International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

BIG DATA FOR PREDICTIVE ANALYTICS OF TRAFFIC VIOLATIONS

Ambar Raghuvanshi[*1] & Akash Raghuvanshi[2]
[*1]ZS Associates Pvt. Ltd., Evanston, Illinois, USA
[2]University of North Carolina, Charlotte, North Carolina, USA

DOI: 10.5281/zenodo.3609528

ABSTRACT

This paper produces observations of the impact of the traffic offence on traffic accidents by using big data analysis techniques dependent on traffic violations information that was updated daily in Montgomery County within the USA. By knowing the reason for traffic accidents, the aim was to recognize infringement. Also, it conceives to use the data to compute protection premiums for insurance agencies. The aimed hypothesis is to predict future violations leading to an accident, predict accident fatality, i.e., personal injury or property damage. Additionally, it attains to utilize the data to calculate insurance premiums for insurance corporations. Several classifications, clustering and regression models are considered in our analysis, like Single Tree, Random Trees, K- Means clustering, multiple regression, and Naive Bayes. The first and second model considers Random Trees and Single Trees as the best algorithms per our business case due to the importance of high sensitivity and a high F1-score. Model III considers K-means clustering and Single Tree classification the best algorithms for having the ability to produce clusters with their numerous violation types and count of injuries.

1. INTRODUCTION

Numerous research and industrial sectors have taken help of different data analytics and mining techniques, during the last decade. [1], [2], [3], [4], [5] . It is a challenge for modern-day scientists to stretch the coverage of those techniques. It is now a common practice to utilize data mining for traffic regulation and safety [1], [2], [3], [4].

The road accidents pose a giant worldwide threat that continues to cause losses, injuries, and casualties on roads leading to an enormous impact at the socio-economic levels. Worldwide 1.27 million casualties and up to fifty million injuries are a result of road accidents annually [5]. Hence, such a global problem needs additional attention to the reduction of frequency and severity of accident prevalence.

The historical data concerning previous traffic violations represents an arduous chance for researchers to acknowledge the essential factors in such violations. The main difficulties in deducing information from this data are its large size and high dimensions [1], [2], [3]. Lately, for the extraction of useful in- formation from massive data sets concerning traffic accidents, a variety of data-mining techniques are efficiently utilized [2], [4], [6]. Road traffic accident and violation researchers widely utilizes data classification for mining. The primary purpose of these strategies was to construct classifiers for prediction of new accidents and their severity.

In this work, we tend to perform our analysis of dataset 'Traffic Violations' following public Cross-Industry standard process for data mining Figure 1 , which permits placing data mining problem into the general problem resolution strategy of a research unit [9], [8]. The project analysis tasks were developed within the project assignment description[...]. The main tasks are, choosing a dataset which is accessible on the internet, study it and draw preliminary conclusions from it, following formulation of the initial hypotheses based on data analysis and development of the business case. The other necessary tasks of the research are exploring dataset and producing visualizations using Tableau, review the hypotheses based on visualizations and produce dataset(s) satisfying the created hypotheses. The ultimate task is to devise three mod- eling techniques for developed hypotheses, implement three algorithms for every model, and provide strategic recommen- dations supported by Based on the CRISP-DM framework and project analysis tasks, research and data understanding, visualization is

per- formed. Following that, data preparation and modeling phases are completed. In line with the assigned analysis task, the modeling part has been completed by applying various well- known data-mining techniques. Besides, the comparison of the efficiency of many data-mining algorithms for the initially developed hypotheses has been done. Finally, Validation and deployment phases are conducted per the CRISP-DM framework. Alongside, the strategic recommendations have been developed.

2. RESEARCH COMPREHENSION

The objective of the project is to investigate Traffic Violations data of Maryland State to enhance awareness of Mary- land Capitol Police (MCP) and to assist them in diminishing the number of violations supported by their location, injury, casualty rates, and variety of property damage.

On the other hand, MCP can share the violation information of out of state drivers with their respective traffic regulation county. MCP hosts the data in a public domain at data.montgomerycountymd.gov [10] to create awareness and provide data enthusiasts primary data to apply science for a safer commute via roadways. We tend to believe that it is necessary to advise Maryland Capitol Police regarding the number of violations taking place at various locations supported by the information and the frequency of specific violations overtime.

Moreover, we can predict the probability of violations arising in an accident that involved injury or property damage. We can predict violations susceptible to the specific season, by uncovering facts like during which season the most violation happened alongside if it involved any personal injury or harm to property. Finally, we can analyze a portion of violations concerned business vehicles. This information can provide suggestions for modification of traffic rules and regulations for commercial vehicles solely. MCP can utilize the analysis which leads to the identification of concerned areas in traffic management and exercise control over them to bring down the number of violations.

3. DATA COMPREHENSION

The chosen dataset “Traffic Violations” was obtained from https://data.montgomerycountymd.gov [10] and updates on a daily frequency. Every electronic violation occurred between 2012-2018 is recorded in the dataset by Montgomery County of Maryland state. This data provides information regarding all the traffic violations in the county, their impact, and also the place where they took place. Moreover, it provides data associated with the vehicle and its driver. To seek out the variables that have the best effect on the accident, we created three research questions by analyzing the attributes of the data set shown within the data search and created a model for verifying the question. Prior to modelling, data search evaluates the impact on accident of parameters like the seat belt, influence of alcohol, hand-held devices, type of vehicle. We devised models to verify the research questions using the algorithms, that big data techniques of tenutilize.

We derived the subsequent initial analysis questions/hypotheses through the dataset.

- Future violations can be predicted using attributes like the probability of accidents, and geo-coordinates of the location using the traffic violations data collected in half a dozen years. This prediction would be helpful for transport authorities, who can revise and reform the present laws of traffic control to enhance traffic safety.
- The chosen dataset can be analyzed to get the data toregard- ing the estimated number of future violations that may result in personal or fatal injury and property damage. Such information is useful for insurance firms, who can provide their quotes supported by the historical statistical data.

Traffic and law enforcement agencies optimize their operations considering the violations, charges, and their frequency.

The primary research questions subjects to update the following data pre-processing and visualization steps. Modelling using different algorithms would be implemented based on these hypotheses.

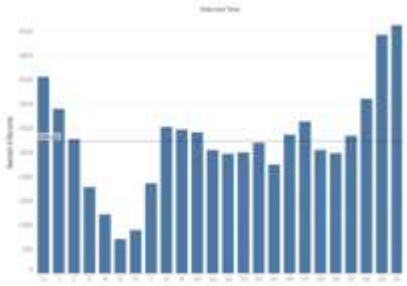
The primary research questions subjects to update the following data pre-processing and visualization steps. Modelling using different algorithms would be implemented based on these hypotheses.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

4. INFORMATION VISUALISATION

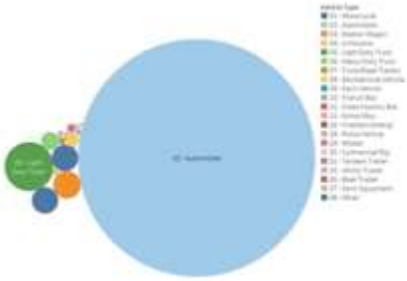
Many visualizations are developed to extend the understand- ing of the initial dataset, communicate insights and fascinating findings. Tableau, a popularly known data analysis tool, performs the required visualization.

Violations versus Hours of the day: Number of violations observed shows that a considerable part of violation took place during midnight compared to any other time of day. Also, during the morning wee hours, it exceeds the average number of traffic violations. The visualization indicates that measures must be in place to take extra caution during these specific hours Graph 1.



Graph 1. Traffic Violations at different hours of the day between 2012- 2017[11]

***Violations per Vehicle Type*:** The packed bubble chart depicts larger the circle, the more the number of violations caused by that vehicle type. Here automobile causes the most violations followed by the light truck on the second spot. Graph2



Graph 2. Traffic Violations by the different type of vehicles between 2012- 2017 [11]

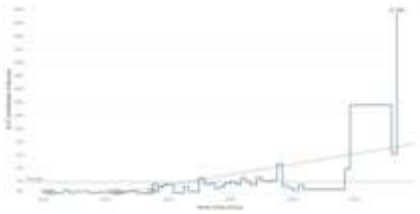
***Traffic violations per seasons*:** The treemap graph repre- sents the rate of violation per four seasons takes place in the USA, where darker the color higher the rate of violation. Theaverage number of violations notices no significant change be- tween the different season. The maximum violations occurred in fall season followed by winters. Graph 3

Graph 3. Traffic Violations summarized at season level between 2012- 2017[11]

***Violations due to Child*:** This line graph shows the number of violations that include the presence of a child.

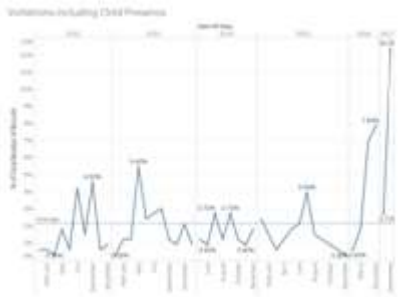
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Based on the description these violations are a result of improper handling of a minor while driving. With a slight dip in 2014 and 2015, it had again regained ground with an all-time high in 2017. Graph4



Graph 4. Traffic Violations that included the involvement of a child between 2012-2017 [11]

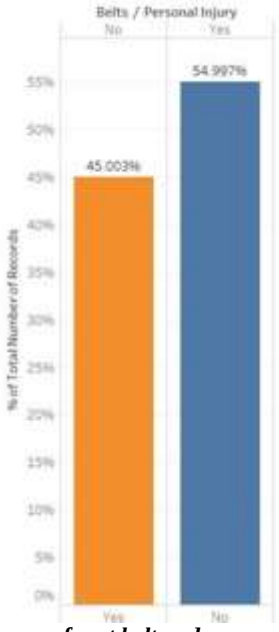
***Violations due to Phone Usage*:** The graph indicates that there is a constant increase in the number of violations caused by the usage of a hand held telephone. The trend line shows that after being below average until 2013 the percentage of total violations maintained a constant upward path which is above the average percentage of violations caused overall. Usage of the handheld telephone has contributed to 29% of violations atthemaximumin2017.Graph5



Graph 5. Traffic Violations that involved the usage of a handheld phone between 2012-2017 [11]

***Personal Injury versus Seat Belt*:** The Visual aid depicts, that seat belt plays a significant role when it comes to whether a violation will result in personal injury or not. Looking at the graph, out of all violation which resulted in injury, 45% of records do not have the seat belt on. It emphasizes on the importance of seat belt during driving and MCP can consider seat belt enforcement. Graph6

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



Graph 6. The relation between usage of seat belt and personal injury between 2012-2017 [11]

The below hypotheses can be conferred once analyzing the visual aids:

- The number of violations involved in phone usage has risen to an extreme level, and the trendline indicate sith as been continuously increasing over the years. What possible measures can be taken by MCP to reduce this trend further?
- Between hours 22:00-01:00 and 08:00-09:00 the number of violations is high higher than the average. What are the main reasons for that and what measures can be taken by MCP to reduce this number?

A close analysis of information and above graphs conjunction with primary research questions produces below three hypotheses:

1) Predict a violation's likelihood to result in an accident, given a set of predictor variables.
2) For given values of predictor variables, predict if the violation led to an accident is fatal.
3) For a given combination of factors, predict the geolocation, where a particular violation is probably going to take place, additionally identifying the locations with the highest number of specific violations.

Data preparation task should be completed, to perform the modelling of these hypotheses. It allows the introduction of the target variables, utilization of predictor variables and treatment of biasing inside the data set.

5. DATA PRE-PROCESSING

A. In consistent data

As a part of data preprocessing, we observed that few a variable has the same value for every record which makes these variable unusable from data analysis viewpoint. These variables are '*Agency*' which represent the agency created the citation, warning or notice and '*Accident*' which indicates accident occurrence at the time of the stop. In case of *'Agency*' since MCP collects all the data, it has only one value whereas '*Accident*' which should have Boolean values 'yes/no' has only one value for all the record, i.e. 'no.' Considering these variables as single-valued, they are removed from the dataset.

B. Missing Data

From data wrangle point of view identifying and managing nulls is a vital step. This can be treated by one of the few manners which are ignored/removing the tuple or adding value manually/automatically. These values can be the mean of an attribute or any constant or most probable value calculated using Bayesian formulae. XLMiner identifies records with null values and processes them. Nulls composed less than one percent of the primary data; thus XLminer discarded them.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

C. Normalization/Standardization

As there was no need for the stable convergence of weight for attributes to support our analysis and hypothesis and most of the variables in the dataset are categorical, the process consumes dataset as it is.

D. New Attributes

Based on the understanding of dataset, there arises a need for inducing new variables. We have introduced three bi- nary variables namely, 'Phone_usage' as a predictor vari- able whereas 'Contribution_to_Accident' and 'Fatal' as the target. These variables have a value '1' for all true cases and '0' for false based on 'Description' attribute value. Like 'Phone_usage' is included in the dataset and is '1' where ever the 'Description' attribute for a record indicates usage of a handheld device otherwise'0'.

E. Outliers

In the initial dataset, outliers were detected and replaced by imputing or deleting the record. The vehicle manufacture year attribute, i.e. 'Year' has a value below 1953 and beyond 2018 for some records. In these cases, all records for 'Year' value less than1953 were replaced by the constant value of 2000 whereas for greater than 2018 records were deleted as it represents data a head of the research performed a year.

F. Data Filtering

Unbiased data is a prerequisite for precise prediction. Filtering can be applied, to attain this goal. We used R script which filters and transforms data into an unbiased dataset.

The primary dataset is filtered to predict target variable '*Contribution_to_Accident'* of the hypothesis where all records with value true for it in conjunction with an equal number of records with value false are selected. Hypothesis three again used the same subset. For hypothesis second the identical process of subset selection repeats, with '*Fatal'* as the target variable.

G. Bins and Dummy variables

For any data analysis method which incorporates classifi- cation, binning the data into intervals and dummy variable creation are essential tasks [9]. Many algorithms pick binned categorical variable over continuous numerical ones. [9].Here, binning remodels variable '*Year'* into a categorical variable with ten equal intervals. For attributes '*Belts*,' '*HAZMAT*,' '*Commercial_Vehicle*,' '*Alcohol*,' '*Phone_usage*,' and binned '*Year'* dummy creation is performed to achieve categorical variables valued '1' and '0' based on the description Appendix A.

6. MODELLING

Modeling section was completed using different modeling techniques and investigating the performance of many classification algorithms in predicting the contribution to an accident, accident fatality, and geolocation, wherever a violation is probably going to take place based on specific predictors chosen from traffic violations records collected by Maryland State Police over the six-year period from 2012 to 2018.

A. ModelI

To predict '*Contribution_to_Accident*' for a particular situ- ation, model I uses three algorithms, i.e., Single Tree, Ran- dom Trees, and Naive Bayes. The model I uses the tar- get variable '*Contribution_to_Accident*' generated during data pre-processing. Categorized variables '*Belts*,' '*Alcohol*,' and "*Phone_usage'* along with binned variables '*Vehicle_Type'* and '*Year'* are used as predictors. The complete tree shows that the best predictors for this model are '*Belts'* and '*Phone_usage'* variables. The performance of the developed model can be evaluated by comparing the parameters shown in Table I.

Table I performance parameters readings for model i obtained using xl miner (percentage) [11]

	Precision	F1-score	Specificity	Sensitivity
SingleTree	62.6	38.3	83.6	27.5
RandomTrees	50.3	66.9	2.3	99.6
NaïveBayes	62.1	37.3	83.8	26.6

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

The table shows that the single tree algorithm has the very best precision. Whereas, Random Trees algorithms have very best sensitivity and F1-score.

It is essential to reollect the meaning of the above parame- ters (precision, sensitivity, specificity, and F1-score) following standard evaluation metrics [4], [6], [7], [5], [12], to assess the performance. We thought-about different performance eval- uators to assess the model and its outcome based on the prediction requisite. Per Table I, the Random Trees algorithm has the highest sensitivity, which shows the proportion of con- tribution to accident cases that are appropriately recognized. When modeling for prediction of '*Contribution_to_Accident*' sensitivity becomes a vital parameter. Informing about the proportion of correctly identified '*Contribution_to_Accident*' casesrandomtreealgorithmwithmaximumF1-score,provides the most effective conjunction of preciseness and sensitivity. At the same time, Single Tree algorithm has the highest preci- sion, which becomes more helpful for general case prediction performance.

B. Model II

Model II uses the same set of algorithms as in model I to predict if the accident is fatal for a provided scenario. Here, the target variable is'*Fatal*.'

In the beginning, alongside the predictor variables are chosen in model I, '*Vehicle Type'* is chosen. Model II uses identical performance evaluators as the model I Table II following a similar analysis. The single tree algorithm notices the best sensitivity and F1-score parameters for model II whereas, Random Trees algorithm has the best precision for model II.

Table II performance parameters readings for model II obtained using xl miner (percentage) [11]

	Precisi	F1-Sco	Sensitivit	Specificit
Single Tree	57.9	66.8	y 79.0	y 40.4
Random Trees	63.9	47. 8	38.3	77.6
Naïve Bayes	57.9	56. 1	54.3	59.0

The full tree shows that the prime predictors for this model are binned *'Year'*, and '*Alcohol'* variables. These results can be possibly obtained because of an outsized number of dataset records, that belong to the ninth and tenth binned year interval. Based on our initial prediction tasks stated, the obtained classification is not very helpful, as a result of it does not provide excellent insights and awareness concerning specific situations that can result in fatal accidents. At the same time, the years of vehicles cannot be used by MCP to take specific measures to reduce the number of accidents.

Variable '*Year'* is neglected to boost the prediction power of model II. Due to biased data, '*Commercial Vehicle'* seems to be the primary predictor of '*Fatal'* in the achieved best- pruned tree. Since '*Fatal'* does not rely upon the '*Vehicle Type'* as per the prediction goals, it is removed. After repeating the modeling process, variables '*Alcohol*,' '*Phone usage*' and '*HAZMAT*' seems as top predictors in the full tree. MCP can become more vigilant and exercise control on specific violations utilizing the outcome of model II. Model II can be assessed utilizing identical performance predictors as in model I, TableIII.

Table III performance parameters readings (update) for model II obtained using xl miner (percentage) [11]

	Precisi	F1-Sco	Sensitivity	Specificit
Single Tree	50.9	67.5	100.0	y 0.0
Random Trees	51.9	67.0	98.0	5.7
Naïve Bayes	50.0	12.0	7.0	92.0

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

C. Model III

In model III, the dataset has been clustered using K-Means clustering and choses '*cluster_ID*' as a target variable. Following that, prediction variables are assigned to figure out '*Fatal'* for a scenario. The same algorithms in conjunction with multiple linear regression as in model I conducts modeling here. Model III utilizes a target variable '*cluster_ID'* to spot the most sought-after cluster for a set of predictors.

The generated prediction model using multiple linear regression algorithm had a bit adjusted R-squared value. Random Trees and Naïve Bayes no longer remain unbiased per the result. Whereas, a single tree algorithm which is being able to perform classification on the formed clusters, shows more determined results. For model III, single tree algorithm is chosen for quality predictions because it features a single node best-pruned tree as an outcome.

The full tree indicates classification score performance for model III. In cluster five, performance indicates that there occurred some violations under the influence of alcohol along personal injuries on the other hand cluster nine received multiple seat belt violations besides personal injuries.

7. CONCLUSION

It concludes from the above-performed research that '*Belts'* and '*Phone_usage'* are two critical parameters. Negligence to them resulted in most violations and further into accidents. Also, '*Phone_usage'* increased chances of an accident being fatal resulting in casualties. Besides, alert notification should be issued to MCP about the significant regions where violations are likely to be caused. Cluster five and Cluster nine are areas where alcohol violations with personal injuries and multiple belt violations with injuries respectively.

MCP alongside traffic control can work on raising awareness around these variables. Additionally, to boost traffic safety, it can opt to reform the existing laws of regulation around the variables.
Similar recommendations can be given to the insurance corporations to think about their policies for drivers in analyzed areas.

REFERENCES

[1] J.Abellan, G. Lopez, D. O ~na, and J., "Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with," Applications,vol.40,pp.6047–6054,2013.
[2] L. Y. Chang and J. T. Chien, "Analysis of driver injury severity in truck- involved accidents using a non-parametric classification tree model," Safety Science, vol. 51, no. 1, pp. 17–22, 2015.
[3] W.H. Chen and P. Jovanis, "Method for identifying factors contributing to driver injury severity in traffic crashes," Transportation Research Record, pp. 1–9,2012.
[4] A. T. Kashani, R. Rabieyan, and M. M. Besharati, "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers," Journal of Safety Research, vol. 51, pp. 93–98, 2014.
[5] D. O ~na, L. J., G., and J. Abellan, "Extracting decision rules from police accident reports through decision trees,"Accident Analysis & Prevention, vol. 50, pp. 1151–1160,2013.
[6] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," Accident Analysis and Prevention, vol. 75, pp. 1–15, 2015.
[7] Y.Xie, Y. Zhang, and F. Liang, "Crash injury severity analysis using Bayesian ordered probit models," Journal of Transportation Engineering ASCE, vol. 135, no. 1, pp. 18–25, 2009.
[8] S. M. S and R. S. T, Loan Credibility Prediction System Based on Decision Tree Algorithm, 9 2015. [Online]. Available: http://dx.doi.org/10.17577/IJERTV4IS090708
[9] "Discovering Knowledge in Data: An Introduction to Data Mining," D. T. C. D, Ed. Larose: Wiley, 2nd edition: Wiley.
[10] "Data Montgomery." [Online]. Available: https://data. montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q
[11] A. Raghuvanshi, D.Mehta, R.Bikmetov, J.Park, S.Pothina, and S. Narsaraj, "Term project for Big data analysis for competitive ad- vantage. UNCC," 2018.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

[12] M. O. Mujalli, O ~na, and J., "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks," Journal of Safety Research, vol. 42, pp. 317–326, 2011.