Investigator(s) Information

Chatrola	Vivekkumar	40059267
Haleem	Michael	40043639

Project's General Information

Title:	**Analysis of Road Transportation Accidents Using Data Mining Techniques**		
Infrastructure Sector	Transportation	**Sub-sector:**	Public Roadways
Municipality:	Ottawa	**Region:**	N.A.

Abstract – *Roadway traffic accidents are major concern for government and citizens. Mortality road due to road accidents is spreading day by day. In order to make driving safe, analysis of transportation collision data is very important to find the factors that are closely related to accidents severity. In this paper, we apply data mining techniques on collision data of the city of Ottawa in order to study this problem. We will try to find out effect of weather, road surface, traffic, lighting on the severity of accidents to generate the rules which leads to certain type of accidents. We also want to see if people drive more cautiously at the location of red-light camera or not. Dataset was divided into similar clusters based on location related factors. Association rule mining technique was used to find relation between different attributes. We trained decision tree model to classify the accident based on given attributes. The decision tree model was evaluated based on 10-fold cross validation technique. Certain suggestion was made based on association rules, classification model and clusters obtained. The results can be used to take some accidental prevention measures in the city of Ottawa.*

1. Introduction and Background

Road Transportation accidents are one of the major causes of losses such as equity losses and life losses. Mortality rate due to road accidents is spreading throughout the world. Although for Canada is the lowest. In Canada , each year there are total 150,000 accidents. Number is almost fixed and is not increasing but we can still try to decrease it as much as we can. This motivates us to inspect the following 4 points: safety of people and equity. In totality we are trying to create a safer environment for the users. Awareness –try to increase traffic road user 's awareness about the most dangerous / not good condition for driving so that they will be more cautioned focussed and try to avoid it leading to lesser losses. Technology – We have the enough tools and data to try to implement the technology or more specifically come up with a model helping us to achieve a decrease in losses. Identification – Identify the most dangerous neighbourhood/intersection for driving. It'll be more of a motivation for government and municipalities to invest more in infrastructure at such perilous intersections.

2. Problem Statement and Objective

Our problem statement will be focussing on Ottawa the capital city of Canada with a population of around 1 Million People. The collision type and road service and user conditions in Ottawa can be described as a prototype for most of the Canadian cities. On an average it experiences 15k cases per year and the number is fixed again but we can still improve it by reducing it further. If we exclude the human behaviour and attitude from the equation, we observe 4 factors that lead to such collisions: Environment, Road Surface, Lightning and Traffic control. So, our scope is from

Ottawa and it is based in data from 2015 to 2017 as this is the only available data. It does not have any human behaviour data, but it has Environmental, Traffic and Road Surface data. Our objective would be trying to first find the condition which would lead to different types of collisions, especially at most frequent accident locations and the second objective is does the presence of red light cameras in the places have any co-relation to the number of collisions as if someone is watching you will that lead to a change in your behaviour?

3. Previous Works

In the previous works, Researchers used Data Mining (DM) techniques to find the relation between the accident severity and other attributes such as collision manner, weather, surface condition, light condition, driver behavior [1, 3-7]. Similarly, Xue-Fei and Lisa [2] tried to find the similar relations but for different group of people in the city of Saskatchewan.

DM methods such as Apriori algorithm [1, 4] and FP-Growth algorithm [3,5] for Association Rule Mining, Decision Tree [2,7] and Naïve Bayes [1,3] for Classification, Artificial Neural Network [7] have been used for accident data analysis with reasonable results by many researchers. Also, they used clustering such as k-means clustering [1,3] and k-mode clustering [5] to find out similar clusters and similar accident location based on accident severity. Ganani and Ramya Devi [3] generated confusion matrix to evaluate the performance of the model they developed. Miao, Ajith and Marcin [7] used 10-fold cross validation to measure the accuracy of the model.

In their results, Liling [1] found that clear cloud condition with daylight has higher fatal rate and human factor such as drunk or not drunk have major effect on accidents. Ganani and Ramya Devi [3] found that rash driving leads to fatalities and accidents are likely to occur in areas with poor lighting. Jianfeng [4] found that wet conditions have a 43% probability resulting in a rollover. Also, they found that when a less experienced driver drives on the slippery muddy road in the rain, there is 56% probability of rear-end collision. Joaquin, Griselda, and Juan [6] used Decision Tree by varying the rood node and they found 70 relevant rules compare to 5 rules obtained from only one root node. In their results, they found that the environment of two-lane rural highways (i.e., safety barriers, shoulders, visibility, lighting, etc.) have high impact on accident severity.

Liling's [1] work was limited to fatal accidents only. They didn't have non-fatal accident data and mileage data. Similarly, Sachin and Durga [5] had very limited information and dataset available. Unavailability of the real time accident data was main limitations in all the studies.

There hasn't been much work done on transportation collision data of the city of Ottawa. To find out relation between accidents and its factor in Ottawa, in this paper, we divided dataset into similar clusters by k-means clustering and then used FP-Growth Algorithm for association rule mining to find the relation between various factors affecting road collisions and used Decision Tree analysis to generate the rules to classify the accidents. To measure the accuracy of the model, we used 10-fold cross validation technique.

4. Methodology

The approach we acquired for our study follows the CRISP-DM data analysis process and its steps are described below. We used RapidMiner software to perform all the steps of CRISP-DM model.

Understanding Data

The data for this study has been acquired from the open data portal of the city of Ottawa. The dataset contains 73,505 collision record from the year 2013-2017 that happened in Ottawa, Ontario. It has total 20 attributes. The attributes of the dataset are mostly categorial in nature. Table 1 shows the brief description of the major attributes of the dataset. This dataset doesn't have any human behavior information. Each reportable accident occurring on public roadways is sent to the City of Ottawa and is validated at least once. Approximately 50% of the records are validated once again by a senior staff. Additionally, many queries are run on the database looking for errors. Collisions are pinned according to information provided by the officer on the Motor Vehicle Collision Report. In some cases, insufficient information was provided, and the collision location is an estimate. From 2013-17 there were total 2.6%, 4.4%, 6%, 6.0%, and 5.3% estimated collision location respectively. There are no known errors in the location of red-light camera.

Table 1 Road Accident Attributes Description (After Preprocessing)

Attribute Name	Type	Values
Time	Polynominal	Morning, Afternoon, Evening, Night
Day	Binominal	Weekend, Weekday
Month	Polynominal	January to December
Environmental	Polynominal	Clear, Snow, Rain, freezing rain, Drifting snow, Strong wind and Fog, mist, smoke, dust
Traffic Control	Polynominal	No control, Traffic signal, Stop sign, Roundabout, Yield sign, Traffic gate, Traffic controller, School bus, Pedestrian cross-over
Red Light Camera	Binominal	Yes, No
Location	Polynominal	Around 13000 different locations of the city
Light	Polynominal	Daylight, Dark, Dust, Dawn, Dark-Artificial, Daylight-Artificial, Dawn-Artificial, Dusk-Artificial
Surface Condition	Polynominal	Dry, Wet, Loose snow, Ice, Slush, Packed Snow, Loose Sand or Gravel, Mud, Spilled liquid
Collision Classification	Polynominal	Fatal Injury, Non-fatal injury, Property damage only
Impact Type	Polynominal	Rear end, SMV other, Angle, Sideswipe, turning movement, SMV unattended vehicle, Approaching
No. of Pedestrians	Integer	0 to 4

Data preparation

Data preparation is one of the most important step of data mining process which involves transferring raw data into an understandable format. Generally Real-world data is incomplete, inconsistent and is likely to contain many errors. In this paper, data preprocessing techniques such as data cleansing and data transformation is used. We removed traffic control condition

attribute (functioning, not functioning, damaged, obscured) since almost half of its value were missing. We also removed X and Y coordinates projected in MTM Zone 9, NAD83 (CSRS) since we already had access to longitude and latitude. Environmental and Traffic control had 1 and 50 missing values respectively. For those attributes, we removed all the datapoints with missing values. For Environmental, Light, Traffic control, Impact type and Surface condition we had values such as unknown and other. Since these values were not specified, we preferred to remove all those datapoints. Instead of replacing these values with mean, mode or median, we removed all the datapoints with missing and unknown values as this data are real world data and editing them manually can change the results. We don't have any inconsistency in our dataset.

Red light camera location data were obtained from the open data portal of the city of Ottawa and then it was added to the main dataset as a Boolean attribute with the help of MS Excel. Several attributes were obtained from the date and time attributes. We discretized date into date, month and year. We found days from the dates and then we mapped it to weekdays=0 or weekend=1 to make it Boolean attribute. Hours of the day were separated from the time and it was further discretized into Morning (6 AM-12 PM), Afternoon (1 PM-4 PM), Evening (5 PM-9 PM) and Night (9 PM-5 AM). All the numeric values were converted into nominal values except location coordinates. After performing all the preprocessing, 71,416 examples and 17 attributes were found suitable for further analysis.

Modeling

A. Clustering Algorithm

Clustering is an unsupervised data mining technique which groups dataset into different cluster in such a way that objects within each group are more similar than the object in other cluster. K-means algorithm is one of the famous clustering algorithms in which centroids are mean of all the data points in each cluster. It divided dataset into k different clusters. Our main aim behind using clustering is to find the different groups based on location characteristics and type of classification and no. of pedestrians involved in each case. In RapidMiner, we used location coordinates, collision classification, no. of pedestrians, red light camera locations and collision id (as an id) to find the clusters. K-means clustering requires data to be in a numeric format and normalized. We converted all the nominal values into numeric and then normalized them. Euclidean distance with numerical measures was used to find the distance. We set the value of maximum run to 10 and maximum optimization steps to 100. Davies-Bouldin Index (DBI) can be used to find the best k. DBI measures the uniqueness of the clusters. Cluster distance performance operator was used to find the DBI for the best k. We then denormalized clustered dataset and Extracted clusters to study the behavior of each cluster and its centroid.

B. Association rule mining

Correlation analysis cannot be applied to categorical data, especially when there is no order in categorical variable. To find the co-occurrence between one item and other item, we need to use association analysis which is also a branch of unsupervised learning. It works based on market basket analysis and finds hidden patterns in dataset by generating rules in the format of {Itemset(s)A} → {Itemset(s)B} where {Itemset(s)A} is antecedent and {Itemset(s)B} is consequent. In RapidMiner, we used FP-Growth algorithm for affinity analysis. Attributes such as collision classification, environmental, light, red light camera, surface condition, traffic condition and collision id (as an id) were selected to find out the relation between them. Our main aim was to find the antecedent which leads to different accident type in its conclusion.

Numerical data were converted into nominal format. Strength of these rules can be measured by various interesting measures such as support, confidence, lift and conviction. Support value indicates the frequency of occurrence of a rule in the dataset whereas confidence indicates reliability of a rule. Hence, rules with high support and high confidence are of main interests. Minimum support of and minimum confidence of were kept to get the results. The number of items per itemset were kept between 1 to infinity. Similarly, number of itemset were kept between 100-1000000 as a criterion to perform the analysis. We lowered down minsup and minconf to get the rules with non-fatal accidents.

C. Decision Tree Analysis

Decision tree analysis is the branch of supervised learning which takes input from independent variables and sort the response variable into different classes. For classification, DT is one of the most frequent used data mining technique. For, DT predictor variable can be categorical or numerical, but target variable must be categorical. To find our most important objective of finding conditions which leads to different types of accident, we developed DT model through RapidMiner. Attributes used in association analysis was again used in the DT to predict the target variable 'collision type'. One of the most important aspect of the DT is to find the impurity. There are many measures to find the impurity such as entropy, Gini-index, information gain, information gain ratio. Different measures build different trees through different bases. In our study, we used information gain ratio which can be utilized for splitting of the tree. Other criteria we set to train the DT were: maximal depth=5, confidence=0.1, minimum leaf size=2, minimal size for split=4. Pre-pruning and Post-pruning were used to simplify the tree and prevent it from overfitting. To find the causes of accidents in top 10 most frequent accident location in the city (having more than 140 accidents), another DT was trained with similar parameters.

Evaluation

We trained our DT model based on Information gain ratio goodness of fit measures. To evaluate the model, n-fold cross validation technique was used. The cross-validation technique shows the ability of the model to hold up to when applied with new data. It generates confusion matrix which shows the different metrics such as precision, recall, accuracy, error and specificity. Apply model and classification performance operators from the RapidMiner were used for this task. We kept the number of folds 10 and used stratified sampling for cross validation purpose. Our focus was to find the accuracy of the both DT models.

5. Results and Discussion

A. K-means Clustering

Total 6 cluster model were generated for k=2 to k=7. DBI for each cluster is shown in table 2. Among them, k=4 cluster model was selected since it has lowest DBI. These 4 clusters obtained by k-means algorithm are explained below.

Table 2 DBI for k-means clustering

k	2	3	4	5	6	7
DBI	0.432	0.463	0.376	0.562	0.469	0.517

Cluster_0 involves 77.55% of the total accidents. This clusters shows accidents with no pedestrian involvement. All the accidents were resulted in property damage only

Cluster_1 involves 16.00% of the total accidents. This clusters shows accidents with no pedestrian involvement. 99.3% accident were resulted in non-fatal injury while 0.7% accident was fatal injury. This is a good cluster to find out reasons for non-fatal accidents.

Cluster_2 involves 2.35% of the total accidents. Only cluster 2 shows accidents with pedestrians. It has only 3% accidents were on the place of red-light camera location. It has 95.7% non-fatal accidents. This cluster can be utilized to find out conditions which leads to pedestrian accidents.

Cluster_3 involves 4.00% of the total accidents. This clusters shows accidents with no pedestrian involvement. All the accidents in this cluster were occurred at red light camera locations. It has 80% property damage accidents. It can be used to find the accident attributes behavior at red light camera locations.

B. Association rule mining

We used minsup of 40% and minconf of 80% to find the rules with property damage as a conclusion. If there is no traffic control and red-light camera located at the place, chances of property damage as an accident type will be high with confidence level of 0.82. Clear environment condition and dry road surface will also lead to property damage type accident with confidence level of 0.84. We lowered down our minsup to 0.15 and minconf to 0.20 to find the rules associated with non-fatal injuries. Also, the day light- light condition and no red-light camera location are more likely to result in non-fatal accidents with confidence level of 0.2. Even after lowering the minsup and minconf, we didn't get any rules which lead to fatal injury. One of the surprise rules we found is that effect of red-light camera on severity of accidents. Red light camera is very important factor for accidents. Increasing it with proper traffic control system would lead to decrease in amount of accidents as people will be more cautious.

C. Decision Tree Analysis

Table 3 shows the list of attributes available in DT model in a descending order of their weights based on maximum depth of 5 and information gain ratio.
Surface condition is the most important attribute for the DT model with 48% weightage. Environment condition is not that much critical for accidents. Based on DT analysis, it can be said that Rear end impact type leads to property damage accidents. Wet surface condition and dusk light leads to non-fatal accidents (confidence=70%). Only one rule with fatal accidents were found with very low confidence:
If Impact type = approaching & Environmental = clear & Surface condition= wet & Light = Dark, then Accident type = fatal injury. (confidence=7.7%)

Table 3 Weightage of attributes for DT

No.	Attribute name	Weight
1	Surface_Condition	0.481
2	Light	0.362
3	Environment	0.0708
4	Traffic_Control	0.0489
5	Impact_Type	0.0363

Table 4 Weightage of attributes - DT (most frequent location)

No.	Attribute name	Weight
1	Surface_Condition	0.443
2	Light	0.222
3	Traffic_Control	0.145
4	Environment	0.106
5	Impact_Type	0.084

Table 4 shows attributes weightage when DT was trained for top 10 most frequent location of the city. It is like the previous one. But traffic control is more important than environmental factor

for accidents occurring at most frequent locations. For most common location, dry surface condition and clear environmental condition more tend to lead to Property damage type accident. Freezing rain and roundabout are more likely to result in non-fatal injuries at those locations.

Decision rules for the DT are generated for non-fatal injuries and fatal injuries. Although, we didn't find any confident rule for fatal injuries. Wet surface condition and dark light are very critical criteria as it can lead to fatal injuries.

D. Performance Evaluation

Figure 1 and 2 shows confusion matrix for DT for full dataset and DT for most frequent station. Accuracy of these models are 81.39% and 83.46% respectively which shows the goodness of measure. Although we didn't find any precision or recall for Fatal injury since our model didn't predict any rules for fatal injuries.

accuracy: 81.39% +/- 0.18% (micro average: 81.39%)

	true 03 - P.D. only	true 02 - Non-fatal injury	true 01 - Fatal injury	class precision
pred. 03 - P.D. only	57160	12577	100	81.85%
pred. 02 - Non-fatal injury	604	962	13	60.92%
pred. 01 - Fatal injury	0	0	0	0.00%
class recall	98.95%	7.11%	0.00%	

Figure 1Confusion matrix for DT

accuracy: 83.46% +/- 0.98% (micro average: 83.46%)

	true 03 - P.D. only	true 02 - Non-fatal injury	true 01 - Fatal injury	class precision
pred. 03 - P.D. only	1459	274	1	84.14%
pred. 02 - Non-fatal injury	16	9	0	36.00%
pred. 01 - Fatal injury	0	0	0	0.00%
class recall	98.92%	3.18%	0.00%	

Figure 2 Confusion matrix for DT-most frequent location

6. Concluding Remarks

A series of studies have been carried out to analyze the causes of accidents in Ottawa using data mining techniques. The objective of this project was to study and analyze the transportation collision data obtained from the open data portal of the city of Ottawa to find the effect of mainly weather, light, road surface and traffic data on the different collision types. Our approach used 71,416 accident records after preprocessing that was occurred in Ottawa during 2013 to 2017. RapidMiner software was used to perform association rule mining, DT analysis, clustering and 10-fold cross validation. We used k-means clustering and obtained 4 clusters based on accident location and type. Further, association rule mining technique was applied to selected parameters to identify some interesting rules with collision types as its conclusion. We then trained Decision tree to classify the accidents based on predictor data and obtained conditions to those classifiers. Evaluation of the Decision tree model shows that it works with 81% accuracy.

The results of this study could be used by the respective authorities to promote road safety and create awareness about risk factors. This work will be useful to Ottawa authorities and citizens.

Our results can be used as an improvement to enhance sustainability of transportation infrastructure.

Although we find some rules, but that was not that much surprising. Most of our rules were for normal conditions of weather. Light and road surface. These rules would have been more interesting if we had access to human behavior data. Also, these data are not real time data. Data analysis on real time data can generate better predictive model.

7. References

[1] L. Li, S. Shrestha and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2017, . DOI: 10.1109/SERA.2017.7965753.

[2] X. Zhang and L. Fan, "A decision tree approach for traffic accident analysis of Saskatchewan highways," in *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2013, . DOI: 10.1109/CCECE.2013.6567833.

[3] G. Janani and N. R. Devi, "ROAD TRAFFIC ACCIDENTS ANALYSIS USING DATA MINING TECHNIQUES." *Journal of Information Technology & Applications*, vol. 7, *(2)*, pp. 84-91, . DOI: 10.7251/JIT1702084J.s

[4] J. Xi *et al*, "A Traffic Accident Causation Analysis Method Based on AHP-Apriori," *Procedia Engineering*, vol. 137, pp. 680, . DOI: 10.1016/j.proeng.2016.01.305.

[5] S. Kumar and D. Toshniwal, "Analysing road accident data using association rule mining," in *2015 International Conference on Computing, Communication and Security (ICCCS)*, 2015, . DOI: 10.1109/CCCS.2015.7374211.

[6] J. Abellan, G. Lopez and J. de Ona, "Analysis of traffic accident severity using Decision Rules via Decision Trees," *Expert Syst. Appl.*, vol. 40, *(15)*, pp. 6047-6054, . DOI: 10.1016/j.eswa.2013.05.027.

[7] Miao Chong, A. Abraham and M. Paprzycki, "Traffic Accident Analysis Using Machine Learning Paradigms." *Informatica (03505596)*, vol. 29, *(1)*, pp. 89-98, .