The Predictions of Road Accidents Causes

Chinmay A. Patane *(Author)*
Dept. of Management and Information Systems
Kent State University,
Kent, Ohio, USA
cpatane@kent.edu

Zi Yang (*Author)*
Dept. of Management and Information Systems
Kent State University
Kent, Ohio, USA
zyang8@kent.edu

Aparnnaa (Author)
Dept. of Mathematical Sciences
Kent State University
Kent, Ohio, USA
aparnna@kent.edu

***Abstract*—Road accidents are the mainly reason for injuries and other health problems, resulting in an estimated 1.3 million deaths and 20-50 million are injured globally each year. In the modern world, road accidents are the major cause of death and injury. The United States is the country experienced higher accident rate. Thus, it is important to reduce accidents to save the life. Data mining is a reliable and efficient technique to analyze and interpret road accidents. Despite all cares and precautions taken road accidents are unavoidable. The data acquired from data.gov shows that there were 341k accidents happened in the United States during the year 2012. The number is nearly constant in 2013 and 2014. As a conclusion, accidents are unavoidable. However, this situation involved people in helping the victims and make them get support after an accident. To do so, we are seeking in the prediction before the accident, include the analysis of the place, time and individual. The prediction regarding the factors description of the accident will help police department in further consideration as well as insurance companies. Including the suggestions to select the appropriate help from the nearest medical institution. Which present the prediction and the description of the contributing factor.**

***Keywords*—Road Accidents, Causes, Data Mining**

I. Introduction

Road traffic accident plays one of the major role in damaged physical health condition and other health-related issues around the world. From Association for Safe International Road Travel, an estimated 1.3 million deaths and 20-50 million are injured would happen each year globally. Road crashes rank as the ninth leading cause of death. Unless actions are taken, road traffic injuries are predicted to become the fifth leading cause of death by 2030 [6]. According to statistics, over, 37,000 people die in road crashes each year in the United States. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person [6]. Hence, the development and application of traffic safety programs are essential for travelers' safety as well as reducing the hazards of road accidents. Traffic experts are still have the problem of factors identification in the incidence affection and the severity of an occurred crash. The traffic safety is subjected to vast and complex dimensions in which to interact together and consequently demands various knowledge and experiences [2].

There are varies of factors make the accidents occurred. However, they happened because of certain actions taken by the driver. These are complex circumstantial relationships between several characteristics (driver, action, car, and disturbance). Under these information, the purpose of this study with mainly nominal values, decision tree and Naive Bayes to figure out the information behind. The data that was used to conduct research was obtained from data.gov. The size of the data was 1,000,000 crashes during the calendar year 2012-2014.

For every road accident that happens across the world, there are factors involved. Factors may be one or more, but the main factor that caused the accident will be the only one. In this project, the model is built based on the factors available at the accident scene by using the data obtained from DATA.GOV and the US government's website for public data. The objectives of the project are make the prediction on two main reasons of the accident:

1. Action of an individual prior to accident
2. Contributing factor of the accident description

The action prior to accident is related to who is driving the vehicle. This prediction gives the action of the driver due to the consideration of the original situation and motivation.
Contributing factors can be many but mainly only 3 factors are considered in the project, which are:

1. Human
2. Environment
3. Vehicle

Again, there can be many factors, but these three are the main reasons. The project further tries to predict the accident causing factor's description, i.e. what could have gone wrong with one of these three factors that caused the accident.

The predictions from the model will be beneficial to the individuals as well as to some insurance companies. The purpose of the project is to predict the action prior to accident. If the action prior to accident can be predicted, this will help the nearest health center or hospital to provide appropriate help to the victims. For example, if the predicted action is "making turn in traffic", the severity of accident can be predicted as high and alert level can be set as high. This will also help the insurance companies to sort out the disputes that can happen later. For example, if the predicted contributing factor description is Animal Action or Unsafe Speed, the insurance companies can deny the insurance because of the violation in the traffic rules by the driver. This kind of prediction can also help he police department to resolve the road accident cases faster.

United States of America has seen almost same number of road accidents in the year 2012 (around 341K), year 2013 (361K) and year 2014 (346K). As compared to the data of road accidents from another nations, the data obtained from DATA.GOV is clearer and takes into account all possible factors that can affect the road accident.

Also, the data obtained is clearly sorted year wise. Almost all attributes are nominal. Every possible description about action prior to accident and contributing factor's description is mentioned. Using this data, with appropriate selection of the attributes, a model is suggested in this project.

II. LITERATURE SURVEY

The focus of this research is to identify the action of driver prior to accident and its contributing factor for taking such an action to avoid the crash. So far there seems to be no other research concentrating on the similar purpose of finding contributing factor based on the action taken prior to the accident.

Various studies have researched on different causes and severities of road accidents. Numerous data-mining techniques and studies have been employed to research and analyze data both locally and globally with results varying based on road infrastructure and socio-economic conditions. Beshah and Hill (2010) used decision tree, naive bayes and k-nearest neighbor classifiers to predict accident severity focusing on the road-related factors in Ethiopia. The accident record data consisted of text, numbers, date and time. From the result, there was approximately 75% property loss, 10% of the time slightly injured, 8% of the time severely injured and very few accidents were fatal.

Pakgoha[2]r, Tabrizi et al. (2010) researched to identify the role of human factor in incidence and severity of road crashes. Attempts were made to find relationship between roles of human factor in severity of road crashes in Iran. Various methods such as descriptive analysis, logistic regression, Classification and Regression Tree were employed. Data was obtained from Database of Traffic Accidents of Iran's Police during 2006. Using logistic regression, 78.57% of all accidents were correctly classified. By using CART approach for determining role of human factor on severity of road crashes 1.91% of accidents have resulted in no injury, 1.8% resulted in injury and 1% of all accidents have resulted in fatal.

Tseng, Nguyen et al. [5] analyzed the data for accidents caused due to distractions and the relationship between driver inattention and motor vehicle accidents. Data was acquired from Fatality Analysis Reporting System from 2000 to 2003. To analyze SPSS Clementine, a data mining software was used to cluster data, decision tree to further classify relationships among the variables and NN model to predict the manner of collision. One such result obtained for decision tree was regardless of being inattentive, if the person has mental/physical condition he or she is more likely to meet with an accident with a fixed object rather than a moving vehicle.

Krishnaveni and Hemalatha [7] researched using various data mining techniques to determine the severity of injury that occurred during traffic accidents. Data was obtained by Transport department of government of Hong Kong. Data set consists of records of 34,575 traffic accidents in the year 2008. Naive Bayes, Bayesian Classifier, ADABoostM1 Meta Classifier, PARTA Rule Classifier, J48 Decision Tree Classifier and Random Forest Tree Classifier. It was seen that Random Forest Tree outperformed the other techniques. Based on accident data set weather attribute gave 90.53% for Naive Bayes, 91.53% for J48, 89.54 for ADABoost M193.34% for PARTA and Random Forest Classifier takes 95.14%.

Yet research made till today focus on the cause of the accidents and try to implement the precautionary measures to avoid the accidents. Research paper [1] shows the relationship between the road conditions and the factor that caused accident, and tries to predict the severity of the accidents. It gives nothing specific about the actions caused the accident. A better prediction about the Human factor that caused accidents is given in [2], but the data shows that there are more than one factor that caused road accidents in US. Another research paper [3] and [4] also shows how the location of the accident and the frequency of the accident can be predicted but fails to mention the exact cause of the accident. Paper [5] gives analysis of driver distraction but again the work is limited up to only human related distraction and does not count any of the other factors. There are many other factors which need to be considered which comes under the distraction category.

III. ATTRIBUTE SELECTION

Original Dataset

The dataset called "Motor Vehicle Crashes – Vehicle Information: Three Year Window" is obtained from the website of DATA.GOV. This dataset contains the reports from NYS DMV with 1048575 instances in total. The original metadata included with 19 attributes, which are: Year, Case Vehicle ID, Vehicle Body Type, Registration Class, Action Prior to Accident, Type / Axles of Truck or Bus, Direction of Travel, Fuel Type, Vehicle Year, State of Registration, Number of Occupants, Engine Cylinders, Vehicle Make, Contributing Factor 1, Contributing Factor 1 Description, Contributing Factor 2, Contributing Factor 2 Description, Event Type, and Partial VIN.

The original data consisted of 19 attributes, after preprocessing the number of useful attributes are 8 that is used for data mining. Data that were removed consisted of Vehicle Id that was numeric and would not produce anything useful after datamining. Other attributes such as Registration class, type of vehicle engine cylinder had lot of missing values. So these were removed to obtain the final data set.

In this study, goal is to predict contributing factor considering the action prior to accident so that the driver can take precaution and know what danger he might be in during such an action. Important features that can affect these are the vehicle body type, number of occupants, direction of travel, vehicle make and so on. Based on the result from action prior to accident, a new attribute containing values from action prior to accident is formed. This new data is used to predict contributing factor 1 description.

Attributes Selected

The purpose is to predict the actions of road accident contributing factor. And its description under the certain circumstances after the road accident has happened, the two attributes "Action Prior to Accident" and "Contributing Factor 1 Description" are selected for the prediction. These two are the most important attributes for predicting the reasons of the motor vehicle crashed. There are 5 remaining attributes are selected for helping this prediction: Vehicle Body Type, Direction of Travel, Number of Occupants, Vehicle Make, Contributing Factor 1.

Reason for Removing Attributes

For the original 19 attributes, there are some reasons that only keep these 5 attributes for the predicting. Firstly, the "Year" attribute is removed. Because all the data are from the year 2012 to 2014 for this dataset. It is only the accidents happened the year, with no more information such as date, time, day in a week, etc. Hence, the "Year" has no relation with the reasons of vehicle crashes. Then, the attributes of "Case Vehicle ID", "Registration Class", "State of Registration", and "Partial VIN" are all basic background information of the accident vehicles, which are codes to find the accident case vehicle but have no help in predicting the accident case reasons. As the result, they were removed for easy predicting for the purpose of this project. Also, 86% of the data from the attribute "State of Registration" is alone from New York, which makes the data biased. This may affect the prediction very bad because the predictor will always predict the state as New York. Thus, this attribute is removed. Thirdly, the attributes called "Type / Axles of Truck or Bus", "Fuel Type", "Vehicle Year", and "Engine Cylinders" are the product information of the accident vehicles which also some basic information of vehicles with less relation with the reason of accidents. Based on the same reason of the previous attributes, these useless attributes are removed. Moreover, the attributes of "Contributing Factor 2" and "Contributing Factor 2 Description" are similar with the attributes that selected for predicting called "Contributing Factor 1" and "Contributing Factor 1 Description". They all describe the detailed reason caused the accidents. However, for the "Contributing Factor 2 Description", it has more missing values than the "Contributing Factor 1 Description". Based on this consideration, we chose to predict the "Contributing Factor 1 Description" and keep the "Contributing Factor 1" for better help in predicting. Lastly, the attribute of "Event Type" is a detailed classification of most of the accident reasons. Although it has a lot help in predicting the objective attributes, similar with "Contributing Factor 2 Description" attribute, the "Event Type" also has a large number of missing values or invalid data. Therefore, the best way to keep the prediction more accurate is removing this attribute.

IV. DATA PREPROCESSING

Data preprocessing is used for cleaning data, merging similar categories of data, and removing the missing values of the selected data. These steps can help keeping the remaining data much more accurate when doing the prediction.

There are two attributes are selected for predicting:

Prediction 1: Predicting Action Prior to Accident
Prediction 2: Predicting Contributing Factor 1 Description

After these two predictions, we hope to find out the most common and frequent reasons caused the motor vehicle crashes. The result may help drivers pay more attention and to keep safe driving and safe life.

Before the predicting, data preprocessing is essential.

At the beginning, using the Filter in Excel Tools to find out the blank cells for attribute "Contributing Factor 1 Description" since this attribute has a lot missing values. This attribute is using for predicting, so it is very important to keep

only valid data, which will cause a different result compare with keeping the missing values. By selecting all categories except "blank" in the Filter for this attribute, simply copy and paste the result worksheet to a new file is the easiest way to remove the missing values.

After the previous step, there are still a lot of different values in each attribute. By trying using WEKA for the predictions, these much different categories make the result looks in a mess. Hence, select and keep the most common categories become very important at this time.

For the first prediction, there are 22 values in the attribute "Action Prior to Accident" by calculating. Since some of the values in this attribute are similar, merging the values is using the way of "Replace" by Excel Tool. For example, replacing "Making Left Turn", "Making Right Turn", and "Making U-Turn" into "Making Turn", replacing "Entering Parked Position", "Parked" and "Starting from Parking" into "Parking/Parked", replacing "Slowing or Stopping", "Starting in Traffic" and "Stopped in Traffic" into "Slowing in Traffic", etc. By doing this step, there are 11 values are kept for the future predicting.

For the second prediction, the different values are much more than the previous one, and most of them have fewer relations with others. Thus, merging the similar values is not a good choice at this time. For the purpose of reducing different values, R Language is a nice tool for dealing with the problem. By using R, it is showing that there are 51 different values in this attribute. After calculating the percentage of each category, the top 10 of them are kept, which are weight about three-fourths of the total percentage. Therefore, the instance is enough for the predicting.

After these data preprocessing steps, the remaining data is now good for predicting.

V. MODEL BUILDING

As discussed in the previous sections the total attributes now to deal with to design a model is:

1. Vehicle body type
2. Direction of travel
3. Number of occupants
4. Vehicle Make
5. Contributing factor

Using the 5 attributes, the two predictions were made:

1. Action Prior to accident
2. Contributing Factor Description

To build a model, WEKA is used as a tool. As the data is labelled, classification methods are used to build the model.

Following three different ways the model is built:

1. Using J48 decision tree algorithm
2. Using Naïve Bayes in Bayes Algorithm
3. Using JRip in Rules classifiers

The results got from above three methods were most accurate, and hence the results from another method like Stacking or Random Forest is scrapped.

The prediction model is built in 2 ways:

1. The two predictions are made independently using the same base data.
2. Using the filtered attributes after pre-processing, first Action Prior to accident is predicted. Then using the first prediction, the contributing factor 1 description is predicted.

The results for second prediction obtained were better in model 2 as compared to model 1, since the extra data is used to make the prediction. However, this better prediction cannot be surely said to be better since in the first prediction the accuracy obtained was not 100% and hence this would affect the second prediction in model 2. Both the models are discussed with their respective methods in following sections.

Data Proportion

The original data contains around 1 million entries. Obviously, these many records cannot be processed on WEKA. Also, the data had so many missing values and the values which were not relevant to the results, so that part of data was removed before processing.

As mentioned in the data pre-processing section, some of the attributes were removed straight way before processing. The total attributes that are considered are as follows:

1. Vehicle body type
2. Direction of travel
3. Total number of occupants
4. Vehicle Make
5. Contributing factor 1

The total number of records after preprocessing reduced to 337214. WEKA still could not process this data, so for model building, some percentage of data is considered. Before model building, a data is passed through unsupervised "RemovePercentage" filter to remove percentage of data randomly and select the remaining one.

The models built by JRip, J48 and Naïve Bayes are built on only 20% data. To study the accuracy of all the methods, all the models were built on same amount of data.

VI. RESULT ANALYSIS

The accuracy obtained for Action Prior to Accident by the different model building algorithms is given below:

Result\Method	J48	JRip	Naïve Bayes
Correctly Classified Instances	*62.05*	62.16	61.36
Kappa Statistics	0.1741	0.1777	0.1756
Confusion Matrix Predicts instances	2/11	3/11	9/11

Table 1: Prediction Model 1 Results

Stratified Cross Validation Process with 10 folds is used. 80% data is used for training and rest is for testing.

The last column of the above table shows how many distinct values the model can predict. Before prediction, the percentage distribution of 11 different 'Action prior to accident' was as below:

1	Avoiding Object in Roadway	0.53823388
2	Backing	5.37522167
3	Backing Unsafely	0.56996447
4	Changing Lanes	4.14365952
5	Going Straight Ahead	58.36145593
6	Making Turn	15.41988174
7	Merging	1.15534942
8	Overtaking/Passing	1.17314228
9	Parking/Parked	1.64761843
10	Police Pursuit	0.02609619
11	Slowing in Traffic	11.58937648

Fig 1: Action Prior to Accident Event Distribution

Since decision trees like J48 goes with the information gain, J48 output has only two predictions, Going Straight Ahead and Making Turn. In association Rule mining, the confidence is used to make predictions and hence it has one more prediction, which is parking/Parked. In predictions of Naive Bayes, since all attributes are considered independently, 9 out of 11 events existed.

Accuracy and Kappa Statistic seen is almost same in all three models. Yet whenever the data distribution is highly unequal, Naïve Base model is better as compared to other two.

Why not use Under-sampling or Over-sampling to improve accuracy?

In most of the datasets, when building a model, under-sampling or oversampling is used to improve the accuracy of model. The practice is common whenever the data is imbalanced.

The distribution of different entries in the Action Prior to Accident before the model building is shown below:

```
# A tibble: 11 x 2
   Action.Prior.to.Accident  percentage
                      <fctr>       <dbl>
 1       Going Straight Ahead 58.36145593
 2              Making Turn 15.41988174
 3        Slowing in Traffic 11.58937648
 4                  Backing  5.37522167
 5           Changing Lanes  4.14365952
 6           Parking/Parked  1.64761843
 7       Overtaking/Passing  1.17314228
 8                  Merging  1.15534942
 9           Backing Unsafely 0.56996447
10 Avoiding Object in Roadway 0.53823388
11           Police Pursuit  0.02609619
```

Fig 2: Initial Distribution of Target Variable

From the figure, data is very much imbalanced, but it is realistic. Replication the data which is less (Oversampling) or removing the data with higher percentage (Under-sampling) will give better accuracy but in real life scenario those results might not be true. The reason behind this is, accident causes are always unbalanced, and some of the reasons are more frequent than the others. To build a model which can be realistic, the model is built on the same proportion of data. And using Naïve Bayes, the results obtained were satisfactory and had predictions about almost every event that is mentioned in original data on which model is built.

Prediction of Second Attribute Without using the first prediction:

For this purpose, same attributes which were used to make first prediction are used. Here also Stratified Cross Validation Process with 10 folds is used. 80% data is used for training and rest is for testing.

Here, data is equally distributed and hence decision tree J48 did better than previous. The J48 algorithm output contains all the events that occurred whereas the JRip fails to yield all. The accuracy and Kappa Statistic is also less for the JRip. Naïve Bayes did equally in both predictions.

Result\Method	J48	JRip	Naïve Bayes
Correctly Classified Instances	*37.11*	34.02	36.85
Kappa Statistics	0.24	0.19	0.23
Confusion Matrix Predicts instances	10/10	5/10	10/10

Table 2: Second Attribute Prediction Accuracy

Initial Distribution:

1	Animal's Action	12.339938
2	Backing Unsafely	6.745272
3	Driver Inattention/Distraction*	17.769428
4	Failure to Yield Right-of-Way	16.792007
5	Following Too Closely	19.007515
6	Passing or Lane Usage Improper	4.444952
7	Pavement Slippery	5.765775
8	Reaction to Backing Unsafely	3.553826
9	Unsafe Lane Changing	3.692611
10	Unsafe Speed	9.888676

Fig. 3: Distribution of Contributing Factor 1 in Original Data(Top 10 Attributes)

Based on results, the model based on JRip is better for first prediction and model based on J48 is best for second prediction. Naïve Bayes works in both way, and can be used as alternative model in both cases.

Using Predictions from First Model to Predict Second attribute:

This method is used to improve the accuracy and J48 and Naïve Bayes method is used to see how much the accuracy improves since J48 was the best method for second prediction.

To do so, predictions for first attribute 'Action Prior to Accident' are saved from WEKA and this extra attribute is added to the data before making predictions for Contributing Factor 2 Description. Before starting the model building, the data is checked on R using Logistic Regression for whether the added new attribute is having any importance in predicting the class. The result obtained had very less P value for Action Prior to Accident-Making Turn, which had obtained from prediction model 1. So that attribute was used to make predictions for second attribute.

As JRip is not that good for making these predictions, J48 and Naïve Base is used. The results obtained are shown below:

Result\Method	**J48**	**Naïve Bayes**
Correctly Classified Instances	40.98	41.13
Kappa Statistics	**0.2964**	**0.297**

Table 3: Accuracy in Predictions of Second Attribute with Use of Model 1 Result

Note that here, for J48 method, results from J48 first prediction are used and for Naïve Bayes, results from Naïve Bayes from first prediction are used.

So, accuracy increased as the extra attribute is added. So the accuracy of second Prediction depends on accuracy of the first one. Any methods as per the data requirement can be used to make first prediction. Same is true for the second prediction.

VII. CONCLUSION

During the pre-processing of the original data, the Excel filter were used to remove the missing values. Then using the tools include R language and Excel Tools to merge the categories of data. After the final data being prepared, the next step is using WEKA to build three models for the prediction, which includes J48, Naïve Bayes, and JRip.

Though data is imbalanced, the accuracy with Naïve Bayes algorithm achieved in the first model. Whenever a data is imbalanced, Naïve Bayes always give better results.

The accuracy of second model may be lower, but Kappa Statistics achieved is almost 0.3, which means the model is satisfactory. Kappa Statistics considers the agreement by chance, and hence it is a better predictor for how good the model is. If time and location are provided in data, the model will be more accurate, comparing the results are differ from including the first model or not.

The predictions of this project can be mainly helpful for insurance companies in deciding the actions of individuals before accident. Using the result from the models built in the analysis, the insurance companies can decide whether offer or deny the insurance.

This project predictions will also be helpful to polices for better resolving the disputes after the accident. After the government or department gain the reliable amount of records information in accidents, it would support the Traffic accident responsibility identification for further investigation and research.

REFERENCES

[1] Beshah, T., & Hill, S. (2010). Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. AAAI Spring Symposium: Artificial Intelligence for Development.

[2] Pakgohar, A., Tabrizi, R. S., Khalili, M., & Esmaeili, A. (2010). Role of human factor in incidence and severity of road crashes based on the CART and LR regression. Retrieved October 19, 2017, from http://www.sciencedirect.com/science/article/pii/S1877050910005016

[3] Kumar, S., & Toshniwal, D. (2016). A Data Mining approach to characterize road accident locations (2nd ed). Retrieved October 19, 2017, from https://doi.org/10.1007/s40534-016-0095-5

[4] Chang, L. & Chen, W. (2005). Data Mining of tree-bases models to analyze freeway accident frequency. Retrieved October 19, 2017, from

http://www.sciencedirect.com/science/article/pii/S0022437505000708

[5] Tseng, W., Nguyen, H., Liebowitz, J., & Agresti, W. (2005). Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. Industrial Management & Data Systems, Vol. 105 Issue: 9, pp.1188-1205. Retrieved October 19, 2017, from https://doi.org/10.1108/02635570510633257

[6] Road Crash Statistics. (n.d.). Retrieved November 28, 2017 from http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

[7] Krishnaveni, S., & Hemalatha, M. (2011). A Perspective Analysis of Traffic Accident using Data Mining Techniques. International Journal of Computer Applications (0975 – 8887) Volume 23– No.7.