International Journal of Advanced Science and Technology
Vol. 29, No. 6, (2020), pp. 116-120

Prediction of Road Accident Severity Using Machine Learning Algorithm

[1]Annie Racheal Rajkumar, [2]Srihari Prabhakar, [3]A Meena Priyadharsini
[1]*Student, Computer Science & Engg., SRM Institute of Science and Technology Chennai, India.*
ar7520@srmist.edu.in
[2]*Student, Computer Science & Engg., SRM Institute of Science and Technology Chennai, India.*
sp9066@srmist.edu.in'
[3]*Assistant Professor, Computer Science & Engg., SRM Institute of Science and Technology Chennai, India. meenapra@srmist.edu.in*

Abstract

Injuries due to road accidents are one of the most prevalent causes of death apart from health related issues. The World Health Organization states that road traffic injuries caused an estimated 1.35 million deaths worldwide in the year 2016. That is, a person is killed every 25 seconds. This calls for the need to analyse road accidents and the factors affecting them and come up with a method to reduce the probability of their occurence. The analysis of road accident severity was done by running an accident dataset through several machine learning classification algorithms to see which model performed the best in classifying the accidents into severity classes such as slight, severe and fatal.. It was observed that logistic regression to perform multilabel classification gave the highest accuracy score. It was also observed that factors such as number of vehicles, lighting conditions and road features played a role in determining the severity of the accident.

Keywords: *Road accident severity, Machine Learning, Classification, Road safety.*

Introduction

A traffic collision, also called a motor vehicle collision (MVC), occurs due to a collision between a vehicle with another vehicle, pedestrian, road debris or other stationary objects. They often result in injury, death, and property damage. The World Health Organisation (WHO) classifies road accidents as a public health issue as it is the leading cause of non medical related deaths globally. Road accidents reported in the UK over the years show a stagnant trend of safety with respect to all types of vehicles. In 2018, a total of 25511 severe injuries were reported of which about 7% were fatal. Although the UK has been battling the crisis by getting into effect several laws and safety measures, severe road accidents have almost always been due to driver neglect and recklessness. This calls for a requirement to analyse and study road accident severity to help combat the increase in road related injuries and deaths.

Road Accidents

Road accidents have proved to be one of the leading causes of severe injury and has been on the increase over the years. With almost double the number of vehicles on the road compared to a few years ago, road accidents have been at an all time high; thus taking a huge toll on health, finance and property. Although various laws and safety measures have come into effect, there is always a probability of an accident occurring due to a variety of reasons. Driver neglect, driver recklessness, road conditions, weather conditions, driving skill and a number of other factors influence the safety of both the vehicle and the surroundings.

Road accident reports in the UK suggest that driver error has been the leading cause of vehicle collision, with the driver failing to look at his surroundings properly. Driver misjudging distance and speed of both

ISSN: 2005-4238 IJAST

same side and oncoming traffic has found to be a close second cause of accidents with about 80% of these collisions occurring on the same side of the road. Driving with poor maneuvering skills, low visibility, loss of control and driving on slippery surfaces also majorly contributed to the occurence of these accidents. With close to about 50000 cases having been reported in the year 2018, a vast majority of these accidents could have been avoided if the driver took the required precautions while on the road.

Accident Severity

Accident severity is determined by the damage resulting from the accident in terms of bodily harm (fatal accidents being the most severe), and helps categorize the collision. Accident severity is directly related to the speed of the vehicle at the time of the collision. Although economic loss and property damage are also abundant with the more severe accidents, physical harm is focused on and taken into account for the sake of uniformity in the accident prediction.

Related Work

There have been works in the prediction of accident severity that have used algorithms such as Random Forest, Naive Bayes, linear regression and other methods to predict the severity of accidents. These methods of road traffic accidents have played a major role in setting up precautionary measures along areas that have been classified as danger zones or potential accident sites.

Road Accident Prediction has been done in various countries using a number of algorithms but one of the biggest issues is the fact that there lies a data imbalance. As all the data collected is of the occurrence of an accident but no record of the absence of an accident. Therefore various methods have been used to perform negative sampling. Another issue is that it is difficult to perform road accident analysis for larger areas. All papers have utilised datasets consisting of only a small area or restricted themselves to a few road segments.

Accident Risk Prediction based on Driving behaviour Feature using XGboost and Cart uses various parameters of driving behavior and are evaluated using which key features depending on correlation to the occurence of the accident is selected. This ensures that only the required features based on contribution to the accident plays a role in prediction and leaves out the redundant measures that have an indirect role to play in the collision.

Using XGBoost to predict the crash using characteristics of collision, time of the accident and the location of the accident and environmental factors showed to have the most accurate results.

For usage of Naive Bayes algorithm it was found that grouping of characteristics into elements such as vehicles, road, human and environment helped get a more accurate result.

I. PROPOSED WORK

Our paper uses a dataset provided on Kaggle containing data collected by the UK government and is made available by the UK Department of Transport. It aims to analyse accident data to make road travel a safer mode of transport for its citizens. We propose to use this dataset to predict the severity of the accident caused due to the various factors that cause it and the conditions prevailing at the time of its occurrence. This will be done by training the data on algorithms such as logistic regression, Naive Bayes and XGBoost classification to see which model performs the best.

Implementation

Data Collection

The traffic accident data was collected by the United Kingdom Government from 2000 to 2016. For our paper, we have used data from 2012 to 2014. As the UK government aimed to improve their fatality rates, they set out to generate one of the most comprehensive datasets on traffic accident data. It contains 33 features and totals to 1.6 million entries. It consists of geospatial features such as easting, northing, latitude and longitude. It also consists of accident severity, number of vehicles and casualties, road features, weather features and details about the police on site.

Data Preprocessing

Data Formatting

The first part is data formatting. The features collected in this study are mostly categorical but not ordinal. However, some machine learning algorithms like logistic regression cannot operate on categorical values directly. They require the input variables and the output variables to be numeric. Therefore, these categorical data underwent label encoding in this study. For example, accident severity has three independent labels including slight, serious and fatal and these were encoding into the values 1, 2 and 3.

Data Cleaning

The dataset contained many unnecessary entries that took up a bulk of the memory. This was rectified by narrowing down to the required parameters that directly factor in road accidents. This included stages such as

1. *Removing duplicate values:* The dataset contained almost 30000 duplicate values which were removed and left us with 430550 unique values to work with.
2. *Converting time to an integer value:* The given date was in the hh:mm format and for this model,
3. *Splitting the data:* The dataset was then split into testing and training sets with a ratio of 75:25.

Table 1: This table contains the positive and negative correlation values of the parameters.

Accident_Severity	1.000000
Number_of_Vehicles	0.075621
2nd_Road_Class	0.062383
Light_Conditions	0.059872
Location_Easting_OSGR	0.037666
Longitude	0.037629
Weather_Conditions	0.023753
2nd_Road_Number	0.021831
Pedestrian_Crossing-Physical_Facilities	0.015904
Road_Surface_Conditions	0.004548
Day_of_Week	0.003434
Year	0.002221
Time	0.001489
1st_Road_Class	-0.000957
Pedestrian_Crossing-Human_Control	-0.003333
1st_Road_Number	-0.008455
Road_Type	-0.021757
Location_Northing_OSGR	-0.034741
Latitude	-0.034828
Number_of_Casualties	-0.064660
Speed_limit	-0.077117
Urban_or_Rural_Area	-0.083904
Name: Accident_Severity, dtype: float64	

Algorithms Used

This data was trained on various machine learning models to observe which model classified the data with the highest accuracy. The algorithms used were:

a) Multilabel Classification using Logistic Regression: Logistic Regression is a classification model that performs predictive analysis. This predictive behavior is required in problem domains like accident prediction, weather outcome, disease outbreaks and more. As logistic regression is normally used for binary classification but our paper consists of multiple labels, multilabel classification was performed.

b) Naive Bayes Classification: This algorithm was chosen as it is very suitable to perform fast classification for a large amount of data which is usually the case with accident data. It uses Bayes theorem of probability to predict the unknown class. It assumes that all the classes are independent of its features and is hence called Naive.
c) XGBoost Classification: eXtreme Gradient Boosting Algorithm was chosen as it is known for its execution speed and is generally the best algorithm for structured data to perform classification and regression.

Results Discussion

The analysis of the dataset containing accident data in the United Kingdom shows that the number of vehicles involved, 2nd road class and light conditions at the time are positively correlated to the severity of the accident. Whereas the number of casualties, speed limit and whether the area is urban or rural have a high negative correlation.

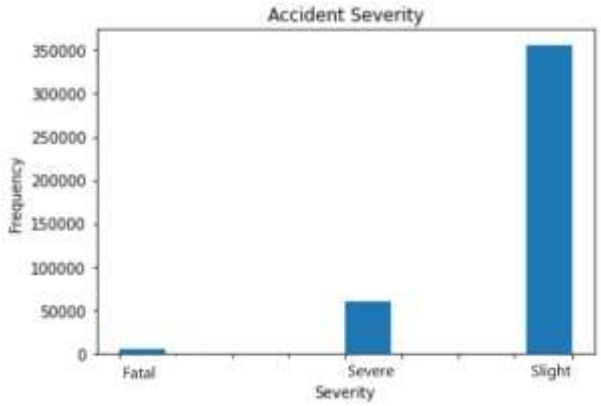

Fig. 1: A bar graph illustrating the frequency of the severity of accidents is given above.

Therefore the data was then trained on a Logistic Regression model to perform multilabel classification. This was achieved with a high mean accuracy of 85%. XGBoost performed the next best and gave a mean accuracy of 84.538% and Naive Bayes gave an accuracy score of 84%.
The regression model also gave a precision of 85 with an F-1 score of 92 having a recall of 1. This means that the model has been trained on acceptable data and provides reliable and accurate results when tested on the testing dataset.

Table 2: This table compares the mean accuracies between the algorithms used.

Algorithm Used	Mean Accuracy Achieved
Multilabel Classification Using Logistic Regression	85%
Naive Bayes	84.483%
XGBoost	84.538%

Conclusion

This paper provides a way to analyse the severity of road accidents and the factors that lead to them. It was observed that factors such as lighting conditions had a high effect on the severity of an accident. Factors like lighting and conditions can be improved upon to make roads safer which can then lead to lower rates of road accidents. Providing a database which contains such a large variety of data such as three classes of accident severity (slight, severe and fatal) and light conditions and details about the police officers at the scene, can be further analysed to provide useful insights and contribute to road safety. Although the occurrence of an accident cannot be controlled, the analysis of this data can enable the government and its citizens to take precautionary steps towards keeping themselves safer.

References

1. M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, Sarawak, Malaysia, Malaysia, 2019, pp. 1-5.
2. Donchenko D., Sadovnikova N., Parygin D. (2020) Prediction of Road Accidents' Severity on Russian Roads Using Machine Learning Techniques. In: Radionov A., Kravchenko O., Guzeev V., Rozhdestvenskiy Y. (eds) Proceedings of the 5th International Conference on Industrial Engineering (ICIE 2019). ICIE 2019. Lecture Notes in Mechanical Engineering. Springer, Cham
3. J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," in IEEE Access, vol. 7, pp. 148059-148072, 2019.
4. Pradhan, Biswajeet & Al-Zuhairi, Maher. "Predicting Injury Severity of Road Traffic Accidents Using a Hybrid Extreme Gradient Boosting and Deep Neural Network Approach", 2019
5. Shi, Xiupeng Wong, Yiik Diew Li, "Accident Risk Prediction Based on Driving Behavior Feature Learning Using Cart and XGBoost", Transportation Research Board 97th Annual Meeting , 2018-1-7
6. Ramya, S. & Reshma, SK & Manogna, V. & Saroja, Y. & Gandhi, Gaurav ." Accident Severity Prediction Using Data Mining Methods." International Journal of Scientific Research in Computer Science, Engineering and Information Technology , 2019.
7. Ting CY., Tan N.YZ., Hashim H.H., Ho C.C., Shabadin A. "Malaysian Road Accident Severity: Variables and Predictive Models." In: Alfred R., Lim Y., Haviluddin H., On C. (eds) Computational Science and Technology. Lecture Notes in Electrical Engineering, vol 603. Springer, Singapore , 2020