International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 937-942 www.ijaem.net ISSN: 2395-5252

Road accident analysis and prediction

Er. Ankit khare

UG student of Department of Information Technology, Shri Ramswaroop Memorial college of Engineering and Management Lucknow, Uttar Pradesh, India

Submitted: 10-04-2022 Revised: 18-04-2022 Accepted: 21-04-2022

ABSTRACT-
There are many things in the automotive industry to design and build vehicle safety systems, but road accidents are inevitable. There are many serious accidents in all urban and rural areas. The patterns involved with different scenarios can be obtained by creating accurate speculative models that will be able to automatically distinguish between different risk scenarios. These collections will be helpful in preventing accidents and improving safety measures. We believe in achieving greater risk reduction opportunities using low budget resources through specific scientific measures. Models built using accident data records can help to understand the features of many things such as driver behavior, road conditions, lighting, weather conditions and more. This can help users to calculate effective safety measures to avoid accidents.

I. INTRODUCTION

According to statistics released by the World Health Organization, the number of annual car accidents worldwide is alarming. Road accidents kill 1.2 million people each year and injure 50 million people. Road accidents are strongly influenced by road geometric features, traffic flow, driver characteristics and road environment. We have therefore designed a model to predict crash waves and to analyze road hazard aspects, including studies on risk tropical identification, risk magnification, and long-term risk analysis. Some studies have focused on the path of danger. Other factors include road weather. This study identified specific data and key target factors that contribute to the severity of road accidents.

The strength of the RTA is one of the areas of research over the past two decades on road safety. Researchers were using intriguing models based on road hazard classification. The authors studied using the traditional mathematical method of modeling. These strategies help to obtain information and identify the main cause of motor vehicle accidents and aspects related to road safety. Nowadays, due to the presence of a large volume of data sets, machine reading exceeds standard calculations based on model prediction.

II. LITERATURE REVIEW

In the vicinity of road protection conventional statistical model based strategies had been used to are expecting coincidence fatal and severity. blended logit modelling method [23, 26], ordered Probity version [54], logit version [11] are few of adopted conventional statistical-based research. some studies believed the traditional statistical model higher pick out structured and independent twist of fate factors [31]. however conventional statistical-based totally method lacks the capability to address multidimensional datasets [16]. to be able to fight conventional statistical fashions limitations; these days many studies used ML approach because of its predictive supremacy, time eating and informative size. In those decade ML approach employed in construction industry [48], occupational twist of fate [41], agriculture [22], academic classification [53], sentiment classification [50] and in banking and insurance [46]. on the other hand, in road twist of fate prediction, many studies finished the usage of data mining, machine studying, and deep gaining knowledge of algorithms. among clustering and classification algorithms: ok-way, guide Vector Machines, ok-Nearest buddies (KNN) decision Tree (DT), Artificial Neural network (ANN), Convolution Neural community (CNN) and Logistic Regression (LR) are in front to construct coincidence severity version. Kwon et al. [28] adopted Nave Bayes (NB) and selection Tree (DT) on California dataset accumulated from 2004 to 2010. Authors used binary regression to evaluate the performance of the developed version however Nave Bayes had been greater touchy to chance factors than the choice Tree version. Sharma et al. [44] analysed road accident statistics using SVM and MLP on a restrained number of datasets (three hundred datasets). except authors used simplest two independent variables (alcohol and velocity) as considering key factors. sooner or later, SVM with RBF kernel gave better accuracy (94%) than MLP (sixty four%). The look at showed driving with

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 937-942 www.ijaem.net ISSN: 2395-5252

excessive pace after drunk changed into the principle cause for coincidence prevalence. Wahab and Jiang [51]performed crash injuries on Ghana dataset using MLP, part, and easy CART intending to evaluate classifiers and to become aware of the main factors for motorbike crash. Authors used Weka tools to evaluate and analyze datasets and information benefit attribute Eval implemented to see the maximum infuential variable for bike crash in Ghana. As a end result easy CART model confirmed better accuracy than other classification models.

III. METHODOLOGY:

The study focuses on classificationon the severity of road accidents. It combines the gradient booster algorithm collection with the division method to get a better result than eachcategory. gradient enhancement is one of the incremental algorithms used to reduce the bias error of the model.and a Decision Nodes are used to make any decision and have many branches, while Leaf nodes are the result of those decisions and have no other branches.Proposed course of work in the study shown in Fig. 1. Large sections of flow chart They are as follows:

3.1 Road accident dataset manipulation

3.1.1Rawtraffic accident data-set

The data set for this study covers 5000 road accidentscollected from the federal traffic police agency reports from2009 to 2015 in India. One of the most challenging parts of this study is collecting sample data sets from an organization. Original database collectedin a personal authority.numerical variables were recorded. In the midst of this variation,Severe Injuries and Minor Injuries). Description of the Full Data setdescribed as follows:

Accident time: This variable means the time at whichtraffic jams occur throughout the day (24hours).

Driver age This difference reflects the age of the driver.Drivers age especially in the 18-80s.

Sex This variable reflects the driver's gender. Driversex (recognized in the data collection) is male ora woman.most of the time the driver is driving the car.

Type of car This flexibility means different typesof cars. Namely: ambulance, car, car, Isuzu,taxi, truck, motorcycle, pick up, bus and minibus.

LocationThis variable indicates where the risk happened: namely: bar area, pubic area, organization, government office, hospital, college, car station, market accommodation, and hospital

Light status This variable indicates the state ofroad at the time of the accident. The variables represent:dry, muddy and wet.

WeatherThis variation reflects the weathersituation at the time of the accident. The variables represent:rainy, sunny, cold and windy.

Causality Category This variable indicates the severity of theclass. The variable represents: driver, passenger,pedestrians, cyclist, and resident.

Gender Causality This variable indicates a separate categorymale or female sex.

This target variableshowsthree classes, namely: fatal, serious injury and minor injury

Type of car This flexibility means different typesof cars. Namely: ambulance, car,truck, motorcycle, pick up, bus and minibus

3.1.2 Pre-processing:

A preliminary processing of data in order to prepare it for the primary processing or for further analysis. The term can be applied to any first or preparatory processing stage when there are several steps required to prepare data for the user. Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms. Raw datasets were dirty, not in a proper format to be understood by computing machines and give incomplete information to use as it is. Using Such datasets will reduce the efficiency of the accident severity prediction model. Therefore, irrelevant datasets need to remove to obtain quality data. In the study before

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 937-942 www.ijaem.net ISSN: 2395-5252

building a model intensive data pre processing technique employed to get meaningful and determinant risk factors Like Data cleaning, missing value handling, outlier treatment, dealing with absolute value—encoding and normalization are carefully purify before using it.

3.1.3 Splitting dataset:

Data classification is often used in machine learning to split data into a train, test, or set of verification. This method allows us to determine the hyper-model parameter and measure the normal performance Green data sets and created k features are divided into train sets and test sets. The set of training mentioned above helps to learn the newly proposed method. On the other hand, a test set is used to measure the performance of the proposed new model. In the study, a ratio of 70:30 is used to separate the raw data. Then 70% is used to train the forecast model, while, 30% of the data is used to test the accuracy of the forecast phase.

3.1.4 Predicting model:

The prediction model is widely used in machine learning strategies to predict future behaviour by analysing Information and historical data.

3.2 K-Nearest Neighbour

The number of nearby neighbours for new unknown variables to be predicted or subdivided is indicated by the 'K' symbol.Let's take a closer look at the real world situation related before we start with this amazing algorithm. We are often informed that you share many traits with your closest peers, be it your way of thinking, good work habits, philosophies, or other aspects. As a result, we make friends with people we consider to be our peers.The KNN algorithm uses the same principle. Its purpose is to locate all nearby neighbours near a new anonymous data centre to determine its category. It is a distance-based approach. Consider the diagram below; it is straightforward and easy for people to identify as "cats" based on their closest partners. This function, however, cannot be performed directly with the algorithm.KNN calculates the distance from all points near unknown data and filters those with the shortest distances to reach. As a result, it is often referred to as a distance-based algorithm.In order to properly classify the results, we must first determine the value of K (Neighbourhood Number).Therefore, you can use the KNN algorithm for applications that require high accuracy but do not require a human-readable model. The level of forecasts depends on the distance rating. Therefore, the KNN algorithm is ideal for applications where sufficient domain information is available.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

3.3 Decision Tree

Decision Tree is a supervised learning method that can be used for both planning and retrospective problems, but is often preferred in solving planning problems. It is a tree-shaped divider, where the internal nodes represent the elements of the database, the branches represent the rules of decision and each leaf node represents the result.In the Decision Tree, there are two nodes, namely Resolution Node and Leaf Node. Decision Nodes are used to make any decision and have many branches, while Leaf nodes are the result of those decisions and have no other branches.
Decisions or tests are made on the basis of the data provided.

It is a metaphorical representation of finding all possible solutions to a problem decision-based situation.It is called a trunks tree because, like a tree, it starts with a root node, stretches out to more branches and builds a tree-like structure.To build a tree, we use the CART algorithm, which represents the Classification and Regression Tree algorithm.The decision tree simply asks a question, and is based on the answer (Yes / No), and also divides the tree into sub-trees.The formula for Entropy is shown below:

$$E(S) = -p_{(+)}\log p_{(+)} - p_{(-)}\log p_{(-)}$$

Here p_+ is the probability of positive class
p_- is the probability of negative class
S is the subset of the training example

3.4 Gradient Boosting

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that errors in machine learning algorithms are broadly divided into two categories namely Bias Error and Difference Error. As gradient enhancement is one of the incremental algorithms used to reduce the bias error of the model.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 937-942 www.ijaem.net ISSN: 2395-5252

In contrast, the Adaboosting algorithm, the basic measurement in the gradient growth algorithm cannot be stated by us. The basic scale of the Gradient Boost algorithm has been adjusted also namely Decision Stack. For example, AdaBoost, we can tune the estimators algorithm for increasing the gradient. However, not to mention the value of estimators, the default estimators value for this algorithm is 100.

The gradient development algorithm can be used to predict not only continuous target variables (such as Regressor) but also target variability (such as Classifier). When used as a regressor, the cost function is Mean Square Error (MSE) and when used as a separator the cost function is a Log loss.

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$

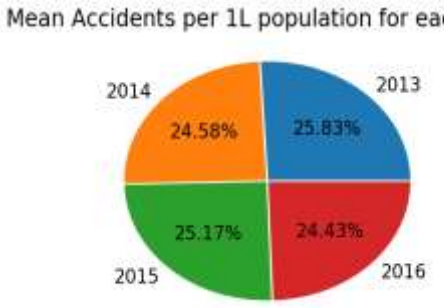
where, α is learning rate and $\sum (y_i - y_i^p)$ is sum of residuals

3.4 Proposed Approach

Nowadays road accident data sets are stored in a large areadatabase. A large number of data sets make up training and testing phase is complex and limitedpredicting efficiency. Therefore, it needs a strong modelto overcome or reduce the complexity of a large valueof the database. We made a hybrid KNN and Decision Tree model to find the most effective forecast model to improve the efficiency and accuracy of the forecast model.Gradient boosting in general, which is an unattended machinea learning algorithm used mainly to find similar groupswithin the database. Although this is unattendedstrategies, k-means can create new features of the train set to improve classifier performance. Consolidationcreates a cluster feature and adds to the training set. Thenrandom forest employed in training data compiled todivide the weight of the RTA. When the combination will produce a powerful model of predictability based on standard performance and precision prediction.

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that errors in machine learning algorithms are broadly divided into two categories namely Bias Error and Difference Error. As gradient enhancement is one of the incremental algorithms used to reduce the bias error of the model.

In contrast, the Adaboosting algorithm, the basic measurement in the gradient growth algorithm cannot be stated by us. The basic scale of the Gradient Boost algorithm has been adjusted also namely Decision Stack. For example, AdaBoost, we can tune the n estimator algorithm for increasing the gradient. However, not to mention the value of n_estimator, the default n_estimator value for this algorithm is 100.



IV. EXPERIMENT, EVALUATION,AND DISCUSSION

In this section, the pre-processing technique applied to the road accident dataset, evaluation metrics, and experimental result analysis presented.

4.1 Dataset manipulation

The Dataset collected is incompleteclear and orderly. The green database is also recorded manuallyprone to injury. However, it should be in an understandable machine format for accurate informationand developing a smart operating system. Roadrisk modelling model depends on qualityof the database. We used different types of pre-processed data testing techniques to clean up the database.

• Invalid value management Invalid value treatment amandatory function in pre-processing data. Before construction the missing model model needs to be delivered using a different strategy. In the database, some attribute values are the sameabsent. Building a fun and efficient pre diction model with incomplete data will not provide a lasting result. He should handle it wisely and carelesslyor it should be completed using a variety of methods to improve iteffect [43]. Ignoring or reducing prices is the way to gomanaging missing prices, but a collapse can lead tom lack of essential information. In the lesson, novalue is not obliged to discard non-existent features. Figure 3indicates the number of non-existent values and their percentagefrom total database. Not value is less

Fifty percent of the population. we have used the substituting feature which means numerical variables and more the standard value (mode) of phase variability [3]. Becausefurther, in Pre-processing our previous work provides detailsinformation, see Ref. [43].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 937-942 www.ijaem.net ISSN: 2395-5252

• Category Coding Value Risk of green trafficdata sets contain segmental and numerical values.However, many machine learning algorithms are requirednumerical values to predict the model. Employment amachine learning algorithm with phase valuesa challenging problem. Therefore, category valuesshould be converted to numeric values orneeds to be removed [32]. In the database, most of thevariables are categorized; Of the 14 variables, 10 arethey are different values and are needed for conversionbe a number format. predictable variables and targetvariables are converted to numbers using the same hot code and label code respectively

4.2 Experimental system set up

The study implemented using python 3.7 on Jupyter notebook as IDM and intel core i7 1.80GHz processor speed CPU, 8Gb RAM, and 1TByte HD system. In this section, different experiments like Choosing an optimal value of k, evaluating the proposed approach, and finally comparing with conventional algorithms with the new approach presented.

4.3 Evaluation metrics

In the study, different types of evaluation, metrics are used to measure the performances of the proposed approach to predict road accident training set as indicated from Eqs.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{FP + TN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$F1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

TP: it shows predictive is positive and it is normally true
TN: it implies predictive is Negative and it is normally True
FP: denotes predictive is positive and it is normally false
FN: represents predictive is negative and it is false. Where TP implies true positive, TN denotes true negative, FP indicates false positive, and FN denotes false negative. in the actual study values are represented by true and false whereas predictive values denoted by positive and negative.

V. EXPERIMENTAL RESULT ANALYSIS AND DISCUSSION

5.1 Train-test split

Using python, jupyter notebook and Scikit learn, pandas and matplot data science libraries we have developed a workflow for processing the dataset and generate the corresponding accident severity prediction models. It is composed of a number of nodes, namely:

1) Dataset: contains the pre-processed data for the experiment

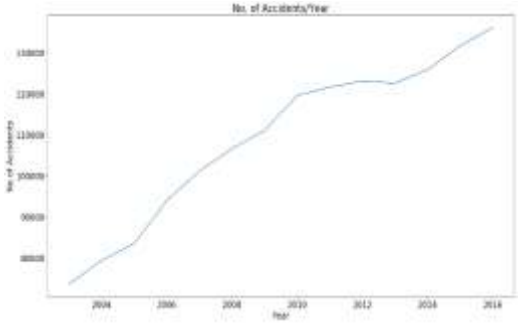
2) Explore Data: is an optional node to help in data exploration and viewing some statistics about the data before modelling.

3) Model: contains the algorithms that will be used for model generation.

4) Apply: where the model is applied to the predictors to generate the required results

5) Predictors: sample dataset for testing the prediction.

6) Prediction: the resulted table after applying the model on the predictorsWhen the database is ready to train the model. It splits in halftraining set and test set. The first one he used to readclassifier, and was later used to test performanceof the forecast model. In research, Used in70:30 average, 70% of the portion used for model training as well30% of the test set used to test the qualified model.



5.2 Model performance evaluation

When the database is ready to train the model. It splits in halftraining set and test set. The first one he used to readclassifier, and was later used to test performanceof the forecast model. In research, Used in70:30 average, 70% portion used for model training as well30% of the closest K test method identifies nearby neighbours from withindatabase and collect them together to form collections. IThe database used contains route name, traffic value andtime interval. Collection of items is done on the basis ofroute name at a different time. Reducing complexitythe process organizes all the data in the correct orderto create custom-compliant collections. Set to test the appropriate modelGradient boosting algorithm is

one of the most powerful algorithms in the field of machine learning. As we know that errors in machine learning algorithms are broadly divided into two categories namely Bias Error and Difference Error. As gradient enhancement is one of the incremental algorithms used to reduce the bias error of the model.

VI. CONCLUSION

This project aims at using Machine Learning classification techniques to predict severity of an accident at any particular location. Machine Learning has enabled us to analyse meaningful data to provide solutions with a greater accuracy than with humans. We have built a model with a accuracy greater than 17% of the conventional system [1].This project can be used by governments to prevent accidents.

VII. FUTURE WORK

With more resources, continuous prediction and alerts can be sent to the police for every location at regular intervals of time to take preventive measures. The model can be incorporated with Google Maps which can be live tracked by the police. A fully-fledged web app for user and police interaction can be published for use in real-time. It can be used for Indian states or cities, if proper data of accidents is provided by the Indian Government.

REFERENCES

[1]. Lu Wendi, Luo Dongyu& Yan Menghua, "A Model of Traffic Accident Prediction"INSPEC Accession Number: 17239218 DOI: 10.1109/ICITE.2017.8056908

[2]. Thin EswaranGunasegaram Yu-N Cheah, "Evolutionary Cross validation"INSPEC Accession Number: 17285520 DOI: 10.1109/ICITECH.2017.8079960

[3]. Simon Bernard, Laurent Heutte and Sebastien Adam, "On the Selection of Decision Trees in Random Forests" INSPEC Accession Number: 10802866 DOI: 10.1109/IJCNN.2009.5178693

[4]. Rafael G.Mantovan,, Ricardo Cerri, Joaquin Vanschoren, "Hyper-parameter Tuning of a Decision Tree Induction Algorithm" INSPEC Accession Number: 16651860 DOI:10.1109/bracis.2016.018

[5]. Fu Huilin, Zhou Yucai, "The Traffic Accident Prediction Based on Neural Network", 2011

[6]. Lin, L., Wang, Q., Sadek, A.W., 2014. Data mining and complex networks algorithms for traffic accident analysis. In: Transportation Research Board 93rd Annual Meeting (No. 14-4172).