International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Accident Severity Prediction

Varun Kumar B
Department of Computer Science and Engineering
JAIN(Deemed-to-be University)
Bengaluru, India
18btrcs085@jainuniversity.ac.in

Reddy Vamsi
Department of Computer Science and Engineering
JAIN(Deemed-to-be University)
Bengaluru, India
18btrcs056@jainuniversity.ac.in

Rajeev Reddy V
Department of Computer Science and Engineering
JAIN(Deemed-to-be University)
Bengaluru, India
18btrcs087@jainuniversity.ac.in

Sai Rohit
Department of Computer Science and Engineering
JAIN(Deemed-to-be University)
Bengaluru, India
18btrcs097@jainuniversity.ac.in

Dr. Somashekar
Department of Computer Science and Engineering
JAIN(Deemed-to-be University)
Bengaluru, India
r.somashekar@jainuniversity.ac.in

***Abstract*---- Traffic safety has always been an important issue in sustainable transportation development, and the prediction of traffic accident severity remains a crucial challenging issue in the domain of traffic safety.Based on the severity we can create awareness of happening the accidents according to the zone or place.So the goal is to predict the accident severity. Here we used classification models to find the f1 score of each model and select the best model with high f1-score to predict the approvals. we take the input parameters, and we training data and output will be the Accident Severity.**

*Keywords---**severity prediction, deep learning,***

I. INTRODUCTION

Accident Severity Prediction is nothing but predicting the occurrence of accidents and it's severity .There is a huge impact on the society due to traffic accidents where there is a great costs of fatalities and injuries .In recent years there is a increase in the researches attention to determine the significantly affect of severity of the drivers injuries which is caused due to the road accident. Accurate and comprehensive accident records depends on some factors like the accuracy of data , data analysis etc. A recent study illustrates that the residential and shopping sites are more hazardous than village areas. A study revealed that the casualty rates among the residential areas are classified as relatively deprived and significantly higher than those from relatively affluent areas .Transport system plays an vital role in the economic resources of the country. Majority of places in the country are connected by roads. Several people use roadways to travel from one place to another. Roads carry more volume of traffic than it was actually designed to carry. This as in turn increased congestion and traffic ,vehicle crashes etc. Accidents pose a serious issue. According to a report by MORTH[1],0.4 million accidents are reported every year. Accidents are unpredictable and they occur in various situation .Hence understanding the factors that leads to an accident can prove to be useful in preventing it

.A. Severity Prediction:

Accident severity is often measured categorically, for instance , the severity level of an accident can be classified as fatal , serious injury or no injury(property damage only).As such , statistical models that are suitable for categorical data , such as logistic models, have been used to analyse accident severity. There were 24,831 serious injuries in road traffic in road traffic accidents reported to the police 2017. However, comparison of
this figure with earlier years should be interpreted with caution due to changes in systems for severity reporting by some police forces. The report contains further information and proposed to account for this discontinuity.

II. LITERATIVE REVIEW

In this section we provide the background knowledge based on several techniques. Many of these are based on the Machine algorithms.

Sarbajit Bhattacharyya, Mrinal Roy, Pinak Paul, Rupanjan Chakraborty, "Accident Analysis and the Suggestion of an Accident Prediction Model for Guwahati [1] The ultimate purpose of scientifically designing a road is to avoid the occurrence of accidents on it. Even after the construction of a road, several measures should be taken to make the road sufficiently safe from accidents. The possibility of accidents occurring on a road depends upon numerous factors. In this paper, our study involved the analysis of those factors affecting number of accidents for Guwahati city by collecting accidental data for various road stretches connected to Jalukbari area of Guwahati which can be considered as a representative of almost the whole city. To avoid the demerits associated with a linear regression model, a log-linear model was developed which could fit the data obtained and was validated by using the model to predict the number of accidents per stretch length for a similar test with considerable accuracy.

Snehal U Bobade, Jalindar R Patil, Raviraj R Sorate [2], Identification of Accident Black spots on National Highway and

Expressways .In 2015, Snehal U Bobade, et al. hintedthat accident-prone locations can be identified by ranking the parameters based on their severity and calculating the severity index. The physical survey was carried out at the actual location for selected stretches of Mumbai-Pune Expressway and Pune-Solapur Highway. The parameters which caused a maximum number of accidents were assigned maximum weightage and top rank. The summation of the weightages was calculated to find out the total severity. The severity Index was then calculated by adding the weightages of each parameter present divided by the total severity .

Chand, M. and Alex, A.P. A comparative analysis of Accident Risk of states in India [3], In 2007, Chand and Alex computed accident risk index and accident severity index (ASI) for different states in India. These indices are based on a set of accident indicators, whichare combined together to form an index. Values of these two indices havebeen computed and compared across the states of India.

Aruna.D.Thube&Dattatraya.T.Thube (2010) [4] ,The objective of this paper was to study the rural highways and finding out various causes of accidents and also to suggest the remedial measures.The identification of the accident-prone areas was done by the PWD as per the data obtained from local police station. These locations were classified based on the severity of accidents. Further the type of remedial measure to be adopted at these location was mentioned.

National Crime Records Bureau (NCRB), Road Accidents in India Report, G. O. I Ministry of Road Transport & Highways Transport Research Wing, New Delhi, 2015[5]. Road accidents have become an alarming issue across the globe. The number of serious as well as minor injuries, human sufferings and the economic loss due caused by accidents is inestimable. Hence road safety is a major concern in the present situation. According to the latest road accident data released by the Ministry of Road Transport and Highway, the total number of accidents increased by 2.5 percent from 4,89,400 in 2014 to 5,01,423 in 2015. The analysis reveals that about 1,374 accidents and 400 deaths take place every day [1]. This implies that every hour 17 people become the victims of a road accident.

Douglas W. Kononen, Carol AC Flannagan, Stewart C. Wang [6] .A multivariate logistic regression model, based upon National Automotive Sampling System Crashworthiness Data System (NASS-CDS) data for calendar years 1999–2008, was developed to predict the probability that a crash-involved vehicle will contain one or more occupants with serious or incapacitating injuries. These vehicles were defined as containing at least one occupant coded with an Injury Severity Score (ISS) of greater than or equal to 15, in planar, non-rollover crash events involving Model Year 2000 and newer cars, light trucks, and vans. The target injury outcome measure was developed by the Centers for Disease Control and Prevention (CDC)-led National Expert Panel on Field Triage in their recent revision of the Field Triage Decision Scheme . The parameters to be used for crash injury prediction were subsequently specified by the National Expert Panel. Model input parameters included: crash direction (front, left, right, and rear), change in velocity (delta-V), multiple vs. single impacts, belt use, presence of at least one older occupant (≥55 years old), presence of at least one female in the vehicle, and vehicle type (car, pickup truck, van, and sport utility). The model was developed using predictor variables that may be readily available, post-crash, from OnStar®-like telematics systems. Model sensitivity and specificity were 40% and 98%, respectively, using a probability cutpoint of 0.20. The area under the receiver operator characteristic (ROC) curve for the final model was 0.84. Delta-V (mph), seat belt use and crash direction were the most important predictors of serious injury. Due to the complexity of factors associated with rollover-related injuries, a separate screening algorithm is needed to model injuries associated with this crash mode.

III.METHODOLOGY

In this step of data pre-processing we will pre-process the data. We will read the data from dataset and replace the null values. We will know the information about all the data types. We will know the mean standard deviation and various metrics regarding to the numerical data. To know the correlation between the numerical attributes we will plot the graphs to visualize the data . Now we need to perform a One Hot Encoding of the categorical variables to prepare the data for classification. We can do this easily by using OneHotEncoder from the sklearn.preprocessing module or simple call get_dummies on a pandas data frame. For simplicity, we will use the later approach. After that the whole dataset is divided into training and testing data using train_test_split from sklearn.model_selection. Standardization of a dataset is common dataset for many machine learning estimators. They might behave badly if the individual features do not more or less look like standard normally distributed data. So we use StandardScalar().

5.1 KNEARESTNEIGHBOURS:

One real world application:

1. KNN can be used to provide recommendations. A real world example of this would be video streaming services such as Netflix or Amazon Prime. If a given user likes an item in the library, similar items that they may like, but are unaware of can be recommended to them by using data from other users and their likes. If it is seen that a similar set of users like two different items, these items are probably similar and to each the respective users taste, and worthy of a recommendation.

Strengths of the model:

1. Easy to understand and implement - not much code is required.

2. KNN is a lazy learner. This means it generalises data during the training phase, not the testing phase. This allows it quickly adapt to changes as it does not expect a generalised data set.

3. KNN is a lazy learner. This means it generalises data during the training phase, not the testing phase. This allows it quickly adapt to changes as it does not expect a generalised data set.

Weaknesses of the model:

KNN gets its information from its input neighbors. As a result of this, localised outliers can affect outcomes significantly when compared with other algorithms which have a generalised view of the data. It is sensitive to localised data.
One of its strengths, lazy-learning, is also one of its weaknesses. As most of the computation is done during testing, rather than during training, this can result in long computation times when dealing with large datasets.
If there is a type of categorey that is present much more than another, classifying an input will result in a bias to this more abundentcategorey. This can be dealt with by adjusting the weights based on occurences, but will still pose a problem near the decision boundary.

Advantages:

The cost of the learning process is zero No assumptions about the characteristics of the concepts to learn have to be done Complex concepts can be learned by local approximation using simple procedures.

Disadvantages:

The model cannot be interpreted (there is no description of the learned concepts) It is computationally expensive to find the k-nearest neighbours when the dataset is very large Performance depends on the number of dimensions that we have (curse of dimensionality)⇒ Attribute Selection

RANDOM FOREST

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).
The Real-World Application Where the model can be applied is Predecting Stock Market Prices.

It works in binary features, it is an Ensemble of Decissiontrees,It works well with large number of training examples.

Advantages:

1. Reduction in over fitting: by averaging several trees, there is a significantly lower risk of over fitting.

2. Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.

Disadvantages :

1. It takes more time to train samples.

5.3 NAIVE BAYES:

Naive Bays is often used for spam filtering.The model performs very well when there are lot of features, and it's simple and easy to understand. However the disadvantage of the model is that it makes a strong assumption about the independence of the features.

The dataset includes several features after one-hot encoding. Because Naive Bayes provides good performance when there are lot of features, we should apply this model.

Advantages:

1. Very simple, easy to implement and fast.
2. If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
3. Even if the NB assumption doesn't hold, it works great in practice.
4. Need less training data.
5. Highly scalable. It scales linearly with the number of predictors and data points.
6. Can be used for both binary and multiclass classification problems.
7. Can make probabilistic predictions.
8. Handles continuous and discrete data.
9. Not sensitive to irrelevant features.

Disadvantages:

1 .The first disadvantage is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be potentially very bad - hence, a "naive" classifier. This is not as terrible as people generally think, because the NB classifier can be optimal even if the assumption is violated, and its results can be good even in the case of sub-optimality.

2.Another problem happens due to data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequents approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results. In this case, you need to smooth in some way your probabilities, or to impose some prior on your data, however you may argue that the resulting classifier is not naive anymore.

3.A third problem arises for continuous features. It is common to use a binning procedure to make them discrete, but if you are not careful you can throw away a lot of information.

5.4 LOGISTIC REGRESSION:

Logistic Regression is very widely used in the case of binary classification problems.
Strengths fast in training and prediction time, gives good results in case of less features.
Weaknesses assumes linear decision boundary, cannot decode complex relationships between the features.
Candidacy problem is of binary classification with clean data, all favourable conditions for logistic regression.

Advantages:

1. Because of its efficient and straightforward nature, doesn't require high computation
2. power, easy to implement, easily interpretable, used widely by data analyst and scientist.
3. Also, it doesn't require scaling of features. Logistic regression provides a probability score for observations.

Disadvantages:

1. Logistic regression is not able to handle a large number of categorical features/variables.
2. It is vulnerable to over fitting. Also, can't solve the nonlinear problem with the logistic regression that is why it requires a transformation of non-linear features.
3. Logistic regression will not perform well with independent variables that are not correlated to the target variable and are very similar or correlated to each other.

5.3 DECISION TREES :

Real world application: Decision Trees and, in general, CART (Classification and Regression Trees) are often used in financial analysis. A concrete example of it is: for predicting which stocks to buy based on past performance.

Strengths:

1. Able to handle categorical and numerical data. Doesn't require much data pre-processing, and can handle data which hasn't been normalized, or encoded for Machine Learning Suitability.Simple to understand and interpret.

Weaknesses:

1. Complex Decision Trees do not generalize well to the data and can result in over fitting.
2. Unstable, as small variations in the data can result in a different decision tree. Hence they are usually used in an ensemble (like Random Forests) to build robustness.

3. Simple to understand and to interpret.
4. Trees can be visualized
5. Requires little data preparation.

1.Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
2.Note however that this module does not support missing values.

The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.

3.Able to handle both numerical and categorical data.

4.Other techniques are usually specialized in analyzing datasets that have only one type of variable.

See algorithms for more information.
5.Able to handle multi-output problems.
6.Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret. Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model. Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

5.6 STOCHASTIC GRADIENT DESCENT (SGD):

One real world application:

Text classification and natural language processing. It is useful as when the given data is sparse, the function can easily scale to problems with more than 10^5 training examples and more than 10^5 features.

Strengths of the model:

1. It is efficient.
2. It is easy to implement and provides a lot of opportunities for code One tuning.

Weaknesses of the model:

1. A number of hyper parameters are required for SGD, such as the number of iterations and the regularisation parameter.
2. It (SGD) is sensitive to feature scaling.

ntages of Stochastic Gradient Descent aEfficiency.
Efficiency
Ease of implementation (lots of opportunities for code tuning).
The disadvantages of Stochastic Gradient Descent include:
SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations.
SGD is sensitive to feature scaling.

The advantages of Stochastic Gradient Descent are:

5.7 ADABOOST:

Face Detection.Very simple to implement, Fairley good generalization.sub optimal solution, sensitive noisy data and outliers.

Adaboost uses a group of weak learners to form a strong learner in prediction. It could work well with less features and amount of data. Cosidering our dataset include few features with limited rows of data. This method would be a good choice to start with.

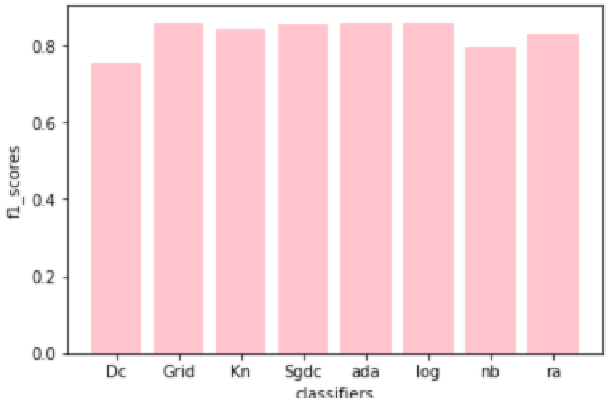
Advantages:
1.Which weak classifier might work best to solve their given classification problem.
2.The number of boosting rounds that should be used during the training phase.

3.The GRT enables a user to add several weak classifiers to the family of weak classifiers that should be used at each round of boosting.
4.AdaBoost algorithm will select the weak classifier that works best at that round of boosting.

Disadvantage:

AdaBoost can be sensitive to noisy data and outliers. In some problems, however, it can be less susceptible to the overfitting problem than most learning algorithms. The GRT AdaBoost algorithm does not currently support null rejection, although this will be added at some point in the near future.

IV. CONCLUSION



Above given bar chart is the comparison of f1- scores of all the chosen classifiers in the project, the first bar being Decision Tree, the second bar being GridsearchCV,the third bar being Kneighbours Classifier ,fourth bar being SGDC, and the sixth bar being the tuned Adaboost,and the next one is logistic Regression,and the bench mAbove given bar chart is the comparison of f1 scores of all the chosen classifiers in the project, the first bar being knn, the second bar being svm,the third bar being decision trees ,fourth bar being Random forest, and the sixth bar being the tuned logistic regression. One can see the performance enhancement in the final bar using random state of 100 as the input training and testing sets.

REFERENCES

[1]. Sarbajit Bhattacharyya, Mrinal Roy, Pinak Paul, Rupanjan Chakraborty, "Accident Analysis and the Suggestion of an Accident Prediction Model for Guwahati city",International Journal of Innovative research in Science Engineering and Technology,2015.

[2]. Snehal U Bobade, Jalindar R Patil, Raviraj R Sorate. Identification of Accident Black spots on National Highway and Expressways, IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) e-ISSN: 2278-1684, p-ISSN: 2320-334X, Volume 12, Issue 3 Ver. I (May. - Jun. 2015), PP 61-67.

[3]. Chand, M. and Alex, A.P. A comparative analysis of Accident Risk of states in India, Highway Research Bulletin, pp.105-116, 2007.

[4]. Accident Black Spots in Rural Highways By Aruna.D.Thubhe and Dattatraya.T.Thubhe.

[5]. National Crime Records Bureau (NCRB), Road Accidents in India Report, G. O. I Ministry of Road Transport & Highways Transport Research Wing, New Delhi, 2015.

[6]. Douglas W. Kononen, Carol AC Flannagan, Stewart C. Wang, "Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes", Acident Analysis and Prevention, vol. 43.1, pp. 112-122, 2011.