UNIVERSITY OF WATERLOO

Department of Management Sciences

MSCI 446 – Data Mining and Warehousing

Data Study on Severity of Vehicle Collisions in Canada, 2017

Prepared by:

Ruhollah Nasim

Aaron Dai

Table of Contents

Abstract

This project aims to explore at a high-level the potential features that affect fatality rates in car collisions. This was achieved using a dataset of 2017 car collisions in Canada and analyzed through a preliminary data exploration process and two machine learning models including a logistic regression model and association mining model using the Apriori algorithm. Overall, the team encountered issues with the imbalanced dataset that made it difficult for the models to extrapolate results regarding the severity of a car collision. As a result, the team was able to understand the importance in managing an imbalanced dataset and how to approach the results from these models due to this issue.

Introduction

Millions of vehicle collisions occur each year worldwide costing billions of dollars in property damage and healthcare. While only a fraction of collisions lead to death, there are still upwards of 1.35 million deaths annually as reported by WHO (2018). From these statistics, the team became interested in identifying relationships between a variety of factors on the fatality rate of vehicle collisions. By understanding factors that have a larger presence in affect fatality rate, it may affect how decisions are made by organizations involved in this industry such as automobile insurance companies on cost rates, or governments on road planning and allocation of resources to reduce fatality rates. The overall goal is to obtain a high level understanding of a variety of features on severity of vehicle collisions using machine learning models and promote improvements of potentially future studies on this topic through lessons learned in this project. The response variable of interest will be a binary severity code indicating whether a person involved in a vehicle collision resulted in a fatality or did not result in a fatality. The full dataset directory and notations are shown in appendix A. This project implemented a logistic regression model to demonstrate the application of supervised learning on this dataset and lessons learned. Logistic regression was an appropriate model to implement due to the binary nature of the response variable which would be fitted to the model. As well, an association rule model was implemented using apriori algorithm to demonstrate the application of unsupervised learning on this dataset. Association is an appropriate model to implement because nearly all features in this dataset are categorical variables. The goal would be to identify correlations amongst sets of features that may lead to interesting insights on the topic. Other models such as clustering would

not be the best choice for categorical variables since conditions within each category do not have any special relationships between each other. While it could be possible to perform clustering of features via the response variable as a medium, it may not be accurate because there is no guarantee that the feature is correlated with the response variable. Overall, the project highlighted the issues with the models on the dataset and suggested improvements for future studies building from the results of this study.

Related Work

Thousands of articles exist on the topic on car collisions however generally most studies tend to focus on one or only a few features. For instance, Abdel-Aty & Abdelwahab (2004) studied traffic fatalities from angle collisions with a focus on vehicle configuration and compatibility. This study also used traffic from the United States and between 1975-2000. Results from this study would be difficult to compare since the intention of this project is an overall view of fatalities by a variety of factors. A study by Assi et al.(2020) predicted crash severity using neural networks as opposed to the logistic regression and association models used in this project. The dataset also used in this study was from the U.K. as opposed to Canada. An interesting note from this study was the limitation of the imbalance dataset that was also encountered in this project that will be discussed. The study used a simple randomized class balancing procedure however it acknowledged more advanced approaches could have been utilized to improve the model accuracy. Although this project did not improve on this issue, it helped to demonstrate the effect of an imbalanced dataset on modeling and that in future projects working with imbalance dataset should place a heavy emphasis on using or developing approaches to work around the issue. Another study by Assi (2020) analyzed collision severity using 16 different features that are similar to the ones to be used in this project. The study however used a dataset from crashes in Victoria, Australia over 5 years and applied neural network models. While the dataset and models are different, having similar parameters provides a basis to compare the models in this project to the study's model. Comparing the confusion matrices in the study to the ones generated in this project, the neural network models were able to predict fatalities accurately much more than the logistic regression models. This suggests that potentially logistic regression models may not be an appropriate choice to model car collision

severity however future studies can perform more in-depth and resource intensive testing to ensure this statement is demonstrated with more evidence.

Data

The dataset used in this project is a list of all police-reported motor vehicle collisions on public roads in Canada (2019). The 2017 dataset was used in the model building; it contains 289841 rows of collision reports that each represents the status of the collision and individual involved. When referring to a vehicle collision - it may involve multiple individuals and in the dataset, multiple rows of data can refer to a single case. As such, the definition of a collision in this report refers strictly to an individual involved in a collision. There are a total 22 different variables presented in a report which will be explored and discussed using the 2017 dataset. (Refer to Appendix A for the full coding of each variable).

The interest of the report is exploring the features that affect the fatality rate of individuals involved in a collision. As such, the response variable will be P_ISEV from the dataset. This variable represents the medical treatment required by an individual as seen in appendix A. Rows of data were removed if the code was non-numeric representing non-applicable, unknown, and missing data. Since the interest lies in individuals, data involving non-applicable entities such as park cars or run-away cars should not be considered in modeling. After removing these entries, a total of 272827 entries remained in the dataset. The numeric values of P_ISEV were then coded into a binary variable to represent an individual's status after a collision where 0 = no fatality and 1 = fatality. As seen in figure 1 below, a significant number of collisions resulted in no fatalities. As such, data visualizations in this section will be presented without comparing the actual count of non-fatalities to fatalities due to the skew in data.

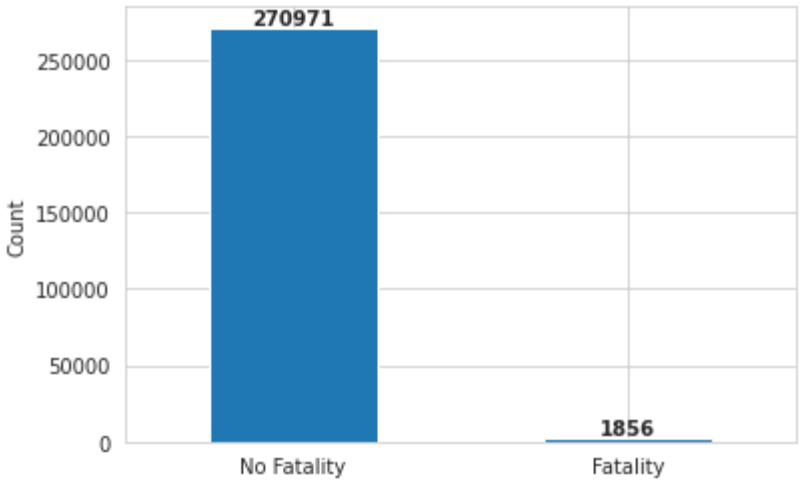

Figure 1: Fatality Count of 2017 Collisions

Figures will consist of two graphs; the left represents the fatality rate of each category (as calculated by the number of fatalities divided by the number of collisions that occurred in that category), while the right shows the fatality counts in each category. This method allows us to visualize variances in the fatality without being obstructed by a skewed y-axis range. The left graph presents the severity of a condition using rate and is necessary to draw initial insights because relying purely on the count can be deceiving. For instance, inspecting weather conditions - clear and sunny has the highest number of fatalities however that is because many collisions occur in that condition. When comparing the rates of each weather condition, the team discovers that visibility limitation has a much higher impact. The 21 other variables will be discussed as follows:

1) Year: This column represents the year of collision. The database contains reports from 1999 until 2017. Older reports will not be used in modeling because the number of fatalities will be skewed towards older data since vehicular safety and road safety was not as developed. As well, the data would also not be as relevant to draw insights from about predicting current day collisions. Only a single dataset is used because there are over 200,000 rows of data to model from and it is the most recent one.

2) Month: This column represents the month of collison. As seen in figure 2, both the distribution of fatalities and fatality rate are higher in the latter half of the year. The highest conditions being July, August, and October. This is a good candidate feature for modelling because there is variance in the data and certain conditions seem to have higher impact on fatality.

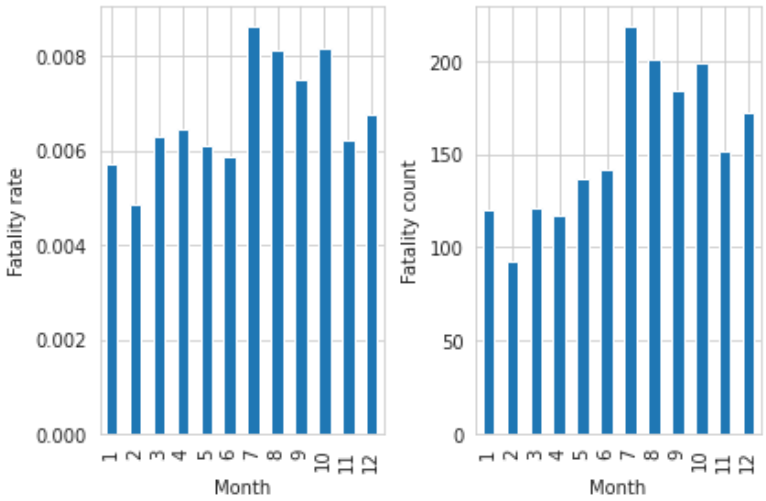

Figure 2: Fatality Histogram By Month

3) Day of week: This column represents the day of week in which collisions occurred. As seen in figure 3, both the distribution of fatalities and fatality rate are higher towards the end of the week where Friday had the highest number of fatalities while Sunday had the highest fatality rate. This is not included in the models because it does not add value to the project objective goals.

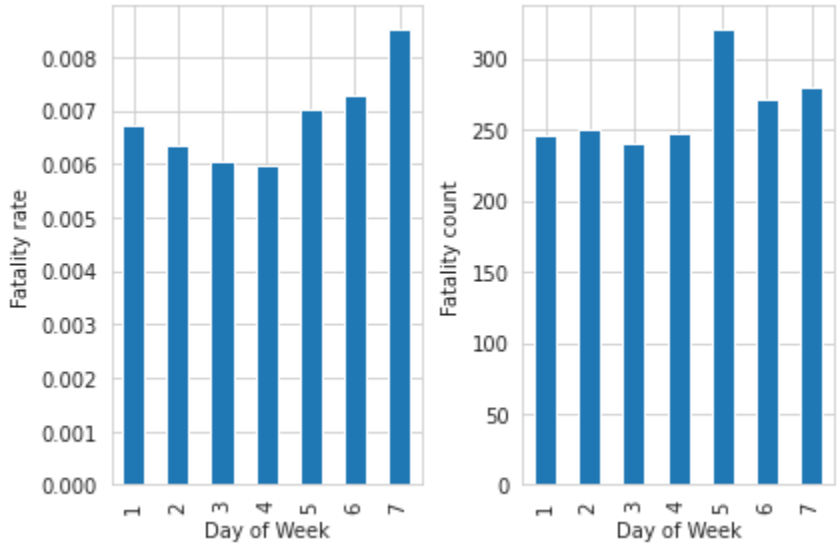

Figure 3: Fatality histogram by day of week

4) Collision hour: This column represents the time of day in which collisions occurred. As seen in figure 4, There are a higher number of fatalities in the afternoon but fatality rate is higher during night time. The team speculates that because there are more people driving during the

day-time which increases the number of accidents but since visibility is higher during the day, drivers may be better at reacting and adjusting to reduce the severity of a collision. There is a fair spread of variance across the conditions however this variable will not be explored in depth since these conditions may change in relation to time of year and thus results would be difficult to interpret when month is already involved in the model.

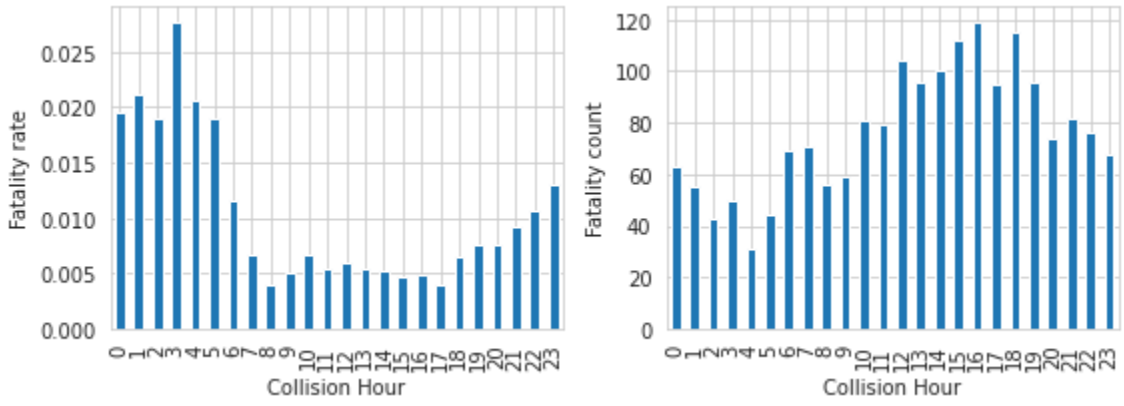

Figure 4: Fatality histogram by Time of Day

5) Collision Severity: This column represents if a collision involves at least 1 fatality or not. This column set will be dropped from the models because it represents fatality on a case basis rather than an individual basis which is the interest of the team.

6) Number of vehicles involved in collision: Similarly to collision severity, this column set will be dropped from the models because the data is presented on a collision case basis rather than an individual basis.

7) Collision configuration: This column represents how the vehicle collision occurred (e.g. code21 = rear-end collision, code 31 = head-on collision). As seen in figure 5, the highest number of fatalities and fatality rate occurred during head-on collisions (code 31) with a spread of fatalities across other conditions. In the database coding of these conditions, configurations are grouped into four categories - single vehicle in motion, two vehicles in motion - same direction of travel, two vehicles in motion - different direction of travel, and two vehicles - hit a parked motor vehicle. The team decided not to group and transform the data in these categories because information is lost when grouping the data. For instance, code 31 (head-on collision) shows the highest fatality count and rate but other conditions in the same category (i.e. code 32,33,34,35,36) does not show high values in comparison. Therefore, this condition would have skewed the data in the grouped and comparisons versus other groups. Choosing not to group the

data will improve model accuracy and information is not lost at the cost of a more complex model and additional work in interpretation.

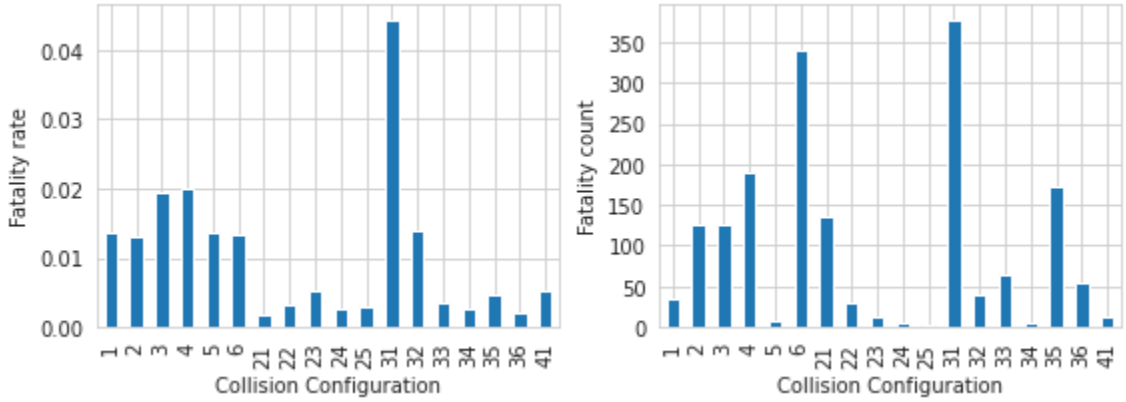

Figure 5: Fatality Histogram by Collision Configuration

8) Roadway configuration: This column represents the type of roadway the collision occurred on (e.g. code 4 represents railroad level crossing). As seen in figure 6, the highest number of fatalities occurred on non-intersections (code 1) while the highest fatality rate occurred on railroad level crossings (code 4). The distributions across the graphs suggest this as a candidate feature to look into since there are differences between the conditions.

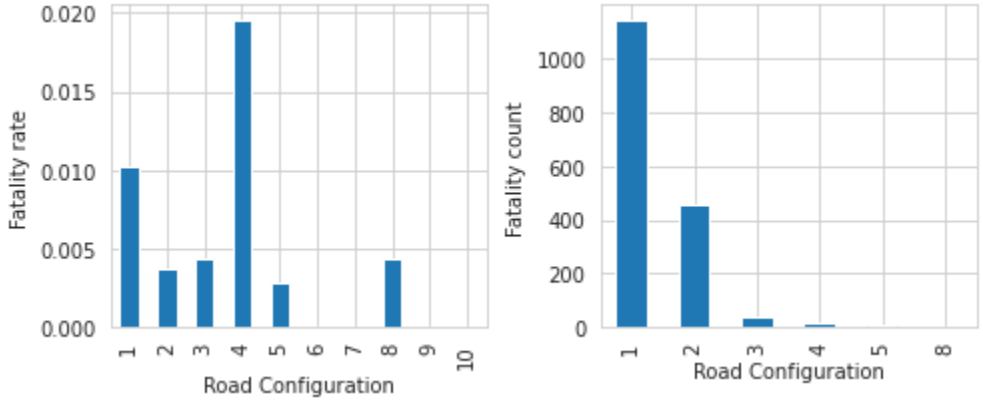

Figure 6: Fatality Histogram by Road Configuration

9) Weather condition: As seen in figure 7, the highest number of fatalities occurred in clear and sunny weather (code 1) but the highest fatality rate occurred in limited visibility conditions (code 6, e.g. fog, smog, dust, smoke, etc.). The team speculates that there are higher fatality counts in clear and sunny weather because there are more often drivers driving in these favourable conditions however when collisions occur, other weather conditions are more likely to have a fatality because they more negatively affect a driver's ability and vehicle control. The

distribution of this data suggests this as a candidate feature in the model due to differences across the conditions.

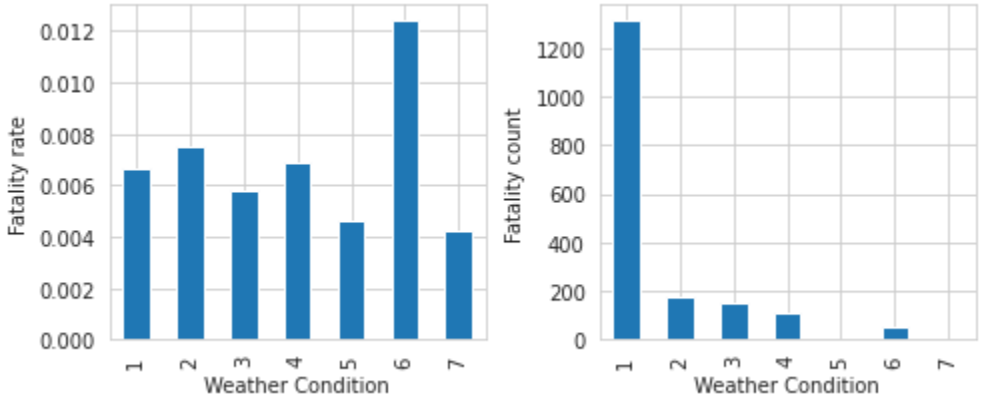

Figure 7: Fatality Histogram by Road Configuration

10) Road surface: This column represents the road surface condition that the collision occurred on. As seen in figure 8, the highest number of fatalities occurred in dry and normal roads (code 1) but the highest fatality rate occurred on muddy road conditions (code 7). Similar to weather conditions, the team speculates higher counts in dry, normal roads because more drivers drive in these conditions but muddy roads have a bigger impact on fatality.

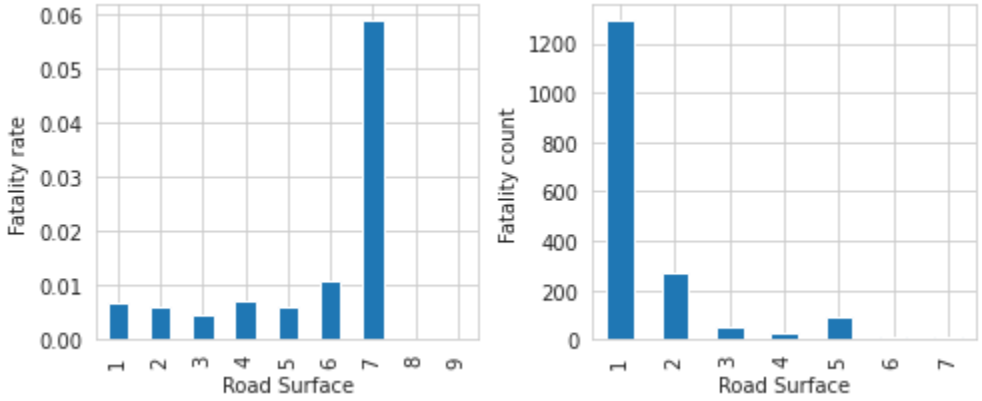

Figure 8: Fatality Histogram by Road Surface

11) Road Alignment: This column refers to the road grade that a collision occurred on. As seen in figure 9, the highest number of fatalities occurred on straight and level roads (code 1) but the highest fatality rates occurred on curved and level roads (code 3) and curved with gradient roads (code 4). The same speculations can be made here as in part 9 and 10. Similarly, this is also a candidate feature.

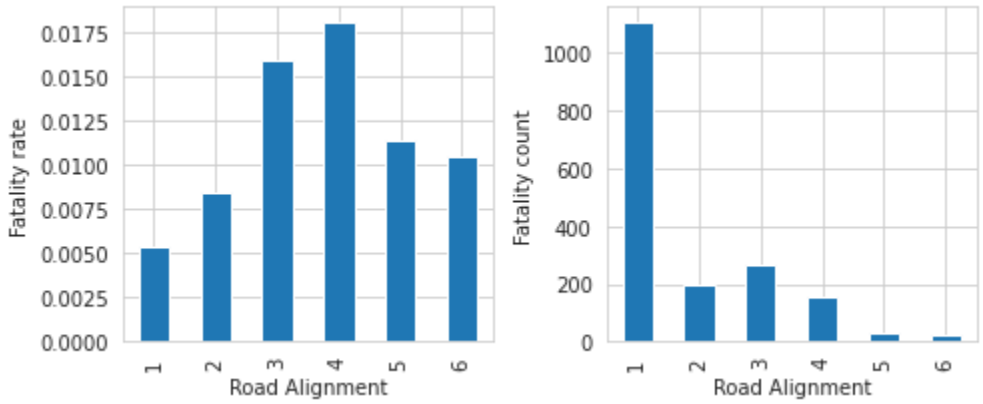

Figure 9: Fatality Histogram by Road Surface

12) Traffic Control: This column refers to the type of traffic control that a collision occurred on (e.g. stop sign, yield sign). As seen in figure 10, the highest number of fatalities occurred when no control was present (code 10) but the highest fatality rates occurred when there were markings on the road (code 12, e.g. no passing) and at railway crossing with signals (code 15). This is not a candidate feature since there are so few fatality counts across other conditions, the skewness in the data may misrepresent the effects of each condition during modeling.

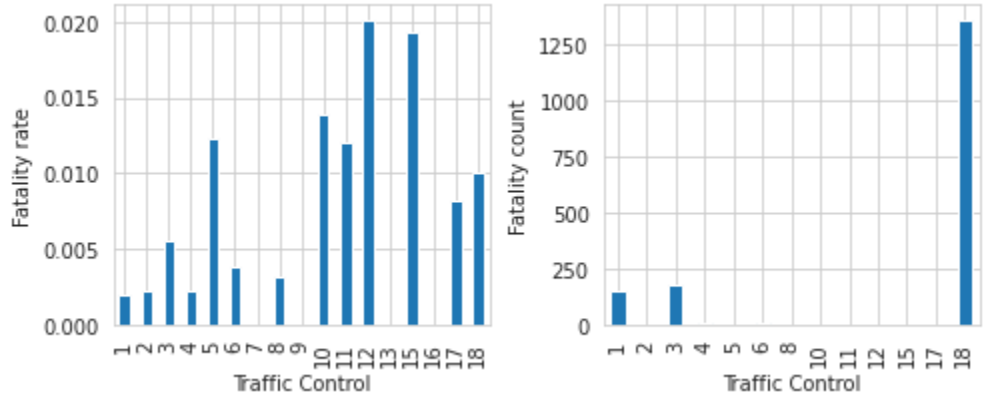

Figure 10: Fatality Histogram by Road Surface

13) Vehicle sequence number: This column represents an identification code for each individual vehicle in a case. Similarly to column 6, this column set will be dropped from the models because the data is presented on a collision case basis rather than an individual basis (There is sequential relationships between rows, not independent).

14) Vehicle Type: As seen in figure 11, the highest number of fatalities occurred in light duty vehicles (code 1) but the highest fatality rate occurred on snowmobiles (code 22). Fatality rate can be misrepresented here however because some conditions have a small sample size so it

must be interpreted carefully. One change in the data made here was re-coding “NN” to 0 because it represented pedestrians (as opposed to a vehicle). It is in the team’s interest to also study fatality of a person involved in a vehicular collision and thus this is a candidate feature for the models.

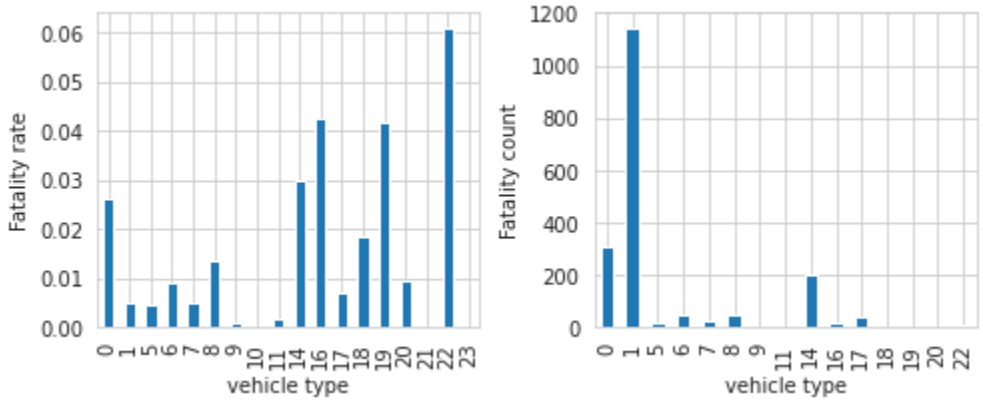

Figure 11: Fatality by Vehicle Type

15) Vehicle Year: This column is removed from the dataset since information on more present data is of interest. Similar to datasets of older years, it is trivial that older cars would have more fatalities since they are not as developed which would skew the data towards older years.

16) Person sequence number: This column is removed from the dataset as it is simply an identification code for individuals in each case. It is not a feature that will affect fatality.

17) Sex: This is a feature candidate in the model since not only is the fatality in males higher, but also the rate. This suggests an effect in the sex of an individual on fatality that should be incorporated in models.

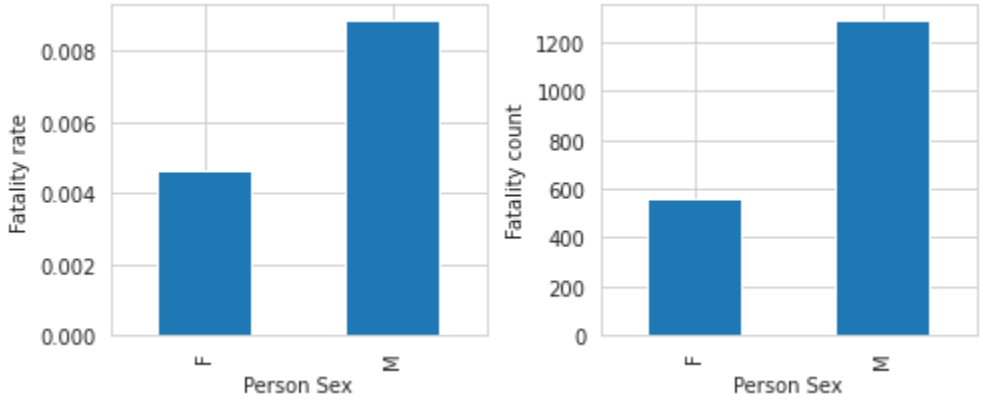

Figure 12: Fatality by Vehicle Type

18) Age: As seen in figure 13, fatality count is highest between 20 to 30 but fatality rate increases the older an individual is. This suggests this is a good candidate feature because there exists a relationship between fatality and age.

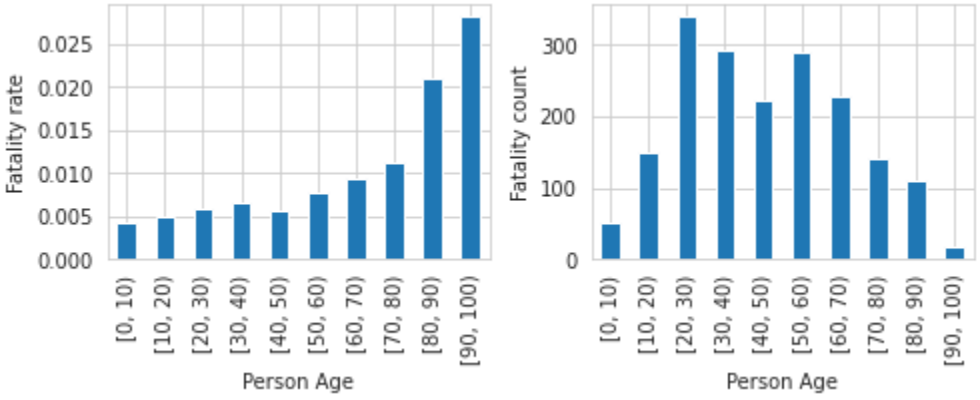

Figure 13: Fatality by Age

19) Person position: This column represents the position of the individual in the vehicle during collision (e.g. code 11 = driver). This was removed from the dataset in models because the number of specific conditions makes it difficult to interpret effects of each condition. It may be applicable for vehicle designer/engineering to know what positions cause higher fatalities however it will not be studied in this report.

20) Person safety: This column represents the type of safety equipment/device that an individual used during a collision (e.g. code 9 = helmet worn). As seen in figure 14, the highest number of fatalities occurred when safety devices were used (code 2, e.g. seat belt) while the highest fatality rate occurred when reflective clothing was worn. This column was removed from the models since it does not add value to the project's objective goals.

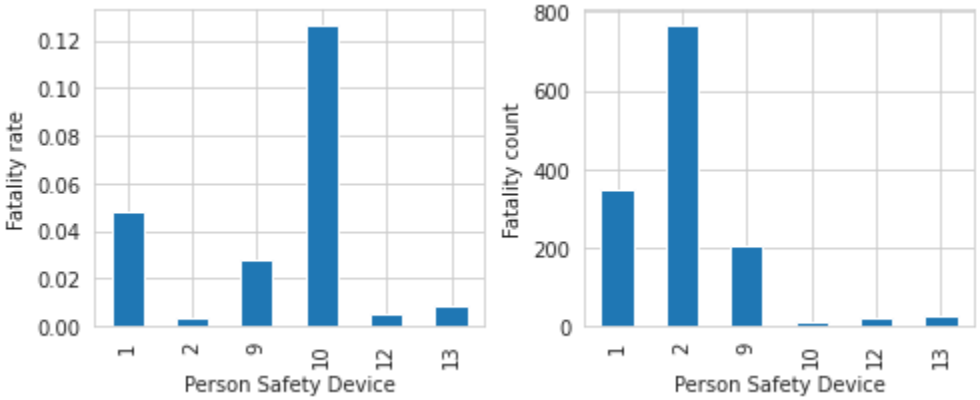

Figure 14: Fatality by Person Safety Usage

21) Road user class: This column represents the category of the individual in the collision (e.g. code 1 = motor vehicle driver). As seen in figure 15, the highest number of fatalities was in motor vehicle drivers while the highest fatality rates were pedestrians (code = 3) and motorcyclists (code = 5).

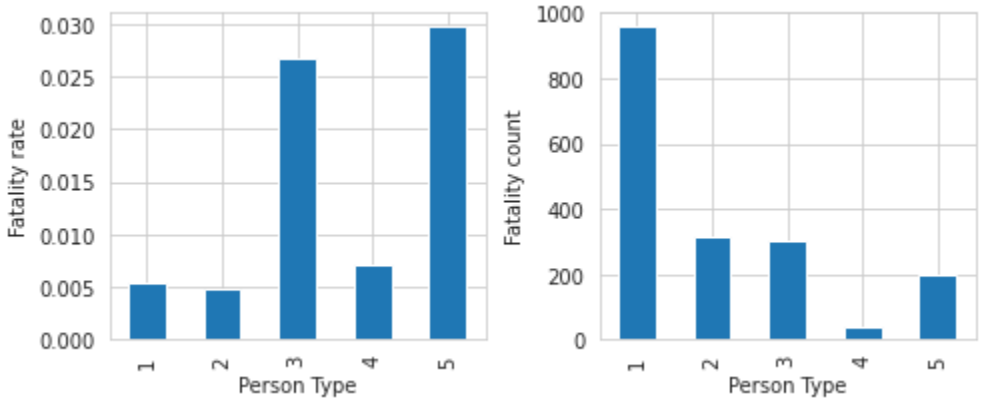

Figure 15: Fatality by Person Type

Chosen Input parameters

The final candidate features selected includes: Month, collision configuration, roadway configuration, weather condition, road surface, road alignment, vehicle type, person sex, person age, and person user class. A limitation of this feature elimination process by using histograms and previous studies is the lack of statistical testing of the models. Future studies can improve on this by comparing statistical significance of categories in different models to demonstrate whether a particular factor has an effect on the binary response variable, fatality.

Results

Logistic Regression

Logistic regression was used on the dataset to demonstrate the use of supervised learning in this project. Since the response variable of interest is a binary variable, logistic regression would be a candidate model to use in predicting the outcomes. Since every variable used in the model (except age) is a categorical variable - one-hot encoding was performed for the model to properly predict the response variable.

The model attempted at first included every feature as described in the previous section (a total of 10 features with 75 coefficients). After fitting the model, the follow results were noted and discussed:

Appendix D shows the full list of coefficients of each variable in this model. Table 1 below shows the list of the 5 most positive and negative coefficients in the model

Variable	CONF31	PUSER3	CONF3	RALN4	CONF4	VTYPE11	CONF36	CONF22	VTYPE7	CONF21
Coefficient	1.68191	1.04235	0.63976	0.58003	0.54279	-0.99979	-1.02801	-1.11865	-1.40330	-1.51667

Table 1: Extreme coefficients of Logistic Regression Model

The interpretation of these results are that the 5 variables most influential in increasing fatality rate are head-on collisions, user class as a pedestrian (as opposed to being on a vehicle), collisions running off the left shoulder, collisions on curved with gradient road grade, and collisions running off the right shoulder respectively. The 5 variables most influential in decreasing fatality rate are rear-end collisions, heavy unit trucks, sideswipe collisions, “other” two-vehicle collisions in different directions of travel, and collisions in an urban and intercity bus.

From these coefficient values, the team made the following findings:

• The three collision configuration conditions corresponding to the top 5 positive coefficients had the highest fatality ratings even though “CONF3” and “CONF4” did not have a high number of collision counts compared to other conditions. This shows it was a good decision to not rely strictly on the counts using histogram but to examine other statistics during initial exploratory phases of the data.

• The user class being a pedestrian (as opposed to being in/on a vehicle during collision) did not have the highest fatality rate and had the lowest fatality count in its category (see figure 15) yet it had the second highest effect on fatality. This is surprising as it would not have been obvious by looking at the histograms. As such, the fact that amongst all these different variables, this

condition manages to stand out indicates the importance of pedestrian safety on roads and that methods should be implemented to help separate pedestrian and vehicles on roads.

• Curved and gradient roads having a significant effect on fatality should indicate these types of roads to be avoided when driving. In terms of road planning, when possible these types of roads should be avoided when possible or spend more resources on remodeling the area to reduce gradient and curved levels if the goal is to reduce fatality rates.

• Person age has a very small coefficient value (0.018) compared to all other conditions. This is surprising as the histogram indicated an increasing fatality rate in seniors and high fatality counts in the 20-30 age range. As well, Age is often a significant factor in insurance rates (Lynch, 2020). This suggests that models used by insurance companies are potentially focused more on collision rates rather than fatality rate however it would be interesting to note if insurance companies factor in fatality rate of ages in their cost models.

• Rear-end collisions having the most negative effect on fatality rate is insightful

• Buses showing a significant negative effect on fatality rate demonstrates the safety of buses on roads and buses should be promoted on roads if the goal is to reduce fatality rates.

After training and testing the dataset on the defaulted 25% testing dataset, the accuracy of the model was given as 99.33%. At first glance, this may suggest the model is good because of the high accuracy however it is misleading. Table 2 below shows the confusion matrix of the model output.

Predicted Actual	0	1
0	51990	0
1	366	0

Table 2: Confusion Matrix of Logistic Regression Model

From the matrix, the team can see that the model actually never predicts a fatality to occur (1). This is because in supervised learning, the model aims to improve accuracy on the training dataset and as such because there is a significant imbalance in the class variable (refer to figure 1), the model will always predict the majority class. As such, the high accuracy rating is

deceptive and the confusion matrix is a better representation of how well the model actually performs - in which case it does not perform well since it never predicted a fatality. The 99% accuracy can be interpreted as a lower bound of accuracy in terms of the dataset and the goal is to find improvements in small quantities.

To verify the model with improved accuracy, a k-fold cross validation technique was performed using k = 10.

Predicted Actual	0	1
0	202132	5892
1	1346	51

Table 3: Confusion Matrix of Logistic Regression Model using 10-fold Cross Validation

Using k-fold cross validation, the model is able to predict fatalities however the number of false positives are fairly high in comparison to the actual number of positive values (1). In future studies, an improvement would be to reperform k-fold cross validation with higher number of folds to reduce bias but also to resample the data such that each sample used in a fold contains a minimum number of the minority class (Brownlee, 2020).

Apriori Algorithm

The Apriori algorithm was used on this dataset to demonstrate unsupervised learning in this report. For this algorithm, the decided parameters that were chosen to partake in this algorithms are as follows:

Parameter	Title
C_MNTH	Collision Month
C_WTHR	Weather Condition
C_RSUR	Road Surface

C_RALN	Road Alignment
C_CONF	Collision Configuration
C_RCFG	Roadway Configuration
V_TYPE	Vehicle Type
P_SEX	Persons Sex
P_ISEV	Medical Treatment
P_USER	Road User Class

Table 4: Parameters used for the Apriori algorithm

For this algorithm, the goal is to observe the likelihood of having a P_ISEV of 'No Injury/Injury' and 'Fatality'. Knowing this, we can assume on further similar datasets that given some combination of parameters excluding P_ISEV, how likely is it that a fatality occurs or does not occur for that collision.

Antecedents in table 5 represent what we are given as initial conditions. Consequent represents the result given the initial condition. Confidence represents the percentage of how likely it is that the combination of antecedent/consequent is for a given collision based on the dataset. For example, from table 5, we can see that if the weather is 'clear and sunny', we are 99% confident that the resulting collision produced no injury or an injury, but not a fatality.

Refer to table 5:highest confidence of 'No Injury/Injury' with a support 0.5 and greater.

Given the dataset, there is an overwhelming number of no injury or injury collisions in comparison to fatalities. Because of this variance, the output is skewed to one side. Given a dataset of equal number of fatalities and/or Injuries/no injuries, the algorithm would produce better likelihoods of collisions being fatal or not.

The example given in table 5 represents data that is supported at a minimum of 0.05 meaning the combination of antecedents/consequents is shown in the dataset at a minimum of 5%. For a minimum support of 0.05, the results ignore the extremely low supported 'Fatality'

consequents and output the highly supported ‘No Injury/Injury’ consequents. In order to view the consequents that include ‘Fatality’, one would need a minimum support value of around 0.00005 which takes a lot of computational power and time. Table 6 below shows roughly the computational time for this dataset for minimum supports of 0.05 and 0.005. For a powerful system that has a lot of computational power, the results obtained from a minimum support of 0.00005 and under would significantly improve the number of combinations of antecedents/consequents, and would result in better data collection and more sophisticated conclusions. One could also reduce the run-time by minimizing the number of parameters used however, the result would not be as varied.

The complete result of the Apriori algorithm based on a minimum support of 0.05 can be found Appendix E.

antecedents	consequents	support	confidence
'Dry, normal', 'Straight and level', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.054676049	0.999828193
'Dry, normal', 'Straight and level', 'Clear and sunny', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.051716524	0.999818363
'Straight and level', 'Clear and sunny', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.057560412	0.999755222
'Straight and level', 'Female', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Rear-end collision'	'No Injury/Injury'	0.051472246	0.999726277
'Dry, normal', 'Light Duty Vehicle', 'At an	'No	0.063554624	0.999704426

intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	Injury/Injury'		
'Dry, normal', 'Clear and sunny', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.059768311	0.999685708
'Straight and level', 'Light Duty Vehicle', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.073250592	0.999679446
'Straight and level', 'Dry, normal', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.057818783	0.999675114
'Dry, normal', 'Straight and level', 'Clear and sunny', 'At an intersection of at least two public roadways', 'Motor Vehicle Driver', 'Rear-end collision'	'No Injury/Injury'	0.054610282	0.999656032
'Female', 'At an intersection of at least two public roadways', 'Rear-end collision', 'Straight and level'	'No Injury/Injury'	0.053473449	0.999648722

Table 5: Highest confidence of 'No Injury/Injury' with a support 0.5 and greater

Minimum Support	Run-Time
0.05	*~ 50 seconds*
0.005	*~ 1321 seconds*

Table 6: Run-time of finding sets vs Minimum Support

Conclusions

The results from the logistic regression model indicate the difficulty in modeling fatality rate using this method. Due to the large imbalance in class data, it is difficult for the model to generate accurate predictions. Given more resources, future studies can improve on this model by performing k-fold cross validation with tighter thresholds and better sampling techniques. As well, datasets from other years can be added to help the model make better predictions but one should be wary about the relevance of data from older years.

The results from the Apriori algorithm proved to be a good un-supervised model because of the overwhelming number of categorical parameters in the dataset. Due to the limitations of computational power and time, as well as the imbalance in class data, it is difficult to measure the likelihood of fatalities/injuries. Given a balanced dataset, the Apriori algorithm provides useful information which can be used to inference the fatality of a collision based on some parameters such as road conditions and weather.

As well, it would have been insightful to remove certain features from the full model and compare the differences between the two. The dataset was overall categorized into collision level, vehicle level, and person level data. Features in each large category could have been modeled alone to have a more in-depth view. Overall, this model should be interpreted as a preliminary study into this topic to provide initial insights on fatality rates rather than as a final assessment on conditions that impact fatality rates.

Persons who work in the insurance industry or in government will benefit from learning about the fatality rates and what affects the likelihood or increase/decreases the likelihood of fatalities. For example, Insurance companies will find it useful to note that motor vehicle drivers are very unlikely to get into a fatal collision. Aslo, government officials can use the newfound data to find areas of interest that can be changed/enhanced such as roadway configurations or road alignments (uphill, downhill, straight road, etc).

References

Abdel-Aty, M., & Abdelwahab, H. (2004). Analysis and prediction of traffic fatalities resulting from angle collisions including the effect of vehicles' configuration and compatibility. *Accident; analysis and prevention*, *36*(3), 457–469. https://doi.org/10.1016/S0001-4575(03)00041-1

Assi, K. (2020). Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models. *International journal of environmental research and public health*, *17*(20), 7598. Retrieved from https://www.mdpi.com/1660-4601/17/20/7598

Assi, K.; Rahman, S.M.; Mansoor, U.; Ratrout, N.(2020). Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. *International Journal of Environmental Research and Public Health*, *17*(15), 5497–. https://doi.org/10.3390/ijerph17155497

Brownlee, J. (2020). 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Canada. (2019). National Collision Database. *Government of Canada - Transport Canada.* Retrieved from https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a#wb-auto-6

Lynch, A. (2020). Other Factors Affecting Car Insurance Rates. *The Zebra.* Retrieved from https://www.thezebra.com/auto-insurance/driver/other-factors/
https://ieeexplore.ieee.org/abstract/document/5164414

WHO. (2018). Global status report on road safety 2018. *World Health Organization. Retrieved from https://www.who.int/publications/i/item/9789241565684*

Appendix A: Dataset Dictionary

(Refer to NCDB_Data_Dictionary.pdf)

Appendix B: Python code for data exploration and logistic regression model

(Refer to Project.ipynb)

Appendix C: Python code for Apriori Algorithm

(Refer to Apriori.ipynb)

Appendix D: Logistic Regression Model Coefficients

Variable	P_AGE	MNTH2	MNTH3	MNTH4	MNTH5	MNTH6	MNTH7	MNTH8	MNTH9	MNTH10	MNTH11	MNTH12
Coefficient	0.0180	-0.3959	-0.2103	-0.2317	-0.3647	-0.3782	-0.0907	-0.1484	-0.1235	-0.0685	-0.2169	-0.0642

Variable	CONF2	CONF3	CONF4	CONF5	CONF6	CONF21	CONF22	CONF23	CONF24	CONF25
Coefficient	-0.1729	0.6398	0.5428	-0.1515	0.0687	-1.5167	-1.1186	-0.4055	-0.9019	-0.4058

Variable	CONF31	CONF32	CONF33	CONF34	CONF35	CONF36	CONF41
Coefficient	1.6819	0.4182	-0.6694	-0.4356	-0.1482	-1.0280	-0.5829

Variable	RCFG2	RCFG3	RCFG4	RCFG5	RCFG6	RCFG7	RCFG8	RCFG9	RCFG10
Coefficient	-0.7734	-0.8893	0.1946	-0.4752	-0.1274	-0.0240	-0.3853	-0.1732	-0.0152

Variable	WTHR2	WTHR3	WTHR4	WTHR5	WTHR6	WTHR7
Coefficient	-0.0756	-0.3055	0.0834	-0.1844	0.5306	-0.2198

Variable	RSUR2	RSUR3	RSUR4	RSUR5	RSUR6	RSUR7	RSUR8	RSUR9
Coefficient	-0.0754	-0.9656	-0.4770	-0.8995	-0.5129	0.3729	-0.0493	-0.0140

Variable	RALN2	RALN3	RALN4	RALN5	RALN6
Coefficient	0.3593	0.4076	0.5800	0.1870	0.3787

Variable	VTYPE1	VTYPE5	VTYPE6	VTYPE7	VTYPE8	VTYPE9	VTYPE10
Coefficient	-0.5660	-0.9278	-0.8838	-1.4033	-0.5100	-0.4613	-0.0338

Variable	VTYPE11	VTYPE14	VTYPE17	VTYPE18	VTYPE21	VTYPE23
Coefficient	-0.9998	0.3268	0.0928	-0.0175	-0.0347	-0.1037

Variable	SEXF	PUSER2	PUSER3	PUSER4	PUSER5
Coefficient	-0.5410	0.1311	1.0423	0.0928	0.3268

Appendix E: Apriori algorithm result based on a minimum support of 0.05

(Refer to out.csv)