality

{Left-hand side} => {IndividualFatality = Fatal}	Lift	Count
Rule1 {SafetyDevices = None, UserClass = Motor vehicle driver} => {P_ISEV = 1}	10.65	7,086
Rule2 {VehiclesInvolved = One, SafetyDevices = None} => {P_ISEV = 1}	9.87	7,365
Rule3 {RoadConfiguration = Non-intersection, SafetyDevices = None} => {P_ISEV = 1}	9.79	8,337
Rule4 {Passenger = Driver, SafetyDevices = None} => {P_ISEV = 1}	9.16	7,379
Rule5 {TrafficControl = No control present, SafetyDevices = None} => {P_ISEV = 1}	9.03	9,849
Rule6 {VehicleType = Light duty vehicles, SafetyDevices = None} => {P_ISEV = 1}	8.66	8,837

with clear skies (*Rule3*). In fact, the combination of only two attributes, head-on collisions, and non-intersection, are dominant enough to result in fatal collisions (*Rule6*). It is also important to note the lack of traffic control present, such as the lack of warning signs, flashing traffic signals, or road marking, at intersected or not intersected areas may lead to fatal head-on collisions (*Rule2*, *Rule4*).

In *Table 5*, the top six lift values range from 10.65 to 8.66 as shown below, whereas the minimum lift value observed in the association rules for individual fatality is 6.94. All data-elements, including collision, vehicle, and personal elements, are considered to find the most significant patterns leading to sudden fatal injuries. It is crucial to note that some individuals that died immediately during the crash or within the time limit had no safety device or child restraint, as shown in all the six rules. Other significant rules were the combinations of a lack of personal safety with one of the followings: a motor vehicle driver (*Rule1*, *Rule4*), a one-vehicle collision (*Rule2*), at a non-intersection area (*Rule3*), no traffic control present (*Rule5*), and light-duty vehicles (*Rule6*). In fact, of the top six lift values presented, the highest number of fatal injuries occur when the roads have no traffic control and the individual has no safety devices (*Rule5*).

5.2 Prediction models

The results of our chosen predictions models are shown in *Table 6* which illustrates the balancing sampling method, prediction model, and performance measures. The accuracy results of our Lasso and XGBoost training models ranged between 81% and 89%, indicating good predictions, with no over-fittings in the test set. We examine that XGBoost performs slightly better than Lasso, in all scenarios including undersampling, oversampling, or combined-sampling. In fact, the best model is a balancing method using combined-sampling followed by an XGBoost training model. This model's predictions on the test set display 83% accuracy, 79% sensitivity, and 83% specificity.

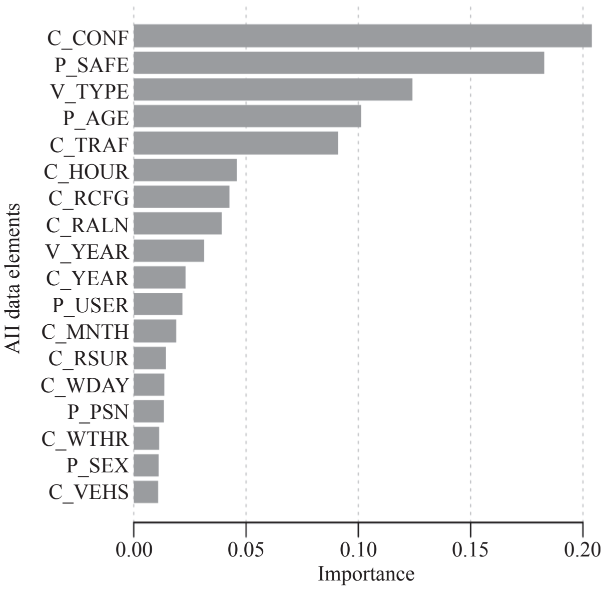

Figure 3 – Feature importance graph of our best performing model, XGBoost

This XGBoost model's most important features are "collision configuration" and "used safety devices" elements, in which each variable has a contribution importance of more than 15% in the model, as seen in *Figure 3*. On the other hand, the least important factors are the "number of vehicles involved" and "person's sex", with each having less than 3% variable contribution importance in predicting the fatality of a person involved in a collision.

6. DISCUSSION AND CONCLUSION

This paper analyses and predicts traffic-related fatalities in Canada through initial analyses, data mining techniques (association rules), and classification models (Lasso Regression and XGBoost).

From our initial analysis, we observe that most collision fatalities took place during the day with clear sunny skies, and straight dry road surfaces, elucidating the large number of collisions in non-extreme conditions. We think that the relatively low collisions in extreme conditions, such as unsafe road curvature, nights, or harsh Canadian weather, might be due to the driver's cautiousness in high perceived risk areas. However, we cannot reach this conclusion in full accuracy, since our dataset does

Table 6 – Prediction results on the test set of individual fatality

Sample	Model	Accuracy	Sensitivity	Specificity
Combined: Under & Over	Lasso	0.81	0.79	0.81
	XGBoost	0.83	0.79	0.83

not include drivers' behaviour-related attributes such as inattentiveness, aggressive driving, high-speed driving, or multitasking.

Based on our results from the data mining algorithms, the leading cause of individual fatalities is the absence of safety devices, such as child restraints, seat belts, helmets, and reflective clothing. We, also, observe that the absence of traffic controls, such as traffic signals, stop signs, yield signs, police officers, school crossings, and railway crossings, is a crucial factor for individual fatalities (*Table 5*). Specifically, the lack of both traffic controls and safety devices results in fatalities, as seen in Rule5 of *Table 5* with the highest count value among the highest lift values. Additionally, as shown in *Table 4*, the prominent factor in fatal collisions is head-on collisions (*Table 4*). Especially, the co-occurrence of head-on collisions and non-intersected areas, as well as the co-occurrence of head-on collisions in areas of no traffic controls lead to collision fatalities, as seen in every rule in *Table 4*.

As our analysis suggests, implementing correct and frequent traffic controls in non-intersection areas may decrease the number of head-on collisions. Also, the use of safety devices in motor vehicles and cyclists may decrease the risk of individual fatality. Furthermore, spreading awareness among motor vehicle drivers, bicyclists, and motorcyclists aged between 16 and 29 is necessary to prevent fatalities, since they are involved in the majority of fatal collisions (*Table 2*).

The predictions conducted by Lasso Regression and XGBoost show promising results. In addition, we conducted our predictions with multiple sample sets and numerous iterations that have proven the stability of our models. However, XGBoost performed slightly better suggesting the consideration of this model for similar datasets. The key features of this prediction model are the "collision configuration" and "safety devices used" attributes.

Missing and inconsistent data records were a threat to internal validity. We mitigated this threat through story-based and algorithm-based imputations. Whereas including the "collision severity" attribute in the prediction of "individual fatality" would result in an external threat to validity, therefore we excluded the collision severity attribute prior to running our prediction models. Lastly, the imbalanced class of our dependent variable could result in a threat to construct validity. We resolved this threat by using balanced sample sets for the training of our binary classification models.

In the future, we plan to work on traffic-related datasets with drivers' behaviour-related attributes from North America. This will allow us to explore different causations of fatalities and to attain a more generalisable model.

DATA AVAILABILITY

The open-source National Collision Database and its data dictionary are available at https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a.

REFERENCES

[1] *Global status report on road safety 2018*. World Health Organisation; 2018. Available from: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/ [Accessed 7th Jan. 2021].

[2] *Canadian Motor Vehicle Traffic Collision Statistics: 2017*. Transport Canada; 2017. Available from: https://tc.canada.ca/en/road-transportation/motor-vehicle-safety/canadian-motor-vehicle-traffic-collision-statistics-2017 [Accessed 7th Jan. 2021].

[3] *Vehicle registrations, by type of vehicle*. Statistics Canada; 2020. Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310006701 [Accessed 7th Jan. 2021].

[4] *Canada's road safety strategy 2025*. Canadian Council of Motor Transport Administrators; 2016. Available from: http://roadsafetystrategy.ca/files/RSS-2025-Report-January-2016-with%20cover.pdf [Accessed 7th Jan. 2021].

[5] *2018 Annual vision zero report*. City of Edmonton; 2018. Available from: https://www.edmonton.ca/transportation/PDF/2018_VisionZero-EdmontonAnnualReport.pdf [Accessed 7th Jan. 2021].

[6] Transportation Services City of Toronto. *Vision zero Toronto's road safety plan*; 2017. Available at: https://www.toronto.ca/wp-content/uploads/2017/11/990f-2017-Vision-Zero-Road-Safety-Plan_June1.pdf [Accessed 7th Jan. 2021].

[7] Venkataraman N, Ulfarsson GF, Shankar VN. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis & Prevention*. 2013;59: 309-318. DOI: 10.1016/j.aap.2013.06.021

[8] Ahmadi A, Jahangiri A, Berardi V, Machiani SG. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety & Security*. 2020;12(4): 522-546. DOI: 10.1080/19439962.2018.1505793

[9] Oikawa S, Matsui Y. Features of serious pedestrian injuries in vehicle-to-pedestrian accidents in Japan. *International Journal of Crashworthiness*. 2017;22(2): 202-213. DOI: 10.1080/13588265.2016.1244230

[10] Evgenikos P, et al. Characteristics and causes of heavy goods vehicles and buses accidents in Europe. *Transportation*

Research Procedia. 2016;14: 2158-2167. DOI: 10.1016/j.trpro.2016.05.231

[11] Gutierrez-Osorio C, Pedraza CA, Characterizing road accidents in urban areas of Bogota (Colombia): A data science approach. In: *2019 2nd Latin American Conference on Intelligent Transportation Systems, 19 March 2019, Bogota, Colombia*. IEEE; 2019. p. 1-6. DOI: 10.1109/ITSLATAM.2019.8721334

[12] Pai CW, Lin HY, Tsai SH, Chen PL. Comparison of traffic-injury related hospitalisation between bicyclists and motorcyclists in Taiwan. *PLoS One*. 2018;13(1): e0191221. DOI: 10.1371/journal.pone.0191221

[13] Novkovic M, et al. Data science applied to extract insights from data-weather data influence on traffic accidents. *INFOTEH-JAHORINA*. 2017;16: 387-392.

[14] Gaurav, Alam Z. Improving Road Safety in India Using Data Mining Techniques. In: Panda B, Sharma S, Roy N. (eds) *Data Science and Analytics. REDSET 2017. Communications in Computer and Information Science, vol 799*. Springer, Singapore; 2018. p. 187-194. DOI: 10.1007/978-981-10-8527-7_17

[15] Shanshal D, Babaoglu C, Başar A. Prediction of Fatal and Major Injury of Drivers, Cyclists, and Pedestrians in Collisions. *Promet – Traffic&Transportation*. 2020;32(1): 39-53. DOI: 10.7307/ptt.v32i1.3134

[16] Vingilis E, Wilk P. Predictors of motor vehicle collision injuries among a nationally representative sample of Canadians. *Traffic Injury Prevention*. 2007;8(4): 411-418. DOI: 10.1080/15389580701626202

[17] Nhan C, Rothman L, Slater M, Howard A. Back-over collisions in child pedestrians from the Canadian Hospitals Injury Reporting and Prevention Program. *Traffic Injury Prevention*. 2009;10(4): 350-353. DOI: 10.1080/15389580902995166

[18] Lécuyer JF, Chouinard A. Study on the effect of vehicle age and the importation of vehicles 15 years and older on the number of fatalities, serious injuries and collisions in Canada. In: *Proceedings of the Canadian Multidisciplinary Road Safety Conference XVI*; 11 June 2006.

[19] Watkins E, Kloc M, Weerasuriya S, El-Hajj M. Collision analysis of driving scenarios. *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 9-11 Jan. 2017, Las Vegas, NV, USA*; 2017. p. 1-7. DOI: 10.1109/CCWC.2017.7868413

[20] Demers S. Survivability factors for Canadian cyclists hit by motor vehicles. *Journal of Community Safety & Well-being*. 2018;3(2): 27-3. DOI: 10.35502/jcswb.66

[21] Government of Canada. National Collision Database [database]; 2019. Available from: https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a [Accessed 7th Jan. 2021].

[22] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work?. *International Journal of Methods in Psychiatric Research*. 2011;20(1): 40-9. DOI: 10.1002/mpr.329

[23] Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. *R Journal*. 2014;6(1): 79-89. Available from: https://www.mclibre.org/descargar/docs/revistas/the-r-journal/the-r-journal-11-en-201406.pdf#page=79 [Accessed 8th Jan. 2020].

[24] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1): 267-88. DOI: 10.1111/j.2517-6161.1996.tb02080.x

[25] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 Aug 2016, San Francisco, California, USA*; 2016. p. 785-794. DOI: 10.1145/2939672.2939785