DATA MINING

ROAD ACCIDENT

Abstract

In many scenarios around the world, it is always difficult to identify nuggets of information in large bodies of data; data mining helps to extract this information to be used in areas such as decision support, predictive analytics, it helps to extract the information with the value from the voluminous amounts of data present. Data mining generates models from historical data that is then used from making critical decisions in various walks of life.

In this paper, the creator applies different data mining techniques like Decision trees, Naïve Bayes and Ensemble techniques to determine the appropriate classification methods of data mining to try to reduce the severity of street mischances in Leeds. This information can be used by architects and transport fashioners to plan for better and secure streets.

Table of Contents

Introduction

The traffic and road accidents have become one of the most serious and frequent incidents in the contemporary midway that also has adopted an image of never-ending situation. Many survey reports have identified that only in the UK 0.4 billion people die every year in massive road accidents. With such a massive rate, the UK has become one of the prominent accident prone countries in the worlds as a particular trend in road accidents have been recorded between 2012 and 2015 (Anderson, 2009). But the road accidents are not guaranteed to be cover or prevented by such trends and it is quite possible that the types of road accidents could occur in the future as well. Here the issues could be realized in terms of the lack of proper systematic invasion in case of resolving the problems and causes related to road accidents in the UK. ---- has stated that in the past the statistical tools and the models were used in order to identify the correlation between the road accidents and the geographical and other reasons behind it. Thus, it has been realized that the large dataset through using traditional statistical models are though prone to evaluate the reason behind road accidents, but it has certain problems such as sparse data presented through large contingency tables (Brennan, 2012). On the other hand, the traditional statistical mo0dfels have also provided instances of falsifying and erroneous outcomes in relation to the explanation of road accidents in the UK.

Due to certain above mentioned identified problems, the present road accidents are being analysed through modern statistical techniques and models which have contributed in delivering proper analysis regarding the massive road accidents taken place in the last few years. In this regard, the concept of data mining could be explained through the terms of a set of techniques that are used to extract the hidden and explicit information from a large quantity of data (Ebrahimi, 2015). In relation to the factors, several types of data mining could be identified association rule mining, clustering along with classification. Thus, the study would be

contributed to evaluating the use of data mining and statistical models in the context of extracting actual information from a large pool of information related to road accidents. Hence, the actual problem could be identified in terms of issues regarding the implementation of statistical models and techniques, through which the contemporary road accidents in the UK could be analysed.

Data Mining Methods

The data mining is one of the swift tools in the contemporary societal scenario that is used to analyse the roads accidents. Thus, in the context of the three types of data mining processes have been realised such as clustering, association rule, and dataset. Therefore, the types of data mining could be discussed in the following manner.

K -means clustering

The clustering data mining techniques is a form of unsupervised data mining tool that refers to the act of accumulating data objects within small clusters instead of groups which is more similar to the group data objects. K- Means the algorithm that appears in the cluster and the data are dependent upon the type and the nature of the data. The prime purpose of the cluster techniques is to accumulate data into small clusters in order to extract the actual information regarding road accidents. Along with its advantages, the issues in relation to this method of data mining is that it is hard to recognize the categories of the accidents and location and decide a level of threshold category for this (Karlis, 2003). Thus, there are a certain number of cluster selection that is meant for data identifying the type of accident and location at the same time. The main issue in relation to cluster data mining is to identify the actual type of accident and location, thus, in this regard, the gap statistics could be considered as effective.

Considering a data set in the context of a road accident,

Dij, i = 1,2,…,*m*, j = 1,2,….,*n*

Here, the *m* data consists with the object value of *n*, and it could be assumed that the d_{xy} is the Euclidean distance between two objects *x* and *y* as per the $d_{xy} = P (X_j - Y_j)^{2.}$ In the case of data set had been clustered in *k*; therefore, the $c_1, c_2,….,c_k$ would have been delivering the value of *k* through $n_i = |c_i|$.

Association Rule Mining

In the case of association rule mining, it is realised as the very popular data mining techniques in which the tools contributes its action towards accumulating interesting data from the market basket and generate an analysis of the variable to point out the differences between the variables. On the other hand, the data mining association can provide a set of rules that would be determining the incorporation of data objects through data set. In the given data set of D, the *n* transaction refers to the transaction set of $T \in D$. Let $I = \{I_1, I_{2.....}, I_n\}$ which appears as a set of items. Apart from this, an A could appear and associate with T if the data set is $A \subseteq T$. $A \rightarrow B$, which is also provided by the $A \subset I, B \subset I$ along with $A \cap B = \emptyset$ (Parmentier G, 2005). Therefore, there are several aspects in relation to data mining association that would help in understanding the set of rules and interesting measures in relation to association data mining in the following manner.

Support:

The rule for support of $A \rightarrow B$ refers to the occurrence of A and B in a data set, and it could be calculated through using Eq.

Confidence

Confidence rule in relation to $A \rightarrow B$ refers to the ratio of occurrence of A and B together while it also focuses upon the occurrence of A only which could be calculated through Eq.

Along with this, other measures could be observed in terms of Conviction. Leverage, Lift which is also contributed to the assessment of the interesting data and data objects that could identify the differentiating reason behind increasing road accidents.

Dataset

The dataset is meant for delivering information and data regarding personal injury during an accident and other data at the time of accident are also recorded. Input to derive examination of road accident in Leeds was acquired from (UK Government, n.d.). The mishap data was part finished every year. The isolated data was analyzed to promise it meets the market essential with the use of information portrayal chronicle. The association amongst qualities gets verified and essential aggregate achieved. The dataset utilized as a part of this investigation was deduced from an aggregate of 18,572 announced car accidents in Leeds, United Kingdom. The crash subtle elements are given by every year. Preparatory investigation was done to guarantee every one of the information is caught amid the arrangement. The dataset was refined utilizing Google refine device to guarantee every one of the anomalies and invalid esteems were precluded.

Catastrophic information elements	Lowest	Highest	Mid- point	Group difference from mean value
Counts of automobiles	1	14	1.865	0.563
Street category	1	4	2.653	1.956

A				
Street Exterior	1	6	1.079	0.450
Flash Situations	1	7	2.895	8.989
Atmospheric Conditions	1	6	1.569	0.762
Casualty Class	1	4	2.365	0.869
Sufferer Severity	1	3	2.856	0.503
Sex of Sufferer	1	2	1.658	0.562
Age of Sufferer	1	99	27.11	15.294
Categories of automobiles	1	9	8.625	3.654

Thus, the above mentioned table has manifested the fact that how many road accidents in the UK. It has been realized that almost 14 vehicle are involved in a road accident in an approximate manner. On the other hand, the mean value for this is that 0.563 which could be calculated through Eq. within association rule mining. Thus, the above classifications are able to exhibit the dataset of UK road accidents yearly and in the following manner the method would be applied in terms of evaluating and analyzing the acute reason of road accidents.

Methodology

After examination of the extensive variety of rationality open the maker has received the Cross Industry Standard Process for Data Mining (CRISP-DM) approach. Each one of the stages in CRISP are clung to in working up the examination which include MarketPerception,

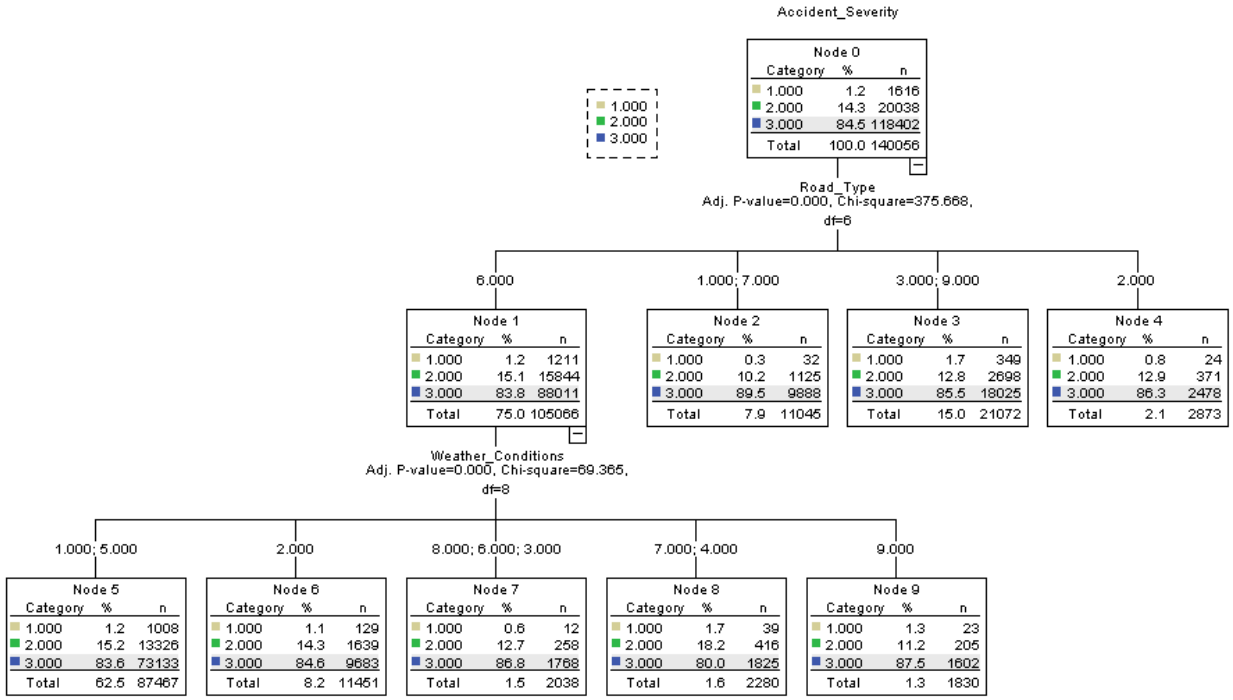
InformationPerception, Information Preparation, Modeling, Assessing and identifying and course of action. The basic techniques join Decision tree, Naive Bayes show, Bayes Boosting framework.

Application

The examination was coordinated under cluster analysis. The cluster analysis frameworks recorded underneath were deduced as a part of this wander i) Decision Tree ii) Naive Bayesian Classifier iii) Ensemble (Bayes Boosting)

Decision Tree classifier

Decision Tree Classifier is a straightforward and broadly utilized order strategy. It applies a strait forward thought to take care of the characterization issue. Decision Tree Classifier represents a progression of precisely created inquiries regarding the qualities of the test record. Each time it gets an answer, a subsequent inquiry is gotten some information about the class name of the record is come to.



From the decision tree, it can be observed that 1.2% of the accident cases can be classified as 'lethal' (coded 1), 14.3% of them can be classified as 'marginal'(coded 2), while 84.5% of the accidents can be classified as 'grave'(coded3). The model is set to predict 'Accident severity'. The best predictor of accident severity is 'road type'. For road category 6, 1.2% of the accident cases are lethal, 15.1% of them are marginal, while 83.8% of them are grave. For road types 1 and 7, 0.3% of the accidents are lethal, 10.2% of them are marginal, while 90.5% of them are grave. For road category 2, 0.8% of the accidents are considered lethal, 12.9% of them are considered to be marginal, while 86.3% of them are grave.

Risk

Estimate	Std. Error
.155	.001

Risk

Estimate	Std. Error
.155	.001

Growing Method: CHAID

Dependent Variable: Accident_Severity

Classification

	Predicted			
Observed	1.00	2.00	3.00	Percent Correct
1.00	0	0	1616	.0%
2.00	0	0	20038	.0%
3.00	0	0	118402	100.0%
Overall Percentage	.0%	.0%	100.0%	84.5%

Growing Method: CHAID

Dependent Variable: Accident_Severity

Random- Over- Sample (ROS)

The random over sampling is a technique through which the data are being classified based on the typical dataset in relation to the information accumulated regarding road accidents in the UK. On the other hand, random over sample also refers to the excess use and gathering of data from the association rule data mining.

Random- Under Sample

In the case of random under sample is also a technique in relation to data mining in which the cluster samples are replaced by the cluster centroids and these are the cluster algorithms that delivers value to K means.

Naïve Bayesian Classifier

Naive Bayes classifiers are exceedingly versatile, requiring various parameters direct in the quantity of factors (highlights/indicators) in a learning issue. Greatest probability preparing should be possible by assessing a shut shape articulation, which takes direct time, as opposed to by costly iterative estimation as utilized for some different sorts of classifiers (Ma, 2008).

p(CjF1Fn) = p(C)p (F1: FnjC) =p (F1: Fn)

Two step cluster

Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	5

Cluster Quality

Multiple Model Accumulation and getting combined effect

While the quantity of part classifiers of a troupe greatly affects the precision of expectation, there is a predetermined number of studies tending to this issue. From the earlier deciding of outfit estimate and the volume and speed of enormous information streams make this much more pivotal for online group classifiers. For the most part, factual tests were utilized for deciding the best possible number of segments. All the more as of late, a hypothetical system proposed that there is a perfect number of segment classifiers for a troupe which having pretty much than this number of classifiers would break down the exactness. It is known as the theory of consistent losses in group development. Their hypothetical system demonstrates that utilizing an

indistinguishable number of autonomous part classifiers from class marks gives the most noteworthy exactness(Sohn, 2003).

Initial Cluster Centers

	Cluster	
	1	2
Accident_Severity	2.00	3.00
Number_of_Casualties	38.00	1.00
Road_Type	6.00	1.00
Weather_Conditions	1.00	9.00

Iteration History[a]

Iteration	Change in Cluster Centers	
	1	2
1	6.331	8.579

2	4.531	.000
3	3.305	.000
4	2.389	.000
5	1.408	.000
6	2.360	.000
7	2.132	.001
8	2.552	.001
9	2.429	.002
10	2.018	.005

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is 2.018. The current iteration is 10. The minimum distance between initial centers is 38.197.

Iteration History[a]

Iteration	Change in Cluster Centers	
	1	2
1	6.331	8.579
2	4.531	.000
3	3.305	.000
4	2.389	.000
5	1.408	.000
6	2.360	.000
7	2.132	.001
8	2.552	.001
9	2.429	.002
10	2.018	.005

Iteration History[a]

Iteration	Change in Cluster Centers	
	1	2
1	6.331	8.579
2	4.531	.000
3	3.305	.000
4	2.389	.000
5	1.408	.000
6	2.360	.000
7	2.132	.001
8	2.552	.001
9	2.429	.002
10	2.018	.005

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is 2.018. The current iteration is 10. The minimum

Final Cluster Centers

	Cluster	
	1	2
Accident_Severity	2.59	2.83
Number_of_Casualties	7.11	1.31
Road_Type	4.65	5.17
Weather_Conditions	1.39	1.51

Number of Cases in each Cluster

Cluster	1	543.000
	2	139513.000
Valid		140056.000
Missing		.000

Estimation

The procedure of affiliation rules with an expansive arrangement of mischance's information to recognize the reasons of street mishaps were utilized. Examination demonstrated that delivering the affiliation rules, makes distinguishing proof of components required in the mischance that happen together, less demanding. It shares a considerable measure in understanding the conditions and reasons for the mischance. So, the affiliation lead mining gives the bearing to further research on the reasons for street mishaps. It encourages government to adjust the movement security strategies with various sorts of mishap and circumstances. The fundamental aftereffect of this investigation is that in spite of the fact that the qualities of humankind and conduct are vital in event of all street mischances however we can comprehend that spatial components and framework assume a noteworthy part in the mishap.

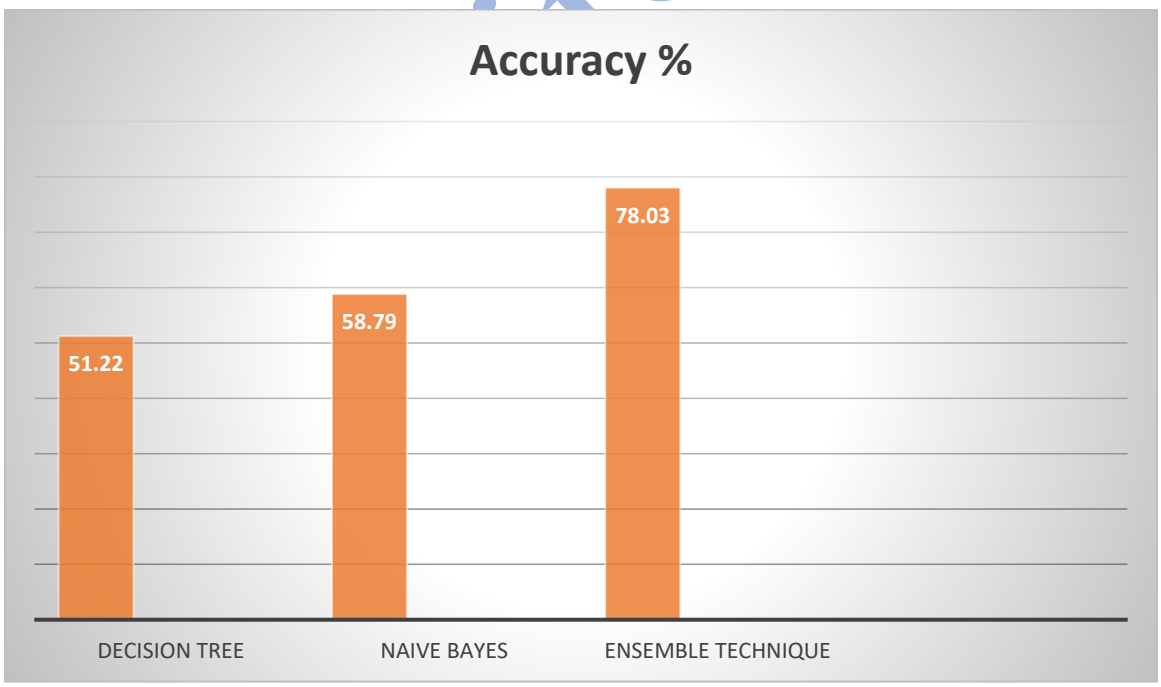


Figure 2: Performance Accuracy of Models

Analysis of the findings:

From the above analysis and data mining methods the findings could be realized and observed through the data analysis. It has been realized that data mining has been proven as a reliable tool of delivering data regarding road accidents in the UK. Different countries and researchers have used several types of data mining techniques in order to deliver a systematic technique of evaluating reasons behind road accidents in the UK. Yearly more than 0.4 billion people die and get heavily injured while getting involved in the roads accidents. Thus, among the several data mining techniques here the cluster analysis data mining techniques has been utilized in order to analyze the acute data regarding road accidents in the UK. The loss of life identified with cars is positioned as the most elevated reason for death in creating nations, where most of the fatalities are in a youthful age section (Wang, 2009). The number of inhabitants in a nation additionally has an influence in the reason for street crashes. The nations with little and mid-level wage have less vehicles. Notwithstanding, they add to 90 percent of street mishaps around the world. This is expected to the absence of street security laws and poor upkeep of streets and insufficient medicinal treatment amid crises. The nations with higher earnings have less commitment towards street mishaps around the world, this could be because of upkeep of streets, and strict street securely laws by giving pace constrains wherever is required. Then again, in very much created nations open transport are exceedingly adulated and mishaps have a tendency to be less. These are by all record not by any means the only inspiration of incidents; the vehicle deliver moreover has a genuine influence in fatalities. In the mid-level countries, the prevailing piece of vehicles is sold with ease and don't meet the principal security benchmarks determined by the road prosperity division. The used vehicle is sold without appropriate overhauling, particularly autos that are sold at an extremely low cost. These autos are the most elevated supporters of street mischances (Romano, 2012). It is noticed that as it were 30 percent of nations around the

world urges individuals to walk or cycle. Actually, street mishaps are not quite recently caused by individuals going in autos, open transport or, on the other hand substantial vehicles, the casualty of mischances are people on foot, cyclist and bikes too. It is evaluated that 20 to 15 rates of the fatalities are helpless street clients. Thus, the cluster analysis have evidently delivered the dataset that proved in an yearly manner more than 18,572 in average witness the road rages. Hence, the road accidents have reduced in the last few years due to the proper data mining techniques implementation. Therefore, depending on the fact organizational or professional decision could be taken in terms of including the cluster data mining techniques within the planning and design of roads and other geographical circumstances would be relevant and effective in case of reducing road accidents in the UK.

Conclusion

The point of this exploration extend was to recognize the best calculation to distinguish the severity of street mischance's in Leeds. The outcome appeared in the usage segment of each calculation had a superior outcome contrasted with different calculations, for example, liner relapse which are most appropriate for classification issues. In this investigation, it is attempted to pick the fascinating and better standards than give a considerable measure of significant data for strategies to give better wellbeing arrangements. A normal of three thousand street crashes happen because of cars at regular intervals. There is different explanation behind mishap to happen which is as yet flighty. Truth be told, a large portion of the mischance commitments around the globe are on the parkways and paths. This article can be a stage towards giving valuable data to thruway architects and transportation fashioners to plan more secure streets. It can be accomplished through recognizing significant designs inside every section and converging into one segment to playing out the forecasts. The greater part of traffic mishap datasets has comparative qualities which gives the same data in a different way. In this manner,

the above talked about determination system can be utilized to recognize the example for the reason and seriousness of the mischance later on.

References

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 359-364.

Brennan, P. (2012). A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection. *Institute of technology Blanchardstown Dublin, Ireland*.

Chang, L.-Y. C.-C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, 365-375.

Driss, M. S.-G. (2013). A fuzzy logic model for identifying spatial degrees of exposure to the risk of road accidents (case study of the wilaya of mascara, northwest of algeria). *Advanced Logistics and Transport (ICALT), 2013 International Conference on, IEEE*, 69-74.

Ebrahimi, M. H. (2015). Sleep habits and road traffic accident risk for iranian occupational drivers. *International journal of occupational medicine and environmental health*, 305-312.

Edwards, J. (1998). The Relationship between Road Accident Severity and Recorded Weather. *Journal of Safety Research*, 249-262.

Fridstrøm L, I. J. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, 1-20.

Harb, R. Y. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, 98-107.

Karlis, D. (2003). An EM Algorithm for Multivariate Poisson Distribution and Related Models. *Journal of Applied Statistics*, 63-77.

Ma, J. K. (2008). A multivariate poisson-lognormal. *Accident Analysis & Prevention*, 964- 978.

Nitesh V. Chawla, K. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321–357.

Parmentier G, C. J. (2005). Road mobility and the risk of road traffic accident as a driver. The impact of medical conditions and life events. *Accident Analysis & Prevention*, 1121-1134.

Romano, E. O. (2012). *Journal of safety research*, 75-82.

Shankar V, M. F. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies.

Sohn, S. Y. (2003). *Safety Science*, 1-14.

Sowmya, M. a. (n.d.).

UK Government. (n.d.). *Road-accidents-safety-data*. Retrieved from https://data.gov.uk: https://data.gov.uk/dataset/road-accidents-safety-data/resource/6d253c0f-caa4-4eafaa85-464dc48252da

Wang, C. Q. (2009). Impact of traffic congestion on road accidents: a spatial analysis of the m25 motorway in England. *Accident Analysis & Prevention*, 798- 808.

Wang, C. Q. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention*, 1979- 1990.

Wu, K.-F. D. (2013). *Journal of Transportation Engineering*, 738-748.