IT in Industry, Vol. 7, No.3, 2019

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

An IOT based Accident Severity Prediction Mechanism using Machine Learning

Aditya Verma
Department of Comp. Sc. & Info. Tech.,
Graphic Era Hill University, Dehradun, Uttarakhand, India 248002,
vermaaditya903@gmail.com

Abstract
The significant number of fatalities and serious injuries caused by traffic accidents around the world is a worrying problem. Developing nations typically bear a heavier weight of casualties. As a result, developing a model to forecast the likelihood of accidents is extremely difficult. However, the application of machine learning algorithms is one of the significant techniques to forecast the seriousness of such events. As a result, the main goal of the suggested thesis is to automate the process of accident detection by evaluating the severity levels and filtering a set of influential factors that could cause a road accident and generating them using IoT. SMOTE's theoretical notions are put into practice in order to address data imbalance and to ensure that the dataset is balanced. In a later step, the dataset is put to use in the process of building a framework that is constructed from five machine learning algorithms and one stacking algorithm. In the final step of the process, a study is conducted using variables such as the state of the weather and the varying degrees of severity that can have a role in the occurrence of traffic accidents. According to the findings of the experimental analysis that was carried out as part of the research project, the random forest model generated a higher level of accuracy than any of the other models that were put into use, achieving 74%.

Keywords: Accident prediction, machine learning, severity, SMOTE

Introduction

Accidents on the road are among the most unwelcome and unanticipated of all the things that might happen to people who use roads. In recent years, a significant number of accidents of this kind have been reported from a variety of locations across the globe. The occurrence of fatalities and injuries has had a significant effect on the economy of the nation as a whole. Not only has this resulted in early deaths, but it has also led to a loss of property and damage on societal levels. According to a study that was carried out by the WHO in the year 2017, [1] it was found that there are 1.5 million people who lose their lives as a result of vehicle accidents and other forms of vehicular violence. In addition to this, they indicated that the number of deaths is more likely to increase by the year 2030 due to the neglect in traffic regulations. According to the numbers provided by the poll, there were 47 people who lost their lives on the roads each and every day, which resulted in a reduction of 3 percent in GDP. According to the findings of a second survey conducted by Michigan Traffic [2], the state of Michigan recorded an estimated total of 314,921 fatalities in 2017 that were the result of car accidents. The calculation resulted in a loss of 230 billion dollars and a significant decline in the economy of the country. Such disastrous numbers have finally become a topic of escalating concern, not only for government officials but also for study academics and accident experts who belong to the same field. This is the case for all of these individuals. Extensive research has been carried out in order to discover all of the components that would be highly responsible for such figures, and the results have been compiled.

Because of their unpredictability and unanticipated nature, accidents provide a

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

considerable obstacle when attempting to draw conclusions from them. In addition to the observations that have been made, coming to conclusions based on facts and outcomes that are accurate to one hundred percent is an extremely difficult undertaking. The thesis makes the proposal that technical methods of machine learning should be used and implemented in order to recognize the severity of the conditions that could lead to traffic accidents. This would allow for the limitation to be overcome. It is believed that ML is one of the most sophisticated and technically advanced methods that may be used to forecast the incidence of such accidents. As a result, the objective is to conduct research and analysis on five different classifiers in order to construct a classification model with the ability to forecast traffic accidents. In the following step, the classifiers would go through a series of processing procedures and arrive at intelligent decisions by gleaning insights from previous data. A basic theory of SMOTE is presented in the thesis in addition to the implementation of the selected classifiers. According to this theory, the data would be balanced in such a way that the classification algorithms would be able to perform on them in a meaningful manner.

In addition to the exterior elements that were just discussed, there are a number of other factors that have not been identified but are believed to contribute to automobile collisions. These kinds of elements are classified as being of a heterogeneous nature. As a result, situations like these lead to the acquisition of an existing dataset that contains many historical records. This is done with the goal of improving the accuracy of predictions made by the suggested model. The following is a list of some of the reasons why people get into car accidents:

- In violation of the speed limit, a vehicle that is going at a higher acceleration may have a greater likelihood of colliding with another vehicle. In addition to this, driving too fast for conditions can lead to an inaccurate assessment of the road's surface.
- Driving while intoxicated is regarded as the second most influential factor that might lead to the occurrence of a road accident, as it is the most common cause of such incidents.
- Distractions: drivers who answer their phones while operating a motor vehicle are more likely to divert their focus away from the road, which impairs their ability to drive safely.
- The majority of drivers have the bad habit of ignoring traffic signals, which increases the risk of getting into an accident at intersections.

On a consistent basis, individuals lose their lives as a result of being involved in traffic collisions. Because of this, the nation suffers a substantial amount of financial damage. Because of this, it is essential to make certain that particular safety procedures are implemented so that the number of traffic accidents and collisions can be reduced. As a result, this becomes a crucial motivating reason for the suggested study, in which all of the relevant criteria are taken into consideration in order to recuperate a significant percentage of the financial loss. A proactive strategy and activities in real time should also be taken into consideration, in addition to the influential aspects. In proactive approaches, the process of resolving road safety issues begins with an examination of the risks connected with it; in real time operations, however, an analysis of scenarios that might result in accidents is included.

Literature Survey

During the process of doing the literature review, it was discovered that the development of statistical methods had been significantly more put into practice in contrast to the methods of machine learning. Statistical algorithms, on the other hand, fared better with less computer complexity; nonetheless, the methods they utilized to analyse crash studies were very

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

praiseworthy. In a research task that the author [3] offered, he analysed the highway data that resulted in car crashes. This work was done. He accomplished this goal by employing the methods of statistical modelling, which included the process of clustering analysis implementation. Because the dataset had information on the causes so as to why a place was more likely to be involved in accidents, the author used this data to analyse the reasons for highway crashes. The utilization of a clustering model assisted in the categorization of particular factors that contributed substantially to the occurrence of an accident. The author analysed the stages in this manner and later provided methodological guidance so that appropriate safety countermeasures could be performed and the overall severity of the crash might be decreased. In addition to clustering analysis, the K-Nearest Neighbour (KNN) algorithm is regarded as an example of a partially non-parametric statistical model that is capable of carrying out regression operations. The authors of the paper [4] contributed their work toward forecasting the real objects that may be viewed as a primary reason for car crashes. The KNN algorithm was used to collect the relevant objects, and then those artifacts were transformed into their equivalent variables. After that, the variables that were formed only responded to observations if they were within a close proximity to the distance between the values that were generated. As soon as the values had been predicted using KNN, an assumption was made, and a local structure of crash data was developed.

However, in order to put into practice statistical modelling, one must first establish certain assumptions in order to construct a probability distribution. This distribution helped further establish a relationship between the variables that are dependent and those that are independent. It was discovered at the same time that the application of machine learning methods increased simultaneously with that of statistical modelling. Because it was not necessary to make any assumptions about the relationships between the variables, the implementation of machine learning did not call for any models of the underlying mechanisms. Nevertheless, statistical modelling and machine learning did intersect in some circumstances, such as the work that was proposed in [5]. This was because both of the notions dealt with analysis that was carried out on the dataset. This was the reason behind this. On the other hand, a significant distinction between the two approaches was the manner in which they drew inferences concerning the variables. Machine learning did not necessitate the establishment of a relationship between the variables, in contrast to statistical analysis, which demanded that a connection be made between them. The process of machine learning does not need the construction of underlying relationships or the making of any assumptions.

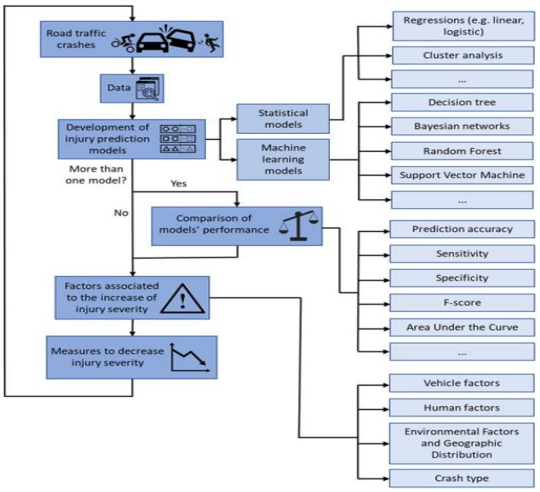

Figure 1: Methodological workflow followed by multiple research scholars

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

One such example of study that was based on machine learning was the work that was proposed by the authors [6]. In this work, the authors evaluated the car accidents that had occurred in all 49 of the states that make up the United States. The data set, which was obtained from the Kaggle repository, consisted of automobile collisions that occurred between February 2018 and June 2020. Because machine learning algorithms function most effectively when applied to enormous datasets, the process of accident detection could be improved by increasing the size of the dataset. Logistic regression, KNN, support vector machines (SVM), decision trees, Naive Bayes, and random forests were the six different classifiers that were utilized. Accuracy and precision were used as evaluation and comparison criteria for the model's overall performance before being deemed satisfactory [7]. In addition to these extraneous factors, such as the availability of ambulance services, we took all of this into account when delivering packages and managing accidents that occurred on the roads. Nevertheless, in contrast to the other six classifiers that were utilized, the SVM gave findings that were more precise and had an accuracy score of 97 percent.

Implementation Details

A Dataset Used

The dataset that was used in the research was obtained from the Kaggle repository. It contains information on automobile collisions that took place between February 2016 and December 2021. The data that is stored in the Kaggle repository was obtained through the use of a number of different traffic APIs. These APIs have a tendency to record accident reports using network sensors and traffic cameras. The dataset contains references to around 4.2 million different reports of automobile collisions. A.CSV file with 47 columns is used to distribute the dataset. This file is used to do the distribution. The headings of each column are highlighted in the table that follows.

Label names of 47 columns			
ID	Source	TMC	Start_Time
End_Time	Start_Lat	Start_Lng	End_Lat
End_Lng	Distance	Description	Number
Street	Side	City	Country
State	Zipcode	Timezone	Airport_Code
Weather_Timestamp	County	Temperature	Wind
Start_Lng	Start_Lng	Start_Lng	Start_Lng
Humidity	Pressure	Weather_Condition	amenity
Pressure	Visibility	Wind_Direction	Wind_Speed
Precipitation	Bump	Astronomical_Twilight	Give_Way
Turning_Loop	No_Exit	Railway	Round_About
Traffic_Signal	Stop	Nautical_Twilight	

B Data Pre-Processing

The raw, unstructured, and imbalanced data that was received from the Kaggle repository are what make up the dataset that was gathered. This data has a tendency to raise the computational complexity of the system, and it also adds on extra training time that is not necessary to the dataset. As a result, we need to get rid of this redundant information in the data. Data pre-processing refers to the process of removing extraneous data from a dataset and plays a vital role in improving the overall computing process of a model. This procedure may be thought of as deleting irrelevant data from the dataset. The data that was obtained from the Kaggle repository were raw, unstructured, and imbalanced; this data is what makes up the dataset that was gathered. This data has a propensity to increase the computational complexity of the system, and it also adds on additional training time that is not necessary to the dataset. In addition, the dataset becomes larger than it needs to be. As a consequence of this, we need to clean up the data by removing any unnecessary or duplicate information. Data pre-processing, also known

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

as the act of deleting unnecessary data from a dataset, is an essential step in optimising the computing process of a model as a whole and plays an important part in achieving this goal. One approach to thinking about this technique is as removing data from the dataset that is irrelevant.

C Data Balance

The data that was collected from the repository is inherently unbalanced, and as a result, it is necessary to rectify the situation before the appropriate algorithms can be applied to it. In order to accomplish this goal, a straightforward method of replicating the minority class is carried out utilising SMOTE. Synthetic Mining Oversampling Technique is what "SMOTE" stands for in its simplified form. The problem of data imbalance in the dataset is resolved by oversampling the minority classes of the dataset; this is done in order to ensure that the generated samples can completely fit into the model and that the dataset is balanced. This kind of replication helps to synthesise the samples that have been made so that they can be matched with the samples that are in the minority.

D Data Train, Test and Split

When the dataset is put through a training and testing procedure, the effectiveness of the model that is subsequently developed may then be determined. In order to accomplish this goal, a collection of algorithms are applied, and a section of the dataset is partitioned. This is done so that the model that has been developed can be improved to the point where it can achieve better levels of accuracy. For this reason, the data is divided in the thesis as follows: eighty percent for the purpose of training, twenty percent for the purpose of testing. In addition to this, the validation strategy can be applied to the model in such a way that it achieves a higher level of resilience. The first step in putting the thesis into action is to resample the unbalanced data using SMOTE so that only a tiny percentage of the data for fatal and serious injuries are duplicated. This step is carried out so that the thesis may be put into action. Following the installation of SMOTE, the model is put through training and then tested and validated using data that has been subjected to 10 fold-cross validation. The data is put through its train, test, and split stages using the algorithms that have been specifically chosen. For the purpose of determining the optimisation accuracy of the model, this thesis makes use of four methods (including one stacking algorithm) in total. In the end, a series of assessment parameters are carried out in order to figure out which algorithm performed better in terms of precision and make a comparison between them.

Experimental Analysis

The confusion matrix that was developed for the suggested thesis was based on the notions of a multi-class classification technique. This technique stipulates that only one output is more likely to be generated from a group of items belonging to a particular class. In light of this, the confusion matrix must include both the actual and the expected values, taking into account all four degrees of severity. However, the outputs that are created in this way must only ever produce a positive prediction for a single severity at a time.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

Figure 2: Confusion Matrix

During the process of putting the suggested thesis into action, a total of four different degrees of severity will be taken into consideration; severity 1 will represent the least severe cause for an accident to occur, and severity 4 will represent the most severe reason for an accident to occur. Because the indexing of a confusion matrix begins with 0, it has been determined that severity 1 will be indexed as "0" for the purposes of this discussion, and so on. When looking at the confusion matrix of random forest, figure 2.1(a), severity "0" yields a forecast of 156982 cases (highlighted in blue) of positive prediction; on the other hand, all the other values in the same column, such as 4803, 19602, and 10225, are wrongly projected by random forest for severity "0." In the same way, random forest accurately predicts 146235 occurrences of positive predictions when the severity is "1," 135765 when the severity is "2," and 118211 when the severity is "3." The rest of the values that have been created by random forest for the severity levels 0, 1, and 3 have all been determined to be incorrect predictions provided by the algorithm.

Conclusions And Future Work

The construction of an automated system that is able to determine the extent of an accident and, in addition, forecast its occurrence is the primary purpose of this part of the thesis. In order to accomplish this, the paper provides an implementation of four machine learning algorithms in addition to one stacking algorithm. Since the data received from the repository was imbalanced and redundant in its nature, this study also incorporates the conceptual theory of SMOTE in order to bring the dataset into better proportion. As a result of doing a literature review, it was discovered that factors such as the amount of alcohol consumed, the circumstances of the weather, the location of the accident, the speed of the vehicle, and any disabilities exhibited by the driver were among those that had a substantial impact on the likelihood that an accident would take place. Consequently, during the entirety of the process of carrying out the thesis, each and every component that could potentially have an impact was carefully considered. On the other hand, it was discovered that random forest produced an accuracy of 78%, making it the most accurate of the four machine learning algorithms. This led to it being ranked as the winner. In addition to this, the stacking method, which merged the ideas of logistic regression and decision trees as estimators and random forest as Meta estimator, provided the greatest accuracy possible, which was 92 percent. The implementation working of the stacking algorithm is therefore chosen to be as the optimised model with the highest generating

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Published Online 30-Dec-2019
DOI: https://doi.org/10.17762/itii.v7i3.811

accuracy. This is because, in comparison to all five methods, it has the highest generating accuracy.

References

[1] Michigan State Police, Michigan Traffic Crash Decade-At-A-Glance, 2018

[2] Pakgohar, A., Tabrizi, R.S., Khalili, M., and Esmaeili, A., (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science, 3, pp.764-769

[3] Yin, C., Xiong, Z., Chen, H., Wang, J., Cooper, D. and David, B., (2015). A literature survey on smart cities. Science China Information Sciences, 58(10), pp.1-18

[4] G. Singh, S.N. Sachdeva, M. Pal Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of Indian highways Advances in Transportation Studies, 45 (2018), pp. 123-140

[5] H. Jeong, Y. Jang, P.J. Bowman, N. Masoud Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data Accident Analysis and Prevention, 120 (2018), pp. 250-261.

[6] Analysis and Study on the Classifier Based Data Mining Methods SN Popat, YP Singh Journal of Advances in Science and Technology| Science & Technology 14 (2)

[7] Efficient Research on the Relationship Standard Mining Calculations in Data Mining SN Popat, YP Singh Journal of Advances in Science and Technology| Science & Technology 14 (2)

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731