National College of Ireland

Analysis of Road accident in Leeds

MSc Research Project
Data Analytics

Syed Ibrahim Kabeer
x15029158

School of Computing
National College of Ireland

Supervisor: Jason Roche

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Syed Ibrahim Kabeer
Student ID:	x15029158
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Jason Roche
Submission Due Date:	12/12/2016
Project Title:	Analysis of Road accident in Leeds
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	21st December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis of Road accident in Leeds

Syed Ibrahim Kabeer
x15029158
MSc Research Project in Data Analytics

21st December 2016

Abstract

An increase in road traffic accidents in Leeds has drawn attention to existing safety measures in place and the requirement for additional safety measures to be implemented. Data relating to traffic accidents has been made available for analysis to determine the best way to increase road safety. The available accident data has been analysed in various dimensions to identify the relationship between the severity's of the accidents. Determining an accurate result from this data set will help the automobile companies and Leeds city council prevent the accident re-occurring in the future. This research paper will determine the best algorithm model to use in order to improve the accuracy and predictions of road accidents in machine language programming. This would also help to improve the accuracy of the classifier in the road accident. The author applies three methods in this research paper, Decision tree, Naive Bayes and the Ensemble technique. This will improve the efficiency of identifying the severity of road crash accidents. The research project will determine the most suitable classification algorithm that can be used in future for prediction of road accidents.
Keywords Road Accident, Decision Tree, Naive Bayes, Ensemble technique (Bayes Boosting)

1 Introduction

It is estimated that in a single year, 2.5 million road accidents occur and nearly 1.25 million people die as a result of these accidents, or are permanently injured. An average of three thousand road crashes occur due to automobiles every 24 hours. There is various reason for accident to happen which is still unpredictable. In fact, most of the accident contributions around the world are on the highways and lanes. The death toll related to automobiles is ranked as the highest cause of death in developing countries, where most of the fatalities are in a young age bracket. The population of a country also plays a part in the cause of road crashs. The countries with small and mid-level income have fewer vehicles. However, they contribute to 90 percent of road accidents worldwide. This is due to the lack of road safety laws and poor maintenance of roads and insufficient medical treatment during emergencies. The countries with higher incomes have less contribution towards road accidents worldwide, this could be due to maintenance of roads, and strict road safely laws by providing speed limits wherever is required. On the other hand, in well-developed countries public transport are highly praised and accidents tend to be

less. In fact, the greater the number of people travel in public transport, the greater the number of fatalities if a road crash occurs. For example, In Nigeria and South Africa, public transport is limited and the private rental vehicles provider fill up the cabins more than the seat capacities and they exceed the speed limits causing more fatalities on the road. These are not the only cause of accidents; the vehicle manufacture also plays a major role in fatalities. In the mid-level countries, the majority of vehicles are sold at low cost and dont meet the basic safety standards specified by the road safety department. The used vehicle are sold without proper servicing, especially cars that are sold at a very low price. These cars are the highest contributors to road accidents. It is noted that only 30 percent of countries around the globe encourages people to walk or cycle.

In fact, road accidents are not just caused by people travelling in cars, public transport or heavy vehicles, the victim of accidents are pedestrians, cyclist and two wheelers also. It is estimated that 20 to 15 percentages of the fatalities are vulnerable road users. Zebra crossings are common accident places for the vulnerable users. This is due to congestion on the road and people want to travel as quickly as possible to their destination. Alcohol and drug user is the highest contributor for accidents involving pedestrians, cyclist and two wheelers. A global road safety issue conference was held in Brasilia where officials from various countries came together to discuss road safety issues. In 2030, the program is to have more safety measures and reduce the number of road accidents. A statistic reveals the highest road accident occur in under developed countries, such as the Dominican Republic. It has the highest traffic death rate with an estimate of 41.7 percent out of 100,000 population. Thailand in second place with 38.1 percentage and Venezuela in the third place with 37.2 out of 100,000 population. This research project is an analysis of Road accident in Leeds, UK. There were 1732 deaths reported in 2015 and 1855 in 2014 and reason for the fatal are still unpredictable. The project is described in the analysis outcome. The details of the algorithm and approach can be found within each section of the analysis.

2 Related Work

The below section will describe the related work with respect to road accidents

2.1 Background

The foremost focus of earlier studies has identified the various attributes such as road segment, intersection, road surface and weather which has a significant issue in cause of crashes. This has led to various improvements in road safety measures. Despite the above said attributes cannot be the major contributor to crashes. Few of the researchers have investigated macro level properties such as cross junction, traffic zones, census. According to the Leeds City Council it has taken various measures to improve the transport strategy to avoid crashes. However, the advancement in statistical and computational model has become more acquiescent to perform severe analysis on various type of data irrespective the type of attributes.

In traffic crash analyses, there are a great many unobserved explanatory variables that affect frequencies and severities of accidents. To identify the hidden pattern is quite challenging. The reason most of the accident data is highly imbalanced due to which analysis becomes quite challenge.However, many research has been done by (Karlis; 2003)analysed the crash severity on the univariant count model. The use of univariant

model ignores the interdependency between the variables which can be used to perform the analysis on macro level information.

2.2 Literature Review

In the past, most of the research on road accidents were done based on only certain attributes to predict the cause of road accidents. shankar in early 90's (Shankar et al.; 1995) analysed the cause of road accident by studying the relationship between the geometric (e.g Horizontal and Vertical) element and weather condition. Similar study was done in the 90s by (Fridstrøm et al.; 1995)) however the both of their finding were biased by the fact that exposure levels of the attribute were not perfectly controlled. On one hand (Edwards; 1999)study the cause of accident using the weather condition. Perhaps all the analysis done in earlier 90s was not accurate to arrive at a conclusion for the cause of the accident. (Parmentier et al.; 2005) studied the cause of accident based on driver behaviour (Chang and Chen; 2005) indicated that the average daily traffic volume determinants the accident frequencies, but (Wang et al.; 2009) examined and stated the accident frequencies are not due to traffic congestion. Further upon, Numerous researches were done based on various attributes and stated traffic congestion is not the cause for the frequent accidents. (Anderson; 2009) identified the cause of accident using K-mean Clustering algorithm by identifying the accident hot spot using GIS information system.However (Romano et al.; 2012) identified the factors affecting the driving and crashes are related to alcohol related crashes (Ebrahimi et al.; 2015)examined the crashes are more high prevalence of poor sleep quality and sleep disorder but the evidence from both were biased. (Driss et al.; 2013) inspected the Geometric information system (GIS) data by integrating with fuzzy logic and neural network technique to explore the cause of the accident.The analysis had a quick response and result was quite satisfactory and predictable. (Wu et al.; 2013)examined the prediction of accident using association rule and artificial neural network and genetic algorithm and moreover they analysed the causality and severity of the accident which was acceptable.In conjunction (Harb et al.; 2009) study examined the causality of accident using Random forest and decision tree. This helped identifying the position of the vehicles, Driver characteristics on crash prevention manoeuvres result were then more precise compared to genetic algorithm. (Wang et al.; 2011) used a two-stage mixed multivariate model in predicting accident frequency at their severity levels which helped in identifying the low frequency accident. (Sowmya and Ponmuthuramalingam; n.d.)predicted the causality severity of the accident using Naive Bayes model technique. The accuracy of the the model was comparatively high than other models. The author (Brennan; 2012) studied the class imbalance of data using various tools and method in producing the accurate results. The data imbalance is a major problem in data mining approach, the author has give an very extensive approach of handling the class imbalance using boosting and other technique.

3 Methodology

After analysis of the wide spectrum of methodology available the author has adopted the Cross Industry Standard Process for Data Mining (CRISP-DM) approach. All the phases in CRISP are adhered to in developing the research which consist of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and testing

and deployment. The primary methods include Decision tree, Naive Bayes model, Bayes Boosting technique.

Figure 1: CRISP-DM Model

3.1 Business understanding

The initial phase of business understanding is to identifying the objective of the research project and convert the objective to business needs. Then the business needs were converted into the data mining problem definition.

3.2 Data Understanding

The data required to perform the analysis of road accident in Leeds was obtained from `https://data.gov.uk/dataset/road-traffic-accidents`. The accident data was split across each year from 2009 to 2015. The extracted data was analysed to ensure it meets the business requirement with the help of a data description document. The relationship between attributes was checked and simple aggregation was done. The exploration of data was done by doing initial hypothesis and by plotting graphs trough unsupervised learning. The dataset and its description are described below.

Attributes	Description	Numerical value
Road class	Motorways	1
	AM Class Road	2
	B Class Road	3
	C Class Road	4
	Unclassified Road	6
Road Surface	Dry	1
	Wet	2
	Snow	3
	Frost	4
	Flood Surface	5
Casualty Class	Driver	1
	Vehicle Passenger	2
	Pedestrian	3
	Rider	4
Casualty Severity	Fatal	1
	Serious	2
	Slight	3
Sex	Male	1
	Female	2
Weather Condition	Fine	1
	Raining	2
	Snowing	3
	Fog or Mist	4
	other	5
	unknown	9
Type of Vehicle	Pedal Cycle	1
	Motor Cycle	2
	Bus	3
	Good Vechile	4
	other Vehicle	5
	Taxi/Car	9
	Tram	7
AccidentDate	Date	YYYY/MM/DD
TIME	24HrsFormat	HH:mm
Number of Vehicles	Provides the No	1 to 9

Table 1: Accident variables

3.3 Data preparation

The dataset used in this study was derived from a total of 18,572 reported traffic crashes in Leeds, United Kingdom. The crash details are provided according to each year. Preliminary analysis was done to ensure all the data is captured during the compilation. The dataset was refined using Google refine tool to ensure all the outliers and null values were omitted. The initial setup is to identify the attributes to fit the modelling, however the data was imbalanced.

Accident Dataset Variables Rows 18572 from 2009 to 2015	Min	Max	Average	Std. Deviation
Number of Vehicles	1	14	1.930	0.825
1st Road Class	1	4	4.373	1.676
Road Surface	1	6	1.279	0.560
Lighting Conditions	1	7	2,146	1.907
Weather Conditions	1	6	1.274	0.790
Casualty Class	1	4	1.806	0.968
Casualty Severity	1	3	2.877	0.350
Sex of Casualty	1	2	1.408	0.491
Age of Casualty	1	98	35.13	18.294
Type of Vehicle	1	9	7.041	3.146
Reference Number				
Grid Ref: Easting		447413		
Grid Ref: Northing		449895		
Accident Date		31/12/2015		
Time (24hr)		HH:MM:SS		

Table 2: Road Accident Safety Data Guidance

3.4 Methods for overcoming imbalance data

The problem with the road accident dataset was the predictive variable class with lower majority class and higher minority class. However, the data mining approach would not best for all the algorithms, certain measures were taken to overcome the class imbalance. The lower majority class variables were compared with higher minority class variables to identify the patterns within the data. The problem was addressed with the help of data fragmentation, divide, and conquer the data. The Brennan,2012 comprehensive studied addressed this issue and the algorithm that would process the imbalanced data. The RIPPER algorithm was used to handle the class imbalance problem. The Random oversampling(ROS) and Random under sampling (RUS) can be performed for class imbalance. Another approach would be to Re-Sampling the data a studied done by author Chawla,2002 wrote on SMOTE algorithm of oversampling the minority class and under sampling the majority call by achieving the ROC space. Another way of addressing the class imbalance is to assign distinct cost to training examples. Many experiments were carried out with different dataset and using machine language algorithm such as decision tree classifier, Ripper, and a Naive Bayes Classifier to improve the class performance over the ROC. The author has used the process of Re-sampling technique such as boosting in Rapid miner to avoid the class imbalance in predictive analysis. The further details on handling imbalance is described in the Data Modelling technique section.
The TN specifies the number of negative samples correctly classified, FP number of negative samples incorrectly classified, FN- positive samples incorrectly classified as negative, TP is the number of positive sample correctly classified. The accuracy of a predictive model is factored by the precision and recall factor. The accuracy metrics best fits for balance dataset when the error rate is -1, whereas in the imbalance dataset its best to use ROC and AUC cure to measure the accuracy.

$$Precision = (TP) = (TP + FP)$$

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 3: Confusion Matrix

$$Recall = (TP) = (TP + FN)$$

Precision is the sum of TP and FN, and Recall is calculated by TP and FN as stated above.

4 Implementation

The experiment to carry out the prediction was done by supervised and unsupervised learning techniques. The unsupervised learning techniques listed below were adopted as part of this project i) Decision Tree ii) Naive Bayesian Classifier iii) Ensemble (Bayes Boosting)

4.1 Decision Tree classifier

The decision tree classification model algorithm was achieved by Rapid miner tool. The decision tree is like an inverted tree which grows from top down. The aim of the model is to create a classification model that predicts target attributes on the causality severity of the given dataset. The implementation of a decision tree is done by three processes through Random Oversampling (ROS), Random Under Sampling (RUS) and then with the complied Raw data. The Causality severity variable was used a predictor variable, in Rapid Miner tool its marked as label. The indepent variable used for prediction were Type of vehicle, Time of the day, Weather, Causality severity, Causality Class,Sex,Age, Lighting Condition, Road Class,

4.2 Random Over sample-ROS

Random oversampling is done by replicating the minority classes which is equal to the majority classes, it was achieved by boosting the label attribute. The Causality severity attributes, Serious, Fatal and Slight were highly imbalanced, with the Fatal of 136 record out of 18572 records. The fatal records where boosted with the help of multiply and Sample Stratified, Select Attributes operators in rapid miner. The boosted records were then appended with the existing Serious, Slight records which increased the record count for 18572 to 27790 records. The records where split into test and train with a ratio of 0.7: 0.3. The train set and test set attributes where processed with 10-fold cross validation process. The decision tree model operator is placed in the training tab, with maximal tree depth as 10, confidence =0.05. The advance parameters such as pruning and apply pre-pruning was selected to avoid unwanted branches (over fitting) in the decision tree creation. The model predicted an accuracy of 45 percentage of accuracy. Figure 14

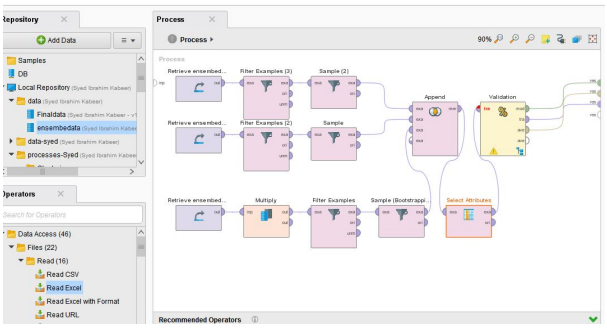

Figure 2: ROS Model Decision Tree

The Fig 2. illustrates the implementation of Decision tree using ROS bootstraping operator.

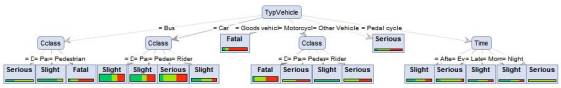

Figure 3: ROS Model Decision Tree

4.3 Random Under sample

The Random under sample was achieved by selecting the majority classes, the matches, and the minority classes. The RUS was then implemented by using the filter operator in rapid miner. The dataset was selected based on an absolute split of 50/50. The model was inducted with 8-fold cross validation. The maximal tree depth = 5 with the confidence level=0.05, with pruning and rep-pruning was selected. The model predicated an accuracy of 51.22 percentage. Further down, then decision tree was applied with the

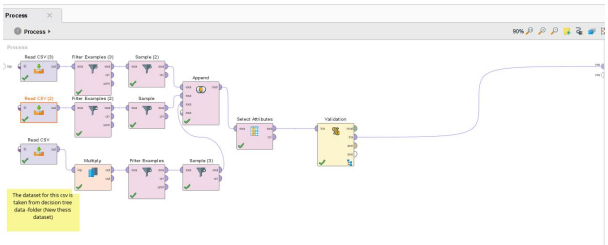

Figure 4: RUS Model Decision Tree

complied dataset with the tree depth of 10, confidence=0.05, pruning and pre-pruning was selected. The model predicted an accuracy of 61.00 Percentage.

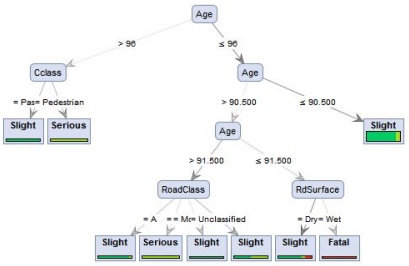

Figure 5: Decision Tree

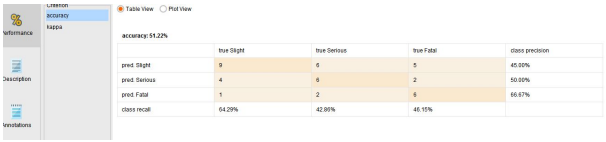

Figure 6: Confusion Matrix

4.4 Confusion matrix

4.5 Naive Bayesian Classifier

Naive Bayesian algorithm has been adopted in many countries to identify the road accident analysis (Ma et al.; 2008) has adopted the algorithm in predicting the severity of the accident which assisted the Road transport department in identifying the crash severity of the accident. Naive Bayesian estimation method is a probabilistic classifier which generally produces a multivariate posterior distribution across all the independent assumption between the attributes. The algorithm uses Bayes theorem to identify the conditional probability

$$p(C|F1Fn) = p(C)p(F1.Fn|C)/p(F1.Fn)$$

ETL Processing: The extracted data was loaded in the comma separated format using the read comma separated operator in rapid miner, and the predictor variable of severity is selected as the label while loading the data. The dataset was extensively processed to ensure there was no missing values. The extracted data was transformed to the store operator which can be referred to in future use. To handle the imbalance in the dataset, the Boot strapping stratified sample method was used. The Stratified sample builds a random subset as sample as the example subset before processing.
Model Execution The Nave Bayes Kernel algorithm was achieved by using the Rapid Miner tool. The tool has the Naive Bayes kernel operator. The parameters Kernel=10, Estimation method =greedy Minimal kernel bandwidth =0.1 and Laplace correction were used to perform the prediction with the 10-fold cross validation, sampling type = Stratified sampling. The Causality severity of the variable was predicted with 78.03 percentage accuracy.

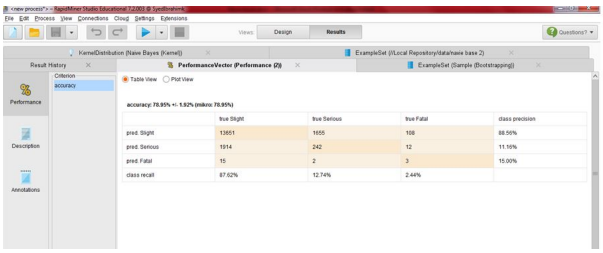

Figure 7: NBayes Confusion Matrix

4.6 Ensemble technique (Bayesian Boosting)

The Bayesian boosting is an ensemble technique which can be used in conjunction with any machine learning algorithm. It is a nested process, where it can handle any process within the sub process to provide a higher level of accuracy. The (Sohn and Lee; 2003) analysed the causality and severity of the accident using the Decision tree and SVM to check the accuracy of the model. The model had high accuracy compared with various classifier the performance of Bayesian boosting was quite impressive and high. The prediction of this model has been used by the Korean transport department to reduce the fatalities in road accidents. The model is currently being used in automobiles industries to identify the crash severity of the vehicle.

The Bayesian boosting technique has been achieved with the help of rapid miner tool. The Bayesian boost operator can handle number of sub process, as part of the analysis decision tree has been implemented in the sub process of Bayesian model. The test and train data was split with a relative ratio of 0.7 and 0.3. A 10-fold cross validation was carried out with sampling type = Stratified sampling. The tree depth of the decision model was limited to 10 by applying the pruning and pre-pruning process. The Bayesian boosting was processed with an iteration of 10 to ensure the model performs as expected. The parameter of skewing of the marginal distribution (P(x)) was set to 1 to ensure the algorithm to perform the learning. The model was executed and displayed an accuracy of 78.3 percentage.

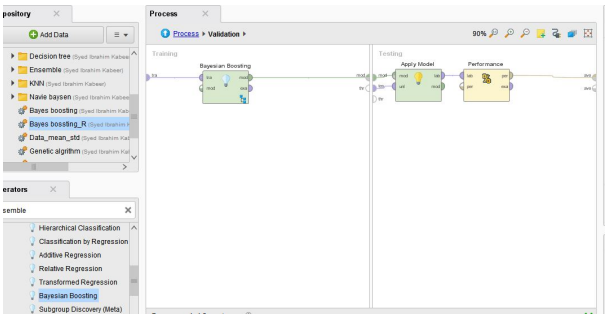

Figure 8: Bayesian Boosting

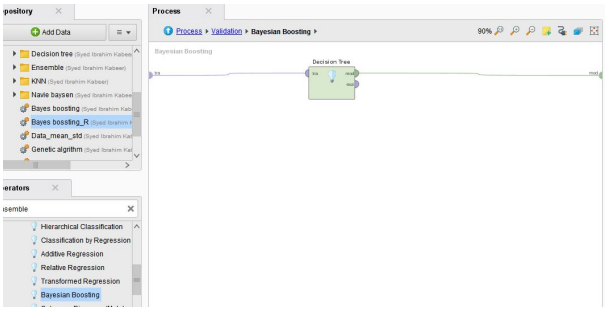

Figure 9: BayesianBoosting Subprocess

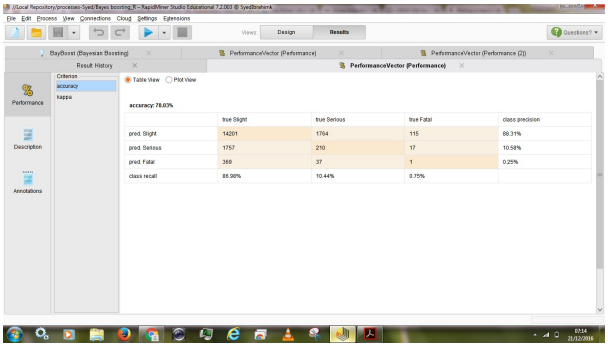

Figure 10: Bayesian Boosting Confusion Matrix

5 Evaluation

The performance accuracy has been evaluated against the each model with the road crash dataset. The below result depict the Ensemble technique Bayesian Boosting provide more accuracy in the result compared to Decision Tree and Naive Bayes.Inspite there are many other upcoming algorithm which can perform better accuracy the change the attributes will just make a difference. The future engineering technique such as Neural networks and genetic algorithm can be applied to get a higher accuracy in the prediction or by applying discover feature engineering technique.

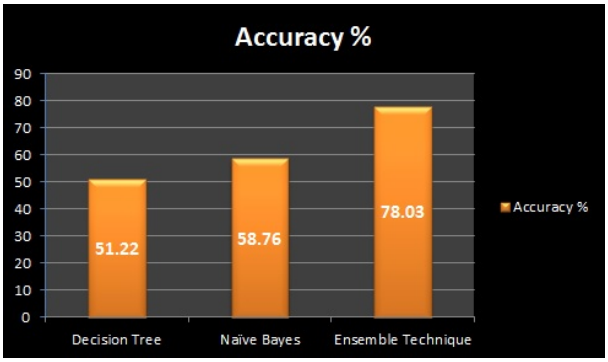

Figure 11: Performance Accuracy of Models

5.1 Experiment / Case Study 1

To identify the type of vehicle which causes maximum crashes in a day based on the causality and severity class ?

Outcome : The analysis predicts that the maximum crashes are caused by Cars, which include Serious, Slight injuires and Fatal. The fatalities are less in Buss compared to other vehicles.Fatalities are greater during the morning and early morning time by the Goods Vehicles . Whereas the fatalities are less in other vehicles. To conclude cars have the highest frequency rate in road crash.

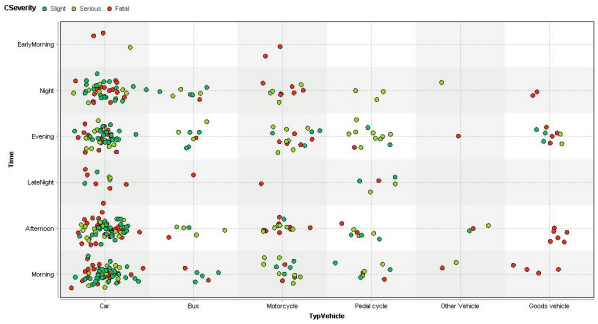

Figure 12: Typevhcle Vs Time of the Day

5.2 Experiment / Case Study 2

To identify the maximum crash in the month of a year in Leeds, UK from the year 2009 to 2015.

Outcome : The analysis predicts more fatalities are prone to occur in the month of July and August in the year 2009 to 2015. The climate is generally warm during the month of July and August which indicates that accidents are more frequent during the above mentioned months. The fatalities are lower in the months of February and March. But the crash severity such as serious and slight sees a small up and down during each month.

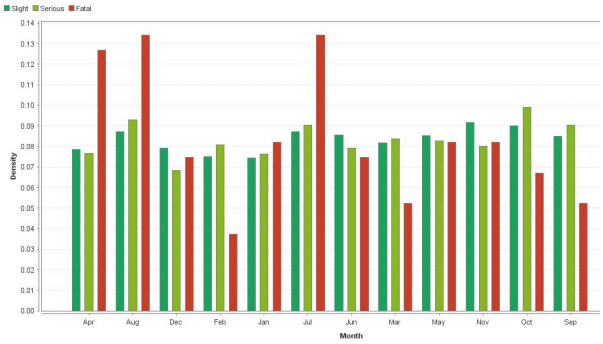

Figure 13: TypeVhcle Vs Year

5.3 Experiment / Case Study 3

To identify the highest frequency of road accident based on the road class?

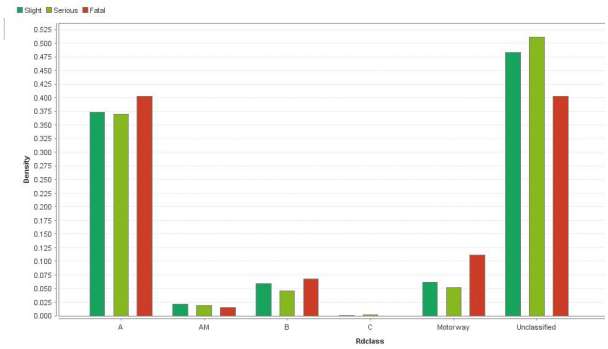

Figure 14: Crash Density Vs Road Class

Outcome : The analysis illustrates that accidents tend to happen more frequently in the unclassified road types, which is generally the backroad in the Leeds city that is intended for local traffic. The next highest frequency of crashes is in the Class A roads which are intended to provide transport to large-scale links between the areas. The accident are much less likely to occur on Road Class C which are minor roads in villages.

6 Conclusion and Future Work

The aim of this research project was to identify the best algorithm to identify the severity of road accidents in Leeds. The result shown in the implementation section of each algorithm had a better result compared to other algorithms such as liner regression which are best suited for classification problems. However, with respect to classification accuracy, the decision tree method was impressive in finding the root attributes and providing classification on the important variable such as Age, Road class, Causality class and Road surface. The tree had the minimal branches required to provide a proper classification approach. In terms of classification actuary, Nave Bayesian algorithm performed better compared to the decision tree with an accuracy of 78 percent and the nearest kernel value of 10. In fact, the ensemble technique of Bayes Boosting showed a marginal improvement in accuracy compared to the other two models. The Bayes model displayed an accuracy of 88.38 percent in predicting the causality and severity of the accident. In conclusion, the causality severity may not be the only attribute which causes accidents. The vehicle also plays a major role in crashes. There is drastic improvement in the data mining algorithm each day, however more the attributes and bigger the dataset. A new concept of Discover Feature Engineering Selection is upcoming in the machine language world in reducing the dimensionality of the dataset. This is achieved by identifying the relevant patterns within each column and merging into one column to performing the predictions. The majority of traffic accident datasets have similar attributes which provides the same information in a different way. Therefore, the above discussed selection technique can be used to identify the pattern for the cause and severity of the accident in the future. Future work: The Future Prediction of road crash can be performed on the the leeds data set, if the data set have the vehicle manufacturing date and the make of model .

This would help to calculate the age of the vehicle which would help in analysis the cause of the accident. The research project has identified most of the accident in Leeds are caused by cars. The car manufactures can also identify trend of the vehicle performance after certain period of time.A current recent study has identified small size cars are most likely to cause accident. `http://www.usatoday.com/story/money/cars/2015/01/29/iihs-driver-death-cars-top-10/22536459/`

Acknowledgements

The Thesis on Analysis of Road Accident in Leeds is done at National College of Ireland. I would like to express my sincere thanks to my supervisor Mr. Jason Roche for his motivation, support and guidance during the tenure of the project. I would also like to thank my Family and friends who supported me during this journey.

References

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots, *Accident Analysis & Prevention* **41**(3): 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. and Chen, W.-C. (2005). Data mining of tree-based models to analyze freeway accident frequency, *Journal of Safety Research* **36**(4): 365–375.

Driss, M., Saint-Gerand, T., Bensaid, A., Benabdeli, K. and Hamadouche, M. A. (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, pp. 69–74.

Ebrahimi, M. H., Sadeghi, M., Dehghani, M. and Niiat, K. S. (2015). Sleep habits and road traffic accident risk for iranian occupational drivers, *International journal of occupational medicine and environmental health* **28**(2): 305–312.

Edwards, J. B. (1999). The relationship between road accident severity and recorded weather, *Journal of Safety Research* **29**(4): 249–262.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R. and Thomsen, L. K. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts, *Accident Analysis & Prevention* **27**(1): 1–20.

Harb, R., Yan, X., Radwan, E. and Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests, *Accident Analysis & Prevention* **41**(1): 98–107.

Karlis, D. (2003). An em algorithm for multivariate poisson distribution and related models, *Journal of Applied Statistics* **30**(1): 63–77.

Ma, J., Kockelman, K. M. and Damien, P. (2008). A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods, *Accident Analysis & Prevention* **40**(3): 964–975.

Parmentier, G., Chastang, J.-F., Nabi, H., Chiron, M., Lafont, S. and Lagarde, E. (2005). Road mobility and the risk of road traffic accident as a driver: the impact of medical conditions and life events, *Accident Analysis & Prevention* **37**(6): 1121–1134.

Romano, E. O., Peck, R. C. and Voas, R. B. (2012). Traffic environment and demographic factors affecting impaired driving and crashes, *Journal of safety research* **43**(1): 75–82.

Shankar, V., Mannering, F. and Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies, *Accident Analysis & Prevention* **27**(3): 371–389.

Sohn, S. Y. and Lee, S. H. (2003). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in korea, *Safety Science* **41**(1): 1–14.

Sowmya, M. and Ponmuthuramalingam, P. (n.d.). Analyzing the road traffic and accidents with classification techniques.

Wang, C., Quddus, M. A. and Ison, S. G. (2009). Impact of traffic congestion on road accidents: a spatial analysis of the m25 motorway in england, *Accident Analysis & Prevention* **41**(4): 798–808.

Wang, C., Quddus, M. A. and Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model, *Accident Analysis & Prevention* **43**(6): 1979–1990.

Wu, K.-F., Donnell, E. T., Himes, S. C. and Sasidharan, L. (2013). Exploring the association between traffic safety and geometric design consistency based on vehicle speed metrics, *Journal of Transportation Engineering* **139**(7): 738–748.

A First Appendix Section

...