BINARY CLASSIFICATION AND LOGISTIC REGRESSION MODELS APPLICATION TO CRASH SEVERITY

Guilhermina A. Torrão, MSc.
Graduate Student, Mechanical Engineering
Centre for Mechanical Technology and Automation and Department of Mechanical Engineering, University of Aveiro
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
Phone: (+351) 234 378 181
E-mail: guilhermina.torrao@ua.pt

Nagui M. Rouphail, Ph.D.
Director, Institute for Transportation Research and Education (ITRE)
Professor of Civil Engineering, North Carolina State University
Raleigh, NC, 27695-8601, USA
Phone: (919) 515-1154
E-mail: rouphail@eos.ncsu.edu

Margarida C Coelho, Ph.D.
Assistant Professor, Mechanical Engineering
Centre for Mechanical Technology and Automation and Department of Mechanical Engineering, University of Aveiro
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
Phone: (+351) 234 370 830
E-mail: margarida.coelho@ua.pt

ABSTRACT

This research explores the factors contributing to crash severity. This study main goal was to develop a binary classification modelling approach to predict the probability that a crash will result in severe outcomes. Real world crash data were collected from the Portuguese Police Republican National Guard records for the Porto metropolitan area, for the period 2006-2010. A total of 1,374 crash observations involving light duty vehicles resulting in injuries and/or fatalities were analysed. In this paper, the effect of vehicle characteristics, such as weight, engine size, wheelbase and registration year (age of vehicle) were analysed using data mining approaches in order to extract patterns from the predictors and relate them to the occurrence of injuries and fatalities in a crash.

Two types of predictive approaches were tested: Classification and Regression Trees (CART) and logistic regression models.. Logistic regression models for two vehicle crashes are not reported in this paper. The response variable FatalSIK was created to signify the probability of a serious injury or fatality, in a crash, and was defined as a binary variable (=1 for a severe crash, 0 otherwise). To conduct the analysis with high imbalanced data, oversampling followed by correction of prior probabilities for the original crash population was applied.

The CART analysis results for two-vehicle collisions show that the weight of the vehicles involved was the most important variable for FatalSIK prediction, followed by the road speed limit. The highest percentage of severe crashes occurred in collisions involving heavier vehicles and in high speed routes (Fisher's exact p-value <4.539E-4). For single-vehicle crashes, the engine size of the vehicle was the most important explanatory variable, followed by the vehicle's age. For those crashes, the highest percentage of severe crashes was associated with vehicles with a larger engine size (>=1588cm^3) and those that were older (=4.5 years) (Fisher's exact p-value <0.0164). The logistic regression model for single vehicle crashes incorporated the vehicle age, engine size, wheelbase, and weather conditions as predicts of severe crash outcomes (Chi-Sq p-value <0.0004). The single vehicle crash model correctly classified 76.4% of the observations.
These research findings provide important information for both public decision makers and automotive industry personnel to continue to produce ecologically friendly vehicles while still achieving improved protection of the vehicle occupants in the event of an injury crash.
Keywords: CART, Logit Model, Crash Severity, Variable Relative Importance, and Rare Events

1. INTRODUCTION

More than 1.2 million people die on the world´s roads every year (WHO, 2009a). Road traffic accidents in the European Union (EU) estimated 34,000 deaths and more than 1.1 million people injured, representing estimated costs of 140 billion Euros (WHO, 2009b).
Hermans et al. (2009) have shown that Portugal had the lowest safety performance score, which suggests that Portugal should invest more in vehicle technology and in new(er) cars. The Portuguese National Authority for Road Safety (ANSR) show that during the year 2011, there have been a total of 32,541 crashes involving injuries and those resulted in 2,436 serious injured and 689 fatalities on the Portuguese roads (ANSR, 2012).
There has been an increase in the amount of consumer interest in the safety performance. However, consumers tend to equate vehicle safety with the presence of specific features or technologies rather than with vehicle crash safety/test results or crashworthiness (Koppel et al., 2008). Crash testing is a resource for consumer regarding vehicle crash safety and credits a car manufacturer for focusing on safety. EuroNCAP discourage consumers from comparing ratings of cars from different segments, and in real crashes, there is obviously no control on the vehicle categories involved (Coelho et al, 2010). Despite the scientific conditions under which crash tests are conducted, they have limitations: first, they do not

account for mass differential between the vehicles involved within the collision; second, on real roads the speed of the crash impact frequently is higher than 64 km/h, which is the speed at the frontal impact takes place in crash testing conducted by EuroNCAP (EuroNCAP, 2009). Therefore, crash testing results must be viewed with some caution when it comes to predict car crashworthiness in crashes involving vehicles of different weights and sizes.
Thus, the two main objectives for this research are:
- To analyse the vehicle characteristics which are more important in the crash severity classification models based on CART methodology, for single-vehicle crashes and two-vehicle collisions.
- To develop a logit model to predict crash severity for single-vehicle crashes.

The information generated in this study can be useful for manufacturers, and to lead clear directions to policymakers about actions needed and which priorities should be set in order to improve vehicle regulations for a significant improvement on road safety level.

The paper is organized as follows: chapter 2 provides a literature review, followed by methodology in chapter 3, results and discussion in chapter 4, and finally conclusions in chapter 5.

2. LITERATURE REVIEW

In the literature, most attention is paid to vehicle type and risk to the drivers, but not to its relation to crashworthiness. There is a lack of a methodology to estimate the effect of vehicles characteristics with the crash severity during vehicles collisions.
A number of studies have attempted to correlate safety and vehicle design features. Evans (2004) explored vehicle mass and size, and concluded that those variables are strongly correlated, which makes it difficult to determine the separate contribution of mass and size on crash risk. Wood and Simms (1997) showed that in collisions between cars of similar size and in single vehicle crashes the fundamental parameters which determine the injury risk are associated to the size, i.e. the length of the vehicle. However, in collision between dissimilar sized cars the fundamental parameters are the weight and the structural energy absorption of the vehicle. Wenzel and Ross (2005) suggested the quality of cars may be more correlated to the risk than weight, but this correlation is not strong. Most of the range in risk in cars must be attributed to vehicle design and to the difficulty to quantify driver characteristics and/or behaviour. Broughton (2008) showed that the driver casualty rate decreases with the size of his car, however the driver casualty increases with the size of the other car involved in the collision. Tolouei and Titheridge (2009) showed that increasing vehicle mass generally decreases the risk of injury to the driver.
Previous studies related to crash analyses have used a broad spectrum of statistical models to reach conclusions. For example, statistical regression models are very popular for analysing contributing factors to injury severity (Li and Bai, 2008, Boufous et al, 2008,

Bedard et al, 2002, Al-Ghamdi, 2002, Chang and Wang, 2006). Kononen et al. (2011) have identified and validated a logistic regression model for predicting serious injuries associated with motor vehicle crashes with the following inputs: crash direction, change in velocity, presence at least on a older occupant and vehicle type. Martin and Lenguerrand (2008) have estimated the driver protection provided by passenger cars for the France vehicles fleet by applying a conditional Poisson regression. Mendez et al. (2010) has evaluated the crashworthiness and the aggressively of the Spanish car fleet by applying two types of regressions: logistic models in a single-crashes and generalized estimating equations models in tow-crash crashes. However, regression models have many assumptions and pre-defined underlying relationships between the dependent and independent variables (Chang and Wang, 2006). A more advanced data mining technique is the Classification and Regression Trees Analysis (CART). Meng and Weng (2012) argued that CART can provide higher prediction accuracy than the conventional binary logit model. CART methods do not require predefined causal relationship between target (dependent variable) and predictors (independent variables). Decision trees provide an excellent introduction to predictive modeling and are useful to predict new cases, select useful inputs and optimizing complexity (SAS Institute Inc., 2007). Chang and Wang (2006) have classified CART as a flexible non-parametric technique which can provide more informative and smart set of models, and its application is a valuable precursor to a more detailed logistic regression analysis in crash injury data.

For real-world crash severity prediction datasets, the target variable (related to serious injuries is predominantly imbalanced, with the majority of instances composed of non-severe crashes (majority class) and only a small percentage of severe crashes (minority class). The classification problems based on imbalanced data occur often in applications when the events of interest are rare (such as severe crashes). The costs of type I and type II errors is dramatically asymmetrical, making an invalid prediction of the minority class more costly than an accurate prediction of the majority class (Crone and Finlay, 2012). Several studies have recommended re-sampling methods to address the problems related to imbalanced classes in several domains (Cieslak and Chawla, 2008, Kotsiantis et al, 2006, He and Garcia, 2009, Japkowicz and Stephen, 2002, and, Crone and Finlay, 2012).

In summary, several studies have addressed the effect of vehicle type to examine the risk of injury levels, with the standard practice being for models to be constructed using samples of the available data. However, there is a gap on how to handle the low frequency of severe crashes. To date, the authors are not aware of any data mining severe crash events and appropriate predictive modeling approach. The strength of the study is that the effect of vehicle characteristics are considered for the decision modeling with rare events, as it happen in the field. Further research is needed to address this imbalanced data to better accident analysis accuracy to reinforce strategies for traffic safety policies and automobile industry smart challenges.

3. METHODOLOGY

One motivation for this research was to focus on the vehicle fleet characteristics and analyze which (if any) has a stronger impact on crash severity. In previous studies, most attention focused on vehicle body type, rather than vehicle specific technical characteristics (Abdel-Aty, 2003, Kononen et al, 2011, Bedard et al, 2002, Al-Ghamdi, 2002). In this study, vehicle technical characteristics, such as weight, engine size, wheelbase and registration year (age of vehicle) were analyzed with data mining methodologies in order to extract patterns from the predictors and relate them to the occurrence of injuries and fatalities in a crash. Figure 1 summarizes the methodology.

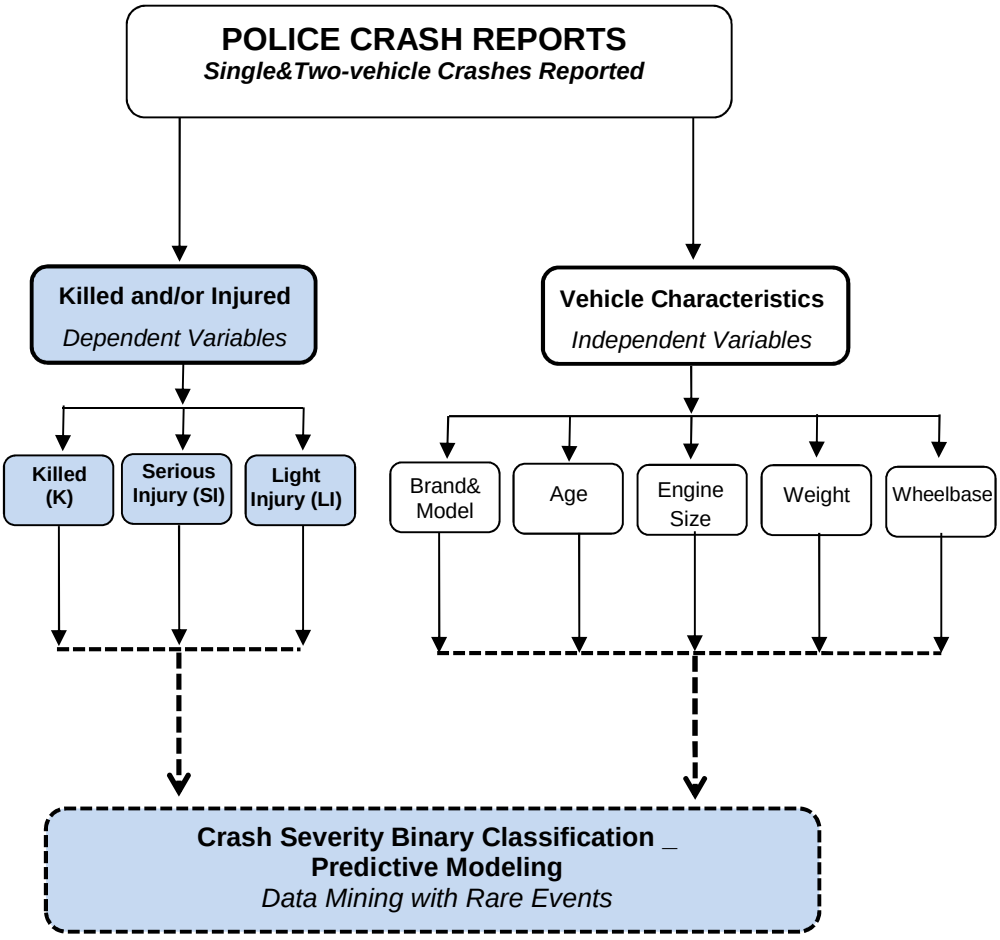

Figure 1 –Methodology overview.

This research focused exclusively on post-crash consequences rather than on pre-crash contributing factors to the event. Drivers' behavioral characteristics, such as age and gender, as well as population socio-demographic factors were out of the scope of this study. Seat belt use is one of the most important predictors of serious injuries (Abdel-Aty, 2003, and Kononen et al 2011). However information on the seat belt use, air bag data, roadway curvature and grade, even though those factors influence crash outcomes, they were not available in the crash reports.

The three most popular methods in predictive modeling are: decision trees, regressions and neural-networks. In the study presented in this paper, CART and logit regression were selected for the development of binary prediction models no analyze crash severity.

3.1 Decision Tree Model

The CART methodology was used for the following reasons. First, traditional statistics have limited utility in the task of variable selection for multiple variable comparisons. Second, predicted variables are rarely satisfactorily distributed. CART has the potential to "uncover complex interaction between predictors which may be impossible to uncover using traditional multivariate techniques" (Crone and Finlay, 2012).
To select useful inputs, trees employ a split search algorithm. The root node is divided into child nodes on the basis of an independent variable (predictors), which creates the best purity in the way that the data in the child note is more homogeneous than in the upper parents node (Kashani, 2011). This process will last until all the data in each node have the most possible homogeneity, leading to the terminal nodes, or terminal leafs. One of the most commonly criteria used in the trees split is the Gini index.
The variable relative importance (VRI) is provided by CART methodology output, and it denotes the surrogate splitting rule using that input in a way that assures the reduction in sum of squares errors (SSE) from the predicted values. The VRI (called Variable Importance Measure) is very useful select useful predictors (Kashani, 2011, and SAS Institute Inc., 2007). The variable that has the most importance to classify the target, has the largest number compared to the others predictors. It must be noted that the variables relative importance may or may not follow the order of the variables selected at the tree structure.

3.2 Logit Model

Regression offers a different approach to prediction compared to decision trees. Regressions, as parametric models, assume a specific structure between inputs (predictors) and target. By contrast, trees as predictive algorithms, do not assume any association structure, they simply isolate concentrations of cases with like-valued target measurements. The logistic regression technique provides information on the parameters estimates, their standard error and their significance.

The logistic function is simple the inverse of logit function. A logistic regression applies a logit transformation (a natural log of the odds) to the probabilities and ensures that the model generates estimated probabilities between 0 and 1. At this function, x has an unlimited range while P (Probability) is restricted to range from 0 to 1.
The logit transformation in the logistic regression model is described by the following equation:

$$logit(P) = \log\left(\frac{P}{1-P}\right) = \beta 0 + \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk} \quad (1)$$

The above equation in terms of probability it can be rewritten into:

$$P = \frac{\exp(\beta 0 \ + \ \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk})}{(1 + \exp(\beta 0 \ + \ \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk}))} \quad (2)$$

The estimated factors for the response variables are provided by the independent variables. Since in the logistic approach the expected value of the target is transformed by a link function, (the likelihood function), the maximum likelihood estimates the values of the parameters that maximize the probability of obtaining the training sample. The parameters estimates from the maximum likelihood estimation are used in the logistic equation model to predict the binary target.

3.3 Data Description

Data were collected from the road traffic Departments of the Portuguese Road Safety Police National Republican Guard (GNR) and the Portuguese Public Safety Police (PSP). Recorded crash reports involving property damage only were excluded. Crash reports that involved injuries and/or fatalities outcomes were exclusively selected. A total of 2270 reports were extracted, as indicated in Table I. Crash reports were gathered for 5-years' time period between 2006 - 2010.

Table I – Relevant Crash Frequencies Gathered in the Study

Data Source	2006	2007	2008	2009	2010	Total by Data Source
GNR Porto, PT	298	548	508	161	184	**1699**
PSP Aveiro, PT	-	65	65	-	-	**130**
PSP Porto PT	-	166	275	-	-	**441**
Total by year	**298**	**779**	**848**	**161**	**184**	**2270**

As a result of the difficulty in matching vehicles to injury outcome, it was decided to reduce the sample size from a desirable 2270 observations to a manageable 1374 crash observations for which vehicle identification number (VIN) requests were made.

Table II shows the description of each variable. This table lists all independent variables and dependent variables in the dataset. Table IIa) identifies the independent variables which were analysed to estimate its impact to the crash severity: 10 independent variables and 18 independent variables, for the single-vehicle crash and two-vehicle collisions, respectively. For the two-vehicle collisions dataset the derivate variables of the combined effect of vehicle V1 and vehicle V2 differential characteristics, are also included. For instance, in a two-vehicle collision, the weight differential between V1 and V2 was expressed by WTV2V1 (kg), which was obtained by subtracting the weight of vehicle V1 to vehicle V2. The same procedure was applied for the vehicle's engine size, wheelbase, and age, leading to the following variables: ccV2V1, WBV2V1, and AgeV2V1.
In Portugal, the injury severity of the vehicle's occupants is recorded in three levels: light Injury (LI), serious injury (SI) and fatality. Table IIb) identifies two categories for the dependent variables. The dependent variable entitled "SIK" was created to signify the sum of the number of serious injured and killed in a crash. Since this study focused on modelling the

crash severity, the response variable SIK was then converted to a binary target, FatalSIK. The response variable FatalSIK was assigned a value of 1 if SIK>0, zero otherwise.

Table II – Description of variables in the crash data set: a) independent original variables and b) dependent variables included in the crash data set

a) Independent Variables	**Description**	**Variables Labels**
Age of Vehicle 1	AgeV1 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	**AgeV1**^{*DT,*LR}
Age of Vehicle 2	AgeV2 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration	**AgeV2**^{*DT}
Age Difference between vehicles (V_2) and (V_1)	AgeV2V1 (yr) stands for age of vehicle V_2 minus the age of vehicle V_1 crash observation.	**AgeV2V1**^{*DT}
Alcohol and/or Drugs	The Driver´s test for alcohol and or drugs is presented as: Code=0, legal; Code=1, illegal	**AlcoholDrugs**^{*DT,*LG}
Crash Type for single vehicles crashes	Ran off road Rollover	**RanOff**^{*DT} **Rollover**$^{*DT, *LG}$
Crash Type for two vehicle- collisions	Rear End Head-On Sideswipe Other	**RearEnd**^{*DT} **HeadOn**^{*DT} **Sideswipe**^{*DT} **Other**^{*DT}
Divided/undivided	Existence or absence of physical median: Code=0, undivided Code=1, divided	**DivisionCode**^{*DT,*LG}
Engine Size of Vehicle 1	Engine size of vehicle (V_1) (cm^3)	**ccV1**^{*DT,*LG}
Engine Size of Vehicle 2	Engine size of vehicle (V_2) (cm^3)	**ccV2**^{*DT}
Engine Size Difference between vehicles (V_2) and (V_1)	ccV2V1 stands for engine size of vehicle V_2 minus the engine size of vehicle V_1 at crash observation, (cm^3).	**ccV2V1**^{*DT}
Road Class	Roads is based in the number of lanes and coded as follows: Code=0, two lanes Code=1, multi-lanes Code=2, motorway	RoadClass
Speed Level	The speed level was coded as follow: If Speed limit<=90km/hr, then code=0 If Speed limit>90, then code=1	**SpeedLevel**^{*DT,*LG}
Wheelbase of Vehicle 1	Wheelbase of vehicle (V_1) (mm)	**WBV1**^{*DT,*LG}
Wheelbase of Vehicle 2	Wheelbase of vehicle (V_2) (mm)	**WBV2**^{*DT}
Wheelbase Difference between vehicles (V_2) and (V_1)	WBV2V1 stands for wheelbase of vehicle V_2 minus the wheelbase of vehicle V_1 at crash observation, (mm).	**WBV2V1**^{*DT}
Weight of Vehicle 1	Weight of vehicle 1 (V1) (kg)	**WTV1**^{*DT,*LG}
Weight of Vehicle 2	Weight of vehicle 2 (V2) (kg)	**WTV2**^{*DT}
Weight Difference between vehicles (V2) and (V1)	WTV2V1 stands for weight of vehicle V2 minus the engine size of vehicle V1 at crash observation (kg).	**WTV2V1**^{*DT}
Weather Conditions	Weather conditions at the moment of the crash: Code=0, Clear and/or dry pavement Code=1, rain and/or wet pavement	**WeatherCode**^{*DT,*LG}
b) Dependent Variables	**Description**	**Variable Label**
Number of Killed (K) plus Serious Injured (SI)	SIK: sum of occupants serious injured (sum SI) + sum of occupants killed (sum K) in a crash SIK	SIK
Serious and/or Fatal SIK	FatalSIK: categorical response for a crash outcome used to predict either a serious injury, or fatality in a crash event. FatalSIK=1, if SI>0 and/or K>0, else, FatalSIK=0	**FatalSIK**^{*DT,*LG}

^{*DT}Variables in bold are used in the decision trees models for single and/or two-vehicle collisions.

*LGVariables in bold are used in the logit model for single-vehicle crashes.

3.4 Modeling Approach for Crash Severity Prediction (with Rare Events)

For real-world crash severity prediction data, the target variable is predominantly imbalanced. The percentage of severe/fatal crashes has been estimated with 2.8% distribution given any level of injury (Abdel-Aty, 2003). Hence severe crashes have lower frequency than non-severe crashes. With regard to binary data classification (severe crash vs. non severe crash), analysis of data containing rare events, poses a great challenge to the machine learning community. The answer to the question "*Why is the data containing rare events a challenge*?" is explained next. Classification models, such as decision trees and logistic models, tend to provide a severely imbalanced degree of accuracy: with the minority class having 100% accuracy, and the minority class (which represent the event of interest) having accuracies in the interval 0-10% (He and Garcia, 2009). When probabilistic statistical methods are used, such as logistic regression, they underestimate the probability of the rare events because they tend to be biased toward the majority class (non-severe crashes), which has significantly higher frequency compared to the minority class (severe crashes). On the other hand, CART tends to perform more poorly with unbalanced data, because the splitting criteria, Gini index, it is skew sensitive (SAS Inc.; 2007 and Cieslak and Chawla, 2008). This occurs because the sampling methods prior to the decision tree induction alter the class distribution driving the bias towards the majority or positive class (Cieslak and Chawla, 2008).

In this study, the target being predicted, crash severity (FatalSIK"1'), had the following distribution: 3.7% and 7.3%, for two and single, vehicles crashes respectively. Hence the dataset clearly qualifies as an imbalanced data.

Random sampling often does not provide enough targets to train a predicted model for rare events, such as, severe crash observations. Following previous work, the original Portuguese crash population was randomly stratified, (Cieslak and Chawla, 2008, Japkowicz and Stephen, 2002, He and Garcia, 2009, and Crone and Finlay, 2012). Hence, all the observations having the rare event (severe crashes) were selected, and only a randomly fraction of the non-event (non-severe crashes) were included (SAS Institute Inc.; 2007).
For the crash dataset, the training samples were stratified to the target level 50/50 proportion, (SAS Institute Inc.; 2007, Cieslak and Chawla, 2008). The bias introduced by resampling (over-representation of target "1") was corrected by adjusting the predicted probabilities with prior probabilities to correct for the original crash distribution (SAS Institute Inc.; 2007). Thus in training sample, the target level "1" (sever crash) is over-represented. The bias introduced by resampling was corrected by adjusting the predicted probabilities with prior probabilities in order to predict the original distribution of target "1" in the original crash data.
For the decision trees the adjustment of prior probabilities was performed with a decision node (SAS Institute Inc.; 2007). Table III summarizes the adjusting prior probabilities for the stratified training samples used in the trees model development. For the logistic regression model development, the bias introduced by re-sampling was corrected by adjusting the cut-off during the model training, and by taking into account the prior probabilities for the severe crashes in the data set.

Table III – Stratified Training Sample for Model Development

Data set	Stratified Levels		Prior Probabilities	Adjusted Prior
	Level	Count		
Two	1	32	0.5	0.037
	0	32	0.5	0.963
Single	1	38	0.5	0.036
	0	38	0.5	0.924

In this study, data mining was performed with the Statistical Analysis Software, SAS® v9.2 and SAS®Enterprise Miner™6.2 software (SAS Institute Inc.; 2007).

Decision trees results were discussed based on VRI (Kashani et al, 2011, SAS Institute Inc., 2007). The assessment of the logistic predictive modeling was evaluated by the event classification table, similarly to the confusion matrix used elsewhere (He and Garcia, 2009, and Kotsiantis et al, 2006). At Enterprise Miner interface, the event classification table provides the assessment score rankings for the model, based on the predicted probabilities of the observed response. This table output provides the number of observations that follow into each of the four classification categories for the target being predicted: False Negative (FN), True Negative (TN), False Negative (FN), and True Positive (TP).

Next, Table IV defines the measures classification criteria developed in this study for the assessment score of the response variable, FatalSIK, used for the evaluation of the candidate models.

Table IV– Event Classification Table for the Target FatalSIK

Target	False Negative (FN)	True Negative (TN)	False Positive (FN)	True Positive (TP)
Predicted Target	FatalSIK"0"	FatalSIK"0"	FatalSIK"1"	FatalSIK"1"
Actual Target	FatalSIK"1"	FatalSIK"0"	FatalSIK"0"	FatalSIK"1"

The accuracy rate in the training model is also equivalent to the percentage of the cases predicted right by the model within the training sample. In this study the terminology "Accuracy Rate" is applied and is defined by the following equation:

$$\%\, Accuracy\ Rate = \frac{(TP + TN)}{(FN + TN + FN + TP)} \quad (3)$$

Following the selection of the best model to predict FatalSIK"1", then the validation was performed based on a procedure adopted by Crone and Finaly, 2011. The performance of accuracy prediction of the final FatalSIK model was evaluated by comparing the model score rates for the original crash dataset with the model score for 10 stratified random samples. For

each sample subset, the observations were randomly excluded from the majority class (non-severe crashes) until to equal number of the minority class (severe crashes). Hence, the final model was evaluated 10 times by score the final model for each of those stratified samples subsets. Then the model accuracy prediction rates for each of those subsets were recorded and the average of those 10 accuracy rates was estimated.

4. RESULTS AND DISCUSSION

This chapter presents descriptive statistics for dependent and independent variables and the results for crash severity (target FatalSIK) prediction based on two modelling approaches: decisions trees and logistic regression models. The results are discussed in the following order. First, descriptive statistics for the entire crash dataset is presented. Second, decision trees models for two-vehicle collisions and single-vehicle crashes are analysed. Finally, a logit model developed to predict the crash severity response in single-vehicle crashes is discussed.

4.1 Descriptive Statistics

The descriptive statistics analysis for the crash data set which includes a total of 1374 crash observations is presented in Table V. For the entire crash population, the injury severity level was as follows: 27 killed k, 61 serious injuries (SI), and 1967 Light Injured. The table also presents the average values for the vehicle technical characteristics: weight, engine size, age and wheelbase. It is interesting to notice that for both vehicles V1 and V2 the mean age was 8.5yr. However, there is a wide range of vehicles age: 1yr to 38 yr, newest and oldest vehicles, respectively.

Table V– Descriptive statistics for selected variables in the crash dataset

Variable name in SAS*	N. Crashes	Mean	S.D.	Minimum	Maximum
SUMLI[1]	1374	1.43	0.87	0	8
SUMSI[2]	1374	0.044	0.25	0	3
SUMK[3]	1374	0.02	0.14	0	2
SIK[4]	1374	0.06	0.30	0	3
WTV1[5](Kg)	1374	1222.34	334.98	640	3200
ccV1[6](cm^3)	1374	1662.65	491.67	599	4104
WBV1[7](mm)	1374	2581.02	256.47	1625	4325
AgeV1[8](yr)	1374	8.48	5.06	1	25
WTV2[9](kg)	874	1262.85	364.46	584	3500
ccV2[10](cm^3)	874	1700.94	522.18	698	4104
WBV2[11](mm)	874	2609.00	289.88	1812	4100
AgeV2[12](yr)	874	8.54	5.26	1	38
WTV2V1[13](mm)	874	28.65	519.87	-2165	2860
ccV2V1[14](cm^3)	874	34.98	719.72	-2905	2909
WBV2V1[15](mm)	874	10.84	396.80	-2213	1918
AgeV2V1[16](yr)	874	<1	7.42	-20	28

1 Sum of LI; **2** Sum of SI; **3** Sum of K; **4** Sum of SI and K; **5** Weight of Vehicle V_1; **6** Engine size of Vehicle V1; **7** Wheelbase of Vehicle V1; **8** Age of vehicle V1; **9** Weight of Vehicle V2; **10** Engine size of Vehicle V2; **11** Wheelbase of Vehicle V2; **12** Age of vehicle V2; **13** Weight Differential between V2 and V1 at two vehicle collision; **14** Engine size differential between V2 and V1 at two vehicle collision; **15** Wheelbase Differential between V2 and V1 at two vehicle collision; **16** Age Differential between V2 and V1 at two vehicle collision.

4.2 Classification and Regression Tree Analysis

The CART methodology was applied to identify the independent variables (vehicles characteristics and crash information) which are more important to classify a crash as severe event. The overall crash severity is expressed by the binary target FatalSIK (as explained previously in section 3.3).

4.2.1 CART Analysis for FatalSIK for Two-vehicle Collisions

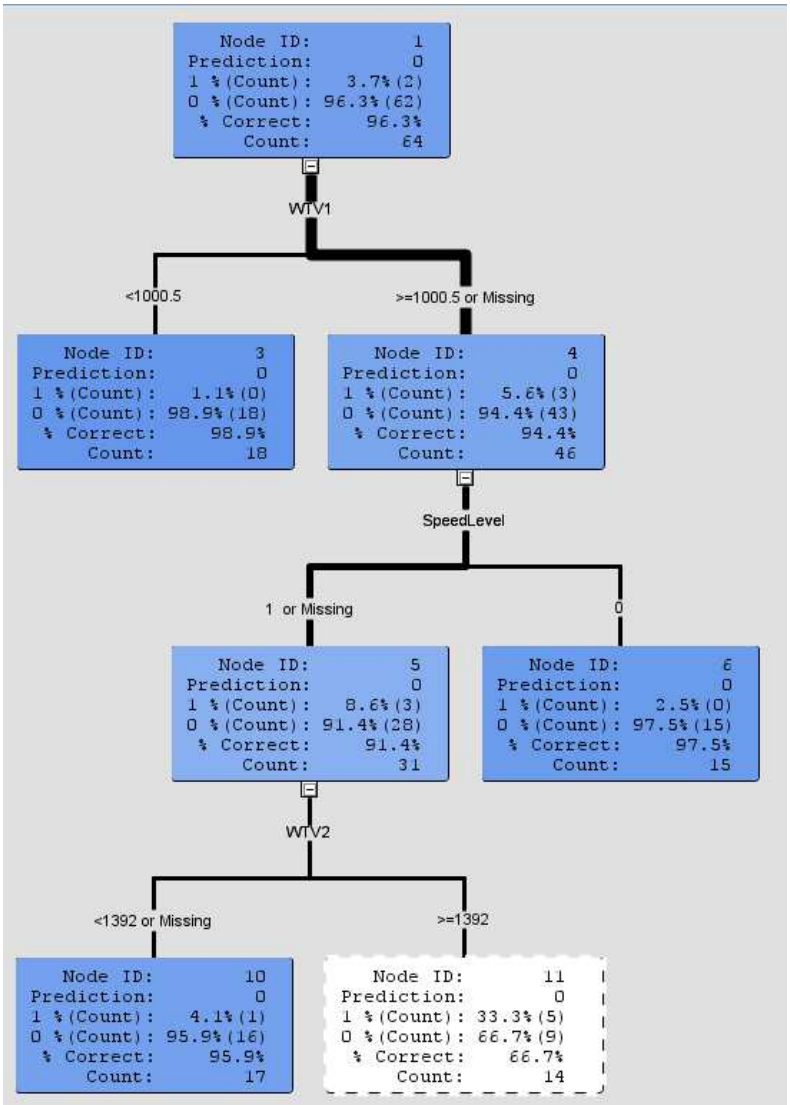

Figure 2 – Classification Tree for FatalSIK for two-vehicle crashes using a balanced sample.

The predictive decision model for 874 two-vehicle collisions is presented in Figure 2. This decision tree was developed with the balanced sample procedure with a 50/50 ratio between the target level “1” (32 counts) and target level “0” (32 counts). To eliminate the bias from the over-reseprentation of the target level “1”, prior probabilities were adjusted for the original

sample distribution, where the 2 counts represent the 3.7% of FatalSIK “1”, and 62 counts represent 96.3% of FatalSIK”0” in the original sample distribution.

In Figure 2 the selected variables were the weight of vehicle V1 (WTV1), weight of vehicle V2 (WTV2) and the speed level. In this figure, the tree variables split order does not follow exactly in the relative importance. The tree parent node was split first by the WTV1, followed by speed level and then by WTV2 level. On the other hand, those selected variables relative importance to minimize the SSE for the FatalSIK was: 1 for WTV1, 0.796 for WTV2 and 0.708 for speed level. The above decision tree shows that when the collision involved a lighter V1, WTV1<1000.5kg, 98.9% of the crashes on the tree leaf were non severe. On the other hand, heavier V1 vehicles with WTV1>=1000.5kg, were associated with a higher percentage of severe crashes (5.6%). Then, this leaf was split by the speed level, showing that higher speeds are associated with a higher percentage of severe crashes (8.6%). Finally, the tree was split by the weight of vehicle V2. Similarly to crashes involving heavier vehicles V1, when a heavier vehicle V2 was involved, WTV2>=1392kg, the expected crash severity was much higher, 33.3%. Fisher’s exact p-value< 4.539E-04 indicated that the FatalSIK cannot be considered independent from the weight of the vehicles involved in the crash neither from the speed level.

4.2.2 CART Analysis for FatalSIK for Single-vehicle Crashes

Figure 3 illustrates the decision tree structure for single vehicle crashes based on the balanced sample approach. During the resampling approach, a 50/50 ratio between the target level “1” (38 counts) and target being level “0” (38 counts) was used. As explained for the previous trees, the CART output results were corrected to reflect the original sample distribution, where 6 counts represented the 7.6% of FatalSIK “1”, and 70 counts represented 92.4% for FatalSIK”0” for the original target level distribution in the single vehicle crashes dataset.

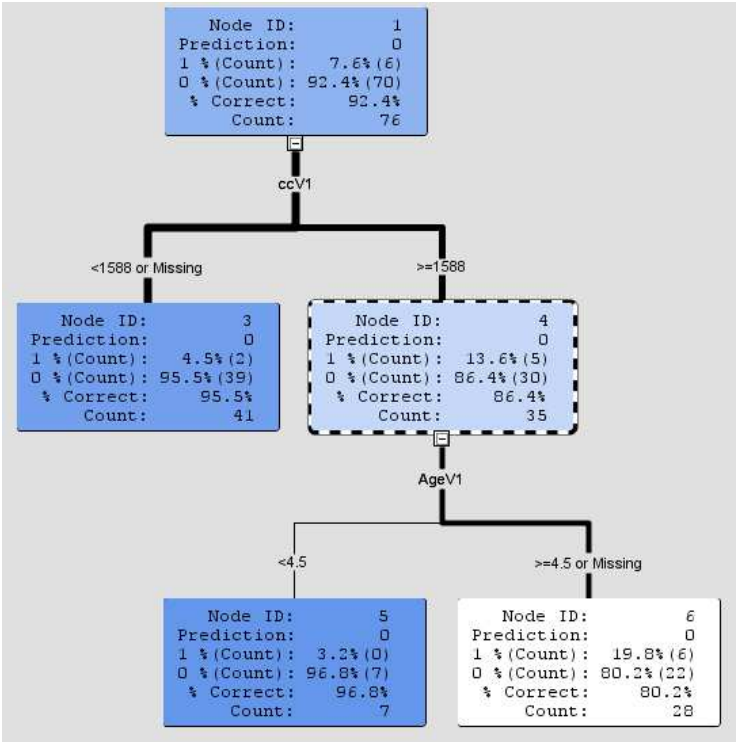

Figure 3 – Classification Tree for FatalSIK for single-vehicle crashes using a balanced sample.

In Figure 3, the tree growth followed the order of the variables relative importance. The selected variables were: engine size (ccV1) and age of the vehicle (AgeV1). For the tree development, ccV1 had a relative importance of 1; meanwhile Age V1 had a relative importance of 0.868. Hence, for lower vehicle engine size, ccV1<1588 cm3, the expected overall crash severity was lower, 4.5%. On the other hand, for vehicles with a higher engine, ccV1>=1588 cm3, 13.6% of the crashes were severe. Following the engine size, the tree was split by the vehicles' age. For crashes with new vehicles, AgeV1<4.5yr, the crash severity was lower, 3.2%. However, when the vehicle was older, AgeV1>=4.5yr, the overall crash severity was 6 times higher than for new vehicles, (19.8% vs. 3.2%). This finding suggests that the lack of safety devices in the other vehicles, such as air-bags may increase the risk of a serious injury for vehicles' occupants. Fisher's exact p-value< 0.0164 suggest that in single vehicle crashes the engine size and age of the vehicle are significant under the caveats previously mentioned.

4.3 Logistic Regression Analysis for Single Vehicle Crashes

The logit model developed to predict FatalSIk has four independent variables: AgeV1, WBV1, ccV1 and WeatherCode. Table VI shows with the exception of the wheelbase of the vehicle, (PrChiSq<0.0593), all the predictors in the model were found to be statistically significant at the 0.05 level (PrChiSq<0.0144, PrChiSq<0.0418, and PrChiSq<0.0031, for

AgeV1, WeatherCode and ccV1, respectively). An increase of the vehicle engine size and vehicle age, as well as good weather conditions are associated with a higher risk of a severe crash, FatalSIk"1". On the other hand, as the vehicle wheelbase increases there is a decrease in the probability of a FatalSIK"1".

Table VI– Fit statistics for FatalSIK model for single-vehicle crashes

Fit Statistics											
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates**								**ASE**	**MISC**
DF	Pr>ChSq	**Parameter**			DF	Estimate	Pr>ChSq				
4	0.0004	Intercept			1	5.1730	0.3023			0.187	0.237
		AgeV1			1	0.1519	0.0144				
		WBV1			1	-0.0045	0.0593				
		WeatherCode (0)			1	0.6879	0.0418				
		ccV1			1	0.0030	0.0031				
Odds Ratio Estimates											
		Effect			Point Estimate						
		AgeV1			1.164						
		WBV1			0.996						
		WeatherCode 0 vs 1			3.958						
		ccV1			1.003						
Accuracy Performance											
Accuracy Rate with Training Sample (N=76)					Accuracy Rate with Original Population (N=500)					Accuracy Performance with 10 Stratified Random Samples	
FN[1]	TN[2]	FP[3]	TP[4]	% AR[5]	TPs[6]	FPs[7]	TNs[8]	FNs[9]	%AR[10]	%A.D[11]	S.D.[12]
10	30	8	28	76.3	17	97	365	21	76.4	14.4	2.4

1 False Negative; 2 True Negative; 3 False Positive ; 4 True Positive; 5 Percentage of Accuracy Rate; 6 True Positives; 7 False Positives; 8 True Negavtives1; 9 False Negatives; 10 Percentage of Accuracy Rate; 11 Average Differential of the Accuracy Rate for the 10 stratified with the Model Accuracy Rate for the score original crash population.; 12 Standard Deviation.

The equation developed for FatalSIK model for single-vehicle crashes is presented next.

$$P\,(FatalSIK = 1) = \frac{\exp(5.1730 + 0.1519 * AgeV1 - 0.0045 * WBV1 + 0.6879 * WeatherCode(=0) + 0.0031 * ccV1)}{1 + \exp(5.1730 + 0.1519 * AgeV1 - 0.0045 * WBV1 + 0.6879 * WeatherCode(=0) + 0.0031 * ccV1)} \quad (4)$$

Where: AgeV1 is the age of the vehicle, WBV1 is the wheelbase of the vehicle, WeatherCode(=0) denotes good weather conditions, and ccV1 is the vehicle's engine size.
In Table VI the odds of a FatalSIK crash in good weather conditions is almost 4 times the odds under poor weather conditions (wet road). As far as the model performance with the training sample, the above model performed very well with a 76.3% AR. In addition, from the 38 severe crashes in the training dataset the model correctly predicted 28 cases. In addition, model FatalSIK shows a very good AR, 76.4%, when predicting new cases in the full population of crashes. Graphical assessment for the FatalSIK model is illustrated in Figures 4 to 6.

In Figure 4, the FatalSIK probability is illustrated a function of the age of the vehicle, while vehicle wheelbase and the engine size were fixed (at 2551mm and 1602 cm3, mean wheelbase and engine size, respectively). A similar approach was used for the plots in Figures 5 and Figure 6.

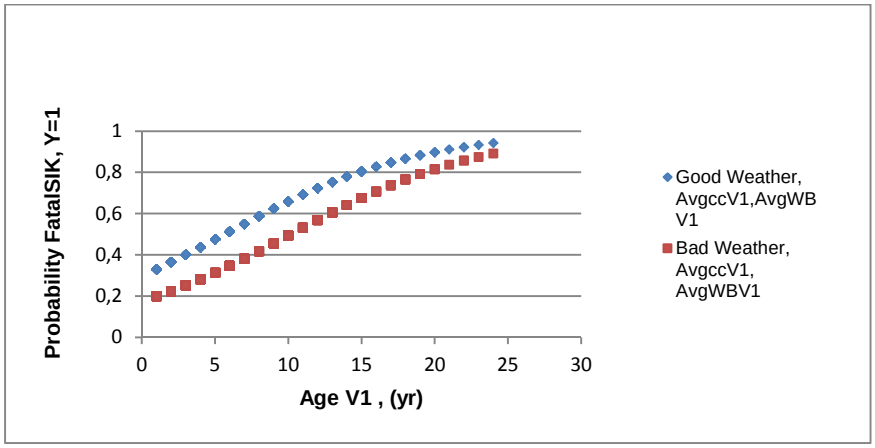

Figure 4 – Probability of a Serious Injury and/or fatality as function of the age of the vehicle, in single-vehicle crashes, using the FatalSIK model.

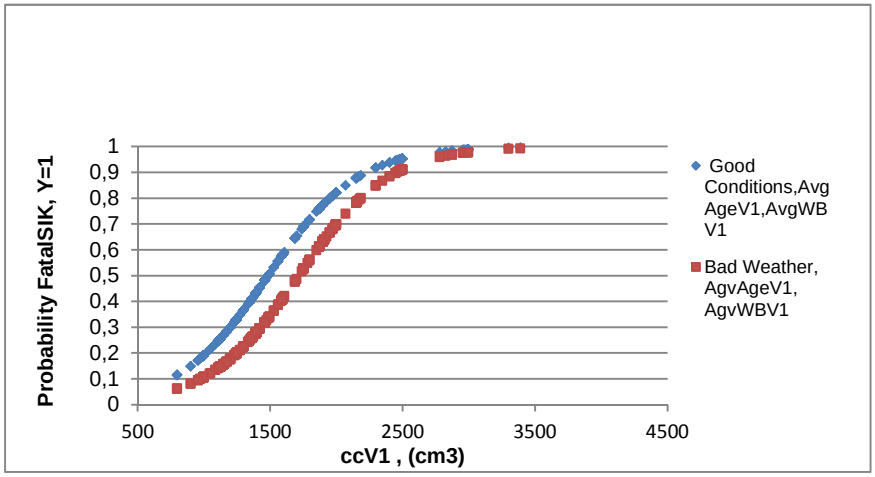

Figure 5 – Probability of a Serious Injury and/or Fatality based in vehicles' engine size, in single-vehicle crashes, using FatalSIK model.

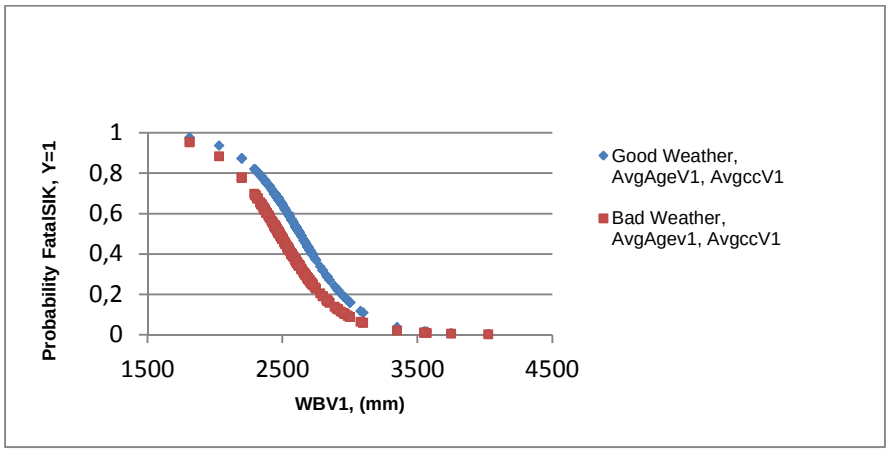

Figure 6 – Probability of a Serious Injury and/or fatality based in vehicles' wheelbase, in single-vehicle crashes, using FatalSIK model.

Figure 4 shows that as the age of the vehicle increases, the probability of a FatalSIK also increases. This trend was expected since newer vehicles have improvements in crashworthiness due to the incorporation of new safety features, such as air bags and

electronic stability control. Also, the logistic curve for the good weather conditions show a higher probability of a FatalSIK than for poor weather conditions, as previously indicated.

Figure 5 shows that as the engine size of the vehicle increases, the probability of a FatalSIK also increases. The effect of the engine size may be interacting with travel speed. It is possible that drivers in cars with a larger engine size (higher power) tend to accelerate more and travel higher speed. Hence, when the crash occurs, the changes in the vehicle velocity associated with the crash event (delta-V), which has been reported as the most significant predictor of a serious injury, (Kononen et al., 2011).

Figure 6 shows that as the wheelbase size of the vehicle decreases, the probability of a FatalSIK also increases. The size of vehicle's wheelbase in the decreasing risk of a serious and or fatal crash may be interpreted by the fact that one of the vehicles attribute most related to the injury severity level is vehicle size, (Evans, 2004, Bedard, et al., 2002). A larger vehicle size, offers a higher area for the energy dissipation following the crash impact force, hence reducing the energy change that the body of the vehicles occupants may be exposed, thus reducing the risk. This finding is consistent with previous work, (Bedard et al., 2002).

5. CONCLUSION

The presented CART and logistic regression results have identified the most important variables contributing to crash severity, expressed by the binary target FatalSIK. The following conclusions can be drawn:

- CART results showed that in two-vehicle collisions the weight of the vehicles involved is the most important factor associated with overall crash severity. A collision of one vehicle heavier than 1000kg involving a second vehicle heavier than 1392kg, had the highest expected crash severity (33.3%). This finding is consistent with previous work (Wood and Simms, 1997, and Tolouei and Titheridge, 2009). CART results for single-vehicle crashes shows that the engine size and the age of the vehicle are associated with a higher risk of a severe crash.
- The developed logit model for crash severity prediction in single-vehicle crashes shows that the engine size, vehicle age and wheelbase, along with good weather conditions; impact the crash severity outcomes for the target FatalSIK. As engine size and vehicle age increase, the risk of a severe crash also increases. Also, the probability of a severe crash decreases as the wheelbase of the vehicle increases, and the risk is lower in poor weather conditions.

It should be noted from the above findings that both CART and logistics regression methodologies have identically identified the explanatory variables that are key predicting a crash severity outcome namely vehicle engine size and vehicle age. This gives more confidence in the robustness of both methods.

In the next research phase, the authors will develop a logit model to predict crash severity in two-vehicle collisions. In terms of future work, the approach developed in this study to predict

crash severity in imbalanced dataset would be useful to apply with other types of motor vehicle crashes.

ACKNOWLEDGMENTS

This work was partially funded by FEDER Funds through the Operational Program "Factores de Competitividade – COMPETE" and by National Funds through FCT – Fundação para a Ciência e Tecnologia within the project PTDC/SEN-TRA/113499/2009, and FLAD – Luso American Foundation (for the Project 91-03/2010, within the program FLAD/NSF – "Project-USA: Networks and Partnerships for Research"). G. A. Torrão also acknowledges the support of FCT for the Scholarship (SFRH/BD/4135/2007). TEMA authors also acknowledge the Strategic Project PEst-C/EME/UI0481/2011. In addition, the authors would like to thank the collaboration and expertise from Dr. David Dickey from NCSU, USA, and Eng. Elson Filho from SAS Portugal. The collaboration between Drs. Coelho and Rouphail was under the auspices of the Luso-American Transportation Impacts Study Group (LATIS-G). Any opinions, findings, conclusions or recommendations expressed in this document are those of the author(s) and do not necessarily reflect the views of FCT, FLAD, or LATIS-G.

REFERENCES

Al-Ghamdi, Al. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Anal Prev. 34 (6), 729-741.

Bedard, M., Guyatt, G.H., Stones, M.J. and Hirdes, J.P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Anal Prev. 34 (6), 717-727.

Boufous, S., Finch, C., Hayen, A. and Williamson, A. (2008). The impact of environmental, vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash. J. Safety Res. 39 (1), 65-72.

Broughton, J. (2008). Car driver casualty rates in Great Britain by type of car. Accident Anal Prev. 40 (4), 1543-1552.

Chang, L. and Wang, H. (2006). Analysis of traffic severity: an application of non-parametric classification tree techniques. Accident Anal Prev. 38 (5), 1019-1027.

Cieslak D. and Chawla N.(2008). Learning Decision Trees for Unbalanced Data, in Machine Learning and Knowledge Discovery in Databases, Part I, Proceedings, W. Daelemans, B. Goethals, and K. Morik, Editors. Springer-Verlag Berlin: Berlin. 241-256.

Coelho, M., J. Andrade, D. Soares, C, Frey, and N. Rouphail. (2010). A Vehicle Energy Use and Safety Information Support System, Presented at 89th Annual Meeting of the Transportation Research Board, Washington, D.C..

Crone S.. and Finlay S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. Int. J Forecasting. 28(1), 224-238.

Das, A., Abdel-Aty, M. and Pande, A. (2009). Using conditional inference forest to identify the factors affecting crash severity on arterial corridors. J. Safety Res. 40 (4), 317-327.

Euro NCAP (2009). Euro NCAP - For safer cars | comparable Cars. European New Car Assessment Program. http://www.euroncap.com/Content-Web-Page/0f3bec79-828b-4e0c-8030-9fa8314ff342/comparable-cars.aspx

Evans, L. (2004). How to make a car lighter and safer. Vehicle mass and size in Traffic Safety. SAE International.

He H. and Garcia E. (2009) Learning from Imbalanced Data. Knowledge and Data Engineering, IEEE Transactions. 21(9), 1263-1284.

Hermans, E., Brijs, T., Wets, G. and Vanhoof, K. (2009). Benchmarking road safety: lessons to learn from data envelopment analysis. Accident Anal Prev. 41 (1), 174-182.

Japkowicz N. and Stephen S.(2002). The class imbalance problem: A systematic study. Intell. Data Anal. 6(5), 429-449.

Kashani, A., Shariat-Mohaymany A. and Ranjbari A. (2011). A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation. 23(1),11-17.

Kononen D., Flannagan C.A.C. and Wang S.C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Anal Prev. 43(1), 112-122.

Koppel, S., Charlton, J., Fildes, B. and Fitzharris, M. (2008). How important is vehicle safety in the new vehicle purchase process? Accident Anal Prev. 40 (3), 994-1004.

Kotsiantis S., Kanellopoulos K. and Pintelas P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 30.

Li, Y. and Bai, Y. (2008). Development of crash-severity-index models for the measurement of work zone risk levels. Accident Anal Prev. 40 (5), 1724-1731.

Meng, Q. and Weng, J. (2012). Classification and Regression Tree Approach for Predicting Drivers' Merging Behavior in Short-Term Work Zone Merging Areas. Journal of Transportation Engineering. 138 (8), 1062-1070.

Portuguese National Authority for Road Safety (ANSR). (2012). Annual Report for the Year 2011. Road Casualties. http://www.ansr.pt/LinkClick.aspx?fileticket=vvB8NPsUJ%2fw%3d&tabid=344&mid=1114&language=pt-PT, Accessed 25 Jul., 2012.

SAS Institute Inc. (2007). Applied Analytics Using SAS9.2®Enterprise MinerTM7.1. Instructor-based training. Cary, NC, USA. ISBN 978-1-59994-515-6.

Tolouei, R. and Titheridge, H. (2009). Vehicle mass as a determinant of fuel consumption and secondary safety performance. Transport Res D, 14 (6), 385-399.

Wenzel, T. and Ross, M. (2005). The effects of vehicle model and driver behavior on risk. Accid Anal Prev. 37 (3), 479-494.

Wood, D. and Simms, C. (1997). Safety and the car size effect: a fundamental explanation. Accid Anal Prev. 29, 139-151.

World Health Organization (WHO), (2009a). Global Status Report on Road Safety. Time for Action. http://whqlibdoc.who.int/publications/2009/9789241563840_eng.pdf

World Health Organization (WHO), (2009b). European Status Report on Road Safety. Towards safer roads and healthier transport choices. http://www.euro.who.int/document/e92789.pdf.